Springer Proceedings in Mathematics 2

Maurício Matos Peixoto
Alberto Adrego Pinto
David A. Rand *Editors*

Dynamics, Games and Science II

Springer

Springer Proceedings in Mathematics

Volume 2

For further volumes:
http://www.springer.com/series/8806

Springer Proceedings in Mathematics

The book series will feature volumes of selected contributions from workshops and conferences in all areas of current research activity in mathematics. Besides an overall evaluation, at the hands of the publisher, of the interest, scientific quality, and timeliness of each proposal, every individual contribution will be refereed to standards comparable to those of leading mathematics journals. It is hoped that this series will thus propose to the research community well-edited and authoritative reports on newest developments in the most interesting and promising areas of mathematical research today.

Mauricio Matos Peixoto • Alberto Adrego Pinto
David A. Rand
Editors

Dynamics, Games and Science II

DYNA 2008, in Honor of Mauricio Peixoto and David Rand, University of Minho, Braga, Portugal, September 8-12, 2008

Springer

Editors

Mauricio Matos Peixoto
Instituto de Matemática
Pura e Aplicada (IMPA)
Estrada Dona Castorina 110
22460-320 Rio de Janeiro
Brazil
peixoto@impa.br

Alberto Adrego Pinto
Universidade do Porto
Departamento de Matemática
Faculdade de Ciências
Rua do Campo Alegre 687
4169-007 Porto
Portugal
aapinto@fc.up.pt

David A. Rand
University of Warwick
Warwick Systems Biology
Coventry House
Coventry CV4 7AL
United Kingdom
d.a.rand@warwick.ac.uk

ISSN 2190-5614
ISBN 978-3-642-14787-6 e-ISBN 978-3-642-14788-3
DOI 10.1007/978-3-642-14788-3
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011925151

Mathematical Subject Classification (2010): 34; 37; 49; 60; 83; 91; 92; 93; 94

Cover design: deblik, Berlin

Printed on acid-free paper

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

To my wife Alcilea Augusto and my four
children and eight grandchildren:
Marta: Daniel, Thomas and Mariana;
Ricardo: Gabriel;
Marcos: Andre and Livia; and
Elisa: Clara and Bruno.

To António Pereira Pinto

To Barbel Finkenstädt and the Rand Kids:
Ben,
Tamsin,
Rupert and
Charlotte.

Mauricio Peixoto

Alberto Pinto has asked me to write about Mauricio Peixoto in this book that honors him as well as David Rand. I am happy to do so. Mauricio is among my oldest friends in mathematics, having met him more than fifty years ago. Moreover he was instrumental in my entry into the field of dynamical systems. So important is this part of my life that my collected works contain four articles that bear on Mauricio in one way or another. That is fortunate since I wrote that material when events were fresher in my mind than they are now. Thus I will borrow freely from these references.

A most important period in my relationship with Mauricio is the summer of 1958 to June of 1960. This is discussed in an article titled "On how I got started in dynamical systems" appearing in the "Mathematics of Time", based on a talk given at a Berkeley seminar circa 1976. There I wrote how I met Mauricio in the summer of 1958 through a mutual friend, Elon Lima, who was a student from Brazil finishing his PhD at Chicago in topology. Through Lefshetz, Peixoto had become interested in structural stability and he explained to me that subject and described his own work in that area. I became immediately enthusiastic, and started making some early conjectures on how to pass from two to higher dimension. Shortly thereafter, Peixoto and Lima invited me and Clara to Rio for a visit to IMPA, or Instituto de Matemática, Pura e Aplicada.

It was during the next six months (January through June, 1960) that I did some of my most well known work, firstly the introduction of the horseshoe dynamical system and its consequences and secondly the proof of Poincare's conjecture in dimensions five or more. I sometimes described these works as having been done on the beaches of Rio; this part of the story is told in two articles in the Mathematics Intelligencer in the 1980's.

Thus we may see here what a big influence Mauricio had on my career. Another impact was his "sending" me a student to write a PhD thesis at Berkeley. That student in fact finished such a thesis and went on to become a world leader in dynamical systems. Jacob Palis' contributions in science go well beyond that. He is a main figure in developing third world science, and mathematics in Brazil in particular.

In the article "What is Global Analysis", based on a talk I gave before the Mathematical Association of America, 1968, I gave a focus to one result as an excellent theorem in global analysis. That result was Peixoto's theorem that structurally stable differential equations on a two dimensional manifold form an open and dense set. Another example of the influence of Mauricio!

I will end on a final note that reinforces all that I have said here. Over the last fifty years I have made fifteen visits to IMPA, the institute founded by Mauricio Peixoto (and Leopoldo Nachbin).

Steve Smale

Alberto Adrego Pinto

I met Alberto a few years ago, in the office of Mauricio Peixoto at IMPA, the Brazilian Institute of Pure and Applied Mathematics. Alberto was on a summer visit, and he wanted to discuss results he had obtained with his former student Diogo Pinheiro on the focal decomposition proposed years earlier by Mauricio.

By sheer accident, I had come across an application of the focal decomposition in finite temperature quantum mechanics. In fact, semi-classical approximations to the problem practically forced one to make use of the focal decomposition, although it was only much later that I became aware of its existence. That was why I was part of the meeting: my mathematician friends were curious about possible applications, and we were eager to collaborate.

Alberto immediately impressed me by his enthusiasm, his genuine interest in science, and by his easy-going style, much appreciated by a "carioca" like myself. Besides, our discussions were lively, and touched upon various conceptual points that seemed quite natural to a physicist, and eventually proved very useful from a mathematical point of view. Our collaboration has been going on ever since, and has already led to a couple of articles.

Alberto has also offered us all with a wonderful event back in 2008, when he organized a conference in honor of David Rand and Mauricio Peixoto in the precious city of Braga. The conference made me appreciate, even more, the versatility and scientific depth of Alberto, as he and his PhD students and postdocs presented seminars that covered a wide variety of subjects.

As a final word about Alberto, it must be said that he is a marvelous host. He showed us the finest of the region of Minho, using a well balanced combination of science, art, good food, good wine, and above all, good humor. That is the reason I always look forward to our next meeting: whether in Brazil or in Portugal, I am sure we will have a pleasant and productive time.

Carlos Alberto Aragão de Carvalho

David Rand

David Rand has had a world-leading influence in dynamical systems theory, in transferring dynamical systems ideas into the sciences, particularly physical and life sciences but also economics, and in developing relevant new mathematics for these areas. Highlights are his theories of the two-frequency route to chaos, invasion exponents in evolutionary dynamics, and robustness of circadian rhythms. He is widely appreciated for his leadership and for his highly pertinent and generous insights into research projects of others.

He was one of the first to bring ideas on dynamical systems with symmetry into fluid mechanics, predicting modulated wave states in circular Couette flow, subsequently confirmed experimentally by Swinney and Gorman.

A major advance was his proposal of a renormalization explanation for observations of asymptotic self-similarity in the transition from quasiperiodic to chaotic dynamics for circle maps. He extended the theory to dissipative annulus maps, providing a complete picture of the breakup of invariant circles in this scenario. Similar analysis of his has been important in understanding the spectrum of quasiperiodic Schrodinger operators.

He put the theory of multifractal scaling for chaotic attractors on a firm footing, including theory for the distribution of Lyapunov exponents.

He contributed significantly to the dynamical theory of evolutionary stability and co-evolution, including the fundamental concept of invasion exponents. He developed pair approximations for spatial ecologies and epidemics, which are now widely used.

With Alberto Pinto, he developed an extensive theory of the smooth conjugacy classes of hyperbolic dynamics in one and two dimensions, surveyed in a recent Springer Monograph in Mathematics.

He made one of the earliest analyses of nonlinear dynamics in an economics context, showing that a duopoly game has chaotic trajectories. Game theory has been a recurrent interest of his, particularly in the contexts of ecology and evolution.

Much of his recent work falls under “systems biology”. He has proposed a theory of the immune system, based on large deviation theory. He has developed theory of the robustness of circadian rhythms, which has generated much interest with experimental collaborators. The work is part of a larger project to develop mathematical tools to aid in the understanding of biological regulatory and signaling networks.

He has played a leading role in establishing Nonlinear Dynamics in the UK, co-founding the Nonlinear Systems Laboratory in Warwick and the journal Nonlinearity. He is doing the same now for Systems Biology, creating the Warwick Systems Biology Centre.

He exudes energy and enthusiasm. So it was a pleasure for me when he attracted me to Warwick. We had great fun setting up and running the Nonlinear Systems Laboratory, building up the applied side to Warwick's Mathematics department and its curriculum, and setting up the Mathematical Interdisciplinary Research Program, which he rebranded as Mathematical Interdisciplinary Research at Warwick and of which I took over directorship from David in 2000. He is a great friend and I have greatly appreciated his insightful comments, suggestions and support for my own work.

Robert S. MacKay

Preface

A couple of years ago Alberto Pinto informed me that he was planning to organize an international conference on dynamical systems and game theory in honor of Mauricio Peixoto and David Rand. I told him that I wholeheartedly support the idea and will ask the International Society of Difference Equations (ISDE) to support the proposed conference which it did later. Through my frequent visits to Portugal, I became aware of the significant contributions in dynamical systems and game theory made by Portuguese mathematicians and have subsequently been involved in fruitful discussions or joint research with a number of them. The growth of dynamical systems and game theory research in Portugal has placed Portuguese mathematicians at the forefront of these emerging fields, bringing worldwide recognition to their contributions. Indeed, in addition to DYNA2008, Portuguese researchers organized two of the last three International conferences on difference equations and applications (ICDEA), which included important talks on dynamical systems and game theory.

The work in this area has unveiled beautiful and deep mathematical theories that capture universal characteristics observed in many apparently unrelated natural phenomena and complex social behavior. Mauricio Peixoto has made lasting contributions in classifying and understanding a variety of behaviors of dynamical systems. Today these problems are the main research focus in diverse yet complementary areas at distinguished research institutions like IMPA, the institute founded by Mauricio Peixoto and Leopoldo Nachbin, and University of Warwick. Alberto Pinto has made notable contributions through his studies on rigidity properties of infinitely renormalizable dynamical systems. In addition, he discovered stochastic universalities in complex natural and social phenomena, e.g. rivers, sunspots and stock market indices, and is developing a theory with Peixoto in semi-classics physics using Peixoto's focal decomposition. David Rand, a world-leading authority, has contributed deeply and broadly to this area by developing theoretical aspects of these two fields, and identifying properties of infinitely renormalizable, universal and chaotic phenomena throughout the sciences - especially in biology, economics and physics. In collaboration with Alberto Pinto, he constructed a fine classification of dynamical systems. Moreover, the research groups led by David Rand and Alberto Pinto have, independently, developed new schools of inquiry using game theoretical and dynamic models applied to biology, economics, finances, psychology and sociology.

The research and survey papers in these volumes, written by leading researchers in their scientific areas, focus on these and many other relevant aspects of dynamical systems, game theory and their applications to science and engineering. The papers in these volumes are based on talks given at the International Conference DYNA2008, in honor of Mauricio Peixoto and David Rand. This conference, held at the University of Minho, was organized by Alberto Pinto and his colleagues and brought together influential researchers from around the world. It is worthwhile to note the warmth and hospitality of the organizers who made sure we enjoyed the beautiful region of Minho with its rich culture and fine cuisine.

Saber Elaydi

Acknowledgments

We thank all the authors for their contributions to this book and the anonymous referees. We are grateful to Steve Smale, Carlos Aragão and Robert MacKay for contributing the forwards for the book. We thank Saber Elaydi, President of the International Society of Difference Equations (ISDE), for contributing the preface for the book and for supporting and dedicating two special issues of the Journal of Difference Equations and Applications to DYNA2008, in honor of Mauricio Peixoto and David Rand, held at University of Minho, Portugal. We thank the Executive Editor for Mathematics, Computational Science and Engineering, Martin Peters, at Springer-Verlag, for his invaluable suggestions and advice throughout this project. We thank João Paulo Almeida, Isabel Duarte, Fernanda Ferreira, Flávio Ferreira, Helena Ferreira, Miguel Ferreira, Patrícia Gonçalves, Rui Gonçalves, Susan Jenkins, José Tenreiro Machado, José Martins, Filipe Mena, Abdelrahim Mousa, Bruno Oliveira, Telmo Parreira, Diogo Pinheiro, Ana Jacinta Soares and Nico Stollenwerk for their invaluable help in assembling these volumes and, especially, to João Paulo Almeida, Helena Ferreira, Miguel Ferreira, Susan Jenkins, and Abdelrahim Mousa for editorial assistance.

At DYNA2008, Mauricio Peixoto and David Rand were honored with the medal of Universidade do Minho, given by the Vice-Chancellor of Universidade do Minho, Leandro Almeida, and with a commemorative plaque given by the Vice-President of Instituto de Matemática Pura e Aplicada, Claudio Landim. We thank the members of the Scientific Committee of DYNA2008 for their invaluable help: Aragão de Carvalho (FRJ, Brazil), Edson de Faria (USP, Brazil), Pablo Ferrari (IME USP, Brazil), Livio Flaminio (U. Lille, France), Claudio Landim (IMPA, Brazil), Marco Martens (SUNYSB, USA), Welington de Melo (IMPA, Brazil), Mauricio Peixoto (IMPA, Brazil), Alberto Adrego Pinto (U. Minho, Portugal), Enrique Pujals (IMPA, Brazil), David Rand (Warwick, UK), Marcelo Viana (IMPA, Brazil) and Athanasios Yannacopoulos (Athens, Greece). We gratefully acknowledge the members of the Organizing Committee of DYNA2008 for their help: Isabel Caiado, Salvatore Cosentino, Wolfram Erlhagen, Miguel Ferreira, Patrícia Gonçalves, Filipe Mena, Alberto Adrego Pinto, Ana Jacinta Soares and Joana Torres.

We thank the Chancellor of Universidade do Minho António Guimarães Rodrigues, the Vice-Chancellors Leandro Almeida and Manuel Mota, the President of Escola de Ciências Graciete Dias, and the President of Escola de Enfermagem

Beatriz Araújo for supporting DYNA2008. We thank Fundação para a Ciência e Tecnologia, Fundação Calouste Gulbenkian, Universidade do Minho, Universidade do Porto, Departamento de Matemática da Escola de Ciências da Universidade do Minho, Departamento de Matemática da Faculdade de Ciências da Universidade do Porto, Centro de Matemática da Universidade do Minho, Laboratório de Inteligência Artificial e Apoio à Decisão and Laboratório Associado Instituto de Engenharia de Sistemas e Computadores for their support.

Contents

1 **Network Control Analysis for Time-Dependent Dynamical States** ... 1
D.A. Rand

2 **Renormalization and Focal Decomposition** ... 25
Carlos A.A. de Carvalho, Mauricio M. Peixoto, Diogo Pinheiro, and Alberto A. Pinto

3 **Micro-Foundations of the Social Change** ... 41
Elvio Accinelli and Leobardo Plata

4 **Singularities, Walrasian Economies and Economic Crisis** ... 55
Elvio Accinelli and Martín Puchet

5 **Probabilistic Methods in Dynamical Analysis: Populations Growths Associated to Models Beta$(\mathbf{p},\mathbf{q})$ with Allee Effect** ... 79
Sandra M. Aleixo, J. Leonel Rocha, and Dinis D. Pestana

6 **Power Indices Applied to Portuguese Parliament** ... 97
José M. Alonso-Meijide, Flávio Ferreira, Mikel Álvarez-Mozos, and Alberto A. Pinto

7 **A Methodological Contribution in the Theory of the Firm Under Uncertainty** ... 107
Alberto A. Álvarez-López and Inmaculada Rodríguez-Puerta

8 **Explosion of Smoothness for Conjugacies Between Unimodal Maps** ... 115
José F. Alves, Vilton Pinheiro, and Alberto A. Pinto

9 **Multidimensional Rovella-Like Attractors** ... 121
Vítor Araújo

10 Robust Heteroclinic Behaviour, Synchronization, and Ratcheting of Coupled Oscillators 125
Peter Ashwin and Özkan Karabacak

11 An Economical Model For Dumping by Dumping in a Cournot Model 141
Nilanjan Banik, Fernanda A. Ferreira, J. Martins, and Alberto A. Pinto

12 Fractional Control of Dynamic Systems 155
Ramiro S. Barbosa and J.A. Tenreiro Machado

13 A Dynamical Point of View of Quantum Information: Discrete Wigner Measures 161
A.T. Baraviera, C.F. Lardizabal, A.O. Lopes, and M. Terra Cunha

14 Dynamics on Spectral Solutions of Forced Burgers Equation 187
Mário Basto, Viriato Semiao, and Francisco Lage Calheiros

15 Area-Preserving Diffeomorphisms from the C^1 Standpoint 191
Mário Bessa

16 A Theoretical, Multidisciplinary View of Catastrophic Regime Change 223
Juan Gabriel Brida, Audrey L. Mayer, Christopher McCord, and Lionello F. Punzo

17 General Relativistic Elasticity: Statics and Dynamics of Spherically Symmetric Metrics 237
Irene Brito and E.G.L.R. Vaz

18 Post-Inflationary Scalar Field Phase Dynamics 243
T. Charters, A. Nunes, and J.P. Mimoso

19 An Application of the SIR Model to the Evolution of Epidemics in Portugal 247
António M. Correia, Filipe C. Mena, and Ana J. Soares

20 Poissonian Tree Constructed from Independent Poisson Point Processes 251
Iesus Carvalho Diniz and José Carlos Simon de Miranda

21 Hamiltonian Systems on Polyhedra 257
Pedro Duarte

22 Bankruptcy Boundaries Determined by Patents 275
M. Ferreira, B.M.P.M. Oliveira, and Alberto A. Pinto

23 Computation of Genus and Braid Index for Renormalizable Lorenz Links 301
Nuno Franco and Luís Silva

24 Statistical Stability for Equilibrium States 317
Jorge Milhazes Freitas and Mike Todd

25 Dynamic Games of Network Effects 323
Filomena Garcia and Joana Resende

26 Exit Times and Persistence of Solitons for a Stochastic Korteweg–de Vries Equation 343
Eric Gautier

27 Optimization Approach to a Simulation Algorithm of a Mixer-Settler System in the Transient State 349
E.F. Gomes and G.A. Pinto

28 Duality Theory, Representation Formulas and Uniqueness Results for Viscosity Solutions of Hamilton–Jacobi Equations 361
Diogo A. Gomes and Enrico Valdinoci

29 Microscopic Dynamics for the Porous Medium Equation 387
Patrícia Gonçalves

30 A Stochastic Model for Wolf's Sunspot Number 393
Rui Gonçalves and Alberto A. Pinto

31 On Topological Classification of Morse–Smale Diffeomorphisms 403
Viacheslav Grines and Olga Pochinka

32 Isentropic Dynamics and Control in an Economic Model for Capital Accumulation 429
Cristina Januário, Clara Grácio, Diana A. Mendes, and Jorge Duarte

33 Hydrodynamic Limit of the Exclusion Process in Inhomogeneous Media 449
Milton Jara

34 Application of Fractional Order Concepts in the Study of Electrical Potential 467
Isabel S. Jesus and J.A. Tenreiro Machado

35 Economics of Bioenergy Crops for Electricity Generation: Implications for Land Use and Greenhouse Gases 471
Madhu Khanna, Hayri Onal, Basanta Dhungana, and Michelle Wander

36 An H-Theorem for Chemically Reacting Gases 503
Gilberto M. Kremer, Filipe Oliveira, and Ana Jacinta Soares

37 Dynamics, Systems, Dynamical Systems and Interaction Graphs 507
Maurício Vieira Kritz and Marcelo Trindade dos Santos

38 The Dynamics of Scalar Fields in Cosmology 543
José Pedro Mimoso

39 The Dynamics of the Nash Map for 2 By 2 Games 549
R.A. Becker, S.K. Chakrabarti, W. Geller, B. Kitchens, and M. Misiurewicz

40 Resort Pricing and Bankruptcy 567
Alberto A. Pinto, Marta Faias, and Abdelrahim S. Mousa

41 On the Hannay–Ozorio De Almeida Sum Formula 575
M. Pollicott and R. Sharp

42 A Fourier Transform Method for Relaxation of Kinetic Equations 591
Manuel Portilheiro

43 Bethe Ansatz Solution of the Finite Bernoulli Matching Model of Sequence Alignment 599
V.B. Priezzhev and G.M. Schütz

44 Fitness Function Evaluation Through Fractional Algorithms 607
Cecília Reis and J.A. Tenreiro Machado

45 An Exponential Observer for Systems on SE(3) with Implicit Outputs 611
Sérgio S. Rodrigues, Naveena Crasta, António Pedro Aguiar, and Fátima Silva Leite

46 Stability of Polydeuces Orbit 635
Dulce C. Pinto and Filipe C. Mena

47 Peixoto Classification of 2-Dim Flows Revisited 639
Antonio R. da Silva

48 Fractional Control of Legged Robots 647
Manuel F. Silva and J.A. Tenreiro Machado

49 Evolutionary Dynamics of the Spatial Prisoner's Dilemma with Single and Multi-Behaviors: A Multi-Agent Application 651
Carla Silva, Welma Pereira, Jan Knotek, and Pedro Campos

50 Euclidean Jordan Algebras and Strongly Regular Graphs 689
Luís Vieira

51 Parameter Estimation in Stochastic Differential Equations 703
G.-W. Weber, P. Taylan, Z.-K. Görgülü, H. Abd. Rahman, and A. Bahar

52 Stochastic Saddle Paths and Economic Theory 735
A.N. Yannacopoulos

Contributors

Elvio Accinelli Facultad de Economía UASLP, Av. Pintores S/N, Fraccionamiento Burocratas del Estado, CP. 78263 San Luis Potosí, México, elvio.accinelli@eco.uaslp.mx

António Pedro Aguiar Institute for Systems and Robotics, Instituto Superior Técnico, Technical University of Lisbon, Lisbon, Portugal, pedro@isr.ist.utl.pt

Sandra M. Aleixo Mathematics Unit, Instituto Superior de Engenharia de Lisboa, Lisbon, Portugal
and
CEAUL, University of Lisbon, Lisbon, Portugal, sandra.aleixo@dec.isel.ipl.pt

José M. Alonso-Meijide Department of Statistics and Operations Research, Faculty of Sciences of Lugo, University of Santiago de Compostela, Santiago de Compostela, Spain, josemaria.alonso@usc.es

Alberto A. Álvarez-López Departamento de Economía Aplicada Cuantitativa II, UNED, Paseo Senda del Rey, 11, Madrid 28040, Spain, aalvarez@cee.uned.es

Mikel Álvarez-Mozos Department of Statistics and Operations Research, Faculty of Mathematics, University of Santiago de Compostela, Santiago de Compostela, Spain, mikel.alvarez@usc.es

José F. Alves Departamento de Matemática Pura, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre 687, 4169-007 Porto, Portugal, jfalves@fc.up.pt

Vítor Araújo Instituto de Matemática, Universidade Federal do Rio de Janeiro, C. P. 68.530, 21.945-970 Rio de Janeiro, Brazil, vitor.araujo@im.ufrj.br
and
Centro de Matemática da Universidade do Porto, Rua do Campo Alegre 687, 4169-007 Porto, Portugal, vdaraujo@fc.up.pt

Peter Ashwin Mathematics Research Institute, Harrison Building, University of Exeter, Exeter EX4 4QF, UK, p.ashwin@ex.ac.uk

Arifah Bahar Faculty of Science, UTM, Skudai, Malaysia, arifah@mel.fs.utm.my

Nilanjan Banik Center for Advanced Financial Studies, Institute for Financial Management and Research, Chennai 600034, India, nilbanik@gmail.com

A.T. Baraviera I.M. – UFRGS, Porto Alegre 91500-000, Brazil, atbaraviera@gmail.com

Ramiro S. Barbosa Institute of Engineering of Porto, Rua Dr. António Bernardino de Almeida, 431, 4200-072 Porto, Portugal, rsb@isep.ipp.pt

Mário Basto IPCA, Barcelos, Portugal, mbasto@ipca.pt

R.A. Becker Department of Economics, Indiana University, Bloomington, IN 47405, USA, becker@indiana.edu

Mário Bessa Departamento de Matemática, Universidade do Porto, Rua do Campo Alegre 687, 4169-007 Porto, Portugal
and
ESTGOH-Instituto Politécnico de Coimbra, Rua General Santos Costa, 3400-124 Oliveira do Hospital, Portugal, bessa@fc.up.pt

Juan Gabriel Brida Free University of Bolzano, Via Sernesi 1, 39100 Bolzano, Italy, JuanGabriel.Brida@unibz.it

Irene Brito Departamento de Matemática para a Ciência e Tecnologia, Universidade do Minho, 4800 058 Guimarães, Portugal, ireneb@mct.uminho.pt

Francisco Lage Calheiros FEUP, Porto, Portugal, jolacam@netcabo.pt

Pedro Campos LIAAD-INESC Porto and Faculty of Economics, University of Porto, Porto, Portugal, pcampos@fep.up.pt

S.K. Chakrabarti Department of Economics, Indiana University – Purdue University Indianapolis, 425 University Blvd., Indianapolis, IN 46202, USA, imxl100@iupui.edu

T. Charters Área Departamental de Matemática, Instituto Superior de Engenharia de Lisboa, Rua Conselheiro Emídio Navarro, 1, 1949-014 Lisbon, Portugal
and
Centro de Astronomia e Astrofísica da Universidade de Lisboa, Campo Grande, Edifício C8 P-1749-016 LISBOA Portugal, tca@cii.fc.ul.pt

António M. Correia Escola EB 2,3 de Celeirós, Av. Sr. da Paciência, Celeirós, 4705-448 Braga, Portugal, amc7761@gmail.com

Naveena Crasta Institute for Systems and Robotics, Instituto Superior Técnico, Technical University of Lisbon, Lisbon, Portugal, ncrasta@isr.ist.utl.pt

M. Terra Cunha D. M – UFMG, Belo Horizonte 30161-970, Brazil, marcelo.terra.cunha@gmail.com

Antonio R. da Silva Instituto de Matemática, Universidade Federal do Rio de Janeiro, CP 68530, Rio de Janeiro, Brazil, ardasilva@ufrj.br

C.A.A. de Carvalho Instituto de Física, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil, aragao@if.ufrj.br

J.C.S de Miranda Universidade de São Paulo, Instituto de Matemática e Estatística, São Paulo, Brazil, simon@ime.usp.br

Basanta Dhungana Department of Agricultural and Consumer Economics, University of Illinois at Urbana-Champaign, 1301 W. Gregory Drive, Urbana, IL 61801, USA

Iesus C. Diniz Departamento de Matemática, Universidade Federal do Rio Grande do Norte, Natal, Brazil, iesus@ufrnet.br

Jorge Duarte Department of Chemistry, Mathematics Unit, ISEL-High Institute of Engineering of Lisbon, Rua Conselheiro Emídio Navarro, 1949-014 Lisbon, Portugal, jduarte@deq.isel.ipl.pt

P. Duarte CMAF/DM-FCUL, 1749-016 Lisbon, Portugal, pedromiguel.duarte@gmail.com

Marta Faias Department of Mathematics, Universidade Nova de Lisboa, Lisbon, Portugal, mcm@fct.unl.pt

Fernanda A. Ferreira ESEIG, Polytechnic Institute of Porto, Rua D. Sancho I, 981, 4480-876 Vila do Conde, Portugal, fernandaamelia@eu.ipp.pt

Flávio Ferreira ESEIG, Instituto Politécnico do Porto, R. D. Sancho I, 981, 4480-876 Vila do Conde, Portugal, flavioferreira@eu.ipp.pt

M. Ferreira Escola Superior de Estudos Industriais e de Getão do Instituto Politécnico do Porto (IPP), Porto, Portugal
and
LIAAD-INESC Porto LA, Escola de Ciências, Universidade do Minho, Braga, Portugal, migferreira2@gmail.com

Nuno Franco CIMA-UE and Department of Mathematics, University of Évora, Rua Romão Ramalho, 59, 7000-671 Évora, Portugal, nmf@uevora.pt

Jorge Milhazes Freitas Centro de Matemática da Universidade do Porto, Rua do Campo Alegre 687, 4169-007 Porto, Portugal, jmfreita@fc.up.pt

Filomena Garcia ISEG – Technical University of Lisbon and UECE, Rua Miguel Lupi, 20 1249-078 Lisbon, Portugal, fgarcia@iseg.utl.pt

Eric Gautier Ecole Nationale de la Statistique et de l'Administration Economique – CREST, 3 avenue Pierre Larousse, 92245 Malakoff Cedex, France, gautier@ensae.fr

W. Geller Department of Mathematical Sciences, Indiana University – Purdue University Indianapolis, 402 N. Blackford Street, Indianapolis, IN 46202, USA, wgeller@math.iupui.edu

Diogo A. Gomes Instituto Superior Técnico, Av. Rovisco Pais, 1049-001 Lisbon, Portugal, dgomes@math.ist.utl.pt

E.F. Gomes Instituto Superior de Engenharia do Porto, Rua Dr. António Benardino de Almeida, 431, Porto, Portugal, efg@isep.ipp.pt

Patrícia Gonçalves Centro de Matemática da Universidade do Minho, Campus de Gualtar, 4710-057 Braga, Portugal, patg@math.uminho.pt

Rui Gonçalves LIAAD-INESC Porto LA and Section of Mathematics, Faculty of Engineering, University of Porto, R. Dr. Roberto Frias s/n, 4200-465 Porto, Portugal, rjasg@fe.up.pt

Zafer-Korcan Görgülü Universität der Bundeswehr München, Munich, Germany, das_lemma@gmx.de

Clara Grácio Department of Mathematics, Universidade de Évora, Rua Romão Ramalho, 59, 7000-585 Évora, Portugal, mgracio@uevora.pt

Viacheslav Grines Nizhny Novgorod State University, 23 Gagarin Ave, Nizhny Novgorod 603950, Russia, vgrines@yandex.ru

Cristina Januário Department of Chemistry, Mathematics Unit, ISEL-High Institute of Engineering of Lisbon, Rua Conselheiro Emídio Navarro, 1949-014 Lisbon, Portugal, cjanuario@deq.isel.ipl.pt

Milton Jara CEREMADE, Université Paris-Dauphine, Place du Maréchal de Lattre de Tassigny, Paris Cedex 75775, France, jara@ceremade.dauphine.fr

Isabel S. Jesus Institute of Engineering of Porto, Rua Dr. António Bernardino de Almeida, 4200-072 Porto, Portugal, isj@isep.ipp.pt

Özkan Karabacak Mathematics Research Institute, Harrison Building, University of Exeter, Exeter EX4 4QF, UK, o.karabacak@ex.ac.uk

Madhu Khanna Department of Agricultural and Consumer Economics, University of Illinois at Urbana-Champaign, 1301 W. Gregory Drive, Urbana, IL 61801, USA, khanna1@uiuc.edu

B. Kitchens Department of Mathematical Sciences, Indiana University – Purdue University Indianapolis, 402 N. Blackford Street, Indianapolis, IN 46202, USA, kitchens@math.iupui.edu

Jan Knotek LIAAD-INESC Porto and Faculty of Economics, University of Porto, Porto, Portugal, jan.knotek@gmail.com

Gilberto M. Kremer Departamento de Física, Universidade Federal do Paraná, Curitiba, Brazil, kremer@fisica.ufpr.br

Maurício Vieira Kritz LNCC/MCT, Av. Getlio Vargas 333, 25651-075, Petrópolis, Rio de Janeiro, Brazil, kritz@lncc.br

C.F. Lardizabal I.M. – UFRGS, Porto Alegre 91500-000, Brazil, carlos.lardizabal@gmail.com

Fátima Silva Leite Department of Mathematics and Institute for Systems and Robotics, University of Coimbra, Coimbra, Portugal, fleite@mat.uc.pt

A.O. Lopes I.M. – UFRGS, Porto Alegre 91500-000, Brazil
arturoscar.lopes@gmail.com

J.A. Tenreiro Machado Department of Electrotechnical Engineering, Institute of Engineering of Porto, Rua Dr. António Bernardino de Almeida, 431, 4200-072 Porto, Portugal, jtm@isep.ipp.pt

J. Martins LIAAD-INESC Porto LA, Department of Mathematics, School of Technology and Management, Polytechnic Institute of Leiria, Campus 2, Morro do Lena – Alto do Vieiro, 2411-901 Leiria, Portugal, jmmartins@estg.ipleiria.pt

Audrey L. Mayer Michigan Technological University, Department of Social Sciences and School of Forest Resources and Environmental Science, 1400 Townsend Dr., Houghton, MI 49931, USA, almayer@mtu.edu

Christopher McCord Department of Mathematical Sciences, University of Cincinnati, P.O. Box 210025, Cincinnati, OH 45221-0025, USA, mccordck@asmail.artsci.uc.edu

Filipe C. Mena Centro de Matemática, Universidade do Minho, Campus de Gualtar, 4710-057 Braga, Portugal, fmena@math.uminho.pt

Diana A. Mendes Department of Quantitative Methods, Instituto Superior de Ciências do Trabalho e da Empresa, Avenida das Forças Armadas, 1649-026 Lisbon, Portugal, diana.mendes@iscte.pt

José P. Mimoso Centro de Astronomia e Astrofísica da Universidade de Lisboa & Departamento de Física, Faculdade de Ciências, Universidade de Lisboa, Ed. C8 Campo Grande, 1749-016 Lisbona, Portugal, jpmimoso@cii.fc.ul.pt

M. Misiurewicz Department of Mathematical Sciences, Indiana University – Purdue University Indianapolis, 402 N. Blackford Street, Indianapolis, IN 46202, USA, mmisiure@math.iupui.edu

Abdelrahim S. Mousa LIAAD-INESC Porto LA e Departamento de Matemática, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, 687, 4169-007, Portugal, abed11@ritaj.ps

A. Nunes Centro de Física da Matéria Condensada da Universidade de Lisboa, Departamento de Física, Campo Grande, Edifício C8 P-1749-016 LISBOA Portugal, anunes@ptmat.fc.ul.pt

B.M.P.M. Oliveira Faculdade de Ciências da Nutrição e Alimentação da, Universidade do Porto, LIAAD-INESC Porto LA, Porto, Portugal, bmpmo@fcna.up.pt

Filipe Oliveira Centro de Matemática e Aplicações, Universidade Nova de Lisboa, Lisbon, Portugal, fso@fct.unl.pt

Hayri Onal Department of Agricultural and Consumer Economics, University of Illinois at Urbana-Champaign, 1301 W. Gregory Drive, Urbana, IL 61801, USA

Leobardo Plata Facultad de Economía UASLP, Av. Pintores S/N, Fraccionamiento Burocratas del Estado, CP. 78263 San Luis Potosí, México, lplata@uaslp.mx

M.M. Peixoto Instituto de Matemática Pura e Aplicada, Rio de Janeiro, Brazil, peixoto@impa.br

Welma Pereira LIAAD-INESC Porto and Faculty of Economics, University of Porto, Porto, Portugal, welma.pereira@gmail.com

Dinis D. Pestana Department of Statistics and Operational Research, Universidade de Lisboa, Lisbon, Portugal
and
CEAUL, University of Lisbon, Lisbon, Portugal, dinis.pestana@fc.ul.pt

D. Pinheiro CEMAPRE, ISEG, Technical University of Lisbon, Lisbon, Portugal, dpinheiro@iseg.utl.pt

Vilton Pinheiro Departamento de Matemática, Universidade Federal da Bahia, Av. Ademar de Barros s/n, 40170-110 Salvador, Brazil, viltonj@ufba.br

Alberto A. Pinto LIAAD-INESC Porto LA e Departamento de Matemática, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, 687, 4169-007, Portugal
and
Centro de Matemática e Departamento de Matemática e Aplicações, Escola de Ciências, Universidade do Minho, Campus de Gualtar, 4710-057 Braga, Portugal, aapinto@fc.up.pt

Dulce C. Pinto Departamento de Matemática, Universidade do Minho, Campus de Gualtar, 4710 Braga, Portugal, dcpinto@iol.pt

G.A. Pinto Instituto Superior de Engenharia do Porto, Rua Dr. António Benardino de Almeida, 431, Porto, Portugal, gap@isep.ipp.pt

Olga Pochinka Nizhny Novgorod State University, 23 Gagarin Ave, Nizhny Novgorod 603950, Russia, olga-pochinka@yandex.ru

M. Pollicott Department of Mathematics, Warwick University, Coventry CV4 7AL, UK, masdbl@warwick.ac.uk

Manuel Portilheiro Universidad Autónoma de Madrid, Campus de Cantoblanco, 28036 Madrid, Spain, manuel.portilheiro@uam.es

V.B. Priezzhev Laboratory of Theoretical Physics, Joint Institute for Nuclear Research, 141980 Dubna, Russia, priezzvb@theor.jinr.ru

Martín Puchet Posgrados en Economía UNAM, Ciudad Universitaria, C.P. 04510, Mexico, anyul@servidor.unam.mx

Lionello F. Punzo Department of Economics, University of Siena, Piazza S. Francesco, 7, 53100 Siena, Italy

and
PPED-INCT, UFRJ, Rio de Janeiro, Brazil, punzo@unisi.it

Haliza Abd. Rahman Faculty of Science, UTM, Skudai, Malaysia, halizarahman@utm.my

D.A. Rand Warwick Systems Biology & Mathematics Institute, University of Warwick, Coventry CV4 7AL, UK, dar@maths.warwick.ac.uk

Cecília Reis Department of Electrotechnical Engineering, Institute of Engineering of Porto, Rua Dr. António Bernardino de Almeida, 431, 4200-072 Porto, Portugal, cmr@isep.ipp.pt

Joana Resende CEF.UP, University of Porto, Rua Dr. Roberto Frias, 4200-464 Porto, Portugal, jresende@fep.up.pt

J. Leonel Rocha Mathematics Unit, Instituto Superior de Engenharia de Lisboa, Lisbon, Portugal
and
CEAUL, University of Lisbon, Lisbon, Portugal, jrocha@deq.isel.ipl.pt

Sérgio S. Rodrigues Department of Mathematics, University of Cergy-Pontoise, UMR CNRS 8088, 95000 Cergy-Pontoise, France, sesiro@gmail.com

Inmaculada Rodríguez-Puerta Área de Métodos Cuantitativos, Departamento de Economía, Métodos Cuantitativos e Historia Económica, Universidad Pablo de Olavide, Carretera de Utrera Km. 1, Sevilla 41013, Spain, irodpue@upo.es

G.M. Schütz Institut für Festkörperforschung, Forschungszentrum Jülich, 52425 Jülich, Germany, g.schuetz@fz-juelich.de

Viriato Semiao IST, Lisbon, Portugal, ViriatoSemiao@ist.utl.pt

R. Sharp Manchester University, Oxford Road, Manchester M13 9PL, UK, sharp@ma.man.ac.uk

Carla Silva Faculty of Economics, University of Porto, Porto, Portugal, cmap.silva@ymail.com

Luís Silva CIMA-UE and Scientific Area of Mathematics, Instituto Superior de Engenharia de Lisboa, Rua Conselheiro Emídio Navarro, 1, 1959-007 Lisbon, Portugal, lfs@dec.isel.ipl.pt

Manuel F. Silva Department of Electrotechnical Engineering, Institute of Engineering of Porto, Rua Dr. António Bernardino de Almeida, 431, 4200-072 Porto, Portugal, mss@isep.ipp.pt

Ana Jacinta Soares Centro de Matemática, Universidade do Minho, Campus de Gualtar, 4710-057 Braga, Portugal
and
Departamento de Matemática, Universidade do Minho, Braga, Portugal, ajsoares@math.uminho.pt

Pakize Taylan Department of mathematics, Dicle University, Diyarbakır, Turkey, ptaylan@dicle.edu.tr

Mike Todd Centro de Matemática da Universidade do Porto, Rua do Campo Alegre 687, 4169-007 Porto, Portugal, mtodd@fc.up.pt

Marcelo Trindade dos Santos LNCC/MCT, Av. Getlio Vargas 333, 25651-075 Petrópolis, Rio de Janeiro, Brazil, msantos@lncc.br

Enrico Valdinoci Department of Mathematics, Universit degli Studi di Roma, Tor Vergata, Rome, Italy, valdinoc@mat.uniroma2.it

E.G.L.R. Vaz Departamento de Matemática para a Ciência e Tecnologia, Universidade do Minho, 4800 058 Guimarães, Portugal, evaz@mct.uminho.pt

Luís Vieira Center of Mathematics of University of Porto, Faculty of Engineering, University of Porto, Portugal, lvieira@fe.up.pt

Michelle Wander Department of Natural Resources and Environmental Sciences, University of Illinois at Urbana-Champaign, 1301 W. Gregory Drive, Urbana, IL 61801, USA

Gerhard-Wilhelm Weber METU, IAM, Ankara, Turkey, gweber@metu.edu.tr

A.N. Yannacopoulos Athens University of Economics and Business, 76 Patission Str, Athens 11434, Greece, ayannaco@aueb.gr

Chapter 1
Network Control Analysis for Time-Dependent Dynamical States

D.A. Rand

Abstract We present an approach to network control analysis that applies to some important time-dependent dynamical states for both autonomous and non-autonomous dynamical systems. In particular, the theory applies to periodic solutions of autonomous and periodically forced differential equations. The key results are summation theorems that substantially generalise previous results. These results can be interpreted as mathematical laws stating the need for a balance between fragility and robustness in such systems. We also present the theory behind what has been called global sensitivity analysis where sensitivities are defined in terms of principal components and principal control coefficients.

1.1 Introduction

In many areas of science it is necessary to consider complex dynamical systems involving high-dimensional state and parameter spaces. The complex regulatory and signalling systems found in systems biology provide important examples of these. To deploy such models rationally and effectively and to understand their design principles we have to increase our ability to analyse their behaviour. In particular, this is necessary to attack two key tasks: firstly, to determine how such systems address the need for robustness and trade off robustness of some aspects against fragility of others and, secondly, to determine the key points of regulation in such systems, aspects of the system that are crucial to its behaviour and control.

Because it identifies which parameters a given particular aspect of the system is most sensitive to, classical sensitivity analysis [1, 10, 11, 13, 19] is a very useful tool that has been used to address both of these aspects. However, apart from some summation theorems about the control coefficients for period and amplitude of free-running oscillators [10] that are analogous to those derived as in metabolic

D.A. Rand
Warwick Systems Biology & Mathematics Institute, University of Warwick,
Coventry CV4 7AL, UK
e-mail: dar@maths.warwick.ac.uk

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_1,

control analysis [5,10,13], there is currently rather little general theory about general non-equilibrium networks. There is a great need to develop tools that give a more integrated picture of all the sensitivities of a system and to develop more coherent universal or widely applicable general principles underlying these sensitivities.

To this end we demonstrate a new summation theorem which substantially generalises previous results. Not only does it apply to non-stationary solutions such as periodic orbits and transient signals but it also holds for non-autonomous systems, for example for forced nonlinear oscillators. We also present the theory behind what has been called global sensitivity analysis [17]. These results have been discussed in the paper [17] but there the emphasis was on methods to compactly represent all the sensitivities of the system whereas here we are concerned with the underlying mathematical details.

Summation theorems are at the heart of metabolic control theory which is a method based on linear perturbation theory for analysing how the control of fluxes and intermediate concentrations in a metabolic pathway is distributed among the different enzymes that constitute the pathway. Originally designed to quantify the concept of rate limitation in complex enzymic systems, rather than assuming the existence of a unique rate-limiting step, it assumes that there is a definite amount of flux control and that this is spread quantitatively among the component enzymes. While there have been attempts to generalise the theory to non-stationary states [2,9,12,16], up to now a comprehensive theory only exists for stationary states. We attempt to partially rectify this situation here. In doing this we point out that the theory is much more general and applies to general dynamical systems and not just those arising as metabolic systems.

The theorems we prove can be interpreted as mathematical laws stating the need for a balance between fragility and robustness in such systems. They contains within them the other known simple summations theorems such as those for stationary solutions and for the period and amplitude of an oscillatory solution of an unforced oscillator. However, they are a substantial generalisation because they relate a set of functions rather than a set of numbers and thus effectively an infinite number of simple summation conditions. Moreover, as already mentioned, unlike the classical summation theorems they apply to non-autonomous systems such as forced oscillators as well as to autonomous systems. It should be particularly useful in analysing non-stationary regulatory, signalling and metabolic networks. Summation relations for some time-dependent solutions have previously been derived in [3,12,16]. However, as is explained in Appendix 4, the summation theorems proved here are quite different.

We also define control coefficients for some non-stationary solutions and, as will be seen in Sect. 1.3, this definition has some subtleties.

We consider general systems of the form

$$\frac{dx}{dt} = f(t, x) \tag{1.1}$$

where t is time and $x = (x_1 \ldots, x_n)$. In general the systems that we consider will depend upon parameters but we only explicitly mention them in formulas when needed.

We assume a specific decomposition of f as follows

$$f_i(t,x) = \sum_{j=1}^{r} N_{ij} v_j(t,x) \tag{1.2}$$

where the $v_j(t,x)$ are C^2 functions of t and x. These functions also depend on parameters but we do explicitly mention them at this stage.

Usually modelled as compartmental systems, metabolic, regulatory or signalling networks often naturally have such a decomposition. They are typically described by such systems where $N = (N_{ij})$ is the network's stoichiometric matrix and v_j is the rate of a process or reaction j which adds N_{ij} particles to compartment i whenever the reaction occurs. Thus the contribution of reaction j to dx_i/dt is $N_{ij}v_j(t,x)$. Nevertheless, this interpretation in terms of reactions, rates and stoichiometric matrices is not necessary for what follows and any system as in (1.1) can be decomposed as in (1.2).

Control analysis is about trying to quantify the extent to which the aspects of the system described by the various terms v_j affect overall system performance. Suppose the term v_j is changed by a factor λ to λv_j. In the cases we consider the solution of interest g will vary smoothly with this new parameter λ when λ is close to 1 (i.e. will be at least C^2 in λ) and the control coefficient of the solution of interest will be defined in terms of the derivative of g with respect to λ at $\lambda = 1$.

In fact it will be convenient to introduce auxiliary parameters $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_r)$ and define what we will call the *parameter-augmented system* of (1.1) $dx/dt = \hat{f}(t,x,\varepsilon)$ where

$$\hat{f}_i(t,x,\varepsilon) = \sum_{j=1}^{r} N_{ij}(1+\varepsilon_j)v_j(t,x). \tag{1.3}$$

When $\varepsilon = 0$ then we have (1.1). The above derivative of g with respect to λ at $\lambda = 1$ is just the partial derivative of g with respect to ε_j at $\varepsilon = 0$ and the control coefficients can all be defined in terms of these partial derivatives.

Suppose that (1.1) depends upon parameters $k = (k_1, \ldots, k_s)$. We say that k is a full set of linear parameters for (1.1) if

$$f(t,x,\rho k) = \rho f(t,x,k)$$

for all $\rho > 0$. For example, if each N_{ij} above is a non-trivial linear sum of the parameters $k_1, \ldots, k_s$ then k is a full set of linear parameters. For the parameter-augmented system $\hat{f}$ the parameters $\lambda_j = (1 + \varepsilon_j)$ form a full set of linear parameters.

Briefly, the main results that we demonstrate are as follows:

1. A summation theorem for the raw perturbations of solutions with a fixed initial condition.
2. A definition of the control coefficients of periodic orbits of autonomous differential equations and a summation theorem for them.

3. A summation theorem for periodic orbits and other solutions of non-autonomous differential equations.

1.2 Summation and Connectivity Equations in the Stationary Case

We start by describing the well-known summation theorem for stationary solutions [4, 7, 8, 10, 13, 14, 18]. In this case (1.1) must be autonomous and we assume that the stationary solution of (1.1) is non-degenerate in that the Jacobian matrix of f at the stationary solution x^* is invertible. In this case the is a unique stationary solution $\hat{x}^* = \hat{x}^*(\varepsilon)$ of the parameter-augmented system $\hat{f}$ for all small ε such that $\hat{x}^*(0) = x_0$. Moreover, $\hat{x}^* = \hat{x}^*(\varepsilon)$ depends smoothly upon ε (i.e. is C^2 in ε). The steady-state fluxes $J_j(\varepsilon) = (1+\varepsilon)v_j(\hat{x}^*(\varepsilon))$ for $\hat{f}$ will also depend smoothly on ε.

Definition 1.1. The *control coefficients* $C^x_{v_j}$ and the *flux control coefficients* $C^J_{v_j}$ are defined by

$$C^x_{v_j} = \left.\frac{\partial \hat{x}^*}{\partial \varepsilon_j}\right|_{x=x^*(0),\varepsilon=0} \quad \text{and} \quad C^J_{v_j} = \left.\frac{\partial J}{\partial \varepsilon_j}\right|_{x=x^*(0),\varepsilon=0}.$$

Thus we can interpret the control coefficients as describing the change in the stationary state x and the fluxes J that results from a relative change in the reaction rate v_k. If $C^x_{v_k}$ is small then the state is relatively insensitive to this reaction and changes in the reaction rate hardly affect the state. If it is large then changes in this reaction change the state significantly. Similarly for $C^J_{v_k}$.

Theorem 1.1. [4, 7, 8, 10, 13, 14, 18] *If the Jacobian of f in (1.1) is non-singular*

$$\sum_j C^x_{v_j} = 0$$

and

$$\sum_j C^J_{v_j} = J$$

These control coefficients are related by the formula

$$C^J_{v_j} = D_{v_j} C^x_{v_j} + v_j(x^*)e_j. \tag{1.4}$$

Here D_{v_j} is the Jacobian matrix with entries $D_{ik} = \partial v_j/\partial x_k$ evaluated at x^*, and e_j is the vector whose entries are all zero except for the jth which is 1. A short proof of these results is given in Appendix 1.

1.3 Summation Theorems for Time-Dependent States

We will be interested in a solution or a class of solutions of (1.1) defined for a specific time range $t_0 \leq t \leq t_0 + T$. For example, for circadian oscillations, the primary object of interest is an attracting periodic orbit of (1.1) and T will be the period of this orbit.

Suppose that (1.1) and the solution of interest $x = g(t,k)$ depends upon parameters $k = (k_1, \dots, k_s)$. Denote the general solution of (1.1) by $\xi(t, t_0, x_0, k)$ i.e. $x(t) = \xi(t, t_0, x_0, k)$ is the solution of (1.1) with initial condition $x(t_0) = x_0$.

The following function will be important in the summation theorems that we prove:

$$\begin{aligned}\Phi(t,t_0) &= (t-t_0) f(t, g(t,k), k) \\ &\quad - \int_{t_0}^{t} (u-t_0) X(u,t) \frac{\partial f}{\partial t}(u, g(u,k), k)\, du. \end{aligned} \tag{1.5}$$

The $n \times n$ matrices $X(s,t)$ referred to here and below are the fundamental solutions of the variational equation

$$\frac{\partial}{\partial t} X(s,t) = D_f(t) \cdot X(s,t) \tag{1.6}$$

where $D_f(t)$ is the Jacobian $d_x f$ evaluated at $x = g(t,k)$, $X(s,t)$ is a $n \times n$ matrix and the initial condition is $X(s,s) = I$. Then the jth column of $X(t_0, t)$ is $\partial\xi/\partial x_j$ evaluated at $(t, t_0, g(t), k)$ (cf. [6] Chaps. IV and XII (Part I)).

Note that in the case where the equation is autonomous the term under the integral is zero and $\Phi(t,t_0) = (t-t_0) f(g(t,k),k)$. Note that $t f(g(t,k)) = t\dot{g}(t,k)$ which is an infinitesimal period change (i.e. the derivative at $\omega = 1$ of $\omega \to g(\omega t)$).

1.3.1 Solutions Defined by a Fixed Initial Condition

We start with the case where $g(t,k) = \xi(t, t_0, x_0, k)$ and x_0 is independent of parameters.

Theorem 1.2. *Suppose that k is a full set of linear parameters for (1.1) and that the initial condition x_0 does not depend upon k. Then*

$$\sum_j k_j \frac{\partial \xi}{\partial k_j}(t, t_0, x_0, k) = \Phi(t, t_0) \tag{1.7}$$

There is an immediate corollary for what I will call the raw control coefficients $C^{\xi}_{v_j}(t,t_0)$. To define these, consider the parameter-augmented system $\hat{f}$ which depends upon the parameters $\varepsilon = (\varepsilon_1, \dots, \varepsilon_r)$. Then the partial derivative of the solution with respect to ε_j gives the effect of increasing the term v_j by a factor $1 + \varepsilon_j$ to $(1 + \varepsilon_j)v_j$.

Definition 1.2. The *raw control coefficients* are given by

$$C^{\xi}_{v_j}(t,t_0) = \frac{\partial \xi}{\partial \varepsilon_j}(t,t_0,x_0,\varepsilon)\bigg|_{\varepsilon=0}$$

For the parameter-augmented system, the parameters $k_j = 1 + \varepsilon_j$ form a full set of linear parameters and therefore, applying the above theorem to $\hat{f}$, we deduce the following corollary

Corollary 1.1.

$$\sum_j C^{\xi}_{v_j}(t,t_0) = \Phi(t,t_0). \tag{1.8}$$

Remark 1.1. 1. This result contains the classical summation theorem. For that result x_0 is a fixed point so that $\xi(t,x_0) \equiv x_0$ and f is autonomous so $\Phi(t,t_0) \equiv 0$.

2. Even for the fixed-point case it gives new results because we can look at solutions that relax to an equilibrium x^* having started at a different initial condition x_0. The control coefficients for this transient solution satisfy (1.8) and have the property that $\sum_j C^{\xi}_{v_j}(t,t_0) \to 0$ as $t \to \infty$.
3. Note that this theorem applies to both autonomous and non-autonomous systems.

1.3.2 Summation Law for Periodic Solutions

1.3.2.1 Control Coefficients for Periodic Orbits of Autonomous Systems

We firstly consider the autonomous case where f is independent of t. Then the dependence of ξ upon t and t_0 is only through $t - t_0$ and therefore we write it as $\xi(t - t_0, x_0, k)$.

In this autonomous case, a solution of (1.1) corresponding to a periodic orbit Γ_k depending upon parameters k is of the form

$$g(t,k) = \xi(t, x_0(k), k) \tag{1.9}$$

where x_0 is a point on Γ_k that depends smoothly on k.

We refer to the way this is written in terms of t and the parameters k as a *parameterisation* of the periodic orbit. This parameterisation of the periodic orbit is not unique. It has to be of the form given in (1.9) but there are infinitely many ways of choosing $x_0(k)$ since it just has to be a point on the periodic orbit that depends smoothly on k. Since

$$\frac{\partial g}{\partial k_j} = \frac{\partial \xi}{\partial x_0}\frac{\partial x_0}{\partial k_j} + \frac{\partial \xi}{\partial k_j}$$

the choice of $x_0(k)$ affects $\partial g/\partial k_j$ and hence the control coefficient.

To deal with this we firstly define the control coefficients of a periodic orbit using a transversal section Σ to the periodic orbit and then show that the way they depend upon Σ is easy to understand and can be factored out to give a control coefficients for the periodic orbit.

We suppose that Γ is a periodic orbit of (1.1) for the reference parameter value k. Choose a point x_0 on Γ and let Σ be any small transversal section to Γ that intersects it precisely in x_0. By transversal we mean that it is transversal to $f(x_0, k)$ at x_0 i.e. its tangent vectors and $f(x_0, k)$ span the whole of $\mathbb{R}^n$.

Suppose that the period of Γ is τ. We say that Γ is non-degenerate if the matrix $X(0, \tau)$ has 1 as a simple eigenvalue. In this case, for all parameters k' near k there is a unique periodic orbit $\Gamma_{k'}$ near Γ_k and the dependence of this periodic orbit on k' is smooth (e.g. Hartman [6]). Thus for k' near k the perturbed periodic orbit $\Gamma_{k'}$ also intersects Σ in a unique point which we denote by $x_0(k')$. This defines a parameterisation of $\Gamma_{k'}$ by $g(t, k') = \xi(t, x_0(k'), k')$.

Now we turn to the definition of the control coefficients. As above we consider the parameter-augmented system $\hat{f}$ which depends upon the parameters $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_r)$. Since the periodic orbit Γ is non-degenerate, for small ε, the perturbed periodic orbit intersects Σ in a unique point $x_0(\varepsilon)$ and thus we have a parameterised family $g(t, \varepsilon) = \xi(t, x_0(\varepsilon), \varepsilon)$ of periodic orbits depending upon ε.

Definition 1.3. If Σ, g, ε and $x_0(\varepsilon)$ are as above, define the *control coefficient* $C_{v_j}^{\Gamma,\Sigma}$ by

$$C_{v_j}^{\Gamma,\Sigma}(t) = \frac{\partial g}{\partial \varepsilon_j}(t)\bigg|_{\varepsilon=0}.$$

The *flux control coefficient* $C_{v_j}^{J,\Sigma}$ can be similarly defined. We can also consider how the period $\tau = \tau(\varepsilon)$ and the time-scaled solution $\gamma(t, \varepsilon) = g(\tau(\varepsilon)t, \varepsilon)$, $0 < t < 1$, vary with ε. Their partial derivatives with respect to ε_j at $\varepsilon = 0$ define auxiliary control coefficients $C_{v_j}^{\tau,\Sigma}$ and $C_{v_j}^{\gamma,\Sigma}$.

Lemma 1.1. *If Σ' is another transversal section to the periodic orbit then*

$$C_{v_j}^{\Gamma,\Sigma'}(t) = C_{v_j}^{\Gamma,\Sigma}(t) + \beta f(g(t, k), k)$$

where β is a linear function $\partial x_0/\partial \varepsilon_j|_{\varepsilon=0}$ which is independent of j and depends only on Σ and Σ'.

This lemma is proved in Appendix 2. The term $\beta f(g(t), k)$ represents a movement along the orbit Γgiven by the vector field f. β depends upon the sections Σ and Σ' and effectively any value is possible. To remove this dependence upon the section we define the control coefficient as follows.

We can regard the $C_{v_j}^{\Gamma,\Sigma}(t)$ as elements of the L^2 Hilbert space $\mathscr{H}$ of $\mathbb{R}^n$-valued functions of t, $0 \leq t \leq T$. The $\mathbb{R}^n$-valued function $f_g(t) = f(g(t), k)$ also belongs to this space. We therefore consider the quotient space $\mathscr{H}_0 = \mathscr{H}/V_f$ obtained by factoring out the 1-dimensional linear space V_f spanned by f_g.

Definition 1.4. The control coefficients for $\mathit{\Gamma}$ are given by

$$C_{v_j}^{\mathit{\Gamma}} = \pi(C_{v_j}^{\mathit{\Gamma},\mathit{\Sigma}})$$

where $\pi : \mathscr{H} \to \mathscr{H}_0$ is the canonical projection.

For those uncomfortable with such an abstract definition we can choose a easily computable representative $C_{v_j}^{\mathit{\Gamma},\mathrm{rep}}(t)$ for $C_{v_j}^{\mathit{\Gamma}}(t)$ as follows:

$$C_{v_j}^{\mathit{\Gamma},\mathrm{rep}}(t) = C_{v_j}^{\mathit{\Gamma},\mathit{\Sigma}} - \left\langle C_{v_j}^{\mathit{\Gamma},\mathit{\Sigma}}, f_g \right\rangle_{L^2} f_g.$$

Using the lemma, it is easy to see that this is independent of the choice of $\mathit{\Sigma}$.

Theorem 1.3.

$$\sum_j C_{v_j}^{\mathit{\Gamma},\mathit{\Sigma}}(t) = tf(g(t),k), \qquad \sum_j C_{v_j}^{\tau,\mathit{\Sigma}} = -1$$
$$\text{and} \quad \sum_j C_{v_j}^{\gamma,\mathit{\Sigma}} = 0. \tag{1.10}$$

Consequently,

$$\sum_j C_{v_j}^{\mathit{\Gamma}}(t) = 0 \qquad \text{and} \qquad \sum_j C_{v_j}^{\mathit{\Gamma},\mathrm{rep}}(t) = 0.$$

In the following theorem we consider the case where (1.1) depends upon a full set of linear parameters. The periodic orbit $\mathit{\Gamma}$ is assumed to be non-degenerate.

Theorem 1.4. *Suppose that* $g(t,k) = \xi(t, x_0(k), k)$ *is the parameterisation determined by* $\mathit{\Sigma}$ *as defined above. Then*

$$\sum_j k_j \frac{\partial \tau}{\partial k_j} = -1 \ , \qquad \sum_j k_j \frac{\partial \gamma}{\partial k_j} = 0 \tag{1.11}$$

$$\text{and} \quad \sum_j k_j \frac{\partial g}{\partial k_j}(t) = tf(g(t,k),k) = \mathit{\Phi}(t,0). \tag{1.12}$$

Remark 1.2. Theorems 1.3 and 1.4 are effectively equivalent as will be seen from their common proof in Appendix 2. Quite different summation relationships for the case of periodic orbits of autonomous systems has been proved in [3, 12, 16]. Their results are discussed in Appendix 3.

1.3.2.2 Non-Autonomous Systems

Now we discuss the non-autonomous case where f in (1.1) does depend upon t. We assume that $f(t+\tau,x) \equiv f(t,x)$ and that the periodic solution $g(t,k)$ is of period τ and is non-degenerate in that the matrix $X(0,\tau)$ does not have 1 as an eigenvalue. This implies that the periodic orbit Γ_k is isolated and therefore the parameterisation of it as $g(t,k) = \xi(t,t_0,x_0(k),k)$ is unique because $x_0(k)$ is unique since $x_0(k) = g(t_0,k)$. Thus we do not have to worry about the multiple parameterisations found in the autonomous case. The non-degeneracy also implies (a) that for all parameters k' near k, for the system parameterised by k', there is a unique periodic orbit $\Gamma_{k'}$ near Γ_k, and (b) that, for k' near k, the period of $\Gamma_{k'}$ equals that of Γ_k.

Definition 1.5. If Γ is a periodic orbit as above the *control coefficients* $C^{\Gamma}_{v_j}$ are defined by

$$C^{\Gamma}_{v_j}(t) = \frac{\partial g}{\partial \varepsilon_j}(t)\bigg|_{\varepsilon=0}.$$

Let

$$\Psi(t,t_0) = X(t_0,t)(I - X_{t_0})^{-1}\Phi(t_0+\tau,t_0) + \Phi(t,t_0) \tag{1.13}$$

where $X_{t_0} = X(t_0,t_0+\tau)$ and Φ is as in (1.5).

Theorem 1.5. *If f and g are as above and $k = (k_1,\dots,k_s)$ is a full set of linear parameters then*

$$\sum_j k_j \frac{\partial g}{\partial k_j}(t) = \Psi(t,t_0) \tag{1.14}$$

and we also have for the control coefficients that

$$\sum_j C^{\Gamma}_{v_j}(t) = \Psi(t,t_0). \tag{1.15}$$

1.4 Principal Control Coefficients

We now consider a general differential equation of the form given in (1.1) and depending upon parameters $k = (k_1,\dots,k_s)$. Ideally we would like to associate a single real number to each parameter k_j as a measure of its global sensitivity. This is clearly not possible because we want this to account for all ways in which the system varies. However, we can do something which usually is nearly as good. Instead of associating a single number measuring the sensitivity of a parameter k_j we will define a set of numbers S_{ij} with the property that to understand the sensitivity of the system to k_j one just has to inspect the S_{ij} for low i. For generic systems the will be a unique set of such numbers satisfying the optimality condition described below which is basically that if $\sigma_i^2 = \sum_j S_{ij}^2$ then the σ_i decrease as fast as possible.

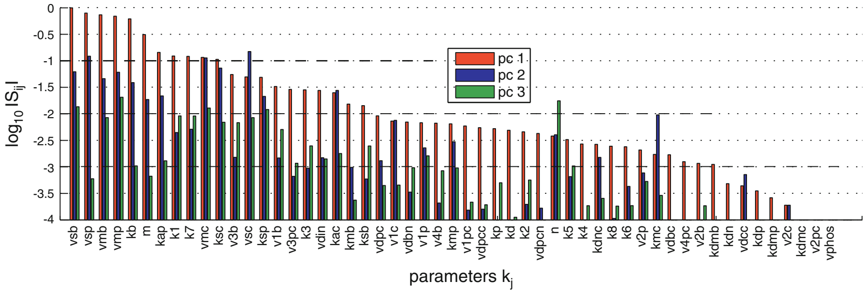


Fig. 1.1 The parameter sensitivity spectrum (pss) for the model of the mammalian circadian clock of Leloup and Goldbeter [15]. Each group of bars corresponds to the value of $\log_{10}|S_{ij}|$ for a particular parameter k_j. These are only plotted for those i for which $|S_{ij}|$ is significant (in this case $i = 1, 2$ and 3). As shown in the legend they are coloured as follows: $i = 1$, *red*; $i = 2$, *blue*; and $i = 3$, *green*. The parameters k_j are ordered by $|S_{1j}|$. See [17] for a discussion of how this can be used to understand the sensitivities of a system like this

The change δg in g caused by a change $\delta k = (\delta k_1, \ldots, \delta k_s)$ in k is

$$\delta g = M\,\delta k + \mathrm{O}(\|\delta g\|^2).$$

where the linear map M is given by

$$M\,\delta k = \sum_j \frac{\partial g}{\partial k_j}\,\delta k_j.$$

We regard M as a map from the parameter space $\mathbb{R}^s$ to the L^2 Hilbert space $\mathscr{H}$ of $\mathbb{R}^n$-valued functions $U(t) = (U_1(t), \ldots, U_n(t))$, $U'(t) = (U_1'(t), \ldots, U_n'(t))$, $0 \leq t \leq T$, with inner product

$$\left\langle U, U'\right\rangle_{L^2} = T^{-1}\int_0^T \sum_{m=1}^{n} U_m(t)U_m'(t)\,dt$$

and norm given by $||U||_{L^2}^2 = \langle U, U\rangle_{L^2}$.

The adjoint operator M^* to M is given by

$$M^*U = (\eta_1, \ldots, \eta_s)$$

swhere

$$\eta_j = \left\langle \frac{\partial g}{\partial k_j}, U\right\rangle_{L^2}.$$

It follows that the ijth element of M^*M is given by

$$\left\langle \frac{\partial g}{\partial k_i}, \frac{\partial g}{\partial k_j}\right\rangle_{L^2}.$$

Since this is self-adjoint it has real positive eigenvalues $\nu_1 \geq \nu_2 \geq \cdots \geq \nu_s$.

Let $\mathscr{H}_0$ denote the subspace spanned by the functions $\partial g/\partial\eta_j(t)$ on $0 \leq t \leq T$.

Theorem 1.6. *There exists a set of numbers $\sigma_I \geq \sigma_2 \geq \cdots \geq \sigma_s$, a set of orthonormal vectors $V_1, \ldots V_s$ of the parameter space $\mathbb{R}^s$, and a set of orthonormal vectors $U_1, \ldots, U_s$ in $\mathscr{H}_0$ such that $MV_i = \sigma_i U_i$, $M^* U_i = \sigma_i V_i$, and the average error given by*

$$e_k^2 = \int_{||v||=1} \|Mv - \sum_{i=1}^{k} \langle Mv, U_i\rangle\, U_i\|^2\, dv$$

is, for all $k \geq 1$, minimised over all orthonormal bases of $\mathscr{H}_0$. At this minimal value $e_k^2 = c\sigma_k^2$ where c is an absolute constant whose value is given in the proof of Theorem 1.6. The σ_i are uniquely determined and the V_i and U_i are respectively eigenvectors of MM^* and M^*M. Thus $\sigma_i = \nu_i$. If they are simple eigenvectors then the U_i and V_i are uniquely determined.

Note that the matrix V whose columns are the vectors V_i is othogonal in that $V^t V = VV^t$ is the identity matrix and that therefore $W = (W_{ij})$ is the inverse of V and $W = V^t$.

Now suppose that $\mathbf{U}' = (U_i')$ is another orthonormal basis of $\mathscr{H}_0$ and define the $s \times s$ matrix $S(\mathbf{U}') = (s_{ij})$ by

$$M\,\delta k = \sum_{i,j} s_{ij}\,\delta k_j\, U_i'.$$

$S(\mathbf{U}')$ is called the *sensitivity matrix* associated to $\mathbf{U}'$ because

$$\|M\,\delta k\| = \|S(\mathbf{U}')\,\delta k\|$$

so that $\|\delta g\| = \|S(\mathbf{U}')\,\delta k\| + \mathrm{O}(\|\delta k\|^2)$. The entries are called *global sensitivities.*

The following corollary is proved in Appendix 3.

Corollary 1.2. *If $\mathbf{U}$ is as in Theorem 1.6 and $\mathbf{U}' = (U_i')$ is another orthonormal basis of $\mathscr{H}_0$ then for all $k = 1, \ldots, s$*

$$\sum_{i\leq k}\sum_j S_{ij}(\mathbf{U})^2 \geq \sum_{i\leq k}\sum_j S_{ij}(\mathbf{U}')^2 \tag{1.16}$$

and

$$\sum_{i>k}\sum_j S_{ij}(\mathbf{U})^2 \leq \sum_{i>k}\sum_j S_{ij}(\mathbf{U}')^2 \tag{1.17}$$

Definition 1.6. The *principal global sensitivities* of f parameterised by k at $k = k_0$ are the numbers $S_{ij} = S_{ij}(\mathbf{U})$. They satisfy the optimality condition expressed in conditions (1.16) and (1.17) of the above corollary. Moreover, $S_{ij} = \sigma_i W_{ij}$ where W is the matrix defined above and therefore $\sum_j S_{ij}^2 = \sigma_i^2 \sum_j W_{ij}^2 = \sigma_i^2$. The elements U_i of the above basis are called *principal components* and the σ_i are called *singular values.*

The principal global sensitivities for a model of the mammalian circadian clock are shown in Fig. 1.1.

Remark 1.3. The use of this terminology is further justified by the following facts:

1. $\|\delta g\|^2 = \|S \cdot \delta k\|^2$
2. $\sum_{i=1}^{k} \sigma_i^2 = \| \sum_{j=1}^{k} S \cdot V_j \|$

Proof of Remark. Since $\delta g = \sum_{i,j} S_{ij} \delta k_j U_i$ up to first order terms,

$$\begin{aligned}\|\delta g\| &= \left(\sum_i \sum_j |S_{ij} \delta k_j|^2 \right)^{1/2} \pm O(\|\delta g\|^2) \\ &= \|S \cdot \delta k\| \pm O(\|\delta g\|^2)\end{aligned}$$

because the U_i are of unit length and orthogonal to each other.

However, $S \cdot \delta k = \mathrm{diag}(\sigma) W \cdot \delta k$. Thus if $W \cdot \delta k$ is the vector u_k whose first k entries are 1 and the rest are zero, then $\|S \cdot \delta k\|^2 = \sum_{i=1}^{k} \sigma_i^2$. In this case $\delta k = V \cdot u_k$ and therefore the jth entry of δk is $\sum_{i=1}^{k} V_{ji}$. Thus $\sum_{i=1}^{k} \sigma_i^2 = \| \sum_{i=1}^{k} V_j \|^2$. □

1.4.1 Principal Control Coefficients

We now define the principal control coefficients $C_{v_j}^{(i)}$ of the system given by (1.1). To define these we consider the parameter-augmented system $dx/dt = \hat{f}(t, x, \varepsilon)$ introduced in (1.3) above. Let $\mathbf{U}$ and $S(\mathbf{U}) = S_{ij}$ be the principal components and global sensitivities of $\hat{f}$ parameterised by $k_j = (1 + \varepsilon_j)$ at $\varepsilon = 0$.

Definition 1.7. The principal control coefficients $C_{v_j}^{(i)}$ are given by the principal global sensitivities S_{ij}.

Theorem 1.7. *(Summation theorem for principal control coefficients)*

$$\sum_{i,j} C_{v_j}^{(i)} U_i(t) = \Psi(t) \tag{1.18}$$

where $\Psi(t)$ is as in Table 1.1.

Table 1.1 The function $\Phi(t, t_0)$ is given in (1.5)

System	Orbit g	$\Psi(t)$
Autonomous	Signal	$tf(g(t))$
Autonomous	Periodic orbit	$tf(g(t))$
Forced	Signal	$\Phi(t, 0)$
Forced, period τ	Periodic orbit	$X_0(I - X_0)^{-1}\Phi(\tau, 0) + \Phi(t, 0)$

Appendix 1: Proof of Theorem 1.1

Proof of Theorem 1.1. We consider the parameter-augmented system $dx/dt = \hat{f}(x,\varepsilon)$ introduced in Sect. 1.1 which is autonomous in the case being considered here since f is. Since the parameters $k_j = 1 + \varepsilon_j$ form a full system of linear parameters for $\hat{f}$, $\sum_j (1+\varepsilon_j)\partial\hat{f}/\partial\varepsilon_j = \hat{f}$. Because the Jacobian $D_f(x_*)$ of f at x_* is invertible, by the implicit function theorem, there is a unique stationary solution $x_*(\varepsilon)$ near the stationary solution $x_*(0) = x_*$ of f and this depends smoothly upon ε with

$$\frac{\partial x_*}{\partial\varepsilon} = D_f(x_*)^{-1}\frac{\partial\hat{f}}{\partial\varepsilon}.$$

Therefore, denoting the column vector consisting of 1's by $\underline{1}$ we have

$$\sum_i \left.\frac{\partial x_*}{\partial\varepsilon_i}\right|_{\varepsilon=0} = \left.\frac{\partial x_*}{\partial\varepsilon}\right|_{\varepsilon=0}\cdot\underline{1} = -D_f(x_*(0))^{-1}\left.\frac{\partial\hat{f}}{\partial\varepsilon}\right|_{\varepsilon=0}\cdot\underline{1}$$
$$= -D_f(x_*(0))^{-1}\hat{f}(x_*(0),0) = 0$$

Equation (1.4) follows from differentiating the equation $J_j = (1+\varepsilon_j)v_j(x^*(\varepsilon))$ with respect to ε. One obtains

$$\frac{\partial J_i}{\partial\varepsilon_j} = \delta_{ij}v_j(x^*(\varepsilon)) + (1+\varepsilon_i)\frac{\partial v_i}{\partial x}\frac{\partial x^*}{\partial\varepsilon_j}$$

or in matrix form

$$\frac{\partial J}{\partial\varepsilon} = v(x^*(\varepsilon)) + \mathrm{diag}(1+\varepsilon)\frac{\partial v}{\partial x}\frac{\partial x^*}{\partial\varepsilon}.$$

The equation $\sum_j C^J_{v_j} = 1$ follows from this and $\sum_j \partial x^*/\partial\varepsilon_j = 0$. □

Appendix 2: Proofs of Theorems 1.2–1.5

We will consider the case where (1.1) depends upon parameters $k = (k_1,\ldots,k_s)$. Suppose that we denote the solution of the differential equation (1.1) with initial condition $x(t_0) = x_0$ and parameters k by $\xi(t,t_0,x,k)$. In Sect. 1.3 above we explained that the derivatives $\partial\xi/\partial x_i$ and $\partial\xi/\partial k_j$ are given by the variational equation

$$\frac{\partial}{\partial t}X(s,t) = D_f(t)\cdot X(s,t) \tag{1.19}$$

where $t \geq s$, $D_f(t)$ is the Jacobian $d_x f$ evaluated at $x = g(t,k)$, $X(s,t)$ is a $n\times n$ matrix and the initial condition is $X(s,s) = I$.

To determine partial derivatives with respect to parameters we consider the associated equation

$$\dot{y}(t) = D_f(t) \cdot y(t) + K_j(t) \tag{1.20}$$

where $K_j(t)$ is the n-dimensional vector $\partial f/\partial k_j$ evaluated at $(t, x) = (t, g(t))$, y is also a n-dimensional vector and the initial condition is $y(t_0) = 0$. Then

$$y(t) = \frac{\partial \xi}{\partial k_j}(t, t_0, g(t), k).$$

If $X(s, t)$ is the solution of (1.19) then by variation of constants,

$$\frac{\partial \xi}{\partial k_j}(t, t_0, g(t), k) = \int_{t_0}^{t} X(s, t) K_j(s)\, ds. \tag{1.21}$$

We firstly consider the case where (1.1) is autonomous. Then $\xi(t, t_0, x_0, k)$ only depends upon t and t_0 through $t - t_0$ so we denote it by $\xi(t - t_0, x_0, k)$.

Lemma 1.2. *If the system is autonomous*

$$X(s, t) \cdot f(g(s)) = f(g(t))$$

Proof. (*cf.* Hartman [6]) Let $x_0 = g(0)$ and let Σ be the normal hyperplane to $f(x_0)$ at x_0. Let $\varphi^t : U \to \mathbb{R}^n$ be the flow of equation (1.1) i.e. $\varphi^t(x) = \xi(t, x)$ and U is some neighbourhood of x_0 in $\mathbb{R}^n$.

Consider the case $s = 0$ first. We can choose U above so that a coordinate system on U is given by (x_0, t) where $x_0 \in \Sigma$ and $|t| < \rho$ for some $\rho > 0$ and the point with coordinates (x_0, t) is $\xi(t, x_0)$. Then in these coordinates the derivative of φ^t ar x_0 is given by the matrix $X(0, t)$ above whose columns are

$$d\xi/dx_{0,k},\ k = 1, \ldots, n-1 \ \text{ and } \ \alpha^{-1} d\xi/dt.$$

Here $x_{0,k}$ is the kth coordinate of x_0.

Consequently,

$$X(0, t) \cdot f(y_0) = X(0, t) \cdot (0, \ldots, 0, \alpha)^* = f(g(t))$$

for all $t \geq 0$. But $X(s, t) f(g(s)) = X(0, t) f(g(0)) = f(g(t))$ because $X(s, t) X(0, s) = X(0, t)$. This is the required result. □

We now consider nonautonomous systems of the form $\dot{x} = f(t, x, k)$. We can rewrite this as an autonomous system by defining $y = (s, x) \in \mathbb{R} \times \mathbb{R}^n$ and letting $F(y) = (\omega, f(s, x, k))$ where ω is a new parameter that we introduce. Then the equation $\dot{y} = F(y)$ is equivalent to

$$\dot{s} = \omega, \qquad \dot{x} = f(s, x, k) \tag{1.22}$$

and therefore, $(\omega t + t_0, x(t))$ is a solution of (1.22) with initial condition (t_0, x_0) precisely when $x(t)$ is a solution of $\dot{x} = f(\omega t, x, k)$ with $x(t_0) = x_0$. For $\omega = 1$ the latter equation is our original equation.

Lemma 1.3. *For autonomous and nonautonomous systems,*

$$X(s,t) f(s, g(s)) = f(t, g(t)) - \int_s^t X(u,t) \frac{\partial f}{\partial t}(u, g(u))\, du$$

Proof. We rewrite the system in the autonomous form (1.22). The Jacobian J_1 of (1.22) at (s, x) is given by

$$J_1 = \begin{pmatrix} J & \partial f/\partial s \\ 0 & 0 \end{pmatrix}$$

where J is the Jacobian of f. Therefore, for (1.22) the fundamental matrices $\hat{X}(s,t)$ are the solution of the equation

$$\begin{aligned} \dot{X}_x &= J(s+t) X_x + \frac{\partial f}{\partial t}(g(s+t)) X_s \\ \dot{X}_s &= 0. \end{aligned}$$

It follows that the first n columns are the columns of the fundamental matrices $X(s,t)$ for (1.19) and the last column is given by

$$\left(\int_s^t X(u,t) \frac{\partial f}{\partial t}(u, g(u))\, du, 1 \right)^*$$

where * indicates a vector transpose to produce a column vector.

By Lemma 1.2

$$\begin{aligned} (f(g(t)), 1)^* &= \hat{X}(s,t) \cdot (f(g(s)), 1)^* \\ &= \left(X(s,t) f(s, g(s)) + \int_s^t X(u,t) \frac{\partial f}{\partial t}(u, g(u))\, du, 1 \right)^* . \end{aligned}$$

which gives the result. □

Proof of Theorem 1.2

We first consider a general system as in (1.1) with parameters $k = (k_1, \ldots, k_s)$ such that $f(t, x, \rho k) = \rho f(t, x, k)$ for all $\rho > 0$.

From Euler's theorem on homogenious functions,

$$\sum_j k_j \frac{\partial f}{\partial k_j} = f.$$

Therefore, from (1.21), if $K_j(s)$ is $\partial f/\partial k_j$ evaluated at $(s, g(s))$,

$$\begin{aligned}\sum_j k_j \frac{\partial \xi}{\partial k_j}(t, t_0) &= \sum_j \int_{t_0}^{t} X(s,t)\cdot k_j K_j(s)\,ds \\ &= \int_{t_0}^{t} X(s,t)\cdot \sum_j k_j K_j(s)\,ds \\ &= \int_{t_0}^{t} X(s,t)\cdot f(s, g(s), k)\,ds \\ &= \int_{t_0}^{t} \left\{ f(t, g(t), k) - \int_{s}^{t} X(u,t)\frac{\partial f}{\partial t}(u, g(u), k)\,du \right\} ds \\ &= (t-t_0) f(t, g(t), k) - \int_{t_0}^{t} uX(u,t)\frac{\partial f}{\partial t}(u, g(u), k)\,du\end{aligned} \tag{1.23}$$

by Lemma 1.3.

Now to apply this to prove Theorem 1.2 we take f to be the parameter augmented system $\hat{f}$ defined in Sect. 1.4.1 above and let $k_j = (1+\varepsilon_j)$. Then $C^{\xi}_{v_j}(t, t_0) = (\partial\xi/\partial k_j)(t, t_0, x_0, k)$ evaluated at $\varepsilon = 0$ and the theorem follows directly from (1.23). □

Proof of Lemma 1.1

Denote Σ and Σ' by Σ_1 and Σ_2 respectively. Fix the point x_0 on the periodic orbit Γ_k. Choose coordinates (t, y_0) on a neighbourhood of x_0 so that $y_0 \in \Sigma_1$ and (t, y_0) corresponds to the point $\xi(t, y_0, k)$ and moreover, in these coordinates, $f(x_0) = (\alpha, \underline{0})$ where $\underline{0} \in \mathbb{R}^{n-1}$. Then near x_0 the points in Σ_2 are of the form $(t(y_0), y_0)$ where $y_0 \in \Sigma_1$. Here $t : \Sigma_1 \to \mathbb{R}$ is a smooth function.

Suppose that k' is a parameter value close to k. If $g_i(t, k')$ has initial condition x_0^i with $x_0^i \in \Sigma_i$ then we denote by $x_0^i(k')$ the points where $\Gamma_{k'}$ intersects Σ_i. Then

$$\begin{aligned}g_2(t, k') &= \xi(t, x_0^2(k'), k') \\ &= \xi(t + t_0(k'), x_0^1(k'), k') = g_1(t + t_0(k'), k')\end{aligned} \tag{1.24}$$

for some smooth function $t_0(k')$. Differentiating (1.24) with respect to k' at $k' = k$, we deduce that

$$\frac{\partial g_2}{\partial k_j}(t, k) = \frac{\partial \xi}{\partial t}(t, x_0^1, k)\cdot \frac{\partial t_0}{\partial k_j}(k) + \frac{\partial g_1}{\partial k_j}(t, k). \tag{1.25}$$

We need to calculate $\partial t_0/\partial k_j(k)$. Consider $\hat{\xi}(k') = \xi(t_0(k'), x_0^1(k'), k')$ and note that $\hat{\xi}(k') = x_0^2(k')$. The derivative with respect to k' at $(0, x_0^1, k)$ is

$$\begin{aligned}\frac{\partial \xi}{\partial k_j}(0, x_0^1, k) &= \frac{\partial \xi}{\partial t}(0, x_0^1, k) \cdot \frac{\partial t_0}{\partial k_j}(k) \\ &\quad + \frac{\partial \xi}{\partial y_0}(0, x_0^1, k) \cdot \frac{\partial x_0^1}{\partial k_j}(k) + \frac{\partial \xi}{\partial k_j}(0, x_0^1, k) \\ &= f(x_0) \cdot \frac{\partial t_0}{\partial k_j}(k) + \frac{\partial x_0^1}{\partial k_j}(k)\end{aligned}$$

since, when $t = 0$, $\partial \xi/\partial t = f$, $\partial \xi/\partial y_0$ is the identity and $\partial \xi/\partial k_j = 0$. Thus, since $f(x_0) = (\alpha, \underline{0})$ and $\partial x_0^1/\partial k_j(k)$ is tangent to Σ_1, the projections on to the first and second components respectively are $\alpha \partial t_0/\partial k_j(k)$ and $\partial x_0^1/\partial k_j(k)$. Since $\hat{\xi}(k') \in \Sigma_s$, $\partial \hat{\xi}/\partial k_j(k)$ is tangent to it at $k' = k$ and therefore

$$\begin{aligned}\alpha \cdot \frac{\partial t_0}{\partial k_j}(k) &= \partial \hat{\xi}_1/\partial k_j = dt(x_0) \cdot \partial \hat{\xi}_2/\partial k_j \\ &= dt(x_0) \cdot \frac{\partial x_0^1}{\partial k_j}(k).\end{aligned}$$

Combining this with (1.25) we deduce that

$$\frac{\partial g_2}{\partial k_j}(t,k) = f(g_1(t,k)) \cdot \alpha^{-1} dt(x_0) \cdot \frac{\partial x_0^1}{\partial k_j}(k) + \frac{\partial g_1}{\partial k_j}(t,k)$$

since $\partial \xi/\partial t(t, x_0^1, k) = f(g_1(t,k))$. This proves Lemma 1.1 with $\beta = \alpha^{-1} dt(x_0)$. □

Proof of Theorem 1.4

This is for periodic orbits of autonomous systems. We again consider a system with parameters k such that $f(x, \rho k) = \rho f(x, k)$ for all $\rho > 0$. For the given parameter value k we fix an initial condition x_0 on the limit cycle and let Σ be the normal hyperplane to $f(x_0, k)$ at x_0. Then for k' near k we let $x_0(k')$ be the unique intersection of the limit cycle with Σ. Therefore, for k' near k the periodic orbit is given by $g(t, k') = \xi(t, 0, x_0(k'), k')$.

It follows from $f(x, \rho k) \equiv \rho f(x, k)$ that

$$\xi(t, x_0(k), \rho k) = \xi(\rho t, x_0(k), k) \tag{1.26}$$

where $\xi(t, x_0, k)$ denotes $\xi(t, 0, x_0, k)$ throughout this proof.

Applying this to the case where $t = \tau(k)$, the period of the limitcycle, we deduce that

$$\tau(\rho k) = \rho^{-1}\tau(k)$$

and hence, by Euler's theorem, that

$$\sum_j k_j \frac{\partial \tau}{\partial k_j} = -\tau. \tag{1.27}$$

Moreover, (1.26) implies that $x_0(k) = x_0(\rho k)$ because this is where both $\xi(t, 0, x_0, k)$ and $\xi(t, 0, x_0, \rho k)$ intersect Σ for $t > 0$. Therefore,

$$\sum_j k_j \frac{\partial x_0}{\partial k_j} = 0. \tag{1.28}$$

Since $g(t, k) = \xi(t, x_0(k), k)$,

$$\sum_j k_j \frac{\partial g}{\partial k_j} = \frac{\partial \xi}{\partial x_0} \cdot \sum_j k_j \frac{\partial x_0}{\partial k_j} + \sum_j k_j \frac{\partial \xi}{\partial k_j} \tag{1.29}$$

where all derivatives etc are evaluated at t, x_0 and k. But the first term on the right-hand side is zero by (1.27) and the second equals $tf(g(t, k), k)$ by (1.23). Thus,

$$\sum_j k_j \frac{\partial g}{\partial k_j}(t) = tf(g(t, k), k) = \Phi(t, 0). \tag{1.30}$$

Finally, let $\gamma(t, k') = g(\bar{\tau}(k')t, k')$ where $\bar{\tau}(k') = \tau(k')/\tau(k)$. Then γ is periodic in t with period $\tau_0 = \tau(k)$ independent of k'.

To prove (1.10) we note that since

$$\gamma(t, k) = \xi(\bar{\tau}t, x_0(k), k) = \xi(t, x_0(k), \bar{\tau}k)$$

it follows that

$$\frac{\partial \gamma}{\partial k_j} = \frac{\partial \xi}{\partial x_0} \frac{\partial x_0}{\partial k_j} + \bar{\tau} \frac{\partial \xi}{\partial k_j} + \frac{\partial \bar{\tau}}{\partial k_j} \sum_i k_i \frac{\partial \xi}{\partial k_i}.$$

Thus

$$\begin{aligned}\sum_j k_j \frac{\partial \gamma}{\partial k_j} &= \frac{\partial \xi}{\partial x_0} \sum_j k_j \frac{\partial x_0}{\partial k_j} + \bar{\tau} \sum_j k_j \frac{\partial \xi}{\partial k_j} \\ &\quad + \sum_j k_j \frac{\partial \bar{\tau}}{\partial k_j} \sum_i k_i \frac{\partial \xi}{\partial k_i} \\ &= 0\end{aligned}$$

by (1.27) and (1.28).

As in the proof of Theorem 1.2, to deduce Theorem 1.4 from what we have proved here, we take f to be the parameter augmented system $\hat{f}$ defined in Sect. 1.4.1 above and let $k_j = (1 + \varepsilon_j)$ so that $C^{\xi}_{v_j}(t, t_0) = (\partial \xi / \partial k_j)(t, x_0, k)$ □

Proof of Theorem 1.5

Again we consider a system as in (1.1) with some parameters k satisfying $f(t, x, \rho k) = \rho f(t, x, k)$ for all $\rho > 0$. We also assume that f is of period $\tau > 0$ in the sense that $f(t + \tau, x, k) \equiv f(t, x, k)$. The solutions are given by $\xi(t, t_0, x_0, k)$.

We suppose that $x = g(t) = g(t, k) = \xi(t, t_0, x_0, k)$ is a periodic solution with period τ. We assume that this solution is non-degenerate in the sense that 1 is not an eigenvalue of X_{t_0}. Here and below X_t denotes $X(t, t + \tau)$. Note that $X_t = X(0, t) X_0 X(0, t)^{-1}$.

Let $y(t) = \sum_j k_j \partial g / \partial k_j(t)$, $t_0 \le t \le t_0 + \tau$. Since the period τ is independent of k' for k' near k, the derivatives $\partial g / \partial k_j$ are periodic in time with period τ and therefore $y(t)$ also has period τ. Moreover, $y(t)$ is a solution of the equation $\dot{y} = D_f(t) y + K(t)$ where $D_f(t)$ is the Jacobian matrix $\partial f / \partial x$ evaluated at $(t, g(t))$ and $K(t) = \sum_j k_j \partial f / \partial k_j (g(t), k)$. The general solution of this equation is

$$y(t) = X(t_0, t)\, c + \int_{t_0}^{t} X(s, t) K(s)\, ds$$

for some vector c. But the last term equals $\sum_j k_j \partial \xi / \partial k_j$ evaluated at $(t, t_0, g(t_0, k))$ and by (1.7) this is $\Phi(t, t_0)$.

If $y(t) = \sum_j k_j \partial g / \partial k_j (t)$, $y(t_0 + \tau) = y(t_0)$ and therefore, since $X(t_0, t_0)$ is the identity, we deduce that $(I - X_{t_0}) c = \Phi(t_0 + \tau, t_0)$. Since 1 is not an eigenvalue of X_{t_0}, $(I - X_{t_0})$ is invertible and $c = (I - X_{t_0})^{-1} \Phi(t_0 + \tau, t_0)$. Therefore,

$$\sum_j k_j \frac{\partial g}{\partial k_j}(t) = X_{t_0} (I - X_{t_0})^{-1} \Phi(t_0 + \tau, t_0) + \Phi(t, t_0). \tag{1.31}$$

As in the previous proofs we get Theorem 1.5 by applying what we have proved here to the parameter augmented system $\hat{f}$. □

Appendix 3: Proof of Theorem 1.6

Let $\mathbf{U} = U_1, U_2, \ldots$ be an orthogonal basis of unit vectors for $\mathscr{H}_0$. Given $\mathbf{U}$ and $v \in \mathbb{R}^s$ consider the error

$$e_k(v) = M v - \sum_{i=1}^{k} \langle M v, U_i \rangle\, U_i$$

of projecting Mv onto the first k basis elements. We seek a basis which minimizes the mean of the L^2 norm of the error for all $k \ge 1$ i.e. minimises $e_k^2 = \int_{||v||=1} ||e_k(v)||^2\,dv$ for all $k \ge 1$. Define $\sigma_i(\mathbf{U})^2 = \int_{||v||=1} \langle Mv, U_i\rangle^2\, dv = \int_{||v||=1} \langle v, M^*U_i\rangle^2\, dv$. Note that if e is any unit vector in $\mathbb{R}^s$ then $c = \int_{||v||=1} \langle v, e\rangle^2\, dv$ is independent of e since the integral is invariant under rotations. Thus

$$\sigma_i(\mathbf{U})^2 = c||M^*U_i||^2.$$

By orthogonality,

$$e_k^2 = \sum_{i>k} \int_{||v||=1} \langle Mv, U_i\rangle^2 = \sum_{i>k} \sigma_i(\mathbf{U})^2.$$

Thus the optimality condition can be expressed as

$$\text{minimise} \sum_{i>k} \sigma_i(\mathbf{U})^2 \text{ for all } k \ge 1. \tag{1.32}$$

But $\langle Mv, Mv\rangle = ||\sum_i \langle Mv, U_i\rangle U_i||^2 = \sum_i \langle Mv, U_i\rangle^2$, and therefore $\sum_{i\ge 0} \sigma_i(\mathbf{U})^2 = \int_{||v||=1} \langle Mv, Mv\rangle\, dv$ which is a positive constant independent of $\mathbf{U}$. Thus (1.32) is equivalent to

$$\text{maximise} \sum_{i=1}^{k} \sigma_i(\mathbf{U})^2 = \sum_{i=1}^{k} c\,||M^*U_i||^2 \text{ for all } k \ge 1. \tag{1.33}$$

To solve this problem consider

$$F(U_1,\dots U_k) = \sum_{i=1}^{k} \int ||M^*U_i||^2\, dv - \sum_{i=1}^{k} \lambda_i(||U_i||^2 - 1).$$

We seek to maximize F; the Lagrange multiplier λ_i is being introduced so as to enforce the condition that the U_is are unit vectors. The partial derivative $d_{\mathbf{U}}F$ of F with respect to $\mathbf{U} = (U_1,\dots,U_k)$ is given by

$$\begin{aligned} &d_{\mathbf{U}}F(U_1,\dots,U_k)\cdot(\delta U_1,\dots,\delta U_k) \\ &= 2\sum_{i=1}^{k} \{\langle M^*U_i, M^*\delta U_i\rangle - \lambda_i\,\langle U_i, \delta U_i\rangle\} \\ &= 2\sum_{i=1}^{k} \langle Y\cdot U_i, \delta U_i\rangle - \lambda_i\,\langle U_i, \delta U_i\rangle \end{aligned}$$

where $Y = MM^* : \mathscr{H} \to \mathscr{H}$. This is a bounded and compact self-adjoint operator.

Clearly, $d_{\mathbf{U}}F(U_1, \dots, U_k) = 0$ if, and only if,

$$Y \cdot U_i = \lambda_i U_i, \qquad 1 \le i \le k$$

and $||U_i|| = 1$. But since Y is bounded, compact and self-adjoint. by the spectral theorem, there exists an orthonormal basis $\mathbf{U} = U_1, U_2, \dots$ of $\mathscr{H}$ consisting of eigenvectors of Y with corresponding real-valued eigenvalues $\lambda_i = \sigma_i^2$ which decrease monotonically to zero. It follows that $\mathbf{U}$ is the required basis with the optimality condition (1.32).

Consider the function $\kappa_i(v) = \langle Mv, U_i \rangle_{L^2}$ defined on the unit sphere of $\mathbb{R}^s$ given by $\|v\| = 1$. Let V_i be a maximum of κ_i. Then V_i is also a critical point of the mapping $\hat{\kappa}_i : \mathbb{R}^s \to \mathbb{R}$ given by

$$\hat{\kappa}_i(v) = \langle M \cdot v, U_i \rangle_{L^2} - \lambda(\|v\|^2 - 1)$$

where the Lagrange multiplier λ is introduced so as to enforce the condition that the v maximising this is a unit vector. Since the derivative of $\hat{\kappa}_i$ at V_i is

$$\delta v \to \langle U_i, M\delta v \rangle - 2\lambda \langle V_i, \delta v \rangle = \langle M^* U_i - 2\lambda V_i, \delta v \rangle$$

the critical point V_i satisfies $M^*U_i = 2\lambda V_i$. Thus since V_i is a unit vector, $4\lambda^2 = \langle M^*U_i, M^*U_i \rangle = \langle U_i, MM^*U_i \rangle = \sigma_i^2$. Therefore, $M^*U_i = \sigma_i V_i$, and therefore, $MV_i = \sigma_i^{-1} MM^*U_i = \sigma_i U_i$. Moreover,

$$\langle V_i, V_j \rangle = (\sigma_i \sigma_j)^{-1} \langle M^*U_i, M^*U_j \rangle = (\sigma_i \sigma_j)^{-1} \langle U_i, MM^*U_j \rangle = (\sigma_j/\sigma_i) \langle U_i, U_j \rangle = \delta_{ij}.$$

Thus we deduce that this gives the required orthonormal basis $v = V_1, \dots, V_s$ such that

$$MV_i = \sigma_i U_i.$$

Since the V_i are eigenvectors of M^*M, if they are simple then this orthonormal basis is unique.

Proof of Corollary to Theorem 1.7

This follows from the fact that $M^*U_i'{=}(\eta_1, \dots, \eta_s)$ where $\eta_j = \langle \partial g/\partial k_j, U_i' \rangle_{L^2} = S_{ij}$. By the proof of the theorem, $\sigma_i(\mathbf{U}')^2 = \|M^*U_i'\|^2 = c\|s_i\|^2$ where s_i is the ith row of $S(\mathbf{U}')$ and, as we have proved, for all $k = 1, \dots, s$,

$$\sum_{i \le k} \sigma_i(\mathbf{U})^2 \ge \sum_{i \le k} \sigma_i(\mathbf{U}')^2.$$

The corollary follows from this. □

Proof of Theorem 1.7

The theorem follows from the above theorems and the fact that if $\hat{f}$ is the parameter-augmented system and $g(t,\varepsilon)$ is the corresponding solution of interest, then

$$\begin{aligned}\sum_{i,j} C_{v_j}^{(i)} U_i(t) &= \sum_{i,j} S_{ij} U_i(t) \\ &= \sum_j \frac{\partial g}{\partial \varepsilon_j}(t,0,g(0),\varepsilon) = \Psi(t,0).\end{aligned}$$

Appendix 4: Previous Summation Relationships

Quite different summation relationships to that in Theorem 1.6 for the case of periodic orbits has been proved by Demin and Westerhoff and Kholodenko in [3], by Ingalls and Sauro in [12] and by Nikolaev, Atlas and Shuler in [16]. To explain why they are different we consider the results of Nikolaev, Atlas and Shuler. The situation for those of Demin et al. and Ingalls and Sauro is similar.

Nikolaev, Atlas and Shuler consider the case of an equation of the form

$$\frac{dx}{dt} = N_R v(t, Lx + T, p) \tag{1.34}$$

which has been obtained from a system with a stoichiometric matrix N with rank r. N_R is the matrix made up of a set of r independent rows of N and the so-called linking matrix L satisfies $N = LN_R$. The original state s is related to the reduced state x by $x = Ls + \bar{x}$ for some constant vector $\bar{x}$. As in the proof of Theorem 1.6 it is straightforward to show that if $g(t,k)$ is a periodic orbit depending smoothly upon parameters k then

$$\begin{aligned}\frac{\partial g}{\partial k_j}(t) &= X(t)(I - X(T))^{-1} \int_0^T X(s,T)\,\partial f_j(s)\,ds \\ &\quad + \int_0^t X(s,t)\,ds\end{aligned} \tag{1.35}$$

$$\begin{aligned}\frac{\partial g}{\partial k_j}(t,t_0) &= X(t,t_0)(I - X_{t_0})^{-1} \int_{t_0}^{t_0+T} X(s,t_0+T)\,\partial f_j(s)\,ds \\ &\quad + \int_{t_0}^t X(s,t)\,\partial f_j(s)\,ds\end{aligned} \tag{1.36}$$

where $\partial f_j(s)$ is $\partial f/\partial k_j$ evaluated at $(s,0,g(0))$. Define the Green's kernels $G(s,t,t_0)$ by

$$G(s,t,t_0) = \begin{cases} Y(s,t,t_0) + X(s,t) \text{ if } t_0 < s < t \\ Y(s,t,t_0) \text{ if } t < s < t_0 + T \end{cases}$$

where $Y(s,t,t_0) = X(t_0,t)(I - X_{t_0})^{-1}X(s,t_0+T)$. Then a straightforward calculation using (1.36) gives that

$$\frac{\partial g}{\partial k_j}(t,t_0) = \int_{t_0}^{t_0+T} G(s,t,t_0)\,\partial f_j(s)\,ds.$$

However, in the case of (1.34) it follows that $\partial f_j(s) = N_R \cdot V_j(s)$ where $V_j(s)$ is $\partial v/\partial k_j$ evaluated at $(s,0,g(0))$. Consequently, if $C(s,t) = LG(s,t,0)N_R$ we can write (1.35) as

$$\frac{\partial g}{\partial k_j}(t) = \int_0^T C(s,t)V_j(t,s,g(0))\,ds.$$

Now if K is a matrix with independent columns such that $N_R K = 0$ then we have

$$\int_0^T C(s,t)K\,ds = 0. \tag{1.37}$$

Moreover, since $dX(s,t)/ds = -X(s,t)D_f(s)$,

$$\int_0^t X(s,t)D_f(s,g(s))\,ds = X(t) - I$$

and therefore, using the above expression for G, we have that

$$\begin{aligned} &\int_0^T C(s,t)D_f(s,g(s))\,ds \\ &= L\int_0^T G(s,t)D_f(s,g(s))\,ds = -L \end{aligned} \tag{1.38}$$

The summation relationships in [16] are of the form in (1.37) and (1.38) or follow from them. They are therefore quite different to that in the above theorems.

References

1. Campolongo, F., Saltelli, A., Sorensen, T., Tarantola, S.: Hitchhiker's guide to sensitivity analysis. In: Saltelli, A., Chan, K., Scott, E.M. (eds.) Sensitivity Analysis, Wiley Series in Probability and Statistics. Wiley, New York (2000)
2. Conradie, R., Westerhoff, H.V., Rohwer, J.M., Hofmeyr, J.H.S., Snoep, J.L.: Summation theorems for flux and concentration control coefficients of dynamic systems. IEE Proc. Syst. Biol. **153**(5), 314–317 (2006)

3. Demin, O.V., Westerhoff, H.V., Kholodenko, B.N.: Control analysis of stationary forced oscillations. J. Phys. Chem. B **103**, 10695–10710 (1999)
4. Fell, D.A.: Metabolic control analysis: a survey of its theoretical and experimental development. Biochem. J. **286**, 313–330 (1992)
5. Fell, D.A.: Understanding the Control of Metabolism. Portand, London (1997)
6. Hartman, P.: Ordinary Differential Equations. Wiley, New York (1964)
7. Heinrich, R., Rapoport, T.A.: Linear theory of enzymatic chains: its application for the analysis of the crossover theorem and of the glycolysis of human erythrocytes. Acta Biol. Med. Germ. **31**, 479–494 (1973)
8. Heinrich, R., Rapoport, T.A.: A linear steady-state treatment of enzymatic chains. general properties, control and effector strength. Eur. J. Biochem. **42**, 89–95 (1974)
9. Heinrich, R., Reder, C.: Metabolic control analysis of relaxation processes. J. Theor. Biol. **151**(3), 343–350 (1991)
10. Heinrich, R., Schuster, S.: The Regulation of Cellular Systems. Chapman and Hall, New York (1996)
11. Hwang, J.T., Dougherty, E.P., Rabitz, S., Rabitz, H.: The green's function method of sensitivity analysis in chemical kinetics. J. Chem. Phys. **69**(11), 5180–5191 (1978)
12. Ingalls, B.P., Sauro, H.M.: Sensitivity analysis of stoichiometric networks: an extension of metabolic control analysis to non-steady state trajectories. J. Theor. Biol. **222**(1), 23–36 (2003)
13. Kacser, H., Burns, J.A., Fell, D.A.: The control of flux. Biochem. Soc. Trans. **23**(2), 341–366 (1973)
14. Kell, D., Westerhoff, H.: Metabolic control theory: its role in microbiology and biotechnology. FEMS Microbiol. Rev. **39**, 305–320 (1986)
15. Leloup, J.C., Goldbeter, A.: Toward a detailed computational model for the mammalian circadian clock. Proc. Natl. Acad. Sci. U.S.A. **100**(12), 7051–7056 (2003)
16. Nikolaev, E.V., Atlas, J.C., Shuler, M.L.: Sensitivity and control analysis of periodically forced reaction networks using the greens function method. J. Theor. Biol. **247**(3), 442–461 (2007)
17. Rand, D.A.: Mapping the global sensitivity of cellular network dynamics: Sensitivity heat maps and a global summation law. J. R. Soc. Interface **5**, S59–S69 (2008)
18. Schuster, P., Heinrich, R.: The definitions of metabolic control analysis revisited. BioSystems **27**, 1–15 (1992)
19. Stelling, J., Gilles, E.D., Doyle, F.J.: Robustness properties of circadian clock architectures. Proc. Natl. Acad. Sci. U.S.A. **101**(36), 13210–13215 (2004)

Chapter 2
Renormalization and Focal Decomposition

Carlos A.A. de Carvalho, Mauricio M. Peixoto, Diogo Pinheiro, and Alberto A. Pinto

Abstract We introduce a renormalization scheme to study the asymptotic dynamical behaviour of a family of mechanical systems with non-isochronous potentials with an elliptic equilibrium. This renormalization scheme acts on a family of orbits of these mechanical systems, all of which are contained on neighbourhoods of the elliptic equilibrium, by rescaling space and shifting time in an appropriate way. We present some new results regarding the properties of this renormalization scheme, and examine the strong connection it has with the focal decomposition for the Euler–Lagrange equation of this family of mechanical systems.

2.1 Introduction

Carvalho et al. [8] have introduced a renormalization scheme that acts on a one dimensional family of orbits of mechanical systems with non-isochronous potentials (see Bolotin and MacKay [4]) with an elliptic equilibrium. This one-dimensional

D. Pinheiro (✉)
CEMAPRE, ISEG, Technical University of Lisbon, Lisbon, Portugal
e-mail: dpinheiro@iseg.utl.pt

M.M. Peixoto
Instituto de Matemática Pura e Aplicada, Rio de Janeiro, Brazil
e-mail: peixoto@impa.br

C.A.A. de Carvalho
Instituto de Física, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil
e-mail: aragao@if.ufrj.br

A.A. Pinto
LIAAD-INESC Porto LA e Departamento de Matemática, Faculdade de Ciências,
Universidade do Porto, Rua do Campo Alegre, 687, 4169-007, Portugal
and
Centro de Matemática e Departamento de Matemática e Aplicações, Escola de Ciências,
Universidade do Minho, Campus de Gualtar, 4710-057 Braga, Portugal
e-mail: aapinto@fc.up.pt

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings
in Mathematics 2, DOI 10.1007/978-3-642-14788-3_2,

family can be characterized through the following two properties: (a) all of its elements have the same initial position, equal to the elliptic equilibrium position; and (b) all of its elements have small enough initial velocities, so that they all lie on the elliptic island surrounding the equilibrium. The renormalization scheme is then defined in such a way that it has the distinguishing feature that time is not rescaled, but rather translated, while the initial velocities and space are appropriately scaled. Since time is translated far away into the the future (or past) standard linearization theory does not apply. Indeed, higher order terms of the period map associated with the non-isochronous potentials must be considered to proceed with the analysis of the renormalized orbits.

The main theorem in Carvalho et al. [8] states that the asymptotic limit of such renormalization scheme is universal: it is the same for all the elements in the considered class of mechanical systems. As a consequence, a universal asymptotic (restricted) focal decomposition for this family of mechanical systems was obtained. This was a first step towards a broader research program, proposed by Peixoto and Pinto, connecting renormalization techniques, focal decomposition of differential equations and semiclassical physics. In Carvalho et al. [9] provide an overview of this research program, describing the sequence of steps that we propose to address in the future which include the convergence of renormalized (restricted) focal decompositions to the universal asymptotic (restricted) focal decomposition and a possible application to semiclassical physics.

In this paper we introduce an extension of the renormalization scheme of Carvalho et al. [8], so that it now acts on all orbits of the family of mechanical systems under consideration and have the initial condition lying on the respective elliptic island. The main motivation for producing this extension is to construct a 4-dimensional universal asymptotic focal decomposition, i.e. one with no restrictions on the base point of the boundary value problem. We also describe the asymptotic limits of this renormalization scheme, their Hamiltonian character and their asymptotic actions. Note that these findings are the foundation for future applications of this theory to semiclassical physics.

Peixoto [15] points out how the focal decomposition is relevant for the computation of the semiclassical quantization via the Feynman path integral method and Carvalho et al. [6, 7] exhibit further relations with quantum statistical mechanics. The concept of focal decomposition is also be relevant to the study of caustic formation by focusing wavefronts in distinct fields of the physical sciences such as optics (see Berry and Upstill [3]), tsunami formation (see Berry [1, 2]) or general relativity (see Friedrich and Stewart [12], Hass et al. [13], Ellis et al. [10] and Ehlers and Newman [11]).

We begin with a review of focal decomposition and renormalization. In Sect. 2.2 we provide definitions and specify notation used in this work. In Sect. 2.3 we introduce a renormalization scheme acting on the orbits of a given family of mechanical systems and state results concerning its asymptotic behaviour, namely, the existence of a two-parameter family of asymptotic trajectories. In Sect. 2.4 we show how to obtain an asymptotic universal focal decomposition from these asymptotic trajectories and how to extend the renormalization scheme to act on focal decompositions.

Section 2.5 is devoted to the computation of the action correspondence of the asymptotic trajectories and a potential application to semiclassical physics. We summarize our conclusions in Sect. 2.6.

2.1.1 Focal Decomposition

The concept of focal decomposition was introduced by Peixoto in [14] and developed by Peixoto and Thom in [16]. For the sake of conciseness, present only the definition of focal decomposition and a notable example due to Peixoto and Thom. Further details can be found in Carvalho et al. [8] and references therein.

Consider the 2-point boundary value problem for ordinary differential equations of the second order

$$\begin{aligned} \ddot{x} &= f(t, x, \dot{x}), t, x, \dot{x} \in \mathbb{R} \\ x(t_1) &= x_1,\ x(t_2) = x_2 \end{aligned} \tag{2.1}$$

and let $\mathbb{R}^4 = \mathbb{R}^2(t_1, x_1) \times \mathbb{R}^2(t_2, x_2)$ be the set of all pairs of points of the (t, x)-plane. To each point (t_1, x_1, t_2, x_2) associate the number $i \in \{0, 1, 2, \ldots, \infty\}$ of solutions of the boundary value problem (2.1) and let $\Sigma_i \subset \mathbb{R}^4$ be the set of points to which the index i has been assigned. Clearly $\mathbb{R}^4$ is the disjoint union of all the sets Σ_i, that is, $\{\Sigma_i\}_i$ is a partition of $\mathbb{R}^4$. This partition is called the *focal decomposition* of $\mathbb{R}^4$ associated with the boundary value problem (2.1):

$$\mathbb{R}^4 = \Sigma_0 \cup \Sigma_1 \cup \ldots \cup \Sigma_\infty.$$

If one of the endpoints in (2.1) is kept fixed, say $(t_1, x_1) = (0, 0)$, then the sets Σ_i induce a decomposition of $\mathbb{R}^2(t_2, x_2)$ by the sets $\sigma_i = \Sigma_i \cap \big(\{(0, 0)\} \times \mathbb{R}^2(t_2, x_2)\big)$. The restricted problem with base point $(0, 0)$ consists of finding the corresponding focal decomposition of $\mathbb{R}^2$ by the sets σ_i:

$$\mathbb{R}^2 = \sigma_0 \cup \sigma_1 \cup \ldots \cup \sigma_\infty.$$

The renormalization scheme of Carvalho et al. [8] was introduced with the goal of studying focal decompositions of $\mathbb{R}^2$ defined by Euler–Lagrange equations associated with elements of a given family of mechanical systems. In the present paper we extend this renormalization scheme in order to deal with focal decompositions of $\mathbb{R}^4$ associated with the same differential equations.

A notable example of a focal decomposition due to Peixoto and Thom [16], is provided by the focal decomposition of the pendulum equation $\ddot{x} + \sin(x) = 0$ with base point $(0, 0)$ (see Fig. 2.1). This focal decomposition contains non-empty sets σ_i with all finite indices. Every set σ_{2k-1}, $k = 1, 2, \ldots$, consists of a 2-dimensional open set plus the cusp-point $(\pm k\pi, 0)$; they all have two connected components. All four connected components of the even-indexed sets σ_{2k} are open-arcs, asymptotic

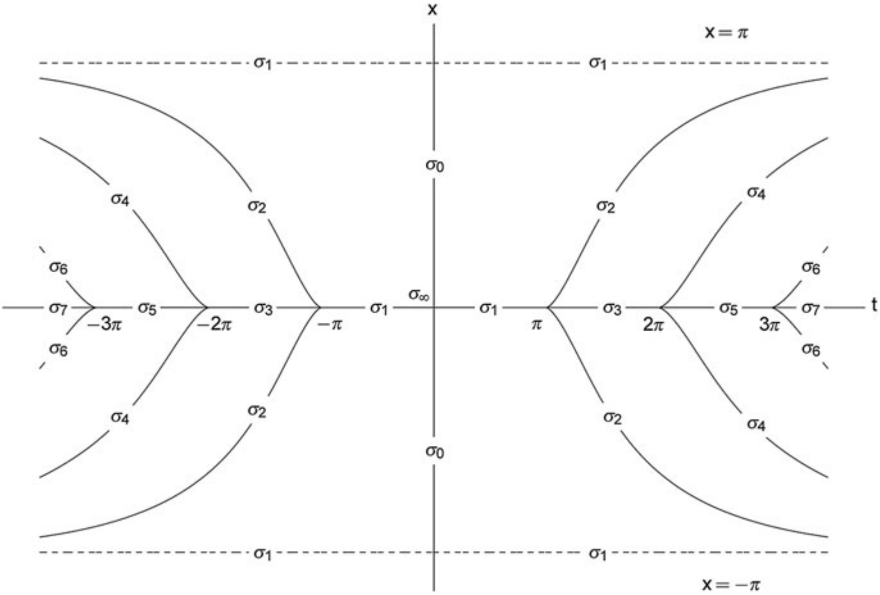


Fig. 2.1 The pendulum's focal decomposition

to one of the lines $x = \pm\pi$ and incident to the cusp-points $(\pm k\pi, 0)$; the lines $x = \pm\pi$ are part of σ_1, except for the points $(0, \pm\pi)$ which belong to σ_0.

2.1.2 *Renormalization*

The main idea behind renormalization is the introduction of an operator – the renormalization operator – on a space of systems whose action on each system is to remove its small scale behaviour and to rescale the remaining variables to preserve some normalization. If a system converges to some limiting behaviour under iteration of the renormalization operator then we say that such behaviour is universal. Since the renormalization operator relates different scales, such universal behaviour is self-similar. See Carvalho et al. [8] and references therein for more details on renormalization.

The main subject of Carvalho et al. [8] is a renormalization scheme acting on the dynamics of a family of mechanical systems that include the pendulum. Our motivation for the introduction of such scheme comes from the restricted focal decomposition with base point $(0, 0)$ of the pendulum equation $\ddot{x} + \sin(x) = 0$ in Fig. 2.1. It turns out that the sequence formed by the even-indexed sets in the pendulum's focal decomposition is approximately self-similar. The renormalization scheme we introduce can then be justified in the following way: for a large integer n, we consider the even-indexed set σ_{2n} and, contrary to previous renormalizations, we do not rescale time but just shift it so that its origin is at $t = n\pi$. We then restrict the initial velocities to a small interval so that that the index corresponding to the shifted even-indexed set is equal to one; we complete the procedure by normalizing space in such way a that the shifted even-indexed set is asymptotic to the lines $x = \pm 1$. Under iteration of this renormalization scheme, we obtain asymptotic trajectories that define an asymptotic focal decomposition. Both the asymptotic trajectories and focal decomposition are universal and self-similar.

2.2 Setting

The purpose of this section is to fix notation and introduce basic definitions which will be used later to state the main results.

We consider mechanical systems defined by a Lagrangian function $\mathscr{L} : \mathbb{R}^2 \to \mathbb{R}$ of the form

$$\mathscr{L}\left(q, \frac{\mathrm{d}q}{\mathrm{d}\tau}\right) = \frac{1}{2}m\left(\frac{\mathrm{d}q}{\mathrm{d}\tau}\right)^2 - \mathscr{V}(q), \tag{2.2}$$

where the potential function $\mathscr{V} : \mathbb{R} \to \mathbb{R}$ is a non-isochronous potential. Furthermore, we assume that the potential $\mathscr{V}$ is a C^κ map ($\kappa \geq 5$) with a Taylor expansion at a point $q^* \in \mathbb{R}$ given by

$$\mathscr{V}(q) = \mathscr{V}(q^*) + \frac{\mathscr{V}''(q^*)}{2}(q - q^*)^2 + \frac{\mathscr{V}^{(4)}(q^*)}{4!}(q - q^*)^4 \pm O\left(|q - q^*|^5\right),$$

where $\mathscr{V}''(q^*) > 0$ and $\mathscr{V}^{(4)}(q^*) \neq 0$. The Euler–Lagrange equation associated with (2.2) is then

$$m\frac{\mathrm{d}^2 q}{\mathrm{d}\tau^2} = -\frac{\mathrm{d}\mathscr{V}}{\mathrm{d}q}(q). \tag{2.3}$$

Alternatively, one can use Hamiltonian formalism, i.e. we consider a Hamiltonian function $\mathscr{H} : \mathbb{R}^2 \to \mathbb{R}$ of the form

$$\mathscr{H}(q, p) = \frac{1}{2m}p^2 + \mathscr{V}(q)$$

and take the symplectic form to be canonical so that the corresponding Hamilton equations are given by

$$\begin{aligned} \frac{\mathrm{d}q}{\mathrm{d}\tau} &= \frac{p}{m} \\ \frac{\mathrm{d}p}{\mathrm{d}\tau} &= -\frac{\mathrm{d}\mathscr{V}}{\mathrm{d}q}(q). \end{aligned} \tag{2.4}$$

The conditions on the potential function $\mathscr{V}$ imply that q^* is an elliptic equilibrium of (2.3) (or equivalently, $(q^*, 0)$ is an elliptic equilibrium of (2.4)) and thus, there is a 1-parameter family of periodic orbits covering a neighbourhood of the equilibrium point.

For the purpose of our study we are interested only on periodic orbits in the elliptic island surrounding the elliptic equilibrium $(q^*, 0)$, i.e. orbits with initial condition (q^0, p^0) on the elliptic island. It turns out to be convenient to express the initial condition on some "modified" polar coordinates that we pass to define.

Since q^* is a local minimum of the potential function $\mathscr{V}$, we obtain that every level set of the Hamiltonian function $\mathscr{H}$ with energy close enough to $\mathscr{H}(q^*, 0)$ is a periodic orbit of the Hamiltonian dynamical system (2.4). Thus, each point (q^0, p^0) in the elliptic island surrounding $(q^*, 0)$ is uniquely determined by the level set of the Hamiltonian function to which the point belongs and by the angle formed by the line joining the equilibrium to the point and the positive horizontal axis, i.e. each point (q^0, p^0) in the elliptic island is uniquely determined by its *energy* $E \in \mathbb{R}_0^+$ and *phase* $\Theta \in \mathrm{S}^1$ given by

$$\mathscr{H}(q^0, p^0) = \mathscr{H}(q^*, 0) + \frac{E^2}{2}\,, \qquad \arctan\left(\frac{p^0}{q^0 - q^*}\right) = \Theta.$$

Note that the coordinates (E, Θ) have a singularity at the point$(q^*, 0)$. Although not necessary, this singularity could be easily removed by requiring E to be strictly positive.

2.3 The Renormalization Operator of Trajectories

In this section we introduce a renormalization operator acting on the trajectories of the Euler–Lagrange equation (2.3) with initial conditions on the elliptic island around $(q^*, 0)$. Moreover, we state some results regarding the asymptotic behaviour of the renormalized trajectories.

2.3.1 Asymptotic Universal Behaviour for the Trajectories

Since q^* is an elliptic equilibrium of (2.3) there is $\epsilon > 0$ such that for all initial conditions with energy $E \in [0, \epsilon]$ and phase $\Theta \in \mathrm{S}^1$ the solutions $q(E, \Theta; \tau)$ of the Euler–Lagrange equation (2.3) are periodic. Thus, the *trajectories* $q : [0, \epsilon] \times \mathrm{S}^1 \times \mathbb{R} \to \mathbb{R}$ of (2.3) are well-defined by $q(E, \Theta; \tau)$ for all $\tau \in \mathbb{R}$, $E \in [0, \epsilon]$ and $\Theta \in \mathrm{S}^1$. Furthermore, there exist $\alpha > 0$ small enough and $N \geq 1$ large enough such that, for every $n \geq N$, the *n-renormalized trajectories* $x_n : [0, 1] \times \mathrm{S}^1 \times [0, \alpha n] \to \mathbb{R}$ are well-defined by

$$x_n(e, \theta; t) = (-1)^n\, \Gamma_{n,t}^{-1}\, \mu^{-1} \left[q\left(\Gamma_{n,t}\, \mu\, \omega\, e, \theta; \frac{n\pi - \ell t}{\omega}\right) - q^* \right],$$

where $\Gamma_{n,t}$ is the (n,t)*-scaling parameter*

$$\Gamma_{n,t} = \left(\frac{8t}{3\pi n}\right)^{1/2}, \tag{2.5}$$

$\ell = \pm 1$ depending on the sign of $\mathscr{V}^{(4)}(q^*)$ and ω and μ are given by

$$\omega = \left(\frac{V''(q^*)}{m}\right)^{1/2}, \qquad \mu = \left(\frac{3!V''(q^*)}{|V^{(4)}(q^*)|}\right)^{1/2}. \tag{2.6}$$

Note that ω^{-1} and μ are the natural time and length scales for the dynamical system defined by (2.3). Furthermore, the variables e and t are dimensionless, as well as the (n,t)-scaling parameter $\Gamma_{n,t}$. Therefore, the n-renormalized trajectories $x_n(e,\theta;t)$ are dimensionless.

Definition 2.1. The *asymptotic trajectories* $X_\ell : [0,1] \times \mathrm{S}^1 \times \mathbb{R}_0^+ \to \mathbb{R}$ are defined by

$$X_\ell(e,\theta;t) = e\ \cos\left(\ell t\left(1-e^2\right)+\theta\right),$$

where $\ell = \pm 1$ depending on the sign of $\mathscr{V}^{(4)}(q^*)$.

The result below generalizes the main result of Carvalho et al. [8] in two ways. On the one hand, we now consider orbits starting form any point in the elliptic island and not just orbits with small velocities starting from the position corresponding to the equilibrium. On the other hand, the convergence below is for the C^2 topology while in Carvalho et al. [8] we only prove it for the C^0 topology. Nevertheless, we already point out in Carvalho et al. [9] that a similar result is true for the C^2 topology.

Theorem 2.1. *Let ℓ be the sign of $\mathscr{V}^{(4)}(q^*)$. The n-renormalized trajectories $x_n(e,\theta;t)$ converge to the asymptotic trajectories $X_\ell(e,\theta;t)$ in the C^2 topology as n tends to infinity.*

2.3.2 The Hamiltonian Character of the Asymptotic Trajectories

The renormalization scheme of the previous section can be extended to act on the *velocities*

$$\dot{q}(E,\Theta;\tau) = \frac{\mathrm{d}q}{\mathrm{d}\tau}(E,\Theta;\tau) \tag{2.7}$$

associated with the periodic solutions $q(E,\Theta;\tau)$ of the Euler–Lagrange equation (2.3) with initial conditions in the elliptic island surrounding the elliptic equilibrium $(q^*,0)$. Similarly to the previous section, the *velocities* $\dot{q} : [0,\epsilon] \times \mathrm{S}^1 \times \mathbb{R} \to \mathbb{R}$ of (2.3) are well-defined by (2.7) for all $\tau \in \mathbb{R}$, $E \in [0,\epsilon]$ and $\Theta \in \mathrm{S}^1$. Furthermore, there exist $\alpha > 0$ small enough and $N \geq 1$ large enough such that, for every $n \geq N$, the *n-renormalized velocities* $y_n : [0,1] \times \mathrm{S}^1 \times [0, \alpha n] \to \mathbb{R}$ are well-defined by

$$y_n(e, \theta; t) = (-1)^n \, \Gamma_{n,t}^{-1} \, \mu^{-1} \omega^{-1} \dot{q} \left(\Gamma_{n,t} \, \mu \, \omega \, e, \theta; \frac{n\pi - \ell t}{\omega} \right),$$

where $\Gamma_{n,t}$ is the *(n,t)-scaling parameter* defined in (2.5) and ω and μ are as given in (2.6). The n-renormalized velocities $y_n(e, \theta; t)$ are also dimensionless.

Definition 2.2. The *asymptotic velocities* $Y_\ell : [0,1] \times \mathrm{S}^1 \times \mathbb{R}_0^+ \to \mathbb{R}$ are defined by

$$Y_\ell(e, \theta; t) = e \; \sin\left(\ell t \left(1 - e^2\right) + \theta\right),$$

where $\ell = \pm 1$ depending on the sign of $\mathscr{V}^{(4)}(q^*)$.

The following result is a natural complement to Theorem 2.1.

Theorem 2.2. *Let ℓ be the sign of $\mathscr{V}^{(4)}(q^*)$. The n-renormalized velocities $y_n(e, \theta; t)$ converge to the asymptotic velocities $Y_\ell(e, \theta; t)$ in the C^2 topology as n tends to infinity.*

There is a strong geometrical and dynamical connection between the asymptotic trajectories $X_\ell(e, \theta; t)$ and the asymptotic velocities $Y_\ell(e, \theta; t)$. As stated in the following theorem, the pair formed by the asymptotic trajectories and the asymptotic velocities is the flow of a canonical Hamiltonian system.

Theorem 2.3. *Let $X_\ell(e, \theta; t)$ and $Y_\ell(e, \theta; t)$ denote, respectively, the asymptotic trajectories and the asymptotic velocities introduced above. The flow $\phi^t(x_0, y_0) = (X_\ell(e, \theta; t), Y_\ell(e, \theta; t))$ with initial condition (x_0, y_0) satisfying the conditions*

$$x_0 = e \cos(\theta) \,, \qquad y_0 = e \sin(\theta)$$

is the Hamiltonian flow of the canonical Hamiltonian system with Hamiltonian function $H_\ell : \mathbb{R}^2 \to \mathbb{R}$ given by

$$H_\ell(x, y) = \ell \left(\left(\frac{x^2 + y^2}{2} \right)^2 - \frac{x^2 + y^2}{2} \right).$$

Similarly, the n-renormalized trajectories $x_n(e, \theta; t)$ and velocities $y_n(e, \theta; t)$ define a flow on a subset of $\mathbb{R}^2$, which we denote by $\phi_n^t(x_0, y_0)$ and call *n-renormalized flow*. For each $n \in \mathbb{N}$, there exists a time-dependent Hamiltonian function $H_n(x, y, t)$, to which we call *n-renormalized Hamiltonian function*, such that $\phi_n^t(x_0, y_0)$ is the Hamiltonian flow of the one-degree of freedom canonical Hamiltonian system determined by $H_n(x, y, t)$. Furthermore, as n tends to infinity, the sequence of n-renormalized Hamiltonian functions H_n converges in the C^1 topology to the asymptotic Hamiltonian function H_ℓ of the previous theorem.

2.4 The Renormalization of Focal Decompositions

In this section we show how to construct the asymptotic universal focal decomposition from the asymptotic trajectories and state a result concerning the convergence of renormalized focal decompositions to the asymptotic universal focal decomposition.

2.4.1 Asymptotic Universal Focal Decomposition

The asymptotic trajectories $X_\ell(e, \theta; t)$ induce an asymptotic universal focal decomposition of $\mathscr{C} \times \mathscr{C} \subset \mathbb{R}^4$, where $\mathscr{C}$ denotes the half-cylinder in $\mathbb{R}^2$ defined by $\mathscr{C} = \mathbb{R}_0^+ \times [-1, 1]$. We briefly describe below how to construct this focal decomposition.

We start by describing the construction of the restricted asymptotic universal focal decomposition of $\mathscr{C}$ with base point $(0, q)$, for some $q \in [-1, 1]$. Let $c^q : I^q \subset\!\!\!\rightarrow [0, 1] \times \mathrm{S}^1$ be the curve in $[0, 1] \times \mathrm{S}^1$ given by $c^q(\lambda) = (e^q(\lambda), \theta^q(\lambda))$ where $e^q(\lambda)$ and $\theta^q(\lambda)$ are such that the relation

$$X_\ell(e^q(\lambda), \theta^q(\lambda); 0) = q$$

holds for all $\lambda \in I^q$ and I^q is the maximal proper subset of $\mathbb{R}$ in such conditions. Let us also define the map $X_\ell^q : I^q \times \mathbb{R}_0^+ \to [-1, 1]$ given by

$$X_\ell^q(\lambda; t) = X_\ell(e^q(\lambda), \theta^q(\lambda); t).$$

The restricted asymptotic universal focal decomposition of $\mathscr{C}$ with base point $(0, q)$ is determined by the sets $\sigma_i^{(0,q)}$ whose elements are pairs $(t, x) \in \mathscr{C}$ such that $X_\ell^q(\lambda; t) = x$ has exactly i solutions $\lambda(t, x) \in I^q$, each distinct solution corresponding to an asymptotic trajectory connecting the points $(0, q) \in \mathscr{C}$ and $(t, x) \in \mathscr{C}$. Therefore, for each $i \in \{0, 1, \dots, \infty\}$, the set $\sigma_i^{(0,q)} \subset \mathscr{C}$ contains all points $(t, x) \in \mathscr{C}$ such that there exist exactly i asymptotic trajectories connecting $(0, q) \in \mathscr{C}$ and $(t, x) \in \mathscr{C}$.

An analogous reasoning would enable us to define the sets $\sigma_i^{(\tau,q)} \subset \mathscr{C}$ containing all points $(t, x) \in \mathscr{C}$ such that there exist exactly i asymptotic trajectories connecting $(\tau, q) \in \mathscr{C}$ and $(t, x) \in \mathscr{C}$. Instead of repeating such construction, we note that by Theorem 2.3 the asymptotic trajectories are solution of a system of autonomous differential equations and, therefore, one could use the invariance of such differential equations under time translations to obtain that the number of asymptotic trajectories connecting $(\tau, q) \in \mathscr{C}$ and $(t, x) \in \mathscr{C}$ is the same as the number of asymptotic trajectories connecting $(0, q) \in \mathscr{C}$ and $(t - \tau, x) \in \mathscr{C}$. Thus, we obtain the restricted asymptotic focal decomposition of $\mathscr{C}$ with base point (t, q) from one with base point $(0, q)$.

To construct the focal decomposition of $\mathscr{C} \times \mathscr{C} \subset \mathbb{R}^4$ we consider the two-dimensional family of sets $\sigma_i^{(\tau,q)} \subset \mathscr{C}$, $(\tau,q) \in \mathscr{C}$, and define

$$\mathscr{C} \times \mathscr{C} = \Sigma_0 \cup \Sigma_1 \cup \ldots \cup \Sigma_\infty,$$

where Σ_i is the set of points $(t_1,x_1,t_2,x_2) \in \mathscr{C} \times \mathscr{C}$ such that there are i asymptotic trajectories connecting $(t_1,x_1) \in \mathscr{C}$ and $(t_2,x_2) \in \mathscr{C}$. The knowledge of the two-dimensional family of restricted focal decompositions of $\mathscr{C}$ determines the focal decomposition of $\mathscr{C} \times \mathscr{C}$.

The following result is a consequence of Theorem 2.1 combined with the discussion above.

Theorem 2.4. *There exists an asymptotic universal focal decomposition of* $\mathscr{C} \times \mathscr{C} \subset \mathbb{R}^4$ *for the Euler–Lagrange equation* (2.3) *induced by the asymptotic trajectories* $X_\ell(e,\theta;t)$.

The restricted asymptotic universal focal decomposition of $\mathscr{C}$ with base point $(0,0)$ is shown in Fig. 2.2. As in the case of the restricted focal decomposition with base point $(0,0)$ of the pendulum equation $\ddot{x} + \sin(x) = 0$ (see Peixoto and Thom [16, pp. 631, 197]), this focal decomposition also exhibits non-empty sets σ_i with all finite indices. For further details on this focal decomposition see Carvalho et al. [8].

If the base point for the restricted asymptotic universal focal decomposition is replaced by a point of the form $(0,q)$ with $q \in (-1,0) \cup (0,1)$, the focal

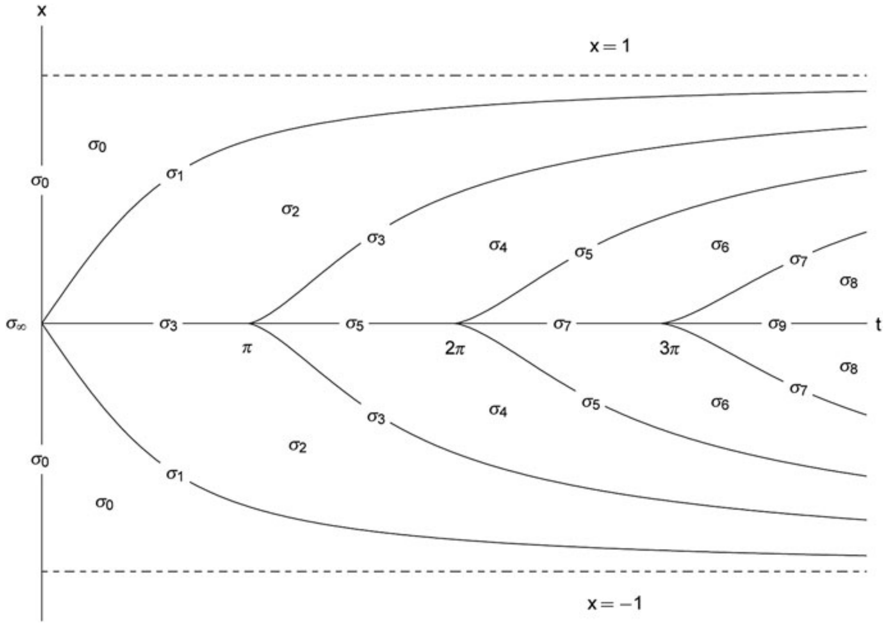


Fig. 2.2 The restricted asymptotic universal focal decomposition with base point $(0,0)$

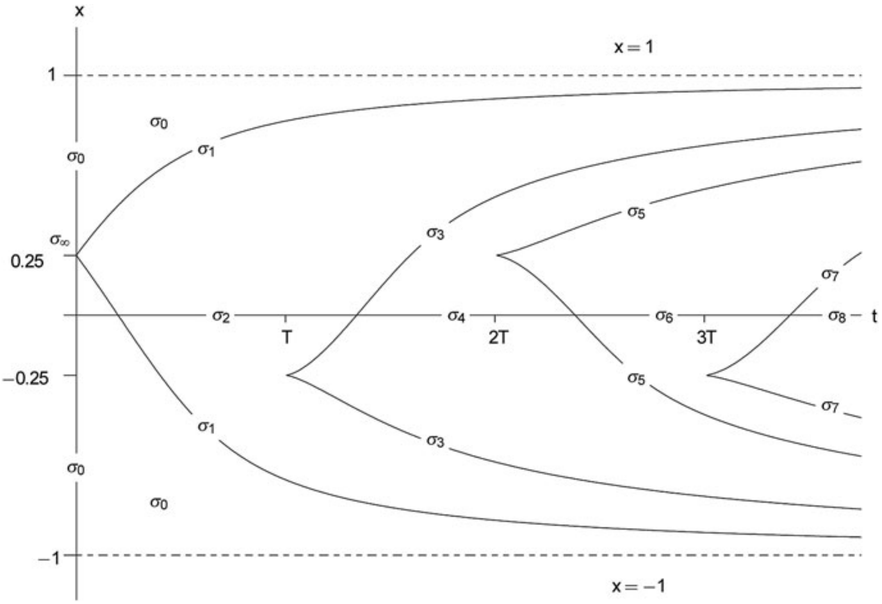


Fig. 2.3 The restricted asymptotic universal focal decomposition with base point $(0, 0.25)$ for $\ell = 1$

decomposition loses some of its symmetry. An example of such focal decomposition is given in Fig. 2.3. For every $k \in \mathbb{N}$, the even-indexed sets σ_{2k} are 2-dimensional connected open sets. The odd-indexed sets σ_{2k-1} are the union of two open arcs, asymptotic to one of the lines $x = \pm 1$ and incident to the cusp-point $p_k = ((k-1)T(q), (-1)^{k-1}q)$, where $T(q)$ denotes the half-period of the asymptotic trajectory starting at $(x, y) = (q, 0)$; The set σ_0 is composed by two connected 2-dimensional open sets, the lines $x = \pm 1$ and the line $t = 0$ except for the base point $(0, q)$ which belongs to σ_∞.

2.4.2 The Renormalized Focal Decompositions

Let us consider the asymptotic universal focal decomposition of $\mathscr{C} \times \mathscr{C}$

$$\mathscr{C} \times \mathscr{C} = \cup_{k=0}^{\infty} \Sigma_k.$$

For each $z = (t_1, x_1, t_2, x_2) \in \Sigma_i$ we define the index $i(z)$ of z equal to i.

Note that each n-renormalized trajectory x_n induces a focal decomposition of $\mathscr{C} \times \mathscr{C}$ by the sets Σ_i^n whose elements are pairs of points $(t_1, x_1, t_2, x_2) \in \mathscr{C} \times \mathscr{C}$ such that there exist a number i of n-renormalized trajectories connecting $(t_1, x_1) \in \mathscr{C}$

to $(t_2, x_2) \in \mathscr{C}$. Therefore, the n-renormalized focal decomposition is given by

$$\mathscr{C} \times \mathscr{C} = \cup_{k=0}^{\infty} \Sigma_k^n .$$

Similarly, for each $z = (t_1, x_1, t_2, x_2) \in \Sigma_i^n$ we define the n-renormalized index $i_n(z)$ of z as the integer i.

The following result states that the sequence of n-renormalized focal decompositions converges to the asymptotic universal focal decomposition.

Theorem 2.5. *For every $z = (t_1, x_1, t_2, x_2) \in \mathscr{C} \times \mathscr{C}$, the n-renormalized index $i_n(z)$ of z converges to the index $i(z)$.*

2.5 The Action of the Asymptotic Trajectories

In this section we show how to compute the action of the asymptotic trajectories, thus defining an *action correspondence*. We finish the section with some brief comments regarding a possible application of the theory reviewed in this paper to semiclassical physics.

2.5.1 The action of the Asymptotic Trajectories

Let N be a smooth manifold, $M = T^*N$ its cotangent bundle, Δ an interval in $\mathbb{R}$ and $H : M \times \Delta \to \mathbb{R}$ a smooth Hamiltonian function. It is well known that a path $\omega : [t_1, t_2] \to M$, from $p_1 \in M$ to $p_2 \in M$, starting at time $t_1 \in \Delta$ and ending at time $t_2 \in \Delta$, is a trajectory of the canonical Hamiltonian system (M, H) if it is a critical point of the action functional in phase space

$$F[\omega] = \int_\omega p \, \mathrm{d}q - H \, \mathrm{d}t.$$

An alternative approach is to write the action functional as the integral

$$F[\omega] = \int_\omega p\dot{q} - H \, \mathrm{d}t$$

and regard the integrand as a Lagrangian function $L : TM \times \Delta \to \mathbb{R}$, obtaining the action functional

$$F[\omega] = \int_\omega L \, \mathrm{d}t = \int_{t_1}^{t_2} L(\dot{\omega}(t), \omega(t), t) \, \mathrm{d}t.$$

Therefore, for any trajectory of the Euler–Lagrange equation determined by L, $\omega : [t_1, t_2] \to M$, one can compute its action $F[\omega]$. Moreover, noticing the dependence of a trajectory ω on its boundary conditions $\omega(t_1) = p_1$ and $\omega(t_2) = p_2$, one can define the *action correspondence* of the trajectories ω connecting (t_1, p_1) to (t_2, p_2) $S : (M \times \Delta)^2 \to \mathbb{R}$ by

$$S(t_1, p_1; t_2, p_2) = \{F[\omega] : \omega(t_1) = p_1,\ \omega(t_2) = p_2\}\,.$$

We remark that S is not a properly defined function, but a correspondence mapping each element of $(M \times \Delta)^2$ to a subset of $\mathbb{R}$. This is due to the fact that there might be more than one trajectory of the Euler–Lagrange equation determined by L connecting (t_1, x_1) to (t_2, x_2).

We denote by $S_\ell(t_1, x_1, t_2, x_2)$ the action correspondence of the asymptotic trajectories $X_\ell(e, \theta; t)$ connecting (t_1, x_1) to (t_2, x_2).

Theorem 2.6. *Let $z = (t_1, x_1, t_2, x_2) \in \mathscr{C} \times \mathscr{C}$. The action correspondence of the asymptotic trajectories $X_\ell(e, \theta; t)$ is given by*

$$S_\ell(z) = \left\{ \frac{e^2}{4} \left(\ell e^2 t + \sin(2(\ell t(1 - e^2) + \theta)) - \sin(2\theta) \right) : (e, \theta) \in \hat{V}(z) \right\},$$

where, for each $z \in \mathscr{C} \times \mathscr{C}$, the set $\hat{V}(z)$ is defined as

$$\begin{aligned}\hat{V}(z) = \big\{&(e(z), \theta(z)) \in [0,1] \times \mathrm{S}^1 : X_\ell(e(z), \theta(z); t_1) \\ &= x_1, X_\ell(e(z), \theta(z); t_2) = x_2\big\}.\end{aligned}$$

We remark that the number of elements in each set $\hat{V}(z)$ is given by the index of the set Σ_i in the asymptotic universal focal decomposition to which $z = (t_1, x_1, t_2, x_2) \in \mathscr{C} \times \mathscr{C}$ belongs. Furthermore, we note that the self-similarities of the asymptotic trajectories and the asymptotic universal focal decomposition are naturally carried over to the action correspondence – see Fig. 2.4 for two examples.

To finish the section we remark that the n-renormalized trajectories $x_n(e, \theta; t)$ and velocities $y_n(e, \theta; t)$, the n-renormalized Hamiltonian function H_n and the n-renormalized focal decomposition induce a renormalization scheme on the actions of the n-renormalized trajectories, i.e. they define n-renormalized action correspondences S_n which converge to the action correspondence of the asymptotic trajectories S_ℓ as n tends to infinity.

2.5.2 Semiclassical Physics

Focal decomposition is in fact a first step towards semiclassical quantization. This was already recognized in the semiclassical calculation of partition functions for quantum mechanical systems, where the need to consider a varying number of classical solutions in different temperature regimes became evident Carvalho et al. [8].

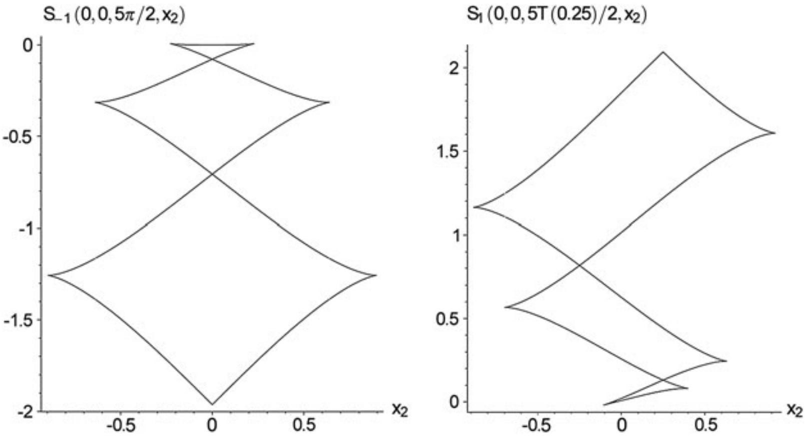


Fig. 2.4 The image on the left is the graph of the action correspondence of the asymptotic trajectories for fixed $t_1 = x_1 = 0$ and $t_2 = 5\pi/2$ and $\ell = -1$. The image on the right is the graph of the action correspondence of the asymptotic trajectories for fixed $t_1 = 0$, $x_1 = 0.25$ and $t_2 = 5T(0.25)/2$ and $\ell = 1$

Either in quantum mechanics or in quantum statistical mechanics, the semiclassical approximation has to sum over all, or part of, the classical paths satisfying fixed point boundary conditions. The number and type of classical trajectories are the very ingredients which lead to a focal decomposition. It should, therefore, be no surprise that the focal decomposition can be viewed as the starting point for a semiclassical calculation.

As for the renormalization procedure, it was introduced to study the behavior of classical trajectories for very short space and very long time separations of the fixed endpoints. It maps those trajectories into n-renormalized ones, whose time separations are shifted by n half-periods, and whose space separations are scaled up to values of order one. As it has been shown in Carvalho et al. [8], this procedure converges to an asymptotic universal family of trajectories that have a well-defined and simple functional form, and which define an asymptotic universal focal decomposition self-similar to the original one.

The natural question to pose is whether the combination of focal decomposition and renormalization can be used to calculate semiclassical expansions for propagators in the short space, long time separation of the endpoints, or analogously, for thermal density matrices for short space separation and low temperatures (long euclidean time $\beta\hbar$ is equivalent to low temperatures $T = 1/(k_B\beta)$) by using the simple asymptotic forms alluded to in the previous paragraph (see Carvalho et al. [8] for a detailed discussion).

The conjecture to be investigated in a forthcoming article is that this can be done in a relatively simple way, thanks to the simple form of the asymptotes. This will bypass a much more difficult calculation involving Jacobi's elliptic functions. Should our expectation be realized, we would obtain semiclassical estimates for both propagators and thermal density matrices in the short space/long time or short space/low temperature limits.

2.6 Conclusions

We have studied the dynamics of a family of mechanical systems that includes the pendulum at small neighbourhoods of an elliptic equilibrium and characterized such dynamical behaviour through a renormalization scheme. We have introduced a renormalization scheme acting on the dynamics of this family of mechanical systems and proved that the asymptotic limit of the renormalization scheme introduced in this paper is universal: it is the same for all the elements in the considered class of mechanical systems. As a consequence we have obtained an universal asymptotic focal decomposition for this family of mechanical systems. We believe that the existence of an universal asymptotic focal decomposition might be useful not only on the theory of boundary value problems of ordinary differential equations but also on several distinct fields of the physical sciences such as quantum statistical mechanics, optics, general relativity and even tsunami formation. Our belief in the utility of this work on such applications is based on the relevance that the concept of focal decomposition may have on the study of caustic formation by focusing wavefronts, of such significance to those fields.

Acknowledgements We thank Robert MacKay and Michael Berry for helpful discussions. We thank the Calouste Gulbenkian Foundation, PRODYN-ESF, FEDER, POFC, POCTI and POSI by FCT and Ministério da Ciência, Tecnologia e Ensino Superior, Centro de Matemática da Universidade do Minho, Centro de Matemática da Universidade do Porto, CEMAPRE and LIAAD-INESC Porto LA for their financial support. We also thank CNPq, FAPERJ, and FUJB for financial support. Part of this research was done during visits by the authors to IMPA (Brazil), The University of Warwick (United Kingdom), Isaac Newton Institute (United Kingdom), IHES (France), CUNY (USA), SUNY (USA) and MSRI (USA), whom we thank for their hospitality. D. Pinheiro's research was supported by FCT – Fundação para a Ciência e Tecnologia grant with reference SFRH/BPD/27151/2006 and program "Ciência 2007".

References

1. Berry, M.V.: Tsunami asymptotics. New J. Phys. **7**, 129 (2005)
2. Berry, M.V.: Focused tsunami waves. Proc. R. Soc. A, **463**, 3055–3071 (2007)
3. Berry, M.V., Upstill, C.: Catastrophe optics: morphologies of caustics and their diffraction patterns. Prog. Optics. **18**, 257–346 (1980)
4. Bolotin, S.V., MacKay, R.S.: Isochronous potentials. In: Vazquez, L., MacKay, R.S., Zorzano, M.P. (eds.) Localization and Energy Transfer in Nonlinear Systems, vol. 217–224, World Scientific, Singapore (2003)
5. Carvalho, C.A.A., Cavalcanti, R.M.: Tunneling catastrophes of the partition function. Brazil. J. Phys. **27**, 373–378 (1997)
6. Carvalho, C.A.A., Cavalcanti, R.M., Fraga, E.S., Jorás, S.E.: Semiclassical series at finite temperatures. Ann. Phys. **273**, 146–170 (1999)

7. Carvalho, C.A.A., Cavalcanti, R.M., Fraga, E.S., Jorás, S.E.: Improved semiclassical density matrix: Taming caustics. Phys. Rev. E. **65**(5), 56112–56221 (2002)
8. Carvalho, C.A.A., Peixoto, M.M., Pinheiro, D., Pinto, A.A.: An asymptotic universal focal decomposition for non-isochronous potentials (submitted)
9. Carvalho, C.A.A., Peixoto, M.M., Pinheiro, D., Pinto, A.A.: Focal decomposition, renormalization and semiclassical physics. J. Differ. Equ. Appl. **17**(6), 1–15 (2011)
10. Ellis, G.F.R., Bassett, B.A.C.C., Dunsby, P.K.S.: Lensing and caustic effects on cosmological distances. Class. Quant. Grav. **15**, 2345–2361 (1998)
11. Ehlers, J., Newman, E.T.: The theory of caustics and wave front singularities with physical applications. J. Math. Phys. **41**, 3344–3378 (2000)
12. Friedrich, H., Stewart, J.M.: Characteristic Initial Data and Wavefront Singularities in General Relativity. Proc. Roy. Soc. Lond. A. **385**, 345–371 (1983)
13. Hasse, W., Kriele, M., Perlick, V.: Caustics of wavefronts in general relativity. Class. Quant. Grav. **13**, 1161–1182 (1996)
14. Peixoto, M.M.: On end point boundary value problems. J. Differ. Equ. **44**, 273–280 (1982)
15. Peixoto, M.M.: Focal decomposition in Geometry, Arithmetic, and Physics. Geometry, Topology and Physics. In: Apanasov, Bradlow, Rodrigues, Uhlenbeck (eds.) de Gruyter & Co. Berlin-New York, pp. 213–231 (1997)
16. Peixoto, M.M., Thom, R.: Le point de vue énumératif dans les problèmes aux limites pour les équations différentielles ordinaires. C. R. Acad. Sci., Paris, Sér. I, **303**, 629–633; erratum, **307**, 197–198 (1988); II **303**, 693–698 (1986)

Chapter 3
Micro-Foundations of the Social Change

Elvio Accinelli and Leobardo Plata

Abstract The aim of this work is to show the relationship between the fundamentals of the economy and social changes in a framework of the General Equilibrium Theory. To analyze this relationship we introduce the Negishi map. This map makes evident the social impact of the efficient reassignments of the economical resources. The social and economic changes occur along the graph of this map. A deeper analysis of this map shows that the social crisis can be perceptible like points in this graph corresponding to the singular economies (from a social point of view). We analyze the possibilities to obtain, in a decentralized way, an equalitarian level of social welfare for an economy with total resources given. This means the possibility to obtain a stable economy in the sense that every agent reach in equilibrium, the same level of utility. Finally we discuss efficiency and equalitarianism in fair economies.

3.1 Introduction

The main concern of this work is to make evident the relationship between economic efficiency and social welfare. We characterize the social structure of the economy by means of a distribution λ on the agents of the economy. This distribution represents the relative weights of the agents in a given social utility function. We show that for each distribution there is a corresponding Pareto efficient allocation and, reciprocally, each efficient allocation has associated to it a distribution of social weights. This assertion is well known. We show that this correspondence between relative social weights and efficient allocations is given by the Negishi map. Furthermore, the graph of this map is a path (or a manifold) of pairs where each pair consists of social weights and its corresponding efficient allocation. The value of this map, at

E. Accinelli (✉) and L. Plata
Facultad de Economía UASLP, Av. Pintores S/N, Fraccionamiento Burocratas del Estado,
CP. 78263 San Luis Potosí, México
e-mail: elvio.accinelli@eco.uaslp.mx, lplata@uaslp.mx

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_3,

each pair, is a measure of the level of social welfare reached by the economy. In this way, the Negishi path, joints two parallel and different points of view of the neo-Walrasian theory. The point of view of the efficiency and the point of view of the social welfare.

In Sect. 3.2 we introduce the model. To give an exact expression for the relationship existing between efficiency and social welfare level we introduce, in Sect. 3.3, the Negishi path. This set is a differential manifold in the cartesian product $S_n \times \mathscr{P}\mathscr{O}$, where S_n is the n-dimensional simplex and $\mathscr{P}\mathscr{O}$ the set of feasible allocation for an economy. We associate a social value to each point (λ, x) in the Negishi path. This value is a measure of the social welfare level reached by a Walrasian economy. A Walrasian economy $\mathscr{E}$ representing a set of consumption spaces, utility functions, endowments and a finite set of agents. We consider the existence of an efficient allocation and the corresponding distribution of social weights, such that all people reach the same level of happiness (meaning that the utilities of every agent evaluate at this allocation are the same).

In Sect. 3.5, we analyze the possibility to reach this allocation in a decentralized way, i.e. without the participation of a central planner. To do this we introduce the excess utility function. In Sect. 3.6 we analyze the stability of an economy and the possibilities that a social crisis appear as a response to a change in the fundamentals of the economy. We understand by a social crisis a big and unforeseeable change in the distribution of the social weights in a framework of continuity. We introduce the definition of singular economy and we show that, if the economy is a singular one, then small changes in the endowments imply big changes in the social structure. In Sect. 3.7 we characterize the social crisis. In Sect. 3.8, we give a definition of a fair economy and its relation with efficiency and equalitarianism in the framework of the general equilibrium theory.

3.2 The Model

We consider a pure exchange economy $\mathscr{E}$ (or Walrasian economy) where

$$\mathscr{E} = \{X, u_i, w_i, I\}\,.$$

Here the set I is a finite set of index, $I = \{1, \ldots, n.\}$ The utilities $u_i : X \to R, i = \{1, \ldots, n\}$ are smooth and strictly concave functions. We assume that the consumption space X for each agent is the positive cone of the space R^l, i.e. $X = R^l_+$, and $w_i \in R^l_+$ represent the endowments of the ith consumer. The total resources are denoted by $W = \sum_{i=1}^n w_i, \in R^l_{++}$, i.e. W is a strictly positive vector in R^l. An allocation is represented by $x = (x_1, \ldots, x_n)$ where x is a vector in R^{ln}. An allocation x is feasible if and only if $\sum_{i=1}^n x_i \leq W$. We denote this set by

$$\mathscr{F} = \left\{ x \in R_+^{ln} : \sum_{i=1}^{n} x_i \leq W. \right\}.$$

We consider the social utility function $U : S_n \times R_+^l \times R$ defined by

$$U(\lambda, x) = \sum_{i=1}^{n} \lambda_i u_i(x_i), \tag{3.1}$$

where S_n is n-dimensional the simplex

$$S_n = \left\{ \lambda \in R^n : \sum_{i=1}^{n} \lambda_i = 1, \lambda_i \geq 0, \forall i \in I \right\}.$$

Each $\lambda \in S_n$ represents a distribution of relative social weights of the agents of the economy. It is a measure of the relative weight of each agent in the market representing the social structure of the economy. For each $\lambda \in S_n$, we consider the social utility function: $U_\lambda : \mathscr{F} \to R$ defined by

$$U_\lambda(x) = \sum_{i=1}^{n} \lambda_i u_i(x_i), \tag{3.2}$$

where $x = (x_1, \ldots, x_n)$ is a feasible allocation.

It is well known (see [9]) that a feasible allocation x^* is a Pareto optimal allocation if and only if there exists $\lambda \in S_n$ such that x^* solves

$$\begin{aligned} &max_x \textstyle\sum_{i=1}^{n} \lambda_i u_i(x_i) \\ &\textstyle\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} w_i. \end{aligned} \tag{3.3}$$

For a fixed λ the solution $x^* = x(\lambda)$ of this problem is a Pareto optimal allocation (see also [7]). This suggest a deeper connection between the social structure and the economic efficiency.

3.3 The Negishi Path

For an economy with fixed total resources $W \in R_{++}^l$, the subset of Pareto optimal allocations will be defined by $\mathscr{PO}$. As it is well known, this set does not depend on the preference representation. For each $\lambda \in S_n$, there exists a Pareto optimal allocation $x(\lambda)$ solving (3.3), and for each optimal allocation $\bar{x}$, there exists $\bar{\lambda} \in S_n$ such that $\bar{x}$ solve (3.3) for $\lambda = \bar{\lambda}$, (see [6]).

Definition 3.1. Let $\mathscr{E}$ be an economy with total resources W. The Negishi map $x : S_n \to \mathscr{P}\mathscr{O}$ is defined by $\lambda \to x(\lambda)$, where the allocation $x(\lambda)$ solves the maximization problem (3.3).

Under the hypothesis that utilities are increasing, if all individual have positive endowments, then $\lambda_i = 0$ if and only if $x_i(\lambda) = 0$. In this case $u_i(w_-) = u_i(0)$ and the i-th consumer is out of the market. We consider that λ is a vector S_{n+} in the interior of the simplex. We are interested only in the set of social weights, that correspond to the set of individual rational Pareto optimal allocations. Here, without loss of generality, we consider the subset of social weights $\lambda \in S_{nw} \subset S_{n+}$, Whose associate allocation $x(\lambda)$ verifies the inequalities $u_i(x_i(\lambda) \geq u_i(w_i), i \in \{1, \dots, n\}$.

Theorem 3.1. *The Negishi map is a differentiable function $x : S_n \to \mathscr{P}\mathscr{O}$.*

Proof. The function $U_\lambda : \mathscr{F} \to R$ is a continuous function defined in a compact set. The first order condition for the maximization problem (3.3) is given by

$$\begin{aligned} &\lambda_i \partial u_i(x_i) - \gamma = 0 \quad i = 1, \dots, n, \\ &\textstyle\sum_{i=1}^n x_i - W = 0, \end{aligned} \tag{3.4}$$

where $\gamma \in R^l$ is the Lagrange multiplier, and $\partial u_i(x_i)$ is the gradient of the utility function of the i-agent evaluated at $x_i \in R^l_+$. Define

$$\eta : S_n \times X^n \times R^l \to R^{nl} \times R^l$$

by

$$n(\lambda, x, \gamma) = (\lambda_1 \partial u_1(x_1) - \gamma, \dots, \lambda_n \partial u_n(x_n) - \gamma, \sum_{i=1}^n x_i - W).$$

This is a differentiable function. Let $(\bar{\lambda}, \bar{x}, \bar{\gamma})$ be a solution of $\eta(\lambda, x, \gamma) = 0$. It follows that $\partial \eta_{x,\gamma}(\bar{\lambda}, \bar{x}, \bar{\gamma})$ is a nonsingular $(nl + l) \times (nl + l)$ matrix. Therefore, by the implicit function Theorem, there exists a neighborhood $U_{\bar{\lambda},\bar{x},\bar{\gamma}} = U_\lambda \times U_{\bar{x}} \times U_{\bar{\gamma}}$ of the solution $(\bar{\lambda}, \bar{x}, \bar{\gamma})$ of (3.4) such that there exists a pair of differentiable functions $x : U_\lambda \to U_{\bar{x}}$ and $\gamma : U_\lambda \to U_{\bar{\gamma}}$ with the property that $x(\bar{\lambda}) = \bar{x}$ and $\gamma(\bar{\lambda}) = \bar{\gamma}$, for all $\lambda \in U_{\bar{\lambda}}$, and the following identities are satisfied

$$\begin{aligned} &\lambda_i \partial u_i(x_i(\lambda)) - \gamma(\lambda) = 0 \quad i = 1, \dots, n, \\ &\textstyle\sum_{i=1}^n x_i(\lambda) - W \qquad = 0. \end{aligned} \tag{3.5}$$

The strict concavity of U_λ show that, for each $\lambda \in S_n$, there exists one and only one allocation $x(\lambda)$ solving (3.3). This solution is given by the Negishi map (see also [7]). □

The welfare program (3.3), can be interpreted as the planning model of a government. In this simple setting, the distribution of the social weights are the policy objectives that the planner must change, but this change can be made only indirectly through a social policy of incentives, taxes or, directly, by means of lump sum transference. A policy reform is equivalent to a change of welfare weights. The maximization program (3.3), shows the necessary elements to analyze the repercussions of a social policy in the whole society. This program make evident the existing relationships between distributions of social weights and efficient allocations. These relations are represented in a geometric way by the Negishi path, i.e. the graph of the Negishi map. Consider the set of pairs $(\lambda, x(\lambda)) \in S_n \times \mathscr{PO}$ where the $x(\lambda)$ is the value of the Negishi map evaluated at λ. The set of these pairs, form a differentiable manifold called the Negishi path.

Definition 3.2. The graph of the Negishi map is the set of pairs $(\lambda, x(\lambda))$, for all $\lambda \in S_n$ and it is a differentiable manifold in $S_n \times R^{nl}$. This manifold will be called the Negishi path (or the Negishi manifold) and will be defined by C_N.

This map does not depend on the distributions of the initial endowments, but only in the total resources W of the economy. This means that all economies, with the same utilities and total resources have the same Negishi map. Consider that the society is represented by the pair $(\lambda, x(\lambda)) \in C_N$ corresponding to a given distribution of the social weights and associate allocation of resources. This is a highly stylized representation, but has all the ingredients to analyze the repercussions of the economics reforms in the society. Suppose that the condition $(\lambda, x(\lambda)) \in C_N$ is satisfied.

If a reform policy in an efficient economy is followed by changes in the social weights, then after the reforms, some consumers will gain, and others will lose. It is not possible for all consumer to gain since the pre-reform situation corresponds to a point in the Negishi map. Furthermore, who gains and who loses, is given by modifications in the social weights. Pre-reformers look for efficiency and equity, both objectives are possible to attain if and only if the initial situation is such that $(\lambda, x(\lambda)) \notin C_N$. Let us consider the function $\mathscr{U} : S_n \to R$ defined by

$$\mathscr{U}(\lambda, x(\lambda)) = \sum_{i=1}^{n} \lambda_i u_i(x_i(\lambda)).$$

The number $\mathscr{U}(\lambda, x(\lambda))$ represents the social value of the efficient allocation $x(\lambda)$. It is possible to assign to each $x \in \mathscr{PO}$ a social value and a distribution $\lambda \in S_n$ of social weights. However, the Pareto criterium only checks whether a consumer gains or losses in terms of utilities, but changes in utilities do not give a cardinal measure of the size of welfare gains and losses for a given consumer or in terms of the social welfare.

3.4 The Negishi Index

Let us considerer the distribution $\lambda^* \in S_n$ corresponding to the solution of the following problem

$$\min_{\lambda \in S_n} \mathscr{U}(\lambda, x(\lambda)). \tag{3.6}$$

Let $x(\lambda^*)$ be the allocation solving the problem of maximization

$$\max_{x \in \mathscr{F}} \sum_{i=1}^{n} \lambda_i^* u_i(x_i). \tag{3.7}$$

The solution of (3.6) exists, because the objective function is a continuous and convex funcion and S_n is a compact set. The existence of the solution for the problem (3.7) is a consequence of the hypothesis on the utility functions and the compactness of $\mathscr{F}$.

Definition 3.3. The value $\mathscr{U}(\lambda^*, x(\lambda^*))$, where λ^* is the solution of (3.6) is the Negishi index.

The following Theorem summarizes this topic.

Theorem 3.2. *Given an economy $\mathscr{E}$, under the hypothesis considered in this work, there exists a pair $(\lambda^*, x(\lambda^*)) \in C_N$, where the Negish index is reached.*

Proof. The Theorem follows from the fact that $\mathscr{U} : S_n \to R$ is a continuous and convex function (see [3]). □

The Negishi index $\mathscr{N}$ corresponds to a pair $(\lambda^*, x(\lambda^*))$ in the Negishi path, and is the same for all economies with the same utilities and total resources, i.e. does no depend on the distribution of the initial resources. Unfortunately, the Negishi index depends on the utility, but the social weight λ^* and the allocation $X(\lambda^*)$ corresponding to this index, for a fair economy, do no depend on the representation of the preferences.

Remark 3.1. The main characteristics of the allocation $x(\lambda^*) = x^*$ are the following ones:

1. It is an efficient allocation maximizing the social utility function U_{λ^*}, for all $x \in \mathscr{F}$, where λ^* is the solution of the problem (3.6).
2. Every agent reach the same level of utility, i.e. $u_i(x_i^*) = u_j(x_j^*)$, for all $i, j = 1, 2, \ldots n$, (see [4]).
3. The utility level of each agent $u_i(x_i^*)$ $i \in 1, 2, \ldots n$ is the same for the social utility level given by $\mathscr{U}(\lambda^*, x(\lambda^*))$.
4. These characteristics of the referred resource allocation are satisfied independently of the representation of the preferences.

We will introduce the concept of democratic distribution of social weights.

Definition 3.4. We will say that λ^* is a *democratic distribution of social weights*, if the associate allocation $x(\lambda^*)$ has the above mentioned four characteristics.

For each economy there exist a *democratic* distribution of social weights. The next question is about if such distribution can be reached in a decentralized way, i.e. without the participation of a central planner.

The point $(\lambda^*, x(\lambda^*)) \in C_N$ where the Negishi index is reached, depends strongly on utilities representing the preferences, but the existence of such point does not depend on this representation. We will focus our attention in the allocation $x(\lambda^*)$ because we understand that its equalitarian properties give to the economy some kind of fairness and stability, in the sense that we will explain in the next sections.

Remark 3.2. Consider $\lambda \in S_n$, then the equality $\lambda_n = 1 - (\sum_{i=1}^{n-1} \lambda_i)$ follows. If $\lambda_i > 0$, for all $i \in \{1, \ldots, n\}$, utilities are strictly concave functions, the Hessian of $\mathcal{U}(\lambda, x(\lambda)$ is a definite negative matrix $H_\lambda \mathcal{U}$ with $n-1$ rows and columns. Since

$$\frac{\partial \mathcal{U}}{\partial \lambda_i}(\lambda, x(\lambda)) = u_i(\lambda, x_i(\lambda)) - u_n(\lambda, x_n(\lambda)),$$

and the Hessian has positive diagonal, for all $i \in \{1, \ldots, n-1\}$ the inequalities

$$\frac{\partial u_i(\lambda, x_i(\lambda))}{\partial \lambda_i} - \frac{\partial u_n(\lambda, x_n(\lambda))}{\partial \lambda_i} > 0 \tag{3.8}$$

hold. Hence,

$$u_i(\lambda, x_i(\lambda)) - u_i(\lambda, x_n(\lambda)) \text{ is increasing with } \lambda_i.$$

This assertion follows from (3.8) and the chain of equalities

$$u_1(\lambda^*, x_1(\lambda^*)) = \ldots = u_n(\lambda^*, x_n(\lambda^*))$$

see remark (3.1,(2)). The above statement means that the inequality

$$\mathcal{U}(\lambda, x(\lambda)) \geq \mathcal{U}(\lambda^*, x(\lambda^*)) \;\; \forall\, \lambda \in S_{n+},$$

does no necessarily imply

$$u_i(\lambda, x_i(\lambda)) > u_i(\lambda^*, x_i(\lambda^*)).$$

To see this note that

$$\begin{aligned}\mathcal{U}(\lambda, x(\lambda)) - \mathcal{U}(\lambda^*, x(\lambda^*)) = \sum_{i=1}^{n} \{\lambda_i \left[u_i(\lambda, x_i(\lambda)) - u_i(\lambda^*, x_n(\lambda^*))\right] \\ + u_n(\lambda^*, x_n(\lambda^*)) \left[\lambda_i - \lambda_i^*\right]\} \geq 0. \end{aligned} \tag{3.9}$$

Hence, from Remark 3.2, it follows that

$$u_i(\lambda, x_i(\lambda)) > u_i(\lambda^*, x_i(\lambda^*)) \Leftrightarrow \lambda_i > \lambda_i^*. \tag{3.10}$$

3.5 Social Equilibria

Reforms following from changes in social weights can be implement in a decentralized way if and only if the after-reforms distribution of social weights are in a particular set, that we will call the social equilibrium set. We introduce in this section, the excess utility function, to characterize the set of vectors $\lambda \in S_n$ and its corresponding allocations $x(\lambda)$ such that can be reached in a decentralized way for an economy $\mathscr{E}$.

For a given economy $\mathscr{E}$, the excess utility function $e_w : S_{n+} \to R^n$ with $e_w(\lambda) = (e_{w1}(\lambda), \dots, e_{wn}(\lambda))$ is defined by

$$e_{wi}(\lambda) = \frac{\partial}{\partial x_i} u_i(x_i(\lambda)) \, [x_i(\lambda) - w_i] \, , \text{for all } i = 1, \dots, n, \tag{3.11}$$

where $x(\lambda)$ is the value of Negishi map evaluated at λ. We introduce the subindex w in the notation $e_w(\lambda)$ to remark that this function depends strongly on the distribution of the initial resources w. The main characteristics of the excess utility function are referred in [1].

We say that a pair $(\lambda, x(\lambda)) \in C_N$ is a social equilibrium for an economy $\mathscr{E}$ if and only if $\lambda \in S_{n+}$ and $e_w(\lambda) = 0$. We denote this subset by

$$\mathscr{S}\mathscr{E}_w = \{(\lambda, x(\lambda)) \in C_N : e_w(\lambda) = 0\} \, .$$

We introduce the following notation to identify the pre-image of zero by the excess utility function e_w:

$$\mathscr{E}\mathscr{Q}_w = \{\lambda \in S_{n+} : e_w(\lambda) = 0.\} \, .$$

The following theorem shows the relationships between the set of social equilibrium $\mathscr{S}\mathscr{E}_w$ and the set of Walrasian equilibria $\mathscr{W}\mathscr{E}$ in a given economy $\mathscr{E}$.

Theorem 3.3. *For every $(\lambda, x(\lambda)) \in \mathscr{S}\mathscr{E}_w$ there exists a vector $p \in R^l$ such that $(p, x(\lambda))$ is a Walrasian equilibrium. Reciprocally, for each Walrasian equilibrium (p, x), there exists $\lambda \in \mathscr{E}\mathscr{Q}_w$ such that $(\lambda, x) \in \mathscr{S}\mathscr{E}_w$.*

Proof. Let $(\bar{\lambda}, x(\bar{\lambda}))$ be a social equilibrium. Consider $\bar{p} = \bar{\lambda}_i \partial u_i(x_i(\bar{\lambda}))$. It follows that $(\bar{p}, x(\bar{\lambda}))$ satisfy preference maximization, under budget constraint and attainability, so it is a Walrasian equilibrium. Reciprocally, if $(\bar{p}, \bar{x})$ is a Walrasian equilibrium then, solving the first order equations, for the problem (3.3) to maximize $U_\lambda(x), \ x \in \mathscr{F}$ evaluated at $x = \bar{x}$, and taking $\gamma = \bar{p}$ the following equalities hold:

$$\lambda_i \partial u_i(\bar{x}_i) - \bar{p} = 0, \text{for all } i = 1, \dots, n.$$

We obtain the corresponding vector $\bar{\lambda} \in \mathscr{E}\mathscr{Q}_w$ satisfying $(\bar{\lambda}, \bar{x}) \in \mathscr{E}\mathscr{S}_w$. □

In contrast with the set of Pareto optimal allocations, the subset of allocations corresponding to the social equilibria depends strongly on the distribution of the initial resources and not, only, on the total resources of the economy. Hence, an economy can reach in a decentralized way the democratic distribution of social weights, λ^* and its corresponding allocation $x(\lambda^*)$, if and only if, the equalities $e_w(\lambda^*) = 0$ are satisfied. This value is reachable only if the central planer is able to implement a policy reform whose consequence is a redistribution of resources such that if w' is the after reform distribution of the initial resources then, $w' \neq w$, $\sum_{i=1}^{n} w'_i = \sum_{i=1}^{n} w_i = W$, and $e_{w'}(\lambda^*) = 0$. It follows that if the consequence of a political reform is a redistribution of initial endowments, then it follows a change in the set of the social equilibria of the economy. Thus, the social expression of a policy reform, implying a redistribution of the initial endowments, is a change in the relative social weights of the agents.

Let $\mathscr{E}$ be a pure interchange economy. Consider a sufficient small real number $\epsilon > 0$ and let $S_{n\epsilon}$ be the set of all the distributions on the n agents of the economy such that the relative weight of each agent λ_i is at least ϵ and such that $S_{nw} \subset S_{n\epsilon}$ where

$$S_{n\epsilon} = \{\lambda \in S_n : 1 - \epsilon \geq \lambda_i \geq \epsilon > 0, \ \forall i \in I\}.$$

Let $\mathscr{P}\mathscr{O}_\epsilon$ be the subset of the Pareto optimal allocations $\mathscr{P}\mathscr{O}$ that can be reached by the Negishi map restricted to $S_{n\epsilon}$. Hence, the image by the Negishi map of this subset is the subset $\mathscr{P}\mathscr{O}_\epsilon$, i.e.

$$x[S_{n\epsilon}] = \mathscr{P}\mathscr{O}_\epsilon \subset \mathscr{P}\mathscr{O}.$$

Let $C_{N\epsilon}$ be the subset of C_N with $\lambda \in S_{n\epsilon}$. The individual rational Pareto optimal allocations is a subset of $\mathscr{P}\mathscr{O}_\epsilon$.

We do not consider agents with initial endowments equal to zero, because in this case the agent is out of the market. The following theorem holds.

Theorem 3.4. *If ϵ is sufficiently small then the set of social equilibrium $\mathscr{E}\mathscr{Q}_w$ are a subset of $C_{N\epsilon}$.*

Proof. The theorem follows from the fact that we assign to a given consumer a social relative weight equal to zero. For instance, for the h-consumer. Let $\lambda_h = 0$. The corresponding coordinate in the Negishi map is equal to zero, i.e. $x_h(\lambda) = 0$. Then the corresponding allocation $x(\lambda)$ is not an equilibrium, because the h-agent prefers his own endowments and we assume that $w_h \neq 0$. □

However, the Negishi index might not be reachable in equilibrium for a given economy. This possibility depends on the characteristics of the endowments of the economy. If the endowments do not entail these appropriate characteristics, then the possibility to obtain, in a decentralized way, an efficient allocation assuring the same level of happiness for every agent of the economy, depends on the redistributions of the endowments. If the aim of a central planner is to obtain some kind of social equality, he can be interested in realizing a transference of resources, but transference imply changes in the distribution of the social weights of the individuals, i.e.

in the social structure of the economy. The next question is how to predict these changes.

3.6 Singular and Regular Economies from a Social Point of View

In this section, we analyze the main characteristics of the changes in the social structure as response to redistributions of the endowments. We introduce the concepts of singular and regular economies from a social point of view.

Definition 3.5. The economy $\mathscr{E}$ is singular from a social point of view, if zero is a singular value of the excess utility function. Otherwise, the economy will be called regular.

The following two properties are satisfied by the excess utility function (see [1]):

1. The *social Walras law*: $\lambda e(\lambda) = 0$ for all $\lambda \in S_n$.
2. The Jacobian matrix of the excess utility function $[Je_w]$ is a linear transformation from S_n into R^{n-1}. For every $\lambda \in S_n$, the dimension of the image of this matrix is, at most, $n-1$ i.e. $dim[Je_w](\lambda) \leq n-1$ for all $\lambda \in S_n$.

The economy $\mathscr{E}$ is:

- Regular if and only if the dimension of the jacobian of the excess utility function evaluated in each $\lambda \in \mathscr{E}\mathscr{Q}_w$, is $n-1$, i.e.

$$dim[Je_w](\lambda) = n-1 \ \forall \lambda \in \mathscr{E}\mathscr{Q}_w.$$

- Singular if and only if there exist at least one $\bar{\lambda} \in \mathscr{E}\mathscr{Q}_w$, such that

$$dim[Je_w](\bar{\lambda}) < n-1.$$

3.7 Main Characteristics of the Social Changes

Consider the set $\mathscr{E}\mathscr{W}$ of all economies with n-agents, consumption spaces $X = R^l_+$, utilities u_i and with total resources fixed $W \in R^l_+$. The economy, $\mathscr{E} \in \mathscr{E}\mathscr{W} = \left\{R^l_+, u_i, W, I\right\}$ if and only its endowments $w = (w_1, \ldots, w_n)$ satisfy the equality $\sum_{i=1}^n w_i = W$.

We will denote by $\mathscr{E}_w$ the elements of $\mathscr{E}\mathscr{W}$. Let $\Omega = \left\{w \in R^{ln} \colon \sum_{i=1}^n w_i = W\right\}$.

The generalized excess demand function $E : S_n \times \Omega \to R^{nl}$, is defined by

$$E_i(\lambda, w_i) = \lambda_i \partial u_i(x_i(\lambda))[x_i(\lambda) - w_i],$$

where $x_i(\lambda)$ represents the bundle set of the i-agent in the Negishi map $x(\lambda)$.

The following Theorem follows as an application of the transversality Theorem to the generalized excess utility function.

Theorem 3.5. *There exists an open and dense subset $\Omega_0 \subset \Omega$ such that an economy $\mathscr{E}_{w'}$ is a regular economy if and only if $w' \in \Omega_0$.*

The keystone to understand the social changes is the set

$$\mathscr{S}\mathscr{E}_W = \{(\lambda, w) \in S_{n+} \times \Omega : E(\lambda, w) = 0\} .$$

Proof. To prove the Theorem 3.5 we introduce the following tools: Let $S_{n+} = \{\lambda \in S_n : \lambda_i > 0\ \forall i\}$ be the set of the relative interior of the manifold S_n. Let $\bar{E} = (E_1, \ldots, E_{n-1})$ be the restricted excess utility function. By the social Walras law, $(\lambda, w) \in E^{-1}(0)$ if and only if $(\lambda, w) \in \bar{E}^{-1}(0)$. Consider $\bar{E} : S_{n+} \times \Omega \to R^{n-1}$. For each $w \in \Omega$ we write $E(w, \cdot) = E_w(\cdot)$. Hence, for each $\lambda \in S_n$ the identity $E_w(\lambda) = (e_{1w}(\lambda), \ldots, e_{nw}(\lambda))$ holds. It follows that for each $w \in \Omega$, $\bar{E}_w : S_{n+} \to R^{n-1}$ is transversal to R^{n-1}. From the transversal Theorem it follows that $\bar{E}$ is transversal for almost every $w \in \Omega$. Hence, 0 is a regular value for $\bar{E}$ for each $w \in \Omega_0$, where Ω_0 is an open and dense subset of Ω. □

The proof of the next corollary is straightforward from the transversality Theorem and shows that in the residual set Ω_0, locally, the set $\mathscr{S}\mathscr{E}_W$ behaves like the space R^{nl}.

Corollary 3.1. *There exists an open and dense subset $\Omega_0 \subseteq \Omega$ such that*

$$\mathscr{S}\mathscr{E}_{W/\Omega_0} = \{(\lambda, w) \in S_{n+} \times \Omega_0 : E(\lambda, w) = 0\}$$

is a manifold of dimension nl embedded in $S_n \times \Omega$. Where $S_{n+} = \{\lambda \in S_n : \lambda_i > 0\ \forall i\}$ is the relative interior of S_n.

Corollary 3.1 has enumerous implications. One of the most important enables one to express the equilibrium distributions of social weights associated with a regular economy as a function of the parameter w. Thus, there exist neighborhoods $V_{w'} \in R^{nl} \subset \Omega_0$, $T_\lambda \in R^{n-1} \subset S_{n+}$ and a function $\Lambda : V_{w'} \to T_{\lambda'}$ such that $\Lambda(w') = \lambda'$ and $E(\Lambda(w), w) = 0$, for all $w \in V_{w'}$.

If the economy $\mathscr{E}_{w'}$ is regular, a redistribution of its endowments give a new economy $\mathscr{E}_{w''}$. If $w'' \in V_{w'} \cap \Omega$ then this new economy is regular and the corresponding new set of social equilibrium $\mathscr{S}\mathscr{E}_{w''}$ is similar to the set $\mathscr{S}\mathscr{E}_{w'}$, in the sense that the respective cardinalities will be the same. If $(\lambda'', x(\lambda'')) \in \mathscr{S}\mathscr{E}_{w''}$ then $\lambda'' \in T_{\lambda'}$ and $|x(\lambda') - x(\lambda'')| < \epsilon$, for $\epsilon > 0$ small enough. From a social point of view, if the endowments w' of a regular economy are redistributed and if this redistribution $w''' \in V_{w'}$, then the respective relative social weights do not change to much by continuity of $\Lambda : W \to S_n$ at w'. The sets $\mathscr{E}\mathscr{Q}_w$ and $\mathscr{E}\mathscr{Q}_{w''}$ do not change to much after the redistribution of the endowments (see [2]).

However, if the economy $\mathcal{E}_{w'}$ is singular then small changes in the endowments give rise to big changes in the social structure. This means that a small redistribution of endowments w' giving place to a new distribution w'', with $|w' - w''| < \epsilon$, is followed by a strong social change, because the distributions of social weights corresponding to the social equilibria of the modified economy can be quite different from the distributions corresponding to the equilibria set $\mathcal{E}\mathcal{Q}_{w'}$ of the original economy. In this case, we say that a change in the endowments of the economy is followed by a social crisis, i.e. an unforeseen and big change in the social participation of the agents of the economy in the social welfare.

3.8 Fair and Unfair Economies

There is not a definition of fair economy in the framework of the General Equilibrium Theory. In this section, we introduce a preliminary (and intuitive) definition of fair economy in this framework.

Definition 3.6. The economy $\mathcal{E}$ is a fair economy, if it is a social regular economy and if the corresponding Negishi index can be reached in equilibrium (i.e. in a decentralized way).

In this case the democratic distribution of social weights λ^* and the corresponding efficient allocation $x(\lambda^*)$ do not depend upon the representation of the preferences.

This definition of fair economy put the emphasis in the distribution of the resources of the economy. Given two economies with the same total resources and same utilities, one of then can be a fair economy and the other no, and this depends exclusively on the distribution of resources. The unfair economies are those whose distribution of resources do not allow reaching an equilibrium, a democratic social structure, i.e. a resource allocation that in equilibrium guarantees the same level of utility for each agent of the economy. In a fair economy there exists the possibility to reach in a decentralized way the pair $(\lambda^*, x(\lambda^*))$, where the level of social welfare reached is the same for all its agents. This means that the distribution of the initial resources w of a fair economy satisfies $e_w(\lambda^*) = 0$.

For a no fair economy $e_w(\lambda^*) = 0$ is not satisfied. A no fair economy can reach its corresponding democratic distribution of social weights, only after a redistribution of the initial endowments. If this economy is a singular one, then large and unexpected social changes can occur in the process of the redistribution. The singular economies are the doors to the social crisis.

The main characteristics of a fair economy are efficiency, in the sense of Pareto, stability and justice in the sense of a democratic distribution of resources can be spontaneously reached, implying the possibility to obtain an equal level of utility for all individuals at equilibrium.

Let $\mathcal{N}$ be the Negishi corresponding to the set of economies $\mathcal{E}\mathcal{W}$. Given a no fair economy $\mathcal{E}_w \subset \mathcal{E}\mathcal{W}$, the value

$$\min_{\lambda \in \mathscr{E}\mathscr{Q}_w} [U(\lambda, x(\lambda)) - \mathscr{N}]$$

allows to define a subjective degree of unfairness of the economy. If this number is equal to zero, then the economy is a fair one and the degree of unfairness increases with this number.

An objective index to measure the degree of unfairness of an economy is given by

$$\min_{\lambda \in \mathscr{E}\mathscr{Q}_w} ||\lambda - \lambda^*||,$$

where λ^* corresponds to the democratic distribution of the social weights, in the set of economies with total resources W. This is a measure of how far a given economy is from the fair economy.

3.9 Conclusions

The Negishi approach allows us to unify two different points of view of the *neo-Walrasian* economy, the point of view of the efficiency and the point of view of the social welfare. The main tool for offering this unified version is the Negishi map, this map characterizes the efficiency and its relationship with the social structure and the possible levels of welfare that an economy can reach.

The allocation $x(\lambda^*)$ corresponding to the solution of the program (3.6) implies that every agent of the economy reach the same level of happiness, this is a form of *no envy*. The solution λ^* corresponds to a distribution of social weights *fair or democratic*. The social level corresponding to this distribution of social weights is the maximum attainable for the economy. The possibility to obtain this level of social welfare for a no fair economy depends on redistributions, but this is a problem for a central planner of a singular economy. In some cases, a redistribution of the initial resources can lead to worse situations than those that are tried to surpass. In particular, in the case of singular economies where redistributions imply big and unforeseen changes in the social welfare. This does not means that the central planner does not try changes to solve problems of the economy, but he needs to know that changes in the fundamentals of the economy can imply undesirable and non reversible changes in the social structure of the economy. He needs to be extremely carefully in the case of economies close to be a singular one.

As a final remark, we observe that the Negishi method can be generalized to infinite dimensional economies. So, using this method, we can analyze, in an unified way, both the finite and infinite dimensional economies (see [1]).

Acknowledgements The authors gratefully acknowledge the support of SEP-CONACYT: Projects 46209 and 82610 and the support of UASLP Secretaría de Posgrado. The authors wish to thank CONACYT for the support provided through the project "82610 de Ciencia Básica".

References

1. Accinelli, E.: Some remarks about uniqueness of the equilibrium for infinite dimensional economies. Estudios de Economía **21**, 315–326 (1994)
2. Accinelli, E., Puchet, M.: Could catastrophe theory become a new tool in understanding singular economies? In: Leskow, J., Puchet, M., Punzo, L. (eds.) New tools of Economic Dynamics, Serie: Lecture Notes in Economics and Mathematical Systems. Springer, vol. 551, Chap. 8 (2005)
3. Accinelli, E., Brida, G., Plata, L., Puchet, M.: Bienestar social, óptimos de Pareto y equilibrios Walrasianos. El Trimestre Económico. vol. LXXV, pp. 97–134 (2008)
4. Accinelli, E., Plata, L., Puchet, M.: The Fenchel duality theorem and the Negishi approach. Brazil. J. Bus. Econ. vol. 7/1, pp. 43–46 (2007)
5. Golubistki, M., Guillemin,V.: Stable Mappings and Their Singularities. Springer (1973)
6. Kehoe, T.J.: Computacion and multiplicity of equilibria. In: Hildenbrand, W., Sonneenschein, H. (eds.) Handbook in Mathematical Economics, vol. IV.. North-Holland, Amsterdam (1991)
7. Mas-Colell, A.: The Theory of General Economic Equilibrium: A Differentiable Approach. Cambridge University Press (1975)
8. Mas-Colell, A., Zame, W.: Equilibrium theory in infinite dimensional economies. In: Hildenbrand, W., Sonneenschein, H. (eds.) Handbook in Mathematical Economics, vol. 4. North-Holland, Amsterdam (1991)
9. Negishi, T.: Welfare economics and existence of an equilibrium for a competitive economy. Metroeconomica **12**, 92–97 (1960)
10. Serra, P.: The excess utility functions and the welfare adjustment process. Econ. Lett. **26**, 1–5 (1988)

Chapter 4
Singularities, Walrasian Economies and Economic Crisis

Elvio Accinelli and Martín Puchet

Abstract We consider pure exchange economies whose consumption spaces are Banach lattices. The utility functions are strictly concave, Gateaux differentiable, and not necessarily separable. Following the Negishi approach and using the excess utility function, we introduce a notion of social equilibria. We show that there exists a bijective correspondence between this set and the set of Walrasian equilibria. By transforming the problem of finding the Walrasian equilibria into an equivalent problem of finding social equilibria, we can use techniques of smooth functional analysis to show that a suitable large subset of economies are regular and its equilibrium set is a Banach manifold. Finally, we focus on the complement of this set, i.e. the set of singular economies, and we analyze its main characteristics, among them, those that are the causes of economic crises.

4.1 Introduction

The main contribution of this paper is to show that the economic crisis can be considered as the result of small perturbations in the fundamentals of a particularly small set of economies, i.e. the singular economies. Analyzing the main characteristics of the singular economies, we will know more on the economic crises and their consequences. Our approach follows the point of view of René Thom who in the 1960s introduce the catastrophe theory as a tool to understand why discontinuities in the behavior of a system can occur, even, in continuous frameworks.

Specifically from the point of the Economic Theory our paper takes as its starting point the Negishi's method. The framework of this work is the Negishi approach. It is important to remark that this approach can be used to analyze the main

E. Accinelli (✉)
Facultad de Economía, UASLP, Av. Pintores S/N, Fraccionamiento Burocratas del Estado, CP. 78263 San Luis Potosí, México
e-mail: elvio.accinell@eco.uaslp.mx

M. Puchet
Posgrados en Economía UNAM, Ciudad Universitaria, C.P. 04510, México
e-mail: anyul@servidor.unam.mx

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_4,

characteristics of theequilibrium set of the economies with a finite set of commodities or to analize economies with infinitely many goods. The main characteristics of this analysis is the same in both cases. Following the Negishi approach, we generalize some of the main results obtained in [1]. As it is well known, much of the financial models are supported on the general equilibrium model, where the assets are represented by square integrable functions in a given measure space (see [20]). Our point of view adds new relevance to the analysis of the infinite dimensional models under the perspective of an analysis of the crisis.

It is well known that the demand function is a tool to deal with the equilibrium manifold in economies such that consumption spaces are subsets of a Hilbert space, in particular, R^l (see [24]). But, if the consumption spaces are subsets of infinite dimensional spaces (not Hilbert spaces), the demand function may not be differentiable (see [7]) or it is not well defined because the price space is very large, or the positive cone where prices are defined has empty interior. Recall that prices are elements of the dual space of the space where the economy is defined. In many cases, following the Negishi approach, it is possible to characterize the equilibria set using the excess utility function. Then, it is possible to introduce differentiable techniques that allow to generalize the results obtained by [16] for smooth infinite dimensional economies to the case with no separable utilities. Since the Negishi approach depends strongly on the existence of Pareto optimal allocations, we do a brief discussion on this topic in Sect. 4.3.

Using methods of singularity theory, we show that the equilibrium set (strictly speaking the social equilibrium set) of an open and dense (residual) subset of economies is a Banach manifold. Our result generalizes the previous one given in [13]. To obtain this result, we assume that the positive cone Ω_+ of the consumption space has non-empty interior. Typically, examples of such spaces are $L^\infty(M, R^n)$ where M is a compact manifold (see for instance [16]). Nevertheless, the set of regular economies is residual (dense), even in some cases where positive cones have empty interior (see [25]). The set of regular economies is large and its complement is a rare set. This claim, in the infinite dimensional framework, is not a consequence of the Debreu theorems. Here, it follows from an alternative approach with particular interest in infinite dimensional cases.

Despite the smallness of the singular economies set (relative to the set of regular economies), it plays a central role to characterize the changes in economics as a response to changes in its fundamentals. In particular, the economic crisis and its social repercussions. In Sect. 4.7, we will focus on this set, i.e. the complement of the set of regular economies that is a rare set. The smallness (in the set of the economies) holds, not only from this topological point of view, but also, if there is a measure in the set of the economies, from a measurable point of view being a set of zero measure.

However, this small set is the origin and also explains can be responsible for the big and unpredictable changes in economics.[1] In singularity theory, one attempts

[1] Note that the only way to consider the existence of the economic crisis in the General Equilibrium Theory is to introduce singularities, because changes of a regular economy as a results of changes in its fundamentals are predictable and smooth.

to classify the possible singularities by producing normal forms. The knowledge of normal forms provide mathematical models of natural or social phenomena. Our main concern in this work is to obtain the normal forms of the singular economies characterizing the economic crisis. We succeed in finding such normal forms only in some cases.

All regular economies have, locally, the same behavior. This means that in a neighborhood of a regular economy there are no big changes and all economy in this neighborhood is also a regular economy. If the economy is regular, small changes in the distribution of the endowments do not imply big changes in the behavior of the economy as a system, and the new economy will be also a regular economy (this means that regular economies are structurally stable). In contrast, singular economies, in contrast with regular ones, characterize the sudden qualitative and unforeseen changes in the economy. In a neighborhood of a singular economy, small changes in the distribution of the endowments imply big changes in the main characteristics of the economy. Hence, the analysis of these economies is the framework to obtain a rigorous theory of the economic crisis.

An economy is singular if zero is a singular value of its excess utility function. Given that utilities appear explicitly in the excess utility function, the strong relationship existing between the characteristics of the agent's preferences, and the behavior of the economy is reflected in this function. The multiplicity of equilibria is a consequence of the existence of the singular economies and this is the clue to understand the importance of the singularities to characterize the crisis. Regular economies with a given number of equilibria are open subsets in the topological space of the economies. The set of regular economies with only one equilibrium and the set of economies with multiple equilibria are separate by singular economies. Hence, singular economies are the boundaries of regular economies with different number of equilibria.

We obtain normal forms for the excess utility functions of the singular economies. We introduce in the set of the singular economies, a partition in equivalence classes, according with the type of normal forms. Roughly speaking, the excess utility functions of economies with the same kind of singularities have the same normal form. The regular economies that a singular economy can adopt after a perturbation in its fundamentals occur are the same for economies with the same normal form. Nevertheless, we can not predict the particular regular form that the economy will adopt after the perturbation occurs from the pre-set of regular economies determined by the normal form.

Singular economies are very sensitive to political and social choices. These possible reactions can be classified and can be understood using normal forms.

The importance of critical equilibria and singular economy has been realized since 1970s with the application of differential topology to general equilibrium models (see the seminal paper [19] and the works of Y. Balasko (see [12, 13]) and Mas-Collel (see [24])).

4.2 The Model

We consider an economy where each agent's consumption set is a subset of a Banach lattice X. The agents will be indexed by $i \in I = \{1, 2, ...n\}$. We denote by X_+ the positive cone of the Banach lattice X. We do not assume separability in the utility functions $u_i : X_+ \to R$. We denote by X_+ the positive cone of X. We denote by X_{++} the set of the elements in X_+ strictly positive. We say that the vector $h \in X$ is *admissible* for $x \in X_+$ if and only if $x + h \in X_+$. Given a point $x \in X_+$ the set of admissible vectors is given by the set $\mathscr{A}_x = \{h \in X : x + h \in X_+\}$. Suppose that $x, y \in X_+$ and $h = \alpha(x-y)$, the vector h is admissible because $w = x+\alpha h \in X_+$, for every $0 \leq \alpha \leq 1$. This property characterize the convex sets. We assume that the utility functions are in $C^2(X_+, R)$, i.e. the set of the functions with continuous second Gateaux derivatives (G-derivatives) in each admissible direction. We assume that they are increasing functions, i.e. each agent prefers to gain more than less. More formally for every $x \in X_+$ the first order G-derivative is positive, where G-derivative is defined in the usual way, i.e.

$$f(x + h) = f(x) + f'(x)h + o(\|h\|),$$

for all $x \in X_+$ and for all admissible h. If the G-derivative of $f : X \to Y$ exists for $x \in X_+$ then $f' : A \subset X \to L(X, Y)$, where $L(X, Y)$ is the set of linear transformations from the Banach space X into de Banach space Y. Let X^* be the dual space of X, i.e. the set of continuous linear transformations (functionals) from X to R. Hence, $u_i'(x) \in X^*$, for every $x \in X_+$ and $i \in I$. The second G-derivative $f''(x)$ exists if and only if the iterated derivative $(f')'(x)$ exists. In this case, we get

$$f''(x)hk = (f')'(x)(h)(k) \ \ \forall h, k \in \mathscr{A}.$$

Similarly, for high order G-derivatives.

In addition, we assume that for all $x \in X_+$ the inverse operator $(u_i^{''})^{-1}(x)$ of the hessian operator $u_i''(x) : X \to X^*$ exists, $(u_i''(x)$ is the bi-linear form $(h, k) \to u''(x)hk)$. Hence, for each $x \in X$ we have

$$u_i''(x)(u_i'')^{-1}(x)h = h \ \ \forall \ h \in X.$$

The consumption set of each agent is the positive cone of the lattice X (the same for each agent). Let $\Omega_+ = (X_+)^n$ be the cartesian product of these n consumption sets. A bundle set for the i-agent is a point $x_i \in X_+$ and an allocation is a vector $x = (x_1, x_2, \ldots, x_n) \in \Omega_+$. The endowments of the i-agent will be denoted by w_i, and $w = (w_1, w_2, ...w_n)$ denotes the initial endowments. We assume that $w \in \Omega_{++}$, where Ω_{++} is the strictly positive cone of Ω. The total amounts of available goods is denoted by $W = \sum_{i=1}^{n} w_i$.

In order to obtain a strictly positive allocation of equilibria, we will assume that the utility functions satisfy at least one of the following two conditions (widely used assumptions in economics [5]).

(a) (Inada condition) $\lim \|u_i'(x_i)\| = \infty, if \ x_i \rightarrow \partial(X_+)$, for every $i = 1, 2, \ldots, n$, and for every utility function, ($\partial(X_+)$ denotes the boundary of the positive cone X). Furthermore, we assume that the marginal utility is infinite for consumption at zero.

(b) All strictly positive allocations are preferable to allocations in the the boundary of Ω_{++}.

The economies are denoted by

$$\mathscr{E} = \{X, u_i, w_i, I\},$$

where X is the consumption set (in our case X_+), u_i is the utility function of the i-agent is, w_i represents the initial endowments of the i-agent, and I the index set of agents. We assume that $I = \{1, 2, \ldots, n\}$ is a finite set. Examples of such economies have the consumption set $X_+ = C_{++}(M, R^n)$ and utility functions $u_i(x) = \int_M u_i(x(t), t)dt$ (see [16] and [5]).

4.3 The Negishi Approach

The Negishi approach is a powerful mathematical way to analyze the behavior of the equilibrium set.

Let Δ denote the social weight set, or distribution function,

$$\Delta = \left\{ \lambda \in R^n : \sum_{i=1}^{n} \lambda_i = 1, \ \ 0 \leq \lambda_i \leq 1 \ \forall i \right\},$$

and $\Delta_+ = int[\Delta]$ denotes the set of $\lambda \in \Delta$, such that $\lambda_i > 0, \forall i \in I$. Every λ_i represents the social weight of the agent in the market, and Ω_+ is the positive cone of the consumption space $\Omega = X^n$. The Negishi approach considers a social welfare function $U_\lambda : \Omega_+ \rightarrow R$ defined by

$$U_\lambda(x) = \sum_{i=1}^{n} \lambda_i u_i(x_i), \tag{4.1}$$

where u_i is the utility function of the agent i, $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_n) \in int[\Delta]$. As it is well known, if $x^* \in \Omega_+$ solves the maximization problem of $U_{\lambda^*}(x)$ for a given λ^*, subject to be a feasible allocation i.e.

$$x^* \in \mathscr{F} = \left\{ x \in \Omega_+ : \sum_{i=1}^{n} x_i \leq \sum_{i=1}^{n} w_i \right\}$$

then x^* is a Pareto optimal allocation. Reciprocally, if a feasible allocation x^* is Pareto optimal, then there exists $\lambda^* \in \Delta$ such that x^* maximizes U_{λ^*} (see [4]).

Without loss of generality, we will consider that $\lambda \in \Delta_+$. Moreover, we are interested in individual rational Pareto optimal allocations, i.e. Pareto optimal allocations such that $u_i(x_i) \geq u_i(w_i)$ for all $i \in I$. If x is an individual rational Pareto optimal allocation, then the corresponding λ is a strictly positive vector belonging to a compact subset of Δ_+.

We follow the Negishi approach to analyze if there are Pareto optimal allocations and if the First Welfare Theorem is satisfied. In our setting this theorem is valid: every walrasian equilibrium define a Pareto optimal allocation. Nevertheless, it is not immediate to recognize the conditions that assure the existence of Pareto optimal allocations. So, we discuss sufficient conditions that assure the existence of these allocations.

4.4 The Existence of Pareto Optimal Allocations in Infinte Dimensional Economies

In [26] it is shown that, if the economy satisfy the *closedness condition*[2] then there exist Pareto optimal allocation. Hence, if this condition is satisfied for every Pareto optimal allocation $\bar{x}$ there exists $\lambda(\bar{x}) \in \Delta_+$ such that $\bar{x}$ solves the maximization program

$$\begin{gathered} \max_{x \in \Omega_+} \textstyle\sum_{i=1}^n \lambda_i(\bar{x}) u_i(x_i), \\ s.t. \textstyle\sum_{i=1}^n x_i \leq \sum_{i=1}^n w_i = W. \end{gathered} \tag{4.2}$$

If the utilities are strictly concave functions, then $\bar{x}$ is the only solution of this program.[3] Closedness condition follows if the attainable set $\mathscr{F}$ is compact in a compatible topology. Every economy with order interval $[0, W]$ weakly compact, satisfy the closedness condition.[4] However, an exchange economy can satisfy closedness condition without order interval being weakly compact.[5] The following theorem

[2] Recall that the closedness condition is satisfied, if and only if the utility set,

$$U = \{(u_1(x_1), u_2(x_2), \dots, u_n(x_n)) : (x_1, x_2, \dots, x_n) \in \mathscr{F}\}$$

is closed (see [26]).

[3] This result follows from the fact that, if the allocation $(\bar{x}_{,} \dots, \bar{x}_n)$ is Pareto optimal, then it solves the maximization problem, $\max_{x_n \in X_+} u_n(x_n), \quad s.t. : u_1(x_1) = u_1(\bar{x}_1), \dots, u_1(x_{n-1}) = u_{n-1}(\bar{x}_{n-1}), \sum_{i=1}^n x_i = W$. The existence of the Lagrange multipliers for this problem, does not depends on the fact that the interior of the positive cone X_+ is or not empty (see [9]).

[4] If the order interval $[0, W]$ is weakly compact, then the set of all individually rational Pareto optimal allocations is a non-empty and it is a weakly compact subset of Ω_+.

[5] It is possible to find *well behaved economy* such that the weak compactness of the interval $[0, W]$ is no satisfied and where there is no Pareto optimal allocations. It follows that these economies do not satisfy the closedness condition (see [8]).

states that if topologies are compatible with a given dual pair,[6] then we can use a closed convex sets without specifying the compatible topology to which we are referring (see [6]).

Theorem 4.1. *All topologies consistent with a given dual pair have the same closed and convex sets.*

Moreover, for normed spaces, the following theorem holds.

Theorem 4.2. *Let (X, n) be a normed space, then the Mackey topology, the strong topology and the norm topology are the same.*

This theorem is a corollary of the well known Alaoglu's Theorem (see [6]).

Recall that the Mackey topology is the strongest topology consistent with a given dual pair. Every consistent topology with a given dual pair is stronger (finer) than the weak topology and weaker (coarser) than the Mackey topology.[7] The following corollary characterize the quasi-concave and upper semi continuous functions in consistent topological spaces.

Corollary 4.1. *All topologies consistent with a given dual pair have the same upper semi-continuous, quasi-concave functions.*

Proof: The corollary follows from the definition of upper-semi continuity and quasi concavity, because $f : C \to R$ is quasi concave and upper semi continuous if and only if $\{x : f(x) \geq \alpha\}$ is convex and closed in a given topology. The result follows because these properties are preserved in all consistent topologies. □

The next theorem summarize the above considerations.

Theorem 4.3. *If utilities are Mackey upper semi-continuous and quasi concave then the closedness condition is a sufficient condition for the existence of an individually rational Pareto optimal allocation.*

Weak compactness of the interval $[0, W]$ imply closedness, but the reciprocal is not true (see [3] and [5]).

Remark 4.1. Assume, in addition to our hypothesis, the closedness condition. Then, the rational Pareto optimal allocations is a non-empty subset contained in the feasible set of allocations.

It is important to keep in mind two important points: First, closedness condition is a technical device to establish a sufficient condition for the existence of the Pareto optimal allocation (it is no necessary to alter the assumption that utilities are continuous in the strong topology). Second, that this condition is strictly weaker than the requirement that the attainable set $\mathscr{F}$ is a compact in some compatible topology.

[6] Recall that a topology τ is consistent with a given dual pair (X, X^*), if the topological dual of X for τ is X^*.

[7] We say that the topology τ' is stronger (or finer) than the topology τ if and only if every τ-open set is τ'-open set. Reciprocally for the concept of weaker topology.

A solution $x(\lambda, W)$, for the maximization program

$$\max_{x\in\Omega_+} \sum_{i=1}^{n} \lambda_i u_i(x_i), s.t. \quad \sum_{i=1}^{n} x_i \le \sum_{i=1}^{n} w_i = W, \tag{4.3}$$

then is a Pareto optimal allocation.

Our next step to characterized the set of Pareto optimal allocations, is to choose the elements x^* in the Pareto optimal set that can be supported by a price p and satisfy $px^* = pw_i$, for all $i = 1, 2, \dots, n$, i.e. an equilibrium allocation.

We denote by $W = \sum_{i=1}^{n} w_i \in int[X_+]$ the aggregate endowments of the economy, and by $w \in \Omega_{++}$ the vector $w = (w_1, w_2, \dots, w_n)$ of the initial endowments such that $w_i > 0$ for all $i \in I$. Suppose that the aggregate endowment of the economy is fixed.

Since we are interested only in individually rational Pareto allocations, then under our hypothesis it is enough to consider $x_i > 0$, and $\lambda_i > 0$, for all $i \in \{1, \dots, n\}$. Let, for every $\lambda \in int[\Delta] = \{\lambda \in \Delta : \lambda_i > 0 \ \ \forall i \in I\}$,

$$x(\lambda, W) = argmax \left\{ \sum_{i=1}^{n} \lambda_i u_i(x_i), \ s.t. \ \sum_{i=1}^{n} x_i \le \sum_{i=1}^{n} w_i = W \right\}. \tag{4.4}$$

4.5 The Excess Utility Function

The excess utility function has similar properties to those of an excess demand function, but its generalization to infinite dimensional economies is straightforward, even in the cases where does not exist continuous excess demand functions. An example of the good properties of the excess utility function to analyze the existence of walrasian equilibria in economies with assets and goods (an infinite dimensional economy) is presented in [22].

Consider the vectorial function $e : int[\Delta] \times \Omega \to R^n$ with coordinates given by

$$e_i(\lambda, w) = u_i'(x_i(\lambda, W))(x_i(\lambda, W)) - w_i), \tag{4.5}$$

The concavity of the utility functions are sufficient for the existence of the Lagrange multiplier γ^* for the problem (4.3), even if the positive cone Ω_+ has an empty interior (see [9]). The excess utility function is defined by:

$$e_i(\lambda, w) = \gamma^*(\lambda, W))(x_i(\lambda, W)) - w_i).$$

Assume that the preferences can be represented by $C^1(X_+)$ utility functions.

Definition 4.1. For every $w \in \Omega_{++}$, we define the set of the *equilibria social weights*

$$\mathscr{E}q(w) = \{\lambda \ \in \ int[\Delta] : e_w(\lambda) \le 0\}.$$

The set of equilibria social weights, does not depend on the utilities representing the preferences. In this sense, it is a robust concept. The social weights is a very useful and powerful mathematical way to analyze the properties of the equilibria set of an economy, even in an infinite dimensional setup. The critical characteristic of a singular economy is reflected by this set (see Sect. 4.9). The response to a perturbation in the fundamentals of a singular economy is a big change in this set. This change implies big changes in the social weights of the economic agents. Under this hypothesis, $u'_i(x_i(\lambda, W)) > 0$, for all $x_i \in X_+$, we can identify the equilibria social weights with the set

$$\mathscr{E}q(w) = \{\lambda \in int[\Delta] : e_w(\lambda) = 0\}.$$

Theorem 4.4. *A distribution λ belong to $\mathscr{E}_q(w)$ if and only if $(x_i(\lambda, W)) - w_i)$ belong to the kernel of the functional $\gamma^*(\lambda, W)$, being $x(\lambda, W)$ the individually rational Pareto optimal allocation solving (4.3) and $\gamma^*(\lambda, W)$, the Lagrange multiplier for the problem (4.3). Symbolically:*

$$\lambda \in \mathscr{E}q(w) \;\Leftrightarrow\; (x_i(\lambda, W)) - w_i) \in Ker[\gamma^*(\lambda, W)].$$

In [4] it is shown that the equilibrium social weights $\mathscr{E}q(w)$ is a non-empty set. Note that, w depends on external influences on the economy.

4.6 The Equilibrium Set as a Banach Manifold

The Equilibrium Social Set is given by

$$\mathscr{E}_q = \{(\lambda, w) \in int[\Delta] \times \Omega_{++} : e(\lambda, w) = 0\}.$$

The allocation $x^* \in \Omega_{++}$ solves (4.3) if and only if there exists $\gamma^* \in X^*$ such that the following identities are satisfied (see [23]):

$$\begin{aligned} &\lambda_i u'_i(x^*_i) - \gamma^* = 0 \\ &\textstyle\sum_{i=1}^n x^*_i - \sum_{i=1}^n w_i = \theta. \end{aligned} \tag{4.6}$$

Both terms in the first equation of (4.6) are linear functionals. The second member denotes the null operator. In the second equation θ represents de null element of X. The functional γ^* is the Lagrange multiplier. For an arbitrary $h \in X$ it follows that:

$$\begin{aligned} &\lambda_i u'_i(x^*_i)h - \gamma^* h = 0 \\ &\textstyle\sum_{i=1}^n x^*_i - W = \theta. \end{aligned} \tag{4.7}$$

These equalities represent the first order conditions for the maximization problem. Under the hypothesis in this work, these are necessary and sufficient conditions for the existence of a solution of the problem (4.3). If for a given $(\tilde{\lambda}, \tilde{W}) \in int[\Delta] \times X_{++}$, there exist $x^* \in X_{++}$ and a functional γ^* solving (4.7), then x^* is a solution for the maximization problem (4.3) with $\lambda = \tilde{\lambda}$ and $W = \tilde{W}$.

Using the implicit function theorem, the function $f : \mathcal{U}_{\tilde{\lambda}} \times U_{\tilde{W}} \to \Omega_{++}$ is well defined by $f(\lambda, W) = x^*$ and the function $g : \mathcal{U}_{\tilde{\lambda}} \times U_{\tilde{W}} \to X^*$ is well defined by $g(\lambda, W) = \gamma^*$, where $\mathcal{U}_{\tilde{\lambda}} \subseteq int[\Delta]$ is an open neighborhood of $\tilde{\lambda}$ and $U_{\tilde{W}} \subseteq X_{++}$ is an open neighborhood of $\tilde{W}$. We recall that $int[\Delta]$ and X_{++} are B-manifolds.

Furthermore, $x^*(\lambda, W)$ and $\gamma^*(\lambda, W)$ C^k.

We use the following notation: $x^*_{i,\lambda_j}(\lambda, W) = \partial x^*_i / \partial \lambda_j(\lambda, W)$ and x^*_{i,w_j} $(\lambda, W) = \partial x^*_i / \partial w_j(\lambda, W)$. The derivatives with respect to w_j follow by the chain rule.[8] The following result summarizes some elementary properties of the excess utility function.

Theorem 4.5. *Let $e : int[\Delta] \times \Omega_{++} \to R^n$ be the excess utility function, then for all $(\lambda, w) \in int[\Delta] \times \Omega_{++}$, $\lambda e(\lambda, w) = 0$ and $e(\alpha\lambda, w) = e(\lambda, w)$, for all $\alpha > 0$.*

The rank of the jacobian matrix $J_\lambda e(\lambda, w)$[9] of the excess utility function $e(\cdot, w) : int[\Delta] \to R^n$ is at most equal to $n - 1$. If $e_i(\lambda, w) = 0$ $\forall i = 1, 2, \ldots, n-1$, then $e_n(\lambda, w) = 0$. We will consider the restricted function $\bar{e} : int[\Delta] \times \Omega_+ \to R^{n-1}$ obtained from the excess utility function removing one of its coordinates, for instance e_n.

The following fundamental result characterizes the Equilibrium Social Set, for some economies, with infinitely many commodities, as a Banach Manifold.

Theorem 4.6. *If the positive cone of the consumption space has a non-empty interior, then there exists an open and dense subset $\Omega_0 \subseteq \Omega_{++}$ such that*

$$\mathcal{E}q/\Omega_0 = \{(\lambda, w) \in int[\Delta] \times \Omega_0 : e(\lambda, w) = 0\}$$

is a Banach manifold.

Proof: There exists a residual set $\Omega_0 \subseteq \Omega$ such that, the mapping $\bar{e} : int[\Delta] \times \Omega_0 \to R^{n-1}$ is a submersion (see [32] vol. 1). In particular that zero is a regular value of e, i.e. for all $(\lambda, w) \in int[\Delta] \times \Omega_0$, $e(\lambda, w) = 0$ is a regular point. For every parameter $w \in \Omega_0$, the mapping $\bar{e}(\cdot, w) : int[\Delta] \to R^{n-1}$ is defined in finite dimensional space and so, is a Fredholm map with index zero. Hence, from Sect. 4.19 of [32], the theorem follows.

We use the notation $e_w(\cdot)$ to denote the function $e(\cdot, w) : \Delta_+ \to R^{n-1}$. The cardinality of the equilibrium social weigh, for every economy with $w \in \Omega_0$, is established in the following corollary.

[8] $x^*_{i,w_j}(\lambda, w) = \frac{\partial x^*(\lambda, W)}{\partial W} \frac{\partial W}{\partial w_j}$.

[9] As $int[\Delta]$ is a B-manifold whose chart map is $X_\phi = R^{(n-1)}$ we can consider the concept of $F - derivative$ of this map, here we represent $e'(\lambda, w)$ by means of the symbol $J_\lambda e(\lambda, w)$,

Corollary 4.2. *For a fixed $w \in \Omega_0$, the equation $e(\lambda, w) = 0$ has at most finitely many solutions, i.e. the subset $\mathscr{E}_q(w) = \{\lambda \in \Delta_+ : e_w(\lambda) = 0\}$ is, for every $w \in \Omega_0$, a finite subset of Δ_+.*

Proof of the corollary: From Theorem (4.6) we know that, for all $w \in \Omega_0$, zero is a regular value for $e(\cdot, w)$. The pre-image of zero is a 0-dimensional manifold. The convergence of $\bar{e}(\lambda_n, w_n) \to 0$ as $n \to \infty$ and convergence of $\{w_n\}$ implies the existence of a convergent subsequence of $\{\lambda_n\} \in int[\Delta]$.[10] Note that under the assumptions in our model and since $w_i > 0$ for all i if $\lambda_n \to \bar{\lambda} \in \partial(\Delta_+)$ then there exists some i such that $\bar{\lambda}_i = 0$ and $x_i(\lambda_n) \to 0$ and $\|u_i'(x((\lambda_n)\| \to \infty$ when $\lambda_n \to \bar{\lambda}$. Hence, $\|e_i(\lambda_n)\| \to \infty$. The pre-image of zero by $\bar{e}_w(\cdot)$ is a finite set of points for all $w \in \Omega_0$. □

The oddness of this solutions is shown in [4].

The economies $\mathscr{E} = \{w_i, u_i, I\}$, where $w \in \Omega_0$ are called *Regular Economies.*

In [25], it is shown that the set of regular economies is an open and dense set in the space of all economies. To obtain this result it is sufficient to allow that w might not be possible. In this work, we need this assumption to characterize the equilibria set as a Banach manifold.

4.7 Singular Economies and Some of its Properties

We describe each economy by its excess utility function $e : int[\Delta] \times \Omega_+ \to R^{n-1}$. The equilibria of an economy are described by the state variables $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_n) \in \mathscr{E}q(w)$. These equilibrium states change with the parameters $w \in \Omega$. These parameters are called external or control parameters. Given w, the set of λ such that $e(\lambda, w) = e_w(\lambda) = 0$ determine the possible states of the economy, i.e. the possible set of equilibria. The parameters w describe the dependence of the system on external forces, the action of these forces cause changes in the states of the economy. Generically these changes are not so big, and the new state is similar to the previous one. This is because generically economies are regular. Nevertheless, a sudden transition can occur from a continuous parameter change. The systematic study of these sudden changes is one of the main concerns of the catastrophe theory. These kind of changes can take place only in a neighborhood of a singular economy. The significance of singularity theory in economics is precisely that the essential changes are connected with singularities. Thus the knowledge of the *canonical forms* gives a deeper inside in the qualitative knowledge of the economic behavior. We classify the economies according with their singularities. We begin classifying economies in two classes: regular and singular. A state $\lambda \in \mathscr{E}q(w)$ is *singular or critical equilibrium* if corank of the jacobian matrix $J_\lambda \bar{e}_w$ is positive, where the corank of $J_\lambda \bar{e}_w(\bar{\lambda})$ is given by

[10] This condition is characterized saying that the family of maps $e(\cdot, w)$ is proper with respect to λ.

$$corank\left[J_{\lambda}e_w(\bar{\lambda})\right] = (n-1) - dim\left[J_{\lambda}\bar{e}_w(\bar{\lambda})\right].$$

Singular economies will be classified in two classes:

Definition 4.2. A singular economy is non-degenerate if for all $\bar{\lambda} \in \mathscr{E}q(w)$ the

$$corankJ_{\lambda}\bar{e}_w(\bar{\lambda}) \leq 1$$

and with strict inequality for at least one $\lambda_c \in \mathscr{E}q(w)$.The equilibria states are called *critical non-degenerate equilibria.* The remain singular economies are called *degenerate.* An equilibrium $\bar{\lambda} \in \mathscr{E}q'(w)$ with $corankJ_{\lambda}\bar{e}_w(\bar{\lambda}) > 1$ are called a *degenerate critical equilibrium.*

The corank is a measure for the degree of the degeneration of the equilibria.

Example 4.1. Let E(W) $= \left\{R^2_+, u_i, w_i; i = 1,2\right\}$ be the set of exchange economies which total endowment $W = (W_1, W_2)$. Let

$$W_j = w_{1j} + w_{2j},\ j = 1,2$$

where w_{ij} is the initial endowment of agent i with respect to the commodity j. Initial endowment may be redistributed but the total endowment can not be modified. Hence, the components of W are constants.

The equilibrium set is:

$$\mathscr{V}_W = \left\{(\lambda, w) \in int[\Delta] \times \Omega_+, : e(\lambda, w) = 0,\ w_{1j} + w_{2j} = W_j; j = 1,2\right\}. \tag{4.8}$$

An equilibrium is a pair (λ, w) such that $e_1(\lambda, w) = 0,\ e_2(\lambda, w) = 0$. Suppose that the excess utility function of the agent 1 is given by

$$e_1(\lambda_1, w_{11}, w_{12}) = 3W_1\lambda_1 - 3w_{11}(\lambda_1)^{\frac{1}{3}} + w_{12}. \tag{4.9}$$

In terms of catastrophe theory λ_1 is the state variable and w_1 are the control parameter. The social equilibria of this economy is given by the set of pairs (λ, w) such that its components $(\lambda_1, w_{11}, w_{12})$ solve the equation $e_1(\lambda_1, w_{11}, w_{12}) = 0$ and by the corresponding $(\lambda_2, w_{21}, w_{22})$ obtained from the former. The set

$$C_F = \{(\lambda_1, w_{11}, w_{12}) \in \mathscr{V}_W : detJ_{\lambda_1}e_1(\lambda_1, w_{11}, w_{12}) = 0\},$$

is the *catastrophe surface.* The economies whose endowments are in this surface are the singular economies. In our case this surface is given by

$$C_F = \left\{(\lambda_1, w_{11}, w_{12}) \in \mathscr{V}_W : \frac{\partial e}{\partial \lambda_1} = 3W_1 - w_{11}\lambda^{-\frac{2}{3}} = 0\right\}.$$

Explicitly,

$$C_F = \left\{ \left(\left(\frac{w_{11}}{3W_1} \right)^{\frac{3}{2}}, \; w_{11}, \; \frac{2w_{11}^{\frac{3}{2}}}{(3W_1)^{\frac{1}{2}}} \right) \right\}.$$

The projection of this set in the space of parameters is called the *bifurcation set.* In our case,

$$B_F = \left\{ \left(w_{11}, \; \frac{2w_{11}^{\frac{3}{2}}}{(3W_1)^{\frac{1}{2}}} \right) \right\}.$$

This set is represented in the space of parameters, (w_{11}, w_{12}) by a parabola. By varying the parameters continuously and crossing this parabola, something unusual occurs: *the number of possible states of equilibria associated with the initial endowments w increases or decreases*. The number of equilibria is given by the sign of δ, where

$$\delta = 27 \left(\frac{w_{11}}{W_1} \right)^2 - 4 \left(\frac{w_{12}}{W_1} \right)^3.$$

Therefore,

- $\delta < 0$ associate with w, there exist three regular equilibria.
- $\delta > 0$ there is one regular equilibrium associate with w.
- $\delta = 0$, *and* $w_{11}w_{22} \neq 0$ there exists one critical (or singular) equilibrium and one regular equilibrium.

The set of regular economies with a unique equilibrium is arc connected in the two agents case, help us to obtain a good geometric representation of economies.

The hessian matrix of the considerate excess utility function (the matrix defined by the second order derivatives of e_w at λ) is singular. Hence, the critical equilibrium is degenerate. Thus, economies with endowments satisfying $\delta = 0$ are degenerate singular economies.

Example 4.2. Consider the economy $\mathscr{E} = \{X, u_{\alpha,i}, w_i, i = 1, 2\}$, with consumption space $X = C^k([0, 1], R_+) \times C^k([0, 1], R_+)$. The utility functions are given by

$$u_{1,\alpha}(x_1) = \int_0^1 [x_{11}(t) - \tfrac{1}{\alpha} x_{12}^{-\alpha}(t)] dt,$$

$$u_{2,\alpha}(x_2) = \int_0^1 [-\tfrac{1}{\alpha} x_{21}^{-\alpha}(t) + x_{22}(t)] dt,$$

where $\alpha \in (0, 1)$ and the endowments (w_{i1}, w_{i2}) are real, strictly positive functions defined in $[0, 1]$. Assume that the equality $W(t) = w_1(t) + w_2(t)$ holds, for all $t \in [0, 1]$. Thus, W is a fixed vectorial continuous field.

Following the Negishi approach (see [28]), we begin solving the optimization problem:

$$\max_{x \in C^k[0,1]_+^4} U_\lambda(x) = \lambda_1 u_{1,\alpha}(x_1, x_2) + \lambda_2 u_{2,\alpha}(x_1, x_2),$$

restricted to the feasible set:

$$\mathscr{F} = \left\{ x \in (C([0,1], R_+))^4 : \sum_{i=1}^{2} x_i(t) \leq \sum_{i=1}^{2} w_i(t) \ \forall \ t \in [0,1] \right\}.$$

The excess utility function is given by

$$e_w(\lambda) = (e_{w_1}(\lambda), e_{w_2}(\lambda)) = \left(\int_0^1 e_{w_1}(t)dt, \int_0^1 e_{w_2}(t)dt \right).$$

Denoting $\lambda_1 = \lambda$ and with $\lambda_2 = 1 - \lambda$, it follows that

$$e_{w_1}(t) = \left(\tfrac{1-\lambda}{\lambda}\right)^{\frac{\alpha}{1+\alpha}} - \left(\tfrac{1-\lambda}{\lambda}\right)^{\frac{1}{1+\alpha}} - w_{12}(t)\left(\tfrac{1-\lambda}{\lambda}\right) + w_{21}(t),$$

$$e_{w_2}(t) = \left(\tfrac{1-\lambda}{\lambda}\right)^{\frac{-\alpha}{1+\alpha}} - \left(\tfrac{1-\lambda}{\lambda}\right)^{\frac{-1}{1+\alpha}} - w_{21}\left(\tfrac{1-\lambda}{\lambda}\right)^{-1} + w_{12}(t).$$

Solving $e_{w_1}(\lambda) = 0$ and $e_{w_2}(\lambda) = 0$, we obtain the social equilibria for the economy characterized by $w = (w_1, w_2)$ (this equilibria depends on w). Taking derivatives in the excess utility functions (with respect to λ). It follows that the catastrophe surface is given by

$$C_F(t) = \left\{ (\lambda, w_{11}, w_{12}) \in \mathscr{V}_W : \int_0^1 w_{12}(t)dt = \frac{\alpha}{1+\alpha} h^{\frac{1}{1+\alpha}} - \frac{1}{1+\alpha} h^{\frac{\alpha}{1+\alpha}} \right\}$$

where $h = \lambda/(1-\lambda)$. The economies $\mathscr{E}$ whose endowments are given by

$$(w_{11}, w_{12}, w_{21}, w_{22}) \in (C[0,1], R_+))^4$$

satisfy, for every $t \in [0,1]$, $W(t) = w_1(t) + w_2(t)$ and

$$\int_0^1 w_{12} = \frac{\alpha}{1+\alpha} h^{\frac{1}{1+\alpha}} - \frac{1}{1+\alpha} h^{\frac{\alpha}{1+\alpha}} \quad \forall \ t \in [0,1]$$

are singular. Solving $e(\lambda, w) = 0$, there exist economies with one equilibrium and economies with three equilibria.

4.8 Catastrophe Theory and Economic Theory

The catastrophe theory was introduced by the mathematician René Thom in the 1960' and it highlights the importance of singularities to understand why a discontinuity in the behavior of a system can occurs even in a smooth or continuous environment. This theory give a deeper insight to understand the sudden changes in economics, especially from a qualitative viewpoint (see [11]). This approach allows to use differential topology in finding the laws of the crises in economy. We focus our analysis in the concept of stable singularities, determined by Hassler Whitney and incorporated in the theory by Rene Thom and Harold Levine in the 1960 Bonn notes.

As Thom and Levine have shown singularity theory can be applied with wide generality in quasi-statical models (models with equilibria states only modified by cause of external forces) in which small changes in its parameters cause sudden changes. When the system is at rest in a position of equilibrium the state variables (the social weights λ in our case) determine the state of the system. The parameters (the initial endowments of the economy) describe the dependence of the system on external forces. The action of these forces can give rise to sudden jumps from an equilibrium position to another. These sudden transitions, when originating from continuous modifications in the parameters are catastrophes. This kind of transitions are observed in a neighborhood of singular economies.

Singularity theory shows that in some cases it is possible to analyze this kind of transition using canonical forms, i.e. taylor expansion up to some order.

We say that a map $f \in C^k(X,Y)$ is *k-equivalent* to a map $g \in C^k(X,Y)$ at the points x_0 and u_0, if there exists local C^k diffeomorphisms $\phi : X \to X$ and $\psi : Y \to Y$ such that $u_0 = \phi(x_0)$ and $g(\phi(x)) = \psi(f(x))$, for all x in a neighborhood of x_0.

Remark 4.2. Let $f : U(p) \subset X \to Y$ be $C^k(X,Y), k \geq 1$, and X and Y Banach manifolds.

(a) Let f be a submersion or an immersion at p and let $g = j^1_k(f)$. Then f is k-equivalent to g at 0.
(b) If $X = R^n$ and $Y = R^m$ and $rank f'(p) = r = \min\{n,m\}$, then f is k-equivalent at p to $h : X \to Y$ given by $h(x_1,\ldots,x_n) = (x_1,\ldots,x_r,0,\ldots,0)$, for all $x \in U(p)$.
(c) Every analytical function $f : X \to Y$, with isolated critical points, is equivalent to a Taylor polynomial of sufficiently high order see ([29]).

The following two theorems are well known in singularity theory, and they will help us to understand some characteristics of the non-degenerate singular economies.

Theorem 4.7. *Let $f : X \to R$ be a smooth function with a non-degenerate singular point p. Then there exists a neighborhood V of p in X such that no other singular point of f are in V, i.e., non-degenerate singular points are isolated.*

We say that a map $f : X \to R$ has singular values if for every two singular points p and q, $p \neq q$ implies $f(p) \neq f(q)$.

The set of Morse functions whose singular values are distinct form a residual set in $C^\infty(X,R)$.

In economics terms, this question takes the following form: Where is it possible to characterize the behavior of an economy from the Taylor expansion, up to some order k, from its excess utility function?

Definition 4.3. We will say that the economy $\mathscr{E} = \{u_i, w_i, i \in I\}$ is k-equivalent at $\lambda^0 \in \mathscr{E}q(w)$ to the economy $\mathscr{E}' = \left\{u_i, w_i', i \in I\right\}$ at $\lambda^1 \in \mathscr{E}'q(w')$ if and only if its respective excess utility functions e_w and $e_{w'}$ are $k -$ *equivalent* functions at λ^0 and λ^1.

From Remark 4.2, every the excess utility function of every regular economy is 1-*determined* in every equilibrium. The regular economies are locally equivalent at their respective equilibria.

Generically, if the economy $\mathscr{E} = \{u_i, w_i, I\}$ is singular non-degenerate, then there exists only one critical equilibrium $\lambda \in \mathscr{E}q(w)$. An extensive analysis of the behavior of the singular economies 2-agent is given in [2].

4.8.1 Two Agents Economies

Define by $\mathscr{E}_{u,w} = \{X_+, u_i, w_i = 1, 2\}$ the set of exchange economies with two agents, whose utility function are denoted by u_i, endowments by w_i, consumption sets by X_+, and $\Omega_+ = X_+ \times X_+$. Let $\bar{e}_w : U_{\lambda_0} \subseteq int[\Delta] \to R^{n-1}$ be the excess utility function of the economy $\mathscr{E}_{u,w}$. If this function is a C^k submersion at λ_0 then there exists a local C^k diffeomorphism ϕ, with $\phi(\lambda_0) = 0$ and $\phi'(\lambda_0) = I$, the following normal local form holds: $\bar{e}(\phi(\lambda_0)) = \bar{e}'(\lambda_0)\lambda$ (see [32] vol. 4). Therefore, this excess utility function corresponds to a regular economy. The main question is: When $\bar{e}_w$ is not a submersion at λ_0? *For an economy with two agents, the Morse lemma is the answer.*

The main characteristics of the excess utility functions given in Theorem 4.5, characterize a two agent economy by only one component of its excess utility function and by only one of the two social weights. Let $e_i : (0, 1) \times \Omega \to R$ be the excess utility function of the agent indexed by i. We classify this kind of economies by looking for the Taylor expansion of $e_{wi} = \bar{e}_w$.

Let $B \subset R$ be open and convex subset. If $g : B \to R$ is a smooth function such that $g(\bar{x}) = g'(\bar{x}) = ... = g^{(k)}(\bar{x}) = 0$ then there is a smooth function $l : B \to R$ such that $g(x) = (x - \bar{x})l(x)$, and $l(\bar{x}) = 0$. However, if $g^{(k)}(\bar{x}) \neq 0$, then there exists a smooth local change of coordinates under which g takes the form x^k for all k odd and $\pm x^k$, if k is odd, and $\pm x^k$ if k is even (see [31]).

By Morse's Lemma in R^n it is possible to reduce the family of the excess utility function of the non-degenerate singular economies, with independent utilities to just 2 simple stereotypes, namely

$$\bar{e}_{\bar{w}i}(\psi(\lambda)) = \pm\lambda_1^2,$$

where $\bar{e}$ is the restricted excess utility function. Degenerate singular economies are characterized by the fact that there exist at least one $\bar{\lambda} \in \mathscr{E}_q(w)$ such that $e''(\bar{\lambda}) = 0$.

4.9 The S_r Classification

We say that a function $f : U_x \to Y$ is *k-determined* if and only if for every function $g : U_x \to Y$ with $j^k f(x) = j^k g(x)$ there exists a local $C^k(X)$ diffeomorphism which satisfies $g(\phi(u)) = j^k f(u)$ in a neighborhood of x.

The k-jet of the excess utility function e_w is given by

$$j^k e_w(\lambda) = (\lambda, 0, e'_w(\lambda), \dots, e_w^{(k)}(\lambda)).$$

If the excess utility function e_w of a given economy $\mathscr{E}$ is k-determined, for some $\lambda \in \mathscr{E}_q(w)$, then every economy $\mathscr{E}'$ whose excess utility function $e_{w'}$ has the same Taylor polynomial up to order k, for some $\lambda' \in \mathscr{E}_q(w')$, show, in some neighborhood of λ', the same qualitative behavior that the economy $\mathscr{E}$ in some neighborhood of λ. Hence, if the excess utility function e_w is k-determined, then the k-jet summarize the essential behavior of every economy in a neighborhood of every of its equilibria.

Let X and Y be n and m dimensional smooth manifolds. Let $f,\ g : X \to Y$ with $f(x) = g(x) = y$, be smooth functions. We say that f and g are equivalent, $f \sim_k g$, if and only if the k-th Taylor expansion of f coincides with the k-th expansion of g at x. The equivalence class of f at x is called the *k-jet of f at x*, and will be denoted by $j_x^k f$. Let $J^k(X,Y)_{x,y}$ denote the set of all equivalence classes $\sim_k$ of maps $f : X \to Y$ with the points that $f(x) = y$. Let $J^k(X,Y) = \bigcup_{(x,y)\in X\times Y} J^k(X,Y)_{x,y}$. Given a smooth map $f : X \to Y$ there is a canonical map $j^k(f) : X \to J^k(X,Y)$. Note that $J^0(X,Y) = X \times Y$ and $j_x^0 f = (x, f(x))$ is the graph of f. It follows that $f \sim_0 g$ at x if and only if $f(x) = g(x)$. We will represent the jacobian matrix of a mapping f by the symbol $(\partial f)_x$. If $\sigma \in J^1(X,Y)_{xy}$ then σ defines a unique linear mapping $T_x X \to T_y Y$, where x is the source of σ and y is the target of σ. Let f be a representative of σ in $C^\infty(X,Y)$. Then $(\partial f)_x$ is the linear mapping. Define $rank(\sigma) = rank(\partial f)_p$ and $corank(\sigma) = \mu - rank(\sigma)$, where $\mu = min(dimX, dimY)$. Let

$$S_r = \{\sigma \in J^1(X,Y) : corank(\sigma) = r\} \tag{4.10}$$

be the subset of the equivalence classes under $\sim_1$ in $C^\infty(X,Y)$ such that the $corank(\partial f) = r$. The subset S_r is a submanifold of $J^1(X,Y)$ satisfying

$$codim\ S_r = (n - \mu + r)(m - \mu + r),$$

(see [21]). Our interest is to study the class $\sigma \in J^1(int[\Delta], R^{n-1})$ with source λ, and target R^{n-1}. It follows that $codimS_r = r^2$.

The set of singularities of $f : X \to Y$ such that the rank of the jacobian matrix drops by r is represented by $S_r(f) = (j^1 f)^{-1}(S_r)$. Then $S_r(f)$ will be, generically, a manifold with the same codimension of (S_r) (see [21]). Since $codimS_r(f) = dimX - dimS_r(f) \geq 0$, there is a relationship between the kind of singularities for every $f \in C^{\infty}(X, Y)$ and the dimension of the manifold. Applying these concepts to economics it follows that

- $S_r(e)$ is the set of critical points of $\bar{e}_w : int[\Delta] \to 0$ where the jacobian matrix of $\bar{e}_w$ drops rank by r. This set is a manifold.
- The set of critical social equilibria is the subset $(\lambda, w) \in S_r(e)$ with $e(\lambda, w) = 0$ and $corank\ (\partial \bar{e}_w)_\lambda = 1$.
- $S_1(e)$ is the set of non-degenerate critical social equilibria that is the set of pairs $(\lambda, w) \in int[\Delta] \times \Omega$ such that $e(w, \lambda) = 0$ and $corank(\partial \bar{e}_w) \geq 1$.

There exists a relationship between the number of agents and normal form of singularities. In others words, the excess utility function can have only some types of singularities and determined by the number of consumers in the economy. If $codim\ S_r(f) > |dimX - dimY|$ then $dimS_r(f) < dimY$. Applying this observation to economics, with $X = int[\Delta], \quad Y = R^{n-1}$, and $f = \bar{e}_w$, it follows that: if n is the number of consumers of the economy then $dimS_r(e) < n - 1$. If $n = 2$, that the singular economies have generically isolated singular equilibrium in $int[\Delta]$.

We note that the topology used in theorems about transversality of maps in $C^{\infty}(X, Y)$ is the Withney topology.

4.10 The Fold and the Cusp in Economics

Let $f : X \to Y$ be a map. We say that a map is *one generic* if $J^1 f$ is transversal to S_1. Recall that $S_1(f)$ denotes the singularities of f of type S_1. By Thom Transversality theorem it follow that the set of $f \in C^{\infty}(X, Y)$ transversal to S_1 is a residual subset of $C^{\infty}(X, Y)$. Hence, $S_r(f) = (j^1 f)^{-1} S_r$.

Remark 4.3. By (4.10), it follows that $S_1(f)$ has codimension 1 and S_2 has codimension 4.

Let p in $S_1(f)$ and $q = f(p)$. Only one of the following two situation can occur

$$\begin{cases} (a)\ T_p S_1(f) \oplus Ker(\partial f)_p = T_p X; \\ \\ (b)\ T_p S_1(f) = Ker(\partial f)_p \end{cases} \tag{4.11}$$

Definition 4.4. *(Submersions with Folds)* Let X and Y be a smooth manifolds with $dim\ X \geq dim\ Y$. Let $f : X \to Y$ be a smooth map, such that $J^1 f$ is transversal to S_1. Then a point $p \in S_1(f)$ is called a fold point if

$$T_p S_1(f) \oplus Ker(\partial f)_p = T_p X.$$

If p is a singularity satisfying (4.11) (a) then p is a *fold*. The first Whitney theorem for maps between 2-manifold give the normal form for fold points:

Theorem 4.8. *If (a) in (4.11) occurs, then one can choose a system of coordinates (x_1, x_2) centered at p and (y_1, y_2) centered at q such that f, in these coordinates is the map $(x_1, x_2) \to (x_1, x_2^2)$.*

Let $f(x_1, x_2) = (x_1, x_2^2)$. Note that

$$Df(x_1, x_2) = \begin{pmatrix} 1 & 0 \\ 0 & 2x_2 \end{pmatrix}.$$

Thus, $S_1(f) = \{(x_1, x_2) \in R^2 : x_2 = 0\}$. Let p be a singular point. It follows that

$$Ker(\partial f)_p = Ker(1, 0) = (0, 1).$$

Hence, $T_p S_1(f) \oplus Ker(\partial f)_p = T_p X$. The normal form given by $(x_1, x_2) \to (x_1, x_2^2)$. This transformation has the following geometric interpretation:

- (x_1, x_2) maps onto the parabolic cylinder (x_1, x_2, x_2^2).
- Then projects onto the (x_1, x_3) plane.

If p is a singularity satisfying (b), then p is a *cusp*.

Definition 4.5. *(Cusp)* Let X and Y be 2-dimensional manifolds, and let $f : X \to Y$ be one generic mapping. We will say that p is a simple cusp, if this zero is a simple zero.

The second main theorem of Whitney states (see [21]).

Theorem 4.9. *If p is a simple cusp then one can find coordinates (x_1, x_2) centered at p and (y_1, y_2) centered at q such that:*

$$\begin{cases} f * y_1 = x_1 \\ f * y_2 = x_1 x_2 + \frac{1}{3} x_2^3. \end{cases} \tag{4.12}$$

Where $f*$ is the pull-back function of f by the homomorphism ϕ associated with the change of coordinates. Note that

$$Df(x_1, x_2) = \begin{pmatrix} 1 & 0 \\ x_2 & x_1 + x_2^2 \end{pmatrix}.$$

Let $S_1 = \{(x_1, x_2) \in R^2 : x_1 = -x_2^2\}$. If $p = 0$ then $T_p S_1$ is generated and this subset is $Ker(\partial f)(p)$. If $p = 0$ then $T_p S_1(f) = Ker(\partial f)_p$. Parameterizing the curve S_1, it follows that

$$f * (\sigma(t)) = (-t^2, -\frac{2}{3} t^2),$$

where $\sigma(t) = (-t^2, t)$. Thus, the image of S_1 by $f*$ is a cusp.

4.11 An Example: 3-Agent Economies

Consider an economy with 3-agents, where the utility functions are given and fixed initial endowment. Then the excess utility function is a mapping between 2-manifolds $e_w : int[\Delta] \to R^2$. By our computation $codimS_1(e_w) = 1$ in Δ and S_2 does not occur because it would have codimension 4. Thus, the only possible singularities are of S_1. Let $\lambda \in \mathscr{E}_q(w)$ be a point in $S_1(e_w)$. Only one of the following two situations can occur:

(a)$T_\lambda S_1(e_w) \oplus Ker(\partial e)_\lambda = T_\lambda \Delta$ (fold).
(b)$T_\lambda S_1(e_w) = Ker(\partial e_w)_\lambda$ (cusp).

Remark 4.4. By Whitney, generically in $C^\infty(X, Y)$ the only singularities are folds and simple cusps.

Let $\bar{\lambda} = (\bar{\lambda}_1, \bar{\lambda}_2, \bar{\lambda}_3) \in \Delta$ be a singular social equilibrium for the economy w.

(1) If (a) holds, this is a system of coordinates (λ_1, λ_2) centered at $(\bar{\lambda}_1, \bar{\lambda}_2) \in S_1(e_w)$ and (e_1, e_2) centered at $e_w(\bar{\lambda}) = 0$ such that e_w is a fold $(\lambda_1, \lambda_2) \to (\lambda_1, \lambda_2^2)$.
(2) If (b) holds, generically singularities, are simple cusps. These are coordinates $(\bar{\lambda}_1, \bar{\lambda}_2)$ centered at $e(\bar{\lambda})$ such that

$$(\bar{\lambda}_1, \bar{\lambda}_2) \to (\bar{\lambda}_1, \bar{\lambda}_1\bar{\lambda}_2 + \bar{\lambda}_2^3).$$

In a neighborhood of a cusp or a fold there exist only regular economies but with different number of equilibria. Recall that in a neighborhood of a regular economy, there are only regular economies with similar characteristics. A regular economy can suffer a change if the perturbation in its endowments is big enough. However, if an economy is singular or it is close to singular one, a small change in its endowments can provoke a big change in the economy.

By Whitney theorems, generically the singular economies for three-agents economies have one of the two forms after transformation of coordinates of the dependent and independent variables by local diffeomorphisms.

4.12 The $S_{r,s}$ Singularities

Let $f : X \to Y$ be a one generic map. We will denote by $S_{r,s}(f)$ the set of points the map $f : S_r(f) \to Y$ drops by rank s, with the property that

$$S_{r,s} \subset \{\sigma \in J^2(X, Y) : corank(\sigma) = r\}$$

i.e. $x \in S_{r,s}(f)$ if and only if $x \in S_r(f)$ and the kernel of $(\partial f)_x$ intersects the tangent space to $S_r(f)$ in a s dimensional subspace. Generically, in cases of economies with $n = 3$, the singularities are $S_{1,0}$ folds, when the singularities of the excess utility function are given by $S_1(e_w) = \{(x_1, x_2) \in R^2 : x_1 = 0\}$ and the image of this set by the excess utility function is $e_w(S_1) = \{(x_1, x_2) \in R^2 : x_1 = 0\}$ or $S_{1,1}$ cusps when these singularities are given by the set $S_1(e_w) = \{(x_1, x_2) \in R^2 : x_1 = -x_2^2, \}$ and $e_w(S_1) = \left\{(x_1, x_2) : x_1 = -\frac{3}{2}x_2^{\frac{2}{3}}\right\}$. Using the Transversality Theorem (see [21]) $j^2 f$ is, generically, transversal to $S_{r,s}$ and $S_{r,s}$ is a submanifold in $J^2(X, Y)$. We define

$$S_{r,s}(f) = (j^2(f))^{-1}(S_{r,s}).$$

Generically, $S_{r,s}(f)$ is a submanifold in X with dimension:

$$dimS_{r,s}(f) = dimX - r^2 - \mu r - (codimS_{r,s}(f)\ in\ S_r(f))\,. \tag{4.13}$$

Furthermore,

$$codimS_{r,s} = \frac{m}{2}(k+1) - \frac{m}{2}(k-s)(k-s+1) - s(k-s), \tag{4.14}$$

where $m = dimY - dimX + k$ $k = r + max(dimX - dimY, 0)$ (see [21]). Hence, the set of possible singularities in economics are strongly related with the number of agents.

Let us consider the economy with $k = r$ and $m = r$, where n is the number of agents, l is the number of commodities an r is the codimension of $(\partial \bar{e}_w)_\lambda$ at the singularity. We have that

$$codimS_{r,s}(e) = -\frac{r^3}{2} + r^2 s + \frac{r}{2}\left[-\frac{s^2}{2} - \frac{3}{2}s\right] + s^2,$$

and

$$dimS_{r,s}(e) = (n-1) - r^2 - codimS_{r,s}.$$

In particular, $codimS_{11} = 1$ and $dimS_{11} = n - 3$. Generically, singularities like $S_{1,2}(e)$ only appear if the number of consumers is $n > 4$.

4.13 Conclusions

The excess utility function considers the weight of consumers in the markets and shows the changes in their relative weights (in equilibrium) when the initial endowments change. Near a regular economy these changes are smooth and there are not qualitative changes. However, in a neighborhood of a singularity sudden and big changes can occur. The economic weights of the agents change drastically, overthrowing the existent order. The uncertainty in the behavior of the economy is a direct result of the existence of singular economies. In a neighborhood of a singular economy, the central planer need to be extremely careful. If he acts according to its experience he can do small changes in endowments, but the perturbed economy can be to much different than the original one. Furthermore, it not possible to go back by means of small changes. This situation is what happen in an economic crisis, sudden and unexpect changes occur with unforeseen repercussion in the social behavior of the economy, reflected in changes in the social weights of equilibrium. This impossibility of prediction is intrinsic to the model and the characteristics of the new economies, even in the case where this economy is very close to the original one in the fundamentals, may be different quite from the original one in some of their main characteristics, like the social equilibria.

Nevertheless, most part of the literature in economics has focused until now, on regular economies whose equilibria change smoothly according to the changes in the endowments. The study of the discontinuous behavior (economic crisis) requires to consider singularities, this led us to study catastrophe theory. This theory refers to drastic changes. However, in spite of being sudden, abrupt and unexpected, this theory shows that these changes have a similar substratum that allow us to do a classification according with its geometric representation. This study requires the theory of singularities to understand the forms (canonical forms) of the unexpected changes in economics. The economies can be characterized by their singularities that capture the essence of their behavior. Economies with the same type of singularities will present the same possible changes.

A final consideration: The excess utility function allows us to extend the analysis of singularities for economies with finite commodities to infinite dimensional economies. Showing that, also in these cases, the catastrophe theory, or singularity theory, might be gate to begin to understand the behavior of an economic system with infinitely many goods in a neighborhood of an economic crises.

Acknowledgements The authors are grateful to CONACYT-Mexico. (Project 46209) and UASLP Secretaría de Posgrado (Grant C07-FAI-11-46.82) for research support.

References

1. Accinelli, E., Plata, L.: Micro-foundations of the social change. In: Peixoto, M., Pinto, A., Rand, D. (eds.) Dynamics, Games and Science, Proceedings in Mathematics Series. Springer (2010)
2. Accinelli, E.: Structural stability, Morse's lemma and singular economies. Appl. Math. Sci. **2**(47), 2297–2308 (2008). Hikary Ltd
3. Accinelli, E.: Existence of GE: Are the cases of non Existence a cause of serious worry. In: Petri, F., Hahn, F. (eds.) General Equilibrium: Problems and perspectives, Routledge Siena Studies in Political Economy (2002)
4. Accinelli, E.: Existence and uniqueness of the equilibrium for infinite dimensional economies. Estud. Econ. **21**, 315–326 (1994)
5. Aliprantis, C.D, Brown, D.J., Burkinshaw, O.: Existence and Optimality of Competitive Equilibrium. Springer (1990)
6. Aliprantis, C.D., Border, K.: Infinite Dimensional Analysis. Springer (1994)

7. Araujo, A.: The non-existence of smooth demand in general banach spaces. J. Math. Econ. **17**, 1–11 (1987)
8. Araujo, A.: A note on existence of pareto optima in topological vector spaces. Econ. Lett. **23**, 5–7 (1985)
9. Araujo, A., Monteiro, P.K.: Notes on programing when the positive cone has an empty interior. J. Optim. Theory Appl. **67**(2), 395–410 (1990)
10. Arnold, V., Varchenko, A., Goussein-Zade, S.: Singularites des Applications Differentiables, Editions MIR, Moscow (1986)
11. Balasko, Y.: Economic equilibrium and catastrophe theory: an introduction. Econometrica **46**, 557–569 (1978)
12. Balasko, Y.: Foundations of the Theory of General Equilibrium, Academic Press, Boston (1988)
13. Balasko, Y.: Equilibrium analysis of the infinite horizont models wit smooth discounted utility functions. J. Econ. Dyn. Control **21**, 783–829 (1997a)
14. Balasko, Y.: The natural projection approach to the infinite horizont models. J. Math. Econ. **27**, 251–265 (1997b)
15. Castrigiano, D., Hayes, S.: Catastrophe Theory. Adisson-Wesley (1993)
16. Chichilnisky, G., Zhou, Y.: Smooth infinte economies. J. Math. Econ. **29**, 27–42 (1988)
17. Debreu, G.: Neighboring economic agents. La Dècision. Colloques Internacionaux du CNRS, vol. 171. Paris (1969)
18. Dierker, E.: Regular economies. In: Arrow, K., Intriligator, M. (eds.) Handbook of Mathematical Economics, Chap. 17, vol. II. North-Holland, Amsterdam (1982)
19. Debreu, G.: Economies with a Finite Set of Equilibria Gerard Debreu. Econometrica Vol. 38, No. 3 (May, 1970), pp. 387–392
20. Duffie, D.: Dynamic Asset Pricnig, 3rd edn Princeton University Press (2008)
21. Golubistki, M., Guillemin, V.: Stable Mappings and Their Singularities. Springer (1973)
22. KARATZAS, I., LAKNER, P., LEHOCZKY J.P., & SHREVE S.E.: (1991HYPERLINK http://www.math.columbia.edu/~ik/KLLS91.pdf) Dynamic equilibrium in a simplified stochastic economy with heterogeneous agents. In Stochastic Analysis: Liber Amicorum for Moshe Zakai, 245–272. Academic Press.
23. Luenberger, D.: Optimization by Vectorial Spaces Methods. Willey (1969)
24. Mas-Colell, A.: The Theory of General Equilibrium, A Differentiable Approach. Cambridge University Press, Cambridge (1985)
25. Mas-Colell, A.: Indeterminaci in incomplete market economies. Econ. Theory **1**, 45–62 (1990)
26. Mas-Colell, A., Zame, W.: Equilibrium theory in infinite dimensional economies. In: 701 Hildenbrand, W., Sonneenschein, H. (eds.) Handbook of Mathematical Economy, Chap. 34, vol. 4 (1991)
27. Milnor, J.: Topology from the Differential Viewpoint. University of Virginia Press, Charlottesville (1965)
28. Negishi, T.: Welfare economics and existence of an equilibrium for a competitive economy. Metroeconomica **12**, 92–97 (1960)
29. Samoilenko, A.M.: The equivalence of a smooth function to a Taylor polynomial in the neighborhood of afinite-type critical point. Funct. Anal. Appl. **2/4**, 63–69 (1968)
30. Thom, R.: Sur la Théorie des Enveloppes. J. Math. XLI (1962)
31. Poston, T., Stewart, l.: Catastrophe Theory and Its Applications. Pitman Publishing Limited (1978)
32. Zeidler, E.: Non Linear Functional Analysis and Its Applications. Springer (1993)

Chapter 5
Probabilistic Methods in Dynamical Analysis: Populations Growths Associated to Models Beta(p, q) with Allee Effect

Sandra M. Aleixo, J. Leonel Rocha, and Dinis D. Pestana

Abstract New populational growth models, proportional to beta densities, with shape parameters p and 2, where $p > 1$, and Malthusian parameter r, are developed. For $p > 2$, these models exhibit natural Allee effect. However, in the case of $1 < p \leq 2$, the proposed models do not include this effect. In order to inforce it, we deduce alternative models and investigate their dynamical behaviour. The Verhulst Model, which is a cornerstone of modern chaos theory, is a special case of those models. The complex dynamical behaviour of these models is analysed in the parameter space (r, p), in terms of topological entropy, using explicit methods of dynamical systems. We emphasize some particular disjoint regions in these parameter space, according to the chaotic behaviour of the models, the main result being the characterization of those disjoint regions. We also present some important results about these modified models.

5.1 Introduction

The logistic Verhulst Model originally introduced as a demographic model by Pierre Franois Verhulst, [23], can be considered the basis of modern chaos theory. It incorporates in its parameters both the Malthusian growth rate and the retroaction due to limitation of the natural resources. This non-linear dynamical equation is a natural candidate to model the dynamic of non-overlapping generations, namely when the unit of time is related to the life span of the individuals in the population. Although

S.M. Aleixo (✉) and J.L. Rocha
Mathematics Unit, Instituto Superior de Engenharia de Lisboa, Lisbon, Portugal
and
CEAUL, University of Lisbon, Lisbon, Portugal
e-mail: sandra.aleixo@dec.isel.ipl.pt, jrocha@deq.isel.ipl.pt

D.D. Pestana
Department of Statistics and Operational Research, Universidade de Lisboa, Lisbon, Portugal
and
CEAUL, University of Lisbon, Lisbon, Portugal
e-mail: dinis.pestana@fc.ul.pt

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_5,

the logistic map has been used with success to model the population growth for some species, it is inadequate for other.

In several numerical studies, the families of unimodal maps have been used, allowing for exhaustive investigations in terms of symbolic dynamics, [2]. The unimodal maps theory can be used in many branches of sciences. In population dynamics, aiming to model the growth of a certain species, the use of those families has been frequent. The Verhulst Model is proportional to the $Beta(2, 2)$ density, [22].

In this work, we present new models proportional to the $Beta(p, 2)$ densities,[1] with $p \in]1, +\infty[$, [3]. These new models can be interesting to model population growth in populations for whom the Verhulst Model fails. We make the theoretical deduction of these models and the characterization of the respective family of unimodal maps, $f_{r,p} : [0, 1] \longrightarrow [0, 1]$, defined by $f_{r,p}(x) = rx^{p-1}(1 - x)$, where $r > 0$ is the Malthusian parameter and $p \in]1, +\infty[$.

We present the results of the dynamical study of the new models and characterize the parameter space $\mathbf{R}$ of this maps family, according to their dynamical behaviour. The chaotic behaviour of the maps is measured in terms of the topological entropy, verifying that for each fixed value of the parameter p, the complexity of the inherent models increases with the Malthusian parameter r, along the seven distinct regions that form the parameter space, Theorem 5.1. The splitting of this parameter space into different regions has been made observing common properties relating to the dynamics that characterizes each region.

After the theoretical deduction of these models and the characterization of the respective family of unimodal maps, we analyze their behaviour as a function of the considered parameters ranges, defining a variation interval to them. Two interesting questions deserve special mention in these models: the negativity of the Schwarz derivative, for $1 < p < 2$, and the natural Allee effect for $p > 2$. The negativity of the Schwarz derivative, plays an important role in unidimensional dynamics, see for example [12]. This condition is violated in a small interval of the maps domain of this family $[0, c[\cup]c, 1]$, where c is the critical point of the map, when $p \in]1, 2[$, but indeed it does not affect the dynamical behaviour of the map.

A weak point of the logistic model and of the models proportional to the $Beta(p, 2)$ densities, with $p \in]1, 2[$, is the inexistence of Allee effect. We propose three models with Allee effect, which therefore can model casual growths of some species in a more realistic way in ecological terms. We investigate their dynamical behaviour and enunciate some important results involving these corrected models.

A detailed study, with complete proofs, can be found in [1].

[1] It is well known that, if X is a random variable with $Beta(p, q)$ distribution, denoted in what follows by $X \frown Beta(p, q)$, with $p > 0$ and $q > 0$, then the corresponding probability density function is

$$f_X(x) = \frac{1}{B(p, q)} x^{p-1}(1 - x)^{q-1} I_{(0,1)}(x),$$

where $B(p, q) = \int_0^1 t^{p-1}(1 - t)^{q-1}\, \mathrm{d}t, \forall p, q > 0$, is the Euler's beta function.

5.2 New Populational Growth Models Proportional to *Beta*$(p, 2)$ Densities, with $p \in]1, +\infty[$

An usual presentation of the Verhulst model starts with an approximation obtained by the linearization of a series expansion, [22]. In a similiar way, we present new models proportional to the *Beta*$(p, 2)$ densities, with $p \in \mathbb{N}\backslash\{1, 2\}$, [2] and [3]. These models can be useful to model populational growth of species whose evolution needs a greater growth rate than the one given by the Verhulst model. Assuming that the population size $N(t)$ has a series representation in the form

$$\begin{aligned}\frac{d}{dt}N(t) = A_0 + A_1 N(t) + \ldots + A_{p-1}N(t)^{p-1} \\ +A_p N(t)^p + A_{p+1}N(t)^{p+1} + \cdots ,\end{aligned}$$

truncating the terms of order smaller than $p - 1$ and the terms of order larger than p, considering that they are irrelevant to the model, i.e., $A_i = 0$, for $i \leq p - 2$ and $i \geq p + 1$, we obtain the simplified model:

$$\frac{d}{dt}N(t) = A_{p-1}N(t)^{p-1}\left(1 + \frac{A_p}{A_{p-1}}N(t)\right) \tag{5.1}$$

with $A_{p-1} > 0$ and $A_p < 0$. Writing $A_{p-1} = r^*$, the coefficient proportional to the instantaneous populational growth, and $K = -\frac{A_p}{A_{p-1}}$, the carrying capacity, the (5.1), which represents the populational growth rate, can be rewritten as follows:

$$\frac{d}{dt}N(t) = r^* N(t)^{p-1}\left(1 - \frac{N(t)}{K}\right),$$

with $p \in \mathbb{N}\backslash\{1, 2\}$. The discretization of these models is made as follows:

$$N(t_{n+1}) = r^* N(t_n)^{p-1}\left(1 - \frac{N(t_n)}{K}\right).$$

Considering that $x_n = \frac{N(t_n)}{K}$ and $r = r^* K^{p-2}$, the discretized model is given by:

$$x_{n+1} = r\, x_n^{p-1}(1 - x_n).$$

So, we consider the family of unimodal maps $f_{r,p} : [0, 1] \rightarrow [0, 1]$, with two parameters $p \in \mathbb{N}\backslash\{1, 2\}$ and $r > 0$ defined by:

$$f_{r,p}(x) = r\, x^{p-1}(1 - x). \tag{5.2}$$

The right-hand side of (5.2) is proportional to the beta density with shape parameters p and 2, denoted by *Beta*$(p, 2)$ density. We can generalize the discrete models given by (5.2), for any $p > 1$, as a natural extension.

Therefore, let us consider the family of unimodal maps $f_{r,p} : [0, 1] \to [0, 1]$, with two parameters p and r, whose maximum variation intervals are given respectively by $p \in]1, p_M]$ and $r \in]0, r(p_M)]$, defined by (5.2), and c is the critical point of $f_{r,p}$, which satisfies the following conditions:

- $f_{r,p} \in C^3([0, 1])$.
- $f'_{r,p}(x) \neq 0, \forall x \neq c$.
- $f'_{r,p}(c) = 0$ and $f''_{r,p}(c) < 0$, meaning that $f_{r,p}$ is strictly increasing in $[0, c\,[$ and strictly decreasing in $]\,c, 1\,]$.
- $f_{r,p}(0) = f_{r,p}(1) = 0$.
- $f_{0,p}(c) = 0$ and $f_{r(p_M),p}(c) = 1$.
- The Schwarz derivative of $f_{r,p}(x)$ is

$$S\left(f_{r,p}(x)\right) = \frac{f'''_{r,p}(x)}{f'_{r,p}(x)} - \frac{3}{2}\left(\frac{f''_{r,p}(x)}{f'_{r,p}(x)}\right)^2 < 0$$

$\forall x \neq c$, with $p > 2$ and $x > x_d$, with $1 < p < 2$.[2]

Note that, the parameter p has to be greater than one, since $f_{r,p}$ is unimodal. In this study, the maximum value considered for the parameter p, denoted by p_M, is the largest value for which we consider that the model can be realistic. The value $r(p_M)$ is the value of the parameter r corresponding to the full shift for $p = p_M$. In these maps $f_{r,p}$, r and p are both shape parameters, which are respectively related to the height and to the skewness of the curve. For any fixed $p > 1$, if $r = 0$ there is no curve; as the value of r increases, we get higher curves, until the value of r corresponds to the full shift, when the height of the curve attains its maximum value 1, see Fig. 5.1. Considering for each p the value of the parameter r for which we obtain the full shift, we conclude that the curve of the map $f_{r,p}$ can have three different patterns of skewness, as shown in Fig. 5.2.

So, in this work, we consider that $1 < p \leq p_M = 20$ and $0 < r \leq 53.001$. Observing Figs. 5.1 and 5.2, we can verify that the unimodal maps $f_{r,p}$, always have the fixed point $x^* = 0$, for any $r > 0$ and $p > 1$. However, seeing the cases of the maps corresponding to $p = 1.1$ and $p = 1.5$ in Fig. 5.2, we can verify that these maps have another positive fixed point besides 0. We can see that there are two more fixed points besides 0, for $p \in [2, p_M]$ and $r > r_1$. To the variation interval of p, r_1 is the first value of the parameter r for which it exists an orbit of period 1. For each fixed parameter value $p > 1$, the critical point of the map $f_{r,p}$ is always given by $c = \frac{p-1}{p}$.

[2] In this case, $S\left(f_{r,p}(x)\right)$ is not always negative in the all interval $I = [0, 1]$. In fact, the interval $[0, x_d]$, where the Schwarz derivative is positive, has a very small range because x_d is near 0, so the positivity of the Schwarz derivative occurs for values in the beginning of the interval, that do not disturb the dynamic behaviour of the map $f_{r,p}$.

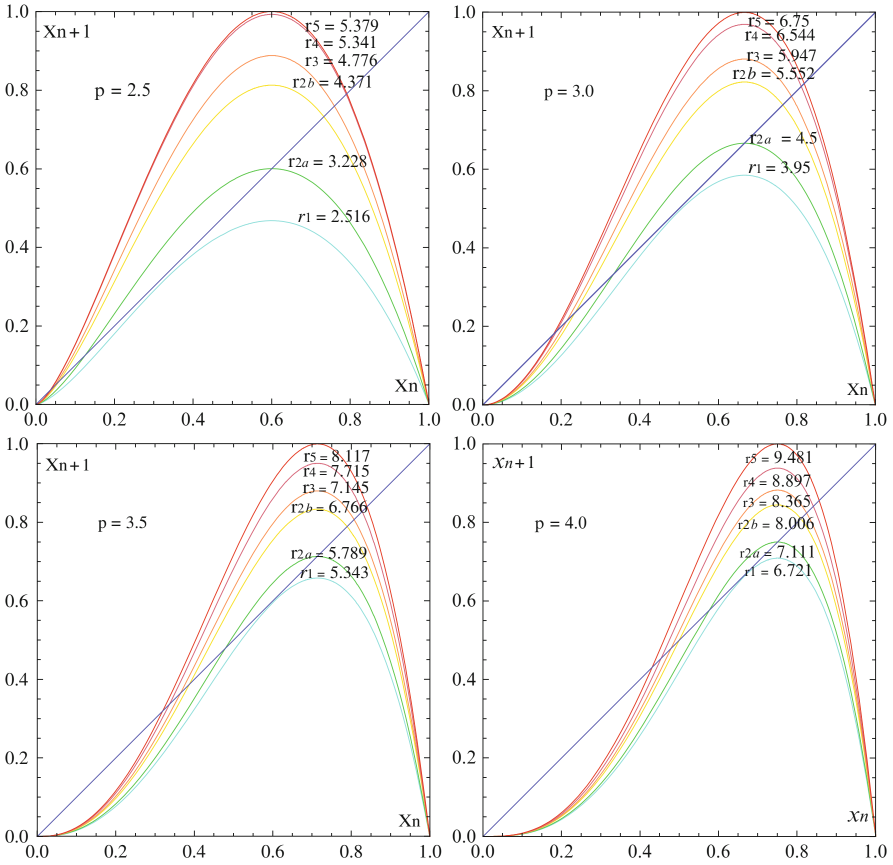


Fig. 5.1 Populational growth rates using models proportional to the *Beta*$(p, 2)$, $p = 2.5, 3, 3.5, 4$ densities

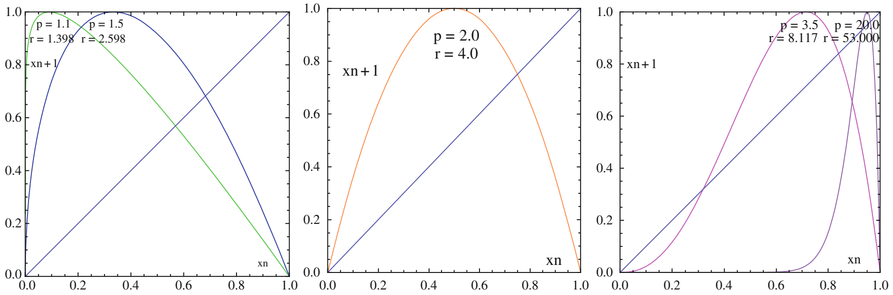


Fig. 5.2 Three types of format to the populational growth rates using the model proportional to the *Beta*$(p, 2)$ density, with $p > 1$: curve skewed to the left, symmetric and skewed to the right

5.3 Dynamical Behaviour on the Parameter Space

In this section, the parameter space, denoted by **R**, is divided into distinct regions, according to the dynamical behaviour of the unimodal maps belonging to the family of functions $f_{r,p}$ proportional to $Beta(p, 2)$ densities, with $1 < p \leq p_M = 20$ and $0 < r \leq 53.001$. Similar works can be seen in [18] and references therein.

The parameter space **R** is split into seven regions, each of them with a distinct behaviour, associated to a certain dynamic of populational evolution of eventual species. Then, for each studied value of p, going through the considered interval of variation for the parameter r, we determine the points $(r_i(p), p)$, with $i = 1, 2a, 2b, 3, 4, 5$. In Fig. 5.1 we can see the maps $f_{r_i(p),p}$, with $i = 1, 2a, 2b, 3, 4, 5$, for $p = 2.5, 3, 3.5, 4$.

In the exhaustive analysis below, we describe how we can define six lines, associating with each ordinate p a corresponding abcissa $r = r(p)$, which delimit seven regions, shown in Fig. 5.3, where iteration takes on different aspects. Observe that for each fixed p^*, the horizontal line (r, p^*) in the parameter space crosses those regions, and henceforth each horizontal line in that graphic corresponds to the summarized information of a Feigenbaum diagram for the map $f_{r,p}$, for a certain fixed p. See an example for $p = 4$ in Figs. 5.3 and 5.4.

5.3.1 Sudden Extinction Region $\mathbf{R_1}$

The first region, is defined by $\mathbf{R_1} = \{(r, p) : 0 < r < r_1(p), 2 \leq p \leq 20\}$. Its right boundary curve, which lies in $\mathbf{R_1}$, is the set of points with ordinates p, for $p \in [2, 20]$, whose abscissas, denoted by $r_1(p)$, are, for each p, the first values for which the iterates of the map $f_{r,p}$ are attracted to an unique positive fixed point.

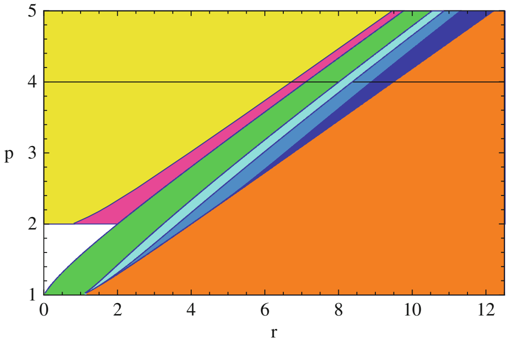


Fig. 5.3 Regions in the parameter space. $\mathbf{R_1}$ is the "triangle" in the upper left corner, $\mathbf{R_6}$ is the "triangle" in the lower right corner, $\mathbf{R_{2a}}$, $\mathbf{R_{2b}}$, $\mathbf{R_3}$, $\mathbf{R_4}$ and $\mathbf{R_5}$ are in between, in the above ordering. $(\{(r, p) : 1 < p < 2, 0 < r < r^*(p)\}$ is investigated in detail [3])

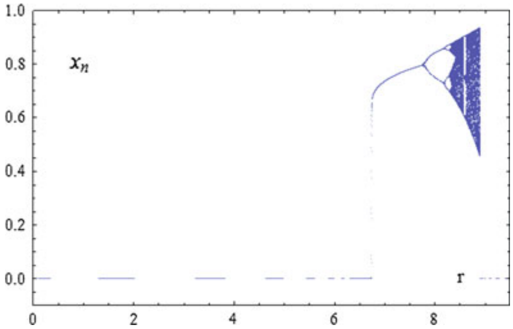


Fig. 5.4 Feigenbaum diagram for the model proportional to the *Beta*$(4, 2)$ density

This function $r_1(p)$, for $p \in [2, 20]$, defines a stable or attraction line. Note that, the curve $r_1(p)$ belong to the next region $\mathbf{R_2}$.

Globally, the iterates of any map $f_{r,p}$, whose parameters values belong to this region $\mathbf{R_1}$, are always attracted to the attractive fixed point $x^* = 0$. So, this is a region of extinction since a map $f_{r,p}$, with $(r, p) \in \mathbf{R_1}$ can model only species that will become extinct: as soon as they appear they are doomed to disappear. The growth rate it is not big enough to stabilize the population size. The unimodal maps $f_{r,p}$ of the region $\mathbf{R_1}$ do not have a chaotic behaviour, its topological entropy is null, [17]. The symbolic sequences associated to the orbits of the critical point $c = \frac{p-1}{p}$ are of the type CL^∞.

5.3.2 Stability or Equilibrium Region $\mathbf{R_2} = \mathbf{R_{2a}} \cup \mathbf{R_{2b}}$

The stability or equilibrium region is $\mathbf{R_2} = \{(r, p) : r_1(p) \leq r < r_2(p), 1 < p \leq 20\}$. In a generic way, we can say that the iterates of any map $f_{r,p}$, whose parameters values belong to the region $\mathbf{R_2}$, converge to the larger positive attractive fixed point (it is unique if $p \in]1, 2]$). So, this is a region of stability or equilibrium, since one map $f_{r,p}$, with $(r, p) \in \mathbf{R_2}$, can model populational evolutions of species whose size is approximately constant in time.

In fact, for each value of the parameter p, a drastic change is observed when $r \in [r_1(p), r_2(p)[$, resulting from the possibility of reaching the equilibrium between the two competitive forces, reproduction on one side and resources limitation on the other. The unimodal maps $f_{r,p}$ of the region $\mathbf{R_2}$ do not exhibit chaotic behaviour, its topological entropy being null, [17].

Remark 5.1. This region contains a super stable or super attractive curve, denoted by $r^*(p)$, whose expression is given by

$$r^*(p) = p\left(\frac{p}{p-1}\right)^{p-2}.$$

Therefore, this curve divides the region $\mathbf{R_2}$ in two sub-regions, denoted by $\mathbf{R_{2a}}$ and $\mathbf{R_{2b}}$, which are delimited by the curves $r_1(p)$, $r^*(p)$ and $r_2(p)$, respectively.

In this region, the symbolic sequences associated to the critical point orbits are of the type CL^{∞} in the region $\mathbf{R_{2a}}$, and in the region $\mathbf{R_{2b}}$ are of the type CR^{∞}. The second boundary curve $r_2(p)$ has points with ordinates p whose abscissas correspond, for each p, to the first value of the parameter r for which we can observe an orbit of period 2. Thus $r_2(p)$ is the curve where period doubling starts. Observe that this curve belongs to the next region $\mathbf{R_3}$.

5.3.3 *Period Doubling Region* $\mathbf{R_3}$

The third region, denominated period doubling region, is $\mathbf{R_3} = \{(r, p) : r_2(p) \leq r < r_3(p),\ 1 < p \leq 20\}$. In other words, the region $\mathbf{R_3}$ shows population dynamics patterns describing the generations of species oscillating in cycles of period 2^n, with $n \in \mathbb{N}$. The unimodal maps $f_{r,p}$ of the region $\mathbf{R_3}$ still do not exhibit chaotic behaviour, its topological entropy being null, [20].

Its right boundary curve $r_3(p)$ is the set of points whose ordinates p correspond to abscissas r where the map $f_{r,p}$ has no longer orbits of period 2^n, i.e., which do not correspond to Feigenbaum points, orbits of other periods starting at those values. Thus it is the chaos starting line, and it belongs to the next region $\mathbf{R_4}$.

5.3.4 *Chaotic Region* $\mathbf{R_4}$

The fourth region, denominated chaotic region and denoted by $\mathbf{R_4}$, is defined for $1 < p \leq 20$ and $r_3(p) \leq r < r_4(p)$. So, the iterates of the maps $f_{r,p}$ whose parameter values belong to the region $\mathbf{R_4}$ origin orbits of the several types, which already present chaotic patterns of behaviour; so its topological entropy is positive. The value of the topological entropy increases with the value of the parameter r, until it attains its maximum value $\ln 2$, [19].

The fourth boundary curve, $r_4(p)$, has the points with ordinates p whose abscissas correspond, for each p, to the first value of the parameter r exhibiting the natural Allee effect, [3]. The curve $r_4(p)$ is thus named line of the Allee effect, and belongs to the region $\mathbf{R_5}$.

5.3.5 *Allee Effect Caused Extinction Region* $\mathbf{R_5}$

The fifth boundary curve $r_5(p)$ has the points with ordinates p, with $p \in]1, 20]$, whose abscissas are the corresponding values of the parameter r where full shift does occur. This full shift line belongs to the closed region $\mathbf{R_5} = \{(r, p) : r_4(p) \leq r \leq r_5(p),\ 1 < p \leq 20\}$.

In a generic way, we can say that the iterates of the maps $f_{r,p}$, whose parameters values belong to this region $\mathbf{R_5}$, are always attracted to the attractive fixed point $x^* = 0$. Therefore, the maps $f_{r,p}$, with $(r, p) \in \mathbf{R_5}$, can model populational evolutions of species that once developed disorderly and now go to extinction, because few individuals remain and eventually they are spatiality far away from each other, so that reproduction chances diminish, leading to the extinction of these species. The unimodal maps $f_{r,p}$ of the region $\mathbf{R_5}$ exhibit chaotic behaviour, with maxim topological entropy $\ln 2$, [20].

5.3.6 Differed Extinction Region $\mathbf{R_6}$

$\mathbf{R_6} = \{(r, p) : r_5(p) < r \leq 53, 1 < p \leq 20\}$ is a differed extinction region. The graphic of any map $f_{r,p}$, with $(r, p) \in \mathbf{R_6}$, is no longer totally included in the invariant interval $[0, 1] \times [0, 1]$. The dynamic completely looses its deterministic component, and the size of the population in successive generations behaves as a random numbers generator device, until ultimate extinction does occur. At this stage, Cantor sets become observable.

The results described above, in order to characterize the topological complexity of the dynamical systems in each region, are stated in the following theorem measured in terms of topological entropy.

Theorem 5.1. *The topological entropy of the family of unimodal maps* $f_{r,p}(x) = rx^{p-1}(1-x)$*, with* $(r, p) \in \mathbf{R}$ *is characterized by:*

1. *In the regions* $\mathbf{R_1}$*,* $\mathbf{R_2}$ *and* $\mathbf{R_3}$*, the topological entropy is null.*
2. *In the region* $\mathbf{R_4}$*, the sets where the topological entropy is constant are connected and indexed in a strictly monotonous and continuous way by this topological invariant, except the null measure sets.*
3. *In the region* $\mathbf{R_5}$ *the topological entropy is constant and equal to its maximum value* $\ln 2$.

Proof. These claims follow easily from the fact that $f_{r,p} : [0, 1] \longrightarrow [0, 1]$ is a family of unimodal maps, having in mind the results in [17, 19, 20], respectively.

Values of the parameters larger than the ones considered above (i.e., $p > 20$) do not have a realistic meaning, since they are not adequate to model satisfactorily any known population. When the values of the parameter r become very large, we observe that:

- The range of the sudden extinction region increases considerably for large values of r.
- The ranges of the equilibrium region, of the period doubling region and of the chaotic region decrease to 0 (i.e., these regions tend to disappear) when r increases.
- The range of the region where the Allee effect exists increases slowly with r;
- The differed extinction region range decreases as a function of p.

Thus, for very large values of the parameter p, the populational growth pattern is extinction.

5.4 Deterministic Populational Growth Models with Allee Effect: Heuristic Approach

The essence of the heuristic approach, on which the deterministic models used to model populations growths of species that exhibit Allee effect are based, dates back at least to Odum and Allee (1954), see [21]. The *per capita* growth rate, dependent on the expected or observed population size, is modeled by a suitable function. We shall consider only one population, whose growth in a homogeneous environment is described by the ordinary differential equation:

$$\frac{dN(t)}{dt} = N(t)\, g\left(N(t)\right) \tag{5.3}$$

or by the difference equation:

$$N_{t+1} - N_t = N_t\, g\left(N_t\right). \tag{5.4}$$

In both cases N is the population size and $g(N)$ denote the *per capita* growth rate dependent of the size N, which is negative for decreasing populations and positive for increasing populations.

In this work, we consider the cases where the Allee effect occurs at small population sizes; the examples where the Allee effect occurs at large population sizes are few, see [14]. The *per capita* growth rate $g(N)$ that describes the Allee effect is an unimodal function, with a long tail; the maximum rate is obtained for only one positive dimension, $N = C > 0$. Below this "optimal" population size, positive effects of the presence of individuals of the same species prevail and $g(N)$ is increasing, while above this "optimal" population size, the negative dependence of the population size dominates and so the *per capita* growth rate, $g(N)$ is decreasing. Most of the models include overcapacity and avoid the indefinite growth by assuming a negative *per capita* growth rate, $g(N) < 0$, for a population size sufficiently large. In the aim of analyzing the stability, the values of the *per capita* growth rate $g(N)$ should have small oscillations near the equilibrium point (slightly increasing until the equilibrium point, and slightly decreasing soon after this value), and should be a continuous function for other values of the population size N, as in the two proposed models, (5.3) and (5.4), see [6]. In the difference equation, (5.4), the *per capita* growth rate should satisfy the condition $g(N_t) \geq -1$ so that we always have $N_t > 0$.

In these heuristic population growth models, (5.3) and (5.4), three basic settings can occur, see [6]:

Unconditional-Extinction (UE): If the Allee effect is too strong, the *per capita* growth rate $g(N)$ is negative for any population size N and the populations will inevitably become extinct, whatever the value of its initial size.

Extinction-Survival (ES): At moderate levels of the Allee effect, the *per capita* growth rate $g(N)$ is positive for intermediate values of the population size N, but is negative for very low or very high values of the population size N. Two equilibrium values exist: the smaller one, which is unstable, denoted by E^i, and the larger E^s which is locally stable. The population size at the instant 0, denoted by E^0, that is locally stable, is called trivial equilibrium. The populations whose dimension at 0 is smaller than the value E^i will become extinct, while those with dimension at 0 greater than the value E^i stabilize at the value E^s.

Unconditional-Survival (US): When the Allee effect gets weaker, the unstable equilibrium E^i vanishes, the trivial equilibrium E^0 becomes unstable every time the *per capita* growth rate $g(N)$ is positive for all the population size $N > 0$, and the population stabilizes in E^s, even though its *per capita* growth rate is still increasing with N at low populations sizes.

The setting ES is the most familiar consequence of the Allee effect, and the issue of extinction or survival of the population, is of utmost practical relevance, see [6].

5.5 Models Based on Maps Proportional to the *Beta*$(p, 2)$ Densities, with $p \in]1, 2[$ and Allee Effect

As it happens in the classical logistic model, and also for the models proportional to *Beta*$(p, 2)$ densities, with $1 < p < 2$, the inexistence of a rarefaction critical dimension E, and consequent inexistence of the Allee effect is a drawback that can be corrected as described below. We follow closely what is usually done for the logistic model.

5.5.1 Logistic Map Modified with Allee Effect

Several criticisms have been made to the logistic model, which is frequently used to model the population growth of certain species. One of these criticisms is related to the fact that this model does not implement the Allee effect. Indeed, the logistic equation assumes that the population always increases, even when its dimension is low; besides, in this case (small population size), this model assumes a fast population increase. At first sight, this could seem acceptable because the environment resources are abundant to the few individuals in the population. However, this assumption is questionable, since for many populations there is a minimal population size (rarefaction critical density), denoted by E, required for reproduction. Below E, the probability that individuals of opposite sexes effectively meet for reproduction is so small that the population can not recover its dimension in order to substitute those who die, and finally becomes extinct. In this case, the instantaneous growth rate r is negative. Above E, the probability of the individuals meeting mates for reproduction is large enough for the population to grow until its carrying capacity K. In this situation, the instantaneous growth rate r is positive. This minimal population size E corresponds to a null growth rate, which allows that the population maintains exactly its dimension at a fixed value. At this density E, the population is incapable to grow up and maintains its equilibrium value, until some disturbance happens, leading either to extinction or to growth. Obviously the rarefaction critical density E is smaller than the carrying capacity K. Between E and K there is a variety of populations dimensions for which the instantaneous growth rate is positive.

The inexistence of a rarefaction critical density E, and consequent inexistence of the Allee effect in the logistic model, is a drawback that can be corrected. Several investigators discussed this issue (see for example the ones mentioned by [6]), suggesting various models for the *per capita* growth rate. The basic idea is to introduce a factor in the classic logistic model, $T\,(N(t))$, forcing the rate $g\,(N(t))$ to be negative as soon as the population size $N(t)$ is smaller than the rarefaction critical density E:

$$g\,(N(t)) = \frac{dN(t)}{dt}\frac{1}{N(t)} = r\left(1 - \frac{N(t)}{K}\right)T\,(N(t))\,.$$

5.5.2 *Models Based on Maps Proportional to* $Beta(p, 2)$ *Densities, with* $p \in\ ]1, 2[$ *and Allee Effect*

Using a similar procedure to the one used in order to correct the logistic model, we are going to deduce three new models for the *per capita* growth rate. So, the basic idea is to introduce in the models proportional to *Beta*$(p, 2)$ densities, with $1 < p < 2$, a new factor $T\,(N(t))$, in such a way that this rate, $g^*\,(N(t))$, is negative as soon as the population size $N(t)$ becomes smaller than the rarefaction critical density E:

$$g^*\,(N(t)) = \frac{dN(t)}{dt}\frac{1}{N(t)} = r^* N(t)^{p-2}\left(1 - \frac{N(t)}{K}\right)T\,(N(t))\,.$$

Therefore, using the same three factors $T\,(N(t))$ suggested by several authors for the logistic model, see [6], we obtain the three maps to model the *per capita* growth rate of a population.

1. Using the factor $T\,(N(t)) = 1 - \frac{E}{N(t)}$ suggested in [7, 9–11, 24] to inforce the Allee effect in the logistic model, we obtain the following function for the *per capita* growth rate:

$$g_1^*\,(N(t)) = r^* N(t)^{p-2}\left(1 - \frac{N(t)}{K}\right)\left(1 - \frac{E}{N(t)}\right) \tag{5.5}$$

and consequently, the corresponding population growth rate is given by:

$$f_1^*\,(N(t)) = N(t)\;g_1^*\,(N(t)) = r^* N(t)^{p-1}\left(1-\frac{N(t)}{K}\right)\left(1-\frac{E}{N(t)}\right). \tag{5.6}$$

The model (5.6) can be discretized, in order to have the properties allowing its study using the symbolic dynamic methods. Two important points: the sign of the Schwarz derivative is not always negative, for $1 < p < 2$, and for certain instantaneous growth rates r the discrete maps that represent the growth rate with Allee effect take values out of the invariant interval, suggesting the need to study Cantor sets. The discretized model, designated by Model 1, can be obtained from the differential equation (5.6), considering that $x_n = \frac{N(t_n)}{K}$ and $r = r^* K^{p-2} > 0$, in the following way:

$$N(t_{n+1}) = f_1^*\,(N(t_n)) \quad\Longleftrightarrow\quad x_{n+1} = r x_n^{p-2}\,(1-x_n)\left(x_n - \frac{E}{K}\right).$$

Therefore, Model 1 corrected with Allee effect is a map $h_1^* : [0,1] \to \mathbb{R}$, defined by:

$$h_1^*(x) = r x^{p-2}\,(1-x)\left(x-\frac{E}{K}\right).$$

2. If we use the factor $T\,(N(t)) = \frac{N(t)}{K} - \frac{E}{K}$ used in [4, 5, 15, 16] to correct the inexistence of the Allee effect in the classical logistic model, we have:

$$g_2^*\,(N(t)) = r^* N(t)^{p-2}\left(1-\frac{N(t)}{K}\right)\left(\frac{N(t)}{K}-\frac{E}{K}\right) \tag{5.7}$$

and therefore, the corresponding population growth rate is given by:

$$f_2^*\,(N(t)) = N(t)\;g_2^*\,(N(t)) = r^* N(t)^{p-1}\left(1-\frac{N(t)}{K}\right)\left(\frac{N(t)}{K}-\frac{E}{K}\right). \tag{5.8}$$

The discretized model, designated by Model 2, can be obtained from (5.8) as in the previous model:

$$N(t_{n+1}) = f_2^*\,(N(t_n)) \quad\Longleftrightarrow\quad x_{n+1} = r x_n^{p-1}\,(1-x_n)\left(x_n - \frac{E}{K}\right).$$

So, Model 2 corrected with Allee effect is a map $h_2^* : [0,1] \to \mathbb{R}$, defined by:

$$h_2^*(x) = r x^{p-1}\,(1-x)\left(x-\frac{E}{K}\right).$$

3. A third possibility is to use the factor $T\left(N(t)\right) = \frac{N(t)}{E} - 1$ suggested in [8, 13], getting the following function to model the *per capita* growth rate:

$$g_3^*\left(N(t)\right) = r^* N(t)^{p-2}\left(1 - \frac{N(t)}{K}\right)\left(\frac{N(t)}{E} - 1\right) \tag{5.9}$$

and so, the corresponding population growth rate is given by:

$$f_3^*\left(N(t)\right) = N(t)\, g_3^*\left(N(t)\right) = r^* N(t)^{p-1}\left(1 - \frac{N(t)}{K}\right)\left(\frac{N(t)}{E} - 1\right). \tag{5.10}$$

The discretized model, designated by Model 3, obtained from (5.10), is given by:

$$N(t_{n+1}) = f_3\left(N(t_n)\right) \quad \Longleftrightarrow \quad x_{n+1} = r x_n^{p-1}\left(1 - x_n\right)\left(\frac{K}{E}x_n - 1\right).$$

So, Model 3 corrected with Allee effect is a map $h_3^* : [0,1] \to \mathbb{R}$, defined by:

$$h_3^*(x) = r x^{p-1}\left(1 - x\right)\left(\frac{K}{E}x - 1\right).$$

5.6 Characterization of the New Models

We now emphasize some properties of the new models presented above. To accomplish it, in this section we established two propositions and state some important notes about the characteristics of these models.

Proposition 5.1. *The models based on the maps proportional to the Beta*$(p, 2)$ *densities, with* $p \in\ ]1, 2[$*, modified with the Allee effect,* h_i^**, with* $i = 1, 2, 3$*, verify the following propositions:*

1. *The conditions of the setting* ES *(Extinction-Survival), which is the more usual consequence of the Allee effect, are satisfied by the three models presented.*
2. *The conditions of the setting* US *(Unconditional-Survival), are satisfied by the Models* 1 *and* 2*, but not by the Model* 3.
3. *None of these models satisfies the conditions of the setting* UE *(Unconditional-Extinction).*

Proof. 1. Having in mind the conditions of the setting ES, it follows that for $1 < p < 2$ the *per capita* growth rates pertaining to any of these models, given respectively by (5.5), (5.7) and (5.9), are positive if and only if $r^* > 0 \wedge 0 < E < N(t) < K$.

2. (a) In what concerns Model 1, the *per capita* growth rate $g_1^*\left(N(t)\right)$, is given by expression (5.5); considering $E = N(0) = 0$, we get

$$g_1^* \left(N(t)\right) = r^* N(t)^{p-2} \left(1 - \frac{N(t)}{K}\right).$$

For $1 < p < 2$, as $0 \leq E \leq K$ and $0 \leq N(t) \leq K$, we have $N(t)^{p-2} \geq 0$ and $0 \leq \frac{N(t)}{K} \leq 1$, so $1 - \frac{N(t)}{K} \geq 0$. As $r* > 0$, it follows that $g_1^* \left(N(t)\right) \geq\ 0$.

(b) In what concerns Model 2, the *per capita* growth rate $g_2^* \left(N(t)\right)$, given by expression (5.7), if $E = N(0) = 0$, then we have

$$g_2^* \left(N(t)\right) = r^* N(t)^{p-2} \left(1 - \frac{N(t)}{K}\right) \frac{N(t)}{K}.$$

For $1 < p < 2$, as $0 \leq E \leq K$ and $0 \leq N(t) \leq K$, then $N(t)^{p-2} \geq 0$ and $0 \leq \frac{N(t)}{K} \leq 1$, therefore $1 - \frac{N(t)}{K} \geq 0$. Having in mind that $r* > 0$, we conclude that $g_2^* \left(N(t)\right) \geq 0$.

(c) In what concerns Model 3, the *per capita* growth rate $g_3^* \left(N(t)\right)$ is (5.9), and for $1 < p < 2$, considering $E = N(0) = 0$, the expression of $g_3^* \left(N(t)\right)$ makes no sense because it includes the ratio $\frac{N(t)}{E}$ which has no meaning. So, the Model 3 does not satisfy the conditions for the setting US.

3. For any of the three models, the *per capita* growth rate $g_i^*(N(t))$, with $i = 1, 2, 3$, is not negative for all the population sizes $N(t)$. In fact, $g_i^*(N(t)) < 0$, for $1 < p < 2$, if and only if $r^* > 0 \wedge 0 < E < K \wedge [(N(t) > K) \vee (0 < N(t) < E)]$.
The condition $N(t) > K$ is impossible, because K corresponds to the carrying capacity. So, if $r^* > 0$ then $g_i^*(N(t))$ is negative only if $0 < N(t) < E$, and therefore it is not negative to any $N(t)$. Therefore, none of those three models satisfy the setting UE.

So, we can state that Models 1 and 2 for the *per capita* growth rate are more flexible than Model 3, because this one only satisfies the conditions of one setting while the other two models satisfy the conditions of two settings. The Schwarz derivatives of these models verify the following result:

Proposition 5.2. *The Schwarz derivatives of the Models* h_i^*, *with* $i = 1, 2, 3$, *do not depend on the value of the Malthusian parameter* r *and satisfy* $S_{h_2^*}(x) = S_{h_3^*}(x)$.

Proof. Having in mind the expressions of the Schwarz derivatives for these models, which are given by:

$$S_{h_1^*}(x) = -\frac{-2EK(-2+p)px\left(3+4p(-1+x)+p^2(-1+x)^2+4x-x^2\right)}{2x^2\left(K(1+p(-1+x))x+E(-2+p+x-px)\right)^2}$$

$$-\frac{E^2(2-3p+p^2)\left(6+p^2(-1+x)^2-p(5-6x+x^2)\right)}{2x^2\left(K(1+p(-1+x))x+E(-2+p+x-px)\right)^2}$$

$$-\frac{K^2(-1+p)px^2(2+p^2(-1+x)^2+4x+p(-3+2x+x^2))}{2x^2\left(K(1+p(-1+x))\,x+E(-2+p+x-px)\right)^2}$$

$$
\begin{aligned}
S_{h_2^*}(x) = S_{h_3^*}(x) = &-\frac{-2EK(-1+p^2)x\left(p^2(-1+x)^2+6x+2p(-1+x^2)\right)}{2x^2\left(Kx\left(p(-1+x)+x\right)+E(-1+p-px)\right)^2}\\
&-\frac{E^2(-1+p)p\left(2+p^2(-1+x)^2+4x+p(-3+2x+x^2)\right)}{2x^2\left(Kx\left(p(-1+x)+x\right)+E(-1+p-px)\right)^2}\\
&-\frac{K^2p(1+p)x^2\left(p^2(-1+x)^2+2x(2+x)+p(-1-2x+3x^2)\right)}{2x^2\left(Kx\left(p(-1+x)+x\right)+E(-1+p-px)\right)^2}
\end{aligned}
$$

we observe that none of them depends on r, and $S_{h_2^*}(x) = S_{h_3^*}(x)$.

Remark 5.2. For any of the models h_i^*, with $i = 1, 2, 3$, the negativity of the Schwarz derivative is not verified in all the interval [0, 1]. This unimodal maps property is not satisfied in a subinterval $[0, x_{di}] \subset [0, 1]$. This positivity of the Schwarz derivative near the origin is due to the fact that the first three derivatives of each one of the models h_i^*, with $i = 1, 2, 3$, go to ∞ when x goes to 0. Moreover, the value of the point x_{di} depends on the model h_i^*, which is associated to a parameter $p \in]1, 2[$, and it also depends on the values of E and K.

Acknowledgements This research has been partially supported by FCT/POCTI, POCI/FEDER, CEAUL and ISEL.

References

1. Aleixo, S.M.: Métodos analíticos em probabilidades e métodos probabilísticos em análise: fractalidade associada aos modelos *Beta*(p, q), evolução de populações e dimensões de Hausdorff, PhD Dissertation, University of Lisbon, (2008)
2. Aleixo, S.M., Rocha, J.L., Pestana, D.D.: Populational growth models in the light of symbolic dynamics. Proceedings of 30th International Conference on Information Technology, ITI 2008, 311–316 (2008)
3. Aleixo, S.M., Rocha, J.L., Pestana, D.D.: Populational growth models proportional to beta densities with Allee effect. Mathematical models in engineering, biology and medicine, AIP Conf. Proc., Amer. Inst. Phys. **1124**, 3–12 (2009)
4. Amarasekare, P.: Allee effects in metapopulation dynamics. Am. Nat. **152**, 298–302 (1998a)
5. Amarasekare, P.: Interactions between local dynamics and dispersal: insights from single species models. Theor. Popul. Biol. **53**, 44–59 (1998b)
6. Boukal, D.S., Berec, L.: Single-species models of the Allee effect: Extinction boundaries, sex ratios and mate encounters. J. Theor. Biol. **218**, 375–394 (2002)
7. Brassil, C.E.: Mean time to extinction of a metapopulation with an Allee effect. Ecol. Model **143**, 9–13 (2001)
8. Courchamp, F., Clutton-Brock, T.H., Grenfell, B.: Inverse density dependence and the Allee effect. Trends Ecol. Evol. **14**, 405–410 (1999a)
9. Courchamp, F., Grenfell, B., Clutton-Brock, T.H.: Population dynamics of obligate cooperators. Proc. R. Soc. Lond. B **266**, 557–563 (1999b)
10. Courchamp, F., Clutton-Brock, T.H., Grenfell, B.: Multipack dynamics and the Allee effect in the African wild dog, Lycaon pictus. Anim. Conservat. **3**, 277–285 (2000a)
11. Courchamp, F., Grenfell, B., Clutton-Brock, T.H.: Impact of natural enemies on obligately cooperative breeders. Oikos **91**, 311–322 (2000b)

12. Graczyk, J., Swiatek G., Sands, D.: La drive schwarzienne en dynamique unimodale. C. R. Acad. Sci. Paris **332**, srie I, 329–332 (2001)
13. Gruntfest, Y., Arditi, R., Dombrovsky, Y.: A fragmental population in a varying environment. J. Theor. Biol. **185**, 539–547 (1997)
14. Kokko, H., Sutherland, W.J.: Ecological traps in changing environments: ecological and evolutionary consequences of a behaviourally mediated Allee effect. Evol. Ecol. Res. **3**, 537–551 (2001)
15. Keitt, T.H., Lewis, M.A., Holt, R.D.: Allee effects, invasion pinning and species' borders. Am. Nat. **157**, 203–216 (2001)
16. Lewis, M.A., Kareiva, P.: Allee dynamics and the spread of invading organisms. Theor. Popul. Biol. **43**, 141–158 (1993)
17. Lind, D., Marcus, B.: An Introduction to Symbolic Dynamics and Codings. Cambridge University Press, Cambridge (1995)
18. Lopez-Ruiz, R., Fournier-Prunaret, D.: Indirect Allee effect, bistability and chaotic oscillations in a predator–prey discrete model of logistic type. Chaos Solitons Fractals **24**, 85–101 (2005)
19. Melo, W., van Strien, S.: One-Dimensional Dynamics. Springer, New York (1993)
20. Milnor, J., Thurston, W.: On iterated maps of the interval. In : Alexander, J.C. (ed.) Proceedings University Maryland 1986–1987. Lecturer Notes in Math, **1342**. Springer, Berlin (1988)
21. Odum, H.T., Allee, W.C.: A note on the stable point of populations showing both intraspecific cooperation and disoperation. Ecology **35**, 95–97 (1954)
22. Pestana, D., Velosa, S.: Introdução à Probabilidade e à Estatística, vol. 1, 3a edição. Fundação Calouste Gulbenkian, Lisboa (2008)
23. Verhulst, P.F.: Recherches mathématiques sur la loi d'accroissement de la population. Nouv. Mém. de l'Academie Royale des Sci. et Belles-Lettres de Bruxelles **18**, 1–41 (1845)
24. Wilson, E.O., Bossert, W.H.: A Primer of Population Biology. Sinauer Associates Sunderland, MA, U.S.A. (1971)

Chapter 6
Power Indices Applied to Portuguese Parliament

José M. Alonso-Meijide, Flávio Ferreira, Mikel Álvarez-Mozos, and Alberto A. Pinto

Abstract In this paper, we apply the following four power indices to the Portuguese Parliament: Shapley–Shubik index, Banzhaf index, Deegan–Packel index and Public Good Index. We also present the main concepts related with simple games and discuss the features of each power index by means of their axiomatic characterizations.

6.1 Introduction

The problem of assessing an a priori distribution of power among the members of decision making bodies is addressed mainly using game theoretical tools. It is indeed a very useful application of mathematics to social sciences. Simple games are used to model decision making bodies. A simple game is a special kind of

J.M. Alonso-Meijide (✉)
Department of Statistics and Operations Research, Faculty of Sciences of Lugo, University of Santiago de Compostela, Santiago de Compostela, Spain
e-mail: josemaria.alonso@usc.es

F. Ferreira
ESEIG, Instituto Politécnico do Porto, R. D. Sancho I, 981, 4480-876 Vila do Conde, Portugal
e-mail: flavioferreira@eu.ipp.pt

M. Álvarez-Mozos
Department of Statistics and Operations Research, Faculty of Mathematics, University of Santiago de Compostela, Santiago de Compostela, Spain
e-mail: mikel.alvarez@usc.es

A.A. Pinto
LIAAD-INESC Porto LA e Departamento de Matemática, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, 687, 4169-007, Portugal
and
Centro de Matemática e Departamento de Matemática e Aplicações, Escola de Ciências, Universidade do Minho, Campus de Gualtar, 4710-057 Braga, Portugal
e-mail: aapinto@fc.up.pt

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_6,

cooperative game where the worth of each coalition is either 1 if the coalition can pass a bill independently of what the remaining voters do, or 0 if they can't pass the bill. In this framework power indices are used as a measure of the ability of each player to transform a losing coalition into a winning one. In the literature one can find many different such power indices, each of them satisfying different sets of properties, and there is little to no consensus on which choice is the most appropriate in a particular context.

The first proposed power index can be found in Shapley and Shubik [11], where the Shapley value [10] is reinterpreted in the context of simple games giving rise to the Shapley–Shubik index. Another important power index is the Banzhaf index, which was proposed by Banzhaf [3]. Both indices are based on the swings of a player. A winning coalition is called a swing for a player if the removal of this player from the coalition would turn the coalition into a losing one. In Banzhaf's model the power of an agent is proportional to his number of swings, whereas the Shapley–Shubik index is a weighted sum of the swings of a player where the weights are sensitive to the size of the coalition.

Other important power indices include the Deegan–Packel index [4] and the Public Good Index [7], which are based on minimal winning coalitions. A winning coalition is a minimal winning coalition when all its members are critical, that is to say, when the removal of any member from the coalition would turn it into a losing coalition. Indeed, the set of minimal winning coalitions is enough to describe a simple game. The Deegan–Packel index assumes that all minimal winning coalitions are equally likely and that all players belonging to a minimal winning coalition have the same power. Alternatively, the Public Good Index is determined by the number of minimal winning coalitions containing a given voter divided by the sum of such numbers across all the voters.

In this paper we first present the above mentioned power indices and discuss the properties that characterize them. The characterizations of power indices are especially interesting since they provide an appealing set of properties that if accepted as reasonable, make each power index unique. Finally, we analyze the distribution of power in the Portuguese Parliament using power indices and present our conclusions.

6.2 Preliminaries

In this section, we provide the main ideas behind simple games and power indices. In particular, we recall the definitions of the Shapley–Shubik index, the Banzhaf index, the Deegan–Packel index and the Public Good Index.

6.2.1 Simple Games

A characteristic function game is a pair (N, v), where $N = \{1, \ldots, n\}$ is the set of players and v, the characteristic function, is a real function on $2^N = \{S : S \subseteq N\}$

with $v(\emptyset) = 0$. A subset $S \subseteq N$ is called a coalition. Shorthand notation will be used and $S \cup \{i\}$ and $S \backslash \{i\}$ will be denoted by $S \cup i$ and $S \backslash i$.

A *null player* in a game (N, v) is a player $i \in N$ such that $v(S \cup i) = v(S)$ for all $S \subseteq N \backslash i$. Two players $i, j \in N$ are *symmetric* in a game (N, v) if $v(S \cup i) = v(S \cup j)$ for all $S \subseteq N \backslash \{i, j\}$.

An important subclass of characteristic function games is the class of simple games. A *simple game* is a characteristic function game (N, v) such that:

- $v(S) \in \{0, 1\}$ for every $S \subseteq N$.
- v is a monotone function, that is, $v(S) \leq v(T)$, for every $S \subseteq T \subseteq N$.
- $v(N) = 1$.

$SI(N)$ denotes the set of simple games with player set N. In a simple game (N, v), a coalition $S \subseteq N$ is *winning* if $v(S) = 1$, and S is *losing* if $v(S) = 0$. $W(v)$ denotes the set of winning coalitions of the game (N, v) and $W_i(v)$ the subset of $W(v)$ formed by coalitions $S \subseteq N$ such that $i \in S$. A winning coalition $S \subseteq N$ is a *minimal winning* coalition if every proper subset of S is a losing coalition, that is, S is a minimal winning coalition in (N, v) if $v(S) = 1$ and $v(T) = 0$ for any $T \subset S$. $M(v)$ denotes the set of minimal winning coalitions of the game (N, v) and $M_i(v)$ the subset of $M(v)$ formed by coalitions $S \subseteq N$ such that $i \in S$.

Given a simple game (N, v), a *swing* for a player $i \in N$ is a coalition $S \subseteq N$ such that $S \backslash i$ is a losing coalition and S is a winning one. $\eta_i(v)$ denotes the set of swings for player $i \in N$. A winning coalition $S \subseteq N$ is a minimal winning coalition if and only if $S \in \eta_i(v)$ for every $i \in S$.

Given a family of simple games $H \subseteq SI(N)$, a *power index* on H is a function f, which assigns to every simple game $(N, v) \in H$ a vector

$$(f_1(N, v), \ldots, f_n(N, v)) \in \mathbb{R}^n,$$

where the real number $f_i(N, v)$ is the "power" of the player i in the game (N, v) according to f. The power index of a simple game can be interpreted as a measure of the ability of the different players to turn a losing coalition into a winning one. It is useful to single out a list of desirable properties a power index may satisfy.

- A power index f satisfies the *null player* property if $f_i(N, v) = 0$ for every $(N, v) \in H$ and every null player $i \in N$.
- A power index f is *symmetric* if $f_i(N, v) = f_j(N, v)$ for every $(N, v) \in H$ and for every pair of symmetric players $i, j \in N$.
- A power index f is *efficient* if $\sum_{i \in N} f_i(N, v) = 1$ for every $(N, v) \in H$.

Young [12] proposed the strong monotonicity property.

- A power index f satisfies *strong monotonicity* if $f_i(N, v) \geq f_i(N, w)$ for every pair of games $(N, v), (N, w) \in H$ and for all $i \in N$ such that $v(S \cup i) - v(S) \geq w(S \cup i) - w(S)$ for all $S \subseteq N \backslash i$.

A set of independent properties (an axiomatic system) is a convenient tool to decide on the use of an index. The indices of Shapley–Shubik, Banzhaf, Deegan–Packel, and Public Good Index are efficient, symmetric, and satisfy the null player property.

6.2.2 Shapley–Shubik Index

Given a simple game (N, v), the Shapley–Shubik power index assigns to each player $i \in N$ the real number

$$\varphi_i(N, v) = \sum_{S \in \eta_i(v)} \frac{(s-1)!\,(n-s)!}{n!},$$

where s is the number of members in S. Given n players, $n!$ is the number of permutations, $(s-1)!\,(n-s)!$ counts the permutations that maintain members of S consecutively.

In the class of simple games, the additivity property introduced by Shapley [10] does not apply because the sum of two simple games is not a simple game. Dubey [5] proposed the transfer property as a substitute of the additivity property and characterized the Shapley value in this class of games.

- A power index f on $H \subseteq SI(N)$ satisfies the *transfer* property if for all (N, v), $(N, w) \in H$ such that $(N, v \vee w)$, $(N, v \wedge w) \in H$, $f(N, v \vee w) + f(N, v \wedge w) = f(N, v) + f(N, w)$ where for all $S \subseteq N$

$$(v \vee w)(S) = \max\{v(S), w(S)\} \text{ and } (v \wedge w)(S) = \min\{v(S), w(S)\}.$$

The characterization is presented below.

- The unique power index on $SI(N)$ that satisfies transfer, null player, symmetry, and efficiency is the Shapley–Shubik index.

6.2.3 Banzhaf Index

Given a simple game (N, v), the non-normalized Banzhaf index assigns to each player $i \in N$ the real number:

$$\beta_i'(N, v) = \frac{|\eta_i(v)|}{2^{n-1}}.$$

Dubey and Shapley [6] characterized the Banzhaf index in a similar way to that introduced by Shapley and Shubik to characterize the Shapley–Shubik index. They use a property of total power instead of efficiency. The total power property states that power of players adds up to the total number of swings divided by the number of coalitions which can join to player $i \in N$.

– A power index f defined on $H \subseteq SI(N)$ satisfies the *total power* property if $\sum_{i \in N} f_i(N, v) = \overline{\eta}(v)/2^{n-1}$, for every simple game $(N, v) \in H$, where $\overline{\eta}(v) = \sum_{i \in N} |\eta_i(v)|$.

- The unique power index on $SI(N)$ that satisfies transfer, null player, symmetry, and total power is the Banzhaf index.

To achieve efficiency a normalized version of the Banzhaf index is considered. Given a simple game (N, v), the Banzhaf index assigns to each player $i \in N$ the real number:

$$\beta_i(N, v) = \frac{\beta_i'(N, v)}{\sum_{j \in N} \beta_j'(N, v)}.$$

6.2.4 Deegan–Packel Index

The power index introduced in Deegan and Packel [4] assumes that

(a) Only minimal winning coalitions will emerge victorious.
(b) Each minimal winning coalition has an equal probability of forming.
(c) Players in a minimal winning coalition divide the "spoils" equally.

These assumptions seem reasonable in a wide variety of situations. The assumptions determine the Deegan–Packel index. Given a simple game (N, v), this index assigns to each player $i \in N$ the real number:

$$\rho_i(N, v) = \frac{1}{|M(v)|} \sum_{S \in M_i(v)} \frac{1}{|S|}.$$

The Deegan–Packel index of a player i is equal to the sum of the inverse of the cardinality of S for the coalitions $S \in M_i(v)$, divided by the cardinality of $M(v)$ in order to achieve normalization.

The Deegan–Packel index does not satisfy the transfer property, but it satisfies the property of DP-mergeability. Two simple games (N, v) and (N, w) are *mergeable* if for all pair of coalitions $S \in M(v)$ and $T \in M(w)$, it holds that $S \nsubseteq T$ and $T \nsubseteq S$. The minimal winning coalitions in game $(N, v \vee w)$ are precisely the union of the minimal winning coalitions in games (N, v) and (N, w). If two games (N, v) and (N, w) are mergeable, the mergeability condition guarantees that $|M(v \vee w)| = |M(v)| + |M(w)|$.

– A power index f on $H \subseteq SI(N)$ satisfies the *DP-mergeability* property if for any pair of mergeable simple games $(N, v), (N, w) \in H$ such that $(N, v \vee w) \in H$, it holds that for every player $i \in N$:

$$f_i\left(N, v \vee w\right) = \frac{\left|M\left(v\right)\right| f_i\left(N, v\right) + \left|M\left(w\right)\right| f_i\left(N, w\right)}{\left|M\left(v \vee w\right)\right|}.$$

This property states that the power in a merged game is a weighted mean of the power in the two component games, where the weights come from the number of minimal winning coalitions in each component game, divided by the number of minimal winning coalitions in the merged game. Deegan and Packel [4] characterized ρ as follows.

- The unique power index on $SI\left(N\right)$ that satisfies DP-mergeability, null player, symmetry, and efficiency is the Deegan–Packel power index.

Lorenzo–Freire et al. [9] characterized the Deegan–Packel index replacing the property of DP-mergeability with the property of DP-minimal monotonicity.

– A power index f defined on $H \subseteq SI(N)$ satisfies the property of *DP-minimal monotonicity* if for any pair of games (N, v), $(N, w) \in H$, it holds that for each player $i \in N$ such that $M_i(v) \subseteq M_i(w)$,

$$f_i(N, w)|M(w)| \geq f_i(N, v)|M(v)|.$$

i.e., if the set of minimal winning coalitions containing a player $i \in N$ in game (N, v) is a subset of minimal winning coalitions containing this player in game (N, w), then the power of player i in game (N, w) is not less than power of player i in game (N, v) (first, this power must be normalized by the number of minimal winning coalitions in games (N, v) and (N, w)).

- The unique power index on $SI(N)$ that satisfies DP-minimal monotonicity, null player, symmetry, and efficiency, is the Deegan–Packel power index.

6.2.5 Public Good Index

The Public Good Index, introduced in Holler [7], considers that only minimal winning coalitions are relevant when it comes to measuring power. Then, given a simple game (N, v), the Public Good Index assigns to each player $i \in N$ the real number:

$$\delta_i\left(N, v\right) = \frac{\left|M_i\left(v\right)\right|}{\sum_{j \in N}\left|M_j\left(v\right)\right|}.$$

The Public Good Index of a player i is equal to the total number of minimal winning coalitions containing player i, divided by the sum of these numbers over all players.

An axiomatic characterization of this index can be found in Holler and Packel [8]. This characterization has qualities similar to the characterization of the Deegan–Packel index with the property of DP-mergeability.

– A power index f defined on $H \subseteq SI(N)$ satisfies the *PGI-mergeability* property if for any pair of mergeable simple games $(N, v), (N, w) \in H$ such that $(N, v \vee w) \in H$, it holds that for all player $i \in N$:

$$f_i\left(N, v \vee w\right) = \frac{f_i\left(N, v\right) \sum_{j \in N} \left|M_j\left(v\right)\right| + f_i\left(N, w\right) \sum_{j \in N} \left|M_j\left(w\right)\right|}{\sum_{j \in N} \left|M_j\left(v \vee w\right)\right|}.$$

- The unique power index on $SI(N)$ that satisfies PGI-mergeability, null player, symmetry, and efficiency is the Public Good Index.

A new characterization of Public Good Index is provided in Alonso-Meijide et al. [1], using a property similar to strong monotonicity [12] instead of PGI-mergeability. This property is named PGI-minimal monotonicity. It describes the relation between the power indices of two games, (N, v) and (N, w), in terms of the sizes of the sets of minimal winning coalitions.

– A power index on $H \subseteq SI(N)$ satisfies the property of *PGI-minimal monotonicity* if for any pair of games (N, v), $(N, w) \in H$, it holds that:

$$f_i(N, w) \sum_{j \in N} \left|M_j(w)\right| \geq f_i(N, v) \sum_{j \in N} \left|M_j(v)\right|,$$

for all player $i \in N$ such that $M_i(v) \subseteq M_i(w)$.

This property states that if the set of minimal winning coalitions containing a player i in game (N, v) is a subset of minimal winning coalitions containing this player in game (N, w), then the power of player i in game (N, w) is not less than power of player i in game (N, v) (first, this power must be normalized by the number of minimal winning coalitions of every player in games (N, v) and (N, w)).

For any two simple games (N, v) and (N, w), and for all $i \in N$ such that $|M_i(v)| = |M_i(w)|$, using the PGI-minimal monotonicity property, it holds that

$$f_i(N, w) \sum_{j \in N} \left|M_j(w)\right| = f_i(N, v) \sum_{j \in N} \left|M_j(v)\right|,$$

that is, a relation between the power index of the player i in the two games is obtained.

- The unique power index on $SI(N)$ that satisfies PGI-minimal monotonicity, null player, symmetry, and efficiency is the Public Good Index.

Table 6.1 Shapley–Shubik index, the Banzhaf index, the Deegan–Packel index and the Public Good Index for the parties of the Portuguese Parliament (IX Term of office; 2002)

Parties	Members	S.S.	Banz.	D.P.	P.G.I.
PPD/PSD	105	0.47	0.46	0.33	0.31
PS	96	0.18	0.18	0.17	0.15
CDS/PP	14	0.18	0.18	0.17	0.15
PCP	10	0.13	0.14	0.20	0.23
BE	3	0.02	0.02	0.07	0.08
PEV	2	0.02	0.02	0.07	0.08

6.3 Application to the Portuguese Parliament

In Table 6.1, we compute the Shapley–Shubik index, the Banzhaf index, the Deegan–Packel index and the Public Good Index for the parties of the Portuguese Parliament (IX Term of office; 2002).

We observe that PS has many more members than CDS/PP, but they have the same power because they are symmetric. We also note that PCP has fewer members than PS, but its Deegan–Packel and Public Good indices are higher. This is due to the fact that in these indices only minimal winning coalitions are taken into account, and PCP belongs to 3 minimal winning coalitions while PS is involved in just 2.

Finally, we claim that the realistic situation is not as simple as we considered in this work. In this simple model we do not take into account the ideology nor the capacity to persuade that each player has. There are several ways to include additional information in the model in order to attain a more realistic index. One approach would be to consider that players are divided into a priori unions and that they cannot form coalitions in which the whole union is not involved, this would give rise to the games with a priori unions. Another way to extend the model is to consider that players can only communicate through a given undirected graph, with this consideration we obtain the games with graph restricted communication.

Acknowledgements This research received financial support from Spanish Ministry for Science and Innovation through Project ECO2008-03484-C02-02, from Xunta de Galicia through Project INCITE09-207-064-PR, and from the FPU program of the Spanish Ministry for Education. F. Ferreira and A. A. Pinto thank ESEIG/IPP, LIAAD-INESC Porto LA, Calouste Gulbenkian Foundation, PRODYN-ESF, FEDER, POFC, POCTI and POSI by FCT and Ministério da Ciência e da Tecnologia, and the FCT Pluriannual Funding Program of the LIAAD-INESC Porto LA and of the Research Centers of Mathematics of University of Minho and University of Porto, for their financial support.

References

1. Alonso-Meijide, J.M., Casas-Méndez, B., Holler, M.J., Lorenzo-Freire, S.: Computing power indices: Multilinear extensions and new characterizations. Eur. J. Oper. Res. **188**, 540–554 (2008)

2. Alonso-Meijide, J.M., Ferreira, F., Mozos, M.A., Pinto, A.A.: Two new power indices based on winning coalitions. Special issue in honour of Mauricio Peixoto and David Rand. J. Differ. Equ. Appl. **17**(6), 71–76 (2011)
3. Banzhaf, J.F.: Weighted voting doesn't work: A mathematical analysis. Rutgers Law Rev. **19**, 317–343 (1965)
4. Deegan, J., Packel, E.W.: A new index of power for simple n-person games. Int. J. Game Theory **7**, 113–123 (1979)
5. Dubey, P.: On the uniqueness of the Shapley value. Int. J. Game Theory **4**, 131–139 (1975)
6. Dubey, P., Shapley, L.S.: Mathematical properties of the Banzhaf power index. Math. Oper. Res. **4**, 99–131 (1979)
7. Holler, M.J.: Forming coalitions and measuring voting power. Polit. Stud. **30**, 262–271 (1982)
8. Holler, M.J., Packel, E.W.: Power, luck and the right index. J. Econ. **43**, 21–29 (1983)
9. Lorenzo-Freire, S., Alonso-Meijide, J.M., Casas-Méndez, B., Fiestras-Janeiro, M.G.: Characterizations of the Deegan–Packel and Johnston power indices. Eur. J. Oper. Res. **177**, 431–434 (2007)
10. Shapley, L.S.: A value for n-person games. In: Tucker, A.W., Kuhn, H.W. (eds.) Contributions to the Theory of Games, pp. 307–317. Princeton University Press, Princeton (1953)
11. Shapley, L.S., Shubik, M.: A method for evaluating the distribution of power in a committee system. Am. Polit. Sci. Rev. **48**, 787–792 (1954)
12. Young, H.P.: Monotonic solutions of cooperative games. Int. J. Game Theory **14**, 65–72 (1985)

Chapter 7
A Methodological Contribution in the Theory of the Firm Under Uncertainty

Alberto A. Álvarez-López and Inmaculada Rodríguez-Puerta

Abstract We show a simple methodology (or scheme to work) to study comparative-static effects in some models of the theory of the firm under uncertainty. We present this methodology in detail for a basic production model with only one decision variable (SANDMO's model). Then we sketch it for a model with two decision variables (HOLTHAUSEN's model with a forward market), and for a model of optimal allocation of production (a two-ends model, of our own).

7.1 Introduction

In recent papers, some results in models of the theory of the firm under uncertainty are frequently proved with the aid of geometrical methods, which have become a useful tool for it.[1] In this short survey we present an alternative methodology (or scheme to work) to study properties and comparative-static effects in models of this theory.

We illustrate the methodology in detail for one of the basic models of the theory: the well-known SANDMO's model, as presented in [7]. This is a model of one decision variable: the amount of output to be produced, and with only one source of uncertainty: the price at which that output will be sold. For this model, we are able to prove easily some important properties and also comparative-static results.

Then we show the methodology for a model which enhances directly that of SANDMO by considering a second decision variable: HOLTHAUSEN's model

A.A. Álvarez-López (✉)
Departamento de Economía Aplicada Cuantitativa II, UNED, Paseo Senda del Rey, 11, Madrid 28040, Spain
e-mail: aalvarez@cee.uned.es

I. Rodríguez-Puerta
Área de Métodos Cuantitativos, Departamento de Economía, Métodos Cuantitativos e Historia Económica, Universidad Pablo de Olavide, Carretera de Utrera Km. 1, Sevilla 41013, Spain
e-mail: irodpue@upo.es

[1] See, for instance, [3].

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_7,

(see [4]). The second variable is the amount of output to be hedged in a forward market which is assumed to exist for the output. We see that a second variable is not a difficulty to obtain results with the same scheme.

Finally, we turn to a one-variable model, this of our own (see [6]), which is of a different kind. Unlike SANDMO's model, the amount of output is fixed, and what the firm decides is the allocation of this output between two possible ends. One of these ends has a certain price, and the other has an uncertain price. The way the firm faces uncertainty is different from that of the previous models. For this model we also have comparative-static results in the same manner.

This methodology is based on analytical methods. We also make use of a lemma adapted from [5], scarcely used in the literature (Lemma 7.1 in Appendix). This lemma gives us useful bounds for products of random variables.

7.2 The Basic Model (One Decision Variable)

We consider a competitive firm which produces a single output and faces uncertainty in the price at which this output will be sold. The firm has to decide the amount of output to be produced *before* the sale date, that is, before knowing the spot price.

For the firm, the price is a non-degenerate random variable $P \geq 0$ with expectation $\mu > 0$. The total cost of producing an amount $q \geq 0$ is given by $C(q) = c(q)+B$, where B is a fixed cost and $c(q)$ stands for variable costs, so that $c(0) = 0$. We assume that the function C is of class $\mathscr{C}^2$ on $\mathbb{R}_+$ and such that $C' > 0$ and $C'' > 0$. The firm's attitude towards risk is modeled by a BERNOULLI utility function u, regular enough (at least of class $\mathscr{C}^2$ on $\mathbb{R}$) and such that $u' > 0$ and $u'' < 0$. In particular, the firm is risk averse.

For each level of output q, the firm's profit is given by $\pi(q) \equiv Pq - C(q)$.[2] The firm seeks to maximize the expected utility of this profit, that is:

$$\max_{q \in \mathbb{R}_+} U(q),$$

where $U(q) \equiv \mathsf{E}\big[u\big(\pi(q)\big)\big]$. The first and second derivatives of U are:

$$U'(q) = \mathsf{E}\big[u'(\pi)\,\big(P - C'(q)\big)\big]$$

and

$$U''(q) = \mathsf{E}\big[u''(\pi)\,\big(P - C'(q)\big)^2\big] - C''(q)\;\mathsf{E}\big[u'(\pi)\big].$$

According to the hypotheses, we have that $U'' < 0$ on $\mathbb{R}_+$, and thus the function U is strictly concave on $\mathbb{R}_+$. Henceforth, we will assume that this maximization

[2] When possible, we will write simply π instead of $\pi(q)$.

problem has a solution q^*, which is not excluded to be null.[3] The strict concavity of U assures that this solution is also unique. Notice that $q^* = 0$ if and only if $U'(0) \leq 0$. In addition, the equality $U'(q) = 0$ is a sufficient condition for $q \geq 0$ to be the unique solution q^*.

Now we consider the following function:

$$F(q) = \frac{\mathsf{E}\big[u'\big(\pi(q)\big)\, P\big]}{\mathsf{E}\big[u'\big(\pi(q)\big)\big]} \ , \quad q \geq 0 \ ,$$

with derivative:

$$F'(q) = \frac{\mathsf{E}\big[u''\big(\pi(q)\big)\, \big(P - C'(q)\big)\, \big(P - F(q)\big)\big]}{\mathsf{E}\big[u'\big(\pi(q)\big)\big]} \ .$$

This function is closely related to the marginal utility U':

$$U'(q) = \big(F(q) - C'(q)\big)\ \mathsf{E}\big[u'\big(\pi(q)\big)\big] \ ; \tag{7.1}$$

in particular, the equality $F(q) = C'(q)$ is a sufficient condition for $q = q^*$. We also have that $F(0) = \mu$, and that $F(q) < \mu$ for all $q > 0$. The latter can be proved by applying Lemma 7.1 with the random variable $X = P - \mu$, and the functions $\psi \equiv 1$ and $\phi(s) = u'\big((s + \mu)q - C(q)\big)$. Indeed, the function ϕ is strictly decreasing when $q > 0$, and we obtain:

$$\mathsf{E}\big[u'(\pi)\,(P - \mu)\big] < \phi(0) \cdot \mathsf{E}[P - \mu] = 0 \ ,$$

and hence $F(q) < \mu$.

The function F and the equality (7.1) are useful to obtain properties of the solution q^*, and also comparative-static results. As a first example, we can give a characterization of the case of corner solution: the optimal level of output q^* is positive if and only if $\mu > C'(0)$. Indeed, writing (7.1) for $q = 0$, and recalling that $F(0) = \mu$, we have:

$$U'(0) = \big(\mu - C'(0)\big)\, u'(-B) \ ,$$

and thus $\mu > C'(0)$ is equivalent to $U'(0) > 0$, but this is equivalent to $q^* > 0$.

From now on, we will assume that the unique optimal solution q^* is positive, so that the optimal solution q^* is characterized by the condition $F(q^*) = C'(q^*)$. According to the well-known fact that, under certainty, the firm decides to produce that level of output for which the marginal cost equals the price, the characterization of q^* gives us a simple interpretation of the number $F(q^*)$: if it were possible for

[3] It can be proved that a sufficient condition of existence of a solution for this problem is: $\lim\limits_{q \to +\infty} C'(q) > \mu$ (with limit possibly infinity).

the firm to sell its product under a certain price, $F(q^*)$ would be the value of that price for which the firm decides to produce exactly q^* units of output.[4]

We can also prove easily the main result obtained by SANDMO in [7]: the optimal production of the firm under price uncertainty is smaller than that when the price is known to be equal to the expected price μ. Indeed: we have: $F(q^*) = C'(q^*)$, and thus $C'(q^*) < \mu$; since the function C' is strictly increasing, q^* is smaller than that value of q for which $C'(q)$ is equal to μ.

Finally, we illustrate a comparative-static effect. If ν denotes a parameter of the model, we can write $F(q;\nu)$ to stand for the further dependence of F on ν, and $\mathrm{d}q^*/\mathrm{d}\nu$ to stand for the corresponding comparative-static effect. From the characterization $F(q^*;\nu) = C'(q^*)$, we can write:

$$\frac{\mathrm{d}q^*}{\mathrm{d}\nu} = -\frac{F'_\nu(q^*;\nu)}{F'(q^*) - C''(q^*)} \ ;$$

since $F'(q^*) = \mathsf{E}\big[u''(\pi)\,\big(P - C'(q^*)\big)^2\big]/\,\mathsf{E}\big[u'(\pi)\big] < 0$, the denominator is negative, so that the sign of $\mathrm{d}q^*/\mathrm{d}\nu$ is the same as that of the numerator $F'_\nu(q^*;\nu)$. For instance, if we focus our attention on the fixed cost B, we have:

$$F'_B(q^*;B) = -\frac{\mathsf{E}\big[u''(\pi^*)\,\big(P - C'(q^*)\big)\big]}{\mathsf{E}\big[u'(\pi^*)\big]} \ ,$$

where $\pi^* \equiv \pi(q^*) = Pq^* - C(q^*)$. Now, if r_u denotes the ARROW–PRATT measure of absolute risk aversion, we can write:

$$-\,\mathsf{E}\big[u''(\pi^*)\,\big(P - C'(q^*)\big)\big] = \mathsf{E}\big[r_u(\pi^*)\,u'(\pi^*)\,\big(P - C'(q^*)\big)\big].$$

Assume that the firm exhibits DARA. Setting:

$$\psi(s) = u'\big((s + C'(q^*))q^* - C(q^*)\big) \quad \text{and} \quad \phi(s) = r_u\big((s + C'(q^*))q^* - C(q^*)\big),$$

thus $\psi > 0$ and ϕ is decreasing; with $X = P - C'(q^*)$, from Lemma 7.1 we obtain:

$$-\,\mathsf{E}\big[u''(\pi^*)\,\big(P - C'(q^*)\big)\big] \le \phi(0)\ \mathsf{E}\big[u'(\pi^*)\,\big(P - C'(q^*)\big)\big] = 0,$$

where the last factor is null due to the first order condition. Hence, if the firm exhibits DARA, then $\mathrm{d}q^*/\mathrm{d}B \le 0$. And, *mutatis mutandis*, we could prove that $\mathrm{d}q^*/\mathrm{d}B = 0$ or $\mathrm{d}q^*/\mathrm{d}B \ge 0$ depending on whether the firm exhibits CARA or IARA, respectively. Both in [7] and [5], we can find the same result, with a different proof in each case.

[4] If the firm exhibits CARA, we can easily see that $F(q)$ does not depend on the cost $C(q)$. This fact lets us give a deeper interpretation of the number $F(q)$ (for any given value $q > 0$) in the CARA case. See [2].

7.3 The Basic Model With a Forward Market (Two Decision Variables)

HOLTHAUSEN, in [4], enhances SANDMO's model by considering a forward market for the output produced by the firm, so that the firm has now a second decision variable: the amount h of output hedged in the forward market, at a certain price b. Now, the firm's profit is given by:

$$\pi(q,h) = Pq + bh - Ph - C(q) = P(q-h) + bh - C(q) \ ,$$

and the firm seeks to maximize the expected utility of this profit, that is, it seeks to maximize $U(q,h) \equiv \mathsf{E}\big[u\big(\pi(q,h)\big)\big]$.[5]

For this model, we consider the following function:

$$F(q,h) = \frac{\mathsf{E}\big[u'\big(\pi(q,h)\big)\, P\big]}{\mathsf{E}\big[u'\big(\pi(q,h)\big)\big]} \ , \quad q \in \mathbb{R}_+ \ , \quad h \in \mathbb{R} \ .$$

If we assume that there is an interior, unique solution (q^*, h^*) for the maximization problem, this optimal solution is characterized by the conditions:

$$F(q^*,h^*) = C'(q^*) \quad \text{and} \quad F(q^*,h^*) = b.$$

We see that $b = C'(q^*)$,[6] which establishes that the optimal output to be produced will not be affected by variations in elements of the model different from the forward price or the marginal cost. Other comparative-static effects for this model could only influence the optimal hedging h^*. We can explore them as we did for SANDMO's model. With analog notations, from the characterization $F(q^*,h^*) = b$ we write:

$$\frac{\mathrm{d}h^*}{\mathrm{d}\nu} = -\frac{F'_\nu(q^*,h^*;\nu)}{F'_h(q^*,h^*)} \ ,$$

where $F'_h(q^*,h^*) = -\mathsf{E}\big[u''(\pi)\,(b-P)^2\big]/\,\mathsf{E}\big[u'(\pi)\big] > 0$, so that the sign of $\mathrm{d}h^*/\mathrm{d}\nu$ is the opposite of that of the numerator $F'_\nu(q^*,h^*;\nu)$. For instance, for $\nu = B$ (a variation of the fixed cost), we are able to prove this result: if the firm exhibits DARA, then $\mathrm{d}h^*/\mathrm{d}B \geq 0$ when $b < \mu$, and $\mathrm{d}h^*/\mathrm{d}B \leq 0$ when $b > \mu$; and the contrary inequalities hold if the firm exhibits IARA.[7]

[5] This is a maximization problem over $\mathbb{R}_+ \times \mathbb{R}$: the variable q is restricted to be non-negative (as in SANDMO's model), but the variable h has no restrictions. From the firm's point of view, the operation in the forward market is interpreted as a sale if $h > 0$, and as a purchase if $h < 0$. For further explanations of the exact meaning of a sale or a purchase in this market, see [4] or [1].

[6] In [4], the author obtains this result simply by adding the two first order conditions.

[7] This effect was not studied by HOLTHAUSEN in [4]. For a complete proof, although with a different method, see [1].

7.4 A Two-Ends Model

Now we consider a different situation.[8] The firm has just produced a known amount q_T of its output, and has two possible ends for it. The firm's decision is which amount of output is allocated for each one. We assume that the total amount q_T is fully distributed between the two ends, so that the firm chooses the quantity q to be allocated for the first one, and the quantity for the second one will be $q_T - q$. The price p for the first end is certain, but the price P for the second one is uncertain. The firm's profit is given by $\pi(q) = pq + P(q_T - q) - B$, where B is here the cost of producing the total amount q_T, and the firm seeks to maximize $U(q)$ over the interval $[0, q_T]$, where $U(q) \equiv \mathsf{E}\big[u\big(\pi(q)\big)\big]$. It follows easily that there is a unique solution q^* for this problem (possibly a corner solution).

Here, we consider a function which formally is the same as that considered in Sect. 7.2:

$$F(q) = \frac{\mathsf{E}\big[u'\big(\pi(q)\big)\, P\big]}{\mathsf{E}\big[u'\big(\pi(q)\big)\big]}\,, \qquad 0 \le q \le q_T\ .$$

Assuming that the optimal solution is interior, the equality $F(q^*) = p$ is a characterization of the optimal solution. From this equality, we can easily obtain comparative-static effects for this model as we did for the models studied in the previous sections. The result about a variation in the cost B is the following: depending on whether the firm exhibits DARA, CARA or IARA, we have: $dq^*/dB \geq 0$, $dq^*/dB = 0$ or $dq^*/dB \leq 0$, respectively. For a proof, see [6], where there is a more general result.

7.5 Concluding Remarks

As we can see, the methodology we present requires to consider an auxiliary function F defined in the form: $\mathsf{E}[u'(\pi)\, P]/\, \mathsf{E}[u'(\pi)]$, where π can have been defined with one or two variables. This function lets us establish a simple characterization of the optimal solution. Next, by applying the Implicit Function Theorem to this characterization, one can obtain easily the comparative-static effects with the aid of Lemma 7.1.

Formally, the three models presented here are studied almost in the same manner. This suggests a possible generalization. In [2] we indeed give a general framework in which these three models are particular cases.

Acknowledgements Previous versions of the material for this work were presented in ASEPUMA Annual Meeting in 2007 and 2008, and in EURO 2010.

The authors would like to thank Alberto Pinto for his kind invitation to write this paper. Alberto A. Álvarez-López would also like to thank the financial support provided by the Spanish

[8] The model in this section is a simplified version of that studied in [6].

Interministerial Commission of Science and Technology (CICYT: Comisión Interministerial de Ciencia y Tecnología), under the Project with the reference number ECO2008-06395-C05-03.

Appendix

The following lemma, with a minor modification, is taken from [5]:

Lemma 7.1. *Let ψ and ϕ be two real functions defined on $\mathbb{R}$ such that $\psi > 0$ and ϕ is increasing. If $\xi = \psi \cdot \phi$, and X is a non-degenerate real random variable such that the expectation $E[X\ \psi(X)]$ is finite, then:*

$$E[X\ \xi(X)] \geq \phi(0)\ E[X\ \psi(X)]\,,$$

and the contrary inequality holds when ϕ is decreasing. In addition, if ϕ is strictly increasing or strictly decreasing, the corresponding inequality also holds strictly.

Proof. See [5], or [6]. □

References

1. Álvarez-López, A.A., Rodríguez-Puerta, I.: Teoría de la empresa bajo incertidumbre con mercado de futuros: el papel de los costes fijos y de un impuesto sobre los beneficios. Rect@ **10**(1), 253–265 (2009)
2. Álvarez-López, A.A., Rodríguez-Puerta, I.: A unified approach for some comparative-static effects in the theory of the firm under uncertainty (2011)
3. Dalal, A.J., Alghalith, M.: Production decisions under joint price and production uncertainty. Eur. J. Oper. Res. **197**, 84–92 (2009)
4. Holthausen, D.M.: Hedging and the competitive firm under price uncertainty. Am. Econ. Rev. **69**(5), 989–995 (1979)
5. Lippman, S.A., McCall, J.J.: The economics of uncertainty: selected topics and probabilistic methods. In: Arrow, K.J., Intriligator, M.J. (eds.) Handbook of Mathematical Economics, vol. 1, Chap. 6. North-Holland, Amsterdam (1982)
6. Rodríguez-Puerta, I., Álvarez-López, A.A.: Optimal allocation of a fixed production under price uncertainty (2010)
7. Sandmo, A.: On the theory of the competitive firm under price uncertainty. Am. Econ. Rev. **61**(1), 65–73 (1971)

Chapter 8
Explosion of Smoothness for Conjugacies Between Unimodal Maps

José F. Alves, Vilton Pinheiro, and Alberto A. Pinto

Abstract Let f and g be C^r unimodal maps, with $r \geq 3$, topologically conjugated by h and without periodic attractors. If h is strongly differentiable at a point p in the expanding set $E(f)$, with $h'(p) \neq 0$, then, there is an open renormalization interval J such that h is a C^r diffeomorphism in the basin $B(J)$ of J, and h is not strongly differentiable at any point in $I \setminus B(J)$. The expanding set $E(f)$ contains all points with positive Lyapunov exponent, and if f has a Milnor's interval cycle attractor A then $E(f)$ has full Lebesgue measure.

8.1 Introduction

Sullivan proved that if a topological conjugacy between analytic uniformly expanding maps of the circle is differentiable at a point then the conjugacy is analytic. De Faria, Jiang and Rand, among others, extended this result in many ways (see Yunping Jiang survey [16]). For instance, Pinto and Ferreira [9] proved that if a topological conjugacy between dynamical systems in hyperbolic basic sets on surfaces

J.F. Alves (✉)
Departamento de Matemática Pura, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre 687, 4169-007 Porto, Portugal
e-mail: jfalves@fc.up.pt

V. Pinheiro
Departamento de Matemática, Universidade Federal da Bahia, Av. Ademar de Barros s/n, 40170-110 Salvador, Brazil
e-mail: viltonj@ufba.br

A.A. Pinto
LIAAD-INESC Porto LA e Departamento de Matemática, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, 687, 4169-007, Portugal
and
Centro de Matemática e Departamento de Matemática e Aplicações, Escola de Ciências, Universidade do Minho, Campus de Gualtar, 4710-057 Braga, Portugal
e-mail: aapinto@fc.up.pt

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_8,

is differentiable at a point, in the basic set, then the conjugacy has a smooth extension to an open set on the surface. The purpose of this paper is, given a topological conjugacy $h : I \to J$ between smooth unimodal maps f and g, to present sufficient conditions in just one point $p \in I$ that imply the differentiability of the topological conjugacy h in an open set O contained in I, such that h has non-zero derivative at every point in O.

8.2 Unimodal Maps

Let I be a compact interval and $f : I \to I$ a C^{1+} map. We say that c is a *non-flat turning point* of f if there exist $\alpha > 1$ and a C^r diffeomorphism ϕ defined in a small neighborhood K of 0 such that

$$f(c+x) = f(c) + \phi(|x|^{\alpha}) \quad \text{for every } x \in K. \tag{8.1}$$

We say that α is the *order* of the turning point c and denote it by $\mathrm{ord}_f(c)$. We say that f is a *unimodal* map, if (a) $f(\partial I) \subset \partial I$; (b) f has only one of turning point c; and (c) the turning point c is non-flat.

A point $p \in I$ is called *nearby expanding*, if there is a sequence of points p_n coverging to p and a sequence of open intervals $V_n \ni p_n$ with the following property: there is $\delta > 0$ and a sequence k_n tending to infinite such that (a) $f^{k_n}|_{V_n}$ is a diffeomorphism and (b) $f^{k_n}(V_n) = B_\delta(f^{k_n}(p_n))$. We denote the set of all nearby expanding points of f by $\widetilde{E}(f)$.

A set $A \subset J$ is said to be *forward invariant* if $f(A) \subset A$. The *basin* $B(A)$ of a positively invariant set A is the set of all points $x \in J$ such that its omega limit set $\omega(x)$ is contained in A. A forward invariant compact set $A \subset J$ is called a *(minimal) attractor*, in Milnor's sense, if the Lebesgue measure of its basin is positive and there is noforward invariant compact set A' strictly contained in A such that $B(A')$ has non zero measure.

A open interval J is a *renormalization interval* of a unimodal map f, if there is $n \geq 1$ such that $f^n|_J$ is also a unimodal map. In this case, the forward orbit $\mathcal{O}^+(J) = J \cup \cdots \cup f^{n-1}(J)$ of J is a positive invariant set. For simplicity, let us denote $B(\mathcal{O}^+(J))$ by $B(J)$ and call it the basin of J. Note that $B(J)$ is exactly the set of points whose forward orbit intersects J. The *accessible boundary* $\partial^* B(J) \subset \partial B(J)$ of the renormalization interval J is the union of the boundary points of all connected components of $B(J)$.

A periodic point p with period $n \in \mathbb{N}$ is called *weak repelling periodic point* if it is a neutral periodic point (i.e., $|Df^n(p)| = 1$) and there is a neighborhood V of p such that $f^n|_V$ is a diffeomorphism with $\lim_{j\to+\infty}(f^n|_V)^{-j}(x) = p$ for all $x \in V$.

The attractors of a C^r non-flat multimodal map are one of the following three types: (a) a periodic attractor; (b) a minimal set with zero Lebesgue measure; or (c) a cycle of intervals such that the omega limit set of almost every point in the cycle is the whole cycle (see [46]).

We note that, a map $h : I \to I$ is a C^1 map if, and only if, for every point $p \in I$

$$\lim_{\substack{x,y \to p \\ x \neq y}} \frac{h(x) - h(y)}{x - y} = h'(p).$$

Hence, we say that h is *strongly differentiable* at a point $p \in I$ if, and only if, the above condition holds for p.

Theorem 8.1. *Let f and g be C^r unimodal maps, $r \geq 3$, topologically conjugated by h, without periodic attractors and neutral periodic points. Assume that h is strong differentiable at a point $p \in \widetilde{E}(f)$, with $h'(p) \neq 0$.*

Then, either

(a) h is a C^r diffeomorphism in the full interval I.
(b) there is a renormalization interval $J \subseteq I$ such that

a. h is a C^r diffeomorphism in the basin $B(J)$.
b. h is not strongly differentiable at any point of $\partial B(J)$.

Taking f, g and h as in the above theorem, one can show that if f is not infinitely renormalizable then $\widetilde{E}(f) = I$. In particular, if f has a absolutely continuous invariant probability then $\widetilde{E}(f) = I$. On the other hand, if f is infinitely renormalizable then $\widetilde{E}(f)$ is a dense set, but with zero Lebesgue measure. In any of the above cases, if h is differentiable at the repeller fixed point $x \in \partial I$ of the unimodal map f, then h is C^r in the full interval I.

Shub and Sullivan [41] proved that if a conjugacy between expanding circle maps is absolutely continuous then it is smooth. They also proved that if the expanding circle maps have the same set of eigenvalues, then the conjugacy is smooth. M. Martens and W. de Melo [30] extended this last result to unimodal maps with attractors that are cycle of intervals. M. Lyubich [26] proved that C^2 unimodal maps with Fibonnaci combinatorics and with the same eigenvalues are C^1 conjugate. In [4], using Theorem 8.1, we prove that if the conjugacy for smooth unimodal maps, with attractors that are cycle of intervals, is absolutely continuous then it is smooth.

Theorem 8.2. *Let f be a C^3 unimodal map without periodic attractors and neutral periodic points and such that the critical point is not pre-periodic. If g is a C^3 unimodal map, topologically conjugated to f by h, with a different order at the critical point, then h is not strongly differentiable at any point $p \in \widetilde{E}(f)$ with $h'(p) \neq 0$.*

For a typical stochastic parameter $\lambda \in [0, 4]$ of the quadratic family $f_\lambda(x) = \lambda x(1 - x)$, with $x \in [0, 1]$, the set $\omega(c)$ intersects the interior of the attractor A. Let g be a C^3 unimodal map topologically conjugated to f_λ, by a homeomorphism h, for some typical stochastic parameter λ. By the above corollary, if $\mathrm{ord}_g(h(c)) \neq 2$ then h is not strong differentiable at any point $p \in [-1, 1]$, with $h'(p) \neq 0$.

In [4], we prove the above results and we extend them to multimodal maps and to non-uniformly expanding maps with singular sets and discontinuities.

Acknowledgements Alberto Pinto would like to thank LIAAD-INESC Porto LA, Calouste Gulbenkian Foundation, PRODYN-ESF, FEDER, POFC, POCTI and POSI by FCT and Ministério da Ciência e da Tecnologia, and the FCT Pluriannual Funding Program of the LIAAD-INESC Porto LA and of the Research Centre of Mathematics of University of Minho, for their financial support.

References

1. Ahlfors, L.V., Beurling, A.: The boundary correspondence under quasiconformal mappings. Acta Math. **96**, 125–142 (1956)
2. Alves, J.F., Bonatti, C., Viana, M.: SRB measures for partially hyperbolic systems whose central direction is mostly expanding. Invent. Math. **140**, 351–398 (2000)
3. Alves, J.F.: Strong statistical stability of non-uniformly expanding maps. Nonlinearity **17**, 1193–1215 (2004)
4. Alves, J.F., Pinheiro, V., Pinto, A.A.: Explosion of smoothness for conjugacies between multimodal maps (in preparation)
5. Blokh, A.M., Lyubich, M.Yu.: Non-existence of wandering intervals and structure of topological attractors of one dimensional dynamical systems II. The smooth case. Ergod. Theory Dyn. Syst. **9**, 751–758 (1989)
6. Carleson, L.: On mappings conformal at the boundary. J. Analyse Math. **19**, 1–13 (1967)
7. Cui, G.: Linear Models of Circle Expanding Maps. Academia Sinica (1994)
8. de Faria, E.: Quasisymmetric distortion and rigidity of expanding endomorphisms of S^1. Proc. Am. Math. Soc. **124**, 1949–1957 (1996)
9. Ferreira, F., Pinto, A.A.: Explosion of smoothness from a point to everywhere for conjugacies between diffeomorphisms on surfaces. Ergod. Theory Dyn. Syst. **23**, 509–517 (2003)
10. Gardiner, F., Sullivan, D.: Lacunary series as quadratic differentials. Proceedings of the Symposium in honor of Wilhelm Magnus at Polytechnic Institute of Brooklyn
11. Gardiner, F., Sullivan, D.: Symmetric structures on a closed curve. Am. J. Math. **114**, 683–736 (1992)
12. Jacobson, M.V., Swiatek, G.: Quasisymmetric conjugacies between unimodal maps. I. Induced expansion and invariant measures. Stony Brook (1991) (preprint)
13. Jiang, Y.: On rigidity of one-dimensional maps. Contemp. Math., AMS Series **211**, 319–431 (1997)
14. Jiang, Y.: On Ulam-von Neumann transformations. Comm. Math. Phys. **172**(3), 449–459 (1995)
15. Jiang, Y.: Geometry of geometrically finite one-dimensional maps. Comm. Math. Phys. **156**(3), 639–647 (1993)
16. Jiang, Y.: Differential Rigidity and Applications in One-Dimensional Dynamics. In: Peixoto, M., Pinto, A.A., Rand, D. (eds.) Dynamics, Games and Science I, Springer Proccedings in Mathematics Series. Springer (2011)
17. Jiang, Y.: Asymptotic differentiable structure on Cantor set. Comm. Math. Phys. **155**(3), 503–509 (1993)
18. Jiang, Y.: Renormalization and Geometry in One-Dimensional and Complex Dynamics. Advanced Series in Nonlinear Dynamics. World Scientific Publishing, River Edge, NJ 10 (1996)
19. Jiang, Y.: Smooth classification of geometrically finite one-dimentional maps. Trans. Am. Math. Soc. **348**(6), 2391–2412 (1996)
20. Jiang, Y.: Metric invariants in dynamical systems. J. Dyn. Differ. Equ. **17**(1), 51–71 (2005)

21. Jiang, Y., Cui, G., Gardiner, F.: Scaling functions for degree two circle endomorphisims. Contemp. Math. AMS Series, **335**, 147–163 (2004)
22. Jiang, Y., Cui, G., Quas, A.: Scaling functions, Gibbs measures, and Teichmuller space of circle endomorphisims. Discrete Continous Dyn. Syst. **5**(3), 535–552 (1999)
23. Keller, G.: Exponents, attractors and Hopf decompositions for interval maps. Ergod. Theory Dyn. Syst. **10**, 717–744 (1990)
24. Liu, P.-D.: Pesin's entropy formula forendomorphisms. Nagoya Math. J. **150**, 197–209 (1998)
25. Lyubich, M.: Almost every real quadratic map is either regular or stochastic. Ann. Math., Second Series, **156**(1), 1–78 (2002)
26. Lyubich, M.: Teichmüller Space of Fibonacci Maps. Stony Brook (1993) (preprint)
27. Lyubich, M., Milnor, J.: The Fibonacci unimodal map. J. Am. Math. Soc. **6**(2), 425–457 (1993)
28. Mañé, R.: Hyperbolicity, sinks and measure in one dimensional dynamics. Commun. Math. Phys. **100** 495–524 (1985), and Erratum. Commun. Math. Phys. **112**, 721–724, (1987)
29. Martens, M.: Distortion results and invariant Cantor Sets of unimodal maps. Ergod. Theory Dyn. Syst. **14**(2), 331–349 (1994)
30. de Melo, W., Martens, M.: The Multipliers of Periodic Point in One Dimensional Dynamics. Nonlinearity **12**(2), 217–227 (1999)
31. de Melo, W., van Strien, S.: One-Dimensional Dynamics. Springer (1991)
32. Milnor, J.: On the concept of attractor. Commum. Math. Phys. **99**, 177–195 (1985a)
33. Milnor, J.: On the concept of attractor: Correction and remarks. Comm. Math. Phys. **102**(3), 517–519 (1985b)
34. Pinto, A.A., Almeida, J.P., Portela, A.: Golden tilings. Trans. Am. Math. Soc. (to appear)
35. Pinto, A.A., Rand, D.A.: Classifying C^{1+} structures on dynamical fractals. II: Embedded trees. Ergod. Theory Dyn. Syst. **15**, 969–992 (1995)
36. Pinto, A.A., Rand, D.A.: Smoothness of holonomies for codimension 1 hyperbolic dynamics. Bull. Lond. Math. Soc. **34**, 341–352 (2002)
37. Pinto, A.A., Sullivan, D.: The circle and the solenoid. DCDS-A **16**(2), 463–504 (2006)
38. Saccck, R.: Sacksteder, The measures invariant under an expanding map. Lecture Notes in Mathematics, vol. 392, pp. 179–194. Springer, Berlin (1972)
39. Shub, M.: Endomorphisms of compact differentiable manifolds. Am. J. Math. **91**, 175–199 (1969)
40. Shub, M., Sullivan, D.: A remark on the Lefschetz fixed point formula for differentiable maps. Topology **13**, 189–191 (1974)
41. Shub, M., Sullivan, D.: Expanding endomorphisms of the circle revisited. Ergod. Theory Dyn. Syst. **5**, 285–289 (1985)
42. Sullivan, D.: Differentiable structures on fractal-like sets, determined by intrinsic scaling functions on dual Cantor sets. Proc. Sympos. Pure Math. **48**, 15–23 (1988)
43. Sullivan, D.: Linking the universalities of MilnorThurston, Feigenbaum and AhlforsBers. In: Goldberg, L., Phillips, A. (eds.) Topological Methods in Modern Mathematics, pp. 543–563. Publish or Perish, Boston, MA (1993)
44. Sullivan, D.: Bounds, quadratic differentials, and renormalization conjectures. American Mathematical Society Centennial Publications, vol. 2: Mathematics into the Twenty-first Century (1988 Centennial Symposium, 8–12 August). American Mathematical Society, Providence, RI (1991)
45. Strebel, K.: On the existence of extremal Teichmüler mappings. J. Anal. Math. **30**, 441–447 (1976)
46. van Strien, S., Vargas, E.: Real bounds, ergodicity and negative Schwarzian for multimodal maps. J. Am. Math. Soc. **17**, 749–782 (2004)

Chapter 9
Multidimensional Rovella-Like Attractors

Vítor Araújo

Abstract In a joint work with A. Castro, V. Pinheiro (both from Federal Univ. of Bahia) and M. J. Pacifico (Federal University of Rio de Janeiro), we construct a multidimensional flow exhibiting a Rovella-like attractor: a compact transitive invariant set with an equilibrium accumulated by regular orbits and a partially hyperbolic splitting of the tangent bundle with a multidimensional non-uniformly expanding direction. Moreover, this attractor has a physical measure with full support which is a u-Gibbs state. As in the 3-dimensional Rovella attractor, this example is not robust. This introduces a class of multidimensional dynamics where Benedicks–Carleson arguments can be applied to get persistent non-uniform expansion.

9.1 Introduction

In [3], Bonatti–Pumarino–Viana define a uniformly expanding map on a k-dimensional torus, suspend it as a time-one map of a flow, and then singularize the flow adding a singularity in a convenient flow-box. This procedure creates a new dynamics on the torus presenting a multidimensional version of the one-dimensional expanding Lorenz-like map. For the Lorenz attractor and singular-hyperbolic attractors in general see e.g. [2]. The quotient of the return map to the global cross-section over the stable directions, see Fig. 9.1 is the one-dimensional Lorenz transformation. The goal here is to construct a flow such that "after the identification by the stable directions", the first return map in a certain cross section M is a multidimensional version of the one-dimensional Rovella-map [5].

V. Araújo
Instituto de Matemática, Universidade Federal do Rio de Janeiro, C. P. 68.530, 21.945-970 Rio de Janeiro, Brazil
and
Centro de Matemática da Universidade do Porto, Rua do Campo Alegre 687, 4169-007 Porto, Portugal
e-mail: vitor.araujo@im.ufrj.br, vdaraujo@fc.up.pt

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_9,

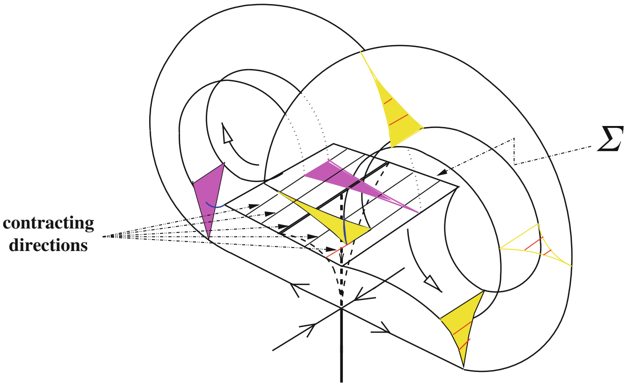
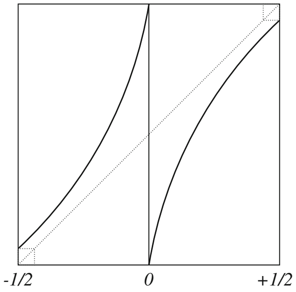


Fig. 9.1 The geometric Lorenz attractor with the contracting directions on the cross-section Σ

Fig. 9.2 The Lorenz one-dimensional transformation



A Rovella-like attractor is the maximal invariant set of a geometric flow whose construction is very similar to the one that gives the geometric Lorenz attractor, [1, 2, 4], except for the fact that the eigenvalues relation $\lambda_u + \lambda_s < 0$ there is replaced by $\lambda_u + \lambda_s > 0$, where $\lambda_u > 0$ and λ_s is the weakest negative eigenvalue at the equilibrium at the origin. We remark that, unlike the one-dimensional Lorenz map obtained from the usual construction of the geometric Lorenz attractor, a one-dimensional Rovella map has a criticality at the origin, caused by the eigenvalue relation $\lambda_u + \lambda_s > 0$ at the singularity. In Fig. 9.3 we present some possible "Rovella one-dimensional maps" obtained through quotienting out the stable direction of the return map to the global cross-section of the attractor, as in Fig. 9.1.

We follow the same strategy of [3]. Nevertheless, since we aim at a multidimensional Rovella-like map, we have to deal with critical regions, that is, regions where the derivative of the return map to a global cross-section vanishes. Because of this, proving the existence of non-trivial attractors for the flow arising from such construction requires a more careful analysis. Indeed, as in the one-dimensional case, depending on the dynamics of the critical region, every attractor for the return map may be periodic (trivial).

Typically, when the critical region is non-recurrent (Misiurewicz maps in one-dimensional dynamics), most of the difficulties introduced by the critical region can

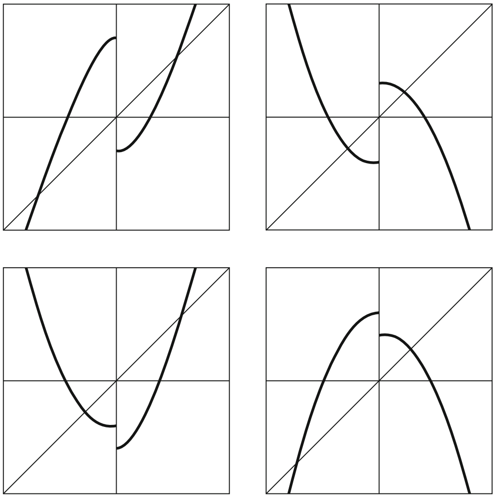


Fig. 9.3 The several cases for the one-dimensional map for the contracting Lorenz model

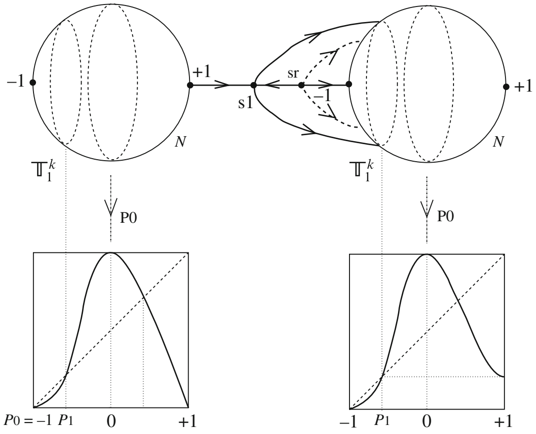


Fig. 9.4 The quotient of the return map of the flow at the upper left, where each parallel is a torus, and the one-dimensional map on the bottom left, obtained quotienting out the parallels. After the introduction of the equilibrium s_1 and source $\hat{s}$ the return map can be seen in the quotient as depicted in the upper half, and the one-dimensional quotient map on tori in the bottom right

be bypassed. That is one of the main reasons for us to construct a kind of multidimensional Misiurewicz dynamics. In general, such critical regions in dimension greater than one are sub-manifolds, and one can not rule out that they intersect each other under the action of the dynamics. Albeit this, we are able to exhibit a class of multidimensional Misiurewicz endomorphisms that appears naturally in a flow dynamics.

9.1.1 Conceptual Description of the Construction

We start with a basic dynamics presenting an expanding invariant torus $\mathbb{T}_1^k$ that will absorve the image of the critical region after the singularization of the associated flow. By topological reasons, this map can not be seen as a time-one map of a suspension flow: locally its degree is not constant. To bypass this new difficulty, we realize this map as a first return map of a singular flow (after identification by stable directions). Afterwards, we singularize a periodic orbit of this flow, introducing a new singularity s_1 of saddle-type, with $(k+1)$-dimensional unstable manifold and l-dimensional stable manifold. Moreover, all the eigenvalues of s_1 are real and, if λ_{s_i} and λ_{u_j} denote the stable and the unstable eigenvalues at s respectively, then $\max\{\lambda_{s_i}\} + \max\{\lambda_{u_j}\} < 0$ for $0 \leq i \leq 3$ and $0 \leq j \leq k+1$. We say that this kind of singularity is a *Rovella-like singularity*. We need also a source $\hat{s}$ to accompany s_1 for topological reasons. The resulting flow will present a multidimensional transitive *Rovella-like attractor*, supporting a physical measure.

The existence of the physical/SRB measure is obtained through a multidimensional extension of arguments of Benedicks–Carleson type, taking advantage of the fact that through identification of stable leaves we can project the dynamics of the first return map of the flow to a global cross-section obtaining a one-dimensional transformation with a Misiurewicz critical point.

Moreover considering the perturbation of this flow along parametrized families, we can show the existence of many parameters for which nearby flows exhibit an attractor with a unique physical measure. In addition, we point out that the analysis of the dynamics of most perturbations of our flow cannot be easily reduced (perhaps not at all) to a one-dimensional model. This indicates that intrinsic multidimensional tools should be developed to fully understand this class of flows.

References

1. Afraimovich, V.S., Bykov, V.V., Shil'nikov, L.P.: On the appearence and structure of the Lorenz attractor. Dokl. Acad. Sci. USSR **234**, 336–339 (1977)
2. Araujo, V., Pacifico, M.J.: Three Dimensional Flows. XXV Brazillian Mathematical Colloquium. IMPA, Rio de Janeiro, (2007)
3. Bonatti, C., Pumariño, A., Viana, M.: Lorenz attractors with arbitrary expanding dimension. C. R. Acad. Sci. Paris Sér. I Math. **325**(8), 883–888 (1997)
4. Guckenheimer, J., Williams, R.F.: Structural stability of Lorenz attractors. Publ. Math. IHES **50**, 59–72 (1979)
5. Rovella, A.: The dynamics of perturbations of the contracting Lorenz attractor. Bull. Braz. Math. Soc. **24**(2), 233–259 (1993)

Chapter 10
Robust Heteroclinic Behaviour, Synchronization, and Ratcheting of Coupled Oscillators

Peter Ashwin and Özkan Karabacak

Abstract This review examines some recent work on robust heteroclinic networks that can appear as attractors for coupled dynamical systems. We focus on coupled phase oscillators and discuss a number of nonlinear dynamical phenomena that are atypical in systems without some coupling structure. The phenomena we discuss include heteroclinic cycles and networks between partially synchronized states. These networks can be attracting and robust to perturbations in parameters and system structure as long as the coupling structure is preserved. We discuss two related effects; extreme sensitivity to detuning (strongly coupled oscillators may lose their frequency synchrony for very small detunings) and heteroclinic ratchet where the sensitivity may only appear for detunings of one sign.

10.1 Introduction

Coupled dynamical systems are a very important source of examples of nonlinear systems that are of interest because of many applications. Additionally, they are of intrinsic interest as structured examples of high dimensional dynamical systems. The applications of coupled dynamical systems are very wide and include in particular solid state physics [2], neuroscience [21] and biological systems generally [37], rather than discuss applications here we refer to these articles. A fundamental concept of use for describing coupled dynamical systems (whether chaotic or not) is that of synchronization in its various forms, and this has been the topic of many papers over the last decade [1, 10, 30, 31].

The topic that we focus on in the review is the robust appearance of dynamics that is neither chaotic nor periodic, but that is intermittent in the sense that repeated switchings are apparent between different saddle states. These "robust heteroclinic cycles" appear naturally in systems ranging from Lotka–Volterra dynamics

P. Ashwin (✉) and Ö. Karabacak
Mathematics Research Institute, Harrison Building, University of Exeter, Exeter EX4 4QF, UK
e-mail: p.ashwin@ex.ac.uk, o.karabacak@ex.ac.uk

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_10,

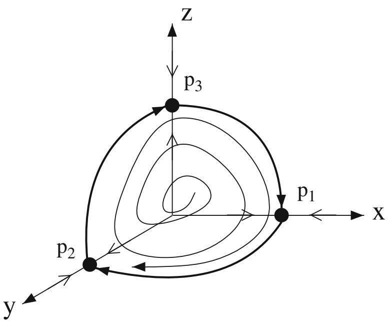


Fig. 10.1 The Guckenheimer–Holmes cycle: (10.1) has an attracting heteroclinic cycle between equilibria p_1, p_2 and p_3 for an open set of parameter values, that is robust to all perturbations that preserve a finite group of symmetries of the vector field

to symmetric systems [14, 16, 20, 27]. Indeed the cycles may be between chaotic saddles in general [8, 16, 29].

The prototype of these cycles is the so-called *Guckenheimer–Holmes cycle* [18] though the same cycle has been studied in a variety of contexts by dos Reis [34], Busse and Clever [12] and others. This can be understood from studying the dynamics of the vector field

$$\begin{aligned} \dot{x} &= \mu x + (ax^2 + by^2 + cz^2)x \\ \dot{y} &= \mu y + (ay^2 + bz^2 + cx^2)y \\ \dot{z} &= \mu z + (az^2 + bx^2 + cy^2)z \end{aligned} \tag{10.1}$$

for the open set of parameters where $\mu > 0$, $a < 0$ and $b < -c < 0$. For this set of parameters (see e.g. [13, p61]) one can verify that the dynamics possesses an attracting heteroclinic cycle whose structure is illustrated in Fig. 10.1. This cycle is robust because the system is preserved under a number of reflection symmetries $(x, y, z) \to (\mp x, \mp y, \mp z)$ and the permutation symmetry $(x, y, z) \to (y, z, x)$ meaning that the axis planes are invariant for the dynamics. Therefore, saddle-to-sink type heteroclinic connections between equilibria on these planes are robust under symmetry-preserving perturbations. Observe that (10.1) can be viewed as a system of three coupled one-dimensional dynamical systems with a particular form of cyclic coupling.

Other families of dynamical systems for which heteroclinic cycles may appear robustly are coupled dynamical systems where the coupling between dynamical units respects to a directed graph (see [14, 17] and the references therein). It is because such families also admit dynamically invariant subspaces, this time forced by the coupling structure rather than the symmetry of the system. The heteroclinic cycles found in such systems are much richer in dynamics due to the lack of symmetry and give rise to a new phenomenon in case of coupled oscillators which we summarize in Sect. 10.4.

Much research has been done on the behaviour of robust heteroclinic cycles in symmetric systems; we will focus only on work that has linked this to coupled oscillators. The paper is organized as follows; in Sect. 10.2 we give an introduction to coupled oscillator dynamics and the reduction to phase oscillators. Section 10.3 discusses examples of robust heteroclinic networks and extreme sensitivity to detuning in such systems. Section 10.4 discusses some recent work on "heteroclinic ratchets" where attractors of the nonlinear system wind in a nontrivial manner around the torus. The final Sect. 10.5 summarizes some open questions in this area and relevance to applications.

10.2 Synchronization Properties of Coupled Oscillators

Many physical processes that are time-periodic in nature can be modelled by nonlinear oscillators, by which we mean dissipative dynamical systems with hyperbolic, attracting limit cycles. When several of these systems are coupled, various phenomena related to the synchronization of oscillators can arise. In this paper we will focus on some synchronization properties of oscillators that are well-modelled by coupled equations for the phases of each oscillator.

10.2.1 From Limit Cycle Oscillators to Phase Oscillators

By a limit cycle oscillator, we mean a dynamical system $\dot{x} = f(x)$ on a manifold M that has an attracting hyperbolic periodic solution $\gamma(t)$. Coupled limit cycle oscillator systems are dynamical systems of the form

$$\dot{x} = F(x, \kappa)\ , \quad x \in M^N \tag{10.2}$$

which reduce to N uncoupled limit cycle oscillators when the coupling strength $\kappa = 0$. In the uncoupled case ($\kappa = 0$), the N-torus defined as the direct product of the limit cycles of each oscillator

$$\tau^N = \{x_i = \gamma_i(t + \theta_i)\colon\ (\theta_1, \ldots, \theta_N) \in \mathbb{T}^N\}$$

is obviously invariant, attracting and normally hyperbolic. Therefore, one can predict that this attracting N-torus persists in the weak coupling case $\kappa \ll 1$. As a result, the asymptotic dynamics of (10.2) can be reduced to the dynamics reduced on this N-torus in the weak coupling case. Note that for a point on this torus, each oscillator can be represented by a phase variable $t + \theta_i \in \mathbb{T}$. Using an averaging technique [9], one can obtain a coupled phase oscillator system of the form

$$\dot{\theta} = \bar{F}(\theta, \kappa)\ , \quad \theta \in \mathbb{T}^N, \tag{10.3}$$

where $\theta_i \in \mathbb{T}$ represents the phase of the oscillator i and F is invariant under the action of $\mathbb{S}^1$ given by $\theta \mapsto \theta + \varepsilon(1, \ldots, 1)$, $\varepsilon \in [0, 2\pi)$ (see [9] for details). This

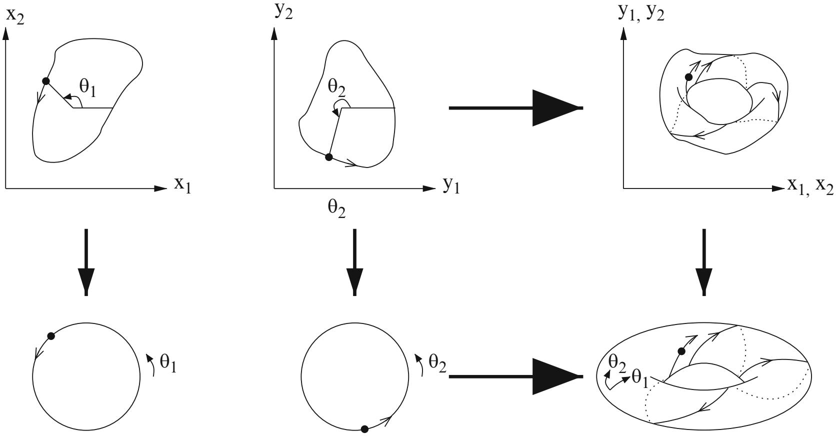


Fig. 10.2 Schematic diagram representing the reduction from limit cycle oscillators (*upper figures*) to phase oscillators (*lower figures*). In the uncoupled case $\kappa = 0$, the direct product of limit cycles (*upper-right*) is invariant and corresponds to the torus (*lower-right*) which is the phase space for the reduced system of coupled phase oscillators

symmetry gives rise to a further reduction of the system on N-torus to a system on the quotient space $\mathbb{T}^N/\mathbb{S}^1$, which is an $(N-1)$-torus,

$$\dot{\phi} = \tilde{F}(\phi, \kappa)\ , \quad \phi \in \mathbb{T}^{N-1}, \tag{10.4}$$

where ϕ_i's can be chosen as independent phase difference variables $\theta_{m_i} - \theta_{n_i}$. In the sequel, we refer to the space of phase difference $\mathbb{T}^{N-1}$ as *phase difference space* of the coupled oscillator system (10.3).

The idea of reducing the coupled phase oscillator systems to limit cycle oscillators was first proposed by Winfree in 1967. However, coupled phase oscillator systems began to be studied widely after Kuramoto's works in 1984 (See [37] and the references therein).

Kuramoto's model consists of N phase oscillators that are coupled globally with a sinusoidal coupling function. That is, the governing equation for each oscillator is

$$\dot{\theta}_i = \omega_i + \frac{\kappa}{N} \sum_{j=1}^{N} \sin(\theta_i - \theta_j), \tag{10.5}$$

where $\theta_i \in \mathbb{T} = [0, 2\pi)$ is the phase and ω_i is the natural frequency of the oscillator i.

Considering an arbitrary coupling structure and a more general coupling function, the coupled phase oscillator dynamics can be written, more generally, as follows:

$$\dot{\theta}_i = \omega_i + \frac{\kappa}{N} \sum_{j=1}^{N} c_{ij} g(\theta_i - \theta_j). \tag{10.6}$$

Here, the connection matrix $\{c_{ij}\}$ represents the coupling between oscillators. $c_{ij} = 1$ if the oscillator i receives an input from the oscillator j and $c_{ij} = 0$ otherwise. The coupling function g is a 2π-periodic function. Therefore, it is natural to consider a Fourier series expansion of g

$$g(x) = \sum_{k=1}^{\infty} r_k \sin(kx + \alpha_k) \tag{10.7}$$

Note that, by scaling the time, we can set $\kappa = N$ and $r_1 = 1$. In this case, the coupling is modulated by the parameters $\alpha_1, \alpha_2, \ldots$ and $r_2, r_3, \ldots$.

Several truncated cases of the general case (10.7) was considered in the literature. Considering the first Fourier term only (as in the Kuramoto model, (10.5)), frequency synchronization and clustering phenomena were analyzed [28, 35]. Hansel et al. used first two Fourier terms and observed a new phenomenon, called slow switching, as a result of the presence of an asymptotically stable robust heteroclinic cycle [19, 25]. Recently, using the first three harmonics an attracting heteroclinic ratchet was found for a nonsymmetric connection structure [23], while four harmonics seem to be necessary to observe chaotic dynamics in four all-to-all coupled oscillators (Timme, 2009, personal communication)

In the literature, there are different definitions for the phase or frequency synchronization of oscillators. Moreover, one can define other concepts related to the synchronization, such as sensitivity to detuning [6]. For an ordered pair of oscillators, we call such properties *synchronization properties* of the oscillator pair and these may include: Phase locking, Phase synchronization, Frequency synchronization, Sensitivity to detuning and Ratcheting. The former three are discussed for example in [31] while the latter two are discussed in [6, 23] and we outline some of the discussion and results from these papers.

10.2.1.1 Phase and Frequency Synchronization

For a solution $\theta(t) = (\theta_1(t), \ldots, \theta_N(t))$ of (10.6), let $\theta^L(t) = (\theta_1^L(t), \ldots, \theta_n^L(t))$ denote the lifted phase variables. We say the oscillator pair (i, j) is *phase synchronized* on the solution $\theta(t)$ if $\theta_i^L(t) - \theta_j^L(t)$ is bounded for all t and *phase locked* if $\lim_{t\to\infty}(\theta_i^L(t) - \theta_j^L(t))$ exists. We say oscillators are *frequency synchronized* if $\lim_{t\to\infty} \frac{\theta_i(t)-\theta_j(t)}{t} = 0$. Note that phase locking implies phase synchronization and phase synchronization implies frequency synchronization. However, the converses are not true in general. For example, on a typical solution approaching to a heteroclinic network, particular pairs of oscillators are never phase locked, but they can be phase synchronized if the heteroclinic network is contractible to the diagonal in $\mathbb{T}^n$. More interesting is the effect of heteroclinic ratchets on the synchronization properties of oscillators. On a solution approaching to a heteroclinic ratchet, some oscillator pairs can be frequency synchronized but not phase synchronized as we will see in Sect. 10.4.

10.2.1.2 Sensitivity to Detuning and Ratcheting

It is known that when the oscillators are synchronized, a mismatch in natural frequencies, that is, detuning may result in loss of synchronization depending on how large the detuning is. Let $\omega_{ij} = \omega_i - \omega_j$ denote the detuning and $\Omega_{ij} = \Omega_i - \Omega_j$ the difference in observed average frequencies. Here $\Omega_i = \lim_{t\to\infty} \frac{\theta_i^L}{t}$. The typical $(\omega_{ij}, \Omega_{ij})$ characteristic of coupled oscillators is as in Fig. 10.3a.

For an ordered oscillator pair (i, j), we generalize notions in [6] to define the *tolerance to positive detuning* and *tolerance to negative detuning* as

$$\Delta_{ij}^{+} := \sup\{\Delta\colon 0 \le \omega_{ij} < \Delta \implies \text{(i,j) is phase synchronized on all attractors of (10.6)}\}$$

$$\Delta_{ij}^{-} := \sup\{\Delta\colon -\Delta < \omega_{ij} \le 0 \implies \text{(i,j) is phase synchronized on all attractors of (10.6)}\},$$

We call $\Delta_{ij} := \min(\Delta_{ij}^{-}, \Delta_{ij}^{+})$ the *tolerance to detuning* of (i, j). If $\Delta_{ij} = 0$ then the oscillator pair (i, j) is said to have extreme sensitivity to detuning. If $\Delta_{ij}^{+} = 0$ but $\Delta_{ij}^{-} > 0$, we say that the oscillator pair (i, j) is *ratcheting* (see Fig. 10.3) (for the details see [22]). Note that ratcheting is an asymmetric relation on the set of oscillators, that is, if (i, j) is ratcheting then (j, i) is not ratcheting. In the following, we will show that heteroclinic networks may result in extreme sensitivity to detuning (Sect. 10.3) and the heteroclinic ratchets give rise to ratcheting of some oscillator pairs (Sect. 10.4).

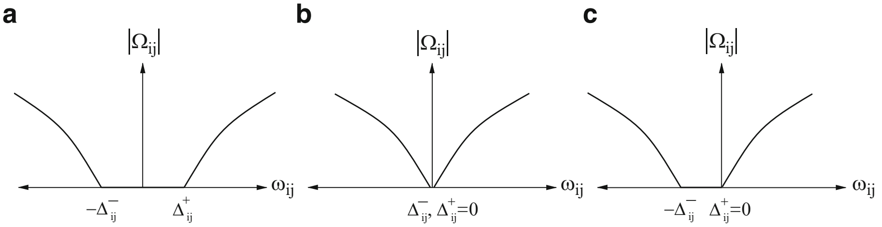


Fig. 10.3 Different $(\omega_{ij}, |\Omega_{ij}|)$-characteristics of coupled oscillators. (**a**) Usual case: Frequency synchronization of the oscillators persist in a certain tolerance range of detuning. (**b**) Extreme sensitivity to detuning: Although there is a dynamically stable frequency synchronized behaviour at $\omega_{ij} = 0$, synchronization is broken by arbitrarily small detuning. This can happen if there is an attracting heteroclinic cycle in state space (see Sect. 10.3). (**c**) Unidirectional extreme sensitivity to detuning (or ratcheting): Under small detuning synchronization is broken only if the detuning is positive

10.3 Heteroclinic Cycles and Extreme Sensitivity

An attracting heteroclinic cycle in the phase difference space of a coupled oscillator system has a strong effect on the synchronization properties of oscillators. For instance, a solution approaching to a heteroclinic cycle implies the absence of phase locking of certain oscillator pairs. Moreover, heteroclinic cycles are related to the extreme sensitivity phenomenon [6].

Heteroclinic cycles induce an intermittent behaviour called *slow switching* where the dynamics stays long time near one cluster and then passes to another cluster. Slow switching behaviour of coupled oscillator systems was first studied by Hansel et al. in [19]. They found heteroclinic cycles for four globally coupled phase oscillator system with a coupling function up to second order Fourier terms ($\alpha_1 = 1.25$, $r_2 = 0.5$). After this work, heteroclinic cycles associated with slow switching were also studied for different oscillator types, such as delayed pulse-coupled integrate-and-fire oscillators [11, 26], limit cycle oscillators [25]. In the following, we will describe the heteroclinic behaviour observed in coupled phase oscillators [4–7], and explain its effect on synchronization properties. This effect has been investigated for fully symmetric (all-to-all coupled) systems but not in many other configurations.

10.3.1 Symmetric Heteroclinic Cycles for All-to-All Coupled Phase Oscillators

All-to-all coupling gives rise to S_N-permutation symmetry. This imposes many dynamically invariant subspaces arising as fixed point subspaces of subgroups of S_N. Therefore, the dynamics is trapped in invariant regions bounded by these fixed point subspaces. Let us choose the phase difference variables as $\phi_i = \theta_1 - \theta_i + 1$, $i = 1, \ldots, N-1$. Then, the invariant regions are $\{\phi \in \mathbb{T}^{N-1} : \phi_{\sigma(1)} \leq \phi_{\sigma(2)} \leq \cdots \leq \phi_{\sigma(N-1)}\}$ where σ is a permutation of oscillators. When σ is identity, this region is called *canonical invariant region* [9]. Since all these regions are symmetric images of each other, it suffices to study the dynamics on the canonical invariant region. Note that, since the dynamics is trapped in these invariant regions in the phase difference space, oscillators are always phase synchronized and therefore frequency synchronized (the subspace $\theta_i = \theta_j$ being invariant implies phase synchronization of oscillators i and j [15]). We will be more interested in the extreme sensitivity properties of oscillators for which the existence of heteroclinic networks are crucial.

For N coupled phase oscillators, heteroclinic behaviour can arise if $N \geq 3$. The case $N = 3$ and $N = 4$ is analyzed in detail by Ashwin et al. in [5]. Considering second order Fourier truncation of the coupling function, they show that for $N = 3$ a heteroclinic cycle appears as a codimension one phenomenon in phase difference space (see Fig. 10.4). This heteroclinic cycle connects the saddles labelled by P and Q on the invariant lines, which have $S_2 \times S_1$ isotropy [5]. Note that, the heteroclinic network on $\mathbb{T}^{N-1}$ formed by these heteroclinic cycles contains winding heteroclinic

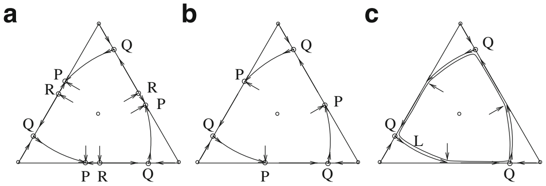


Fig. 10.4 Schematic diagrams illustrating a bifurcation of all-to-all coupled 3-oscillator system in the canonical invariant regions. The edges of the triangles represent the fixed point subspaces of the form $\{\theta_i = \theta_j\}$. On these lines two equilibria P and R (**a**) join together by a saddle-node bifurcation (**b**) and disappear giving birth to a periodic orbit in the interior of the canonical invariant region (**c**). At the bifurcation point (**b**), a heteroclinic cycle appears connecting the saddles P and Q on the invariant lines. (Adapted from [5])

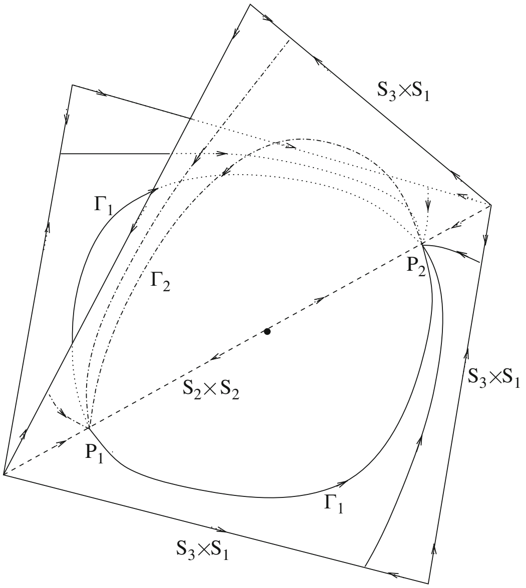


Fig. 10.5 A robust heteroclinic cycle for the all-to-all coupled 4-oscillator system. The heteroclinic cycle consists of two saddle equilibria P_1 and P_2 with $S_2 \times S_2$ isotropy and two connections Γ_1 and Γ_2 on the two dimensional invariant subspaces. The invariant subspaces are embedded in a cube that represents a unit cell for the torus of phase difference space- in this representation all vertices of the cube represent in-phase solutions where all oscillators are synchronized. (Adapted from [5])

cycles in each $\theta_i - \theta_j$ direction. Therefore any detuning Δ_{ij} gives rise to a periodic orbit that breaks the synchronization of the oscillators i and j (see [6] for details). As a result, this heteroclinic network leads to extreme sensitivity to detuning (see [6]). However, this phenomenon is not robust for $N = 3$ as it occurs at a bifurcation point.

For the case $N = 4$, one can observe robust heteroclinic cycles (see Fig. 10.5). In this case the canonical invariant region is a tetrahedron whose lines have either

$S_2 \times S_2$ or $S_3 \times S_1$ isotropy. The heteroclinic cycle shown in Fig. 10.5 exists robustly for an open set in the parameter space (see [5] for details). This time the heteroclinic network formed by these heteroclinic cycles in different invariant regions does not contain any winding heteroclinic cycle, except for the critical case when the heteroclinic cycles first appear and lie on the invariant lines. As a result, although the heteroclinic behaviour is robust when $N = 4$, the extreme sensitivity phenomenon is again not robust.

Robust extreme sensitivity behaviour arises when one considers an all-to-all coupled oscillator system with $N \geq 5$. It is numerically shown in [6] that for $N = 5$, the extreme sensitivity is robust. In [4], a heteroclinic network for the 5-oscillator all-to-all coupled system is shown to exist on the phase difference space $\mathbb{T}^4$. In this case, the heteroclinic network contains winding heteroclinic cycles in any direction breaking the frequency synchronization of oscillators, and this happens robustly under small parameter changes. This robust extreme sensitivity behaviour is bidirectional due to the presence of full permutation symmetry.

10.4 Heteroclinic Ratchets for Nonsymmetric Coupling

The heteroclinic cycles described in Sect. 10.3 are robust for $N > 3$ because they are contained in invariant subspaces forced by the symmetries of the coupling structure in such a way that connections are saddle-to-sink type in each subspace. However, these symmetries impose some restrictions on the types of possible robust heteroclinic cycles. Namely, such a cycle necessarily has the symmetries that are related to the invariant subspaces which contain parts of the heteroclinic cycle. For instance, in the case of all-to-all coupled oscillators, the heteroclinic network found in [4] have S_5 permutation symmetry. Therefore, the dynamics near the heteroclinic network is the same for each oscillator. This means one expects the same synchronization properties for all pairs of oscillators.

On the other hand, as shown in Sect. 10.1, one can find nonsymmetric robust heteroclinic cycles. These are contained in invariant subspaces not forced by the symmetry but by the balanced equivalence relations of the underlying graph [3, 17]. The balanced equivalence relations result in invariant subspace without having much restrictions on the overall dynamics as symmetry. Therefore, it is possible to find richer dynamics in such systems.

For coupled oscillators an example of a nonsymmetric heteroclinic network is discussed in [23]. This robust heteroclinic network induces different effects on different oscillators, since it includes heteroclinic cycles winding in one direction around the torus and no other cycles winding in the opposite direction. This type of heteroclinic network is called heteroclinic ratchet as its dynamical consequences are similar to a mechanical ratchet, a device that allows rotary motion on applying a torque in one direction but not in the opposite direction:

Definition 10.1. [23] For a system on $\mathbb{T}^N$, a heteroclinic network is a *heteroclinic ratchet* if it includes a heteroclinic cycle with nontrivial winding in one direction but no heteroclinic cycles winding in the opposite direction. More precisely, we say

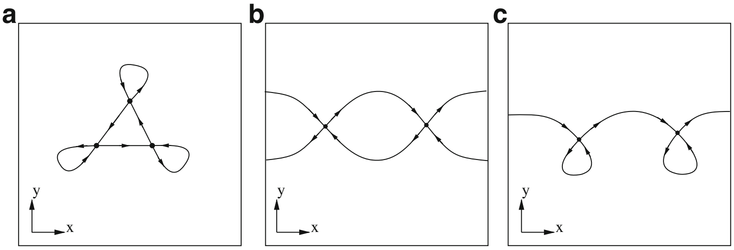


Fig. 10.6 Three different heteroclinic networks on $\mathbb{T}^2$ containing (**a**) no winding heteroclinic cycle (**b**) winding heteroclinic cycles in opposite directions $+x$ and $-x$ (**c**) one winding heteroclinic cycle in $+x$ direction. Therefore, only the network in (**c**) is a heteroclinic ratchet

a heteroclinic cycle $C \subset \mathbb{T}^N$ parametrized by $x(s)$ ($x\colon [0,1) \to \mathbb{T}^N$) has *nontrivial winding in some direction* if there is a projection map $P\colon \mathbb{R}^N \to \mathbb{R}$ such that the parametrization $\bar{x}(s)$ ($\bar{x}\colon [0,1) \to \mathbb{R}^N$) of the lifted heteroclinic cycle $\bar{C} \subset \mathbb{R}^N$ satisfies $\lim_{s\to 1} P(\bar{x}(s)) - P(\bar{x}(0)) = 2k\pi$ for some positive integer k. A heteroclinic cycle winding in the opposite direction would satisfy the same condition for a negative integer k.

In Fig. 10.6, three heteroclinic networks on a 2-torus are shown. The first network is not a heteroclinic ratchet because it does not contain a winding heteroclinic cycle. The second network contains a heteroclinic cycle winding around $+x$ direction, but it also contains a cycle winding in the opposite direction $-x$. Therefore, the only heteroclinic ratchet in the figure is the third one, which has a winding cycle in $+x$ direction.

10.4.1 A Simple Example of Ratcheting

Heteroclinic ratchets have strong effects on the synchronization properties of oscillators. An example of a heteroclinic ratchet in coupled oscillator systems is first introduced and analyzed in [23].

The coupled oscillator system considered in [23] is given by

$$\begin{aligned}
\dot{\theta}_1 &= \omega_1 + f(\theta_1; \theta_2, \theta_3) \\
\dot{\theta}_2 &= \omega_2 + f(\theta_2; \theta_1, \theta_4) \\
\dot{\theta}_3 &= \omega_3 + f(\theta_3; \theta_1, \theta_2) \\
\dot{\theta}_4 &= \omega_4 + f(\theta_4; \theta_1, \theta_2),
\end{aligned} \tag{10.8}$$

which has a coupling structure as in Fig. 10.7a. Here, the coupling function f is chosen as in (10.7). We first assume identical oscillators, that is

$$\omega = \omega_1 = \cdots = \omega_4. \tag{10.9}$$

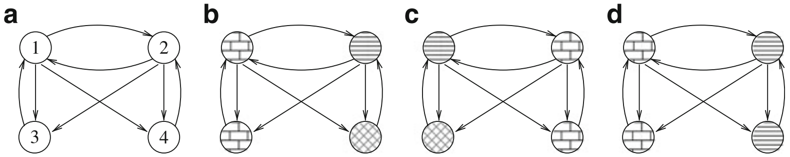


Fig. 10.7 The graph showing the connection structure of the system (10.8) (**a**) and three of its balanced colourings (**b**–**d**). Colors are represented by different filling patterns

Since then the oscillators are identical, one can use the balanced colouring method to find the invariant subspaces imposed by the coupling structure. A coloring of cells, that is, a partition of the set of all cells into a number of groups or colors is called *balanced* if each pair of cells with the same color receive same number of inputs from the cells with any given color. Three balanced colorings (beside the others) of the graph in Fig. 10.7a are shown in Fig. 10.7b–d. These give three invariant subspaces:

$$
\begin{aligned}
V_1 &= \{\theta \in \mathbb{T}^4 \colon \theta_1 = \theta_3\}\\
V_2 &= \{\theta \in \mathbb{T}^4 \colon \theta_2 = \theta_4\}\\
\bar{V} &= V_1 \cap V_2
\end{aligned}
$$

Using the phase-shift symmetry $(\{\theta_1, \ldots, \theta_4\} \to \{\theta_1 + \varepsilon, \ldots, \theta_4 + \varepsilon\}_{\text{mod } 2\pi})$ of (10.8), one can reduce the dynamics to a phase difference system on 3-torus. Defining the new variables as $(\phi_1, \phi_2, \phi_3) := (\theta_1 - \theta_3, \theta_2 - \theta_4, \theta_3 - \theta_4)$, the phase difference dynamics can be written as

$$
\begin{aligned}
\dot{\phi}_1 &= f(\phi_1; \phi_2 - \phi_3, 0) - f(0; \phi_1, \phi_2 - \phi_3)\\
\dot{\phi}_2 &= f(\phi_2; \phi_1 + \phi_3, 0) - f(0; \phi_1 + \phi_3, \phi_2)\\
\dot{\phi}_3 &= f(\phi_3; \phi_1 + \phi_3, \phi_2) - f(0; \phi_1 + \phi_3, \phi_2).
\end{aligned}
\tag{10.10}
$$

Note that the invariant subspaces V_1, V_2 and $\bar{V}$ correspond to the planes $\phi_1 = 0$, $\phi_2 = 0$ and the line $\phi_1 = \phi_2 = 0$, respectively. Therefore, it is possible that a robust heteroclinic network exists on these invariant subspaces in the phase difference space. In fact, there exists a robust heteroclinic ratchet for the parameter values $(\alpha_1, r_2, r_3) = (1.4, 0.3, -0.1)$ (see Fig. 10.8).

Heteroclinic ratchets have two main effects on the synchronization properties of oscillators: ratcheting via noise and ratcheting via detuning as described in Sect. 10.2.1.2. Both arbitrary small noise and arbitrary small detuning in a certain direction give rise to perpetual one-directional phase slips, and therefore, loss of frequency synchronization.

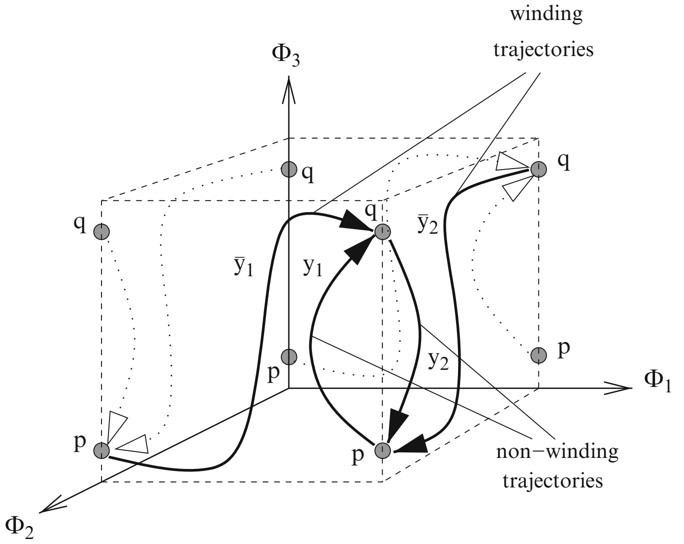


Fig. 10.8 A heteroclinic ratchet for the system (10.10). The heteroclinic network consists of two equilibria (p and q) and four heteroclinic trajectories ($y_1, y_2, \bar{y}_1, \bar{y}_2$). It contains three winding heteroclinic cycles: $(p, \bar{y}_1, q, y_2)$, $(p, \bar{y}_2, q, y_1)$ and $(p, \bar{y}_1, q, \bar{y}_2)$. (Adapted from [23])

10.4.2 Ratcheting Forced by Noise

We consider as in [23] the dynamics of (10.10) near the heteroclinic ratchet shown in Fig. 10.8. On applying small noise to the system, phase differences between oscillators grow in certain directions such that for some pairs one oscillator always has a larger average frequency than the other. Therefore, the effect of noise is not homogeneous for oscillators even though the added noise is homogeneous.

Figure 10.9 depicts a solution of (10.10) under small noise with amplitude 10^{-6}. Although the noise is homogeneous with a zero mean, phase slips occur only in $+\phi_1$ and $+\phi_2$ directions. Recall that $\phi_1 = \theta_1 - \theta_3$ and $\phi_2 = \theta_2 - \theta_4$. Therefore, the oscillator pairs $(1, 3)$ and $(2, 4)$ lose frequency synchronization such that the first oscillator has as greater average frequency then the second oscillator.

10.4.3 Ratcheting Forced by Detuning

The effect of detuning, setting $\Delta_{ij} = \omega_i - \omega_j$ nonzero, on a heteroclinic ratchet is similar to the effect of noise. A system with a heteroclinic ratchet winding in some direction on the phase difference space $\mathbb{T}^{N-1}$, say $+\phi_i = \theta_{m_i} - \theta_{m_j}$, responds to a positive detuning $\Delta_{m_i n_i} > 0$ by breaking frequency synchronization, whereas a small enough negative detuning, $\Delta_{m_i n_i} < 0$, leaves the frequency synchronization unchanged. We call this phenomenon *unidirectional extreme sensitivity to detuning*.

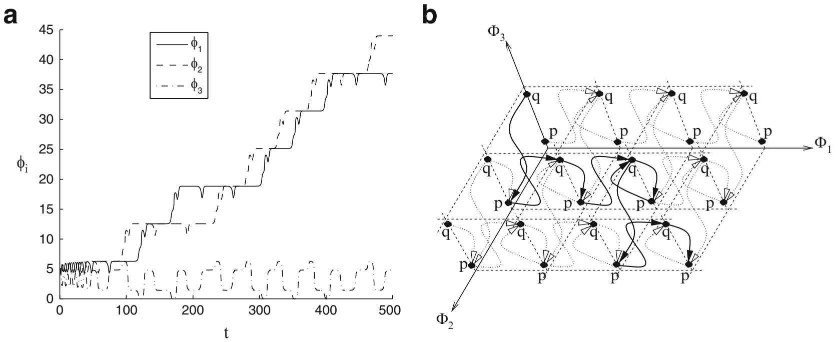


Fig. 10.9 The figure in (**a**) shows a time series solution of the system (10.10) under white noise with amplitude 10^{-6}. For the first half of the solution in (**a**) the switchings between saddle states are shown in (**b**). (Adapted from [23])

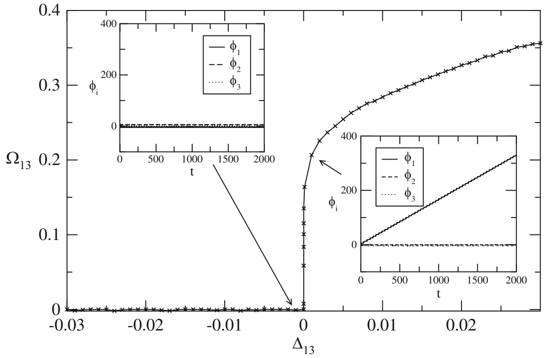


Fig. 10.10 Difference in observed average frequencies of the oscillators 1 and 3 are plotted for different detuning values $\Delta_{13} = \omega_1 - \omega_3$. Insets show time series solutions of (10.10) for small negative and positive detuning. (Adapted from [23])

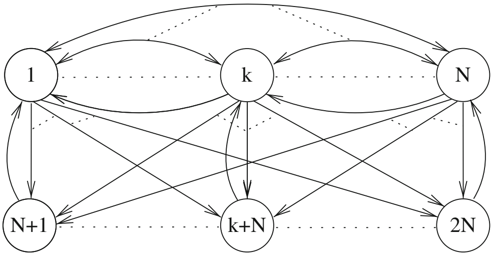


Fig. 10.11 Schematic diagram of coupling for a $2N$-cell network that admits heteroclinic ratchets. Upper cells are all-to-all coupled between themselves and each upper cell receives an extra input from the cell below itself. A lower cell receives one input from each upper cell

As shown in [23] for the system (10.10), oscillators 1 and 3 (2 and 4) loose frequency synchronization when $\Delta_{13} > 0$ ($\Delta_{24} > 0$) (see Fig. 10.10). It is also noted in [23] that a $2N$-cell coupled oscillator system as in Fig. 10.11 can admit heteroclinic ratchets ratcheting in $\theta_k - \theta_{k+N}$, $k = 1, \ldots, N$ directions. A solution of such coupled systems for $2N = 6$ is illustrated in Fig. 10.12.

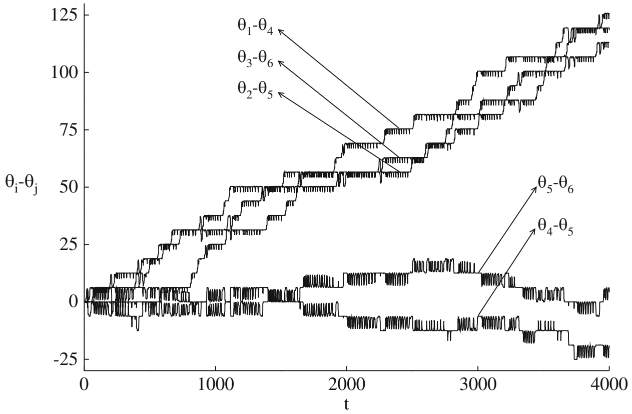


Fig. 10.12 A solution of coupled 6-oscillator system coupled as in Fig. 10.11 (with $2N = 6$) under small noise with amplitude 10^{-6}. Coupling function is chosen as $f(x) = \sin(x + 1.15) + 0.3 \sin 2x - 0.1 \sin 3x$. One directional phase slips in $\theta_k - \theta_{k+3}$, $k = 1, 2, 3$ directions suggest the existence of an attracting heteroclinic ratchet on $\mathbb{T}^6$. (Adapted from [23])

10.5 Discussion

In summary, we have reviewed some of the basic properties of robust heteroclinic networks that arise in coupled systems of nonlinear oscillators; these can manifest themselves as intermittent switching between various different partially synchronized states. Such dynamics have been observed in various systems including coupled chemical reactors [24, 38] and models of neural activity [32].

These systems provide a rich set of examples of nontrivial dynamical behaviours where the dynamics of individual systems, the topology of the network and the nature of the coupling can give networks within phase space of surprising richness. These heteroclinic networks may wind around the torus which is the natural phase space for coupled oscillator systems to give rise to topologically nontrivial networks (leading to robust extreme sensitivity to detuning) and to with nontrivial unidirectional winding in such networks (leading to heteroclinic ratcheting).

Robust extreme sensitivity is of interest in that it can only appear in globally coupled systems of five or more coupled identical oscillators and in that sense it is a truly high dimensional phenomenon. It remains to be seen to what extent it can be found systems that are not globally coupled. The heteroclinic ratcheting can be understood by a careful analysis of the nonlinear dynamics in phase space. We conjecture they may be of interest as analogues of brownian ratcheting systems in molecular dynamics with a significant difference that they are based on detuning or noise perturbed dissipative systems rather than diffusing systems in a modulated periodic potential [33]. They may also be of use as a possible "circuit elements" or "dynamical motifs" in neural computational systems [36, 39] where we suggest that the presence of such a network can lead to a robust clamping of one oscillator frequency to be above another.

Acknowledgements We thank Alberto Pinto for the invitation to contribute this paper to the volume, David Rand for his insights and inspiration in applying nonlinear dynamics to biological systems and Sasha Burylko, Mike Field, Yuri Maistrenko and many others who have contributed to the research discussed here.

References

1. Abrams, D.M., Strogatz, S.H.: Chimera states for coupled oscillators. Phys. Rev. Lett. **93**(17), 174102 (2004)
2. Acebron, J.A., Bonilla, L.L., Vicente, C.J.P., Ritort, F., Spigler, R.: The Kuramoto model: A simple paradigm for synchronization phenomena. Reviews of Modern physics **77**(1), 137–185 (2005)
3. Aguiar, M.A.D., Ashwin, P., Dias, A.P.S., Field, M.: Robust heteroclinic cycles in coupled cell systems: identical cells with asymmetric inputs, to appear in: J. Nonlinear Science (2011)
4. Ashwin, P., Borresen, J.: Encoding via conjugate symmetries of slow oscillations for globally coupled oscillators. Phys. Rev. E **70**(2), 026203 (2004)
5. Ashwin, P., Burylko, O., Maistrenko, Y.: Bifurcation to heteroclinic cycles and sensitivity in three and four coupled phase oscillators. Physica D **237**, 454–466 (2008)
6. Ashwin, P., Burylko, O., Maistrenko, Y., Popovych, O.: Extreme sensitivity to detuning for globally coupled phase oscillators. Phys. Rev. Lett. **96**(5), 054102 (2006)
7. Ashwin, P., Orosz, G., Wordsworth, J., Townley, S.: Dynamics on networks of clustered states for globally coupled phase oscillators. SIAM J. Appl. Dyn. Sys. **6**(4), 728–758 (2007)
8. Ashwin, P., Rucklidge, A.M., Sturman, R.: Cycling chaotic attractors in two models for dynamics with invariant subspaces. Chaos **14**(3), 571–582 (2004)
9. Ashwin, P., Swift, J.W.: The dynamics of n weakly coupled identical oscillators. J. Nonlinear Sci. **2**(1), 69–108 (1992)
10. Boccaletti, S., Kurths, J., Osipov, G., Valladares, D.L., Zhou, C.S.: The synchronization of chaotic systems. Phys. Rep. **366**, 1–101 (2002)
11. Broer, H., Efstathiou, K., Subramanian, E.: Heteroclinic cycles between unstable attractors. Nonlinearity **21**(6), 1385–1410 (2008)
12. Busse, F.H., Clever, R.M.: Nonstationary convection in a rotating system. In: Müller, U., Roesner, K.G., Schmidt, B. (eds.) Recent Developments in Theoretical and Experimental Fluid Dynamics, pp. 376–385. Springer, Berlin (1979)
13. Field, M.J.: Lectures on bifurcations, dynamics and symmetry, vol. 356 of Pitman Research Notes in Mathematics Series. Longman, Harlow (1996)
14. Field, M.J.: Dynamics and symmetry, vol. 3 of ICP Advanced Texts in Mathematics. Imperial College Press, London (2007)
15. Golubitsky, M., Josic, K., Shea-Brown, E.: Winding numbers and average frequencies in phase oscillator networks. J. Nonlinear Sci. **16**(3), 201–231 (2006)
16. Golubitsky, M., Stewart, I.: The Symmetry Perspective. Birkhäuser Verlag, Basel (2002)
17. Golubitsky, M., Stewart, I.: Nonlinear dynamics of networks: the groupoid formalism. Bull. Am. Math. Soc. (N.S.) **43**(3), 305–364 (electronic), (2006)
18. Guckenheimer, J., Holmes, P.: Structurally stable heteroclinic cycles. Math. Proc. Camb. Phil. Soc. **103**, 189–192 (1988)
19. Hansel, D., Mato, G., Meunier, C.: Clustering and slow switching in globally coupled phase oscillators. Phys. Rev. E **48**(5), 3470–3477 (1993)
20. Hofbauer, J., Sigmund, K.: Evolutionary Games and Population Dynamics. Cambridge University Press, Cambridge (1998)
21. Hoppensteadt, F.C., Izhikevich, E.M.: Weakly connected neural networks, vol. 126 of Applied Mathematical Sciences. Springer, New York (1997)
22. Karabacak, O.: PhD thesis, University of Exeter, 2010

23. Karabacak, O., Ashwin, P.: Heteroclinic ratchets in networks of coupled oscillators. J. Nonlinear Sci., pages DOI:10.1007/s00332-009-9053-2 (2009)
24. Kiss, I.Z., Rusin, C.G., Kori, H., Hudson, J.L.: Engineering complex dynamical structures: sequential patterns and desynchronization. Science **316**, 1886–1889 (2007)
25. Kori, H., Kuramoto, Y.: Slow switching in globally coupled oscillators: robustness and occurence through delayed coupling. Phys. Rev. E **63**(046214) (2001)
26. Kori, H., Kuramoto, Y.: Slow switching in a population of delayed pulse-coupled oscillators. Phys. Rev. E **68**(021919) (2003)
27. Krupa, M.: Robust heteroclinic cycles. J. Nonlinear Sci. **7**(2), 129–176 (1997)
28. Kuramoto, Y.: Chemical Oscillations, Waves and Turbulence. Springer, Berlin (1984)
29. Kuznetsov, A.S., Kurths, J.: Stable heteroclinic cycles for ensembles of chaotic oscillators. Phys. Rev. E (3) **66**(2), 026201, 4, (2002)
30. Mosekilde, E., Maistrenko, Y., Postnov, D.: Chaotic Synchronization: Applications to Living Systems. World Scientific Publishing, River Edge, NJ (2002)
31. Pikovsky, A., Rosenblum, M., Kurths, J.: Synchronization, a Universal Concept in Nonlinear Sciences. Cambridge University Press, Cambridge (2001)
32. Rabinovich, M.I., Huerta, R., Varona, P., Afraimovich, V.S.: Generation and reshaping of sequences in neural systems. Bio. Cyb. **95**, 519–536 (2006)
33. Reimann, P.: Brownian motors: noisy transport far from equilibrium. Phys. Rep.-Rev. Sec. Phys. Lett. **361**(2-4), 57–265 (2002)
34. dos Reis, G.L.: Structural Stability of Equivariant Vector-Fields on 2-Manifolds. Trans. Am. Math. Soc. **283**(2), 633–643 (1984)
35. Sakaguchi, H., Kuramoto, Y.: A soluble active rotator model showing phase transitions via mutual entrainment. Prog. Theor. Phys. **76**(3), 576–581 (1986)
36. Sporns, O., Kötter, R.: Motifs in brain networks. PLoS Biol. **2**(11), 1910–1918 (2004)
37. Strogatz, S.H.: From Kuramoto to Crawford: exploring the onset of synchronization in populations of coupled oscillators. Physica D **143**, 1–20 (2000)
38. Zhai, Y.M., Kiss, I.Z., Daido, H., Hudson, J.L.: Extracting order parameters from global measurements with application to coupled electrochemical oscillators. Physica D **205**, 57–69 (2005)
39. Zhigulin, P.Z.: Dynamical motifs: building blocks of complex dynamics in sparsely connected random networks. Phys. Rev. Lett. **92**(23), 238701 (2004)

Chapter 11
An Economical Model For Dumping by Dumping in a Cournot Model

Nilanjan Banik, Fernanda A. Ferreira, J. Martins, and Alberto A. Pinto

Abstract We consider an international trade economical model where two firms of different countries compete in quantities and can use three different strategies: (i) repeated collusion, (ii) deviation from the foreigner firm followed by punishment by the home country and then followed by repeated Cournot, or (iii) repeated deviation followed by punishment. In some cases (ii) and (iii) can be interpreted as dumping. We compute the profits of both firms for each strategy and we characterize the economical parameters where each strategy is adopted by the firms.

11.1 Introduction

In an international trade where one firm from the home country is competing with another firm from a foreign country, the phenomena of dumping happens often for several reasons. The foreign firm profit increases in the periods of dumping while

N. Banik (✉)
Center for Advanced Financial Studies, Institute for Financial Management and Research,
Chennai 600034, India
e-mail: nilbanik@gmail.com

F.A. Ferreira
ESEIG, Polytechnic Institute of Porto, Rua D. Sancho I, 981, 4480-876 Vila do Conde, Portugal
e-mail: fernandaamelia@eu.ipp.pt

J. Martins
LIAAD-INESC Porto LA, Department of Mathematics, School of Technology and Management,
Polytechnic Institute of Leiria, Campus 2, Morro do Lena – Alto do Vieiro, 2411-901 Leiria,
Portugal
e-mail: jmmartins@estg.ipleiria.pt

A.A. Pinto
LIAAD-INESC Porto LA e Departamento de Matemática, Faculdade de Ciências,
Universidade do Porto, Rua do Campo Alegre, 687, 4169-007, Portugal
and
Centro de Matemática e Departamento de Matemática e Aplicações, Escola de Ciências,
Universidade do Minho, Campus de Gualtar, 4710-057 Braga, Portugal
e-mail: aapinto@fc.up.pt

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_11,

the home firm profit decreases. As a response, the domestic firm can try to impose a penalty by lobbying its government to impose a tariff on the foreign firm. There are two ways in which the domestic firm can induce its government to impose a tariff. First, the domestic firm can strategically alter its behavior (trying to make the foreign firm deviate) and thereby influence antidumping outcome in the second stage of the game. Ethier and Fischer [6], Fischer [14], Staiger and Wolak [24] and Reitzes [22] mention this 'behavioral' aspect of the domestic firm. Second is by mounting political pressure. For instance, Moore [17, 18], DeVault [3], and Hansen and Prusa [15,16] have shown that industries with production facilities in the districts of legislators fare are better in terms of receiving antidumping protection.

In this work, we will study three different strategies taken by the home firm and the foreign firm in an infinitely repeated game. The first strategy involves collusion, where both firms cooperate in every period of the game, to their mutual benefit. However, after a period of collusion the foreign firm may decide to dump, thereby deviating from the collusion equilibrium. As a consequence, the foreign firm realizes a higher profit compared to collusive profit and the home firm realizes a smaller one. Hence, the home firm can lobby its government to impose a punishment tariff on the foreign firm, in the period after the deviation. These two periods of deviation-punishment can be repeated forever or can be followed by a Cournot competition, where each firm plays to maximize its own profit.

11.2 The Duopoly Model

We consider an economy consisting of a duopoly in which both firms, F_1 the domestic and F_2 the foreign firm, compete on quantities rather than price [23] of production for a certain good. Let q_i denote the produced quantities for firm F_i, $i = 1, 2$, and p_i the selling prices. We suppose that the utility function is quadratic [25]

$$U(q_1, q_2) = \alpha_1 q_1 + \alpha_2 q_2 - \frac{1}{2}\left(\beta_1 q_1^2 + 2\gamma q_1 q_2 + \beta_2 q_2^2\right) \tag{11.1}$$

giving the linear inverse demand functions [4,5]

$$\begin{array}{l} p_1 = \alpha_1 - \beta_1 q_1 - \gamma q_2 \\ p_2 = \alpha_2 - \gamma q_1 - \beta_2 q_2 \end{array}. \tag{11.2}$$

We consider that $\beta_i > 0$ and $\beta_1\beta_2 \geq \gamma^2$. The value of γ is the measure of the substitutability of the produced goods. These can be substitutes, independent, or complements according to whether $\gamma > 0$, $\gamma = 0$ or $\gamma < 0$. The goods are identical if $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2 = \gamma$. When the goods are nonidentical, the firm with the net absolute advantage in demand will enjoy of a higher α_i value. We assume that both firms have constant and equal marginal cost $c_i < \alpha_i$. Hence, the profit for the firm F_i is given by

$$\pi_i = (p_i - c_i)q_i = (\alpha_i - \beta_i q_i - \gamma q_j - c_i)q_i.$$

Without loss of generality, we can set the marginal costs equal to zero, replacing $\alpha_i - c_i$ by α_i again, and therefore the profit function for the firm F_i is given by

$$\pi_i = (\alpha_i - \beta_i q_i - \gamma q_j)q_i. \tag{11.3}$$

We compute the profits of both firms in collusion, Cournot and deviation followed by punishment strategies.

11.2.1 Collusion

In the collusion game, we consider that both firms cooperate for their mutual benefit and therefore, each firm will produce the quantities that maximizes the joint profit

$$\pi_1 + \pi_2 = \alpha_1 q_1 + \alpha_2 q_2 - \beta_1 q_1^2 - \beta_2 q_2^2 - 2\gamma q_1 q_2. \tag{11.4}$$

In the following Lemma, we present the equilibrium for the quantities and the profits of both firms when they play a collusion game. We consider that both firms are competing in every period of the game and therefore the produced quantities of the good are strictly positive. We will assume that

$$A_{i,j} = \alpha_i \beta_j - \alpha_j \gamma > 0\ ,$$

and similar assumptions are made throughout the article.

Lemma 11.1. *The equilibrium of the collusion game is attained at*

$$q_i = \frac{\alpha_i \beta_j - \alpha_j \gamma}{2\,(\beta_1 \beta_2 - \gamma^2)} \tag{11.5}$$

for the firm F_i and the correspondent profit is given by

$$\pi_i = \frac{\alpha_i}{4} \frac{\alpha_i \beta_j - \alpha_j \gamma}{\beta_1 \beta_2 - \gamma^2}. \tag{11.6}$$

See the proof of Lemma 11.1 in [1].

In the case of having $\alpha_i \beta_j = \alpha_j \gamma$, the amount of the good produced by the firm F_i is zero. Hence, the other firm F_j produces the monopoly quantity. The monopoly quantity is the one that maximizes the profit

$$\pi_j = (\alpha_j - \beta_j q_j)q_j, \tag{11.7}$$

and therefore is given by

$$\frac{\partial \pi_j}{\partial q_j} = 0 \Leftrightarrow q_j = \frac{\alpha_j}{2\beta_j}. \tag{11.8}$$

By (11.7), the monopoly profit for firm F_j is

$$\pi_j = \frac{\alpha_j^2}{4\beta_j}. \tag{11.9}$$

11.2.1.1 The identical goods limiting case

When the goods produced by the home firm and the foreign firm are identical, i.e. $\beta_1 = \beta_2 = \gamma = \beta$ and $\alpha_1 = \alpha_2 = \alpha$, the collusion equilibrium is not well defined. Indeed, the join profit is now given by

$$\begin{aligned} \pi_1 + \pi_2 &= \alpha q_1 + \alpha q_2 - \beta q_1^2 - \beta q_2^2 - 2\beta q_1 q_2 \\ &= \alpha(q_1 + q_2) - \beta(q_1 + q_2)^2. \end{aligned} \tag{11.10}$$

Hence, taking the amount derivative of $\pi_1 + \pi_2$ for both firms, we obtain the same equation $\alpha - 2\beta(q_1 + q_2)$ giving

$$q_1 + q_2 = \frac{\alpha}{2\beta}, \tag{11.11}$$

that defines an infinite number of equilibria in the collusion strategy. We observe that $\alpha/2\beta$ is the monopoly quantity computed in (11.8). In this case, the joint profit for both firms is given by

$$\begin{aligned} \pi_1 + \pi_2 &= \alpha\frac{\alpha}{2\beta} - \beta\left(\frac{\alpha}{2\beta}\right)^2 \\ &= \frac{\alpha^2}{4\beta}, \end{aligned} \tag{11.12}$$

which also corresponds to the monopoly profit, as presented in (11.9). Since the firms cooperate, in their mutual benefit, we assume that both firms produce the same quantities of the good and therefore each one produce

$$q_i = \frac{\alpha}{4\beta}. \tag{11.13}$$

Under this symmetric hypothesis, the profit for the home and the foreign firm follows immediately as

$$\pi_i = \frac{\alpha^2}{8\beta}. \tag{11.14}$$

11.2.2 Cournot

In the Cournot game, firms decide simultaneously and independently of each other the quantities of goods to be produced in order to maximize their own profit.

Lemma 11.2. *In the Cournot strategy the Nash equilibrium is given by*

$$q_i = \frac{2\alpha_i\beta_j - \alpha_j\gamma}{4\beta_1\beta_2 - \gamma^2} \tag{11.15}$$

and the corresponding profits are given by

$$\pi_i = \beta_i \left(\frac{2\alpha_i\beta_j - \alpha_j\gamma}{4\beta_1\beta_2 - \gamma^2}\right)^2 . \tag{11.16}$$

See the proof of Lemma 11.2 in [1].

11.2.3 Deviation Followed by Punishment

We suppose that after a period of collusion, when both firms produce under cooperation the quantity given by

$$q_i = \frac{\beta_j\alpha_i - \alpha_j\gamma}{2\left(\beta_i\beta_j - \gamma^2\right)},$$

that the foreign firm deviates from this quantity to maximize its own profit. The deviation from collusion can be understood as dumping and a period of punishment can be imposed. Once the foreign firm deviates from the collusion equilibrium expanding its production with the marginal costs unchanged, and if the home firm does not respond in the same period, the foreign firm's profit rises comparably to the collusion profit. This profit increase constitutes an incentive to the foreign firm to deviate from collusion. However, since the price being charged for the good in the home market is lower than the price charged before, this deviation can be interpreted as dumping. Assuming that the foreign firm deviates optimally to maximize its profit, the amount of the good produced is given by

$$\begin{aligned}\frac{\partial\pi_2}{\partial q_2} = 0 \Leftrightarrow q_2^D &= \frac{1}{2}\frac{\alpha_2 - \gamma q_1}{\beta_2} \\ &= \frac{1}{2\beta_2}\left(\alpha_2 - \gamma\frac{\beta_2\alpha_1 - \alpha_2\gamma}{2\beta_1\beta_2 - 2\gamma^2}\right)\end{aligned}$$

$$= \frac{1}{4\beta_2(\beta_1\beta_2 - \gamma^2)} \left(2\alpha_2\beta_1\beta_2 - 2\alpha_2\gamma^2 - \gamma\beta_2\alpha_1 + \alpha_2\gamma^2\right)$$

$$= \frac{2\alpha_2\beta_1\beta_2 - \alpha_2\gamma^2 - \gamma\beta_2\alpha_1}{4\beta_2(\beta_1\beta_2 - \gamma^2)}.$$

Hence, in the deviation period the profit of the home firm is given by

$$\pi_1 = \left(\alpha_1 - \beta_1 \frac{\beta_2\alpha_1 - \alpha_2\gamma}{2(\beta_1\beta_2 - \gamma^2)} - \gamma \frac{2\alpha_2\beta_1\beta_2 - \alpha_2\gamma^2 - \gamma\beta_2\alpha_1}{4\beta_2(\beta_1\beta_2 - \gamma^2)}\right) \frac{\beta_2\alpha_1 - \alpha_2\gamma}{2(\beta_1\beta_2 - \gamma^2)}$$

$$= \left(2\alpha_1\beta_1\beta_2^2 - 3\beta_2\alpha_1\gamma^2 + \alpha_2\gamma^3\right) \frac{\beta_2\alpha_1 - \alpha_2\gamma}{8\beta_2(\beta_1\beta_2 - \gamma^2)^2} \tag{11.17}$$

and the profit of the foreign firm given by

$$\pi_2 = \beta_2 \left(\frac{2\alpha_2\beta_1\beta_2 - \alpha_2\gamma^2 - \gamma\beta_2\alpha_1}{4\beta_2(\beta_1\beta_2 - \gamma^2)}\right)^2. \tag{11.18}$$

Consequently, in the deviation period the profit realized by the home firm is smaller than the collusion profit. To prevent the foreign firm from dumping in the short run and to recover from the unfair practices, the home firm will try to lobby its government to impose antidumping duties on the foreign firm. Antidumping duties can be imposed in the next period, only if the domestic firm has not increased its output in the period in which the foreign firm deviates. The rationale for this assumption is that domestic firm must prove material injury, and this would be difficult to prove under expanded production. In the punishment phase that occurs in the period after deviation, we assume that the home firm has successfully lobbied the government to the price L, hence

$$\pi_1 = (\alpha_1 - \beta_1 q_1 - \gamma q_2)\, q_1 - L.$$

As punishment, the government imposes a prohibitive tariff τ, per unit, to the foreign firm

$$\pi_2 = (\alpha_2 - \beta_2 q_2 - \gamma q_1 - \tau)\, q_2.$$

This prohibitive tariff ensures that the foreign firm earns zero profit in the home market during the punishment phase, $\pi_2 = 0$, producing nothing. Hence, the domestic firm produces the monopoly quantity

$$q_1 = \frac{\alpha_1}{2\beta_1},$$

leading to the profit given by

$$\begin{aligned}\pi_1 &= \left(\alpha_1 - \beta_1 \frac{\alpha_1}{2\beta_1} - \gamma 0\right) \frac{\alpha_1}{2\beta_1} - L \\ &= \frac{\alpha_1^2}{4\beta_1} - L.\end{aligned}$$

11.3 Infinitely Repeated Games

Now, we consider the situation where both firms have to choose their produced quantities over several periods. We will consider the following possible strategies:

1. Collusion strategy (COL) when both firms cooperate in every periods maximizing the join profit of both firms.
2. Deviation-Punishment strategy (DP) when the foreign firm deviates from collusion maximizing its own profit followed, in the next period, by a punishment strategy of the home firm. The punishment comes from the home firm lobbying its own government, resulting in a prohibitive tariff on the foreign firm during the second period. After these two periods of the game, we consider two different possible repeated strategies than can happen:
   a. Deviation-Punishment Repeated strategy (DPR), where the strategy taken by the firms in the previous two periods will keep being repeated.
   b. Deviation-Punishment followed by a Cournot strategy (DPC), where after taking the deviation-punishment strategy in the past two periods, the firms do not cooperate and adopt a Cournot strategy.

11.3.1 COL Strategy

Let $\delta \in (0, 1)$ denote the rate of discount. Let $\pi_{T,1}^{COL}$ denote the total profit of the home firm, when both the home and the foreign firms play a collusion strategy in every period of the game, which is given by

$$\begin{aligned}\pi_{T,1}^{COL} &= (1-\delta)\left(\pi_1^{COL} + \delta\pi_1^{COL} + \delta^2\pi_1^{COL} + ...\right) \\ &= (1-\delta)\frac{1}{1-\delta}\pi_1^{COL} \\ &= \pi_1^{COL}, \qquad (11.19)\end{aligned}$$

where π_1^{COL} denotes the home firm profit in one period of the collusion strategy. We consider the $(1-\delta)$ factor in every expression for the total profit in order to simplify the final value. Applying for π_1^{COL} the corresponding expression presented in Lemma 11.1, we obtain

$$\pi_{T,1}^{COL} = \frac{\alpha_1\left(\alpha_1\beta_2 - \alpha_2\gamma\right)}{4(\beta_1\beta_2 - \gamma^2)}. \tag{11.20}$$

Similarly, for the foreign firm we obtain

$$\pi_{T,2}^{COL} = \frac{\alpha_2\left(\alpha_2\beta_1 - \alpha_1\gamma\right)}{4(\beta_1\beta_2 - \gamma^2)}, \tag{11.21}$$

as the total value of the profit in the repeated collusion strategy.

11.3.2 DPR Strategy

For the case of the Deviation-Punishment Repeated strategy being practiced by both firms, the home firm profit is given by

$$\pi_{T,1}^{DPR} = (1-\delta)\left(\pi_1^D + \delta\pi_1^P + \delta^2\pi_1^D + \delta^3\pi_1^P + ...\right), \tag{11.22}$$

where π_1^D denotes the home firm profit in one period in which the foreign firm deviates from collusion and π_1^P denotes the profit in the punishment phase that corresponds to the monopoly profit minus the lobby price. Hence,

$$\pi_{T,1}^{DPR} = \frac{\pi_1^D + \delta\pi_1^P}{1+\delta},$$

and using the profits in (11.17) and (11.9) we obtain

$$\pi_{T,1}^{DPR} = \frac{\left(2\alpha_1\beta_1\beta_2^2 - 3\beta_2\alpha_1\gamma^2 + \alpha_2\gamma^3\right)\frac{\beta_2\alpha_1 - \alpha_2\gamma}{8\beta_2(\beta_1\beta_2-\gamma^2)^2} + \delta\left(\frac{\alpha_1^2}{4\beta_1} - L\right)}{1+\delta}. \tag{11.23}$$

The total profit for the foreign firm in the Deviation-Punishment Repeated strategy is given by

$$\begin{aligned}\pi_{T,2}^{DPR} &= (1-\delta)\left(\pi_2^D + \delta 0 + \delta^2\pi_2^D + \delta^3 0 + ...\right)\\ &= \frac{\pi_2^D}{1+\delta},\end{aligned} \tag{11.24}$$

where π_2^D denotes the foreign firm profit when it deviates from collusion, given in (11.18). Hence,

$$\pi_{T,2}^{DPR} = \frac{\beta_2}{1+\delta}\left(\frac{2\alpha_2\beta_1\beta_2 - \alpha_2\gamma^2 - \gamma\beta_2\alpha_1}{4\beta_2(\beta_1\beta_2-\gamma^2)}\right)^2. \tag{11.25}$$

11.3.3 DPC Strategy

When the strategies played in the game consist in Deviation-Punishment in the first two periods followed by a Cournot competition in subsequent periods, the total profit for the home firm is given by

$$\begin{aligned}\pi_{T,1}^{DPC} &= (1-\delta)\left(\pi_1^D + \delta\pi_1^P + \delta^2\pi_1^{CN} + \delta^3\pi_1^{CN} + ...\right)\\ &= (1-\delta)\left(\pi_1^D + \delta\pi_1^P\right) + \delta^2\pi_1^{CN}, \end{aligned} \tag{11.26}$$

where π_1^{CN} denotes the home firm profit in the Cournot strategy, given by Lemma 11.2. Hence, $\pi_{T,1}^{DPC}$ is given by

$$\begin{aligned}\pi_{T,1}^{DPC} &= (1-\delta)\left(\left(2\alpha_1\beta_1\beta_2^2 - 3\beta_2\alpha_1\gamma^2 + \alpha_2\gamma^3\right)\frac{\beta_2\alpha_1 - \alpha_2\gamma}{8\beta_2(\beta_1\beta_2-\gamma^2)^2}\right.\\ &\left.+\delta\left(\frac{\alpha_1^2}{4\beta_1} - L\right)\right) + \delta^2\beta_1\left(\frac{2\alpha_1\beta_2 - \gamma\alpha_2}{4\beta_1\beta_2-\gamma^2}\right)^2. \end{aligned} \tag{11.27}$$

In this same strategy, the profit for the foreign firm is given by

$$\begin{aligned}\pi_{T,2}^{DPC} &= (1-\delta)\left(\pi_2^D + \delta 0 + \delta^2\pi_2^{CN} + \delta^3\pi_2^{CN} + ...\right)\\ &= (1-\delta)\pi_2^D + \delta^2\pi_2^{CN}\\ &= (1-\delta)\beta_2\left(\frac{2\alpha_2\beta_1\beta_2 - \alpha_2\gamma^2 - \gamma\beta_2\alpha_1}{4\beta_2(\beta_1\beta_2-\gamma^2)}\right)^2 + \delta^2\beta_2\left(\frac{2\alpha_2\beta_1 - \gamma\alpha_1}{4\beta_1\beta_2-\gamma^2}\right)^2. \end{aligned} \tag{11.28}$$

11.3.4 The Optimal Strategy

We observe that the foreign firm makes the decision between choosing the collusion strategy (COL) and the deviation-punishment strategy (DP). The home firm makes the decision between choosing the repeated deviation-punishment strategy (DPR) and the deviation-punishment followed by Cournot strategy (DPC). Hence, we have the following possible optimal strategies:

Strategy 1: If

$$\pi_{T,2}^{DPC} \geq \pi_{T,2}^{COL} \quad and \quad \pi_{T,1}^{DPC} \geq \pi_{T,1}^{DPR}$$

the best repeated strategy for the game is DPC (Deviation-Punishment followed by Cournot).

Strategy 2: If

$$\pi_{T,2}^{DPR} \geq \pi_{T,2}^{COL} \quad and \quad \pi_{T,1}^{DPR} \geq \pi_{T,1}^{DPC}$$

the best repeated strategy for the game is DPR (Deviation-Punishment Repeated).

Strategy 3: If

$$\pi_{T,2}^{COL} \geq \pi_{T,2}^{DPC} \quad and \quad \pi_{T,1}^{DPC} \geq \pi_{T,1}^{DPR},$$

or

$$\pi_{T,2}^{COL} \geq \pi_{T,2}^{DPR} \quad and \quad \pi_{T,1}^{DPR} \geq \pi_{T,1}^{DPC},$$

the best repeated strategy for the game is COL (Collusion).

For the symmetric model where, $\alpha_1 = \alpha_2 = \alpha$ and $\beta_1 = \beta_2 = 1$, we compare the profits of the different repeated strategies. Let $0 \leq L_0 \leq 1$ be the percentage value of the monopoly profit, paid by the home firm, for making lobby in the government. Let $L = L_0\alpha^2/4$, be the lobby price paid by the home firm. The home firm profits in the repeated DPR and DPC strategy coincides at the curve δ_1^{DP} defined by

$$\delta_1^{DP} = \frac{8\alpha^2(1-\gamma)(\gamma+1)^2 - \alpha^2(\gamma+2)^2(-3\gamma^2+2+\gamma^3)}{8(1-L_0)\frac{\alpha^2}{4}(\gamma+2)^2(1-\gamma)(\gamma+1)^2 - 8\alpha^2(1-\gamma)(\gamma+1)^2}. \quad (11.29)$$

Comparing the foreign firm profit in the COL repeated strategy with the profit in the DPR strategy, we observe that they are equal at the curve

$$\pi_{T,2}^{COL} = \pi_{T,2}^{DPR} \Leftrightarrow \delta_2^{COLDPR} = \frac{\gamma^2}{4(\gamma+1)}. \quad (11.30)$$

Similarly, the profits under the COL and DPC strategies coincide at the curve

$$\pi_{T,2}^{COL} = \pi_{T,2}^{DPC} \Leftrightarrow \delta_2^{COLDPC} = \frac{-B - \sqrt{B^2 - 4AC}}{2A}, \quad (11.31)$$

where

$$A = \frac{1}{(\gamma+2)^2} \quad (11.32)$$

$$B = -\frac{1}{16}\left(\frac{\gamma}{\gamma-1} + \frac{2}{\gamma^2-1}\right)^2 \quad (11.33)$$

$$C = \frac{1}{16}\left(\frac{\gamma}{\gamma-1} + \frac{2}{\gamma^2-1}\right)^2 - \frac{1}{4(\gamma+1)}. \quad (11.34)$$

In Fig. 11.1 left, we present the curve δ_1^{DP} and the curves δ_2^{COLDPR} and δ_2^{COLDPC}, considering $L_0 = 0$. We observe that, for values of $\delta < \delta_1^{DP}$, the home firm prefers

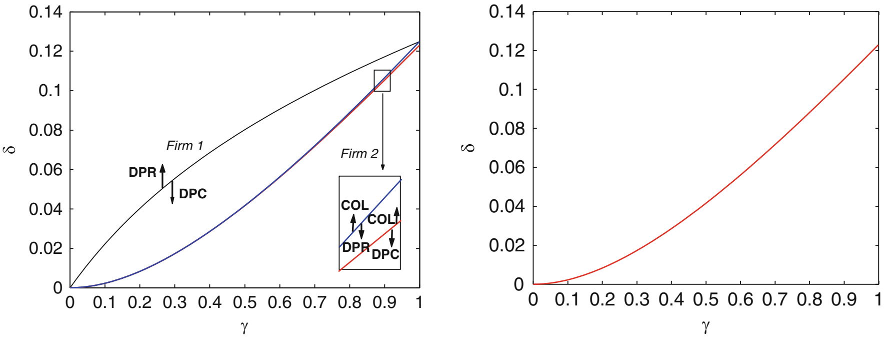


Fig. 11.1 The curves $\delta(\gamma)$ dividing the δ, γ plane in the regions that correspond to the preferred strategy for each firm. In the left figure, we have the curve δ_1^{DP}, separating the type of DP strategy that the home firm (Firm 1) will choose, and the curves δ_2^{COLDPR} and δ_2^{COLDPC} for the foreign firm (Firm 2) chooses. In the right figure, we have the curve δ_2^{COLDPC} that separates the decision over the joint strategy of both firms

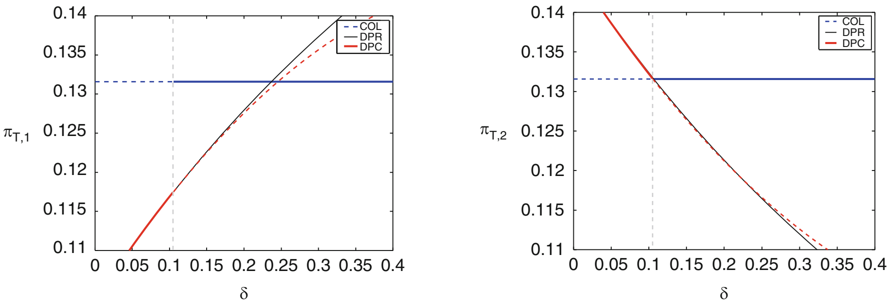


Fig. 11.2 Profits of the home firm (*left*) and the foreign firm (*right*) in the COL, DPR and DPC strategies for different values of δ. The parameters considered were $\gamma = 0.9$, $\alpha_1 = \alpha_2 = 1$, $\beta_1 = \beta_2 = 1$ and $L_0 = 0$

the DPC repeated strategy and, for values of $\delta > \delta_1^{DP}$, prefers the DPR strategy. The foreign firm prefers the DPR strategy for values of $\delta < \delta_2^{COLDPR}$. By Fig. 11.1 left, $\delta_2^{COLDPR} < \delta_1^{DP}$, and so the home firm never allows the DPR strategy to occur. For values of $\delta < \delta_2^{COLDPC}$, the foreign firm prefers the DPC strategy and, for values of $\delta > \delta_2^{COLDPC}$, the foreign firm prefers the COL strategy. Since $\delta_2^{COLDPC} < \delta_1^{DP}$, for values of $\delta < \delta_2^{COLDPC}$, both firms adopt the DPC strategy and, for values of $\delta > \delta_2^{COLDPC}$, both firms adopt the COL strategy (see Fig. 11.1 right).

Observing the profits of both firms, for a fixed value of γ, under the adopted repeated strategy, we observe a discontinuity in the profit of the home firm at δ_2^{COLDPC} and a discontinuity of the time derivative in the profit of the foreign firm at the same point. These discontinuities are illustrated in Fig. 11.2, for $\gamma = 0.9$. The discontinuities occur at $\delta_2^{COLDPC} = 0.132$.

Conclusions

In the symmetric case of the model we show that only two strategies can occur: repeated collusion or deviation followed by punishment followed by repeated Cournot. We characterise the parameter space where each one of these strategies occurs.

Acknowledgements We thank LIAAD-INESC Porto LA, Calouste Gulbenkian Foundation, PRODYN-ESF, FEDER, POFC, POCTI and POSI by FCT and Ministério da Ciência e da Tecnologia, and the FCT Pluriannual Funding Program of the LIAAD-INESC Porto LA and of the Research Centre of Mathematics of University of Minho and of Centro de Matemática da Universidade do Porto, for their financial support.

References

1. Banik, N., Ferreira, F.A., Martins, J., Pinto, A.: Dumping in a Cournot model. (submitted).
2. Brida, J., Defesa, M., Faias, M., Pinto, A.: Strategic choice in tourism with differentiated crowding types. Econ. Bull. **30** (2) 1509–1515 (2010)
3. DeVault, J.M.: Economics and the international trade commission. South. Econ. J. **60** 463–78 (1993)
4. Dixit, A.K.: Anti-dumping and countervailing duties under oligopoly. Eur. Econ. Rev. **32** 55–68 (1988)
5. Eaton, J., Grossman, G.M.: Optimal trade and industrial policy for the US automobile industrial policy under oligopoly. Q. J. Econ. **100** 383–406 (1986)
6. Ethier, W.J., Fischer, R.D.: The new protectionism. J. Int. Econ. Integr. **2** 1–11 (1987)
7. Ferreira, F.A., Ferreira, F., Pinto, A.A.: Bayesian price leadership. In: Tas, K., et al. (eds.) Mathematical Methods in Engineering, Springer 359–369 (2007)
8. Ferreira, F.A., Ferreira, F., Pinto, A.A.: Flexibility in stackelberg leadership. In: Machado, J.A.T., et al. (eds.) Intelligent Engineering Systems and Computational Cybernetics, Springer 399–405 (2008)
9. Ferreira, F.A., Ferreira, F., Pinto, A.A.: 'Own' price influences in a Stackelberg leadership with demand uncertainty. Brazil. J. Bus. Econ. **8** (1) 29–38 (2008)
10. Ferreira, F., Ferreira, F.A., Pinto, A.A.: Price-setting dynamical duopoly with incomplete information. In: Tenreiro Machado, J.A., et al. (eds.) Nonlinear Science and Complexity, Springer 397–404 (2010)
11. Ferreira, F.A., Ferreira, F., Ferreira, M., Pinto, A.A.: Quantity competition in a differentiated duopoly. In: Machado, J.A.T., et al. (eds.) Intelligent Engineering Systems and Computational Cybernetics, Springer 365–374 (2008)
12. Ferreira, F.A., Ferreira, F., Pinto, A.A.: Uncertainty on a Bertrand duopoly with product differentiation. In: Machado, J.A.T., et al. (eds.) Nonlinear Science and Complexity, Springer 389–396 (2010)
13. Ferreira, F.A., Ferreira, F., Pinto, A.A.: Unknown costs in a duopoly with differentiated products. In: Tas, K., et al. (eds.) Mathematical Methods in Engineering, Springer 371–379 (2007)
14. Fischer, D.R.: Endogenous probability of protection and firm behavior. J. Int. Econ. **32** 149–163 (1992)
15. Hansen, L.W., Prusa, T.J.: Cumulation and ITC decision making, The sum of the parts is greater than the whole. Econ. Enquiry **34** 746–769 (1996)
16. Hansen, L.W., Prusa, T.J.: The economics and politics of trade policy, An empirical analysis of ITC decision making. Rev. Int. Econ. **5** 230–245 (1997)

17. Moore, M.O.: Rules or politics? An empirical analysis of ITC antidumping decisions. Econ. Enquiry **30** 449–466 (1992)
18. Moore, M.O.: The political economy of trade protection. In: Krueger, Anne, O. (eds.) University of Chicago Press (1996)
19. Pinto, A.A.: Game Theory and Duopoly Models. Interdisciplinary Applied Mathematics Series. Springer (2011)
20. Pinto, A.A., Oliveira, B.M.P.M., Ferreira, F.A., Ferreira, M.: Investing to survive in a duopoly model. In: Machado, J.A.T., et al. (eds.) Intelligent Engineering Systems and Computational Cybernetics, Springer 407–414 (2008)
21. Pinto, A.A., Oliveira, B.M.P.M., Ferreira, F.A., Ferreira, F.: Stochasticity favoring the effects of the R&D strategies of the firms. In: Machado, J.A.T., et al. (eds.) Intelligent Engineering Systems and Computational Cybernetics, Springer 415–423 (2008)
22. Reitzes, D.J.: Antidumping policy. Int. Econ. Rev. **34** 745–763 (1993)
23. Singh, N., Vives, X.: Price and quantity competition in a differentiated duopoly. RAND J. Econ. **15** 546–554 (1984)
24. Staiger, R.W., Wolak, F.A.: The effect of domestic antidumping law in the presence of foreign monopoly. J. Int. Econ. **32** 265–287 (1992)
25. Vives, X.: Duopoly information equilibrium, Cournot and Bertrand. J. Econ. Theory **34** 546–554 (1984)

Chapter 12
Fractional Control of Dynamic Systems

Ramiro S. Barbosa and J.A. Tenreiro Machado

Abstract The concepts involved with fractional calculus (FC) theory are applied in almost all areas of science and engineering. Its ability to yield superior modeling and control in many dynamical systems is well recognized. In this article, we will introduce the fundamental aspects associated with the application of FC to the control of dynamic systems.

12.1 Introduction

Fractional calculus (FC) is the area of mathematics that extends derivatives and integrals to an arbitrary order (real or, even, complex order) and emerged at the same time as the classical differential calculus. FC generalizes the classical differential operator $D_t^n \equiv d^n/dt^n$ to a fractional operator D_t^α, where α can be a complex number [4, 8]. However, its inherent complexity delayed the application of the associated concepts.

Nowadays, the FC is applied in science and engineering, being recognized its ability to yield a superior modeling and control in many dynamical systems. We may cite its adoption in areas such as viscoelasticity and damping, diffusion and wave propagation, electromagnetism, chaos and fractals, heat transfer, biology, electronics, signal processing, robotics, system identification, traffic systems, genetic algorithms, percolation, modeling and identification, telecommunications, chemistry, irreversibility, physics, control, economy and finance [2, 6].

R.S. Barbosa (✉)
Institute of Engineering of Porto, Rua Dr. António Bernardino de Almeida, 431, 4200-072 Porto, Portugal
e-mail: rsb@isep.ipp.pt

J.A.T. Machado
Department of Electrotechnical Engineering, Institute of Engineering of Porto, Rua Dr. António Bernardino de Almeida, 431, 4200-072 Porto, Portugal
e-mail: jtm@isep.ipp.pt

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_12,

In what concerns the area of control systems the application of the FC concepts is still scarce and only in the second-half of the last century appeared the first applications [1,3,6].

12.2 Fundamentals of Fractional-Order Control Systems

In general, a fractional-order control system can be described by a Linear Time Invariant (LTI) fractional-order differential equation of the form:

$$\begin{aligned} &a_n D_t^{\beta_n} y(t) + a_{n-1} D_t^{\beta_{n-1}} y(t) + \cdots + a_0 D_t^{\beta_0} y(t) \\ &\quad = b_m D_t^{\alpha_m} u(t) + b_{m-1} D_t^{\alpha_{m-1}} u(t) + \cdots + b_0 D_t^{\alpha_0} u(t) \end{aligned} \tag{12.1}$$

or by a continuous transfer function of the form:

$$G(s) = \frac{b_m s^{\alpha_m} + b_{m-1} s^{\alpha_{m-1}} + \cdots + b_0 s^{\alpha_0}}{a_n s^{\beta_n} + a_{n-1} s^{\beta_{n-1}} + \cdots + a_0 s^{\beta_0}} \tag{12.2}$$

where $\beta_k,\ \alpha_k\ (k = 0,\ 1,\ 2,\ \ldots)$ are real numbers, $\beta_k > \cdots > \beta_1 > \beta_0,\ \alpha_k > \cdots > \alpha_1 > \alpha_0$ and $a_k,\ b_k\ (k = 0,\ 1,\ 2,\ \ldots)$ are arbitrary constants.

A discrete transfer function of (12.2) can be obtained by using a discrete approximation of the fractional-order operators, yielding:

$$G(z) = \frac{b_m \left[w\left(z^{-1}\right)\right]^{\alpha_m} + b_{m-1} \left[w\left(z^{-1}\right)\right]^{\alpha_{m-1}} + \cdots + b_0 \left[w\left(z^{-1}\right)\right]^{\alpha_0}}{a_n \left[w\left(z^{-1}\right)\right]^{\beta_n} + a_{n-1} \left[w\left(z^{-1}\right)\right]^{\beta_{n-1}} + \cdots + a_0 \left[w\left(z^{-1}\right)\right]^{\beta_0}} \tag{12.3}$$

where $w\left(z^{-1}\right)$ denotes the discrete equivalent of the Laplace operator s, expressed as a function of the complex variable z or the shift operator z^{-1}.

The generalized operator ${}_aD_t^{\alpha}$, where a and t are the limits and α the order of operation, is usually given by the Riemann–Liouville definition ($\alpha > 0$):

$${}_aD_t^{\alpha} x(t) = \frac{1}{\Gamma(n-\alpha)} \frac{d^n}{dt^n} \int_a^t \frac{x(\tau)}{(t-\tau)^{\alpha-n+1}} d\tau, \quad n-1 < \alpha < n \tag{12.4}$$

where $\Gamma(z)$ represents the Gamma function of z. Another common definition is that given by the Grünwald–Letnikov approach ($\alpha \in \Re$):

$${}_aD_t^{\alpha} x(t) = \lim_{h \to 0} \frac{1}{h^{\alpha}} \sum_{k=0}^{\left[\frac{t-a}{h}\right]} (-1)^k \binom{\alpha}{k} x(t-kh) \tag{12.5}$$

where h is the time increment and $[v]$ means the integer part of v.

The fractional-order derivatives can also be defined in the transform domain. It is shown that the Laplace transform (L) of definitions (12.4) and (12.5), under null initial conditions, is given by:

$$L\left\{D^{\alpha}x\left(t\right)\right\} = s^{\alpha}X\left(s\right) \tag{12.6}$$

where $X\left(s\right) = L\left\{x\left(t\right)\right\}$.The Laplace transform reveals to be a valuable tool for the analysis and design of fractional-order control systems.

12.3 Fractional-Order Controllers and its Implementation

The fractional-order controllers were introduced by Oustaloup, who developed the so-called *Commande Robuste d'Ordre Non Entier* (CRONE) controller [5]. More recently, Podlubny proposed a generalization of the PID controller, the PI$^{\lambda}$D$^{\mu}$-controller, involving an integrator of order λ and a differentiator of order μ [8]. The transfer function $G_c\left(s\right)$ of such a controller has the form:

$$G_c\left(s\right) = \frac{U\left(s\right)}{E\left(s\right)} = K_P + K_I s^{-\lambda} + K_D s^{\mu}, \quad \lambda,\ \mu > 0 \tag{12.7}$$

where $(K_P,\ K_I,\ K_D)$ are the proportional, integral, and derivative gains of the controller, respectively. The transfer function (12.7) is represented by a fractional integro-differential equation of type:

$$u\left(t\right) = K_P e\left(t\right) + K_I D^{-\lambda} e\left(t\right) + K_D D^{\mu} e\left(t\right) \tag{12.8}$$

Taking $(\lambda,\ \mu) \equiv (1,\ 1)$ gives a classical PID controller, $(\lambda,\ \mu) \equiv (1,\ 0)$ gives a PI controller, $(\lambda,\ \mu) \equiv (0,\ 1)$ gives a PD controller and $(\lambda,\ \mu) \equiv (0,\ 0)$ gives a P controller. All these classical types of PID controllers are the particular cases of the fractional PI$^{\lambda}$D$^{\mu}$-controller. Thus, the PI$^{\lambda}$D$^{\mu}$-controller is more flexible and gives the possibility of adjusting more carefully the dynamical properties of a control system [9].

As shown by the above expressions, the fractional-order operators are characterized by having irrational continuous transfer functions in the Laplace domain or infinite dimensional discrete transfer functions in time domain. These properties preclude their direct utilization both in time and frequency domains. Therefore, the usual approach for analysing fractional-order systems is the development of continuous and discrete integer-order approximations to these operators [10].

In order to implement the operator s^{α} $(\alpha \in \Re)$, a frequency-band limited approximation may be used by cutting out both high and low frequencies of transfer $\left(s/\omega_u\right)^{\alpha}$ to a given frequency range $\omega \in [\omega_b,\ \omega_h]$, distributed geometrically around the unit gain frequency $\omega_u = \left(\omega_b\omega_h\right)^{1/2}$ [7]. The resulting continuous transfer function of such approximation is given by the formula:

$$D_N(s) = \left(\frac{\omega_u}{\omega_h}\right)^{\alpha} \prod_{k=-N}^{N} \frac{1 + s/\omega'_k}{1 + s/\omega_k} \tag{12.9}$$

where the zero and pole of rank k can be evaluated, respectively, as:

$$\omega'_k = \left(\frac{\omega_h}{\omega_b}\right)^{\frac{k+N+\frac{1}{2}-\frac{\alpha}{2}}{2N+1}} \omega_b, \quad \omega_k = \left(\frac{\omega_h}{\omega_b}\right)^{\frac{k+N+\frac{1}{2}+\frac{\alpha}{2}}{2N+1}} \omega_b \tag{12.10}$$

On the other hand, the usual approach for obtaining discrete equivalents of continuous operators of type s^{α} ($\alpha \in \Re$) adopts the Euler, Tustin and Al-Alaoui generating functions.

It is well known that the continued fraction expansions (CFE) is a method of evaluation of functions, that frequently converges much more rapidly than power series expansions, and converges in a much larger domain in the complex plane. A method for obtaining discrete equivalents of the fractional-order operators, which combines the well known advantages of the trapezoidal rule (commonly designated as the Tustin method in the control theory) and the advantages of the CFE uses as generating function [10]:

$$\left(w\left(z^{-1}\right)\right)^{\pm\alpha} = \left(\frac{2}{T}\frac{1-z^{-1}}{1+z^{-1}}\right)^{\pm\alpha} \tag{12.11}$$

The application of the CFE of (12.11) results in the discrete transfer function, approximating fractional-order operators, expressed as:

$$\begin{aligned} D^{\pm\alpha}(z) &= \frac{Y(z)}{X(z)} = \left(\frac{2}{T}\right)^{\pm\alpha} \mathrm{CFE}\left\{\left(\frac{1-z^{-1}}{1+z^{-1}}\right)^{\pm\alpha}\right\}_{m,n} \\ &= \left(\frac{2}{T}\right)^{\pm\alpha} \frac{P_m\left(z^{-1}\right)}{Q_n\left(z^{-1}\right)} = \left(\frac{2}{T}\right)^{\pm\alpha} \frac{p_0 + p_1 z^{-1} + \cdots + p_m z^{-m}}{q_0 + q_1 z^{-1} + \cdots + q_n z^{-n}} \end{aligned} \tag{12.12}$$

where T is the sampling period, CFE$\{u\}$ denotes the function from applying the continued fraction expansion to the function u, $Y(z)$ is the Z transform of the output sequence $y(nT)$, $X(z)$ is the Z transform of the input sequence $x(nT)$, m and n are the orders of the approximation, and P and Q are polynomials of degrees m and n, correspondingly, in the variable z^{-1}.

References

1. Barbosa, R.S., Machado, J.A.T., Ferreira, I.M.: Tuning of PID controllers based on Bode's ideal transfer function. Nonlinear Dyn. **38**, 305–321 (2004)
2. Hilfer, R.: Applications of Fractional Calculus in Physics. World Scientific, Singapore (2000)
3. Machado, J.A.T.: Analysis and design of fractional-order digital control systems. Systems Analysis-Moddelling-Simulation. Gordon and Breach Science Publishers **27**(2-3), 107–122 (1997)
4. Oldham, K.B., Spanier, J.: The Fractional Calculus. Academic, New York (1974)
5. Oustaloup, A.: La Commande CRONE: Commande Robuste d'Ordre Non Entier. Editions Hermès, Paris (1991)
6. Oustaloup, A.: La Dérivation Non Entière: Théorie, Synthèse et Applications. Editions Hermès, Paris (1995)
7. Oustaloup, A., Levron, F., Mathieu, B., Nanot, F.M.: Frequency-band complex noninteger differentiator: Characterization and synthesis. IEEE Trans. Circuits Syst. I Fund. Theory Appl. **47**(1), 25–39 (2000)
8. Podlubny, I.: Fractional Differential Equations. Academic, San Diego (1999)
9. Podlubny, I.: Fractional-order systems and $\mathrm{PI}^{\lambda}\mathrm{D}^{\mu}$-controllers. IEEE Trans. Automat. Contr. **44**(1), 208–214 (1999)
10. Vinagre, B.M., Podlubny, I., Hernández, A., Feliu, V.: Some approximations of fractional order operators used in control theory and applications. FCAA Fractional Calculus Appl. Anal. **3**(3), 231–248 (2000)

Chapter 13
A Dynamical Point of View of Quantum Information: Discrete Wigner Measures

A.T. Baraviera, C.F. Lardizabal, A.O. Lopes, and M. Terra Cunha

Abstract We describe some well known properties of Wigner measures and then analyze some connections with Quantum Iterated Function Systems.

13.1 Discrete Weyl Relations

This section follows parts of [3]. Consider the Hilbert space $\mathscr{H} = \mathbb{C}^N$. Let $\{|k\rangle\}_{k=0}^{N-1}$ be an orthonormal basis. Fix $\alpha_u, \alpha_v \in [0,1]$ and define the following matrices $U_N, V_N \in M_N(\mathbb{C})$:

$$U_N := e^{\frac{2\pi}{N} i\alpha_u} \sum_{k=0}^{N-1} e^{\frac{2\pi}{N} ik} |k\rangle\langle k|, V_N := e^{\frac{2\pi}{N} i\alpha_v} \sum_{k=0}^{N-1} |k\rangle\langle k-1| \tag{13.1}$$

together with the identification $|j\rangle = |j \bmod N\rangle$. Such operators are unitary and we have

$$U_N|l\rangle = e^{\frac{2\pi}{N} i(\alpha_u + l)}|l\rangle, V_N|l\rangle = e^{\frac{2\pi}{N} i\alpha_v}|l+1\rangle \tag{13.2}$$

Defining $n := (n_1, n_2) \in \mathbb{Z}^2$, we have that U_N and V_N satisfy the *discrete Weyl relations*

$$U_N^{n_1} V_N^{n_2} = e^{\frac{2\pi}{N} i n_1 n_2} V_N^{n_1} U_N^{n_2} \tag{13.3}$$

Also, inspired in the continuous case, we define the *discrete Weyl operators*:

$$W_N(n) := e^{-i\frac{\pi}{N} n_1 n_2} U_N^{n_1} V_N^{n_2} \tag{13.4}$$

A.T. Baraviera, C.F. Lardizabal (✉), and A.O. Lopes
I.M. – UFRGS, Porto Alegre 91500-000, Brazil
e-mail: atbaraviera@gmail.com, carlos.lardizabal@gmail.com, arturoscar.lopes@gmail.com

M. Terra Cunha
D. M – UFMG, Belo Horizonte 30161-970, Brazil
e-mail: marcelo.terra.cunha@gmail.com

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_13,

Such operators satisfy

$$W_N^*(n) = W(-n) \tag{13.5}$$

and

$$W_N(n)W_N(m) = e^{i\frac{\pi}{N}\sigma(n,m)}W_N(n+m) \tag{13.6}$$

where $\sigma(n,m) := n_1 m_2 - n_2 m_1$.

When normalized, the discrete Weyl operators form an orthonormal basis for $M_N(\mathbb{C})$. In fact, using (13.2) and (13.4), we have

$$\begin{aligned} tr(W_N(n)) &= \sum_{l=0}^{N-1} e^{-i\frac{\pi}{N}n_1 n_2}\langle l|U_N^{n_1}V_N^{n_2}|l\rangle \\ &= \sum_{l=0}^{N-1} e^{-i\frac{\pi}{N}(n_1 n_2 + 2n_1(\alpha_u + l) - 2n_2\alpha_v)}\langle l|l+n_2\rangle \\ &= \delta_{n_2,0}\sum_{l=0}^{N-1} e^{-\frac{2\pi i n_1}{N}(\alpha_u + l)} = N\delta_{n,0} \end{aligned} \tag{13.7}$$

This allows us to obtain

$$tr(W_N^*(n)W_N(m)) = N\delta_{n,m} \tag{13.8}$$

and therefore for all $A \in M_N(\mathbb{C})$,

$$A = \frac{1}{N}\sum_{n\in\mathbb{Z}_N^2} tr\Big(W_N^*(n)A\Big)W_N(n) \tag{13.9}$$

where $\mathbb{Z}_N^2 := \{n = (n_1, n_2) : 0 \le n_i \le N-1\}$.

13.2 Introduction to the Wigner Function

This section follows parts of [10]. Given a quantum system, we are interested in obtaining another form of representing the wave function $\Psi(x)$. Such object will be the Wigner function, which will depend on two variables, moment and position. In order to understand such functions, we need to study the structure of phase spaces.

The Wigner function consists of a special way of describing density operators. In principle, we could say that density operators are a more fundamental structure than its Wigner representation. For instance, the Wigner representation is unable to describe the density operators associated to two-level systems. However, due to its simplicity, we will see that an understanding of the Wigner distribution gives us insight on certain aspects of density operators.

Definition 13.1. Given a wave function $\Psi(x)$, the *Wigner distribution function* is

$$W(q,p) = W_\Psi(q,p) := \frac{1}{2\pi\hbar}\int_{-\infty}^{\infty} e^{isp/\hbar}\langle q - \frac{s}{2}|\Psi\rangle\langle\Psi|q + \frac{s}{2}\rangle ds \tag{13.10}$$

where above we are using Dirac notation

$$\langle q - \frac{s}{2}|\Psi\rangle = \Psi(q - \frac{s}{2}) \tag{13.11}$$

$$\langle\Psi|q + \frac{s}{2}\rangle = \Psi^*(q + \frac{s}{2}) \tag{13.12}$$

Define the change of coordinates

$$x = q + \frac{s}{2}, \;\; x' = q - \frac{s}{2} \tag{13.13}$$

and then we obtain

$$W(q,p) = \frac{1}{2\pi\hbar}\int_{-\infty}^{\infty} e^{\frac{i}{\hbar}p(x-x')}\langle x'|\Psi\rangle\langle\Psi|x\rangle ds \tag{13.14}$$

That is, the Wigner distribution is obtained by calculating the product $\Psi(x')\Psi^*(x)$ and then applying the Fourier transform on $s = x - x'$. Such distribution has the following properties:

$$\int_{-\infty}^{\infty} W(q,p)dp = \langle q|\Psi\rangle\langle\Psi|q\rangle = |\Psi(q)|^2 \tag{13.15}$$

$$\int_{-\infty}^{\infty} W(q,p)dq = \langle p|\Psi\rangle\langle\Psi|p\rangle = |\tilde{\Psi}(p)|^2 \tag{13.16}$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} W(q,p)dpdq = 1 \tag{13.17}$$

where $\tilde{\Psi}$ is the moment representation of the wave function Ψ.

The Wigner function is real, but can assume negative or positive values. In this sense, it is not a density, but it is a kind of joint distribution of the position and momentum distributions.

Now, note that (13.14) can be written as

$$W(q,p) = \frac{1}{2\pi\hbar}\int_{-\infty}^{\infty} e^{\frac{i}{\hbar}p(x-x')}\langle x'\big(|\Psi\rangle\langle\Psi|\big)x\rangle ds = \frac{1}{2\pi\hbar}\int_{-\infty}^{\infty} e^{\frac{i}{\hbar}p(x-x')}\langle x'|\rho|x\rangle ds \tag{13.18}$$

where

$$x = q + \frac{s}{2}, \;\; x' = q - \frac{s}{2} \tag{13.19}$$

where we define the density operator associated to a pure state as

$$\rho := |\Psi\rangle\langle\Psi| \tag{13.20}$$

The general definition for ρ includes pure and mixed states:

$$\rho = \sum_i p_i |\Psi_i\rangle\langle\Psi_i| \tag{13.21}$$

where $p_i \geq 0$ and $\sum_i p_i = 1$. Such equation describes ρ as an incoherent superposition of pure state density operators $|\Psi_i\rangle\langle\Psi_i|$, where Ψ_i is a wave function, but not necessarily an energy eigenstate. On (13.21) the p_i denote the probabilities of finding the system on the state $|\Psi_i\rangle$.

Hence, besides the usual probabilistic interpretation for finding a particle described by a certain wave function at some position, we also have a probability distribution that such a particle can be found in different states.

13.3 Discrete Wigner Function

This section follows parts of [7] and [11]. In dimension 1, the *continuous Wigner function* is in 1–1 correspondence with a density matrix ρ and is defined by

$$W_\rho(q,p) = W(q,p) := \frac{1}{2\pi\hbar}\int_{-\infty}^{\infty} e^{i\lambda p/\hbar}\langle q - \frac{\lambda}{2}|\rho|q + \frac{\lambda}{2}\rangle d\lambda \tag{13.22}$$

Such function is uniquely defined by the following properties: [7, 11]:

1. $W(q,p) \in \mathbb{R}$
2. If ρ_1 and ρ_2 are two density states then

$$tr(\rho_1\rho_2) = 2\pi\hbar \int W_1(q,p)W_2(q,p)dqdp \tag{13.23}$$

3. (Projection property) The integral along a line on phase space, described by $a_1q + a_2p = a_3$, is the probability density that the measurement of the observable $a_1\hat{Q} + a_2\hat{P}$ gives a_3 as a result.

Remark Note that the Wigner function is always associated to a density matrix. It would be more appropriate to use the notation W_ρ instead of W. When there is no possibility of confusion we will denote W. The projection property stated above means, in other words, that the projection of the Wigner function along any direction of the phase space is equal to the probability distribution of a certain observable $a_1q + a_2p$, associated to that direction. Two special cases of this property are well-known:

$$\int W(q,p)dq \tag{13.24}$$

is the probability distribution for the moment, and

$$\int W(q,p)dp \tag{13.25}$$

is the probability distribution for position. For more details on these properties, see [11].

We can write W as the expected value of a Fano operator, so we have

$$W(q,p) = tr(\rho\hat{A}(q,p)) \tag{13.26}$$

where $\hat{A}$ can be written as

$$\hat{A}(q,p) = \frac{1}{(2\pi\hbar)^2}\int exp\Big[-\frac{\lambda}{\hbar}(\hat{P}-p)+i\frac{\lambda'}{\hbar}(\hat{Q}-q)\Big]d\lambda d\lambda' \tag{13.27}$$

$$= \frac{1}{(2\pi\hbar)^2}\int \hat{D}(\lambda,\lambda')exp\Big[-\frac{i}{\hbar}(\lambda' q-\lambda p)\Big]d\lambda d\lambda' \tag{13.28}$$

where

$$\hat{D}(\lambda,\lambda') := exp\Big[-\frac{i}{\hbar}(\lambda\hat{P}-\lambda'\hat{Q})\Big] \tag{13.29}$$

Also we can write $\hat{A}$ as

$$\hat{A}(q,p) = \frac{1}{\pi\hbar}\hat{D}\hat{R}\hat{D}^* \tag{13.30}$$

where above we write $\hat{D} = \hat{D}(q,p)$ and $\hat{R}$ is an operator acting on positive eigenstates such that $\hat{R}|x\rangle = |-x\rangle$.

The proof that W satisfies properties 1–3 stated above follows from simple phase space properties. The fact that $W(q,p) \in \mathbb{R}$ is a consequence of the fact that $\hat{A}(q,p)$ is hermitian. As for property 2, we can show that

$$tr\Big(\hat{A}(q,p)\hat{A}(q',p')\Big) = \frac{1}{2\pi\hbar}\delta(q-q')\delta(p-p') \tag{13.31}$$

As a consequence, it is possible to invert (13.26) so we can write

$$\rho = 2\pi\hbar\int W(q,p)\hat{A}(q,p)dqdp \tag{13.32}$$

Property 2 follows from the formula above. As for property 3, note that by integrating $\hat{A}(q,p)$ along a line on phase space gives us a projection operator. Therefore

$$\int \delta(a_1q+a_2p-a_3)\hat{A}(q,p)dqdp = |a_3\rangle\langle a_3| \tag{13.33}$$

where $|a_3\rangle$ is an eigenstate of the operator $a_1\hat{Q} + a_2\hat{P}$ with eigenvalue a_3. Later we will describe the proof of this property for the discrete case.

Now we are interested in defining the Wigner function in the discrete case. The first step is to define a discrete phase space. Consider a Hilbert space of dimension N and define a basis

$$B_x = \{|n\rangle, n = 0, \dots, N-1\},$$

which will be seen as a *discrete position basis*. Now we define a basis of moments

$$B_p = \{|k\rangle, k = 0, \dots, N-1\}$$

A natural way of introducing the moment base from the position base is via the *discrete Fourier transform*. Then we can obtain the states of B_p from the states in B_x in the following way:

$$|k\rangle = \frac{1}{\sqrt{N}} \sum_n \exp[2\pi ink/N]|n\rangle \tag{13.34}$$

Therefore, as in the continuous case, position and moment are related by the Fourier transform.

Remark We can relate the dimension of the Hilbert space with the Planck constant in the following way. We are supposing that the phase space has a finite area, which we can suppose equal to 1. In this area we can have N orthogonal states. If each state fills an area equal to $2\pi\hbar$, we have $N = 1/2\pi\hbar$. So N plays the role of the inverse of the Planck constant and the limit as N goes to infinity can be seen as the semiclassical limit [7].

Given position and moment bases, we can define their respective displacement operators. For discrete systems, we can define translation operators $\hat{U}$ and $\hat{V}$, in a way which is similar to what we have in (13.1) and (13.2), Sect. 13.1:

$$\hat{U}^m|n\rangle := |n+m\rangle, \;\; \hat{U}^m|k\rangle := \exp[-2\pi imk/N]|k\rangle \tag{13.35}$$

where the vector sums are mod N. In a similar way the operator $\hat{V}$ is a shift on moment basis, and it is diagonal on positions:

$$\hat{V}^m|k\rangle := |k+m\rangle, \;\; \hat{V}^m|n\rangle := \exp[2\pi imn/N]|n\rangle \tag{13.36}$$

Then it is possible to show that

$$\hat{V}^p\hat{U}^q = e^{2\frac{\pi}{N}ipq}\hat{U}^q\hat{V}^p, \tag{13.37}$$

the discrete Weyl relations (13.3), seen on Sect. 13.1. Let us also define a reflection operator as $\hat{R}|n\rangle := |-n\rangle$. We have that

$$\hat{U}\hat{R} = \hat{R}\hat{U}^{-1}, \;\; \hat{V}\hat{R} = \hat{R}\hat{V}^{-1} \tag{13.38}$$

The reflection operator is related to the Fourier transform in the following way. Denote by U_{FT} the discrete Fourier transform, that is the operator whose entries on basis B_x are

$$\langle n'|U_{FT}|n\rangle = \exp[2\pi i n n'/N] \tag{13.39}$$

Then we have

$$\hat{R} = U_{FT}^2 \tag{13.40}$$

In order to define the discrete Wigner function, we still have to define a translation operator $\hat{T}$ and a point operator $\hat{A}$, corresponding to the Fano operator defined in the continuous case. This is what we will do next. Define

$$\hat{T}(q,p) := \hat{U}^q \hat{V}^p \exp[i\pi q p/N] \tag{13.41}$$

Such operators satisfy

$$\hat{T}(\lambda q, \lambda p) = \hat{T}^\lambda(q,p) \tag{13.42}$$

Remark In $\mathbb{R}^2$ we define the translation operator with position q and moment p as

$$\hat{T}(q,p) = e^{-\frac{i}{\hbar}(q\hat{P} - p\hat{Q})} \tag{13.43}$$

Instead of definitions (13.35) and (13.36) given for $\hat{U}$ and $\hat{V}$ we could, in principle, define $\hat{U}$ and $\hat{V}$ as the exponential of two operators $\hat{Q}$ and $\hat{P}$, defined as being diagonal in B_x and B_p. However, infinitesimal operators $\hat{Q}$ and $\hat{P}$ satisfying the canonical commutation relations (CCR) cannot be defined over a discrete Hilbert space [4, 11]. Because of that we will use the finite cyclic shifts, given by (13.35) and (13.36).

Remark Due to technicalities, the phase-space can be taken to be a $N \times N$ or a $2N \times 2N$ grid [7]. Typically we will be interested in phase spaces with even dimension and we will use the $2N \times 2N$ grid (for instance, if $N = 2$ the phase space has 16 points). Our following definitions will follow this choice as well.

Let $\alpha = (q,p)$ be a point of the discrete phase space, with q and p assuming values between 0 and $2N-1$. Define

$$\hat{A}(\alpha) := \frac{1}{(2N)^2} \sum_{\lambda,\lambda'=0}^{2N-1} \hat{T}(\lambda,\lambda') \exp\left[-2\pi i \frac{(\lambda' q - \lambda p)}{2N}\right] = \frac{1}{2N} \hat{U}^q \hat{R} \hat{V}^{-p} e^{i\pi p q/N} \tag{13.44}$$

We can express the translation operator in terms of $\hat{A}(\alpha)$ by inverting the above definition and then we obtain the Fourier transform of $\hat{A}$:

$$\tilde{T}(n,k) = \sum_{q,p=0}^{2N-1} \hat{A}(q,p) \exp[-i \frac{2\pi}{2N}(np - kq)] \tag{13.45}$$

Note that as we defined the point operators over a lattice of $2N \times 2N$ points, we get a total of $4N^2$ operators. However, such set is not independent. That is, we can show that

$$\hat{A}(q+\sigma_q N, p+\sigma_p N) = \hat{A}(q,p)(-1)^{\sigma_p q+\sigma_q p+\sigma_q\sigma_p N} \tag{13.46}$$

for $\sigma_q, \sigma_p = 0, 1$. So we have that N^2 operators define the remaining ones. Define

$$G_N := \{\alpha = (q,p) : 0 \leq q, p \leq N-1\}$$

And the set G_{2N} will denote the entire lattice of order $2N$.

A relation between $\hat{A}$ and $\hat{T}$ is the following:

$$\hat{A}(\alpha)\hat{A}(\alpha') = \hat{T}(\alpha-\alpha')\frac{\exp[i(\pi/N)(q_\alpha p_{\alpha'} - q_{\alpha'} p_\alpha)]}{4N^2} \tag{13.47}$$

By taking the trace of the above equation we get

$$tr(\hat{A}(\alpha)\hat{A}(\alpha')) = \frac{1}{4N}\delta_N(q'-q)\delta_N(p'-p) \tag{13.48}$$

where α and α' are in G_N and

$$\delta_N(q) := \frac{1}{N}\sum_{n=0}^{N-1} e^{-2\pi iqn/N} \tag{13.49}$$

is the periodic Dirac delta function, which is equal to zero unless $q \equiv 0 \bmod N$.

Definition 13.2. The *discrete Wigner function* is

$$W(\alpha) = W_\rho(\alpha) := tr(\hat{A}(\alpha)\rho) \tag{13.50}$$

where $\alpha \in G_{2N}$.

These $4N^2$ values are not independent because in a similar way to what we have for the operator $\hat{A}$, we have

$$\hat{W}(q+\sigma_q N, p+\sigma_p N) = \hat{W}(q,p)(-1)^{\sigma_p q+\sigma_q p+\sigma_q\sigma_p N} \tag{13.51}$$

for $\sigma_q, \sigma_p = 0, 1$. As the operators $\hat{A}(\alpha)$ form a complete set, we can write the density operator as a linear combination of the $\hat{A}(\alpha)$. So we can show that

$$\rho = 4N\sum_{\alpha\in G_N} W(\alpha)\hat{A}(\alpha) = N\sum_{\tilde{\alpha}\in G_{2N}} W(\tilde{\alpha})\hat{A}(\tilde{\alpha}) \tag{13.52}$$

Remark It is possible to show that the discrete Wigner function defined above satisfies properties 1 to 3 stated in the beginning of this section. Property 1 is a consequence of the fact that $\hat{A}(q, p)$ are hermitian operators. Property 2 follows from the completeness of the set $\hat{A}(\alpha)$, which allows us to show that

$$tr(\rho_1\rho_2) = N \sum_{\alpha \in G_{2N}} W_1(\alpha) W_2(\alpha) \tag{13.53}$$

The proof of the third property requires a brief analysis of the lattice G_N and we refer the reader to [7] for details.

Conclusions We have defined the Wigner function for systems over a Hilbert space of dimension $N < \infty$. The Wigner functions is defined as the expected value of the operator $\hat{A}(\alpha)$ defined over the phase space given by (13.44). The definition is such that $W(\alpha) \in \mathbb{R}$ and is such that we can calculate the inner product between states and gives the correct marginal distributions along any line over the phase space, which is the lattice G_{2N} with $4N^2$ points. Also, the values of $W(\alpha)$ on the sublattice G_N are enough to determine W in the entire space.

13.4 Calculating Wigner Functions

In order to calculate the Wigner function of a quantum state, we will use (13.35), (13.36) and (13.44) se we can write W in the following convenient form:

Lemma 13.1.

$$W(q, p) = \frac{1}{2N} \sum_{n=0}^{N-1} \langle q - n|\rho|n\rangle \exp\left[\frac{2\pi i}{N} p(n - q/2)\right] \tag{13.54}$$

Proof: In the following calculations, recall that the inner product is linear on the second variable. We have that

$$\begin{aligned}
W(q, p) = tr(A\rho) &= \frac{1}{2N} \exp[i\pi pq/N] tr(U^q R V^{-p} \rho) \\
&= \frac{1}{2N} \exp[i\pi pq/N] \sum_{i=0}^{N-1} \langle n|U^q R V^{-p} \rho|n\rangle \\
&= \frac{1}{2N} \exp[i\pi pq/N] \sum_{i=0}^{N-1} \langle U^{-q} n|R V^{-p} \rho|n\rangle \\
&= \frac{1}{2N} \exp[i\pi pq/N] \sum_{i=0}^{N-1} \langle n - q|R V^{-p} \rho|n\rangle \\
&= \frac{1}{2N} \exp[i\pi pq/N] \sum_{i=0}^{N-1} \langle q - n|V^{-p} \rho|n\rangle
\end{aligned}$$

$$= \frac{1}{2N} \exp[i\pi pq/N] \sum_{i=0}^{N-1} \langle V^p(q-n)|\rho|n\rangle$$
$$= \frac{1}{2N} \exp[i\pi pq/N] \sum_{i=0}^{N-1} \exp[-2\pi i p(q-n)/N]\langle q-n|\rho|n\rangle$$

Also, note that

$$i\pi pq/N - 2\pi i p(q-n)/N = \frac{ip\pi}{N}(q-2(q-n)) = \frac{ip\pi}{N}(2n-q) = \frac{2\pi i p}{N}(n-q/2)$$

Hence,

$$W(q,p) = \frac{1}{2N} \sum_{n=0}^{N-1} \langle q-n|\rho|n\rangle \exp[\frac{2\pi i p}{N}(n-q/2)] \qquad \square$$

We believe there is a misprint in [7] in the expression corresponding to the $W(q,p)$ described by the claim of the above lemma.

One can ask how W_ρ changes with ρ. Suppose first ρ is a projector from a wave ψ which has norm 1 in $\mathscr{L}^2$. Suppose $(a\phi_1 + b\phi_2) = \psi$, where ψ, ϕ_1, ϕ_2 have norm 1, and $\rho = |\psi >< \psi|$. Then, $W_\psi \neq aW_{\phi_1} + bW_{\phi_1}$. The linearity occurs only when $\rho = \sum_i c_i |i >< i|$, that is, when ρ is diagonal. This in general do not happen for operators $|\psi >< \psi|$ induced by a wave ψ. However, if $\rho = (a\rho_1 + b\rho_2)$, where ρ, ρ_1, ρ_2 are density matrices, then $W_\rho = aW_{\rho_1} + bW_{\rho_2}$.

Example 13.1. Let $N = 2$, and let $|\psi\rangle = a|0\rangle + b|1\rangle$ be a state superposition. Let $W_1(\alpha)$ and $W_2(\alpha)$ be the Wigner functions for $|0\rangle$ and $|1\rangle$, respectively. We have that the Wigner function W for $|\psi\rangle$ is such that

$$W(\alpha) = |a|^2 W_1(\alpha) + |b|^2 W_2(\alpha) + 2Re\{ab^*\langle 1|A(\alpha)|0\rangle\} \tag{13.55}$$

In fact, note that

$$W(\alpha) = tr(A(\alpha)\rho) = tr\Big(A(\alpha)(|a|^2|0\rangle\langle 0| + |b|^2|1\rangle\langle 1| + ab^*|0\rangle\langle 1| + a^*b|1\rangle\langle 0|)\Big)$$
$$= |a|^2 W_1(\alpha) + |b|^2 W_2(\alpha) + ab^* tr(A(\alpha)|0\rangle\langle 1|) + a^* b tr(A(\alpha)|1\rangle\langle 0|)$$
$$= |a|^2 W_1(\alpha) + |b|^2 W_2(\alpha) + ab^* tr(\langle 1|A(\alpha)|0\rangle) + a^* b tr(\langle 0|A(\alpha)|1\rangle)$$

so the result follows.

Let us remark a few properties of the Wigner function for a pure state ρ. In this case by expanding ρ in terms of the phase space operators as in (13.52) and by imposing the condition $\rho^2 = \rho$, we get

$$W(\alpha) = 4N^2 \sum_{\beta,\gamma \in G_N} \Gamma(\alpha,\beta,\gamma) W(\beta) W(\gamma) \tag{13.56}$$

where the function $\Gamma(\alpha, \beta, \gamma)$, which depends on 3 points (i.e., a triangle) is given by

$$\Gamma(\alpha, \beta, \gamma) := tr(\hat{A}(\alpha)\hat{A}(\beta)\hat{A}(\gamma)) = \frac{1}{4N^3} \exp\left[\frac{2\pi i}{N} S(\alpha, \beta, \gamma)\right], \qquad (13.57)$$

of 2 of the 3 point (α, β, γ) contain even q and p coordinates. Otherwise we define

$$\Gamma(\alpha, \beta, \gamma) := 0, \qquad (13.58)$$

and in the above expression, valid for even N, the value $S(\alpha, \beta, \gamma)$ is the area of the triangle formed by these points (measured in units of the elementary triangle formed by 3 points which are one position apart from each other).

Now we calculate the Wigner function for a position eigenvalue

$$\rho_{q_0} = |q_0\rangle\langle q_0| \qquad (13.59)$$

We obtain the following closed expression for W:

$$\begin{aligned} W_{q_0}(q, p) &= \frac{1}{2N}\langle q_0|\hat{U}^q \hat{R}\hat{V}^{-p}|q_0\rangle e^{i\pi pq/N} \\ &= \frac{1}{2N}\delta_N(q - 2q_0)(-1)^{p[(q-2q_0) \bmod N]} \end{aligned} \qquad (13.60)$$

We can also write the Wigner function of a state which is a linear superposition:

$$|\psi\rangle = \frac{1}{\sqrt{2}}(|q_0\rangle + e^{-i\phi}|q_1\rangle) \qquad (13.61)$$

Again, we can obtain a closed expression for W, which is

$$W(q, p) = \frac{1}{2}\Big(W_{q_0}(q, p) + W_{q_1}(q, p) + \Delta W_{q_0,q_1}(q, p)\Big) \qquad (13.62)$$

where the interference term is

$$\Delta W_{q_0,q_1}(q, p) := \frac{1}{N}\delta_N(\tilde{q})(-1)^{\tilde{q}p} \cos\Big(\frac{2\pi}{\lambda}p + \phi\Big) \qquad (13.63)$$

where

$$\tilde{q} = q_0 + q_1 - q, \lambda = \frac{2N}{q_0 - q_1} \qquad (13.64)$$

This is an explicit expression for the calculation seen in Example 13.1.

Now we make a few considerations on the time evolution of quantum systems on phase space. If U is the unitary operator which determines the evolution of a state, then the associated density matrix evolves in the following way,

$$\rho(t+1) = U\rho(t)U^* \tag{13.65}$$

By this fact, we can show that the Wigner function evolves in the following way:

$$W(\alpha, t+1) = \sum_{\beta\in G_{2N}} Z_{\alpha\beta} W(\beta, t) \tag{13.66}$$

where the matrix $Z_{\alpha\beta}$ is defined as

$$Z_{\alpha\beta} := Ntr\Big(\hat{A}(\alpha)U\hat{A}(\beta)U^*\Big) \tag{13.67}$$

Therefore the time evolution in phase space is represented by a linear transformation, which is a consequence of Schrödinger's equation. The unitarity imposes a few restrictions on the matrix $Z_{\alpha\beta}$. In fact, since purity of states is preserved, the time evolution has to preserve the restriction given by (13.56). Therefore, the matrix has to leave the function $\Gamma(\alpha, \beta, \gamma)$ invariant, that is,

$$\Gamma(\alpha', \beta', \gamma') = \sum_{\alpha,\beta,\gamma} Z_{\alpha'\alpha} Z_{\beta'\beta} Z_{\gamma'\gamma} \Gamma(\alpha, \beta, \gamma) \tag{13.68}$$

The real matrix $Z_{\alpha\beta}$ contains all the information on the time evolution of the system. In general, such matrix relates a point α with several other points β. So the evolution will be, in general, nonlocal, a unique property of quantum mechanics. In classical systems, the value of the classical distribution function $W(\alpha, t+1)$ is equal to the value $W(\beta, t)$ for some point β, which consists of a well defined function of α and t. However, we have in [7] a few examples of unitary operators which generate a local dynamical evolution on the phase space.

To conclude this section, we calculate the Wigner function for a quantum channel Λ, as the ones considered for our analysis of QIFS. This is a straightforward calculation. Let V_i be linear operators, $i = 1, \ldots, k$ such that $\sum_i V_i^* V_i = I$. Then $\Lambda(\rho) = \sum_i V_i \rho V_i^* \in \mathscr{M}_N$. Hence,

$$\begin{aligned} W_{\Lambda(\rho)}(q,p) &= \frac{1}{2N}\sum_{n=0}^{N-1} \langle q-n|\Lambda(\rho)|n\rangle \exp\Big[\frac{2\pi i}{N} p(n-q/2)\Big] \\ &= \frac{1}{2N}\sum_{n=0}^{N-1}\sum_{i=1}^{k} \langle q-n|V_i\rho V_i^*|n\rangle \exp\Big[\frac{2\pi i}{N} p(n-q/2)\Big] \\ &= \frac{1}{2N}\sum_{n=0}^{N-1}\sum_{i=1}^{k} \langle (q-n)V_i|\rho|V_i^*(n)\rangle \exp\Big[\frac{2\pi i}{N} p(n-q/2)\Big] \end{aligned} \tag{13.69}$$

Writing $\rho = \sum_{j=0}^{N-1} \rho_j |j\rangle\langle j|$, $\sum_j \rho_j = 1$, we get

$$W_{\Lambda(\rho)}(q, p) = \frac{1}{2N} \sum_{n,j=0}^{N-1} \sum_{i=1}^{k} \rho_j \langle (q-n) V_i | j \rangle \langle j | V_i^{*}(n) \rangle \exp\left[\frac{2\pi i}{N} p(n - q/2)\right] \tag{13.70}$$

Therefore the Wigner function of $\Lambda(\rho)$ is obtained in a simple way from the function for ρ.

13.5 Some Properties of the Discrete Wigner Function

We have seen in Sect. 13.3 that the discrete Wigner function

$$W(\alpha) = tr(\hat{A}(\alpha)\rho) \tag{13.71}$$

satisfies properties 1 and 2. Now let us prove property 3. Let $\rho = \sum_i p_i |i\rangle\langle i|$, $\sum_i p_i = 1$ be a density operator. Denote by

$$B_x = \{|n\rangle, n = 0, \ldots, N-1\},$$

a position basis and

$$B_p = \{|k\rangle, k = 0, \ldots, N-1\}$$

a moment basis, as before, where

$$|k\rangle = \frac{1}{\sqrt{N}} \sum_n \exp[2\pi i n k/N] |n\rangle \tag{13.72}$$

To prove property 3, we must show that as we sum the operators $\hat{A}(q, p)$ over the point of the phase space which lie over a line L, we obtain a projection operator. This implies that by summing the values of the Wigner function over all the points of a line we get a positive number, which can be interpreted as a probability.

We begin by defining a line on the phase space. A line L is a set of point of the lattice, defined as

$$L = L(n_1, n_2, n_3) = \{(q, p) \in G_{2N} : n_1 p - n_2 q = n_3, 0 \leq n_i \leq 2N - 1\} \tag{13.73}$$

Also, we say that two lines as parallel if they are parameterized by the same integers n_1 and n_2.

Now, let us show that as we sum the point operators A over a line, we get projection operators. So we are interested in the operator

$$\hat{A}_L = \sum_{(q,p)\in L} \hat{A}(q, p) \tag{13.74}$$

Since $\delta_N(q) = \frac{1}{N}\sum_{n=0}^{N-1} e^{-2\pi iqn/N}$, we can rewrite such operator as

$$
\begin{aligned}
A_L &= \sum_{q,p=0}^{2N-1} \hat{A}(q,p)\delta_{2N}(n_1 p - n_2 q - n_3) \\
&= \frac{1}{2N}\sum_{\lambda=0}^{2N-1}\sum_{q,p=0}^{2N-1} \hat{A}(q,p)\exp[-i\frac{2\pi}{2N}\lambda(n_1 p - n_2 q - n_3)] \\
&= \frac{1}{2N}\sum_{\lambda=0}^{2N-1} \hat{T}^{\lambda}(n_1,n_2)\exp[i\frac{2\pi}{2N}n_3\lambda] \qquad (13.75)
\end{aligned}
$$

where we use the Fourier transform of $\hat{A}$ to obtain the last equality. Since $\hat{T}$ is unitary, we have N eigenvectors $|\phi_j\rangle$ with eigenvalues $\exp[-2\pi i\phi_j/N]$. Besides, such operator is cyclic and satisfies $\hat{T}^N = I$. Therefore as its eigenvalues are $N-th$ roots of unity, the ϕ_j are integers. So we can rewrite (13.75) as

$$
\begin{aligned}
\hat{A}_L &= \frac{1}{2N}\sum_{\lambda=0}^{2N-1}\sum_{j=0}^{N}\exp[-i\frac{2\pi}{2N}(2\phi_j - n_3)\lambda]|\phi_j\rangle\langle\phi_j| \\
&= \sum_{j=0}^{N}\delta_{2N}(2\phi_j - n_3)|\phi_j\rangle\langle\phi_j| \qquad (13.76)
\end{aligned}
$$

Hence we have that $\hat{A}_L$ is a projection operator over a subspace generated by a subset of eigenvectors of the translation operator $\hat{T}(n_1,n_2)$.

Example 13.2. For a line L_q defined by $q = n_3$ (that is, $n_1 = 1, n_2 = 0$), the Wigner function summed over all point of L_q is

$$\sum_{(q,p)\in L_q} W_\rho(q,p) = \sum_p W_\rho(n_3,p) = \langle n_3/2|\rho|n_3/2\rangle \qquad (13.77)$$

if n_3 is even, and equal to zero otherwise.

More precisely, we have the following proposition:

Proposition 13.1. *Let N be even and let ρ be a density operator. Then*

$$\sum_{p=0}^{2N-1} W_\rho(2q,p) = \langle q|\rho|q\rangle, q = 0,2,\ldots,N-1$$

and

$$\sum_{p=0}^{2N-1} W_\rho(2q+1,p) = 0, q = 0,2,\ldots,N-1$$

Proof: First, to see why the case q odd implies that the Wigner function equals zero, consider the expression for W given by

$$W_\rho(q,p) = \frac{1}{2N}\sum_{n=0}^{N-1}\langle q-n|\rho|n\rangle \exp\left[\frac{2\pi i}{N}p(n-q/2)\right] \tag{13.78}$$

Write $\rho = \sum_j c_j|j\rangle\langle j|$, $c_j > 0$. Then

$$\langle q-n|\rho|n\rangle = \sum_j c_j\langle q-n|j\rangle\langle j|n\rangle \tag{13.79}$$

which is $\neq 0$ if and only if $j = q-n = n$ for some j. In particular, in order to have a nonzero inner product above, we must have that q is even, because $q-n = n$ implies $q = 2n$.

Now suppose that $q = 2q_0$. By the analysis above, we see that in the sum of the terms forming the Wigner function (13.78), we only have to sum the indices such that the equation

$$q-n = n \Leftrightarrow 2q_0 - n = n \tag{13.80}$$

is satisfied (recall that all calculations are made modulo N). Such equation has two solutions, namely $n = q_0$ and $n = q_0 + N/2$. To see that there are no other solutions for (13.80), we proceed in the following way. From $2q_0 - n = n$ we get $2(q_0 - n) = 0$. We know that $n = 0$ and $n = q_0 + N/2$ are solutions. Also, note that $x = 0$ and $x = N/2$ are solutions of $2x = 0$. Now, if y is a solution of $2x = 0$ then $y - N/2$ also is. Clearly if y is an element between 0 and $N/2$ then $2y$ will be at most equal to $2N-2$, hence $2y \neq 0$. Finally, let y be an element between $N/2$ and N and by contradiction suppose that $2y = 0$. Then by the remark above we have that $z = 2y - N/2$ is also a solution and z is between 0 and $N/2$. But there are no solutions for $2x = 0$ between 0 and $N/2$. This shows that $2x = 0$ admits only the solutions stated above.

Now note that if n equals q_0 then

$$\exp\left[\frac{2\pi i}{N}p(n-q/2)\right] = 1 \tag{13.81}$$

If $n = q_0 + N/2$, we have that the exponential above is equal to ±1, being positive or negative if p is even or odd, respectively. Therefore, for N even and $q = 2q_0$, we have

$$W_\rho(2q_0,p) = \frac{1}{2N}(\langle q_0|\rho|q_0\rangle \pm \langle q_0 + N/2|\rho|q_0 + N/2\rangle) \tag{13.82}$$

where the sign ± depends on p. For fixed q and considering all possible p (i.e., $p = 0,\dots,2N-1$), we have that the second inner product above will have a plus sign in front of it in the N possibilities in which p is even and will have a negative sign in the N remaining possibilities. So

$$\sum_p W_\rho(2q_0, p) = \langle q_0|\rho|q_0\rangle \tag{13.83}$$

This concludes the proof. □

Corollary 13.1. *If q is odd then $W_\rho(q, p) = 0$, for any p and any ρ density operator.*

Proof: Follows from the first paragraph of the proof above. □

Definition 13.3. Let ψ be a state. The *W-transform* of ψ is

$$\phi(p) := \sum_{q=0}^{2N-1} W_\psi(q, 2p) \tag{13.84}$$

for $p = 0, \ldots, 2N - 1$.

Let ϕ be the W-transform of ψ, and let $\mathscr{F}\psi$ be the discrete Fourier transform of ψ.

Question:

$$|(\mathscr{F}\psi)(p)|^2 \overset{?}{=} \phi(p), p = 0, 1, \ldots, N - 1 \tag{13.85}$$

Answer For $N = 2$ and $\psi = |0\rangle$ or $|1\rangle$, the answer is yes. In fact, let $|\psi\rangle = |0\rangle = (1, 0)$. Then

$$\mathscr{F}|0\rangle = \frac{1}{\sqrt{2}}\sum_j \exp[2\pi i j 0/2]|j\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$$
$$\Rightarrow (\mathscr{F}|0\rangle)(0) = \frac{1}{\sqrt{2}} \Rightarrow |(\mathscr{F}|0\rangle)(0)|^2 = \frac{1}{2}$$

And

$$\phi(0) = \sum_q W_{|0\rangle}(q, 0) = \frac{1}{4} + 0 + \frac{1}{4} + 0 = \frac{1}{2}$$

Also

$$(\mathscr{F}|0\rangle)(1) = \frac{1}{\sqrt{2}} \Rightarrow |(\mathscr{F}|0\rangle)(1)|^2 = \frac{1}{2}$$

And

$$\phi(1) = \sum_q W_{|0\rangle}(q,2) = \frac{1}{4} + 0 + \frac{1}{4} + 0 = \frac{1}{2}$$

Therefore in this case

$$|(\mathscr{F}\psi)(p)|^2 = \phi(p), p = 0, 1 \tag{13.86}$$

Now let $|\psi\rangle = |1\rangle = (0, 1)$. Then

$$\mathscr{F}|1\rangle = \frac{1}{\sqrt{2}}\sum_j \exp[2\pi ij/2]|j\rangle = \frac{1}{\sqrt{2}}(|0\rangle + \exp[2\pi i/2]|1\rangle) = \frac{1}{\sqrt{2}}(|0\rangle - |1\rangle)$$
$$\Rightarrow (\mathscr{F}|1\rangle)(0) = \frac{1}{\sqrt{2}} \Rightarrow |(\mathscr{F}|0\rangle)(0)|^2 = \frac{1}{2}$$

And

$$\phi(0) = \sum_q W_{|1\rangle}(q,0) = \frac{1}{4} + 0 + \frac{1}{4} + 0 = \frac{1}{2}$$

Also

$$(\mathscr{F}|1\rangle)(1) = -\frac{1}{\sqrt{2}} \Rightarrow |(\mathscr{F}|0\rangle)(0)|^2 = \frac{1}{2}$$

And

$$\phi(1) = \sum_q W_{|1\rangle}(q,2) = \frac{1}{4} + 0 + \frac{1}{4} + 0 = \frac{1}{2}$$

Therefore

$$|(\mathscr{F}\psi)(p)|^2 = \phi(p), p = 0, 1 \tag{13.87}$$

Now let us write an example in which the state considered is mixed. Let $\psi = 1/\sqrt{2}(|0\rangle + |1\rangle)$. Then

$$\mathscr{F}|\psi\rangle = \frac{1}{\sqrt{2}}(\mathscr{F}|0\rangle + \mathscr{F}|1\rangle) = \frac{1}{\sqrt{2}}\Big[\frac{1}{\sqrt{2}}(|0\rangle + |1\rangle) + \frac{1}{\sqrt{2}}(|0\rangle - |1\rangle)\Big] = |0\rangle \tag{13.88}$$

Then $|(\mathscr{F}\psi)(0)|^2 = 1$ e $|(\mathscr{F}\psi)(1)|^2 = 0$. Now let us calculate $\phi(p)$, $p = 0, 1$. By definition, we have $\phi(p) = \sum_q W_\psi(q, 2p)$. We can use the expression (13.62):

$$W_\psi(q,0) = \frac{1}{2}\Big(W_{|0\rangle}(q,0) + W_{|1\rangle}(q,0) + \Delta_{0,1}(q,0)\Big) \tag{13.89}$$

$$W_\psi(q,2) = \frac{1}{2}\Big(W_{|0\rangle}(q,2) + W_{|1\rangle}(q,2) + \Delta_{0,1}(q,2)\Big) \tag{13.90}$$

Then

$$\begin{aligned}
W_{\psi}(0,0) &= \frac{1}{2}\left(\frac{1}{4}+\frac{1}{4}+0\right) = \frac{1}{4}\\
W_{\psi}(1,0) &= \frac{1}{2}\left(0+0+\frac{1}{2}\right) = \frac{1}{4}\\
W_{\psi}(2,0) &= \frac{1}{2}\left(\frac{1}{4}+\frac{1}{4}+0\right) = \frac{1}{4}\\
W_{\psi}(3,0) &= \frac{1}{2}\left(0+0+\frac{1}{2}\right) = \frac{1}{4}
\end{aligned}$$

which implies $\phi(0) = 1 = |(\mathscr{F}\psi)(0)|^2$. Similarly,

$$\begin{aligned}
W_{\psi}(0,2) &= \frac{1}{2}\left(\frac{1}{4}+\frac{1}{4}+0+0\right) = \frac{1}{4}\\
W_{\psi}(1,2) &= \frac{1}{2}\left(0+0-\frac{1}{2}\right) = -\frac{1}{4}\\
W_{\psi}(2,2) &= \frac{1}{2}\left(\frac{1}{4}+\frac{1}{4}+0+0\right) = \frac{1}{4}\\
W_{\psi}(3,2) &= \frac{1}{2}\left(0+0-\frac{1}{2}\right) = -\frac{1}{4}
\end{aligned}$$

and so $\phi(1) = 0 = |(\mathscr{F}\psi)(1)|^2$.

Inspired in the calculation above, we prove the following lemma, valid for pure states only. After that, we will prove the result for density operators.

Lemma 13.2. *Let $\psi = |m\rangle \in \{|0\rangle, \ldots, |N-1\rangle\}$, N even. Then*

$$|(\mathscr{F}\psi)(p)|^2 = \phi(p),\ p = 0, 1, \ldots, N-1 \tag{13.91}$$

Proof: We have

$$\mathscr{F}|m\rangle = \frac{1}{\sqrt{N}}\sum_{j=0}^{N-1}\exp\left[2\pi i jm/N\right]|j\rangle$$

So

$$(\mathscr{F}|m\rangle)(p) = \frac{1}{\sqrt{N}}\exp\left[2\pi i pm/N\right] \Rightarrow |(\mathscr{F}|m\rangle)(p)|^2 = \frac{1}{N}$$

Let us calculate $\phi(p) = \sum_{q=0}^{2N-1} W_{|m\rangle}(q, 2p)$. By the Corollary 13.1, we only have to sum the even q. Then $\phi(p) = \sum_{q=0}^{N-1} W_{|m\rangle}(2q, 2p)$. By Proposition 13.1 we get, using expression (13.82), that

$$W_\rho(2q_0, p) = \frac{1}{2N}(\langle q_0|\rho|q_0\rangle + \langle q_0 + N/2|\rho|q_0 + N/2\rangle) \tag{13.92}$$

where the sign of the second inner product is positive because $2p$ is even. Now note that only one of the inner products above can be nonzero, because ρ is pure, by assumption. Moreover, ρ pure implies that such inner products are equal to 1. Finally, since q varies between 0 and $2N - 1$ we have exactly two nonzero terms in the sum of $\phi(p)$ namely, the terms corresponding to the m and $m + N/2$ indices. Hence,

$$\phi(p) = 1/2N + 1/2N = 1/N = |(\mathscr{F}|m\rangle)(p)|^2$$

This concludes the proof. □

The following result, inspired in the previous one, completes Proposition 13.1, which related the discrete Wigner function with the basis of position vectors. Now we do the corresponding work for the basis of momentum vectors.

Proposition 13.2. *Let N be even and let ρ be a density operator. Let $|p\rangle$ be a vector of the momentum basis, that is, obtained via the discrete Fourier transform of a position basis vector:*

$$|p\rangle = \frac{1}{\sqrt{N}} \sum_{j=0}^{N-1} \exp[2\pi i j p/N]|j\rangle \tag{13.93}$$

Then

$$\sum_{q=0}^{2N-1} W_\rho(q, 2p) = \langle p|\rho|p\rangle, \quad p = 0, 1, \ldots N - 1 \tag{13.94}$$

$$\sum_{q=0}^{2N-1} W_\rho(q, 2p + 1) = 0, \quad p = 0, 1, \ldots N - 1 \tag{13.95}$$

Proof: Let us calculate $\phi(p) = \sum_{q=0}^{2N-1} W_\rho(q, 2p)$. By Corollary 13.1, we only have to sum the even q indices. Then $\phi(p) = \sum_{q=0}^{N-1} W_\rho(2q, 2p)$. By Proposition 13.1 we get, using expression (13.82), that

$$W_\rho(2q, 2p) = \frac{1}{2N}(\langle q|\rho|q\rangle + \langle q + N/2|\rho|q + N/2\rangle) \tag{13.96}$$

where the sign of the second inner product is a plus because $2p$ is even. Write $\rho = \sum_i c_i |i\rangle\langle i|$. Take, for instance, $q = 0$. Then

$$\begin{aligned} W_\rho(0,2p) &= \frac{1}{2N}(\langle 0|\rho|0\rangle + \langle 0+N/2|\rho|0+N/2\rangle) \\ &= \frac{1}{2N}(\sum_i c_i \langle 0|i\rangle\langle i|0\rangle + \langle N/2|i\rangle\langle i|N/2\rangle) = \frac{1}{2N}(c_0 + c_{N/2}) \end{aligned} \tag{13.97}$$

As we know, $W_\rho(1,2p) = 0$. Take $q = 2$, then

$$W_\rho(2,2p) = \frac{1}{2N}(c_1 + c_{N/2+1}) \tag{13.98}$$

and so on (noting that we always have zeroes when q is odd). In this way, we end up summing all c_i coefficients c_i twice (because q varies between 0 and $2N-1$) and we get that

$$\phi(p) = \sum_{q=0}^{2N-1} W_\rho(q,2p) = \frac{1}{N}(c_0 + c_1 + \cdots + c_{2N-1}) = \frac{1}{N} \tag{13.99}$$

By the calculation above, we only have to calculate $\langle p|\rho|p\rangle$ and show that such number equals $1/N$. Recall that the inner product we consider is linear on the second variable, so we write $\rho = \sum_m c_m |m\rangle\langle m|$ and then:

$$\begin{aligned} \langle p|\rho|p\rangle &= \sum_m c_m \frac{1}{N} \sum_{j=0}^{N-1} \exp[-2\pi ijp/N] \sum_{l=0}^{N-1} \exp[2\pi ilp/N]\langle j|m\rangle\langle m|l\rangle \\ &= \sum_m c_m \frac{1}{N} \sum_{j=0}^{N-1} \exp[-2\pi ijp/N]\exp[2\pi imp/N]\langle j|m\rangle = \frac{1}{N}\sum_m c_m = \frac{1}{N} \end{aligned} \tag{13.100}$$

□

Conclusion By Propositions 13.1 and 13.2 we have for the discrete Wigner transform that if N is even and ρ is a density operator then

$$\sum_{p=0}^{2N-1} W_\rho(2q,p) = \langle q|\rho|q\rangle, \quad \sum_{p=0}^{2N-1} W_\rho(2q+1,p) = 0, \quad q = 0,1,\ldots,N-1 \tag{13.101}$$

and if

$$|p\rangle = \frac{1}{\sqrt{N}} \sum_{j=0}^{N-1} \exp[2\pi ijp/N]|j\rangle \tag{13.102}$$

then

$$\sum_{q=0}^{2N-1} W_\rho(q,2p) = \langle p|\rho|p\rangle, \quad \sum_{q=0}^{2N-1} W_\rho(q,2p+1) = 0, \quad p = 0,1,\ldots N-1 \tag{13.103}$$

Such expressions are the discrete analog of the result we have for the continuous Wigner function, namely the result that relates the marginals with the Fourier transform $\mathscr{F}$: if $\rho = |\psi\rangle\langle\psi|$ then

$$\int W_\rho(q,p)dp = |\psi(q)|^2, \quad \int W_\rho(q,p)dq = |\mathscr{F}\psi(p)|^2 \tag{13.104}$$

See [5] for more details.

Example 13.3. Denote by $\mathscr{W}_\rho$ the matrix with entries $W_\rho(q,p)$ for $q,p = 0,\ldots,2N-1$. For instance, if $N = 2$ and writing $|0\rangle = (1,0)$ and $|1\rangle = (0,1)$, we have that $\mathscr{W}_\rho$ contains the image of the Wigner function for each point of the phase space. We immediately notice that the integral over all space equals 1:

$$\mathscr{W}_{|0\rangle\langle 0|} = \begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 & 0 \\ \frac{1}{4} & -\frac{1}{4} & \frac{1}{4} & -\frac{1}{4} \\ 0 & 0 & 0 & 0 \end{pmatrix}, \mathscr{W}_{|1\rangle\langle 1|} = \begin{pmatrix} \frac{1}{4} & -\frac{1}{4} & \frac{1}{4} & -\frac{1}{4} \\ 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{13.105}$$

Example 13.4. Denote by $\mathscr{W}_\rho$ the matrix with entries $W_\rho(q,p)$ for $q,p = 0,\ldots,2N-1$. Let $N = 4$, and writing $|0\rangle = (1,0,0,0)$, $|1\rangle = (0,1,0,0)$, $|2\rangle = (0,0,1,0)$, $|3\rangle = (0,0,0,1)$, we have, in a similar way as seen in the previous example, that

$$\mathscr{W}_{|0\rangle\langle 0|} = \begin{pmatrix} \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{8} & -\frac{1}{8} & \frac{1}{8} & -\frac{1}{8} & \frac{1}{8} & -\frac{1}{8} & \frac{1}{8} & -\frac{1}{8} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$\mathscr{W}_{|1\rangle\langle 1|} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{8} & -\frac{1}{8} & \frac{1}{8} & -\frac{1}{8} & \frac{1}{8} & -\frac{1}{8} & \frac{1}{8} & -\frac{1}{8} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$\mathscr{W}_{|2\rangle\langle 2|} = \begin{pmatrix} \frac{1}{8} & -\frac{1}{8} & \frac{1}{8} & -\frac{1}{8} & \frac{1}{8} & -\frac{1}{8} & \frac{1}{8} & -\frac{1}{8} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$\mathscr{W}_{|3\rangle\langle 3|} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{8} & -\frac{1}{8} & \frac{1}{8} & -\frac{1}{8} & \frac{1}{8} & -\frac{1}{8} & \frac{1}{8} & -\frac{1}{8} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Remark 1 What occurs in general for pure states: the Wigner function $W_{|q_0\rangle\langle q_0|}$ is zero except in two lines, located in $q \equiv 2(modN)$. When $q = 2q_0$, W assumes the value $1/2N$, and when $q = 2q_0 \pm N$, W assumes the value $1/2N$ for even values of p and $-1/2N$ for odd values. Such oscillations are typical of interference fringes and can be interpreted as arising from the interference between the line $q = 2q_0$ and a mirror image formed at a distance of $2N$ from $2q_0$, induced by the periodic boundary conditions [7].

Remark 2 The fact that the Wigner function assumes negative values in the interference line is essential for one to be able to recover the correct marginal distributions. Summing the values $W(q, p)$ along a vertical line gives us the probability of measuring $q/2$, which should be equal to 1 if $q = 2q_0$, and equal to zero, otherwise.

A natural question is to try to understand the action of the operator which defines QIFS in the dual variables p. This is the purpose of the next results.

Lemma 13.3. *Let $\Lambda(\rho) = \sum_i V_i \rho V_i^*$ and define $F(\rho) = \mathscr{F} \rho \mathscr{F}^*$, where $\mathscr{F}$ is any unitary map. Then there is $G : \mathscr{M}_N \to \mathscr{M}_N$ such that the above diagram commutes:*

$$\begin{tikzcd} \mathscr{M}_N \arrow[r, "F"] \arrow[d, "\Lambda"'] & \mathscr{M}_N \arrow[d, "G"] \\ \mathscr{M}_N \arrow[r, "F"] & \mathscr{M}_N \end{tikzcd} \tag{13.106}$$

Proof: First, note that $F^{-1}(\rho) = \mathscr{F}^* \rho \mathscr{F}$. Also $\mathscr{F}$ is unitary, therefore we have $\mathscr{F}^{-1} = \mathscr{F}^*$. Define $G = F \circ \Lambda \circ F^{-1}$. Explicitly,

$$\begin{aligned} G(\rho) &= F(\sum_i V_i \mathscr{F}^* \rho \mathscr{F} V_i^*) = \mathscr{F}\Big[\sum_i V_i \mathscr{F}^* \rho \mathscr{F} V_i^*\Big]\mathscr{F}^* \\ &= \sum_i \mathscr{F} V_i \mathscr{F}^* \rho \mathscr{F} V_i^* \mathscr{F}^* = \sum_i \tilde{V}_i \rho \tilde{V}_i^* \end{aligned}$$

where $\tilde{V}_i = \mathscr{F} V_i \mathscr{F}^*$. And a simple inspection shows that

$$F(\varLambda(\rho)) = G(F(\rho)) = \sum_i \mathscr{F} V_i \rho V_i^* \mathscr{F}^* \qquad \square$$

Example 13.5. Consider $N = 2$. Then the discrete Fourier transform is given by

$$\mathscr{F} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \tag{13.107}$$

In this case we have $\mathscr{F}^{-1} = \mathscr{F}$. Let

$$V_1 = \begin{pmatrix} \sqrt{p_{11}} & 0 \\ 0 & 0 \end{pmatrix}, \quad V_2 = \begin{pmatrix} 0 & \sqrt{p_{12}} \\ 0 & 0 \end{pmatrix}, \tag{13.108}$$

$$V_3 = \begin{pmatrix} \sqrt{p_{21}} & 0 \\ 0 & 0 \end{pmatrix}, \quad V_4 = \begin{pmatrix} 0 & 0 \\ 0 & \sqrt{p_{22}} \end{pmatrix} \tag{13.109}$$

where the p_{ij} form a column stochastic matrix P. Then Lemma 13.3 for this example shows that $G(\rho) = \sum_i \tilde{V}_i \rho \tilde{V}_i^*$, where

$$\tilde{V}_1 = \mathscr{F} V_1 \mathscr{F}^* = \frac{1}{2}\sqrt{p_{11}}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad \tilde{V}_2 = \mathscr{F} V_2 \mathscr{F}^* = \frac{1}{2}\sqrt{p_{12}}\begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix}$$

$$\tilde{V}_3 = \mathscr{F} V_3 \mathscr{F}^* = \frac{1}{2}\sqrt{p_{21}}\begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix}, \quad \tilde{V}_4 = \mathscr{F} V_4 \mathscr{F}^* = \frac{1}{2}\sqrt{p_{22}}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

Then, from $p_{11} + p_{21} = 1$, $p_{12} + p_{22} = 1$ and writing

$$\rho = \begin{pmatrix} \rho_{11} & \rho_{12} \\ \rho_{21} & 1 - \rho_{11} \end{pmatrix}$$

we get from Lemma 13.3 the expression

$$\begin{aligned} F(\varLambda(\rho)) &= G(F(\rho)) = \sum_i \mathscr{F} V_i \rho V_i^* \mathscr{F}^* \\ &= \begin{pmatrix} \frac{1}{2} & p_{11}\rho_{11} + p_{12}(1-\rho_{11}) - \frac{1}{2} \\ p_{11}\rho_{11} + p_{12}(1-\rho_{11}) - \frac{1}{2} & \frac{1}{2} \end{pmatrix} \end{aligned} \tag{13.110}$$

In the case that the vector $\pi = (\rho_{11}, 1 - \rho_{11})$ is fixed for the stochastic matrix P, we can rewrite the expression above as

$$F(\Lambda(\rho)) = G(F(\rho)) = \begin{pmatrix} \frac{1}{2} & \rho_{11} - \frac{1}{2} \\ \rho_{11} - \frac{1}{2} & \frac{1}{2} \end{pmatrix} \qquad (13.111)$$

Lemma 13.4. *Define $\Lambda : \mathscr{M}_N \to \mathscr{M}_N$, $\Lambda(\rho) = \sum_i V_i \rho V_i^*$, with V_i linear, $\sum_i V_i^* V_i = I$ and let $W_{\Lambda(\rho)}$ be the associated discrete Wigner function. Then given (q, p) there are $M_i = M_i(q, p)$ such that*

$$W_{\Lambda(\rho)}(q, p) = \sum_i tr(M_i \rho M_i^*)$$

Proof: First, as $A(q, p)$ is hermitian, we have a decomposition

$$A = UDU^{-1}$$

where U is unitary and D is diagonal (and real). Then

$$A^{1/2} = UD^{1/2}U^{-1}$$

where $(A^{1/2})^2 = A$, $D^{1/2}$ is the diagonal matrix whose entries are the positive square roots of the entries of D. Then

$$\begin{aligned} W_{\Lambda(\rho)}(q, p) &= tr(\hat{A}(q, p)\Lambda(\rho)) = tr(\hat{A}\sum_i V_i \rho V_i^*) = \sum_i tr(\hat{A} V_i \rho V_i^*) \\ &= \sum_i tr(A^{1/2} V_i \rho V_i^* A^{1/2}) = \sum_i tr(UD^{1/2}U^{-1} V_i \rho V_i^* UD^{1/2}U^{-1}) \end{aligned} \qquad (13.112)$$

Defining $M_i = UD^{1/2}U^{-1}V_i$ and noting that $U^{-1} = U^*$, we can write

$$W_{\Lambda(\rho)}(q, p) = \sum_i tr(M_i \rho M_i^*) \qquad \square$$

References

1. Baraviera, A., Lardizabal, C.F., Lopes, A.O., Terra Cunha, M.: A thermodynamic formalism for density matrices in quantum information. Appl. Math. Res. Express **1**, 63–118 (2010)
2. Baraviera, A., Lardizabal, C.F., Lopes, A.O., Terra Cunha, M.: Quantum stochastic processes, quantum iterated function systems and entropy. São Paulo J. Math. Sci. (2010) (To appear)
3. Benatti, F.: Dynamics, Information and Complexity in Quantum Systems. Springer (2009)
4. García-Mata, I., Saraceno, M.: Spectral approach to chaos and quantum-classical correspondence in quantum maps. Mod. Phys. Let. B **19**(7, 8) (2005)

5. de Gosson, M.: Symplectic Geometry and Quantum Mechanics. Birkhauser (2006)
6. Lozinski, A., Życzkowski, K., Słomczyński, W.: Quantum iterated function systems. Phys. Rev. E **68**, 04610 (2003)
7. Miquel, C., Paz, J.P., Saraceno, M.: Quantum computers in phase space. Phys. Rev. A. **65**, 062309 (2002)
8. Słomczyński, W., Życzkowski, K.: Quantum Chaos: an entropy approach. J. Math. Phys. **32**(1), 5674–5700 (1994)
9. Słomczyński, W.: Dynamical Entropy, Markov Operators and Iterated Function Systems. Jagiellonian University Press (2003)
10. Tannor, D.J.: Introduction To Quantum Mechanics: A Time-dependent Perspective. University Science Books (2006)
11. Wootters, W.K.: A Wigner-Function formulation of finite-state quantum mechanics. Ann. Phys. **176**, 1–21 (1987)

Chapter 14
Dynamics on Spectral Solutions of Forced Burgers Equation

Mário Basto, Viriato Semiao, and Francisco Lage Calheiros

Abstract Burgers equation $\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = \delta \frac{\partial^2 u}{\partial x^2} + f\ (x)$ is one of the simplest partial nonlinear differential equation which can develop discontinuities, being the driven equation used to explore unidimensional "turbulence". For low values of the viscosity coefficient δ, by discretization through spectral collocation methods, oscillations in Burgers equation can occur. For the Dirichlet problem and under a dynamic point of view, several bifurcations and stable attractors can be observed. Periodic orbits, nonperiodic and strange attractors may arise. Bistability can also be observed. Numerical simulations indicate that the loss of stability of the asymptotic solution of Burgers equation must occur by means of a supercritical Hopf bifurcation. Many nonlinear phenomena are modeled by spatiotemporal systems of infinite or very high dimension. Coupling and synchronization of spatially extended dynamical systems, periodic or chaotic, have many applications, including communications systems, chaos control, estimation of model parameters and model identifications. For the unidirectionally linear coupling, numerical studies show the presence of identical or generalized synchronization for different values of spacial points and different values of the viscosity coefficient δ. Also, nonlinear coupling by a convex linear combination of the drive and driven variables corresponding to the waves velocity, can achieve identical or generalized synchronization.

M. Basto (✉)
IPCA, Barcelos, Portugal
e-mail: mbasto@ipca.pt

V. Semiao
IST, Lisbon, Portugal
e-mail: ViriatoSemiao@ist.utl.pt

F.L. Calheiros
FEUP, Porto, Portugal
e-mail: jolacam@netcabo.pt

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_14,

14.1 Dynamics in Spectral Solutions of Burgers Equation

Burgers equation with small values of the viscosity coefficient δ, develops waves with sharp slopes, leading to the appearance of discontinuities for values $\delta \to 0$. By discretization through spectral collocation methods, oscillations in Burgers equation can occur. Under a dynamic point of view, these oscillations may be related to bifurcations and atractors arising to the discretized equation.

Dang-Vu and Delcarte [1] and Basto et al. [2] provided numerical studies of the following Dirichlet problem for Burgers equation with homogeneous boundary conditions, by Chebyshev collocation method,

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = \delta \frac{\partial^2 u}{\partial x^2} + f(x) \quad -1 \leq x \leq 1 \tag{14.1}$$

with $f(x) = \pi \sin(\pi x)[\cos(\pi x) + \delta\pi]$.

The following equation with $N-1$ degrees of freedom is obtained by discretization of (14.1) with $N+1$ points x_j, $0 \leq j \leq N$, by Chebyshev collocation method,

$$\frac{du_i}{dt} = -u_i D_i^{(1)} u + \delta D_i^{(2)} u + f_i, \quad 1 \leq i \leq N-1 \tag{14.2}$$

where $u_1 = u(x_1, t)$, $u_2 = u(x_2, t)$, ..., $u_{N-1} = u(x_{N-1}, t)$, $u = [u_1, u_2, ..., u_{N-1}]^T$, $f_i = f(x_i)$ and $D^{(i)}$, $1 \leq i \leq 2$ are the Chebyshev differentiation matrices of order i. The problem is then reduced to a system of ordinary differential equations of order $N-1$.

Dang-Vu and Delcarte [1] found out a critical value of the viscosity δ for (14.1), where a Hopf bifurcation took place and a periodic orbit around the critical point arose.

Besides the trapping region found by Dang-Vu and Delcarte [1], arising from the loss of stability of the periodic orbits emerging from Hopf bifurcation, other phenomena and bifurcation can be observed [2]. In fact, it is observed the existence of torus type attractors or strange attractors, for lower values of δ, before the dynamics becomes unbounded. Also, bistability is observed. In this case, both the coexistence of two periodic attractors, a periodic and a nonperiodic one (torus type or strange attractor), or even two nonperiodic attractors can be observed. In addition, other stable equilibrium points can occur, diverse from the ones corresponding to the asymptotic solution of Burgers equation. For few cases, positive values yielded by the largest Lyapunov exponent for some nonperiodic motions, provide evidence of chaotic attractors.

As an example of bistability, for $N = 16$ and $\delta = 0.0061$, a nonperiodic attractor and a periodic one is observed in Fig. 14.1.

To further investigate the dynamical behavior of the spectral solutions of Burgers equation, more studies were made involving different functions f [3].

Burgers equation is a wave nonlinear equation where the convection u is active since it depends on the solution of the equation. As the speed of the waves is given by the solution itself, it increases when u increases and decreases when u decreases.

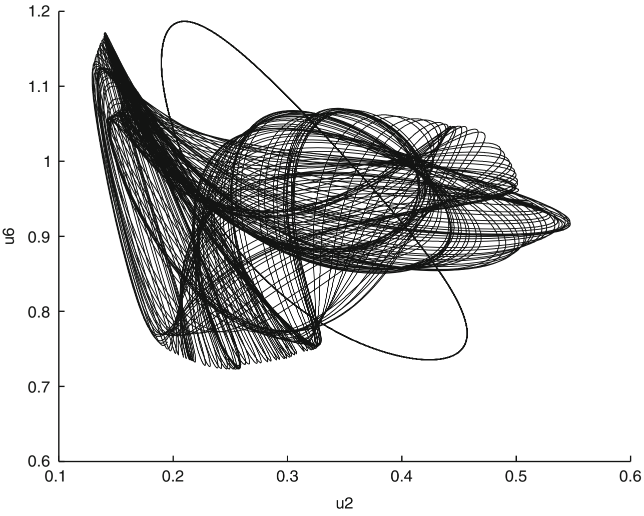


Fig. 14.1 $N = 16$. Bistability for $\delta = 0.0061$, with a periodic orbit and a quasiperiodic one

The higher points of the nonlinear wave travel at a higher speed and shocks and discontinuities for low values of δ will tend to appear in the intervals where u is decreasing. The instabilities observed in the forced Burgers equation will then tend to appear first at intervals where the asymptotic solution is decreasing. These branches of the solution are the ones that are fixed by one extremity to each fixed boundary for the Dirichlet problem. By numerically studying several examples, one argues that this fact, together with the nonexistence of such branches where discontinuities tend to appear not fixed to the boundaries, is a necessary condition to keep the asymptotic equilibrium solution stable for lower values of δ, giving time for the first loss of stability to be signed by the emergency of a supercritical Hopf bifurcation [3].

14.2 Dynamics in Coupled Spectral Solutions of Burgers Equation

Coupling and synchronization of spatially extended dynamical systems is an area of intensive research, concerning communications systems, chaos control and estimation of model parameters. Besides synchronization of periodic signals, it has been shown that it is also possible to synchronize chaotic dynamical systems [4].

14.2.1 Linear Coupling

Consider two unidirectionally coupled discretized Burgers equations for the Dirichlet problem with homogeneous boundary conditions, and a linear coupling between them:

$$\frac{du_i}{dt} = -u_i D_i^{(1)} u + \delta_u D_i^{(2)} u + f_i \tag{14.3}$$

$$\frac{dv_i}{dt} = -v_i D_i^{(1)} v + \delta_v D_i^{(2)} v + f_i + \alpha (u_i - v_i) \tag{14.4}$$

Numerical experiments for synchronization with parameter mismatch between drive and driven equations $\delta_u \neq \delta_v$, by means of the auxiliar system approach and the negativeness of the conditional Lyapunov exponents of the response equation, confirm for this case the possibility of generalized synchronization for an adequate coupling strength α [3].

14.2.2 Nonlinear Coupling

The procedure consists of replacing the discretized response variable v by $v + \alpha (u - v)$, where u represents the drive and α the coupling parameter. For $\alpha = 1$ one reaches a situation of partial replacement.

With parameter mismatch it is observed that coupling only at the position corresponding to the waves velocity v can lead to generalized synchronization, but generally not allowing values of $\alpha = 1$. This means that the partial replacement in certain locations may not lead to synchronization, but the convex linear combination do [3].

The tools used to perform this study were MATLAB [5] and MATCONT [6, 7].

References

1. Dang-Vu, H., Delcarte, C.: Hopf bifurcation and strange attractors in Chebyshev spectral solutions of the Burgers equation. Appl. Math. Comput. **73**, 99–113 (1995)
2. Basto, M., Semiao, V., Calheiros, F.: Dynamics in spectral solutions of Burgers equation. J. Comput. Appl. Math. **205**, 296–304 (2006)
3. Basto, M., Semiao, V., Calheiros, F.: Dynamics and synchronization of numerical solutions of the Burgers equation. J. Comput. Appl. Math. **231**, 793–806 (2009)
4. Pecora, L.M., Carroll, T.L., Johnson, G.A., Mar D.J., Heagy, J.F.: Fundamentals of synchronization in chaotic systems, concepts, and applications. Chaos Interdisciplinary J. Nonlinear Sci. **7**, 520–543 (1997)
5. MATLAB, The Mathworks Inc., http://www.mathworks.com (verified on 01/02/2011)
6. Dhooge, A., Govaerts, W., Kuznetsov, Y.A.: MATCONT, a graphical Matlab package, http://www.matcont.ugent.be/matcont.html (verified on 01/02/2011)
7. Dhooge, A., Govaerts, W., Kuznetsov, Y.A.: MATCONT: A MATLAB Package for Numerical Bifurcation Analysis of ODEs. ACM Trans. Math. Software (TOMS) **29**, 141–164 (2003)

Chapter 15
Area-Preserving Diffeomorphisms from the C^1 Standpoint

Mário Bessa

Abstract More than thirty years have passed since Newhouse (Am. J. Math. 99:1061–1087, 1977) published a dichotomy on C^1 area-preserving diffeomorphisms. Here we revisit some central results on surface conservative C^1-diffeomorphisms by presenting, in particular, a new proof of Newhouse's theorem and also by proving some, although folklore, not yet proved results on this setting. We intend that this exposition can be used by a large audience as an introduction to the concept of dominated splitting and its relevance to the theory of C^1-stability of area-preserving diffeomorphisms.

15.1 Introduction

Let M be a compact, connected, boundaryless, Riemannian surface and let ω be an area-form on M. Denote by $\mathrm{Diff}^1_\omega(M)$ the space of diffeomorphisms on M, of class C^1, such that $f_*\omega = \omega$, that is, any Lebesgue measurable subset $\mathscr{M} \subset M$ satisfy $Leb(\mathscr{M}) = Leb(f(\mathscr{M}))$, where $Leb(\cdot)$ denotes the Lebesgue measure induced by the two-form ω. We endow the set $\mathrm{Diff}^1_\omega(M)$ with the Whitney C^1 topology (see Sect. 15.2.1). The set $(\mathrm{Diff}^1_\omega(M), C^1)$ is a *Baire space*, hence every intersection of countably many C^1-dense and C^1-open sets is C^1-dense.

These area-preserving (or conservative) diffeomorphisms in surfaces are a traditional object of study from Classical Mechanics, see e.g. [5]. Despite being outside the scope of our text we recall the Kolmogorov, Arnold and Moser (KAM) theorem, see e.g. [37], which gives prevalence of dynamically invariant circles supporting irrational rotations.

M. Bessa
Departamento de Matemática, Universidade do Porto, Rua do Campo Alegre 687, 4169-007 Porto, Portugal
and
ESTGOH-Instituto Politécnico de Coimbra, Rua General Santos Costa, 3400-124 Oliveira do Hospital, Portugal
e-mail: bessa@fc.up.pt

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_15,

The concept of periodic points plays a central role in dynamical systems and so we recall that a point x is said to be *periodic* for the diffeomorphism $f\colon M \to M$ if

$$f^n(x) = x \text{ where } f^n(x) = \overbrace{f \circ f \circ \ldots \circ f}^{n\text{-times}}(x), \text{ for } n \in \mathbb{N},$$

and the least of these positive integers is called the *period* of x. Moreover, it is well known that the knowledge of the behavior of the derivative of f, Df, along periodic orbits gives us a deep understanding of the local dynamics of f.

Given a periodic point x of period n of a diffeomorphism f if the n-iterated tangent map of f at x, denoted by Df^n_x, has its spectrum in $\mathbb{S}^1 \setminus \mathbb{R}$, then x is called *elliptic*. On the other hand if the spectrum does not intersect $\mathbb{S}^1$ then the point x is called *hyperbolic*. We recall that a periodic point is said to be *Lyapunov stable* if the iterates of all nearby points remain bounded for all time. So, KAM's theorem, implies abundance of Lyapunov stable elliptic points.

In spite that KAM theory needs higher order of differentiability of the diffeomorphisms it is our purpose to study systems with only C^1 regularity; which means closeness up to the first derivative.

The aim of this paper is to understand the typical dynamics for the elements $f \in \text{Diff}^1_\omega(M)$. Some property could be considered to be *typical* if it holds for an open and dense subset, or even for some dense subset. However, the notion of *typical* that we are going to use here means that for a *generic* (or *residual*) set some property holds. Let us make this idea more precise; we say that the property $\mathscr{P}$ holds in a C^1-residual set of $\text{Diff}^1_\omega(M)$ if $\mathscr{P}$ contains a G_δ, that is, a countable intersection of C^1-open and C^1-dense sets. In particular, as we already mention, by Baire's theorem (see e.g. [16]), any G_δ is dense in $\text{Diff}^1_\omega(M)$.

Let us display some capital results on C^1-generic conservative diffeomorphisms in surfaces:

(A) Every periodic point is hyperbolic or elliptic.
(B) M is the closure of the set of periodic points.
(C) The diffeomorphism is transitive, that is, it has a dense orbit.

The property (A) is a consequence of Thom's transversality theorem and was proved by Robinson [32], actually, this is a C^r-generic property, $r \geq 2$. Property (C) is a corollary of an theorem by Bonatti and Crovisier [11]. Item (B) is the so-called *general density theorem* proved by Pugh and Robinson (see [31]) and says that for a C^1-generic set $\mathscr{G} \subset \text{Diff}^1_\omega(M)$, we have that the set of periodic points for $f \in \mathscr{G}$ is dense in the *nonwandering set*[1] of f denoted by $\Omega(f)$.

We say that $x \in M$ is an f*-recurrent point* if given any neighborhood U of x, there exists n such that $f^n(x) \in U$. Poincaré's recurrence theorem (see e.g. [22]) states that for $f \in \text{Diff}^1_\omega(M)$ Lebesgue almost every point is recurrent. Hence,

[1] Recall that $x \in \Omega(f)$ if for every neighborhood U of x there exists $n \in \mathbb{N}$ such that $f^n(U) \cap U = \emptyset$.

we conclude that Lebesgue almost every point is nonwandering and that, in the conservative class, C^1-generically the closure of the set of periodic points is the entire manifold M.

At this point we ask, given a C^1-generic area-preserving diffeomorphism, how often we find elliptic periodic orbits? And hyperbolic ones?

Recall that, due to the Hartman–Grobman theorem (see e.g. [32]), we have that hyperbolic periodic points are topological conjugated to its derivative, and so its local dynamics is simple. The hyperbolicity reveals stable also for sets (see e.g. [35, Chap. 8]). We say that a surface diffeomorphism is *completely hyperbolic*, or *Anosov*, if there exists $0 < \lambda < 1$ such that, for all $x \in M$, the tangent space decomposes into two one-dimensional subbundles on which the derivative contracts backward by a factor of λ in one subbundle and contracts under positive iterates by a factor of λ in the other direction. These geometric and dynamical properties imply a topological restriction in the manifold; the only surfaces that supports Anosov diffeomorphisms are the tori (see [15]). Another relevant property is that the Anosov diffeomorphisms are open (see [35]), thus the set of Anosov diffeomorphisms in $\text{Diff}_\omega^1(M)$ is also open in $\text{Diff}_\omega^1(M)$.

In the mid-1970s, (see [24]), Newhouse proved a result on area-preserving diffeomorphisms. He presented a C^1-generic set $\mathscr{R} \subset \text{Diff}_\omega^1(M)$ such that for any $f \in \mathscr{R}$ either f is Anosov or else the elliptic points are dense in M. As a corollary of this result and of the aforementioned topological restriction, we obtain that, for example, in any surface aside from the torus, C^1-generic area-preserving diffeomorphisms have dense elliptic orbits.

In this paper we will give a new proof of Newhouse's theorem based in the perturbation techniques *à la Mañé* (see [20,21]). These perturbations were first developed, in the conservative setting, by Bochi in [9] to prove the so-called Bochi–Mañé Theorem (see Theorem 15.6 and the references wherein).

Let us stress that, since $\text{Diff}_\omega^1(M)$ is not C^1 dense among the set of C^1 dissipative diffeomorphisms in surfaces, our perturbations are more rigid and some careful is needed to perform them.

In order to obtain Newhouse's dichotomy we apply some perturbation results [2, 4], jointly with the approach in [1, 7] and by making use of the above-mentioned Bochi–Mañé theorem. Mañé's ideas are an intrinsic part of this exposition and a recurrent influence.

The main dynamical ingredient is to use the absent of a hyperbolic behavior to perturb, in the C^1 topology and along a large period orbit, in order to transform this hyperbolic periodic orbit into an elliptic one with the same period. One crucial fact can be taken in account; we need to take small neighborhoods of the periodic hyperbolic orbit, and that is why we are restricted to the C^1 topology. The C^1 topology allow us to *rescale* the support of the perturbation with no implication to the size of the perturbation (see Lemma 15.4). However, the attempt to replace the C^1 topology by higher order ones is very difficult because the size of the perturbations increases if we decrease the support of the perturbation. These are the main difficulties which are the base of one of the most challenging problems in the modern theory of dynamical systems; the C^r*-closing lemma* (for $r \geq 2$), see [12] A.1 for details.

We recall that Newhouse's proof of [24, Theorem 1.1] (see Theorem 15.2) uses the concept of *homoclinic point* (see Sect. 15.1.1 for the definitions). Actually, in [24, Lemma 4.1], it is proved that a homoclinic tangency $T \in M$ associated to a hyperbolic periodic point for $f \in \text{Diff}_{\omega}^{1}(M)$ has a g-elliptic periodic point near T for g C^1-close to f. Then, Newhouse apply [31, 36] and the Birkhoff norm form to perturb $f \in \text{Diff}_{\omega}^{1}(M)$ in order to obtain that the homoclinic points of the perturbed diffeomorphism are dense in M. Finally, if the original diffeomorphism f is not Anosov, then there exists g, C^1-close to f, and exhibiting an elliptic orbit passing through any pre-fixed open set $U \subset M$.

15.1.1 Statement of the Main Results

We start with Newhouse's dicothomy for area-preserving diffeomorphisms.

Theorem 15.1. *There exists a residual set $\mathscr{R} \subset Diff_{\omega}^{1}(M)$ such that for $f \in \mathscr{R}$*

- *Either f is Anosov.*
- *Or else the elliptic points are dense in M.*

This theorem is a consequence of the following result.

Theorem 15.2. *Given any non Anosov diffeomorphism $f \in Diff_{\omega}^{1}(M)$, $\epsilon > 0$ and any non empty open subset U of M, then there exists $g \in Diff_{\omega}^{1}(M)$ ϵ-C^1-close to f and exhibiting an elliptic orbit passing through U.*

Previous theorems were proved by Newhouse (see [24];Theorems 1.1 and 1.3). Saghin and Xia (see [34, Theorem 2]), proved a general $2n$ symplectic perturbation results which allowed them to obtain the higher dimensional version of Theorem 15.2. Let us stress that the perturbation results used by these authors were already explored by Bochi and Viana in [10] and also that, in [3], Arnaud obtained the four dimensional counterpart of Theorem 15.2. We point out that these results are restricted to the symplectic context, and not to the broader setting of the volume-preserving diffeomorphisms, because the stability of elliptic points (which is false for volume-preserving diffeomorphisms on dimension ≥ 3) plays a crucial role in the arguments.

We say that a diffeomorphism $f: M \to M$ is *transitive* if there exists a dense orbit $x \in M$, that is, $\overline{\cup_{n\in\mathbb{N}} f^n(x)} = M$ where $\overline{A}$ stands for the closure of the set A. Moreover, a diffeomorphism $f: M \to M$ is said to be C^1*-robustly transitive* (in the conservative class) if it is transitive and every sufficiently C^1-close and conservative one is also transitive. Classical examples are the area-preserving Anosov diffeomorphisms. Actually, in dimension two these are the only examples. In Sect. 15.7.1 we will present another proof of [2, Theorem 5.1] by making use of a KAM-type theorem.

Theorem 15.3. *If $f \in Diff_{\omega}^{1}(M)$ is C^1-robustly transitive, then f is Anosov.*

Let $f \in \mathrm{Diff}_{\omega}^{1}(M)$ we say that f is a *conservative star-diffeomorphism* if there exists a neighborhood $\mathscr{V}$ of f in $\mathrm{Diff}_{\omega}^{1}(M)$ such that any $g \in \mathscr{V}$, has all the periodic orbits hyperbolic. We denote this set by $\mathscr{F}_{\omega}^{1}(M)$. We define analogously the set $\mathscr{F}^{1}(M)$ in the broader set of dissipative diffeomorphisms $\mathrm{Diff}^{\,1}(M)$.

Let $\mathscr{A}_{\omega}^{2}$ denote the set of conservative Anosov diffeomorphisms on the surface M. Recall that the set $\mathscr{A}_{\omega}^{2}$ is open in $\mathrm{Diff}_{\omega}^{1}(M)$. Moreover, if $A \in \mathscr{A}_{\omega}^{2}$, then $f \in \mathscr{F}_{\omega}^{1}(M)$. In the next result we obtain the converse.

Theorem 15.4. *If $f \in \mathscr{F}_{\omega}^{1}(M)$, then f is Anosov.*

We recall that the *dissipative* version of previous result was proved by Mañé (see [19]), loosely speaking, any $f \in \mathscr{F}^{1}(M)$ has a hyperbolic-type behavior. A diffeomorphism is said to be C^1-*structurally stable* if there is a C^1-neighborhood of f on $\mathrm{Diff}_{\omega}^{1}(M)$ such that any $g \in \mathrm{Diff}_{\omega}^{1}(M)$ in this neighborhood is topologically conjugate to f, i.e., there exists a global homeomorphism h such that $h \circ f = g \circ h$. As we already pointed out the Anosov systems are structurally stable (see [35]), and in Theorem 15.10 we will obtain the converse.

Given a periodic hyperbolic orbit $\mathscr{O}$ and $p \in \mathscr{O}$ let W_p^s (respectively W_p^u) denote the stable (respectively unstable) manifold of p that is:

$$W_p^s := \left\{ x \in M \colon dist(f^n(x), f^n(p)) \underset{n \to +\infty}{\to} 0 \right\}$$

and

$$W_p^u := \left\{ x \in M \colon dist(f^{-n}(x), f^{-n}(p)) \underset{n \to +\infty}{\to} 0 \right\}.$$

There exists a very complete theory about these invariant manifolds (see [35]).

We say that $\mathscr{O}$ has a *homoclinic tangency* at $q \neq p$ if:

- $T_q W_p^s \cap T_q W_p^u$ contains a nonzero vector and
- $T_q W_p^s \oplus T_q W_p^u \neq T_q M$.

We say that q is a *transversal homoclinic point* if it is not a homoclinic tangency.

The next result, that will be proved in Sect. 15.7.3, is in the spirit of Palis' conjecture [28] and with respect to the C^1-topology (see [24, (6) on page 1075]).

Theorem 15.5. *Any $f \in Diff_{\omega(M)}^{1}$ can be C^1-approximated by another one $g \in Diff_{\omega(M)}^{1}$ satisfying one of the following properties:*

1. *g is Anosov or else,*
2. *g has a homoclinic tangency associated to a hyperbolic periodic orbit.*

In Sect. 15.2 we set up notation, terminology and standard facts on uniform hyperbolic theory. Section 15.4 provides a detailed exposition of the perturbations that we will use in order to go on with the main proofs. In Sect. 15.6 we present the proof of Theorem 15.2. Theorem 15.1 shall be proved in Sect. 15.3 assuming Theorem 15.2. In Sect. 15.5 we will be concerned with the creation of elliptic periodic orbits by C^1 small perturbations. Finally, in Sect. 15.7 we will restrict our attention to some results about robust transitivity, stability, bifurcations on periodic points and some questions about the coexistence of two different definitions of chaos in the C^1 sense (see Theorem 15.12).

15.2 Preliminaries and Basic Definitions

15.2.1 Charts and Neighborhoods

By compactness of M we can use Darboux's theorem (see e.g. [5]) and obtain a finite atlas $\mathscr{A} = \{\varphi_i : U_i \to \mathbb{R}^2\}$, for $i = 1, \ldots, k$ and thus define local coordinates such that the pullback of the two form ω by φ_i is the canonical area in the plane, i.e., $(\varphi_i)_*\omega = dx \wedge dy$. Note that we can switch the metric associated to the Riemannian structure of M at $x \in M$ by the metric $\|\cdot\| = \|D(\varphi_{i(x)})_x(\cdot)\|$ where $i(x)$ is uniquely defined and associated to each $x \in M$. For this reason we will not use the Riemannian metric *a priori* fixed on M. Denote by $dist(\cdot,\cdot)$ the distance inherit from the Riemannian structure in M and the pre-fixed charts; that is, given $x, y \in M$ with $y \in U_{i(x)}$, $d(x, y) := \|\varphi_{i(x)}(x) - \varphi_{i(x)}(y)\|$.

We sometimes consider balls in M defined by

$$B(x, r) := \varphi^{-1}_{i(x)}[B(\varphi_{i(x)}(x), r)],$$

where $r > 0$ is chosen to be sufficiently small in order to have each ball contained in the open set U_i for $i = 1, \ldots, k$.

Given any 1-linear map $A \in \mathscr{L}(\mathbb{R}^2)$ we consider the norm

$$\|A\| := \sup_{v \neq \mathbf{0}} \frac{\|A \cdot v\|}{\|v\|}. \qquad (15.1)$$

This norm will be used to estimate distances between two maps and it will be the one fixed in the preceding paragraph. In the sequel we will also use another norm which will reveal to be useful when dealing with estimates (see Sect. 15.2.2).

As a consequence, every time we compute distances between two maps we use Darboux's theorem to translate the scenario to $\mathbb{R}^2$. So let us define properly the distance we are going to consider. Given $f \in \text{Diff}^1_\omega(M)$, a finite atlas $\{\varphi_i\}_{i \in F}$, compact sets $K_i \subset U_i$ such that $f(K_i) \subset U_i$ for all $i \in F$ and $\epsilon > 0$, we say that $U(f, \varphi, K_i, \epsilon)$ is an ϵ-C^1 basic neighborhood of f in the *Whitney C^1-topology* if it is formed by those maps $g \in \text{Diff}^1_\omega(M)$ such that:

- $g(K_i) \subset U_i$ and
- (C^0-closeness) $\sup_{x \in \varphi_i(K_i)} \{\|\varphi_{i(f(x))} f \varphi^{-1}_{i(x)}(x) - \varphi_{i(g(x))} g \varphi^{-1}_{i(x)}(x)\|\} < \epsilon$ and

- (C^1-closeness) $\sup_{x\in\varphi_i(K_i)} \{\|D(\varphi_{i(f(x))} f \varphi^{-1}_{i(x)})(x) - D(\varphi_{i(g(x))} g \varphi^{-1}_{i(x)})(x)\|\} < \epsilon$.

In this way we obtain what we shall call the ϵ-C^1-neighborhood of f and we denote it by $\mathcal{N}^{\omega}_{\epsilon}(f)$.

15.2.2 Some Elementary Linear Algebra

15.2.2.1 The Linear Group $SL(2,\mathbb{R})$, Angles and Eigenvalues

We say that a 2×2 matrix A belongs to $SL(2,\mathbb{R})$ if $\det(A) = 1$. Moreover, if the eigenvalues of $A \in SL(2,\mathbb{R})$ are real and distinct we say that the matrix is *hyperbolic*. A matrix $A \in SL(2,\mathbb{R})$ is *elliptic* if the eigenvalues are different complex conjugates. Finally, we call $A \in SL(2,\mathbb{R})$ *parabolic* if it is not hyperbolic neither elliptic. It is easy to see that stability (with respect to the norm defined in (15.1)), within these three classes of matrices, holds both for hyperbolic and elliptic matrices whilst the parabolic ones are unstable.

Given a hyperbolic matrix $A \in SL(2,\mathbb{R})$, let $\sigma > 1$ be the upper eigenvalue and $\theta > 0$ be the angle between its eigenspaces. We define the function

$$\eta_\theta(\sigma) = \|A - Id\|, \tag{15.2}$$

where Id denotes the identity in $\mathbb{R}^2$. Of course that $\eta_\theta(\cdot)\colon]1,+\infty[\to]0,+\infty[$ defined by $\sigma \mapsto \eta_\theta(\sigma)$ is a strictly increasing diffeomorphism. On the other hand, the map $\eta_{(\cdot)}(\sigma)\colon]0,\pi/2[\to]\sigma,+\infty[$ defined by $\theta \mapsto \eta_\theta(\sigma)$ is a strictly decreasing diffeomorphism.

15.2.2.2 A New Norm

Let be given $x, y \in M$, a linear map $A\colon T_xM \to T_yM$ and two invariant 1-dimensional splittings $E^1_x \oplus E^2_x = T_xM$ and $E^1_y \oplus E^2_y = T_yM$ that is $A(E^i_x) = E^i_y$ for $i = 1, 2$. We define four linear actions as:

$$a_{11}\colon E^1_x \to E^1_y,\ a_{12}\colon E^2_x \to E^1_y,\ a_{21}\colon E^1_x \to E^2_y \text{ and } a_{22}\colon E^2_x \to E^2_y,$$

and let $v = v_1 + v_2$ where $v_i \in E^i_x$ for $i = 1, 2$. Let

$$A \cdot v = (a_{11} + a_{21})v_1 + (a_{12} + a_{22})v_2. \tag{15.3}$$

The linear map A can be represented by the matrix

$$\widetilde{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \tag{15.4}$$

related to these previous splittings. We define a new norm by

$$\|A\|_m = \max\{|a_{11}|, |a_{12}|, |a_{21}|, |a_{22}|\},$$

and we call it the *norm of the maximum.*

Example 15.1. Let us consider a linear map in the plane represented by a conservative hyperbolic matrix (in the canonical base of $\mathbb{R}^2$),

$$A = \begin{pmatrix} 2 & 1,000 \\ 0 & \frac{1}{2} \end{pmatrix}.$$

This matrix has eigendirections associated to the vectors $b = \{(1,0), (-2,000,3)\}$ (associated to eigenvalues 2 and $1/2$). We observe that the angle θ between the eigendirections is close to zero. If we consider the diagonalized matrix with respect to the base of eigenvectors, then we get the matrix

$$\widetilde{A} = \begin{pmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{pmatrix}.$$

When we compute the norm of A, related to the usual metric in $\mathbb{R}^2$ we get $\|A\| = 1,000$, on the other hand the norm of the maximum of $\widetilde{A}$ is 2. In Lemma 15.1 we will obtain a relation between these quantities, namely that $\|A\| \leq 4\|A\|_m \sin^{-1}\theta$. In fact, in this example

$$\theta = \arccos\left(\frac{(1,0)\cdot(-2,000,3)}{\|(1,0)\|\|(-2,000,3)\|}\right) \approx 0.0015,$$

and we get an estimate since $1,000 \leq 8\sin^{-1}(0.0015) \approx 5334$.

Let us now show how we can relate the usual norm and the norm of the maximum.

Lemma 15.1. *Given $A \in \mathscr{L}(\mathbb{R}^2)$ as above, if $\angle(E^1_\sigma, E^2_\sigma) > \theta$ for $\sigma = x, y$, then A satisfies:*

1. $\|A\| \leq 4\sin^{-1}\theta\|A\|_m$.
2. $\|A\|_m \leq \sin^{-1}\theta\|A\|$.

Proof. We follow [10, Lemma 4.5]. Let $v = v_1 + v_2$ where $v_i \in E^i_x$ for $i = 1, 2$. Using elementary geometry it is easy to see that

$$\|v_i\| \leq \|v\|\sin^{-1}\theta, \text{ for } i = 1, 2.$$

Hence, using (15.3) and the preceding inequality

$$\begin{aligned} \|A\cdot v\| &\leq \|a_{11}v_1\| + \|a_{11}v_2\| + \|a_{22}v_1\| + \|a_{22}v_2\| \\ &= |a_{11}|\|v_1\| + |a_{11}|\|v_2\| + |a_{22}|\|v_1\| + |a_{22}|\|v_2\| \\ &\leq 4\|A\|_m\|v\|\sin^{-1}\theta. \end{aligned}$$

Therefore, by definition (15.1) we obtain (1).

Given $v_1 \in E_x^1$ using (15.3) we write $A \cdot v_1 = a_{11}v_1 + a_{21}v_1 \in E_y^1 \oplus E_y^2$ and so,

- $|a_{11}|\|v_1\| = \|a_{11}v_1\| \leq \|A \cdot v_1\| \sin^{-1}\theta \leq \|A\|\|v_1\| \sin^{-1}\theta$.
- $|a_{21}|\|v_1\| = \|a_{21}v_1\| \leq \|A \cdot v_1\| \sin^{-1}\theta \leq \|A\|\|v_1\| \sin^{-1}\theta$.

Analogously, given $v_2 \in E_x^2$ we write $A \cdot v_2 = a_{12}v_2 + a_{22}v_2 \in E_y^1 \oplus E_y^2$ and so, $|a_{12}|\|v_2\| \leq \|A\|\|v_2\| \sin^{-1}\theta$ and $|a_{22}|\|v_2\| \leq \|A\|\|v_2\| \sin^{-1}\theta$ and therefore (2) follows directly.

Finally, we present a simple lemma that will not be needed until Sect. 15.5.

Lemma 15.2. *([9, Lemma 3.9]) Given $\theta > 0$, there exists $c > 1$ such that for any linear map $A\colon \mathbb{R}^2 \to \mathbb{R}^2$ satisfying $\|A \cdot s\|.\|A \cdot u\|^{-1} > c$, where u, s are unit vectors, we can find a nonzero vector v such that $\angle(v,u) < \theta$ and $\angle(A \cdot v, A \cdot s) < \theta$.*

15.2.2.3 Orthogonal Decompositions

Sometimes we need to consider orthogonal decompositions in order to proceed with the estimates in a more treatable way. Consider the same map $A\colon T_xM \to T_yM$ as before and two new orthogonal decompositions $E_x^1 \oplus (E_x^1)^\perp$ and $E_y^1 \oplus (E_y^2)^\perp$ of T_xM and T_yM respectively. Denote by θ_x (respectively θ_y) the angle between E_x^1 and E_x^2 (respectively E_y^1 and E_y^2). Identify, using a rotation, the directions E_x^1 and E_y^1 with the direction $\mathbb{R}(1,0)$, the direction E_x^2 with $\mathbb{R}(\cos\theta_x, \sin\theta_x)$ and the direction E_y^2 with $\mathbb{R}(\cos\theta_y, \sin\theta_y)$. The $SL(2,\mathbb{R})$ matrix

$$\Psi_x := \begin{pmatrix} \sin^{-1}\theta_x & \cos\theta_x \\ 0 & \sin\theta_x \end{pmatrix},$$

maps E_x^1 into E_x^1 and $(E_x^1)^\perp$ into E_x^2, thus performs a conservative change from the decomposition $E_x^1 \oplus (E_x^1)^\perp$ into $E_x^1 \oplus E_x^2$. In the same way we define the matrix

$$\Psi_y := \begin{pmatrix} \sin^{-1}\theta_y & \cos\theta_y \\ 0 & \sin\theta_y \end{pmatrix},$$

mapping E_y^1 into E_y^1 and $(E_y^1)^\perp$ into E_y^2.

We now represent the linear action A in a new coordinate system by

$$A^\perp := \Psi_y^{-1} \circ A \circ \Psi_x. \tag{15.5}$$

We point out that every time we perform these change of coordinates we can keep track of the constants of estimation using the following inequality:

$$\|A^\perp\| \leq \frac{\|A\|}{(\sin\theta_x \sin\theta_y)}. \tag{15.6}$$

In conclusion, if the angle is bounded from bellow from zero, then it is possible to control the norm and thus to use this orthogonal splitting (see hypothesis (1) of Lemma 15.6).

15.2.3 Hyperbolicity and Dominated Splitting

Given a diffeomorphism f, a compact f-invariant set $\Lambda \subset M$ is said to be *hyperbolic* if there is $m \in \mathbb{N}$ such that, for every $x \in \Lambda$, there is a Df-invariant continuous splitting $T_xM = E^u_x \oplus E^s_x$ such that we have:

1. $\|Df^m_x|_{E^s_x}\| \leq \frac{1}{2}$ and
2. $\|(Df^m_x)^{-1}|_{E^u_x}\| \leq \frac{1}{2}$.

There are several ways to weaken the definition of uniform hyperbolicity. Here we use the one introduced independently by Mañé [18, 19], Liao [17] and Pliss [29] around the 1970s when motivated by the desire to prove the stability conjecture [27]. Given $m \in \mathbb{N}$, a compact f-invariant set $\Lambda \subset M$ is said to have an *m-dominated splitting* if there is, over Λ, a Df-invariant continuous splitting $TM = E^u \oplus E^s$ such that for all $x \in \Lambda$ we have:

$$\|Df^m_x|_{E^s_x}\|.\|Df^m_x|_{E^u_x}\|^{-1} \leq \frac{1}{2}. \tag{15.7}$$

It is worth pointing out that both subbundles may expand. However, E^u expands more than E^s. If both subbundles contract, E^u is less contracting than E^s. Like in the uniform hyperbolicity, the angle between the subbundles is uniformly bounded away from zero. This follows because the splitting is continuous and the base set is compact. Moreover, the dominated splitting extends to the closure of Λ. See [12] for the complete proofs of these properties.

Example 15.2. For $\mu > 1$ let us define

$$A := \begin{pmatrix} 1 & 0 \\ 0 & \mu \end{pmatrix} \text{ and } B := \begin{pmatrix} 1 & 0 \\ 0 & \mu^{-1} \end{pmatrix}.$$

The matrices A and B are not hyperbolic. However, A has an m-dominated splitting $E^u = \mathbb{R}(0,1)$ and $E^s = \mathbb{R}(1,0)$, and B has also an m-dominated splitting $E^u = \mathbb{R}(1,0)$ and $E^s = \mathbb{R}(0,1)$, where $m \geq \frac{\log 2}{\log \mu}$. It is immediate that μ close to 1 implies m very large.

Given $p \in Per(f)$, if p is hyperbolic and E^u_x and E^s_x are the Df-invariant subbundles, then the real numbers

- $\lambda_u(p) := \lim\limits_{n\to\pm\infty} \frac{1}{n} \log \|Df^n_p|_{E^u_x}\|$ and
- $\lambda_s(p) := \lim\limits_{n\to\pm\infty} \frac{1}{n} \log \|Df^n_p|_{E^s_x}\| < \lambda_u(p)$,

are called the *upper Lyapunov exponent* and the *lower Lyapunov exponent* respectively. By the celebrated Oseledet's theorem [25] (see [30] for a proof on dimension two) these numbers exist for Lebesgue almost every point in M and not necessarily a periodic point.

A central result about the Lyapunov exponents of C^1-generic conservative surface diffeomorphisms is the following result of Bochi based on a conjecture of Mañé.

Theorem 15.6. *(Bochi–Mañé [9, 20, 21]) There exists a C^1-generic subset $\mathscr{R}$ of $Diff^1_{\omega(M)}$ such that if $f \in \mathscr{R}$, then f is Anosov or else Lebesgue almost every point in M has zero Lyapunov exponents.*

As we will see, this result will play an important role in the proof of our results.

As a consequence of Oseledets' theorem we obtain the equality,

$$\lambda_u(p) + \lambda_s(p) = \lim_{n\to\pm\infty} \frac{1}{n} \log|\det Df^n_p|. \tag{15.8}$$

Then, by the area-preserving property, $|\det Df^n_p| = 1$ for every p, and so we obtain that $\lambda_u(p) = -\lambda_s(p)$. Therefore, if $\lambda_u(p) = 0$, then the spectrum of Df^τ_p lies in $\mathbb{S}^1$, where τ denotes the period of p. Otherwise, the real eigenvalues $\sigma^{\pm 1}$ of the map Df^τ_p, satisfy

$$e^{\lambda_u(p)\tau} = |\sigma| > 1 > |\sigma^{-1}| = e^{-\lambda_u(p)\tau}.$$

Let $Per_{\mathrm{hyp}}(f)$ denote the subset of all hyperbolic periodic points in $Per(f)$. Note that if $x \in Per_{\mathrm{hyp}}(f)$, then x has a dominated splitting, but in general we have that $m(x)$ is unbounded. Also, the weak hyperbolic behavior relates with the splitting angle being close to zero.

Since M is compact and the hyperbolic splitting varies continuously, given a uniformly hyperbolic invariant set $\Lambda \subset \overline{Per_{\mathrm{hyp}}(f)}$, the splitting angle between E^u and E^s, denoted by $\angle(E^u, E^s)$, is bounded away from zero over Λ.

Given $f \in \mathrm{Diff}^1_\omega(M)$, we define

$$\Delta_m(f) := \left\{ x \in Per_{\mathrm{hyp}}(f)\colon\ \|Df^m_x|_{E^s_x}\|.\|Df^m_x|_{E^u_x}\|^{-1} \geq \frac{1}{2} \right\},$$

and

$$\Lambda_m(f) := \left\{ x \in Per_{\mathrm{hyp}}(f)\colon \|Df^m_{f^n(x)}|_{E^s_x}\|.\|Df^m_{f^n(x)}|_{E^u_x}\|^{-1} \leq \frac{1}{2} \ \text{ for all } n \in \mathbb{N} \right\}.$$

Since $\overline{\Lambda_m(f)}$ has m-dominated splitting, and M is a surface, then, by the area-preserving property, $\overline{\Lambda_m(f)}$ is a hyperbolic set (see Lemma 15.3 below). Of course that we have

$$Per_{\text{hyp}}(f) = \Lambda_m(f) \dot{\bigcup} \left(\underset{n \in \mathbb{N}}{\cup} f^n(\Delta_m(f)) \right).$$

The following simple lemma, which only holds because M is a surface, with be useful in the sequel.

Lemma 15.3. *Let f be an area-preserving diffeomorphism and $\Lambda \subset M$ a compact f-invariant set. If Λ has a dominated splitting, then this splitting is hyperbolic.*

Proof. Since f admits a dominated splitting over Λ one gets that there exists $m \in \mathbb{N}$ such that

$$\Delta(x,m) := \|Df_x^m|_{E_x^s}\| \|Df_{f^m(x)}^{-m}|_{E^u_{f^m(x)}}\| \leq \frac{1}{2}, \ \forall x \in \Lambda,$$

where E^s and E^u are Df-invariant and one-dimensional.

For any $i \in \mathbb{N}$ we have $\Delta(x, im) \leq 1/2^i$. For every $n \in \mathbb{R}$ we may write $n = im + r$, for $0 \leq r < m$, and since $\|Df^r\|$ is bounded, say by L, take $C = 2^{\frac{r}{m}} L^2$ and $\sigma = 2^{-\frac{1}{m}}$ to get $\Delta(x,n) \leq C\sigma^n$, for every $x \in \Lambda$ and $n \in \mathbb{N}$.

Denote by α_n the angle between $E^s_{f^n(x)}$ e $E^u_{f^n(x)}$. We already know, by domination, that this angle is bounded bellow from zero, say by β.

Since f is area-preserving and the subbundles are both one-dimensional we have that

$$\sin \alpha_0 = \|Df_x^n|_{E_x^s}\| \|Df_x^n|_{E_x^u}\| \sin \alpha_n.$$

So

$$\|Df_x^n|_{E_x^s}\|^2 = \frac{\sin \alpha_0}{\sin \alpha_n} \Delta(x,n) \leq \Delta(x, im + r) \sin^{-1} \beta \leq \sigma^n C \sin^{-1} \beta.$$

Analogously we get

$$\|Df_x^{-n}|_{E_x^u}\|^2 = \frac{\sin \alpha_n}{\sin \alpha_0} \Delta(x,n) \leq \Delta(x, im + r) \sin^{-1} \beta \leq \sigma^t C \sin^{-1} \beta.$$

These two inequalities show that Λ is hyperbolic for Df completing the proof of the lemma.

15.3 Two Proofs of Theorem 15.1

First proof of Theorem 15.1: Assuming Theorem 15.2 we give now the proof of Theorem 15.1 and we postpone the proof of Theorem 15.2 to Sect. 15.6. Recall that $\mathscr{A}^2_\omega \subset \text{Diff}^1_\omega(M)$ denotes the open set of Anosov area-preserving diffeomorphisms and let $\overline{\mathscr{A}^2_\omega}$ be its C^1-closure. We define the open set $\mathscr{N} := \text{Diff}^1_\omega(M) \setminus \overline{\mathscr{A}^2_\omega}$. Consider the C^1-topology in $\text{Diff}^1_\omega(M)$, the topology inherited by the Riemannian metric in M, $dist(\cdot,\cdot)$, and the usual euclidean distance in $\mathbb{R}$. Let $\mathscr{H}$ be the subset of $\text{Diff}^1_\omega(M) \times M \times \mathbb{R}^+$ of all triples (f, x, ϵ) such that f has a closed elliptic orbit going through the ball $B(x,\epsilon) \subset M$. Finally, we endow $\mathscr{H}$ with the product topology. Since M is two-dimensional we get that the elliptic orbits are stable concluding that $\mathscr{H}$ is open.

Given any open set $\mathscr{U} \subseteq \mathscr{N}$ consider the following (also open) set

$$\mathscr{H}(\mathscr{U},x,\epsilon) := \{\, g \in \mathscr{U} \ : \ (g,x,\epsilon) \in \mathscr{H}\,\}.$$

It follows directly from Theorem 15.2 that if we take $\epsilon > 0$, $x \in M$ and an open set $\mathscr{U} \subseteq \mathscr{N}$, then $\mathscr{H}(\mathscr{U},x,\epsilon)$ is an open and dense subset of $\mathscr{U}$.

Using the smooth charts $\varphi_i\colon U_i \to \mathbb{R}^2$ for $i = 1,\dots,k$ we take k dense sequences in $\varphi_i(U_i) \subset \mathbb{R}^2$ and so, using φ_i^{-1}, we define $\{x_n\}_n$ to be a dense sequence in M. Let $\{\epsilon_n\}_n > 0$ be a sequence converging to zero. Defining recursively

$$\mathscr{U}_0 = \mathscr{N} \qquad \text{and} \qquad \mathscr{U}_{n+1} = \mathscr{H}(\mathscr{U}_n, x_n, \epsilon_n) \quad \text{for } n \geq 1,$$

the residual set $\mathscr{R} = \cap_{n=1}^{\infty}\mathscr{U}_n$ is such that for all $g \in \mathscr{R}$, the elliptic closed orbits of g are dense in M. Then $\mathfrak{R} = \mathscr{A}_\omega^2 \cup \mathscr{R}$ is the residual subset of $\text{Diff}_\omega^{\,1}(M)$, announced in Theorem 15.1.

Second proof of Theorem 15.1: We could also obtain another proof of Theorem 15.1 from Theorem 15.2 by using the elegant arguments explored in [24]. Denote by $\mathscr{E}_N(f)$ the set of elliptic periodic points (of the diffeomorphism f) of period less than N. Consider now the function

$$\begin{aligned} P_N : \text{Diff}_\omega^{\,1}(M) &\longrightarrow \quad \mathfrak{M} \\ f \quad &\longmapsto \mathscr{E}_N(f), \end{aligned}$$

where $\text{Diff}_\omega^{\,1}(M)$ is endowed with the C^1-topology and $\mathfrak{M}$ denotes the set of all closed subsets of M endowed with the Hausdorff metric. By the stability of the elliptic periodic points it follows that P_N is a continuous function. Hence we obtain that $P = \sup_{N\in\mathbb{N}}\{P_N\}$ is a lower semi-continuous function (see [16]). Actually, we have

$$\begin{aligned} P : \text{Diff}_\omega^{\,1}(M) &\longrightarrow \quad \mathfrak{M} \\ f \quad &\longmapsto \overline{\mathscr{E}(f)}, \end{aligned}$$

where $\overline{\mathscr{E}(f)}$ denotes the closure of the set of the elliptic periodic points of f.

Using [32, Proposition 26] we obtain that there exists a residual $\mathscr{R} \subset \text{Diff}_\omega^{\,1}(M)$ formed by continuity points of P.

Therefore, if $f \in \mathscr{R}$ is not Anosov, it is an immediate consequence of Theorem 15.2 that the elliptic points are dense in M and Theorem 15.1 is proved.

15.4 Perturbation Lemmas

In order to achieve our goal we will need to perform some perturbations of the tangent map. One of the main perturbation tools will induce rotations in the tangent bundle and so the next basic lemma will be very useful. We emphasize that a more or less general result will be stated (see Theorem 15.7 and Remark 15.3), however the advantage of presenting the proof of Lemma 15.4 lies in the fact that it sheds some light in the nice properties of the C^1 topology and for this reason we decide to state and prove it nevertheless.

Lemma 15.4. *If $f \in \mathit{Diff}^1_\omega(M)$ and $\epsilon > 0$, there exists $\beta_0 > 0$ such that for any $x \in M$, $r \in (0,1)$ and $\beta < \beta_0$ there exists $g \in \mathscr{N}^\omega_\epsilon(f)$ such that (in local charts):*

(a) $Dg_x = Df_x \cdot R_\beta(x)$, where $R_\beta(x)$ denotes the rotation of angle β centered in x, and
(b) $g = f$ outside the ball $B(x,r)$.

Proof. We first prove the result for $r = 1$. Using the aforementioned charts (see Sect. 15.2.1) we assume $x = \mathbf{0}$. Let $\alpha\colon [0,\infty) \to [0,\infty)$ be a C^∞ function such that $\alpha(t) = 0$ for $t \geq 1$, $\alpha(t) = 1$ for $0 \leq t \leq 1/2$ and $|\alpha'| \leq 2$. Take $y \in B(0,1)$ and $\beta > 0$. Let $h_\beta(y) = R_{\alpha(\|y\|)\beta}(\mathbf{0})$. We define $g_\beta = f \circ h_\beta$. Computing the derivative of g_β at y we obtain that $(Dg_\beta)_y = Df_{h_\beta(y)} \cdot (Dh_\beta)_y$. Therefore:

- If $y \in B(x,1/2)$, then $(Dg_\beta)_y = Df_{h_\beta(y)} \cdot R_\beta(x)$.
  In particular $(Dg_\beta)_x = Df_{h_\beta(x)} \cdot R_\beta(x) = Df_x \cdot R_\beta(x)$ which gives item (a).
- If y lies outside $B(x,1)$, then $g_\beta = f$ and we get item (b).

Since $\det(Dg_\beta)_x = 1$ for all x, our final goal is to to prove that g_β is ϵ-C^1-close to f. The C^0-closeness is obvious. Let us prove that $(Dh_\beta)_y$ is C^0-close to the identity. In local coordinates we can write:

$$h_\beta(y) = (\cos(\phi_y)y_1 - \sin(\phi_y))y_2, \sin(\phi_y)y_1 + \cos(\phi_y)y_2),$$

where $\phi_y = \alpha(\|y\|)\beta$ and $y = (y_1, y_2)$. Taking derivatives we obtain:

$$(Dh_\beta)_y = A_y + \begin{pmatrix} \cos(\phi_y) & -\sin\phi_y \\ \sin\phi_y & \cos(\phi_y) \end{pmatrix},$$

where

$$A_y = \begin{pmatrix} -\frac{\partial\phi_y}{\partial y_1}\sin(\phi_y)y_1 - \frac{\partial\phi_y}{\partial y_1}\cos(\phi_y)y_2 & -\frac{\partial\phi_y}{\partial y_2}\sin(\phi_y)y_1 - \frac{\partial\phi_y}{\partial y_2}\cos(\phi_y)y_2 \\ \frac{\partial\phi_y}{\partial y_1}\cos(\phi_y)y_1 - \frac{\partial\phi_y}{\partial y_1}\sin(\phi_y)y_2 & \frac{\partial\phi_y}{\partial y_2}\cos(\phi_y)y_1 - \frac{\partial\phi_y}{\partial y_2}\sin(\phi_y)y_2 \end{pmatrix}.$$

It is clear that, if β is chosen to be small, then $(Dh_\beta)_y - A_y$ is arbitrarily close to the identity.

We just have to prove that, for a suitable β, A_y is close to the null matrix. For that we first compute the gradient of ϕ_y.

$$\nabla\phi_y = \left(\frac{\partial\phi_y}{\partial y_1}, \frac{\partial\phi_y}{\partial y_2}\right) = \alpha'(\|y\|)\beta\|y\|^{-1}(y_1, y_2).$$

Recall that $|\alpha'| \leq 2$, $\left|\frac{\partial \phi_y}{\partial y_i} y_i\right| = |\alpha'(\|y\|)\beta\|y\|^{-1} y_i|$ and $\|y\|^{-1} y_i \leq 1$. Hence, we obtain that $\left|\frac{\partial \phi_y}{\partial y_i}\right| \leq 2\beta$.

Therefore, given $\epsilon > 0$, there exists $\beta_0 > 0$ and $g \in \mathrm{Diff}^1_\omega(M)$ such that $g \in \mathcal{N}^\omega_\epsilon(f)$ and $g = g_{\beta_0}$ satisfies (a) and (b). Finally, for $r \in (0,1]$, we consider the r-homothethy and $h = h_{\beta_0}$ associated to $r = 1$ established above. We define the new h_r and g_r (associated to r) by $rh(y/r)$ and $f \circ h_r$ respectively. Clearly $D(rh(y/r)) = Dh(y/r)$ which is C^0-close to the identity and the lemma is proved.

Remark 15.1. A slight change in the proof of Lemma 15.4 allow us to obtain a version where, in (a), we switch from $Dg_x = Df_x \cdot R_\beta(x)$ to $Dg_x = R_\beta(x) \cdot Df_x$. The details are left to the reader.

In [2] it was proved a weak pasting lemma for diffeomorphisms which, in rough terms, allow us to replace the area-preserving diffeomorphism f by another area-preserving diffeomorphism g such that g is equal to the first order linear approximation of f in a small neighborhood $\mathcal{U}$ of a given point, and equal to f outside a set containing $\mathcal{U}$. Let us present the formal statement.

Theorem 15.7. *([2, Theorem 3.6]) If $f \in \mathit{Diff}^2_\omega(M)$ and $x \in M$, then for any $0 < \alpha < 1$ and $\epsilon > 0$, there exists $\tilde{\epsilon} > 0$ such that any $A_x \in SL(2,\mathbb{R})$ which is $\tilde{\epsilon}$-close to Df_x satisfies the following; there exists $g \in \mathcal{N}^\omega_\epsilon(f)$ of class $C^{1+\alpha}$ such that for small neighborhoods $U \supset V$ of x we have, in local charts, that:*

- *$g|_V = A_x$ and*
- *$g = f$ outside the set U.*

Actually, in [2] they proved that $g|_V = Df_x$ by constructing a perturbation $h(y) = \rho(y)(f(x) + Df_x(y - x)) + (1 - \rho(y))f(y)$ where ρ is a *bell-function* over the annulus $B(x,r) \setminus B(x,r/2)$. Then they make use of a cleaver application of a theorem of Dacorogna and Moser [13] to obtain a new $\tilde{h}$ which will be area-preserving. Theorem 15.7 is obtained in the same way by switching Df_x by A_x.

Remark 15.2. We can take, in Theorem 15.7, $A_x = Df_x \cdot S_x$, where S_x is $\frac{\tilde{\epsilon}}{C}$-close to the identity, where $C := \max_{x\in M} \|Df_x\|$.

Remark 15.3. A similar version of Lemma 15.4 can be obtained directly from Theorem 15.7 if we take f of class C^2.

Finally, we recall the conservative C^1-closing lemma of Arnaud (see [4]), which in particular assures properties (a) and (b) bellow. This result is an upgrade of the C^1-closing lemma [31] and Mañé's ergodic closing lemma [19] and states that the orbit of a f-recurrent point x (denoted by $R(f)$) can be approximated for a very long time $\pi > 0$ by a periodic orbit of an area-preserving diffeomorphism $g \in \mathcal{N}^\omega_\epsilon(f)$.

Let $f \in \mathrm{Diff}^1_\omega(M)$ and let $\Sigma(f)$ be the set of points $x \in M$ such that for any C^1-neighborhood $\mathscr{U} \subset \mathrm{Diff}^1_\omega(M)$ of f and every $\epsilon > 0$, there exists $g \in \mathscr{U}$ and a periodic point of period π, $p \in M$, such that:

(a) $dist\big(f^i(x), g^i(p)\big) < \epsilon$ for all $i \in \{0, \dots, \pi\}$.
(b) $g = f$ in $M \setminus \bigcup_{0 \le i \le \pi} B(f^i(x), \epsilon)$.

Theorem 15.8. *(Arnaud's C^1-closing lemma [4]) The set $\Sigma(f)$ is a countable intersection of open subsets of the set $R(f)$ and is dense in this set.*

15.5 Creating Elliptic Points

15.5.1 Mixing the Eigendirections-Part I

We start by proving the following result.

Lemma 15.5. *Given a hyperbolic matrix $A \in SL(2, \mathbb{R})$, let $\theta = \angle(E^s, E^u)$ be the angle between the matrix A eigendirections. Assume that the rotation R_θ of angle θ takes the unstable direction onto the stable direction of A, i.e., $R_\theta(E^u) = E^s$. Then the matrix $A \cdot R_\theta$ is elliptic.*

Proof. Let $B := A \cdot R_\theta$. Consider the action of the matrices A and B on the projective line $\mathbb{P}^1 = \mathbb{R}/\pi\,\mathbb{Z}$, described by the diffeomorphisms $f_A : \mathbb{P}^1 \to \mathbb{P}^1$ and $f_B : \mathbb{P}^1 \to \mathbb{P}^1$. Lift these maps to diffeomorphisms $F_A : \mathbb{R} \to \mathbb{R}$ and $F_B : \mathbb{R} \to \mathbb{R}$ such that $F_A(x + \pi) = F_A(x) + \pi$ and $F_B(x + \pi) = F_B(x) + \pi$, for all $x \in \mathbb{R}$. As $\det(A) = \det(B) = 1$ we get that F_A and F_B are increasing functions. The definition of θ shows that the lifting F_B can be chosen to satisfy the relation $F_B(x) = F_A(x + \theta)$, for all $x \in \mathbb{R}$. Since A is hyperbolic, f_A has two fixed points: an expanding fixed point x^u, and a contracting fixed point x^s. We can choose the lifting F_A so that it has two families of fixed points, $x^u + k\,\pi$ and $x^s + k\,\pi$, with $k \in \mathbb{Z}$, and we may assume that the fixed points x^s, $x^u \in \mathbb{R}$ satisfy $|x^s - x^u| = \theta$. In order to prove that B is elliptic it is enough to show that f_B has non zero rotation number, which amounts to say that $F_B(x) - x$ keeps a constant sign as x runs through $\mathbb{R}$. Two cases may occur: $x^s < x^u$ and $x^u < x^s$. Assume first that $x^s < x^u$. Then $-\theta < F_A(x) - x < 0$ for all $x \in]x^s, x^u[$, and $F_A(x) - x > 0$ for all $x \in]x^u, x^s + \pi[$. This implies that $F_A(x) - x > -\theta$, for all $x \in \mathbb{R}$. Therefore, $F_B(x) - x = F_A(x + \theta) - (x + \theta) + \theta > -\theta + \theta = 0$, for every $x \in \mathbb{R}$, proving that B is elliptic. Assume now that $x^u < x^s$. In this case $0 < F_A(x) - x < \theta$ for all $x \in]x^u, x^s[$, and $F_A(x) - x < 0$ for every $x \in]x^s, x^u + \pi[$. But this implies that $F_A(x) - x < \theta$, for all $x \in \mathbb{R}$. Accordingly, $F_B(x) - x = F_A(x + \theta) - (x + \theta) + \theta < -\theta + \theta = 0$, for every $x \in \mathbb{R}$, which proves that B is elliptic.

We easily deduce the following result from Lemmas 15.5 and 15.4.

Proposition 15.1. *Given $\epsilon > 0$ and $f \in \mathit{Diff}^1_\omega(M)$, there exists $\theta > 0$ such that given any $x \in Per_{hyp}(f)$ with period $\tau > 1$, and such that $\angle(E^u_y, E^s_y) < \theta$ for some y in the f-orbit of x, then there is some perturbation $g \in \mathcal{N}^\omega_\epsilon(f)$ such that x is an elliptic periodic point for g with period τ.*

As we saw in the preceding proposition mixing eigendirections by rotations reveals to be useful to create elliptic periodic orbits for maps near the original one. However, we are only allowed to perform a small perturbation and this can be difficult, or maybe impossible, if the angle between eigendirections is far from zero. In the next lemma we assume some hypotheses under which it will be possible to achieve that objective and, its proof, although easier, follows closely the one in [9, Lemma 3.8].

Lemma 15.6. *Given $f \in \mathit{Diff}^r_\omega(M)$, $r \geq 1$ and $\epsilon > 0$ let $\theta(f, \epsilon) = \theta > 0$ be given by Lemma 15.4 (with $\theta < \beta_0$). There is $m_0 \in \mathbb{N}$ such that for every $m \geq m_0$, if $x \in Per_{hyp}(f)$ has period $\tau > m$ and satisfies*

1. *$\angle(E^u_{f^n(x)}, E^s_{f^n(x)}) > \theta$, for all $n \in \{1, \dots, \tau\}$ and*
2. *we have $f^n(x) \in \Delta_m(f)$ for some $n \in \{1, \dots, \tau\}$,*

then there exist a C^r conservative map $g \in \mathcal{N}^\omega_\epsilon(f)$ and $y = f^k(x)$ ($k \in \{1, \dots, \tau\}$) such that $Dg^m_x(E^u_y) = E^s_{f^m(y)}$.

Proof. Let $C := \max\limits_{x \in M} \|Df_x\|$ and $c > C^2$ depending on the angle θ and obtained according to Lemma 15.2.

Let $x \in Per_{\text{hyp}}(f)$ with period $\tau > m > m_0$ and satisfying (1) and (2). The number m_0 will be very large and will be defined below. By (2) there exists y in the f-orbit of x such that $y \in \Delta_m(f)$, i.e.,

$$\|Df^m_y|_{E^s_y}\|.\|Df^m_y|_{E^u_y}\|^{-1} \geq 1/2. \tag{15.9}$$

$\boxed{\text{Case I}}$

Suppose that for any $i, j \in \{0, 1, \dots, m\}$, where $i < j$, we have

$$\|Df^{j-i}_{f^i(y)}|_{E^s_y}\|.\|Df^{j-i}_{f^i(y)}|_{E^u_y}\|^{-1} \leq c. \tag{15.10}$$

Noting that $E^{(\cdot)}_y$ (for $(\cdot) = u/s$) are one-dimensional and using (15.9) and (15.10) we get

$$\frac{\|Df^{j-i}_{f^i(y)}|_{E^s_y}\|}{\|Df^{j-i}_{f^i(y)}|_{E^u_y}\|} = \frac{\|Df^{m-j}_{f^i(y)}|_{E^u_y}\|.\|Df^m_y|_{E^s_y}\|.\|Df^i_y|_{E^u_y}\|}{\|Df^{m-j}_{f^i(y)}|_{E^s_y}\|.\|Df^m_y|_{E^u_y}\|.\|Df^i_y|_{E^s_y}\|} \geq \frac{1}{2c^2}. \tag{15.11}$$

Using (15.10) again we obtain, for $H := 2c^2$, that

$$\frac{1}{H} \leq \frac{\|Df^{j-i}_{f^i(y)}|_{E^s_y}\|}{\|Df^{j-i}_{f^i(y)}|_{E^u_y}\|} \leq H. \tag{15.12}$$

Using (1) we can make a conservative change of coordinates as it was explained in Sect. 15.2.2.3 keeping the control on the estimated (depending on $\sin^2\theta$). Hence, by conservativeness, for any $j \in \{0, 1, \dots, m\}$, we have $\|Df^j_y|_{E^s_y}\|.\|Df^j_y|_{E^u_y}\| = \det Df^j_y = 1$. Therefore, using (15.12) we get that $\|Df^j_y|_{E^{(\cdot)}_y}\| \leq 2H = 4c^2$ for $(\cdot) = u/s$ and every j. This implies that for every $j \in \{0, 1, \dots, m\}$ we have $\|Df^j_y\| \leq 2H$.

For some $\gamma > 0$ very small, let $\{\theta_j\}_{j=0}^{m-1}$ be such that $0 < \theta_j \leq \gamma$ (for all j) and $\angle(E^s_y, E^u_y) = \sum_{j=0}^{m-1}\theta_j$. We define, for every $j = \{0, 1, \dots, m-1\}$, linear maps $S_j \colon T_{f^j(y)}M \to T_{f^{j+1}(y)}M$ by $S_j := Df^{j+1}_y \cdot R_{\theta_j} \cdot (Df^j_y)^{-1}$. It is straightforward to see that

$$S_{m-1} \cdot S_{m-2} \cdot \ldots \cdot S_1 \cdot S_0(E^u_y) = Df^m_y \cdot R_{\angle(E^s_y, E^u_y)}(E^u_y) = E^s_{f^m(y)}.$$

Using Theorem 15.7 we realize[2] these perturbation by m conservative maps g_j in m small self-disjoint balls $B_j := B(f^j(y), r_i)$, $r_i > 0$. Then we define a conservative map g by being equal to g_i in B_i and equal to f outside the union of these balls.

Observe that, since H is fixed, $\|S_i - Id\|$ is small as long as θ_i is close to zero which is equivalent to take γ very small.

We leave it to the reader to verify that, since we have a control on the norm of Df^j_y, g can be chosen ϵ-close to f and we just have to take m_0 be any positive integer such that $m_0 \geq \frac{2\pi}{\gamma}$.

$\boxed{\text{Case II}}$

We now turn to the case where (15.10) is false, i.e., there exists $i, j \in \{0, 1, \dots, m\}$, where $i < j$, such that

$$\|Df^{j-i}_{f^i(y)}|_{E^s_y}\|.\|Df^{j-i}_{f^i(y)}|_{E^u_y}\|^{-1} > c. \tag{15.13}$$

It is understood that $j - i > 1$ because $c > C^2$. Take unit vectors $s \in E^s_{f^i(y)}$ and $u \in E^u_{f^i(y)}$. By (15.13) we are in the hypotheses of Lemma 15.2 for the linear map $Df^{j-i}_{f^i(y)}$, therefore we can find a nonzero vector $v \in T_{f^i(y)}M$ such that $\angle(v, u) < \theta$ and $\angle(Df^{j-i}_{f^i(y)} \cdot v, E^s_{f^j(y)}) < \theta$. By making two perturbations at $f^i(y)$ (using Lemma 15.4) and at $f^{j-1}(y)$ (using Remark 15.1) we can obtain $g \in \mathscr{N}^\omega_\epsilon(f)$,

[2] In order to use Theorem 15.7 f must be of class C^2. The important point to note here is that we can perturb slightly, using [38], and obtain a C^2 conservative map having the same properties (1) and (2) of Lemma 15.6 for the analytic continuation of the hyperbolic point x.

such that:

$$Dg_{f^i(y)} := Df_{f^i(y)} \cdot R_{\angle(v, E^u_{f^i(y)})} \text{ and}$$
$$Dg_{f^{j-1}(y)} := R_{\angle(Df^{j-i}_{f^i(y)} \cdot v, E^s_{f^j(y)})} \cdot Df_{f^{j-1}(y)}.$$

Moreover, $g = f$ outside two small balls around $f^i(y)$ and $f^j(y)$. Is is easy to verify that by concatenating the tangent maps of g along $\{f^n(y)\}_{n=i}^j$ we complete the proof.

15.5.2 Mixing the Eigendirections-Part II

Our purpose now is to prove the next proposition and its proof will be divided into two main steps; Lemma 15.6 above and Lemma 15.7 below.

Proposition 15.2. *Given $f \in Diff^1_\omega(M)$, $\epsilon > 0$ and $\theta > 0$, there exist $m \in \mathbb{N}$ and $T \in \mathbb{N}$ ($T > m$) such that given a periodic hyperbolic point $x \in M$ with period $\tau > T$, satisfying the conditions (1) and (2) of Lemma 15.6, then there is some perturbation $g \in \mathcal{N}^\omega_\epsilon(f)$ such that x is an elliptic periodic point for g with period τ.*

The following result allows us, once in the hypotheses of Proposition 15.2, to obtain some control on the growth of the norm of Dg^τ for a large τ, where $g \in \mathcal{N}^\omega_\epsilon(f)$.

Lemma 15.7. *Let $f \in Diff^1_\omega(M)$, $\epsilon > 0$ and $\theta > 0$ be given. Let $m = m(\epsilon, \theta) \in \mathbb{N}$ be given by Lemma 15.6. Then there exists $K = K(\theta, m) \in \mathbb{R}$ such that given any hyperbolic periodic point x with period $\tau > m$ satisfying (1) and (2) of Lemma 15.6, then there exists $g \in \mathcal{N}^\omega_\epsilon(f)$ such that x is also a periodic orbit for g with period τ and $\|Dg^\tau_y\| < K$, for some y in the g-orbit of x.*

Proof. For $f \in \mathrm{Diff}^1_\omega(M)$ and $\epsilon > 0$ given, there exists $C > 1$ such that, if $g \in \mathcal{N}^\omega_\epsilon(f)$ then $\|Dg\| \leq C$. We define

$$K(m(\theta)) := 4C^{m+2} \sin^{-2} \theta. \tag{15.14}$$

Take any hyperbolic periodic point x with period $\tau > m$. Let $g \in \mathrm{Diff}^1_\omega(M)$ be the perturbation provided by Lemma 15.6, corresponding to the same ϵ and θ of this lemma. We assume that the point y given in Lemma 15.6 is $y = x$. According to Sect. 15.2.2 we take matrix representations diagonalizing the hyperbolic decomposition and along the orbit.

Given $k \in \{1, \ldots, \tau - m\}$ let $y = f^{-k}(x)$. Take a finite sequence $\{F(y,i)\}_{i=1}^k \subset \mathbb{R}$ such that the matrix Df^i_y written in the diagonal[3] form associated

[3] In fact, we are abusing the notation since we should denote this representation by $\widetilde{Df^i_y}$ instead of Df^i_y.

to the eigendirections is,

$$Df_y^i = \begin{pmatrix} F(y,i) & 0 \\ 0 & F(y,i)^{-1} \end{pmatrix},$$

and let $\sigma = F(y,\tau) > 1$. Observe that by Lemma 15.1 (2), for $i \in \{1,\dots,\tau\}$, we have

$$\begin{aligned} \max\{|F(y,i)|, |F(y,i)|^{-1}\} = \|Df_y^i\|_m &\leq \|Df_y^i\| \sin^{-1}\theta \\ &\leq \prod_{j=0}^{i-1} \|Df_{f^j(y)}\| \sin^{-1}\theta \leq C^i \sin^{-1}\theta. \end{aligned}$$

We will consider two cases:

$\boxed{\text{Case I}}$ If $\sigma \leq C^{m+1}\sin^{-1}\theta$ then observing that $\|Df_y^\tau\|_m = \sigma$ and applying Lemma 15.1 (1), we obtain

$$\|Df_y^\tau\| \leq 4\sin^{-1}\theta \|Df_y^\tau\|_m = 4\sigma\sin^{-1}\theta \leq 4C^{m+1}\sin^{-2}\theta \leq K,$$

and the lemma is proved by just choosing $g = f$.

$\boxed{\text{Case II}}$ On the other hand, if $\sigma > C^{m+1}\sin^{-1}\theta$, we will use the following calculus lemma whose proof we postpone to the end of the proof of Lemma 15.7.

Lemma 15.8. *Given $\tau, m \in \mathbb{N}$, $\tau > m$, $C > 1$ and $\{a_i\}_{i=1}^\tau$ such that $|a_i|^{\pm 1} < C$ we define $\sigma := |\Pi_{i=1}^\tau a_i|$. If $\sigma > C^{m+1}$, then there exists $k \in \{1,\dots,\tau-m\}$ such that*

$$\left|\frac{\Pi_{i=k+m}^\tau a_i}{\Pi_{i=1}^k a_i}\right|^{\pm 1} \leq C^2.$$

We feed Lemma 15.8 with $a_i = F(y,i)$ and let $k \in \{1,\dots,\tau-m\}$ be given by this lemma. Take $y = f^{-k}(x)$. Since

$$Df_y^\tau = Df_{f^m(x)}^{\tau-m-k} \cdot Df_x^m \cdot Df_y^k$$

we may write Df_y^τ as the following diagonal matrix product representation

$$\begin{pmatrix} F(f^m(x),\tau-m-k) & 0 \\ 0 & \frac{1}{F(f^m(x),\tau-m-k)} \end{pmatrix} \begin{pmatrix} F(x,m) & 0 \\ 0 & \frac{1}{F(x,m)} \end{pmatrix} \begin{pmatrix} F(y,k) & 0 \\ 0 & \frac{1}{F(y,k)} \end{pmatrix}. \tag{15.15}$$

Recall that g, given by Lemma 15.6, is a conservative perturbation of f, supported in a small neighborhood of $\{f^i(x): i \in \{0,\dots,m\}\}$, and such that $Dg_x^m(E_x^u) = E_{f^m(x)}^s$. Taking in account the notation of Sect. 15.2.2.2 we get that $\xi: E_x^u \to E_{f^m(x)}^u$ must be the null map, where

$$Dg_x^m := \begin{pmatrix} \xi & \alpha \\ \beta & \gamma \end{pmatrix}, \tag{15.16}$$

for some constants ξ, α, β and γ. That is, the unstable component of the image by Dg_x^m of E_x^u must be zero and so $\xi = 0$.

Now, one just replaces the middle matrix in (15.15) and we obtain that

$$Dg_y^\tau = Df_{f^m(x)}^{\tau-s-m} \cdot Dg_x^m \cdot Df_y^k,$$

is given by

$$Dg_y^\tau = \begin{pmatrix} 0 & \alpha \frac{F(f^m(x),\tau-m-k)}{F(y,k)} \\ \beta \frac{F(y,k)}{F(f^m(x),\tau-m-k)} & \gamma \frac{1}{F(y,k)F(f^m(x),\tau-m-k)} \end{pmatrix}.$$

Notice that,

$$\begin{aligned} \frac{1}{F(y,k)F(f^m(x),\tau-m-k)} &= \frac{F(x,m)}{\sigma} \le \frac{\|Df_x^m\|_m}{\sigma} \le \frac{\|Df_x^m\|}{\sigma \sin\theta} \\ &\le \frac{C^m}{\sigma \sin\theta} < \frac{1}{C}. \end{aligned}$$

Moreover, by Lemma 15.1 (2)

$$\max\{|\alpha|, |\beta|, |\gamma|\} = \|Dg_x^m\|_m \le \sin^{-1}\theta \|Dg_x^m\| \le C^m \sin^{-1}\theta.$$

Using Lemma 15.8 we get $\|Dg_y^\tau\|_m \le \max\{|\alpha|, |\beta|, |\gamma|\}C^2 \le C^{m+2}\sin^{-1}\theta$. Finally, using Lemma 15.1 (1) we get

$$\|Dg_y^\tau\| < 4\sin^{-1}\theta \|Dg_y^\tau\|_m < 4C^{m+2}\sin^{-2}\theta = K,$$

and the lemma is proved.

Proof. (of Lemma 15.8) For $k = 1$ since we have $\sigma > C^{m+1}$ we obtain

$$\left| \frac{\Pi_{i=m+1}^\tau a_i}{a_1} \right| = \frac{\sigma}{|a_1 \Pi_{i=1}^m a_i|} \ge \frac{\sigma}{C^{m+1}} > 1.$$

For $k = \tau - m$ we have

$$\left| \frac{a_\tau}{\Pi_{i=1}^{\tau-m} a_i} \right| = \frac{|a_\tau \Pi_{i=\tau-m+1}^\tau a_i|}{\sigma} \le \frac{C^{m+1}}{\sigma} < 1.$$

Let

$$\Phi(k) = \left| \frac{\Pi_{i=k+m}^\tau a_i}{\Pi_{i=1}^k a_i} \right|.$$

We chose $k \in \{1, \dots, \tau - m - 1\}$ such that $\Phi(k) > 1$ and $\Phi(k+1) < 1$. Since $\Phi(k)^{-1} < 1 < C^2$ we are left to the task of proving that $\Phi(k) \leq C^2$.

$$\Phi(k) = \left| \frac{\Pi_{i=k+m}^{\tau} a_i}{\Pi_{i=1}^{k} a_i} \right| = \Phi(k+1)|a_{k+1}||a_{k+m}| \leq C^2.$$

Remark 15.4. The important thing to note here is that Lemma 15.7 allows us to fix a uniform bound K such that we can pick a periodic hyperbolic point with very large period and, nevertheless, the tangent map (on the period) is bounded by K for a C^1-arbitrarily close conservative map.

Proof. (of Proposition 15.2) We know that for any diffeomorphism f_1 C^1-close to f any hyperbolic periodic point x of f has an analytic continuation y for the diffeomorphism f_1 (see e.g. [35]). Moreover, by [38], $\text{Diff}_{\omega}^2(M)$ is C^1-dense in $\text{Diff}_{\omega}^1(M)$. Hence, for a diffeomorphism $f_1 \in \text{Diff}_{\omega}^2(M)$ arbitrarily C^1-close to f, by Lemma 15.6, we take $m_0(f_1)$ (larger than $m_0(f)$ if necessary) such that, if y is a hyperbolic periodic point of period $\tau > m$ for any $m \geq m_0(f_1)$ satisfying

1. $\angle(E^u_{f_1^n(y)}, E^s_{f_1^n(y)}) \geq \theta$ for all $n \in \{1, \dots, \tau\}$ and
2. $f_1^n(y) \in \Delta_m(f_1)$ for some $n \in \{1, \dots, \tau\}$,

then there exist $f_2 \in \text{Diff}_{\omega}^2(M) \cap \mathcal{N}_{\epsilon}^{\omega}(f)$ and $z = f_1^k(y)$, for $k \in \{1, \dots, \tau\}$, such that $(Df_2{}^m)_y(E_z^u) = E^s_{f_1^m(z)}$.

Fix $f_2 \in \text{Diff}_{\omega}^2(M)$ and any $x \in M$. By Theorem 15.7 followed by Remark 15.2, for $\epsilon > 0$, there exists $\zeta_0 > 0$ such that any $S_x \in SL(2, \mathbb{R})$ which is ζ-close to the identity (with $\zeta < \zeta_0$) satisfies the following; there exists $g \in \mathcal{N}_{\epsilon}^{\omega}(f_2)$ such that for small neighborhoods $U \supset V$ of x we have, in local charts, that:

- $g|_V = (Df_2)_x \cdot S_x$ and
- $g = f_2$ outside the set U.

Take $K := K(m(\theta))$ according to Lemma 15.7 and depending on $f_1 \in \text{Diff}_{\omega}^2(M)$, on ϵ, $m_0(f_1)$ and on θ. Now, for ζ_0 and θ fixed above, set $\sigma := (\eta_\theta)^{-1}(\zeta_0)$, where $\eta_\theta(\zeta_0)$ was defined in (15.2). By definition, the number $\sigma > 1$ has the following property: Given any $\varphi \geq \theta$, we can pick hyperbolic matrices $S \in SL(2, \mathbb{R})$ such that:

(a) $\|S - Id\| \leq \zeta_0$.
(b) σ and σ^{-1} are the eigenvalues of S.
(c) S has an angle φ between its eigenspaces.

Finally, we take $T \in \mathbb{N}$ such that $\sigma^T \geq K$. Now, let $\Gamma = \{f_1^n(x) : n \in \{1, \dots, \tau\}\}$ be any hyperbolic periodic orbit, with period $\tau > T$, satisfying (1) and (2) of Lemma 15.6. Let $g \in \mathcal{N}_{\epsilon}^{\omega}(f_1)$ be the diffeomorphism provided by Lemma 15.7 satisfying $\|Dg_y^{\tau}\| < K$ for some point $y \in \Gamma$. We take $i \in \{0, \dots, \tau - 1\}$ and we define $x_i := f_1^i(y)$ and $x_i' := f_1(x_i) = f_1^{i+1}(y)$. Take the linear maps $Dg_{f_1^i(y)} := Dg_i : \mathbb{R}^2_{x_i} \to \mathbb{R}^2_{x_i'}$ and let $\theta_i \geq \theta$ be the angle between the eigenspaces

$E^u_{x_i}$ and $E^s_{x_i}$ of the map Dg_i^τ, for each $i \in \{0, \dots, \tau - 1\}$. Take now $S_i \in SL(2, \mathbb{R})$ such that $\|S_i - I\| \leq \zeta_0$, and S_i has eigenspace $E^u_{x_i}$ with eigenvalue σ^{-1}, and has eigenspace $E^s_{x_i}$ with eigenvalue σ. Observe that these eigenspaces do make an angle equal to θ_i. The product linear map $(Dg_i \cdot S_i)\colon \mathbb{R}^2_{x_i} \to \mathbb{R}^2_{x_i'}$ takes the decomposition $\mathbb{R}^2_{x_i} = E^u_{x_i} \oplus E^s_{x_i}$ onto the decomposition $\mathbb{R}^2_{x_i'} = E^u_{x_i'} \oplus E^s_{x_i'}$. Moreover, we have $\|Dg_i \cdot S_i|_{E^u_{x_i}}\| = \|Dg_i|_{E^u_{x_i}}\|\sigma^{-1}$ and $\|Dg_i \cdot S_i|_{E^s_{x_i}}\| = \|Dg_i|_{E^s_{x_i}}\|\sigma$.

Consider a family of smooth deformations of the identity into S_i, that is, let $\{S_{i,t}\}_{i=0,t\in[0,1]}^{\tau-1}$ be defined analogously to S_i but with eigenvalues σ^t and σ^{-t}, where for $t = 0$ we get the identity and for $t = 1$ we get S_i.

By a direct application of Theorem 15.7 we can obtain a family of C^1 area-preserving diffeomorphisms $(h_i)_t$ such that $(h_i)_t \in \mathcal{N}_\epsilon^\omega(g)$, $g = (h_i)_t$ outside a small neighborhood of the point x_i, and $[D(h_i)_t]_{x_i} = Dg_i \cdot S_{i,t}$. But, since we can produce these perturbations with self-disjoint support, we can glue them into a single conservative C^1 perturbation h_t $(t \in [0, 1])$ of g such that $h_t \in \mathcal{N}_\epsilon^\omega(g)$ and $g = h_t$ outside a small neighborhood of Γ. By way of construction, the area-preserving diffeomorphism h_t has the same invariant decomposition as g. Moreover, using that $\|Dg_y^\tau\| < K$ and also the unidimensionality of E^u, we have

$$\varphi(t) := \|D(h_t)_y^\tau|_{E^u_y}\| = \|Dg_y^\tau|_{E^u_y}\|\,\sigma^{-\tau t} < K\,\sigma^{-\tau t}\ , \tag{15.17}$$

while, on the other hand, $\|D(h_t)_y^\tau)|_{E^s_y}\| > K\,\sigma^{\tau t}$. For $t = 0$ we have

$$\varphi(0) = \|Dg_y^\tau|_{E^u_y}\| > 1.$$

But, since $\sigma^\tau \geq K$ (recall that $\tau > T$), for $t = 1$ we get $\varphi(1) < 1$. Therefore, there is some $t_0 \in]0, 1[$ such that $\varphi(t) = 1$. For such t_0 we must have $\|D(h_{t_0})_y^\tau\| = 1$.

Finally, applying[4] Lemma 15.4 to the periodic orbit y of h_{t_0} we get a conservative C^1 perturbation h of h_{t_0} such that $h \in \mathcal{N}_\epsilon^\omega(h_{t_0})$ and y is an elliptic periodic orbit of h.

Going back and replacing ϵ by $\epsilon/5$ along the proof enables us to conclude the proof of the proposition.

The absence of elliptic periodic orbits for all nearby perturbations implies uniform bounds on hyperbolic orbits with large enough period. This is an easy consequence of the two previous Propositions 15.1 and 15.2 which we state for future reference.

Corollary 15.1. *Let $f \in \mathit{Diff}^{\,1}_\omega(M)$ and $\epsilon > 0$ be given and set $\theta = \theta(\epsilon, f)$, $m = m(\epsilon, \theta)$ and $T = T(m)$ given by Propositions 15.1 and 15.2.*

Assume that all area-preserving maps g which are ϵ-C^1-close to f do not admit elliptic periodic orbits. Then for every such g all closed orbits with period larger

[4] If the point is parabolic we can perform a small rotation in the tangent space in order to make it elliptic.

than T are hyperbolic, m-dominated and with angle between its stable and unstable directions bounded from below by θ.

15.6 Proof of Theorem 15.2

In this section we present the proof of Theorem 15.2. Let $f \in \mathrm{Diff}_\omega^1(M)$ be a non Anosov diffeomorphism $\epsilon > 0$ and U any open subset of M, we will prove that there exists an area-preserving map $g \in \mathscr{N}_\epsilon^\omega(f)$ and which exhibits an elliptic orbit passing through U.

Let $\mathscr{P}$ be the residual set given by the general density theorem (see [31]), that is $\mathscr{P}$ is the set of all area-preserving maps f such that $\Omega(f)$ is the closure of the set of periodic orbits, all of them hyperbolic or elliptic, and $\Omega(f) = M$ by the Poincaré recurrence theorem.

We take any $f \in \mathrm{Diff}_\omega^1(M)$ which is not approximated by an Anosov area-preserving map. Then by a small C^1 perturbation we can and will assume that f belongs to $\mathscr{P}$ and that f is still *not* approximated by an Anosov conservative map. We fix some open set $U \subset M$ and $\epsilon > 0$.

If some elliptic periodic orbit of f intersects U there is nothing to prove, just choose $f = g$. Otherwise we must consider three cases:

$\boxed{\text{Case I}}$ All periodic orbits of f which intersect U are hyperbolic, and some of them has a small angle, less than $\theta = \theta(\epsilon, f)$ provided by Proposition 15.1, between the stable and unstable eigendirections at one point of the orbit.

$\boxed{\text{Case II}}$ All periodic orbits of f which intersect U are hyperbolic, with angle between the stable and the unstable directions bounded from bellow by θ, but some of them, with period larger than T, do not admits any m-dominated splitting, where $m = m(\epsilon, \theta)$ and $T = T(m)$ are given by Proposition 15.2, and $\theta = \theta(\epsilon, f)$ was given as before by Proposition 15.1.

$\boxed{\text{Case III}}$ All periodic orbits of f which intersect U and have period larger than T are hyperbolic, with m-dominated splitting, and with the angle between the stable and unstable directions bounded from bellow by θ, where $m = m(\epsilon, \theta)$ and $T = T(m)$ are given by Proposition 15.2, and $\theta = \theta(\epsilon, f)$ was given as before by Proposition 15.1.

Using Proposition 15.1 the Case I implies the desired conclusion for some area-preserving diffeomorphism $g \in \mathscr{N}_\epsilon^\omega(f)$. Analogously for Case II by the choice of the bounds m, T and by Proposition 15.2.

Finally, we use Theorem 15.6 to show that if f is in Case III and we assume that every C^1-nearby area-preserving map g does not admit elliptic periodic orbits through U, then we get a contradiction. This establishes the statement of Theorem 15.2.

If f is in Case III, then from Corollary 15.1 we know that every periodic orbit intersecting U, for area-preserving diffeomorphism $g \in \mathscr{N}_\epsilon^\omega(f)$, with period larger than T, is hyperbolic with uniform bounds on m and θ.

From Theorem 15.6, since f is not approximated by an Anosov area-preserving map, there exists an area-preserving map g, which is $\frac{\epsilon}{2}$-C^1-close to f, admitting a full Lebesgue measure subset $\mathscr{Z}$ where all the Lyapunov exponents for g are zero. Moreover, we can assume that g is aperiodic, that is the set of all periodic orbits has zero Lebesgue measure.[5]

Let $\hat{U} \subset U$ be a measurable set with positive Lebesgue measure. Let $R \subset \hat{U}$ be the set given by Poincaré Recurrence Theorem with respect to g. Then every $x \in R$ returns to $\hat{U}$ infinitely many times under g and is not a periodic point. Denote by $\mathscr{T}$ the set of positive return times to $\hat{U}$ under g.

Given $x \in \mathscr{Z} \cap R$ and $0 < \delta < \log 2/2m$, from the Oseledets' theorem there exists $n_x \in \mathbb{R}$ such that the upper Lyapunov exponent is near zero, formally, for every $n \geq n_x$ we have

$$e^{-\delta n} < \|Dg^n_x\| < e^{\delta n}.$$

Let us choose $\tau \in \mathscr{T}$ such that $\tau > \max\{n_x, T\}$.

Now, by Arnaud's closing lemma (Theorem 15.8), given a g-recurrent point x, $\epsilon > 0$ and a neighborhood $\mathscr{N}^{\omega}_{\epsilon/2}(g)$, there exists a periodic orbit p of $h \in \mathscr{N}^{\omega}_{\epsilon/2}(g)$ with period π such that

(a) $dist\big(g^i(x), h^i(p)\big) < \epsilon$ for all $i \in \{0, \dots, \pi\}$.
(b) $h = g$ except on the ϵ-neighborhood of the h-orbit of p.

Letting $\epsilon > 0$ be small enough we obtain also that

$$e^{-\delta\pi} < \|Dh^{\pi}_p\| < e^{\delta\pi} \quad \text{with} \quad \pi > T. \tag{15.18}$$

Now it is easy to see that $h \in \mathscr{N}^{\omega}_{\epsilon}(f)$, so that the orbit of p under h satisfies the conclusion of Corollary 15.1. In particular we have that

$$\frac{\|Dh^m_x \mid E^s_x\|}{\|Dh^m_x \mid E^u_x\|} \leq \frac{1}{2} \quad \text{for all } x \text{ in the } h\text{-orbit of } p,$$

for otherwise we would use Proposition 15.2 and produce an elliptic periodic orbit for an area-preserving map in $\mathscr{N}^{\omega}_{\epsilon}(f)$. Since the subbundles E^s and E^u are one-dimensional we write $p_i := h^{im}(p)$ for $i = 0, \dots, \lfloor \pi/m \rfloor = \ell$ with $\lfloor z \rfloor$ denoting the largest integer less or equal than z and

$$\frac{\|Dh^{\pi}_p \mid E^s_p\|}{\|Dh^{\pi}_p \mid E^u_p\|} = \frac{\|Dh^{\pi-m\ell} \mid E^s_{p_\ell}\|}{\|Dh^{\pi-m\ell} \mid E^u_{p_\ell}\|} \cdot \prod_{i=0}^{\ell-1} \frac{\|Dh^m \mid E^s_{p_i}\|}{\|Dh^m \mid E^u_{p_i}\|} \leq L(p,h) \cdot \left(\frac{1}{2}\right)^{\ell}, \tag{15.19}$$

where

[5] Actually, by the conservative version of the Kupka–Smale theorem (see [32]) we obtain a residual where the periodic points are countable, hence of zero Lebesgue measure.

$$L(p,h) = \sup_{i \in \{0,\dots,m\}} \left(\frac{\|Dh^i \mid E^s_p\|}{\|Dh^i \mid E^u_p\|} \right)$$

depends continuously on h in the C^1 topology. Therefore, there exists a uniform bound on $L(p,h)$ for all maps $h \in \mathcal{N}^{\omega}_{\epsilon}(f)$.

We note that we can take $\pi > T$ arbitrarily large by letting $\epsilon > 0$ be small enough in the above arguments. Therefore (15.19) ensures that

$$\frac{1}{\pi} \log \left\| Dh^{\pi} \mid E^s_p \right\| \leq \frac{1}{\pi} \log L(p,h) + \frac{\ell}{\pi} \log \frac{1}{2} + \frac{1}{\pi} \log \left\| Dh^{\pi} \mid E^u_p \right\|.$$

Moreover, since h is area-preserving and recalling (15.8), we have that the sum of the Lyapunov exponents along the h-orbit of p is zero, that is (we recall that π is the period of p)

$$\frac{1}{\pi} \log \|Dh^{\pi} \mid E^s_p\| = -\frac{1}{\pi} \log \|Dh^{\pi} \mid E^u_p\|.$$

The constants in (15.19) are independent of π so taking the period very large and noting that $\|Dh^{\pi}_p\| = \|Dh^{\pi} \mid E^u_p\|$ we deduce that

$$\frac{1}{\pi} \log \|Dh^{\pi}_p\| \geq \frac{1}{2m} \log 2 > \delta.$$

This contradicts (15.18) and Theorem 15.2 follows.

15.7 More Results on Area-Preserving Diffeomorphisms

15.7.1 Robust Transitivity

Here we present an alternative proof of Theorem 15.3 using the next well-known theorem (see for example [33, Theorem 5.2]).

Theorem 15.9. *(KAM) Let $f \in Diff^{\infty}_{\omega}(M)$, p a periodic elliptic orbit with period π and assume that the two eigenvalues of Df^{π}_p, denoted by λ_1 and λ_2, are such that $\lambda_1 = e^{2\pi i \theta}$ and $\lambda_1 = e^{-2\pi i \theta}$ for $\theta \in \mathbb{R} \setminus \mathbb{Q}$. Then, there exists a sequence $\{f_k\}_{k \in \mathbb{N}} \in Diff^{\infty}_{\omega}(M)$ such that $f_k \underset{k \to +\infty}{\longrightarrow} f$ (in the C^1-topology) such that each f_k has an elliptic periodic orbit p_k admitting a f_k-invariant tori.*

Proof. (of Theorem 15.3) Assume that $f \in \mathrm{Diff}^1_{\omega}(M)$ is non Anosov and C^1-robustly transitive. Hence, there exists a C^1-neighborhood of f, $\mathcal{V} \subset \mathrm{Diff}^1_{\omega}(M)$, such that every $h \in \mathcal{V}$ is transitive. By Theorem 15.2 given a non Anosov diffeomorphism $f \in \mathrm{Diff}^1_{\omega}(M)$, $\epsilon > 0$, $x \in M$ and any open subset U of M, then there exists $g \in \mathcal{N}^{\omega}_{\epsilon}(f)$ and exhibiting an elliptic orbit passing through U. Choose, ϵ

such that $g \in \mathscr{V}$. Since elliptic orbits are stable, we use Zehnder's Theorem [38] and we take $\tilde{g} \in \text{Diff}^{\infty}_{\omega}(M) \cap \mathscr{V}$ and exhibiting an elliptic orbit passing through U.

If the eigenvalues of this elliptic point are in $\mathbb{Q}$, then by using Lemma 15.4, we can perturb in order to get these eigenvalues in $\mathbb{R} \setminus \mathbb{Q}$.

Therefore, we are in the conditions of Theorem 15.9. So, there exists a sequence $\{f_k\}_{k\in\mathbb{N}} \in \text{Diff}^{\infty}_{\omega}(M)$ such that $f_k \underset{k\to+\infty}{\to} \tilde{g}$ (in the C^1-topology) such that each f_k has an elliptic periodic orbit p_k admitting a f_k-invariant tori. Of course that, for $k \geq k_0$, we have $f_k \in \mathscr{V}$ and the property of having f_k-invariant tori contradicts the C^1-robust transitivity.

We say that $f \in \text{Diff}^1_{\omega}(M)$ is *ergodic* if given any measurable f-invariant set it has full or zero Lebesgue measure. Stable ergodicity means persistence of the ergodicity for perturbations of f. It is easy to see that stable ergodicity implies robust transitivity within the conservative context. However, we note that this implication is false if the (stable) ergodicity is with respect to some atomic invariant measure (c.f. the next example).

Example 15.3. Consider the gradient flow on $\mathbb{S}^2 \subset \mathbb{R}^3$ generated by the height function $h(x, y, z) = -z$. The points $N = (0, 0, 1)$ and $S = (0, 0, -1)$ are a source and a sink respectively. The Dirac measure δ_N (or δ_S) is ergodic, however the flow is non-transitive.

Corollary 15.2. *If $f \in \text{Diff}^1_{\omega}(M)$ is C^1-stably ergodic, then f is Anosov.*

As we said in the introduction the KAM phenomena contrasts with stable ergodicity, since it prevails persistence of invariant tori with positive Lebesgue measure.

We end this section with the following yet unknown problem.

Question: Is ergodicity C^1-generic among conservative surface diffeomorphisms?

15.7.2 Area-Preserving Star Diffeomorphisms

Let $f \in \text{Diff}^1_{\omega}(M)$ be a conservative star-diffeomorphism, that is, there exists a neighborhood $\mathscr{V}$ of f in $\text{Diff}^1_{\omega}(M)$ such that any $g \in \mathscr{V}$, has all the periodic orbits hyperbolic. We denote this set by $\mathscr{F}^1_{\omega}(M)$ and, as we said in Sect. 15.3, $\mathscr{A}^2_{\omega}$ denotes the set of conservative Anosov diffeomorphisms on the surface M.

It is clear that $\mathscr{F}^1(M) \cap \text{Diff}^1_{\omega}(M) \subset \mathscr{F}^1_{\omega}(M)$; Theorem 15.4 implies that

$$\mathscr{F}^1(M) \cap \text{Diff}^1_{\omega}(M) = \mathscr{F}^1_{\omega}(M) = \mathscr{A}^2_{\omega}.$$

As a consequence of Theorem 15.4 we also obtain the following result.

Corollary 15.3. *The boundary of $\mathscr{A}^2_{\omega}$ has no isolated points.*

A diffeomorphism $f \in \mathrm{Diff}^1_\omega(M)$ is said to be C^1-structurally stable in the conservative setting if there exists a C^1 neighborhood, $\mathscr{V}$, of f in $\mathrm{Diff}^1_\omega(M)$ such that every $g \in \mathscr{V}$ is topological equivalent to f (see [33]).

Combining Theorem 15.4 with Theorem 15.1 we are able to obtain the next result.

Theorem 15.10. *If f is a C^1-structurally stable surface area-preserving diffeomorphism, then f is Anosov.*

We assume Theorem 15.4 for a moment and we conclude the proof of Theorem 15.10 but before that we present an abstract result about finite product of $SL(2,\mathbb{R})$ matrices that will be used in the proof of Theorem 15.10.

Lemma 15.9. *([11, Lemme 6.6]) For all $\epsilon > 0$, there exists $N \geq 1$ such that, for all $n \geq N$ and every family $\{A_i\}_{i=1}^n \subset SL(2,\mathbb{R})$, there exists $\{\alpha_i\}_{i=1}^n$ (where each $\alpha_i \in]-\epsilon,\epsilon[$) satisfying the following property: For all $i \in \{1,\dots,n\}$ we denote $B_i = R_{\alpha_i} \cdot A_i$ and we have that*

$$B_n \cdot B_{n-1} \cdot \ldots \cdot B_1$$

has real eigenvalues.

Proof. (of Theorem 15.10) Let us fix a C^1-structurally stable area-preserving diffeomorphism in $\mathrm{Diff}^1_\omega(M)$ and choose a neighborhood $\mathscr{V}$ of f whose elements are topologically equivalent to f. If $f \notin \mathscr{A}^2_\omega = \mathscr{F}^1_\omega(M)$, then it follows that $\mathscr{V} \cap \mathscr{A}^2_\omega = \emptyset$. Using Theorem 15.1 one gets that there exists a residual subset $\mathscr{R} \subset \mathscr{V}$ such that for every $f_0 \in \mathscr{R}$ the set of elliptic periodic orbits is dense in M. Let us fix $f_0 \in \mathscr{R}$ and choose a small neighborhood of f_0, $\mathscr{W} \subset \mathscr{V}$.

Let x be an elliptic periodic point of large period, say π (given by Lemma 15.9) depending on ϵ (depending on $\mathscr{V}$) and on $A_i := Df_{f^i(x)}$ for $i = 1,\dots,\pi$. Define, for $t \in [0,1]$, $B_{i,t} := R_{t\alpha_i} \cdot A_i$. By Lemma 15.9 we obtain that

$$B_1^\pi := B_{\pi,1} \cdot B_{\pi-1,1} \cdot \ldots \cdot B_{1,1}$$

has real eigenvalues. Since $B_0^\pi = A^\pi = Df_x^\pi$ has complex eigenvalues, there must be $t_0 \in]0,1[$ such that $B_{t_0}^\pi$ has a parabolic behavior. Finally, we apply Lemma 15.4 several times, in order to realize an area-preserving map $f_1 \in \mathscr{V}$ exhibiting a parabolic periodic orbit. Since the existence of a parabolic point prevents structural stability and $f_1 \in \mathscr{W}$ we get a contradiction. Therefore $f \in \mathscr{A}^2_\omega$, which ends the proof.

Proof. (of Theorem 15.4) We observe that $\mathscr{F}^1_\omega(M)$ is C^1 open in $\mathrm{Diff}^1_\omega(M)$. Let $f \in \mathscr{F}^1_\omega(M) \setminus \mathscr{A}^2_\omega$.

We recall Corollary 15.1 and we consider a C^1-neighborhood $\mathscr{V}$ of f in $\mathscr{F}^1_\omega(M)$ where any $g \in \mathscr{V}$ do not admit elliptic closed orbits. Then, from Corollary 15.1 there exist constants $\theta = \theta(\epsilon, g)$, $m = m(\epsilon,\theta)$ and $T = T(m)$ such that, for each periodic orbit with period greater than T, one has:

- m-dominated splitting and
- Angle between its stable and unstable directions bounded from below by θ.

Observe that, since $g \in \mathscr{F}^1_\omega(M)$, these periodic orbits are hyperbolic.

We will get a contradiction with the fact that there exists a positive measure set without domination. For that we consider the following claim.

Claim. For all $m \in \mathbb{N}$, there exists an f-invariant and positive Lebesgue measure set $\varGamma_m \subset M$ without m-dominated splitting.

If the claim was false, then there would exist $m \in \mathbb{N}$ and $\varLambda_m \subset M$ such that $Leb(M \setminus \varLambda_m) = 0$ and $\varLambda_m$ has an m-dominated splitting. Since the m-dominated splitting extends to the closure and we are considering the Lebesgue measure it follows that M has an m-dominated splitting. But the existence of an m-dominated splitting implies, by Lemma 15.3, that f is Anosov which contradicts our assumption.

Now, we recall the core of the dynamical principle involved in the proof of Theorem 15.6; given any $\epsilon > 0$, there exists (a sufficiently large) $m \in \mathbb{N}$ such that for any $\eta > 0$ arbitrarily close to 0, for a.e. $x \in \varGamma_m$ there exists g, ϵ-C^1-close to f, such that $e^{-n\eta} < \|Dg^n_x\| < e^{n\eta}$, for every arbitrarily large $n \in \mathbb{N}$.

Repeating the arguments in the proof of Theorem 15.2 we get a periodic point with period π for an area-preserving map $h \in \mathscr{V}$ and such that:

$$e^{-\delta\pi} < \|Dh^\pi_p\| < e^{\delta\pi}, \tag{15.20}$$

and in the same way we obtain a contradiction. Therefore, f has a dominated splitting over M and, by Lemma 15.3, we conclude that f is Anosov.

Proof. (of Corollary 15.3) Take an isolated point f in the interior of the boundary of $\mathscr{A}^2_\omega$ and a small neighborhood $\mathscr{V}$ of f such that any $g \in \mathscr{V}$ is Anosov. The diffeomorphism f must satisfy Claim 15.7.2 otherwise f is Anosov. We follow the proof of Theorem 15.4 and we conclude that under a small C^1-perturbation we find $g \in \mathscr{V}$ exhibiting an elliptic periodic orbit which is a contradiction.

15.7.3 Homoclinic Tangencies

For surface area-preserving diffeomorphisms the existence of smooth invariant curves is associated to the existence of elliptic points. Actually, Mora and Romero [23] developed a mechanism to create open sets containing a dense set of maps exhibiting homoclinic tangencies once one has a smooth invariant curve. A key step to prove this result is [23, Proposition 7]. To state this proposition let us define

$$\mathbb{A} = \{(\theta, r) \colon \theta \in \mathbb{S}^1,\ r \in \mathbb{R}\} \text{ and } \mathbb{A}_\delta = \{(\theta, r) \colon \theta \in \mathbb{S}^1,\ r \in]-\delta, \delta[\}.$$

Theorem 15.11. *Let $f \colon \mathbb{A}_\delta \to \mathbb{A}$ be a C^∞ area-preserving map of the annulus leaving invariant some C^∞ curve*

$$\Lambda = \{(\theta, \Phi(\theta)), \theta \in \mathbb{S}^1\},$$

where $\Phi: \mathbb{S}^1 \to \mathbb{R}$, and such that $f|_\Lambda$ has an irrational rotation number. Then, for $s \geq 1$ and $\epsilon > 0$, f can be ϵ-C^s-approximated by an area-preserving g exhibiting homoclinic tangencies such that for some $\delta' < \delta$ we have

$$g|_{\mathbb{A}_\delta \setminus \mathbb{A}_{\delta'}} = f|_{\mathbb{A}_\delta \setminus \mathbb{A}_{\delta'}}.$$

Let $f_0 \in \mathrm{Diff}^1_\omega(M)$ be such that it cannot be C^1-approximated by a diffeomorphism in $\mathscr{A}^2_\omega$. Using Theorem 15.1, we approximate, in the C^1-topology, f_0 by $f_1 \in \mathrm{Diff}^1_\omega(M)$ such that the elliptic points of f_1 are dense on the surface. Now, using Zehnder Theorem [38] and the stability of elliptic orbits, we approximate, in the C^1-topology, f_1 by $f_2 \in \mathrm{Diff}^\infty_\omega(M)$ having an elliptic point p of period π.

Now we consider the linear action $Df_2^\pi: T_pM \to T_pM$ defined by the rotation R_θ, in a small neighborhood of the orbit, and a direct application of Theorem 15.7 allows us to C^1-approximate f_2 by $f_3 \in \mathrm{Diff}^\infty_\omega(M)$ such that p is still an elliptic point of period π and there exists an f_3-invariant neighborhood $\mathfrak{T}$ where the first return map at p (not the tangent map) is a rotation of angle θ. We can assume that θ is irrational, otherwise, we could perturb f_3, by using Lemma 15.4, in order to get $f_4 \in \mathrm{Diff}^\infty_\omega(M)$, C^1-close to f_3, with the same properties but with irrational rotation angle. This area-preserving diffeomorphism is in the hypotheses of Theorems 15.11 and 15.5 is proved.

We end this section by observing that, for dimension $d \geq 3$, the author and Rocha proved in [8] that any volume-preserving d-dimensional diffeomorphism can be C^1-approximated by an Anosov volume-preserving diffeomorphism or else by a volume-preserving diffeomorphism exhibiting a heterodimensional cycle.

15.7.4 Lots of Chaos or Lack of It?

We recall one of the most common definitions of chaos due to Devaney (see [14, Definition 8.5]): $f: M \to M$ is *chaotic* if:

(a) f is transitive.
(b) The periodic points are dense in M.
(c) f is *sensitive to the initial conditions*, i.e., there exists $\delta > 0$ such that for all $x \in M$ and all neighborhood of x, V_x, there exists $y \in V_x$ and an integer n where $dist(f^n(y), f^n(x)) > \delta$.

In this case we also say that f is *chaotic in the topological sense*.

It was proved in [6] that (a) and (b) implies (c), and so in order to be chaotic in the sense of Devaney the system only has to satisfy the transitivity property and the density of periodic points.

The other definition of chaotic map that we are going to use is the one that says that there are no zero Lyapunov exponents for Lebesgue almost every point.

When, in our conservative surface setting, we have two non-zero (thus symmetric) Lyapunov exponents we say that f is *chaotic in the measurable sense*.

Theorem 15.12. *Let M is any closed surface aside from the two-torus. There exists a C^1-residual $\mathscr{R} \subset \mathrm{Diff}^1_\omega(M)$ such that, if $f \in \mathscr{R}$, then f is chaotic in the topological sense and nonchaotic in the measurable sense.*

Proof. As an outcome of [11] we obtain that there exists a residual subset $\mathscr{R}_1$ of $\mathrm{Diff}^1_\omega(M)$ such that if $f \in \mathscr{R}_1$, then f is transitive. Furthermore, by the general density theorem [31] we get there exists a residual subset $\mathscr{R}_2$ of $\mathrm{Diff}^1_\omega(M)$ such that if $f \in \mathscr{R}_2$, then the periodic points of f are dense in M. Therefore, defining $\mathscr{R}_3 = \mathscr{R}_1 \cap \mathscr{R}_2$ and recalling [6] we conclude that there exists a residual subset $\mathscr{R}_3$ of $\mathrm{Diff}^1_\omega(M)$ such that if $f \in \mathscr{R}_3$, then f is chaotic in the topological sense.

By Franks' classical result about the rigidity of Anosov diffeomorphisms (see [15]) we know that the only surfaces that support Anosov diffeomorphisms are the tori. Therefore, if M is any closed surface except the two-torus, then by Theorem 15.6, there exists a C^1-residual subset $\mathscr{R}_4$ of $\mathrm{Diff}^1_\omega(M)$ such that, if $f \in \mathscr{R}_4$, then f has zero Lyapunov exponents for almost every points, thus is nonchaotic in the measurable sense.

Finally, $\mathscr{R} := \mathscr{R}_3 \cap \mathscr{R}_4$ is the residual set required by the statement of the theorem.

Acknowledgements The author is greatly indebted to Vítor Araújo, Jairo Bochi, Pedro Duarte and Jorge Rocha for all the knowledge obtained during the conversations and discussions about these issues. The author was partially supported by FCT through CMUP (SFRH/BPD/20890/2004).

References

1. Araújo, V., Bessa, M.: Dominated splitting, singularities and zero volume for incompressible 3-flows. Nonlinearity **21**, 1637–1653 (2008)
2. Arbieto, A., Matheus, C.: A pasting lemma and some applications for conservative systems. Ergod. Theory Dyn. Syst. **27**(5), 1399–1417 (2007)
3. Arnaud, M.-C.: The generic symplectic C^1-diffeomorphisms of four-dimensional symplectic manifolds are hyperbolic, partially hyperbolic or have a completely elliptic periodic point. Ergod. Theory Dyn. Syst. **22**(6), 1621–1639 (2002)
4. Arnaud, M.-C.: Le "closing lemma" en topologie C^1. Mm. Soc. Mat. Fr. (N.S.) **74**, vi+120 (1998)
5. Arnold, V.I.: "Mathematical Methods of Classical Mechanics". Springer (1978)
6. Banks, J., Brooks, J., Cairns, G., Davis, G., Stacy, P.: On devaney's definition of chaos. Am. Math. Montly **99**, 332–334 (1992)
7. Bessa, M., Duarte, P.: Abundance of elliptic dynamics on conservative three-flows, Dyn. Syst. Intl. J. **23**(4), 409–424 (2008)
8. Bessa, M., Rocha, J.: Anosov versus Heterodimensional cycles: A C^1 dichotomy for conservative maps. `http://cmup.fc.up.pt/cmup/bessa/` Preprint (2009)
9. Bochi, J.: Genericity of zero Lyapunov exponents, Ergod. Theory Dyn. Syst. **22**(6), 1667–1696 (2002)
10. Bochi, J., Viana, M.: The Lyapunov exponents of generic volume-preserving and symplectic maps. Ann. Math. 2 **161**(3), 1423–1485 (2005)

11. Bonatti, C., Crovisier, S.: Récurrence et généricité, Invent. Math. **158**(1), 33–104 (2004)
12. Bonatti, C., Díaz, L.J., Viana, M.: "Dynamics beyond uniform hyperbolicity. A global geometric and probabilistic perspective". Encycl. Math. Sci. **102**. Math. Phys. 3. Springer (2005)
13. Dacorogna, B., Moser, J.: On a partial differential equation involving the Jacobian determinant. Ann. Inst. Henri Poincaré **7**(1), 1–26 (1990)
14. Devaney, R.: "An Introduction to Chaotic Dynamical Systems". Addison-Wesley (1989)
15. Franks, J.: Anosov diffeomorphisms. Global Analysis (Proc. Sympos. Pure Math., vol. XIV, Berkeley, California, 1968) 61–93. American Mathematical Society, Providence, R.I. 1070
16. Kuratowski, K.: "Topology, vol. 1". Academic (1966)
17. Liao, S.D.: On the stability conjecture. Chinese Ann. Math. **1**, 9–30 (1980)
18. Mañé, R.: Persistent manifolds are normally hyperbolic. Bull. Am. Math. Soc. **80**, 90–91 (1974)
19. Mañé, R.: An ergodic closing lemma. Ann. Math. **116**, 503–540 (1982)
20. Mañé, R.: Oseledec's theorem from the generic viewpoint. In: Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Warsaw, 1983), pp. 1269–1276. Warsaw, 1984. PWN
21. Mañé, R.: The Lyapunov exponents of generic area preserving diffeomorphisms. In: International Conference on Dynamical Systems (Montevideo, 1995), vol. 362 Pitman Res. Notes Math. Ser., pp. 110–119. Longman, Harlow, 1996
22. Mañé, R.: "Ergodic Theory and Differentiable Dynamics". Springer, Berlin (1987)
23. Mora, L., Romero, N.: Persistence of homoclinic tangencies for area-preserving maps. Ann. Fac. Sci. Toulouse Math. **6**(4), 711–725 (1997)
24. Newhouse, S.: Quasi-elliptic periodic points in conservative dynamical systems. Am. J. Math. **99**, 1061–1087 (1977)
25. Oseledets, V.I.: A multiplicative ergodic theorem: Lyapunov characteristic numbers for dynamical systems. Trans. Moscow Math. Soc. **19**, 197–231 (1968)
26. Palis, J., Takens, F.: "Hyperbolicity and Sensitive Chaotic Dynamics at Homoclinic Bifurcations. Fractal Dimensions and Infinitely Many Attractors". Cambridge Studies in Advanced Mathematics, vol. 35. Cambridge University Press, Cambridge (1993)
27. Palis, J., Smale, S.: Structural stability theorems. 1970 Global Analysis (Proc. Sympos. Pure Math., Vol. XIV, Berkeley, California, 1968) pp. 223–231. American Mathematical Society, Providence, R.I., 1968
28. Palis, J.: Open questions leading to a global perspective in dynamics. Nonlinearity **21**, 37–43 (2008)
29. Pliss, V.A.: On a conjecture of Smale. Differ. Uravnenija **8**, 268–282 (1972)
30. Pollicott, M.: "Lectures on Ergodic Theory and Pesin Theory on compact manifolds". London Mathematical Society Lecture Notes Series, vol. 180. London Mathematical Society, Cambridge (1993)
31. Pugh, C., Robinson, C.: The C^1 closing lemma, including Hamiltonians. Ergod. Theory Dyn. Syst. **3**, 261–313 (1983)
32. Robinson, C.: Generic properties of conservative systems. Am. J. Math. **92**, 562–603 (1970)
33. Robinson, C.: Dynamical Systems. Stability, Symbolic Dynamics, and Chaos, 2nd edn. Stud. Adv. Math. CRC, Boca Raton, FL (1999)
34. Saghin, R., Xia, Z.: Partial Hyperbolicity or dense elliptical periodic points for C^1-generic symplectic diffeomorphisms. Trans. AMS **358**, 5119–5138 (2006)
35. Shub, M.: "Global Stability of Dynamical Systems". Springer, New York (1987)
36. Takens, F.: Homoclinic points in conservative systems. Invent. Math. **18**, 267–292 (1972)
37. Yoccoz, J.C.: Travaux de Herman sur les Tores invariants. Astérisque **206**(4), Exp:754, 311–344 (1992)
38. Zehnder, E.: Note on smoothing symplectic and volume-preserving diffeomorphisms. In: Proceedings III Latin American School of Mathematics, Inst. Mat. Pura Aplicada CNPq, Rio de Janeiro, 1976, vol. 597 Lecture Notes in Math., pp. 828–854. Springer, Berlin (1977)

Chapter 16
A Theoretical, Multidisciplinary View of Catastrophic Regime Change

Juan Gabriel Brida, Audrey L. Mayer, Christopher McCord, and Lionello F. Punzo

Abstract Dynamic regime theory is used in a growing number of disciplines to understand, manage, and predict system behavior. A variety of mathematical models have been developed for seemingly disparate systems, however the similarity of these models suggests that the systems could be approached as a collection of samples. A multidisciplinary meta-analysis of dynamic regime models could yield several benefits. Given the difficulty of replication and experimentation in real-world systems, a collection of dynamic systems across disciplines and scales could serve as much-needed replicates. If endogenous variables behave similarly regardless of the source of exogenous pressures, and of the scale at which the system is define, then general models, rules and coded behaviors can be developed. Furthermore, if the same basic theory regarding system behavior (including rapid regime change) applies across disciplines at multiple spatiotemporal scales, then models developed from these theories may help manage those systems which, at larger scales, cross traditional disciplinary lines. This result would emphasize the need to collaborate across disciplines to study the sustainability of dynamic systems. Here we discuss the mathematical basis for common dynamic regime models, and then

J.G. Brida (✉)
Free University of Bolzano, Via Sernesi 1, 39100 Bolzano, Italy
e-mail: JuanGabriel.Brida@unibz.it

A.L. Mayer
Michigan Technological University, Department of Social Sciences and School of Forest Resources and Environmental Science, 1400 Townsend Dr., Houghton, MI, 49931, USA
e-mail: almayer@mtu.edu

C. McCord
Department of Mathematical Sciences, University of Cincinnati, P.O. Box 210025, Cincinnati, OH 45221-0025, USA
e-mail: mccordck@asmail.artsci.uc.edu

L.F. Punzo
Department of Economics, University of Siena, Piazza S. Francesco, 7, 53100 Siena, Italy
and
PPED-INCT, UFRJ, Rio de Janeiro, Brazil
e-mail: punzo@unisi.it

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_16,

describe their application to sociological, ecological, and economic systems, in a scale-explicit manner.

16.1 Introduction

The increasing visibility of publications investigating catastrophic regime change (e.g., [1,35,38,41,46]) suggests that dynamic systems and catastrophe theories are increasingly central to a wide variety of disciplines. To study complex dynamic systems, researchers collect data or construct models to determine which variables and parameters may be useful to identify regimes and their boundaries. In many cases, knowledge about the distance to a regime boundary (and hence a regime shift) is needed to either forestall regime change or bring about regime change (to a more desirable regime). While specific variables may differ across disciplines, models are based on the same fundamental principles.

The similarity in behaviors across systems would suggest that a mathematical foundation could be used to study the systems as a unified set [44], increasing the sample size for more quantitative approaches to dynamic systems research. Likewise, if modeled behaviors are scale-invariant, then each system can potentially serve as one of several samples as boundaries are expanded to larger spatial and temporal scales. The scale at which the system's boundaries are drawn has a large impact on the regimes that can be observed, and the differentiation between endogenous and exogenous forces. To investigate the similarity between systems and across scales, we describe mathematical research on basin boundaries and multiple time-scale dynamics, demonstrating the fundamental mechanisms and properties of regime shifts. Then, using several examples, we identify themes across disciplines and develop a common terminology.

16.1.1 What is a Regime?

A regime is a dynamic model with its own associated multidimensional domain, in which state variables exhibit characteristic behaviors or structures. Those structures can be defined either by inherent dynamic behavior (e.g., a basin of attraction) or by the observable manifestation of them (e.g., an oligotrophic lake vs. a eutrophic one). The state space of a system can encompass multiple regimes of a variety of basin sizes and attraction strength, which in some disciplines is referred to as resilience [6]. Identifying the boundaries that separate regimes is essential to understanding the regimes themselves. However, boundaries may be poorly defined, vary with exogenous and endogenous parameters, and evolve in time in response to changes in these parameters.

A regime shift occurs when a system moves across regime boundaries, which are influenced by a variety of exogenous and endogenous mechanisms. These are qualitatively different from phase transitions, which are driven solely by changes in external conditions (e.g., liquid water into ice as temperature drops), even through the phrases are sometimes used synonymously in some disciplines. Exogenously generated regime shifts can take several forms, including: continuous changes in parameters or in the functional form of the dynamics; or randomly distributed shocks which change the values of the state variables, the parameters and/or the very rule of motion (i.e., stochastic or random motion that is layered on top of a deterministic system). Endogenously generated shifts depend fundamentally on mechanisms internal to the system, inbuilt in its architecture or relational wiring. Shifts often occur when the system reaches and overshoots some frontier values in its state space and/or in the parameter space.

The various mechanisms and their interactions can produce behavior that is a compounding of smooth evolution within a given regime, and sudden qualitative discontinuities in behavior across regimes – regime shifts (also "catastrophes", [44]). Although the dynamics of the system may be linear or otherwise easily predictable within a regime, cross-regime dynamics shifts can be reversible or irreversible. However, the evidence that such a shift has occurred or is about to occur may be subtle, complicating predictions of when a system is in or about to enter a regime change, and complicating the decision processes for controlling regime shifts.

16.2 Mathematical Models

To analyze the basic properties of multiple regime phenomena, it is useful to capture the basic properties in a mathematical model. Usually a dynamical model is represented by a system of difference and/or differential equations. Several tasks have to be solved:

Identify distinct dynamic regimes and the boundaries separating their domains. In relatively simple dynamic systems, regimes are associated with the presence (and the properties) of the set of the system's steady states or attractors. (This requires the set of equilibria to be finite, i.e. embedded in a zero-dimensional space, which of course is rare if the state space is large) When this is the case, their description preliminarily requires the identification of the basins of attraction. When two or more attractors are present in a given system, each of them has its own basin of transient states lying on trajectories that lead asymptotically to it. Basin boundaries are thin sets (i.e., Lebesque measure zero) separating basins and commonly have a complicated fractal structure [28, 33]. With one-dimensional systems and related to generating partitions, one can identify regimes with increasing-decreasing intervals [26].

The analogous situation in greater dimension can be given by piecewise defined systems [39]. In more general cases, attractors can have a more complicated set structure. The simplest structures are closed loops, but any compact subset of the state space can act as an attractor. In this case, well-defined domains are more difficult to isolate, or may not exist at all. We have the general case of multiregime dynamics, whereby a domain can be associated with a given local rule or mathematical model, without this implying any stability property in the classical sense. We have a dynamics segmented through state space, with regularly or irregularly distributed jumps between nearby domains (but not necessarily nearby states belonging to adjacent domains). In this case, a partition in regimes reflects our understanding or our hypotheses upon the various local rules; it is an exogenous as opposed to the endogenous partition of the previous case. The result is a regular, a mildly irregular, or a seemingly chaotic regime dynamics, which can be coupled to any type of local point dynamics.

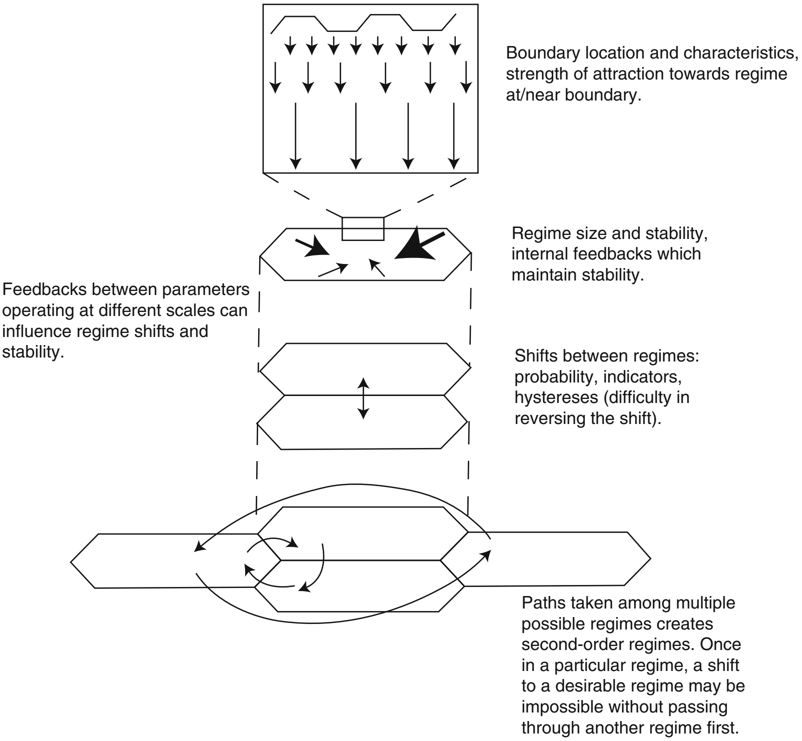
Describe the basic properties of the dynamic regimes. Once we have identified the portfolio of regimes, we can study their stability and reversibility, compute the dimension of the regime, and investigate for the presence of hysteresis and the elasticity and amplitude of resilience [47]. A regime is highly dependent upon the scale at which the system is observed; at smaller scales, variability within a regime may appear as large, discontinuous shifts. Therefore, one of the properties of a regime is the scale at which it exists. Here, scale includes both spatial and temporal dimensions, and refers to the resolution of the observations (i.e., how often do they occur) and their extent (usually dictated by the system boundaries). It is also possible to study the dependence of regime boundaries and other characteristics on the parameters of the system.

Identify the mechanism of regime change. If the system has only stable states, then a shift across regimes may require an external perturbation, or an abrupt change in model parameters. Nonlinear functions of the state can represent the effect of cumulative causes and self-reinforcing mechanisms bringing about qualitative changes [5]. These exogenous shifts can be modeled, for example, with Markov chains [5]. Again, the scale at which the system is observed is important. Exogenous mechanisms which can cause regime shifts at one scale may not do so at larger scales, at which the same mechanism is now endogenous. Furthermore, forcing mechanisms can interact across scales, complicating efforts to identify each mechanism and its relationship to the system.

Describe dynamics across regimes. Multiple regime models exhibit a twofold dynamics: within a regime and across regimes. While dynamics within regimes are represented in the classical form of differential or difference equations, dynamics across regimes can be represented via symbolic and coded dynamics [4, 5]. Hybrid systems may be modified to represent multi-regime models (see [20]), describing an interaction mechanism between discrete (representing regime shifts) and continuous dynamics (representing dynamics within a regime). At this macroscale, the "regime" is now the path between regimes at a smaller scale; some smaller-scale regimes may never be visited by the system if the system starts in a particular regime or becomes ensconced in a particular loop between a subset of all possible regimes (Fig. 16.1).

16.3 Scale Dependency

Depending upon the scale (spatial or temporal) at which a system is observed, the boundaries of a regime and shifts between regimes may not be obvious or relevant. The scale-dependent definition of a regime is important to note when collecting



Boundary location and characteristics, strength of attraction towards regime at/near boundary.

Regime size and stability, internal feedbacks which maintain stability.

Feedbacks between parameters operating at different scales can influence regime shifts and stability.

Shifts between regimes: probability, indicators, hystereses (difficulty in reversing the shift).

Paths taken among multiple possible regimes creates second-order regimes. Once in a particular regime, a shift to a desirable regime may be impossible without passing through another regime first.

Fig. 16.1 Dynamic regimes research at increasingly larger spatial and temporal scales. Arrow represent attracting forces at small scales, and system behavior (direction of movement) at larger scales

and using data to describe and manage a system. Here, we begin with a simple two-regime example in sociology at the scale of a romantic couple, and scale up to ecological and climate systems with mechanisms operating at several scales to form multiple possible regimes.

16.3.1 Sociology

Sociologists and psychologists have used nonlinear dynamics to develop testable hypotheses for describing and predicting patterns in human behavior and interaction [13, 27, 49]. Romantic relationships, in particular, have received considerable attention from this perspective [13]. A very simple model identifies the stability of a couple using two differential equations (one for each partner) with three variables: the attraction each partner feels towards the other, the degree to which a partner reciprocates affection, and the durability of each partner's memory of these two variables [36]. Stable couples have one regime in which they tend to continuously return affection and remain interested, while unstable couples have two regimes, one characterized by positive feelings and one by negative. The regime to which unstable couples are attracted depends on the initial conditions at the start of the relationship. This model was meant to describe the behavior of couples at the scale of months or years, and therefore ignores fluctuations at shorter timescales or lifetime stability of the relationship [36].

More complicated relationship models include both partners plus an exogenous force, such as expectations from the community or negative life events [27, 49]. Again, the endogenous mechanism influencing whether or not a relationship will be formed, is the preferences and attractions internal to each potential partner. However, the strength of this mechanism is now also dependent upon an exogenous force, social pressure, which can influence just how attractive a potential mate needs to be before a relationship is formed (Fig. 16.2). Without strong social pressure to stay in a relationship (such as taboos against divorce), the probability of forming a partnership is almost linearly related to the attractiveness of the individual as a potential partner [43] (Fig. 16.2).

In fact, with no social pressure these two regimes are indistinct. As social pressure increases for relationship maintenance, the probability of entering a

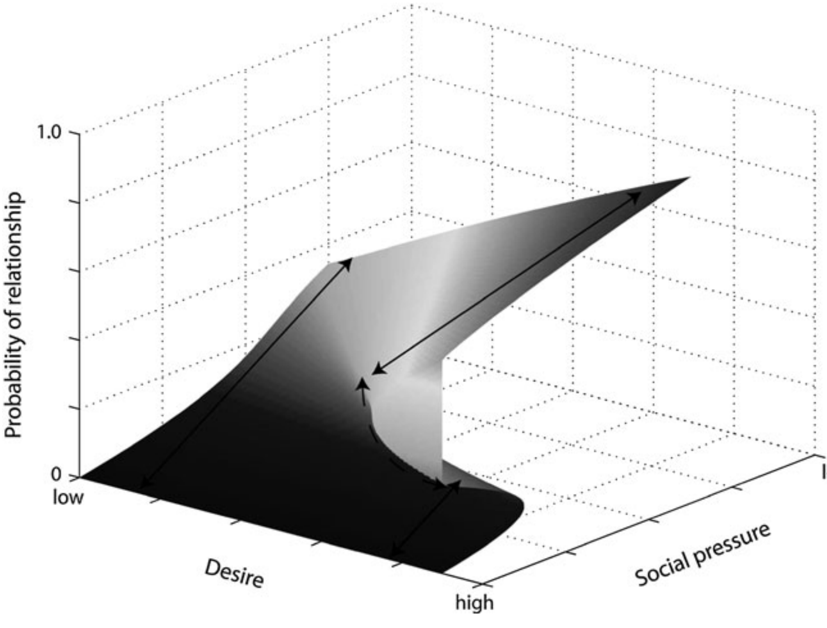


Fig. 16.2 With little social pressure (either positive or negative) regarding dating, the probability that a person will enter a relationship regime is linearly related to the desirability of the potential partner (*back wall of cube, arrow on left*). As social pressure increases, a hysteresis (catastrophe) develops. The desirability of the potential partner must be very high before a relationship is started, and once in a relationship, partner desirability must decrease dramatically to end the relationship regime (*arrows on right*). {Modified from [27]}

relationship forms a hysteresis between being in and out of a relationship; relationships are entered and exited more tentatively [27]. In this way, societal pressure (and the susceptibility of individuals to this pressure) can increase the resilience of both relationship/no relationship regimes, and can determine the nonlinearity of the shift between regimes. In the presence of strong social pressure, potential mates must be very attractive to begin a relationship, and problems must be severe (causing a partner to become very unattractive) to dissolve a relationship.

While the stability of a relationship is dependent upon the internal behavior of the couple along with exogenous factors, the appearance of stability is highly dependent upon the scale at which behavioral observations are made. Some variables (such as the amount of irritability directed towards the partner) may appear to fluctuate randomly during the day or over several weeks, suggesting instability. However, over longer time periods these same variables demonstrate a more ordered, predictable pattern, indicating a considerably stable relationship [49]. Variables which fluctuate over much longer timescales, such as overall satisfaction with the relationship, may seem stationary at short timescales but after crossing a threshold may precipitate a (catastrophic) end to the relationship.

16.3.2 Ecology

In most dynamic systems, several endogenous and exogenous mechanisms interact to form a variety of possible regimes, and a variety of ways to shift between regimes. Endogenous and exogenous causes for regime shifts (and feedbacks which stabilize regimes) have been observed in a broad range of ecosystems, from terrestrial to freshwater to marine [38]. However, few ecosystems have well-developed mathematical models to describe system behavior, primarily due to the difficulty of obtaining empirical observations and finding numerous system replicates. For ecosystems with vague boundaries (e.g., marine, [21]), low number of replicates relative to the potential types of exogenous forces (e.g., coral reefs, [29]), or significant spatial heterogeneity [45], models developed for systems from other disciplines may prove especially helpful, especially for cases for which data collection is easier and replicates can be constructed [21, 31].

Shallow, temperate freshwater lakes represent one of the most studied ecological cases of regime shifts. The large number of relatively isolated replicates and long history of human interactions have allowed for rigorous data collection and modeling [6]. These lakes typically persist in one of two states: an oligotrophic one with low algal biomass, high biomass of rooted plants, and low phosphorus recycling between the sediments and the water; and a eutrophic one with high algal biomass (and frequent blooms), few rooted plants, and high phosphorus recycling. An increase in phosphorus inputs from outside of the lake, such as sewage from human settlements or runoff from fertilized agricultural fields, can push a lake from an oligotrophic to eutrophic state. As water quality becomes murkier (due to high algal biomass), sunlight penetrates less deeply and rooted plants die. Plant loss reduces the removal of phosphorus from the water for plant tissues, and the loss of their roots destabilizes lake sediments, allowing a greater exchange of phosphorus between the water and the sediments. Shifts from oligotrophic to eutrophic conditions can occur rapidly at the threshold at which lake sediments can no longer absorb additional phosphorus, promoting algal blooms and precipitating rooted plant loss.

Mechanistic models and statistical models based on time-series data have demonstrated pronounced hysteresis between these regimes. Even if phosphorus addition to the lake is dramatically reduced, phosphorus concentration in the lake can remain high for long periods of time, indicating a high resilience of the eutrophic regime, (particularly in shallow lakes or those with low rates of hydrologic flushing). Regime shifts in shallow lakes can be irreversible on shorter time scales, as the loss of key animal and plant species remove a mechanism of phosphorus removal even when phosphorus inputs are dramatically reduced [6].

16.3.3 Climatology

The largest dynamic system on Earth is the climate system, and although the sun exogenously supplies energy to drive climate processes, variability in the energy reaching the planet does not explain the variability in global climate [37]. Instead, feedbacks between several endogenous mechanisms maintain climate regimes: the concentration of several atmospheric gases; freshwater input into the ocean and resulting circulation patterns; and the albedo of the planet surface governed by the extent of ice and snow [9, 32]. The atmosphere may be the most unstable variable, changing more quickly than other global climate factors [16]. Early differential equation models of these three variables indicated the potential for catastrophic shifts in global climate [48]. However, as more climate data and higher computing power became available, climate models became increasingly complex to account for the many physical and biological feedbacks between different components and regions [16]. Despite their complexity, more recent models such as General Circulation Models (GCM), still demonstrate the existence of multiple regimes in the global climate system, as do long-term empirical data [9]. These complex models have also identified possible regime boundaries, especially with respect to atmospheric concentrations of carbon dioxide [16]. Boundaries have also been from empirical data through the structure of the noise of the system; variability in the system increases as the system nears a regime boundary [19]. Hysteresis between multiple stable regimes have been observed and quantified for regional climate systems (e.g., [34]).

Feedbacks between endogenous and exogenous variables can influence (and even create) regime boundaries. Although at regional scales global climate mechanisms are exogenous, they are influenced by ecosystem processes, such as through surface albedo (decreased by vegetation through absorption of solar radiation) or precipitation patterns (increased by vegetation through evapotranspiration, [22]). While thermohaline ocean currents exogenously determine ecosystem regimes (particularly along warm currents), feedbacks between vegetation and regional climate can create hysteresis at regime boundaries [8, 15]. For example, in the Amazon basin in Brazil, high temperatures combine with high humidity created by evapotranspiration from tropical forests to stimulate cumulous cloud formation, resulting in ecosystem-generated precipitation [40]; forest fires, more common in arid systems, decrease cloud formation and precipitation [2] and hence forest regeneration. Savanna vegetation in the Sahel in Africa modifies albedo and holds soil in place, delaying a shift to desert beyond the threshold at which a shift in ocean currents would cause desertification (and vice versa; [12, 15]). In some of these systems, human activities can trigger regime shifts or maintain regimes [8, 9], although the strength of human influence may not be capable of dominating non-human drivers in all systems [12].

16.3.4 Economics

Multiregime dynamics (dynamics ranging across regimes) can be seen as a generalization of the notion of a business cycle or any other economic oscillation comprised of interconnected, well-defined phases. The generalization is two-fold: more and qualitatively different phases are permitted in the conventional business cycle; and such phases may not be interconnected through a well defined mechanism, yielding a qualitatively regular or predictable sequence (such as ups and downs). The latter is a consequence of the possibility of numerous regimes through which a system may pass, the general term regime replacing phase to indicate the absence of a necessary, ordered concatenation.

We may distinguish two approaches, though they share an evolutionary viewpoint and are rooted in the economics tradition of Schumpeter, with his theory of dynamics as implying intermittent economic change of a qualitative type. One approach takes a long-term view on how phases may link up together in the time evolution of societies in an evolutionary sequel, or how this may break down due to internal or external mechanisms (e.g. migration, overpopulation, the evolution or import of technologies, imperial expansion). In such multi-phase dynamics [10], sudden phase shifts take place at an ideal borderline between the domains of two adjacent local models of the system, but they in a sense come one at a time. This is similar to the other approach, if limited to the idea of structural change. In fact, conceived as a qualitative phenomenon affecting the overall behavior of a system, structural change has to be thought of (and modeled) as a sudden change in the very model of the system behavior, if model stands for the set of rules governing it, and the representation of inner mechanisms resulting from its wiring and architecture. In stimulating episodes of structural change (in this sense, singular regime shifts) a variety of mechanisms, not solely of a strict economic nature, may be working together. This suggests a more comprehensive and interdisciplinary view of economics as societal evolution.

The other approach produces a uniform account of a phenomenon emerging from the comparative literature on growth: the variety of patterns exhibited by different countries, or regions, or even sectors within the same country (horizontal variety); and at the same time the appearance of repeated, though irregular, qualitative changes in the history of any one exemplary system (vertical variety). A uniform framework to accommodate and account for both types of phenomena has to be based upon the acceptance of the implications of relevant empirical evidence. It has proved difficult to discover the emergence of the same dynamic model across countries; moreover there is no evidence of a unique model dominating or being typical in any system's history. In other words, embedded in any actual history are repeated episodes of structural change. These findings seem to contradict predictions of well established growth theory, as they violate its two postulates: of convergence to a unique type of behavior; and stability of the unique implied equilibrium.

The framework of multi-regime dynamics [5], where a regime is a model plus its own domain in the system state space, was created to handle this sort of double variability. Mathematically, it converts ordinary dynamics in a continuous space of real valued states into a dynamics over a discrete space, thinly populated by regimes. In addition to growth theory and empirics, the framework can be applied to inflationary processes that at one point enter an hyperinflationary state, after a sudden jump in the rate of price changes, and how they may be tamed with various policy measures or simply get exhausted endogenously. It an also be applied to the behavior of stock market and asset prices, where the key theorem asserts that the latter follow a white noise process. In the former example there are two states, inflation and hyperinflation (indexed by the values of the derivative of a price index, against the second or acceleration derivative). In the latter the states are bull and bear markets, but these states are themselves processes with their own dynamical laws or models. The notion of states as qualitative behaviors, with multiplicity and inherent instability, has played an increasing role in monetary theory and analysis, in rates of exchange, and in many other similar settings, including some game theoretic ones. In principle, they are all amenable to a reformulation in terms of the present multi-regime approach, and with regimes identified with specific point attractors as special cases. In the theoretical analysis of the very complex situations involved, where various dynamical layers (of adjustment and structural change) may take place at the same time, computational experiments with a variety of modeling settings have to be developed and play a key role [14].

In cross country empirical analysis, a Markov-type of approach has been developed to identify convergence to different long run behaviors defined as regimes, rather than the single attractors of the current literature [5]. As applied to a single system's history and time series data, it proved useful to use a coding or symbolic dynamics technique, which transforms irregular time series data into a sequence of symbols from a chosen finite alphabet, each symbol being assigned to a distinct regime domain. Coding renders the dimension of the original state space irrelevant, and does away with the need to search for dimension reducing conditions (e.g., stationarity of time series, the related co-integration technique). At the same time, in symbolic representation we may discover (near) regularities that may not be obvious in the original real valued and vectorial time series. At the symbolic or regime level of dynamics, the technique produces a deterministic view of statistical irregularity, upon which measures of embodied or inherent stochasticity can be developed [5].

To illustrate one implication of the approach as applied to economic growth, it is useful to pre-assume exogenously (on the basis of theory alone) the presence of six regimes, and code them accordingly. Empirical screening of actual data shows in many cases that six is too many, hence suggesting a more appropriate, endogenous partition into regimes. In most applications, such as the inflation example above, or to the variability of asset prices, two regimes can most often be identified endogenously, and therefore a binary representation suffices.

16.4 Conclusions

A multidisciplinary meta-analysis of dynamic regime models yields several benefits. If endogenous variables behave similarly regardless of the source of exogenous pressures, and of the scale at which the system is define, then general models, rules and coded behaviors can be developed. For example, do droughts exogenously cause grasslands to shift to shrubland in the same manner that oil supply shocks cause recessions in national economies? The identification of transdisciplinary, scale-invariant rules for the order in which systems can pass through multiple regimes would help advance systems research within and across many disciplines. Of course, these comparisons must first ascertain whether the systems are in fact best described through the regime concept; some models, such as periodic forcing, can also result in a system shifting between two or more states [30].

Furthermore, if the same basic theory regarding system behavior (including rapid regime change) applies across disciplines at multiple spatiotemporal scales, then models developed from these theories may help manage those systems which, at larger scales, cross traditional disciplinary lines. Indeed, the scale at which the system boundaries are delineated will determine which variables are considered exogenous [18]. For example, while the resilience of a particular forest stand as a forest may be dependent on hydrologic and climate factors, over a longer time period evolutionary changes or the extinction of dominant tree species may eliminate the existence of particular communities as potential regimes. Over a larger area, forests are influenced by a variety of both environmental and anthropogenic forces, including international trade patterns [24]. The resilience of forests at the country level are exogenously influenced by exchange rates between trading partners, economic growth (which influences demand for wood products), and technological advancements in forestry sectors which can increase or decrease the amount of wood needed to make those products. At this scale, hydrology and climate have now become endogenous to the system, particularly when they are manipulated by human activities. Similar multidisciplinary forces exist for lakes [7].

More than two regimes are possible for many dynamic systems (including those above), although for some systems observing shifts between more than two regimes may require too long of a time period relative to human lifetimes or historic records. However, human systems are within the time scale of our observation, so that for instance economies and governments provide examples of systems with well-known multiple regimes and shifts between them. In these examples, shifts between particular combinations or sequences of regimes may often seem to be the norm, rather than a random exploration among all possible regimes.

Although in many articles the application of regime theory to systems is theoretical and abstract, there is considerable interest in the application of regime theory to system management. For ecosystems, this utility and implications of this approach have been widely investigated [23, 30, 35, 42]. Ecologists have used the theory to identify both the internal mechanisms which can increase the resilience of a particular regime (such as a clear lake [6] or a grassland suitable for grazing [3]), and also the thresholds at which external pressures (usually anthropogenic) can overwhelm these internal stabilizing mechanisms and cause a regime shift. Dynamic regime concepts may help develop more successful marriage counseling methods, as counselors can help couples develop new, more stable attractors after significant periods of instability [13, 49]. The political stability of nations may also be best described as stable or instable regimes, particularly with respect to the types of external perturbations the system of government is able to withstand [17], although this theoretical knowledge has not been applied to real systems. Dynamic regimes theory offers many disciplines a robust and innovative framework for understanding and managing complex systems.

Acknowledgements This paper was initially conceived at a 2005 workshop in Cincinnati, Ohio, sponsored by the US Environmental Protection Agency, Office of Research and Development. JGB was supported by the Free University of Bolzano (project "Dynamical Regimes in Economics: modelling and statistical tools").

References

1. Alley, R.B., Marotzke, J., Nordhaus, W.D., Overpeck, J.T., Peteet, D.M., Pielke Jr., R.A., Pierrehumbert, R.T., Rhines, P.B., Stocker, T.F., Talley, L.D., Wallace, J.M.: Abrupt climate change. Science **299**, 2005–2010 (2003)
2. Andreae, M.O., Rosenfeld, D., Artaxo, P., Costa, A.A., Frank, G.P., Longo, K.M., Silva-Dias, M.A.F.: Smoking rain clouds over the Amazon. Science **303**, 1337–1342 (2004)
3. Bestelmeyer, B.T., Herric, J.E., Brown, J.R., Trujillo, D.A., Havstad, K.M.: Land management in the American Southwest: A state-and-transition approach to ecosystem complexity. Environ. Manage. **34**, 38–51 (2004)
4. Brida, J.G.: A model of inflation and unemployment with multiple regimes. Intl. Math. Forum **1**, 1125–1144 (2006)
5. Brida, J.G., Punzo, L.F.: Symbolic time series analysis and dynamic regimes. Struct. Change Econ. Dyn. **14**, 159–183 (2003)
6. Carpenter, S.R.: Regime Shifts in Lake Ecosystems: Pattern and Variation. International Ecology Institute, Oldendorf/Luhe, Germany (2003)
7. Carpenter, S., Walker, B., Anderies, J.M., Abel, N.: From metaphor to measurement: Resilience of what to what? Ecosystems **4**, 765–781 (2001)
8. Chapin III, F.S., Callaghan, T.V., Bergeron, Y., Fukuda, M., Johnstone, J.F., Juday, G., Zimov, S.A.: Global change and the boreal forest: thresholds, shifting states or gradual change? Ambio **33**, 361–365 (2004)
9. Clark, P.U., Pisias, N.G., Stocker, T.F., Weaver, A.J.: The role of thermohaline circulation in abrupt climate change. Nature **415**, 863–869 (2002)

10. Day, R.H.: Multi-phase economic dynamics. In: Maruyama, T., Takahashi, W. (eds.) Lecture Notes in Economics and Mathematical Systems, pp. 25–46. Springer, Berlin (1995)
11. Eriksson, E.: Air–Ocean–Icecap interactions in relations to climatic fluctuations and glaciation cycles. Meteor. Mon. **8**, 68–92 (1968)
12. Giannini, A., Saravana, R., Chang, P.: Oceanic forcing of Sahel rainfall on interannual to interdecadal time scales. Science **302**, 1027–1030 (2003)
13. Gottman, J.M., Murray, J.D., Swanson, C.C., Tyson, R., Swanson, K.R.: The Mathematics of Marriage: Dynamic Nonlinear Models. MIT, Cambridge, MA (2002)
14. Tesfatsion, L., Judd, KL. (eds). Handbook of Computational Economics, vol. 2: Agent-Based Computational Economics. Elsevier/North-Holland: Amsterdam (2006)
15. Higgins, P.A.T., Mastrandrea, M.D., Schneider, S.H.: Dynamics of climate and ecosystem coupling: abrupt changes and multiple equilibria. Philos. Trans. R. Soc. Lond. B Biol. Sci. **357**, 647–655 (2002)
16. IPCC (Intergovernmental Panel on Climate Change): Climate Change 2007: The physical science basis. In: Solomon, S., Quin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K.B., Tignor, M., Miller, H.L. (eds.) Cambridge University Press, Cambridge, UK and New York, (2007)
17. King, G., Zeng, L.: Improving forecasts of state failure. World Polit. **53**, 623–658 (2001)
18. Kinzig, A.P., Ryan, P., Etienne, M., Allison, H., Elmqvist, T., Walker, B.H.: Resilience and regime shifts: assessing cascading effects. Ecol. Soc. **11**, 20 [online] (2006)
19. Kleinen, T., Held, H., Petschel-Held, G.: The potential role of spectral properties in detecting thresholds in the Earth system: application to the thermohaline circulation. Ocean Dyn. **53**, 53–63 (2003)
20. Liberzon, D.: Switching in Systems and Control. Birkhäuser, Boston, MA (2003)
21. Mantua, N.: Methods for detecting regime shifts in large marine ecosystems: a review with approaches applied to North Pacific data. Prog. Oceanogr. **60**, 165–182 (2004)
22. Marchant, R., Hooghiemstra, H.: Rapid environmental change in African and South American tropics around 4000 years before present: review. Earth-Sci. Rev. **66**, 217–260 (2004)
23. Mayer, A.L., Rietkerk, M.: The dynamic regime concept for ecosystem management and restoration. BioScience **54**, 1013–1020 (2004)
24. Mayer, A.L., Kauppi, P.E., Angelstam, P.K., Zhang, Y., Tikka, P.M.: Importing timber, exporting ecological impact. Science **308**, 359–360 (2005)
25. Miller, G., Mangan, J., Pollard, D., Thompson, S., Felzer, B., Magee, J.: Sensitivity of the Australian Monsoon to insolation and vegetation: Implications for human impact on continental moisture balance. Geology **33**, 65–68 (2005)
26. Milnor, J., Thurston, W.: On iterated maps of the interval. Lect. Notes Math. **1342**, 465–563 (1988)
27. Nowak, A., Vallacher, R.R.: Dynamical Social Psychology. Guilford, NY (1998)
28. Nusse, H.E., Yorke, J.A.: The structure of basins of attraction and their trapping regions. Ergod. Theory Dyn. Syst. **17**, 463–481 (1997)
29. Nyström, M., Folke, C., Moberg, F.: Coral reef disturbance and resilience in a human-dominated environment. Trends Ecol. Evol. **15**, 413–417 (2000)
30. Overland, J.E., Percival, D.B., Mofjeld, H.O.: Regime shifts and red noise in the North Pacific. Deep-Sea Res. I Ogeanogr. Res. Pap. **53**, 582–588 (2006)
31. Peters, D.P.C., Pielke Sr., R.A., Bestelmeyer, B.T., Allen, C.D., Munson-McGee, S., Havstad, K.M.: Cross-scale interactions, nonlinearities, and forecasting catastrophic events. Proc. Natl. Acad. Sci. USA **101**, 15130–15135 (2004)
32. Pierrehumbert, R.T.: Climate change and the tropical Pacific: the sleeping dragon wakes. Proc. Natl. Acad. Sci. USA **97**, 1355–1358 (2000)
33. Poon, L., Campos, J., Ott, E., Grebogi, C.: Wada basin boundaries in chaotic scattter. Int. J. Bifurcat. Chaos. **6**, 251–265 (1996)
34. Rahmstorf, S., Crucifix, M., Ganopolski, A., Goosse, H., Kamenkovich, I., Knutti, R., Lohmann, G., Marsh, R., Mysak, L.A., Wang, Z., Weaver, A.J.: Thermohaline circulation hysteresis: A model intercomparison. Geophys. Res. Lett. **32**, L23605 (2005)

35. Rietkerk, M., Dekker, S.C., de Ruiter, P.C., van de Koppel, J.: Self-organized patchiness and catastrophic shifts in ecosystems. Science **305**, 1926–1929 (2004)
36. Rinaldi, S., Gragnani, A.: Love dynamics between secure individuals: a modeling approach. Nonlinear Dyn. Psychol. Life Sci. **2**, 283–301 (1998)
37. Rind, D.: The sun's role in climate variations. Science **296**, 673–677 (2002)
38. Scheffer, M., Carpenter, S., Foley, J.A., Folke, C., Walker, B.: Catastrophic shifts in ecosystems. Nature **413**, 591–596 (2001)
39. Srinivasu, P.D.N.: Regime shifts in eutrophied lakes: a mathematical study. Ecol. Modell **179**, 115–130 (2004)
40. Sternberg, L.dS.L.: Savanna-forest hysteresis in the tropics. Glob. Ecol. Biogeogr. **10**, 369–37 (2001)
41. Stewart, I.: Regime change in meteorology. Nature **422**, 571–573 (2003)
42. Suding, K.N., Gross, K.L., Houseman, G.R.: Alternative states and positive feedbacks in restoration ecology. Trends Ecol. Evol. **19**, 46–53 (2004)
43. Tesser, A., Achee, J.: Aggression, love, conformity, and other social psychological catastrophes. In: Vallacher, R.R., Nowak, A. (eds.) Dynamical Systems in Social Psychology, pp. 96–109. Academic, San Diego (1994)
44. Thom, R.: Structural Stability and Morphogenesis: An Outline of a General Theory of Models. Westview, Perseus, New York (1972)
45. van Nes, E.H., Scheffer, M.: Implications of spatial heterogeneity for catastrophic regime shifts in ecosystems. Ecology **86**, 1797–1807 (2005)
46. Venegas, J.G., Winkler, T., Musch, G., Vidal Melo, M.F., Layfield, D., Tgavalekos, N., Fischman, A.J., Callahan, R.J., Bellani, G., Harris, R.S.: Self-Organized Patchiness in Asthma as a Prelude to Catastrophic Shifts. Nature **434**, 777–782 (2005)
47. Walker, B., Holling, C.S., Carpenter, S.R., Kinzig, A.: Resilience, adaptability and transformability in social–ecological systems. Ecol. Soc. **9**, 5 [online] (2004)
48. Weart, S.R.: The Discovery of Global Warming. Harvard University Press, Cambridge, MA (2003)
49. Weigel, D., Murray, C.: The paradox of stability and change in relationships: What does chaos theory offer for the study of romantic relationships? J. Soc. Pers. Relat. **17**, 425–449 (2000)

Chapter 17
General Relativistic Elasticity: Statics and Dynamics of Spherically Symmetric Metrics

Irene Brito and E.G.L.R. Vaz

Abstract An introduction is provided to the theory of elasticity in general relativity. Important tensors appearing in this context are presented. In particular, attention is focussed on the elasticity difference tensor, for which an algebraic analysis is performed. Applications are given to static and non-static spherically symmetric configurations. For the latter, dynamical equations are obtained characterizing the space-time in the context of general relativistic elasticity.

17.1 General Relativistic Elasticity

General relativistic elasticity was formulated in the mid-twentieth century due to the necessity to study astrophysical problems such as deformations of neutron star crusts. Relevant contributions to the theory of general relativistic elasticity were given by Carter and Quintana [1], Kijowski and Magli [2], Beig and Schmidt [3], Karlovini and Samuelsson [4] and by many other authors.

The theory is based on a configuration mapping

$$\Psi : M \longrightarrow X,$$

a C^k $(k > 1)$ mapping from space-time M, endowed with a Lorentz metric g of signature $(-,+,+,+)$ and assumed to be time-orientable, to the material space X. The material space is a three-dimensional manifold, whose points represent the particles of the material. The material metric K defined on X measures the distances between particles in the locally relaxed state of the material. Coordinates on M are here denoted by $\{\omega^a\}$, $a = 0,1,2,3$, and coordinates on X by $\{\xi^A\}$, $A = 1,2,3$. Associated with Ψ are the pull-back operator Ψ^* and the push-forward operator

I. Brito (✉) and E.G.L.R. Vaz
Departamento de Matemática para a Ciência e Tecnologia, Universidade do Minho 4800 058 Guimarães, Portugal
e-mail: ireneb@mct.uminho.pt, evaz@mct.uminho.pt

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_17,

Ψ_* which give rise to a 3×4 matrix, the relativistic deformation gradient, whose entries are $\xi^A_a = \frac{\partial \xi^A}{\partial \omega^a}$. The velocity field of the matter, u^a, satisfies the following conditions: $u^0 > 0$, $u^a u_a = -1$ and $u^a \xi^A_a = 0$. The pulled-back material metric $k_{ab} = \xi^A_a \xi^B_b K_{AB}$ is such that $k_{ab}u^a = 0$ and $\mathscr{L}_u k_{ab} = 0$. It is used to construct other relativistic elastic tensors. Let n_1^2, n_2^2, n_3^2 be the eigenvalues of $k^a{}_b$, then one can write $k_{ab} = n_1^2 x_a x_b + n_2^2 y_a y_b + n_3^2 z_a z_b$, where x, y and z denote the eigenvectors of k and n_1, n_2, n_3 represent the linear particle densities (see [4]). Considering the orthonormal tetrad $\{u, x, y, z\}$, then the space-time metric takes the form $g_{ab} = -u_a u_b + x_a x_b + y_a y_b + z_a z_b$ and $h_{ab} = x_a x_b + y_a y_b + z_a z_b$ is the projection tensor.

The relativistic strain tensor $s_{ab} = \frac{1}{2}(h_{ab} - k_{ab})$ contains information about the local state of strain of the matter. The material is said to be locally relaxed at a particular point of space-time if $s_{ab} = 0$.

The elasticity difference tensor $S^a{}_{bc}$ introduced by [4] can be expressed as

$$S^a_{bc} = \frac{1}{2}k^{-am}(D_b k_{mc} + D_c k_{mb} - D_m k_{bc}), \tag{17.1}$$

where k^{-1am} is such that $k^{-1am}k_{mb} = h^a_b$ and D_b is the spatially projected connection defined by $D_a t^{b\dots}{}_{c\dots} = h^d{}_a h^b{}_e \dots h^f{}_c \dots \nabla_d t^{e\dots}{}_{f\dots}$, where $t^{b\dots}{}_{c\dots}$ is an arbitrary tensor field, and it satisfies $D_a h_{bc} = 0$. A mathematical analysis of the elasticity difference tensor is presented in [5]. It is decomposed along the eigenvectors of $k^a{}_b$ as follows

$$S^a{}_{bc} = \underset{1}{M}_{bc} x^a + \underset{2}{M}_{bc} y^a + \underset{3}{M}_{bc} z^a; \tag{17.2}$$

and for the three second-order symmetric tensors $\underset{1}{M}$, $\underset{2}{M}$ and $\underset{3}{M}$ the eigenvalue-eigenvector problem is studied. In particular, conditions are investigated for the three eigenvectors, x, y, z, of the pulled-back material metric to be eigenvectors for $\underset{1}{M}$, $\underset{2}{M}$ and $\underset{3}{M}$.

Here, the algebraic analysis of the elasticity difference tensor is carried out for a static and a non-static spherically symmetric space-time.

17.2 Applications to Static and Dynamical Configurations

17.2.1 Static Spherically Symmetric Space-time

Consider a static spherically symmetric space-time with g given by the line-element

$$ds^2 = -e^{2\nu(r)}dt^2 + e^{2\lambda(r)}dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\phi^2$$

Table 17.1 Eigenvectors and eigenvalues for $\underset{1}{M}$, $\underset{2}{M}$ and $\underset{3}{M}$

	Eigenvectors	Eigenvalues
$\underset{1}{M}$	x	$\mu_1 = \frac{e^{-\lambda}}{n_1} n_1'$
	y	$\mu_2 = \frac{e^{-\lambda}}{r} - \frac{e^{-\lambda}}{r} \frac{n_2^2}{n_1^2} - e^{-\lambda} \frac{n_2}{n_1^2} n_2'$
	z	$\mu_3 = \mu_2$
$\underset{2}{M}$	$x + y$	$\mu_4 = \frac{e^{-\lambda}}{n_2} n_2'$
	$x - y$	$\mu_5 = -\frac{e^{-\lambda}}{n_2} n_2'$
	z	$\mu_6 = 0$
$\underset{3}{M}$	$x + z$	$\mu_7 = \mu_4$
	$x - z$	$\mu_8 = \mu_5$
	y	$\mu_9 = 0$

and with coordinates $\omega^a = \{t, r, \theta, \phi\}$. The space-time can be specified by the orthonormal tetrad $\{u, x, y, z\}$ using the basis vectors $u^a = \left[\frac{1}{e^{\nu(r)}}, 0, 0, 0\right]$, $x^a = \left[0, \frac{1}{e^{\lambda(r)}}, 0, 0\right]$, $y^a = \left[0, 0, \frac{1}{r}, 0\right]$ and $z^a = \left[0, 0, 0, \frac{1}{r \sin\theta}\right]$. Due to the spherical symmetry, on X the coordinates are $\xi^A = \{\tilde{r}, \tilde{\theta}, \tilde{\phi}\}$ where $\tilde{r} = \tilde{r}(r)$, $\tilde{\theta} = \theta$ and $\tilde{\phi} = \phi$. The non-zero components of the deformation gradient are $\frac{d\xi^1}{d\omega^1} = \tilde{r}'$, $\frac{d\xi^2}{d\omega^2} = 1$, $\frac{d\xi^3}{d\omega^3} = 1$, where a prime represents a derivative with respect to r, and the line-element of the pulled-back material metric is $ds^2 = \tilde{r}'^2\, dr^2 + \tilde{r}^2\, d\theta^2 + \tilde{r}^2\, sin^2\theta\, d\phi^2$. Calculating the eigenvalues of ${k^a}_b$, one obtains $n_1^2 = \tilde{r}'^2\, e^{-2\lambda}$ and $n_2^2 = n_3^2 = \frac{\tilde{r}^2}{r^2}$. The strain tensor has three non-zero components: s_{rr}, $s_{\theta\theta}$, $s_{\phi\phi}$, and it vanishes if and only if $\tilde{r} = c\, e^{\int \frac{e^\lambda}{r} dr}$, $c > 0$. Solving the eigenvalue-eigenvector problem for $\underset{1}{M}$, $\underset{2}{M}$ and $\underset{3}{M}$, building up the elasticity difference tensor in (17.2), leads to the results listed in Table 17.1.

17.2.2 Non-Static Spherically Symmetric Space-Time

Consider a non-static spherically symmetric space-time, whose metric g is given by the line-element $ds^2 = -e^{2\nu(t,r)} dt^2 + e^{2\lambda(t,r)} dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\phi^2$. On M the coordinates are $\omega^a = \{t, r, \theta, \phi\}$. The space-time can be specified by defining the orthonormal tetrad $\{u, x, y, z\}$ with the following basis vectors: $u^a = \left[e^{-\nu}\, \gamma, -e^{-\nu}\, \frac{\dot{\tilde{r}}}{\tilde{r}'}\, \gamma, 0, 0\right]$, $x^a = \left[-e^{\lambda-2\nu}\, \frac{\dot{\tilde{r}}}{\tilde{r}'}\, \gamma, e^{-\lambda}\, \gamma, 0, 0\right]$, $y^a = \left[0, 0, \frac{1}{r}, 0\right]$ and $z^a = \left[0, 0, 0, \frac{1}{r \sin\theta}\right]$, where $\gamma = \sqrt{\frac{e^{2\nu}\, \tilde{r}'^2}{e^{2\nu}\, \tilde{r}'^2 - e^{2\lambda}\, \dot{\tilde{r}}^2}}$ and a dot represents a derivative with respect to t. In this case, the coordinates on X are $\xi^A = \{\tilde{r}, \tilde{\theta}, \tilde{\phi}\}$, where $\tilde{r} = \tilde{r}(t, r)$, $\tilde{\theta} = \theta$ and $\tilde{\phi} = \phi$, so that the non-zero components of the relativistic deformation gradient take the form $\frac{\partial\xi^1}{\partial\omega^0} = \dot{\tilde{r}}$, $\frac{\partial\xi^1}{\partial\omega^1} = \tilde{r}'$, $\frac{\partial\xi^2}{\partial\omega^2} = 1$, $\frac{\partial\xi^3}{\partial\omega^3} = 1$. The line-element of the pulled-back material metric is given by

Table 17.2 Eigenvectors and eigenvalues for $\underset{1}{M}$, $\underset{2}{M}$ and $\underset{3}{M}$

	Eigenvectors	Eigenvalues
$\underset{1}{M}$	x	$\mu_1 = \frac{e^{2\nu}\tilde{r}'n_1' - e^{2\lambda}\dot{\tilde{r}}\dot{n}_1}{e^{\lambda+\nu}n_1}\sqrt{\frac{1}{e^{2\nu}\tilde{r}'^2 - e^{2\lambda}\dot{\tilde{r}}^2}}$
	y	$\mu_2 = \frac{rn_2(e^{2\lambda}\dot{\tilde{r}}\dot{n}_2 - e^{2\nu}\tilde{r}'n_2') + \tilde{r}'e^{2\nu}(n_1^2 - n_2^2)}{e^{\lambda+\nu}rn_1^2}\sqrt{\frac{1}{e^{2\nu}\tilde{r}'^2 - e^{2\lambda}\dot{\tilde{r}}^2}}$
	z	$\mu_3 = \mu_2$
$\underset{2}{M}$	$x+y$	$\mu_4 = -\frac{e^{2\lambda}\dot{\tilde{r}}\dot{n}_2 - e^{2\nu}\tilde{r}'n_2'}{e^{\lambda+\nu}n_2}\sqrt{\frac{1}{e^{2\nu}\tilde{r}'^2 - e^{2\lambda}\dot{\tilde{r}}^2}}$
	$x-y$	$\mu_5 = \frac{e^{2\lambda}\dot{\tilde{r}}\dot{n}_2 - e^{2\nu}\tilde{r}'n_2'}{e^{\lambda+\nu}n_2}\sqrt{\frac{1}{e^{2\nu}\tilde{r}'^2 - e^{2\lambda}\dot{\tilde{r}}^2}}$
	z	$\mu_6 = 0$
$\underset{3}{M}$	$x+z$	$\mu_7 = \mu_4$
	$x-z$	$\mu_8 = \mu_5$
	y	$\mu_9 = 0$

$ds^2 = -\dot{\tilde{r}}'^2\,dt^2 + 2\dot{\tilde{r}}\,\tilde{r}'\,dt\,dr + \tilde{r}'^2\,dr^2 + \tilde{r}^2\,d\theta^2 + \tilde{r}^2\,sin^2\theta\,d\phi^2$. Calculating the eigenvalues of $k^a{}_b$, one obtains $n_1^2 = \tilde{r}'^2\,e^{-2\lambda} - \dot{\tilde{r}}^2\,e^{-2\nu}$ and $n_2^2 = n_3^2 = \frac{\tilde{r}^2}{r^2}$. The strain tensor has three more components than in the static case: s_{tt}, s_{tr}, s_{rr}, $s_{\theta\theta}$, $s_{\phi\phi}$, and it vanishes if and only if the following condition involving the functions and λ, μ and the material radius is satisfied: $\tilde{r}'^2 e^{-2\lambda} - \dot{\tilde{r}}^2 e^{-2\nu} = \frac{\tilde{r}^2}{r^2}$. Solving the eigenvalue-eigenvector problem in this case, one obtains the results listed in Table 17.2.

17.2.3 Concluding Remarks

Comparing the results obtained for the static case and for the non-static case, the following conclusions and remarks can be drawn.

For spherically symmetric space-times, passing from a static to a non-static configuration preserves the behaviour of the eigenvectors of the pulled-back material metric k for the tensors $\underset{1}{M}$, $\underset{2}{M}$ and $\underset{3}{M}$ building up the elasticity difference tensor: x, y, z are eigenvectors for $\underset{1}{M}$; $x+y$, $x-y$, z are eigenvectors for $\underset{2}{M}$; $x+z$, $x-z$, y are eigenvectors for $\underset{3}{M}$. In particular, the eigenvectors y and z of k remain the same for both configurations, only x changes. Furthermore, in the non-static case we can observe that the velocity field of matter u depends on the material radius; all relativistic elastic quantities (k_{ab}, n_1^2, n_2^2, s_{ab}, $S^a{}_{bc}$) are time-dependent through λ, ν and the material radius $\tilde{r}$; the condition to be satisfied for the strain tensor to vanish involve the functions ν and $\dot{\tilde{r}}$ in addition to λ and $\tilde{r}'$.

References

1. Carter, B., Quintana, H.: Proc. R. Soc. A **331**, 57 (1972)
2. Kijowski, J., Magli, G.: J. Geom. Phys. **9**, 207–223, (1992)
3. Beig, R., Schmidt, B.: Class. Quantum Grav. **20**, 889–904 (2003)
4. Karlovini, M., Samuelsson, L.: Class. Quantum Grav. **20**, 3613–3648 (2003)
5. Vaz, E.G.L.R., Irene, B.: Gen. Rel. Grav. **40**, 1947–1966 (2008)

Chapter 18
Post-Inflationary Scalar Field Phase Dynamics

T. Charters, A. Nunes, and J.P. Mimoso

Abstract We present a brief summary of the results of Charters et al. [1] where a simple model of a massive inflation field ϕ coupled to another scalar filed χ with interaction term $g^2\phi^2\chi^2$ for the first stage of preheating, and we give a full description of the dynamics of the χ field modes, including the behaviour of the phase, in terms of the iteration of a simple family of circle maps.

18.1 Scalar Field Dynamics

The reheating mechanism was proposed as a period, immediately after inflation, during which the inflation field ϕ oscillates and transfers its energy into ultra-relativistic matter and radiation, here modelled by another scalar field χ. Consider the potential $V(\phi) = 1/2m_{\phi}^2\phi^2$ and interaction potential [2–5] $V_{int}(\phi, \chi) = g^2\phi^2\chi^2$. The evolution of the flat Friedman–Robertson–Walker (FRW) universe is given by

T. Charters (✉)
Área Departamental de Matemática, Instituto Superior de Engenharia de Lisboa, Rua Conselheiro Emídio Navarro, 1, 1949-014 Lisbon, Portugal
and
Centro de Astronomia e Astrofísica da Universidade de Lisboa, Campo Grande, Edifício C8 P-1749-016 LISBOA Portugal
e-mail: tca@cii.fc.ul.pt

A. Nunes
Centro de Física da Matéria Condensada da Universidade de Lisboa, Departamento de Física, Campo Grande, Edifício C8 P-1749-016 LISBOA Portugal
e-mail: anunes@ptmat.fc.ul.pt

J.P. Mimoso
Centro de Astronomia e Astrofísica da Universidade de Lisboa, Departamento de Física, Campo Grande, Edifício C8 P-1749-016 LISBOA Portugal
e-mail: jpmimoso@cii.fc.ul.pt

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_18,

$$3H^2 = \frac{8\pi}{m_{pl}^2}\left(\frac{1}{2}\dot{\phi}^2 + V(\phi) + \frac{1}{2}\dot{\chi}^2 + g^2\phi^2\chi^2\right), \tag{18.1}$$

where $H = \dot{R}/R$ and R is the FRW scalar factor. The equations of motion in a FRW universe for a homogeneous scalar field ϕ coupled to the k-mode of the χ field are given by

$$\ddot{\phi} + 3H\dot{\phi} + \left(m_\phi^2 + g^2\chi_k^2\right)\phi = 0, \tag{18.2}$$

$$\ddot{\chi}_k + 3H\dot{\chi}_k + \omega_k^2(t)\chi_k = 0, \tag{18.3}$$

where $\omega_k^2(t) = k^2/R^2 + g^2\phi^2$.

In Minkowski space-time we set $H = 0$ and $R = 1$ in (18.2) and (18.3). To approximate in the broad resonance regime the solution of (18.3) in Minkowski space-time we set

$$\ddot{\chi}_k + \omega_k^2(t)\chi_k = 0, \tag{18.4}$$

where $\omega_k^2(t) = a_k + b\sin^2(t)$ with $a_k = k^2/m_\phi^2$ and $b = g^2A^2/m_\phi^2$, A is the constant amplitude of the field ϕ, and the time variable is now $t \to m_\phi t$. Typical values of the parameters [2, 6] are $g^2 \leq 10^{-6}$, $m = 10^{-6}m_{pl}$, $A = \alpha m_{pl}$, where $0 < \alpha < 1$, and thus $b \leq \alpha^2 \times 10^6$. In the broad ressonance regime we have, $\sqrt{b} \gg 1$, and it is possible to construct an approximate global solution

$$\chi_k^j(t; \alpha_k^j, \beta_k^j) = \frac{\alpha_k^j}{\sqrt{2\omega_k(t)}}e^{-i\int_0^t \omega_k(s)ds} + \frac{\beta_k^j}{\sqrt{2\omega_k(t)}}e^{i\int_0^t \omega_k(s)ds}, \tag{18.5}$$

which is valid except in the neighbourhood of $t_j = j\pi$, $j = 0, 1, \ldots$ (where $\phi \sim 0$). The parameters (α_k^j, β_k^j) for consecutive j are determined by the behaviour of the solution of (18.3) in Minkowski space-time for t close to t_j. In terms of the phase $\nu_k^j = \arg\beta_k^j + \theta_k^j$ of the field χ_k when $t = t_j$, one gets [2]

$$\nu_k^{j+1} = \theta(b,\kappa) + \arg\left(\sqrt{1+\rho_\kappa^2}e^{-i\varphi_\kappa}e^{i\nu_k^j} - i\rho_\kappa e^{-i\nu_k^j}\right), \tag{18.6}$$

where $\kappa^2 = a_k/\sqrt{b}$, $\theta(b,\kappa) = \int_0^\pi \omega(s)ds$, with $\omega_\kappa^2(s) = \kappa^2 + \sin^2(s)$, $\rho_\kappa = \exp(-\pi\kappa^2/2)$, $\varphi_\kappa = \arg(\Gamma((1+i\kappa^2)/2)) + \kappa^2/2(1+\ln(2/\kappa^2))$, and $\kappa = k/\sqrt{Agm_\phi}$. Since $n_k = |\beta_k|^2$, the growth index μ_κ^j, defined by $n_k^{j+1} = n_k^j\exp(2\pi\mu_\kappa^j)$, is given in [2], in terms of the phase $\nu_k^j = \arg\beta_k^j + \theta_k^j$ of the field χ_k when $t = t_j$,

$$\mu_\kappa^j = \frac{1}{2\pi}\ln\left(1 + 2\rho_\kappa^2 - 2\rho_\kappa\sqrt{1+\rho_\kappa^2}\sin(-\varphi_\kappa + 2\nu_k^j)\right). \tag{18.7}$$

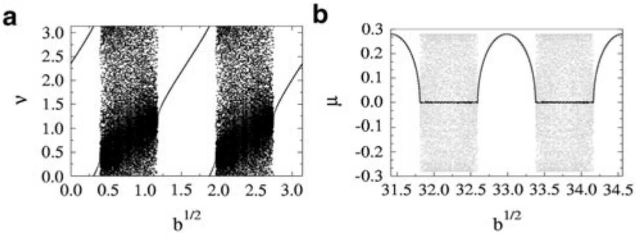


Fig. 18.1 (**a**) Bifurcation diagram of the family of circle maps (18.8) for $\sqrt{b} \in [0, \pi]$. (**b**) The asymptotic value of μ_0 as a function of b for $\sqrt{b} \in [10\pi, 11\pi]$ computed analytically (*full line*) and numerically (*dotted line*). Also shown (*in grey*) are all the values of μ_0^j, $j = 100, 101, \ldots, 200$

The dynamics of the growth rate depends on the phase dynamics but the dynamics of ν can be studied separately.

The properties of (18.6) are best understood by looking at the behavior (see Fig. 18.1) of the family $P_{b,0}(\nu)$ parametrised by $\sqrt{b}$ [1], that corresponds to $k = 0$ modes

$$P_{b,0}(\nu) = 2\sqrt{b} + \arctan \frac{\sqrt{2}\sin\nu - \cos\nu}{\sqrt{2}\cos\nu - \sin\nu}. \qquad (18.8)$$

The map has a two dynamical regimes, a strongly attractive fixed point, and random oscillations around a mean value (see Fig. 18.1). It turns out that for finite k the behaviour of the phase dynamics is similar and essentially governed by the parameter b.

So, in static universe and in the broad resonance regime, small perturbations are exponentially amplified or not according to the amplitude of the adimensional parameter b. Let us now examine what happens in a FRW universe after inflation.

In the first stage of preheating, that ends when $n_\chi(t) \approx m_\phi^2 \Phi(t)/g$, where $\Phi(t)$ is the varying amplitude of the inflation field ϕ, (18.1), (18.2) decouple from (18.3), and the evolution of the inflation field and of the scale factor $R(t)$ is given by [2] $\phi(t) = \Phi(t)\sin t$, $\Phi(t) = m_{pl}/(3(\pi/2 + t))$, $R(t) = (2t/\pi)^{2/3}$. With the change of variable $X_k = R^{3/2}\chi_k$, (18.3) can be reduced to the form of an oscillator with a time dependent frequency $\varpi^2 = k^2/(m_\phi^2 R(t)^2) + g^2\phi(t)^2/m_\phi^2 + \delta/m_\phi^2$ and where $\delta/m_\phi^2 \ll 1$.

The preheating period ends when $g\Phi(t)/m_\phi \simeq 1$, and so, during preheating, the rate of variation of those parameters and the oscillations of the inflation field are much slower than the oscillations of the χ_k modes. As pointed out in [2], the basic assumptions for the approximation developed for Minkowski space time are thus still valid in preheating, and the changes in occupation numbers n_k will occur at $t = j\pi$ with exponential growth rate given by (18.7), provided that the decreasing amplitude of the perturbations and the redshift of the wave numbers are taken into account, $\kappa_j = k/(R(t_j)\sqrt{g m_\phi \Phi(t_j)})$ and $\sqrt{b_j} = g\Phi(t_j)/m_\phi$ (see Fig. 18.2).

Equations (18.6), (18.7), provide an alternative to numerical integrations of the full equations do compute the occupation number of a given mode as a function

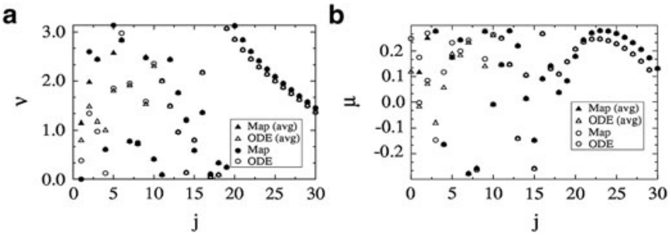


Fig. 18.2 For $b_0 = 5 \times 10^3$ and $\kappa_0 = 0.1$, the phase (**a**), growth factor (**b**) with initial conditions corresponding to $n_k^0 = 1/2$ and $\nu_k^0 = 0$. The values obtained from the iteration of equations (18.6), (18.7) are plotted as full circles, and the values given by numerical integration of equations of motion for the same initial conditions and parameter values are plotted as open circles. Iteration and integration were carried out until the end of preheating when $\sqrt{b(t_j)} \approx 1$. Also shown are the values of these same quantities averaged over the initial phase ν_k^0 (full triangles for the iterated maps (18.6), (18.7) and open triangles for the numerical values)

of time. The phase and growth factor for a typical orbit as obtained from the iteration of (18.6), (18.7) are shown in Fig. 18.2. We see "reminiscences" of the phase dynamics of the Minkowski model. In particular, the fixed point regime interval is clearly visible after the first few ϕ oscillations.

18.2 Conclusion

We consider first the broad resonance regime in Minkowski space-time and use the theory of scattering in parabolic potentials developed in [2] to obtain the map whose iteration governs the phase dynamics of the modes χ_k that coupled to the inflation field. We show that the features of this phase dynamics are given by the properties of a simple family of circle maps. We then consider the case of an expanding universe and show that the equations for the phase dynamics and the growth number derived for Minkowski space time still provide a good approximation of the true solutions, once the decay of the inflation amplitude is taken into account. The qualitative behaviour of the phase and growth number evolution is reminiscent of the behaviour found in the case without expansion, in the sense that it can be interpreted as a random phase regime followed by a slowly varying phase regime where occupation number growth is approximately exponential. These two regimes occur as the inflation decay slows down and the perturbation amplitude crosses more and more slowly the intervals that give rise to fixed phase behaviour.

References

1. Charters, T., Nunes, A., Mimoso, J.P.: Phys. Rev. D **71**, 083515 (2005)
2. Kofman, L., Linde, A.D., Starobinsky, A.A.: Phys. Rev. D **56**, 3258 (1997)
3. Dolgov, A.D., Linde, A.D.: Phys. Lett. B **116**, 329 (1982)
4. Abbott, L.F., Farhi, E., Wise, M.B.: Phys. Lett. B **117**, 29 (1982)
5. Traschen, J.H., Brandenberger, R.H.: Phys. Rev. D **42**, 2491 (1990)
6. Linde, A.D.: Phys. Lett. **108B**, 389 (1982)

Chapter 19
An Application of the SIR Model to the Evolution of Epidemics in Portugal

António M. Correia, Filipe C. Mena, and Ana J. Soares

Abstract We apply the SIR model to study the evolution of Measles and Hepatitis C in Portugal using data from 1996 until 2007. We use our results to forecast the evolution of those viruses in subsequent years.

19.1 Introduction and the SIR Model

The well-known SIR model was introduced by Kermack and McKendrick in 1927 (see e.g. [1, 4]) to study the propagation of epidemics. The model describes the dynamics of a population divided into three classes of individuals: susceptible (S), infected (I) and recovered (R). It assumes a spatially homogeneous population in each class (for $S, I, R : \mathbb{R} \to \mathbb{R}$ of class C^1) whose evolution is given by:

$$\begin{cases} \dfrac{dS}{dt} = (\lambda - \mu)S - \beta SI \\[2ex] \dfrac{dI}{dt} = \beta SI - (\mu + \alpha)I \\[2ex] \dfrac{dR}{dt} = \alpha I - \mu R, \qquad t \geq 0, \end{cases} \tag{19.1}$$

A.M. Correia (✉)
Escola EB 2,3 de Celeirós, Av. Sr. da Paciência, Celeirós, 4705-448 Braga, Portugal
e-mail: amc7761@gmail.com

F.C. Mena
Centro de Matemática, Universidade do Minho, Campus de Gualtar, 4710-057 Braga, Portugal
e-mail: fmena@math.uminho.pt

A.J. Soares
Centro de Matemática, Universidade do Minho, Campus de Gualtar, 4710-057 Braga, Portugal
and
Departamento de Matemática, Universidade do Minho, Braga, Portugal
e-mail: ajsoares@math.uminho.pt

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_19,

where $\alpha, \beta, \lambda, \mu \in \mathbb{R}^+$ and $\lambda > \mu$ are, respectively, the disease death rate, the infection coefficient, the birth rate and the natural death rate. It is easy to see that the first two equations can be directly integrated to give:

$$(\lambda - \mu)(\ln I - \ln I_0) - \beta(I - I_0 + S - S_0) + (\mu + \alpha)(\ln S - \ln S_0) = 0, \quad (19.2)$$

where the subscript 0 denotes evaluation at $t = 0$. It turns out that I has extreme values, which are often used as indicators of the epidemics strength [1], for

$$S = \frac{\mu + \alpha}{\beta}.$$

Despite the mathematical simplicity of the SIR model, it has been used in the past to study the evolution of epidemics in a variety of scenarios (see e.g. [1] and references therein).

19.2 Application to Recent Data of the Portuguese Health System

We have applied the SIR model briefly described in the previous section as a toy model to study the evolution of the Measels (M) and Hepatitis C (HC) in Portugal from 1996 until 2007. Part of this work is included in the master thesis of Correia [2]. The data we have studied was obtained from the webpage of the portuguese health system [3] and refers to *monthly* values of the number of infected individuals in each case. In order to find the best SIR fit to the data we have obtained numerically the optimal values for the parameters α, β, λ and μ corresponding to the minimum average error ϵ and maximum of the correlation coefficient r given by:

$$r^2 = 1 - \frac{\sum_{j=1}^{N} \left(d_j - i_j\right)^2}{\sum_{j=1}^{N} \left(d_j - \overline{d}\right)^2} \quad (19.3)$$

where $d_1, \ldots, d_N$ denote the observed values, $i_1, \ldots, i_N$ the adjusted values and $\overline{d}$ is the average of the observed values.

We found that the optimal values for the parameters are $\alpha_M = 0.9$, $\beta_M = 0.02$, $\lambda_M = 0.09$, $\mu_M = 0.01$ for the case of Measels, and $\alpha_{HC} = 0.039$, $\beta_{HC} = 0.001$, $\lambda_{HC} = 0.006$, $\mu_{HC} = 0.001$ for the case of Hepatitis C, which give $r_M = 0.79$ and $r_{HC} = 0.8$, respectively. From the numerical integration of (19.1), referred to the dynamics of both Measels and Hepatitis C, we have found the fittings shown in Figs. 19.1 and 19.2. The left frames of these figures show the data for infected individuals and the curves obtained from the SIR model. The right frames show the corresponding error curves.

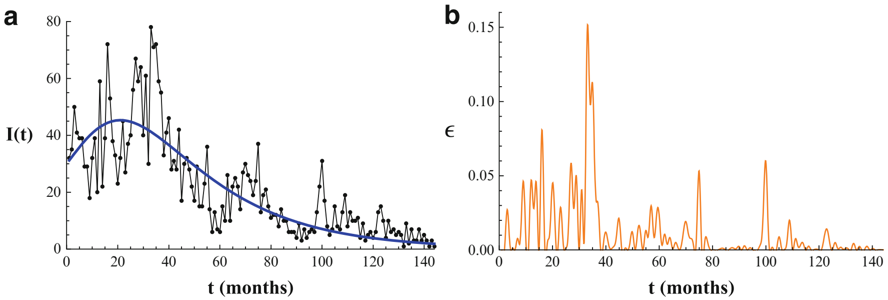


Fig. 19.1 *Monthly data – Hepatitis C.* SIR model for $\alpha = 0.039$, $\beta = 0.001$, $\lambda = 0.006$, $\mu = 0.001$. Curve of the model **(a)** and corresponding error curve **(b)**

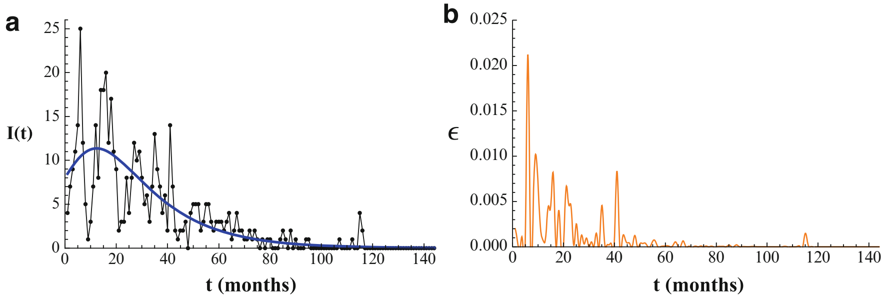


Fig. 19.2 *Monthly data – Measles.* SIR model for $\alpha = 0.9$, $\beta = 0.02$, $\lambda = 0.09$ and $\mu = 0.01$. Curve of the model **(a)** and corresponding error curve **(b)**

We have also considered *annual* data, and by performing a similar analysis, we have obtained $\alpha_M = 0.9$, $\beta_M = 0.002$, $\lambda_M = 0.09$, $\mu_M = 0.01$ and $\alpha_{HC} = 0.24$, $\beta_{HC} = 0.002$, $\lambda_{HC} = 0.09$, $\mu_{HC} = 0.01$, which give $r_M = 0.95$ and $r_{HC} = 0.91$. The dynamics is obtained from the numerical integration of the corresponding differential systems and is represented by the fittings shown in Fig. 19.3.

The annual data shows lower volatility than the monthly data and therefore, as expected, we have found a better fitting for the former data translated into higher correlation coefficients.

In general, the available data from the portuguese health system [3] seems too scarce to make feasible predictions for the evolution of virus epidemics, although the particular case of Hepatitis C seem to be the one with more complete data. Thus, in this case, we have used our previous results to forecast the number of infected individuals for the four subsequent years using polynomial interpolation. For the years 2009 and 2011 we have obtained:

$$I_{HC}(2009) \approx 36, \qquad I_{HC}(2011) \approx 15.$$

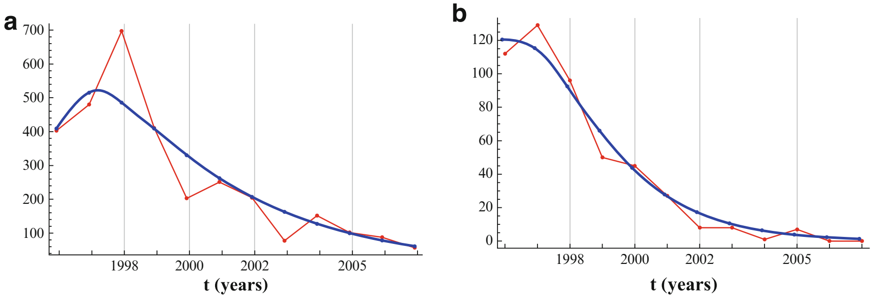


Fig. 19.3 *Annual data* – Infection early rates and approximation curve obtained from the SIR model. (**a**) Hepatitis C: $\alpha = 0.24$, $\beta = 0.002$, $\lambda = 0.09$ and $\mu = 0.01$. (**b**) Measles: $\alpha = 0.9$, $\beta = 0.002$, $\lambda = 0.09$ and $\mu = 0.01$

We conclude that although the SIR model (1) is quite simple and the 1996–2007 data from the portuguese health system is scarce, it can give us some useful insight about the evolution of the Measles and Hepatitis C viruses. In turn, this can be used as a forecast for the number of infected individuals in subsequent years and we have applied this idea to forecast the evolution of Hepatitis C infections up to 2011.

Acknowledgements FCM and AJS thank to the research centre CMAT, University of Minho, for financial support.

References

1. Brauer, F., Castillo-Chvez, C.: Mathematical Models in Population Biology and Epidemiology. Springer, New York (2001)
2. Correia A.M.: Qualitative Theory of Differential Equations and Apllication to Biology, MSc. Thesis (in portuguese), University of Minho, 2009
3. Direcção-Geral da Saúde (http://www.dgs.pt): Estatísticas de Saúde - Publicações - Doenças de Declaração Obrigatória 1996–2000, 1998–2002, 2003–2007
4. Edelstein-Keshet, L.: Mathematical Models in Biology. Random House, New York (1988)

Chapter 20
Poissonian Tree Constructed from Independent Poisson Point Processes

Iesus Carvalho Diniz and José Carlos Simon de Miranda

Abstract In this work a connected graph without cycles and with a single infinite self-avoiding path, i.e., a tree with an end, is constructed. The vertices of the tree are points of an infinite sequence of independent Poisson point processes defined on $\mathbb{R}^d$, such that for every $k \geq 1$, the rate of kth process X_k is λ_k. This graph will be called a *One-Ended Poissonian Tree*. The algorithm of construction of the Poissonian Tree is given, as well as the definition of its elements. This algorithm will be called algorithm $\mathbb{A}$. We also give a sufficient condition for the generation of a unique tree. In the case where the sequence of rates is such that $\liminf \lambda_k = 0$, for processes defined on $\mathbb{R}$, we prove that algorithm $\mathbb{A}$ generates a One-Ended Poissonian Tree.

20.1 The Model and Related Results

A connected graph without cycles and with a single infinite self-avoiding path, i.e., a tree with an end, is constructed. The vertices of the tree are points of an infinite sequence of independent Poisson point processes defined on $\mathbb{R}^d$, such that for every $k \geq 1$, the rate of kth process X_k is λ_k. We will call such a graph a *One-Ended Poissonian Tree*.

Let a_k and b_k be two arbitrary points in X_k. Their respective nearest points a_{k+1} and b_{k+1} in X_{k+1} will be called their *ancestors*. Let us call algorithm $\mathbb{A}$ the procedure that takes the infinite sequence of independent realizations of Poisson point processes $(X_k)_{k\geq 1}$ to a graph whose vertices are the points of these processes realizations and each edge links a point $\xi_k \in X_k$ to its ancestor $\xi_{k+1} \in X_{k+1}$.

I.C. Diniz (✉)
Departamento de Matemática, Universidade Federal do Rio Grande do Norte, Natal, Brazil
e-mail: iesus@ufrnet.br

J.C.S. de Miranda
Universidade de São Paulo, Instituto de Matemática e Estatística, São Paulo, Brazil
e-mail: simon@ime.usp.br

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_20,

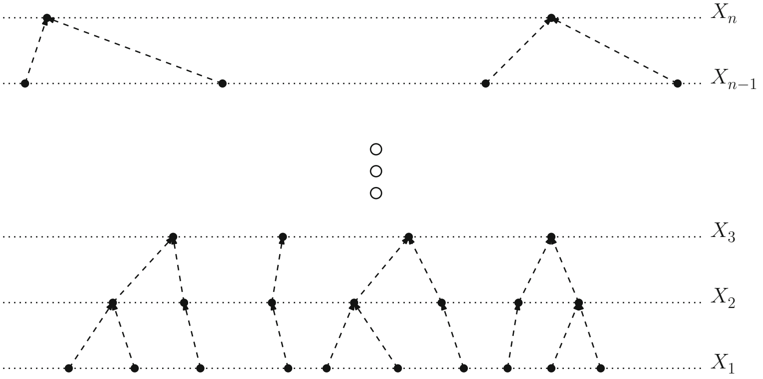


Fig. 20.1 Succession line of a sequence of unidimensional independent poisson point processes

This procedure is illustrated in Fig. 20.1. Consider that each point also determines its position and that $D_k = |b_k - a_k|$ is the distance between the kth ancestors of a_1 and b_1, chosen arbitrarily in X_1.

A Poissonian Tree with a unique end is constructed in [4] for the points of a stationary Poisson process when it is defined in $S \subset \mathbb{R}^d$, for $d \leq 3$. It is also shown that that for $d \geq 4$ the graph has infinite many components, a forest. In [5] a One-Ended Poissonian Tree is constructed in a deterministic isometry-invariant way for any d-dimensional Poisson process.

In order to prove the almost sure existence of a unique infinite self-avoiding path, in the graph generated by algorithm $\mathbb{A}$, it is sufficient to prove that given any two points in X_1 (a_1 and b_1), the following condition is satisfied:

$$\lim_{k\to\infty} \mathbb{P}(D_k \neq 0) = 0$$

This will be called *coalescense in probability* and we will say that a_1 and b_1k *coalesce in probability*. Since we could have chosen an arbitrary index in the sequence of Poisson Point Process instead of 1, i.e., we could have chosen points of the realization of X_j for an arbitrary j, we will denote the condition above by

$$a_\kappa \overset{\mathbb{P}}{=} b_\kappa := \lim_{k\to\infty} \mathbb{P}(D_k \neq 0) = 0 \qquad (20.1)$$

Clearly, this condition is equivalent to $\mathbb{P}\left(\lim_{k\to\infty} D_k = 0\right) = 1$.

20.2 Poissonian Tree Formed from Unidimensional Poisson Processes

The proofs of the following propositions and of Theorem 20.1 can be found in [2].

Proposition 20.1. *Let $X_{k+1}^{a_{k+1}}$ be the position of the ancestor of $X_k^{a_k}$ in X_{k+1}. The distribution of $X_{k+1}^{a_{k+1}}$ conditioned on $X_k^{a_k}$ is given by*

$$f_{X_{k+1}^{a_{k+1}}|X_k^{a_k}}(x) = \lambda_{k+1} \exp(-2\lambda_{k+1}|x - a_k|) \tag{20.2}$$

From Proposition 20.1, we obtain the two following propositions, that jointly with Proposition 20.4, are used to prove Theorem 20.1.

Proposition 20.2 (Coalescing Conditional Probability Law). *For all $a_k, b_k \in X_k$*

$$\mathbb{P}(D_{k+1} = 0|a_k, b_k) = e^{-2\lambda_{k+1}D_k}(1 + \lambda_{k+1}D_k).$$

Proposition 20.3. *For all $k \geq 1$, $\mathbb{E}(D_{k+1}) = \mathbb{E}(D_k) = D_1$.*

Proposition 20.4. *For all $k \geq 1$, $\mathbb{P}(D_k = 0) \leq \mathbb{P}(D_{k+1} = 0)$.*

Theorem 20.1. *Let $(X_k)_{k\geq 1}$ be a sequence of independent Poisson point processes defined on $\mathbb{R}$ and λ_k be the rate of X_k. Suppose $\liminf \lambda_k = 0$. Then, almost surely, algorithm $\mathbb{A}$ constructs a One-Ended Poissonian Tree with all the points of all processes.*

20.2.1 Determination Criteria for a One-Ended Poissonian Tree as a Function of the Rates Sequence of the Processes

Proposition 20.2 describes the coalescing probability of a_k and b_k as a function of the rate of the $(k+1)$th process and the distance between a_k and b_k. The theorems that follow show that, depending on the rates of the sequence of processes, the graph generated by algorithm $\mathbb{A}$ may be either connected with probability 1, see Theorem 20.2, or disconnected with positive probability, see Theorem 20.3. Moreover, Theorem 20.4 guarantees that the probability of getting a non connected graph may be chosen to be greater than or equal to any prescribed probability level; this will require a sufficiently increasing sequence of rates. The proofs of these results may be found in [2].

Let $G : R^+ \times \mathbb{N} \to [0, 1]$ be defined in the following manner.

$$G(\lambda_{k+1}, k) := \mathbb{P}(a_{k+1} = b_{k+1}|D_k \neq 0) = \int_{0^+}^{\infty} e^{-2\lambda_{k+1}u}\,(1+\lambda_{k+1}u)\, f_{D_k}(u)du$$

Theorem 20.2. *If $G(\lambda_{k+1}, k) > \frac{1}{k}$, then with probability one, algorithm $\mathbb{A}$ generates a One-Ended Poissonian Tree consisting of all the points of the processes $(X_k)_{k \geq 1}$.*

Theorem 20.3. *If $G(\lambda_{k+1}, k) < \frac{1}{e^k}$, then there is a positive probability that algorithm $\mathbb{A}$ generates a graph which is not connected and has no cycles.*

Theorem 20.4. *For any $p \in (0,1)$ it is always possible to obtain a sequence of rates $(\lambda_k)_{k \geq 1}$ such that the probability of not having a One-Ended Poissonian Tree is larger than p.*

20.3 One-Ended Poissonian Tree Formed from Multidimensional Poisson Processes

For processes defined on $\mathbb{R}^d$, the main difficulties that appear are:

1. Differently from what happens in Proposition 20.2, we are not able to express $\mathbb{P}(D_{k+1} = 0 | a_k, b_k)$ in a "closed form".
2. The distribution of D_k does not have the property described in proposition 20.3.

A lower bound for the coalescing conditional probability, which is obtained from the distance distribution between a point and its ancestor, and a deterministic rescale, $d_k := (\alpha)^{\frac{k}{d}} D_k$ of the process D_k, will be the alternative to these difficulties.

Let $\mathscr{L}_d = f(v_d(1), \alpha, \beta)$ be a positive constant that depends on: the volume of the d-dimensional unity ball $(v_d(1))$, the ratio α of decay of rates of the processes and the value $\beta \in (\alpha^{\frac{1}{d}}, 1)$ given in (20.3) associated to the "mean drift" of the rescaled process d_k. We have,

$$\mathbb{E}(d_{k+1} | d_k) < \beta d_k \text{ if } d_k \in (\mathscr{L}_d, \infty) \tag{20.3}$$

From this, Theorem 20.5 will ensure that, almost surely, $d_k \leq \mathscr{L}_d$ for infinite many $k's$. This fact and the condition given in Lemma 20.1, which establishes a positive lower bound for the limit of coalescing conditional probability, will be enough to prove that algorithm $\mathbb{A}$ generates a One-Ended Poissonian Tree.

Lemma 20.1. *The coalescing conditional probability limit is larger than a positive constant ε that depends on d, α and β.*

$$\lim_{k \to +\infty} \mathbb{P}(d_{k+1} = 0 | d_k \in [0, \mathscr{L}_d]) \geq \exp(-\alpha (\mathscr{L}_d)^d v_d(1)) = \varepsilon(\alpha, \beta, d) > 0$$

The proof of this lemma is in [1].

Theorem 20.5. *Let $S_0 > C$ and, for some $\varepsilon > 0$ and for all $n \geq 0$,*

$$\textit{If}\, \mathbb{E}(\widetilde{S}_{n+1} | \mathbb{F}_n) \leq \widetilde{S}_n - \varepsilon 1\{\tau > n\}\, a.s., \;\; \textit{then}\, \mathbb{E}(\tau) < \frac{S_0}{\varepsilon} < \infty \tag{20.4}$$

The proof of this theorem is given in [3], p. 17.

Theorem 20.6. *Let $(X_k)_{k \geq 1}$ be a sequence of independent Poisson point processes defined on $\mathbb{R}^d$, and λ_k, the rate of X_k, be such that $\lambda_k = (\alpha)^k$. Then almost surely algorithm $\mathbb{A}$ constructs a One-Ended Poissonian Tree consisting of the points of all processes.*

The proof of Theorem 20.6 may be found in [2].

Acknowledgements The first author was supported by FAPESP grant 04/15864-1.

References

1. Diniz, I.C., in *A Limit of an Improper Integral Depending on One Parameter*, Crux Mathematicorum with Mathematical Mayhem. Can. Math. Soc. Problem Solving J. (Digital Supplement) **34**(8), 478–480 (2008) http://journals.cms.math.ca/CRUX/
2. Diniz, I.C., de Miranda, J.C.S.: Poissonian tree constructed from independent Poisson point processes, 16p, 2008, Technical Report, Department of Statistics, Institute of Mathematics and Statistics, University of São Paulo. RT-MAE-2008 number 12
3. Fayolle, G., Malyshev, V.A., Menshikov, M.V.: Topics in the Constructive Theory of Countable Markov Chains. Cambridge University Press, NY (1995)
4. Ferrari, P., Landim, C., Thorisson, H.: Poisson trees, succession lines and coalescing random walks. Ann. Inst. H. Poincar Probab. Statist. **40**(2), 141–152 (2004)
5. Holroyd, A., Yuval, P.: Trees and matchings from point processes. Electron. Comm. Probab. **8**(2), 17–27 (2003)

Chapter 21
Hamiltonian Systems on Polyhedra

Pedro Duarte

Abstract We describe a class of Hamiltonian systems on simple polyhedra, which includes several models from game dynamics (e.g., conservative Lotka–Volterra systems). A technique to detect complex dynamical behaviour along the polyhedron edges is explained.

21.1 Flows on Polyhedra

Let Γ^d be a simple polyhedron with dimension d. We say that a vector field X on Γ^d is tangent to $\partial\Gamma^d$ if X is tangent to every face σ of Γ^d, i.e., $X(p) \in T_p\sigma$ at each point $p \in \sigma$. We denote by $\mathscr{X}(\Gamma^d)$ the vector space of all analytic vector fields X on Γ^d which are tangent to $\partial\Gamma^d$. For any given $X \in \mathscr{X}(\Gamma^d)$ the flow $\phi_X^t : \Gamma^d \to \Gamma^d$ of X is complete and every face of Γ^d is invariant under ϕ_X^t. In particular, the vertices of Γ^d are singularities of the vector field X, and many edges will consist of single orbits flowing from one boundary vertex to the other. Our goal is, for some rather large class of "*regular*" vector fields $X \in \mathscr{X}(\Gamma^d)$, to encapsulate the dynamics of ϕ_X^t along heteroclinic cycles on $\partial\Gamma^d$ in a simple and "computable" dynamical system, that we refer as *the skeleton vector field on the dual cone of* Γ^d.

Before continuing we give precise definitions of the concepts of polyhedron, dimension, face, vertex, edge and simplicity, while introducing the notation used in the sequel. A subset Γ of some Euclidean space $\mathbb{R}^N$ is called a *polyhedron* if it is a compact convex set which can be represented as a finite intersection of closed half-spaces. Denote by $E(\Gamma)$ the smallest affine subspace of $\mathbb{R}^N$ that contains Γ. The *dimension of a polyhedron* Γ is defined to be the dimension of $E(\Gamma)$. From now on Γ^d will denote a polyhedron of dimension d, which for the sake of simplicity we assume, unless otherwise said, to live in $E(\Gamma^d) = \mathbb{R}^d$. We call *supporting hyperplane of* Γ^d to any affine hyperplane $H \subset \mathbb{R}^d$ such that $H \cap \Gamma^d \neq \emptyset$,

P. Duarte
CMAF/DM-FCUL, 1749-016 Lisbon, Portugal
e-mail: pedromiguel.duarte@gmail.com

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_21,

and Γ^d is contained in one of the two closed half-spaces determined by H. The intersection of Γ^d with any of its supporting hyperplanes is another polyhedron, called a *face of* Γ^d, or an r-face when its dimension is equal to r. As usual, a *vertex* is any 0-face, and an *edge* is any 1-face of Γ^d. Capital letters A, B, C will denote vertices of Γ^d, while γ will denote a generic edge of Γ^d. By default, the term "face" shall always refer to a $(d-1)$-face, and σ will represent a generic such $(d-1)$-face. We represent by V the set of all vertices, by E the set of all edges, and by F the set of all $(d-1)$-faces of Γ^d (Figs. 21.1, 21.2, 21.3, and 21.4).

Definition 21.1. A family of functions $\{ f_\sigma : \mathbb{R}^d \to \mathbb{R} \}_{\sigma\in F}$ is called a *defining family for* Γ^d if for every face $\sigma \in F$,

1. $f_\sigma : \mathbb{R}^d \to \mathbb{R}$ is an affine function.
2. $f_\sigma(p) = 0$ for all $p \in \sigma$.
3. $f_\sigma(p) \geq 0$ for all $p \in \Gamma^d$.
4. $\Gamma^d = \bigcap_{\sigma\in F}\{f_\sigma \geq 0\}$.

We assume a defining family $\{f_\sigma\}_{\sigma\in F}$ for Γ^d is fixed once and for all.

We call d-simplex to the convex hull of any $d+1$ affinely independent points. These are the simplest polyhedra. A polyhedron Γ^d is called *simple* if each vertex is incident with exactly d faces (edges). This amounts to the supporting hyperplanes $\{f_\sigma = 0\}$ intersecting each other in general position. A d-simplex is of course

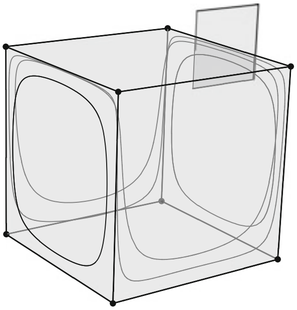


Fig. 21.1 The dynamics near the edges for a flow ϕ_X^t on the polyhedron $\Gamma^3 = [0,1]^3$

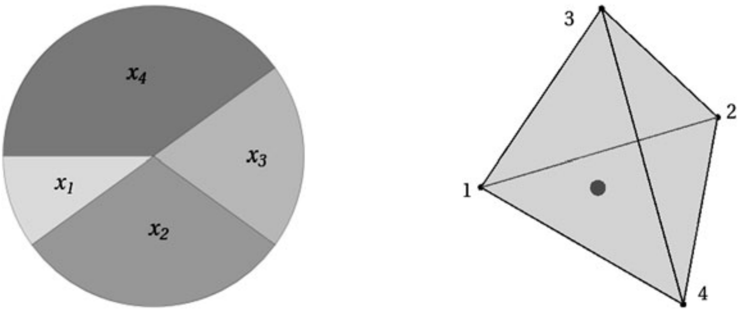


Fig. 21.2 A point in Δ^3 is a probability vector in $\{1, 2, 3, 4\}$

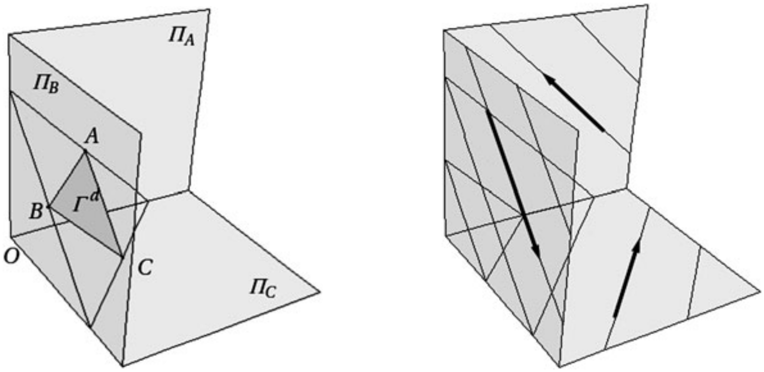
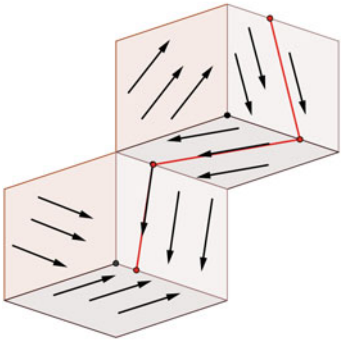


Fig. 21.3 The dual cone of a triangle polyhedron and a skeleton vector field on it

Fig. 21.4 A finite orbit of a skeleton vector field



simple in this sense. A polyhedron Γ^d is simple if and only if every face of its dual polyhedron is a $(d-1)$-simplex.

21.2 Game Dynamics

Systems as these include many interesting classes from Game Dynamics, for instance the *replicator equation* see [5]. Within a population individuals interact using one of n possible strategies. The time evolution of a population distribution $(x_1, \ldots, x_n) \in \Delta^{n-1}$ is ruled by

$$\frac{x_i'}{x_i} = f_i(x_1, \ldots, x_n) - \sum_{k=1}^{n} x_k\, f_k(x_1, \ldots, x_n)\,, \tag{21.1}$$

where Δ^{n-1} stands for the usual $(n-1)$-simplex $\{\,(x_1, \ldots, x_n) \,:\, x_i \geq 0, \sum_{i=1}^{n} x_i = 1\,\}$. The value $f_i(x_1, \ldots, x_n)$ measures the *absolute fitness* of strategy i

for the population distribution $(x_1, \ldots, x_n) \in \Delta^{n-1}$. Likewise, the right-hand-side in (21.1) expresses the *relative fitness* of strategy i within the same population. In the replicator equation model, strategies in the population thrive or recede proportional to their relative fitnesses. When the functions $f_i(x)$ are linear, say $f_i(x_1, \ldots, x_n) = \sum_{j=1}^{n} a_{ij}\, x_j$, the system is determined by a matrix $A = \left(a_{ij}\right)$ called the payoff matrix. The payoffs a_{ij} are the eigenvalues of the singularities at the vertices, for the associated replicator flow or vector field.

An important class of equations which reduces to the (linear) replicator equation are the so called Lotka–Volterra equations. They govern the time evolution of a n-species ecosystem $y = (y_1, \ldots, y_n) \in \mathbb{R}_+^n$

$$\frac{y_i'}{y_i} = r_i + \sum_{j=1}^{n} a_{ij}\, y_j\,, \tag{21.2}$$

where y_i measures the size of species i within the ecosystem, a_{ij} is an interaction coefficient between species i and j, while r_i models the interaction of species i with environment. Every Lotka–Volterra system is equivalent to a replicator system in the sense that the underlying vector fields are equivalent. The equivalence is given by the algebraic map defined by

$$x = (x_0, \ldots, x_n) \in \Delta^n \;\longleftrightarrow\; y = (y_1, \ldots, y_n) = \left(\frac{x_1}{x_0}, \ldots, \frac{x_n}{x_0}\right) \in \mathbb{R}_+^n\ ,$$

which maps the interior of the simplex Δ^n onto the the interior of $\mathbb{R}_+^n$. In the new coordinates $x = (x_0, \ldots, x_n) \in \Delta^n$ the system becomes (up to a time reparametrization)

$$\frac{x_i'}{x_i} = \sum_{j=0}^{n} \tilde{a}_{ij}\, x_j - \sum_{j,k=0}^{n} \tilde{a}_{kj}\, x_k\, x_j \tag{21.3}$$

which is a linear replicator with payoff matrix $\widetilde{A} = \left(\tilde{a}_{ij}\right)$, where $\tilde{a}_{ij} = a_{ij}$ when $i, j \geq 1$, $\tilde{a}_{i0} = r_i$ and $\tilde{a}_{0i} = 0$. This reduction, due to J. Hofbauer [4], consists roughly in letting the n species together with the environment play the roles of $n+1$ strategies.

Another important class which falls within the scope of this work is that of asymmetric games, where two groups of individuals within a population, e.g. males and females, interact using different sets of strategies, say n strategies for males and m strategies for females. The phase space of an asymmetric game system is a polyhedron, product of simplices $\Delta^{n-1} \times \Delta^{m-1}$, and the time co-evolution of two population distributions $(x, y) \in \Delta^{n-1} \times \Delta^{m-1}$ is governed by

$$\begin{aligned} \frac{x_i'}{x_i} &= f_i(y_1, \ldots, y_m) - \textstyle\sum_{k=1}^{n} x_k\, f_k(y_1, \ldots, y_m) \\ \frac{y_j'}{y_j} &= g_j(x_1, \ldots, x_n) - \textstyle\sum_{k=1}^{m} x_k\, g_k(x_1, \ldots, x_n) \end{aligned} \tag{21.4}$$

The value $f_i(y)$ measures the absolute fitness of a male strategy i in a female population $y \in \Delta^{m-1}$, while $g_j(x)$ measures the absolute fitness of a female strategy j in a male population $x \in \Delta^{n-1}$. The right-hand-sides in (21.4) express, respectively, the relative fitnesses of a male strategy i, and of a female strategy j, within the populations of opposite gender. Once more, in this asymmetric game model strategies in the male and female populations thrive or recede proportional to their relative fitnesses. When the functions $f_i(y)$ and $g_j(x)$ are both linear, say $f_i(y_1, \dots, y_m) = \sum_{j=1}^m a_{ij}\, y_j$ and $g_j(x_1, \dots, x_n) = \sum_{i=1}^n b_{ji}\, x_i$, the system is determined by a pair of matrices $A = (a_{ij})$ of order $n \times m$ and $B = (b_{ji})$ of order $m \times n$, called the payoff matrices. Again, the payoffs a_{ij} and b_{ji} are related to the eigenvalues of the singularities at the vertices, for the associated asymmetric game flow or vector field.

21.3 Skeletons and Dual Cones

Assume for a while $\Gamma^d \subset \mathbb{R}^{d+1} - \{0\}$ and the cone $\widehat{\Gamma}^{d+1} = \{\, t\, X \,:\, t \geq 0,\ X \in \Gamma^d \,\}$ has dimension $d+1$. In Convex Analysis the dual cone of Γ^d is defined to be

$$(\Gamma^d)^* = \{\, Y \in \mathbb{R}^{d+1} \,:\, Y \cdot X \geq 0,\ \forall\, X \in \Gamma^d \,\}.$$

Here we shall call *dual cone of* Γ^d to the boundary of this set, $\mathscr{C}^*(\Gamma^d) = \partial(\Gamma^d)^*$. We give an alternative description of the dual cone, which is more convenient for our purposes. Denote by Σ^d the dual of the polyhedron Γ^d. We can identify $V^* = V(\Sigma^d) \equiv F$ and $F^* = F(\Sigma^d) \equiv V$. By duality each vertex $A \in V$ stands for a $(d-1)$-face in Σ^d, each face $\sigma \in F$ represents a vertex of Σ^d, and the relation $A \in \sigma$ in Γ^d is equivalent to $\sigma \in A$ in Σ^d. We define

$$\mathscr{C}(\Sigma^d) := \{x \in \mathbb{R}^{V^*} : \exists A \in F^* \text{ for all } \sigma \in V^*,\ x_\sigma \geq 0 \text{ and } x_\sigma = 0 \text{ if } \sigma \notin A\},$$

and for each face ρ of Σ^d we set

$$\Pi_\rho := \{\, x \in \mathbb{R}^{V^*} \,:\, \text{ for all } \ \sigma \in V^*,\ x_\sigma \geq 0 \ \text{ and } \ x_\sigma = 0 \ \text{ if } \ \sigma \notin \rho \,\}.$$

Then the following properties hold for all faces ρ, ρ' of Σ^d:

1. $\dim \Pi_\rho = \dim_{\Sigma^d}(\rho) + 1$.
2. $\Pi_\rho \subseteq \Pi_{\rho'} \quad \Leftrightarrow \quad \rho \subseteq \rho'$ in Σ^d.
3. $\Pi_\rho \cap \Pi_{\rho'} = \Pi_{\rho \cap \rho'}$.

Because Γ^d is simple, by duality, every r-face of Σ^d is a $(r-1)$-simplex, i.e., it has exactly r vertices. This implies item 1. Properties 2 and 3 are obvious consequences of definitions. Realizing the dual polyhedron Σ^d as a transversal section to the cone $(\Gamma^d)^*$, we can identify $(\Gamma^d)^*$ with $\widehat{\Sigma}^{d+1} = \{\, t\, X \,:\, t \geq 0,\ X \in \Sigma^d \,\}$. Whence, the faces of $\partial(\Gamma^d)^*$ satisfy the exact same properties 1–3 above. In fact, the three models $\mathscr{C}(\Sigma^d)$, $\partial(\Gamma^d)^*$ and $\partial\widehat{\Sigma}^{d+1}$ are piecewise-linear isomorphic. From now on we consider the dual cone of Γ^d to be $\mathscr{C}^*(\Gamma^d) := \mathscr{C}(\Sigma^d)$. Properties 1–3 above can be re-interpreted in terms of Γ^d's faces. Since each r-face ρ of Γ^d corresponds to a $(d-1-r)$-face of Σ^d, we have for all faces ρ, ρ' of Γ^d:

1. $\dim \Pi_\rho = d - \dim_{\Gamma^d}(\rho)$,
2. $\Pi_\rho \subseteq \Pi_{\rho'} \;\;\Leftrightarrow\;\; \rho' \subseteq \rho$ in Γ^d,
3. $\Pi_\rho \cap \Pi_{\rho'} = \Pi_{\rho\vee\rho'}$,

where $\rho \vee \rho'$ stands for smallest face of Γ^d containing $\rho \cup \rho'$. In particular, the dual cone $\mathscr{C}^*(\Gamma^d)$ has a face Π_A for each vertex A of Γ^d, and the intersection $\Pi_A \cap \Pi_B$ of any two meeting faces corresponds to an edge of Γ^d connecting A to B.

A skeleton vector field is a piecewise constant vector field on the dual cone $\mathscr{C}^*(\Gamma^d)$, i.e., one which is constant on each face Π_A, $A \in V$. Any skeleton vector field is given by the finite data $\chi = (\chi_\sigma^A)_{A\in V,\sigma\in F}$ with $\chi_\sigma^A = 0$ whenever $A \notin \sigma$. We write χ^A for the vector $(\chi_\sigma^A)_{\sigma\in F}$ in the tangent space to Π_A. Orbits of a skeleton vector field χ are defined to be the polygonal curves whose intersection with each face Π_A of $\mathscr{C}^*(\Gamma^d)$ is a line segment parallel to χ^A on Π_A. Notice that orbit continuation is essentially unique, because as an orbit through Π_A reaches the intersection $\Pi_\gamma = \Pi_A \cap \Pi_B$ of two faces Π_A and Π_B of $\mathscr{C}^*(\Gamma^d)$ at some point p interior to Π_γ, there is at most one possible continuation on Π_B, because Π_B is the unique face which meets Π_A at p.

Of course some orbits will end in finite time. This definition gives us an incomplete piecewise linear flow on $\mathscr{C}^*(\Gamma^d)$. Vertices and edges of Γ^d are classified w.r.t. the skeleton vector field χ as Figs. 21.5 and 21.6 indicate.

Definition 21.2. Given a vertex $A \in V$, we say that A is a

1. χ-attractor $\;\Leftrightarrow\;\; -\chi^A \in \Pi_A$
2. χ-repellor $\;\Leftrightarrow\;\; \chi^A \in \Pi_A$
3. χ-saddle $\;\;\Leftrightarrow\;\; \chi^A \notin \Pi_A$ and $-\chi^A \notin \Pi_A$

Because we will be looking for recurrent behavior, χ-saddle vertices are the interesting ones, for if a vertex A is a χ-repellor, respectively a χ-attractor, then Π_A is forward, respectively backward, invariant by the flow of χ^A.

Let γ be an edge connecting two vertices $A, B \in V$. Take $\sigma, \rho \in F$ to be the unique faces such that $\gamma \cap \sigma = \{A\}$ and $\gamma \cap \rho = \{B\}$.

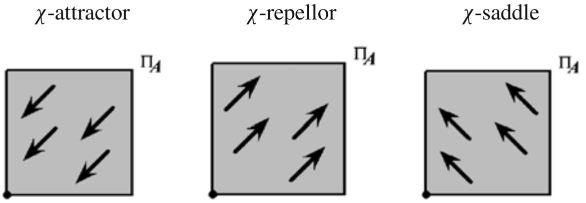


Fig. 21.5 The classification of vertices for a skeleton vector field

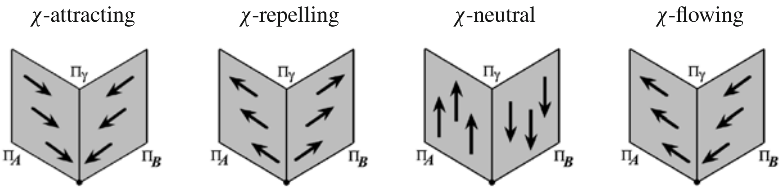


Fig. 21.6 The classification of edges for a skeleton vector field

Definition 21.3. We say that γ is

1. χ -attracting $\Leftrightarrow$ $\chi_\sigma^A < 0$ and $\chi_\rho^B < 0$
2. χ -attracting $\Leftrightarrow$ $\chi_\sigma^A > 0$ and $\chi_\rho^B > 0$
3. χ -flowing $\Leftrightarrow$ $\chi_\sigma^A \, \chi_\rho^B < 0$.

All other edges are said to be χ-undefined.

We shall not consider skeleton vector fields with χ-undefined edges. When all vertices are χ-saddles and all edges are either χ-neutral or χ-flowing then some recurrence occurs. This will be the case of the Hamiltonian systems introduced below. Note the flowing edges are naturally oriented, from a source vertex, we denote by $s(\gamma)$, to a target vertex, denoted by $t(\gamma)$. Let $G_\chi(\Gamma^d)$ be the oriented graph consisting of all vertices, and all oriented edges of χ-flowing type of Γ^d. The dynamics of a skeleton vector field can be described in terms of piecewise linear return maps. Fixing an edge γ of $G_\chi(\Gamma^d)$ we can define the return map $R_\gamma^\chi : \Pi_\gamma \to \Pi_\gamma$. These return maps satisfy:

(1) The domain of R_γ^χ splits into a finite or countable number of open convex cones Π_ξ, each associated with a cycle ξ of $G_\chi(\Gamma^d)$ starting and ending with γ, and not passing through γ in between.
(2) The restriction of R_γ^χ to each cone Π_ξ is a linear map.
(3) The linear branches of R_γ^χ, as well as their domains, are computable.

The return maps R_γ^χ and their domains Π_ξ can be expressed in terms of matrices in $\mathbb{R}^{F\times F}$ whose coefficients are functions of the data χ_σ^A. Given an edge $\gamma \in G_\chi(\Gamma^d)$, let $A = s(\gamma)$ be the source of γ, and $\sigma_0 \in F$ be the unique face such that $\sigma_0 \cap \gamma = \{A\}$. We associate the following $F \times F$ matrix to the edge γ,

$$M_\gamma = \left(\delta_{\sigma,\sigma'} - \frac{\chi_\sigma^A}{\chi_{\sigma_0}^A}\, \delta_{\sigma_0,\sigma'} \right)_{(\sigma,\sigma')\in F\times F} .$$

A sequence $\xi = (\gamma_0, \gamma_1, \gamma_2, \ldots, \gamma_n)$ is called a *chain* if $s(\gamma_i) = t(\gamma_{i-1})$, for every $i = 1, \ldots, n$. We call sub-chain of ξ to any initial subsequence $\xi_i = (\gamma_0, \ldots, \gamma_i)$ of ξ with $1 \le i \le n$. For each chain $\xi = (\gamma_0, \gamma_1, \ldots, \gamma_n)$ we define the product matrix $M_\xi = M_{\gamma_n} \cdots M_{\gamma_1}$. Note M_{γ_0} is excluded from this product. The matrix M_ξ defines a linear operator on $\mathbb{R}^F$, which projects $\mathbb{R}^F$ onto the linear subspace spanned by the cone Π_{γ_n}. The chain $\xi = (\gamma_0, \gamma_1, \dots, \gamma_n)$ is called a *cycle* when $\gamma_n = \gamma_0$, in which case we have for every $X \in \Pi_\xi$, $R^\chi_{\gamma_0} X = M_\xi X$. The open convex cone Π_ξ can be characterized as the set of all $X \in \Pi^\chi_{\gamma_0}$ such that for each sub-chain $\xi_i = (\gamma_0, \dots, \gamma_i)$ of ξ the vector $M_{\xi_i} X$ is interior to Π_{γ_i}.

21.4 Main Results

We are going to rescale the vector field X around the singularities at the vertices using some type of logarithmic coordinates. In [1] we single out a class of vector fields, that we call *regular vector fields*, for which these coordinates around the vertex singularities can be glued along the edges to obtain a global rescaling mapping. Regular vector fields include generic ones, with hyperbolic singularities at the vertices, but they also comprise many others with non-hyperbolic singularities. This generality is essential to embrace the Hamiltonian systems in which we are interested. Given $A \in V$ and $\sigma \in F$ such that $A \in \sigma$ we denote by $\gamma = \gamma_{A,\sigma}$ the edge opposed to σ at A, which is characterized by $\sigma \cap \gamma = \{A\}$. We refer to the pair (A, σ) as an *end corner of* γ. Notice each edge has exactly two end corners. Let $e_{A,\sigma} \in T_A \Gamma^d$ denote the unit vector tangent to the edge $\gamma_{A,\sigma}$ at A. To each vector field $X \in \mathscr{X}(\Gamma^d)$, $X \neq 0$, we associate an *order function* $\nu_X : F \to \mathbb{N}$

$$\nu_X(\sigma) = \max\{k \in \mathbb{N} \, : \, D(f_\sigma)_p D^i X_p \equiv 0, \ \forall\, i < k, \ \forall\, p \in \sigma\},$$

with the order of the first non-zero derivative at some of the face's vertices. Remark each face has finite order because the vector field X is analytic. Then we define the *character of* X *at the corner* (A, σ) by $\chi^A_\sigma = -\frac{1}{\nu!} D(f_\sigma)_A \, D^\nu X_A \cdot {e_{A,\sigma}}^{(\nu)}$, where $\nu = \nu_X(\sigma)$. We set $\chi^A_\sigma = 0$ if $A \notin \sigma$. The data $\chi = (\, \chi^A_\sigma \,)_{A \in V, \sigma \in F}$ determines a skeleton vector field we shall call the *skeleton of* X.

Definition 21.4. We say that a vector field $X \in \mathscr{X}(\Gamma^d)$ is *regular* iff for every edge γ of Γ^d, either $X = 0$ along γ or else $X \neq 0$ in the interior of γ and X has non-zero character at both end corners (A, σ) and (A', σ') of γ.

In particular, for the skeleton χ of a regular vector field, every edge γ of Γ^d is either χ-neutral or χ-flowing.

For each order function $\nu : F \to \mathbb{N}$ we define a one-parameter family of rescaling co-ordinates $\Psi^\nu_\varepsilon : \Gamma^d - \partial\Gamma^d \to \mathscr{C}^*(\Gamma^d)$ $(\varepsilon > 0)$ by $\Psi^\nu_\varepsilon(p) = (\Psi^\sigma_\varepsilon(p))_{\sigma \in F}$, where

$$\Psi^\sigma_\varepsilon(p) := \begin{cases} -\varepsilon \log f_\sigma(p) & \text{if } \ \nu_X(\sigma) = 1 \\ -\varepsilon \, \frac{1}{\nu - 1} \left(1 - \frac{1}{f_\sigma(p)^{\nu-1}} \right) & \text{if } \ \nu_X(\sigma) \geq 2 \end{cases}$$

Actually, we take the domain of Ψ^ν_ε to be the union of a family of neighborhoods N_A, one for each vertex $A \in V$ (Fig. 21.7).

The mapping Ψ^ν_ε *zooms in* a neighborhood of the union of all edges of Γ^d. The first theorem says, given $X \in \mathscr{X}(\Gamma^d)$, the rescaling limit of the flow ϕ^t_X is exactly

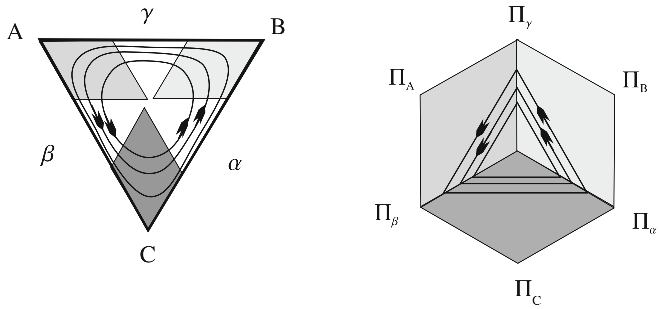


Fig. 21.7 The rescaling coordinates in the dual cone $\mathscr{C}^*(\Gamma^d)$

the piecewise linear flow of the skeleton χ of X. Given a cycle ξ of χ, starting and ending with $\gamma \in G_\chi(\Gamma^d)$, we denote by P_ξ^X the Poincaré return map along ξ. This map is defined in a small cross section of ϕ_X^t which is mapped by every Ψ_ε^ν into the face $\Pi_\gamma \subset \mathscr{C}^*(\Gamma^d)$.

Theorem 21.1. *If $X \in \mathscr{X}(\Gamma^d)$ is a regular vector field with order ν, skeleton χ, and ξ is a cycle in $G_\chi(\Gamma^d)$ which starts and ends with γ, then for every compact subset $K \subset \Pi_\xi$, $(\Psi_\varepsilon^\nu) \circ P_\xi^X \circ (\Psi_\varepsilon^\nu)^{-1}$ converges to $R_\gamma^\chi : \Pi_\xi \to \Pi_\gamma$, in the C^∞-topology, uniformly over K, as $\varepsilon \to 0^+$.*

We consider in [1] the vector space, denoted by $\mathscr{H}(\Gamma^d)$, of analytic functions $h : \Gamma^d - \partial\Gamma^d \to \mathbb{R}$ such that for each face $\sigma \in F$, either h is essentially analytic on σ, or else dh has a pole of finite order along σ. We say that h is *essentially analytic on σ* if h has an analytic extension to a neighborhood of σ minus the union of all other faces $\sigma' \in F$, $\sigma' \neq \sigma$. A similar definition is given for analytic 1-forms. We say that dh has a *pole of order k along σ* iff there is a 1-form λ and function g, both analytic in $\Gamma^d - \partial\Gamma^d$ and essentially analytic on σ, such that $dh = \lambda + g\,\frac{df_\sigma}{(f_\sigma)^k}$. It follows from this definition that g is constant on σ. Each function $h \in \mathscr{H}(\Gamma^d)$ can be represented as

$$h = G + \sum_{\sigma\in F} c_{1,\sigma} \log f_\sigma + \frac{c_{2,\sigma}}{f_\sigma} + \cdots + \frac{c_{k_\sigma,\sigma}}{(f_\sigma)^{k_\sigma - 1}}\,, \tag{21.5}$$

where G is an analytic function on Γ^d, each $c_{i,\sigma}$ is a real constant, and $c_{k_\sigma,\sigma} \neq 0$. The function $\kappa : \sigma \mapsto k_\sigma$ is called *the order of h*.

We define now the *skeleton of $h \in \mathscr{H}(\Gamma^d)$* to be the piece-wise linear function $\lambda_h : \mathscr{C}^*(\Gamma^d) \to \mathbb{R}$, $\lambda_h(u_\sigma)_{\sigma\in F} = \sum_{\sigma\in F} c_{k_\sigma,\sigma}\, u_\sigma$, where $c_{k_\sigma,\sigma}$ is the main coefficient in (21.5). A function $h \in \mathscr{H}(\Gamma^d)$ with order κ is called *regular* if $\kappa(\sigma) \geq 1$, and all faces of order $\kappa(\sigma) \geq 2$ are pairwise disjoint. The second theorem states that the rescaling limit of a function $h \in \mathscr{H}(\Gamma^d)$ is precisely its skeleton λ_h.

Theorem 21.2. *Given $h \in \mathscr{H}(\Gamma^d)$ regular with order κ, and $A \in V$, respectively $\gamma \in E$, as $\varepsilon \to 0^+$ the rescaled function $h \circ (\Psi_\varepsilon^\kappa)^{-1} : \mathscr{C}^*(\Gamma^d) \to \mathbb{R}$ tends in the*

C^∞-topology and uniformly on compact subsets in the interior of Π_A, respectively Π_γ, to the skeleton function $\lambda_h : \mathscr{C}^*(\Gamma^d) \to \mathbb{R}$.

The class of Hamiltonian systems on polyhedra we are about to introduce uses Hamiltonian functions in the space $\mathscr{H}(\Gamma^{2d})$ and the class of algebraic symplectic structures we now discuss. Consider the finite dimensional space $\Omega^2(\Gamma^{2d})$ of algebraic 2-forms

$$\omega = \sum_{(\sigma_1,\sigma_2)\in F\times F} \omega_{\sigma_1,\sigma_2} \frac{df_{\sigma_1} \wedge df_{\sigma_2}}{f_{\sigma_1}\, f_{\sigma_2}}\,, \tag{21.6}$$

where $\Omega = (\,\omega_{\sigma_1,\sigma_2}\,)_{(\sigma_1,\sigma_2)\in F\times F}$ is a skew-symmetric matrix such that $\omega_{\sigma_1,\sigma_2} = 0$ whenever σ_1 and σ_2 are disjoint faces. Any algebraic form $\omega \in \Omega^2(\Gamma^{2d})$ determines the linear 2-form $\widehat{\omega} : \mathbb{R}^F \times \mathbb{R}^F \to \mathbb{R}$, $\widehat{\omega}(X,Y) = X^T \Omega Y$, which by restriction induces a piecewise linear 2-form on $\mathscr{C}^*(\Gamma^{2d})$ still denoted by $\widehat{\omega}$. Conversely, assume we are given a continuous piecewise linear 2-form $\widehat{\omega}$ on $\mathscr{C}^*(\Gamma^{2d})$. This is a family of linear 2-forms $\widehat{\omega}^A : \Pi_A \times \Pi_A \to \mathbb{R}$, one on each face Π_A with $A \in V$, such that $\widehat{\omega}^A = \widehat{\omega}^B$ on Π_γ, for every pair of vertices $A, B \in V$ connected by some edge γ. Under such conditions the piecewise linear 2-form $\widehat{\omega}$ is determined by a skew-symmetric matrix $\Omega = (\,\omega_{\sigma_1,\sigma_2}\,)_{(\sigma_1,\sigma_2)\in F\times F}$ as above, and is therefore associated to an algebraic 2-form $\omega \in \Omega^2(\Gamma^{2d})$. Given an algebraic 2-form $\omega \in \Omega^2(\Gamma^{2d})$, if ω is non-degenerate at every point interior to Γ^{2d} then ω is a symplectic structure on the interior of Γ^{2d}, that we refer as an algebraic symplectic structure. The third theorem says the symplectic gradient of a function in $\mathscr{H}(\Gamma^{2d})$ w.r.t. an algebraic symplectic structure in $\Omega^2(\Gamma^{2d})$ is, up to time reparametrization, a regular vector field in $\mathscr{X}(\Gamma^{2d})$.

Theorem 21.3. *Given an algebraic symplectic structure $\omega \in \Omega^2(\Gamma^{2d})$, and a regular function $h \in \mathscr{H}(\Gamma^{2d})$ of order κ, the symplectic gradient X_h of h w.r.t. ω is equivalent to the regular vector field $X = p\, X_h$ on Γ^{2d} with the same order $\nu_X = \kappa$, where $p = \prod_{\sigma\in F} (f_\sigma)^{\kappa(\sigma)-1} \geq 0$.*

Given an order function $\nu : F \to \mathbb{N}$ and $\omega \in \Omega^2(\Gamma^{2d})$ given by (21.6), we define the *reduced algebraic form*

$$\omega^\nu = \sum_{(\sigma_1,\sigma_2)\in F\times F} \omega^\nu_{\sigma_1,\sigma_2} \frac{df_{\sigma_1} \wedge df_{\sigma_2}}{f_{\sigma_1}\, f_{\sigma_2}}\,,$$

where

$$\omega^\nu_{\sigma_1\,\sigma_2} = \begin{cases} \omega_{\sigma_1\,\sigma_2} & \text{if } \nu(\sigma_1) = \nu(\sigma_2) = 1 \\ 0 & \text{otherwise} \end{cases} .$$

Next theorem says the rescaling limit of an algebraic form $\omega \in \Omega^2(\Gamma^{2d})$ is the piecewise linear reduced form $\widehat{\omega^\nu}$.

Theorem 21.4. *Given an order function $\nu : F \to \mathbb{N}$, $\omega \in \Omega^2(\Gamma^{2d})$, and $A \in V$, then as $\varepsilon \to 0^+$ the rescaled form $\varepsilon^2\, [(\Psi^\nu_\varepsilon)^{-1}]^*\omega$ tends in the C^∞-topology* *and uniformly on compact subsets in the interior of Π_A to the piecewise linear 2-form $\widehat{\omega^\nu}$.*

Corollary 21.1. *Consider $\omega \in \Omega^2(\Gamma^{2d})$, $h \in \mathscr{H}(\Gamma^{2d})$ and $X \in \mathscr{X}(\Gamma^{2d})$ as above. The skeleton χ of X is, up to some constant, the gradient of the skeleton λ_h w.r.t. $\widehat{\omega^\nu}$, i.e., for every $A \in V$ and $u \in \Pi_A$, $\lambda_h(\chi^A) = 0$ and $\widehat{\omega^\nu}(\chi^A, u) = p(A)\,\lambda_h(u)$, where p is the function referred in Theorem 21.3.*

Corollary 21.2. *Under the same assumptions, if all components of λ_h have the same sign (positive or negative), then every $A \in V$ is a χ-saddle and almost all orbits of χ are defined for all time.*

Two important subclasses of Lotka–Volterra systems, already studied by Volterra, are the so called dissipative and conservative systems. A Lotka–Volterra system, with interaction matrix A, is said to be *conservative* if there is a positive diagonal matrix D such that $A\,D$ is skew-symmetric. On even dimensions, if conservative system (21.2) is Hamiltonian with respect to the symplectic structure on $\mathbb{R}^{2d}_+$

$$\omega = \sum_{i,j=1}^{2d} a^{-1}_{i,j}\, \frac{dx_i \wedge dx_j}{x_i\, x_j}\;,$$

where $a^{-1}_{i,j}$ is the coefficient of the inverse matrix A^{-1}. In general, a conservative Lotka–Volterra system is Hamiltonian with respect to the Poisson structure on $\mathbb{R}^d_+$

$$\{f, g\} = \frac{1}{2} \sum_{i,j=1}^{d} a_{i,j}\, x_i\, x_j \left(\frac{\partial f}{\partial x_i}\, \frac{\partial g}{\partial x_j} - \frac{\partial g}{\partial x_i}\, \frac{\partial f}{\partial x_j} \right).$$

In any case, if q is a solution of the equation $r + A\,q = 0$, where r and A are the Lotka–Volterra coefficient matrices, then the Hamiltonian function $h : \mathbb{R}^d_+ \to \mathbb{R}$ is

$$h(x_1, \ldots, x_d) = \sum_{i=1}^{d} (x_i - q_i\, \log x_i), \qquad (21.7)$$

which is, of course, a first integral for (21.2).

A Lotka–Volterra is called *dissipative* if there is a diagonal matrix $D > 0$ such that $A\,D \leq 0$. In this case, the system admits the global Lyapounov function (21.7). In [2] we have proved a result which further motivates the study of conservative Lotka–Volterra systems:

Theorem 21.5. *Every stably dissipative Lotka–Volterra system, with a singularity interior to $\mathbb{R}^d_+$, has a global attractor where the dynamics is that of a conservative Lotka–Volterra system.*

The non-zero entries of the matrix A determine a *food chain graph* $G(A)$ with the eating relations within the ecosystem $\{1, 2, \ldots, d\}$. A Lotka–Volterra system is said to be *stably dissipative* iff every nearby Lotka–Volterra system with the same food chain graph is still dissipative. Next theorem states that all linear replicator systems (21.3), in the simplex Δ^{2d}, which come from a conservative Lotka–Volterra system, fall in the scope of Theorem 21.3, i.e., they are time reparametrizations of symplectic gradients of functions in $\mathscr{H}(\Delta^{2d})$ w.r.t. algebraic symplectic structures.

Theorem 21.6. *The replicator equation on Δ^{2d} corresponding to a conservative Lotka–Volterra system on $\mathbb{R}^{2d}$ given by some invertible coefficient matrix, is, up to equivalence, the symplectic gradient of a regular function $h \in \mathscr{H}(\Delta^{2d})$ of the form*

$$h(x_0, \ldots, x_{2d}) = \sum_{i=1}^{2d} \frac{x_i}{x_0} - q_i \, \log \frac{x_i}{x_0}$$

w.r.t. to some algebraic symplectic structure $\omega \in \Omega^2(\Delta^{2d})$.

21.5 An Application

In [2] we have analyzed the following Lotka–Volterra system, a four species food chain which couples two independent predator–prey systems

$$\begin{cases} y_1' = y_1\,(-1 + y_2) \\ y_2' = y_2\,(1 - y_1 + \delta\, y_3) \\ y_3' = y_3\,(-1 - \delta\, y_2 + y_4) \\ y_4 = y_4\,(1 - y_3) \end{cases} \tag{21.8}$$

where the coupling strength is controlled by the parameter δ. The coefficient matrices of this system are

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & \delta & 0 \\ 0 & -\delta & 0 & 1 \\ 0 & 0 & -1 & 0 \end{pmatrix} \quad \text{and} \quad r = \begin{pmatrix} -1 \\ 1 \\ -1 \\ 1 \end{pmatrix}.$$

We prove in [2] system (21.8) is non-integrable for any $\delta \neq 0$. There we pay special attention to a family of periodic orbits $\Gamma = \Gamma(\delta, E)$ defined, for all δ, as the intersection of the energy level $\{h = E\}$ with the following invariant 2-plane

$$\Pi = \{\, (y_1, y_2, y_3, y_4) \in \mathbb{R}^4_+ \,:\, y_1 = (1+\delta)\, y_3,\ y_4 = (1+\delta)\, y_2 \,\}\,.$$

There is a 3-plane containing Π, slicing transversally all energy levels in 2-spheres. The orbit Γ splits each of these 2-spheres in two disks transversal to the flow. The first return map, along the flow, to any of these disks is a continuous map which determines the dynamics in that energy level. Finally, the periodic orbit Γ has rotation number which tends to $+\infty$ with the energy level E, and its character alternates between stable (elliptic) and unstable (hyperbolic), as δ varies in $(0, +\infty)$. Furthermore, there is a sequence of small intervals of the parameter δ, where as $E \to +\infty$, the periodic orbit Γ becomes hyperbolic with arbitrary large trace.

In [1] we pursue the analysis of this system proving that

Theorem 21.7. *For* $0 < \delta < 1$*, the Lotka–Volterra system (21.8) has, in all sufficiently large energy level* $\{\, h = E \,\}$*, a non-trivial invariant hyperbolic basic set of saddle type.*

To prove this theorem we consider the replicator vector field $X \in \mathscr{X}(\Delta^4)$ in (21.3), associated with the Lotka–Volterra system (21.8). We denote by σ_i the face of Δ^4 opposed to vertex i, and by $\gamma_{i,j}$ the edge connecting the vertices i and j. We have $\nu_X(\sigma_0) = 2$ and $\nu_X(\sigma_i) = 1$, for $i = 1, 2, 3, 4$. Let χ be the skeleton of X (Fig. 21.8).

We can compute the following chains for χ, where $*$ stands for the chain concatenation operation.

$$
\begin{array}{ll}
\xi^0 = (\gamma_{4,0}, \gamma_{0,1}) & \xi^1 = (\gamma_{0,1}, \gamma_{1,2}, \gamma_{2,0}, \gamma_{0,1}) \\
\xi^2 = (\gamma_{0,1}, \gamma_{1,2}, \gamma_{2,3}, \gamma_{3,4}, \gamma_{4,0}) & \xi^3 = (\gamma_{0,1}, \gamma_{1,2}, \gamma_{2,0}, \gamma_{0,3}, \gamma_{3,4}, \gamma_{4,0}) \\
\xi^4 = (\gamma_{4,0}, \gamma_{0,3}, \gamma_{3,4}, \gamma_{4,0}) & \\
\xi^5_n = \xi^0 * (\xi^1)^n * \xi^2 & \xi^6_n = \xi^0 * (\xi^1)^n * \xi^3 \qquad (n \geq 0)
\end{array}
$$

There are exactly four families of χ-cycles which start and end with $\gamma_{4\,0}$ but do not pass through this edge in between. They are $\{\xi^4\}$, $\{\, \xi^5_n \,:\, n \geq 0 \,\}$ and $\{\, \xi^6_n \,:\, n \geq 0 \,\}$. Whence the first return map $R^{\chi}_{\gamma_{4,0}}$ to $\Pi_{4\,0}$ is given by

$$
R^{\chi}_{\gamma_{4,0}}(X) = \begin{cases} M_{\xi^4}\, X & \text{if } \; X \in \Pi_{\xi^4} \\ M_{\xi^5_n}\, X & \text{if } \; X \in \Pi_{\xi^5_n}\,, \quad n \geq 0 \\ M_{\xi^6_n}\, X & \text{if } \; X \in \Pi_{\xi^6_n}\,, \quad n \geq 0 \end{cases} ,
$$

whose domain, the union of the open convex cones $\Pi_{\xi^4} \cup \bigcup_{n=0}^{\infty} \Pi_{\xi^5_n} \cup \bigcup_{n=0}^{\infty} \Pi_{\xi^6_n}$, can be characterized as follows.

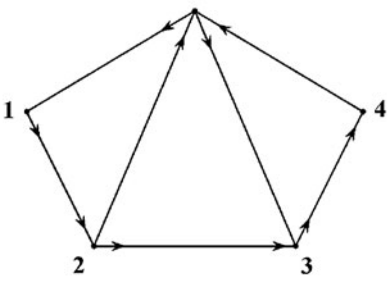


Fig. 21.8 The oriented graph $G_\chi(\Delta^4)$ consists of the 7 edges

Proposition 21.1. *The open cones Π_{ξ^4}, $\Pi_{\xi^5_n}$ and $\Pi_{\xi^6_n}$ $(n \geq 0)$ are defined by the following inequalities:*

1. *Π_{ξ^4} by $u_0 = u_4 = 0$, $-u_1 + u_3 < 0$, $u_1 > 0$ and $u_2 > 0$.*
2. *$\Pi_{\xi^5_n}$ by $u_0 = u_4 = 0$, $u_1 > 0$, $u_2 > 0$ and*

$$\frac{-u_1 + u_3}{(1+\delta)(u_1+u_2)} - \frac{\delta}{1+\delta} < n < \frac{-u_1 + u_3}{(1+\delta)(u_1+u_2)} \,.$$

3. *$\Pi_{\xi^6_n}$ by $u_0 = u_4 = 0$, $u_1 > 0$, $u_2 > 0$ and*

$$\frac{-u_1 + u_3}{(1+\delta)(u_1+u_2)} - 1 < n < \frac{-u_1 + u_3}{(1+\delta)(u_1+u_2)} - \frac{\delta}{1+\delta} \,.$$

A simple computation shows that

$$M_{\xi^4} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 1 & 1+\delta & \delta \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$M_{\xi^5_n} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ * & -n - \frac{n+1}{\delta} & -n - \frac{n}{\delta} & \frac{1}{\delta} & 0 \\ * & (n+1)\delta & n\delta & 0 & \delta \\ * & 2n+2+\frac{n+1}{\delta} & 2n+1+\frac{n}{\delta} & -\frac{1}{\delta} & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

and

$$M_{\xi^6_n} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ * & n+2+(n+1)\delta & n+1+(n+1)\delta & -1 & 0 \\ * & -(n+1)-(n+1)(\delta+\delta^2) & -n-(n+1)(\delta+\delta^2) & 1+\delta & \delta \\ * & -(n+1)\delta & -(n+1)\delta & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

The "$*$" entries are not important, since we are only interested in the action of these matrices on the 3-plane $u_0 = u_4 = 0$ spanned by the cone $\Pi_{\gamma_{0,4}}$. Actually, the action of these matrices on $\Pi_{\gamma_{0,4}}$ is determined by the inner 3×3 submatrices of the above ones. By Theorem 21.6 this system is, up to a time reparametrization, the symplectic gradient of the following Hamiltonian

$$\begin{aligned} h(x_0,\ldots,x_4) = {} & \frac{x1 + x2 + x3 + x4}{x_0} + (1+\delta)\log\frac{x_1}{x_0} + \log x_2 \\ & + \log x_3 + (1+\delta)\log\frac{x_4}{x_0} \end{aligned}$$

w.r.t. some algebraic symplectic structure. Whence by Corollary 21.1, $\lambda_h : \mathscr{C}^*(\Gamma^d) \to \mathbb{R}$ is invariant under the flow of χ. We have $\lambda_h(u) = (1+\delta)\, u_1+u_2+u_3$, for every $u \in \Pi_{\gamma_{0,4}}$. Consider now the 2-simplex $\Delta^2 = \{ u \in \Pi_{\gamma_{0,4}} : \lambda_h(u) = 1\}$, which is invariant under $R^{\chi}_{\gamma_{0,4}}$, and denote by $T : \Delta^2 \to \Delta^2$ the restriction of $R^{\chi}_{\gamma_{0,4}}$ to this simplex. For each cycle ξ through $\gamma_{0,4}$ we define $\Delta_\xi = \{ u \in \Pi_\xi : \lambda_h(u) = 1\}$. Each restriction $T_\xi = T|_{\Delta^\xi}$ is an affine map, which we can compute explicitly, as well as its domain Δ_ξ and range $T_\xi(\Delta_\xi)$, for every cycle ξ through $\gamma_{4,0}$. With this notation, Δ^2 is the disjoint union (mod 0) of the polygons

$$\Delta_{\xi^4},\ \Delta_{\xi^5_0},\ \Delta_{\xi^6_0},\ \Delta_{\xi^5_1},\ \Delta_{\xi^6_1},\ \Delta_{\xi^5_2},\ \Delta_{\xi^6_2},\ \ldots .$$

Figure 21.9 shows these polygons, as well as their T-images, labeled in this order.

We can check that the affine map $T_\xi : \Delta_\xi \to \Delta^2$ is

1. Parabolic for $\xi = \xi^4$, for all $0 < \delta < 1$.
2. Elliptic for $\xi = \xi^6_n$, $n \geq 0$, for some $0 < \delta < 1$.
3. Hyperbolic with negative trace for $\xi = \xi^5_n$, $n \geq 0$, $0 < \delta < 1$.

For $0 < \delta < 1$ we compute the following two hyperbolic fixed points:

1. $P_0 = \left(\frac{1}{2+3\delta}, \frac{\delta}{2+3\delta}, \frac{1+\delta}{2+3\delta}\right) \in \Delta_{\xi^5_0}, \qquad P_0 = T_{\xi^5_0}(P_0)$, and
2. $P_1 = \left(\frac{1-\delta}{3+4\delta-\delta^2}, \frac{2\delta}{3+4\delta-\delta^2}, \frac{2+2\delta}{3+4\delta-\delta^2}\right) \in \Delta_{\xi^5_1}, \quad P_1 = T_{\xi^5_1}(P_1)$.

We define the local invariant manifolds of these hyperbolic fixed points as follows: $W^s_{\text{loc}}(P_i)$ is the intersection of the line through P_i parallel to the contracting eigenspace of P_i, w.r.t. the linear part of $T_{\xi^5_i}$, with the polygon $\Delta_{\xi^5_i}$, while $W^u_{\text{loc}}(P_i)$ is the intersection of the line through P_i parallel to the expanding eigenspace of P_i with the image polygon $T_{\xi^5_i}(\Delta_{\xi^5_i})$. Using them we define the global manifolds

$$W^s(P_i) = \bigcup_{n\geq 0} T^{-n} W^s_{\text{loc}}(P_i) \quad \text{and} \quad W^u(P_i) = \bigcup_{n\geq 0} T^{n} W^u_{\text{loc}}(P_i)\ .$$

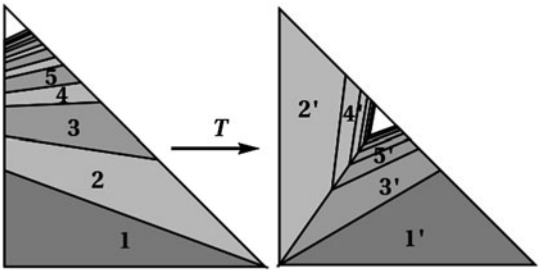


Fig. 21.9 Domain and range of the return map $T : \Delta^2 \to \Delta^2$

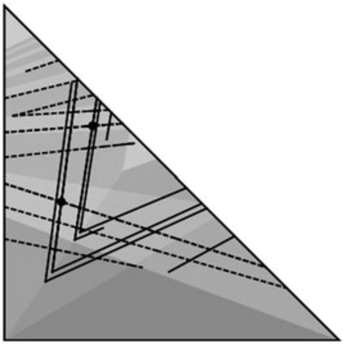


Fig. 21.10 Heteroclinic intersections of the return map $T : \Delta^2 \to \Delta^2$

Then we can prove that

Proposition 21.2. *For all $\delta \in (0,1)$,*

$$W^s_{loc}(P_0) \cap W^u(P_1) \neq \emptyset \quad \textit{and} \quad W^s_{loc}(P_1) \cap W^u_{loc}(P_0) \neq \emptyset\,,$$

with transversal intersections.

In Fig. 21.10, the filled lines represent unstable manifolds of P_0 and P_1, while the dashed lines represent stable manifolds.

By Proposition 21.2, the map T has a transversal heteroclinic cycle formed of two heteroclinic orbits. Because these orbits accumulate on the fixed points they stay at positive distance of the boundaries $\partial\Delta_{\xi_i^5}$ $(i = 0, 1)$. Using them we can construct an invariant hyperbolic basic set of saddle type $\Lambda \subset \Delta_{\xi_0^5} \cup \Delta_{\xi_1^5}$, for the map T, still at a positive distance of $\partial\Delta_{\xi_0^5} \cup \partial\Delta_{\xi_1^5}$. By Theorem 21.1, in all sufficiently large energy level surface the system must have a conjugate invariant hyperbolic basic set of saddle type $\Lambda_E \subset \{h = E\}$, which concludes the argument for Theorem 7.

21.6 Conclusions

We finish with some related questions and possible generalizations.

The analyticity assumption was mainly aesthetic, everything works fine for smooth systems. One can also adapt the argument to work with compact manifolds with boundary, instead of simple polyhedra. Recall that a compact manifold with boundary, say M^d of dimension d, is one which at every point is locally diffeomorphic to a model $\left(\mathbb{R}^k \times \mathbb{R}^{d-k}_+, 0\right)$, for some $0 \le k \le d$. The integer k is called the index of M^d at that point. The set of all points with index k, denoted by $\partial_k(M^d)$, is exactly the union of all interiors of k-dimensional faces of the manifold M^d.

General algorithms can be developed to facilitate the analysis of a skeleton vector field's dynamics. In [2] we describe how to derive the skeleton vector field components from the payoff matrix of a replicator system. Similar relations can be driven for other Game Theory systems.

Vicinity relations of a cone domain Π_ξ should translate to symbolic kneading relations of the corresponding chain, or cycle, ξ. Such a kneading theory would be a very useful instrument of analysis.

Given a skeleton vector field, can we realize it as the edge asymptotics of some regular vector field? This realization is important to construct examples with prescribed dynamical behavior along the edges. For general regular vector fields the answer to this problem is positive. Every regular skeleton vector field χ in $\mathscr{C}^*(\Gamma^d)$ is the skeleton of some regular vector field $X \in \mathscr{X}(\Gamma^d)$. For conservative skeleton vector fields, the answer is yes locally, in a neighborhood of the 1-dimensional skeleton of Γ^{2d}. If a skeleton vector field χ of $\mathscr{C}^*(\Gamma^d)$ is the symplectic gradient of a skeleton function $\lambda : \mathscr{C}^*(\Gamma^d) \to \mathbb{R}$ w.r.t. a piecewise linear symplectic structure $\widehat{\omega}$ on $\mathscr{C}^*(\Gamma^d)$ then λ is the skeleton of a function $h \in \mathscr{H}(\Gamma^{2d})$ and $\widehat{\omega}$ is associated to some algebraic form $\omega \in \Omega^2(\Gamma^{2d})$. Whence, the symplectic gradient X_h of h w.r.t. ω is, as in Theorem 21.3, equivalent to a regular vector field X whose skeleton will be χ. The problem with this approach is that it's not clear if ω is non degenerate everywhere, i.e., if ω is a symplectic structure on the interior of Γ^{2d}. In this case the gradient X_h may not be defined everywhere in Γ^{2d}. This raises the question of characterizing the subset of symplectic structures in $\Omega^2(\Gamma^{2d})$. We can avoid this problem dealing with Poisson structures instead of symplectic ones. We believe that a concept of "algebraic Poisson structure" can be defined on the polyhedron Γ^d, as well as a class of Hamiltonian systems with Hamiltonians in $\mathscr{H}(\Gamma^d)$ w.r.t. such algebraic Poisson structures, which up to equivalence give rise to regular vector fields in $\mathscr{X}(\Gamma^d)$. Then Theorems 21.3, 21.4 and 21.6 should generalize to arbitrary dimensions.

Skeleton vector field's bifurcations is another interesting subject of study. These bifurcations are caused by changes in the geometry and combinatorics of the domain and image partitions of the return map $R^\chi_\gamma : \Pi_\gamma \to \Pi_\gamma$, respectively $\{ \Pi_\xi \}_\xi$ and $\{ R^\chi_\gamma(\Pi_\xi) \}_\xi$ where ξ varies on the set of all χ-cycles which start and end with γ but do not pass through γ in between. Considering skeletons of regular vector fields in $\mathscr{X}(\Gamma^d)$, it should be possible to relate these skeleton bifurcations with the bifurcations of the underlying vector field.

In Theorem 21.7, for simplicity, we have assumed $\delta \in (0, 1)$, but we believe that the same holds for all $\delta > 0$. The reason we made such restriction is that for $\delta > 1$ the dynamics is harder to analyze due to the presence of the elliptic fixed point P_0 in the main branch $T_{\xi_0^5} : \Delta_{\xi_0^5} \to \Delta^2$.

Figure 21.11 shows ten different orbits, with a couple of hundred iterates each, for a particular parameter. The shaded regions represent the polygon $\Delta_{\xi_0^5}$, on the left, and its image $T_{\xi_0^5}(\Delta_{\xi_0^5})$, on the right. The invariant curves break up as they touch the boundary of their domains. Outside these curves, the dynamics seems to be chaotic, which indicates the presence of hyperbolicity. Concerning the parameter interval $(0, 1)$, we pose some more questions. Are there elliptic periodic points for parameters $0 < \delta < 1$? Is this true for many parameters? The Newhouse phenomenon, of persistent homoclinic tangencies associated with large thickness hyperbolic sets,

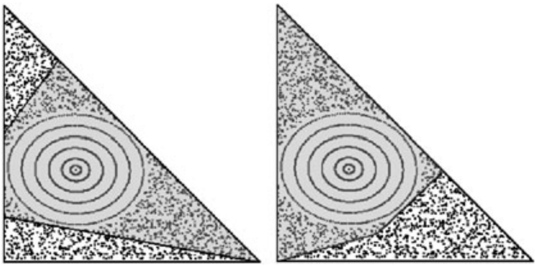


Fig. 21.11 An elliptic fixed point P_0 at $\delta = 3.7$

is a mechanism for the appearance of many elliptic structures in the dynamics of the underlying Hamiltonian vector field. See for instance [3]. As $\delta \to 0^+$ can one find large uniformly hyperbolic basic sets with very large thickness? Is this also a mechanism for the creation of many elliptic periodic points of the skeleton vector field? It would be interesting to understand, for conservative skeleton vector fields, the mechanism for the creation of elliptic structures, and then relate it with the corresponding homoclinic bifurcation mechanism of the underlying dynamics of vector fields in $\mathscr{X}(\Gamma^{2d})$.

References

1. Duarte, P.: Dynamics along the edges of simple polyhedrons (2006) (Preprint). http://cmaf.ptmat.fc.ul.pt/preprints/preprints.html
2. Duarte, P., Fernandes, R.L., Oliva, W.: Dynamics on the attractor of Lotka–Volterra equations. J. Differ. Equ. **149**, 143–189 (1998)
3. Duarte, P.: Persistent homoclinic tangencies for maps near the identity. Ergod. Theory Dyn. Syst. **20**, 393–438 (2000)
4. Hofbauer, J.: On the occurrence of limit cycles in the Lotka–Volterra equation. Nonlinear Anal. **5**, 1003–1007 (1981)
5. Hofbauer, J., Sigmund, K.: Evolutionary Games and Dynamical Systems. Cambridge University Press, Cambridge (1998)

Chapter 22
Bankruptcy Boundaries Determined by Patents

M. Ferreira, B.M.P.M. Oliveira, and Alberto A. Pinto

Abstract We use a new R&D investment function in a Cournot competition model inspired in the logistic equation. We present the full characterization of the associated game and study the short and long term economical effects derived from using this new R&D investment function. We observe the existence of four different Nash investment equilibria regions and fully characterize the boundaries of these regions.

22.1 Introduction

We consider a Cournot competition model where two firms invest in R&D projects to reduce their production costs. This competition is modeled, as usual, by a two stage game (see d'Aspremont and Jacquemin [2]). In the first subgame, two firms choose, simultaneously, the R&D investment strategy to reduce their initial production costs. In the second subgame, the two firms are involved in a Cournot competition with production costs equal to the reduced cost determined by the R&D investment program. We use an R&D cost reduction function inspired in the logistic equation (see Equation (2) in [7]) that was first introduced in Ferreira et al. [7].

M. Ferreira (✉)
Escola Superior de Estudos Industriais e de Gestão do Instituto Politécnico do Porto (IPP), LIAAD - INESC Porto LA, Escola de Ciências, Universidade do Minho, Braga, Portugal
e-mail: migferreira2@gmail.com

B.M.P.M. Oliveira
Faculdade de Ciências da Nutrição e Alimentação da, Universidade do Porto, LIAAD-INESC Porto LA, Porto, Portugal
e-mail: bmpmo@fcna.up.pt

A.A. Pinto
LIAAD-INESC Porto LA e Departamento de Matemática, Faculdade de Ciências,
Universidade do Porto, Rua do Campo Alegre 687, 4169-007 Porto, Portugal
and
Centro de Matemática e Departamento de Matemática e Aplicações, Escola de Ciências,
Universidade do Minho, Campus de Gualtar, 4710-057 Braga, Portugal
e-mail: aapinto@fc.up.pt

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_22,

The main differences to the standard R&D cost reduction function (see [2]) are explained therein. For the first subgame, consisting of an R&D investment program, we observe the existence of four different Nash investment equilibria regions that we define as follows (see [7]): a competitive Nash investment region C where both firms invest; a single Nash investment region S_1 for firm F_1, where just firm F_1 invests; a single Nash investment region S_2 for firm F_2, where just firm F_2 invests; and a nil Nash investment region N, where neither of the firms invest.

The nil Nash investment region N is the union of four disjoint sets: the set N_{LL} consisting of all production costs that are low for both firms; the set N_{LH} (resp. the set N_{HL}) consisting of all production costs that are low for firm F_1 (resp. F_2) and high for firm F_2 (resp. F_1); and the set N_{HH} consisting of all production costs that are high for both firms.

The single Nash investment region S_i can be decomposed into two disjoint regions: a *single favorable Nash investment region* S_i^F where the production costs, after investment, are favorable to firm F_i; and a *single recovery Nash investment region* S_i^R where the production costs, after investment are, still, favorable to firm F_j but firm F_i recovers, slightly, from its initial disadvantage. In the single recovery region S_i^R, the production costs of firm F_j are too low for the firm F_j to be willing further decrease its production costs and, therefore, firm F_i is able to decrease its production costs by investing. The single favorable region S_i^F can also be decomposed into three regions: the *single duopoly region* S_i^D; the *single monopoly region* S_i^M, and the *single monopoly boundary region* S_i^B. The single monopoly region S_i^M consists of all production costs such that, after firm F_i's investment, the new production costs are in the monopoly region of firm F_i. The single monopoly boundary region S_i^B consists of all production costs such that, after firm F_i's investment, the new production costs are in the boundary between the monopoly region and the duopoly region of firm F_i. The single duopoly region S_i^D consists of all production costs such that, after the firm F_i's investment, the new production costs are still in the duopoly region of firm F_i (see Fig. 22.3).

The competitive Nash investment region determines the region where the production costs of both firms evolve over time. The single Nash investment region S_1 determines the set of production costs where the production cost of firm F_2 is constant, over time, and just the production costs of firm F_1 evolve. Similarly, the single Nash investment region S_2 determines the set of production costs where the production cost of firm F_1 is constant, over time, and just the production costs of firm F_2 evolve. The nil Nash investment region N determines the set of all production costs that are fixed by the dynamics.

In this paper we characterize, in detail, the boundaries of each of these Nash investment regions computed in the research papers [7] and [8].

22.2 The Model

The Cournot competition with R&D investment programs to reduce the production costs consists of two simultaneous subgames. The first subgame is an R&D investment program, where both firms have initial production costs and simultaneously choose their R&D investment strategies to obtain lower new production costs. The second subgame is a typical Cournot competition on quantities with production costs equal to the reduced costs determined by the R&D investment program. As it is well known, the second subgame has a unique perfect Nash equilibrium. The analysis of the first subgame is of higher complexity and is covered with detail in Ferreira et al. [7].

22.2.1 New Production Costs

The sets of possible new production costs for firms F_1 and F_2, given initial production costs c_1 and c_2 are, respectively,

$$A_1 = A_1(c_1, c_2) = [b_1, c_1] \quad \text{and} \quad A_2 = A_2(c_1, c_2) = [b_2, c_2],$$

where $b_i = c_i - \epsilon(c_i - c_L)$, for $i \in \{1, 2\}$.

The R&D programs a_1 and a_2 of the firms determine a bijection between the *investment region* $R_0^+ \times R_0^+$ of both firms and the *new production costs region* $A_1 \times A_2$, given by the map

$$\begin{aligned} \mathbf{a} = (a_1, a_2)\ : R_0^+ \times R_0^+ &\longrightarrow \qquad A_1 \times A_2 \\ (v_1, v_2)\ &\longmapsto (a_1(v_1), a_2(v_2)) \end{aligned}$$

where

$$a_i(v_i) = c_i - \frac{\eta_i v_i}{\lambda + v_i}.$$

We denote by $W = (W_1, W_2) : \mathbf{a}\left(R_0^+ \times R_0^+\right) \to R_0^+ \times R_0^+$

$$W_i(a_i) = \frac{\lambda(c_i - a_i)}{a_i - c_i - \eta_i}$$

the inverse map of $\mathbf{a}$.

The new production costs region can be decomposed, at most, in three disconnected economical regions characterized by the optimal output level of the firms (see Fig. 22.1):

M_1 The *monopoly region* M_1 of firm F_1 that is characterized by the optimal output level of firm F_1 being the monopoly output and, therefore, the optimal output level of firm F_2 is zero.

D The *duopoly region* D that is characterized by the optimal output levels of both firms being non-zero and, therefore, below their monopoly output levels.

M_2 The *monopoly region* M_2 of firm F_2 that is characterized by the optimal output level of firm F_2 being the monopoly output and, therefore, the optimal output level of firm F_1 is zero.

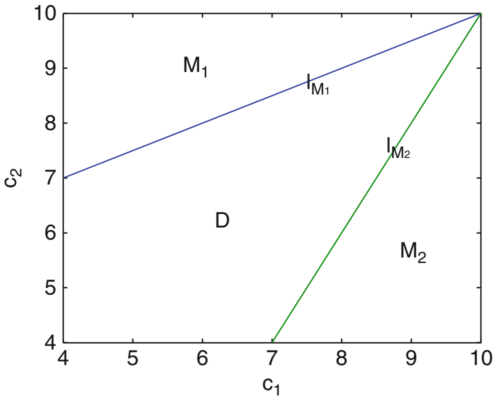


Fig. 22.1 We exhibit the duopoly region D and the monopoly regions M_1 and M_2 for firms F_1 and F_2, respectively, in terms of their new production costs (a_1, a_2); l_{M_i} with $i \in \{1,2\}$ are the boundaries between M_i and D

The boundary between the duopoly region D and the monopoly region M_i is l_{M_i} with $i \in \{1,2\}$.

The explicit expression characterizing l_{M_i}, the boundary between the monopoly region M_i and the duopoly region D, is presented in [7].

22.2.2 Best R&D Investment Response Functions

To determine the *best investment response function* $V_1(v_2)$ of firm F_1 to a given investment v_2 of firm F_2, we study, separately, the cases where the new production costs $(a_1(v_1,v_2), a_2(v_1,v_2))$ belong to (a) the monopoly region M_1; (b) the duopoly region D; or (c) the monopoly region M_2.

If there is $v_1 \in R_0^+$ such that $(a_1(v_1), a_2(v_2)) \in M_1$, we select the best response v_{M_1} of firm F_1, restricted to $(a_1(v_{M_1}), a_2(v_2)) \in M_1$, to the investment v_2 of firm F_2 as follows: Let Z_{M_1} be the set of solutions v_1 of the following equation

$$\frac{\partial \pi_{1,M_1}}{\partial v_1} = 0,$$

such that $(a_1(v_1), a_2(v_2)) \in M_1$. Let F_{M_1} be the set of v_1 such that $(a_1(v_1), a_2(v_2)) \in l_{M_1}$. The best response v_{M_1} of firm F_1 in M_1 is given by

$$v_{M_1} = \arg \max_{v_1 \in Z_{M_1} \cup F_{M_1}} \pi_{1,M_1}(a_1(v_1), a_2(v_2)).$$

The set F_{M_1} is given explicitly in Lemma 1 in [7]. Since the investment v_2 is fixed, let us characterize the set Z_{M_i}.
Let $L_i = 6\beta\lambda^2 - \lambda\eta_i^2 - \eta_i\lambda(\alpha - c_i)$ and $N_i = 2\beta\lambda^3 - \eta_i\lambda^2(\alpha - c_i)$.

Theorem 22.1. *Let v_i be such that $(a_i(v_i), c_j) \in M_i$. The set Z_{M_i} is the set of zeros of the following polynomial:*

$$2\beta v_i^3 + 6\beta\lambda v_i^2 + L_i v_i + N_i = 0.$$

The order of the polynomial is three, and so the set Z_{M_1} can be explicitly computed.

The proof is in [7].

If there is $v_1 \in R_0^+$ such that $(a_1(v_1), a_2(v_2)) \in D$, we select the best response v_D of firm F_1, restricted to $(a_1(v_D), a_2(v_2)) \in D$, to the new production cost a_2 of firm F_2 as follows: Let Z_D be the set of zeros v_1 of the following polynomial

$$\frac{\partial \pi_{1,D}}{\partial v_1} = 0,$$

such that $(a_1(v_1), a_2(v_2)) \in D$. The best response v_D of firm F_1 in D is given by

$$v_D = \arg \max_{v_1 \in Z_D \cup F_{M_1}} \pi_{1,D}(a_1(v_1), a_2(v_2)).$$

Let us characterize the set Z_D. Let us define the following parameters

- $A_i = -4\beta^2\eta_i\lambda F_i$; $B_i = -4\beta^2\lambda\eta_i$.
- $C = \left(4\beta^2 - \gamma^2\right)^2$; $E_i = \alpha - c_i + \eta_i$.
- $F_i = 2\beta E_i - \gamma E_j$; $G_i = -2\beta\eta_i\lambda_i$ and $H_i = \gamma\eta_j\lambda$.

Theorem 22.2. *Let (v_1, v_2) be such that $(a_1(v_1), a_2(v_2)) \in D$. The set Z_D is the set of zeros of the following polynomials:*

$$C W_i^3 W_j + A_i W_i W_j + B_i W_j - (B_i/\lambda) H_i W_i = 0 \tag{22.1}$$

where $W_i = \lambda + v_i$ and $W_j = \lambda + v_j$.

Proof. (See [7]). □

If there is $v_1 \in R_0^+$ such that $(a_1(v_1), a_2(v_2)) \in M_2$, the best response v_{M_2} of firm F_1, restricted to $(a_1(v_{M_1}), a_2(v_2)) \in M_2$, is given by firm F_1 to invest zero, i.e. not investing. Hence, $V_1(v_2)$ is given by

$$V_1(v_2) = \arg \max_{v_1 \in F} \pi_1(a_1(v_1), a_2(v_2)),$$

where $V_1 \in F = Z_{M_1} \cup F_{M_1} \cup Z_D \cup \{0\}$.

Theorem 22.3. *The best investment response function $V_i : R_0^+ \to R_0^+$ of firm F_i is explicitly computed.*

We note that, the best investment response function $V_i : R_0^+ \to R_0^+$ can be multi-valued.

Proof. (See [7]). □

22.3 Nash Investment Equilibria

Let c_L be the minimum attainable production cost and α the market saturation. Given production costs $(c_1, c_2) \in [c_L, \alpha] \times [c_L, \alpha]$, the *Nash investment equilibria* $(v_1, v_2) \in R_0^+ \times R_0^+$ are the solutions of the system

$$\begin{cases} v_1 = V_1(v_2) \\ v_2 = V_2(v_1) \end{cases}$$

where V_1 and V_2 are the best investment response functions computed in the previous sections (See [7]).

All the results presented are consistent with [7] and hold in an open region of parameters $(c_L, \epsilon, \alpha, \lambda, \beta, \gamma)$ containing the point $(4, 0.2, 10, 10, 0.013, 0.013)$.

The Nash investment equilibria consists of one, two or three points depending upon the pair of initial production costs. The set of all Nash investment equilibria form the *Nash investment equilibrium set* (see Fig. 22.2):

C The *competitive Nash investment region* C that is characterized by both firms investing.

S_i The *single Nash investment region* S_i that is characterized by only one of the firms investing.

N The *nil Nash investment region* N that is characterized by neither of the firms investing.

In Fig. 22.2, the Nil Nash investment region is the union of N_{LL}, N_{LH}, N_{HL} and N_{HH} and the Single Nash investment region is the union of S_i^F and S_i^R. The economical meaning of the subregions of N and S_i is explained in the next sections.

Denote by $R = [c_L, \alpha] \times [c_L, \alpha]$ the region of all possible pairs of production costs (c_1, c_2). Let $A^c = R - A$ be the complementary of A in R and let $R_{A \cap B}$ be the intersection between the Nash investment region A and the Nash investment region B.

22.4 Single Nash Investment Region

The *single Nash investment region* S_i consists of the set of production costs (c_1, c_2) with the property that the Nash investment equilibrium set contains a pair (v_1, v_2) with the Nash investment $v_i = V_i(0) > 0$ and the Nash investment $v_j = V_j(v_i) = 0$, for $j \neq i$.

The single Nash investment region S_i can be decomposed into two disjoint regions: a *single favorable Nash investment region* S_i^F where the production costs, after investment, are favorable to firm F_i, and in a *single recovery Nash investment region* S_i^R where the production costs, after investment are, still, favorable to firm F_j but firm F_i recovers a little from its disadvantageous (see Fig. 22.3).

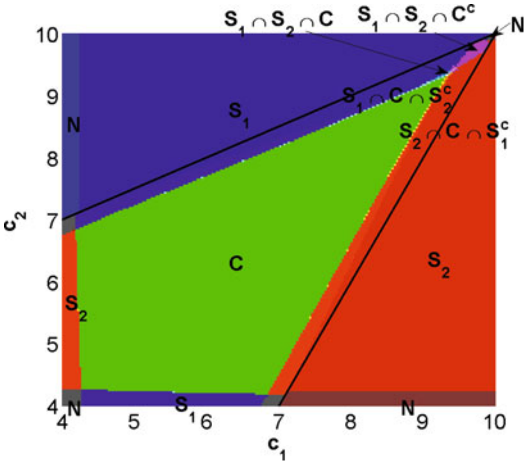


Fig. 22.2 Full characterization of the Nash investment regions in terms of the firms' initial production costs (c_1, c_2). The monopoly lines l_{M_i} are colored black. The nil Nash investment region N is colored grey. The single Nash investment regions S_1 and S_2 are colored blue and red, respectively. The competitive Nash investment region C is colored green. The region where S_1 and S_2 intersect is colored pink, the region where S_1 and C intersect is colored lighter blue and the region where S_2 and C intersect is colored yellow. The area where the regions S_1, S_2 and C intersect is colored lighter grey

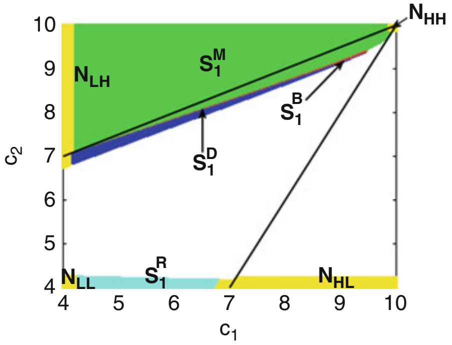


Fig. 22.3 Full characterization of the single Nash investment region S_1 and of the nil Nash investment region N in terms of the firms' initial production costs (c_1, c_2). The subregions N_{LL}, N_{LH}, N_{HL} and N_{HH} of the nil Nash investment region N are colored yellow. The subregion S_1^R of the single Nash investment region S_1 is colored lighter blue. The subregion S_1^F of the single Nash investment region S_1 is decomposed in three subregions: the *single Duopoly region* S_i^D colored blue, the *single Monopoly region* S_i^M colored green and the *single Monopoly boundary region* S_i^B colored red

The single favorable Nash investment region S_i^F can be decomposed into three regions: the *single Duopoly region* S_i^D, the *single Monopoly region* S_i^M and the *single Monopoly boundary region* S_i^B (see Fig. 22.3). For every cost $(c_1, c_2) \in S_i^F$, let $(a_1(v_1), a_2(v_2))$ be the Nash new investment costs obtained by the firms F_1 and

F_2 choosing the Nash investment equilibrium (v_1, v_2) with $v_2 = 0$. The single duopoly region S_i^D consists of all production costs (c_1, c_2) such that for the Nash new investment costs $(a_1(v_1), a_2(v_2))$ the firms are in the duopoly region D (see Fig. 22.3). The single monopoly region S_i^M consists of all production costs (c_1, c_2) such that for the Nash new costs $(a_1(v_1), a_2(v_2))$ the Firm F_i is in the interior of the Monopoly region M_i. The single monopoly boundary region S_i^B consists of all production costs (c_1, c_2) such that the Nash new investment costs $(a_1(v_1), a_2(v_2))$ are in the boundary of the Monopoly region l_{M_i}.

Theorem 22.4. *If the initial production cost* (c_1, c_2) *belongs to the single monopoly region* S_1^M *then* $v_i = V_i(0; c_1, c_2)$ *does not depend upon the value* c_j*, with* $i \neq j$.

Proof. The maximum profit for firm F_1 is attained at a point in the interior of the domain of π_{1,M_1}. Since π_{1,M_1} does not depend upon c_2, we get that v_1 does not depend upon c_2 either. □

We are going to characterize the boundary of the single monopoly region S_1^M (which, due to of the symmetry holds, a similar characterization for S_2^M). We study the boundaries of S_1^M by separating it in four distinct boundaries: the *upper boundary* $U_{S_1}^M$, that is the union of a vertical segment line $U_{S_1}^l$ with a curve $U_{S_1}^c$, the *intermediate boundary* $I_{S_1}^M$, the *lower boundary* $L_{S_1}^M$ and the *left boundary* $Le_{S_1}^M$ (see Fig. 22.4). The left boundary of the single monopoly region $Le_{S_1}^M$ is the right boundary d_1 of the nil Nash investment region N_{LH} that will be characterized in Sect. 22.5.

The boundary of the single monopoly boundary region S_1^B is the union of an *upper boundary* $U_{S_1}^B$ and a *lower boundary* $L_{S_1}^B$ (see Fig. 22.8).

The boundary of the single duopoly region S_1^D is the union of an *upper boundary* $U_{S_1}^D$, a *lower boundary* $L_{S_1}^D$ and a *left boundary* $Le_{S_1}^D$ (see Fig. 22.9). The left boundary of the single duopoly region $Le_{S_1}^D$ is the right boundary d_3 of the nil Nash investment region N_{LH} that will be characterized in Sect. 22.5.

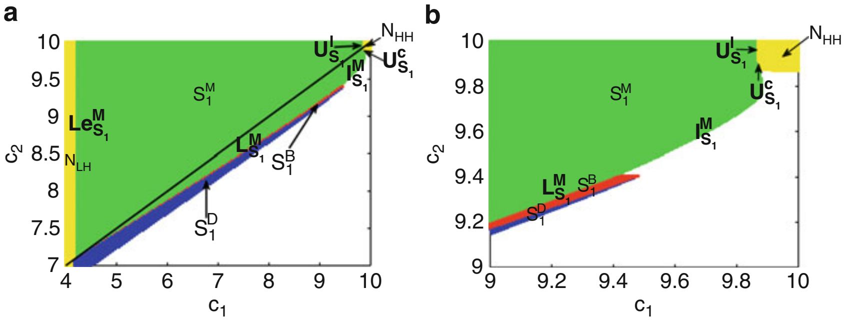


Fig. 22.4 (**a**) Full characterization of the boundaries of the single monopoly region S_1^M: the upper boundary $U_{S_1}^C$ is the union of a vertical segment line $U_{S_1}^l$ with a curve $U_{S_1}^c$; the lower boundary $L_{S_1}^M$; and the left boundary $Le_{S_1}^M$; (**b**) close up of the upper part of figure (**a**) where the boundaries $U_{S_1}^C$ and $U_{S_1}^l$ can be seen in more detail

The single recovery Nash investment region S_1^R has three boundaries: the *upper boundary* $U_{S_1}^R$, the *left boundary* $Le_{S_1}^R$, and the *right boundary* $R_{S_1}^R$ (see Fig. 22.11).

22.4.1 Boundary of the Single Monopoly Region S_1^M

In the following Lemmas we characterize, the boundaries of the single monopoly region S_1^M. Let us characterize the boundary $U_{S_1}^l$ between the single monopoly region S_1^M and the nil Nash investment region N_{HH} with initial production costs (c_1, c_2) in the Monopoly region M_1. The boundary $U_{S_1}^l$ is a vertical line segment corresponding to initial production costs (c_1, c_2) such that the profit $\pi_{1,M_1}(0,0;c_1,c_2) = \pi_{1,M_1}(v_1,0;c_1,c_2)$ where $v_1 = V_1(0)$ is the best investment response of firm F_1 to a zero investment of firm F_2 (see Fig. 22.5). In Lemma 22.1, we give the algebraic characterization of $U_{S_1}^l = \{c_1^M\} \times [l_{S_1}^M(c_1^M), 10]$ by determining the value c_1^M. The value $l_{S_1}^M(c_1^M)$ such that $(c_1^M, l_{S_1}^M(c_1^M)) \in l_{M_1}$ is computed using Lemma 1 in [7]. Let

- $K_1 = -(8\beta\lambda - \epsilon^2(c_1 - cL)^2 - 2\epsilon(\alpha - c_1)(c_1 - c_L))/(8\beta)$.
- $K_2 = (4\beta\lambda^2 - 2\epsilon\lambda(\alpha - c_1)(c_1 - c_L))/(64\beta)$.
- $K_3 = -(4\beta\lambda^2 - 2\epsilon\lambda(\alpha - c_1)(c_1 - c_L))/(4\beta)$.

Lemma 22.1. *The initial production costs $c_1 = c_1^M$ of firm F_1, such that $(c_1^M, c_2) \in U_{S_1}^l$ and the best investment response $v_1 = V_1(0)$ of firm F_1 to a zero investment of firm F_2 are implicitly determined as solutions of the following polynomial equations:*

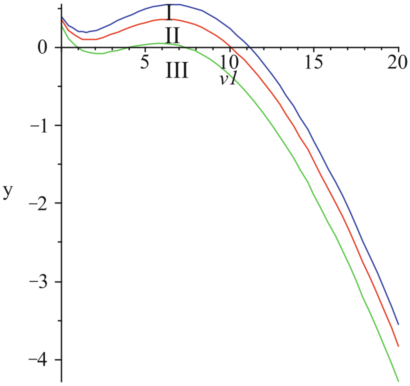


Fig. 22.5 Each of the plots corresponds to the profit π_1 of Firm F_1 when Firm F_2 decides not to invest, i.e. $\pi_1(v_1, 0; c_1, c_2)$. The plot in red (II) corresponds to a pair of production costs $(c_1, c_2) \in U_{S_1}^l$, the plot in blue (I) corresponds to a pair of production costs (c_1, c_2) that are in the single monopoly region S_1^M and the plot in green (III) corresponds to a pair of production costs (c_1, c_2) that are in the nil region N_{HH}

$$2\beta v_1^3 + 6\beta\lambda v_1^2 + L_1 v_1 + N_1 = 0 \tag{22.2}$$

$$K_2^2 - K_1^2 + K_3 + 2V_1 K_1 - V_1^2 = 0 \tag{22.3}$$

Proof. By Theorem 22.1, $\partial\pi_{1,M_1}(v_1, 0; c_1, c_2)/\partial v_1 = 0$ can be written as equality (22.2). From $\pi_{1,M_1}(0, 0; c_1, c_2) = \pi_{1,M_1}(v_1, 0; c_1, c_2)$, we get

$$(\alpha - c_1)^2 = \left(\alpha - c_1 + \frac{\epsilon(c_1 - c_L)v_1}{\lambda + v_1}\right)^2 - 4\beta v_1$$

that leads to

$$4\beta v_1^2 + (8\beta\lambda - \epsilon^2(c_1 - cL)^2 - 2\epsilon(\alpha - c_1)(c_1 - c_L))v_1 \\ + (4\beta\lambda^2 - 2\epsilon\lambda(\alpha - c_1)(c_1 - c_L)) = 0.$$

Choosing the positive solution of the above equality, we get

$$v_1 = K_1 + \sqrt{K_2^2 + K_3}$$

that is equivalent to equality (22.3). By Theorem 22.1, $\partial\pi_{1,M_1}(v_1, 0; c_1^M(c_2), c_2)/\partial v_1 = 0$ can be written as equality (22.2). □

Let us characterize the boundary $U^c_{S_1}$ between the single monopoly region S_1^M and the nil Nash investment region N_{HH} with initial production costs (c_1, c_2) in the Monopoly region M_1. The boundary $U^c_{S_1}$ is a curve corresponding to initial production costs (c_1, c_2) such that the profit $\pi_{1,M_1}(0, 0; c_1, c_2) = \pi_{1,M_1}(v_1, 0; c_1, c_2)$ where $v_1 = V_1(0)$ is the best investment response of firm F_1 to a zero investment of firm F_2 (see Fig. 22.6). In Lemma 22.4 we give the algebraic characterization of the curve $U^c_{S_1} = \{c_1(c_2) : c_2 \in [B(U^C_{S_1}; I^M_{S_1}), l^M_{S_1}(c_1^M)]\}$. The value $l^M_{S_1}(c_1^M)$ is such that $(c_1^M, l^M_{S_1}(c_1^M)) \in l_{M_1}$ is computed, as before, using Lemma 1 in [7]. Let $B(U^C_{S_1}; I^M_{S_1})$ be the common boundary $U^C_{S_1} \cap I^M_{S_1}$ between the boundaries of the single monopoly region $U^C_{S_1}$ and $I^M_{S_1}$. The point $B(U^C_{S_1}; I^M_{S_1})$ is determined as a solution of the polynomial equations presented in Lemmas 22.1 and 22.2. Let

- $K_4 = (\beta(2\beta(\alpha - c_1) - \gamma(\alpha - c_2))^2)/(4(\beta^2 - \gamma^2))$.
- $K_5 = 4\beta K - (\alpha - c_1)^2 + 8\beta K - \epsilon^2(c_1 - c_L)^2 - 2\epsilon(\alpha - c_1)(c_1 - c_L)$.
- $K_6 = 8\beta\lambda K - 2\lambda(\alpha - c_1)^2 + 8\beta\lambda^2 - 2\epsilon\lambda(\alpha - c_1)(c_1 - c_L)$.

Lemma 22.2. *The initial production costs $c_1 = c_1^M(c_2)$ of firm F_1 such that $(c_1^M(c_2), c_2) \in U^c_{S_1}$, and the best investment response $v_1 = V_1(0)$ of firm F_1 to a zero investment of firm F_2 are implicitly determined as solutions of the following polynomial equations:*

$$2\beta v_1^3 + 6\beta\lambda v_1^2 + L_1 v_1 + N_1 = 0 \tag{22.4}$$

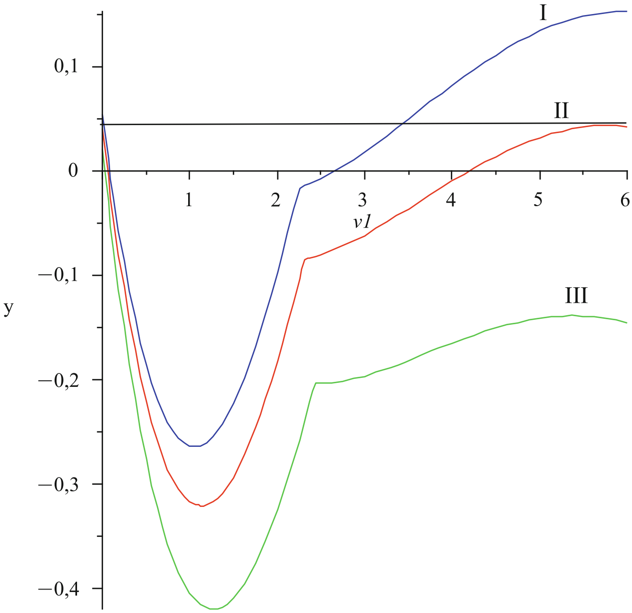


Fig. 22.6 Each of the plots corresponds to the profit π_1 of Firm F_1 when Firm F_2 decides not to invest, i.e. $\pi_1(v_1, 0; c_1, c_2)$. The plot in red (II) corresponds to a pair of production costs $(c_1, c_2) \in U^c_{S_1}$, the plot in blue (I) corresponds to a pair of production costs (c_1, c_2) that are in the single monopoly region S^M_1 and the plot in green (III) corresponds to a pair of production costs (c_1, c_2) that are in the nil region N_{HH}

and

$$4\beta v_1^3 + K_5 v_1^2 + K_6 v_1 + 4\beta K - (\alpha - c_1)^2 = 0 \tag{22.5}$$

Proof. From $\pi_{1,D}(0, 0; c_1, c_2) = \pi_{1,M_1}(v_1, 0; c_1, c_2)$ we get

$$\frac{\beta^2(2\beta(\alpha - c_1) - \gamma(\alpha - c_2))^2}{(\beta^2 - \gamma^2)} = \left(\alpha - c_1 + \frac{\epsilon(c_1 - c_L)v_1}{\lambda + v_1}\right)^2 - 4\beta v_1$$

that leads to (22.5). By Theorem 22.1, $\partial\pi_{1,M_1}(v_1, 0; c_1^M(c_2), c_2)/\partial v_1 = 0$ can be written as equality (22.4) □

Let us characterize the boundary $I^M_{S_1}$ between the single monopoly region S^M_1 and the single Nash investment region S^M_2 with initial production costs (c_1, c_2) in the Monopoly region M_1 (see Fig. 22.7). The intermediate boundary $I^M_{S_1}$ of the single monopoly region S^M_1 is characterized by the best investment response $V_2(V_1(0))$ of firm F_2 to the best investment response $V_1(0)$ of firm F_1 to zero, to be a set with two elements. One of the elements V_2^- of $V_2(V_1(0))$ is zero and the other element V_2^+ is greater than zero. In Lemma 22.3 we give the algebraic characterization of the

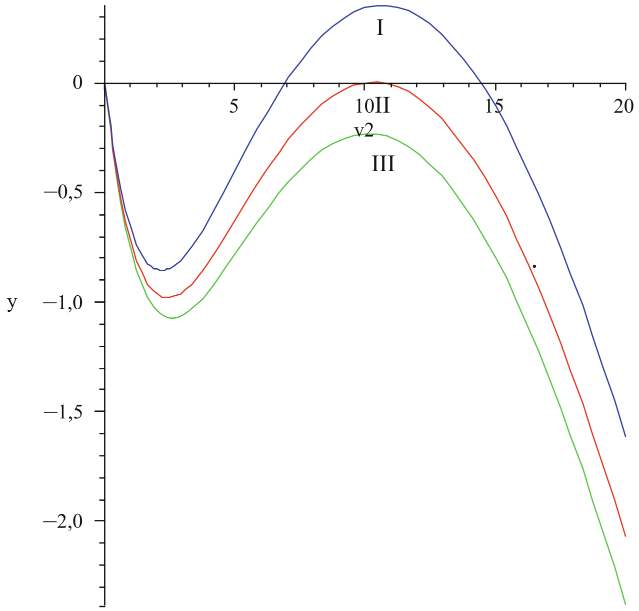


Fig. 22.7 Each of the plots corresponds to the Profit π_2 of Firm F_2 when Firm F_1 decides to invest v_1 and Firm F_2 has two possible best responses $V_2(v_1) = \{v_2; 0\}$ with $v_2 > 0$, i.e. $\pi_2(V_1(0), v_2; c_1, c_2)$. The plot in red (II) corresponds to a pair of production costs $(c_1, c_2) \in I^M_{S_1}$, the plot in blue (I) corresponds to a pair of production costs (c_1, c_2) that are in the single monopoly region S^M_1 and the plot in green (III) corresponds to a pair of production costs (c_1, c_2) that are in the single monopoly region S^M_2

curve $I^M_{S_1} = \{c_1(c_2) : c_2 \in [B(I^M_{S_1}; L^M_{S_1}), B(U^c_{S_1}; I^M_{S_1})]\}$. The point $B(I^M_{S_1}; L^M_{S_1})$ is implicitly determined as a solution of the polynomial equations presented in Lemmas 22.3 and 22.2. The point $B(U^C_{S_1}; I^M_{S_1})$ is determined, as before, as a solution of the polynomial equations presented in Lemmas 22.1 and 22.2.

Let L_1 and N_1 be as in Theorem 22.1. Let C and A_2, B_2 and H_2 be as in Theorem 22.2. Let

- $K_7 = -4\beta\gamma(c_1 - \epsilon(c_1 - c_L))(c_2 - \epsilon(c_2 - c_L))$.
- $K_8 = -4\beta\gamma\epsilon\lambda(c_2 - c_L)(c_1 - \epsilon(c_1 - c_L))$.
- $K_9 = -4\beta\gamma\epsilon\lambda(c_1 - c_L)(c_2 - \epsilon(c_2 - c_L))$.
- $K_{10} = -4\beta\gamma\epsilon^2\lambda^2(c_1 - c_L)(c_2 - c_L)$.
- $K_{11} = 4\beta^2 c_1^2 + \epsilon^2(c_1 - c_L) - 2\epsilon c_1(c_1 - c_L) + \gamma c_2^2 + \epsilon^2(c_2 - c_L) - 2\epsilon c_2(c_2 - c_L) + c_1(8\beta^2\alpha + 4\beta\alpha\gamma) - \epsilon(c_1 - c_L)(8\beta^2\alpha + 4\beta\alpha\gamma) + c_2(-2\alpha\gamma^2 + 4\beta\gamma\alpha) - \epsilon(c_2 - c_L)(-2\alpha\gamma^2 + 4\beta\gamma\alpha) + 4\beta^2\alpha^2 + \gamma^2\alpha^2 - 4\beta\gamma\alpha^2 + (\lambda(4\beta^2 - \gamma^2)^2)/\beta$.
- $K_{12} = -2\lambda\epsilon^2(c_1 - c_L) + 2\epsilon\lambda c_1(c_1 - c_L) + \lambda\epsilon(c_1 - c_L)$.
- $W_1 = v_1 + \lambda$; $W_2 = v_2 + \lambda$.

Lemma 22.3. *The initial production costs $c_1 = c^M_1(c_2)$ of firm F_1 such that $(c^M_1(c_2), c_2) \in I^M_{S_1}$, the best investment $v_1 = V_1(0)$ of firm F_1 to a zero investment* *of firm F_2 and the best investment of firm F_2 $V_2^+ \in V_2(V_1(0))$ are implicitly determined as solutions of the following polynomial equations:*

$$\begin{aligned} K_7 W_1^4 W_2^4 &+ K_8 W_1^4 W_2^3 + K_9 W_1^3 W_2^4 + K_{10} W_1^3 W_2^3 - ((4\beta^2-\gamma^2)^2/\beta) W_2^3 W_1^2 + \\ &+ K_{11} W_1^2 W_2^2 + K_{12} W_1 W_2^2 + K_{13} W_2 W_1^2 + \\ &+ \lambda^2 \epsilon^2 (c_2 - c_L) W_1^2 + \lambda^2 \epsilon^2 (c_1 - c_L) W_2^2 = 0 \end{aligned} \tag{22.6}$$

and

$$C W_2^3 W_1 + A_2 W_2 W_1 + B_2 W_1 - (B_2/\lambda) H_2 W_2 = 0 \tag{22.7}$$

and

$$2\beta(W_1-\lambda)^3 + 6\beta\lambda(W_1-\lambda)^2 + L_1(W_1-\lambda) + N_1 = 0. \tag{22.8}$$

Proof. From $\pi_{2,D}(v_1, v_2; c_1, c_2) = 0$, we get

$$\frac{\beta(2\beta(\alpha-a_1) - \gamma(\alpha-a_2))^2}{(4\beta^2-\gamma^2)^2} - v_2 = 0.$$

The equality above can be written as

$$\begin{aligned} 4\beta^2 a_1^2 &+ \gamma a_2^2 + (-8\beta^2\alpha + 4\beta\alpha\gamma) a_1 + (-2\alpha\gamma^2 + 4\beta\gamma\alpha) a_2 - 4\beta\gamma a_1 a_2 + \\ &+ (4\beta^2\alpha^2 + \gamma^2\alpha^2 - 4\beta\gamma\alpha^2) - ((4\beta^2-\gamma^2)^2/\beta) v_2 = 0. \end{aligned}$$

Substituting $a_i = c_i - (\eta_i v_i)/(\lambda + v_i)$ and manipulating algebraically, we get equality (22.6). By Theorem 22.2, we have that $\partial\pi_{2,D}(v_1, v_2; c_1^M(c_2), c_2)/\partial v_2 = 0$ can be written as equality (22.7). By Theorem 22.1 $\partial\pi_{1,M_1}(v_1, 0; c_1^M(c_2), c_2)/\partial v_1 = 0$ can be written as equality (22.8). □

Let us characterize the boundary $L^M_{S_1}$ between the single monopoly region S_1^M and the single monopoly boundary region S_1^B with initial production costs (c_1, c_2) in the Monopoly region M_1. In Lemma 22.4 we give the algebraic characterization of the curve $L^M_{S_1} = \{c_1(c_2) : c_2 \in [B(L^M_{S_1}; Le^M_{S_1}), B(I^c_{S_1}; L^M_{S_1})]\}$. The point $B(L^M_{S_1}; Le^M_{S_1})$ is implicitly determined as a solution of the polynomial equations presented in Lemma 22.4 and Theorem 22.6. The point $B(I^M_{S_1}; L^M_{S_1})$ is determined, as before, as a solution of the polynomial equations presented in Lemmas 22.3 and 22.4. Let L_1 and N_1 be as in Theorem 22.1.

Lemma 22.4. *The initial production costs $c_1 = c_1^M(c_2)$ of firm F_1 such that $(c_1^M(c_2), c_2) \in L^M_{S_1}$, and the best investment $v_1 = V_1(0)$ of firm F_1 to a zero investment of firm F_2 are implicitly determined as solutions of the following polynomial equations:*

$$2\beta v_1^3 + 6\beta\lambda v_1^2 + L_1 v_1 + N_1 = 0, \tag{22.9}$$

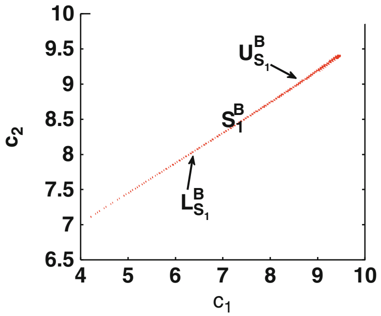


Fig. 22.8 Full characterization of the boundaries of the single monopoly boundary region S_1^B: the upper boundary $U^B_{S_1}$ and the lower boundary $L^B_{S_1}$

where

$$v_1 = \frac{\gamma\lambda(c_2-\alpha) - 2\beta\lambda(c_1-\alpha)}{2\epsilon\beta(c_L-c_1) + 2\beta(c_1-\alpha) - \gamma(c_2-\alpha)} \tag{22.10}$$

Proof. By Theorem 22.1, $\partial\pi_{1,M_1}(v_1, 0; c_1, c_2)/\partial v_1 = 0$ can be written as (22.9). Take $a_1 = c_1 - (\epsilon(c_1 - c_L)v_1)/(\lambda + v_1)$ and $a_2 = c_2$, by Lemma 1 in [7], we get

$$\left(\frac{\gamma}{2\beta}(c_2\alpha) - (c_1-\alpha)\right)(\lambda + v_1) = \epsilon(c_L - c_1)v_1 \tag{22.11}$$

Thus (22.10) follows from (22.11). □

22.4.2 Boundary of the Single Monopoly Boundary Region S_1^B

The upper boundary of the single monopoly boundary region $U^B_{S_1}$ is the lower boundary of the single monopoly region $L^M_{S_1}$ and has already been characterized in Sect. 22.4.1. Let us characterize the boundary $L^B_{S_1}$ between the single monopoly boundary region S_1^B and the single duopoly region S_1^D for initial production costs (c_1, c_2) in the monopoly region M_1. In Lemma 22.5 we give the algebraic characterization of the curve $L^B_{S_1} = \{c_1(c_2) : c_2 \in [B(L^D_{S_1}; L^B_{S_1}), B(L^B_{S_1}; d_3)]\}$. The point $B(L^D_{S_1}; L^B_{S_1})$ is implicitly determined as a solution of the polynomial equations presented in Lemmas 22.5 and 22.6. The point $B(L^B_{S_1}; d_3)$ is implicitly determined as a solution of the polynomial equations presented Lemma 22.5 and Theorem 22.5. Let A_1, E_1, F_1, G_1 and H_1 be as in Theorem 22.5.

Lemma 22.5. *The initial production costs $c_1 = c_1^M(c_2)$ of firm F_1 such that $(c_1^M(c_2), c_2) \in L^B_{S_1}$ and the best investment $v_1 = V_1(0)$ of firm F_1 to a zero investment of firm F_2 are implicitly determined as solutions of the following polynomial equations:*

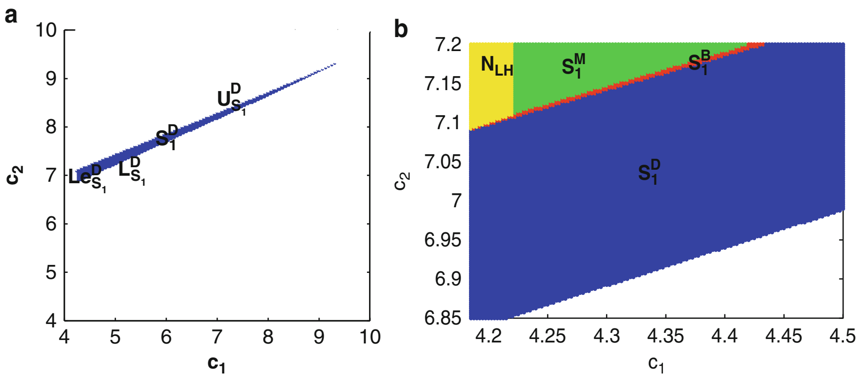


Fig. 22.9 (**a**) Full characterization of the boundaries of the single duopoly region S_1^D: the upper boundary $U^D_{S_1}$; the lower boundary $L^D_{S_1}$; and the left boundary $Le^D_{S_1}$; (**b**) close up of the lower part of $Le^D_{S_1}$

$$A_1c_1^2 + E_1c_1c_2 + F_1c_1 + G_1c_2 + H_1 = 0 \tag{22.12}$$

where

$$v_1 = \frac{\gamma\lambda(c_2 - \alpha) - 2\beta\lambda(c_1 - \alpha)}{2\epsilon\beta(c_L - c_1) + 2\beta(c_1 - \alpha) - \gamma(c_2 - \alpha)} \tag{22.13}$$

Proof. By Theorem 22.5, $\partial\pi_{1,D}(v_1, 0; c_1, c_2)/\partial v_2 = 0$ can be written as (22.12). We get (22.13) as in Lemma 22.4. □

22.4.3 Boundary of the Single Duopoly Region S_1^D

The upper boundary of the single duopoly region $U^D_{S_1}$ is the lower boundary of the single monopoly boundary region $L^B_{S_1}$ and has already been characterized in Sect. 22.4.2. The left boundary of the single duopoly region $Le^D_{S_1}$ is the right boundary d_3 of the nil Nash investment region N_{LH} that will be characterized in Sect. 22.5.

Let us characterize the boundary $L^D_{S_1}$ between the single duopoly region S_1^D and the competitive region C for initial production costs (c_1, c_2) in the monopoly region M_1 (see Fig. 22.10). In Lemma 22.6 we give the algebraic characterization of the curve $L^D_{S_1} = \{c_1(c_2) : c_2 \in [B(L^B_{S_1}; L^D_{S_1}), B(L^D_{S_1}; d_3)]\}$. The point $B(L^B_{S_1}; L^D_{S_1})$ is implicitly determined as a solution of the polynomial equations presented in Lemmas 22.5 and 22.6. The point $B(L^B_{S_1}; d_3)$ is implicitly determined as a solution of the polynomial equations presented in Lemma 22.6 and Theorem 22.5.

Lemma 22.6. *The initial production costs $c_1 = c_1^M(c_2)$ of firm F_1 such that $(c_1^M(c_2), c_2) \in L^D_{S_1}$, the best investment $v_1 = V_1(0)$ of firm F_1 to a zero investment*

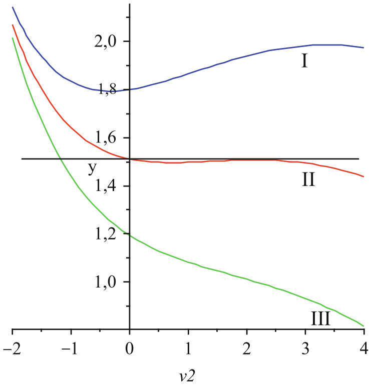


Fig. 22.10 Each of the plots corresponds to the Profit π_2 of Firm F_2 when Firm F_1 decides not to invest. The plot in red (II) corresponds to a pair of production costs $(c_1, c_2) \in L^D_{S_1}$, the plot in blue (I) corresponds to a pair of production costs (c_1, c_2) that are in the competitive region C and the plot in green (III) corresponds to a pair of production costs (c_1, c_2) that are in the single duopoly region S^D_2

of firm F_2 and the best investment of firm F_2 $V_2^+ \in V_2(V_1(0))$ are implicitly determined as solutions of the following polynomial equations:

$$K_7 W_1^4 W_2^4 + K_8 W_1^4 W_2^3 + K_9 W_1^3 W_2^4 + K_{10} W_1^3 W_2^3 - ((4\beta^2 - \gamma^2)^2/\beta) W_2^3 W_1^2 + \\ + K_{11} W_1^2 W_2^2 + K_{12} W_1 W_2^2 + K_{13} W_2 W_1^2 + \\ + \lambda^2 \epsilon^2 (c_2 - c_L) W_1^2 + \lambda^2 \epsilon^2 (c_1 - c_L) W_2^2 = 0 \tag{22.14}$$

and

$$C W_2^3 W_1 + A_2 W_2 W_1 + B_2 W_1 - (B_2/\lambda) H_2 W_2 = 0 \tag{22.15}$$

and

$$C W_1^3 W_2 + A_1 W_1 W_2 + B_1 W_2 - (B_1/\lambda) H_1 W_1 = 0 \tag{22.16}$$

Proof. We get (22.14) as in Lemma 22.3.
By Theorem 22.2, $\partial \pi_{2,D}(v_1, v_2; c_1, c_2)/\partial v_2 = 0$ can be written as (22.15). By Theorem 22.2, $\partial \pi_{1,D}(v_1, v_2; c_1, c_2)/\partial v_1 = 0$ can be written as equality (22.16). □

22.4.4 Boundary of the Single Recovery Region S_1^R

The single recovery region S_1^R (which, due to the symmetry, holds a similar characterization for S_2^R) has three boundaries: the *upper boundary* $U^R_{S_1}$, the *left boundary*

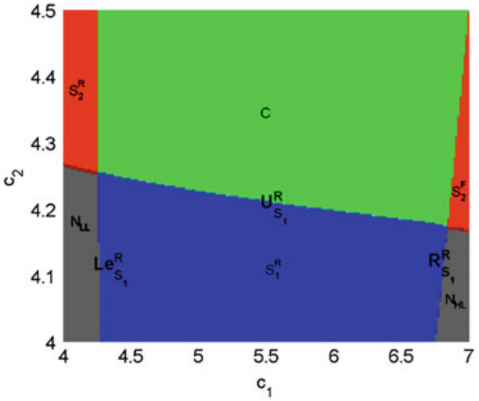


Fig. 22.11 Full characterization of the boundaries of the single recovery region S_1^R: the upper boundary $U^R_{S_1}$; the right boundary $R^R_{S_1}$; and the left boundary $Le^R_{S_1}$. In green the competitive Nash investment region C, in grey the nil Nash investment region N, in red the single Nash investment region S_2 for firm F_2 and in blue the single recovery region S_1^R for firm F_1

$L^R_{S_1}$, and the *right boundary* $R^R_{S_1}$. We are now going to characterize the upper boundary $U^R_{S_1}$ of the single recovery region S_1^R and will leave the left and right boundaries of the single recovery region, that are also boundaries of the Nil Nash investment region, to be characterized in Sect. 22.5 (see Fig. 22.12). In Lemma 22.7 we give the algebraic characterization of the curve $U^R_{S_1} = \{c_1(c_2) : c_2 \in [Q; P3)]\}$ where the point Q is characterized by being in the intersection between the competitive region C and the nil region N_{LL} and the point P_3 is characterized by being in the intersection between the competitive region C and the nil region N_{HL}.

Lemma 22.7. *The initial production costs $c_1 = c_1^R(c_2)$ of firm F_1 such that $(c_1^R(c_2), c_2) \in U^R_{S_1}$ are implicitly determined as solutions of the following polynomial equations:*

$$A_2c_2^2 + E_2c_1c_2 + F_2c_2 + G_2c_1 + H_2 = 0 \qquad (22.17)$$

and

$$A_1c_1^2 + E_1c_1c_2 + F_1c_1 + G_1c_2 + H_1 = 0 \qquad (22.18)$$

Proof. By Theorem 22.5, we get (22.17) and (22.18). □

22.5 Nil Nash Investment Region

The *nil Nash investment region* N is the set of production costs $(c_1, c_2) \in N$ with the property that $(0, 0)$ is a Nash investment equilibrium. Hence, the nil Nash investment region N consists of all production costs (c_1, c_2) with the property that the new

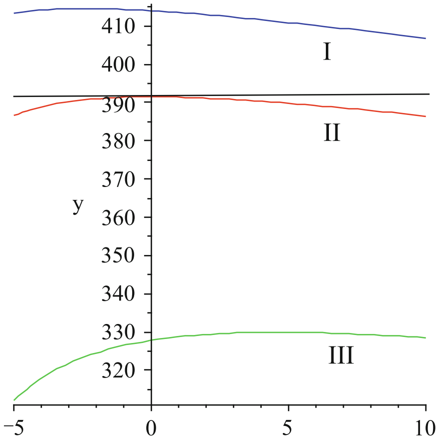


Fig. 22.12 Each of the plots corresponds to the profit π_2 of Firm F_2 when Firm F_1 decides not to invest, i.e. $\pi_{2,D}(V_1(0), v_2; c_1, c_2)$. The plot in red (II) corresponds to a pair of production costs $(c_1, c_2) \in U_{S_1}^R$, the plot in blue (I) corresponds to a pair of production costs (c_1, c_2) that are in the nil region N_{HL} and the plot in green (III) corresponds to a pair of production costs (c_1, c_2) that are in the single recovery region S_1^R

production costs $(a_1(v_1), a_2(v_2))$, with respect to the Nash investment equilibrium $(0, 0)$, are equal to the production costs (c_1, c_2).

The nil Nash investment region N is the union of four disjoint sets: the set N_{LL} consisting of all production costs that are low for both firms (see Fig. 22.13a); the set N_{LH} (respectively N_{HL}) consisting of all production costs that are low for firm F_1 (respectively F_2) and high for firm F_2 (respectively F_1) (see Fig. 22.13b); and the set N_{HH} consisting of all production costs that are high for both firms (see Fig. 22.13c).

The set N_{LH} (respectively N_{HL}) is the union of the sets N_{LH}^M (respectively N_{HL}^M) and N_{LH}^D (respectively N_{HL}^D). The set N_{LH}^M (respectively N_{HL}^M) consists of all production costs in the region N_{LH} (respectively N_{HL}) such that firm F_1 (respectively firm F_2) is in monopoly or equivalently, firm F_2 is out of the market. The set N_{LH}^D (respectively N_{HL}^D) consists of all production costs in the region N_{LH} (respectively N_{HL}) such that both firms have positive outputs, i.e. both firms are in the duopoly region D (see Fig. 22.1).

In this section, we characterize the boundaries of these Nil Nash investment regions. The boundaries of the Nash investment region N_{HH} have been characterized in the previous section. The left boundary $Le_{N_{HH}}$ of the nil Nash investment region N_{HH} coincides with the upper boundary of the single monopoly region $U_{S_1}^M$ (see Lemmas 22.1 and 22.2) and the lower boundary $L_{N_{HH}}$ of the nil Nash investment region N_{HH} coincides with the upper boundary of the single monopoly region $U_{S_2}^M$. To characterize all the other boundaries of the nil regions, we will use the following theorems:

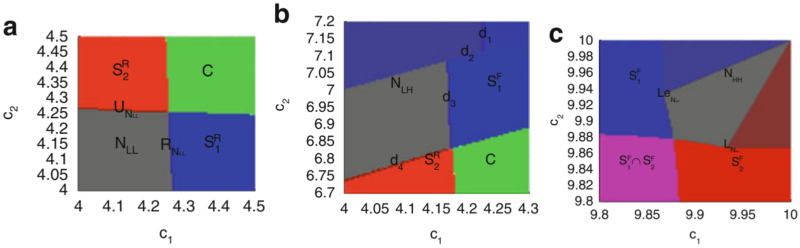


Fig. 22.13 Full characterization of the nil Nash investment region N in terms of the firms' initial production costs (c_1, c_2): (**a**) The subregion N_{LL} of the nil Nash investment region N is colored grey corresponding to initial production cost such that the firms do not invest and do not produce; (**b**) The subregion N_{LH} of the nil Nash investment region N is colored grey corresponding to initial production cost such that the firms do not invest and do not produce and dark blue corresponding to cases where the firms do not invest but firm F_1 produces a certain amount q_1 greater than zero; (**c**) The subregion N_{HH} of the nil Nash investment region N is colored grey corresponding to initial production cost such that the firms do not invest and do not produce; dark blue corresponding to cases where the firms do not invest but firm F_1 produces a certain amount q_1 greater than zero and dark red corresponding to cases where the firms do not invest but firm F_2 produces a certain amount q_2 greater than zero

Let us define the following parameters

- $I_i = 4\beta^2/\lambda$; $A_i = -2I_i\epsilon\beta$; $E_i = I_i\epsilon\gamma$.
- $G_i = -I_i\epsilon\gamma c_L$; $F_i = 2I_i\epsilon\beta\alpha + 2I_i\epsilon c_L\beta - I_i\epsilon\gamma\alpha$.
- $K = (4\beta^2 - \gamma^2)^2$; $H_i = -2I_i\epsilon c_L\beta\alpha + I_i\epsilon c_L\gamma\alpha - K$.

Theorem 22.5. *The solutions of* $\partial\pi_{i,D}(0, 0; c_1, c_2)/\partial v_i = 0$ *are contained in*

$$A_i c_i^2 + E_i c_i c_j + F_i c_i + G_i c_j + H_i = 0.$$

Proof. Let us compute

$$\frac{d\pi_{i,D}}{dv_i} = \frac{\partial\pi_{i,D}}{\partial a_i}\frac{\partial a_i}{\partial v_i} + \frac{\partial\pi_{i,D}}{\partial a_j}\frac{\partial a_j}{\partial v_i} + \frac{\partial\pi_{i,D}}{\partial v_i}. \tag{22.19}$$

We have that

$$\begin{aligned}
\frac{\partial\pi_{i,D}}{\partial a_i} &= -\frac{4\beta^2(2\beta(\alpha - a_i) + \gamma(a_j - \alpha))}{(4\beta^2 - \gamma^2)^2} \\
\frac{\partial a_i}{\partial v_i} &= \frac{\eta_i\lambda}{(\lambda + v_i)^2} \\
\frac{\partial\pi_{i,D}}{\partial a_j} &= -\frac{2\beta_i\gamma(2\beta_j(\alpha_i - a_i) + \gamma(a_j - \alpha_j))}{(4\beta_i\beta_j - \gamma^2)^2} \\
\frac{\partial\pi_{i,D}}{\partial v_i} &= -1.
\end{aligned}$$

Hence, $d\pi_{i,D}/dv_i = 0$ if, and only if,

$$\frac{4\beta^2\eta_i\lambda(2\beta(\alpha - a_i) + \gamma(a_j - \alpha))}{\lambda^2} = K \tag{22.20}$$

Taking $a_i = c_i$ and $a_j = c_j$, we get that $d\pi_{i,D}/dv_i = 0$ if, and only if,

$$I_i\eta_i(2\beta(\alpha - c_i) + \gamma(c_j - \alpha)) - K = 0$$

After algebric manipulations, we get

$$2I_i\eta_i\beta\alpha - 2I_i\eta_i\beta c_i + I_i\eta_i\gamma c_j - I_i\eta_i\gamma\alpha - K = 0$$

which leads to

$$A_ic_i^2 + E_ic_ic_j + F_ic_i + G_ic_j + H_i = 0.$$

□

Let $Q = \epsilon(\alpha + c_L)$ and $R = -\epsilon\alpha c_L - 2\beta\lambda$.

Theorem 22.6. *The solution of* $\partial\pi_{i,M_i}(0, 0; c_1, c_2)/\partial v_i = 0$, *is contained in*

$$c_i = (-Q + \sqrt{Q^2 - 4PR})/(-2\epsilon). \tag{22.21}$$

Proof. Let us compute

$$\frac{d\pi_{i,M_i}}{dv_i} = \frac{\partial\pi_{i,M_i}}{\partial a_i}\frac{\partial a_i}{\partial v_i} + \frac{\partial\pi_{i,M_i}}{\partial v_i}.$$

Since

$$\partial\pi_{i,M_i}(v_i, 0; c_1, c_2)/\partial v_i = (\epsilon\lambda(\alpha - a_i)(c_i - c_L)) \,/\, \left(2\beta(\lambda + v_i)^2\right) - 1,$$

$d\pi_{i,M_i}(v_i, 0; c_1, c_2)/dv_i = 0$ if, and only if,

$$\epsilon\lambda(\alpha - a_i)(c_i - c_L) = 2\beta(\lambda + v_i)^2.$$

Letting $v_i = 0$ ($a_i = c_i$), we get

$$\epsilon\lambda(\alpha - c_i)(c_i - c_L) = 2\beta\lambda^2$$

that can be written as

$$-\epsilon\lambda c_i^2 + \epsilon\lambda(\alpha + c_L)c_i - \epsilon\lambda\alpha c_L - 2\beta\lambda^2 = 0.$$

We choose

$$c_i = (-Q + \sqrt{Q^2 - 4PR})/(-2\epsilon).$$

□

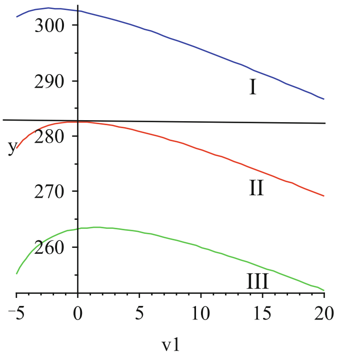


Fig. 22.14 Each of the plots corresponds to the profit π_1 of Firm F_1 when Firm F_2 decides not to invest, i.e. $\pi_{1,D}(V_1(0), 0; c_1, c_2)$. The plot in red (II) corresponds to a pair of production costs $(c_1, c_2) \in R_{N_{LL}}$, the plot in blue (I) corresponds to a pair of production costs (c_1, c_2) that are in the nil region N_{LL} and the plot in green (III) corresponds to a pair of production costs (c_1, c_2) that are in the single recovery region S_1^R

We begin by characterizing the boundary of the Nil Nash investment region N_{LL} that is composed by a *right boundary* $R_{N_{LL}}$ and a *upper boundary* $U_{N_{LL}}$. The right boundary of the Nil Nash investment region N_{LL} (see Fig. 22.13a) is given by the curve (see Theorem 22.5 and Fig. 22.14)

$$\frac{\partial \pi_{1,D}}{\partial v_1}(0, 0; c_1, c_2) = 0.$$

Furthermore, the upper boundary of the region N_{LL} is given by the curve (see Theorem 22.5 and Fig. 22.15)

$$\frac{\partial \pi_{2,D}}{\partial v_2}(0, 0; c_1, c_2) = 0.$$

We will refer to the boundaries of the region N_{LH}^M as d_1 and d_4 (see Fig. 22.13b). The arc d_1 is given by the curve (see Theorem 22.6 and Fig. 22.16)

$$\frac{\partial \pi_{1,M_1}}{\partial v_1}(0, 0; c_1, c_2) = 0.$$

The arc d_2 is a line segment l_{M_1} characterized in Appendix 1. The boundaries of the region N_{LH}^D are d_2, d_3 and d_4. The arc d_2 is described above. The arc d_3 is given by the curve (see Theorem 22.5 and Fig. 22.17)

$$\frac{\partial \pi_{1,D}}{\partial v_1}(0, 0; c_1, c_2) = 0,$$

and the arc d_4 is given by the curve (see Theorem 22.5 and Fig. 22.18)

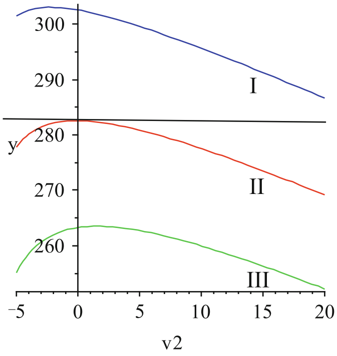


Fig. 22.15 Each of the plots corresponds to the profit π_2 of Firm F_2 when Firm F_1 decides not to invest, i.e. $\pi_{2,M_2}(0, V_2(0); c_1, c_2)$. The plot in red (II) corresponds to a pair of production costs $(c_1, c_2) \in U_{N_{LL}}$, the plot in blue (I) corresponds to a pair of production costs (c_1, c_2) that are in the nil region N_{LL} and the plot in green (III) corresponds to a pair of production costs (c_1, c_2) that are in the single recovery region S_2^R

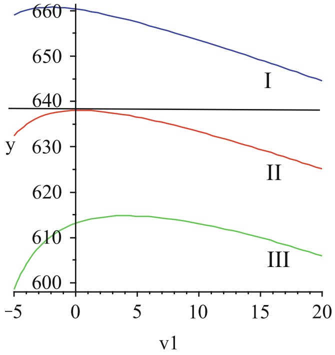


Fig. 22.16 Each of the plots corresponds to the profit π_1 of Firm F_1 when Firm F_2 decides not to invest, i.e. $\pi_1(V_1(0), 0; c_1, c_2)$. The plot in red (II) corresponds to a pair of production costs $(c_1, c_2) \in d_1$, the plot in blue (I) corresponds to a pair of production costs (c_1, c_2) that are in the nil region N_{LH} and the plot in green (III) corresponds to a pair of production costs (c_1, c_2) that are in the single favorable region S_1^F

$$\frac{\partial \pi_{2,D}}{\partial v_2}(0, 0; c_1, c_2) = 0.$$

22.6 Competitive Nash Investment Region

The *competitive Nash investment region* C consists of all production costs (c_1, c_2) with the property that there is a Nash investment equilibrium (v_1, v_2) with the property that $v_1 > 0$ and $v_2 > 0$. Hence, the new production costs $a_1(v_1, v_2)$ and

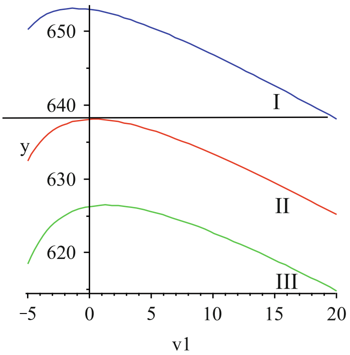


Fig. 22.17 Each of the plots corresponds to the profit π_1 of Firm F_1 when Firm F_2 decides not to invest, i.e. $\pi_{1,D}(V_1(0), 0; c_1, c_2)$. The plot in red (II) corresponds to a pair of production costs $(c_1, c_2) \in d_3$, the plot in blue (I) corresponds to a pair of production costs (c_1, c_2) that are in the nil region N_{LH} and the plot in green (III) corresponds to a pair of production costs (c_1, c_2) that are in the single favorable region S_1^F

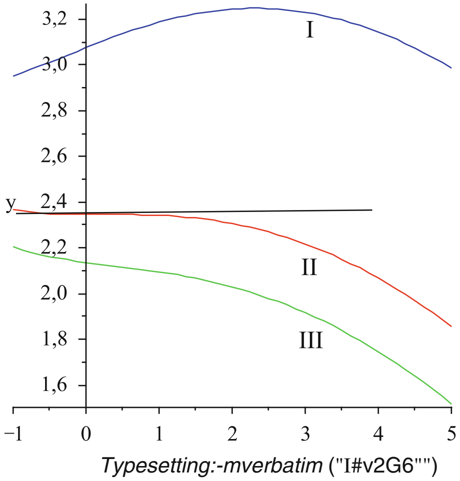


Fig. 22.18 Each of the plots corresponds to the profit π_2 of Firm F_2 when Firm F_1 decides not to invest, i.e. $\pi_{2,D}(0, V_2(0); c_1, c_2)$. The plot in red (II) corresponds to a pair of production costs $(c_1, c_2) \in d_4$, the plot in blue (I) corresponds to a pair of production costs (c_1, c_2) that are in the single recovery region S_2^R and the plot in green (III) corresponds to a pair of production costs (c_1, c_2) that are in the nil region N_{LH}

$a_2(v_1, v_2)$ of firms F_1 and F_2 are smaller than the actual production costs c_1 and c_2 of the firms F_1 and F_2, respectively.

In Fig. 22.2, the boundary of region C consists of four piecewise smooth curves: The curve C_1 is characterized by $a_1(v_1) = c_1$ i.e. $v_1 = 0$; the curve C_2 is characterized by $a_2(v_2) = c_2$ i.e. $v_2 = 0$; the curve C_3 corresponds to points (c_1, c_2) such that the Nash investment equilibrium $(a_1(v_1), a_2(v_2))$ has the property

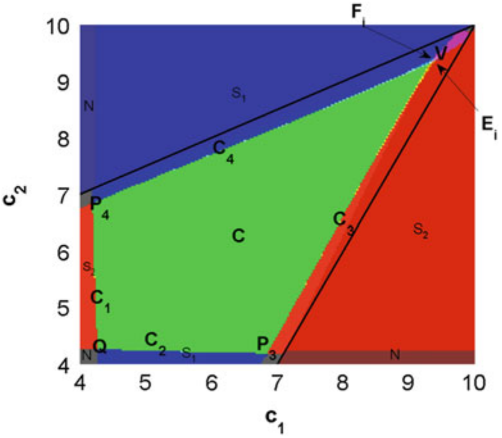


Fig. 22.19 Firms' investments in the competitive Nash investment region. The competitive Nash investment region is colored green, the single Nash investment region S_1 (respectively S_2) is colored blue (respectively red) and the nil Nash investment region N is colored grey

that $\pi_1(a_1, a_2) = \pi_1(a_1, c_2)$; and the curve C_4 corresponds to points (c_1, c_2) such that the Nash investment equilibrium $(a_1(v_1), a_2(v_2))$ has the property that $\pi_1(a_1, a_2) = \pi_1(c_1, a_2)$.

The curve C_2 (respectively C_1) is the common boundary between the competitive region C and the single recovery region S_2^R (respectively S_1^R). The boundary C_3 can be decomposed in three parts C_3^D, C_3^B and C_3^M. The boundary C_3^D consists of all points in C_3 between the points P_3 and E_3 (see Fig. 22.19). The boundary $C_3^D - \{P_3\}$ has the property of being contained in the lower boundary of the single duopoly region S_2^D of firm F_2. The boundary C_3^B consists of all points in C_3 between the points E_3 and F_3 (see Fig. 22.19). The boundary C_3^B has the property of being contained in the lower boundary of the single monopoly boundary region S_2^B of firm F_2. The boundary C_3^M consists of all points in C_3 between the points F_3 and V (see Fig. 22.19). The boundary C_3^M has the property of being contained in the lower boundary of the single monopoly boundary region S_2^B of firm F_2. Due to the symmetry, a similar characterization holds for the boundary C_4. The points P_3, P_4, Q and V are the corners of the competitive region C (see Fig. 22.19). The point Q is characterized by being in the intersection between the competitive region C and the nil Nash region N_{LL}. The point P_3 (respectively P_4) is characterized by being in the intersection between the competitive region C and the nil region N_{HL}^D (respectively N_{LH}^D). The point E_3 in the boundary of the competitive region C is characterized by belonging to the boundaries of the single duopoly region S_2^D and the single monopoly boundary region S_2^B (see Fig. 22.19). The point F_3 in the boundary of the competitive region C is characterized by belonging to the boundaries of the single monopoly boundary region S_2^B and the single monopoly region S_2^M (see Fig. 22.19).

22.7 Conclusions

The following conclusions are valid in some parameter region of our model. We described four main economic regions for the R&D deterministic dynamics corresponding to distinct perfect Nash equilibria: a competitive Nash investment region C where both firms invest, a single Nash investment region for firm F_1, S_1, where just firm F_1 invests, a single Nash investment region for firm F_2, S_2, where just firm F_2 invests, and a nil Nash investment region N where neither of the firms invest.

The nil Nash investment region has four subregions: N_{LL}, N_{LH}, N_{HL} and N_{HH}. The single Nash investment region can be divided into four subregions: the single favorable region for firm F_1, S_1^F, the single recovery region for firm F_1, S_1^R, the single favorable region for firm F_2, S_2^F, the single recovery region for firm F_2, S_2^R. The single favorable region S_1^F (due to the symmetry the same characterization holds for S_2^F) is the union of three disjoint regions: the single duopoly region S_1^D where the production costs, after the investments, belong to the duopoly region D; the single monopoly boundary region S_1^B where the production costs, after the investments, belong to the boundary of the monopoly region l_{M_1}; and the single monopoly region S_1^M where the production costs, after the investments, belong to the monopoly region M_1.

We exhibited regions where the Nash investment equilibrium are not unique: the intersection $R_{S_1 \cap S_2}$ between the single Nash investment region S_1 and the single Nash investment region S_2 is non-empty; the intersection $R_{S_i \cap C}$, between the single Nash investment region S_i and the competitive Nash investment region C is non-empty; the intersection $R_{S_1 \cap C \cap S_2}$ between the single Nash investment region S_1, the single Nash investment region S_2 and the competitive Nash investment region C is non-empty.

Acknowledgements Previous versions of this work were presented in the International Congresses of Mathematicians ICM 2006 and 2010, International Congress for Industrial and Applied Mathematics 2007, the Second Brazilian Workshop of the Game Theory Society in honor of John Nash and EURO 2010, ICDEA 2009, LAMES 2008, SAMES 2006, VII JOLATE and Palestine Conference in Modern Trends in Mathematics and Physics 2008. This work was highlighted in the article [13] after being presented in ICM 2010. We thank LIAAD-INESC Porto LA, Calouste Gulbenkian Foundation, PRODYN-ESF, FEDER. POFC, POCTI and POSI by FCT and Ministério da Ciência e da Tecnologia, and the FCT Pluriannual Funding Program of the LIAAD-INESC Porto LA and of the Research Centre of Mathematics of the University of Minho, for their financial support. Bruno Oliveira gratefully acknowledges financial support from PRODEP III by FSE and EU and Miguel Ferreira gratefully acknowledges financial support from Fundação para a Ciência e a Tecnologia (FCT) given through a PhD scholarship.

References

1. Amir, R., Evstigneev, I., Wooders, J.: Noncooperative versus cooperative R&D with endogenous spillover rates. Core Discussion Paper 2001/50, Louvain-la-Neuve, Belgium
2. Aspremont, C.d'., Jacquemin, A.: Cooperative and noncooperative R&D in duopoly with spillovers. Am. Econ. Rev. **78**, 1133–1137, Erratum. Am. Econ. Rev. **80**, 641–642 (1988)

3. Bischi, G.I., Gallegati, M., Naimzada, A.: Symmetry-breaking bifurcations and representative firm in dynamic duopoly games. Ann. Oper. Res. **89**, 253–272 (1999)
4. Brander, J.A., Spencer, B.J.: Strategic commitment with R&D: the symmetric case. Bell J. Econ. **14**, 225–235 (1983)
5. Cournot, A.: Recherches sur les Principes Mathématiques de la Théorie des Richesses. Paris, 1838. English edition: Researches into the Mathematical Principles of the Theory of Wealth. In: Bacon, N. (ed.). Macmillan, New York (1897)
6. DeBondt, R.: Spillovers and innovative activities. Int. J. Indust. Organ. **15**, 1–28 (1997)
7. Ferreira, M., Oliveira, B., Pinto, A.A.: Patents in new technologies. J. Differ. Equ. Appl. **15**, 1135–1149 (2009)
8. Ferreira, M., Oliveira, B., Pinto, A.A.: Piecewise R&D Dynamics on costs. Fasciculi Mathematici (2009)
9. Ferreira, F.A, Ferreira, F., Ferreira, M., Pinto, A.A.: Quantity competition in a differentiated duopoly. Intelligent Engineering Systems and Computational Cybernetics. Springer, Netherlands (2008)
10. Kamien, M., Muller, E., Zang, I.: Research joint ventures and R&D cartels. Am. Econ. Rev. **82**, 1293–1306 (1992)
11. Kamien, M., Zang, I.: Competing research joint ventures. J. Econ. Manag. Strategy **2**, 23–40 (1993)
12. Katz, M.: An analysis of cooperative research and development. RAND J. Econ. **17**, 527–543 (1986)
13. Mudur, G.S.: Maths for movies, medicine & markets. The Telegraph Calcutta, India, 20/09/2010
14. Pinto, A.A.: Game Theory and Duopoly Models. Interdisciplinary Applied Mathematics. Springer (2011)
15. Pinto, A.A., Oliveira, B., Ferreira, F.A., Ferreira, M.: Investing to survive in a duopoly model. Intelligent Engineering Systems and Computational Cybernetics. In: Tenreiro Machado, J.A., Patkai, B., Rudas, I.J. (eds.) Springer, Netherlands, Chapter 23, 407–414 (2008)
16. Qiu, D.L.: On the dynamic efficiency of Bertrand and Cournot equilibria. J. Econ. Theory **75**, 213–229 (1997)
17. Ruff, L.: Research and technological progress in a Cournot economy. J. Econ. Theory **1**, 397–415 (1997)
18. Singh, N., Vives, X.: Price and quantity competition in a differentiated duopoly. RAND J. Econ. **15**, 546–554 (1984)

Chapter 23
Computation of Genus and Braid Index for Renormalizable Lorenz Links

Nuno Franco and Luís Silva

Abstract We present some recent results concerning the structure of renormalizable Lorenz links. Then we use these results to derive formulae for the computation of knot and link invariants. Finally we analyze the complexity of the algorithms obtained and compare it with the complexity of the algorithms derived from the definitions, obtaining a reduction from exponential complexity, in the classic algorithms, to linear in our algorithms.

23.1 Introduction

Let ϕ_t be a flow on S^3 with countably many periodic orbits $(\tau_n)_{n=1}^{\infty}$. We can look to each closed orbit as a knot in S^3. It was R.F. Williams, in 1976, who first conjectured that non trivial knotting occur in the Lorenz system [12]. In 1983, Birmann and Williams introduced the notion of template, in order to study the knots and links (i.e. finite collections of knots, taking into account the knotting between them) contained in the geometric Lorenz attractor [2].

To study these families of knots and links we consider related families of invariants. In this paper we will be concerned with the genus and the braid index. These invariants are easy to compute using symbolic dynamics. Unfortunately the complexity of these computations increases very fast (sometimes exponentially) with the length of the symbolic sequence considered. Since the length of the sequences related with n-renormalizable maps increases exponentially with n, this makes it impossible to compute these invariants for n-renormalizable maps, even for small values of n, using the classic algorithms (constructed directly from the definitions).

N. Franco (✉)
CIMA-UE and Department of Mathematics, University of Évora, Rua Romão Ramalho,
59, 7000-671 Évora, Portugal
e-mail: nmf@uevora.pt

L. Silva
CIMA-UE and Scientific Area of Mathematics, Instituto Superior de Engenharia de Lisboa,
Rua Conselheiro Emídio Navarro, 1, 1959-007 Lisbon, Portugal
e-mail: lfs@dec.isel.ipl.pt

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_23,

In this paper we present two algorithms based on the formulae presented in [10], that allow us to compute the braid index and the genus for n-renormalizable maps with very large n. These algorithms give us an exceptional tool to compute those invariants, as dramatically reduce the complexity and computing time.

We define a *Lorenz flow* as a semi-flow that has a singularity of saddle type with a one-dimensional unstable manifold and an infinite set of hyperbolic periodic orbits, whose closure contains the saddle point (see [8]). A Lorenz flow, together with an extra geometric assumption (see [13]) is called a *geometric Lorenz flow*. The dynamics of this type of flows can be described by the iteration of one-dimensional first-return maps $f : [a, b] \setminus \{c\} \to [a, b]$ with one discontinuity at $c \in]a, b[$, increasing in the continuity intervals $[a, c[$ and $]c, b]$ and boundary anchored (i.e. $f(a) = a$ and $f(b) = b$), see [8]. These maps are called Lorenz maps and sometimes we denote them by $f = (f_-, f_+)$, where f_- and f_+ correspond, respectively, to the left and right branches.

23.2 Symbolic Dynamics of Lorenz Maps

Symbolic dynamics is a very useful combinatoric tool to study the dynamics of one-dimensional maps.

Let $f^j = f \circ f^{j-1}$, $f^0 = id$, be the jth iterate of the map f. We define the *itinerary* of a point x under a Lorenz map f as $i_f(x) = (i_f(x))_j, j = 0, 1, \ldots,$ where

$$(i_f(x))_j = \begin{cases} L \text{ if } f^j(x) < 0 \\ 0 \text{ if } f^j(x) = 0 \\ R \text{ if } f^j(x) > 0 \end{cases}.$$

It is obvious that the itinerary of a point x will be a finite sequence in the symbols L and R with 0 as its last symbol, if and only if x is a pre-image of 0 and otherwise it is one infinite sequence in the symbols L and R. So we consider the symbolic space Σ of sequences $X_0 \cdots X_n$ on the symbols $\{L, 0, R\}$, such that $X_i \neq 0$ for all $i < n$ and: $n = \infty$ or $X_n = 0$, with the lexicographic order relation induced by $L < 0 < R$.

It is straightforward to verify that, for all $x, y \in [-1, 1]$, we have the following:

1. If $x < y$ then $i_f(x) \leq i_f(y)$.
2. If $i_f(x) < i_f(y)$ then $x < y$.

We define the *kneading invariant* associated to a Lorenz map $f = (f_-, f_+)$, as

$$K_f = (K_f^-, K_f^+) = (L i_f(f_-(0)), R i_f(f_+(0))).$$

We say that a pair $(X, Y) \in \Sigma \times \Sigma$ is *admissible* if $(X, Y) = K_f$ for some Lorenz map f. Denote by Σ^+, the set of all admissible pairs.

Consider the *shift map* $s : \Sigma \setminus \{0\} \to \Sigma$, $s(X_0 \cdots X_n) = X_1 \cdots X_n$. The set of admissible pairs is characterized, combinatorially, in the following way (see [7]).

Proposition 23.1. *Let $(X, Y) \in \Sigma \times \Sigma$, then $(X, Y) \in \Sigma^+$ if and only if $X_0 = L$, $Y_0 = R$ and, for $Z \in \{X, Y\}$ we have:*

(1) If $Z_i = L$ then $s^i(Z) \leq X$;
(2) If $Z_i = R$ then $s^i(Z) \geq Y$;

with inequality (1) (respectively (2)) strict if X (respectively Y) is finite.

23.2.1 Renormalization and $*$-Product

In the context of Lorenz maps, we define renormalizability on the following way, see for example [13]:

Definition 23.1. Let f be a Lorenz map, then we say that f is renormalizable if there exist $n, m \in \mathbb{N}$ with $n + m \geq 3$ and points $P < y_L < 0 < y_R < Q$ such that

$$g(x) = \begin{cases} f^n(x) & \mathit{if} : y_L \leq x < 0 \\ f^m(x) & \mathit{if} : 0 < x \leq y_R \end{cases}$$

is a Lorenz map.

The map $R_{(n,m)}(f) = g = (f^n, f^m)$ is called the (n, m)-renormalization of f.

Let $|X|$ be the length of a finite sequence $X = X_0 \cdots X_{|X|-1}0$, it is reasonable to identify each finite sequence $X_0 \cdots X_{|X|-1}0$ with the corresponding infinite periodic sequence $(X_0 \cdots X_{|X|-1})^\infty$, this is the case, for example, when we talk about the knot associated to a finite sequence.

It is easy to prove that a pair of finite sequences

$$(X_0 \dots X_{|X|-1}0, Y_0 \dots Y_{|Y|-1}0)$$

is admissible if and only if the pair of infinite periodic sequences

$$((X_0 \cdots X_{|X|-1})^\infty, (Y_0 \cdots Y_{|Y|-1})^\infty)$$

is admissible.

We define the $*$-product between a pair of finite sequences $(X, Y) \in \Sigma \times \Sigma$, and a sequence $U \in \Sigma$ as

$$(X, Y) * U = \overline{U}_0 \overline{U}_1 \cdots \overline{U}_{|U|-1} 0,$$

where

$$\overline{U}_i = \begin{cases} X_0 \dots X_{|X|-1} & \mathit{if}\ U_i = L \\ Y_0 \dots Y_{|Y|-1} & \mathit{if}\ U_i = R \end{cases} .$$

Now we define the $*$-product between two pairs of sequences, $(X, Y), (U, T) \in \Sigma \times \Sigma$, X and Y finite, as

$$(X, Y) * (U, T) = ((X, Y) * U, (X, Y) * T).$$

The next theorem states that the reducibility relative to the $*$-product is equivalent to the renormalizability of the map. The proof can be found, for example, in [7].

Theorem 23.1. *Let f be a Lorenz map, then f is renormalizable with renormalization $R_{(n,m)}(f)$ iff there exist two admissible pairs (X,Y) and (U,T) such that $|X| = n$, $|Y| = m$, $K_f = (X,Y) * (U,T)$ and $K_{R_{(n,m)}(f)} = (U,T)$.*

We know from [7] that $(X,Y) * (U,T) \in \Sigma^+$ if and only if both $(X,Y) \in \Sigma^+$ and $(U,T) \in \Sigma^+$, so for each finite admissible pair (X,Y), the subspace $(X,Y) * \Sigma^+$ is isomorphic to the all space Σ^+, this provides a self-similar structure in the symbolic space of kneading invariants. It is straightforward to verify that the $*$-product of kneading invariants is associative, consequently this self-similar structure is nested. The following proposition states that the order structure is reproduced at each level of renormalization.

Proposition 23.2. *Let (X,Y) be one admissible pair of finite sequences, and $Z < Z'$, then $(X,Y) * Z < (X,Y) * Z'$.*

The proof is straightforward.

23.3 Lorenz Knots and Links

Let $n > 0$ be an integer. We denote by B_n the braid group on n strings given by the following presentation:

$$B_n = \left\langle \sigma_1, \sigma_2, \ldots, \sigma_{n-1} \;\middle|\; \begin{array}{ll} \sigma_i \sigma_j = \sigma_j \sigma_i & (|i-j| \geq 2) \\ \sigma_i \sigma_{i+1} \sigma_i = \sigma_{i+1} \sigma_i \sigma_{i+1} & (i = 1, \ldots, n-2) \end{array} \right\rangle.$$

Where σ_i denotes a crossing between the strings occupying positions i and $i+1$, such that the string in position i crosses (in the up to down direction) over the other, analogously σ_i^{-1}, the algebraic inverse of σ_i, denotes the crossing between the same strings, but in the negative sense, i.e., the string in position i crosses under the other. A *positive braid* is a braid with only positive crossings. A simple braid is a positive braid such that each two strings cross each other at most once. So there is a canonical bijection between the permutation group Σ_n and the set S_n, of simple braids with n strings, which associates to each permutation π, the braid b_π, where each point i is connected by a straight line to $\pi(i)$, keeping all the crossings positive.

Let X be a periodic sequence with least period k and let $\varphi \in \Sigma_k$ be the permutation that associates to each i, the position occupied by $s^i(X)$ in the lexicographic ordering of the k-tuple $(s(X), \ldots, s^k(X))$ $(s^k(X) = X)$. Define $\pi \in \Sigma_k$ to be the permutation given by $\pi(\varphi(i)) = \varphi(i \mod k + 1)$, i.e., $\pi(i) = \varphi(\varphi^{-1}(i) + 1)$. We associate to π the corresponding simple braid $b_\pi \in B_k$ and call it the *Lorenz braid* associated to X. Since X is periodic, this braid represents a knot, and we call it the *Lorenz knot* associated to X.

Example. Let $X=(LRRLR)^{\infty}$. Hence we have $s^5(X)=X$, $s(X)=(RRLRL)^{\infty}$, $s^2(X) = (RLRLR)^{\infty}$, $s^3(X) = (LRLRR)^{\infty}$ and $s^4(X) = (RLRRL)^{\infty}$. Now, ordering the $s^i(X)$ we obtain $s^3(X) < s^5(X) < s^2(X) < s^4(X) < s(X)$ and $\varphi = (1,5,2,3)$ written as a disjoint cycle. Finally we obtain $\pi = (1,4,2,5,3)$ and $b_{\pi} = \sigma_2\sigma_1\sigma_3\sigma_2\sigma_4\sigma_3$.

We can also generalize the previous algorithm to be used in the case of a p-tuple of symbolic periodic sequences $(X^1,\dots,X^p)$ with periods $(k_1,\dots,k_p)$. In this case the permutation $\varphi \in \Sigma_{k_1+\dots+k_p}$ is the permutation that describes the lexicographic ordering of the $(k_1+\dots+k_p)$-tuple $(s(X^1),\dots,s^{k_1}(X^1),\dots,s(X^p)\dots,$ $s^{k_p}(X^p))$ and $\pi \in \Sigma_{k_1+\dots+k_p}$ is defined by $\pi(\varphi(i)) = \varphi(i+1)$ if there is no q such that $i = k_1+\dots+k_q$ and $\pi(\varphi(i)) = \varphi(k_1+\dots+k_{q-1}+1)$ if $i = k_1+\dots k_q$, assuming $k_0 = 0$ (Fig. 23.1).

Remark 23.1. What we are doing here is simply to mark in two parallel lines, $k_1 + \dots + k_p$ points, corresponding in a ordered way, to the sequences $s^{i_j}(X^j)$, $j = 1,\dots,p$, $i_j = 1,\dots,k_j$ and connect by straight lines the points corresponding to $s^{i_j}(X^j)$ with the points corresponding to $s^{i_j+1}(X^j)$, keeping the crossings positive.

A *template* is a compact branched two-manifold with boundary and a smooth expansive semiflow built locally from two types of charts: joining charts and splitting charts. Each chart carries a semiflow, endowing the template with an expanding semiflow, and the gluing maps between charts must reflect the semiflow and act linearly on the edges.

Following [4], we can take a semigroup structure on braided templates. The generators of the braided template semigroup are (see Fig. 23.2):

1. $\sigma_i^{\pm}$, a positive (respectively negative) crossing between the strips occupying the ith and $(i+1)$th positions.
2. $\tau_i^{\pm}$, a half twist in the strip occupying the ith position, in the positive (respectively negative) sense.
3. $\beta_i^{\pm}$, a branch line chart with the ith and $(i+1)$th strips incoming , 2 outgoing strips and either a positive (β_i) or negative (β_i^{-}) crossing at the branch line.

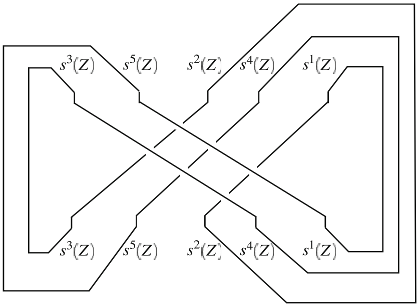


Fig. 23.1 The Lorenz knot associated to $X = (LRRLR)^{\infty}$

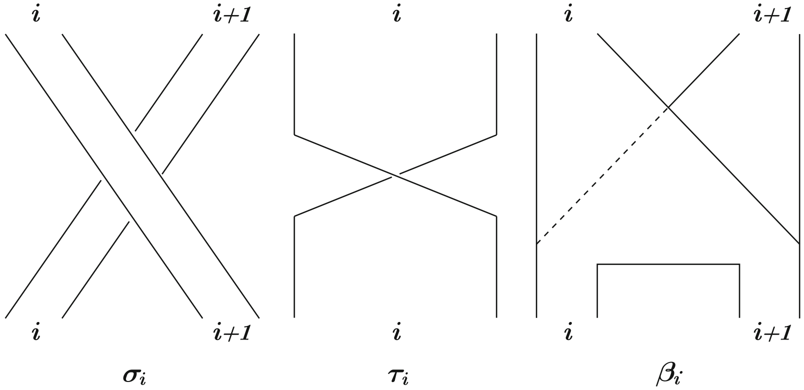


Fig. 23.2 Generators of the braided template semigroup

Given any pair of finite admissible sequences (X, Y), we define the *tail's length* $m(X, Y)$ as

$$m(X, Y) = \min\{i \geq 0 : X_{|X|-1-i} \neq Y_{|Y|-1-i}\}$$

For a finite sequence S, let $n_L(S) = \#\{S_i : 0 \leq i < |S| \text{ and } S_i = L\}$, $n_R(S) = \#\{S_i : 0 \leq i < |S| \text{ and } S_i = R\}$.

Let (X, Y) be a finite admissible pair and $j = \phi(|X| - m(X, Y))$ be the relative position of $X_{|X|-m(X,Y)}$, then we associate to (X, Y) a subtemplate $R(X, Y)$, the *renormalization subtemplate* associated to (X, Y), substituting each string of the braid associated to (X, Y) by a strip and adding $\beta_j^{\pm}$ according if $X_{|X|-m(X,Y)-1} = L$ or $X_{|X|-m(X,Y)-1} = R$, respectively.

The next theorem describes the structure of renormalizable Lorenz links and it was demonstrated in [10].

Theorem 23.2. *Let* (X, Y) *be one admissible pair of finite sequences and* $(Z^1, \ldots, Z^n)$ *be a* n*-tuple of sequences whose associated Lorenz link haves braid word* $\sigma_{p_1} \cdots \sigma_{p_k}$*, then the Lorenz link associated to* $((X, Y) * Z^1, \ldots, (X, Y) * Z^n)$ *is the Lorenz link contained in* $R(X, Y)$ *with:*

1. $|Z^1| + \cdots + |Z^n|$ *strings in each strip if* $s^i(X^\infty) = Y^\infty$ *for some* $i < |X|$.
2. $n_L(Z^1) + \cdots + n_L(Z^n)$ *strings in each strip associated to* X *and* $n_R(Z^1) + \cdots + n_R(Z^n)$ *strings in each strip associated to* Y *if* $s^i(X^\infty) \neq Y^\infty$ *for all* $i < |X|$.

In both cases, the braid word of the restriction to the branch line chart β_j *(respectively* β_j^-*) is* $\sigma_{q+p_1} \cdots \sigma_{q+p_k}$ *(respectively* $\sigma_{q+p_1}^{-1} \cdots \sigma_{q+p_k}^{-1}$*), where* $q + 1$ *is the index of the left-most string getting in* β_j.

23.4 The Formulas

The previous theorem allow us to deduce formulas to compute the trip number and genus of renormalizable Lorenz knots and links. These formulas where all obtained in [10].

We start introducing some terminology, following [2].

Let β be a Lorenz braid, then:

1. The *string index* is the number n of strings in β. It is the sum of the word lengths.
2. The *braid index* of a knot is the minimum string index among all closed braid representatives of that knot.
3. The *crossing number* c is the number of double points in the projected image of the Lorenz braid β.
4. The *linking number* $l(X, Y)$ is the number of crossings between one string from the knot associated to X and one string from the knot associated to Y.
5. The *genus* g of a link L is the genus of M, where M is an orientable surface of minimal genus spanned by L.
6. The *trip number*, t, of a finite sequence X, is the number of syllables in X, a syllable being a maximal subword of X, of the form $L^a R^b$.

Remark 23.2. Birmann and Williams conjectured in [2] that, for the case of a Lorenz knot τ, $b(\tau) = t(\tau)$, where $t(\tau)$ is the trip number of the finite sequence associated to τ. In [11], following a result obtained by Franks and Williams in [6], Waddington observed that this conjecture is true. So our computations will be done about the trip number t.

From now on we freely identify the Lorenz knots (respectively links) with the corresponding periodic sequences (respectively n-tuples of periodic sequences). Denote $n_L(S) = \#\{S_i : 0 \le i < |S| \text{ and } S_i = L\}$, $n_R(S) = \#\{S_i : 0 \le i < |S| \text{ and } S_i = R\}$.

Let (X, Y) and (S, W) be two Lorenz links, defined by the corresponding symbolic sequences, and $(A(n), B(n)) = (X, Y) * (S, W)^n = (X, Y) * (S, W)^{n-1} * (S, W)$. We must consider four distinct cases:

- **Case 1:** $X_{|X|-m(X,Y)-1} = S_{|S|-m(S,W)-1} = L$. In this case we have that $A(n)_{|A(n)|-m(A(n),B(n))-1} = L$ for all n.
- **Case 2:** $X_{|X|-m(X,Y)-1} = L$ and $S_{|S|-m(S,W)-1} = R$. In this case we have that

$$A(n)_{|A(n)|-m(A(n),B(n))-1} = \begin{cases} L \text{ if } n \text{ is even} \\ R \text{ if } n \text{ is odd} \end{cases}.$$

- **Case 3:** $X_{|X|-m(X,Y)-1} = R$ and $S_{|S|-m(S,W)-1} = L$. In this case we have that $A(n)_{|A(n)|-m(A(n),B(n))-1} = R$ for all n.
- **Case 4:** $X_{|X|-m(X,Y)-1} = S_{|S|-m(S,W)-1} = R$. In this case we have that

$$A(n)_{|A(n)|-m(A(n),B(n))-1} = \begin{cases} L \text{ if } n \text{ is odd} \\ R \text{ if } n \text{ is even} \end{cases}.$$

23.4.1 Trip Number

If $X_{|X|-1} = Y_{|Y|-1}$, then

$$t((X,Y) * S) = n_L(S)t(X) + n_R(S)t(Y).$$

If $X_{|X|-1} \neq Y_{|Y|-1}$, then

$$t((X,Y) * S) = n_L(S)t(X) + n_R(S)t(Y) \pm t(S),$$

where we take the signal $+$ in $t(S)$ if $X_{|X|-1} = L$ and signal $-$ otherwise.

Now for each $n \in \mathbb{N}$ we have:

1. If $X_{|X|-1} = Y_{|Y|-1}$, then

$$\begin{bmatrix} t((X,Y) * (S,W)^{n-1} * S) \\ t((X,Y) * (S,W)^{n-1} * W) \end{bmatrix} = \begin{bmatrix} n_L(S) & n_R(S) \\ n_L(W) & n_R(W) \end{bmatrix}^n \begin{bmatrix} t(X) \\ t(Y) \end{bmatrix}.$$

2. If $X_{|X|-1} \neq Y_{|Y|-1}$ and $S_{|S|-1} \neq W_{|W|-1}$, then

$$\begin{aligned}\begin{bmatrix} t((X,Y) * (S,W)^{n-1} * S) \\ t((X,Y) * (S,W)^{n-1} * W) \end{bmatrix} = \begin{bmatrix} n_L(S) & n_R(S) \\ n_L(W) & n_R(W) \end{bmatrix}^n \begin{bmatrix} t(X) \\ t(Y) \end{bmatrix} \\ + \textstyle\sum_{i=0}^{n-1} a_i \begin{bmatrix} n_L(S) & n_R(S) \\ n_L(W) & n_R(W) \end{bmatrix}^i \begin{bmatrix} t(S) \\ t(W) \end{bmatrix}\end{aligned}$$

   In Case 1 $a_i = 1$ for all i; in Case 2 $a_i = (-1)^{i+n+1}$; in Case 3 $a_i = -1$ for all i; in Case 4 $a_i = (-1)^{i+n}$.

3. If $X_{|X|-1} \neq Y_{|Y|-1}$ and $S_{|S|-1} = W_{|W|-1}$, then

$$\begin{aligned}&\begin{bmatrix} t((X,Y) * (S,W)^{n-1} * S) \\ t((X,Y) * (S,W)^{n-1} * W) \end{bmatrix} \\ &= \begin{bmatrix} n_L(S) & n_R(S) \\ n_L(W) & n_R(W) \end{bmatrix}^{n-1} \begin{bmatrix} t((X,Y) * S) \\ t((X,Y) * W) \end{bmatrix} \\ &= \begin{bmatrix} n_L(S) & n_R(S) \\ n_L(W) & n_R(W) \end{bmatrix}^{n-1} \left(\begin{bmatrix} n_L(S) & n_R(S) \\ n_L(W) & n_R(W) \end{bmatrix} \begin{bmatrix} t(X) \\ t(X) \end{bmatrix} \pm \begin{bmatrix} t(S) \\ t(W) \end{bmatrix} \right)\end{aligned}$$

   where we take the signal $+$ in the last summand if $X_{|X|-1} = L$ and the signal $-$ otherwise.

23.4.2 Genus

We start again by presenting the formula for $n = 1$:

$$g((X,Y) * S) = \frac{c(X)n_L(S)^2+c(Y)n_R(S)^2+l(X,Y)n_L(S)n_R(S)-n_L(S)|X|-n_R(S)|Y|+1\pm c(S)}{2},$$

where we take the signal $+$ in $c(S)$ if $X_{|X|-m} = L$ and the signal $-$ otherwise.

Now, considering the matrices

$$A_{33} = \begin{bmatrix} n_L(S)^2 & n_L(W)^2 & 2n_L(S)n_L(W) \\ n_R(S)^2 & n_R(W)^2 & 2n_R(S)n_R(W) \\ n_L(S)n_R(S) & n_L(W)n_R(W) & n_L(W)n_R(S) + n_L(S)n_R(W) \end{bmatrix}$$

and

$$B_{13} = \begin{bmatrix} (n_L(S) + n_L(W))^2 \\ (n_R(S) + n_R(W))^2 \\ (n_L(S) + n_L(W))(n_R(S) + n_R(W)) \end{bmatrix},$$

for each $n \in \mathbb{N}$ we have:

$$g((X,Y) * (S,W)^n) = \frac{1}{2}\left(\begin{array}{c} \left([c(X)c(Y)l(X,Y)]\, A_{33}^{n-1} + [c(S)c(W)l(S,W)] \sum_{i=0}^{n-2} a_i A_{33}^i \right) B_{13} \\ +\alpha c(S,W) - [|X||Y|] \begin{bmatrix} n_L(S) & n_L(W) \\ n_R(S) & n_R(W) \end{bmatrix}^n \begin{bmatrix} 1 \\ 1 \end{bmatrix} \end{array} \right).$$

In Case 1 $a_i = 1 = \alpha$ for all i; in Case 2 $a_i = (-1)^{i+n}$ and $\alpha = (-1)^{n+1}$; in Case 3 $a_i = -1 = \alpha$ for all i; in Case 4 $a_i = (-1)^{i+n+1}$ and $\alpha = (-1)^n$.

23.5 The Algorithms

We start presenting two formulas that will be used through this section.

Theorem 23.3 (Birman and Williams [3]). *Let L be a non-separable link of μ components, presented as a positive braid on N strands with c crossings. Then $g(L)$, the genus of L, is given as*

$$g(L) = \frac{c - N - \mu}{2} + 1.$$

Remark 23.3. In [3] it is presented a formula to compute a braid representative of the lorenz link with a minimal number of strings. We could be tempted, in a first attempt, to use this simpler way to obtain a braid and hence to compute the genus of the Lorenz link. The problem with this approach is that the complexity of computing the Lorenz permutation of a Lorenz link $(X, Y) * (S, W)^n$ grows exponentially with n. The major advantage of our formulae is to skip this expensive step.

Lemma 23.1. *Let Z be a Lorenz link with Lorenz permutation π_Z and Lorenz braid b_Z, then*

$$c(b_Z) = \sum_{i=1}^{n_L(Z)} \pi_Z(i) - i.$$

Proof. Each string that goes from i to $\pi_Z(i)$ will cross exactly $\pi_Z(i)$ strings arriving from the right side minus the number of strings that previously arrived to position $j < \pi_Z(i)$ and departed from position $k \leq i$. This gives the formula since $1 \leq k \leq i$. □

We will now present the algorithms to be tested.

The Trip Number Definition Algorithm (TD)

INPUT: A pair of Lorenz links (X, Y), (S, W) and an integer n.

1. Compute the star product $(X, Y) * (S, W)^n$.
2. Compute the vector $T((X, Y) * (S, W)^n) = (t((X, Y) * (S, W)^{n-1} * S, t((X, Y) * (S, W)^{n-1} * W)))$ using the definition.

OUTPUT: $T((X, Y) * (S, W)^n)$.

The Genus Definition Algorithm (GD)

INPUT: A pair of Lorenz links (X, Y), (S, W) and an integer n.

1. Compute the star product $(X, Y) * (S, W)^n$.
2. Compute $c((X, Y) * (S, W)^n)$.
3. Compute $|(X, Y) * (S, W)^n|$.
4. Compute the genus $G = g((X, Y) * (S, W)^n)$ using the formula from Theorem 23.3.

OUTPUT: The integer G, the genus of $(X, Y) * (S, W)^n$.

The Trip Number Formulae Algorithm (TF)

INPUT: A pair of Lorenz links $L_1 = (X_1, X_2)$, $L_2 = (X_3, X_4)$ and an integer n.

1. Compute $n_L(X_i), n_R(X_i), c(X_i), c(L_j)$ for $i = 1, \ldots, 4$ and $j = 1, 2$.
2. Compute the vector $T = t(L_1 * L_2^n)$ using the formulas in the previous section.

OUTPUT: $T = t(L_1 * L_2^n)$.

The Genus Formulae Algorithm (GF)

INPUT: A pair of Lorenz links $L_1 = (X_1, X_2), L_2 = (X_3, X_4)$ and an integer n.

1. Compute $n_L(X_i), n_R(X_i), c(X_i), c(L_j)$ for $i = 1, \dots, 4$ and $j = 1, 2$.
2. Compute the genus $G = g(L_1 * L_2^n)$ using the formulas in the previous section.

OUTPUT: The integer G, the genus of $L_1 * L_2^n$.

Now we will evaluate the (worst case) complexity of the previous algorithms in order to obtain upper bounds for them.

Proposition 23.3. *Given a pair of Lorenz links, (X, Y), (S, W), then the algorithms TD and GD exhibit exponential complexity with n and the algorithms TF and GF exhibit linear complexity.*

Proof. We start by analyzing the complexity of the $*$-product. Given two links $L_1 = (X, Y)$ and $L_2 = (S, W)$, in order to compute $L = L_1 * L_2$ we just replace each appearance of L (respectively R) in L_2 by X (respectively Y). This is done with complexity $O(|S| + |W|)$, so to compute $L = L_1 * L_2^n$ we have complexity $O((|S| + |W|)^n)$. The length of L is bounded by $(|X| + |Y|)(|S| + |W|)^n$. The complexity of computing the crossings of the Lorenz link $Z = (Z_1, \dots, Z_k)$ (using the Lorenz permutation) depends, by Lemma 23.1, linearly on the complexity of computing the Lorenz permutation π_Z. To obtain π_Z we must order the elements $s^i(Z_j)$ with $i = 1, \dots, |Z_j|$ and $j = 1, \dots, k$. This ordering can be done, classically, with complexity $O((\sum_{i=1}^k |Z_i|) \log(\sum_{i=1}^k |Z_i|))$. Hence the complexity for computing $C(Z)$ is $O(n_L(Z)(\sum_{i=1}^k |Z_i|) \log(\sum_{i=1}^k |Z_i|))$. Notice that we can compute both $n_L(X)$ and $n_R(X)$ at the same time in $O(|X|)$ time. Counting the number of syllables of type $L^j R^i$ takes also linear time in the length of the sequence. The complexity for computing A^n, where A is a fixed square matrix is $O(n)$. So we have the following complexities:

- *Algorithm TD:* $O((1 + |X| + |Y|)(|S| + |W|)^n)$.
- *Algorithm TF:* $O(|X|^2 \log(|X|)) + O(|Y|^2 \log(|Y|)) + O(|S|^2 \log(|S|)) + O(|W|^2 \log(|W|)) + O((|X| + |Y|)^2 \log(|X| + |Y|)) + O((|S| + |W|)^2 \log(|S| + |W|)) + O(nT_{X,Y,S,W}) = O((|X| + |Y|)^2 \log(|X| + |Y|)) + O((|S| + |W|)^2 \log(|S| + |W|)) + O(nT_{X,Y,S,W})$.
- *Algorithm GD:* $O((|S| + |W|)^n) + O((|X| + |Y|)(|S| + |W|)^{2n} \log((|X| + |Y|)(|S| + |W|)^n)) = O((\log(|X| + |Y|))(|X| + |Y|)(|S| + |W|)^{2n} + n(|X| + |Y|)(|S| + |W|)^{2n+1})$.
- *Algorithm GF:* $O(|X|^2 \log(|X|)) + O(|Y|^2 \log(|Y|)) + O(|S|^2 \log(|S|)) + O(|W|^2 \log(|W|)) + O((|X| + |Y|)^2 \log(|X| + |Y|)) + O((|S| + |W|)^2 \log(|S| + |W|)) + O(nG_{X,Y,S,W}) = O((|X| + |Y|)^2 \log(|X| + |Y|)) + O((|S| + |W|)^2 \log(|S| + |W|)) + O(nG_{X,Y,S,W})$.

Where $T_{X,Y,S,W}$ (respectively $G_{X,Y,S,W}$) represents the cost of the matrix multiplications in the Trip Number (respectively Genus) formula.

To illustrate the effectiveness of our algorithms we selected randomly 1,000 pairs of Lorenz links L_1, L_2 and computed $L_1 * L_2^n$ with $n = 1, \ldots, 6$. The Lorenz links are built from irreducible sequences up to level 10 in the Farey tree (see [9]). We computed the average of the running times for the trip number and the genus (only using the formula), sorted by trip number (see Tables 23.1, 23.2 and 23.3). It was not possible to use the classical algorithm to compute the genus, under renormalization, because the lack of computer power and memory. In fact for a random pair of Lorenz links we get systematically "out of memory" message. This is due to the fact that computing the $*$-product and reordering all the shifts, for computing the Lorenz permutation, takes too much resources.

Table 23.1 Results of the computations using algorithm *TF* (time in seconds)

Trip	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$
2	0.002325	0.003600	0.003500	0.003100	0.005150	0.005450
3	0.003237	0.002895	0.002447	0.002868	0.002895	0.005737
4	0.002678	0.003023	0.004172	0.003943	0.007149	0.007483
5	0.002918	0.004071	0.003988	0.003694	0.005424	0.004765
6	0.002833	0.004402	0.003576	0.004750	0.005295	0.006598
7	0.003020	0.003444	0.004131	0.005687	0.006303	0.006505
8	0.002367	0.002677	0.004567	0.004642	0.007687	0.005963
9	0.003543	0.002391	0.004880	0.004935	0.004076	0.006076
10	0.002785	0.003514	0.003776	0.005075	0.004075	0.006224
11	0.002350	0.003133	0.004050	0.004133	0.005417	0.009683
12	0.002818	0.004515	0.003530	0.005455	0.005652	0.010879
13	0.001778	0.000889	0.004444	0.005278	0.005278	0.007722
14	0.004813	0.004875	0.005813	0.006875	0.005813	0.011688
15	0.001632	0.004158	0.004105	0.006579	0.005000	0.011474
16	0.005643	0.004429	0.006714	0.004429	0.004429	0.013429

Table 23.2 Results of the computations using algorithm *TD* (time in seconds)

Trip	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$
2	0.000400	0.006250	0.029250	0.151325	0.992575	5.908250
3	0.001605	0.011947	0.091579	0.484079	3.848842	18.043000
4	0.000724	0.014333	0.132850	1.201874	10.989299	25.686759
5	0.001259	0.030682	0.202459	1.778259	16.486611	30.499059
6	0.001280	0.023228	0.257932	3.012310	24.085773	35.740773
7	0.000949	0.047424	0.331495	3.689444	27.710505	32.048919
8	0.001978	0.042709	0.414291	5.384129	36.054769	⋆
9	0.000859	0.046098	0.516076	6.699478	36.390554	⋆
10	0.001607	0.040150	0.598673	9.000421	34.995813	⋆
11	0.001300	0.047383	0.807850	12.755267	31.692850	⋆
12	0.002091	0.058348	1.031470	15.913545	30.815955	⋆
13	0.004222	0.057944	0.944556	17.046333	32.592167	⋆
14	0.003813	0.063375	1.205063	20.203875	⋆	⋆
15	0.003316	0.069842	1.255579	26.708316	⋆	⋆
16	0.006714	0.078143	1.531143	35.499286	⋆	⋆

Table 23.3 Results of the computations using algorithm *GF* (time in seconds)

Trip	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$	$n = 7$	$n = 8$	$n = 9$	$n = 10$
2	0.01400	0.01480	0.01688	0.02068	0.02933	0.02410	0.03310	0.03193	0.03430	0.03795
3	0.01120	0.01478	0.01595	0.02103	0.02140	0.02505	0.02883	0.03090	0.03548	0.03743
4	0.01783	0.01520	0.01709	0.02406	0.02382	0.02642	0.03540	0.03293	0.03859	0.04227
5	0.01995	0.01901	0.02389	0.02151	0.02392	0.02766	0.03273	0.03257	0.04154	0.03829
6	0.01491	0.01665	0.01959	0.02238	0.02523	0.03159	0.03311	0.03444	0.03707	0.04741
7	0.01489	0.01751	0.01915	0.02342	0.02523	0.03040	0.03123	0.04570	0.03969	0.04063
8	0.01916	0.02055	0.02265	0.02329	0.02505	0.02941	0.03207	0.03953	0.03774	0.04464
9	0.01841	0.01853	0.02097	0.02270	0.02689	0.02974	0.03592	0.03904	0.03892	0.04395
10	0.01705	0.01986	0.02203	0.02730	0.03022	0.03053	0.03412	0.03623	0.04204	0.04484
11	0.02037	0.02096	0.02294	0.02482	0.02780	0.03094	0.03594	0.04392	0.03945	0.04810
12	0.01867	0.02053	0.02323	0.02656	0.02846	0.03725	0.04021	0.03775	0.04512	0.04463
13	0.01915	0.02042	0.02281	0.02819	0.03173	0.03181	0.03473	0.04188	0.05165	0.04681
14	0.03313	0.02404	0.02546	0.02854	0.03067	0.03383	0.04675	0.04208	0.04088	0.04621
15	0.02655	0.03510	0.02655	0.02965	0.03200	0.03430	0.03595	0.04130	0.04520	0.05835
16	0.02350	0.03150	0.03100	0.03150	0.03100	0.03900	0.03100	0.04650	0.04700	0.04650

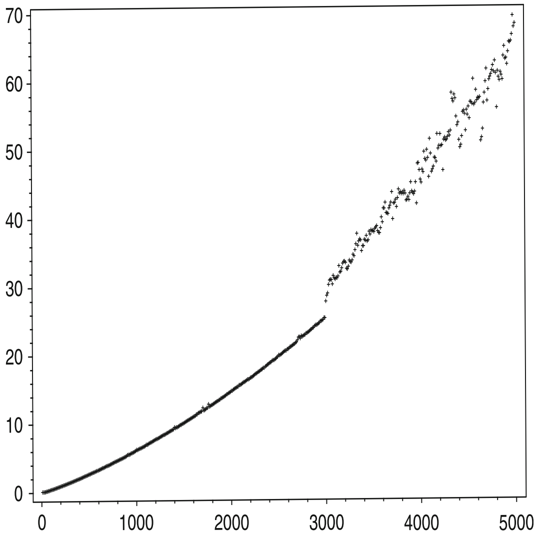


Fig. 23.3 Computations of the Genus

In order to test the real possibilities of the formulas we selected a random pair of links and computed genus (see Fig. 23.3) using the formula for $n = 1, \ldots, 5{,}000$ (time is in seconds). The values of the genus reach as high as $0.51167 \times 10^{10{,}659}$.

We also computed the trip number (see Fig. 23.4) using the formula for $n = 1, \ldots, 300{,}000$ (time is in seconds). The values of the trip number reach as high as $0.17187 \times 10^{288{,}839}$

There are two main conclusions to take out of these computations. The first one is that the algorithms arising from the formulae are much more effective in

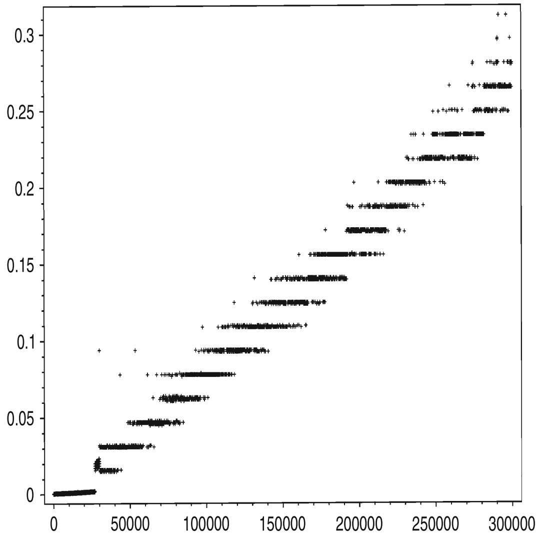


Fig. 23.4 Computations of the Trip Number

the computations of these invariants than the ones arising from the definitions. The second conclusion is that this formulae allow us to, effectively, compute these invariants in much deeper regions of renormalization.

These are powerful tools to compute these combinatorial invariants and, in some future work, they may help us to understand better how these invariants, and it's associated Lorenz links, behave assymptotically as we dive in to deeper regions of renormalization.

References

1. Artin, E.: Theory of braids. Ann. Math. **48**(2), 101–126 (1947)
2. Birmann, J., Williams, R.F.: Knotted periodic orbits in dynamical systems I: Lorenz's equations. Topology **22**, 47–82 (1983)
3. Birmann, J., Williams, R.F.: Knotted periodic orbits in dynamical systems II: knot holders for fibered knots. Cont. Math. **20**, 1–60 (1983)
4. Ghrist, R., Holmes, P., Sullivan, M.: Knots and Links in Three-Dimensional Flows. Lecture Notes in Mathematics. Springer (1997)
5. Holmes, P.: Knotted periodic orbits in suspensions of Smale's horseshoe: period multiplying and cabled knots. Physica **21D**, 7–41 (1986)
6. Franks, J., Williams, R.F.: Braids and the Jones polynomial. Trans. Am. Math. Soc **303**, 97–108 (1987)
7. Silva, L., Sousa Ramos, J.: Topological invariants and renormalization of Lorenz maps. Physica D **162**(3-4), 233–243 (2002)
8. de Melo, W., Martens, M.: Universal models for Lorenz maps. Ergod. Theory Dyn. Syst. **21**, 833–860 (2001)

9. Franco, N., Silva, L.: Effective computation of the multivariable Alexander polynomial of Lorenz links. Physica D **237**, 3322–3328 (2008)
10. Franco, N., Silva, L.: Genus and braid index associated to sequences of renormalizable Lorenz maps (Submitted)
11. Waddington, S.: Asymptotic formulae for Lorenz and horseshoe knots. Comm. Math. Phys. **176**(2), 273–305 (1996)
12. Williams, R.: The structure of Lorenz attractors. In: Chorin, A., Marsden, J., Smale, S. (eds.) Turbulence Seminar, Berkeley 1976/77, Lecture Notes in Mathematics, vol. 615, pp. 94–116. Springer (1977)
13. Williams, R.: The structure of Lorenz attractors. Publ. Math. I.H.E.S. **50**, 73–99 (1979)

Chapter 24
Statistical Stability for Equilibrium States

Jorge Milhazes Freitas and Mike Todd

Abstract We consider multimodal interval maps with at least polynomial growth of the derivative along the critical orbit. For these maps Bruin and Todd showed the existence and uniqueness of equilibrium states for the potential $\phi_t : x \to -t log|Df(x)|$, for t close to 1. We show that for certain families of this type of maps the equilibrium states vary continuously in the weak* topology, when we perturb the map within the respective family. Moreover, in the case $t = 1$, when the equilibrium states are absolutely continuous with respect to Lebesgue, we show that the densities also vary continuously in the $L^1 - norm$.

24.1 Introduction

One of the main goals in the study of Dynamical Systems is to understand how the behaviour changes when we perturb the underlying dynamics. We examine the persistence of statistical properties of a multimodal interval map (I, f). In particular we are interested in the behaviour of the Cesaro means $\frac{1}{n}\sum_{k=0}^{n-1}\phi\circ f^k(x)$ for a potential $\phi : I \to \mathbf{R}$ for "some" points x, as $n \to \infty$. If the system possesses an invariant *physical measure* μ, then part of this statistical information is described by μ since, by definition of physical measure, there is a positive Lebesgue measure set of points $x \in I$ such that

$$\overline{\phi}(x) := \lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1}\phi\circ f^k(x) = \int \phi\, d\mu.$$

If for nearby dynamics these measures are proven to be close, then the Cesàro means do not change much under small deterministic perturbations. This motivated Alves

J.M. Freitas (✉) and M. Todd
Centro de Matemática da Universidade do Porto, Rua do Campo Alegre 687, 4169-007 Porto, Portugal
e-mail: jmfreita@fc.up.pt, mtodd@fc.up.pt

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_24,

and Viana [1] to propose the notion of *statistical stability*, which expresses the persistence of statistical properties in terms of the continuity, as a function of the map f, of the corresponding physical measures.

However, the study of Cesàro means is not confined to the analysis of these measures. The study of other classes can be motivated through the encoding of these statistical properties by "multifractal decomposition", see [14] for a general introduction. Given $\alpha \in \mathbf{R}$, we define the sets

$$B_\phi(\alpha) := \{x \in I : \overline{\phi}(x) = \alpha\}, \ B'_\phi := \{x \in I : \overline{\phi}(x) \text{ does not exist}\}.$$

Then the multifractal decomposition in this case is

$$I = B'_\phi \cup \left(\bigcup_\alpha B_\phi(\alpha) \right).$$

Understanding the nature of this decomposition gives us information about the statistical properties of the system. This can be studied via "equilibrium states". See [16] for a fuller account of these ideas, where the theory is applied to subshifts of finite type. To define equilibrium states, given a potential $\phi : I \to \mathbf{R}$, we define the *pressure* of ϕ to be

$$P(\phi) := \sup \left\{ h_\mu + \int \phi \, d\mu \right\},$$

where this supremum is taken over all invariant ergodic probability measures. Here h_μ denotes the metric entropy of the system (I, f, μ). Any such measure μ which "achieves the pressure", i.e. $h_\mu + \int \phi \, d\mu = P(\phi)$, is called an *equilibrium state* for (I, f, ϕ).

For a given map f, we are interested in the equilibrium state μ_t of the "natural" potential $\phi_t : x \mapsto -t \log |Df(x)|$ for different values of t. For a multimodal map f and an f-invariant probability measure μ we denote the *Lyapunov exponent of* μ by $\lambda(\mu) := \int \log |Df| \, d\mu$. For any f in the class of multimodal maps $\mathscr{F}$ which we define below, Ledrappier [12] showed that for $t = 1$, there is an equilibrium state μ_1 with $\lambda(\mu_1) > 0$ if and only if μ_1 is absolutely continuous with respect to Lebesgue. We then refer to μ_1 as an acip. In this setting any acip is also a physical measure.

Using tools developed by Keller and Nowicki in [11], Bruin and Keller [6] further developed this theory, showing that for unimodal Collet–Eckmann maps there is an equilibrium state μ_t for ϕ_t for all t close to 1. This range of parameters was extended to all t in a neighbourhood of $[0, 1]$ for a special class of Collet–Eckmann maps by Pesin and Senti [15]. Bruin and Todd showed similar results for the non-Collet Eckmann multimodal case in [8].

The *Lyapunov exponent of a point* $x \in I$ is defined as $\overline{\phi}_1(x)$, if this limit exists. So the set of points with the same Lyapunov exponent is $B_{\phi_1}(\alpha)$. If f is transitive and there exists an acip then by the ergodic theorem $\mu_1(B_{\phi_1}(\lambda(\mu_1))) = 1$. As shown in [20], under certain growth conditions on f, for a given value of α, close to

$\lambda(\mu_1)$, there is an equilibrium state μ_t supported on $B_{\phi_1}(\alpha)$ for some t close to 1. Therefore, to understand the statistics of the system with potential ϕ_1, it is useful to study the properties of the equilibrium states μ_t. We mention the pioneer work of Bohr and Rand [5] that considered the multifractal spectrum for Lyapunov exponent of non-uniformly expanding interval maps.

24.2 Statement of Results

Let Crit $=$ Crit(f) denote the set of critical points of f. We say that $c \in$ Crit is a *non-flat* critical point of f if there exists a diffeomorphism $g_c : \mathbf{R} \to \mathbf{R}$ with $g_c(0) = 0$ and $1 < \ell_c < \infty$ such that for x close to c, $f(x) = f(c) \pm |g_c(x-c)|^{\ell_c}$. The value of ℓ_c is known as the *critical order* of c. We define $\ell_{max}(f) := \max\{\ell_c : c \in \text{Crit}(f)\}$. Throughout $\mathscr{H}$ will be the collection of C^2 interval maps which have negative Schwarzian (that is, $1/\sqrt{|Df|}$ is convex away from critical points) and all critical points non-flat.

For ease of exposition, we will assume that maps in $\mathscr{H}$ are non-renormalisable with only one transitive component Ω of the non-wandering set, a cycle of intervals. We also assume that for any $f \in \mathscr{H}$, $f^j(\text{Crit}) \cap f^k(\text{Crit}) \neq \emptyset$ implies $j = k$. For maps failing this assumption, either $f^k(\text{Crit}) \cap \text{Crit} \neq \emptyset$ for some $k \in \mathbf{N}$, in which case we could consider these relevant critical points in a block; or some critical point maps onto a repelling periodic cycle, which we exclude here for ease of exposition since our method is particularly tailored to case of more interesting maps where the critical orbits are infinite. It is also convenient to suppose that there are no points of inflection.

Let $\mathscr{H}_{r,\ell} \subset \mathscr{H}$ denote the set of maps $f \in \mathscr{H}$ with r critical points with $\ell_{max}(f) \leq \ell$. We will consider families of maps in $\mathscr{H}$ which satisfy the following conditions. The first one is the Collet–Eckmann condition: For any $r \in \mathbf{N}$, $\ell \in (1, \infty)$ and $C, \alpha > 0$, the class $\mathscr{F}_e(r, \ell, C, \alpha)$ is the set

$$f \in \mathscr{H}_{r,\ell} \text{ such that } |Df^n(f(c))| \geq Ce^{\alpha n} \quad \text{for all} \quad c \in \text{Crit} \quad \text{and} \quad n \in \mathbf{N}. \tag{24.1}$$

Secondly we consider maps satisfying a polynomial growth condition: For any $r \in \mathbf{N}$, $\ell \in (1, \infty)$, $C > 0$, and any $\beta > 2\ell$, the class $\mathscr{F}_p(r, \ell, C, \beta)$ is the set

$$f \in \mathscr{H}_{r,\ell} \quad \text{such that} \quad |Df^n(f(c))| \geq Cn^{\beta} \text{ for all } c \in \text{Crit} \quad \text{and} \quad n \in \mathbf{N}. \tag{24.2}$$

We will take a map $f_0 \in \mathscr{F}$ where we suppose that either $\mathscr{F} = \mathscr{F}_e(r, \ell, C, \alpha)$ or $\mathscr{F} = \mathscr{F}_p(r, \ell, C, \beta)$, and consider the continuity properties of equilibrium states for maps in $\mathscr{F}$ at f_0.

We will consider equilibrium states for maps in these families. Suppose first that $\mathscr{F} = \mathscr{F}_e(r, \ell, C, \alpha)$. Then by [8, Theorem 2], there exists an open interval $U_{\mathscr{F}} \subset \mathbf{R}$ containing 1 and depending on α and r so that for $f \in \mathscr{F}$ and $t \in U_{\mathscr{F}}$ the potential $\phi_{f,t} : x \mapsto -t \log |Df(x)|$ has a unique equilibrium state $\mu = \mu_f$. We note that by [13], the fact that there are r critical points gives a uniform upper bound $\log(r+1)$ on the topological entropy, which plays an important role in the computations which determine $U_{\mathscr{F}}$ in [8]. If instead we assume that $\mathscr{F} = \mathscr{F}_p(r, \ell, C, \beta)$ then by [8, Theorem 1] we have the same result but instead $U_{\mathscr{F}}$ is of the form $(t_{\mathscr{F}}, 1]$ where $t_{\mathscr{F}}$ depends on β, ℓ and r.

We choose our family $\mathscr{F}$, fix $t \in U_{\mathscr{F}}$ and denote $\phi_{f,t}$ by ϕ_f. For every sequence $(f_n)_n$ of maps in $\mathscr{F}$ we let $\mu_{n,t} = \mu_{f_n,t}$ denote the corresponding equilibrium state for each n with respect to the potential ϕ_{f_n}. We fix $f_0 \in \mathscr{F}$ and say that $\mu_{0,t}$ is *statistically stable within the family* $\mathscr{F}$ if for any sequence $(f_n)_n$ of maps in $\mathscr{F}$ such that $\|f_n - f_0\|_{C^2} \to 0$ as $n \to \infty$, we have that $\mu_{0,t}$ is the weak* limit of $(\mu_{n,t})_n$.

Theorem 24.1 ([10]). *Let $\mathscr{F} \subset \mathscr{H}$ be a family satisfying (24.1) or (24.2) with potentials $\phi_{f,t}$ as above. Then, for every fixed $t \in U_{\mathscr{F}}$ and $f \in \mathscr{F}$, the equilibrium state $\mu_{f,t}$ as above is statistically stable within the family $\mathscr{F}$.*

Although the definition of statistical stability involves convergence of measures in the weak* topology, when we are dealing with acips, it makes sense to consider a stronger type of stability due to Alves and Viana [1]: for a fixed $f_0 \in \mathscr{F}$, we say that the acip μ_{f_0} is *strongly statistically stable* in the family $\mathscr{F}$ if for any sequence $(f_n)_n$ of maps in $\mathscr{F}$ such that $\|f_n - f_0\|_{C^2} \to 0$ as $n \to \infty$ we have

$$\int \left| \frac{d\mu_{f_n}}{dm} - \frac{d\mu_{f_0}}{dm} \right| dm \to 0, \tag{24.3}$$

as $n \to \infty$ where m denotes Lebesgue measure and μ_{f_n} and μ_{f_0} denote the acips for f_n and f_0 respectively. As a byproduct of the proof of Theorem 24.1 we also obtain:

Theorem 24.2 ([10]). *Let $\mathscr{F} \subset \mathscr{H}$ be a family satisfying (24.1) or (24.2). Then, for every $f \in \mathscr{F}$, the acip μ_f is strongly statistically stable.*

For uniformly hyperbolic maps, it is known that the measures do not merely vary continuously with the map, but actually vary differentiably in the sense of Whitney. For example, if $f_0 : M \to M$ is a C^3 Axiom A diffeomorphism of a manifold M with an unique physical measure μ_0, and the family $t \mapsto f_t$ is C^3, then the map $t \mapsto \int \psi \, d\mu_t$ is differentiable at $t = 0$ for any real analytic observable $\psi : M \to \mathbf{R}$, see [17]. We would like to emphasise that the situation for non-uniformly hyperbolic maps is quite different. For example if $\mathscr{F}$ is the class of quadratic maps for which acips exist then it was shown in [19] that these measures are not even continuous everywhere in this family, although as proved in [21] they are continuous on a positive Lebesgue measure set of parameters. It has been conjectured in [3] that if $\mathscr{F}$ is the set of quadratic maps with some growth along the critical orbit then the acips should be at most Hölder continuous in this class, see also [4]. For a positive result in that direction [18] proved the Hölder continuity of the densities of the acips as in (24.3) for Misiurewicz parameters. Later, in [9], strong statistical stability was proved for Benedicks–Carleson quadratic maps, which are unimodal and satisfy condition (24.1). Hence, Theorem 24.2 provides a generalisation of this last result. We would like to point out that for continuous potentials $\upsilon : I \to \mathbf{R}$ the theory of statistical stability has been studied in [2].

References

1. Alves, J.F., Viana, M.: Statistical stability for robust classes of maps with non-uniform expansion. Ergod. Theory Dyn. Syst. **22**, 1–32 (2002)
2. Araújo, V.: Semicontinuity of entropy, existence of equilibrium states and continuity of physical measures. Discrete Contin. Dyn. Syst. **17**, 371–386 (2007)
3. Baladi, V.: On the susceptibility function of piecewise expanding interval maps. Comm. Math. Phys. **275**, 839–859 (2007)
4. Baladi, V.: Linear response despite critical points. Nonlinearity **21**, T81–T90 (2008)
5. Bohr, T., Rand, D.: The entropy function for characteristic exponents. Phys. D **25**, 387–398 (1987)
6. Bruin, H., Keller, G.: Equilibrium states for S-unimodal maps. Ergod. Theory Dyn. Syst. **18**, 765–789 (1998)
7. Bruin, H., Luzzatto, S., van Strien, S.: Decay of correlations in one–dimensional dynamics. Ann. Sci. École Norm. Sup. **36**, 621–646 (2003)
8. Bruin, H., Todd, M.: Equilibrium states for interval maps: the potential $-t \log |Df|$. Ann. Sci. École Norm. Sup. **42**(4), 559–600 (2009)
9. Freitas, J.M.: Continuity of SRB measure and entropy for Benedicks–Carleson quadratic maps. Nonlinearity **18**, 831–854 (2005)
10. Freitas, J.M., Todd, M.: Statistical stability of equilibrium states for interval maps. Nonlinearity **22**, 259–281 (2009)
11. Keller, G., Nowicki, T.: Spectral theory, zeta functions and the distribution of periodic points for Collet–Eckmann maps. Comm. Math. Phys. **149**, 31–69 (1992)
12. Ledrappier, F.: Some properties of absolutely continuous invariant measures on an interval. Ergod. Theory Dyn. Syst. **1**, 77–93 (1981)
13. Misiurewicz, M., Szlenk, W.: Entropy of piecewise monotone mappings. Studia Math. **67**, 45–63 (1980)
14. Pesin, Y.: Dimension Theory in Dynamical Systems. University of Chicago Press, Chicago, IL (1997)
15. Pesin, Y., Senti, S.: Equilibrium measures for maps with inducing schemes. J. Mod. Dyn. **2**, 397–427 (2008)
16. Pesin, Y., Weiss, H.: The multifractal analysis of Birkhoff averages and large deviations. Global Analysis of Dynamical Systems, pp. 419–431. Institute of Physics, Bristol (2001)
17. Ruelle, D.: Differentiation of SRB states. Comm. Math. Phys. **187**, 227–241 (1997)
18. Rychlik, M., Sorets, E.: Regularity and other properties of absolutely continuous invariant measures for the quadratic family. Commun. Math. Phys. **150**, 217–236 (1992)
19. Thunberg, H.: Unfolding of chaotic unimodal maps and the parameter dependence of natural measures. Nonlinearity **14**, 323–337 (2001)
20. Todd, M.: Multifractal analysis for multimodal maps. Preprint, arXiv:0809.1074
21. Tsujii, M.: On continuity of Bowen–Ruelle–Sinai measures in families of one dimensional maps. Commun. Math. Phys. **177**, 1–11 (1996)

Chapter 25
Dynamic Games of Network Effects

Filomena Garcia and Joana Resende

Abstract Network effects occur when the benefit that agents derive from a good or service depends on how many other agents adopt the same good or service. This strategic complementarity between consumers' actions has several implications on the behavior of firms: For instance, firms need to gain advantage from early marketing stages. Network effects are intrinsically a dynamic phenomenon: past consumption of the good influences the utility of present consumers. This effect can be either direct, when consumers value interaction with their peers, and/or indirect, through an increase in the quality of the good. This chapter surveys the literature on dynamic network effects. First we provide general formulations for the modelization of network effects in a dynamic setting. Second, we analyse recent developments in the literature on firms' strategies in the context of dynamic network effects. We survey the literature both on monopoly and oligopoly markets. In the case of oligopoly markets, we distinguish between situations in which firms produce horizontally and vertically differentiated goods. Main results on pricing and evolution of market shares are exposed.

25.1 Introduction

Network effects exist when agents are better off by choosing a certain action when their peers also do so. Several are the examples of these effects: joining a certain club, purchasing an operating system, buying a specific brand of clothes, using a telephone or a fax machine. Economists and sociologists have long recognized this

F. Garcia (✉)
ISEG – Technical University of Lisbon and UECE, Rua Miguel Lupi, 20 1249-078 Lisbon, Portugal
e-mail: fgarcia@iseg.utl.pt

J. Resende
CEF.UP, University of Porto, Rua Dr. Roberto Frias, 4200-464 Porto, Portugal
e-mail: jresende@fep.up.pt

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_25,

pattern of consumption in which agents derive utility from the good depending on the size of the demand for the good. Among economic sociologists, this phenomenon has been made famous by Veblen [63], who argued that a considerable part of households' consumption is what he designated by "conspicuous consumption" i.e., consumption that has the sole purpose of indicating the level of wealth of consumers.[1] A first economic analysis of this interdependence of demands is due to Pigou [56]. However his analysis is rather succinct and concentrates solely on discussing the existence of equilibrium, leaving out many of the details that later on have been found to be relevant in this literature.[2,3] Once the foundations have been laid, many have explored the effects of conspicuous consumption or interdependent demand. Considering the former, and on a more sociological aspect, authors have concentrated on the emergence of trends and fads, and on explaining the phenomena of fashion and snobbism.[4] Regarding the latter, the concept of interdependent demand has been in the background for the development of network economics. The first to analyse the effect of demand interdependence in the communications market was Rohlfs [62] who developed the exact same concept put forth by Pigou [56], stressing how prices are affected by these effects.[5] Followers of this work are Katz and Shapiro, Farrell and Saloner, Arthur, Liebowitz and Margolis and more recently, Colla and Garcia, Gabszewicz and Garcia, Markovich, Driskill, Cabral and Laussel and Resende. Within these papers several issues have been treated. Among these we find: pricing, standardization and compatibility, adoption of technology, innovation, antitrust, path dependence and *lock in* to name a few.

Seminal literature includes mainly a static treatment of network effects, i.e., consumers' utility depends positively on the number of other consumers who contemporaneously choose the same good or service. As research evolved, it became clear that a dynamic analysis of network effects is fundamental to capture the essence of the issue at hand. Network effects are intrinsically dynamic: agents' utility is increasing in the overall number of other agents, both contemporaneous and past, who acquire the good. Moreover agents consider the evolution of the networks while choosing in the present, in the sense that they forecast what network will have more users in the future. All these considerations are better captured in a dynamic framework. Surprisingly, only recently did the dynamic aspects of networks start

[1] John Rae was the first sociologist to identify the "conspicuous consumption" phenomenon, in writings dated back to 1834. It was the treatment of Veblen, however, that made the notion so popular. See Leibenstein [49] for a discussion.

[2] The interdependence of consumption has also been analysed from the point of view of welfare by Meade [55], Pigou [57].

[3] See Leibenstein [49] for a formal analysis of the different aspects of consumption, namely: functional consumption and non-functional consumption which includes the Veblen effect, the snob effect and the bandwagon effect.

[4] See, for example, Bikhchandani et al. [8], Bernheim [9], Corneo and Jeanne [15], and Corneo and Jeanne [16].

[5] Surprisingly, Rohlfs was unaware that some economics literature had already focused on the interdependence of demands, hence he does not cite the work of Pigou [56].

to be in the research agenda of scholars.[6] Some of these papers are Bensaid and Lesne [7], Colla and Garcia [14], Gabszewicz and Garcia [32, 33], etc.

While in a static setup, consumers must form expectations concerning the size of competing networks in the present, in the dynamic model expectations might become irrelevant, depending on the type of network effects considered. If we assume that it takes some periods of time for users to fully enjoy the benefits of the network, then past consumption is the only information that users must have prior to their choice. Obviously, depending on whether we consider a static or a dynamic context, different equilibrium concepts for the game must be used.

The notion of equilibrium used in the static framework is rational, or fulfilled expectations equilibrium. The problem associated to this notion is that multiple fulfilled expectations equilibria may exist and hence a selection procedure becomes crucial. In the dynamic setup, without forward looking behavior, both Subgame Perfect Equilibrium and Markov Perfect Equilibrium have been used. With forward looking behavior a unique equilibrium may be selected using the tools of global games and considering equilibria with switching strategies.[7] In fact, network effect models belong to the class of games of strategic complementarities (as also, with minor additions, to the class of global games). When an individual chooses to join a network, the utility for others to join the same network also increases. Hence, the actions of joining a network are strategic complements among agents. Games of strategic complementarities, also known as supermodular games, have a structure that allows to derive some interesting results, namely, regarding existence of pure strategy Nash Equilibria and monotonicity of extremal equilibria. (see Amir [2] for a survey on supermodular games and Amir and Lazzati [3] for a static model of network effects, in which the game is supermodular).

Another important topic in network economics is the difference between direct and indirect network effects. Direct network effects arise from interacting with agents who consume the same goods. Indirect network effects arise when the consumption of non-interacting agents affects the quality or availability of the good. Indirect network effects also take place when the consumption of related goods has a positive effect on the utility of consumers. For instance, direct network effects occur when an agent uses a telephone and is able to reach a wider network of users. Indirect network effects occur when the phone device or the interconnection improved due to past usage and experience. Another type of network effects that have become of increasing interest for scholars occurs in the so-called two-sided markets. We refer to these effects as cross network effects. In markets where cross network effects are present, the utility of consumers on one side of the market is positively affected by the consumption on the other side of the market. This is the case of customers going to a shopping mall and having higher utility if more shops are available in this shopping mall or the case of dating sites, which become more

[6] Even if, it had long been recognized that only a dynamic setup could encompass all the features of network effects. See, for instance, Leibenstein [49], Arthur and Rusczcynski [5] and Hanson [41].

[7] See Colla and Garcia [14].

valuable for men when more women join the platform.[8] Literature in this field has flourished in the last years (see Armstrong [6] and Rochet and Tirole [60]), however, as in regular network effects, little attention has been given to the dynamics of two sided markets. Hence, we will not develop further this aspect in the present chapter.

In what follows we will survey the different types of dynamic models used to study network effects and summarize the main results regarding pricing under monopoly and competition. First we introduce precise definitions of network effects which are appropriate to study the phenomenon in a dynamic setting. These definitions can be applied to different types of network goods. Then we concentrate on the problem of dynamic pricing of network industries under different market structures and different models of product differentiation, namely monopoly, oligopoly and horizontal versus vertical product differentiation.

25.2 The Dynamic Nature of Network Effects

In this section, we distinguish between the different dynamic characterizations of network effects that have been considered in the literature. We will use the following notation: U represents the utility of a representative agent, $x^t_{i,j}$ represents the consumption of agent i for good j at time t, $D^t_j = \sum_i x^t_{i,j}$ represents the demand for good j at time t and m_i represents other factors which might influence consumer i's choice at time t. The function U, for agent i, has the following arguments at time t: $U^t_i(\{\mathbf{x}\}^t_{i,j}, m_i)$, where $\mathbf{x}$ represents the vector of agent's consumptions.[9,10]

25.2.1 Non Durable and Backwards Network Effects

The utility of each agent is affected by choices of agents who precede him (or her). There is no contemporaneous interaction effect, meaning that agents always regard past choices. Also, non durability implies that after one period, past choices become irrelevant for the utility of present agents. The lag in the network effect has been studied in different contexts, and it corresponds to goods for which there is some

[8] See Caillaud and Jullien [12] for a treatment of competing matchmakers; Rochet and Tirole [59] for a study of competition on the credit card market and Gabszewicz et al. [36] and Ferrando et al. [30] for the analysis of media markets.

[9] In this section, we focus on network effects which result from the consumption of incompatible goods, and we will disregard potential cross network effects. These definitions are easily extendable to encompass these situations. For an example see Garcia and Vergari [39] and Laussel and Resende [48].

[10] For simplicity, we assume throughout that the utility function is differentiable. The same effects could be expressed without resorting to differentiability.

learning period or *word of mouth* phenomenon.[11] We say that there are non durable and backward network effects in consumption of good j if

$$\frac{\partial U_i^t\left(\{\mathbf{x}\}_{i,j}^t, m_i\right)}{\partial D_j^{t-1}} > 0.$$

Papers that have considered this formulation are Bensaid and Lesne [7], Doganoglu [20], and Garcia and Resende [38]. These papers typically focus on the study of goods such as fashion and reputation goods. The utility derived from these goods is increasing in the number of users. Since consumers may not know the number of consumers who at the present acquire the good, they regard the immediate predecessors' choice, in order to make their consumption decisions. In fact we can interpret past consumption as a signal of the reputation of the good that consumers take into account when deciding what to acquire. This effect is non-cumulative since the number of consumers who was using the good far-off in the past can hardly influence or be informative of the current value of a fashion good or reputation good. Some examples of goods in which this type of effects arises quite often include clothes, shoes, restaurants, and touristic destinations.

25.2.2 Durable Backward Network Effects

These effects correspond to the situation in which the utility function of the agents depends on all past choices of the peers. Once again, it takes at least one period of time for the network to be constituted or for agents to enjoy the network effect and as such, there is no contemporaneous effect on the utility function. The cumulative effect can be interpreted in terms of goods whose quality increases with the total number of past users, such as in software industries or any industry where improvements can arise from the experience of past consumers.[12] These network effects have been studied by Arthur [4], Cabral et al. [11] and Gabszewicz and Garcia [32].[13] The utility function for cumulative network effects should have the following formulation.

$$\frac{\partial U_i^t(\{\mathbf{x}\}_{i,j}^t, m_i)}{\partial \sum_{k=1}^{t-1} D_j^k} > 0.$$

[11] See, for instance, Bensaid and Lesne [7], Gabszewicz and Garcia [33] and Garcia [38].

[12] We can include here the situation in which agents discount past users as it is less likely to interact with them or to benefit from their consumption.

[13] Also Markovich [53], Markovich and Moenius [52] analyse a model with backward cumulative network effects in the context of aftermarkets. The primary market in their case is hardware, whereas the secondary market is software. The utility that consumers derive from hardware depends on the availability of software. Dhebar and Oren [18], in their seminal article, analyse a cumulative network within continuous time model.

25.2.3 Contemporaneous Network Effects

When a certain good or service has contemporaneous network effects, this implies that consumers benefit from the present consumption of their peers, hence, the utility function has the following property:

$$\frac{\partial U_i^t(\{\mathbf{x}\}_{i,j}^t, m_i)}{\partial D_j^t} > 0.$$

As mentioned earlier, it is usual to consider that agents base their choices on the expected network effect that they will obtain from consumption. As such it is common to identify contemporaneous network effects with:

$$\frac{\partial U_i^t(\{\mathbf{x}\}_{i,j}^t, m_i)}{\partial \widetilde{D}_j^t} > 0,$$

where $\widetilde{D}_j^t$ is the expected demand of good j in period t. Seminal papers on networks focused mostly on this type of effects. To name a few papers that consider this type of static network we have: Katz and Shapiro [42], Economides [23], Grilo, Shy and Thisse [40].[14]

Up to this point, we considered the situation in which each agent obtains the network benefits upon joining the network and there are no additional payoffs in later periods. However, many network industries concern durable goods or services and agents enjoy benefits for longer periods. In those cases, the consumption choice depends also on the number of other agents which are expected to adopt the network. Here we distinguish between the situation in which agents only consider the future network effects and the situation in which both future, present and past consumption is taken into account in the choice of adoption. The former we designate by forward durable network effects, whereas the latter we designate by generalized network effect.

25.2.4 Forward Durable Network Effects

Once we move from the world where agents receive their payoffs upon joining the network and start considering that agents benefit from remaining in the network in the future (as for instance, when joining a club), we need to incorporate forward looking behavior in consumption choice. We now define U_i^t to be the intertemporal utility of consumer i, in the moment of choice t. In other words, U_i^t can be seen

[14] Mitchell and Skrzypacz [54] consider a setup where consumers regard contemporaneous and past adoption of the network.

as the stream of discounted benefits that the agent who decides to adopt in period t obtains throughout his permanence in the network. Algebraically, we say that there are forward network effects when

$$\frac{\partial U_i^t(\{\mathbf{x}\}_{i,j}^t, m_i)}{\partial \tilde{D}_{i,j}^k} > 0, k = t, ...\infty.$$

This network effect is used in the modelling of the so-called *aftermarkets*. In aftermarkets the utility of buying in the primary market (e.g. the utility of buying a printer) depends on the number of users of the secondary market (e.g. cartridges), because these affect availability, quality and price. Some papers that have assumed this network effect are Cabral [10], Laussel and Resende [48].

25.2.5 Generalized Network Effect

The generalized network effect refers to the situation in which the peers' choices, in all moments affect the utility of adopting the good. Algebraically we assume that:

$$\frac{\partial U_i^t(\{\mathbf{x}\}_{i,j}^t, m_i)}{\partial D_i^\tau} = \lambda_i^\tau > 0$$

This modelization has been used by Colla and Garcia [14] who were among the first to consider the importance of forward looking behavior in network choices. The different aspects of network effects here explained imply that firms undertake different strategies to maximize their profit. In fact, investment on network formation has been shown to occur through pricing and compatibility decisions. In what follows, we will explore the main results in dynamic pricing for the different types of network effects.

25.3 Strategic Behavior with Dynamic Network Effects

In the previous sections, we have underlined that in network industries, firms' market shares may endogenously change over time as a consequence of strategic complementarities among consumers. Even when firms do not behave strategically, network effects *per se* might be sufficient to engender a mechanical (non-strategic) effect that generates a reinforcement of firms' position in the market. Obviously, the nature and the magnitude of this non-strategic snowball effect can be affected by firms' strategies aiming to accelerate or dampen demands' interdependence entailed by network effects. For example, at the beginning of the product's life cycle, firms offering goods which generate network effects may be interested in over-investing in advertising or quality in order to make their products more attractive to future consumers. Similarly, firms may adjust their pricing strategies, trading off profits across periods, in order to make the maximum possible profit along the products' life cycle.

This section concentrates on the strand of literature dealing with pricing dynamics in network industries. Following the seminal works of Rohlfs [62]; Katz and Shapiro [42]; Farrell and Saloner [24], the early theoretical literature on pricing in network industries has mostly focused on firms' pricing strategies in a static context. More recently, the theoretical literature on dynamic pricing on network industries has flourished and a number of works have concentrated on the investigation of dynamic pricing strategies in network industries. This boost has been caused by the general acceptance of the idea that network effects constitute an intrinsically dynamic phenomenon together with some technical developments related to the widespread use of dynamic equilibrium concepts, such as Markov Perfect Equilibrium (see, for example, Maskin and Tirole [51]) and the extensive use of dynamic optimization tools (such as dynamic programming or optimal control).

So far, the main research questions addressed in this flourishing literature include: (a) the description of firms' optimal pricing strategies and the corresponding market shares' trajectories; (b) the analysis of the responsiveness of firms' prices to firms' installed base of customers in order to test in a dynamic setting whether larger firms are able to charge higher prices, obtaining a "network premium"; (c) the study of the persistence of firms' dominance in network industries; (d) the dynamic analysis of the welfare impact of firms' optimal pricing strategies in network industries.

In the following sections, we present the main results of the literature in relation to these four questions. First, we concentrate on the literature on dynamic pricing policies in monopoly industries, addressing how the monopolist's pricing strategies are affected by its incentives to trade-off profits across periods along the product's lifecycle. Afterwards, we deal with the recent literature on dynamic price competition in oligopoly markets, which studies how strategic competition affects optimal pricing paths of network goods.

25.3.1 Dynamic Pricing in Monopoly Markets with Network Effects

When a monopolist firm sells a good that generates network effects, there is an interdependence of the demands faced by the monopolist at each point of time. For example, in the case of backwards network effects (either durable or not), the future attractiveness of the network good depends on the size of the network at the present moment since early consumers allow the monopolist firm to boost the network benefit generated by its good in later periods. Accordingly, in markets with such characteristics, by lowering its current price, the monopolist is able to enhance both its present and future sales.

This type of dynamic channels of demand interdependence is based on the so-called introductory pricing strategies. In the context of the literature on network effects, the term introductory pricing strategies is used to designate pricing paths in which firms charge low prices at earlier periods, increasing their prices as time goes by. Introductory pricing strategies are very frequent on monopoly network industries but they also tend to emerge in many other contexts. This is for example the case of the so called reputation goods (see Rogerson [61]), or goods which require a learning period. Clarke et al. [13] also conclude that the optimal pricing path adopted by a monopolist participating in a market with experience effects in demand is increasing through time, corresponding to an introductory pricing strategy.

In the literature on dynamic pricing in monopoly industries with network effects, the optimal pricing strategies often correspond to introductory pricing strategies (see, for example, Dhebar and Oren [18, 19], Ackere and Reyniers [1], or, more recently, Gabszewicz and Garcia [32, 33]). Empirical evidence suggests that introductory pricing strategies are very frequent in network industries: Banks usually offer low rates for new clients, new software is usually offered at very low price (or, even, given for free), whereas updates are expensive; similarly, phone companies or network providers have low price deals for new customers (see Ackere and Reyniers [1]).

Dhebar and Oren [18, 19] build upon Rolhfs [62] to develop a continuous time model with infinite time horizon. In the context of such model, the authors investigate the optimal linear and non-linear path of prices adopted by a monopolist firm offering a network good. Their analysis reveals that the monopolist's equilibrium price-cost margins are increasing through time, confirming the idea that the dynamic channels of demand interaction in monopoly network industries may lead to introductory pricing strategies. Dhebar and Oren [18,19] have also studied the possibility of uncertainty regarding the size of the monopolist's network. In this case consumers are uncertain about the choices of their peers and they need to formulate expectations regarding the future network growth. The authors perform some comparative statics regarding the impact of the degree of optimism of consumers on the growth of the network. As expected they conclude that if consumers are more optimistic in relation to the network growth, the monopolist is able to charge higher prices.

Gabszewicz and Garcia [32,33] develop a discrete time model of monopoly provision of a good generating durable backward network effects. The authors are able to derive the closed form solution for the optimal path of prices for a monopolist operating during a finite number of periods. The results of Gabszewicz and Garcia [32,33] also reveal that introductory pricing strategies tend to arise in monopoly network industries. The authors show that pricing levels are increasing through time: as time evolves, the network of the monopolist becomes wider, which enables the monopolist firm to charge a higher price for the network good. This result is consistent with the idea of a network premium: as the monopolist's network expands, consumers' willingness to pay for the network good increases and the monopolist is able to charge a higher price. The authors also perform some comparative statics regarding the role played by the intensity of network effects or the length of the monopolist's time horizon. Namely, in the case of very low intensity of network effects and/or very wide time horizon length, Gabszewicz and Garcia [32, 33] show that the monopolist will offer the good at zero price in the first period, as an attempt to boost its network.

Cabral et al. [11] also studied introductory pricing strategies in monopoly network industries. They conclude that when consumers are “large”, there are multiple equilibria and it is possible to construct examples in which discounted prices rise over time, by selecting among these multiple equilibria. When consumers are “small”, Cabral et al. [11] showed that introductory pricing only arises in cases of incomplete information about demand or asymmetric information about costs.[15]

Fudenberg and Tirole [31] studied dynamic pricing strategies adopted by the monopolist supplier of a network good, when there is the threat of entry. They show that in the case of network industries, the installed base of users can constitute an effective mechanism to deter the entry of new rivals. Accordingly, the threat of entry may create incentives for the incumbent monopolist to charge lower prices in order to expand its network and prevent the entry of new competitors.

More recently, Driskill [21] has investigated the properties of the Markov Perfect Equilibrium arising in industries with dynamic network effects. Comparing the outcomes under monopoly and perfectly competitive supply of the network good, Driskill [21] argued that the level of monopoly output may be greater than the steady-state level of output with perfect competition. For that to be the case, it is necessary to have increasing marginal production costs and sufficiently intense network effects.[16]

25.3.2 Dynamic Pricing in Oligopoly Markets with Network Effects

The literature on dynamic pricing in oligopoly markets with network effects is a relatively recent literature and a number of exciting contributions are still under development. The dynamic analysis of strategic price competition in markets with network industries constitutes a complex economic problem since firms must take into consideration the dynamic interdependence of their demands across periods as well as the existence of multiple channels of strategic interaction among competing firms, who must account for both strategic competition within periods and strategic competition across periods.[17] The results on the nature of strategic competition in network industries are very scarce. Garcia and Resende [38] propose a simple

[15] According to Cabral et al. [11] a consumer is “large” if her purchase decision has a non-negligible effect on other buyers’ payoffs and decisions; and she is “small” if her purchase decision has no effect on the payoff to other buyers or on the monopolist’s pricing strategy.

[16] This result is in line with the result obtained by Driskill and McCafferty [22] considering the case of addictive goods.

[17] This aspect is also present in monopoly settings (see the previous section), often leading to introductory pricing strategies.

three-period model, in the context of which they conclude that both price strategic complementarity and strategic substitutability arise along the equilibrium path.

To the best of our knowledge, there are no results on the nature of strategic competition in network industries in more general settings than the one proposed by Garcia and Resende [38]. In a more general setting there would be even more dynamic channels of strategic interaction and probably the nature of strategic competition in network industries is expected to become more and more intricate.

In the light of the degree of complexity of the economic problem faced by firms operating in oligopoly markets with network effects, the computation of equilibrium price paths when firms strategically interact in markets with network effects also constitutes a difficult technical problem. Dynamic models of oligopoly interaction tend to be considerably complex and very often it is not even possible to explicitly obtain the optimal path of prices. In the literature, such issues have been overcame either by imposing additional structure on the theoretical model (see for example, Doganoglu [20], Laussel et al. [47], Mitchell and Skrzypacz [54], Laussel and Resende [48]) or by relying on numerical methods to unveil some predictions of the model (see for example Chen et al. [17], Markovich and Moenius [53] and Markovich [52]); or even a combination of both (as in Cabral [10]).

In the remainder of the section, we put forward the main results of the recent literature on the dynamics of price competition in oligopoly markets with network effects. To facilitate the exposition, we start with the literature on dynamic pricing competition in network industries with horizontal differentiation. Afterwards, we address the literature on dynamic strategic interaction in network industries with vertically differentiated goods.

25.3.2.1 Dynamic Price Competition in Network Industries: Horizontal Differentiation

This section deals with the literature on dynamic price competition in network industries with horizontal differentiation. In these industries, when all available goods generate similar network benefits and firms quote equal prices, we observe that consumers are not unanimous in relation to the good whose intrinsic characteristics match more closely their own tastes.

The strand of economics literature addressing this type of problems is relatively recent. Two significant exceptions are the works of Arthur and Rusczcynski [5] and Hanson [41]. These works studying dynamic pricing in industries with increasing returns on market shares generated by network effects. They study the impact of network effects on markets' equilibrium configuration, concluding that if discount rates are large enough, the initially dominant firm adopts a less aggressive pricing policy, losing its dominance. In contrast if discount rates are low enough, the large firm prefers to quote lower prices, reinforcing its dominance in the forthcoming periods.

At present, the study of the dynamics of price competition in network industries with horizontal differentiation has been attracting the increasing interest of economic scholars. In fact, a number of recent papers have developed dynamic games of price competition in network industries with horizontal differentiation. Most of these papers aim to characterize the optimal path of prices in network markets with horizontally differentiated products. To this end, these works concentrate on pricing paths corresponding to the Markov Perfect Equilibrium of the dynamic game under consideration.

In general, the dynamic models of network effects with horizontal differentiation correspond to discrete choice models, in which the utility obtained by consumer i when buying good j at moment t is given by:

$$U_{ij}^{t} = n_{ij}^{t} + u_{ij}^{t} - p_{j}^{t}, \tag{25.1}$$

where n_{ij}^{t} denotes the network benefit entailed by good j at period t (which may differ from consumer to consumer); p_{j}^{t} denotes the price of good j at moment t; and u_{ij}^{t} denotes the stand alone value of good j for consumer type i at moment t. When network goods are horizontally differentiated, even if $n_{ij}^{t} = n_{ik}^{t}$ and $p_{j}^{t} = p_{k}^{t}$, we observe that different consumers rank good k and good j differently.

Although the majority of papers dealing with dynamic competition in network industries with horizontal differentiation is based on the utility specification in (25.1), it is still not possible to talk of a unified body of literature. In fact, existing papers dealing with this issue differ along a number of aspects, such as methodological issues, the nature of network effects, or the timing of consumers' entry and exit in the market.

In relation to methodological differences we observe that some papers rely on analytical methods, imposing additional structure on the model in order to obtain closed form solutions for equilibrium pricing paths. For example, Doganoglu [20], Laussel et al. [47], Laussel and Resende [48] or Driskill [21] focus on Linear Markov Perfect Equilibrium pricing strategies. Under these assumptions, the authors obtain a linear quadratic differential game in which the equilibrium prices are assumed to be an affine function of the state variable (market share), enabling the authors to obtain the explicit expression of firms' optimal path of prices. Mitchell and Skrzypacz [54] are able to derive some general results without imposing linearity of pricing strategies. To illustrate these results, the authors present an example, in the context of which they also assume that equilibrium prices are affine functions of the state variable. The analytical models dealing with forward-looking agents, like Laussel et al. [47], Laussel and Resende [48] or Driskill [21] tend to impose even more structure to the model. Often, in these models it is assumed that agents' expectations about future market shares are also linear in the state variable.

A different strand of literature relies on numerical methods to characterize equilibrium price paths (see for example Cabral [10] that provides a combination of analytical results and numerical analysis). Although these papers often do not derive the closed-form solution for optimal pricing trajectories, they have the advantage of being less restrictive in relation to the structure of the model.

From a methodological point of view, it is also possible to distinguish between continuous time dynamic models (such as Laussel et al. [47], Laussel and Resende [48] or Driskill [21]) and discrete time dynamic models (such as Doganoglu [20]; Mitchell and Skrzypacz [54] or Cabral [10]). The former rely on the dynamic optimization toolset of optimal control theory, while the latter rely on dynamic programming techniques to derive equilibrium pricing strategies. In both cases, existing models of dynamic competition in network industries with horizontal differentiation consider an infinite horizon time setting. The simpler model proposed by Garcia and Resende [38] constitutes an exception to this trend.

Another aspect that distinguishes recent models of dynamic price competition in network industries with horizontal differentiation refers to the type of network effects considered in each paper. In other words, recent works have important differences in relation to the term n_{ij}^{t} in the utility specification presented in (25.1).

For example, Doganoglu [20] studies the dynamics of price competition when goods are horizontally differentiated à la Hotelling and network effects take the form of non-durable backward network effects. Mitchell and Skrzypacz [54] consider a model of horizontal differentiation à la Hotelling in which the network effect depends on the number of contemporaneous consumers buying a certain good as well as on the number of consumers who bought the good in the preceding period (in other words, Mitchell and Skrzypacz [54] consider a combination of non-durable backward network effects with contemporaneous network effects). Laussel et al. [47], Driskill [21] and Cabral [10] consider dynamic pricing competition in network industries with forward-looking agents. Cabral [10] proposes a model of dynamic competition in which consumers have a privately known preference for each network and durable forward network effects take place as consumers are forward-looking agents who care about firms' future market shares. Driskill [21] considers an overlapping generations model in which consumers have heterogeneous views on how their lifetime earnings are affected by the purchase of the good. Laussel et al. [47] propose a duopoly model with Hotelling differentiation and negative network effects arising in the form of durable forward network effects. Laussel and Resende [48] build upon Laussel et al. [47] to develop a model in which forward-looking consumers choose between two horizontally differentiated equipment whose value depends on the future availability of complementary goods and services (which in turn depends on firms' future equipment sales).

Finally, the papers under analysis also differ in relation to the consumers' entry and exit process. Doganoglu [20], Laussel et al. [47], Driskill [21] and Laussel and Resende [48] consider infinitely lived agents that face a constant (exogenous) probability of death. At each instant of time, entry rates and exit rates coincide so that the size of the market is stationary. Mitchell and Skrzypacz [54] consider the case of consumers that live for one period and then leave the market. The total size of the market is stationary as in each period the mass of consumers is exogenous and fixed at one. Cabral [10] considers that at each period of time a new consumer enters the market. Consumers live for infinitely many periods but they die with a constant hazard rate. The birth and death processes are stochastic.

In the light of the distinct modeling options considered in the papers under scrutiny in this section, it is not surprising that these papers put forward different predictions regarding the characteristics of optimal pricing strategies in oligopoly markets with horizontally differentiated network goods. This diversity of results reflects the early stage of this literature, which started to flourish very recently.

Doganoglu [20] concludes that firms with a larger installed user base tend to charge higher prices, benefiting from a network premium. In contrast, firms with a smaller installed user base tend to price more aggressively to compensate consumers for the lower network benefit generated by their product. In the light of such pricing strategies, the model predicts a symmetric steady state in which firms share the market evenly. Regarding the convergence to the symmetric steady state, Doganoglu [20] concludes that the convergence process may be significantly slow, especially when the oligopoly market exhibits strong network effects.

Laussel et al. [47] focus on dynamic price competition in a market with negative network effects. They obtain that the higher is firms' current market share the lower is their current price (i.e. larger firms offer a "congestion discount"). In line with Doganoglu [20], in the steady state equilibrium firms have symmetric market shares. However, the existence of negative consumption network effects is shown to soften price competition. Laussel et al. [47] also consider the possibility of entry, showing that the price of an entrant decreases gradually after entry, while the price of the incumbent firm increases. These results are substantially different from the ones obtained by Laussel and Resende [48]. The reduced form of the model suggested by Laussel and Resende [48] can be seen as a problem of dynamic price competition in markets with increasing returns on the size of firm's networks.[18] Due to this increasing returns effect, the results of Laussel and Resende [48] considerably depart from those in Doganoglu [20] or Laussel et al. [47]. In fact, the authors obtain that firms with a larger installed user base charge lower prices than the ones quoted by smaller firms, whose goods entail smaller network benefits. Even though larger firms tend to adopt more aggressive pricing policies in the context of the model by Laussel and Resende [48], they also face higher exit rates. Accordingly, when the conditions for the existence of a unique Linear Markov Perfect Equilibrium are met, the steady state properties are similar to the ones obtained by Doganoglu [20]: firms share the market evenly and network effects lead to lower steady state prices.

Driskill [21] also obtains that in imperfectly competitive industries with network effects, steady state prices tend to be lower. In particular, in the presence of strong positive dynamic network effects, the steady state price may be less than the marginal cost (a result that is also obtained by Laussel and Resende [48]).

As already mentioned, Driskill [21] also compares market outcomes under monopoly and oligopoly provision of a network good, concluding that the steady state industry profits may be lower with fewer firms in the industry.

[18] To be more precise, the utility specification of Laussel and Resende [48] departs from the standard linear separable version of the utility specification in (25.1), considering increasing returns with respect to the term $n^t_{i,j}$.

Mitchell and Skrzypacz [54] extend the previous literature by introducing the possibility of divergence in steady state market shares. When the discount factor is sufficiently low they conclude that market over-tightness keeps market shares from becoming too skewed. When network effects are very strong, there is always an equilibrium in which firms' market shares diverge. Mitchell and Skrzypacz [54] also allow their model to accommodate the possibility of quality differences in the available network goods. In line with the previous literature on quality improvement in network industries, they conclude that an inferior product may take the entire market due to network effects.

Cabral [10] relies on a combination of analytical and numerical methods to study dynamic pricing strategies in network industries with horizontal differentiation. In the case of a symmetric equilibrium, Cabral [10] shows that firms' network sizes are generally asymmetric since both the birth and death processes are stochastic. In general, a larger network is more likely to attract new consumers. Moreover, if network effects are sufficiently strong, the larger network reinforces its dominant position as time evolves (strong market dominance). An exception to this result occurs if the market share of the dominant firm is already close to 100%. As the model does not allow for tipping and eviction, when the market share of the dominant firm reaches 100%, the dominant firm tends to decrease its size.

25.3.2.2 Dynamic Price Competition in Network Industries: Vertical Differentiation

Vertical differentiation corresponds to the situation in which goods have different intrinsic qualities. Contrary to horizontal differentiation, other things alike, all consumers prefer the higher quality. Vertical differentiation, first introduced by Gabszewicz and Thisse [37], is widely used to model the choice between competing networks of different quality or different degrees of innovation. In particular they have been used to analyse predatory pricing, entry deterrence, standardization and lock in. In what follows, we survey the models of price competition with vertically differentiated firms, highlighting the features of the network modelization which drive the main results.

The problem of dynamic competition with quality differences has been studied extensively by several authors. Farrell and Saloner [25] study a dynamic model where consumers opt between a lower quality network good and a higher quality one. Consumers do not differ in their preference for the quality, only on the arrival time to the economy. The network effect is modelled as a forward durable network effect, following our specification above. Their main conclusions are that network effects may inhibit innovation if the new, high quality standard is incompatible with the old one. They point out the important strategic implications of this result. Namely, installed firms have interest in increasing their market shares in initial periods, and hence practice introductory pricing. As retaliation, innovating firms may engage in strategic preannouncement of their goods, so that consumers wait to buy the innovation rather than the old product. In their model, these features are not fully explored, because firms operate in a competitive industry. In an opposite side of the analysis, Katz and Shapiro [45] argue that there might exist rushing towards new network goods, rather than excess inertia. In their model, where forward durable network effects are assumed, the timing of introduction of the innovation and the pricing becomes endogenous and drives the aforementioned results. These papers are all concerned with the problem of quality differentiation and the introduction of new qualities in the presence of network effects. However, a common feature is that consumers are assumed to be homogeneous in the preference for quality. This implies that as soon as the new quality becomes available all consumers (or none) adopt it, depending on the magnitude of the network effect at the time of innovation. By contrast, a series of more recent papers assume that consumers are heterogeneous in their valuations of the quality of the good. Specifically, this heterogeneity allows that, at identical prices, some consumers prefer the low quality network good with an installed base, while others prefer the new technology without installed base. This series of papers includes Gabszewicz and Garcia [33], Chen et al. [17] and Driskill [21]. Gabszewicz and Garcia [33] conclude about the existence of predatory pricing when the quality differentiation between an incumbent and an entrant is high, compared to the network effect. In a two period model, they consider that the network effects are backward and that the high quality good is introduced only in the second period, lacking installed base. In their model, the high quality network good is incompatible with the good produced by the incumbent. Chen et al. [17] consider a model where firms avoid market dominance by their rivals through pricing and the choice of compatibility. They conclude that, when firms have similar installed bases, they choose to make their products compatible in order to expand the market. When firms have asymmetric installed bases, the larger firm has interest in rendering its product incompatible.[19] The authors obtain the result that strategic pricing precludes installed base differential from expanding to the point of incompatibility. Their results are obtained numerically and a sensitivity analysis is undertaken. Driskill [21] presents a section on dynamic price competition in a continuous time model with overlapping generations. This papers investigates the properties of Markov Perfect Equilibria that is the suitable equilibrium concept to study the model with network effects. Main results point to the existence of steady state prices less than marginal cost and disadvantageous market power, in the sense that industry profits may not be maximized under monopoly.

Finally, recent developments in the area of oligopoly competition with network effects have focused on indirect network effects as the ones provided for hardware users by software development. Markovich [52] and Markovich and Moenius [53] focus on this type of network effects. It is necessary to distinguish between direct and indirect network effects. As mentioned earlier, direct effects are related to the increase in the quality of the product due to an increment in the number of directly relatable users. Indirect effects entail broader benefits stemming from

[19] Garcia and Vergari [39] show that under some specific circumstances, it might be the case that firms with higher intrinsic quality may be interested in rendering their product compatible if network effects are very intense.

remote adoptions or from adoptions of related products. Even though direct interaction with remote consumers is not possible, the fact that a technology has had many previous users increases its value for the consumer: there is higher likelihood that the technology has less flaws, technical assistance might be prompter and more and better components might be available (see Liebowitz and Margolis [50] for a complete characterization of indirect network effects). In a setup where firms repeatedly invest in quality upgrades, compete in the product market and make exit and entry decisions, Markovich [52] studies the emergence of standardization and its persistence through time. The main conclusion is that, in general, excess inertia does not occur and innovation speed may drive standardization. In a related article, Markovich and Moenius [53] study the determinants of competition dynamics in markets with indirect network effects. They conclude that market structure is the main determinant of competition dynamics: a successful software developer raises the value of all firms who operate under the same platform, i.e. the system in which the software runs. In this paper, there is evidence for increasing competition across platforms, for different market structures. This contrasts to the tipping result in the literature, under which the market tips over for one platform, which becomes dominant.

25.4 Conclusion

The literature on network effects has grown extensively over the two last decades. The objective of this chapter is to survey recent contributions to the study of network effects in a dynamic setup. The main contribution is to formalize the different possible forms of network effects that correspond to different market situations. The theoretical literature on network effects has demonstrated that, in the presence of network effects, the standard economic theory might provide an inaccurate description/prediction of economic behaviour. In this context, the theoretical literature on network effects has been providing notable contributions, improving our understanding of a wide range of economic problems such as:

(a) The specifics of optimal price strategies in the presence of (simple) network effects (Rohlfs [62], Katz and Shapiro [42, 46]).
(b) The phenomenon of proprietary networks and the impact of compatibility between rival networks (Garcia and Vergari [39]), Farrell and Saloner [25]).
(c) The latent trade-off between quality provision and the intensity of network effects (Gabszewicz and Garcia [32]).
(d) The problem of standard wars and the trade-off between competition for the market and competition in the market.
(e) The impact of network effects on competition policy.

Despite the outstanding contributions already made available by the theoretical literature on network effects, this field is far from being exhausted and very challenging questions are still under investigation. Namely, issues related to the dynamic strategic interaction of multiple firms in network industries and in multiple sided markets.

Acknowledgements Filomena Garcia gratefully acknowledges financial support for this research from Fundação para a Ciência e a Tecnologia under project n. PTDC/ECO/65856/2006 and from UECE. Joana Resende acknowledges financial support from Cef.up, University of Porto and Fundação para a Ciência e Tecnologia.

References

1. Ackere, A., Reyniers, D.J.: Trade-ins and introductory offers in a monopoly. RAND J. Econ. **26**(1), 58–74 (1995)
2. Amir, R.: Supermodularity and complementarity in economics: An elementary survey. South. Econ. J. **71**, 636–660 (2005)
3. Amir, R., Lazzati, N.: Network effects, market structure and industry performance, mimeo (2009)
4. Arthur, B.: Competing technologies, increasing returns, and lock-in by historical events. Econ. J. **99**, 116–131 (1989)
5. Arthur, B., Rusczcynski, A.: Strategic pricing in markets with increasing returns, in Increasing returns and path dependence in the economy, The University of Michigan Press (1992)
6. Armstrong, M.: Competition in two-sided markets. RAND J. Econ. **37**(3), 668–691 (2006)
7. Bensaid, B., Lesne, J.-P.: Dynamic monopoly pricing with network externalities. Int. J. Ind. Organ. **14**(6), 837–855 (1996)
8. Bikhchandani, S., Hirshleifer, D., Welch, I.: A theory of fads, fashion, custom, and cultural change as informational cascades. J. Polit. Econ. **100**(5), 992–1026 (1992)
9. Bernheim, D.: A theory of conformity. J. Polit. Econ. **102**(5), 841–877 (1994)
10. Cabral, L.: Dynamic Price Competition with Network Effects, mimeo (2010)
11. Cabral, L, Salant, D., Woroch, G.: Monopoly pricing with network externalities. Int. J. Ind. Organ. **17**(2), 199–214 (1999)
12. Caillaud, B., Jullien, B.: Chicken and egg: competition among intermediation service providers. RAND J. Econ. **34**, 309–328 (2003)
13. Clarke, F., Darrough, M., Heineke, J.: Optimal pricing policy in the presence of experience effects. J. Bus. **55**(4), 517–530 (1982)
14. Colla, P., Garcia, F.: Technology Adoption With Forward Looking Agents", *mimeo* CORE Discussion Paper 2004/41 (2004)
15. Corneo, G., Jeanne, O.: Snobs, bandwagons, and the origin of social customs in consumer behavior. J. Econ. Behav. Organ. **32**, 333–347 (1997)
16. Corneo, G., Jeanne, O.: Conspicuous consumption, snobbism and conformism. J. Public Econ. **66**, 55–71 (1998)
17. Chen, J., Harrington, J., Doraszelski, U.: Avoiding market dominance: product compatibility in markets with network effects. RAND J. Econ. **49**(3), 455–485 (2009)
18. Dhebar, A., Oren, S.: Optimal dynamic pricing for expanding networks. Marketing Sci. **4**(4), 336–351 (1985)
19. Dhebar, A., Oren, S.: Dynamic nonlinear pricing in networks with interdependent demand. Oper. Res. **34**(3), 384–94 (1985a)
20. Doganoglu, T.: Dynamic price competition with consumption externalities. Netnomics **5**(1), 43–69 (2003)
21. Driskill, R.: Monopoly and Oligopoly Supply of a Good with Dynamic Network Externalities, mimeo (2007)
22. Driskill, R., McCafferty, S.: Monopoly and oligopoly provision of addictive goods. Int. Econ. Rev. **42**(1), 43–72 (2001)

23. Economides, N.: Network externalities, complementarities, and invitations to enter. Eur. J. Polit. Econ. **12**, 211–233 (1996)
24. Farrell, J., Saloner, G.: Standardization, compatibility, and innovation. RAND J. Econ. **16**, 70–83 (1985)
25. Farrell, J., Saloner, G.: Installed base and compatibility: innovation, product preannouncement, and predation. Am. Econ. Rev. **76**, 940–955 (1986)
26. Farrell, J., Saloner, G.: Standardization and variety. Econ. Lett. **20**, 71–74 (1986)
27. Farrell, J., Saloner, G.: Economic issues in standardization. In: Miller, J. (ed.) Telecommunications and Equity. North Holland, Amsterdam (1986)
28. Farrell, J., Saloner, G.: Competition, Compatibility and Standards: The Economics of Horses, Penguins, and Lemmings. In: Landis Gabel (ed.) Product Standardization and Competitive Strategy. North Holland (1987)
29. Farrell, J., Saloner, G.: Converters, compatibility, and the interfaces. J. Ind. Econ. **40**(1), 9–36 (1992)
30. Ferrando, J., Gabszewicz, J.J., Laussel, D., Sonnac, N.: Intermarket network effects and competition: An application to the media industry. Int. J. Econ. Theory **4**(3), 357–379 (2008)
31. Fudenberg, D., Tirole, J.: Pricing a network good to deter entry. J. Ind. Econ. **48**, 373–390 (2000)
32. Gabszewicz, J.J., Garcia, F.: A note on expanding networks and monopoly pricing. Econ. Lett. **98**(1), 9–15 (2008)
33. Gabszewicz, J.J., Garcia, F.: Optimal monopoly price paths with expanding networks. Rev. Network Econ. **6**(1), 42–49 (2007)
34. Gabszewicz, J.J., Garcia, F.: Intrinsic quality improvements and network externalities. Int. J. Econ. Theory **3**(4), 261–278 (2007a)
35. Gabszewicz, J.J., Garcia, F.: Quality improvements optimal monopoly price paths with expanding networks. Rev. Network Econ. **6**(1), 42–49 (2007b)
36. Gabszewicz, J.J., Laussel, D., Sonnac, N.: Press advertising and the ascent of the 'Pensée Unique'. Eur. Econ. Rev. **45**, 641–651 (2001)
37. Gabszewicz, J.J., Thisse, J.-F.: Price competition, quality, and income disparities. J. Econ. Theory **20**, 340–359 (1979)
38. Garcia, F., Resende, J.: Conformity based behavior and the dynamics of price competition: a new rational for fashion shifts, mimeo (2010)
39. Garcia, F., Vergari, C.: Compatibility Choice in Vertically Differentiated Technologies, CORE Discussion Papers (2008)
40. Grilo, I., Shy, O., Thisse, J.: Price competition when consumer behavior is characterized by conformity or vanity. J. Public Econ. **80**, 385–408 (2001)
41. Hanson, W.A.: Bandwagons and orphans: Dynamic pricing of competing technological systems subject to decreasing costs, mimeo Stanford University (1983)
42. Katz, M., Shapiro, C.: Network externalities, competition and compatibility. Am. Econ. Rev. **75**(3), 424–440 (1985)
43. Katz, M., Shapiro, C.: Technology adoption in the presence of network externalities. J. Polit. Econ. **94**, 822–841 (1986a)
44. Katz, M., Shapiro, C.: Product compatibility choice in a market with technological progress. Oxf. Econ. Pap. **38**, 146–165 (1986b)
45. Katz, M., Shapiro, C.: Product introduction with network externalities. J. Ind. Econ. **40**(1), 55–84 (1992)
46. Katz, M., Shapiro, C.: Systems competition and network externalities. J. Econ. Perspect. **8**(2), 93–115 (1994)
47. Laussel, D., Montmarin, M., Van Long, N.: Dynamic duopoly with congestion effects. Int. J. Ind. Organ. **22**(5), 655–677 (2004)
48. Laussel, D., Resende, J.: Does the absence of competition in the market foster competition for the market? A dynamic approach to aftermarkets, mimeo (2009)
49. Leibenstein, H.: Bandwagon, snob, and veblen effects in the theory of consumers' demand. Q. J. Econ. **64**(2), 183–207 (1950)

50. Liebowitz, S.J., Margolis, S.E.: Network externality: an uncommon tragedy. J. Econ. Perspect. **8**, 133–50 (1994)
51. Maskin, E., Tirole, J.: A theory of dynamic oligopoly, I: overview and quantity competition with large fixed costs. Econometrica **56**(3), 549–69 (1988)
52. Markovich, S.: Snowball: a dynamic oligopoly model with indirect network effects. J. Econ. Dyn. Control **32**, 909–938 (2008)
53. Markovich, S., Moenius, J.: Winning while losing: competition dynamics in the presence of indirect network effects. Int. J. Ind. Organ. (2007) (forthcoming)
54. Mitchell, M., Skrzypacz, A.: Network externalities and long-run market shares. Econ. Theory **29**(3), 621–648 (2006)
55. Meade, J.E.: Mr. Lerner on the economics of control. Econ. J. **55**(217), 51–56 (1945)
56. Pigou, A.C.: The interdependence of different sources of demand and supply in a market. Econ. J. **23**(89), 19–24 (1913)
57. Pigou, A.C.: The Economics of Welfare. Macmillan, New york (1929)
58. Rae, J.: Some New Principles on the Subject of Political Economy Exposing the Fallacies of Free Trade and Some Other Doctrines Maintained in the "Wealth of Nations". Hilliard, Gray and Co, Boston (1834)
59. Rochet, J.-C., Tirole, J.: Platform competition in two-sided markets. J. Eur. Econ. Assoc. **1**, 990–1029 (2003)
60. Rochet, J.-C., Tirole, J.: Two-sided markets: a progress report. RAND J. Econ. **37**, 645–667 (2006)
61. Rogerson, W.: Reputation and product quality. Bell J. Econ. **14**(2), 508–516 (1983)
62. Rohlfs, J.: A theory of interdependent demand for a communication service. Bell J. Econ. **5**, 16–37 (1974)
63. Veblen, T.: The Theory of the Leisure Class. Macmillan, New York (1899)

Chapter 26
Exit Times and Persistence of Solitons for a Stochastic Korteweg–de Vries Equation

Eric Gautier

Abstract The Korteweg–de Vries equation is a model of nonlinear shallow water long waves of small amplitude that admit soliton solutions. Solitons are a family of solutions which are progressive localized waves that propagate with constant speed and shape. These waves are stable in many ways against perturbations or interactions. We consider random perturbations by an additive noise of small amplitude. It is common in Physics to approximate the solution in the presence of noise, corresponding to an initial datum generating a soliton in the deterministic system, by a randomly modulated soliton (the parameters of the soliton fluctuate randomly). The validity of such an approximation has been proved by A. de Bouard and A. Debussche. We present results obtained in a joint work with A. de Bouard where we study in more details the exit time from a neighborhood of the soliton and randomly modulated soliton and obtain the scaling in terms of the amplitude of the noise for each approximation. This allows to quantify the gain of an approximation of the form of a randomly modulated soliton in describing the persistence of solitons.

26.1 Introduction

The Kordeweg–de Vries (KdV) equation

$$\partial_t u + \partial_x^3 u + \partial_x(u^2) = 0, \ x \in \mathbb{R}, \ t \in \mathbb{R}^+ \tag{26.1}$$

is a model for nonlinear, shallow water (and many other systems), unidirectional long waves of small amplitude. u is proportional to the relative vertical displacement of the fluid and x and t are dimensionless distance and time, where distance is measured in a moving frame $x = x' - t$ if x' is the original coordinate. Due to two

E. Gautier
Ecole Nationale de la Statistique et de l'Administration Economique – CREST, 3 avenue Pierre Larousse, 92245 Malakoff cedex, France
e-mail: gautier@ensae.fr

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_26,

opposite effects: dispersion and nonlinearity, it has soliton solutions of the form

$$\varphi_c(x - ct + x_0) \tag{26.2}$$

where c is the velocity, x_0 is the initial position and

$$\varphi_c(x) = \frac{3c}{2\cosh\left(\sqrt{c}x/2\right)}. \tag{26.3}$$

Equation (26.1) has an infinite number of invariant quantities. Two are important for the study of the stability of solitons, they are:

1. The mass

$$\mathbf{M}(u^{u_0}(t)) = \frac{1}{2}\int_{\mathbb{R}} (u^{u_0})^2(t,x)dx = \mathbf{M}(u_0),\ \forall t \geq 0,$$

2. The Hamiltonian

$$\mathbf{H}(u^{u_0}(t)) = \frac{1}{2}\int_{\mathbb{R}} (\partial_x(u^{u_0}))^2(t,x)dx - \frac{1}{3}\int_{\mathbb{R}} (u^{u_0})^3(t,x)dx = \mathbf{H}(u_0),\ \forall t \geq 0$$

where u^{u_0} is the solution of (26.1) with initial datum u_0. Indeed φ_c is a critical point of $Q_c = \mathbf{H} + c\mathbf{M}$. Global well-posedness in the space H^1 where the two quantities are defined is proved in [7]. Solitons are considered as highly stable states of motions and several aspects of stability have been studied. For example these waves are stable with respect to perturbation of the initial datum when it is a soliton profile (of the form (26.2) for $t = 0$). However note that an initial very small change in c implies propagation at different speed and thus for large times a notable difference in position. A first notion introduced in [1] is that of orbital stability where for $\epsilon > 0$ there exists $\delta > 0$ such that

$$\|u_0 - \varphi_c\|_{\mathrm{H}^1} \leq \delta \Rightarrow \forall t \geq 0,\ d(u^{u_0}(t,\cdot), \varphi_c(\cdot - ct)) \leq \epsilon$$

with $d(u,v) = \inf\{\|u(\cdot) - v(\cdot - s)\|_{\mathrm{H}^1},\ s \in \mathbb{R}\}$. A stronger notion of asymptotic stability has been obtained in [8, 9]. However the convergence to a function of the form (26.2) is only obtained in a weak topology and it is not expected to have strong convergence.

When random pressure acts at the surface of the fluid a random perturbation of (26.1) could be considered

$$du + \left(\partial_x^3 u + \partial_x(u^2)\right)dt = \epsilon dW(t) \tag{26.4}$$

where $W(t)$ is a Wiener process in H^1 (i.e. of the form $W(t) = \sum_{i\in\mathbb{N}} \beta_i(t)\Phi e_i$ where $(e_i)_{i\in\mathbb{N}}$ is a complete othonormal system of L^2 and Φ is Hilbert-Schmidt from L^2 to H^1, we write later $\Phi \in \mathscr{L}_2^{0,1}$). Mild solutions to (26.4) are solutions of the following integral equation

$$u^{\epsilon,u_0}(t) = S(t)u_0 - \int_0^t S(t-s)\partial_x((u^{\epsilon,u_0})^2)ds + \epsilon \int_0^t S(t-s)dW(s) \quad (26.5)$$

where $(S(t))_{t\in\mathbb{R}}$ is the Airy group on H^1 associated to the unbounded operator $(-\partial_x^3, \mathrm{H}^3)$. The following theorem is proved in [2].

Theorem 26.1. *There exists a mild solution to* (26.4)*, a.s. continuous in time with values in* H^1 *defined for all* $t \geq 0$*. For any* $T > 0$*, the solution is unique among those having paths in some space* $X_T \subset \mathrm{C}([0,T];\mathrm{H}^1)$.

Note that the above result is obtained for a noise which is colored in space. However, though we cannot give a mathematical justification to it since $S(t)$ has only local smoothing properties, the space-time white noise is often considered in physics. In the following we consider a sequence of equations corresponding to a sequence of noises such that

$$\Phi_n = \left(I - \Delta + \frac{1}{n}(x^2 I - \Delta)^k\right)^{-1/2}$$

They are Hilbert–Schmidt for k large enough and uniformly bounded as operators from L^2 to H^1 (we write later their norm as $\|\Phi_n\|_{\mathscr{L}_2^{0,1}}$, they are less than 1 for every n). This allows to consider noises such that the larger is n the more the noise mimics the space-time white noise and is a central assumption to allow to obtain lower bounds on the probabilities involving exit times. We denote the solutions of mild solutions of (26.4) by $u^{n,\epsilon,\varphi_{c_0}}$. The aim of the two next sections is to study the stability of the shape of the soliton in the presence of noise and when the initial datum is a soliton profile.

26.2 Exit from the Vicinity of the Soliton

In this section we study the exit from a neighborhood of the soliton (the solution of the deterministic equation).

Definition 26.1. We define the exit time off a neighborhood of the soliton by

$$\tilde{\tau}_\alpha^{n,\epsilon} = \inf\left\{t \in [0,\infty) : \left\|u^{n,\epsilon,\varphi_{c_0}}(t,\cdot + c_0 t) - \varphi_{c_0}(\cdot)\right\|_{\mathrm{H}^1} \geq \alpha\right\}.$$

We expect this approach to yield a poor description of the persistence of the soliton in the presence of noise for the same argument that motivated the introduction of the notion of orbital stability. Also the following heuristic is given in [3]. The operator arising from the linearization of $u \to -\partial_x^3 u - \partial_x(u^2)$ around φ_{c_0} is $\partial_x L_{c_0}$ where $L_{c_0} = Q''_{c_0}(\varphi_{c_0}) = -\partial_x^2 + c_0 - 2\varphi_{c_0}$. It has no unstable eigenvalue but a null space: $\mathrm{span}\left\{\partial_c\varphi_c|_{c=c_0}, \partial_x\varphi_{c_0}\right\}$. Thus if we formally replace the infinite dimensional system by a linear system of SDEs such that 0 is a degenerate simple eigenvalue corresponding to a Jordan block:

$$\begin{cases} dX_1 = X_2 dt + \epsilon dW_1(t) \\ dX_2 = \epsilon dW_2(t) \end{cases}$$

where W_1 and W_2 are independent Brownian motions, then $X_1(T) = \epsilon \int_0^T W_2(s) ds + \epsilon W_1(T)$ thus $\mathrm{var}(X_1(T)) \propto \epsilon^2 T^3$ for large T. Such an analogy suggests that the solution stays in the neighborhood of the soliton for times of the order of $\epsilon^{-2/3}$. Using a large deviations result as well as studying the associated variational problem we are able to justify this heuristic in [4].

Proposition 26.1. *Take T, $c_0 > 0$, then for $\alpha_0 > 0$, $\forall 0 < \alpha < \alpha_0$, $\exists C(\alpha, c_0)$ such that*

$$\underline{\lim}_{n\to\infty}\underline{\lim}_{\epsilon\to 0}\epsilon^2 T^3 \log \mathbb{P}\left(\tilde{\tau}_\alpha^{n,\epsilon} \le T\right) \ge -C(\alpha, c_0).$$

It means that such an approximation by the soliton is valid up to times at most of the order of $\epsilon^{-2/3}$. Note that this order is the same as the one obtained in [5] for the tails of the position of the soliton in the stochastic nonlinear Schrödinger (NLS) equation. We do not supplement this lower bound by an upper bound as we are after a better description of the persistence of solitons. We expect that by allowing the position (or more parameters) of the soliton to fluctuate we might obtain stability of the shape of the soliton on a diffusive time scale (see, e.g. [5, 6, 10]).

26.3 Exit from the Vicinity of the Modulated Soliton

The following result from [3] justifies the collective coordinate approach often used in physics to approximate the solution of the stochastic KdV (or NLS) equation starting from a soliton profile.

Theorem 26.2. $\exists \alpha_0 > 0 : \forall \alpha \in (0, \alpha_0]$, $\exists \tau_\alpha^\epsilon > 0$ *a.s. stopping time,* $\exists c^\epsilon(t), x^\epsilon(t)$ *semi-martingales defined a.s. for $t \le \tau_\alpha^\epsilon$ with values in* $(0, \infty)$ *and* $\mathbb{R}$ *such that a.s.* $\forall t \le \tau_\alpha^\epsilon$,

$$u^{\epsilon, u_0}(t, \cdot + x^\epsilon(t)) - \varphi_{c^\epsilon(t)} = \epsilon \eta^\epsilon(t),$$

$$\int_{\mathbb{R}} \eta^\epsilon(t,x)\varphi_{c_0}(x)dx = \int_{\mathbb{R}} \eta^\epsilon(t,x)\partial_x\varphi_{c_0}(x)dx = 0, \tag{26.6}$$
$$\|\epsilon\eta^\epsilon(t)\|_{\mathrm{H}^1} \le \alpha, \quad |c^\epsilon(t) - c_0| \le \alpha.$$

Moreover, $\exists C > 0 : \forall T > 0$, $\forall \alpha \le \alpha_0$, $\exists \epsilon_0 > 0 : \forall \epsilon < \epsilon_0$,

$$\mathbb{P}\left(\tau_\alpha^\epsilon \le T\right) \le \frac{C\epsilon^2 T \|\Phi\|_{\mathscr{L}_2^{0,1}}}{\alpha^4}. \tag{26.7}$$

Equations for c^ϵ, x^ϵ and the first order of η^ϵ are also derived therein.

In [4] we are able to improve on the upper bound (26.7) and to obtain an exponential upper bound in the case where Φ is fixed and a uniform such upper bound for the sequence of noises defined in terms of the sequence of operators Φ_n.

Proposition 26.2. *For $T > 0$ and $0 < \alpha < \alpha_0$ and n fixed, $\exists C(\alpha, c_0)$ and $\epsilon_0 > 0$ such that $\epsilon_0^2 T$ is small enough (depending on $\|\Phi_n\|_{\mathscr{L}_2^{0,1}}$ and α), such that $\forall \epsilon < \epsilon_0$,*

$$\mathbb{P}\left(\tau_\alpha^{n,\epsilon} \le T\right) \le \exp\left(-\frac{C(\alpha, c_0)}{\epsilon^2 T}\right). \tag{26.8}$$

We also provide an exponential lower bound, with a similar technique as for the proof of Proposition 26.1, with the same scaling in ϵ and T

Proposition 26.3. *For $T, \alpha > 0$, $\exists C(\alpha, c_0)$:*

$$\underline{\lim}_{n\to\infty}\underline{\lim}_{\epsilon\to 0}\epsilon^2 T \log \mathbb{P}\left(\tau_\alpha^{n,\epsilon} \le T\right) \ge -C(\alpha, c_0).$$

References

1. Benjamin, T.B.: The stability of solitary waves. Proc. Roy. Soc. Lond. A **328**, 153–183 (1972)
2. de Bouard, A., Debussche, A.: On the stochastic Korteweg–de Vries equation. J. Funct. Anal. **154**, 215–251 (1998)
3. de Bouard, A., Debussche, A.: Random modulation of solitons for the stochastic Korteweg-de Vries equation. Ann. Inst. H. Poincaré Anal. Non Linéaire. **24**, 251–278 (2007)
4. de Bouard, A., Gautier, E.: Exit problems related to the persistence of solitons for the Korteweg-de Vries equation with small noise. Discrete Contin. Dyn. Syst. **26**, 857–871 (2010)
5. Debussche, A., Gautier, E.: Small noise asymptotic of the timing jitter in soliton transmission. Ann. Appl. Probab. **18**, 178–208 (2008)
6. Garnier, J.: Long-time dynamics of Korteweg-de Vries solitons driven by random perturbations. J. Stat. Phys. **105**, 789–833 (2001)
7. Kenig, C.E., Ponce, G., Vega, L.: Well-posedness and scattering results for the generalized Korteweg–de Vries equation via the contraction principle. Comm. Pure Appl. Math. **46**, 527-620 (1993)
8. Martel, Y., Merle, F.: Asymptotic stability of solitons for subcritical generalized KdV equation. Arch. Ration. Mech. Anal. **157**, 219–254 (2001)
9. Pego, R.L., Weinstein, M.I.: Asymptotic stability of solitary waves. Comm. Math. Phys. **164**, 305–349 (1994)
10. Wadati., M.: Stochastic Korteweg–de Vries equations. J. Phys. Soc. Jpn. **52**, 2642–2648 (1983)

Chapter 27
Optimization Approach to a Simulation Algorithm of a Mixer-Settler System in the Transient State

E.F. Gomes and G.A. Pinto

Abstract In this paper we describe a parameter optimization approach to a simulation algorithm of a mixer-settler system in the transient state. The model we are using for the shallow-layer settler, in a mixer-settler system, is able to describe the hydrodynamic phenomena of the transient state of a liquid–liquid system. Its mathematical model includes parameters of the drop transport process as well as of the drop–drop and drop-interface coalescence with the active interface. The most adequate values of these parameters are unknown. In order to tune the model parameters we have linked the mixer-settler simulation algorithm to an optimization procedure. We have used the Hooke–Jeeves optimization algorithm to fit these parameters to given experimental results.

27.1 Introduction

Mixer-settler equipments are extremely useful for developing liquid-liquid extraction processes. Liquid–liquid extraction, also called solvent extraction, is a process that allows the separation of two or more components due to their unequal solubilities in two immiscible liquid phases. The liquid–liquid dispersion is created during the mixing step (in the mixer unit) and is separated by gravity (in the settler unit) in a second step. The mixing and separation steps constitute one stage of extraction. The importance of separation of immiscible liquid–liquid systems is well known in many industrial fields, such as wastewater treatment and the crude oil industry [9]. Due to the high complexity and cost of the direct experimentation using such equipments, computer simulation becomes very attractive.

In previous work [6] we have proposed a model for the shallow-layer settler unit, in a mixer-settler system, which is able to describe the hydrodynamic phenomena of

E.F. Gomes (✉) and G.A. Pinto
Instituto Superior de Engenharia do Porto, Rua Dr. António Benardino de Almeida, 431, Porto, Portugal
e-mail: efg@isep.ipp.pt, gap@isep.ipp.pt

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_27,

the transient state of a liquid–liquid system. The mathematical model used includes parameters of the drop transport process as well as of the drop–drop and drop-interface coalescence with the active interface. In order to find the most appropriate values for the model parameters, we have coupled the mixer-settler simulation algorithm with an optimization algorithm. A direct numerical resolution technique had already been proposed by the authors [4] for the simulation of liquid–liquid systems. The underlying mathematical model is also described in this work. Previous computational results indicate that the model provides qualitative predictions of the settler's dynamic behavior. In particular, the length of the band changes plausibly when input parameters are affected by step changes. Moreover, the system converges to steady state.

In order to tune these parameter values of the mixer-settler model we have used the Hooke–Jeeves algorithm [8]. This algorithm uses a deterministic pattern search method which is widely used for non-smooth objective functions due to its simplicity and robustness. The objective function (to be minimized) is defined as the sum of squares of the differences between the computed and given target values for the thickness of the dispersion band. Such target values are being determined in experiments conducted in our liquid–liquid systems laboratory.

27.2 The Mathematical Model

In previous work [4, 6], we have already proposed a mathematical model of the transient state for the dispersion band in the settler. This mathematical approach models the phenomena that occur in each volume element $(H \times w \times \Delta x)$, or vertical slice, of the dispersion band, in the settler, where x is its longitudinal position in the dispersion band, H is the thickness of the dispersion band and w is the width of the settler (Fig. 27.1). The modeling of the longitudinal drop transport is set-up by taking into account the gravitational instability caused in the dispersion band by the non-uniform thickness of the band. The viscous character of the draining of the continuous phase is also taken into account in this model.

This model for the shallow layer gravity settler assumes there is a large enough horizontal area, such that, given the inflow at one end and the outflows at the other, a dispersion wedge is formed by settling effect. This wedge does not cover the entire phase boundary between the two liquids. In this wedge, we assume:

- Binary drop-drop coalescence.
- Negligible drop breakage due to low turbulence within the dispersion band.
- Drops-interface coalescence.
- Gravitational drainage of the continuous phase to the passive interface.
- Drop movement away from the entrance essentially in plug flow movement.
- Negligible mass transfer between phases due to the low interaction between drops, the low specific surface and the short residence time of both phases in the dispersion band.

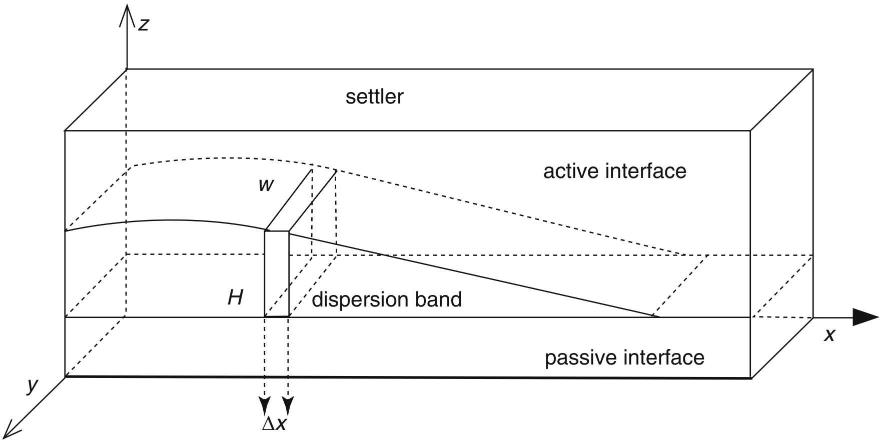


Fig. 27.1 Schematics of the dispersion band

- Good representation of drop size distribution at any position at the slice by a continuous distribution of drop volumes.
- Uniform longitudinal drop velocity in each slice (there are no wall effects).
- Negligible changes of the physical properties of the dispersion band in the vertical direction.

The calculation of the drop volume variation in time is performed for each volume element. For each time step, the events that occur in each volume element of the dispersion band are computed. They are (see (27.1)):

- The dispersion fed from the mixer unit into the settler.
- The coalescence between drops of the dispersed phase.
- The coalescence of drops with the active interface.
- The draining of the continuous phase to the passive interface.
- The transport of the dispersion between consecutive volume elements in the dispersion band.

The dispersion flows between consecutive slices from the thicker to the thinner slice. The following differential equation describes these phenomena in terms of the volume variation of the dispersion for each slice at the position x over time

$$\begin{aligned}\frac{\delta V(x,t)}{\delta t} &= F(x,t)w\Delta x - \frac{\delta C(x,t)}{\delta t} - \frac{\delta D(x,t)}{\delta t}\\ &\quad +\alpha w\Delta x\left[\frac{H(x-\Delta x,t)-H(x,t)}{\Delta x} - \frac{H(x,t)-H(x+\Delta x,t)}{\Delta x}\right]\end{aligned} \tag{27.1}$$

where w is the width of the settler, α is an unknown parameter describing the effects of inner friction on the longitudinal movement of the dispersion, F is the velocity of the dispersion entering the settler from the mixer, H is the thickness of the dispersion band at position x and time t and $\delta C(x,t)/\delta t$ represents the volume rate of the drops coalescing to the active interface, which, in her pioneer work, Ruiz [12] expresses as Equation (27.2). In that work, Ruiz defined $n(v,x)dv$ as the number of drops of size from v to $v + dv$ (volume class v) per unit volume of the dispersion band at position x, as the fraction of dispersion phase projected onto the surface and assumed close to 1, as the average projected area of drops and $N(x)$ as the total number of drops per unit volume at position x. $\lambda^*(v)$ is the drop-interface coalescence frequency which another unknown parameter. We use this expression yet in our work:

$$\begin{aligned}\frac{\delta C(x,t)}{\delta t} &= \frac{w\eta^*(x)\lambda^*(v)n(v,x)dv}{A_p(x)N(x)}\Delta x dv \\ &= \frac{w\Delta x\eta^*(x)\lambda^*(v)n(v,x)dv}{\frac{\pi}{4}\left(\frac{6}{\pi}\right)^{\frac{2}{3}}\int_0^\infty v^{\frac{2}{3}}n(v,x)\,dv}\end{aligned} \tag{27.2}$$

Where $\delta D(x,t)/\delta t$ represents the volume variation of the dispersion due to the draining to the passive interface. To describe this variation, we use

$$\frac{\delta D(x,t)}{\delta t} = kV(x,t)\left[(1-\eta(x,t)) - (1-\eta_M)\right] \tag{27.3}$$

In (27.3) we have another unknown constant, k, which describes the effects of the friction between the continuous and the dispersed phases. V is the volume of the dispersion in the slice at time t, η is the present local hold up and η_M is the maximum hold up (corresponds to the maximal compactation of the drops of the dispersion). Using the expression in (27.4), we can calculate the volume of the slice at position x and we can obtain the thickness of the dispersion band, H, along the dispersion band.

$$V(x,t) = H(x,t)w\Delta x \tag{27.4}$$

The hold-up, η, is calculated directly as the ratio of the volume of dispersed phase to the total volume of the dispersion. Further, besides the volume of the dispersed phase, we need to know the volume of the continuous phase. We decompose the expression of the volume variation into two terms, one describing the volume variation of dispersed phase, $V_d(x,t)$, as can be seen in (27.5), and the other describing the variation of the continuous phase volume, $V_c(x,t)$, as can be seen in (27.6).

$$\begin{aligned}&\frac{\delta V_d(x,t)}{\delta t} = F_d(x,t)w\Delta x \\ &+\alpha w\Delta x\eta(x,t)\left[\frac{H(x-\Delta x,t)-H(x,t)}{\Delta x} - \frac{H(x,t)-H(x+\Delta x,t)}{\Delta x}\right] \\ &-\frac{\delta C(x,t)}{\delta t}\end{aligned} \tag{27.5}$$

$$\frac{\delta V_c(x,t)}{\delta t} = F_c(x,t)w\Delta x$$
$$+\alpha w\Delta x(1-\eta(x,t))\left[\frac{H(x-\Delta x,t)-H(x,t)}{\Delta x} - \frac{H(x,t)-H(x+\Delta x,t)}{\Delta x}\right]$$
$$-\frac{\delta D(x,t)}{\delta t} \tag{27.6}$$

As we have already referred, we assume that breakage, given the low turbulence, is negligible. However, we consider two types of coalescence, drop–drop coalescence and drop-interface coalescence. To represent the volume of drops coalescing to the active interface per unit time we use (27.2). To represent the volume rate of the drops disappearing and appearing by drop–drop coalescence within the element of dispersion volume, we use (27.7) and (27.8) respectively.

$$\frac{wH(x)\Delta x}{N(x)}\left[\int_0^\infty \lambda(v,v')n(v,x)n(v',x)\,dv'\right]dv$$
$$= \frac{wH(x)\Delta x}{\int_0^\infty n(v,x)\,dv}\left[\int_0^\infty \lambda(v,v')n(v,x)n(v',x)\,dv'\right]dv \tag{27.7}$$

$$\frac{wH(x)\Delta x}{2N(x)}\left[\int_0^v \lambda(v-v',v')n(v-v',x)n(v',x)\,dv'\right]dv$$
$$= \frac{wH(x)\Delta x}{2\int_0^\infty n(v,x)\,dv}\left[\int_0^v \lambda(v-v',v')n(v-v',x)n(v',x)\,dv'\right]dv \tag{27.8}$$

According to Ruiz [12], $\lambda(v,v')$ in (27.7) and (27.8) represents the drop–drop coalescence frequency (27.9). λ_0 is a constant of the equation, and it is another unknown parameter.

$$\lambda(v,v') = \lambda_0(v^{-1/3} + v'^{-1/3})^2 \tag{27.9}$$

27.3 The Numerical Approach

In previous works [4–6] the authors have used a direct numerical approach by means of an adequate space-time discretization. In that approach, the discretized form of the population balance equation is solved by a first-order finite difference method with careful control of variable, time and space integration steps. The phase space coordinates are the position of the slice in the dispersion band and the size (volume) of the drops. For each time step we compute the dispersion changes in each volume element (slice) of the dispersion band, taking into account the variations of the volume of the continuous phase and of the properties of the dispersed phase due to the acting discrete phenomena. In this algorithm, the variation of the volume of the dispersed phase and the volume of the continuous phase are calculated separately. The continuous phase in each slice of the dispersion band is analyzed from the point of view of volume variation alone and is calculated from the balance of the volume of the continuous phase. The dispersed phase in each slice, at each time step, is analyzed from both the volume variation and drop size composition points of view. For greater precision of drop size representation, we adopted a logarithmic grid of volumes (of drops), v_k, in order to simultaneously obtain enough information about the smaller drop sizes and to achieve a reasonable calculation time. Like Ribeiro [11] we used the maximum drop volume, v_{max}, and minimum drop volume, v_{min}, observed in the dispersion (depending on physical–chemical characteristics of the liquid–liquid system) and the number of drop classes allowed must be selected. Equation (27.10) gives the drop volume grid.

$$v_k = v_{min} \left(\frac{v_{max}}{v_{min}} \right)^{\left(\frac{j-1}{nclasses-1} \right)} \quad j = 1, 2, \ldots, nclasses \tag{27.10}$$

Knowing the variation of the volume of the dispersed and continuous phases, we can obtain the total volume by using a numerical method of ordinary differential equations such as a careful implementation of the Euler method. The calculation of the hold up, η, and of the thickness of the dispersion band, H, along the settler can be obtained from the definitions. From the volume variation of the two phases, we know the total variation of the volume of the slice at the position x of the dispersion band, and we may calculate the hold up and the thickness of the dispersion band at position x. Since a direct numerical method of resolution of ordinary differential equations is used, the choice of the space (length) and time integration step, as well as of the convergence criterion, is very important for the successful implementation of the algorithm [5]. At each time integration step, a convergence criterion is used for the space discretization of the ordinary differential equation to evaluate each solution.

27.4 Strategies for Parameter Optimization

Given the complexity of the model, the values for the newly introduced parameters of the drop transport process and coalescence need careful study. In order to identify the most adequate parameter values for the model we use our simulation algorithm for the mixer-settler unit as a sub-routine in the Hooke and Jeeves direct search nonlinear optimization algorithm. It is a simple method that does not require derivatives of the objective function. Therefore, the method has the advantage that the objective function is not required to be continuous nor differentiable. In our case, the objective function is defined as the sum of squares of the differences between the computed and given target results for the thickness of the dispersion band, H, each centimeter (100 data points), along the dispersion band. The program, written in C, needs a starting guess for the parameter values. With this set of parameters, the simulation program runs the transient state until a steady state is reached. At this point, the optimization program evaluates the objective function. Next, it tries to minimize the objective function by finding a new set of parameters and running the simulation again. The optimization algorithm stops when the value of the objective function is below a given threshold or a maximum number of iteration is reached.

27.5 Results

We have run the optimization program for estimating the best values for the four unknown parameters α, k, λ^*, λ_0. In order to simulate the mixer-settler for the stationary and transient state it is very important to know the hydrodynamic phenomena in the mixer and settler units. In order to obtain such knowledge, sophisticated experiments are required to define the parameters and validate the simulation algorithms developed. We have a purpose-built acrylic mixer-settler system at our liquid–liquid systems laboratory at ISEP (Instituto Superior de Engenharia do Porto) where experiments have been conducted during the last few years. Another important tool in the modeling and validation of the simulation of the hydrodynamics and mass transfer in liquid–liquid systems is the calculation of the variation of the diameter of the drops along the settler. We have been working in image processing for the automatic counting of the drops in the image frames obtained at our lab, minimizing the errors for major reliability of the experimental results [1, 2]. Since the aim of the work presented in this paper is mainly to show that the optimization program can fit the unknown parameters to given target data, in this section we show the results obtained with the optimization program for a given set of data. The simulated liquid–liquid system was composed by an organic phase (kerosene) and aqueous phase (water) with an organic flow rate of 2.96 l/min of both phases and a hold up of 0.5. The stirring speed in the mixer is constant at 200 rpm.

Figure 27.2 shows the evolution of the shape of the dispersion band, as the optimization algorithm searches for better values of the parameters. The starting value for each one of the parameters is 0.01, 0.03, 0.1 and 0.0001, respectively. At this point the objective function has approximately the value of 5.93.

We can see, in Fig. 27.2, some shapes of the dispersion band in the settler, starting with the lower curve, and changing the parameters the fit to the target results (the highest curve).

In Table 27.1 we can see the evolution of the parameter value during the optimization process. In the first line, iteration 0, we have the start guess given to the optimization program. In the lines corresponding to the iterations 1 and 2, we can see the first choices from the program for the α parameter and the respective values for the objective function f (almost the same). In iteration 3 the optimization program starts to change the λ^* parameter to better fit the target results. In iteration 5 it starts to fit the k parameter and in iteration 7 the λ_0 parameter is changed.

The value of 1.57 for the objective function is reached in a few time steps, only 164. The choice of the starting point is very important in this kind of methods.

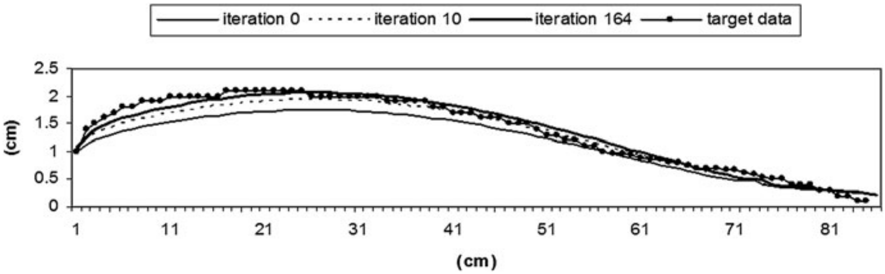


Fig. 27.2 Band thickness for different parameters values obtained from the optimization program

Table 27.1 Variation of the parameters values and objective function over time

Iteration	α	λ^*	k	λ_0	f
0	10^{-2}	3×10^{-2}	10^{-1}	10^{-4}	5.93153
1	1.05×10^{-2}	3×10^{-2}	10^{-1}	10^{-4}	5.95738
2	9.5×10^{-3}	3×10^{-2}	10^{-1}	10^{-4}	5.90571
3	9.5×10^{-3}	3.05×10^{-2}	10^{-1}	10^{-4}	6.02712
4	9.5×10^{-3}	2.95×10^{-2}	10^{-1}	10^{-4}	5.79112
5	9.5×10^{-3}	2.95×10^{-2}	1.005×10^{-1}	10^{-4}	6.22646
6	9.5×10^{-3}	2.95×10^{-2}	9.95×10^{-1}	10^{-4}	5.37141
7	9.5×10^{-3}	2.95×10^{-2}	9.95×10^{-1}	1.05×10^{-4}	7.05557
8	9.5×10^{-3}	2.95×10^{-2}	9.95×10^{-1}	9.5×10^{-5}	3.87476
9	8.5×10^{-3}	2.90×10^{-2}	9.90×10^{-1}	9.0×10^{-5}	2.34631
10	8.5×10^{-3}	2.85×10^{-2}	9.90×10^{-1}	9.0×10^{-5}	2.29587
50	8.493×10^{-3}	2.8493×10^{-2}	9.8493×10^{-1}	8.493×10^{-5}	1.60812
100	8.493×10^{-3}	2.8493×10^{-2}	9.8453×10^{-1}	8.453×10^{-5}	1.58935
164	8.36×10^{-3}	2.8361×10^{-2}	9.8361×10^{-1}	8.361×10^{-5}	1.57313

27.6 Other Parameters to Optimize

Our algorithm for the simulation of the transient state of the mixer-settler system is, in fact, composed by two connected algorithms, the mixer and the settler. The dispersion entering the settler is the dispersion leaving the mixer as Fig. 27.3 illustrates (the equipment conception must obey this principle). One change in the operating variables of the mixer (hold-up, flow rate or stirring speed), has an impact on the dispersion band of the settler over time. Therefore, it is very important to find the best values for the parameters of the hydrodynamic models for the breakage and drop–drop coalescence of the dispersion in the mixer unit. The tuning of these parameters through the present optimization procedure will be done in the future.

In the mixer, we adopted [5, 11] a simplified Coulaloglou and Tavlarides [3] model for the coalescence and breakage of the drops. The model of the mixer unit is prepared for the mass transfer phenomena but we do not consider this phenomenon at present. To model the breakage of the drops, this model takes into account the breakage frequency, $g(v)$, the size distribution of the daughters drops

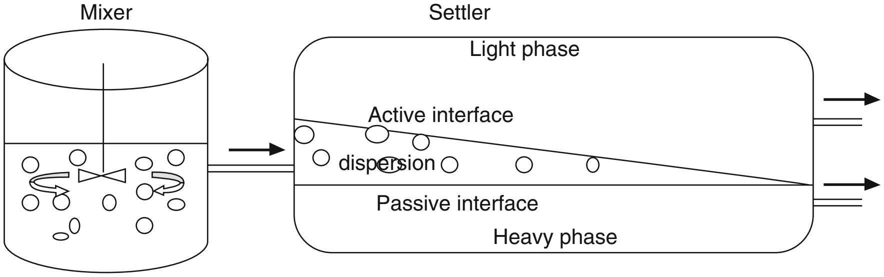


Fig. 27.3 Scheme of the mixer-settler system

and the number of the resulting drops of one drop.

$$g(v) = C_1 \frac{\varepsilon^{1/3} v^{-2/9}}{1+\Phi} \exp\left(-\frac{C_2 \sigma (1+\Phi)^2}{\rho_d \varepsilon^{2/3} v^{5/9}}\right) \tag{27.11}$$

The symbols C_1 and C_2 are dimensionless constants whose value must be determined experimentally. The ε, σ, ρ_d and Φ parameters represent, respectively, the agitation power of the mixer, the interfacial tension, the density of the dispersed phase, the hold up (in the mixer), and v is the drop volume. The distribution of the daughter drops, $\beta(v|v')$, is defined in (27.12) ones:

$$\beta(v|v') = \frac{2.402}{v} \exp\left(-\frac{4.5(2v'-v)^2}{v^2}\right) \tag{27.12}$$

To model the coalescence of the drops we have (27.13) and (27.14) describing the collision frequency, $h(v, v')$, and the coalescence efficiency, $\lambda(v, v')$, respectively.

$$h(v, v') = C_3 \varepsilon^{1/3} (v^{2/3} + v'^{2/3})(v^{2/9} + v'^{2/9})^{1/2} \tag{27.13}$$

$$\lambda(v, v') = \exp\left[-C_4 \frac{\mu_c \rho_c \varepsilon}{\sigma^2} \left(\frac{v^{1/3} v'^{1/3}}{v^{1/3} + v'^{1/3}}\right)^4\right] \tag{27.14}$$

C_3 and C_4 are two parameters of the model whose value must be provided and ρ_c and μ_c represents respectively, the density and viscosity of the continuous phase.

27.7 Conclusions

In this paper we have presented an optimization approach for tuning a set of four parameters in a shallow-layer settler simulation algorithm. We have used the Hooke–Jeeves direct search optimization algorithm to minimize an objective function that evaluates the fit of the simulation results to given target values. Our results show that the optimization approach is able to significantly reduce the value of the objective function, even if we start from a relatively good initial guess. In previous studies of our simulation program we have seen that the computational results were found to be meaningful and indicate that the model and the simulation provide adequate qualitative predictions of the settler's dynamic behavior and we knew the possible range for these four parameters. In the future, given more and better sets of target values, we will be able to work on finer details of the simulation algorithms. We have now a workable optimization program to fit these model parameters to experimental work.

27.8 Future Work

As pointed out before, this work has a number of lines that are worth pursuing. We intend to increase the number of parameters to be tuned with the optimization procedure. Namely we will extend the parameter optimization to models of the hydrodynamic phenomenon in the mixer unit. We hope to obtain the presently lacking experimental data necessary for a more accurate validation of the mathematical models after solving the current crud problem in the band dispersion. Since we are simulating a transient state, validation should be done taking time into account as well. In the present work, however, due to the problems with experimental data, we only validate the final state of the simulated process. To improve our simulation model we also need more information about the distribution of the drops entering the settler from the mixer. This will also be studied both experimentally and theoretically. It is also important to know the variation of the diameter of the drops along the settler. As we mentioned already, this can be improved using image acquisition and processing [1, 2].

References

1. Brás, L.M.R., Gomes, E.F., Ribeiro, M.M.M., Guimarães, M.M.L.: Drop distribution determination in a liquid–liquid dispersion by image processing. Int. J. Chem. Eng. (2009). Article ID:746439, doi:10.1155/2009/7464
2. Brás, L.M.R., Gomes, E.F., Ribeiro, M.M.M.: Image processing for the estimation of drop distribution in agitated liquid–liquid dispersion. In: Tenreiro Machado, J.A., Luo, A.C.J., Barbosa, R.S., Silva, M.F., Figueiredo, L.B. (eds.) Nonlinear Science and Complexity. Springer (2011)
3. Coulaloglou, C.A., Tavlarides, L.L.: Description of interaction processes in agitated liquid–liquid dispersions. Chem. Eng. Sci. **32**, 1289 (1977)
4. Gomes, E.F., Pinto, G.A., Guimarães, M.M.L., Ribeiro, L.M.: Coupling optimization to the algorithm for a Mixer-Settler system. In: Papadrakakis, M., Topping, B.H.V. (eds.) Proceedings of the Sixth International Conference on Engineering Computational Technology, paper 119. Civil-Comp Press, Stirlingshire, United Kingdom (2008)

5. Gomes, E.F., Guimarães, M.M.L., Ribeiro, L.M.: Numerical modelling of a gravity settler in dynamic conditions. J. Adv. Eng. Software, Elsevier **38**, 810–817 (2007)
6. Gomes, E.F., Guimarães, M.M.L., Pinto, G.A., Ribeiro, L.M.: Dynamical model for a settler unit using kinetic formulation. In: Topping, B.H.V., Montero, G., Montenegro, R. (eds.) Proceedings of The Fifth International Conference on Engineering Computational Technology, paper 176. Civil-Comp Press, Stirlingshire, United Kingdom (2006)
7. Gomes, E.F., Madureira, C.M.N., Guimarães, M.M.L., Ribeiro, L.M.: Numerical modelling of a gravity settler in dynamic conditions. In: Topping, B.H.V., Mota Soares, C.A. (eds.) Proceedings of the Fourth International Conference on Engineering Computational Technology, paper 59. Civil-Comp Press, Stirling, United Kingdom (2004)
8. Hooke, R., Jeeves, T.A.: Direct search solution of numerical and statistical problems. J. ACM **8**(2), 212–229 (1961)
9. Jeelani, S.A.K., Hartland, S.: The continuous separation of liquid/liquid dispersions. Chem. Eng. Sci. 1993; **48**(2), 239–254 (2004)
10. Ribeiro, M.M.M., Guimarães, M.M.L., Madureira, C.M.N., Cruz-Pinto, J.J.C.: Non-invasive system and procedures for the characterization of liquid–liquid dispersions. Chem. Eng. J. **97**, 173–182 (2004)
11. Ribeiro, L.M., Regueiras, P.F.R., Guimarães, M.M.L., Cruz-Pinto, J.J.C.: The dynamic behaviour of liquid–liquid agitated dispersions-I. The hydrodynamics. Comput. Chem. Eng. **19**(3), 333–343 (1995)
12. Ruiz, M.C.: Mathematical Modelling of a Gravity Settler, Ph.D. Thesis, University of Utah (1985)

Chapter 28
Duality Theory, Representation Formulas and Uniqueness Results for Viscosity Solutions of Hamilton–Jacobi Equations

Diogo A. Gomes and Enrico Valdinoci

Abstract In this paper we review some representation formulas for viscosity solutions in terms of certain variational problems, following an approach due to (Fleming and Vermes, SIAM J Control Optim 27(5):1136–1155, 1989).We consider both the discounted cost infinite horizon problem and the terminal value problem. These formulas are obtained using a relaxed control formulation and then applying duality theory.

28.1 Introduction

The objective of this paper is to review some representation formulas for viscosity solutions of Hamilton Jacobi equations using generalized Mather measures and infinite dimensional linear programming. Formulas of this type were first obtained in [6] in case no boundary was present.

We will consider both the discounted cost infinite horizon problem and the terminal value optimal control problem and we will recall some representation formulas for viscosity solutions in terms of certain variational problems, see Theorems 28.1 and 28.2.

As a by-product of these representation formulas, the uniqueness of viscosity solutions easily follows.

In further detail, we deal with two basic models. The first one is the discounted cost infinite horizon problem, driven by the Hamilton–Jacobi equation

$$\begin{aligned} &\alpha u + H(D^2u, Du, x) = 0 \qquad \text{in } \Omega, \\ &u = \psi \qquad \text{on } \partial\Omega. \end{aligned} \tag{28.1}$$

D.A. Gomes (✉)
Instituto Superior Técnico, Av. Rovisco Pais, 1049-001 Lisbon, Portugal
e-mail: dgomes@math.ist.utl.pt

E. Valdinoci
Department of Mathematics, Universit degli Studi di Roma, Tor Vergata, Rome, Italy
e-mail: valdinoc@mat.uniroma2.it

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_28,

In this case, $\alpha > 0$, Ω is an open subset of $\mathbb{R}^n$, not necessarily bounded, $\partial\Omega$ is its boundary, possibly empty, and $\psi \in C(\partial\Omega)$ is a bounded function. The function $H \in C(\mathrm{Sym}^n \times \mathbb{R}^n \times \mathbb{R}^n, \mathbb{R})$ is the Hamiltonian, where Sym^n is the space of the $(n \times n)$-symmetric matrices, and $u : \overline{\Omega} \to \mathbb{R}$ is a viscosity solution of (28.1) (see, e.g., [2] or [3, 5]).

We work with the stochastic control setting in which the dynamics is given by a diffusion coefficient σ and a drift coefficient f. We take $\sigma,\ f \in C(\overline{\Omega} \times \mathbb{R}^n, \mathbb{R}^n)$ and we suppose that

$$\lim_{\substack{|(x,v)| \to +\infty \\ (x,v) \in \Omega \times \mathbb{R}^n}} \frac{\big(|f| + |\sigma|^2 + L\big)}{1 + |x|^q + |v|^q} = 0, \tag{28.2}$$

for some $q > 1$. We also assume that there exists a bounded progressively measurable control ϑ (in many cases it suffices to take $\vartheta = 0$), such that the stochastic differential equation

$$dx = f(x, \vartheta)dt + \sigma(x, \vartheta)dW_t,$$

has a solution defined for all times up to the hitting time $T_{\partial\Omega}$ of $\partial\Omega$ (which can be $+\infty$). We assume further that

$$E \int_0^{T_{\partial\Omega}} L(x, \vartheta)dt < \infty. \tag{28.3}$$

In this control theory setting (see [2] for deterministic control problems or [5] for stochastic control) the Hamiltonian H is the generalized Legendre transform of a Lagrangian, namely there exists a lower semicontinuous function $L : \Omega \times \mathbb{R}^n \to \mathbb{R}$, bounded from below, for which

$$H(x, \zeta, M) = \sup_{v \in \mathbb{R}^n} \left[-f(x, v) \cdot \zeta - \frac{\sigma\sigma^T}{2}(x, v) : M - L(x, v) \right], \tag{28.4}$$

for any $x \in \Omega$, $\zeta \in \mathbb{R}^n$ and $M \in \mathrm{Sym}^n$. We suppose further that H has the following uniform continuity property: for any sequences $(M_n, p_n), (\tilde{M}_n, \tilde{p}_n) \in \mathrm{Sym}^n \times \mathbb{R}^n$ and any sequence $x_n \in \Omega$ (not necessarily convergent) such that $(M_n - \tilde{M}_n, p_n - \tilde{p}_n) \to 0$ we have

$$|H(M_n, p_n, x_n) - H(\tilde{M}_n, \tilde{p}_n, x_n)| \to 0. \tag{28.5}$$

We assume that any viscosity solution $u \in C(\overline{\Omega})$ can be approximated by smooth subsolutions, namely that for any $\epsilon > 0$ there exists $u^\epsilon \in C^2(\overline{\Omega}) \cap W^{2,\infty}(\overline{\Omega})$ such that

$$\begin{aligned}
&\alpha u^\epsilon + H(D^2u^\epsilon(x), Du^\epsilon(x), x) \leq \epsilon \qquad \text{for any } x \in \Omega,\\
&u^\epsilon(x) = \psi^\epsilon(x) \qquad \text{for any } x \in \partial\Omega\\
&\qquad \text{and } \lim_{\epsilon \to 0+} u^\epsilon(x) = u(x) \qquad \text{for any } x \in \overline{\Omega}\,.
\end{aligned} \tag{28.6}$$

We further assume that $\psi^\epsilon \to \psi$ uniformly in $\partial\Omega$.

We should point out that these approximation hypotheses are quite general. For instance, in the first order case one has a-priori Lipschitz bounds for the viscosity solution, and since the Hamiltonian is convex, by convolving the solution with a standard mollifier one obtains a solution with the required properties. For second order equations one can use the inf/sup convolution to obtain semiconcave subsolutions which can then be convolved with standard mollifiers to produce v^ϵ, see [5] for details.

We also suppose that the growth of u is, at most, logarithmic, in the sense that

$$\lim_{\substack{|x| \to +\infty \\ x \in \Omega}} \frac{u(x)}{\ln(1 + |x|^2)} = 0. \tag{28.7}$$

Under the above hypothesis, the following representation result holds:

Theorem 28.1. *The function u in* (28.1) *may be represented as*

$$u(x) = \inf_{\mu,\nu} \int_{\overline{\Omega} \times \mathbb{R}^n} L \, d\mu + \int_{\partial\Omega} \psi \, d\nu, \tag{28.8}$$

where the infimum is taken over all measures μ *on* $\overline{\Omega} \times \mathbb{R}^n$ *and* ν *on* $\partial\Omega$ *satisfying the constraint*

$$\int_{\overline{\Omega} \times \mathbb{R}^n} f \cdot D\phi + \frac{\sigma\sigma^T}{2} : D^2\phi - \alpha\phi \, d\mu = \int_{\partial\Omega} \phi \, d\nu - \phi(x) \tag{28.9}$$

for any $\phi \in C^2(\overline{\Omega}) \cap W^{2,\infty}(\overline{\Omega})$.

Also, u admits also the following dual representation:

$$\begin{aligned}
u(x) = &\sup_{\phi \in C^2(\overline{\Omega}) \cap W^{2,\infty}(\overline{\Omega})} \inf_{T \geq 0} \inf_{y \in \overline{\Omega}} \frac{e^{-\alpha T} - 1}{\alpha} \Big(\alpha\phi(y) + H(D^2\phi(y), D\phi(y), y)\Big)\\
&+ e^{-\alpha T} \inf_{z \in \partial\Omega} \Big(\psi(z) - \phi(z)\Big) + \phi(x)\,.
\end{aligned} \tag{28.10}$$

The second problem we consider is the parabolic problem

$$\begin{aligned}
&-v_t + H(D^2v, Dv, v, x) = 0 \qquad \text{in } \mathbb{R}^n \times (0, T),\\
&v(x, T) = \psi(x) \qquad \text{for } x \in \mathbb{R}^n\,.
\end{aligned} \tag{28.11}$$

As before, $H : \mathrm{Sym}^n \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ and $v : \mathbb{R}^n \times [0,T] \to \mathbb{R}$, is a viscosity solution of (28.11).

As usual, $v_t = \partial_t v(x,t)$ denotes the time-derivative, while $Dv = (\partial_{x_1} v, \dots, \partial_{x_n} v)$ is the spatial gradient. We suppose that $v \in C(\mathbb{R}^n \times [0,T])$, and that v may be approximated by smooth subsolutions $v^\epsilon \in C^2(\mathbb{R}^n \times [0,T]) \cap W^{2,\infty}(\mathbb{R}^n \times [0,T])$ for any $\epsilon > 0$, that is

$$\begin{aligned} & -v_t^\epsilon + H(D^2 v^\epsilon(x,t), Dv^\epsilon(x,t), x) \le \epsilon \qquad \text{for any } (x,t) \in \mathbb{R}^n \times (0,T), \\ & v^\epsilon(x,T) = \psi^\epsilon(x) \qquad \text{for any } x \in \mathbb{R}^n \\ & \quad \text{and } \lim_{\epsilon \to 0^+} v^\epsilon(x,t) = v(x,t) \qquad \text{for any } (x,t) \in \mathbb{R}^n \times [0,T]. \end{aligned} \tag{28.12}$$

We also suppose that the growth of v at infinity is less than logarithmic, that is

$$\lim_{\substack{|x| \to +\infty \\ x \in \mathbb{R}^n}} \frac{\sup_{s \in [0,T]} |v(x,s)|}{\ln(1+|x|^2)} = 0. \tag{28.13}$$

In this framework, the following result holds true:

Theorem 28.2. *The function v in* (28.11) *may be represented as*

$$v(x,t) = \inf \int_{\mathbb{R}^n \times [t,T]} L \, d\mu + \int_{\mathbb{R}^n \times \{T\}} \psi \, d\nu \tag{28.14}$$

where the infimum is taken over all the measures μ on $\mathbb{R}^n \times [t,T] \times \mathbb{R}^n$ and ν on $\mathbb{R}^n \times \{T\}$ satisfying the constraint

$$\int_{\mathbb{R}^n \times [t,T] \times \mathbb{R}^n} f \cdot D\phi + \frac{\sigma\sigma^T}{2} : D^2\phi + \phi_t \, d\mu = \int_{\mathbb{R}^n \times \{T\}} \phi(y,T) \, d\nu - \phi(x,t) \tag{28.15}$$

for any $\phi \in C^2(\mathbb{R}^n \times [t,T]) \cap W^{2,\infty}(\mathbb{R}^n \times [t,T])$.

Moreover, we can also represent v as

$$\begin{aligned} v(x,t) = & \sup_{\phi \in C^2(\mathbb{R}^n \times [t,T]) \cap W^{2,\infty}(\mathbb{R}^n \times [t,T])} \inf_{\substack{y \in \mathbb{R}^n \\ s \in [t,T]}} (T-t) \\ & \times \Big(\phi_t(y,s) - H(D^2\phi(y,s), D\phi(y,s), y) \Big) \\ & + \inf_{\zeta \in \mathbb{R}^n} \Big(\psi(\zeta) - \phi(\zeta,T) \Big) + \phi(x,t). \end{aligned} \tag{28.16}$$

We point out that Theorems 28.1 and 28.2 imply that the viscosity solutions of (28.1) and (28.11) are unique, under our assumptions, since they have to agree with (28.8) and (28.14), respectively. In this sense, the representation formulas of Theorems 28.1 and 28.2 give a different proof of the standard uniqueness results for viscosity solutions (see, for instance, [3, 5]).

Linear programming methods for deterministic and stochastic control have been used by several authors, see for instance [4, 7, 8] and references therein. The representation formulas in (28.10) and (28.16), which were first obtained in [6] in case no boundary was present, will be derived via a general duality formulation, developed in Sect. 28.3, which we believe is interesting in itself and which may lead to further developments.

The paper is organized in the following way: Sect. 28.2 heuristically motivates the duality formulas we present; Sect. 28.3 develops an abstract duality theory for a generalized Mather problem, which is then applied in Sects. 28.4 and 28.5 to the proofs of Theorems 28.1 and 28.2.

28.2 Heuristic Motivations

We make some comments to motivate the representation formulas (28.10) and (28.16) of Theorems 28.1 and 28.2. To simplify the presentation we consider the first-order calculus of variations setting, that is $\sigma = 0$, and $f = v$. Furthermore, since the argument is analogous for the terminal value problem, we only describe the discounted cost infinite horizon problem.

It is well known that the viscosity solution u in (28.1) can be expressed as the solution to the optimal control problem

$$u(x) = \inf_{\mathbf{x}(0)=x,\, T\geq 0} \int_0^T e^{-\alpha t} L(\mathbf{x}, \dot{\mathbf{x}}) ds + e^{-\alpha T} \psi(\mathbf{x}(T)) \,. \tag{28.17}$$

The quantity T in (28.17) is the "terminal time" $T_{\partial\Omega}(\mathbf{x}) \in [0, +\infty]$ for which the trajectory $\mathbf{x}$ exits Ω (this is the case discussed, for instance, in [2] or [5]).

In this sense, (28.8) and (28.9) is a relaxed form of the optimal control problem (28.17). Indeed, if we define

$$\int_{\overline{\Omega}\times\mathbb{R}^n} \phi(y, v)\mu(y, v) = \int_0^T \phi(\mathbf{x}(s), \mathbf{v}(s))\, e^{-\alpha s}\, ds$$

and

$$\int_{\partial\Omega} \phi(y))\, d\nu(y) = \phi(\mathbf{x}(T))\, e^{-\alpha T},$$

we have

$$\begin{aligned}\int_{\overline{\Omega}\times\mathbb{R}^n} v \cdot D\phi(y) - \alpha\phi(y)\, d\mu(y, v) &= \int_0^T \frac{d}{ds}\Big(\phi(\mathbf{x}(s))\, e^{-\alpha s}\Big)\, ds \\ &= \phi(\mathbf{x}(T))\, e^{-\alpha T} - \phi(x) \\ &= \int_{\partial\Omega} \phi(y)\, d\nu(y) - \phi(x),\end{aligned}$$

thence the optimal control trajectories are compatible with the constraint in (28.9) and (28.8) may thus be seen as a relaxation of (28.17).

By heuristically using the minimax principle the dual representation of u claimed in Theorem 28.1 then follows. More precisely, we can write

$$u(x) = \inf_{\mu,\nu} \sup_{\phi} \int_{\overline{\Omega}\times\mathbb{R}^n} L + vD_x\phi - \alpha\phi d\mu + \int_{\partial\Omega} (\psi - \phi)d\nu + \phi(x),$$

where the infimum is taken over all measures μ and ν as in Theorem 28.1, and similarly for the supremum. By (formally) exchanging the supremum with the infimum we obtain, claiming that the minimax principle holds,

$$\begin{aligned} u(x) &= \sup_{\phi} \inf_{\mu,\nu} \int_{\overline{\Omega}\times\mathbb{R}^n} L + vD_x\phi - \alpha\phi d\mu + \int_{\partial\Omega} (\psi - \phi)d\nu + \phi(x) \\ &= \sup_{\phi} \inf_{0\le\lambda\le1} \frac{\lambda}{\alpha} \inf_{y\in\overline{\Omega}} -H(D\phi(y), y) + \lambda \inf_{z\in\partial\Omega} \psi(z) - \phi(x) + \phi(x), \end{aligned}$$

where we took into account that pair (μ, ν) satisfying (28.9) will also satisfy

$$\alpha \int d\mu + \int d\nu = 1,$$

by choosing in (28.9) $\phi = 1$.

28.3 A General Duality Theory

A main tool to establish the results in this paper is the Legendre–Fenchel–Rockafellar duality theory (see, for instance, [9]). In this section we give a general dual formulation that will simplify the proofs in our examples.

28.3.1 The Generalized Mather Problem

Let Ω_i be closed (possibly unbounded) subsets of $\mathbb{R}^n$, for $i = 0, \dots, N$.

Let L_i be lower semicontinuous functions on Ω_i such that

$$\inf_{1\le i\le N} \inf_{z_i\in\Omega_i} L_i(z_i) > -\infty. \tag{28.18}$$

Suppose that there exists $\gamma_i \in C\big(\Omega_i, [1, +\infty)\big)$ such that

$$\lim_{\substack{|z_i|\to+\infty\\ z_i\in\Omega_i}} \frac{L_i(z_i)}{\gamma_i(z)} = +\infty, \tag{28.19}$$

for $i = 0, \dots, N$.

We denote by $C_0^{\gamma_i}(\Omega_i)$ the set of continuous functions ϕ_i that satisfy

$$\|\phi_i\|_{\gamma_i} = \sup_{\Omega_i}\left|\frac{\phi_i}{\gamma_i}\right| < \infty, \qquad \lim_{\substack{|z_i|\to+\infty\\ z_i\in\Omega_i}} \frac{\phi_i(z)}{\gamma_i(z)} = 0.$$

For any $1 \le i \le N$, let

$$R_i = \left\{\mu_i \text{ signed measures on } \Omega_i \text{ with } \int_{\Omega_i} \gamma_i\, d|\mu_i| < \infty\right\}. \tag{28.20}$$

The set R_i is the dual of the set $C_0^{\gamma_i}(\Omega_i)$. Let $\mathscr{M} = \prod_i R_i$. Let X be a vector space and let $A_i : X \to Y_i \subseteq C_0^{\gamma_i}(\Omega_i)$, for $i = 0, \dots, N$ be linear operators.

Let P be a (non-empty) convex subset of $\mathscr{M}$ of non-negative measures such that

$$\sup_{(\mu_1,\dots,\mu_N)\in P} \int_{\Omega_i} d\mu_i < +\infty, \tag{28.21}$$

and the following separation property holds: for any nonnegative $(\mu_1, \dots, \mu_N) \in \mathscr{M} \setminus P$, there exist $c_o > 0$ and $\psi_{o,1} \in C_0^{\gamma_1}(\Omega_1), \dots, \psi_{o,N} \in C_0^{\gamma_N}(\Omega_N)$ in such a way that

$$\sum_{i=1}^N \int_{\Omega_i} \psi_{o,i}\, d\mu_i - \sup_{(\hat\mu_1,\dots,\hat\mu_N)\in P} \sum_{i=1}^N \int_{\Omega_i} \psi_{o,i}\, d\hat\mu_i \ge c_o\,. \tag{28.22}$$

We remark that

$$\text{if } (\mu_1, \dots, \mu_N) \in P \quad \text{then} \quad \mu_1 \ge 0, \dots, \quad \text{and} \quad \mu_N \ge 0, \tag{28.23}$$

Also, we observe that both (28.21) and (28.22) are automatically satisfied if P is defined by either

$$\sum_{i=1}^N \lambda_i \int_{\Omega_i} d\mu_i = 1 \qquad (\text{or} \le 1) \tag{28.24}$$

or

$$\int_{\Omega_i} d\mu_i = \lambda_i \qquad (\text{or} \le \lambda_i) \tag{28.25}$$

for suitable $\lambda_i > 0$.

To check this claim, it suffices to take $\psi_{0,i} = \pm\lambda_i$, in case (28.24). In case (28.25), assume that the identity j fails. Then we take $\psi_{0,i} = 0$, for $i \neq j$, and $\psi_{o,j} = \pm 1$.

Fix a measure $\mu_0 \in R_0$. The generalized Mather problem consists in

$$\inf \sum_{i=1}^{N} \int_{\Omega_i} L_i \, d\mu_i \tag{28.26}$$

where the infimum is taken over all measures μ_i on Ω_i satisfying the following two constraints:

$$(\mu_1, \ldots, \mu_N) \in P$$

and

$$\sum_{i=1}^{N} \int_{\Omega_i} A_i \varphi \, d\mu_i = \int_{\Omega_0} A_0 \varphi_0 \, d\mu_0,$$

for any $\varphi \in X$.

From now on, we will suppose that the constraints are non-void, meaning that there exists $(\underline{\mu}_1, \ldots, \underline{\mu}_N) \in P$ such that

$$\sum_{i=1}^{N} \int_{\Omega_i} A_i \varphi \, d\underline{\mu}_i = \int_{\Omega_0} A_0 \varphi \, d\mu_0 \tag{28.27}$$

for any $\varphi \in X$.

We will also suppose that the problem in (28.26) makes sense, that is that there exists $(\mu_1^\star, \ldots, \mu_N^\star) \in P$ for which

$$\sum_{i=1}^{N} \int_{\Omega_i} L_i \, d\mu_i^\star < +\infty. \tag{28.28}$$

These last two hypothesis in our problems will be guaranteed by the use of (28.3).

28.3.2 Duality

We start by recalling the Legendre–Fechel–Rockafellar Theorem (see, for instance, [9]). For that, let E be a Banach space with dual E'. The pairing between E and E' is denoted by $(\cdot, \cdot)$. Suppose that $h : E \to (-\infty, +\infty]$ is a lower semicontinuous convex function.

The Legendre–Fenchel transform $h^* : E' \to [-\infty, +\infty]$ of h is defined by

$$h^*(y) = \sup_{x \in E} \left(-(x, y) - h(x)\right), \tag{28.29}$$

for all $y \in E'$. In a similar way, if $g : E \to [-\infty, +\infty)$ is concave and upper semicontinuous, we define

$$g^*(y) = \inf_{x \in E} \left(-(x, y) - g(x)\right). \tag{28.30}$$

Theorem 28.3 (Legendre–Fenchel–Rockafellar). *Let E be a locally convex, Hausdorff topological vector space over $\mathbb{R}$ with dual E'. Let $h : E \to (-\infty, +\infty]$ be a convex lower semicontinuous function, and $g : E \to [-\infty, +\infty)$ a concave upper semicontinuous function. Then*

$$\sup_{x \in E} [g(x) - h(x)] = \min_{y \in E'} \left[h^*(y) - g^*(y)\right], \tag{28.31}$$

provided that there exists a point x_0 where g and h are finite, and that at this point at least one of them is continuous.

Note that it is part of the theorem that the right-hand side of (28.31) is in fact a minimum.

28.3.2.1 Identification of Dual Problems

We make use of the Legendre–Fenchel–Rockafellar's Theorem to compute the dual of the generalized Mather problem.

Let

$$\begin{aligned} \mathscr{M}_\star = \Big\{ &(\mu_1, \ldots, \mu_N) : \mu_i \in R_i\,, \\ &\sum_{i=1}^{N} \int_{\Omega_i} A_i\varphi\, d\mu_i = \int_{\Omega_0} A_0\varphi\, d\mu_0 \ \ \forall \varphi \in X \Big\}. \end{aligned} \tag{28.32}$$

We also use a vector-like notation, by setting

$$\begin{aligned} \gamma &= (\gamma_1, \ldots, \gamma_N), \\ \Omega &= \Omega_1 \times \cdots \times \Omega_N, \end{aligned}$$

and

$$C_0^\gamma(\Omega) = C_0^{\gamma_1}(\Omega_1) \times \cdots \times C_0^{\gamma_N}(\Omega_N).$$

For $\phi = (\phi_1, \dots, \phi_N) \in C_0^\gamma(\Omega)$ we denote

$$\|\phi\|_\gamma = \sum_{i=1}^N \|\phi_i\|_{\gamma_i}$$

and

$$h(\phi) = \sup_{(\mu_1,\dots,\mu_N)\in P} \sum_{i=1}^N \int_{\Omega_i} \big(-\phi_i - L_i\big)\, d\mu_i. \tag{28.33}$$

Since h is the supremum of convex, and in fact linear, functions of ϕ, we have that

$$h \text{ is convex.} \tag{28.34}$$

Lemma 28.1. *Let $\phi_o \in C_0^\gamma(\Omega)$ and $\mu_o \in P$. Suppose that there exists $\kappa > 0$ such that*

$$h(\phi_o) \le \kappa + \sum_{i=1}^N \int_{\Omega_i} \big(-\phi_{o,i} - L_i\big)\, d\mu_{o,i}. \tag{28.35}$$

Then there exists $C(\|\phi_o\|_\gamma, \kappa)$ such that

$$\sum_{i=1}^N \int_{\Omega_i} \gamma_i \, d\mu_{o,i} \le C(\|\phi_o\|_\gamma, \kappa).$$

Proof. By (28.33) and (28.28),

$$\begin{aligned}
-h(\phi_o) &\le \sum_{i=1}^N \int_{\Omega_i} \big(\phi_{o,i} + L_i\big)\, d\mu_i^\star \\
&\le \sum_{i=1}^N \int_{\Omega_i} \big(\|\phi_{o,i}\|_{\gamma_i}\gamma_i + L_i\big)\, d\mu_i^\star \\
&\le \|\phi_o\|_\gamma \sum_{i=1}^N \int_{\Omega_i} \gamma_i \, d\mu_i^\star + \sum_{i=1}^N \int_{\Omega_i} L_i \, d\mu_i^\star \\
&\le C_o(\|\phi_o\|_\gamma),
\end{aligned}$$

for a suitable $C_o(\|\phi_o\|_\gamma) > 0$.

Thus, from (28.35),

$$\sum_{i=1}^N \int_{\Omega_i} \big(\phi_{o,i} + L_i\big)\, d\mu_{o,i} \le \kappa + C_o(\|\phi_o\|_\gamma). \tag{28.36}$$

Also, by (28.19), there exists $R(\|\phi_o\|_\gamma)$ such that

$$L_i(z_i) \geq (1 + \|\phi_o\|_\gamma)\gamma_i(z_i), \tag{28.37}$$

for any $z_i \in \Omega_i$ such that $|z_i| \geq R(\|\phi_o\|_\gamma)$.

Let

$$c(\|\phi_o\|_\gamma) = (1 + \|\phi_o\|_\gamma) \sum_{i=1}^{N} \left(\sup_{\substack{|z_i| \leq R(\|\phi_o\|_\gamma) \\ z_i \in \Omega_i}} \gamma_i(z_i) + |\inf_{\Omega_i} L_i| \right).$$

Note that $c(\|\phi_o\|_\gamma)$ is finite, due to (28.18), and that

$$L_i(z_i) \geq (1 + \|\phi_o\|_\gamma)\gamma_i(z_i) - c(\|\phi_o\|_\gamma),$$

for any $z_i \in \Omega_i$, because of (28.37).

Accordingly,

$$\begin{aligned} \sum_{i=1}^{N} \int_{\Omega_i} \big(\phi_{o,i} + L_i\big)\, d\mu_{o,i} &\geq \sum_{i=1}^{N} \int_{\Omega_i} \big(-\|\phi_{o,i}\|_{\gamma_i}\gamma_i + L_i\big)\, d\mu_{o,i} \\ &\geq \sum_{i=1}^{N} \int_{\Omega_i} \big(\gamma_i - c(\|\phi_o\|_\gamma)\big)\, d\mu_{o,i} \\ &\geq \sum_{i=1}^{N} \int_{\Omega_i} \gamma_i\, d\mu_{o,i} - \tilde{c}(\|\phi_o\|_\gamma), \end{aligned}$$

for a suitable $\tilde{c}(\|\phi_o\|_\gamma) > 0$, which is finite thanks to (28.21).

This estimate and (28.36) give the desired result. □

Consider the sets

$$\begin{aligned} &\mathscr{C}_o = \Big\{\phi = (\phi_1, \ldots, \phi_N) \,:\, \phi_i = A_i\varphi\,,\ \varphi \in X\,,\ i = 0, \ldots, N\Big\}, \\ &\mathscr{C} = \operatorname{cl}\mathscr{C}_o \end{aligned} \tag{28.38}$$

where cl denotes the closure in $C_0^\gamma(\Omega)$. Note that, since A_i is linear,

$$\mathscr{C} \text{ is a convex set.} \tag{28.39}$$

Let

$$g(\phi) = \begin{cases} -\sum_{i=1}^{N} \int_{\Omega_i} \phi_i\, d\underline{\mu}_i & \text{if } \phi \in \mathscr{C} \\ -\infty & \text{otherwise,} \end{cases} \tag{28.40}$$

where $\underline{\mu}_1, \ldots, \underline{\mu}_N$ are given by (28.27).

It follows from (28.39) that

$$g \text{ is concave and upper semicontinuous.} \tag{28.41}$$

Furthermore, by (28.30), (28.20) and (28.40),

$$\begin{aligned} g^*(\mu) &= \inf_{\phi \in C_0^\gamma(\Omega)} \left(-\sum_{i=1}^N \int_{\Omega_i} \phi_i \, d\mu_i - g(\phi) \right) \\ &= \inf_{\phi \in \mathscr{C}} \left(-\sum_{i=1}^N \int_{\Omega_i} \phi_i \, d\mu_i + \sum_{i=1}^N \int_{\Omega_i} \phi_i \, d\underline{\mu}_i \right). \end{aligned} \tag{28.42}$$

Such formula may be more conveniently written as follows:

Lemma 28.2. *We have*

$$g^*(\mu) = \inf_{\varphi \in X} \left(\int_{\Omega_0} A_0 \varphi \, d\mu_0 - \sum_{i=1}^N \int_{\Omega_i} A_i \varphi \, d\mu_i \right).$$

Proof. Of course, by (28.42), (28.38) and (28.27),

$$\begin{aligned} g^*(\mu) &\le \inf_{\phi \in \mathscr{C}_o} \left(-\sum_{i=1}^N \int_{\Omega_i} \phi_i \, d\mu_i + \sum_{i=1}^N \int_{\Omega_i} \phi_i \, d\underline{\mu}_i \right) \\ &= \inf_{\varphi \in X} \left(-\sum_{i=1}^N \int_{\Omega_i} A_i \varphi \, d\mu_i + \sum_{i=1}^N \int_{\Omega_i} A_i \varphi \, d\underline{\mu}_i \right) \\ &= \inf_{\varphi \in X} \left(-\sum_{i=1}^N \int_{\Omega_i} A_i \varphi \, d\mu_i + \int_{\Omega_0} A_0 \varphi \, d\mu_0 \right). \end{aligned}$$

We now prove the reverse inequality. Fix $\epsilon > 0$ and use (28.42) to obtain $\phi_\epsilon \in \mathscr{C}$ for which

$$\left| g^*(\mu) + \sum_{i=1}^N \int_{\Omega_i} \phi_{\epsilon,i} \, d\mu_i - \sum_{i=1}^N \int_{\Omega_i} \phi_{\epsilon,i} \, d\underline{\mu}_i \right| \le \frac{\epsilon}{2}. \tag{28.43}$$

By (28.38), there exists a sequence of functions

$$\phi_\epsilon^{(j)} \in \mathscr{C}_o \tag{28.44}$$

converging to ϕ_ϵ in the topology of $C_0^\gamma(\Omega)$, as $j \to +\infty$.

Accordingly,

$$\lim_{j\to+\infty}\int_{\Omega_i}\phi_{\epsilon,i}^{(j)}\,d\mu_i=\int_{\Omega_i}\phi_{\epsilon,i}\,d\mu_i$$
$$\text{and}\quad\lim_{j\to+\infty}\int_{\Omega_i}\phi_{\epsilon,i}^{(j)}\,d\underline{\mu}_i=\int_{\Omega_i}\phi_{\epsilon,i}\,d\underline{\mu}_i$$

and therefore, recalling (28.43), we have that there exists $j_\epsilon>0$ such that, if $j\geq j_\epsilon$, we have

$$\left|g^*(\mu)+\sum_{i=1}^N\int_{\Omega_i}\phi_{\epsilon,i}^{(j)}\,d\mu_i-\sum_{i=1}^N\int_{\Omega_i}\phi_{\epsilon,i}^{(j)}\,d\underline{\mu}_i\right|\leq\epsilon\,.$$

That is, by (28.44) and (28.38),

$$\left|g^*(\mu)+\sum_{i=1}^N\int_{\Omega_i}A_i\varphi_\epsilon^{(j)}\,d\mu_i-\sum_{i=1}^N\int_{\Omega_i}A_i\varphi_\epsilon^{(j)}\,d\underline{\mu}_i\right|\leq\epsilon$$

for a suitable $\varphi_\epsilon^{(j)}\in X$.

Therefore, by (28.27),

$$\left|g^*(\mu)+\sum_{i=1}^N\int_{\Omega_i}A_i\varphi_\epsilon^{(j)}\,d\mu_i-\int_{\Omega_0}A_0\varphi_\epsilon^{(j)}\,d\mu_0\right|\leq\epsilon$$

as long as $j\geq j_\epsilon$.

In particular,

$$\begin{aligned}\epsilon+g^*(\mu)&\geq\int_{\Omega_0}A_0\varphi_\epsilon^{(j_\epsilon)}\,d\mu_0-\sum_{i=1}^N\int_{\Omega_i}A_i\varphi_\epsilon^{(j_\epsilon)}\,d\mu_i\\&\geq\inf_{\varphi\in X}\int_{\Omega_0}A_0\varphi\,d\mu_0-\sum_{i=1}^N\int_{\Omega_i}A_i\varphi\,d\mu_i\,.\end{aligned}$$

By sending $\epsilon\to0^+$, we end the proof of Lemma 28.2. □

We plan to show that the dual of

$$\sup_{\phi\in C_0^\gamma(\Omega)}\quad g(\phi)-h(\phi)\tag{28.45}$$

agrees with the generalized Mather problem (this will be achieved in Proposition 28.2 and Theorem 28.4).

We start by computing the Legendre transforms of h and g. For $\mu \in \mathscr{M}$, we write $\mu \geq 0$ as a short-hand notation for $\mu_i \geq 0$ for all $i = 1, \dots, N$. We recall that $P \subseteq \{\mu \geq 0\}$ by (28.23).

Proposition 28.1. *For any $\mu = (\mu_1, \dots, \mu_N) \in \mathscr{M}$, we have*

$$h^*(\mu) = \begin{cases} \sum_{i=1}^N \int_{\Omega_i} L_i \, d\mu_i & \text{if} \quad \mu \in P \\ +\infty & \text{otherwise,} \end{cases} \tag{28.46}$$

and

$$g^*(\mu) = \begin{cases} 0 & \text{if} \quad \mu \in \mathscr{M}_\star \\ -\infty & \text{otherwise.} \end{cases} \tag{28.47}$$

The proof of Proposition 28.1 requires some preliminary work.

By (28.29),

$$h^*(\mu) = \sup_{\phi \in C_0^\gamma(\Omega)} \left(-\sum_{i=1}^N \int_{\Omega_i} \phi_i \, d\mu_i - h(\phi) \right). \tag{28.48}$$

First we prove that if $\mu \ngeq 0$, then $h^*(\mu) = \infty$.

Lemma 28.3. *If there exists i, $1 \leq i \leq N$ for which $\mu_i \ngeq 0$, then $h^*(\mu) = +\infty$.*

Proof. We fix $M > 0$. If, say, $\mu_1 \ngeq 0$ then we can choose a non-negative function $\phi_* \in C_0^{\gamma_1}(\Omega_1)$ such that

$$-\int_{\Omega_1} \phi_* \, d\mu_1 \geq M. \tag{28.49}$$

Also, since

$$-\phi_* - L_j \leq 0 - \inf_{1 \leq j \leq N} \inf_{\Omega_j} L_j \leq C$$

for some universal $C > 0$, thanks to (28.18), we deduce from (28.21) and (28.33) that $h(\phi_*, 0, \dots, 0) \leq C'$ for some universal $C' > 0$.

Thus, by plugging $(\phi_*, 0, \dots, 0)$ as a test in the left hand side of (28.48) and recalling (28.49),

$$h^*(\mu) \geq -\int_{\Omega_1} \phi_* \, d\mu_1 - C' \geq M - C'.$$

Since M can be taken arbitrarily large, we get that $h^*(\mu) = +\infty$. □

Lemma 28.4. *If $\mu \geq 0$ then*

$$h^*(\mu) \geq \sum_{i=1}^N \int_{\Omega_i} L_i \, d\mu_i + \sup_{\psi \in C_0^\gamma(\Omega)} \left(\sum_{i=1}^N \int_{\Omega_i} \psi_i \, d\mu_i - \sup_{\hat{\mu} \in P} \sum_{i=1}^N \int_{\Omega_i} \psi_i \, d\hat{\mu}_i \right). \tag{28.50}$$

Proof. Since L_i is lower semicontinuous, using the Yosida regularization (see, e.g., Theorem 2.64 in [1]), we have that L_i may be approximated monotonically from below by continuous functions $\tilde{L}_{i,j}$, as $j \to +\infty$.

Let $\tau_j \in C_0^\infty(B_{j+1})$, $0 \leq \tau_j \leq 1$, $\tau_j(z) = 1$ for any $z \in B_j$. Let $L_{i,j}(z_i) = \tau_j(z_i)\tilde{L}_{i,j}(z_i)$, for each $i = 1, \dots, N$ and $j \in \mathbb{N}$. Then, $L_{i,j}$ is also increasing pointwise towards L_i as $j \to +\infty$.

Accordingly, any ϕ in $C_0^\gamma(\Omega)$ can be written as $\phi_i = -L_{i,j} - \psi_i$, for some $\psi = (\psi_1, \dots, \psi_N)$ also in $C_0^\gamma(\Omega)$. Thus, by (28.33),

$$\begin{aligned}
&\sup_{\phi \in C_0^\gamma(\Omega)} \left(-\sum_{i=1}^N \int \phi_i d\mu_i - h(\phi) \right) \\
&\quad = \sup_{\psi \in C_0^\gamma(\Omega)} \left(\sum_{i=1}^N \int L_{i,j}\, d\mu_i + \sum_{i=1}^N \int \psi_i d\mu_i \right. \\
&\qquad \left. - \sup_{\hat{\mu} \in P} \sum_{i=1}^N \int (L_{ij} + \psi_i - L_i) d\hat{\mu}_i \right) \\
&\quad \geq \sup_{\psi \in C_0^\gamma(\Omega)} \left(\sum_{i=1}^N \int L_{i,j}\, d\mu_i + \sum_{i=1}^N \int \psi_i d\mu_i - \sup_{\hat{\mu} \in P} \sum_{i=1}^N \int \psi_i d\hat{\mu}_i \right).
\end{aligned}$$

By the monotone convergence theorem

$$\lim_{j \to +\infty} \int L_{i,j}\, d\mu_i = \int L_i\, d\mu_i$$

and so

$$\begin{aligned}
&\sup_{\phi \in C_0^\gamma(\Omega)} \left(-\sum_{i=1}^N \int \phi_i d\mu_i - h(\phi) \right) \\
&\quad \geq \sup_{\psi \in C_0^\gamma(\Omega)} \left(\sum_{i=1}^N \int L_i\, d\mu_i + \sum_{i=1}^N \int \psi_i d\mu_i - \sup_{\hat{\mu} \in P} \sum_{i=1}^N \int \psi_i d\hat{\mu}_i \right),
\end{aligned}$$

giving the desired result via (28.48). □

Lemma 28.5. *If $\mu \in P$, then*

$$h^*(\mu) \leq \sum_{i=1}^N \int_{\Omega_i} L_i\, d\mu_i.$$

Proof. Fix $\mu \in P$. Then, by (28.33),

$$h(\phi) \geq \sum_{i=1}^{N} \int (-\phi_i - L_i)\, d\mu_i$$

for any $\phi \in C_0^\gamma(\Omega)$.

This and (28.48) yield the claim. □

Lemma 28.6. *Let $\mu \geq 0$, $\mu \in \mathscr{M} \setminus P$. Then, there exists a sequence $\psi^{(j)} \in C_0^\gamma(\Omega)$ in such a way that*

$$\lim_{j \to +\infty} \left(\sum_{i=1}^{N} \int_{\Omega_i} \psi_i^{(j)}\, d\mu_i - \sup_{\hat{\mu} \in P} \sum_{i=1}^{N} \int_{\Omega_i} \psi_i^{(j)}\, d\hat{\mu}_i \right) = +\infty.$$

Proof. We take $\psi_i^{(j)} = j\, \psi_{o,i}$ in (28.22) and the desired claim follows at once. □

Proof (End of the proof of Proposition 28.1). We start by checking (28.46). If $\mu \ngeq 0$, then (28.46) follows from Lemma 28.3, thus we may focus on the case in which $\mu \geq 0$.

Accordingly, by choosing $\psi = 0$ in (28.50) we get that

$$h^*(\mu) \geq \sum_{i=1}^{N} \int_{\Omega_i} L_i\, d\mu_i. \tag{28.51}$$

Since the reverse inequality holds if $\mu \in P$, due to Lemma 28.5, it follows that (28.46) holds true if $\mu \in P$.

Therefore, we focus on the proof of (28.46) for any $\mu \geq 0$, $\mu \in \mathscr{M} \setminus P$. For such μ, we exploit Lemmata 28.4 and 28.6, together with (28.18) and (28.21), to obtain

$$\begin{aligned} h^*(\mu) \geq & \sum_{i=1}^{N} \int_{\Omega_i} L_i\, d\mu_i \\ & + \lim_{j \to +\infty} \left(\sum_{i=1}^{N} \int_{\Omega_i} \psi_i^{(j)}\, d\mu_i - \sup_{\hat{\mu} \in P} \sum_{i=1}^{N} \int_{\Omega_i} \psi_i^{(j)}\, d\hat{\mu}_i \right) \\ = & +\infty. \end{aligned}$$

This shows that (28.46) holds also when $0 \leq \mu \in \mathscr{M} \setminus P$. The proof of (28.46) is thus completed.

We now prove (28.47).

By (28.32), if $\mu \notin \mathscr{M}_\star$ then there exists $\hat{\varphi} \in X$ such that

$$\int_{\Omega_0} A_0 \hat{\varphi}\, d\mu_0 - \sum_{i=1}^{N} \int_{\Omega_i} A_i \hat{\varphi}\, d\mu_i \neq 0. \tag{28.52}$$

Let us denote by $\hat{c} \neq 0$ the quantity in (28.52). Then, fixed any $M > 0$, we define $\hat{\varphi}_M = -(\hat{c}/|\hat{c}|)M\hat{\varphi}$. Then $\hat{\varphi}_M \in X$ and so, by Lemma 28.2,

$$\begin{aligned} g^*(\mu) &\leq \int_{\Omega_0} A_0\hat{\varphi}_M \, d\mu_0 - \sum_{i=1}^{N} \int_{\Omega_i} A_i\hat{\varphi}_M \, d\mu_i \\ &= -|c|M. \end{aligned}$$

By taking M arbitrary large, we conclude that $g^*(\mu) = -\infty$. This proves (28.47) when $\mu \notin \mathscr{M}_\star$.

If, on the other hand, then $\mu \in \mathscr{M}_\star$, (28.47) plainly follows from Lemma 28.2 and (28.32).

This completes the proof of Proposition 28.1. □

Proposition 28.2. *We have that*

$$\sup_{\phi \in C_0^\gamma(\Omega)} (g(\phi) - h(\phi)) = \inf_{\mu \in \mathscr{M}} (h^*(\mu) - g^*(\mu)), \tag{28.53}$$

Proof. We will prove that

$$h \text{ is a continuous function.} \tag{28.54}$$

From this, the result follows from Legendre–Fenchel–Rockafellar's theorem, recalling (28.41) and (28.34).

To check (28.54), let $\phi_k \to \phi$ in C_0^γ, as $k \to +\infty$.

Fix $\epsilon > 0$ and use (28.33) to obtain $\mu_\epsilon \in P$ in such a way that

$$h(\phi) - \epsilon \leq \sum_{i=1}^{N} \int_{\Omega_i} \big(- \phi_i - L_i \big) \, d\mu_{\epsilon,i} \, .$$

Then,

$$h(\phi_k) \geq \sum_{i=1}^{N} \int_{\Omega_i} \big(- \phi_{k,i} - L_i \big) \, d\mu_{\epsilon,i}$$

and therefore

$$\begin{aligned} \limsup_{k \to +\infty} h(\phi) - h(\phi_k) &\leq \epsilon + \limsup_{k \to +\infty} \sum_{i=1}^{N} \int_{\Omega_i} |\phi_{k,i} - \phi_i| \, d\mu_{\epsilon,i} \\ &= \epsilon. \end{aligned}$$

By taking ϵ as small as we wish, we obtain

$$\limsup_{k \to +\infty} h(\phi) - h(\phi_k) \leq 0 \, . \tag{28.55}$$

Conversely, fixed $\epsilon \in (0,1)$, we can use (28.33) to obtain $\mu_{\epsilon,k} \in P$ in such a way that

$$h(\phi_k) - \epsilon \leq \sum_{i=1}^{N} \int_{\Omega_i} \big(-\phi_{k,i} - L_i \big)\, d\mu_{\epsilon,k,i}\,. \tag{28.56}$$

We can also assume that

$$\|\phi_k\|_\gamma \leq \|\phi_k - \phi\|_\gamma + \|\phi\|_\gamma \leq \|\phi\|_\gamma + 1. \tag{28.57}$$

Due to (28.56) and (28.57), we can now use Lemma 28.1 and we thus obtain that

$$\sum_{i=1}^{N} \int_{\Omega_i} \gamma_i \, d\mu_{\epsilon,k,i} \leq C(\|\phi\|_\gamma).$$

As a consequence,

$$\begin{aligned} h(\phi) - h(\phi_k) + \epsilon &\geq \sum_{i=1}^{N} \int_{\Omega_i} (-\phi_i - L_i)\, d\mu_{\epsilon,k,i} - \sum_{i=1}^{N} \int_{\Omega_i} (-\phi_{k,i} - L_i)\, d\mu_{\epsilon,k,i}\,. \\ &= \sum_{i=1}^{N} \int_{\Omega_i} \big(-\phi_i + \phi_{k,i} \big)\, d\mu_{\epsilon,k,i} \\ &\geq -\|\phi - \phi_k\|_\gamma \sum_{i=1}^{N} \int_{\Omega_i} \gamma_i \, d\mu_{\epsilon,k,i} \\ &\geq -C(\|\phi\|_\gamma) \|\phi - \phi_k\|_\gamma\,. \end{aligned}$$

Therefore,

$$\liminf_{k \to +\infty} h(\phi) - h(\phi_k) + \epsilon \geq 0,$$

and so, since ϵ is arbitrary,

$$\liminf_{k \to +\infty} h(\phi) - h(\phi_k) \geq 0.$$

This and (28.55) yield (28.54), as desired. □

We are now in the position to obtain the duality result for the generalized Mather problem:

Theorem 28.4. *The quantity in* (28.26) *equals*

$$\sup_{\varphi \in X} \left(\inf_{\substack{m_1,\ldots,m_N \geq 0 \\ m_i = \int_{\Omega_i} d\mu_i\,,\ \mu \in P}} \sum_{i=1}^{N} \big(m_i \inf_{\Omega_i} (L_i + A_i \varphi) \big) - \int_{\Omega_0} A_0 \varphi \, d\mu_0 \right).$$

Proof. The result will follow from (28.53).

By Proposition 28.1, the right hand side of (28.53) is equal to

$$\inf_{\mu \in P \cap \mathscr{M}_\star} \sum_{i=1}^N \int_{\Omega_i} L_i \, d\mu_i ,$$

which is exactly (28.26).

We now compute the left hand side of (28.53). Using (28.40) and (28.33), the left hand side of (28.53) is

$$\sup_{\phi \in \mathscr{C}} \left(- \sum_{i=1}^N \int_{\Omega_i} \phi_i \, d\underline{\mu}_i - \sup_{\mu \in P} \sum_{i=1}^N \int_{\Omega_i} (-\phi_i - L_i) \, d\mu_i \right).$$

By arguing as in Lemma 28.2, we see that the above quantity equals

$$\sup_{\phi \in \mathscr{C}_o} \left(- \sum_{i=1}^N \int_{\Omega_i} \phi_i \, d\underline{\mu}_i - \sup_{\mu \in P} \sum_{i=1}^N \int_{\Omega_i} (-\phi_i - L_i) \, d\mu_i \right).$$

In the light of (28.38) this is equal to

$$\sup_{\varphi \in X} \left(- \sum_{i=1}^N \int_{\Omega_i} A_i \varphi \, d\underline{\mu}_i - \sup_{\mu \in P} \sum_{i=1}^N \int_{\Omega_i} (-A_i \varphi - L_i) \, d\mu_i \right)$$

which, by (28.27), is the same as

$$\sup_{\varphi \in X} \left(- \int_{\Omega_0} A_0 \varphi \, d\mu_0 - \sup_{\mu \in P} \sum_{i=1}^N \int_{\Omega_i} (-A_i \varphi - L_i) \, d\mu_i \right), \tag{28.58}$$

By taking μ_i supported at a single point (i.e. convenient multiples of Dirac deltas) we see that

$$\sup_{\substack{\mu_i \\ \mu \in P}} \int_{\Omega_i} (-A_i \varphi - L_i) \, d\mu_i = \int_{\Omega_i} d\mu_i \, \sup_{\Omega_i} (-A_i \varphi - L_i),$$

therefore (28.58) becomes

$$\sup_{\varphi \in X} \left(- \int_{\Omega_0} A_0 \varphi \, d\mu_0 - \sup_{\substack{m_1, \dots, m_N \geq 0 \\ m_i = \int_{\Omega_i} d\mu_i , \, \mu \in P}} \sum_{i=1}^N \Big(m_i \sup_{\Omega_i} (-A_i \varphi - L_i) \Big) \right),$$

as desired. □

28.4 Proof of Theorem 28.1

For any fixed $x \in \Omega$, define $\tilde{u}(x)$ to be the infimum on the right hand side of (28.8).

We now apply Theorem 28.4 with $A_0 := -\mathrm{Id}$, $\mu_0 := \delta_{\{x\}}$, $\Omega_0 := \{x\}$, $A_1 := f \cdot D + (\sigma\sigma^T)/2 : D^2 - \alpha$, $L_1 := L$, $\mu_1 := \mu$, $\Omega_1 := \overline{\Omega} \times \mathbb{R}^n$, $\gamma_1 := 1 + |x|^{q/2} + |v|^{q/2}$, $A_2 := -\mathrm{Id}$, $L_2 := \psi$, $\Omega_2 := \partial\Omega$, $\gamma_2 := |x|^q$. Here, q is the exponent in (28.2).

Also, we take P as the set of all pairs (μ, ν) of non-negative measures satisfying

$$\alpha \int_{\overline{\Omega}\times\mathbb{R}^n} d\mu + \int_{\partial\Omega} d\nu = 1. \tag{28.59}$$

Notice that the set P here is of the form requested in (28.24). We observe that (28.9) implies (28.59), by taking $\phi := 1$ and that if (μ, ν) satisfies (28.59) then

$$\int_{\partial\Omega} d\nu \le 1 \;\text{ and }\; \int_{\overline{\Omega}} d\mu \le \frac{1}{\alpha}.$$

Finally, let $X := C^2(\overline{\Omega}) \cap W^{2,\infty}(\overline{\Omega})$. Thus, we are in the position of applying Theorem 28.4, from which we conclude that

$$\begin{aligned}\tilde{u}(x) = \sup_{\phi\in X} \inf_{\substack{m_2\in[0,1]\\ \alpha m_1+m_2=1}} m_1 \inf_{\overline{\Omega}\times\mathbb{R}^n} \Big(L + f\cdot\phi + \frac{\sigma\sigma^T}{2} : D^2\phi - \alpha\phi\Big)\\ + m_2 \inf_{\partial\Omega}\Big(\psi - \phi\Big) + \phi(x)\,.\end{aligned}$$

Hence, from (28.4),

$$\tilde{u}(x) = \sup_{\phi\in X} \inf_{\substack{m_2\in[0,1]\\ \alpha m_1+m_2=1}} \inf_{\overline{\Omega}} m_1\Big(-H - \alpha\phi\Big) + m_2 \inf_{\partial\Omega}\Big(\psi - \phi\Big) + \phi(x)\,. \tag{28.60}$$

It is suggestive rewrite (28.60) using the new variable $T := (1/\alpha)\ln(1/m_2)$: if we do this, we obtain from (28.60) and (28.59) that

$$\begin{aligned}\tilde{u}(x) = \sup_{\phi\in X} \inf_{T\ge 0} \inf_{y\in\overline{\Omega}} \frac{e^{-\alpha T}-1}{\alpha}\Big(\alpha\phi(y) + H(D^2\phi(y), D\phi(y), y)\Big)\\ + e^{-\alpha T} \inf_{z\in\partial\Omega}\Big(\psi(z) - \phi(z)\Big) + \phi(x)\,.\end{aligned} \tag{28.61}$$

Then, the proof of Theorem 28.1 will be accomplished once we show that $\tilde{u} = u$. To this end, given $\epsilon > 0$, we take u^ϵ as in (28.6) and we conclude that

$$\begin{aligned}\tilde{u}(x) \geq & \inf_{T\geq 0}\inf_{y\in\overline{\Omega}} \frac{e^{-\alpha T}-1}{\alpha}\Big(\alpha u^{\epsilon}(y) + H(D^2u^{\epsilon}(y), Du^{\epsilon}(y), y)\Big)\\ & + e^{-\alpha T}\inf_{z\in\partial\Omega}\Big(\psi(z) - u^{\epsilon}(z)\Big) + u^{\epsilon}(x)\\ \geq & \inf_{T\geq 0} -\frac{\epsilon(1-e^{-\alpha T})}{\alpha} - C\epsilon e^{-\alpha T} + u^{\epsilon}(x)\\ = & -C\epsilon + u^{\epsilon}(x)\,.\end{aligned}$$

By sending ϵ to zero, we obtain

$$\tilde{u}(x) \geq u(x)\,. \tag{28.62}$$

We now prove the reverse inequality. For this, let us observe that the left hand side of (28.61) is not changed if we replace ϕ with $\phi - c$, for any $c \in \mathbb{R}$. Therefore, (28.61) may be written as

$$\begin{aligned}\tilde{u}(x) = & \sup_{\substack{\phi\in X\\ \phi(x)=u(x)}} \inf_{T\geq 0}\inf_{y\in\overline{\Omega}} \frac{e^{-\alpha T}-1}{\alpha}\Big(\alpha\phi(y) + H(D^2\phi(y), D\phi(y), y)\Big)\\ & + e^{-\alpha T}\inf_{z\in\partial\Omega}\Big(\psi(z) - \phi(z)\Big) + \phi(x)\\ = & \sup_{\substack{\phi\in X\\ \phi(x)=u(x)}} \inf_{T\geq 0}\inf_{y\in\overline{\Omega}} \frac{e^{-\alpha T}-1}{\alpha}\Big(\alpha\phi(y) + H(D^2\phi(y), D\phi(y), y)\Big)\\ & + e^{-\alpha T}\inf_{z\in\partial\Omega}\Big(\psi(z) - \phi(z)\Big) + u(x)\,.\end{aligned}$$

Fix $\epsilon > 0$. Let $\phi^{\epsilon} \in X$ with $\phi^{\epsilon}(x) = u(x)$, and such that

$$\begin{aligned}\tilde{u}(x) - \epsilon \leq & \inf_{T\geq 0}\inf_{y\in\overline{\Omega}} \frac{e^{-\alpha T}-1}{\alpha}\Big(\alpha\phi^{\epsilon}(y) + H(D^2\phi^{\epsilon}(y), D\phi^{\epsilon}(y), y)\Big)\\ & + e^{-\alpha T}\inf_{z\in\partial\Omega}\Big(\psi(z) - \phi^{\epsilon}(z)\Big) + u(x)\,.\end{aligned} \tag{28.63}$$

Let $\delta > 0$ be a small parameter (possibly smaller than ϵ) and let

$$\Phi^{\epsilon,\delta}(y) = \phi^{\epsilon}(y) - \delta\ln(1+|y-x|^2),$$

for any $y \in \overline{\Omega}$.

From (28.7), we have that $u - \Phi^{\epsilon,\delta}$ attains its minimum at a point $x^{\epsilon,\delta} \in \overline{\Omega}$. We distinguish two cases: either $x^{\epsilon,\delta} \in \Omega$ or $x^{\epsilon,\delta} \in \partial\Omega$.

If $x^{\epsilon,\delta} \in \Omega$, the fact that u is a viscosity solution implies that

$$\alpha u(x^{\epsilon,\delta}) + H(D^2\Phi^{\epsilon,\delta}(x^{\epsilon,\delta}), D\Phi^{\epsilon,\delta}(x^{\epsilon,\delta}), x^{\epsilon,\delta}) \geq 0\,. \tag{28.64}$$

Since

$$(u - \Phi^{\epsilon,\delta})(x^{\epsilon,\delta}) \leq (u - \Phi^{\epsilon,\delta})(x) = 0$$

we thus see that

$$\phi^{\epsilon}(x^{\epsilon,\delta}) \geq \Phi^{\epsilon,\delta}(x^{\epsilon,\delta}) \geq u(x^{\epsilon,\delta})\,.$$

The latter estimate and (28.64) imply that

$$\alpha\phi^{\epsilon}(x^{\epsilon,\delta}) + H(D^2\Phi^{\epsilon,\delta}(x^{\epsilon,\delta}), D\Phi^{\epsilon,\delta}(x^{\epsilon,\delta}), x^{\epsilon,\delta}) \geq 0\,. \tag{28.65}$$

Since

$$\sum_{i=1,2} |D^i\Phi^{\epsilon,\delta}(x^{\epsilon,\delta}) - D^i\phi^{\epsilon}(x^{\epsilon,\delta})|_{\infty} \leq C\,\delta$$

for some $C > 0$, we thus get, using (28.5), that

$$|H(D^2\Phi^{\epsilon,\delta}(x^{\epsilon,\delta}), D\Phi^{\epsilon,\delta}(x^{\epsilon,\delta}), x^{\epsilon,\delta}) - H(D^2\phi^{\epsilon}(x^{\epsilon,\delta}), D\phi^{\epsilon}(x^{\epsilon,\delta}), x^{\epsilon,\delta})| \leq \epsilon$$

as long as δ is conveniently small, possibly depending on ϵ. Therefore

$$\phi^{\epsilon}(x^{\epsilon,\delta}) + H(D^2\phi^{\epsilon}(x^{\epsilon,\delta}), D\phi^{\epsilon}(x^{\epsilon,\delta}), x^{\epsilon,\delta}) \geq -\epsilon\,,$$

for small δ, due to (28.65).

Thus, by taking $y = x^{\epsilon,\delta}$ in (28.63), we conclude that

$$\alpha\tilde{u}(x) - \epsilon \leq \frac{\epsilon(1 - e^{-\alpha T})}{\alpha} + e^{-\alpha T} \inf_{z\in\partial\Omega}\Big(\psi(z) - \phi^{\epsilon}(z)\Big) + u(x)\,.$$

for any $T > 0$.

By sending $T \to +\infty$ and then $\epsilon \to 0^+$, we conclude that $\tilde{u}(x) \leq u(x)$. This information and (28.62) complete the proof of Theorem 28.1 when $x^{\epsilon,\delta} \in \Omega$.

If, on the other hand, $x^{\epsilon,\delta} \in \partial\Omega$, we proceed as follows. First, we observe that

$$\begin{aligned}\psi(x^{\epsilon,\delta}) - \phi^{\epsilon}(x^{\epsilon,\delta}) &= u(x^{\epsilon,\delta}) - \Phi^{\epsilon,\delta}(x^{\epsilon,\delta}) - \delta\ln(1 + |x^{\epsilon,\delta} - x|^2)\\ &\leq u(x) - \Phi^{\epsilon,\delta}(x)\\ &= u(x) - \phi^{\epsilon}(x) = 0.\end{aligned}$$

As a consequence,

$$\inf_{z\in\partial\Omega}\Big(\psi(z) - \phi^{\epsilon}(z)\Big) \leq 0.$$

Thence, by taking $T = 0$ as candidate,

$$\begin{aligned}&\inf_{T\geq 0}\inf_{y\in\Omega}\frac{e^{-\alpha T} - 1}{\alpha}\Big(\alpha\phi^{\epsilon}(y) + H(D^2\phi^{\epsilon}(y), D\phi^{\epsilon}(y), y)\Big)\\ &\quad + e^{-\alpha T}\inf_{z\in\partial\Omega}\Big(\psi(z) - \phi^{\epsilon}(z)\Big)\end{aligned}$$

$$\begin{aligned}&\le 0 + \inf_{z\in\partial\Omega}\Big(\psi(z) - \phi^{\epsilon}(z)\Big)\\ &\le 0\,.\end{aligned}$$

Consequently, from (28.63),

$$\tilde{u}(x) - \epsilon \le u(x)$$

and so, by taking $\epsilon \to 0$, we conclude that $\tilde{u}(x) \le u(x)$. Recalling (28.62), this completes the proof of Theorem 28.1 also when $x^{\epsilon,\delta} \in \partial\Omega$.

28.5 Proof of Theorem 28.2

Given $x \in \mathbb{R}^n$, $t \in (0, T)$, we define $\tilde{v}(x,t)$ to be the infimum on the right hand side of (28.14).

We now apply Theorem 28.4 with $A_0 := -\mathrm{Id}$, $\mu_0 := \delta_{\{(x,t)\}}$, $\Omega_0 := \{(x,t)\}$, $A_1 := f \cdot D + (\sigma\sigma^T)/2 : D^2 + \partial_t$, $L_1 := L$, $\mu_1 := \mu$, $\Omega_1 := \mathbb{R}^n \times [t,T] \times \mathbb{R}^n$, $\gamma_1 := 1 + |x|^{q/2} + |v|^{q/2}$, $A_2 = -Id$, $L_2 := \psi$, $\Omega_2 := \mathbb{R}^n \times \{T\}$, $\gamma_2 := |x|^q$. Here, q is the exponent in (28.2).

We also take

$$X := C^2(\mathbb{R}^n \times [t,T]) \cap W^{2,\infty}(\mathbb{R}^n \times [t,T]).$$

Consider the set P to be the set of pairs of measures (μ, ν) on $\mathbb{R}^n \times [t,T] \times \mathbb{R}^n$ and on $\mathbb{R}^n \times \{T\}$, respectively, satisfying

$$\int_{\mathbb{R}^n\times[t,T]\times\mathbb{R}^n} d\mu = T - t \quad\text{and}\quad \int_{\mathbb{R}^n\times\{T\}} d\nu = 1. \tag{28.66}$$

Note that here P is of the form prescribed by (28.25). We remark that (28.15) implies (28.66), by taking $\phi(x,s) := 1$ and $\phi(x,s) := s$.

Thus, from Theorem 28.4, we obtain

$$\begin{aligned}\tilde{v}(x,t) = \sup_{\phi\in X}\,&(T-t)\inf_{\mathbb{R}^n\times[t,T]\times\mathbb{R}^n}\Big(L + f\cdot D\phi + \frac{\sigma\sigma^T}{2} : D^2\phi + \phi_t\Big)\\ &+ \inf_{\zeta\in\mathbb{R}^n}\Big(\psi(\zeta) - \phi(\zeta,T)\Big) + \phi(x,t)\,.\end{aligned}$$

Whereupon, from (28.4), we conclude

$$\begin{aligned}\tilde{v}(x,t) = \sup_{\phi\in X}\,&\inf_{\substack{y\in\mathbb{R}^n\\ s\in[t,T]}}(T-t)\Big(\phi_t(y,s) - H(D^2\phi(y,s), D\phi(y,s), y)\Big)\\ &+ \inf_{\zeta\in\mathbb{R}^n}\Big(\psi(\zeta) - \phi(\zeta,T)\Big) + \phi(x,t)\,.\end{aligned} \tag{28.67}$$

Thus, the proof of Theorem 28.2 will be completed once we show that $\tilde{v} = v$.

To this end, given $\epsilon > 0$, we consider v^ϵ as in (28.12) to be a candidate in (28.67) and we conclude, by sending ϵ to zero, that

$$\tilde{v}(x,t) \geq v(x,t), \tag{28.68}$$

and so, we now need to prove the reverse inequality.

Fix any $\epsilon > 0$, and let $\tilde{\phi}^\epsilon \in X$ be such that

$$\begin{aligned} \tilde{v}(x,t) - \epsilon \leq & \inf_{\substack{y\in\mathbb{R}^n \\ s\in[t,T]}} (T-t)\Big(\tilde{\phi}^\epsilon_t(y,s) - H(D^2\tilde{\phi}^\epsilon(y,s), D\tilde{\phi}^\epsilon(y,s), y)\Big) \\ & + \inf_{\zeta\in\mathbb{R}^n} \Big(\psi(\zeta) - \tilde{\phi}^\epsilon(\zeta,T)\Big) + \tilde{\phi}^\epsilon(x,t)\,. \end{aligned} \tag{28.69}$$

Note that, due to the property in (28.66) of the measures in P, the left hand side of (28.69) is not changed if, in the right hand side of (28.69), we would have replaced $\tilde{\phi}^\epsilon(y,s)$ with $\tilde{\phi}^\epsilon(y,s) + a + bs$, for any $a, b \in \mathbb{R}$.

Thence, suppose a^ϵ is fixed and set

$$b^\epsilon = \frac{v(x,t) - \tilde{\phi}^\epsilon(x,t) - a^\epsilon}{T-t},$$

and

$$\phi^\epsilon(y,s) = \tilde{\phi}^\epsilon(y,s) + a^\epsilon + b^\epsilon(T-s).$$

Then, we have that

$$\phi^\epsilon(x,t) = v(x,t)\,, \tag{28.70}$$

and if we choose a^ϵ suitably,

$$\inf_{\zeta\in\mathbb{R}^n} \psi(\zeta) - \phi^\epsilon(\zeta,T) = \epsilon. \tag{28.71}$$

Then, (28.70) and (28.71) imply that

$$\begin{aligned} \tilde{v}(x,t) - \epsilon \leq & \inf_{(y,s)\in\mathbb{R}^n\times[t,T]} \Big(\phi^\epsilon_t(y,s) - H(D^2\phi^\epsilon(y,s), D\phi^\epsilon(y,s), y)\Big) \\ & + \epsilon + v(x,t)\,. \end{aligned} \tag{28.72}$$

We now take $\delta > 0$ to be a small parameter (possibly smaller than ϵ) and let

$$\Phi^{\epsilon,\delta}(y,s) = \phi^\epsilon(y,s) - \delta\ln(1+|y-x|^2)$$

for any $y \in \mathbb{R}^n$ and $s \in [t,T]$.

By (28.13), we have that $v - \Phi^{\epsilon,\delta}$ takes its minimum at a point $(x^{\epsilon,\delta}, t^{\epsilon,\delta}) \in \mathbb{R}^n \times [t,T]$.

In fact, $t^{\epsilon,\delta} \neq T$, because, from (28.11), (28.71) and (28.70), for any $y \in \mathbb{R}^N$,

$$\begin{aligned} v(y,T) - \Phi^{\epsilon,\delta}(y,T) &= \psi(y) - \phi^{\epsilon}(y,T) + \delta \ln(1 + |y-x|^2) \\ &\geq \epsilon + 0 > 0 = v(x,t) - \Phi^{\epsilon,\delta}(x,t). \end{aligned}$$

This shows that $(x^{\epsilon,\delta}, t^{\epsilon,\delta}) \in \mathbb{R}^n \times [t,T)$.

Now we take two additional small positive parameters η and θ and we define

$$\Phi^{\epsilon,\delta,\eta}(y,s) = \Phi^{\epsilon,\delta}(y,s) - \eta \ln(1 + |y - x^{\epsilon,\delta}|^2 + |s - t^{\epsilon,\delta}|^2)$$

and

$$\Phi^{\epsilon,\delta,\eta,\theta}(y,s) = \Phi^{\epsilon,\delta,\eta}(y,s) - \frac{\theta}{s-t}.$$

Then, $v - \Phi^{\epsilon,\delta,\eta}$ has a strict minimum at $(x^{\epsilon,\delta}, t^{\epsilon,\delta})$ and so $v - \Phi^{\epsilon,\delta,\eta,\theta}$ has a local minimum at a point $(x^{\epsilon,\delta,\eta,\theta}, t^{\epsilon,\delta,\eta,\theta})$ close to $(x^{\epsilon,\delta}, t^{\epsilon,\delta})$ for small θ. In particular, by construction, $t^{\epsilon,\delta,\eta,\theta} \in (t,T)$.

Hence, since v is a viscosity solution,

$$-\Phi_t^{\epsilon,\delta,\eta,\theta}(\cdot) - \frac{\theta}{(s-t)^2} + H(D^2\Phi^{\epsilon,\delta,\eta,\theta}(\cdot), D\Phi^{\epsilon,\delta,\eta,\theta}(\cdot), \cdot) \geq 0$$

at the point $(x^{\epsilon,\delta,\eta,\theta}, t^{\epsilon,\delta,\eta,\theta}) \in \mathbb{R}^n \times (t,T)$. In particular,

$$-\Phi_t^{\epsilon,\delta,\eta,\theta}(\cdot) + H(D^2\Phi^{\epsilon,\delta,\eta,\theta}(\cdot), D\Phi^{\epsilon,\delta,\eta,\theta}(\cdot), \cdot) \geq 0$$

at the point $(x^{\epsilon,\delta,\eta,\theta}, t^{\epsilon,\delta,\eta,\theta}) \in \mathbb{R}^n \times (t,T)$.

Consequently, taking into account (28.72),

$$\tilde{v}(x,t) - 2\epsilon \leq v(x,t)$$

as long as the parameters $\theta \ll \eta \ll \delta$ are suitably small with respect to ϵ.

By sending $\epsilon \to 0^+$ and recalling (28.71), we conclude that $\tilde{v}(x,t) \leq v(x,t)$ and so, by (28.68), this completes the proof of Theorem 28.2.

Acknowledgements Diogo A. Gomes was partially supported by CAMGSD/IST through FCT Program POCTI-FEDER and by FCT grants POCI/FEDER/MAT/55745/2004 and PTDC/EEA-ACR/67020/2006. Enrico Valdinoci was partially supported by FIRB Project Analysis and Beyond.

References

1. Attouch, H.: Variational convergence for functions and operators. Applicable Mathematics Series. Pitman (Advanced Publishing Program), Boston, MA (1984)
2. Bardi, M., Capuzzo-Dolcetta, I.: Optimal control and viscosity solutions of Hamilton-Jacobi-Bellman equations. Birkhäuser Boston Inc., Boston, MA (1997). With appendices by Maurizio Falcone and Pierpaolo Soravia

3. Crandall, M.G., Ishii, H., Lions, P.-L.: User's guide to viscosity solutions of second order partial differential equations. Bull. Am. Math. Soc. (N.S.) **27**(1), 1–67 (1992)
4. Cho, M.J., Stockbridge, R.H.: Linear programming formulation for optimal stopping problems. SIAM J. Control Optim. **40**(6), 1965–1982 (electronic) (2002)
5. Fleming, W.H., Soner, H.M.: Controlled Markov Processes and Viscosity Solutions. Springer, New York (1993)
6. Fleming, W.H., Vermes, D.: Convex duality approach to the optimal control of diffusions. SIAM J. Control Optim. **27**(5), 1136–1155 (1989)
7. Helmes, K., Stockbridge, R.H.: Linear programming approach to the optimal stopping of singular stochastic processes. Stochastics **79**(3–4), 309–335 (2007)
8. Stockbridge, R.H.: Characterizing option prices by linear programs. In: Mathematics of Finance, volume 351 of Contemp. Math., pp. 349–359. AMS, Providence, RI (2004)
9. Villani, C.: Topics in optimal transportation, volume 58 of Graduate Studies in Mathematics. AMS, Providence, RI (2003)

Chapter 29
Microscopic Dynamics for the Porous Medium Equation

Patrícia Gonçalves

Abstract In this work, I present an interacting particle system whose dynamics conserves the total number of particles but with gradient transition rates that vanish for some configurations. As a consequence, the invariant pieces of the system, namely, the hyperplanes with a fixed number of particles can be decomposed into an irreducible set of configurations plus isolated configurations that do not evolve under the dynamics. By taking initial profiles smooth enough and bounded away from zero and one and for parabolic time scales, the macroscopic density profile evolves according to the porous medium equation. Perturbing slightly the microscopic dynamics in order to remove the degeneracy of the rates the same result can be obtained for more general initial profiles.

29.1 Introduction

The purpose of this work is to present the hydrodynamic limit for an non-ergodic interacting particle system. The non ergodicity translates by saying that each hyperplane with a fixed number of particles (which is a conserved quantity of the system) can be decomposed into a irreducible set of configurations plus isolated configurations that do not evolve under the dynamics. In contrast with erdogic systems it is not possible to pick randomly one configuration η from a certain hyperplane and get to any other configuration in the same hyperplane with jumps that are allowed by the dynamics. This is the main difficulty when establishing the hydrodynamic limit for this class of processes. The process considered here belongs to the class of *kinetically constrained lattice gases* (KCLG) which are used in physical literature to model liquid/glass and more general jamming transitions. In this context, the constraints are devised to mimic the fact that the motion of a particle in a dense medium can be inhibited by the geometrical constraints induced by the neighboring particles.

P. Gonçalves
Centro de Matemática da Universidade do Minho, Campus de Gualtar, 4710-057 Braga, Portugal
e-mail: patg@math.uminho.pt

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_29,

Here I present the hydrodynamic limit for a particle system associated to the porous medium equation. The process is of gradient type and is one of the simplest models in the KCLG class. The porous medium equation is given by $\partial_t \rho(t,u) = \partial_u^2 \rho^2(t,u)$ and it can be written in divergence form as $\partial_t \rho(t,u) = \nabla(D(\rho(t,u))\nabla(\rho(t,u)))$ with diffusion coefficient given by $D(\rho(t,u)) = 2\rho(t,u)$ and thus the equation looses the parabolic character as $\rho \to 0$. One of the properties of the solutions is that they can be compactly supported at each fixed time. A second observation is that the solutions of the equation can be continuous on the domain of definition, without being smooth at the boundary, see [4]. In the next section I will present a Markov process whose macroscopic density behavior $\rho : [0,T]\times\mathbb{T} \to [0,1]$ evolves according to the partial differential equation above, the so called hydrodynamic equation. Here $\mathbb{T}$ denotes the one-dimensional torus.

29.2 Markov Process

Let η_t be a continuous time Markov process with space state $\{0,1\}^{\mathbb{T}_N}$, where $\mathbb{T}_N$ denotes the one-dimensional discrete torus. For a site x on the microscopic space, $\eta(x)$ denotes the number of particles at that site and $\eta(x) = 1$ will have the physical meaning as the site x being occupied by a particle, while $\eta(x) = 0$ will denote a vacancy at that site. For a configuration η, $c(x,y,\eta)$ denotes the rate at which a particle jumps from x to y. We restrict to the case of nearest-neighbor jumps, so that $c(x,y,\eta) = 0$ if $|x-y| > 1$ and the exclusion rule, a particle at site x jumps to y if the site y is empty otherwise the jump is suppressed. The jump rates are degenerate and of gradient type, in fact we consider $c(x,x+1,\eta) = \eta(x-1) + \eta(x+2)$ and $c(x,x+1,\eta) = c(x+1,x,\eta)$. This Markov process has generator given on local functions $f : \{0,1\}^{\mathbb{T}_N} \to \mathbb{R}$ by

$$(\mathscr{L}_P f)(\eta) = \sum_{\substack{x,y\in\mathbb{T}_N\\|x-y|=1}} c(x,y,\eta)\eta(x)(1-\eta(y))(f(\eta^{x,y}) - f(\eta)). \tag{29.1}$$

In order to have a non-trivial temporal evolution of the density profile the process is evolving on the parabolic time scale tN^2. Since the jump rates are symmetric, the Bernoulli product measures $(\nu_\alpha)_\alpha$ in $\{0,1\}^{\mathbb{T}_N}$ are invariant and in fact reversible. This chosen rates define a *gradient system* since the instantaneous current $W_{0,1}(\eta) = c(0,1,\eta)\Big[\eta(0)(1-\eta(1)) - \eta(1)(1-\eta(0))\Big]$ can be rewritten as the gradient of a local function, namely $W_{0,1}(\eta) = h(\eta) - \tau_1 h(\eta)$, with $h(\eta) = \eta(0)\eta(1) + \eta(0)\eta(-1) - \eta(-1)\eta(1)$. The relation between h and the hydrodynamic equation is that $\partial_t \rho(t,u) = \partial_u^2 \tilde{h}(\rho(t,u))$ where $\tilde{h}(\rho) = E_{\nu_\rho}(h(\eta)) = \rho^2$.

29.3 Decomposition of the Space State

By the definition of the dynamics, the number of particles is obviously a preserved quantity, and as a consequence the state space can be decomposed into hyperplanes with a fixed number of particles, namely $\Sigma_{N,k} = \{\eta \in \{0,1\}^{\mathbb{T}_N} : \sum_{x \in \mathbb{T}_N} \eta(x) = k\}$. It is said that $\mathscr{O}$ is an irreducible component of $\Sigma_{N,k}$ if for every η, $\xi \in \mathscr{O}$ it is possible to go from η to ξ by jumps that are allowed by the dynamics. Since the dynamics is defined by the presence of particles in the neighboring positions to the site where the particle jumps, it is natural to have a critical density for which in a regime under that critical density some configurations, that do not evolve under the dynamics, arise. For this process if $k > N/3$, each hyperplane with k particles is not decomposable into smaller ergodic subsets; however, for $k \leq N/3$, each hyperplane is decomposable into an irreducible component (the set of configurations that contain at least one couple of particles at distance at most two) plus many irreducible sets: configurations that do not evolve under the dynamics – to which we call frozen.

29.4 Hydrodynamic Limit

To investigate the hydrodynamic limit, define the empirical measure by:

$$\pi_t^N(du) = \pi^N(\eta_t, du) = \frac{1}{N} \sum_{x \in \mathbb{T}_N} \eta_t(x) \delta_{\frac{x}{N}}(du). \tag{29.2}$$

Fix an initial profile $\rho_0 : \mathbb{T} \to [0,1]$ and denote by $(\mu^N)_N$ a sequence of probability measures on $\{0,1\}^{\mathbb{T}_N}$.

Definition 29.1. A sequence $(\mu^N)_N$ is associated to an initial profile ρ_0, if for every continuous function $H : \mathbb{T} \to \mathbb{R}$ and for every $\delta > 0$

$$\lim_{N \to +\infty} \mu^N \Big[\Big| \frac{1}{N} \sum_{x \in \mathbb{T}_N} H\Big(\frac{x}{N}\Big) \eta(x) - \int_{\mathbb{T}} H(u) \rho_0(u) du \Big| > \delta \Big] = 0. \tag{29.3}$$

We can translate the definition above by saying that a sequence of measures $(\mu^N)_N$ is associated to a profile ρ_0 if a Law of Large Number (in the weak sense) holds for the empirical measure at time $t = 0$ under the probability μ^N. We can rewrite (29.3) as

$$\lim_{N \to +\infty} \mu^N \Big[\Big| \int_{\mathbb{T}} H(u) \pi_0^N(du) - \int_{\mathbb{T}} H(u) \rho_0(u) du \Big| > \delta \Big] = 0. \tag{29.4}$$

The goal in hydrodynamic limit consists in showing that if at time $t = 0$ the empirical measures are associated to some initial profile ρ_0, then at time t they are associated to a profile ρ_t, where ρ_t is the solution of the hydrodynamic equation, then if a Law of Large Numbers holds for the empirical measure at time $t = 0$ then it holds at any time t. The hydrodynamic limit can be derived in two different ways. One is known as the Relative Entropy Method and it was first introduced by Yau [5], when proving the hydrodynamic limit for Ginzburg–Landau models. This method requires the existence of smooth solutions of the hydrodynamic equation. The second one is known as the Entropy Method and it is due to Guo et al. [2]. In contrast with the first method, this requires the uniqueness of weak solutions of the hydrodynamic equation. Before proceeding we recall the definition of a weak solution of the porous medium equation.

Definition 29.2. Fix a bounded profile $\rho_0 : \mathbb{T} \to \mathbb{R}$. A bounded function $\rho : [0,T] \times \mathbb{T} \to \mathbb{R}$ is a weak solution of the hydrodynamic equation, if for every function $H : [0,T] \times \mathbb{T} \to \mathbb{R}$ of class $C^{1,2}([0,T] \times \mathbb{T})$

$$\begin{aligned}\int_0^T dt \int_{\mathbb{T}} du \Big\{ \rho(t,u)\partial_t H(t,u) + (\rho(t,u))^2 \partial_u^2 H(t,u) \Big\} \\ + \int_{\mathbb{T}} \rho_0(u) H(0,u) du = \int_{\mathbb{T}} \rho(T,u) H(T,u) du.\end{aligned} \tag{29.5}$$

29.4.1 The Relative Entropy Method

Fix $\epsilon > 0$ and let $\rho_0 : \mathbb{T} \to [0,1]$ be a profile of class $C^{2+\epsilon}(\mathbb{T})$. By a well known result, the porous medium equation admits a solution denoted by $\rho(t,u)$ of class $C^{1+\epsilon,2+\epsilon}(\mathbb{R}_+ \times \mathbb{T})$. In order to apply the method, there is a technical condition that has to be assumed: the existence of a constant $\delta_0 > 0$ such that the profile is bounded away from 0 and 1: $\forall u \in \mathbb{T}$ it holds that $\delta_0 \le \rho_0(u) \le 1 - \delta_0$. Let $\nu^N_{\rho_0(.)}$ be the product measure in $\{0,1\}^{\mathbb{T}_N}$ such that $\nu^N_{\rho_0(.)}\{\eta, \eta(x) = 1\} = \rho_0(x/N)$. This means that for a fixed site $x \in \mathbb{T}_N$, $\eta(x)$ has Bernoulli distribution of parameter $\rho(0, x/N)$ and $(\eta(x))_x$ are independent. For two measures μ and ν in $\{0,1\}^{\mathbb{T}_N}$ define the relative entropy of μ with respect to ν as:

$$H(\mu/\nu) = \sup_f \Big\{ \int f d\mu - \log \int e^f d\nu \Big\}.$$

The supremum is taken over all continuous functions.

Theorem 29.1. *(G.L.T. [1]) Let* $\rho_0 : \mathbb{T} \to [0,1]$ *be a initial profile of class* $C^{2+\epsilon}(\mathbb{T})$ *that satisfies:*

$$\exists \delta_0 > 0 : \forall u \in \mathbb{T}, \quad \delta_0 \le \rho_0(u) \le 1 - \delta_0. \tag{29.6}$$

Let $(\mu^N)_N$ be a sequence of probability measures on $\{0,1\}^{\mathbb{T}_N}$ such that:

$$\lim_{N\to+\infty} \frac{H(\mu^N/\nu^N_{\rho_0(.)})}{N} = 0. \tag{29.7}$$

Then, for each $t \geq 0$

$$\pi^N_{tN^2}(du) \xrightarrow[N\to+\infty]{} \rho(t,u)du \tag{29.8}$$

in probability, where $\rho(t,u)$ is a smooth solution of the porous medium equation.

We remark that in the last result, there was made two assumptions on the initial profile in order to obtain the result (1) the bound condition

$$\exists \delta_0 > 0 : \forall u \in \mathbb{T}, \quad \delta_0 \leq \rho_0(u) \leq 1 - \delta_0 \tag{29.9}$$

and (2) the smoothness of class $C^{2+\epsilon}(\mathbb{T})$. This is too restrictive, since one could want to analyze profiles that are (for example) indicator functions over a certain set. On the other hand, the Entropy Method relies on the full irreducibility of the Markov process when restricted to a hyperplane. We have seen that the process defined above when restricted to a hyperplane with a low density of particles it is not fully irreducible – the frozen states arise. To overcome this problem, the idea is to perturb slightly the dynamics in such a way that the frozen states disappear but the macroscopic density profile still evolves according to the porous medium equation.

29.4.2 The Entropy Method

In this section we present the hydrodynamic limit for a slightly different dynamics, in which each hyperplane is a unique ergodic piece and whose macroscopic density profile still evolves according to the porous medium equation. Here we follow the strategy described in [3]. Due to the non-ergodicity, the main difficulty to overcome using this approach is the Replacement Lemma. The interested reader can find more details in [1]. For $\theta > 0$, consider a Markov process with generator given by

$$\mathscr{L}_\theta = \mathscr{L}_P + N^{\theta-2}\mathscr{L}_S$$

where $\mathscr{L}_P$ was defined above and $\mathscr{L}_S$ is the generator of the Symmetric Simple Exclusion process:

$$(\mathscr{L}_S f)(\eta) = \sum_{\substack{x,y\in\mathbb{T}_N\\|x-y|=1}} \frac{1}{2}\eta(x)(1-\eta(y))(f(\eta^{x,y}) - f(\eta)),$$

The Bernoulli product measures $(\nu_\alpha)_\alpha$ are invariant for $\mathscr{L}_\theta$ since they are invariant measures for $\mathscr{L}_P$ and $\mathscr{L}_S$. Its also easy to show that the Markov processes with generators $\mathscr{L}_P$ and $\mathscr{L}_\theta$, have the same hydrodynamic equation as long as $\theta < 2$. This restriction on θ comes from the fact that we want to perturb sightly the dynamics microscopically in order to destroy the frozen configurations, but we do not want to see the effect of this perturbation macroscopically, and for that the Symmetric Simple Exclusion Process has to be speeded up on a time scale less than the parabolic one. Nevertheless, while the former process has each hyperplane (on the low density regime) decomposed in many ergodic or irreducible pieces, which is a consequence of the existence of the frozen states, the latter has each hyperplanes $\{\Sigma_{N,k} : k = 0, \dots, N\}$ as a unique ergodic piece. Then the Entropy Method can be applied to the process with generator $\mathscr{L}_\theta$.

For that, denote by $\mathbb{P}_\mu$ the probability measure on $D([0,T], \{0,1\}^{\mathbb{T}_N})$, induced by the Markov process with generator $\mathscr{L}_\theta$, speeded up by N^2 and with initial measure μ.

Theorem 29.2. *(G.L.T. [1])*

Let $\rho_0 : \mathbb{T} \to [0,1]$ and $(\mu^N)_N$ be a sequence of probability measures on $\{0,1\}^{\mathbb{T}_N}$ ***associated to the profile*** *ρ_0. Then, for every $0 \le t \le T$, for every continuous function $H : \mathbb{T} \to \mathbb{R}$ and for every $\delta > 0$,*

$$\lim_{N\to+\infty} \mathbb{P}_{\mu^N}\Big[\Big|\frac{1}{N}\sum_{x\in\mathbb{T}_N} H\Big(\frac{x}{N}\Big)\eta_t(x) - \int_{\mathbb{T}} H(u)\rho(t,u)du\Big| > \delta\Big] = 0, \quad (29.10)$$

where $\rho(t,u)$ is the unique weak solution of the porous medium equation.

References

1. Gonçalves, P., Landim, C., Toninelli, C.: Hydrodynamic Limit for a Particle System with degenerate rates. Ann. Inst. Henri Poincaré: Prob. Stat. **45**(4), 887–909 (2009)
2. Guo, M.Z., Papanicolau, G.C., Varadhan, S.R.S.: Nonlinear diffusion limit for a system with nearest neighbor interactions. Commun. Math. Phys. **118**, 31–59 (1988)
3. Kipnis, C., Landim, C.: Scaling Limits of Interacting Particle Systems. Springer, New York (1999)
4. Vazquez, J.L.: An introduction to the mathematical theory of the porous medium equation. In: Delfour, M.C., Sabidussi, G. (eds.) Shape Optimization and Free Boundaries, pp. 261–286. Kluwer, Dordrecht (1992)
5. Horng-Tzer, Y.: Relative Entropy and Hydrodyamics of Ginzburg-Landau Models. Lett. Math. Phys. **22**, 63–80 (1991)

Chapter 30
A Stochastic Model for Wolf's Sunspot Number

Rui Gonçalves and Alberto A. Pinto

Abstract We present a simplified cycle model, using the available data, for the monthly sunspot number random variables $\{X_t\}_{t=1}^{133}$, where 133 is taken as the mean duration of the Schwabe's cycle. We present a fit for the mean and standard deviation of X_t. In the descending and ascending phases, we analyse the probability histogram of the monthly sunspot number fluctuations.

30.1 Introduction

A sunspot is a region on the Sun's surface (photosphere) that is marked by a lower temperature than its surroundings and has intense magnetic activity. The sunspots are an indicator of solar activity especially of ultraviolet emission that heats up the Earth's atmosphere expanding it and consequently increasing the drag force on satellites. In 1848, the Swiss astronomer Johann Rudolph Wolf introduced a daily measurement of sunspot number. His method, which is still used today, counts the total number of spots visible on the face of the sun and the number of groups into

R. Gonçalves (✉)
LIAAD-INESC Porto LA
and
Section of Mathematics, Faculty of Engineering, University of Porto,
R. Dr. Roberto Frias s/n, 4200-465 Porto, Portugal
e-mail: rjasg@fe.up.pt

A.A. Pinto
LIAAD-INESC Porto LA e Departamento de Matemática, Faculdade de Ciências,
Universidade do Porto, Rua do Campo Alegre, 687, 4169-007, Portugal
and
Centro de Matemática e Departamento de Matemática e Aplicações, Escola de Ciências,
Universidade do Minho, Campus de Gualtar, 4710-057 Braga, Portugal
e-mail: aapinto@fc.up.pt

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings
in Mathematics 2, DOI 10.1007/978-3-642-14788-3_30,

which they cluster. The observed number of sunspot cycles, up to now, is 23.[1] The solar maximum and solar minimum refer, respectively, to epochs of maximum and minimum sunspot counts. The rule for the maximum and minimum (end of the cycle) is the maximum and minimum, respectively, of the smoothed sunspot numbers. Each cycle is divided in an ascending phase (rise) ranging from the start of the cycle to solar maximum with an average duration of 4.7 years and a descending phase (fall) from solar maximum to minimum with an average duration of 6.3 years. It was Schwabe [19] who first suggested a probable period of 10 years (i.e. at every 10th year the number of spots reached a maximum). The average duration of the sunspot cycle is 133 months (11.08 years). The physical basis of the solar cycle was studied by George Ellery Hale and co-workers. Babcock [1] proposed a qualitative model for the dynamics of the solar outer layers giving an explanation for the appearance of sunspots. Lu et al. [2, 15] observed some universal properties of the solar magnetic activity. In [12], the average duration of the sunspots cycle is used to present a simplified heuristic model for what we call the monthly sunspot number random variables $\{X_t\}_{t=1}^{133}$ characterizing the sunspot number at month t. Our starting point is the beginning of the monthly sunspot number count, i.e. January of 1749. Curiously, we observed, for this simplified model, that the mean of the monthly sunspot number random variables X_t, along the cycle, consists of two almost equally sized descending and ascending phases. These phases can be well fitted to two lines that are close to orthogonal. This can be an indication that there is a well defined periodic, or quasi-periodic, cycle with period close to 133 in the physical phenomenon that creates the sunspot numbers and the apparent different durations from cycle to cycle are indeed oscillations (see also [16, 18]). The standard deviation of the monthly sunspot number random variables X_t is well fitted by the first subharmonic of the Fourier series. In the descending phase, the histograms of the monthly sunspot numbers fluctuations is close to the BHP pdf (see Bramwell et al. [3–5] and, for other applications of the BHP pdf, see [6–12] and [14]). In the ascending phase, two periods will be considered. The first ascending period occurs for the cycle months between 62 and 100, where the standard deviation period is higher than the mean cycle. The second ascending period occurs for the cycle months between 101 and 133, where the standard deviation is smaller than the mean period. In the first ascending period, the histogram of the monthly sunspot number fluctuations is close to the BHP pdf, but deviating for values close to the center of the universal BHP distribution. In the second ascending period, the histogram is close to the BHP. Since the ascending and descending phases have different characteristics, we analyze separately the two phases. After synchronizing the beginning of the ascending phases and the beginning of the descending phases, for all cycles, we show the data collapse of the fluctuations to the BHP pdf (see [14]). In this chapter, we survey the results presented in [12] and [14].

[1] The data and related information on the sunspot numbers is available at the Solar Data Services site, http://www.ngdc.noaa.gov/stp/SOLAR/SSN/ssn.html. of the *National Geophysical Data Center*

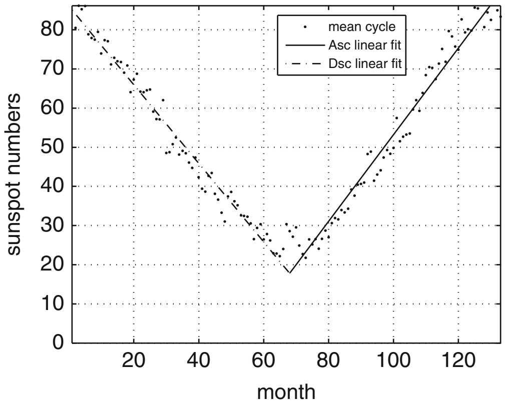


Fig. 30.1 Sunspot numbers mean period $w_\mu(t)$

30.2 Universality in Sunspot Number Fluctuations

We start by estimating the mean of the monthly sunspot number random variable X_t, using the *monthly sunspot number mean* $w_\mu(t)$ given by

$$w_\mu(t) = \frac{1}{T}\sum_{j=0}^{T-1} w(t + j * 133), \tag{30.1}$$

where $T = 23$ is the number of observed cycles.

In Fig. 30.1, we show the sunspot number mean period $w_\mu(t)$ with the mean square lines for the descending and ascending phase, respectively. The fits to the ascending and descending phase of the mean period curve $h(t)$ are two, approximately, orthogonal lines. The equation for the descending phase fit is $y = -1.0136t + 86.131$ with $R^2 = 0.98$ (percentage of variance explained by the line), and the equation for the ascending phase fit is $y = 1.0767t - 54.046$ with $R^2 = 0.96$. The small region in the neighborhood of the minimum values, approximately between the cycle months 65 and 73, was not used in the fitting process. We estimate the standard deviation of the monthly sunspot number random variable X_t, using the *monthly sunspot number standard deviation* $w_\sigma(t)$ given by

$$w_\sigma(t) = \sqrt{\frac{\sum_{j=0}^{T-1} w(t + j * 133)^2}{T} - w_\mu(t)^2}, \tag{30.2}$$

where T is the number of observed cycles. In Fig. 30.2, we fit the sunspot number standard deviation period curve with the first sub-harmonic of the Fourier series $y(t) = 38.3 - 1.0670 * \sin(0.0472t) + 16.77 * \cos(0.0472t)$ with $R^2 = 0.988$. The standard deviation curve attains its minimum at approximately the cycle month 40, and its maximum at, approximately, the cycle month 108. We define the *monthly sunspot number fluctuations* $w_f(t)$ by

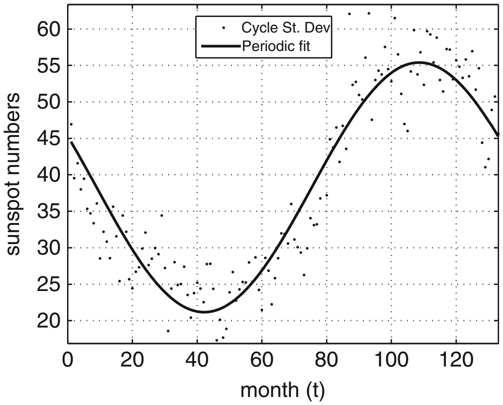


Fig. 30.2 Sunspot number standard deviations $w_\sigma(t)$

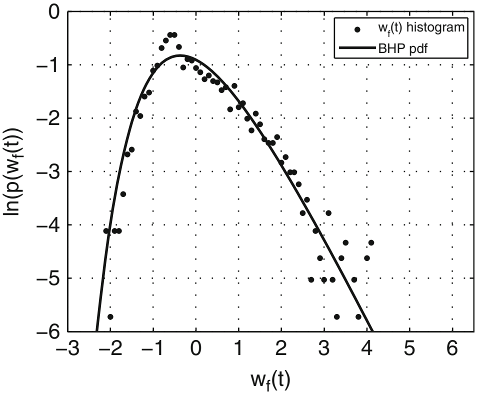


Fig. 30.3 Histogram of the sunspots number fluctuations for the mean period

$$w_f(t) = \frac{w(t) - w_\mu(t)}{w_\sigma(t)}. \tag{30.3}$$

We use the histogram of the monthly sunspot number fluctuations $w_f(t)$ as an approximation of the normalized monthly sunspot number random variable X_t pdf. In Fig. 30.3, we observe that the histogram of the monthly sunspot number fluctuations show some differences to the BHP pdf. When compared to the BHP pdf, small negative fluctuations are more common and small positive fluctuations are less common.

In Fig. 30.4, we show the data collapse of the sunspot number fluctuations histogram to the BHP pdf, in the descending phase occurring for the cycle months between 1 and 61.

In the ascending phase, two periods will be considered. The first ascending period occurs for the cycle months between 62 and 100, where the standard deviation period is higher than the mean cycle, and the second ascending period occurs for the cycle months between 101 and 133, where the standard deviation is smaller than the mean period (see Fig. 30.2). In the first ascending period, we observe

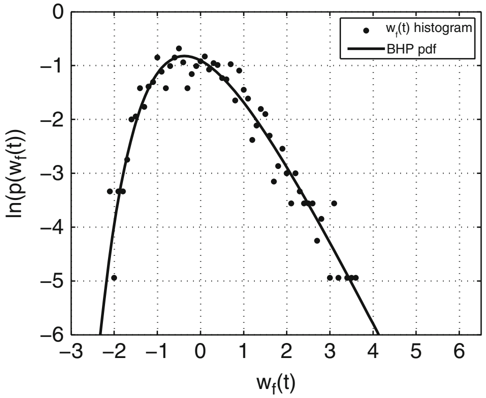


Fig. 30.4 Histogram of the sunspots number fluctuations for months in the range 1–61 of the cycle with the BHP pdf on top, in a semi-log plot

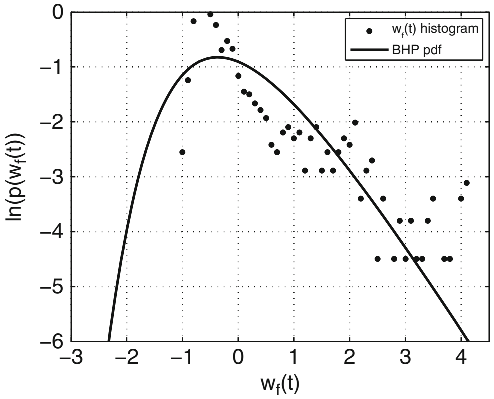


Fig. 30.5 Histogram of the sunspots number fluctuations for months in the range 62–100 of the cycle with the BHP pdf on top, in a semi-log plot

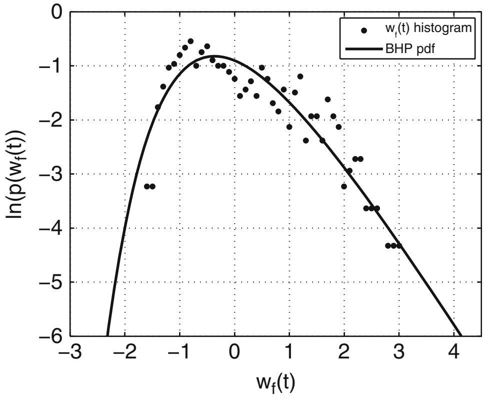


Fig. 30.6 Histogram of the sunspot number fluctuations for months in the range 101–133 of the cycle with the BHP pdf on top, in a semi-log plot

that the histogram of the sunspot number fluctuations deviates from the BHP pdf for small positive fluctuations (see Fig. 30.5). In the second ascending period, we observe that the histogram of the sunspot number fluctuations is close to the BHP pdf (see Fig. 30.6). Since the histogram of the sunspot number fluctuations for the ascending and descending phases have different characters, we next analyze separately the two phases, by synchronizing the beginning of the ascending phases and the beginning of the descending phases, for all cycles. Let M_k be the *month corresponding to the maximum value* X_{M_k} of the monthly sunspot number kth cycle for $k \in \{1, \ldots, 23\}$. Let m_k be the *month corresponding to the minimum value* X_{m_k} of the monthly sunspot number kth cycle. The *duration* a_k of the ascending phase of the kth sunspot cycle is given by $a_k = m_k - M_k$. The *kth ascending phase variable* A_t^k is defined by

$$A_t^k = X_{t+m_k},$$

where $t \in \{0, \ldots, a_k\}$. Let $\mathscr{A}(t)$ denote the set of all k's such that the ascending phase A_t^k has durations a_k higher than t, i.e.

$$\mathscr{A}(t) = \{k : t \leq a_k\}.$$

Let $\mathscr{T}^a$ be the *minimum t* subjected to $\#\mathscr{A}(t) > 1$, i.e.

$$\mathscr{T}^a = \max\{t : \#\mathscr{A}(t) > 1\}.$$

Hence, there are at least two ascending phases t months long, for every $t \leq \mathscr{T}^a$. We define the *ascending mean* μ_t^a by

$$\mu_t^a = \frac{1}{\#\mathscr{A}(t)} \sum_{k \in \mathscr{A}(t)} A_t^k,$$

where $t \in \{0, \ldots, \mathscr{T}^a\}$ (see Fig. 30.7). We define the *ascending standard deviation* σ_t^a by

$$\sigma_t^a = \sqrt{\frac{1}{\#\mathscr{A}(t)} \sum_{k \in \mathscr{A}(t)} (A_t^k - \mu_t^a)^2},$$

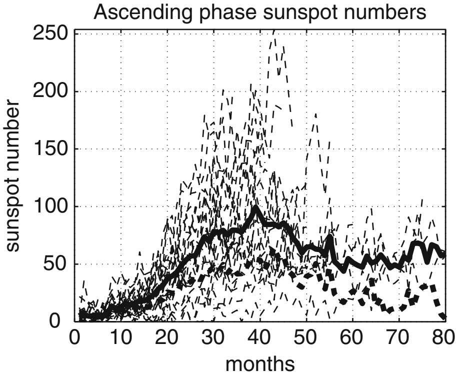


Fig. 30.7 Ascending phases A_t^k of the sunspot cycles and respective mean (*full line*) and standard deviation (*dotted line*)

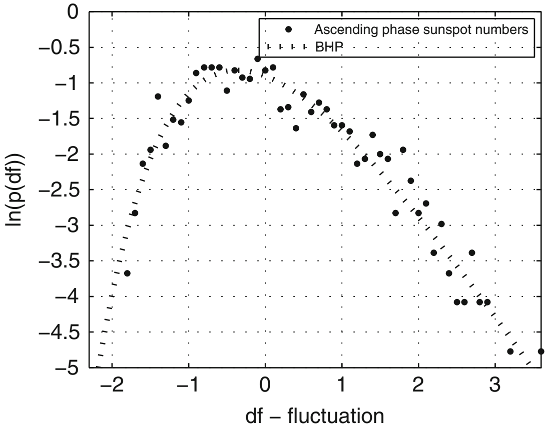


Fig. 30.8 Histogram of the aggregated ascending fluctuations $A^f_{t,k}$ of the sunspot cycles

where $t \in \{0, \ldots, \mathscr{T}^a\}$. For each $t \in \mathscr{T}^a$, we define the *ascending fluctuation variables* $A^f_{t,k}$ by

$$A^f_{t,k} = \frac{A^k_t - \mu^a_t}{\sigma^a_t},$$

where $k \in \{1, \ldots, 23\}$ and $t \in \{0, \ldots, \mathscr{T}^a\}$. The ascending fluctuation variables $A^f_{t,k}$ measure the deviations of the sunspot's ascending phases A^k_t to the ascending mean μ_t in standard deviation σ^a_t units. Surprisingly, the histogram of the aggregated ascending observed fluctuations shows a data collapse to the universal nonparametric BHP pdf (see Fig. 30.8).[2] In particular, the histogram of the ascending fluctuation variables $A^f_{t,k}$ do not follow a gaussian distribution, exhibiting heavy tails and a universal non-zero skewness. The highest observed positive fluctuation $A^f_{t,k}$ is equal to 3.604 and the lowest observed negative fluctuation $A^f_{t,k}$ is -1.894 showing the asymmetries of the histogram. We get an estimator for the sunspot number

$$A^k_t = \sigma^a_t A^f_{t,k} + \mu^a_t, \tag{30.4}$$

using the ascending mean μ^a_t and the ascending standard deviation σ^a_t (see Fig. 30.7) and noting that $A^f_{t,k}$ follows the universal nonparametric BHP pdf (see Fig. 30.8). We observe that the highest ascending means μ^a_t occur together with the highest standard deviations σ^a_t, for values of t close to 44. Hence, by (30.4), the highest sunspot numbers A^k_t occur for values of t close to 44.

30.3 Conclusions

We gave a simplified heuristic model for the monthly sunspot number random variables $\{X_t\}_{t=1}^{133}$. We observed that the mean of X_t, along the cycle, consists of two approximately equally sized descending and ascending phases. These phases are

[2] The results are similar for the descending case (see [14]).

well fitted by two almost orthogonal lines. The standard deviation of the random variables X_t is well fitted by the first Fourier subharmonic. In the descending phase, we discovered that the histogram of the monthly sunspot number fluctuations is close to the BHP pdf. In the ascending phase, two periods were considered. The histogram of the monthly sunspot number fluctuations is closer to the BHP pdf in the second period than the first period. Since the ascending and descending phases have different characters, we analyzed separately, the two phases. After synchronizing the beginning of the ascending phases and the beginning of the descending phases, for all cycles, we observed the data collapse of the fluctuations to the BHP pdf (see [14]).

Acknowledgements We thank Nico Stollenwerk, in particular, for showing us the relevance of the BHP distribution. We would like to thank Robert Wilson for giving us very useful information on the sunspot numbers. We thank LIAAD-INESC Porto LA, Calouste Gulbenkian Foundation, PRODYN-ESF, FEDER, POFC, POCTI and POSI by FCT and Ministério da Ciência e da Tecnologia, and the FCT Pluriannual Funding Program of the LIAAD-INESC Porto LA and of the Research Centre of Mathematics of the University of Minho, for their financial support.

References

1. Babcock, H.W.: The topology of the sun's magnetic field and the 22-year cycle. Astrophys. J. **133**(2), 572–587 (1961)
2. Bak, P., Tang, C., Wiesenfeld, K.: Self-organized criticality. Phys. Rev. Lett. **A38**, 364–374 (1988)
3. Bramwell, S.T., Christensen, K., Fortin, J.Y., Holdsworth, P.C.W., Jensen, H.J., Lise, S., López, J.M., Nicodemi, M., Sellitto, M.: Universal fluctuations in correlated systems. Phys. Rev. Lett. **84**, 3744–3747 (2000)
4. Bramwell, S.T., Fortin, J.Y., Holdsworth, P.C.W., Peysson, S., Pinton, J.F., Portelli, B., Sellitto, M.: Magnetic Fluctuations in the classical XY model: the origin of an exponential tail in a complex system. Phys. Rev. E **63**, 041106 (2001)
5. Bramwell, S.T., Holdsworth, P.C.W., Pinton, J.F.: Universality of rare fluctuations in turbulence and critical phenomena. Nature **396**, 552–554 (1998)
6. Gonçalves, R., Ferreira, H., Stollenwerk, N., Pinto, A.A.: Universal fluctuations of AEX index. Physica A **389**(21), 4776–4784 (2010)
7. Gonçalves, R., Ferreira, H., Pinto, A.A., Stollenwerk, N.: Universality in nonlinear prediction of complex systems. J. Dif. Equ. Appl. **15**, 1067–1076 (2009)
8. Gonçalves, R., Ferreira, H., Pinto, A.A., Stollenwerk, N.: Universality in the Stock Exchange Market. J. Dif. Equ. Appl. **17**(6), 35–39 (2011)
9. Gonçalves, R., Ferreira, H., Pinto, A.A.: Universality in energy sources. (submitted)
10. Gonçalves, R., Ferreira, H., Pinto, A.A.: Universal fluctuations of the Dow Jones. (submitted)
11. Gonçalves, R., Ferreira, H., Pinto, A.A.: A qualitative and quantitative Econophysics stock market model (submitted)
12. Gonçalves, R., Pinto, A.A.: BHP universality and gaussianity in sunspot numbers fluctuations. arXiv:0802.2880 4 physics.data-an (2008)
13. Gonçalves, R., Pinto, A.A.: Negro and Danube are mirror rivers. J. Dif. Equ. Appl. (to appear)
14. Gonçalves, R., Pinto, A.A., Stollenwerk, N.: Cycles and universality in sunspot numbers fluctuations. Astrophys. J. **691**, 1583–1586 (2009)
15. Lu, E., Hamilton, R.: Avalanches and distributions of solar flares. Astrophys. J. **380**, L89–L92 (1991)

16. Paluš, M., Novotná, D.: Sunspot cycle, a nonlinear oscillator?, Phys. Rev. Lett. **83**(17), 3406–3409 (1999)
17. Pinto, A.A.: Game Theory and Duopoly Models. Interdisciplinary Applied Mathematics Series. Springer (in conclusion)
18. Rabin, D., Wilson, R., Moore, R.: Bimodality of the solar cycle. Geophys. Res. Lett. **13**, 352–354 (1986)
19. Schwabe, H.: Die Sohne. Astron. Nachr. **20**, 283–286 (1843)

Chapter 31
On Topological Classification of Morse–Smale Diffeomorphisms

Viacheslav Grines and Olga Pochinka

Abstract Well-known results on topological classification of Morse–Smale flows were obtained by Leontovich and Maier (Dokl Akad Nauk 103(4):557–560, 1955) for flows on the two dimensional sphere, and by Mauricio Peixoto (Ann Math 69:199–222, 1959; On a classification of flows on 2-manifolds. Proc. Symp. Dyn. Syst. Salvador 389–492, 1973) for flows on any closed surfaces. Since 1980s rather great progress was achieved in classification of Morse–Smale diffeomorphisms on surfaces. For such diffeomorphisms with finite number heteroclinic orbits there is complete invariant in the form of a graph (similar to that introduced by Peixoto for flows). This graph is defined by taking into account the heteroclinic intersections and it is equipped with a graph automorphism induced by the given diffeomorphism (see for example the surveys (Bonatti et al. Comput Appl Math 20(1–2):11–50, 2001; Grines J Dyn Control Syst 6(1):97–126, 2000) for references and details). Describing of Morse–Smale diffeomorphisms with infinite set of heterolinic orbits uses Markov chains endowed by additional information (see Bonatti et al. Comput Appl Math 20(1–2):11–50, 2001). A progress in dimension 3 is based on rather recent results on finding new topological knot and link invariants which describe (possibly, wild) embedding of invariant manifolds of saddle periodic points into the ambient manifold. These invariants allowed to discover a principal distinctive phenomenon of Morse–Smale diffeomorphisms in dimension 3: the existence of a countable set of non-conjugate Morse–Smale diffeomorphisms with isomorphic Peixoto graphs. The main goal of the present survey is to give an exposition of recent results on classification of Morse–Smale diffeomorphis on 3-manifolds.

V. Grines (✉) and O. Pochinka
Nizhny Novgorod State University, 23 Gagarin Ave, Nizhny Novgorod 603950, Russia
e-mail: vgrines@yandex.ru, olga-pochinka@yandex.ru

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_31,

31.1 Morse–Smale Systems

31.1.1 Introduction

Following to A.A. Andronov and L.S. Pontryagin [1], S. Smale [53] introduced in 1960 a class of dynamical systems, later called Morse–Smale systems. This class was introduced as a candidate for all structurally stable flows in dimension $n \geq 2$ and moreover forming open dense set in the space of vector fields equipped by C^1-topology. S. Smale understood himself very soon that his hypotheses are not true for flows on manifolds with dimension $n > 2$ but however the class of Morse–Smale systems are intensively investigated as a class of structurally stable systems with a simplest behavior of trajectories. The concept of structural stability is a generalization of the concept of roughness which was introduced by A.A. Andronov and L.S. Pontryagin in [1] where necessary and sufficient conditions for roughness for flows on bounded part of the plane was obtained and was shown that the set of rough flows is dense in the space of C^1 flows. M.C. and M.M. Peixoto introduced in [45] the concept of structural stability and in 1958–1962 M. Peixoto gives necessary and sufficient conditions for the structural stability of flows on two-dimensional manifolds [46, 47] and proved that such flows form open and dense set in the space of C^1-flows. According to the Andronov–Pontryagin–Peixoto results, the non-wandering set of a structurally stable vector field on a compact 2-manifold consists of a finitely many hyperbolic fixed points and hyperbolic periodic orbits. Moreover, there are no separatrix connections (including loops). It is well known that characterization of structurally stable flows in higher dimensions is more complicated and was finished much later thanks to results by Anosov, Smale, Palis, Robbin, Robinson, Mañe, Hayashi.

Let us recall some definitions and concepts which we will use in our survey (see [2, 42, 51, 54] for more detail information). Let $f : M \to M$ be a diffeomorphism of a closed n-manifold M. An *invariant set* of f is a subset $\Lambda \subset M^n$ such that $f(\Lambda) = \Lambda$.

A point $x \in M^n$ is *non-wandering* if for any neighborhood U of x, $f^k(U) \cap U \neq \emptyset$ for infinitely many integers k. Then $\Omega(f)$, the *non-wandering set* of f, defined as the set of all non-wandering points, is an f-invariant closed set. Obviously, the set *Per* (f) of periodic points belongs to $\Omega(f)$.

Suppose now that M is smooth orientable Riemann manifold endowed with a metric ρ. An invariant set Λ is called *hyperbolic* is there is a continuous df-invariant splitting of the tangent bundle TM_Λ into *stable* and *unstable bundles* $E^s_\Lambda \oplus E^u_\Lambda$, $\dim E^s_x + \dim E^u_x = n = \dim M$ $(x \in \Lambda)$, with

$$\|\mathrm{d}f^i(v)\| \leq C_s\lambda^i\|v\|, \;\; v \in E^s_\Lambda, \; \|\mathrm{d}f^{-i}(w)\| \leq C_u\lambda^i\|w\|, \;\; w \in E^u_\Lambda, \; i \in \mathbb{N}$$

for some fixed $C_s > 0$, $C_u > 0$, $0 < \lambda < 1$. The hyperbolic structure implies the existence of stable and unstable manifolds $W^s(x)$, $W^u(x)$ respectively passing through any point $x \in \Lambda$:

$$W^s(x) = \{y \in M : \lim_{j \to \infty} \rho(f^j(x), f^j(y)) \to 0\},$$

$$W^u(x) = \{y \in M : \lim_{j \to \infty} \rho(f^{-j}(x), f^{-j}(y)) \to 0\}$$

which are smooth, injective immersions of the E^s_x and E^u_x into M. Moreover, $W^s(x)$, $W^u(x)$ are homeomorphic to Euclidean spaces $\mathrm{R}^{\dim W^s(x)}$, $\mathrm{R}^{\dim W^u(x)}$ and tangent to E^s_x and E^u_x at x respectively. Invariant manifolds depend continuously on initial conditions on compact sets. We set $W^s_A = \bigcup\limits_{a \in A} W^s(a)$ and $W^u_A = \bigcup\limits_{a \in A} W^u(a)$ for any subset A of the hyperbolic set Λ.

A diffeomorphism $f : M \to M$ satisfies *axiom A* (f is A-diffeomorphism) if its non-wandering set $\Omega(f)$ is hyperbolic, and the set $Per(f)$ is dense in $\Omega(f)$.

Moreover f is said to satisfy the *strong transversality condition* if, for all points $x, y \in \Omega(f)$, the stable manifold $W^s(x)$ is transverse (at all the intersection points) to the unstable manifold $W^u(y)$.

Let *Diff*^{r} (M) be the space of C^r diffeomorphisms endowed with the uniform C^r topology. Two diffeomorphisms $f : M \to M, g : M \to M$ are called *topologically conjugate* if there is a homeomorphism $h : M \to M$ such that $h \circ f = g \circ h$. A diffeomorphism f is said to be *structurally stable* if there exists a neighborhood U of f in *Diff*(M) such that every $g \in U$ is conjugate to f.

As well-known now axiom A and strong transversality condition are necessary and sufficient conditions for structural stability of diffeomorphism.

Among structurally stable diffeomorphisms Morse–Smale diffeomorphisms have the most simple type of structure of trajectories. Namely, a diffeomorphism f is called *Morse–Smale*, if $\Omega(f)$ is hyperbolic and finite (hence, $\Omega(f) = Per\ (f)$) and $W^s(x)$ is transverse to $W^u(y)$ for every $x, y \in \Omega(f)$.

A hyperbolic periodic point p is called *source (repellent)*, *sink (attractive)* if $\dim\ W^u(p) = n$, $\dim\ W^u(p) = 0$, accordingly. In opposite case, p is called *saddle point*. A connected component of $W^u(p) \setminus p$ ($W^s(p) \setminus p$) is called *unstable (stable) separatrix* of saddle point p.

Classification of structurally stable cascades on circles was obtained by A. Mayer in [36]. Results of that paper were are independently repeated by V. Arnold [5] and V. Pliss [49]. The basic achievement was the proof of that non-wandering set structurally stable diffeomorphism of circle consists of finite number of repellent and attractive periodic points. Thus, topological classification of such diffeomorphism is rather trivial from the modern point of view.

Let $n \geq 2$ and f be a Morse–Smale diffeomorphism. Let p, q be saddle periodic points of f for which $W^u(p) \cap W^s(q) \neq \emptyset$, then according to [54] we write $q \prec p$ and call a diffeomorphism f *gradient-like*, if the condition $q \prec p$ implies $\dim\ W^s(p) < \dim\ W^s(q)$.

If $W^u(p) \cap W^s(q) \neq \emptyset$ and $\dim\ W^s(p) = \dim\ W^s(q)$, then from the transversality of the intersection of $W^u(p)$ with $W^s(q)$ it follows that $W^u(p) \cap W^s(q)$ is a countable set. Each point of this set is called a *heteroclinic point* of the diffeomorphism f (see Fig. 31.1).

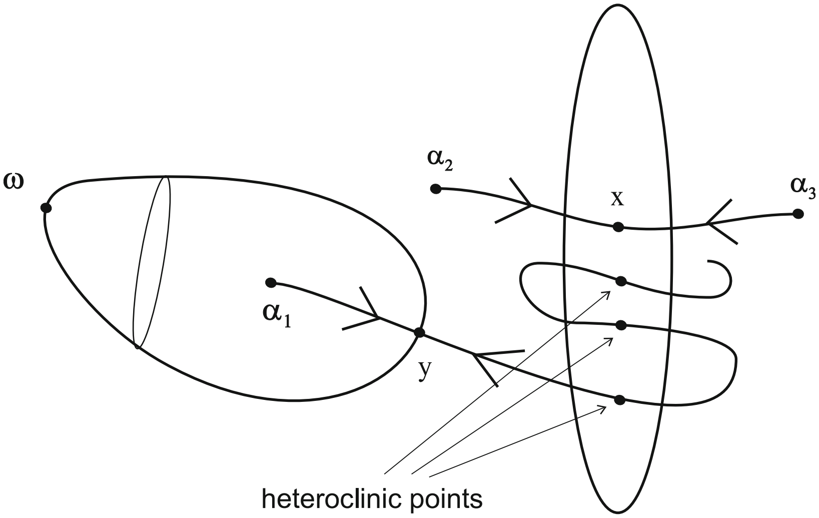


Fig. 31.1 Heteroclinic points

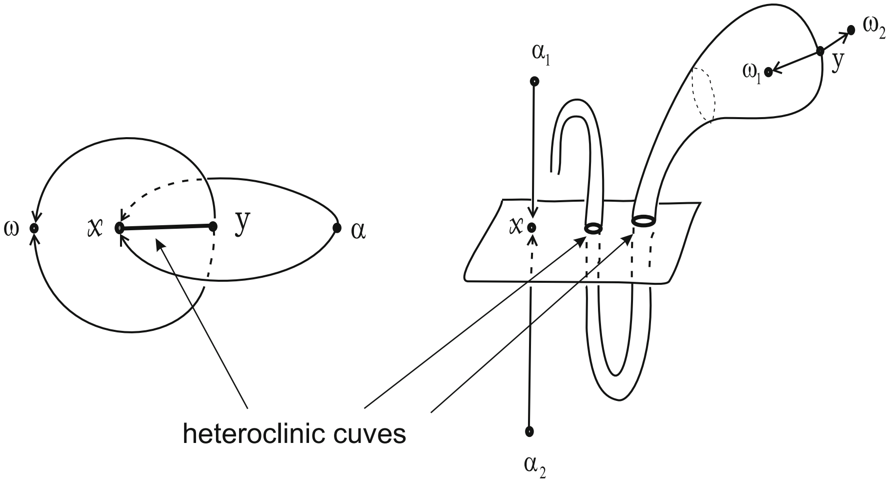


Fig. 31.2 Heteroclinic curves

From the transversality of the intersection of $W^u(p)$ with $W^s(q)$ it follows that $dim\ W^s(p) \leq dim\ W^s(q)$, hence any Morse–Smale diffeomorphism which does not contain heteroclinic points is a gradient-like diffeomorphism. If $W^u(p) \cap W^s(q) \neq \emptyset$ and $dim\ W^s(p) < dim\ W^s(q)$, then a connected component of the intersection $W^u(p) \cap W^s(q)$ is called a *heteroclinic submanifold* and for $n = 3$ is called *heteroclinic curve* (see Fig. 31.2).

The topological classification of preserving orientation gradient-like diffeomorphisms given on closed orientable surfaces (obtained by A. Bezdenezhnyh and V.

Grines in [8]–[9]) is closely connected with the topological classification of Morse–Smale flows on surfaces (in Peixoto's style) and with Nilsen's classification of periodic maps of surfaces [41].

Following to S. Smale [54] and J. Palis [43], we say that sequence of different periodic points $q = p_0, \dots, p = p_k$, $k \geq 1$ forms *chain* if $W^u(p_i) \cap W^s(p_{i+1}) \neq \emptyset$ for all $i \in \{0, \dots, k-1\}$ and there is no a periodic point p_* such, that $W^u(p_i) \cap W^s(p_*) \neq \emptyset$, $W^u(p_*) \cap W^s(p_{i+1}) \neq \emptyset$. The number k is called *length of chain* $q = p_0, \dots, p = p_k$ and it is denoted by *beh* $(q|p)$. If p, q are saddle point then chain is called *heteroclinic chain*. The length of the maximum heteroclinic chain is denoted by *beh* (f). By definition, *beh* $(f) = 0$ means that any separatrices of saddle periodic points are disjoint.

If *beh* $(f) = 1$ then the number of heteroclinic trajectories is finite. For surfaces it is possible to describe the pattern of the intersection of stable and unstable separatrices in a rather simple way. For diffeomorphisms given on closed orientable surfaces A. Bezdenezhyh and V. Grines considered at first a special case where the intersection of separatrices is orientable for any saddle periodic points p, q. They proved (in [7], see also the surveys [3, 4, 28]) that the condition *beh*$(f) = 1$ is a corollary of the orientability of the intersection and obtained a complete set of invariants for such diffeomorphisms. Then V. Grines and R. Langevin introduced independently (in [27, 35] see also [3, 4, 28]) invariants describing the pattern of the intersection of separatrices of saddle periodic points p, q of Morse–Smale diffeomorphisms for which *beh*$(q|p) = 1$ and obtained a complete set of invariants for such diffeomorphisms.

If *beh*$(p|q) > 1$ then the description of the intersection of the separatrices is more complicated. In fact the classification of Morse–Smale diffeomorphisms in this case is no easier than the classification of the general structurally stable diffeomorphisms. These two cases have been solved by Ch. Bonatti and R. Langevin using Markov chains endowed by additional information (see [20]).

In dimension 3, except heteroclinic intersection, Morse–Smale diffeomorphisms can have wildly embedded separatrices (see Sect. 31.1.3). Such effect very complicates their topological classification and requires new topological invariants, because Peixoto's graph is not complete invariant. The first example of such a nontrivial embedding was constructed by D. Pixton [48]. Then Ch. Bonatti and V. Grines [10] gave a complete topological classification of Pixton's class of diffeomorphisms. As consequence they get that there exist infinitely many diffeomorphisms from Pixton's class which are not topologically conjugate. Ch. Bonatti, V. Grines, V. Medvedev and O. Pochinka [19] investigated bifurcations connected with changes of embedding of separatrices for class above. The results which they have obtained are based on the fact that the space of "North pole – South pole" diffeomorphisms equipped by C^1-topology is connected. They shown also in [17] that this fact does not take place, for example, in dimension 6 (see Sect. 31.1.4).

Different interrelations between Morse–Smale diffeomorphism and topology of ambient manifold are contained in Sect. 31.1.5.

In Sect. 31.1.6 we give topological classification of gradient-like difeomorphisms on 3-manifolds, which was obtained by Ch. Bonatti, V. Grines, V. Medvedev and

E. Pecou, firstly in [11] without heteroclinic curve and then for all gradient-like difeomorphisms in [13].

By first step in study of Morse–Smale diffeomorphisms with heteroclinic orbits on 3-manifolds was done in the papers [14, 50] in which were obtained necessary and sufficient conditions of topological conjugacy of diffeomorphisms given on 3-manifolds, whose nonwandering set consists of exactly six points and wandering set does not contain heteroclinic curves. Ambient manifolds for such diffeomorphisms are only one of the following manifolds: $\mathbb{S}^3$, $\mathbb{S}^2 \times \mathbb{S}^1$, $\mathbb{S}^2 \times \mathbb{S}^1 \# \mathbb{S}^2 \times \mathbb{S}^1$. By development of ideas of these papers have become the papers [15, 16] in which Ch. Bonatti, V. Grines, and O. Pochinka obtained complete topological classification of Morse–Smale diffeomorphisms with a finite set of heteroclinic orbits and without heteroclinic curves on 3-manifolds. Moreover they got the classification of Morse–Smale diffeomorphisms with the chain of saddles of arbitrary length on 3-manifolds in [18] (see Sect. 31.1.7).

Finally we show in Sect. 31.1.8 that Pexoto's graph is again complete topological invariant for some class of Morse–Smale diffeomorphisms on manifolds of dimension $n > 3$. It is a result from the recent paper [30].

We start our survey from Sect. 31.1.2 which is devoted to description of general dynamic properties of Morse–Smale diffeomorphisms. Everywhere we suppose that M is closed smooth orientable manifold of dimension $n \geq 2$, a Morse–Smale diffeomorphism is preserving orientation as and a conjugating homeomorphism.

31.1.2 Global Dynamic of Morse–Smale Diffeomorphisms

The simplest Morse–Smale diffeomorphism is "source-sink" diffeomorphism $f : M \to M$, whose non-wandering set consists of exactly two points: sink and source. Such diffeomorphisms have trivial dynamics: all points which are distinct from fixed points, are wandering and go by f from the source to the sink (see Fig. 31.3a). Ambient manifold for "source-sink" diffeomorphism is homeomorphic to n-dimensional sphere $\mathbb{S}^n$, and space of wandering orbits (the orbit space of action of group $F = \{f^k, k \in \mathbb{Z}\}$ on $\mathbb{S}^n \setminus \Omega(f)$) is homeomorphic to $\mathbb{S}^{n-1} \times \mathbb{S}^1$. Thanks to such clear dynamics, it is easy to show, that any such diffeomorphisms are topological conjugated.

Studying more complicated Morse–Smale diffeomorphisms it is natural to try to present their dynamics like to "source-sink" diffeomorphism where roles of "source" and "sink" already play invariant sets (simple as possible from the topological point of view), one of them $\mathscr{A}_f^+$ is attractor and other $\mathscr{A}_f^-$ is repeller (see Fig. 31.3b) in the next sense.

An invariant set $\Lambda \subset M$ is an *attractor* if it has *attracting neighborhood* , that is a compact neighborhood $N \neq M$ of Λ such that $f(N) \subset int\ N$, and $\Lambda = \bigcap\limits_{k \geq 0} f^k(N)$. By definition, *repeller* for f is attractor for f^{-1}.

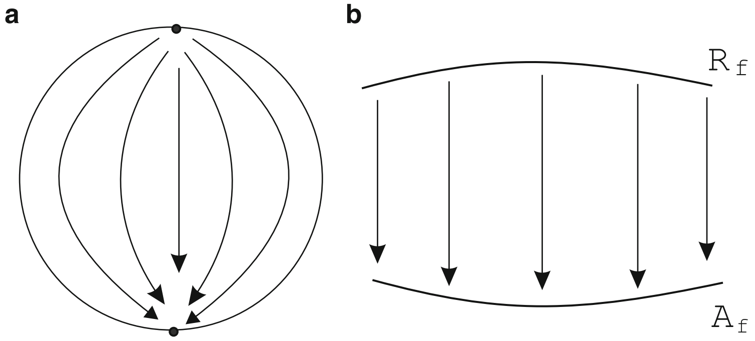


Fig. 31.3 "Source-sink" diffeomorphism (**a**) and it's generalization (**b**)

To form attractor $\mathscr{A}_f^+$ and repeller $\mathscr{A}_f^-$ we recall basic properties of Morse–Smale diffeomorphism in proposition below, which follows from Theorem 2.3 in [54].

Proposition 31.1. *Let $f : M \to M$ be Morse–Smale diffeomorphism. Then*

(1) $M = \bigcup\limits_{p \in NW(f)} W^u(p) = \bigcup\limits_{p \in NW(f)} W^s(p)$

(2) $W^s(p)$ $(W^u(p))$ *is smooth submanifold of* M *which is diffeomorphic to* $R^{\dim E_p^s}$ $(R^{\dim E_p^u})$ *for each periodic point* $p \in NW(f)$

(3) f *has at least one source and at least one sink*

(4) if f *has at least one saddle point then for any sink* ω *there is a saddle* σ *such that* $\omega \in clos(W^u(\sigma))$

(5) $clos(\ell_q^u) \setminus (\ell_q^u \cup q) = \bigcup\limits_{p \in NW(f): \ell_q^u \cap W^s(p) \neq \emptyset} W^u(p)$ *for connected component* ℓ_q^u *of* $W^u(q) \setminus q$, $q \in NW(f)$

(6) if ℓ_q^u *has no heteroclinic intersection then* $clos(\ell_q^u) \setminus (\ell_q^u \cup q) = \{\omega\}$*, where* ω *is a sink, if* $\dim W^u(q) = 1$ *then* $\{q\} \cup \ell_q^u \cup \{\omega\}$ *is an arc in* M*, if* $\dim W^u(q) \geq 2$ *then* $\{q\} \cup \ell_q^u \cup \{\omega\}$ *is topologically embedded in* M *sphere* $S^{\dim W^u(q)} \subset M$.

Let $f : M \to M$ be Morse–Smale diffeomorphism on closed orientable n-manifold M $(n \geq 2)$. Denote by Δ_f^+, Δ_f^-, Σ_f the set of sinks, source, saddle, accordingly. Set $\Delta_f = \Delta_f^+ \cup \Delta_f^-$. Let us represent the set $Per(f)$ as union $Per(f) = Per_f^+ \cup Per_f^-$ of disjoint subsets, where Per_f^+ (Per_f^-) consists of all periodic points $p \in Per(f)$ such that $\dim W^u(p) \leq 1$ $(\dim W^u(p) > 1)$. Set $\mathscr{A}_f^+ = W^u_{Per_f^+}$ and $\mathscr{A}_f^- = W^s_{Per_f^-}$.

It is possible to show that for f the set $\mathscr{A}_f^+$ is connected attractor of dimension ≤ 1 and the set $\mathscr{A}_f^-$ is repeller, connected for $n \geq 3$. We called the sets $\mathscr{A}_f^+$ and $\mathscr{A}_f^-$ *global attractor and repeller of Morse–Smale diffeomorphism* of f.

Set $\mathscr{M}_f = M^n \setminus (\mathscr{A}_f^+ \cup \mathscr{A}_f^-)$. Denote $V_f = \mathscr{M}_f / f$ the orbit space of action of the group $F = \{f^k, k \in \mathbb{Z}\}$ on $\mathscr{M}_f$ and $p_f : \mathscr{M}_f \to V_f$ the natural projection.

Next lemma follows from Proposition 3.5.7 in [56].

Lemma 31.1. *The orbit space V_f is closed n-manifold (connected for $n \geq 3$) and $p_f : \mathscr{M}_f \to V_f$ is open cover.*

Information on topology of the space V_f and embedding (possibly knotted) images under projection p_f invariant manifolds (of dimension more than 1) of saddle periodic points creates prerequisite for introducing new topological invariants which can be used for topological classification of Morse–Smale diffeomorphisms on 3-manifolds with different assumptions (see results of the papers [11–16] and [18], exposition of which we give below).

31.1.3 Wild Objects in Three-Dimensional Dynamic

The principle difference in topological classification of Morse–Smale diffeomorphisms on 3-manifolds in comparison with one on 2-manifolds is possibility of wild embedding of separatrices. In more details.

An *embedding* of one topological space in another space Y is a homeomorphism of X onto a subspace of Y. Two embeddings $\lambda, \lambda' : X \to Y$ are *equivalent* if there exists a homeomorphism $\theta : Y \to Y$ such that $\theta\lambda = \lambda'$. Moreover we will suppose that X and Y are triangulated manifolds or equivalently *PL* (piecewise linear) manifolds.

If there are embeddings of X in the Y that are homotopic but not equivalent, then X is said *to knot* in Y. It is often possible to identify a distinguished class of PL embeddings of X in Y that are considered to be *unknotted*; any PL embedding that is not equivalent to an unknotted embedding is then said to be *knotted*.

An embedding $\lambda : X \to Y$ is said to be a *tame embedding* if it is equivalent to a PL embedding; the others are called *wild*. If $X \subset Y$ then X is said to be *tame (wild)* if the inclusion $i : X \to Y$ is tame (wild) embedding. In other words, a manifold $X \subset Y$ is tame if there exists a homeomorphism $\theta : Y \to Y$ such that $\theta(X)$ is a subpolyhedron; is wild in the opposite case.

A topological embedding $\lambda : N \to M$ of an n-dimensional manifold N into an m-dimensional manifold M is *locally flat at $x \in N$* if there exists a neighborhood U of $\lambda(x)$ in M such that $(U, U \cap \lambda(N)) \approx (\mathrm{R}^m, \mathrm{R}^n)$ or $(U, U \cap \lambda(N)) \approx (\mathrm{R}^m, \mathrm{R}^n_+)$. An embedding is said to be *locally flat* if it is locally flat at each point x of its domain.

A classic theorem in piecewise linear topology assures that N tamely embedded in M is locally flat if $m - n \neq 2$ (see [24]). Since tameness implies local flatness for embeddings of manifolds in all codimensions $(m - n)$ except two, we will say that an embedding $\lambda : N \to M$ for $m - n \neq 2$, is wild at $\lambda(x)$ when $\lambda(N)$ fails to be locally flat at $\lambda(x)$. Concerning codimension two it is known that similar fact takes place for $n = 3$ and $m = 1$. Namely any arc tamely embedded in a 3-manifold is locally flat (see [34]). If $n = 2$ any arc and, hence, one-dimensional separatrix is always tame (see corollary 5, Sect. 4, Chap. 2 in [34]). According to [21], there are no arcs with one point of wildness on the manifold of dimension greater than 3. An example (not connected with dynamic) of wild arc in $\mathbb{S}^3$ with one point of wildness firstly was constructed by E. Artin and R. Fox in [6]. We explain this construction below.

For representation of a set of smooth arcs it is convenient to use their *orthogonal projection* on a plane. So *projection plane* needs to be chosen such that following conditions were satisfied:

(1) Projection of a tangent line to any arc in any point is a straight line (i.e. projection of a tangent line not degenerates in a point)
(2) More than two different points of arcs are not projected to the same point of a plane
(3) Set of *crossroads* (points in plane which are projections of two points of arc) is finite and projections of tangent lines at corresponding points of arcs do not coincide.

Let's consider in $\mathbb{R}^3$ a three-dimensional ring V, defined in spherical coordinates as $\frac{1}{2} \leq \rho \leq 1$ and homothety $\phi : \mathbb{R}^3 \to \mathbb{R}^3$ given by the formula $\phi(\rho, \varphi, \theta) = (\frac{1}{2}\rho, \varphi, \theta)$. Set $V_{\frac{1}{2}} = \{(\rho, \varphi, \theta) \in \mathbb{R}^3 : \rho = \frac{1}{2}\}$ and $V_1 = \{(\rho, \varphi, \theta) \in \mathbb{R}^3 : \rho = 1\}$. Then $\partial V = V_{\frac{1}{2}} \cup V_1$. Let $\alpha; \beta; \gamma \in V$ be simple pairwise closed arcs with end points $\alpha_1, \alpha_2; \beta_1, \beta_2; \gamma_1, \gamma_2$, accordingly, satisfying to following conditions:

(1) $\alpha_1, \alpha_2, \gamma_1 \subset V_1$ and $\beta_1, \beta_2, \gamma_2 \subset V_{\frac{1}{2}}$
(2) $\phi(\alpha_1) = \gamma_2, \phi(\alpha_2) = \beta_1, \phi(\gamma_1) = \beta_2$
(3) the plane Ox_1x_2 is a plane of a projection for arcs α, β, γ and their projection looks like to Fig. 31.4a.

If we identify $V_{\frac{1}{2}}$ and V_1 by the diffeomorphism ϕ then V/ϕ is diffeomorphic to $\mathbb{S}^2 \times \mathbb{S}^1$. Denote $p_\phi : V \to \mathbb{S}^2 \times \mathbb{S}^1$ the natural projection. Set $\hat{\ell} = p_\phi(\alpha \cup \beta \cup \gamma)$.

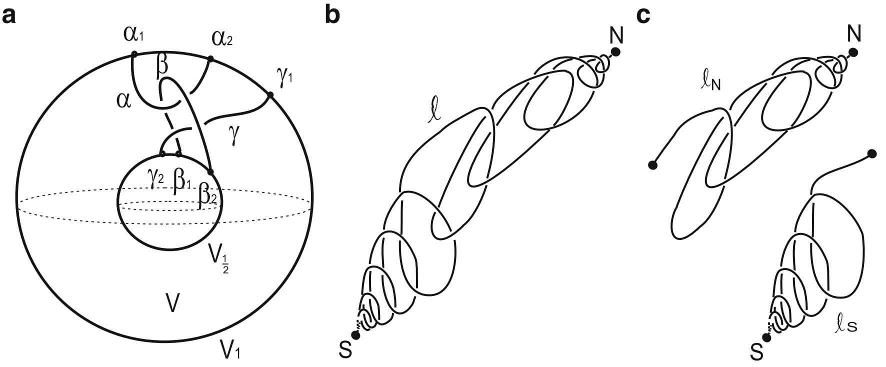


Fig. 31.4 Construction of wild arcs in $\mathbb{S}^3$

It has been proved B. Mazur [38], that $\hat{\ell}$ is knotted circle, that is $\hat{\ell}$ has the same homotopic class 1 as circle $\{x\} \times \mathbb{S}^1$ for $x \in \mathbb{S}^2$ but is not equivalent to it.

Set $\tilde{\ell} = \bigcup\limits_{k \in \mathbb{Z}} \phi^k(\alpha \cup \beta \cup \gamma)$ and $\ell = \vartheta_+^{-1}(\tilde{\ell}) \cup N \cup S$ (see Fig. 31.4b), where ϑ_+ is the stereographic projection and N, S are northern and south poles of sphere $\mathbb{S}^3$.

It has been proved in [6], that ℓ is wildly embedded in $\mathbb{S}^3$ and has exactly two points of wildness N and S. Arc ℓ_N (ℓ_S) which is represented on Fig. 31.4c is a part of arc ℓ from point $\vartheta_+^{-1}(\alpha_1)$ to point N (from point $\vartheta_+^{-1}(\alpha_1)$ to point S). According to [34] (Chap. 4, Sect. 2, Example 2.4), the arc ℓ_N (ℓ_S) is wildly embedded in $\mathbb{S}^3$ and has exactly one point of wildness N (S).

If we thicken arcs in Fig. 31.4b or c, we will get 2-sphere S^2 which is wildly embedded in $\mathbb{S}^3$ and has exactly one point of wildness at pole (see [34], Chap. 4, Sect. 2, Examples 2.1(b), 2.4). Moreover, set $\mathbb{S}^3 \setminus S^2$ consists of two connected components A_1 and A_2, each of them is homeomorphic to *int* $\mathbb{D}^3$, *clos*(A_1) is homeomorphic to $\mathbb{D}^3$ and *clos*(A_2) is not homeomorphic to $\mathbb{D}^3$.

Now let $f : M \to M$ be a gradient-like diffeomorphism of 3-manifold M. According to Prop. 31.1, the closure $clos(\ell)$ of any one-dimensional unstable separatrix ℓ of saddle point σ is homeomorphic to a segment which consists of ℓ and two points: σ and a sink ω. Moreover, $\ell \cup \sigma$ is smooth submanifold of M. Thus $clos(\ell)$ may be wild only at ω. We say that the separatrix ℓ is *wild* or *wild embedded* in M if arc $clos(\ell)$ is wild at ω. In opposite case we say that the separatrix ℓ is *tame* or *tame embedded* in M. Similarly the concept of wild embedding is generalized on stable one-dimensional and two-dimensional separatrices of a gradient-like diffeomorphism. Due to Prop. 31.1, $W^s(\omega)$ is homomorphic to $\mathbb{R}^3$. Then, according to [34], the tameness of ℓ is equivalent to the existence of a homeomorphism $\psi : W^s(\omega) \to \mathbb{R}^3$ such that $\psi(\omega) = O$, where O is the origin and $\psi(\bar{\ell} \setminus \sigma)$ is a ray starting from O.

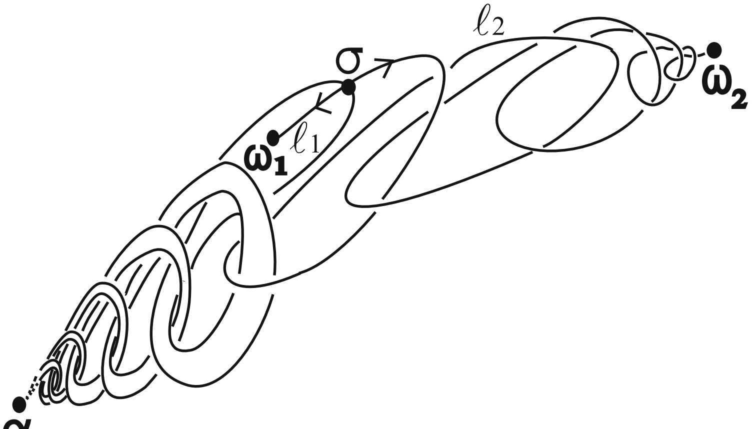


Fig. 31.5 Pixton's example

First example of Morse–Smale diffeomorphism with wildly embedded separatrices was constructed by D. Pixton in [48] (see Fig. 31.5). In the Pixton's example, the non-wandering set of $f : \mathbb{S}^3 \to \mathbb{S}^3$ consists of exactly four fixed points: one source α, two sinks ω_1, ω_2, one saddle σ whose one unstable separatrix ℓ_1 is tamely embedded and the other ℓ_2 is wildly embedded (see Fig. 31.5). This example disproved hypothesis of M. Shub [52] and F. Takens [55] on existence of an energy function for any Morse–Smale diffeomorphism in following sense. *An energy function* for a dynamic system on M is a smooth function $\varphi : M \to \mathbb{R}$ which strictly decreases along orbits outside of the chain recurrent set, is constant on the chain components of the system and set of critical points of φ coincides with the chain recurrent set of the system. As noticed J. Franks in [26], application of W. Wilson's results [57] to Conley's [23] construction gives an existence of an energy function for any smooth flow with hyperbolic chain recurrent set. But question on existence of an energy function for a diffeomorphism is an open even for Morse–Smale systems. Namely, Pixton proved the following results.

- For any Morse–Smale diffeomorphism given on a surface there is an energy Morse function (function with non degenerate critical point).
- The diffeomorphism on Fig. 31.5 has no energy Morse function.

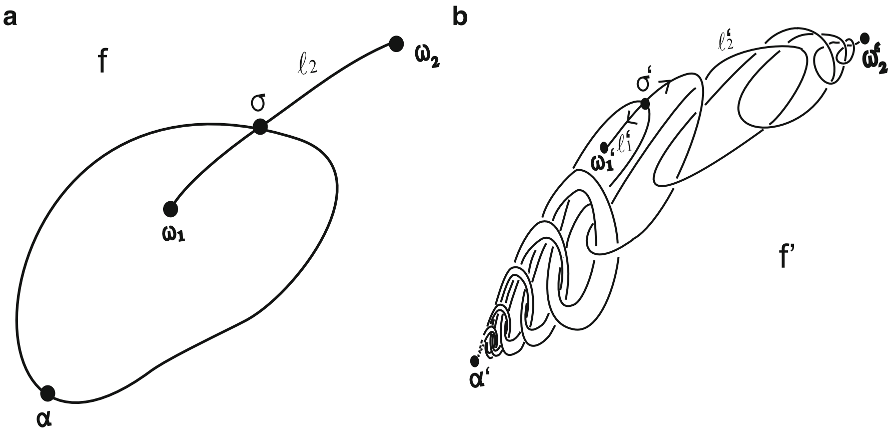
Recently V. Grines, F. Laudenbach and O. Pochinka [31, 32] obtained necessary and sufficient conditions to the existence of energy Morse function for Morse–Smale diffeomorphisms on 3-manifolds.

We denote $\mathcal{G}_4$ the Pixton's class that is class of diffeomorphisms on S^3 whose nonwandering set is fixed and hyperbolic and consists of exactly one source, two sinks and one saddle.

31.1.4 Classification and Bifurcation in Pixton's Class

The Pixton's class $\mathcal{G}_4$ was considered in [10]. It was also shown that the topological classification of diffeomorphisms from $\mathcal{G}_4$ is reduced to the embedding classification of the one-dimensional separatrix. Hence there exist infinitely many diffeomorphisms from $\mathcal{G}_4$ which are not topologically conjugate. Notice that all diffeomorphisms from $\mathcal{G}_4$ have isomorphic Peixoto's graph. Now we represent the main results of paper [10].

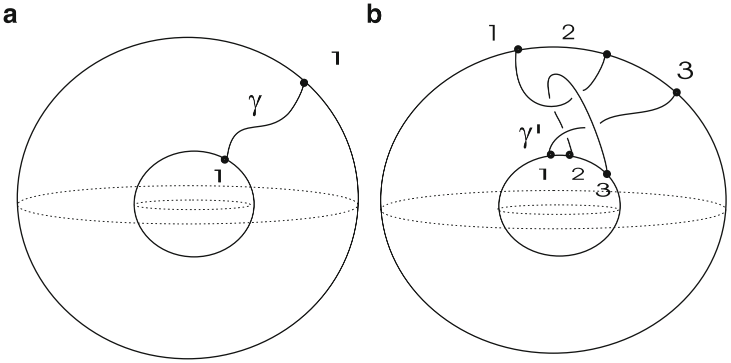
Let $f \in \mathcal{G}_4$. Denote by α the fixed source, by σ the fixed saddle point and by ω_1, ω_2 the fixed sinks belonging to the nonwandering set of f. Denote by L the stable separatrix and by ℓ_1, ℓ_2 the unstable separatrices of the point σ. Then we have that the closure $clos\ (L)$ of two dimensional (stable) separatrix of the point σ is homeomorphic to the sphere $\mathbb{S}^2$ and consists of this separatrix and source α. The closure $clos\ (\ell_i)$ $(i = 1, 2)$ of one-dimensional (unstable) separatrix of the point σ is homeomorphic to a closed simple arc and consists of this separatrix and two points: the point σ and a sink. Moreover, the separatrices ℓ_1 and ℓ_2 contain different sinks in the closure. Let us assume for definiteness that the point ω_i belongs to $clos\ (\ell_i)$. We introduce in the set $W^s(\omega_i) \setminus \omega_i$ the equivalence relation, assuming



f is a diffeomorphism with tame embedded separatrices

f' is a diffeomorphism with wild embedded separatrices

Fig. 31.6 Phase portraits of diffeomorphisms from Pixton's class



knot associated with the diffeomorphism f

knot associated with the diffeomorphism f'

Fig. 31.7 Topological invariants of diffeomorphisms from Pixton's class

that two points are equivalent if they belong to the same orbit of the diffeomorphism f. Denote by N_{ω_i} the factor space obtained by this equivalence relation and by $\pi_{\omega_i} : W^s(\omega_i) \setminus \omega_i \to N_{\omega_i}$ the natural projection.

We call smooth submanifold $\gamma \subset \mathbb{S}^2 \times \mathbb{S}^1$ by *a knot* if it is homeomorphic to $\mathbb{S}^1$ and has homotopic class 1. We say that two knots γ and γ' in the manifold $\mathbb{S}^2 \times \mathbb{S}^1$ are *equivalent*, if there exists diffeomorphism $\hat{w} \in Diff(\mathbb{S}^2 \times \mathbb{S}^1)$ such that $\gamma' = \hat{w}(\gamma)$. Put $\gamma_0 = \{(s, \rho) \in \mathbb{S}^2 \times \mathbb{S}^1 : s = (0, 0, -1)\}$. We call the knot γ *unknotted* if it is equivalent to γ_0 and *knotted* in the opposite case. On the Fig. 31.7 there is three-dimensional rings and arcs. After identifying of boundaries of the rings we get an unknotted knot γ and knotted knot γ' (Fig. 31.6).

Lemma 31.2. *([10], Lemma 1.1) There is a preserving orientation diffeomorphism $\phi_{\omega_i} : N_{\omega_i} \to \mathbb{S}^2 \times \mathbb{S}^1$ such that the set $\gamma_i = \phi_{\omega_i}(\pi_{\omega_i}(L_i))$ is a knot in $\mathbb{S}^2 \times \mathbb{S}^1$.*

Remark 31.1. If $\check{\phi}_{\omega_i} : N_{\omega_i} \to \mathbb{S}^2 \times \mathbb{S}^1$ is a diffeomorphism such that the set $\check{\gamma}_i = \check{\phi}_{\omega_i}(\pi_{\omega_i}(L_i))$ is a knot in $\mathbb{S}^2 \times \mathbb{S}^1$ which is different from γ_i then the diffeomorphism $\hat{w} = \check{\phi}_{\omega_i} \circ \phi_{\omega_i}^{-1} \in Diff(S^2 \times S^1)$ realizes the equivalence of the knots γ_i and $\check{\gamma}_i$, that is $\hat{w}(\gamma_i) = \check{\gamma}_i$.

Theorem 31.1. *([10], Theorem 1) At least one of two knots γ_1, γ_2 is unknotted.*

Let us assume for definiteness that the knot γ_1 is unknotted.

Theorem 31.2. *([10], Theorem 3) Two diffeomorphism $f, f' \in \mathcal{G}_4$ are topologically conjugated if and only if the knots γ_2 and γ_2' are equivalent.*

Theorem 31.3. *([10], Theorem 2) For any knot γ in $\mathbb{S}^2 \times \mathbb{S}^1$ there exists a diffeomorphism f_γ from the class $\mathcal{G}_4$ given on the sphere $\mathbb{S}^3$ and such that γ_2 for f_γ is equivalent to γ.*

Remark 31.2. It follows from Propositions 31.2 and 31.3 that the ambient manifold for diffeomorphisms from the class $\mathcal{G}_4$ is homeomorphic to the sphere $\mathbb{S}^3$.

In connection with detection of new topological invariants there is a natural problem on finding of elementary bifurcations allowing to pass from one class of topological conjugacy of diffeomorphisms to another. It was shown in [19] that any two diffeomorphisms from the class $\mathcal{G}_4$ can be joined by a smooth arc containing two bifurcations of the saddle-node type. Let us notice that this result concerns of decision of the problem posed by J. Palis and C. Pugh in [44] about finding of a smooth arc with some good properties (for example, with the finite number bifurcations) joining two structurally stable dynamic systems (two flows or two diffeomorphisms). S. Newhouse and M. Peixoto have proved in [40] that any Morse–Smale flows on closed manifold can be joined by arc with the finite number bifurcations. From another hand, as have proved S. Matsumoto in [37], any oriented closed surface admits two isotopic Morse–Smale diffeomorphisms which can not be joined by similar arc. Beginning from dimension 3 this problem is not trivial even for simplest diffeomorphisms of the type “North pole – South pole”. It is rather easy follows from Milnor’s result [39] that there are two “North pole – South pole” diffeomorphisms on S^6 which can not be joined a smooth arc. From another hand, due to J. Cerf [22], for any two preserving orientation diffomorphisms (and consequently for any two “North pole – South pole” diffeomorphisms) of $\mathbb{S}^3$ there is a smooth arc their joining. We will show that this arc may be chosen consisting of “North pole – South pole” diffeomorphisms.

Now we represent the main results of paper [19].

Let N, M be orientable smooth manifolds. A map $f : N \to M$ is *smooth embedding of N to M* if f is a diffeomorphism from N to $f(N)$, where $f(N)$ is a smooth submanifold of M. Two embeddings $f, f' : N \to M$ are *smoothly isotopic* if there is a smooth map $F : N \times [0, 1] \to M$ (smooth isotopy) such that f_t given by the formula $f_t(x) = F(x,t)$ is an embedding for each $t \in [0,1]$ and $f_0 = f$, $f_1 = f'$. We say that the family $\{f_t\}$ is *a smooth arc joining the embeddings f, f'*. Let $C(N,M)$ be the space of all embeddings $N \to M$ with C^1-topology. We say that a subset $A \subset C(N,M)$ is *connected* if for any embeddings $f, f' \in A$ there is a smooth arc $\{f_t \in A\}$ connecting their.

Denote by $Diff(M)$ the space of all C^r-diffeomorphisms on M ($r \geq 2$), by $Diff_+(M) \subset Diff(M)$ the space of all orientation preserving diffeomorphisms and by $Diff_0(M) \subset Diff_+(M)$ the space of diffeomorphisms which are smoothly isotopic to the identical map. Let

$$\mathbb{S}^n = \{(x_1,\ldots,x_{n+1}) \in \mathbb{R}^{n+1} : \sum_{i=1}^{n+1} x_i^2 = 1\}.$$

Denote by $J(\mathbb{S}^n) \subset Diff_+(\mathbb{S}^n)$ the class of diffeomorphisms whose nonwandering set consists of exactly two hyperbolic fixed points: the source, the sink and by $NS(\mathbb{S}^n) \subset J(\mathbb{S}^n)$ the class of "North pole – South pole" diffeomorphisms that is diffeomorphisms which have the source in the point $N(\underbrace{0,\ldots,0}_{n},1)$ and the sink in the point $S(\underbrace{0,\ldots,0}_{n},-1)$.

Theorem 31.4. *([19], Theorem 1) For any diffeomorphisms $f, f' \in J(\mathbb{S}^3)$ there is a smooth arc $\{f_t \in J(\mathbb{S}^3)\}$ joining these diffeomorphisms.*

Theorem 31.5. *([19], Theorem 2) In the class $NS(\mathbb{S}^6)$ there are diffeomorphisms which are can not be joined a smooth arc.*

Theorem 31.6. *([19], Theorem 3) For any diffeomorphisms $f, f' \in \mathcal{G}_4$ there is a smooth arc $\{f_t \in Diff_+(\mathbb{S}^3)\}$ and numbers t_1, t_2 such that:*

(1) $f_0 = f$, $f_1 = f'$
(2) $f_t \in \mathcal{G}_4$ *for all* $t \in [0,t_1) \cup (t_2,1]$
(3) $f_t \in J(\mathbb{S}^3)$ *for all* $t \in (t_1,t_2)$
(4) The nonwandering set of the diffeomorphism f_{t_i}, $i = 1,2$ consists of two hyperbolic fixed points: source and sink and one non hyperbolic fixed point of the type saddle-node.

31.1.5 Interrelation Between Morse–Smale Diffeomorphism and Topology of Ambient Manifold

In this section, firstly, we attempt to characterize those manifolds which admit Morse–Smale diffeomorphisms without heteroclinic curves.

Theorem 31.7. *([12], Theorem) Let M be a three-dimensional closed, connected, orientable manifold. There exists a Morse–Smale diffeomorphism without heteroclinic curve on M admitting k saddle periodic points and l sinks and sources if and only if M is the sphere if $k = l - 2$, or M is the connected sum of $(k - l + 2)/2$ copies of $S^2 \times S^1$.*

Recall that the *connected sum* $M_1 \# M_2$ of two oriented connected manifolds M_1 and M_2, is the manifold obtained by choosing disks $D_i \subset M_i$ and by gluing the manifolds $M_i \setminus D_i$ $(i = 1, 2)$, by a diffeomorphism between the boundaries which reverses the natural orientation on the boundaries.

The result of the Theorem 31.7 is well known for Morse–Smale vectorfields without heteroclinic curves and without periodic orbits on 3-manifolds, and so our result can be interpreted as follows: the 3-manifolds admitting Morse–Smale diffeomorphisms without heteroclinic curves are the same which admit Morse–Smale vectorfields without heteroclinic curves and periodic orbits. This is in some way surprising because Morse–Smale diffeomorphisms present a very different behavior:

- The invariant manifolds of the saddles, without any heteroclinic points or curves, may induce wild arcs and wild spheres (see Sect. 31.1.4).
- Here we allow infinitely many orbits of heteroclinic points, and we avoid only heteroclinic curves.

The key of the proof of the Theorem will 31.7 is the following result which describes the topological nature of the neighborhoods of the wild spheres which appear in our context.

Proposition 31.2. *([12], Proposition 0.1) Let η: $S^2 \to M$ be a topological embedding of the two-sphere which is a smooth immersion everywhere, except at one point and let $\Sigma = \eta(S^2)$. Then any neighborhood of Σ contains an open neighborhood K such that $\Sigma \subset K$ and clos (K) is diffeomorphic to $S^2 \times [0, 1]$.*

The theorem above is sufficient condition of existence of a heteroclinic curve for given Morse–Smale diffeomorphism.

- For example, if non-wandering set $\Omega(f)$ of Morse–Smale diffeomorphism $f : S^3 \to S^3$ consists of two saddles, one sink and one source then wandering set of such diffeomorphism contains a heteroclinic curve. Moreover, in this case there is at least one noncompact heteroclinic curve.
- If ambient manifold of Morse–Smale diffeomorphism f is not homeomorphic to the connected some of copies of $\mathrm{S}^2 \times \mathrm{S}^1$ (for example if ambient manifold is the torus T^3) then $\Omega(f)$ contains at least one heteroclinic curve.

Secondly, we study topology of ambient 3-manifold admitting gradient-like difeomorphisms. Let $f : M \to M$ be a gradient-like diffeomorphism and $L(\omega)$ be the union of all unstable one-dimensional separatrices of saddles which contain ω in their closure. The collection $L(\omega)$ is called *tame* if there is a homeomorphism $\psi : W^s(\omega) \to \mathrm{R}^3$ such that $\psi(\omega) = O$, where O is the origin and $\varphi(clos\ (\ell) \setminus \sigma)$

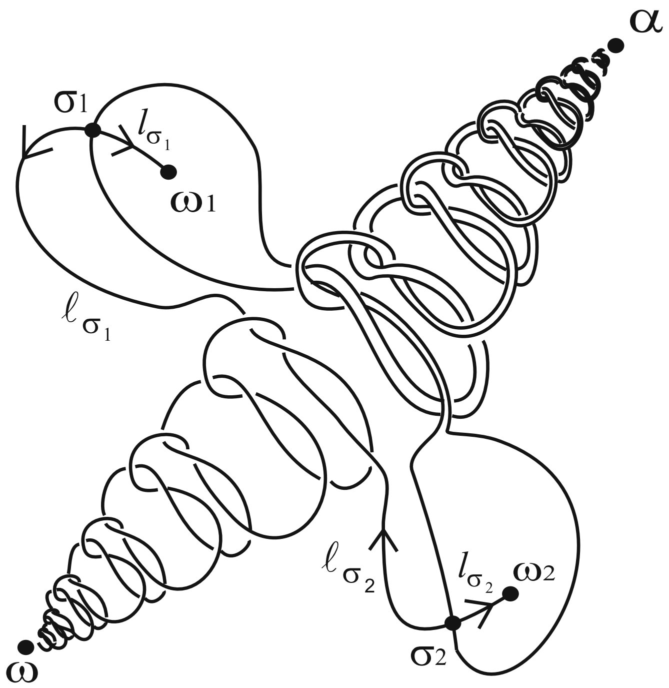


Fig. 31.8 Phase portrait of the diffeomorphism on $\mathbb{S}^3$ with mildly wild frame of separatrices

is a ray starting from O for any separatrix $\ell \in L(\omega)$. In the opposite case the set $L(\omega)$ is *wild*.

Notice that the tameness of each separatrix $\ell \in L(\omega)$ does not imply the tame property of $L(\omega)$. In [25] there is an example of a wild collection of arcs in $\mathbb{R}^3$ where each arc is tame. Using this example and methods of realization of Morse–Smale diffeomorphisms suggested in [10] and [16], it is possible to construct a gradient-like diffeomorphisms on $\mathbb{S}^3$ having a wild bundle $L(\omega)$ (see Fig. 31.8).

Theorem 31.8. *([33], Theorem 4.1) Let M be a three-dimensional closed, connected, orientable manifold and $f : M \to M$ be gradient-like diffeomorphism with k saddle periodic points and l sinks and sources such that for any sink ω (source α) the set $L(\omega)$ ($L(\alpha)$) are tame. Then the manifold M^3 admits the Heegaard splitting of genus $g = \frac{k-l+2}{2}$.*

Let us form the global attractor $\mathscr{A}_f^+$ and repeller $\mathscr{A}_f^-$ for diffeomorphism f as in Sect. 31.1.2. Recall that Δ_f^+ (Δ_f^-) is the set of all sinks (sources), $\Delta_f = \Delta_f^+ \cup \Delta_f^-$, Σ_f^+ (Σ_f^-) is the set of saddle points with one-dimensional unstable (stable) invariant manifolds, $\Sigma_f = \Sigma_f^+ \cup \Sigma_f^-$, L_f^+ (L_f^-) is a union of one-dimensional separatrices of saddle points from Σ_f^+ (Σ_f^-), $\mathscr{A}_f^+ = \Delta_f^+ \cup L_f^+ \cup \Sigma_f^+$, $\mathscr{A}_f^- = \Delta_f^- \cup L_f^- \cup \Sigma_f^-$, $L_f = L_f^+ \cup L_f^-$, $\mathscr{M}_f = M^3 \setminus (\mathscr{A}_f^+ \cup \mathscr{A}_f^-)$.

According to Sect. 31.1.2, the set $\mathscr{A}_f^+$ ($\mathscr{A}_f^-$) is connected attractor (repeller) of dimension $\leq$ 1. Due to Lemma 31.1, the space of orbits $V_f = \mathscr{M}_f/f$ is closed 3-manifold. If $\mathscr{A}_f^+$ and $\mathscr{A}_f^-$ are tame then it is possible to recognize the topological structure of $\mathscr{M}_f$ and V_f next way.

Corollary 31.1. *The space $\mathscr{M}_f$ is diffeomorphic to $S_{g_f} \times \mathbb{R}$ and the manifold V_f is diffeomorphic to $S_g \times \mathbb{S}^1$, where S_g is orientable surface of genus g from Theorem 31.8.*

31.1.6 Topological Classification of Gradient-Like Diffeomorphisms on 3-Manifolds

Denote G_0^3 class of gradient-like diffeomorphisms on 3-manifold M. In this section we give complete topological classification of diffeomorphisms from this class by means of topological invariant named *global scheme*.

For a diffeomorphism $f \in G_0^3$ we keep denotation of Sects. 31.1.2 and 31.1.5, where we formed the global connected attractor $\mathscr{A}_f^+$ and repeller $\mathscr{A}_f^-$, the wandering space $\mathscr{M}_f = M^3 \setminus (\mathscr{A}_f^+ \cup \mathscr{A}_f^-)$ and the space of orbits $V_f = \mathscr{M}_f/f$. It follows from corollary 31.1 that in the case of tame embedding of $\mathscr{A}_f^+$ and $\mathscr{A}_f^-$ the manifold V_f is diffeomorphic to $S_g \times \mathbb{S}^1$, where S_g is orientable surface of genus $g = \frac{|\Sigma_f|-\Delta_f|+2}{2}$. In general case the manifold V_f is obtained from 3-dimensional cobordism $(K, P_{\tilde{g}}^1, P_{\tilde{g}}^2)$, where $P_{\tilde{g}}^i$ is the boundary of handlebody of genus $\tilde{g} \geq g$, by identifying of $P_{\tilde{g}}^1$ and $P_{\tilde{g}}^2$ by means f.

Denote $p_{_f} : \mathscr{M}_f \to V_f$ the natural projection and $\alpha_{_f} : \pi_1(V_f) \to \mathbb{Z}$ the epimorphism corresponding to cover $p_{_f}$. $\alpha_{_f}$ has following property: any curve in $\mathscr{M}_f$ joining some point x with the point $f^k(x)$ is projected to the closed loop c on V_f such that $\alpha_{_f}([c]) = k$.

Set $\hat{\lambda}_\sigma^s = p_{_f}(W^s(\sigma)\backslash\sigma)$ ($\hat{\lambda}_\sigma^u = p_f(W^u(\sigma)\backslash\sigma)$) for any $\sigma \in \Sigma_f^+$ ($\sigma \in \Sigma_f^-$) and $\hat{\Lambda}_f^s = \bigcup\limits_{\sigma\in\Sigma_f^+} \hat{\lambda}_\sigma^s$ ($\hat{\Lambda}_f^u = \bigcup\limits_{\sigma\in\Sigma_f^-} \hat{\lambda}_\sigma^u$). As $f^{per(\sigma)}|_{W_\sigma^s}$ ($f^{per(\sigma)}|_{W_\sigma^u}$) is conjugated with contraction (expansion) on $\mathbb{R}^2$ then $\hat{\lambda}_\sigma^s$ ($\hat{\lambda}_\sigma^u$) is two-dimensional torus, if $f^{per(\sigma)}|_{W_\sigma^s}$ ($f^{per(\sigma)}|_{W_\sigma^u}$) preserves orientation and Klein bottle in the opposite case. Moreover, $\alpha_{_f}(\pi_1(\hat{\lambda}_\sigma^\delta)) = per(\sigma)\mathbb{Z}$, connected components of the set $\hat{\Lambda}_f^\delta$ are pairwise disjoint for $\delta \in \{u, s\}$, but for all that $\hat{\Lambda}_f^u$ and $\hat{\Lambda}_f^s$ can have transversal intersection at projection of heteroclinic curves.

Definition 31.1. A collection $S_f = (V_f, \alpha_{_f}, \hat{\Lambda}_f^u, \hat{\Lambda}_f^s)$ is called *global scheme* of diffeomorphism $f \in G_0^3$.

On Fig. 31.9 is represented a phase portrait of gradient-like diffeomorphism $f : \mathbb{S}^3 \to \mathbb{S}^3$. For its global scheme S_f, V_f is diffeomorphic to three-dimensional torus

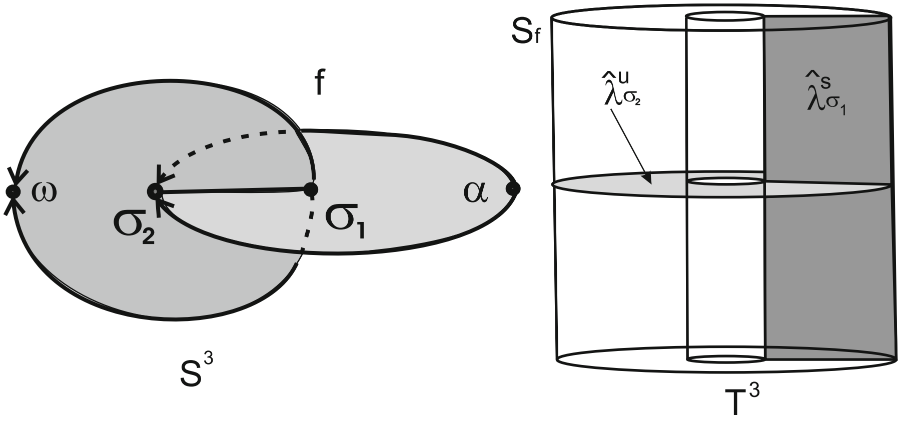


Fig. 31.9 A global scheme of a diffeomorphism $f \in G_0^3$

and $\hat{\Lambda}^u_f$, $\hat{\Lambda}^s_f$ are two-dimensional tore. To get S_f we have to identify lateral faces and bases of cylinder on Fig. 31.9 on the right.

Definition 31.2. Global schemes $S_f = (V_f, \alpha_f, \hat{\Lambda}^u_f, \hat{\Lambda}^s_f)$ and $S_{f'} = (V_{f'}, \alpha_{f'}, \hat{\Lambda}^u_{f'}, \hat{\Lambda}^s_{f'})$ of diffeomorphisms $f, f' \in G_0^3$ are called *equivalent* if there is a preserving orientation homeomorphism $\hat{\varphi} : V_f \to V_{f'}$ such that $\alpha_f = \alpha_{f'}\hat{\varphi}_*$ and $\hat{\varphi}(\hat{\Lambda}^u_f) = \hat{\Lambda}^u_{f'}, \hat{\varphi}(\hat{\Lambda}^s_f) = \hat{\Lambda}^s_{f'}$.

Theorem 31.9. *([13], Theorem 2) Diffeomorphisms $f, f' \in G_0^3$ are topological conjugated if and only if their global schemes S_f, $S_{f'}$ are equivalent.*

For the solution of the problem of realisation we introduce a concept of *perfect scheme*. Let V be a smooth closed orientable 3-manifold, whose fundamental group admit an epimorphism $\alpha : \pi_1(V) \to \mathbb{Z}$. Let $\hat{\Lambda}^u$, $\hat{\Lambda}^s \subset V$ be sets of smoothly embedded tore and Klein bottles such that elements from $\hat{\Lambda}^\delta$ are pairwise disjoint, $\alpha(\pi_1(\hat{\lambda}^\delta)) \neq 0$ for any element $\hat{\lambda}^\delta \in \hat{\Lambda}^\delta$ and sets $\hat{\Lambda}^u$, $\hat{\Lambda}^s$ can have transversal intersection.

For each component $\hat{\lambda}^\delta \in \hat{\Lambda}^\delta$ the fundamental group $\pi_1(\hat{\lambda}^\delta)$ admit a system of generatrices (a, b) such, that $\alpha([a]) > 0$ and $\alpha([b]) = 0$. Let $N(\hat{\lambda}^\delta)$ be a tubular neighborhood of $\hat{\lambda}^\delta$, $\bar{V} = V \setminus int\, N(\hat{\lambda}^\delta)$ and $\bar{\alpha} : \pi_1(\bar{V}) \to \mathbb{Z}$ be epimorphism which is induced by α. By the construction $\bar{V}$ is a compact manifold whose boundary consists of two tore if $\hat{\lambda}^\delta$ is tores or of one torus if $\hat{\lambda}^\delta$ is Klein bottle. Besides, the system of generatrices (a, b) of element $\hat{\lambda}^\delta$ induces on each connected component of $\partial\bar{V}$ system of generatrices (a, b) or (a^2, b), accordingly, and $\bar{\alpha}([a]) > 0, \bar{\alpha}([b]) = 0$. Denote $(V, \hat{\lambda}^\delta)$ the closed 3-manifold which is obtained from $\bar{V}$ by gluing of solid tore to each connected component of $\partial\bar{V}$ such that meridian of solid tore sticks together with b (we will notice, that this construction does not depend on a choice of gluing diffeomorphism). We say that $(V, \hat{\lambda}^\delta)$ is a manifold which is obtained from V by *cutting and pasting along* $\hat{\lambda}^\delta$ (see Fig. 31.10).

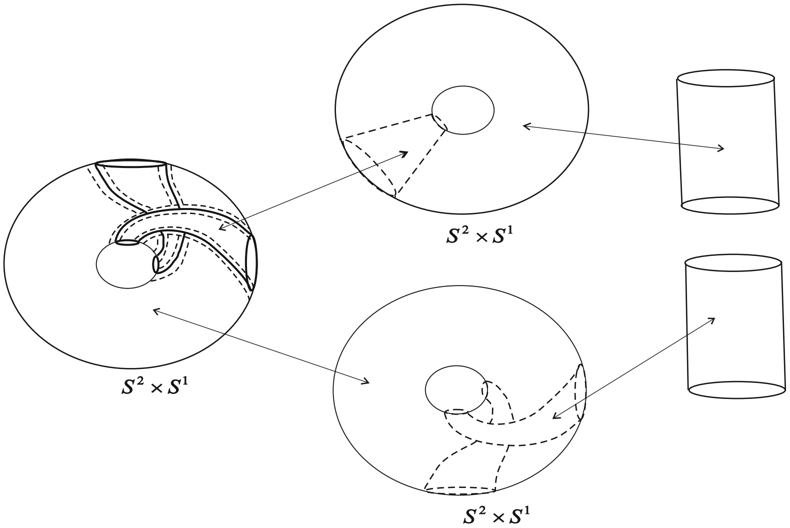


Fig. 31.10 Cutting and pasting along a torus

Notice that the result of cutting and pasting V along $\hat{\Lambda}^{\delta}$ does not depend on an order of operations. Denote $(V, \hat{\Lambda}^{\delta})$ the manifold which is obtained from V by *cutting and pasting along* $\hat{\Lambda}^{\delta}$.

Definition 31.3. A collection $S = (V, \alpha, \hat{\Lambda}^s, \hat{\Lambda}^u)$ is called *perfect scheme* if each connected component of $(V, \hat{\Lambda}^s)$ and $(V, \hat{\Lambda}^u)$ is diffeomorphic to $\mathbb{S}^2 \times \mathbb{S}^1$.

Theorem 31.10. *([13], Proposition 2.2) For any perfect scheme S there is a diffeomorphism $f \in G_0^3$, whose global scheme is equivalent to S.*

31.1.7 Topological Classification of Non Gradient-Like Diffeomorphisms on 3-Manifolds

In this section firstly we give a complete classification of a class H_1^3 of Morse–Smale diffeomorphisms f with the finite number of heteroclinic orbits and without heteroclinic curves on 3-manifold M.

Similar to gradient-like diffeomorphism we form the global connected attractor $\mathscr{A}_f^+$ and repeller $\mathscr{A}_f^-$, the wandering space $\mathscr{M}_f = M^3 \setminus (\mathscr{A}_f^+ \cup \mathscr{A}_f^-)$, the space of orbits $V_f = \mathscr{M}_f/f$, the natural projection $p_{_f} : \mathscr{M}_f \to V_f$ and epimorphism $\alpha_{_f} : \pi_1(V_f) \to \mathbb{Z}$ corresponding to cover $p_{_f}$.

Set $\hat{\lambda}_{\sigma}^s = p_{_f}(W^s(\sigma)\setminus\sigma)$ $(\hat{\lambda}_{\sigma}^u = p_{_f}(W^u(\sigma)\setminus\sigma))$ for any $\sigma \in \Sigma_f^+$ $(\sigma \in \Sigma_f^-)$ and $\hat{\Lambda}_f^s = \bigcup\limits_{\sigma \in \Sigma_f^+} \hat{\lambda}_{\sigma}^s$ $(\hat{\Lambda}_f^u = \bigcup\limits_{\sigma \in \Sigma_f^-} \hat{\lambda}_{\sigma}^u)$. The set $\hat{\Lambda}_f^s$ $(\hat{\Lambda}_f^u)$ consists of two-dimensional laminations, which are called *heteroclinic lamination*.

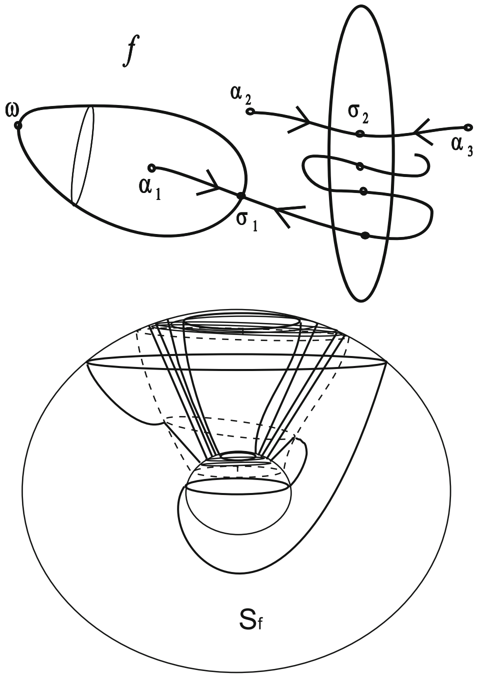


Fig. 31.11 A global scheme of a diffeomorphism $f \in H_1^3$

Definition 31.4. A collection $S_f = (V_f, \alpha_f, \hat{\Lambda}^u_f, \hat{\Lambda}^s_f)$ is called *global scheme* of diffeomorphism $f \in H_1^3$.

On Fig. 31.11 is represented a phase portrait of diffeomorphism $f : \mathbb{S}^3 \to \mathbb{S}^3$ and its global scheme S_f.

Definition 31.5. Global schemes $S_f = (V_f, \alpha_f, \hat{\Lambda}^u_f, \hat{\Lambda}^s_f)$ and $S_{f'} = (V_{f'}, \alpha_{f'}, \hat{\Lambda}^u_{f'}, \hat{\Lambda}^s_{f'})$ of diffeomorphisms $f, f' \in H_1^3$ are called *equivalent* if there is a preserving orientation homeomorphism $\hat{\varphi} : V_f \to V_{f'}$ such that $\alpha_f = \alpha_{f'}\hat{\varphi}_*$ and $\hat{\varphi}(\hat{\Lambda}^u_f) = \hat{\Lambda}^u_{f'}, \hat{\varphi}(\hat{\Lambda}^s_f) = \hat{\Lambda}^s_{f'}$.

Theorem 31.11. *([16] Theorem 2.1) Diffeomorphisms $f, f' \in H_1^3$ are topological conjugated if and only if their global schemes S_f, $S_{f'}$ are equivalent.*

Similar to gradient-like diffeomorphisms it is defined perfect schemes and is proved realization theorem.

Secondly, we give a complete topological classification of Morse–Smale diffeomorphisms f on 3-manifold M belonging to a class Q_n^3 $(n \geq 0)$ of diffeomorphisms satisfying to the next conditions:

(1) Nonwandering set $\Omega(f)$ consists of fixed points
(2) The number of saddle points is equal to $n + 1$

(3) All saddle points $\sigma_0, \dots, \sigma_n \in \Omega(f)$ have the same Morse index[1] and form n-chain (connecting σ_0 and σ_n). For definiteness we will suppose that Morse index of saddles is equal to 2 (the case when all saddles of a diffeomorphism f have Morse index 1 reduces to our case by consideration of the diffeomorphism f^{-1}).

Complete topological classification of diffeomorphisms from the class Q_0^3 is contained in Sect. 31.1.4. To each diffeomorphism of this class corresponds a knot embedded in the manifold $\mathbb{S}^2 \times \mathbb{S}^1$ and classification such diffeomorphisms is equivalent to classification of corresponding knots. Diffeomorphisms of the class Q_1^3 are contained in the class H_1^3. To each diffeomorphism of this class corresponds a heteroclinic lamination in the manifold $\mathbb{S}^2 \times \mathbb{S}^1$ and classification such diffeomorphisms is equivalent to classification of corresponding heteroclinic laminations.

In this paper we consider class Q_n^3 for $n \geq 2$.

Theorem 31.12. *([18], Theorem 1) Nonwandering set of any diffeomorphism $f \in Q_n^3$ consists of $2n+4$ fixed points: one sink ω_0, $n+2$ sources $\alpha_0, \dots, \alpha_{n+2}$, $n+1$ saddles $\sigma_0, \dots, \sigma_n$ and ambient manifold M is homeomorphic to the manifold S^3.*

For a diffeomorphism $f \in H_n^3$ the global attractor $\mathscr{A}_f^+$ consists of exactly one point ω_0 and, hence, the orbits space V_f is diffeomorphic to $\mathbb{S}^2 \times \mathbb{S}^1$. To each diffeomorphism $f \in H_n^3$ we assign the orbit space $\Lambda_n(f)$ of the diffeomorphism f action on the set $\bigcup\limits_{i=0}^{n} W^u(\sigma_i) \setminus \bigcup\limits_{i=0}^{n} W^s(\sigma_i)$. The set $\Lambda_n(f)$ is a torus heteroclinic lamination of order n (see Fig. 31.12). In Fig. 31.12 in the center a three-dimensional annulus is represented. After gluing its boundary spheres, the needed manifold $\mathbb{S}^2 \times \mathbb{S}^1$ and lamination $\Lambda_2 = \mathscr{T}_0 \cup \mathscr{T}_1 \cup \mathscr{T}_2$ are obtained. Below the union $\mathscr{T}_0 \cup \mathscr{T}_1 \cup \mathscr{T}_2$ is represented after gluing.

Theorem 31.13. *([18], Theorem 3) Diffeomorphisms $f, f' \in Q_n^3$ are topological conjugate if and only if $\Lambda_n(f)$ and $\Lambda_n(f')$ are equivalent.*

Theorem 31.14. *([18], Theorem 3) For any torus heteroclinic lamination Λ_n of order n there is a diffeomorphism $f \in Q_n^3$ such that $\Lambda_n(f)$ and Λ_n are equivalent.*

31.1.8 Peixoto' Graph is Complete Invariant Again

Let H_0^n is the class of Morse–Smale diffeomorphisms on manifold M of dimension $n \geq 4$ such that for any $f \in H_0^n$ the set of unstable separatrixes has dimension 1 and does not contain the heteroclinic orbits. We will associated with any $f \in H_0^n$ the oriented graph Γ_f which is similar to graph introduced by Peixoto for structural

[1] Morse index of a periodic point is dimension of its unstable manifold.

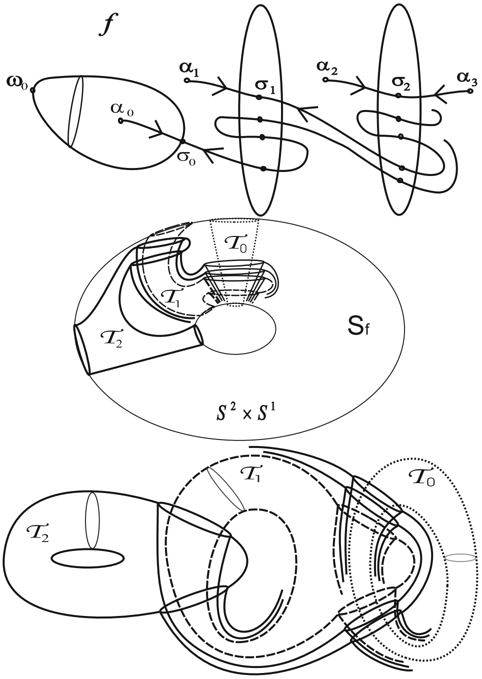


Fig. 31.12 A torus heteroclinic lamination of order 2 for a diffeomorphism $f \in Q_2^3$

stable flow and, hence, for gradient-like diffeomorphisms on two-dimensional manifolds. The set of vertices of Γ_f is isomorphic to the non-wandering set $\Omega(f)$, the set of edges of Γ_f is isomorphic to the set of separatrixes of saddle periodic points.

Theorem 31.15. *([29], Theorem 1) Diffeomorphisms $f, f' \in H_0^n$ are topologically conjugated iff graphs $\Gamma_f, \Gamma_{f'}$ are isomorphic.*

In case $n = 2$ this result follows from results of the papers [8]–[9] and [27]. In case $n = 3$ it contrasts with results of Sect. 31.1.4 where, in particular, it is shown, that there are countable set topologically non-conjugated Morse–Smale diffeomorphisms with isomorphic graphs.

We give a representation of each topological conjugacy subclass of H_0^n.

We say that connected orientable graph Γ is *admissible* if the set of vertixes of Γ may be represented as a union of three non-empty disjoint subsets $\Gamma_1^0 = \{a_1^1\}$, $\Gamma_2^0 = \{a_2^1, \dots, a_2^k\}$, $\Gamma_3^0 = \{a_3^1, \dots, a_3^{k+1}\}$ such that:

(1) For any $i \in \{1, \dots, k\}$ the vertix a_2^i is incident for exactly three edges: one edge joins a_2^i with a_1^1 and two edges join a_2^i with two different vertixes from set Γ_3^0
(2) There are no edges joining any two vertixes from Γ_3^0 and there are no edges joining a_1^1 with a vertix from Γ_3^0
(3) For any $j \in \{1, \dots, k\}$ the edge (a_1^1, a_2^j) is oriented from a_1^1 to a_2^j

(4) For any couple $i \in \{1, \dots, k\}$, $j \in \{1, \dots, k+1\}$ such that vertixes a_2^i, a_3^j are incident the edge (a_2^i, a_3^j) is oriented from a_2^i to a_3^j

(5) Graph $\Gamma \setminus a_1^1$ is connected.

Lemma 31.3. *([29], Theorem 2) Graph $\Gamma(f)$ of diffeomorphism $f \in H_0^n$ is admissible.*

Theorem 31.16. *([30], Theorem 3) Let P is any orientation preserving automorphism of admissible graph Γ. Then there is diffeomorphism $f \in H_0^n$ such that $\Gamma(f) = \Gamma$ and automorphism $P(f)$ of $\Gamma(f)$ induced by f coincides with P.*

Acknowledgements The research was partially supported by RFFI grants 08-01-00547, 11-01-00730 and grant of government of Russian Federation No 11.G34.31.0039.

References

1. Andronov, A., Pontrjagin, L.: Rough systems. Dokl. Akad. Nauk. **14**, 247–250 (1937)
2. Anosov, D.V.: Structurally stable systems. Tr. Mat. Inst. im. V.A. Steklova Akad. Nauk SSSR **169**, 59–93(1985); Engl. transl. Proc. Steklov Inst. Math. **169**, 61–95 (1985)
3. Aranson, S., Grines, V.: Topological classification of cascades on closed two-dimensional manifolds. Usp. Mat. Nauk. **45**(4), 3–32 (1990). Engl. transl. in: Russ. Math. Surv. **45**(1) (1990), 1–35
4. Aranson, S., Grines, V.: Cascades on surfaces. Dynamical Systems-9. VINITI. Moscow. (1991), Itogi Nauki Tekh., Ser.: Sovrem. Probl. Mat., Fundam. Napravl. **66**, 148–187; Engl. transl. Dynamical Systems IX (Springer, Berlin, 1995), Encycl. Math. Sci. **66**, 141–175
5. Arnold, V.: Small denominators I. Mapping of the circle onto itself. Izvestia AN SSSR, ser. matem. **25**, 21–86(1961) (Russian); MR 25#4113
6. Artin, E., Fox, R.H.: Some wild cells and spheres in three-dimensional space. Ann. Math. **49**, 979–990 (1948)
7. Bezdenezhnykh, A., Grines, V.: Diffeomorphisms with orientable heteroclinic sets on two-dimensional manifolds. (Russian) In: Qualitative Methods of the Theory of Differential Equations, pp. 139–152. Gorky (1985)
8. Bezdenezhnykh, A., Grines, V.: Dynamical properties and topological classification of gradient-like diffeomorphisms on two-dimensional manifolds. Part 1. Sel. Math. Sov. **11**(1), 1–11 (1992a)
9. Bezdenezhnykh, A., Grines, V.: Realization of gradient-like diffeomorphisms of two-dimensional manifolds. Sel. Math. Sov. **11**(1), 19–23 (1992c)
10. Bonatti, Ch., Grines, V.: Knots as topological invariant for gradient-like diffeomorphisms of the sphere S^3. J. Dyn. Control Syst. **6**, 579–602 (2000)
11. Bonatti, Ch., Grines, V., Medvedev, V., Pecou, E.: On topological classification of gradient-like diffeomorphisms without heteroclinic curves on 3-manifolds. Dokl. RAN. **377**(2), 151–155 (2001b)
12. Bonatti, Ch., Grines, V., Medvedev, V., Pecou, E.: Three-manifolds admitting Morse-Smale diffeomorphisms without heteroclinic curves. Topol. Appl. **117**, 335–344 (2002b)
13. Bonatti, Ch., Grines, V., Medvedev, V., Pecou, E.: Topological classification of gradient-like diffeomorphisms on 3-manifolds. Topology **43**, 369–391 (2004)
14. Bonatti, Ch., Grines, V., Pochinka, O.: Classification of the simplest non gradient-like diffeomorphisms on three-manifolds. Contemporary mathematics and its applications. Institute of Cybernetics of Academy of Science of Georgia. **7**, 43–71 (2003). Transl. in J. Math. Sci. (N. Y.) **126**(4), 1267–1296 (2005)

15. Bonatti, Ch., Grines, V., Pochinka, O.: Classification of Morse-Smale diffeomorphisms with a finite set of heteroclinic orbits on 3-manifolds. (Russian) Dokl. Akad. Nauk. **396**(4), 439–442 (2004)
16. Bonatti, Ch., Grines, V., Pochinka, O.: Classification of Morse-Smale diffeomorphisms with finite number heteroclinic orbits on 3-manifolds. Trudy MIAN **250**, 5–53 (2005)
17. Bonatti, Ch., Grines, V., Pochinka, O.: On Existence of a Smooth Arc Joining "North Pole-South Pole" Diffeomorphisms. Prepubl. Inst. Math. Bourgogne. http://math.u-bourgogne.fr/topologie/prepub/bifs$_4$.pdf (2006)
18. Bonatti, Ch., Grines, V., Pochinka, O.: Classification of Morse-Smale diffeomorphisms with the chain of saddles on 3-manifolds. Foliations 2005, pp. 121–147. World Scientific, Singapore, (2006)
19. Bonatti, Ch., Grines, V., Medvedev, V., Pochinka, O.: Bifurcations of Morse-Smale diffeomorphisms with wildly embedded separatrices. Trudy MIAN **256**, 54–69 (2007)
20. Bonatti, C., Langevin, R.: Difféomorphismes de Smale des surfaces. Astérisque **250**, Société mathématique de France (1998)
21. Cantrell, J.: n-frames in Euclidean k-space. Proc. Am. Math. Soc. **15**(4), 574–578 (1964)
22. Cerf, J.: Sur les diffeomorphismes de la sphere de dimension trois ($\Gamma_4 = 0$). Lecture Notes in Math., vol. 53. Springer, Berlin (1968)
23. Conley, C.: Isolated invariant sets and morse index. CBMS Regional Conference Series in Math., vol.38. AMS, Providence, RI (1978)
24. Daverman, R., Venema, G.: Embeddings in Manifolds. Graduate studies in Mathematics, vol. 106. AMS, Providence, RI (2009)
25. Debrunner, H., Fox, R.: A mildly wild imbedding of an n-frame. Duke Math. J. **27**, 425–429 (1960)
26. Franks, J.: Nonsingular Smale flow on S^3. Topology **24**(3), 265–282 (1985)
27. Grines, V.: Topological classification of Morse-Smale diffeomorphisms with a finite set of heteroclinic trajectories on surfaces. Math. Zametki. **54**(3), 3–17 (1993)
28. Grines, V.: On topological classification of A-diffeomorphisms of surfaces. J. Dyn. Control Syst. **6**(1), 97–126 (2000)
29. Grines, V., Gurevich, E.: About Morse-Smale diffeomorphisms on manifolds of dimention greater then three. Dokl. Akad. Nauk. **416**(1), 15–17 (2007)
30. Grines, V., Gurevich, E., Medvedev, V.: Peixoto's graph for Morse-Smale diffeomorphisms on manifolds of dimensions greater than 3. Trudy MIAN **261**, 61–86 (2008a)
31. Grines, V., Laudenbach, F., Pochinka, O.: An energy function for gradient-like diffeomorphisms on 3-manifolds. (Russian) Dokl. Akad. Nauk. **422**(3), 299–301 (2008b)
32. Grines, V., Laudenbach, F., Pochinka, O.: Self-indexing energy function for Morse-Smale diffeomorphisms on 3-manifolds. Moscow Math. J. **4**, 801–821 (2009)
33. Grines, V. and Medvedev, V., Zhuzhoma, E.: New relation for Morse-Smale systems with trivially embedded one-dimensional separatrices. Sb. Math. **194**, 979–1007 (2003)
34. Keldysh, L.: Topological embeddings in euclidean space. Proceedings of Math. Inst. V. A. Steklova., vol. 81. Nauka, Moscow (1966)
35. Langevin, R.: Quelques nouveaux invariants des difféomorphismes de Morse-Smale d'une surface. Ann. Inst. Fourier, Grenoble. **43**, 265–278 (1993)
36. Maier, A.G.: Rough map of circle to circle. Uch. Zap. GGU. Gorky. Izd-vo GGU. **12**, 215–229 (1939)
37. Matsumoto, S.: There are two isotopic Morse-Smale diffeomorphism which can not be joined by simple arcs. Invent. Math. **51**, 1–7 (1979)
38. Mazur, B.: Note on some contractible 4-manifolds. Ann. Math. Second Ser. **73**(1), 221–228 (1961)
39. Milnor, J.: On manifolds homeomorphic to 7-sphere. Ann. Math. **64**(2), 399–405 (1956)
40. Newhouse, S., Peixoto, M.: There is a simple arc joining any two Morse-Smale flows. Asterisque **31**, 15–41 (1976)
41. Nielsen, J.: Die Structure periodischer Transformation von Flachen. Det. Kgl. Dansk Videnskaternes Selskab. Math.-Phys. Meddelerser. **15** (1937)

42. Nitecki, Z.: Differentiable dynamics. MIT, Cambridge (1971)
43. Palis, J.: On Morse-Smale dynamical systems. Topology **8**(4), 385–404 (1969)
44. Palis, J., Pugh, C.: Fifty problems in dynamical systems. Lecture Notes in Math. **468** (1975), 345–353.
45. Peixoto, M.C., Peixoto, M.M.: Structural stability in the plane with enlarged boundary conditions. An. Acad. Brasil. Cienc. **31**(2), 135–160 (1958)
46. Peixoto, M.M.: On structural stability. Ann. Math. **69**, 199–222 (1959)
47. Peixoto, M.M.: Structural stability on two-dimensional manifolds. Topology **1**, 101–120 (1962); A further remark. Topology **2**, 179–180 (1963)
48. Pixton, D.: Wild unstable manifolds. Topology **16**, 167–172 (1977)
49. Pliss, V.: On roughness of differential equations given on torus. Vestnik LGU. Ser. Math. Mech. **13**(3), 15–23 (1960)
50. Pochinka, O.: On topological conjugacy of the simplest Morse-Smale diffeomorphisms with a finite number of heteroclinic orbits on S^3. Prog. nonlin. Sci. **1**, 338–345 (2002)
51. Robinson, C.: Dynamical Systems: stability, symbolic dynamics, and chaos. Studies in Advanced Mathematics, 2nd edn. CRC, Boca Raton, FL (1999)
52. Shub, M.: Morse-Smale diffeomorphism are unipotent on gomology. In: Peixoto, M. (ed.) Dynamical Systems, pp. 489–491. Academic, New York (1973)
53. Smale, S.: Morse inequalities for a dynamical system. Bull. Am. Math. Soc. **66**, 43–49 (1960)
54. Smale S.: Differentiable dynamical systems. Bull. Am. Math. Soc. **73**, 747–817 (1967)
55. Takens, F.: Tolerance stability. In: Dynamical Systems. Warwick 1974, pp. 293–304. Springer, Berlin (1975)
56. Thurston, W.: The geometry and topology of three-manifolds. Lecture Notes. Princeton University, Princeton (1976–1980)
57. Wilson, W.: Smoothing derivatives of functions and applications. Trans. Am. Math. Soc. **139**, 413–428 (1969)

Chapter 32
Isentropic Dynamics and Control in an Economic Model for Capital Accumulation

Cristina Januário, Clara Grácio, Diana A. Mendes, and Jorge Duarte

Abstract The study of economic models has generated deep interest in exploring the complexity of our society. The primary purpose of this article is to study the chaotic dynamical behavior of an economic growth model describing capital accumulation presented by Böhm and Kaas in (J Econ Dyn Control 24:965–980, 2000). To start with, we use the techniques of symbolic dynamics to explore several properties, with the explicit computation of two topological invariants, which are associated with the discrete dynamical system in consideration. The analysis of these results allows us to understand the dynamics of the economical model and to distinguish different scenarios of complexity, namely in situations of isentropic dynamics. Finally, we show that the chaotic behavior arising from the discrete model can be controlled without changing its original properties and the dynamics can be turned into a desired attracting time periodic motion (a stable steady state or a regular cycle). The orbit stabilization is illustrated by a analytical control technique. This study tends to integrate and interrelate different methods in order to illustrate how our understanding of economic models can be enhanced by the theory of nonlinear dynamical systems.

C. Januário and J. Duarte (✉)
Department of Chemistry, Mathematics Unit, ISEL-High Institute of Engineering of Lisbon,
Rua Conselheiro Emdio Navarro, 1949-014 Lisbon, Portugal
e-mail: cjanuario@deq.isel.ipl.pt, jduarte@deq.isel.ipl.pt

C. Grácio
Department of Mathematics, Universidade de Évora, Rua Romão Ramalho, 59, 7000-585 Évora, Portugal
e-mail: mgracio@uevora.pt

D.A. Mendes
Department of Quantitative Methods, Instituto Superior de Ciências do Trabalho e da Empresa,
Avenida das Forças Armadas, 1649-026 Lisbon, Portugal
e-mail: diana.mendes@iscte.pt

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_32,

32.1 Introduction

Nonlinear dynamics in economical models leads to potential complexity and unpredictability which are significant obstacles in understanding the qualitative behavior of such dynamical systems. In the last years, the application of tools from nonlinear analysis, in particular chaos theory to the study of complex economical systems, seems to be relevant and has generated extensive research programmes (for instance: [8, 10, 14]). However, from the point of view of economists, chaos is not clearly understood.

Dynamic economic growth models have often considered the standard one-sector neoclassical model by Ramsey (1924) or the Solow–Swan model (1956). In both cases the economic models are characterized by a monotonically convergence to the steady state so no periodic fluctuations or complex dynamics can be observed. Nevertheless, at the same time, other one-sector growth models were developed with the capacity of generating multiple and unstable steady states, particularly those introduced by Kaldor and Pasinetti. In 2000, in a very interesting paper, Bohm and Kaas [2], analyze the role of differential savings behavior as proposed by Kaldor (1956) and its consequences regarding the stability of stationary equilibria in a discrete-time Solow growth model. These authors encountered a very rich dynamical behavior, characterized by stable/unstable equilibrium points, fluctuations and even topological chaos, when the income distribution varies sufficiently and if shareholders save more than workers. More recently, in 2007, Brianzoni et al. [3] studied the Bohm and Kaas model considering a different production function and a non-constant labor force growth.

Due to important developments in nonlinear dynamics, there has been a considerable research effort into the analysis of chaotic systems. For instance, control, targeting, synchronization and forecasting of chaotic motion have proved well established results in the fields of applied mathematics, economy, physics and engineering. In particular, since the publication of the seminal paper of Ott, Grebogi and Yorke in 1990 [19], several methods have been proposed to control chaotic dynamics, with applications, for example, to economy, biochemistry, cardiology, communications, physics laboratories and turbulence. This pioneering work showed that very small changes of a parameter, when performed in a convenient way, can effectively control a chaotic dynamic. Other controlling methods have been published since that time using proportional feedback, small periodic perturbations of a parameter or regulator pulses on a variable in order to achieve the desired regular motion (see [5] and references therein).

In the context of economy, practical methods of this new and exciting field can be applied to show that the presence of chaotic motion in economic processes does not necessarily need to be interpreted as a curse for economic theory and economic policy (for instance: [6,7,20,21]). Particularly, in order to control economic chaotic motion, we do not need to change the fundamental characteristics of the system. We can eliminate large business cycles, leaving the main features unchanged.

The aim of this paper is to provide a contribution for the detailed analysis of the chaotic behavior of the neoclassical one-sector growth model with differential savings in the sense of Kaldor–Pasinetti, as presented in the paper of Bohm and Kass (see [2]). Using results of symbolic dynamics theory, we compute a quantifier for the complex orbit structure – an attribute used to define chaos – the topological entropy. This topological invariant is related to the exponential orbital growth and gives us a finer distinction between different chaotic states. For certain type of maps, the study of this measure of the amount of chaos leads to situations of isentropic dynamics. The characterization of the isentropic maps becomes possible with the introduction of another topological invariant that allows us to distinguish different scenarios of complexity. We exhibit numerical results about the relation of this particular topological invariant and each of the control parameters. It is interesting to notice that, although the concept of entropy was originally developed in a thermodynamic context, it has been adapted in other different fields of research, including thermoeconomics, information theory, evolution and string theory. For instance in [9], a number of interconnected issues involving superstring theory, entropy and the particle content of the standard model of high energy physics, have been analysed. The identification of chaotic states can be efficiently used to apply chaos control strategies. In this context, we examine the effects of periodic proportional pulses on the stabilization of chaotic trajectories performed by the economic model. This control method was presented by Chau in [4] based on the work carried out by Matias and Güémez in [11].

32.2 Description of the Model

We consider a one-sector growth model in the sense of Kaldor and Pasinetti, where there are two types of agents: workers and shareholders. They may have possibly different but constant savings propensities. A single investment/consumption commodity is produced from labor and capital input with constant returns to scale. The production function, $f : R_+ \to R_+$, with the property that transforms capital per worker k into output per worker y, satisfies the weak Inada condition, that is, f is C^2, strictly monotonically increasing, strictly concave, and such that

$$\lim_{k\to\infty} \frac{f(k)}{k} = 0 \text{ and } \lim_{k\to 0} \frac{f(k)}{k} = \infty. \tag{32.1}$$

The labor force growth at rate $n \geq 0$ (as usually assumed in standard economic growth theory) and capital depreciates at rate $0 < \delta \leq 1$. The wage rate is characterized by the following relation

$$w(k) = f(k) - kf'(k), \tag{32.2}$$

where $f'(k)$ is the marginal product of capital received by the shareholders and $kf'(k)$ represents the total capital income per worker. The constant saving rates for workers, s_w, and shareholders, s_r, are both limited between 0 and 1.

It follows that the one dimensional map, which describes the capital accumulation, is given by

$$k_{t+1} = G(k_t) = \frac{1}{n+1}\left((1-\delta)k_t + s_w w(k_t) + s_r k_t f'(k_t)\right), \tag{32.3}$$

and is continuously depending on f and on the parameters s_w, s_r, n, δ. If the two savings propensities are equal, then the standard growth model of Solow is obtained. In this case, there exists a unique globally stable equilibrium point, which is not optimal, that is, it does not maximize long-run consumption per capita. Dynamic behavior different from stable steady state can be obtained by considering several production function in (32.3) and varying the model parameters.

Among the production functions pointed out in [2], we consider the concave production function which is taken as an approximation of the Leontief technology, that is,

$$f(k) = a\left(k + \alpha \ln\left(\frac{1+e^{-b/(\alpha a)}}{1+e^{(ak-b)/(\alpha a)}}\right)\right) + c, \tag{32.4}$$

and study the dynamics of the one-dimensional map presented in (32.3) for this specified f. The real parameters a, b, c and α are all positive. For more details the reader is referred to the paper [2] and references therein. In our numerical investigation, we will use throughout the following standard parameter calibration: $n = 0.0$, $s_w = 0.4$, $a = 0.2$, $b = 1$, $c = 0.01$ and $\alpha = 0.01$ and consider δ and s_r as control parameters. A typical map of G is shown in Fig. 32.1, which is a bimodal map with turning points c_1 (corresponding to the relative maximum) and c_2 (corresponding to the relative minimum).

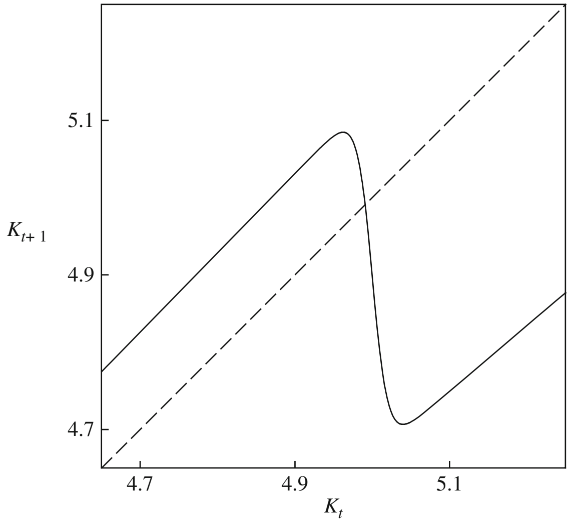


Fig. 32.1 Graphical representation of G. In this case $\delta = 0.148$ and $s_r = 0.87$

32.3 Coping with Complex Behavior and Isentropic Dynamics

It is well known that topological and metric invariants like entropy, Lyapunov exponents, information and correlation dimensions, are fundamental in the local and global characterization of the behavior of a nonlinear dynamical system. For example, the entropy, which is considered one of the most powerful invariant [13], was already successfully applied to quantify information and uncertainty in financial data, or in the prediction of chemical processes, among others. Moreover, there are some conjectures that affirm that the direction of economic change may have as much to do with the entropies of neighboring macrostates as with any of the other dynamical factors now recognized [12]. In another paper, Montrucchio and Sorger [18], derived a simple relationship between the topological entropy of the optimal policy function of a concave dynamic program and the underlying discount factor. They obtained a relationship reflecting that solutions with very complicated dynamics can only occur in models with small discount factors. Motivated by all these results, we will give some importance to the computation and interpretation of topological entropy in our economical model.

In this section we will use techniques of symbolic dynamics, in particular some results concerning Markov partitions associated with bimodal maps in order to compute the topological entropy of the capital accumulation model. We describe briefly this powerful symbolic tool (for more details see [15] and [16]).

Let us consider that the bimodal map G is defined on the interval $I = [c_0, c_3]$. G is a piecewise monotone map, where I is subdivided into three subintervals:

$$I_L = [c_0, c_1[, \quad I_A = \{c_1\}, \quad I_M =]c_1, c_2[, \quad I_B = \{c_2\}, \quad I_R =]c_2, c_3], \tag{32.5}$$

in such way that the restriction of G to each interval I_L or I_R is strictly increasing and in the other interval I_M is strictly decreasing (see for instance Fig. 32.1). Each such maximal intervals on which the function G is monotone is called a *lap* of G, and the number $\ell = \ell(G)$ of distinct laps is called the lap number of G.

Denoting by c_1 and c_2 the two turning points of G, we obtain the orbits

$$O(c_1) = \left\{x_i : x_i = G^i(c_1),\ i \in \mathbb{N}\right\} \text{ and } O(c_2) = \left\{y_i : y_i = G^i(c_2),\ i \in \mathbb{N}\right\}. \tag{32.6}$$

With the aim of studying the topological properties of these orbits we associate with each orbit $O(c_i)$ a sequence of symbols $S = S_1 S_2 ... S_j ...$ where

$$\begin{cases} S_j = L \text{ if } G^j(c_i) < c_1, \\ S_j = A \text{ if } G^j(c_i) = c_1, \\ S_j = M \text{ if } c_1 < G^j(c_i) < c_2, \\ S_j = B \text{ if } G^j(c_i) = c_2, \\ S_j = R \text{ if } G^j(c_i) > c_2. \end{cases} \tag{32.7}$$

The critical points c_1 and c_2 play an important role since the dynamics of the bimodal map on the interval is fully characterized by the symbolic sequences associated with the orbits of these points.

We denote by $n_M(S)$ the frequency of the symbol M in S and we define the M-parity of this sequence,

$$\rho(S) = (-1)^{n_M(S)}, \tag{32.8}$$

according to whether $n_M(S)$ is even or odd. Thus, in the first case we have $\rho(S) = +1$ and in the second $\rho(S) = -1$. In our study we use an order relation defined in $\Sigma = \{L, A, M, B, R\}^{\mathbb{N}}$ that depends on M-parity. Thus, for two of such sequences, P and Q in Σ, let i be such that $P_i \neq Q_i$ and $P_j = Q_j$ for $j < i$. If the M-parity of the block $P_1 ... P_{i-1} = Q_1 ... Q_{i-1}$ is even (that is, $\rho(P_1 ... P_{i-1}) = +1$), we say that $P < Q$ if $P_i < Q_i$ in the order $L < A < M < B < R$. If the M-parity of the same block is odd (that is, $\rho(P_1 ... P_{i-1}) = -1$), we say that $P < Q$ if $P_i < Q_i$ in the order $R < B < M < A < L$. If no such index i exists, then $P = Q$.

If a finite symbolic sequence S has n symbols, it is common to write $|S| = n$. When $O(c_i)$ is a k-periodic orbit we obtain a sequence of symbols that can be characterized by a block of length k, that is

$$S^{(k)} = S_1 ... S_{k-1} C_i, \text{ with } i = 1, 2. \tag{32.9}$$

In what follows, we restrict our study to the case where the two critical points are periodic (respectively, eventually periodic), $O(c_1)$ is p-periodic and $O(c_2)$ is q-periodic (respectively, $G^p(c_1) = c_2$ or $G^q(c_2) = c_1$). Note that $O(c_1)$ is realizable if the block $P = P_1 ... P_{p-1} A$ is maximal, that is, $\sigma^i(P) \leq P$, where $1 \leq i \leq p$ and

$$\sigma(P_i P_{i+1} P_{i+2} ...) = P_{i+1} P_{i+2} ... \tag{32.10}$$

is the usual shift operator. On the other hand, $O(c_2)$ is realizable if the block $Q = Q_1 ... Q_{q-1} B$ is minimal, that is, $\sigma^j(Q) \geq Q$, where $1 \leq j \leq q$. Finally, note that the pair of sequences that are realizable satisfies the following conditions

$$\sigma^i(P) \geq Q, \ 1 \leq i \leq p \text{ and } \sigma^j(Q) \leq P, \ 1 \leq j \leq q. \tag{32.11}$$

The set of such pair of sequences is denoted by $\Sigma_{(A,B)}$.

We designate by kneading data the pairs $(P^{(p)}, Q^{(q)}) \in \Sigma_{(A,B)}$, where

$$P^{(p)} = P_1 ... P_{p-1} A \text{ and } Q^{(q)} = Q_1 ... Q_{q-1} B. \tag{32.12}$$

The bistable sequence is denoted by $P_1 ... P_{p-1} B Q_1 ... Q_{q-1} A$, and the eventually periodic sequences are given by

$$P_1 ... P_{p-1} B Q_1 ... Q_{q-1} B \text{ or } Q_1 ... Q_{q-1} A P_1 ... P_{p-1} A. \tag{32.13}$$

To each value of the control parameters, the dynamics is characterized using the kneading data. This kneading data determines a Markov partition of the interval, considering the orbits $O(c_1) = \{x_i\}_{i=1,2,\dots,p}$ and $O(c_2) = \{y_i\}_{i=1,2,\dots,q}$, and ordering the elements x_i, y_i of these orbits. With this procedure we obtain the partition $\{I_k = [z_k, z_{k+1}]\}_{k=1,2,\dots,p+q}$ of the interval $I = [y_1, x_1]$. The transitions between the subintervals are represented by a matrix $\mathscr{M}(G)$. According to the above description, the topological entropy of G, denoted by $h_{top}(G)$, can be given by

$$h_{top}(G) = \log \lambda_{\max}(\mathscr{M}(G)) = \log s(G), \tag{32.14}$$

where $\lambda_{\max}(\mathscr{M}(G))$ is the spectral radius of the transition matrix $\mathscr{M}(G)$ and $s(G)$ is the growth rate of the number of intervals on which G^k is monotone.

To illustrate the previous considerations, we discuss the following example.

Example 32.1. Let us consider the map of Fig. 32.1. The orbits of the turning points define the pair of sequences $(RLLMA, LLA)^{\infty}$. Putting the points of the orbits in order we obtain:

$$y_1 < x_2 < y_2 < x_3 < c_1 = x_0 < x_4 < c_2 = y_0 < x_1. \tag{32.15}$$

The correspondent transition matrix is

$$\mathscr{M}(G) = \begin{bmatrix} 0&0&1&0&0&0&0 \\ 0&0&0&1&0&0&0 \\ 0&0&0&0&1&0&0 \\ 0&0&0&0&0&1&1 \\ 0&0&0&0&1&1&1 \\ 1&1&1&1&0&0&0 \\ 1&0&0&0&0&0&0 \end{bmatrix}$$

which has the characteristic polynomial $p(\lambda) = -t^3(-1-t-t^2-t^3+t^4)$. Therefore the value of the topological entropy is $h_{top}(G) = 0.656256\ldots.$

To see the long term behavior for different values of the parameters, we plot, in Figs. 32.2 and 32.3 typical bifurcation diagrams. It is interesting to observe from the bifurcation diagrams, and by other hand obvious to confirm, that when the capital depreciation rate, δ, is growing then the capital accumulation is decreasing in a quite abrupt way and when the saving rate for holders, s_r, is increasing then the capital accumulation is increasing too.

With these diagrams it is easier to understand Figs. 32.4 and 32.5 that present some numerical results of the Lyapunov exponents with each of the parameters δ and s_r, in some regions of the parameter space. Notice that the depicted values of the Lyapunov exponents above the zero line correspond to chaotic behavior which is associated with positive topological entropy. More precisely, the variation of the topological entropy follows the behavior of the Lyapunov exponents. These variations are non-monotone and are characterized by several fluctuations (successive

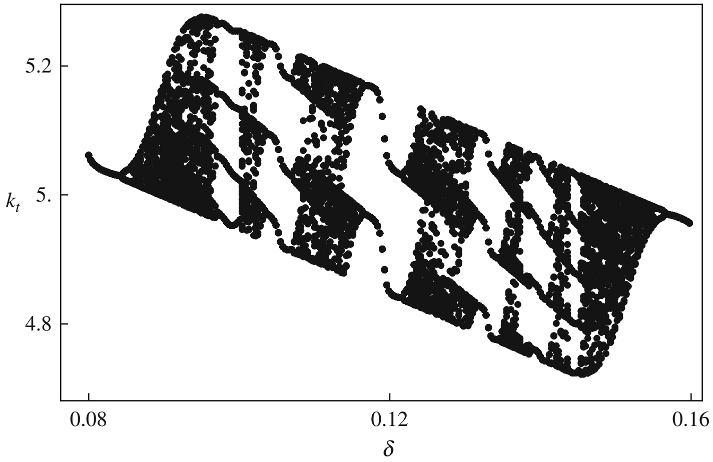


Fig. 32.2 Bifurcation diagram for k_t as a function of δ, with $s_r = 0.8$ and $0.08 \leq \delta \leq 0.16$

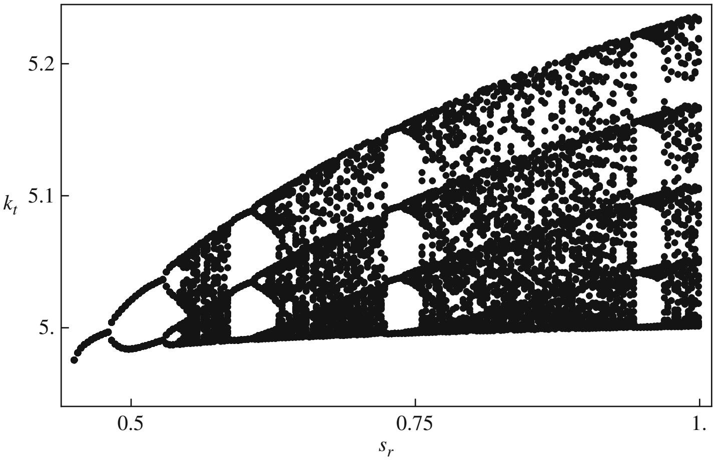


Fig. 32.3 Bifurcation diagram for k_t as a function of s_r, with $\delta = 0.09$ and $0.45 \leq s_r \leq 1.0$

increasing and decreasing) in the topological entropy when the capital depreciation rate and the saving rate for holders are increased. These numerical results are perfectly connected with the intuitive idea generated by the bifurcation diagrams. Moreover, we can observe situations of isentropic dynamics (that is, dynamics with the same entropy) that can raise interesting questions.

In order to illustrate the idea of isentropic dynamics, we are going to search for some capital accumulation maps with the same topological entropy, that is, $\log(1.927561...) = 0.656256...$, when the parameters δ and s_r are varied. In our study we consider the bimodal map G restricted to its invariant region $\Omega \in \mathbb{R}^2$ (see Fig. 32.6), given by

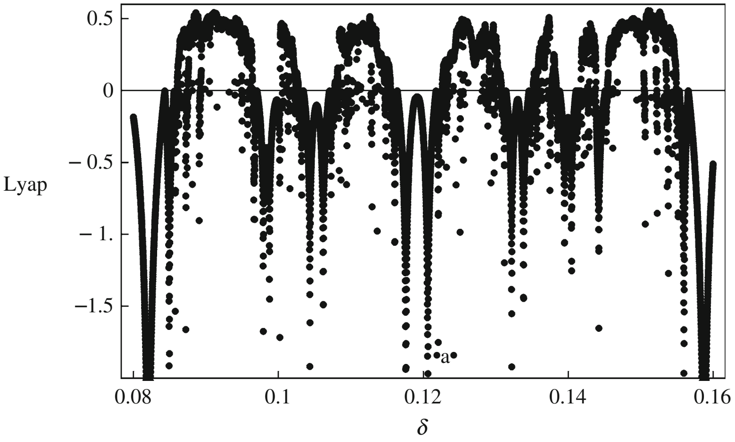


Fig. 32.4 Lyapunov exponents of the map G as a function of δ, with $s_r = 0.8$ and $0.08 \leq \delta \leq 0.16$

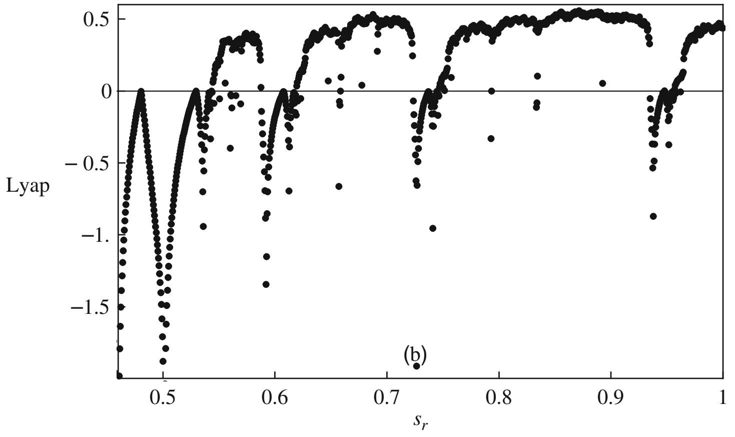


Fig. 32.5 Lyapunov exponents of the map G as a function of s_r, with $\delta = 0.09$ and $0.45 \leq s_r \leq 1.0$

$$\Omega = \{(\delta, s_r) \in \mathbb{R}^2 : G(G(c_1)) < G(c_2) \text{ and } G(G(c_2)) > G(c_1) \\ \text{and } G(c_2) > c_2 \text{ and } G(c_1) < c_1\}. \tag{32.16}$$

With the procedure presented above, we can compute the topological entropy for these maps. The following tables show the kneading data and the characteristic polynomial associated with each map. It is important to notice that the common factor $(-1 - t - t^2 - t^3 + t^4)$ is fundamental since determines the same spectral radius 1.92756.... and, therefore, the same topological entropy $h_{top}(G) = \log(1.92756...) = 0.656256...$for each considered map.

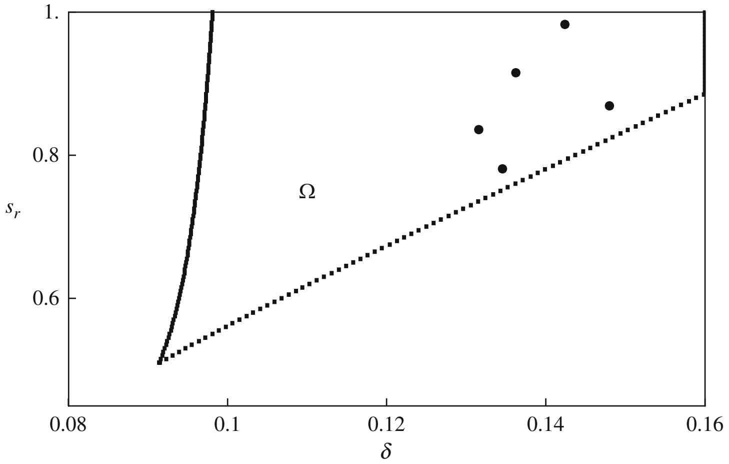


Fig. 32.6 Representation of points (δ, s_r) in the Ω region. To each point corresponds a map with topological entropy $h_{top}(G) = 0.656256...$

(δ, s_r)	Kneading data of G
$(0.148, 0.87)$	$(RLLMA, LLA)$
$(0.13156, 0.83565)$	(RLB, LA)
$(0.13459, 0.78086)$	(RL^2MLM^2LMA, L^2M^2LA)
$(0.136238, 0.915037)$	$(RLRLRLM^3A, LMRLM^5B)$
$(0.14242417, 0.982589)$	$(RLRLRLMLB, LMRLM^4LMA)$

(32.17)

(δ, s_r)	Characteristic polynomial of $M(G)$
$(0.148, 0.87)$	$-t^3(-1 - t - t^2 - t^3 + t^4)$
$(0.13156, 0.83565)$	$-1 - t - t^2 - t^3 + t^4$
$(0.13459, 0.78086)$	$-t^6(-1 + 2t + t^5)(-1 - t - t^2 - t^3 + t^4)$
$(0.136238, 0.915037)$	$(1 - t)(1 - t + t^2 - t^3 + t^4)(1 + t + t^2 + t^3 + t^4)$ $(-1 + t + t^5 + t^6)(-1 - t - t^2 - t^3 + t^4)$
$(0.14242417, 0.982589)$	$(1 - t)(-1 + t + t^5 + t^6)(-1 + t^2 + t^4 + t^6 + t^8)$ $(-1 - t - t^2 - t^3 + t^4)$

(32.18)

At this point of our study, we emphasize that in all the considered cases we have chaotic behavior and the topological entropy has exactly the same positive value. One question appears naturally: how can we distinguish these isentropic maps? In the following lines we contribute with an answer to this question.

First at all, it is interesting to exhibit a numerical result about the isentropic maps studied here. We can observe from Fig. 32.6 that to each pair of points (δ, s_r) (represented by black spots in Fig. 32.6) it is corresponding a map with the same topological entropy, namely $h_{top}(G) = 0.656256\ldots$. This means that the topological entropy by itself is no longer sufficient to classify these maps and will be necessary to consider another topological invariant in order to distinguish them.

The study of topological classification for bimodal maps G leads to the introduction of two topological invariants: one of them is the well known growth number $s(G) = e^{h_{top}(G)}$ and the other numerical quantity, denoted by r, is associated with the relative positions of the turning points of the map. The topological invariant r is introduced using the hypothesis $s(G) > 1$ and the Milnor–Thurston results about the topologically semi-conjugate by λ of G to a piecewise linear map $F_{e,s}$ having slope $\pm s(G)$ everywhere (see [1,16,17]). The map $F_{e,s}$ is unique and it is defined by

$$F_{e,s} : [0,1] \longrightarrow [0,1] \qquad \text{so that} \qquad F_{e,s}\left(\lambda\left(x\right)\right) = \lambda\left(G\left(x\right)\right) \tag{32.19}$$

for every $x \in [0,1]$ such that

$$F_{e,s}\left(y\right) = \begin{cases} s\ y & \text{if } \ 0 \leq y < \lambda\left(c_1\right) \\ -s\ y + e & \text{if } \ \lambda\left(c_1\right) \leq y < \lambda\left(c_2\right) \\ s\ y + 1 - s & \text{if } \ y \geq \lambda\left(c_2\right) \end{cases} \tag{32.20}$$

where $\lambda\left(c_1\right) = e/(2s)$, $\lambda\left(c_2\right) = e/(2s) + (s-1)/(2s)$ (see Fig. 32.7).

Now, the new invariant, $r(G)$, is given by

$$r(G) = \frac{4s\lambda(c_1) - 1 - s}{2} = \frac{4s\lambda(c_2) + 1 - 3s}{2} \tag{32.21}$$

$$\text{with} \quad \lambda\left(c_1\right) = \sum_{i=1}^{n_L+1} v_i \quad \text{and} \quad \lambda\left(c_2\right) = \sum_{i=1}^{n_L+n_M+2} v_i \tag{32.22}$$

where n_L (respective n_M) denote the number of symbols L (respective symbols M). The vector v is the Perron eigenvector associated with the eigenvalue $\lambda_{\max} = s$, $\mathscr{M}v = \lambda_{\max}v$, where $\mathscr{M}$ is the transition matrix with the extreme intervals $I_0 = [0, z_1]$ and $I_{p+q} = [z_{p+q}, 1]$ included. It is important to note that $r(G)$ is in fact a topological invariant because all the variables $\lambda(c_1)$, $\lambda(c_2)$ and $s(G)$ that lead to $r(G)$ are topological invariants (see [1]). In the piecewise linear case, $F_{e,s}$, the parameter $r(G)$ is the invariant that distinguish isentropic dynamics and $r \in \left[\frac{s-3}{2}, \frac{3-s}{2}\right]$. Regarding the previous considerations, we derive the following result,

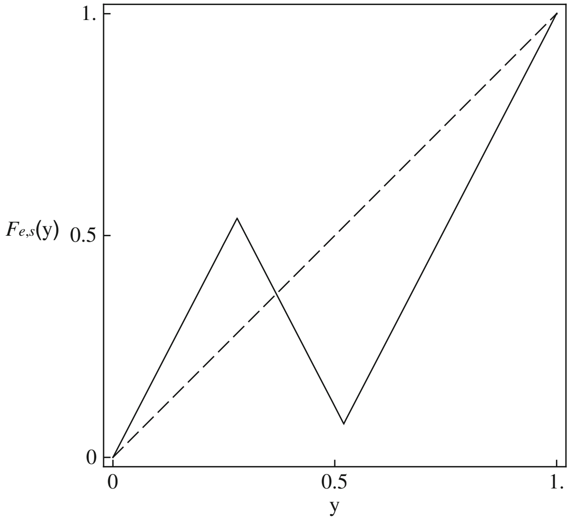


Fig. 32.7 Piecewise linear map for $s = 1.9275619...$ and $r = -0.385686...$

Proposition 32.1. *The maps G can be topologically classified by the pair of topological invariants (s, r), where s is the lap growth number, $s(G) = e^{h_{top}(G)}$, and r is the invariant given by*

$$r(G) = \frac{4s\lambda(c_1) - 1 - s}{2} = \frac{4s\lambda(c_2) + 1 - 3s}{2}, \tag{32.23}$$

and where λ is the map defined by the semi-conjugacy to the piecewise linear map $F_{e,s}$.

We discuss the following example which illustrates very well the situation presented above.

Example 32.2. For the kneading data $(RLLMA, LLA)^\infty$ (as in Example 32.1) we can apply the previous algorithm to compute the topological invariants associated with this sequence. The transition matrix is given by

$$\mathscr{M}(G) = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

and satisfies the equation of Perron eigenvector, $\mathscr{M}v = \lambda_{\max}v$. Then we have

$$\lambda(c_1) = \sum_{i=1}^{5} v_i = 0.279653... \text{ and } \lambda(c_2) = \sum_{i=1}^{7} v_i = 0.520257... \qquad (32.24)$$

(with v normalized to the unit interval). We obtain $s = 1.92756...$ and $r = -0.385686....$ The semi-conjugate piecewise linear map associated with this kneading data is given in the Fig. 32.7.

To each kneading data $(P^{(p)}, Q^{(q)})$ corresponds one and only one value of r. For the set of points studied, we present in Figs. 32.8 and 32.9 some numerical results of the variation of the topological invariant r with each of the parameters δ and s_r. We can see that for the same topological entropy, if we increase the capital depreciation rate the invariant r is oscillating in a different and not correlated way than when we vary the saving rate for holders. It is quite interesting, since the several capital accumulation bimodal maps, considered for different parameter settings, and showing the same topological entropy can still be differentiate by another topological invariant.

32.4 Control of the Capital Accumulation

There are several methods available to control chaos in one-dimensional dynamical systems as can be seen in [4, 5] and [19] . In this section, we restrict our analysis and show that periodic proportional pulses, applied to the chaotic dynamics of the capital accumulation model given by the map (32.3), that is,

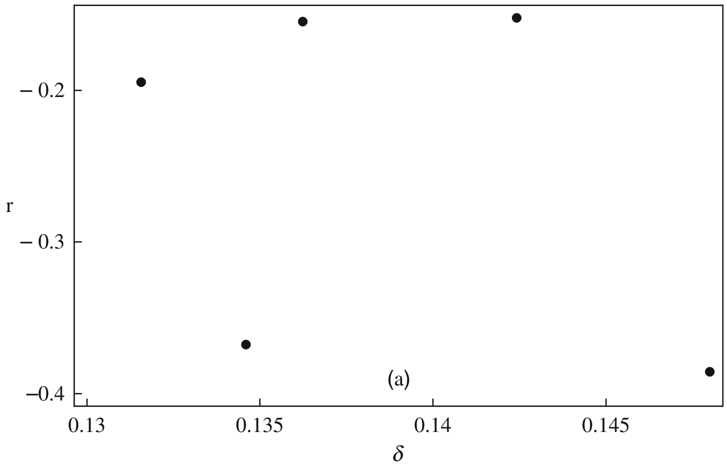


Fig. 32.8 Variation of the topological invariant r with δ

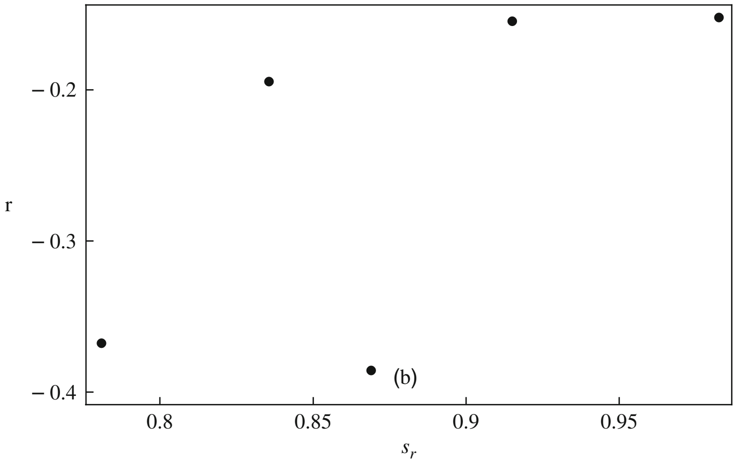


Fig. 32.9 Variation of the topological invariant r with s_r

$$k_{t+1} = G(k_t) = \frac{1}{n+1}\left((1-\delta)\,k_t + s_w w\,(k_t) + s_r k_t\, f'\,(k_t)\right) \tag{32.25}$$

can stabilize the dynamics at a desired periodic orbit. For more details on the method and its application to the Hénon map the reader is directed to [4].

In order to control the chaos in this discrete dynamical system, instantaneous pulses will be applied to the map variable, k_t, at every p iterations such that

$$k_i \longrightarrow q\;k_i \;\;(i \text{ is a multiple of } p) \tag{32.26}$$

where q is a constant to be determined and p denotes the period of the orbit in the dynamics. A fixed point of period one, k_s, of $k_{t+1} = G(k_t)$ is such that $k_s = G(k_s)$, and it is called stable if and only if the modulus of the first order derivative is lower than 1, that is,

$$\left|\frac{dG(k_s)}{dk}\right| < 1. \tag{32.27}$$

Now, we kick the dynamics by multiplying its values with a factor q, at every p iterations, by considering

$$G^*(k) = qG^p(k), \tag{32.28}$$

where G^p is the composition of the map G with itself p times. A fixed point of G^* satisfies the equation

$$qG^p(k_s) = k_s, \tag{32.29}$$

where the fixed point k_s is locally stable if

$$\left|q\frac{dG^p(k_s)}{dk}\right| < 1. \tag{32.30}$$

A stable fixed point of G^* is viewed as a stable periodic point of period p of the original dynamics, *kicked by the control method*. Considering parameter values such that the map (32.3) is chaotic and willing to control it into a stable periodic orbit of period p, we must find a point k_s and a factor q satisfying (32.29) and (32.30).

Defining the function $C^p(k)$ by

$$C^p(k) = q\frac{dG^p(k_s)}{dk}, \tag{32.31}$$

and taking q from (32.29), that is,

$$q = \frac{k_s}{G^p(k_s)}, \tag{32.32}$$

we obtain

$$C^p(k) = \frac{k_s}{G^p(k_s)}\frac{dG^p(k_s)}{dk} \tag{32.33}$$

and (32.30) becomes

$$|C^p(k_s)| < 1 \Longleftrightarrow \left|\frac{k_s}{G^p(k_s)}\frac{dG^p(k_s)}{dk}\right| < 1. \tag{32.34}$$

We emphasize the importance of the previous inequality: if the fixed point k_s satisfies (32.34), then with the kicking factor q defined by (32.32), the control is switched on and will stabilize the dynamics at a periodic orbit of period p, passing through the given point. For particular details and remarks about this control procedure see [4].

In the context of economy, it has been accepted that chaos control methods constitute interesting applications when they can lead to stable periodic orbits of low periods, namely, $p = 1$ and $p = 2$, which represent short-time predictable behavior. For illustrative purposes, we fix the parameter values $\delta = 0.148$ and $s_r = 0.87$, where the system exhibits positive topological entropy (see Example 32.1). The functions $C^1(k)$ and $C^2(k)$, when their values are between -1 and 1, are shown in Figs. 32.10 and 32.11.

Concerning the function $C^1(k)$, fixed points of period 1 can be stabilized for every k_s in two ranges. When $p = 2$, the orbit of period 2 can be stabilized in six ranges. In fact, the control ranges become smaller as the periodicity increases. Figures 32.12 and 32.13 show two examples of stabilizing the economic map at period 1 and at period 2. The values k_s have been selected by examining Figs. 32.10 and 32.11 and the values of q were calculated using (32.32).

In both examples, the convergence was very fast (see Figs. 32.12 and 32.13). It is interesting to observe that for values of k_s such that $|C^p(k_s)|$ is near the unity the convergence is slower. As far as an economic system is concerned, it is convenient to obtain a fast convergence in order to reach the desired behavior. As an example of a slower convergence see Figs. 32.14 and 32.15.

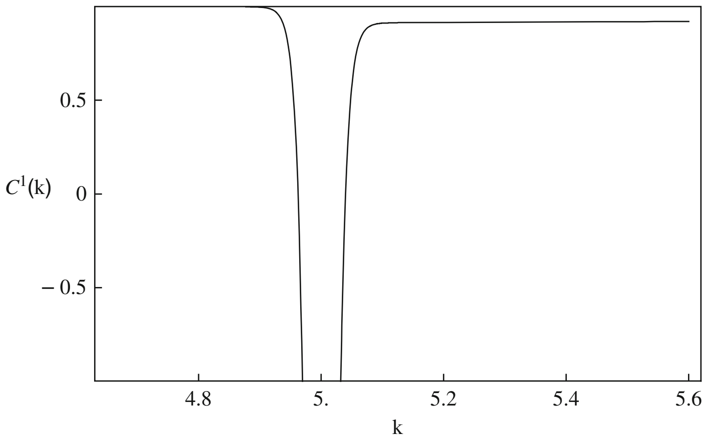


Fig. 32.10 The control curves C^p, $p = 1$, for the economic map when $\delta = 0.148$ and $s_r = 0.87$. In each case, the range is restricted to $-1 < C^p(k) < 1$

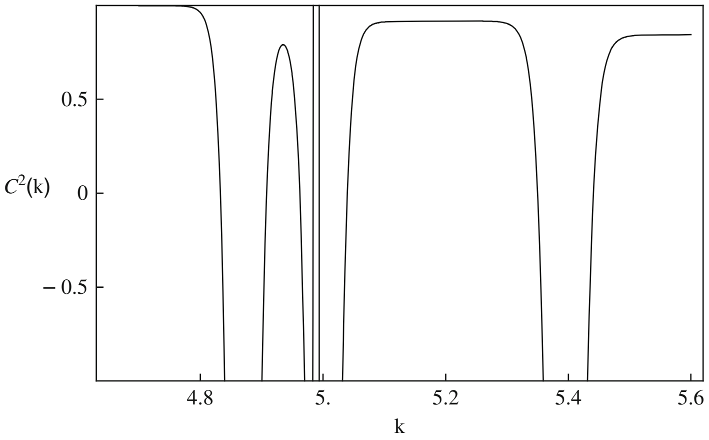


Fig. 32.11 The control curves C^p, $p = 2$, for the economic map when $\delta = 0.148$ and $s_r = 0.87$. In each case, the range is restricted to $-1 < C^p(k) < 1$

Note that the system can be stabilized to many different points on or even out of the basin of attraction of the attractor (see the work of Chau [4]).

32.5 Concluding Remarks

In this article we have analyzed in detail some aspects of the dynamics of a one-sector growth model with differential savings and with a Leontief production function as introduced by Böhm and Kaas. The rich and complex behavior of this

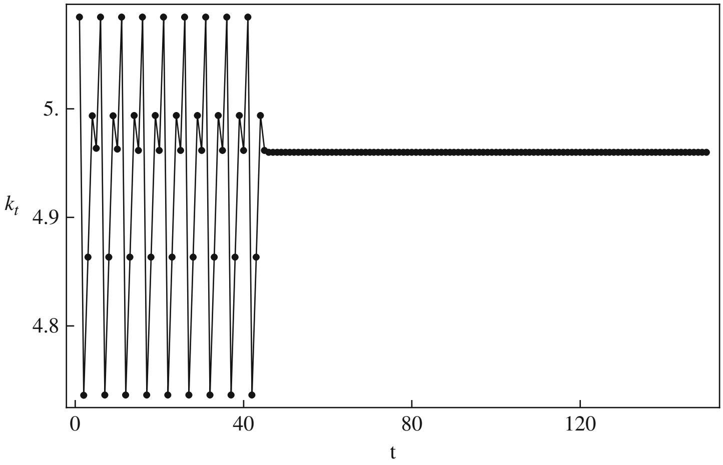


Fig. 32.12 An example of the effect of controlling the economic map to periodic orbits (of periods 1 and 2). For period 1: $k_s = 4.96$ and $q = 0.975502...$ (the control was switched on at $t = 45$)

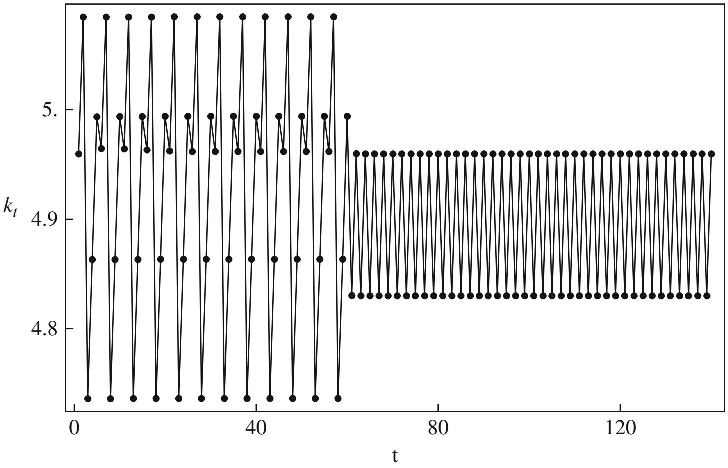


Fig. 32.13 An example of the effect of controlling the economic map to periodic orbits (of periods 1 and 2). For period 2: $k_s = 4.83$ and $q = 0.949952...$ (the control was switched on at $t = 60$)

model allowed us to apply different theoretical and numerical approaches. More specifically, we analyzed the model in terms of symbolic dynamics theory and in terms of applicability of chaos control theory.

In the theory of business cycles the use of powerful tools for the study of dynamic models, such as the symbolic dynamics stands out to be very effective for the explicit computation of important numerical invariants that characterize the chaotic behavior. In fact, each one of the chaotic windows identified by the Lyapunov exponents, that occurs with the variation of the capital depreciation rate, δ, and the savings

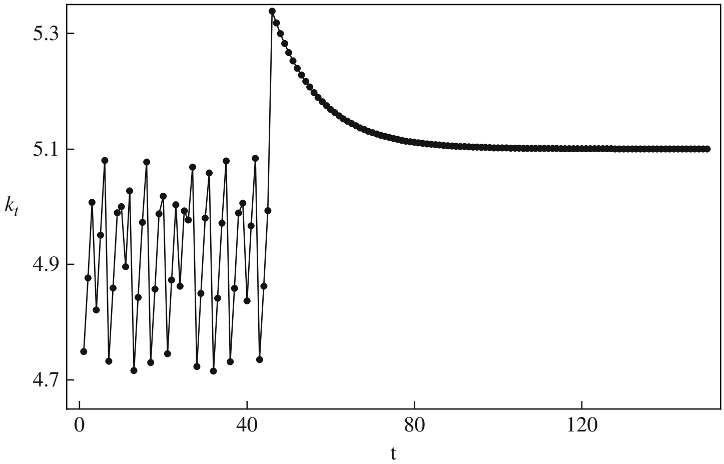


Fig. 32.14 An example of a slow convergence to the desired periods. For period 1: $k_s = 5.1$ and $q = 1.07386...$ (the control was switched on at $t = 45$)

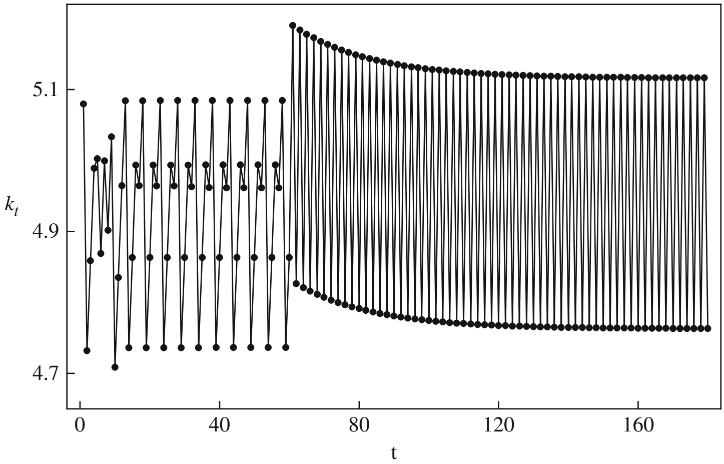


Fig. 32.15 An example of a slow convergence to the desired periods. For period 2: $k_s = 4.95$ and $q = 1.04608...$ (the control was switched on at $t = 60$)

rate for holders, s_r, is associated with positive values of the topological entropy. This important numerical invariant is related to the exponential orbit growth and its analysis revealed situations of isentropic dynamics. The complete topological classification of the bimodal maps became possible with the introduction of the numerical invariant r. In the context of economic models what is the meaning of this measure of complexity? This is an open and challenging question for which there is no answer yet.

Motivated by the chaotic structure of the map and the central role of regular cycles in economy, we have applied the periodic proportional pulses control method in order to obtain predictable behavior – the stabilized period-one orbit and the stabilized period-two orbit. Indeed, if there is some form of relevant nonlinearity in an economic structure, then the control of such structures may benefit a lot from the understanding of what chaos control is all about. We showed, that the complicated motion which emerges from the dynamics of the capital accumulation model can be controlled by applying instantaneous pulses to the system's dynamical variable k, at every periodic iteration. The analytical representation of the control functions $C^p(k)$, allowed us to exhibit the control ranges of the capital accumulation in each case of periodicity. We emphasize that, with the application of the chaos control technique, the model performs different times of efficiency in the convergence process. The chaotic dynamics could be converted, by using just periodic proportional pulses, to motion on the desired period orbits.

References

1. Almeida, P., Lampreia, J.P., Sousa Ramos, J.: Topological invariants for bimodal maps. In: Iteration Theory, pp. 1–8. World Science Publishing, Singapore (1992)
2. Böhm, V., Kaas, L.: Differential savings, factor shares, and endogenous growth cycles. J. Econ. Dyn. Control **24**, 965–980 (2000)
3. Brianzoni, S., Mammana, C., Michetti, E.: Complex dynamics in the neoclassical growth model with differential savings and non-constant labor force growth. Stud. Nonlin. Dyn. Econom. **11**(3), (2007)
4. Chau, N.P.: Controlling chaos by periodic proportional pulses. Phys. Lett. A **234**, 193–197 (1997)
5. Chau, N.P.: Stabilizing effect of periodic or eventually periodic constant pulses on chaotic dynamics. Phys. Rev. E. **57**(6), 7317–7320 (1998)
6. Chen, L., Chen, G.: Controlling chaos in an economic model. Phys. A **374**, 349–358 (2007)
7. Mendes, D.A., Mendes, V.: Control of chaotic dynamics in an OLG economic model. J. Phys. Conf. Ser. **23**, 158–181 (2005)
8. Day, R.H., Zhang, M.: Classical economic growth theory: a global bifurcation analysis. Chaos Solitons Fractals **7**(12), 1969–1988 (1996)
9. Naschie, E.L.: Superstrings, entropy and the elementary particles content of the standard model. Chaos Solitons Fractals **29**, 48–54 (2006)
10. Fanti, L., Manfredi, P.: Chaotic business cycles and fiscal policy: an IS-LM model with distributed tax collection lags. Chaos Solitons Fractals **32**(2), 736–744 (2007)
11. Güémez, J., Matias, M.A.: Control of chaos in unidimensional maps. Phys. Lett. A **181**, 29–32 (1993)
12. Jaynes, E.T.: How should we use entropy in economics? http://citeseer.ist.psu.edu/ 82786.html. (1991)
13. Katok, A., Hasselblat, B.: An Introduction to the Modern Theory of Dynamical Systems. University of Cambridge, Cambridge (1999)
14. Lorenz, H.-W.: Nonlinear dynamical economics and chaotic motion. Springer, Berlin (1993)
15. Lampreia, J.P. , Sousa Ramos, J.: Symbolic dynamics of bimodal maps. Portugal. Math. **54**(1), 1–18 (1997)
16. Milnor, J., Thurston, W.: On iterated maps of the interval I and II. In: Lecture Notes in Mathematics, **1342**, 465–563. Springer, New York (1988)

17. Martins, N., Severino, R., Sousa Ramos, J.: Isentropic real cubic maps. Int. J. Bifur. Chaos Appl. Sci. Eng. **13**(7), 1701–1709 (2003)
18. Montrucchio, L., Sorger, G.: Topological entropy of policy functions in concave dynamic optimization models. J. Math. Econ. **25**(2), 181–194 (1996)
19. Ott, E., Grebogi, C., Yorke, J.A.: Controlling chaos. Phys. Rev. Lett. **64**, 1196–1199 (1990)
20. Paula, A.S., Savi, M.A.: A multiparameter chaos control method based on OGY approach. Chaos Solitons Fractals (2007). doi:10.1016/j.chaos.2007.09.056
21. Salarieh, H., Alasty, A.: Chaos control in an economic model via minimum entropy strategy. Chaos Solitons Fractals (2007). doi:10.1016/j.chaos.2007.08.045.

Chapter 33
Hydrodynamic Limit of the Exclusion Process in Inhomogeneous Media

Milton Jara

Abstract We obtain the hydrodynamic limit of a simple exclusion process in an inhomogeneous environment of divergence form. Our main assumption is a suitable version of Γ-convergence for the environment. In this way we obtain an unified approach to recent works on the field.

33.1 Introduction

Since the seminal paper [9], the theory of hydrodynamic limit of interacting particle systems has evolved into a powerful tool in the study of non-equilibrium properties of statistical systems of many components (see the book [13] for a comprehensive exposition). Recently, and due to the influence of physical and mathematical works about random walks in random environment, an increasing attention has been posed into particle systems evolving in random environments. Despite the early works [6, 14, 17], we mention [1–5, 7, 8, 10, 15, 17]. In [7, 10] the *corrected empirical density* was introduced, which is nothing but a microscopic version of the compensated compactness lemma of Tartar [18]. Roughly speaking, when the inhomogeneous environment (random or not) has a divergence form and has a Γ-limit, space homogenization of the environment and time homogenization of the interaction decouples, and the standard tools from the theory of hydrodynamic limit can be used to obtain the asymptotic behavior of the density of particles in a family of models, including the exclusion process and the zero-range process.

In this review, we give an unified approach to this problem, recovering previous results in [1–3, 10, 15]. In order to concentrate our efforts in the influence of the inhomogeneous environment on the asymptotics of the density of particles, we consider the simplest model of interacting particle systems, which is the symmetric

M. Jara
CEREMADE, Université Paris-Dauphine, Place du Maréchal de Lattre de Tassigny, Paris Cedex 75775, France
e-mail: jara@ceremade.dauphine.fr

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_33,

exclusion process η_t^n in an unoriented graph. In this process, particles perform symmetric random walks on a graph $\{X_n\}_n$ with some rates $\omega^n = \{\omega_{x,y}^n; x, y \in X_n\}$, conditioned to have at most one particle per site. We think of $\{X_n\}_n$ as a sequence of graphs embedding in some metric space X, and we are interested in the evolution of the measure $\pi_t^n(dx)$ in X, obtained by giving a mass a_n^{-1} to each particle.

This article is organized as follows. In Sect. 33.2 we give precise definitions of the exclusion process, the inhomogeneous environment and we state our main result. We also define what we mean by an approximation $\{X_n\}_n$ of X and by Γ-convegence of the environment. In Sect. 33.3 we introduce the corrected empirical density and we prove our main theorem. In Sect. 33.4 we introduce the concept of energy solutions of the hydrodynamic equation, we prove uniqueness of such solutions and we obtain a substantial improvement of the main theorem. The material of this section is new and it gives a better understanding of the relation between Γ-convergence of the environment and hydrodynamic limit of the particle system. In Sect. 33.5 we discuss how to reobtain previous results in the literature relying in our main theorem.

33.2 Definitions and Results

In this section we define the exclusion process in inhomogeneous environment and we recall some notions of Γ-convergence that will be necessary in order to obtain the hydrodynamic limit of this process.

33.2.1 Partitions of the Unity and Approximating Sequences

In this section we fix some notation and we define some objects which will be useful in the sequel. Let $(X, \mathscr{B})$ be a Polish space. We assume that X is σ-compact. We say that a sequence of functions $\{\mathscr{U}_i; i \in I\}$ is a *partition of the unity* if:

(1) For any $i \in I$, $\mathscr{U}_i : X \to [0,1]$ is a continuous function.
(2) For any $x \in X$, $\sum_{i \in I} \mathscr{U}_i(x) = 1$.
(3) For any $x \in X$, the set $\{i \in I; \mathscr{U}_i(x) > 0\}$ is finite.

We say that the partition of the unity $\{\mathscr{U}_i; i \in I\}$ is *regular* if supp $\mathscr{U}_i$ is compact for any $i \in I$, and additionally $\mathscr{U}_i(X) = [0,1]$. We denote by $\mathscr{M}_+(X)$ the set of Radon, positive measures in X. The symbol $\{x_n\}_n$ will denote a sequence of elements x_n in some space, indexed by the set $\mathbb{N}$ of positive integers.

Let $\{\mathscr{U}_i\}_i$ be a regular partition of the unity. We say that a sequence $\{x_i; i \in I\}$ in X is a *representative* of $\{\mathscr{U}_i\}_i$ if $\mathscr{U}_i(x_i) = 1$ for any $i \in I$. Notice that we have $x_i \neq x_j$ for $i \neq j$.

Let $\{\mathscr{U}_i^n; i \in I_n\}_n$ be a sequence of regular partitions of the unity. We say that a measure $\mu \in \mathscr{M}_+(X)$ is the scaling limit of the sequence $\{\mathscr{U}_i^n\}_n$ if there exists a sequence $\{a_n\}_n$ of positive numbers such that for any sequence $\{x_i^n; i \in I_n\}$ of representatives of $\{\mathscr{U}_i^n\}_n$ we have

$$\lim_{n\to\infty} \frac{1}{a_n} \sum_{i\in I_n} \delta_{x_i^n} = \mu$$

with respect to the vague topology, where δ_x is the Dirac mass at $x \in X$. We call $\{a_n\}_n$ the *scaling* sequence.

From now on, we fix a sequence $\{\mathscr{U}_i^n\}_n$ of regular partitions of the unity with scaling limit μ, scaling sequence $\{a_n\}_n$ and we assume that $\mu(A) > 0$ for any non-empty, open set $A \subseteq X$. Fix a sequence $\{x_i^n; i \in I_n\}$ of representatives of $\{\mathscr{U}_i^n\}_n$. Define $X_n = \{x_i^n; i \in I_n\}$. Since $\{\mathscr{U}_i^n\}$ is a partition of the unity, the induced topology in X_n coincides with the discrete topology. For $x = x_i^n$, we will denote $\mathscr{U}_x^n = \mathscr{U}_i^n$. Define

$$\mu_n(dx) = \frac{1}{a_n} \sum_{x\in X_n} \delta_x(dx).$$

By definition, $\mu_n \to \mu$ in the vague topology. We denote by $\mathscr{L}^2(\mu_n)$ the Hilbert space of functions $f : X_n \to \mathbb{R}$ such that $\sum_{x\in X_n} f(x)^2 < +\infty$, equipped with the inner product

$$\langle f, g\rangle_n = \frac{1}{a_n} \sum_{x\in X_n} f(x)g(x).$$

We define $\mathscr{L}^2(\mu)$, $\mathscr{L}^1(X_n)$ and $\mathscr{L}^1(\mu)$ in the analogous way and we denote $\langle f, g\rangle = \int fgd\mu$. We denote by $\mathscr{C}_c(X)$ the set of continuous functions $f : X \to \mathbb{R}$ with compact support. In the same spirit, we denote by $\mathscr{C}_c(X_n)$ the set of functions $f : X_n \to \mathbb{R}$ with finite support. We define the projection $S_n : \mathscr{C}_c(X) \to \mathscr{C}_c(X_n)$ by taking

$$\big(S_n G\big)(x) = a_n \int G\mathscr{U}_x^n d\mu.$$

This operator, under suitable conditions, can be extended to a bounded operator from $\mathscr{L}^2(X)$ to $\mathscr{L}^2(X_n)$. Notice that $\int S_n G d\mu_n = \int G d\mu$. Therefore S_n is continuous from $\mathscr{L}^1(\mu)$ to $\mathscr{L}^1(X_n)$.

33.2.2 *Γ*-Convergence

Define $\bar{\mathbb{R}} = [-\infty, +\infty]$. Let $(Y, \mathscr{F})$ be a topological space, and let $F_n, F : Y \to \bar{\mathbb{R}}$. We say that F_n is Γ-convergent to F if:

(1) For any sequence $\{y_n\}_n$ in Y converging to $y \in Y$,

$$F(y) \le \liminf_{n\to\infty} F_n(y_n).$$

(2) For any $y \in Y$ there exists a sequence $\{y_n\}_n$ converging to y such that

$$\limsup_{n\to\infty} F_n(y_n) \le F(y).$$

An important property of Γ-convergence is that it implies *convergence of minimizers* in the following sense:

Proposition 33.1. *Let $F_n, F : Y \to \bar{\mathbb{R}}$ be such that F_n is Γ-convergent to F. Assume that there exists a relatively compact set $K \subseteq Y$ such that for any n,*

$$\inf_{y\in Y} F_n(y) = \inf_{y\in K} F_n(y).$$

Then,

$$\lim_{n\to\infty} \inf_{y\in K} F_n(y) = \min_{y\in Y} F(y).$$

Moreover, if $\{y_n\}_n$ is a sequence in K such that $\lim_n (F_n(y_n) - \inf_K F_n) = 0$, *then any limit point y of $\{y_n\}_n$ satisfies $F(y) = \min_Y F$.*

A useful property that follows easily from the definition is the stability of Γ-convergence under continuous perturbations:

Proposition 33.2. *Let $F_n, F : Y \to \bar{\mathbb{R}}$ be such that F_n is Γ-convergent to F. Let $G_n : Y \to \mathbb{R}$ be such that G_n converges uniformly to a continuous limit G. Then, $F_n + G_n$ is Γ-convergent to $F + G$.*

33.2.3 The Exclusion Process in Inhomogeneous Environment

In this section we define the exclusion process in inhomogeneous environment as a system of particles evolving in the set X_n. Let $\omega^n = \{\omega^n_{x,y}; x, y \in X_n\}$ be a sequence of non-negative numbers such that $\omega^n_{x,x} = 0$ and $\omega^n_{x,y} = \omega^n_{y,x}$ for any $x, y \in X_n$. We call ω^n the *environment*. We define the exclusion process η^n_t with environment ω^n as a continuous-time Markov chain of state space $\Omega_n = \{0,1\}^{X_n}$ and generated by the operator

$$L_n f(\eta) = \sum_{x,y\in X_n} \omega^n_{x,y}\big[f(\eta^{x,y}) - f(\eta)\big],$$

where η is a generic element of Ω_n, $f : \Omega_n \to \mathbb{R}$ is a function which depends on $\eta(x)$ for a finite number of elements $x \in X_n$ (that is, f is a *local function*) and $\eta^{x,y} \in \Omega_n$ is defined by

$$\eta^{x,y}(z) = \begin{cases} \eta(y), & \text{if } z = x \\ \eta(x), & \text{if } z = y \\ \eta(z), & \text{if } z \neq x, y. \end{cases}$$

In order to have a well-defined Markovian evolution for any initial distribution η_0^n, we assume that $\sup_x \sum_{y \in X_n} \omega_{x,y}^n < +\infty$. We interpret X_n as a set of sites and $\eta_t^n(x)$ as the number of particles at site $x \in X_n$ at time t. Since $\eta_t^n(x) \in \{0,1\}$, there is at most one particle per site at any given time: this is the so-called *exclusion rule*. Notice that the dynamics is conservative in the sense that no particles are annihilated or destroyed.

Our interest is to study the collective behavior of particles for the sequence of processes $\{\eta_\cdot^n\}_n$. In order to do this, we introduce the *empirical density of particles* as the measure-valued process π_t^n defined by

$$\pi_t^n(G) = \frac{1}{a_n} \sum_{x \in X_n} \eta_t^n(x) S_n G(x)$$

for any $G \in \mathscr{C}_c(X)$. Using Riesz's theorem, it is not difficult to check that π_t^n is effectively a positive Radon measure in X. Observe that when $\eta_0^n(x) = 1$ for any $x \in X_n$, then $\eta_t^n(x) = 1$ for any $x \in X_n$ and any $t \geq 0$. In this situation, the empirical process π_t^n is identically equal to the measure μ. Notice that the random variable π_t^n defined in this way corresponds to a process defined in the space $\mathscr{D}([0,\infty), \mathscr{M}_+(X))$ of càdlàg paths with values in $\mathscr{M}_+(X)$. For functions $G : X_n \to \mathbb{R}$, we define $\pi_t^n(G) = a_n^{-1} \sum_x \eta_t^n(x) G(x)$.

33.2.4 *Γ*-Convergence of the Environment

In this section we will make a set of assumptions on the environment $\{\omega^n\}_n$ which will allows us to obtain an asymptotic result for the sequence $\{\pi_\cdot^n\}_n$. We start with two assumptions about the sequence of partitions of the unity $\{\mathscr{U}_x^n\}_n$. Our first assumption corresponds to a sort of ellipticity condition on the partitions of the unity $\{\mathscr{U}_x^n\}_n$:

(H1) There exists $\Theta < +\infty$ such that

$$\sup_{x \in X_n} a_n \int \mathscr{U}_x^n d\mu \leq \Theta \text{ for any } n > 0.$$

Under this condition, the projection S_n satisfies $||S_n G||_\infty \leq \theta ||G||_\infty$, and by interpolation S_n can be extended to a continuous operator from $\mathscr{L}^2(\mu)$ to $\mathscr{L}^2(X_n)$. Our second condition states that S_n is close to an isometry when $n \to \infty$:

(H2) For any $F \in \mathscr{L}^2(\mu)$, we have

$$\lim_{n \to \infty} \langle S_n F, S_n F \rangle_n = \langle F, F \rangle.$$

Now we are ready to discuss on which sense we will say that the environment ω^n converges. For a given function $F : X_n \to \mathbb{R}$ of finite support, we define $\mathscr{L}_n F$ by

$$\mathscr{L}_n F(x) = \sum_{y \in X_n} \omega^n_{x,y}\big(F(y) - F(x)\big).$$

It turns out that $\mathscr{L}_n$ can be extended to a non-positive operator in $\mathscr{L}^2(X_n)$. In fact, for any function F of finite support, the *Dirichlet form*

$$\langle F, -\mathscr{L}_n F\rangle_n = \frac{1}{2a_n} \sum_{x,y \in X_n} \omega^n_{x,y}\big(F(y) - F(x)\big)^2$$

is clearly non-negative. For a function $G \in \mathscr{L}^2(\mu)$, define $\mathscr{E}_n(G) = \langle S_n G, -\mathscr{L}_n S_n G\rangle$. Notice that $\mathscr{E}_n : \mathscr{L}^2(\mu) \to \bar{\mathbb{R}}$ is a quadratic form. Now we are ready to state our first hypothesis about the environment:

(H3) There exists a non-negative, symmetric operator $\mathscr{L} : D(\mathscr{L}) \subseteq \mathscr{L}^2(\mu) \to \mathscr{L}^2(\mu)$ such that $\mathscr{E}_n$ is Γ-convergent to $\mathscr{E}$, where $\mathscr{E}(G) = -\int G\mathscr{L}G d\mu$.

Our second hypothesis about the environment ω^n concerns to its Γ-limit $\mathscr{L}$:

(H4) There exists a dense set $\mathscr{K} \subseteq \mathscr{C}_c(X)$ such that $\mathscr{K}$ is a kernel for the operator $\mathscr{L}$, and for any $G \in \mathscr{K}$, $\mathscr{L}G$ is continuous and $\int |\mathscr{L}G| d\mu < +\infty$.

33.2.5 Hydrodynamic Limit of η^n_t

In this section we explain what we understand as the hydrodynamic limit of η^n_t. We say that a sequence $\{\nu_n\}_n$ of distributions in Ω_n is *associated* to a function $u : X \to \mathbb{R}$ if for any function $G \in \mathscr{C}_c(X)$ and any $\epsilon > 0$ we have

$$\lim_{n\to\infty} \nu_n\bigg\{\bigg|\frac{1}{a_n}\sum_{x\in X_n} \eta(x)G(x) - \int G(x)u(x)\mu(dx)\bigg| > \epsilon\bigg\} = 0.$$

Notice that we necessarily have $0 \leq u(x) \leq 1$ for any $x \in X$, since $\eta(x) \in \{0,1\}$. Fix an initial profile $u_0 : X \to [0,1]$ and take a sequence of distributions $\{\nu_n\}$ associated to u_0. Let η^n_t be the exclusion process with initial distribution ν_n. We denote by $\mathbb{P}_n$ the law of η^n_t in $\mathscr{D}([0,\infty), \Omega_n)$ and by $\mathbb{E}_n$ the expectation with respect to $\mathbb{P}_n$. The fact that $\{\nu_n\}_n$ is associated to u_0 can be interpreted as a law of large numbers for the empirical measure π^n_0: $\pi^n_0(dx)$ converges in probability to the deterministic measure $u_0(x)\mu(dx)$. We say that the hydrodynamic limit of η^n_t is given by the equation $\partial_t u = \mathscr{L}u$ if for any $t > 0$, the empirical measure $\pi^n_t(dx)$ converges in probability to the measure $u(t,x)\mu(dx)$, where $u(t,x)$ is the solution of the equation $\partial_t u = \mathscr{L}u$ with initial condition u_0. Before stating our main result in a more precise way, we need some definitions.

For $F, G \in D(\mathscr{L})$, define the bilinear form $\mathscr{E}(F,G) = -\int F\mathscr{L}G d\mu$. Notice that $\mathscr{E}(F,G)$ is still well defined if only $G \in D(\mathscr{L})$. We say that a function $u : [0,T] \times X \to [0,1]$ is a weak solution of (33.1) with initial condition u_0 if

$\int_0^T \int u_t^2 d\mu dt < +\infty$ and for any differentiable path $G : [0, T] \to \mathscr{K}$ such that $G_T \equiv 0$ we have

$$\langle u_0, G_0 \rangle + \int_0^T \Big\{ \langle \partial_t G_t, u_t \rangle - \mathscr{E}(G_t, u_t) \Big\} dt = 0.$$

Theorem 33.1. *Let $\{\nu_n\}_n$ be associated to u_0 and consider the exclusion process η_t^n with initial distribution ν_n. Assume that $\int \pi_0^n(dx)$ is uniformly finite:*

(H5)

$$\lim_{M \to \infty} \sup_n \nu_n \Big\{ \frac{1}{a_n} \sum_{x \in X_n} \eta(x) > M \Big\} = 0.$$

Then, the sequence of processes $\{\pi_\cdot^n(dx)\}_n$ is tight and the limit points are concentrated on measures of the form $u(t, x)\mu(dx)$, where $u(t, x)$ is a weak solution of the hydrodynamic equation

$$\begin{cases} \partial_t u = \mathscr{L} u, \\ u(0, \cdot) = u_0(\cdot). \end{cases} \tag{33.1}$$

If such solution is unique, the process $\pi_\cdot^n(dx)$ converges in probability with respect to the Skorohod topology of $\mathscr{D}([0, \infty), \mathscr{M}_+(X))$ to the deterministic trajectory $u(t, x)\mu(dx)$.

Usually in the literature, hydrodynamic limits are obtained in finite volume, since the pass from finite to infinite volume is non-trivial. Assumption **(H5)** is in this spirit: it is automatically satisfied when the cardinality of X_n is of the order of a_n (on which case $\mu(X) < +\infty$), and it is very restrictive when X_n is infinite. For simplicity, we restrict ourselves to the case on which **(H5)** is satisfied.

33.3 Hydrodynamic Limit of η_t^n: Proofs

In this section we obtain the hydrodynamic limit of the process η_t^n. The strategy of proof of this result is the usual one for convergence of stochastic processes. First we prove tightness of the sequence of processes $\{\pi_\cdot^n\}_n$. Then we prove that any limit point of this sequence is concentrated on solutions of the hydrodynamic equation. Finally, a uniqueness result for such solutions allows us to conclude the proof. However, the strategy outlined above will not be carried out for $\{\pi_\cdot^n\}_n$ directly, but for another process $\hat{\pi}_\cdot^n$, which we call the *corrected* empirical process.

33.3.1 The Corrected Empirical Measure

In this section we define the so-called corrected empirical measure, relying on the Γ-convergence of the environment. First we need to extract some information about convergence of the operators $\mathscr{L}_n$ to $\mathscr{L}$ from the Γ-convergence of the associated Dirichlet forms.

Take a general Hilbert space $\mathscr{H}$ and let $\mathscr{A}$ be a non-negative, symmetric operator defined in $\mathscr{H}$. By Lax–Milgram theorem, we know that for any $\lambda > 0$ and any $g \in \mathscr{H}$, the equation $(\lambda + \mathscr{A})f = g$ has a unique solution in $\mathscr{H}$. Moreover, the solution f is the minimizer of the functional $f \mapsto \langle f, \mathscr{A} f\rangle + \lambda ||f||^2 - 2\langle f, g\rangle$. Fix $\lambda > 0$. For a given function $G \in \mathscr{L}^2(\mu)$, define the functionals

$$\mathscr{E}_n^G(F) = \mathscr{E}_n(F) + \lambda\langle S_n F, S_n F\rangle_n - 2\langle S_n F, S_n G\rangle_n,$$

$$\mathscr{E}^G(F) = \mathscr{E}(F) + \lambda\langle F, F\rangle - 2\langle F, G\rangle.$$

By Proposition 33.2, $\mathscr{E}_n^G$ is Γ-convergent to $\mathscr{E}^G$. In particular, a sequence of minimizers F_n of $\mathscr{E}_n^G$ converge to the minimizer F of $\mathscr{E}^G$. Notice that F_n is not uniquely defined in general, although $S_n F_n$ it is. By the discussion above, $(\lambda - \mathscr{L}_n) S_n F_n = S_n G$ and $(\lambda - \mathscr{L})F = G$. Since the operator norm of S_n is bounded by Θ, we conclude that the $\mathscr{L}^2(X_n)$-norm of $S_n F_n - S_n F$ converges to 0 as $n \to \infty$. By **(H2)**, we conclude that $\mathscr{E}_n(F_n)$ converges to $\mathscr{E}(F)$.

Now we are ready to define the corrected empirical measure $\hat{\pi}_t^n$. Take a function $G \in \mathscr{K}$ and define $H = (\lambda - \mathscr{L})G$. Define G_n as a minimizer of $\mathscr{E}_n^H$. Notice that in this way $S_n G_n$ is uniquely defined. Then we define

$$\hat{\pi}_t^n(G) = \frac{1}{a_n} \sum_{x \in X_n} \eta_t^n(x) S_n G_n(x).$$

In order to prove that $\hat{\pi}_t^n(G)$ is well defined, we need to prove that $\sum_x S_n G_n(x)$ is finite. Remember that $(\lambda - \mathscr{L}_n) S_n G_n = S_n H$. Consider the continuous-time random walk with jump rates $\omega_{x,y}^n$. Remember that the condition $\sup_x \sum_y \omega_{x,y}^n$ ensures that this random walk is well defined. Let $p_t^n(x, y)$ be its transition probability function. An explicit formula for $S_n G_n$ in terms of $p_t^n(x, y)$ is

$$S_n G_n(x) = \int_0^\infty e^{-\lambda t} \sum_{y \in X_n} p_t^n(x, y) S_n H(y) dt.$$

Since $\sum_x p_t(x, y) = 1$ for any $y \in X_t$, we conclude that

$$\frac{1}{a_n} \sum_{x \in X_n} S_n G_n(x) = \frac{1}{\lambda} \int H d\mu$$

and in particular $S_n G_n$ is summable. We conclude that $\hat{\pi}_t^n(G)$ is well defined. Notice that it is not clear at all if $\hat{\pi}_t^n$ is well defined as a measure in X.

33.3.2 Tightness of $\{\pi_\cdot^n\}_n$ and Proof of Theorem 33.1

In this section we prove tightness of $\{\pi_\cdot^n\}_n$ and we prove Theorem 33.1. As we will see, we rely on the corrected empirical measure, which turns out to be the right object to be studied. By **(H5)**, we have

$$\lim_{n\to\infty} \mathbb{P}_n\Big(\sup_{0\le t<+\infty} \big|\pi_t^n(G) - \hat{\pi}_t^n(G)\big| > \epsilon\Big) = 0.$$

Notice that **(H5)** can be substituted by the following condition, which can be sometimes proved directly.

(H5') For any $G \in \mathscr{K}$,

$$\lim_{n\to\infty} \frac{1}{a_n} \sum_{x\in X_n} \big|S_n G_n(x) - S_n G(x)\big| = 0.$$

In particular, $\{\pi_\cdot^n(G)\}_n$ is tight if and only if $\{\hat{\pi}_\cdot^n(G)\}_n$ is tight. The usual way of proving tightness of $\{\hat{\pi}_\cdot^n(G)\}_n$ is to use a proper martingale decomposition. A simple computation based on Dynkin's formula shows that

$$\mathscr{M}_t^n(G) = \hat{\pi}_t^n(G) - \hat{\pi}_0^n(G) - \int_0^t \pi_s^n(\mathscr{L}_n S_n G_n) ds \tag{33.2}$$

is a martingale. The quadratic variation of$\mathscr{M}_t^n(G)$ is given by

$$\langle \mathscr{M}_t^n(G)\rangle = \int_0^t \frac{1}{a_n^2} \sum_{x,y\in X_n} \big(\eta_s^n(y) - \eta_s^n(x)\big)^2 \omega_{x,y}^n \big(S_n G_n(y) - S_n G_n(x)\big)^2 ds.$$

In particular, $\langle \mathscr{M}_t^n(G)\rangle \le t a_n^{-1} \mathscr{E}_n(G_n)$. At this point, the convenience of introducing the corrected empirical process becomes evident. By definition, $\mathscr{L}_n S_n G_n = S_n \mathscr{L} G + \lambda(S_n G_n - S_n G)$. Since $H = (\lambda - \mathscr{L})G$, the function G is the minimizer of $\mathscr{E}^H$. Therefore, G_n converges to G in $\mathscr{L}^2(X)$. By **(H2)**, the $\mathscr{L}^2(X_n)$-norm of $S_n G_n - S_n G$ goes to 0 and $\mathscr{E}_n(G_n)$ converges to $\mathscr{E}(G)$.

We conclude that $\mathscr{M}_t^n(G)$ converges to 0 as $n \to \infty$, and in particular the sequence $\{\mathscr{M}_\cdot^n(G)\}_n$ is tight. In the other hand, the integral term in (33.2) is equal to $\int_0^t \pi_s^n(\mathscr{L}G) ds$.

Notice that $\pi_s^n(\mathscr{L}G) \le \int |\mathscr{L}G| d\mu$ for any $t \ge 0$, from where we conclude that the integral term is of bounded variation, uniformly in n. Tightness follows at once. Since $\{\hat{\pi}_0^n(G)\}_n$ is tight by comparison with $\{\pi_0^n(G)\}_n$, we conclude that $\{\hat{\pi}_\cdot^n(G)\}_n$ is tight, which proves the first part of Theorem 33.1. As a by-product, we have obtained tightness for $\{\pi_\cdot^n\}_n$ as well, and the convergence result

$$\lim_{n\to\infty} \Big\{ \pi_t^n(G) - \pi_0^n(G) - \int_0^t \pi_s^n(\mathscr{L}G) ds \Big\} = 0$$

for any $G \in \mathscr{K}$. Notice that we have exchanged $\hat{\pi}_t^n(G)$ by $\pi_t^n(G)$. Let $\pi_\cdot$ be a limit point of $\{\pi_\cdot^n\}_n$. Then, $\pi_\cdot$ satisfies the identity

$$\pi_t(G) - \pi_0(G) - \int_0^t \pi_s(\mathscr{L}G)ds = 0$$

for any function $G \in \mathscr{K}$. By hypothesis, $\pi_0(dx) = u_0(x)\mu(dx)$. Repeating the arguments for a function $G_t(x) = G_0(x) + tG_1(x)$ with $G_0, G_1 \in \mathscr{K}$, we can prove that

$$\pi_t(G_t) - \pi_0(G_0) - \int_0^t \pi_s((\partial_t + \mathscr{L})G_s)ds = 0$$

for any piecewise-linear trajectory $G_\cdot : [0,T] \to \mathscr{K}$. The same identity holds by approximation for any smooth path $G_\cdot : [0,T] \to \mathscr{C}_c(X)$, which proves that the process $\pi_\cdot$ is concentrated on weak solutions of the hydrodynamic equation. When such solutions are unique, the process π is just a δ-distribution concentrated on the path $u(t,x)\mu(dx)$. Since compactness plus uniqueness of limit points imply convergence, Theorem 33.1 is proved.

33.4 Energy Solutions and Energy Estimate

In this section we define what we mean by *energy solutions* of (33.1), we prove that any limit point of the empirical measure $\{\pi_\cdot^n\}$ is concentrated on energy solutions of (33.1) and we give a simple criterion for uniqueness of such solutions.

33.4.1 Energy Solutions

Let $\mathscr{E} : H \to \bar{\mathbb{R}}$ be a quadratic form defined over a Hilbert space H of inner product $\langle \cdot, \cdot \rangle$. We say that $\mathscr{E}$ is *closable* if for any sequence $\{f_n\}_n$ converging in H to some limit f such that $\mathscr{E}(f_n - f_m)$ goes to 0 as $n, m \to \infty$, we have $f = 0$. Let $\mathscr{E} : H \to \bar{\mathbb{R}}$ be closable. We define $\mathscr{H}_1 = \mathscr{H}_1(\mathscr{E})$ as the closure of the set $\{f \in H; \mathscr{E}(f) < +\infty\}$ under the norm $||f||_1 = (\mathscr{E}(f) + \langle f, f \rangle)^{1/2}$.

We say that a dense set $K \subseteq H$ is a *kernel* of $\mathscr{E}$ if $\mathscr{H}_1$ is equal to the closure of K under the norm $|| \cdot ||_1$. We say that a symmetric operator $\mathscr{L} : D(\mathscr{L}) \subseteq H \to H$ generates $\mathscr{E}$ if $\mathscr{E}(f) = \langle f, -\mathscr{L}f \rangle$ for $f \in D(\mathscr{L})$ and $D(\mathscr{L})$ is a kernel of $\mathscr{E}$.

Fix $T > 0$. For a function $u : [0,T] \to H$ we define the norm

$$||u||_{1,T} = \Big(\int_0^T ||u_t||_1^2 dt \Big)^{1/2}$$

and we define $\mathscr{H}_{1,T}$ as the Hilbert space generated by this norm. Given a closable form $\mathscr{E}$ generated by the operator $\mathscr{L}$, we say that a trajectory $u : [0,T] \to H$

is an *energy solution* of (33.1) if $u \in \mathscr{H}_{1,T}$ and for any differentiable trajectory $G : [0,T] \to \mathscr{H}_1$ with $G(T) = 0$ we have

$$\langle G_0, u_0\rangle + \int_0^T \Big\{\langle \partial_t G_t, u_t\rangle - \mathscr{E}(G_t, u_t)\Big\} dt = 0.$$

In other words, an energy solution of (33.1) is basically a weak solution belonging to $\mathscr{H}_{1,T}$. In fact, by taking suitable approximations of G, it is enough to prove this identity for trajectories G such that $G_t \in K$ for any $t \in [0,T]$, where K is any kernel of $\mathscr{E}$ contained in $D(\mathscr{L})$. Notice that the norm in $\mathscr{H}_{1,T}$ is stronger than the norm $\int_0^T u_t^2 dt$, and therefore a weak solution is effectively weaker than an energy solution of (33.1).

33.4.2 The Energy Estimate

In this section we prove that the limit points of the empirical measure are concentrated on energy solutions of (33.1). For simplicity, we work on finite volume. From now on we assume that X is compact. Therefore, there exists a constant κ such that the cardinality of X_n is bounded by κa_n. We have the following estimate.

Theorem 33.2. *Fix $T > 0$. Let $\{H^i : X_n \times X_n \times [0,T] \to \mathbb{R}; i = 1,\dots,l\}$ be a finite sequence of functions. There exists a constant $C = C(T)$ such that*

$$\mathbb{E}_n\Big[\sup_{i=1,\dots,l} \int_0^T \Big\{\frac{2}{a_n} \sum_{x,y\in X_n} \omega^n_{x,y} H^i_{x,y}(t)\big(\eta^n_t(y) - \eta^n_t(x)\big) - \frac{1}{a_n} \sum_{x,y\in X_n} \omega^n_{x,y} (H^i_{x,y})^2 \eta^n_t(x)\Big\} dt\Big] \le C + \frac{\log l}{a_n}. \quad (33.3)$$

Proof. Before starting the proof of this theorem, we need some definitions. Fix $\rho > 0$. Denote by ν^ρ the product measure in Ω_n defined by

$$\nu^\rho\big(\eta(x_1) = 1, \dots, \eta(x_k) = 1\big) = \rho^k.$$

It is not difficult to check that the measure ν^ρ is left invariant under the evolution of η_t. For two given probability measures P_1, P_2, we define the entropy $H(P_1|P_2)$ of P_1 with respect to P_2 as

$$H(P_1|P_2) = \begin{cases} +\infty, & \text{if } P_1 \text{ is not absolutely continuous with respect to } P_2 \\ \int \log \frac{dP_1}{dP_2} dP_1 & \text{otherwise.} \end{cases}$$

For $\eta \in \Omega_n$, denote by δ_η the Dirac measure at η. It is not difficult to see that $H(\delta_\eta|\nu^\rho) \leq C(\rho)a_n$ for any $\eta \in \Omega_n$, where $C(\rho)$ is a constant that can be chosen independently from n. Let us denote by $\mathbb{P}^\rho$ the distribution in $D([0,T],\Omega_n)$ of the process η_t^n with initial distribution ν^ρ. By the convexity of the entropy, $H(\mathbb{P}_n|\mathbb{P}^\rho) \leq C(\rho,T)a_n$ for a constant $C(\rho,T)$ not depending on n. The following arguments are standard and can be found in full rigor in [13]. Let us denote by $F^i(s)$ the function (depending on $H^i(s)$ and η_s^n) under the time integral in (33.3). By the entropy estimate,

$$\mathbb{E}_n\Big[\sup_{i=1,\dots,l}\int_0^T F^i(t)dt\Big] \leq \frac{H(\mathbb{P}_n|\mathbb{P}^\rho)}{a_n} + \frac{1}{a_n}\log\mathbb{E}^\rho\Big[\exp\big\{\sup_{i=1,\dots,l} a_n \int_0^T F^i(t)dt\big\}\Big].$$

In order to take the supremum out of the expectation, we use the inequalities $\exp\{\sup_i b_i\} \leq \sum_i \exp\{b_i\}$ and $\log\{\sum_i b_i\} \leq \log l + \sup_i \log b_i$, valid for any real numbers $\{b_i, i = 1,\dots,l\}$. In this way we obtain the bound

$$\begin{aligned}\mathbb{E}_n\Big[\sup_{i=1,\dots,l}\int_0^T F^i(t)dt\Big] &\leq C(\rho,T) + \frac{\log l}{a_n}\\ &\quad + \sup_{i=1,\dots,l}\frac{1}{a_n}\log\mathbb{E}^\rho\Big[\exp\big\{a_n\int_0^T F^i(t)dt\big\}\Big].\end{aligned} \tag{33.4}$$

Therefore, it is left to prove that the last supremum is not positive. It is enough to prove that the expectation $\mathbb{E}^\rho\big[\exp\big\{\int_0^T F^i(t)dt\big\}\big]$ is less or equal than 1 for any function F^i. From now on we drop the index i. By Feynman–Kac's formula plus the variational formula for the largest eigenvalue of the operator $F(t)+L_n$, we have

$$\frac{1}{a_n}\log\mathbb{E}^\rho\Big[\exp\big\{a_n\int_0^T F(t)dt\big\}\Big] \leq \int_0^T \sup_f\big\{\langle F(t), f^2\rangle_\rho - \langle f, -L_n f\rangle_\rho\big\},$$

where we have denoted by $\langle\cdot,\cdot\rangle_\rho$ the inner product in $\mathscr{L}^2(\nu_\rho)$ and the supremum is over functions $f \in \mathscr{L}^2(\nu_\rho)$. A simple computation using the invariance of ν_ρ shows that

$$\langle f, -L_n f\rangle_\rho = \sum_{x,y\in X_n}\omega_{x,y}^n\int\big[f(\eta^{x,y}) - f(\eta)\big]^2\nu_\rho(d\eta).$$

Recall the expression for $F(t)$ in terms of H. We will estimate each term of the form $2a_n^{-1}\langle H_{x,y}(\eta(y)-\eta(x)), f^2\rangle_\rho$ separatedly:

$$\frac{2}{a_n}\langle H_{x,y}(\eta(y)-\eta(x)), f^2\rangle_\rho = \frac{2}{a_n}H_{x,y}\langle \eta(x), f(\eta^{x,y})^2 - f(\eta)^2\rangle_\rho$$

$$\leq \frac{2}{a_n}\Big\{\frac{(H_{x,y})^2\beta^n_{x,y}}{2}\langle\eta(x),(f(\eta^{x,y})+f(\eta))^2\rangle_\rho$$
$$+\frac{1}{2\beta^n_{x,y}}\langle\eta(x),(f(\eta^{x,y})-f(\eta))^2\rangle_\rho\Big\}.$$

Choosing $\beta^n_{x,y} = 1/\omega^n_{x,y}$ and putting this estimate back into (33.4), we obtain the desired estimate. □

Take $G^i \in \mathscr{K}$ and take $H^i_{x,y} = S_n G^i_n(y) - S_n G^i_n(x)$, with G^i_n defined as in Sect. 33.3.1. Recall the identity $\mathscr{L}_n S_n G^i_n = S_n \mathscr{L} G^i + \lambda(S_n G^i_n - S_n G^i)$. The energy estimate (33.3) gives

$$\mathbb{E}_n\Big[\sup_{i=1,\dots,l}\int_0^T \big(2\hat{\pi}^n_t(\mathscr{L}G^i) - \mathscr{E}_n(G^i_n)\big)dt\Big] \leq C(\rho,T) + C_1(l,n),$$

where $C_1(l,n)$ is a constant that goes to 0 when l is fixed and $n \to \infty$. Take a limit point of the sequence $\{\pi^n_\cdot\}_n$. We have already seen that $\hat{\pi}^n_t(\mathscr{L}G^i)$ converges to $\pi_t(\mathscr{L}G)$. Therefore, the process $\pi_\cdot$ satisfies

$$E\Big[\sup_{i=1,\dots,l}\int_0^T \big(2\pi_s(\mathscr{L}G^i) - \mathscr{E}(G^i)\big)dt\Big] \leq C(\rho,T).$$

Similar arguments prove that for piecewise linear trajectories $\{G^i_t; i = 1,\dots,l\}$ in $\mathscr{K}$, we have

$$E\Big[\sup_{i=1,\dots,l}\int_0^T \big(2\pi_s(\mathscr{L}G^i(t)) - \mathscr{E}(G^i(t))\big)dt\Big] \leq C(\rho,T).$$

Since l is arbitrary and piecewise linear trajectories with values in $\mathscr{K}$ are dense in $\mathscr{H}_{1,T}$, we conclude that $E[||\pi_\cdot||^2_{1,T}] < +\infty$, from where we conclude that $||\pi_\cdot||_{1,T}$ is finite $a.s.$ We establish this result as a theorem.

Theorem 33.3. *Let η^n_t an exclusion process as in Theorem 33.1. If one of the following conditions is satisfied,*

(i) X is compact
(ii) Assumption **(H5')** *holds and the entropy density is finite:*

$$\sup_n \frac{H(\mathbb{P}_n|\mathbb{P}^\rho)}{a_n} < +\infty,$$

then any limit point of the sequence $\{\pi^n_\cdot(dx)\}_n$ is concentrated on energy solutions of the hydrodynamic equation (33.1). In particular, since such energy solutions are unique, the sequence $\{\pi^n_\cdot(dx)\}_n$ is convergent.

33.4.3 Uniqueness of Energy Solutions

In this section we prove uniqueness of energy solutions for (33.1). Since the equation is linear, it is enough to prove uniqueness for the case $u_0 \equiv 0$. Let u_t be a solution of (33.1) with $u_0 \equiv 0$. Then,

$$\int_0^T \big\{\langle \partial_t G_t, u_t\rangle - \mathscr{E}(G_t, u_t)\big\} dt = 0$$

for any differentiable trajectory in $\mathscr{H}_{1,T}$ with $G_T = 0$. Take $G_t = -\int_t^T u_s ds$. Then $\partial_t G_t = u_t$ and the first term above is equal to $\int_0^T \langle u_t, u_t\rangle dt$. An approximation procedure and Fubini's theorem shows that the second term above is equal to

$$\frac{1}{2}\mathscr{E}\Big(\int_0^T u_t dt\Big).$$

Both terms are non-negative, so we conclude that $\int_0^T \langle u_t, u_t\rangle dt = 0$ and $u_t \equiv 0$.

33.5 Applications

In this section we give some examples of systems on which Theorems 33.1 and 33.3 apply. In the literature, the sequence ω^n is often referred as the set of *conductances* of the model. Unless stated explicitely, in these examples, X will be equal to $\mathbb{R}^d$ or the torus $\mathbb{T}^d = \mathbb{R}^d/\mathbb{Z}^d$. The set X_n will be equal to $n^{-1}\mathbb{Z}^d$ and we construct the partitions $\{\mathscr{U}_x^n\}$ in the canonical way, taking $\mathscr{U}_x^n$ as a continuous, piecewise linear function with $\mathscr{U}_x^n(x) = 1$ and $\mathscr{U}_x^n(y) = 0$ for $y \in X_n$, $y \neq x$.

33.5.1 Homogenization of Ergodic, Elliptic Environments

Let $(\Omega, \mathscr{F}, P)$ be a probability space. Let $\{\tau_x; x \in \mathbb{Z}^d\}$ be a family of $\mathscr{F}$-measurable maps $\tau_x : \Omega \to \Omega$ such that

(1) $P(\tau_x^{-1} A) = P(A)$ for any $A \in \mathscr{F}$, $x \in \mathbb{Z}^d$.
(2) $\tau_x \tau_{x'} = \tau_{x+x'}$ for any $x, x' \in \mathbb{Z}^d$.
(3) If $\tau_x A = A$ for any $x \in \mathbb{Z}^d$, then $P(A) = 0$ or 1.

In this case, we say that the family $\{\tau_x\}_{x \in \mathbb{Z}^d}$ is ergodic and invariant under P. Let $a = (a_1, \dots, a_d) : \Omega \to \mathbb{R}^d$ be an $\mathscr{F}$-measurable function. Assume that there exists $\epsilon_0 > 0$ such that

$$\epsilon_0 \leq a_i(\omega) \leq \epsilon_0^{-1} \text{ for all } \omega \in \Omega \text{ and } i = 1, \dots, d.$$

We say in this situation that the environment satisfies the *ellipticity condition*. Fix $\omega \in \Omega$. Define ω^n by $\omega^n_{x,x+e_i/n} = \omega^n_{x+e_i/n,x} = n^2 a_i(\tau_n x\omega)$, $\omega^n_{x,y} = 0$ if $|y-x| \neq 1/n$. Here $\{e_i\}_i$ is the canonical basis of $\mathbb{Z}^d$. In this case, $a_n = n^d$ and μ is the Lebesgue measure in $\mathbb{R}^d$. In [16], it is proved that there is a positive definite matrix A such that the quadratic form $\mathscr{E}_n$ associated to ω^n is Γ-convergent to $\mathscr{E}(f) = \int \nabla f \cdot A\nabla f dx$, $P-a.s.$ In particular, Theorem 33.1 applies with $\mathscr{L} f = \text{div}(A\nabla f)$. This result was first obtained in [7].

33.5.2 The Percolation Cluster

Let $e = \{e^i_x; x \in \mathbb{Z}^d, i = 1, \cdots, d\}$ be a sequence of i.i.d. random variables, with $P(e^i_x = 1) = 1 - P(e^i_x = 0) = p$ for some $p = (0,1)$. Define for $x, y \in X_n$, $\omega^n_{x,x+e_i/n} = \omega^n_{x+e_i/n,x} = n^2 e^i_{nx}$, $\omega^n_{x,y} = 0$ if $|y-x| \neq 1/n$. Fix a realization of e. We say that two points $x, y \in X_n$ are connected if there is a finite sequence $\{x_0 = x, \dots, x_l = y\} \subseteq X_n$ such that $|x_{i-1} - x_i| = 1/n$ and $\omega^n_{x_{i-1},i} = 1$ for any i. Denote by $\mathscr{C}_0$ the set of points connected to the origin. It is well known that there exists $p_c \in (0,1)$ such that $\theta(p) = P(\mathscr{C}_0$ is infinite $)$ is 0 for $p < p_c$ and positive for $p > p_c$. Fix $p > p_c$. Define $a_n = n^d$ and $\mu_0(dx) = \theta(p)dx$. In [3], it is proved that there exists a constant D such that, $P-a.s$ in the set $\{\mathscr{C}_0$ is infinite $\}$, the quadratic form $\mathscr{E}_n$ associated to the environment ω^n restricted to $\mathscr{C}_0$ is Γ-convergent to $\mathscr{E}(f) = \theta(p)D\int(\nabla f)^2 dx$. Theorem 33.1 applies with $\mathscr{L} = D\Delta$, assuming that the initial measures ν_n put mass zero in configurations with particles outside $\mathscr{C}_0$. This result was first obtained in [3], relying on a duality representation of the simple exclusion process.

33.5.3 One-Dimensional, Inhomogeneous Environments

In dimension $d = 1$, the Γ-convergence of $\mathscr{E}_n$ can be studied explicitely. For nearest-neighbors environments ($\omega^n_{x,y} = 0$ if $|x-y| = 1$), Γ-convergence of $\mathscr{E}_n$ is equivalent to convergence in distribution of the measures

$$W_n(dx) = \frac{1}{n}\sum_{x\in\mathbb{Z}}(\omega^n_{x,x+1})^{-1}\delta_{x/n}(dx).$$

Let $W(dx)$ be the limit. We assume that $W(dx)$ gives positive mass to any open set. For simplicity, suppose that $W(\{0\}) = 0$. Otherwise, we simply change the origin to another point with mass zero. For two functions $f, g : \mathbb{R} \to \mathbb{R}$ we say that $g = df/dW$ if

$$f(x) = f(0) + \int_0^x g(y)W(dy).$$

Then $\mathscr{E}_n$ is Γ-convergent to the quadratic form defined by $\mathscr{E}(f) = \int (df/dW)^2 dW$. In this case, $\mathscr{L} = d/dxd/dW$. A technical difficulty appears if $W(dx)$ has atoms. In that case, there is no kernel $\mathscr{K}$ for $\mathscr{L}$ contained in $\mathscr{C}_c(\mathbb{R})$. To overcome this point, we define for $x \leq y$, $d_W(x,y) = d_W(y,x) = W((x,y])$. The function d_W is a metric in $\mathbb{R}$, and in general $\mathbb{R}$ is *not* complete under this metric: an increasing sequence x_n converging to x is always a Cauchy sequence with respect to d_W, but $d_W(x_n, x) \geq W(\{x\})$, which is non-zero if x is an atom of W. Define $\mathbb{R}_W = \mathbb{R} \cup \{x-; W(\{x\}) > 0\}$. It is easy to see that $\mathbb{R}_W$ is a complete, separable space under the natural extension of d_W, and that continuous functions in $\mathbb{R}_W$ are in bijection with càdlàg functions in $\mathbb{R}$ with discontinuity points contained on the set of atoms of $W(dx)$. It is not difficult to see that the set of W-differentiable functions in $\mathscr{C}_c(\mathbb{R}_W)$ is a kernel for $\mathscr{L}$ and that Theorems 33.1 and 33.3 apply to this setting. In [2], the remarkable case on which $W(dx)$ is a *random*, self-similar measure (an α-stable subordinator) was studied in great detail.

33.5.4 Finitely Ramified Fractals

Let us consider the following sequence of graphs in $\mathbb{R}^2$. Define $a_0 = (0,0)$, $a_1 = (1/2, \sqrt{3}/2)$, and $a_2 {=} (1,0)$ and define $\varphi_i : \mathbb{R}^2 {\to} \mathbb{R}^2$ by taking $\varphi_i(x) = (x + a_i)/2$. Define $X_0 = \{a_0, a_, a_2\}$ and $X_{n+1} = \cup_i \varphi_i(X_n)$ for $n \geq 0$. For $x, y \in X_0$ we define $\omega^0_{x,y} = 1$, we put $\omega^0_{x,y} = 0$ if $\{x,y\} \subsetneq X$ and inductively we define

$$\omega^{n+1}_{x,y} = 5 \sum_i \omega^n_{\varphi_i^{-1}(x), \varphi^{-1}(y)}.$$

The set X_n is a discrete approximation of the Sierpinski gasket X defined as the unique compact, non-empty set X such that $X = \cup_i \varphi_i(X)$. Here we are just saying that $\omega^n_{x,y} = 5^n$ if x, y are neighbors in the canonical sense. In this case $a_n = 3^n$ and μ is the Hausdorff measure in X. It has been proved [12] that the quadratic forms $\mathscr{E}_n$ converge to a certain Dirichlet form $\mathscr{E}$ which is used to define an abstract Laplacian in X. In particular, Theorems 33.1 and 33.3 apply to this model. This result was obtained in [11] in the context of a zero-range process. The same result can be proved for general *finitely ramified fractals*, in the framework of [12].

References

1. Faggionato, A.: Bulk diffusion of 1D exclusion process with bond disorder. Markov Process. Relat. Fields **13**(3), 519–542 (2007)
2. Faggionato, A., Jara, M., Landim, C.: Hydrodynamic behavior of 1D subdiffusive exclusion processes with random conductances. Probab. Theory Relat. Fields **144**(3–4), 633–667 (2009)
3. Faggionato, A.: Random walks and exclusion processes among random conductances on random infinite clusters: homogenization and hydrodynamic limit. Electron. J. Probab. **13**, 2217–2247 (2008)

4. Faggionato, A., Martinelli, F.: Hydrodynamic limit of a disordered lattice gas. Probab. Theory Relat. Fields **127**(4), 535–608 (2003)
5. Franco, F., Landim, C.: Hydrodynamic limit of gradient exclusion processes with conductances. Arch. Rat. Mech. Anal. **195**, 409–439 (2010)
6. Fritz, J.: Hydrodynamics in a symmetric random medium. Comm. Math. Phys. **125**(1), 13–25 (1989)
7. Gonçalves, P., Jara, M.: Scaling limits for gradient systems in random environment. J. Stat. Phys. **131**(4), 691–716 (2008)
8. Gonçalves, P., Jara, M.: Scaling limits of a tagged particle in the exclusion process with variable diffusion coefficient. J. Stat. Phys. **132**(6), 1135–1143 (2008)
9. Guo, M.Z., Papanicolaou, G.C., Varadhan, S.R.S.: Nonlinear diffusion limit for a system with nearest neighbor interactions. Comm. Math. Phys. **118**(1), 31–59 (1988)
10. Jara, M.D., Landim, C.: Nonequilibrium central limit theorem for a tagged particle in symmetric simple exclusion. Ann. Inst. H. Poincaré Probab. Statist. **42**(5), 567–577 (2006)
11. Jara, M.: Finite-dimensional approximation for the diffusion coefficient in the simple exclusion process. Ann. Probab. **34**(6), 2365–2381 (2006)
12. Kigami, J.: Analysis on fractals, volume 143 of Cambridge Tracts in Mathematics. Cambridge University Press, Cambridge (2001)
13. Kipnis, C., Landim, C.: Scaling limits of interacting particle systems, volume 320 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer, Berlin (1999)
14. Koukkous, A.: Hydrodynamic behavior of symmetric zero-range processes with random rates. Stoch. Process. Appl. **84**(2), 297–312 (1999)
15. Nagy, K.: Symmetric random walk in random environment in one dimension. Period. Math. Hung. **45**(1–2), 101–120 (2002)
16. Papanicolaou, G.C., Varadhan, S.R.S.: Diffusions with random coefficients. In: Statistics and probability: essays in honor of C. R. Rao, pp. 547–552. North-Holland, Amsterdam (1982)
17. Quastel, J.: Bulk diffusion in a system with site disorder. Ann. Probab. **34**(5), 1990–2036 (2006)
18. Tartar, L.: Homogénéisation et compacité par compensation. In: Séminaire Goulaouic-Schwartz (1978/1979), pages Exp. No. 9, 9. École Polytech., Palaiseau (1979)

Chapter 34
Application of Fractional Order Concepts in the Study of Electrical Potential

Isabel S. Jesus and J.A. Tenreiro Machado

Abstract The Maxwell equations, expressing the fundamental laws of electricity and magnetism, only involve the integer-order calculus. However, several effects present in electromagnetism, motivated recently an analysis under the fractional calculus (FC) perspective. In fact, this mathematical concept allows a deeper insight into many phenomena that classical models overlook. On the other hand, genetic algorithms (GA) are an important tool to solve optimization problems that occur in engineering. In this work we use FC and GA to implement the electrical potential of fractional order. The performance of the GA scheme and the convergence of the resulting approximations are analyzed.

34.1 Introduction

A fresh look into several phenomena present in electrical systems [1] induced an approach supported by the fractional calculus (FC). Some authors [2–4] verified that well-known expressions for the electrical potential are related through integer-order integrals and derivatives and have proposed its generalization, leading to the concept of fractional-order poles. Nevertheless, the mathematical generalization towards FC lacks a comprehensive method for a practical implementation.

This article addresses the synthesis of fractional-order multipoles and is organized as follows. In Sect. 34.2 we recall the classical expressions for the static electrical potential and we analyze them in the perspective of FC. Based on this re-evaluation we develop a GA scheme for implementing fractional-order electrical potential approximations. Finally, in Sect. 34.3 we outline the main conclusions.

I.S. Jesus (✉)
Institute of Engineering of Porto, Rua Dr. António Bernardino de Almeida, 4200-072 Porto, Portugal
e-mail: isj@isep.ipp.pt

J.A.T. Machado
Department of Electrotechnical Engineering, Institute of Engineering of Porto, Rua Dr. António Bernardino de Almeida, 431, 4200-072 Porto, Portugal
e-mail: jtm@isep.ipp.pt

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_34,

34.2 Integer and Fractional Electrical Potential

For homogeneous, linear and isotropic media, the electric potential φ at a point P produced by a single charge, a dipole, a quadrupole, an infinite straight filament, two opposite charged filaments, and a planar surface, is given by $\varphi = \frac{q}{4\pi\varepsilon_0}\frac{1}{r} + C$, $\varphi = \frac{ql\cos\theta}{4\pi\varepsilon_0}\frac{1}{r^2} + C,\ r{>>}l, \varphi = \frac{ql^2(3\cos^2\theta - 1)}{4\pi\varepsilon_0}\frac{1}{r^3} + C,\ r{>>}l, \varphi = -\frac{\lambda}{2\pi\varepsilon_0}\ln r + C$, $\varphi = \frac{\lambda l\cos\theta}{2\pi\varepsilon_0}\frac{1}{r} + C,\ r{>>}l,\ \varphi = -\frac{\sigma}{2\varepsilon_0}r + C$, respectively, where $C \in \Re$, ε_0 represents the permittivity, q the electric charge, λ the density of charges per length, σ the density of charges per surface, l the length, r the radial distance and θ the corresponding angle [5].

Analyzing these expressions we verify the relationship $\varphi \sim r^{-3}$, r^{-2}, r^{-1}, $\ln r$, r corresponding to the successive application of integer-order derivatives and integrals.

The integer-order differential nature of the expressions motivated several authors [4,6] to propose its generalization in a FC perspective. Therefore, a fractional multipole produces at point P a potential $\varphi \sim r^{\alpha}$, $\alpha \in \Re$. Nevertheless, besides the abstract manipulation of mathematical expressions, the truth is that there is no practical method for establishing the fractional potential [3,4,6].

Inspired by the integer-order approximations of fractional transfer functions [7,8], with recursive poles and zeros, we adopt a genetic algorithm (GA) [9,10] for implementing the fractional potential using the multipole integer counterpart. In fact, similarly to what occur with transfer function, the electrical integer-order potential has a *global* nature and fractional potentials can have only a *local* nature. By other words, fractional potentials are possible to capture only in a restricted region of the space. This observation leads to an implementation approach conceptually similar to the one described in [6–8,11] that is, to an approximation scheme based on a recursive superposition of integer potentials.

In this line of thought, we develop a one-dimensional GA that determines n charges q_i at the positions x_i. Our goal is to compare the approximate potential $\varphi_{app} = \sum_{i=0}^{n-1}\frac{q_i}{4\pi\varepsilon_0|x - x_i|}$, where n is the total number of charges, that mimics the desired reference potential $\varphi_{ref} = kx^{\alpha}$ in a given interval $x_{mim} < x < x_{max}$.

The experiments consist on executing the GA, in order to generate a combination of charges and positions that lead to an electrical potential with fractional slope similar to the desire reference potential. In the first case of study, the values of GA parameters are: population number P $=$ 40, crossover C(%) $=$ 85.0%, mutation M(%) $=$ 1.0% and an elitist strategy ES(%) $=$ 10.0%. The chromosome has $2n$ genes: the first n genes correspond to the charges q_i and the last n genes indicate their positions x_i $(i = 0, \ldots, n-1)$. The gene codifications adopts a Gray code with a string length of 16 bits. The optimization fitness function corresponds to the minimization of the index $J = \sum\limits_{k=1}^{m}\left(\ln\left|\frac{\varphi_{app}}{\varphi_{ref}}\right|\right)^2,\ \min\limits_{i}(J),\ i = 0,\ 1, \ldots,\ n-1$, where m is the number of sampling points along the interval $x_{mim} < x < x_{max}$. We establish a maximum number of iterations $I_{Max} = 100$ and a stoping scheme when $J < 10^{-10}$ for the best individual (i.e., solution) of the GA population.

In the following experiments the results have a scale factor of $\times(4\pi\varepsilon_0)^{-1}$.

Figure 34.1a shows the electrical potential φ when $\varphi_{ref} = 1.0\ x^{-1.5}$, $n = 5$ and $0.2 < x < 0.8$, leading to $\{q_1, q_2, q_3, q_4, q_5\} = \{0.737, 0.846, -0.777, 0.382, -0.225\}$ (C), located at $\{x_1, x_2, x_3, x_4, x_5\} = \{-0.06, 0.092, 0.147, -0.106, 0.117\}$ (m). In this case, the GA needs $I = 51$ iterations to satisfy the fitness function stoping threshold. The results show a good fit between φ_{ref} and φ_{app}. Repeating the GA execution, due to its stochastic nature, we verify that it is possible to find more than one 'good' solution (Fig. 34.1b).

With the proposed method it is also possible to have a reference potential with other slope values α [6, 11]. Therefore, we apply the GA with identical parameters, for $0.2 < x < 0.8$ (m) while varying α, namely from $\alpha \geq -2.0$ up to $\alpha \leq -0.5$.

Figure 34.2a shows a $\varphi_{ref} = 1.0\ x^{-1.3}$, $n = 5$ and $0.2 < x < 0.8$, leading to $\{q_1, q_2, q_3, q_4, q_5\} = \{0.471, 0.464, 0.578, -0.371, -0.173\}$ (C), located at $\{x_1, x_2, x_3, x_4, x_5\} = \{-0.125, 0.029, 0.037, 0.132, 0.152\}$ (m). Figure 34.2b

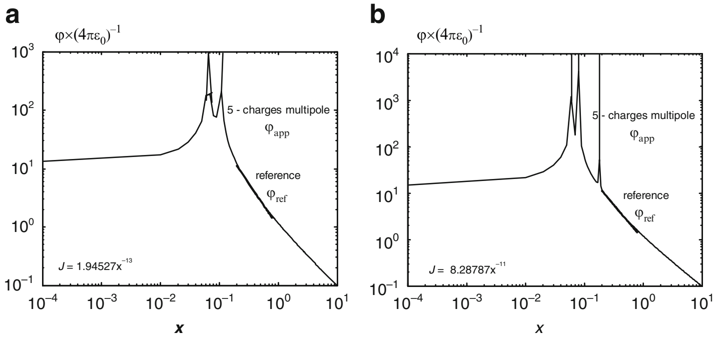


Fig. 34.1 Comparison of the electric potential φ_{app} and φ_{ref} *versus* the position x for $\varphi_{ref} = 1.0\ x^{-1.5}$ (volt), $n = 5$ charge approximation and $0.2 < x < 0.8$ (m)

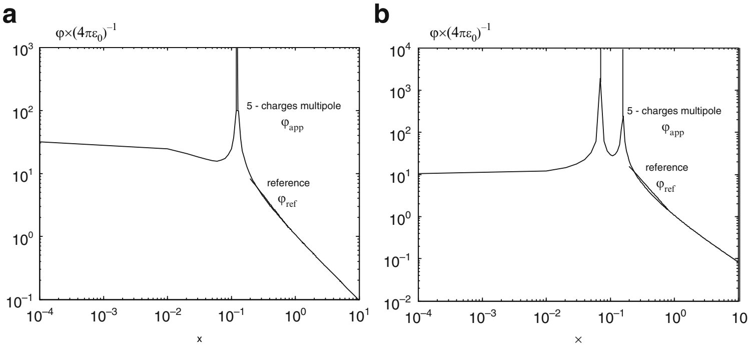


Fig. 34.2 Comparison of the electrical potential φ_{app} and φ_{ref} *versus* the position x for (**a**) $\varphi_{ref} = 1.0\ x^{-1.3}$ (volt) and (**b**) $\varphi_{ref} = 1.0\ x^{-1.7}$ (volt), $n = 5$ charge approximation and $0.2 < x < 0.8$ (m)

shows the case of $\varphi_{ref} = 1.0\ x^{-1.7}$, $n = 5$ and $0.2 < x < 0.8$, leading to $\{q_1, q_2, q_3, q_4, q_5\} = \{0.753, 0.535, 0.429, -0.218, -0.681\}$(C), located at $\{x_1, x_2, x_3, x_4, x_5\} = \{-0.157, -0.070, 0.171, 0.188, 0.200\}$ (m).

34.3 Conclusions

This paper addressed the problem of implementing a fractional-order electrical potential through a GA. The GA establisher a good compromise between the approximation accuracy and the computational time. Furthermore, the proposed technique leads to good results for different values of the fractional order.

References

1. Engheta, N.: On fractional calculus and fractional multipoles in electromagnetism. IEEE Trans. Antennas Propag. **44**, 554–566 (1996)
2. Machado, J.T., Jesus, I.S., Galhano, A.: Electric fractional order potential. Proceedings of the XII International Symposium on Electromagnetics Fields in Mechatronics, Electrical and Electronic Engineering (ISEF'05), Spain (2005)
3. Machado, J.T., Jesus, I., Galhano, A., Cunha, J.B.: Fractional order electromagnetics. Signal Processing (EURASIP/Elsevier) Special Issue on Fractional Calculus Applications in Signals and Systems, vol. 86, pp. 2637–2644 (2006)
4. Jesus, I.S., Machado, J.A.T., Cunha, J.B.: Application of genetic algorithms to the implementation of fractional electromagnetic potentials. Proceedings of The Fifth International Conference on Engineering Computational Technology (ECT'06), Spain (2006)
5. Bessonov, L.: Applied Electricity for Engineers. MIR, Moscow (1968)
6. Jesus, I.S., Machado, J.T.: An Evolutionary approach for the synthesis of fractional potentials. Fractional Calculus and Applied Analysis – FCAA, An International Journal for Theory and Applications, vol. 3, pp. 237–248 (2008)
7. Oldham, K.B., Spanier, J.: The Fractional Calculus: Theory and Application of Differentiation and Integration to Arbitrary Order. Academic, New York (1974)
8. Oustaloup, A.: La Commande CRONE: Commande Robuste d'Ordre Non Entier. Hermes, France (1991)
9. Goldberg, D.E.: Genetic Algorithms in Search Optimization and Machine Learning. Addison-Wesley, Reading, MA (1989)
10. Mitchell, M.: An Introduction to Genetic Algorithms. MIT, Cambridge, MA (1998)
11. Jesus, I.S., Machado, J.A.T., Barbosa, R.S.: Implementing an electrical fractional potential through a genetic algorithm. Proceedings of The 2nd Conference on Nonlinear Science and Complexity (NSC'08), Portugal (2008)

Chapter 35
Economics of Bioenergy Crops for Electricity Generation: Implications for Land Use and Greenhouse Gases

Madhu Khanna, Hayri Onal, Basanta Dhungana, and Michelle Wander

Abstract This chapter develops a dynamic linear optimization framework to examine the optimal land allocation for two perennial crops, switchgrass and miscanthus, that can be co-fired with coal for electricity generation. Detailed spatial data at county level is used to determine the heterogeneous costs of producing and delivering biomass to power plants in Illinois over a 15-year period. A transportation module is incorporated in the model to link power plants to perennial crop growing areas such that power plants obtain their biomass input from the cheapest sources. A supply curve for bioenergy is thereby generated and the implications of various levels of production for farm income, subsidy payments and for the environment are analyzed. The environmental benefit in the form of reduced carbon-dioxide emissions from co-firing biomass with coal is determined by conducting a lifecycle analysis of carbon-dioxide emissions from electricity generated by co-firing bioenergy crops as compared to that generated from coal only. The lifecycle analysis includes the soil carbon sequestered by perennial grasses and the carbon emissions displaced by these grasses due to both conversion of land from row crops and co-firing the grasses with coal. Spatial variability in land use and in soil carbon sequestration potential of land use choices, and their policy implications are discussed.

35.1 Introduction

The U.S. greenhouse gas emissions have increased by approximately 1% each year in the last decade. More than a quarter of the emissions are generated by coal-based electricity production (see [64]). Concerns about climate change have led to growing

M. Khanna (✉), H. Onal, and B. Dhungana
Department of Agricultural and Consumer Economics, University of Illinois
at Urbana-Champaign, 1301 W. Gregory Drive, Urbana, IL 61801, USA
e-mail: khanna1@uiuc.edu

M. Wander
Department of Natural Resources and Environmental Sciences, University of Illinois
at Urbana-Champaign, 1301 W. Gregory Drive, Urbana, IL 61801, USA

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_35,

interest in renewable fuels for electricity generation and many states in the U.S. have established Renewable Portfolio Standards (RPS)[1] to encourage utilities to generate a minimum percentage of their electricity from renewable sources. Moreover, consumer willingness to pay for 'green electricity' is leading to an expansion in programs offered by utilities that allow consumers to purchase some portion of their power supply from renewable sources. One such renewable fuel source is biomass from bioenergy crops such as willow, short rotation woody crops and herbaceous perennials. As compared to coal, these fuel sources reduce carbon dioxide emissions, produce virtually no sulfur dioxide emissions and contain low amounts of ash and mercury (see Tillman [54]). Moreover, compared to the traditional row crops they displace, the production of bioenergy crops requires considerably less fossil fuel energy and can result in much higher soil carbon sequestration (see McLaughlin and Walsh [41], Turhollow and Perlack [6]).[2]

Despite extensive efforts by the U.S. Department of Energy (USDOE) in the last 15 years to sponsor demonstration projects to determine the feasibility of co-firing biomass by utilities, the share of total electricity generated from biomass remains small. The Energy Information Administration projects that biomass will generate only 0.3% of the electricity generated in 2020 in the absence of any Renewable Portfolio Standard and climate policy (see [63]). The findings of studies evaluating the economic viability of co-firing willow (see Tharakan et al. [53]), wood and waste fuels (see McGowin and Wiltsee [36]), corn stover (see Hitzhusen and Abdallah [24]), woody biomass (see Nienow et al. [42]) and switchgrass (see Qin et al. [47]) with coal in a power plant suggest that substantial incentives in the form of tax credits, subsidies and emission reduction credits would be needed to make bioenergy crops competitive with coal. These studies analyze scenarios with a representative biomass production cost and a representative power plant.

The focus of this chapter is to examine the extent to which it would be profitable to allocate cropland to two bioenergy crops, switchgrass and miscanthus, for co-firing in coal-based power plants in Illinois and the spatial variability in the allocation of that land at various bioenergy prices. Our analysis recognizes that the costs of growing these bioenergy crops and their yields vary both spatially (depending on soil and weather conditions) and temporally (depending on the age of the perennial crop). The opportunity costs of using land for energy crops also vary spatially depending on the foregone profitability of alternative uses such as row crops. Furthermore, transportation costs, which constitute a significant component of the

[1] By mid 2006, 22 states and the District of Columbia had adopted RPS, that impose mandatory or voluntary goals that electricity suppliers generate a minimum percentage of their electricity from renewable sources The RPS in Illinois sets a goal of producing 10% of Illinois' electricity using renewable energy sources by 2015. (*http://www.commerce.state.il.us/dceo/News/pr08222006.htm*).

[2] Perennial cropping eliminates soil carbon losses caused by annual physical disturbance associated with annual crops (planting, cultivation, fertilizer addition) and soil erosion by keeping soils covered with vegetation throughout the year and by developing prolific root systems that stabilize soil structure (see Lewandowski et al. [31], McLaughlin et al. [40] and Paustian et al. [43]).

delivered cost of bioenergy crops, vary spatially depending on the location of fields producing the crops and the power plants to which they are delivered.

The two bioenergy crops considered here, switchgrass and miscanthus, are perennial grasses that can be grown on cropland and are being promoted by the USDOE [62]. Switchgrass was identified by the USDOE as a "model" crop due to its relatively high yields, adaptability to a wide range of growing conditions, and environmental benefits (see McLaughlin and Kszos [37]). Miscanthus (miscanthus giganteus) has been studied and grown extensively in Europe for bioenergy generation and is being grown experimentally in the US since 2002 following establishment of field trials at the University of Illinois Agricultural Research and Education Centers in 2002 (see Heaton et al. [22]). We also quantify the economically viable potential for bioenergy crops while recognizing their potential to reduce greenhouse gas emissions through sequestration of carbon in the soil and by displacement of coal. Our analysis incorporates both the spatial and temporal variability in the soil carbon accumulation process.

To incorporate the features described above, we develop a dynamic optimization model using detailed spatial data on costs of producing and delivering bioenergy crops for co-firing in existing coal-based power plants in Illinois. These costs are determined using a biophysical crop productivity model which simulates bioenergy crop yields depending on soil conditions and climate. This framework is used to examine potential changes in land allocation between bioenergy crops and row crops over a 15-year horizon from 2003 to 2017.[3] The model includes a transportation module that links power plants to bioenergy crop growing areas such that power plants obtain their biomass from the cheapest sources. We obtain a supply curve for biomass for Illinois and analyze the implications of growing bioenergy crops for farm income, subsidy payments and the environment. The second important objective of the analysis here is to determine the soil carbon sequestration levels resulting from switching some crop land to biomass production. For this we use estimates of county-level stocks of soil carbon and develop soil carbon accumulation functions under alternative land uses that incorporate saturation limits to soil carbon accumulation. We also conduct a lifecycle analysis of CO_2 emissions from electricity generated by co-firing bioenergy crops as compared to that from coal to examine the emission reduction benefits of bioenergy crops.

The chapter is organized as follows. The next section discusses the existing literature and the main contributions of this chapter to that literature. In Sect. 35.3 we present the theoretical model followed by a description of the data set in Sect. 35.4. The empirical results of the model are presented in Sect. 35.5. Finally, we discuss the conclusions and policy implications of the study in Sect. 35.6.

[3] Biomass co-firing involves combining biomass material with coal in existing coal-fired boilers. Coal-fired boilers can handle a pre-mixed combination of coal and biomass in which the biomass is combined with the coal in the feed lot and fed through an existing coal feed system. Alternatively, boilers can be retrofitted with a separate feed system for the biomass such that the biomass and coal actually mix inside the boiler (*http://www.eia.doe.gov/oiaf/analysispaper/biomass/index.html*).

35.2 Literature Review

Several studies have estimated the costs of producing switchgrass in the U.S. (see Duffy and Nanhou [10], Epplin [12]) and for miscanthus in Europe under representative conditions (see review in Khanna et al. [30]). These studies find that the production cost of switchgrass is lower than the costs of other herbaceous crops (see Hallam et al. [20], Turhollow [56] and Walsh et al. [66]) and woody crops such as willow and poplar (see Downing and Graham [8]). Cropland allocation at regional level in the U.S. for large scale production of switchgrass, willow and poplar at various farmgate prices for these crops is examined by Walsh et al. [66]. That study, however, does not consider specific end-uses of these crops, the cost of transportation to processing facilities, and the environmental implications. Graham et al. [19], develop a GIS-based model to examine the cost of delivering feedstock to ethanol facilities but do not analyze its implications for land use allocation.

This chapter makes several contributions to this emerging literature on the economics of bioenergy production. First, we develop a spatially disaggregated micro-economic framework using detailed geospatial data on crop yields, input applications and transportation costs to analyze the extent to which cropland in Illinois can be allocated to bioenergy crops. This will be done under various assumptions about the technical potential to co-fire biomass with coal and levels of subsidy for the use of bioenergy by power plants. Second we use lifecycle analysis to estimate the greenhouse gas reduction benefits from allocating land to bioenergy crops. We incorporate not only the energy consumed during production and transportation of bioenergy crops but also the energy saved by replacing row crops and the additional soil carbon sequestration achieved thereby. Since each of these components is location specific, the greenhouse gas mitigation benefits depend on where the bioenergy crops are grown. Third, our estimation of the soil carbon sequestration potential of bioenergy crops recognizes that it varies spatially (depending on the land use history and soil and climatic conditions) and temporally (depending on the amount of carbon already present in the soil) (see West et al. [69]). Moreover, there is an upper limit on the amount of carbon that can be stored in soil and the annual sequestration rate diminishes over time as the soil carbon level approaches the equilibrium level established by the land use practice applied (see Six et al. [52]). Our analysis shows the bioenergy prices needed to provide incentives to landowners to switch land from annual row crops to perennial bioenergy crops and the extent to which renewable energy subsidies would be needed to make bioenergy competitive with coal given the current coal prices.

35.3 The Model

The model developed here assumes a social planner that aims to maximize the total returns from all row crops and perennial crops while achieving specified targets for biomass production and soil carbon sequestration. The study area is divided into sub-regions where each sub-region is assumed to be represented by a single decision maker (an aggregate producer) who is endowed with the productive resources available in that region. It can be shown mathematically that the optimal choices for the social planner would coincide with the voluntary land allocation decisions made by independent representative producers under the assumptions of perfect competition and rational behavior (profit maximization) if appropriate incentives are provided to the individual producers. Such incentives can be derived from the shadow price information obtained from the model solutions as will be discussed later.

The sub-regions differ in terms of crop productivity and the profitability of alternative land uses. They also differ in their proximity and therefore the costs of transporting biomass to existing power plants. We examine the optimal allocation of land among various annual row crops with alternative management practices (rotations and tillage choices) and perennial grasses that can be used for either forage or co-fired with coal in power plants such that the discounted present value of aggregate profits over a specified time horizon is maximized. The price of bioenergy paid by all power plants is assumed to be the same and dependent on the energy content of the biomass relative to coal. Thus, the farmgate price received by bioenergy crop producers in each sub-region differs depending on the proximity to the power plant to which the crop is delivered. All input and crop output prices are assumed to be constant over time, but they may differ across sub-regions depending on their distances to major markets. Various constraints on crop rotations, land availability and ease of conversion from one use to another are included as described below. We use this framework to develop a supply curve for bioenergy crop production and to examine the spatial allocation of land for bioenergy crops. We then examine the implications of optimal land allocation for soil carbon sequestration and life-cycle carbon emissions from power plants. The soil carbon sequestration of alternative land uses depends on the existing stock of soil carbon in each sub-region, the capacity for additional carbon sequestration with each land use alternative in each sub-region, and the length of time a particular land use/practice is maintained continuously. Moreover, the costs and yields of perennials in any sub-region also vary with the age of the perennial. The model developed here explicitly accounts for all these aspects.

The indices, parameters and variables used in the algebraic model are defined, respectively, in Tables 35.1, 35.2 and 35.3 (see the Appendix). We use lower case letters and Greek letters to denote exogeneously given parameters and upper case letters to represent endogeneous variables.

The mathematical model representing the social planner's problem is as follows

$$\begin{aligned}\text{Maximize} \sum_{t=1}^{T} \beta^{t} \left[\sum_{i} \left\{ \sum_{\{j,jr:\delta_{j,jr}=1\},m} \pi r_{i,j,jr,m,t}.RO_{i,j,jr,m,t} \right.\right. \\ + \sum_{jp,a} \pi p_{i,jp,a} PA_{i,jp,a,t}\end{aligned}$$

$$-\sum_{l} tc.d_{i,l}SB_{i,l,t} - \sum_{jp,a} sc\Delta.PA_{i,jp,a,t}\Bigg\}\Bigg] + \sum_{t=T+1}^{\infty}\beta^{t}\sum_{i,j,jr,m}(\pi r_{i,j,jr,m}RO_{i,j,jr,m,t})$$

$$+\sum_{i,jp,a}\left[\sum_{a'>a}^{a'=el}\beta^{(T+a'-1)}\pi p_{i,jp,a'} + \sum_{t=T+el-a+1}^{\infty}\beta^{t}\left(\rho\sum_{a=1}^{el}\beta^{a}\pi p_{i,jp,a}\right)\right]PA_{i,jp,a,t}$$

$$-\sum_{t=T+1}^{\infty}\beta^{t}\sum_{i,l} tc.d_{i,l}SB_{i,l,t}, \quad \forall\, a', a \in A. \tag{35.1}$$

subject to

$$\sum_{i} SB_{i,l,t} \le q_{l} \quad \text{for every} \quad l, t \tag{35.2}$$

$$\sum_{l} SB_{i,l,t} \le \sum_{jp,a} y_{i,jp,a}.PA_{i,jp,a,t} \quad \forall i, t, jp \in \{\text{bioenergy crops}\} \tag{35.3}$$

$$\sum_{jr,m} RO_{i,j,jr,m,t}.\delta_{j,jr} \le r\bar{a}_{i,j\in Jr}|_{t=1} + RA_{i,j\in Jr,t-1}|_{t>1} - \Delta RA_{i,j\in Jr,t} + \sum_{a}\Delta PA_{i,j\in Jp,a-1,t} \quad \forall i, j, t \tag{35.4}$$

$$RA_{i,jr,t} = \sum_{jr,m} RO_{i,j,jr,m,t}.\delta_{j,jr} \tag{35.5}$$

$$NT_{i,a,t} = n\bar{t}_{i,a-1}|_{t=1} + NT_{i,a-1,t-1}|_{t>1} + \Delta CT_{i,t}|_{a=1} - \Delta NT_{i,a-1,t} \quad \forall i, a, t \tag{35.6}$$

$$\sum_{j,jr.\delta_{j,jr}=1,m} RO_{i,j,jr,m=1,t} = \sum_{a} NT_{i,a,t} \quad \forall i, t \tag{35.7}$$

$$\sum_{a} NT_{i,a,t} \le 0.8 \times \sum_{\{j,jr.\delta_{j,jr}=1\}m} RO_{i,j,jr,m,t} \quad \forall i, t \tag{35.8}$$

$$PA_{i.jp,a,t} = p\bar{a}_{i,jp,a-1}|_{t=1} + PA_{i,jp,a-1,t-1}|_{t>1} - \Delta PA_{i,jp,a-1,t} \quad \forall i, a > 1, jp, t \tag{35.9}$$

$$\sum_{jp} PA_{i,jp,a=1,t} = \sum_{jr}\Delta RA_{i,jr,t} \quad \forall i, t \tag{35.10}$$

$$0.90 \times \left(r\bar{a}_{i,j}|_{j\in Jr} + \sum_{a} p\bar{a}_{i,j,a}|_{j\in Jr}\right) \le RA_{i,j,t}|_{j\in Jr} + \sum_{a} PA_{i,j,a,t}|_{j\in Jr} \le 1.10 \times \left(r\bar{a}_{i,j}|_{j\in Jr} + \sum_{a} p\bar{a}_{i,j,a}|_{j\in Jr}\right) \quad \forall i, j, t \tag{35.11}$$

$$\sum_{jr} RA_{i,jr,t} + \sum_{a,jp} PA_{i,jp,a,t} = \sum_{jr} r\bar{a}_{i,jr} + \sum_{a,jp} p\bar{a}_{i,jp,a} \quad \forall i,t \tag{35.12}$$

$$RA_{i,jr,t}; RO_{i,j,jr,m,t}; PA_{i,jp,a,t}; \Delta NT_{i,a,t}; \Delta CT_{i,t}; \Delta PA_{i,jp,a,t}; \Delta RA_{i,jr,t}; \\ SB_{i,l,t} \geq 0 \; \forall i,j,a,t \tag{35.13}$$

The objective function (35.1) represents the discounted aggregate profits over a finite planning horizon of T years, which is to be maximized by allocating land across various crops, rotations and management practices. The first term in parenthesis in (35.1) represents the discounted net returns from production of both row crops and perennials over T years. The second and third terms capture the terminal value of a unit of land at the end of the planning horizon which is represented by the return to that land if it were to remain permanently in that land use in year T. The second term is the discounted returns from row crop production in perpetuity while the third term is the terminal value for the land under perennials in year T. The latter reflects the value of remaining economic life of the standing crop in year T followed by a return in perpetuity for growing that perennial on that land (see McCarl et al. [35]).

Equation (35.2) constrains the supply of biomass from all sub-regions to power plant l not to exceed the power plant's technical capacity to co-fire biomass with coal.[4] Power plants have the flexibility to acquire biomass from any sub-region. Incorporation of the biomass transportation costs in the objective function ensures that each power plant acquires its biomass input from the most economical sub-region subject to the availability of biomass in that sub-region. Equation (35.3) constrains the total supply of biomass from a sub-region to all power plants not to exceed the total production of biomass in that sub-region.

Equations (35.4) and (35.5) govern dynamic changes in the acreage of row crops resulting from conversion of land across different tillage options, row crops and perennial crops. Constraint (35.4) limits the land available for row crop j in period t based on the land planted in period $t-1$ for crops that can precede j given the allowable crop rotation possibilities, plus the land converted from perennials, minus the acreage that switches to perennials. Equation (35.5) is an accounting equation that relates the total acres of each row crop (by sub-region and period) to the rotation activities producing that crop in that year.

Equation (35.6) reflects the dynamics of total acreage under conservation till practice. It states that in each sub-region the current year's allocation of land for conservation till for each age category is equal to the previous year's land under

[4] In each period the maximum amount of biomass that power plant l can utilize is $q_l = \mu.z_l.f.\phi$ where μ is the limit on the percentage of the heat energy required by power plant that can be met by biomass; z_l is the capacity of the power plant to generate electricity in kwh, f is the amount of coal required per kwh of electricity and ϕ is the relative heat content of a unit of biomass compared to that of coal.

conservation till (that was one year younger) plus the land converted from conventional till to conservation till minus the amount of land that switches from conservation till to conventional till. Equation (35.7) is an accounting equation that relates the total land under conservation till to the rotation activities that can use this tillage option.

Equation (35.8) limits the total acreage under conservation till (in each subregion and period) to a specified maximum (here specified as 80%) of the total acreage under row crops. Analogous to equation (35.6), constraints (35.9) and (35.10) govern the dynamics of land conversion between perennials and row crops. These two equations jointly state that the current year allocation of land to each perennial crop of a given age group is equal to the previous year's land allocation to the same perennial crop (of one year younger age group) plus the acreage that switches from row crops to perennials minus the acreage that switches from perennial crops to row crops.

To prevent large scale and abrupt changes in land use, we incorporate lower and upper bounds for land allocation to each row crop and perennial crop in order to reflect farmers' inflexibility towards changing crop patterns (based on historically observed behavior). Equation (35.11) states that the allocation of land to a particular crop in a sub-region should not exceed the initial allocation of land to that crop by more than 10% or fall below 90% of it.

Equation (35.12) ensures that the total allocation of land among different land use choices should not exceed the total availability of land in the initial period. This implicitly assumes that the land availability is constant throughout the planning horizon. Finally, equation (35.13) states non-negativity conditions for the endogenous variables.

The simulation is run in annual time steps for the 15-year period, 2003–2017. By solving the model repeatedly at different prices of biomass, a supply function of biomass and land allocation to bioenergy crops and alternative crops is obtained under different assumptions about the co-firing limits on power plants.

35.4 Data

The model described above is solved using the county level data for the state of Illinois. The crop choices included four row crops (corn, soybeans, wheat, and sorghum), grown using either conventional or conservation tillage practices, and three perennial crops including pasture for forage and switchgrass and miscanthus for biomass that can be co-fired with coal. The biomass can be delivered to 24 existing coal based electricity-generating plants in Illinois.[5] Thirty-four different

[5] There are 48 primarily coal-fired power plants in Illinois. We combined power plants that are located at the same ZIP code which resulted in 24 unique power plant locations for the purpose of modeling. The corresponding ratio of switchgrass yields to miscanthus yields when the two were grown side by side at three locations in Illinois ranges from 8% to 37% (see Heaton et al. [21]).

rotation possibilities among the row crops are considered. Pasture involves a five-year rotation consisting of four years of continuous alfalfa for hay and a year of corn for silage. Switchgrass and miscanthus have low input requirements, particularly energy and fertilizers, and a tolerance for the cool temperatures in the Midwest. They can be grown on a broad range of land types using conventional farming practices. Switchgrass is assumed to have a productive life of 10 years while miscanthus has a life of 20 years, both of which are assumed to be replanted or converted to row crops beyond those times. Crop productivity models as well as field trials in Illinois indicate that miscanthus has relatively high yields, more than twice the yield of switchgrass and higher than miscanthus yields observed in Europe (see Heaton, et al. [21, 22]).

Four types of data are compiled for these crop choices for each of the 102 counties that comprise approximately 9.4 million hectares of cropland in Illinois (see USDA/NASS [61]). These include data on crop yields, rotation- and tillage-specific costs of production for row crops, age-specific costs of production for perennials, and data on location and capacity of coal-fired power plants. Each county is assumed to be a land use decision making unit with relatively homogenous production characteristics.

35.4.1 Crop Yields

The perennial grasses considered here, switchgrass and miscanthus, are suitable for growing on the Midwest farmland using conventional farming practices and have relatively low need for water and fertilizer inputs. Because of the absence of long term observed yield data, we simulate the miscanthus yield in Illinois using a process-based crop productivity simulation model, MISCANMOD, that runs on a daily time step at a 2×2 km scale (see Clifton-Brown et al. [6]). The model is applied to Illinois using long term historical data on climate, weather and soil moisture as described in Khanna et al. [30]. Simulated yields are lowest in northern Illinois and increase as one moves south. Both the pattern of yield distribution and the average simulated yield closely agree with those obtained in field experiments (see Heaton et al. [22]). For switchgrass, we use the results of field experiments in Iowa and Illinois. The average yield of switchgrass is about 25% of the average yield for miscanthus predicted by MISCANMOD.[6] We assume that this ratio holds in each county and use it to obtain the county level yields for switchgrass. Yields for corn, soybean, wheat, sorghum, and pasture for each county are set at their five year (1998–2002) historical averages obtained from NASS/USDA.

[6] The corresponding ratio of switchgrass yelds to miscanthus yelds when the two were grown side by side at three locations in Illinois ranges from 8% to 37% (see Heaton et al. [22]).

35.4.2 Crop Production Costs and Revenues

Crop budgets that itemize costs of production for each of the perennial crops and row crops for each county that vary by tillage and rotation choice are developed. Production costs include: costs of chemicals, fertilizers and seeds; costs of equipment for land preparation and harvest operations; costs of drying and crop insurance for row crops; costs of storage and transportation of biomass; and interest payments on all variable input costs. Application levels for nitrogen, phosphorus, potassium and seed under conventional till for each row crop and for alfalfa are based on application rates recommended by the University of Illinois Extension (see Schnitkey [50]). Costs of machinery include: repair and maintenance costs; fuel and lube costs; wages for hired labor; and depreciation and interest on investment (see Schnitkey [50]). Costs of fertilizers and machinery under conservation tillage differ from those under conventional tillage and were obtained from the USDA data (see Wu et al. [70]) as weighted averages of the costs of different types of conservation tillage practices.[7]

Perennial grasses typically take a year to establish, with no harvestable yield in the first year. The second year yield is about two thirds of the maximum yield for switchgrass and half of the maximum yield for miscanthus (see Ugarte et al. [58]). Beyond that, yields are assumed to remain constant at their second-year levels throughout the life of the plant. Costs and revenues of perennials are therefore age-specific. A detailed description of the assumptions underlying the determination of these costs and revenues for this study can be found in Khanna et al. [30]. We also include the cost of switching land from perennials to row crops due to the use of herbicides to control weeds (see Duffy and Nanhou [9]).[8] The costs of land, overhead (such as farm insurance and utilities), building repair and depreciation, and the farmer's own labor are not included in the costs of perennials or row crops since they are assumed to be the same for all crops and do not affect the crop choice. Transportation costs from each county to each coal-fired power plants in Illinois are calculated using the ”great circle” distance method based on geo-referenced data on location of county centers and power plants (see Sinnott [51]).[9]

[7] The CPS data shows that machinery costs under conservation tillage are approximately 23%–35% lower than those under conventional tillage for most crop-rotation choices. However, fertilizer costs are lower for some crops/rotations and higher for others. For example, these costs are 39% lower for a hay-corn rotation but 80% higher for a corn-soybean rotation with conservation till as compared to conventional till. Pesticide cost assumptions are based on personal communication with Gary Schnitkey (2004). They are also supported by Uri [60] who reports 16%–21% higher average chemical costs for corn with conservation till compared to conventional till based on the 1987 Farm Costs and Return Survey of corn farms conducted by the NASS/USDA. His econometric analysis also shows that there exists a statistically significant positive relationship between chemical costs and adoption of conservation till for corn.

[8] This requires 2 qt of Roundup^{TM}/acre at \$9.39/qt. and a machine to spray it at \$4.30/acre (in 2000 prices).

[9] Since our purpose was to obtain a proxy for transportation costs from a hypothetical field located in the county center to the power plants rather than exact distance, we did not use actual road

For expected crop prices[10] we use the county level loan rates for corn, soybean, wheat and sorghum.[11] The price of alfalfa is the average price reported for Illinois by NASS/USDA[12] and assumed to be the same across all sub-regions. For perennial grasses used for bioenergy production, we assumed that the price that a power plant would be willing to pay for biomass would depend on the cost of coal and the energy content of the biomass. The relevant data were obtained from different sources, particularly from McLaughlin [39] and USDOE/EIA [65]. We examine the effects of alternative subsidy levels for bioenergy on the production of bioenergy crops assuming that the use of biomass by power plants is constrained by the capacity of the power plant to co-fire biomass with coal without adversely affecting the thermal efficiency of the plant. We consider alternative specifications for the co-firing capacity in sensitivity runs. Experience from co-firing in Europe and the U.S. shows that 5–15% biomass (on energy basis) can be co-fired in coal plants without loss of thermal efficiency and problems of corrosion, fuel handling and fuel feeding.[13] In the sensitivity analysis we consider 5%, 15% and 25% rates but the results

distance. The *great circle* distance between two locations with (ϕ_1, λ_1) and (ϕ_2, λ_2) as their latitude and longitudes is $r\Delta\sigma$, where r is the great-circle radius of the earth's sphere which is 3,963 statute miles and $\Delta\sigma$ is as defined below with $\Delta\lambda$ representing the difference in the longitudes of the two locations:

$$\Delta\sigma=2\arctan\sqrt{\sin\{\frac{\phi_2-\phi_1}{2}^{2}\}+\cos\phi_1\cos\phi_2\sin\{\frac{\Delta\lambda}{2}\}^2/\cos\{\frac{\phi_2-\phi_1}{2}\}^2-\cos\phi_1\cos\phi_2\sin\{\frac{\Delta\lambda}{2}\}^2}.$$

[10] Studies differ in their approach to the estimation of the expected price of a crop. Just and Rausser [28] and Gardner [17] argue in favor of futures price based on rational expectations assumption. Chavas and Holt [4] assume adaptive expectations and use lagged market prices to obtain expected future prices. Chavas, Pope and Kao [5] investigated the role of future prices, lagged market prices and support prices in their econometric analysis of acreage supply response of corn and soybeans in the US. They found that government's corn support program plays a major role in corn and soybean production decision and that future prices are not good proxies for expected prices in the presence of government program. Wu and Segerson [71] use the higher of the current target price and a linear function of previous year's market price as a measure of expected price for program crops.

[11] When market prices are low, farmers can receive the difference between the price designated as the loan rate and the market price per ton sold as a direct payment from the government; these prices, therefore, serve as a price floor These loan rates are obtained from FSA/USDA for 2003 *(http://www.fsa.usda.gov/dafp/psd/LoanRate.htm).* These support prices play a major role in corn and soybean acreage decisions (see Young and Westcott [72]) and have been found to be better proxies for expected future cash prices than futures prices in the presence of government programs (see Chavas et al. [5]).

[12] See *http://nas.usda.gov/statistics_by_state/Illinois?Publications/Farm_Reports/2005/ifr0504.pdf*. Since corn silage is typically not marketed, we determine its implicit price by estimating the foregone revenue per acre by growing corn silage instead of corn and the additional cost of fertilizer replacement that is needed for corn silage, using the method in FBFM (*http://www.farmdoc.uiuc.edu*).

[13] In practice, this price could be lower if the power plant has to make investments in processing the biomass, such as converting it to pellets, before co-firing, or retrofitting equipment to co-fire

are not reported for brevity. The model assumes that the land under bioenergy crops at the end of the planning horizon will remain in that land use permanently. All costs and revenues are discounted to the beginning of the simulation horizon.

35.4.3 Soil Carbon Sequestration

Soil carbon sequestration rates are calculated by assuming a negative exponential time path for sequestration with saturation limits depending on the land use. The two land uses considered here are conservation tillage and planting perennial grasses. The annual rate of sequestration with a given land use depends on the existing stock of carbon in the soil and on the sequestration potential for that land use. We use the following non-linear functions (see INRA [25]) to determine the carbon sequestration rates achieved by switching to conservation tillage from row crops and perennial grasses:

$$rs_{i,a} = (s_i^e - s_i^0)(1 - e^{-ka}) \text{ for every } i, a$$

$$ps_{i,jp,a} = (s_{i,jp}^e - s_i^0)(1 - e^{-ka}) \text{ for every } i, jp, a$$

where the terms $rs_{i,a}$ and $ps_{i,jp,a}$ represent the cumulative amount of carbon stored by switching to conservation tillage from row crops and perennial crops of age a, respectively, s_i^0 is the initial amount of carbon stock in the in sub-region i, and s_i^e and $s_{i,jp}^e$ are the long-run equilibrium levels of carbon that can be stored in the soil in sub-region i by perennial crop j and conservation tillage, respectively. Carbon accumulation depends on each region's site specific characteristics, specifically the existing level of soil carbon, the long-run equilibrium level of soil carbon and the natural growth rate of carbon accumulation (denoted by k) (as in INRA [25]). Carbon accumulation rates obtained in this study and methods used to obtain them are described in Table 35.4 (see the Appendix). The annual sequestration rates differ across land uses and counties due to the variability in existing levels of accumulated carbon and in sequestration potential with alternative land uses.

35.4.4 Carbon-Dioxide Emission Mitigation Through Co-Firing

We estimate the CO_2 emissions per kilowatt hour (kwh) mitigated by displacing a portion of the coal used for electricity generation by biomass by including the emissions generated in the process of production and transportation of biomass,

grasses and/or if there is a loss in boiler efficiency with co-firing of grasses. Alternatively this price could be higher if the government needs to subsidize power plants for switching to renewable energy or if power plants include a value for savings due to avoided costs of SO_2 permits, NO_2 permits and CO_2 permits, since the substitution of these grasses for coal reduces these emissions.

the carbon sequestered in the soil during bioenergy crop production and the carbon emissions resulting from land use changes (such as when bioenergy crops replace row crops, or vice versa). These emissions are computed using life cycle analysis which involves estimating carbon emissions based on energy used by machinery in the production of each crop, the energy used to produce other inputs such as fertilizers and herbicides, and the energy used directly in the form of gasoline, diesel, liquefied propane gas and electricity. The estimates for a representative field in Illinois are presented in Table 35.5 (see the Appendix). The input application rates for each crop (given in the footnotes to the table) are multiplied by the CO_2 equivalent greenhouse emissions (CO_2e) generated in the production of that input (obtained from Farrell et al. [15]).

Each hectare of land converted from a corn-soybean rotation to switchgrass or miscanthus is estimated to reduce corresponding emissions from corn and soybean by 1,962 Kg CO_2e per hectare. This includes emissions of 2,867 kg CO_2 per hectare from corn and 1,056 kg CO_2 per hectare from soybeans. Farrell et al. [15] estimate emissions from corn in the 'ethanol today scenario' to be 2,703 kg CO_2e per hectare. They assume a lower fertilizer application rate but much higher use of gasoline and diesel per hectare for corn production as compared to our study. However, their estimate of energy consumed per hectare in producing corn is 18,297 million joules (MJ) while the corresponding estimate in this study is 15,641 (MJ) because of the high energy content of these fuels. The corresponding estimate obtained by Hill et al. [23] is 18,920 MJ and is also higher than that in our study for the same reason. The fuel use per hectare for corn production assumed in our study is based on the rate provided for Illinois by Shapouri et al. [49] and is lower than the national average.

We use the input application rates for switchgrass and miscanthus to estimate the energy requirements (see Khanna et al. [30]). Fossil fuel energy requirements for harvesting and post-harvesting operations for switchgrass and miscanthus are based on Elsayed et al. [11] which are based on an extensive review of European studies. The production of switchgrass and miscanthus is estimated to generate carbon emissions equivalent to 1,662 Kg CO_2e per hectare and 1,575 Kg CO_2e per hectare, respectively. Farrell et al. [15] estimate the carbon emissions from switchgrass production as 971 Kg CO_2e per hectare while Tilman et al. [55] estimate the carbon emissions from low-input high-diversity grassland biomass as 324 Kg CO_2e per hectare. Our estimate is higher than those because of two reasons. First, we assume a 15-year economic life of machinery (as in Hill et al. [23]) instead of 30 years of economic life assumed by Tilman et al. [55].[14] Second, we assume a much higher fertilizer application rate (based on (McLaughlin and Kszos [38])) and a higher yield per hectare than Tilman et al. [55].

[14] The machinery inputs used to estimate embodied energy in machinery inputs may not necessarily match the machinery inputs used for calculating crop budget because the two are based on different sources. However, we believe any discrepancy as a result of this would be minimal since embodied energy and CO_2e emissions from farm machinery use has a small share in over all energy and CO_2e budget.

Over a 20-year period, switchgrass and miscanthus are assumed to sequester 3,117 $KgCO_2e$ per hectare per year and 3,777 $KgCO_2e$ per hectare per year, respectively on average. Tilman et al. [55] assume that grassland biomass can sequester 4,033 $KgCO_2e$ per hectare per year. Our overall estimate for the carbon emissions mitigated by switchgrass is 12,899 $KgCO_2e$ per hectare. This is higher than the estimate by Tilman et al. [55] of 10,088 $KgCO_2e$ per hectare, because we also include the emissions reduced by displacing corn and soybeans from cropland converted to switchgrass. Miscanthus is estimated to mitigate 34,998 $KgCO_2e$/ha if used for electricity generation instead of coal. Thus, electricity generated using switchgrass and miscanthus reduces emissions by 1,300 $KgCO_2e$/Mwh and 1,080 $KgCO_2e$/Mwh, respectively, and results in a negative net emission in comparison to coal-based electricity which releases 964 $KgCO_2e$/Mwh.

35.5 Results

We first determined the profit maximizing land allocation in 2003 (including the land under conservation tillage and pasture) without any bioenergy subsidy, which we call the 'business as usual' (BAU) scenario. In this scenario we find that about 45% of the total cropland (9.4 million hectares) would be under conservation tillage, about 3% under pasture, and the rest under conventional tillage by the 15th year (2017). This represents 6% and 10% greater allocation to conservation tillage and pasture, respectively, than the observed allocation in 2002 in Illinois. The results are presented in the first column of Table 35.6 (see the Appendix).

Next we examined the land that would be allocated to biomass production at various bioenergy prices and with various assumptions about the technical maximum capacity of a power plant to co-fire biomass with coal. Results obtained under the assumption that the latter is 15% are reported in Table 35.7 (see the Appendix). Results with a co-firing capacity of 5% or 25% are not reported for brevity. We find that the minimum price of bioenergy needed to induce landowners to produce miscanthus is \$2.4/GJ. At this price miscanthus would be planted on 660 hectares of cropland irrespective of the potential for co-firing. If power plants pay a coal energy-equivalent price for bioenergy (which would be \$20.22 per ton of biomass or \$1.12 per GJ at the current coal price in Illinois) the minimum subsidy that would be needed to make miscanthus profitable is found to be \$1.23 / GJ. This subsidy rate, however, would result in an insignificant amount of miscanthus production (generating less than 0.01% of the electricity supply from coal).[15]

As the subsidy rate is increased biomass production increases but very inelastically. At a bioenergy price of \$2.8 per GJ, the acreage under miscanthus increases to 1.7% of the cropland and the biomass produced thereby is sufficient to generate 5%

[15] We assume that existing power plants in Illinois would supply 75% of the total name plate capacity each year. However, the actual production may vary depending on the demand for electricity each year.

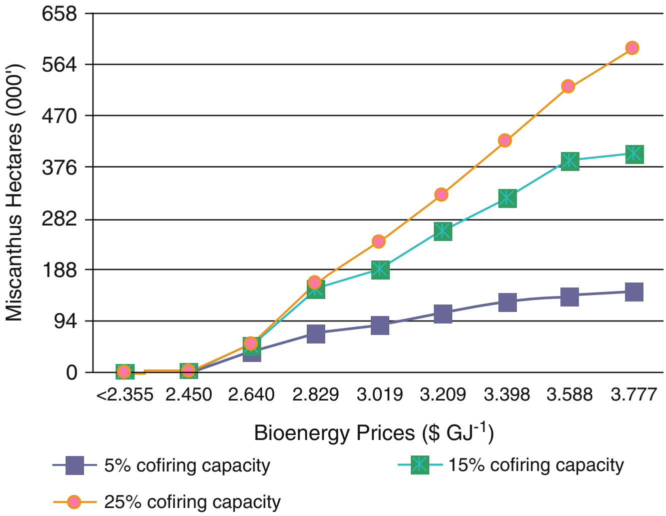


Fig. 35.1 Acreage response to bioenergy price

of the electricity produced in Illinois. As shown in Table 35.6 (see the Appendix), a 30% increase in the bioenergy price (from $2.8 / GJ to $3.6 / GJ) would expand the acreage under miscanthus by 2.5%. However, this more than doubles the share of coal-based electricity generated from biomass. Figure 35.1 shows the miscanthus acreage in response to bioenergy prices and co-firing capacity.

As the price of bioenergy is increased, we observe a reduction in the acreage of row crops with both types of tillage practices and under pasture. The land under conservation tillage decreases more than the land under conventional tillage because the former is less profitable than conventional till in many counties. Similarly, as the co-firing capacity is increased, the amount of cropland under miscanthus increases and that under conservation tillage falls.

Not all power plants find it profitable to co-fire biomass with coal to the technically maximum capacity because of the limitation on the availability of biomass at the coal energy equivalent price. A power plant can be expected to first exhaust the supply potential for biomass from the lowest cost source which may not necessarily be the closest to the power plant. Supply potential is determined by the distribution of cropland under various rotation and tillage practices and constrained by the ease with which land use can be changed across rotations and tillage practices in a given year. We find that typically power plants would obtain their biomass from multiple counties. Of all power plants, about one-third co-fire miscanthus at a level close to the maximum 15% level. These power plants are located in the southwest region where costs of production of miscanthus are relatively low. Allocation of land to miscanthus in these counties is constrained by the power plant's technical potential to co-fire. At the bioenergy price of $2.8/GJ, five power plants would not be able to co-fire any biomass with coal because of inadequate biomass supply. These power

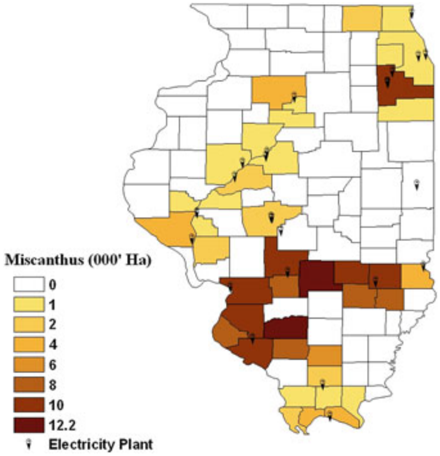


Fig. 35.2 Area of Miscanthus at $2.8/GJ and 15% co-firing limit

plants are located in the north-eastern, central and eastern regions. As the price of bioenergy increases to $3.6 / GJ, all plants are able to obtain biomass for co-firing. However, the co-firing percentage in power plants located in the north-eastern region remains less than 1%.

The spatial distribution of miscanthus production favors counties where there is a power plant in close proximity. As shown in Fig. 35.2, miscanthus production would be heavily concentrated in the southern counties located near power plants at the bioenergy price of $2.8/ GJ. Only 41 of the 102 counties in Illinois would find it profitable to produce miscanthus with the minimum share of county cropland dedicated to miscanthus being 0.2%. About one-third of the miscanthus producing counties would supply more than 70% of the total miscanthus produced in the state. This is mainly due to the low opportunity cost of producing miscanthus in these counties; primarily due to the relatively low yields per acre of corn and soybean and the high yield of miscanthus in these counties. However, even among these, only four counties find it profitable to allocate 10% of their cropland to miscanthus. Three of those counties do not have a power plant located within the county, but they would supply biomass to a power plant in the fourth county which is in close proximity. The cost advantage of even the southern counties gets rapidly eroded as the transportation cost increases. With a bioenergy price of $2.8/GJ, the maximum distance that miscanthus could be profitably transported is 35 miles while the average distance is 15 miles. The presence of power plants in central and north-eastern counties, thus lower transportation costs, makes it profitable to produce some miscanthus in those counties as well, despite the relatively low miscanthus yield and high opportunity cost of land. The delivered cost of biomass to power plants in these counties is lower for nearby counties than it is for counties in southern Illinois. As Fig. 35.3 shows, increases in the price of bioenergy increase the area under miscanthus in counties near power plants. Moreover, biomass could now be delivered to power plants located further away; with a bioenergy price of $3.6/GJ the maximum distance to which bioenergy is delivered increases to 73 miles.

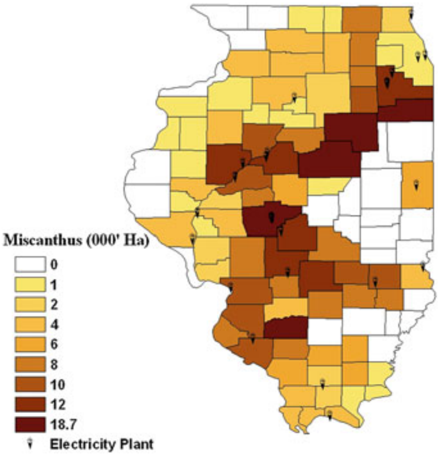


Fig. 35.3 Area of Miscanthus at $3.6/GJ and 15% Co-firing Limit

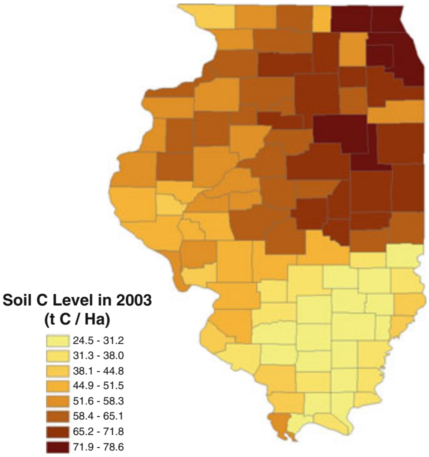
An increase in the co-firing capacity to 25% would have a modest impact on the maximum distance biomass is transported and the number of counties that produce miscanthus. Only three power plants would co-fire biomass at their 25% capacity if the bioenergy price is $2.8 per GJ. At the price of $3.6 per GJ, 6% of the cropland would be allocated to miscanthus and the share of bioenergy based electricity would be 18% of the total electricity generated.

The present value of the subsidy payment needed over 15 years to induce 1.7% of the cropland to switch to miscanthus production a bioenergy price of $2.4 / GJ and a 15% co-firing limit is $1074 Million. The subsidy amount increases more than three-folds as the bioenergy price increases to $3.6 per GJ and biomass production increases by 2.4 times. The provision of a subsidy for bioenergy, with all other crop prices fixed, reduces the profitability of row crops and increases the profits from miscanthus production. Since this shift in production only occurs because it is profitable to do so, the discounted present value of farm profits increase. A subsidy of $1.2 per GJ which raises the bioenergy price to $2.8/GJ increases the discounted value of farm profits by $218 Million but costs $1074 Million to the taxpayers and generates a deadweight loss of $855 Million. This deadweight loss increases to $1.7 Billion if the subsidy-driven bioenergy price increases to $3.6/GJ.

35.5.1 Greenhouse Gas Mitigation

We quantify the greenhouse gas mitigation achieved by producing and co-firing bioenergy crops through soil carbon sequestration, displacement of coal and displacement of row crops on cropland converted to miscanthus. Carbon accumulated on land previously under conservation tillage and pasture is assumed to be released back to the atmosphere if this land switches to miscanthus.

Fig. 35.4 Soil carbon level in 2003 under BAU scenario



Under the BAU scenario, the aggregate carbon stock is estimated to be 16 Million metric tons (Mt) in 2003 and to increase to 33 Mt by 2017. There is wide variation in existing carbon stock across counties, ranging from 24.5 tons to 78.6 tons per hectare. Carbon stocks are typically higher in the northeastern and central regions of Illinois (see Fig. 35.4).[16] This spatial distribution of carbon stocks is similar to the pattern estimated by Alexander and Darmody [1] for Illinois for 1991 (see Fig. 35.4). Carbon stock increases by 17 Mt through sequestration; 93% of this is achieved by land acres under conservation tillage and the rest by land under pasture. This would mitigate 4.3% of the expected carbon emissions by coal-fired power plants in Illinois over the period 2003-2017 (see Table 35.6 in the Appendix).

As the bioenergy price increases to \$2.8/GJ and land is converted to miscanthus, carbon accumulation increases to 18 Mt over the 15 year period. However, 82% of this is achieved by acres under conservation tillage and 12% by acres under miscanthus. As shown in Fig. 35.5 large percentages of central Illinois continue to choose conservation tillage even with a subsidy to miscanthus. As bioenergy price increases further to \$3.6/GJ, the share of sequestration by conservation tillage and miscanthus is 62% and 33%, respectively, of the total 21 Mt sequestered over the 15-year period (see Fig. 35.6). There are two reasons for this decline in share of the soil carbon sequestered by conservation tillage and pasture as the bioenergy price increases. First, it increases the land area that switches from conservation tillage and pasture to miscanthus which results in a net loss of soil carbon relative to the BAU level on these acres. The second reason is that the land that is converted from conservation tillage to miscanthus acres sequesters more soil carbon per hectare. This increased

[16] In Fig. 35.4, we computed the area weighted average of gain in soil carbon level under conservation till and pasture land at the end of the first year of the simulation run, 2003, under the BAU scenario. We normalize the conservation till and pasture acreage in each county by their sum (i.e. conservation till + pasture). The net weighted gain in soil carbon is then added to the adjusted 1991 Carbon level reported in Alexander and Darmody [1].

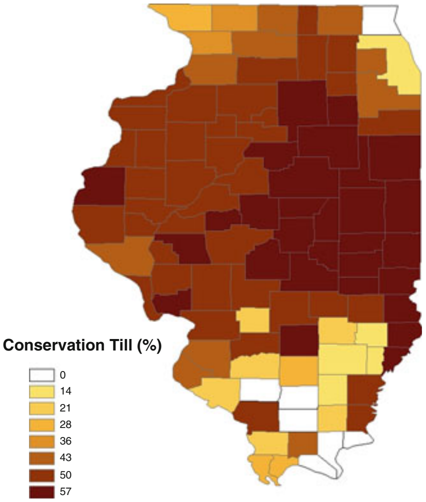


Fig. 35.5 Share of conservation till acreage relative to county cropland with 15% co-firing limit and at $2.8/GJ

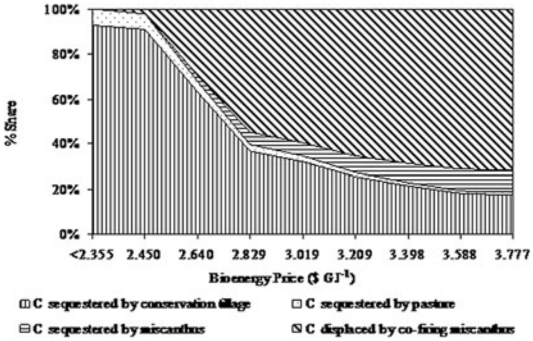


Fig. 35.6 Contribution to alternative approaches to carbon mitigation with 15% co-firing limit

rate of sequestration on acres under miscanthus more than compensates for the initial loss of soil carbon that occurs due to switching of land from conservation tillage and pasture to miscanthus.

Miscanthus not only accumulates carbon in soil, it also displaces carbon (above the ground) by replacing coal in the power plant and by replacing carbon intensive corn production on land. After accounting for these displacement effects, we find that miscanthus mitigates above ground carbon by 21 MT over the 15 year period. The total amount of greenhouse gas mitigation due to sequestration and displacement is 39 MT; this is equivalent to an 11% reduction in the total greenhouse gas emissions by power plants over 2003-2017 period. Thus, 54% of the mitigation is due to displacement of coal and conversion of land use from row crops to miscanthus and 46% is due to soil sequestration at the bioenergy price of $2.8 per GJ. As the bioenergy price increases to 3.6/GJ, the total carbon reduction achieved is 72 Mt (a 20% reduction in carbon emissions by power plants), 71% of this is the result of replacing coal with miscanthus and 28% is due to sequestration, the bulk of which is due to conservation tillage. Thus soil carbon sequestration by miscanthus is a relatively small fraction of the total mitigation benefits it provides. Figure 35.6 shows that as the bioenergy price increases, the share of mitigation benefits provided by soil carbon sequestration through conservation tillage declines steeply and the share of mitigation benefits through displacement of coal increases commensurately. Soil carbon sequestration by miscanthus and pasture provide only a small share of the overall mitigation benefits and these decline over time as the soil gets saturated with carbon.

35.5.2 Sensitivity Analysis

We examine the sensitivity of our results to various assumptions underlying our numerical model. In particular, we consider the impact of increasing (i) row crop yields, (ii) row crop prices (iii) the discount rate, (iv) the ease of conversion of land to biomass production, (v) biomass crop yield, (vi) production cost of biomass crops. In each case, we keep all other assumptions the same as in the case of 15% co-firing capacity with bioenergy price at $2.8/GJ (see Table 35.6 in the Appendix). We find that acreage planted under miscanthus is sensitive to assumptions about row crop yields, row crop prices, discount rate and costs of producing biomass crops. A 10% increase in row crop yields or a 50% increase in row crop prices, a 25% increase in biomass crop production costs and a doubling of the discount rate would reduce the cropland share of miscanthus from 1.65% to less than 1% and in some cases make it close to zero. The reduction in land under miscanthus in these alternative scenarios leads to a corresponding increase in land under conventional tillage while the share of cropland under conservation tillage does not change much. The maximum distance that miscanthus is transported remains in the 20–35 miles range across various scenarios, although the number of power plants that co-fire biomass changes considerably across these scenarios.

Increasing the ease of conversion of land to biomass crops or increasing biomass crop yields by 10% does not have a large impact on share of cropland under miscanthus. It does, however, increase the number of power plants that co-fire biomass and the share of electricity generated from biomass in Illinois. We find that our results are not very sensitive to changes in the following factors: a 10% increase in miscanthus yield, a 25% decrease in transportation cost of biomass, and a 2% reduction in thermal efficiency of power plant boilers. Miscanthus acreage response to each of these factors is in the range of 0.6% to 8%.

35.6 Conclusions

This chapter analyzes the cost of supplying bioenergy for co-firing with coal in existing coal-fired power plants in Illinois using a dynamic, linear programming framework. It takes into account the location specific returns from growing traditional row crops and perennial grasses, and the cost of transporting and storing biomass and identifies the optimal mix of land use with 5%, 15% and 25% co-firing capacity of power plants. The framework developed here accounts for the inter-temporal feedback effects of crop and tillage choices and keeps track of their age structure to compute the site- and age- specific sequestration contribution of each of the alternative land use choices. We assess the spatially and temporally varying life cycle carbon emission reduction benefit resulting from the process of soil sequestration, production, transportation and co-firing of perennial grasses at the county level while estimating overall carbon reduction benefits from land use choices. When the per ton cost of producing biomass is much higher than the power plants' willingness to pay coal-equivalent price to farmers, the magnitude of biomass production depends on the amount of bioenergy subsidy provided either to power plants or to farmers.

Our main findings are as follows: At the current coal-equivalent energy price, a relatively large bioenergy subsidy would be needed to make it profitable for farmers to grow miscanthus. Decisions about allocation of land to miscanthus are strongly influenced by: the location of production sources relative to power plants (due to transportation costs), the capacity of power plants to co-fire biomass and the price of bioenergy. At a price of $2.8/GJ, it is profitable to grow miscanthus on only 1.65% of 9.4 million hectares of cropland (with 15% co-firing limit). Miscanthus production occurs in one-third of counties with more than 70% of production concentrated in southern counties and counties located within a 35 miles radius from the existing power plants. These counties differ in their cropland allocation, which ranges between 0.19% and 10%, and in their share of miscanthus, which ranges between less than 1% to almost 8% of the total biomass supply. About four-fifth of power plants would utilize biomass in the ranges between 0.07% and 15% of their production capacity, replacing 5.5% of the total electricity supply by co-fired power plants in Illinois.

A subsidy payment of $1,074 million is needed to replace this 5.5% of the coal-based electricity by bioenergy. This payment can be designed either to pay farmers in the form of a miscanthus supply subsidy or to pay power plants in the form of a bioenergy subsidy, or some combination of both. This subsidy, if transferred to farmers, would increase annual farm profits by $57/ ha on cropland allocated to miscanthus production. The carbon mitigation benefits of the land use changes considered here are substantial. Conservation tillage and perennials together could mitigate 10.6% of the cumulative carbon emissions by power plants over the 15 years period, 2003–2017 in response to the bioenergy price of $2.8/GJ. Of the total carbon mitigated, 55% is due to the displacement of coal and the rest is due to soil carbon sequestration.

Our results have several policy implications. They show that with low coal prices the market incentive to divert land from conventional row crops to biomass crops in Illinois is virtually non-existent. Large bioenergy subsidies per unit of energy either to power plants or to farmers would be needed to encourage them to switch even as little as less than 2% of cropland to bioenergy crops which could produce less than 6% of the electricity generated by coal-fired units in Illinois. Both the amount of land allocated to bioenergy crops and the subsidy needed are sensitive to the assumption about the constraint on power plants' technical capacity to co-fire.

The bioenergy subsidy paid to replace 5.5% of coal-based electricity by bioenergy implies a discounted carbon payment of $45 per ton of carbon which is equivalent to a $12.3 per ton CO_2e price. This is comparable to the historically observed CO_2 price traded in the European Climate Exchange. Thus, valuation not only of the energy content of biomass crops but also of their greenhouse gas mitigation benefits is critical to make bioenergy competitive with coal.

Appendix: Tables

Note on Tables 35.1, 35.2 and 35.3. We use lower case letters and Greek letters to denote exogeneously given parameters and upper case letters to represent endogeneous variables.

Table 35.1 Indices used in the mathematical model of Sect. 35.3

Indices	Definition
$Jr = \{jr\}$	set of row crops
$Jp = \{jp\}$	set of perennials
$J = \{j\} = Jr \cup Jp$	set of all crops
$T = \{t\}$	set of time periods
$I = \{i\}$	set of sub-regions
$m = \{1, 2\}$	set of tillage practices, 1 = conservation till, 2 = conventional till
$A = \{a\}$	set of ages of perennials and conservation tillage
$L = \{l\}$	set of power plants

Table 35.2 Parameters used in the mathematical model of Sect. 35.3

Parameters	Definition
$y_{i,jr}$	Yield of row crop *jr* in sub-region *i* in metric tons
$y_{i,jp,a}$	Yield of perennial crop *jp* of age *a* in sub-region *i* in metric tons
el	Economic life of a perennial crop
$\pi r_{i,j,jr,m,t}$	Profit per unit acreage from rowcrop *jr* followed after crop *j* in sub-region *i* with practice *m* in year *t*
$\pi P_{i,jp,a}$	Profit per unit acreage from perennial crop *jp* of age *a* in sub-regions *i*
$d_{i,l}$	Distance between region *i* and power plant *l* in kilometers
tc	Transportation cost per unit quantity and per unit distance
sc	Per hectar cost of switching from perennials to row crops
$\beta = \frac{1}{(1+\rho)}$	Discount factor where ρ is the discount rate
$\delta_{j,jr}$	Binary parameter, 1 if crop *j* is followed by row crop *jr*, 0 otherwise
$r\bar{a}_{i,jr}$	Initial acreage under rowcrop *jr* in sub-region *i*
$n\bar{t}_{i,a}$	Initial acreage under conservation till of age *a* in sub-region *i*
$p\bar{a}_{i,jp,a}$	Initial acreage under perennial crop *jp* of age *a* in sub-region *i*
q_l	Demand for biomass by power plant *l*

Table 35.3 Variables used in the mathematical model of Sect. 35.3

Variables	Definition
$RO_{i,j,jr,m,t}$	Acreage under row crop *jr* following crop *j* in sub-region *i* with practice *m* in year *t*
$RA_{i,jr,t}$	Total acreage rowcrop *jr* in sub-region *i* in year *t*
$NT_{i,a,t}$	Acreage under conservation till of age *a* in sub-region *i* in period *t*
$\Delta NT_{i,a,t}$	Acreage under conservation till of age *a* switching back to conventional till in sub-region *i* in period *t*
$\Delta CT_{i,t}$	Acreage converted from conservation till to conventional till in sub-region *i* in period *t* (for row crops)
$PA_{i,jp,a,t}$	Acreage under perennial crop *jp* of age *a* in sub-region *i* in period *t*
$\Delta PA_{i,jp,a,t}$	Acreage under perennial crop *jp* of age *a* switching to row crops in sub-region *i* in period *t*
$\Delta RA_{i,jr,t}$	Acreage converted from rowcrop *jr* to perennials in sub-region *i* in period *t*
$SB_{i,l,t}$	Amount of biomass shipped from sub-region *i* to plant *l* in period *t*

Table 35.4 Carbon sequestration rates

Land Use	This Study[1] (t C /ha in 20 years)	Previous studies (t C/ ha in 20 years)	References
Conservation till	3.46–10.43	5.93–9.88	Wander, Bidart-Bouzat and Aref [67]; Dick et al. [7]; Robertson, Paul and Harwood [48]
Pasture	5.19–15.64	7.91–24.71	Robertson et al. [48]; Eve et al. [13], Puget et al. [46]
Switchgrass	7.93–23.99	13.84–22.24	Gebhart et al. [18]; McLaughlin et al. [38]
Miscanthus	9.69–29.21	18.78–27.68	Beuch et al. [2]; Kahle et al. [29]; Matthews and Grogan [34]

[1] This is the range of estimates obtained across the different counties in Illinois.

We obtained estimates for the percentage of soil organic matter (SOM) for major soil series and the percentage of total county land in that soil series in each county in Illinois from Alexander and Darmody [1]. Data on total cropland acres in each county were obtained from USDA's and multiplied by the percentage of SOM in each soil series to obtain the acres of land in each soil series in each county. We assigned soil organic matter (SOM) and acreage in each soil series in descending order to the cropland acres in each county, assuming that land with the highest SOM were more likely to be in agricultural use. We then computed a weighted average of the percentage of SOM in the cropland obtained from USDA's NASS database for each county with the weights being the share of county cropland in each soil series and obtained the average soil organic carbon (in metric tons per hectare) in each county, using the method in Bowman and Peterson [3]. This method assumes that there is 0.52% of soil carbon in each 1% of SOM and that there is 2.24 million kg of surface soil (to a depth of 30 cm) per hectare. We update the county-specific carbon stock from the level in 1991 estimated by Alexander and Darmody [1] to the level at the end of 2002 by assuming that carbon accumulation only occurred on land that was under conservation tillage and pasture between 1991-2002. We obtained data on acreage under pasture and under conservation till from Conservation Tillage Information Center (http://ctic.purdue.edu). We assumed that the smallest acreage under conservation till and pasture in each county in each year between 1992 and 2002 has been in that land use/practice for the entire period 1992–2002. Thus, it is only on this land that carbon stocks will be likely to have changed since 1992. The remaining land area in conservation tillage in 2002 is allocated equi-proportionately to each of the ages 1 through 9. A similar exercise is conducted to assign duration of time to land under pasture over this period.

The theoretical maximum capacity of the soil to accumulate carbon is determined by assuming that the stock of adjusted soil carbon at the start of 2003 is 60% of the theoretical maximum capacity. Several studies suggest that 40% of soil carbon might have been lost on currently farmed agricultural land during the last century (Flach et al. [16]; Mann [33]; Paustian et al. [44]; Unger [59]). Following (Paustian et al. [44]; Six et al. [52]), we assume that conservation-till and pasture can achieve 70% and 75% of the maximum capacity, respectively, and that switchgrass and miscanthus achieve 83% and 88% of the maximum capacity, respectively. We determine annual sequestration rates for each land use by assuming that carbon accumulation occurs in a non-linear manner with rapid increase in soil carbon in the first 10 years and then a gradual leveling off (Ismail et al. [27]; Liu et al. [32]; Prueger et al. [45]; West et al. [68]). Finally, we assume that discontinuation of a particular land use results in a loss of all the carbon accumulated by that land use over time.

Table 35.5 Balance sheet of representative CO_2e emissions by co-fired electricity generation

Sources of emissions and sinks	Unit	Switchgrass	Miscanthus
Carbon emissions during production of energy crops (a)[1]	KgCO_2e /ha/yr	1662	1575
Carbon sequestration by energy crops (b)[2]	KgCO_2e /ha/yr	3117	3777
Carbon emissions displaced by energy crops replacing corn-soybeans (c)[3]	KgCO_2e /ha/yr	1962	1962
Carbon emissions displaced by energy crops replacing coal (d)[4]	KgCO_2e /ha/yr	9482	30834
Net mitigation (sink) by energy crop production ($e = b - a + c + d$)	KgCO_2e /ha/yr	12899	34998
Net reduction of carbon per ton of energy crop ($f = e$/yield)[5]	KgCO_2e/ t DM	2164	1806
Net reduction of carbon per kilowatt hour	KgCO_2e/ Mwh	1300	1080

Coments to Table 35.5.

[1] We assume zero nitrogen, 34 kg P2O5/ha and 45 kg K2O/ha application rates in the first year of switchgrass establishment. In subsequent years, 112 kg N/ha, 0.17 kg P_2O_5 /t DM and 0.72 kg K_2O/t DM are applied. Additionally, 3.5 L/ha of Atrazine and 1.75 L/ha of 2,4-D herbicides are required to control weeds in the first two years of switchgrass establishment and probability of reseeding switchgrass in the second year is expected to be 25% (see Duffy and Nanhou [10]). Lime application rate is based on Turhollow [56]. For miscanthus we assume 60 Kg N/ha for rhizome development and 50 kg N/ha in subsequent years to maintain soil fertility; 0.3 and 0.8 kg/t DM, respectively of P2O5 and K2O (Lewandowski et al. [31]). Same amount of herbicides and lime per hectare are applied for miscanthus as for switchgrass in the first year and no herbicides are used in subsequent years. Fossil fuel energy required per hectare of biomass production is 118 MJ for land preparation, 109 MJ for planting, 125 MJ each for fertilizer and herbicide applications, 52.8 MJ / t DM for cutting, swathing, baling, loading, carting and transferring and 42.9 MJ / t DM for handling of biomass (see Elsayed, Mathews and Mortimer [11]). For nitrogen fertilizer, emissions rates account for not only CO_2 emissions from energy used to produce fertilizers but also N_2O emissions from nitrification and denitrification process in soil. Similarly, emissions from lime use include those due to energy required for the production of lime and the carbon released due to the soil reaction with lime. CO_2e emissions associated with these inputs are estimated by aggregating the major greenhouse gases emitted namely carbon dioxide (CO_2), methane (CH_4), and nitrous oxide (N_2O) using their 100-year global warming potential factors. These are 1 for CO_2, 23 for CH_4, and 296 for N_2O (see IPCC [26]). Rate of energy per unit of input and CO_2e emissions are obtained from Farrell et al. [15]. However, we did not account for carbon emission from seed input and packaging across all crops for lack of reliable, comparable data.

Assumptions about the farm machinery and equipment needed for corn and soybean are based on Hill et al. [23] and for switchgrass and miscanthus production are based on Tilman et al. [55]. We assume that the tractor size used for bioenergy crops (reported in Tilman et al. [55]) can also be used for row crops. Weight of potato planter for miscanthus is obtained from *http://www.jjbroch.com/patata/i_pinza.htm*. Each piece of machinery and equipment is assumed to entirely consist of steel for the purpose of calculating its embodied energy. It is assumed that 25 MJ energy is needed to produce each kilogram of steel with an additional 50% energy for assembly (see Hill et al. [23], Tilman et al. [55] and citations there in). All machinery items are assumed to have a 15 year life (see Hill et al. [23]). Average size of farm is assumed to 151.36 ha (=374 acres) for purposes of calculating the per hectare energy use for machinery and equipment.

[2] This is based on the assumption of 0.85 t C / ha / yr for switchgrass and 1.03 metric ton C / ha / yr for miscanthus. These rates represent the 20 year average estimated across counties in Illinois.

[3] We assume 159.13 kg N, 70.60 kg P, 44.83 kg K and 1120.63 kg lime/ha application rates for corn and 47.07 kg P, 72.84 kg K and 1120.63 kg lime / ha application rate for soybeans for representative farm in Illinois for 2003 (Schnitkey [50]). Since

these rates are yield dependent, they vary by county. We assume 855.48 MJ and 87.21 MJ / ha energy equivalent herbicides and insecticides, respectively, for corn. The herbicide rate is based on a weighted average of active ingredients Atrazin, Cyanazani, Metolachlor and Acetochlor based on their share of use in Illinois as reported in Wang et al. [49]. For soybeans, we used U.S. average application rate of 1.12 liter/ha estimated by West and Marland [68]. Rates of gasoline, diesel, LPG, and electricity required per hectare of corn are as in Shapouri et al. [49]. For soybean production, diesel requirement of 34.61 liter/ha is obtained from FBFM (2005) and other fuel requirements were unavailable for Illinois, thus the national average application rates reported in Hill et al. [23] were used as proxy values.

[4] 96 t DM/ha for switchgrass, 19.38 t DM / ha for miscanthus. We account for harvesting loss of 20% for switchgrass and 33% for miscanthus for December harvest. We also assume 7% loss of biomass during storage. Both switchgrass and miscanthus feedstock are assumed to contain 15% moisture at the time of transportation and storage (see Khanna et al. [30] and reference there in).

[5] The CO_2 emissions displaced by each metric ton of biomass are calculated based on the net emissions per unit of coal-based electricity in Illinois. The average emissions rate for the 48 coal fired power plants in Illinois is 964.48 kg CO_2 / MWh and each metric ton of biomass is assumed to generate 1.65 MWh (18 GJ heat input with 33% thermal conversion efficiency). Biomass therefore results in a reduction of 1,591 kg CO_2e per ton of DM. 5It is assumed that representative annualized yield is 358 bu/ha for corn, 124 bu/ha for soybean.

Table 35.6 Response of change in biomass prices to land uses and environment

Biomass co-firing capacity (%)	BAU	15% co-firing capacity		
Bioenergy price ($/ GJ)	<$2.4	$2.8	$3.2	$3.6
Land under conservation till (%)	45.07	44.09	43.17	42.03
Land under miscanthus (%)	0.00	1.65	2.78	4.15
Biomass supply (Mt with 15% moisture)	0.00	4.24	7.02	10.18
Electricity generated with biomass (%)	0.00	5.53	9.16	13.27
Average distance to power plants from counties producing miscanthus (miles)	0.00	15.11	20.81	28.16
Total amount of carbon mitigated in 15 Years (Mt)	15.85	38.86	54.12	71.64
-coal displacement by biomass	0.00	21.29	35.27	51.05
-sequestration by miscanthus	0.00	2.05	3.97	6.65
-sequestration by conservation till	14.72	14.37	13.82	12.98
-sequestration by pasture	1.13	1.15	1.06	0.96
% of carbon emission mitigated in 15 years	4.32	10.59	14.75	19.53
Discounted present value of bioenergy subsidy ($M)	0.00	1074.22	2173.00	3721.10
Discounted NPV of farm profit ($M)	48100.47	48319.13	49038.64	50171.91

Note: Baseline annual carbon (not CO_2) emissions from coal-fired power plants are 24.46 million metric tons. We denote Mt for million metric tons and GJ for Giga Joule and $B for billion dollars.

Table 35.7 Sensitivity Analysis with 15% Co-firing Capacity and Bioenergy Price at $2.8 /GJ

Parameters	Increase in row crop yield by 10%	50% Increase in crop price	Doubling discount rate from 4% to 8%	Change in land flexibility constraint ±5%	Increase in miscanthus yield by 10%	Increase in production cost of miscanthus by 25%	25% Increase in biomass harvesting cost	25% Increase in hauling cost
Land under Conservation Till (%)	47.98	44.01	44.45	45.75	44.06	45.07	45.07	44.09
Land under miscanthus (%)	0.99	0.01	0.94	1.84	1.66	0.00	0.01	1.32
Biomass Supply (Mt with 15% moisture)	2.86	0.04	2.47	4.72	5.18	0.00	0.02	3.73
Electricity generated with biomass (%)	3.39	0.05	3.22	6.16	6.12	0.00	0.03	4.42
Average hauling distance (miles)	13.50	13.83	13.10	13.88	15.04	0.00	7.55	13.72
Total amount of C mitigated (Mt)	31.29	14.23	31.29	42.81	41.15	15.91	16.01	34.19
-coal displacement by biomass	13.04	0.20	13.04	23.71	23.54	0.00	0.10	17.01
-sequestration by miscanthus	1.22	0.01	1.22	2.37	2.15	0.00	0.01	1.65
-sequestration by conservation till	15.89	13.15	15.89	15.39	14.32	14.77	14.78	14.39
-sequestration by pasture	1.14	0.86	1.14	1.35	1.14	1.13	1.13	1.13
% of carbon mitigated at the 15th years	8.53	3.88	8.53	11.67	11.22	4.34	4.36	9.32
Discounted present value of bioenergy subsidy ($M)	657.55	10.12	0.00	1194.76	1188.84	0.00	4.96	857.28

Acknowledgements This research was completed while Basanta Dhungana was a Ph.D student in the Dept. of Agricultural and Consumer Economics, University of Illinois at Urbana-Champaign. We thank John Clifton-Brown and Lyubov Kurkalova for valuable input in this research. Funding from the Illinois Council on Food and Agricultural Research and the Dudley Smith Initiative is gratefully acknowledged.

References

1. Alexander, J.D., Darmody, R.G..: Extent and Organic Matter Content of Soils in Illinois Soil Associations and Counties. University of Illinois at Urbana-Champaign (1991)
2. Beuch, S., Boelcke, B., Belau, L.: Effect of Organic Residues of Miscanthus X Giganteus on the Soil Organic Matter Level of Arable Soils. J. Agron. Crop Sci. **183**, 111–119 (2000)
3. Bowman, R.A., Peterson, M.: Soil Organic Matter Levels in the Central Great Plains. USDA-ARS and NRCS (1997)
4. Chavas, J.-P., Holt, M.T.: Acreage Decisions under Risk: The Case of Corn and Soybeans. Am. J. Agric. Econ. **72**(3), 529–538 (1990)
5. Chavas, J.-P., Pope, R.D., Kao, R.S.: An Analysis of the Role of Future Prices, Cash Prices and Government Programs in Acreage Response. West. J. Agric. Econ. **8**(1), 27–33 (1983)
6. Clifton-Brown, J.C., Stampfl, P.F., Jones, M.B.: Miscanthus Biomass Production for Energy in Europe and Its Potential Contribution to Decreasing Fossil Fuel Carbon Emissions. Glob. Chang. Biol. **10**(4), 509–518 (2004)
7. Dick, W.A., Belvins, R.L., Frye, W.W., Peters, S.E., Christenson, D.R., Pierce, F.J., Vitosh, M.L.: Impacts of Agricultural Management Practices on C Sequestration in Forest – Derived Soils of the Eastern Corn Belt. Soil Tillage Res. **47**(3–4), 235–244 (1998)
8. Downing, M., Graham, R.L.: The Potential Supply and Cost of Biomass from Energy Crops in the Tennessee Valley Authority Region. Biomass Bioenergy **11**(4), 283–303 (1996)
9. Duffy, M.D., Nanhou, V.Y.: Costs of Producing Switchgrass for Biomass in Iowa, PM 1866 Revised April Edition. Ames, Iowa State University, University Extension (2001)
10. Duffy, M.D., Nanhou, V.Y.: In: Janick, J., Whipkey, A. (eds.) Costs of Producing Switchgrass for Biomass in Southern Iowa, pp. 267–275. ASHS Press, Alexandria, VA (2002)
11. Elsayed, M.A., Matthews, R., Mortimer, N.D.: Carbon and Energy Balances for a Range of Biofuels Options. Energy Technology Support Unit (2003)
12. Epplin, F.M.: Cost to Produce and Deliver Switchgrass Biomass to an Ethanol Conversion Facility in the Southern Plains of the United States. Biomass and Bioenergy **11**(6), 459–467 (1996)
13. Eve, M.D., Sperow, M., Howerton, K., Paustian, K., Follett, R.F.: Predicted Impact of Management Changes on Soil Carbon Storage for Each Cropland Region of the Conterminous United States. J. Soil Water Conserv. **58**, 196–204 (2002)
14. FBFM, 2005. Fuel cost by tillage type [Online]. Urbana: UI Extension, University of Illinois at Urbana-Champaign. Available: http://www.farmdoc.uiuc.edu/manage/newsletters/fefo06_07/fefo06_07.pdf [Accessed].
15. Farrell, A.E., Plevin, R.J., Turner, B.T., Jones, A.D., O'Hare, M., Kammen, D.M.: Ethanol Can Contribute to Energy and Environmental Goals. Science **311**(27), 506–509 (2006)
16. Flach, K.W., Barnwell, T.O., Crosson, P.: In: Paul, E.A., et al. (eds.) Impacts of Agriculture on Atmospheric Carbon Dioxide. CRC, Boca Raton, FL (1997)
17. Gardner, B.L.: Futures prices in supply analysis. Am. J. Agric. Econ. **58**(1), 81–84 (1976)
18. Gebhart, D.L., Johnson, H.B., Mayeux, H.S., Polley, H.W.: The Crop Increases Soil Organic Carbon. J. Soil Water Conserv. **49**, 488–492 (1994)
19. Graham, R.L., English, B.C., Noon, C.E.: A Geographic Information System-Based Modeling System for Evaluating the Cost of Delivered Energy Crop Feedstock. Biomass and BioEnergy **18**(4), 309–329 (2000)

20. Hallam, A., Anderson, I.C., Buxton, D.R.: Comparative economic analysis of perennial, annual, and intercrops for biomass production. Biomass Bioenergy **21**, 407–424 (2001)
21. Heaton, E.A., Clifton-Brown, J., Voigt, T., Jones, M.B., Long, S.P.: Miscanthus for renewable energy generation: European union experience and projections for Illinois. Mitigation Adapt. Strateg. Glob. Change **9**, 433–451 (2004)
22. Heaton, E.A., Voigt, T.B., Long, S.P.: Miscanthus and Switchgrass Trials in Illinois. Urbana, Department of Plant Biology, University of Illinois at Urbana Champaign, pp. 1–37. (2006)
23. Hill, J., Nelson, E., Tillman, D., Polasky, S., Tiffany, D.: From the cover: environmental, economic, and energetic costs and benefits of biodiesel and ethanol biofuels. Proc. Natl. Acad. Sci. USA **103**, 11206–11210 (2006)
24. Hitzhusen, F.J., Abdallah, M.: Economics of electrical energy from crop residue combustion with high sulfure coal. Am. J. Agric. Econ. **62**(3), 416–425 (1980)
25. INRA: Mitigation of the Greenhouse Effect. Increasing Carbon Stocks in French Agricultural Soils. French Institute for Agricultural Research (INRA) (2002)
26. IPCC Climate Change 2001: In: Houghton, J.T., et al. (eds.) The Scientific Basis. Cambridge University Press, New York (2001)
27. Ismail, I., Blevins, R.L., Frye, W.W.: Long-term no-tillage effects on soil properties and continuous corn yields. Soil Sci. Soc. Am. J. **58**, 193–198 (1994)
28. Just, R.E., Rausser, G.C.: Commodity price forecasting with large-scale econometric models and the futures market. Am. J. Agric. Econ. **63**(2), 197–208 (1981)
29. Kahle, P., Beuch, S., Boelcke, B., Leinweber, P., Schulten, H.R.: Cropping of miscanthus in central Europe: biomass production and influence on nutrients and soil organic matter. Eur. J. Agron. **15**, 171–184 (2001)
30. Khanna, M., Dhungana, B., Clifton-Brown, J.: Costs of producing switchgrass and miscanthus for bioenergy in Illinois. Biomass Bioenergy **32**(6), 482–493
31. Lewandowski, I., Clifton-Brown, J.C., Andersson, B., Basch, G., Christian, D.G., Jrgensen, U., Jones, M.B., Riche, A.B., Schwarz, K.U., Tayebi, K., Teixeira, F.: Environment and harvest time affects the combustion qualities of miscanthus genotypes. Agron. J. **95**(5), 1274–1280 (2003)
32. Liu, S., Liu, J., Loveland, T.R.: Spatial-temporal carbon sequestration under land use and land cover change. 12th International Conference on Geoinformatics – Geospatial Information Research: Bridging the Pacific and Atlantic (2004). 7–9 June 2004
33. Mann, L.K.: A regional comparison of carbon in cultivated and uncultivated alfisols and mollisols in the central United States. Geoderma **36**, 241–253 (1985)
34. Matthews, R.B., Grogan, P.: Potential C sequestration rates under short-rotation coppiced willow and miscanthus biomass crops: a modeling study. Asp. Appl. Biol. **65**, 303–312 (2001)
35. McCarl, B.A., Adams, D.M., Alig, R.J., Chmelik, J.T.: Competitiveness of biomass fueled electrical power plants. Ann. Oper. Res. **94**(1), 37–55 (2000)
36. McGowin, C.R., Wiltsee, G.A.: Strategic analysis of biomass and waste fuels for electric power generation. Biomass Bioenergy **10**(2–3), 167–175 (1996)
37. McLaughlin, S.B., Kszos, L.A.: Development of switchgrass (Panicum Virgatum) as a bioenergy feedstock in the United States. Biomass Bioenergy **28**, 515–535 (2005a)
38. McLaughlin, S.B., Kszos, L.A.: Development of switchgrass (Panicum Virgatum) as a bioenergy feedstock in the United States. Biomass Bioenergy **28**, 515–535 (2005b)
39. McLaughlin, S.B., Samson, R., Bransby, D.I., Weislogel, A.: Evaluating Physical, Chemical, and Energetic Properties of Perennial Grasses as Biofuels. 15–20 Sept 1996
40. McLaughlin, S.B., Ugarte, D.G., Garten, C.T., Lynd, L.R., Sanderson, M.A., Tolbert, V.R., Wolf, D.D.: High-value renewable energy from prairie grasses. Environ. Sci. Technol. **36**, 2122–2129 (2002)
41. McLaughlin, S.B., Walsh, M.E.: Evaluating environmental consequences of producing herbaceous crops for bioenergy. Biomass Bioenergy **14**(4), 317–324 (1998)
42. Nienow, S., McNamara, K.T., Gillespie, A.R., Preckel, P.V.: A model for the economic evaluation of plantation biomass production for co-firing with coal in electricity production. Agric. Resour. Econ. Rev. **28**(1), 106–118 (1999)

43. Paustian, K., Cole, C.V., Sauerbeck, D., Sampson, N.: CO_2 mitigation by agriculture: an overview. Clim. Change **40**, 135–162 (1998)
44. Paustian, K., Collins, H.P., Paul, E.A.: In: Paul, E.A. et al. (eds.) Management Controls on Soil Carbon. CRC, Boca Raton, FL (1997)
45. Prueger, J.H., Hatfield, J.L., Parkin, T.B., Kustas, W.P., Kaspar, T.C.: Carbon dioxide dynamics during a growing season in midwestern cropping systems. Environ. Manag. **33**(1), 330–343 (2004)
46. Puget, P., Lal, R., Izaurralde, C., Post, M., Owens, L.: Stocks and distribution of total and corn-derived soil organic carbon in aggregate and primary particle fractions for different land use and soil management practices. Soil Sci. **170**(4), 256–279 (2005)
47. Qin, X., Mohan, T., El-Halwagi, M., Cornforth, G., McCarl, B.A.: Switchgrass as an alternate feestock for power generation: an integrated environmental, energy and economic life-cycle assessment. Clean Technol. Environ. Policy **8**, 233–249 (2006)
48. Robertson, G.P., Paul, E.A., Harwood, R.R.: Greenhouse gases in intensive agriculture: contribution of individual gases to the radiative forcing of the atmosphere. Science **289**, 1922–1924 (2000)
49. Shapouri, H., Duffield, J., McAloon, A., Wang, M.: The 2001 Net Energy Balance of Corn-Ethanol. Proceedings of the Conference on Agriculture as a Producer and Consumer of Energy. 24–25 June 2004
50. Schnitkey, G.: Estimated cost of crop production in Illinois. Department of Agricultural and Consumer Economics, University of Illinois at Urbana-Champaign (2003)
51. Sinnott, R.W.: Virtues of the Haversine. Sky Telescope **68**(2), 159 (1984)
52. Six, J., Conant, R.T., Paul, E.A., Paustian, K.: Stabilization mechanisms of soil organic matter: implications for C-saturation of soils. Plant Soil **241**(2), 155–176 (2002)
53. Tharakan, P.J., Volk, T.A., Lindsey, C.A., Abrahamson, L.P., White, E.H.: Evaluating the impact of three incentive programs on the economics of cofiring willow biomass with coal in New York state. Energy Policy **33**, 333–347 (2005)
54. Tillman, D.A.: Biomass co-firing: the technology, the experience, the combustion consequences. Biomass Bioenergy **19**, 365–384 (2000)
55. Tilman, D., Hill, J., Lehman, C.: Carbon-negative biofuels from low-input high diversity grassland biomass. Science **314**, 1598–1600 (2006)
56. Turhollow, A.: Costs of Producing Biomass from Riparian Buffer Strips. Prepared by the Oak Ridge National Laboratory for the U.S. Department of Energy, ORNL/TM-1999/146 (2000)
57. Turhollow, A.F., Perlack, R.D.: Emissions of C02 from energy crop production. Biomass Bioenergy **1**(3), 129–135 (1991)
58. Ugarte, D.G., Walsh, M., Shapouri, H., Slinsky, S.: The Economic Impacts of Bioenergy Crop Production on U.S. Agriculture. U.S. Department of Agriculture (2003)
59. Unger, P.W.: In: Lal, R. (ed.) Total Carbon, Aggregation, Bulk Density, and Penetration Resistance of Cropland and Grassland Soils. vol. Special Publication 57 Madison, WI, Soil Science Society of America, pp. 72–92 (2001)
60. Uri, N.D.: Conservation tillage and the use of energy and other inputs in us agriculture. Energy Econ. **20**(4), 389–410 (1998)
61. USDA/NASS: Agricultural Statistics Database. (2003)
62. USDOE: Breaking the Biological Barriers to Cellulosic Ethanol: A Joint Research Agenda, Doe/Sc-0095. U.S. Department of Energy, Office of Science, Washington, D.C. (2006)
63. USDOE/EIA: Annual Energy Outlook 2002. DOE/EIA-0383(2002), U.S. Department of Energy (2001)
64. USDOE/EIA: Emissions of Greenhouse Gases in the United States 2005. U.S. Department of Energy, Energy Information Administration (2006)
65. USDOE/EIA: State Electricity Profiles 2002. U.S Department of Energy, Energy Information Administration (2002)
66. Walsh, M., Ugarte, D.G., Shapouri, H., Slinsky, S.P.: Bioenergy crop production in the United States: potential quantities, land use changes, and economic impacts on the agricultural sector. Environ. Resour. Econ. **24**(4), 313–333 (2003)

67. Wander, M.M., Bidart-Bouzat, G., Aref, S.: Tillage impacts on depth distribution of total and particulate organic matter in three illinois soil. Soil Sci. Soc. Am. J. **62**, 1740–1711 (1998)
68. West, T.O., Marland, G.: A synthesis of carbon sequestration, carbon emissions, and net carbon flux in agriculture: comparing tillage practices in the United States. Agric. Ecosyst. Environ. **91**, 217–232 (2002)
69. West, T.O., Marland, G., King, A.W., Post, W.M., Jain, A.K., Andrasko, K.: Carbon management response curves: estimates of temporal soil carbon dynamics. Environ. Manag. **33**(4), 507–518 (2004)
70. Wu, J., Adams, R.M., Kling, C.L., Tanaka, K.: From microlevel decisions to landscape changes: an assessment of agricultural conservation policies. Am. J. Agric. Econ. **86**(1), 26–41 (2004)
71. Wu, J., Segerson, K.: The impact of policies and land characteristics on potential groundwater pollution in Wisconsin. Am. J. Agric. Econ. **77**(4), 1033–1147 (1995)
72. Young, C.E., Westcott, P.C.: How decoupled is U.S. agricultural support for major crops? Am. J. Agric. Econ. **82**(3), 762–767 (2000)

Chapter 36
An H-Theorem for Chemically Reacting Gases

Gilberto M. Kremer, Filipe Oliveira, and Ana Jacinta Soares

Abstract The trend to equilibrium of a quaternary mixture undergoing a reversible reaction of bimolecular type is studied in a quite rigorous mathematical picture within the framework of Boltzmann equation extended to chemically reacting gases. A characterization of the reactive summational collision invariants, equilibrium Maxwellian distributions and entropy inequality allow to prove two main results under the assumption of uniformly boundedness and equicontinuity of the distribution functions. The first establishes the tendency of the reacting mixture to evolve to an equilibrium state as time becomes large. The other states that the solution of the Boltzmann equation for the chemically reacting mixture of gases converges in strong L^1-sense to its equilibrium solution.

36.1 The Model Equations

In this section, we describe a model for a mixture of four species undergoing elastic and reactive (binary) collisions of type

$$\mathrm{A}_1 + \mathrm{A}_2 \rightleftharpoons \mathrm{A}_3 + \mathrm{A}_4.$$

For $\alpha \in \{1,2,3,4\}$, we set m_α, c_α and $f_\alpha(x, c_\alpha, t)$ the mass, velocity, and distribution function of the α species respectively.

A.J. Soares (✉)
Centro de Matemática, Universidade do Minho, Campus de Gualtar, 4710-057 Braga, Portugal
and
Departamento de Matemática, Universidade do Minho, Braga, Portugal
e-mail: ajsoares@math.uminho.pt

G.M. Kremer
Departamento de Física, Universidade Federal do Paraná, Curitiba, Brazil
e-mail: kremer@fisica.ufpr.br

F. Oliveira
Centro de Matemática e Aplicações, Universidade Nova de Lisboa, Lisbon, Portugal
e-mail: fso@fct.unl.pt

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_36,

To describe this system, we consider the Boltzmann-like equation

$$\frac{\partial f_\alpha}{\partial t} + \sum_i c_i^\alpha \frac{\partial f_\alpha}{\partial x_i} = \sum_{\beta=1}^{4} \mathscr{Q}_{\alpha\beta}^E + \mathscr{Q}_\alpha^R, \tag{36.1}$$

where $\mathscr{Q}_{\alpha\beta}^E$ and $\mathscr{Q}_\alpha^R$ are the production terms with respect to Elastic and Reactive collisions. These terms are given by:

$$\mathscr{Q}_{\alpha\beta}^E = \int \left(f_\alpha' f_\beta' - f_\alpha f_\beta \right) g_{\beta\alpha} \sigma_{\beta\alpha} d\Omega_{\beta\alpha} d\mathbf{c}_\beta, \tag{36.2}$$

where $g_{\beta\alpha} = |c_\beta - c_\alpha|$, $d\Omega_{\beta\alpha}$ is an element of solid angle and $\sigma_{\alpha\beta}$ a differential elastic cross section. The models of hard sphere and Maxwell molecules [1] are commonly adopted in literature for $\sigma_{\alpha\beta}$. Moreover,

$$\mathscr{Q}_{1(2)}^R = \int \left[f_3 f_4 \left(\frac{m_{12}}{m_{34}} \right)^3 - f_1 f_2 \right] \sigma_{12}^\star g_{21} d\Omega d\mathbf{c}_{2(1)}, \tag{36.3}$$

and

$$\mathscr{Q}_{3(4)}^R = \int \left[f_1 f_2 \left(\frac{m_{34}}{m_{12}} \right)^3 - f_3 f_4 \right] \sigma_{34}^\star g_{43} d\Omega' dc_{4(3)}. \tag{36.4}$$

Here, $m_{\alpha\beta} = \frac{m_\alpha m_\beta}{m_\alpha + m_\beta}$ and the quantities $\sigma_{12}^\star$ and $\sigma_{34}^\star$ are differential reactive cross sections for forward and backward reactions, respectively.

In the expressions (36.3) and (36.4) it was considered the micro-reversibility principle which gives a relationship between $\sigma_{12}^\star$ and $\sigma_{34}^\star$, namely,

$$\sigma_{34}^\star = \left(\frac{m_{12}}{m_{34}} \right)^2 \left(\frac{g_{21}}{g_{43}} \right)^2 \sigma_{12}^\star. \tag{36.5}$$

36.2 Collisional Invariants

For a reactive collision the conservation laws of mass, linear momentum and total energy read

$$\begin{cases} m_1 + m_2 = m_3 + m_4, \\ m_1 c_1 + m_2 c_2 = m_3 c_3 + m_4 c_4, \\ \epsilon_1 + \frac{1}{2} m_1 c_1^2 + \epsilon_2 + \frac{1}{2} m_2 c_2^2 = \epsilon_3 + \frac{1}{2} m_3 c_3^2 + \epsilon_4 + \frac{1}{2} m_4 c_4^2. \end{cases} \tag{36.6}$$

Above, m_α denotes the mass of molecule $\alpha = 1, \ldots, 4$ whereas (c_1, c_2) are the velocities of the reactants, (c_3, c_4) the velocities of the products of the forward reaction and ϵ_α is the formation energy of a molecule of constituent α. In a certain sense, these are the only invariants for system (36.1). Indeed, if we define a summational collisional invariant as a function ψ which obeys to the constraints

$$\psi_\alpha + \psi_\beta = \psi'_\alpha + \psi'_\beta, \qquad \psi_1 + \psi_2 = \psi_3 + \psi_4, \tag{36.7}$$

where the first constraint refers to elastic collisions whereas the second one to reactive interactions, then we have the following result:

Theorem 36.1. *Let ψ_α be a smooth function of c_i^α, of class C^2. This function is a summational collision invariant if and only if*

$$\psi_\alpha = A_\alpha + B_i m_\alpha c_i^\alpha + C\left(\frac{1}{2}m_\alpha c_\alpha^2 + \epsilon_\alpha\right), \quad \alpha = 1, \ldots, 4, \tag{36.8}$$

where A_α and C are arbitrary scalars with $A_1 + A_2 = A_3 + A_4$, and B_i an arbitrary vector that do not depend on c_i^α.

36.3 Trend to Equilibrium

36.3.1 The Equilibrium Solution

The equilibrium solution for the present reacting mixture is characterized, at the molecular level, by the vanishing of the collision terms (36.3–36.4) on the *r.h.s.* of the reactive Boltzmann equation (36.1). Hence, the equilibrium distribution functions $f_\alpha^{(0)}$ are obtained when the equalities

$$\begin{cases} f_\alpha^{(0)} f_\beta^{(0)} = f_\alpha'^{(0)} f_\beta'^{(0)}, \\[2ex] \dfrac{f_1^{(0)}}{m_1^3}\dfrac{f_2^{(0)}}{m_2^3} = \dfrac{f_3^{(0)}}{m_3^3}\dfrac{f_4^{(0)}}{m_4^3}. \end{cases} \tag{36.9}$$

hold almost everywhere in the velocity space.

By noticing that $\ln(f_\alpha(0))$ is a summational collisional invariant, one can use Theorem 36.1 to derive an explicit expression for the equilibrium functions (see [2]). We obtain the well-known Maxwellian

$$f_\alpha^{(0)} = n_\alpha \left(\frac{m_\alpha}{2\pi k T}\right)^{3/2} e^{-\frac{m_\alpha}{2kT}\xi_\alpha^2} \tag{36.10}$$

with number densities n_α subjected to the mass action law

$$\frac{n_1^{\text{eq}} n_2^{\text{eq}}}{n_3^{\text{eq}} n_4^{\text{eq}}} = \left(\frac{m_1 m_2}{m_3 m_4}\right)^{3/2} \exp\left(\frac{E}{k_B T}\right). \tag{36.11}$$

36.3.2 An $\mathscr{H}$-Function for System (36.1)

We now define an $\mathscr{H}$-function for our system.

Definition 36.1. We put

$$\mathscr{H} = \sum_{\alpha=1}^{4} \int f_\alpha \ln\left(\frac{f_\alpha}{m_\alpha^3}\right) dc_\alpha, \tag{36.12}$$

The following results were proven in [2]:

Theorem 36.2. *For all $t \in [0; +\infty[$, $\frac{d\mathscr{H}}{dt}(t) \leq 0$. Furthermore, let $\mathscr{H}_E$ denote the $\mathscr{H}$-function referred to equilibrium Maxwellian distributions $f_\alpha^{(0)}$. Then*

$$\forall t \in [0, +\infty[, \;\; \mathscr{H} - \mathscr{H}_E \geq 0, \tag{36.13}$$

36.3.3 Convergence Results

Finally, we state the following result concerning the convergence of the distribution functions to equilibrium:

Theorem 36.3. *Assuming that $\mathscr{H}$ is a continuously differentiable function, $\mathscr{H} \in \mathscr{C}^1([0; +\infty[)$, and that every f_α is uniformly bounded and equicontinuous in t, then*

$$\lim_{t\to+\infty} \mathscr{H}(t) = \mathscr{H}_E.$$

Under these conditions, f_α converges in strong L^1-sense to $f_\alpha^{(0)}$.

Acknowledgements The paper is partially supported by Brazilian Research Council (CNPq) and by Portuguese Foundation for Science and Technology (FCT) through the research Centres CMA-UNL and CMAT, and Project PTDC/MAT/68615/2006.

References

1. Cercignani, C.: The Boltzmann Equation and its Application. Springer, New York (1988)
2. Kremer, G., Oliveira, F., Soares, A.J.: $\mathscr{H}$-Theorem and trend to equilibrium of Chemically Reacting Mixture of Gases, Kinetic and Related Models, **2**, 333–343 (2009)

Chapter 37
Dynamics, Systems, Dynamical Systems and Interaction Graphs

Maurício Vieira Kritz and Marcelo Trindade dos Santos

Abstract Graphs and Dynamical Systems are well established and mature mathematical disciplines. For a long while, their employ went along rather parallel paths. Recently, problems in the life sciences are challenging an encounter between them, particularly in what concerns understanding the interplay between the inherent organisation of living entities and their ever present dynamical character.

We review part of these recent accomplishments from a perspective broad enough to contemplate the joint statement of problems involving interaction graphs and dynamical systems. This perspective allows for a sounder understanding of mathematical descriptions of the dynamics of natural systems. It is shown that graph properties may retract characteristics of the dynamical behaviour of systems. Moreover, dynamical characteristics associated to graphs properties apply to any dynamical systems sharing the same interaction graph. This stand highlights perspectives that enrich both the study of dynamical systems and their possible applications in the life and socio-economic sciences.

37.1 Introduction

> There is a remarkable interaction between theoretical constructions in Science and conceptual notions in Mathematics: the same idea—or the same idea, disguised, may arise both in science and Mathematics. S. Mac Lane, 1985 [38, Chap. IX].

Differential equations (dynamical systems) and graphs have been used for investigating natural phenomena since long. Notwithstanding, their employ to enlighten other scientific disciplines remained restricted to physics and chemistry until the beginnings of last century, with humble incursions in the life and engineering

M.V. Kritz (✉) and M. Trindade dos Santos
LNCC/MCT, Av. Getlio Vargas 333, 25651-075, Petrópolis, Rio de Janeiro, Brazil
e-mail: kritz@lncc.br, msantos@lncc.br

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_37,

sciences. Resourcing to Warren Weaver's classification of scientific problems [75], it was restricted to problems of simplicity and of disorganised complexity.

In problems of *simplicity*, few objects interact under simple rules casted by common directives. In problems of *disordered complexity* an unmanageable number of objects interact under simple rules bearing the same directives for all interactions. Problems in both these classes can be handled by mathematical methods requiring a relatively small amount of computation by resourcing to limits, infinities, infinitesimals, perturbations and stochastic approximations. Furthermore, the distinction of boundaries, that is, of what pertains or not to the phenomenon under appreciation, is relatively straightforward for these phenomena [31]. However, Weaver also pointed to a third class of problems, where a large but not-as-huge number of unities interact under variegated rules but in an organised manner. These are problems of *organised complexity*. Although always loosely stated, *organisation* is often exemplified by living entities and phenomena [43, 55]. For phenomena in the latter class, the distinction of boundaries is often not as straightforward as in the former two; their cohesion arising mostly from the interactions among their components.

During revolutionising decades in middle of the last century, the usefulness of graphs and dynamical systems was enormously widened by the need to apply scientific methodology to problems arising from alternative disciplines, like those in economic and environmental sciences. Of no less importance was the development of computers and new computational methods that appeared in this period, which paved the road allowing mankind to face ever larger problems. Moreover, these larger problems contained components of ever greater complexity, a major source of complexity in them being biological components whenever present.

These achievements boosted the appearance and fast spreading of new concepts and approaches that greatly enlarged our vision of dynamics, systems, interactions and mathematical descriptions [59]. They levered the employ of scientific methodology and of formal reasoning to a new standard. Among these concepts one may count: control, observation, independence, state, regulation, feedback, feedforward, cybernetics, signal, noise, information and a "new" concept of system often named *general systems* [26,27,32,73,76]. Roughly speaking, the concept of system sprung from physics and chemistry is associated to delimitations of a phenomenon in space-time, while the idea of general systems neglect these distinctions in favour of distinguishing connections among objects enduring a phenomenon.

In the life and socio-economic sciences, observations about relations abound while about variations are meagre, due to restrictions imposed by present day observation methodologies. Observations about how interrelations and dynamical changes affect each other are even rarer, possible as a result of the long standing dissociation of dynamical and relational descriptions in the life sciences. Hence, problems of network complexity have been taken for problems of organised complexity, even if there is no consensual definition of organisation [33,34]. The present need to understand organisation in living systems requires new forms of analysis and reasoning, of which those leading to a deeper understanding about causations between component interactions and the dynamics of a system come in the forefront.

Mathematics has always profited from inter-playing with other scientific disciplines, 'feeding' and 'being fed' with problems, ideas and approaches (see [44, Editor's Foreword], for an account in the context of graphs and molecular genetics). From an internal stand, the interplay between mathematical disciplines themselves has also been fruitful since long [2, 38], presenting unexpected outbursts.

The purpose of this work is to address the interplay between Dynamical Systems and the Theory of Graphs, pointing to possible repercussions of their common development in the Life Sciences. It provides evidence that graph characteristics do reflect dynamical constrains and that using relational and dynamical approaches conjointly shall greatly enhance our understanding of life phenomena. Furthermore, it is argued that concepts centred around the terms systems and general systems are intrinsically related by the essence of the underlying natural phenomena they delimit, being two sides of the same coin.

In the next section we review certain characteristics of scientific descriptions of nature, particularly when expressed mathematically, discussing the two concepts of system and pointing to their tight entanglement. In the third section, we describe a generic procedure, grounded on interaction graphs, to connect the dynamical and relational descriptions of natural phenomena. In the fourth section, we review literature about results and methods that associate properties of dynamical systems to properties of their interaction graphs. In the fifth section, we argue that graph theory methods may provide information about behaviour of dynamical systems, by means of a working example. In the sixth, we discuss how mathematical knowledge about associations between properties of dynamical systems and their interaction graphs could greatly improve the advancement of life sciences, particularly with respect to the daring problems of finding the right equations in ecology and systems biology. Finally, concluding remarks and future prospects are presented in the last section.

Bibliography is meant to be only illustrative. The review of the literature about conjoint studies of interaction graphs and dynamical systems, in section fourth, is far from being complete due to the sources being multi-disciplinary.

37.2 Science and Systems

The term system, often used in a loose manner, may be taken from a scientific stand as a specification as precise and formal as possible of whatever is involved in the phenomenon under study [5, 36].

Natural phenomena of scientific interest are grounded on reproducible events or on recurring enchainments of events [22, 46, 50, 77]. Events are distinguished by clearly identifiable changes in our perceptions regarding the world. Such changes result from interactions among entities of various types: molecules, billiard balls, bodies, chemical substances, magnetic and electrical fields, processes, organisms, organisations, as well as, more complex things. Entities causing events are distinguished through features, aspects[1], or patterns in changes that unequivocally single out these entities or their environment.

Thus, phenomena of scientific interest always relate to things (objects and entities), changes, and regularities in behaviour. Where behaviour is any pattern in changes that occur in a phenomenon. Dynamic behaviour is either described by listing individual events that compose a phenomenon or by summarising them through principles and rules that unequivocally differentiate their entailment. Therefore, scientific inquiries jiggle around the study of variations (Var).

Six time-proven questions guide scientific enquiries [41,46]. "Where, when" and "how" things happen? "What" is happening? "Who" makes things happen? And, "why" do things happen in the way they do?

The questions "where, when," and "who" concern mostly things in a phenomenon. Variations in their observations typically occur in time (Var_t), along space ($\mathsf{Var}_{\mathbf{x}}$) or form ($\mathsf{Var}_{\wp(\mathbf{x})}$), and on "who" intervenes and plays a role in the investigated phenomenon (Var_{who}). Variation with respect to "whos" fall into two categories: the "whos" themselves may transform into "whos" of another nature, or just their aspects may change. Variations in the first category are exemplified by chemical reactions, where substances while moving combine giving rise to other chemical substances, or by interactions among elementary particles, which transform the interacting particles into particles of a different kind. Those in the second category are typified by a dying planet or the solidification of a liquid metal.

The questions "what, how," and "why" refer to changes that occur rather than things. They aim to gain knowledge about which changes are occurring, how these changes are performed and why at all is the phenomena occurring in the way it is, in the given situation. They, thus, centre on interactions. Curiously, interactions proper are usually relegated to a low priority during the observation phase. Interactions are instead indirectly described and investigated by analysing and comparing observations about the state of a phenomenon just before and after interactions take place. This is grounded on an implicit assumption: interactions must remain alike, or indistinguishable, and unchanged throughout the observation of a phenomenon.

As a consequence, scientific phenomena are represented by a collection of observable aspects, variables, and parameters[2] that may be arranged in two domains: an event- or entity-space and an aspect-space; the usual phase- or state-spaces being special cases of the latter. Since material things are univocally associated to positions, the event-space records answers to the questions "when, where" and "who." Thus, variables and parameters in the event-space reflect time, space

[1] *Aspects* stand for anything that may be 'perceived' about the elements of a phenomenon, particularly perceptions characterising them. Features characterise the "whos" more intrinsically, along structure and organisation, and may not be observable.

[2] Variables record changes in aspects that do not remain still. Parameters register aspects that *may be changed* according to necessity of inferences or experiments [32, Chap. 8].

and environmental conditions. It also localises the “whos” of a phenomenon[3], registering how close or apart from the boundary or from one another they are.

The aspect-space registers observations about other perceptions and describes dynamical characteristics of “what” is happening and “how.” It is not so easily or simply described, due to the large number of possible interacting entities, their constitutional variety, and their manyfold ways of interacting; all of which possess aspects to be recorded. The nature of the aspect-space is affected by the number and type of features and aspects of interest. Its structure is often difficult to distinguish, since possible interactions affect the notion of proximity in this space. Notwithstanding, arising from the same phenomenon, both spaces are definitely entangled by the phenomenon elements and any adequate scientific description must contemplate both.

Variations in the event-space relate basically to the appearance, disappearance and displacement of entities. Historically, displacements where the first changes to fascinate mankind and to receive a pertinent symbolic vestiture. The dynamics of displacements is often ruled by conservation of mass, energy or the number of existing entities. As a result, they are expressed in terms of balance equalities. The mathematical representation of such changes is based on differences or differentials (see Sect. 37.3), and leads to dynamical systems — expressed as differential equations, if time and space are continuous, or as iterative and recurrent equations, if time and space are considered discrete.

Variations in the aspect-space relate to the nature and character of intervening “whos” and their interactions, describe “what” is going on and “how” changes in the aspects of “whos” take place. The “whys” remain a matter of interpretation of representations and their properties and are not contained in either space. In the life and social sciences, the “whys” are attached to evolutionary, purpose or teleological matters [41, 58, 60]. Correctly constructed, the aspect-space contains a plethora of information about possibilities (and impossibilities) of interactions among the entities (“whos”) intervening in a phenomenon. For instance, while describing chemical interactions, properties of the aspect space should reflect the impossibility of reaction between certain substances and the easiness degree of reaction among others.

By and large, natural and artificial phenomena rely on material things and are mathematically described by dynamical systems, whenever the perception of observables being conserved allow for their description as equations. Therefore, dynamical systems are of general applicability, appearing in a form or another as models for the behaviour inherent in natural and artificial phenomena throughout scientific disciplines.

Nevertheless, the descriptive role of events and the event-space is often neglected in disciplines where the number of interacting entities is great and interactions difficult to grasp and observe; while in disciplines where the number of interacting entities is small and interactions recurrent and simple is the other way around — it is

[3] Thereby, it may contain portions of what is called configuration-space in mechanics.

the richness of information registered in the aspect-space that is neglected. As a consequence, different viewpoints have developed around each of these concerns, resulting in different and often dissociated methods of analysis as well as in two different, allegedly unrelated, concepts of system.

One, the system concept of physics and chemistry, centres in distinguishing a region where things happen and what are the exchanges between the system, taken as a single unity, and its environment. The region is determined by establishing a boundary that neatly separates what pertains to the system and what does not — defining the environment by exclusion. Examples of this type of systems are: a set of interacting billiard balls over a table, particles moving around a point, penduli on a wall, particles bouncing and colliding in closed regions, chemical substances interacting in containers, chemostats etc.

The term *system* in physics is used somewhat loosely, referring either to physical objects, coordinates or mathematical equations. Nevertheless, it always refers to a collection of units interacting or interdependent in a way or another [16]. In chemistry as well, a *system* refers to a collection of substances confined somewhere and interacting. In this sense, the term also conforms to the definition of system given by Klir [32] for *general systems*.

The second concept, widely known as general systems and intrinsic to the systemic perspective [32, 73, 76], centres in distinguishing relations and possible interactions among the system's components. This approach barely considers alterations in interactions caused by changes in vicinity, focusing in qualitative characteristics of interactions. A system is then depicted as a collection of elements with specific behaviours that react to stimuli received from other components in the system. The interchanged stimuli are treated as signals instantaneously reaching other components (see Fig. 37.1), even when delays are taken into consideration. That is, no signal propagation is represented. Their boundary and exchange with the environment are frequently difficult to establish and depicted either as special elements of the interaction network, not strictly composing the system, or as lost signals.

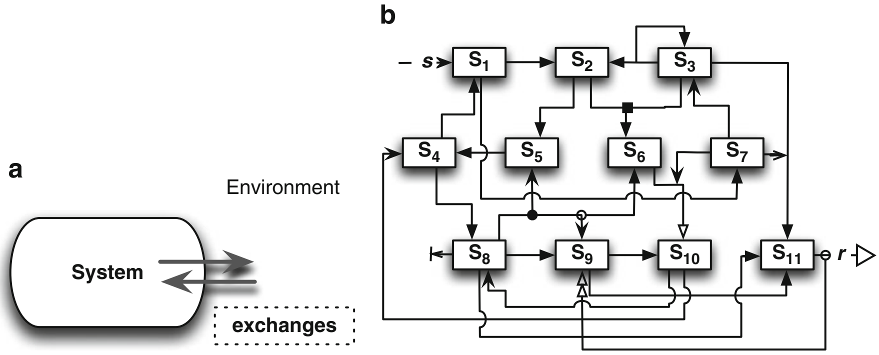


Fig. 37.1 Two kind of systems—a distinction in emphasis: (**a**) entity-based, (**b**) aspect-based

The latter concept is mostly used in the life, economic and social sciences. In all these domains and for all possible phenomenological scales we may find motor systems, distribution systems, production systems, decomposition systems, filtering systems, information processing systems, transducer systems, sensorial systems, memory systems and so on [43], that may interact and form components of other, more aggregated and complex, systems.

Matter undergo change in essentially three manners: change in position (displacement), change in form and change in substance or organisation (as in chemical transformations). Change in form fall in two categories: geometric-topological (deformations), or changes in structural organisation (as in phase transitions). For phenomena involving a collection of distinct entities, changes in interactions are mostly relevant. Interactions and their changes are ruled by two types of closeness: proximity and affinity. Ergo, any encompassing description of material phenomena requires both system concepts to be depicted in a convenient manner.

Since all natural and artificial phenomena are ultimately grounded on material things, changes are likely to happen as described in any conceivable phenomenon. Hence, these two concepts are essentially one and only one intellectual tool. Together, they allow for the description of all sort of variations in a phenomenon.

Developing a common view and a methodology for analysing systems that embrace both perspectives would reinforce and enrich this tool, as well as deepen our understanding of systems as delimitations of natural units we consider for study. The boundary of a system in the event-space, easily described, may help determining interaction boundaries in the aspect-space; while the more conspicuous stand of interactions in the aspect-space may help distinguishing events and changes due to complex interactions. The next sections indicate some ways of achieving this.

37.3 Bridging Dynamical Systems and Interaction Graphs

Although having no strict sub-sections, this section organises around three divisions or moments: dynamical systems, interaction graphs and their connection. Let us start by examining what will be understood by dynamical systems.

As detailed above, physics and chemistry focus on describing and explaining changes in matter involving simple situations, where the number of intervening objects and substances is small. Furthermore, by hypothesis, all possible transformations among these substances are known in advance. That is, no matter how the dynamics alters the elements in a phenomenon, the transformation of substances into another remains confined to a finite collection of possibilities that are known *a priori*. In general, tracking displacements of all points in a volume will tell us everything about changes in its form. Hence, from a mathematical standpoint, we need to find relations between Var_t, $\mathsf{Var}_{\mathbf{x}}$ and Var_{who} in such a way that (we feel) they explain the phenomenon.

Since the *whos* of physics and chemistry are substances, or (material) objects made out of them, let $\{\mathscr{S}_1, \ldots, \mathscr{S}_p\}$ denote the substances composing all objects intervening in a phenomenon and $\mathscr{A} = \{a_1, \ldots, a_n, n \geq p\}$ a collection of observable aspects. A central concept in our understanding of dynamical behaviour is the *state*. A *state* is essentially given by a collection of aspects and information which observation fully distinguish the dynamical behaviour of something. It is, in a certain sense, a snapshot of objects and change tendencies in a phenomenon at a given instant of time or in a time-interval.

Remark 37.1. As long as an aspect is not associated to more than one substance, the transformation of a substance ($\mathscr{S}_k$) into another ($\mathscr{S}_{k'}$) is a change in state. It will change aspects associated to $\mathscr{S}_k$ into aspects associated to $\mathscr{S}_{k'}$ in an indisputable manner, as long as, two collection of aspects associated to substances do not possess aspects in common. □

Therefore, for properly chosen aspects, to know what happens with $a_j, j = 1, \ldots, n$ amounts to know what happens and how in the phenomenon; as well as where and when changes occur. Being so, we may disregard the description of substances $\{\mathscr{S}_1, \ldots, \mathscr{S}_p\}$ and of objects formed with them in favour of describing and studying aspects associated to objects.

Note 37.1. It is a fundamental ontological assumption in science that the collection $\mathscr{A}$ of observable aspects identifies uniquely the state of a phenomenon. That is [56], if s_1 and s_2 stand for states of a phenomenon, then

$$\mathsf{s}_1 \neq \mathsf{s}_2 \Longrightarrow (\exists a^\star \in \mathscr{A}) \text{ such that } [a^\star(\mathsf{s}_1) \neq a^\star(\mathsf{s}_2)]. \tag{37.1}$$

It is implicitly required that different states associated to the same aspects $\{a_1, \ldots, a_n\}$ must be considered indistinguishable and so treated in theories. □

The choice of aspects to be observed must conform to the available observation capabilities. This establishes a window of perception and distinguishability, delimiting what can be scientifically observed, represented and studied in a phenomenon. Hence, different observation possibilities often require that specially tailored symbolic descriptions be employed. Even if disassociated at a first sight, these descriptions can often be attached to each other at a higher level of abstraction, as this section shall illustrate.

In physical systems, the aspects a_j stand mostly for displacements, changes and rates of change in points and forms, and energy related aspects. The final mathematical expression in each case studied can nevertheless be qualitatively and structurally quite different, depending whether their components are particles or continua (bodies, fluids and fields). For instance, aspects may be the mass, position, momenta and kinetic energy of each ball on a table; or of gas molecules in a container; or yet current, inductance and voltage in an electrical circuit. Whatever varies coherently. The same aspects occur in continuous objects but their mathematical expression may become rather sophisticated, due to the tracking of what is happening at each point in the continuum and to mutual influences and exchanges between aspects at nearby points. In physical interactions, analogous aspects are exchanged between objects of a phenomenon or between their points, essentially transferring intensities.

For instance, in particle interactions, energy and momenta are exchanged among intervening particles.

Physical inquiries do not consider changes in substance, while chemical ones do. Chemical systems are mostly continua since transformations of just a few molecules are never considered in this domain. In chemical systems, some aspects a_j stand also for concentrations, which vary along space. Homogeneity assumptions, however, simplify their description leading to lumped expressions. Whatever is represented in the event-space, substances may be seen as compartments, or *sorts* in mathematical terms, and changes in substance can be depicted as a migration of aspects from one compartment (or sort) into another.

Without loss of generality, we shall consider in the sequel that some aspects are associated to exactly one compartment in an unique manner, so that objects and substances are identified by a collection of aspects (see Remark (37.1) and Note (37.1)). We shall then not distinguish among aspects, objects and substances. Any entity, though, may have more than one aspect associated to it.

Summing up, the dynamic behaviour of phenomena may be described by a relation:

$$\mathscr{R}(\mathsf{Var}_t(\mathscr{A}), \mathsf{Var}_{\mathbf{x}}(\mathscr{A}), \mathsf{Var}_{\wp(\{\mathbf{x}\})}(\mathscr{A}), \mathscr{I}_{\boldsymbol{who}}(\mathscr{A})) \tag{37.2}$$

where $\mathscr{A}$ is a list of aspects and $\mathscr{I}_{\boldsymbol{who}}$ denotes exchanges between them. For aspects grounded on material things, when mass and other conservation laws apply, this relation results from reckoning the change in aspects due to changes in position, form and exchange of aspects, at each point in the event-space occupied by material objects.

What are then *dynamical systems*? In this writing, dynamical systems are symbolic descriptions of the dynamical behaviour of natural systems by means of their aspects; that is, of physical, chemical, biological, ecological, social or artificial systems. Whenever the aspects in $\mathscr{A}$ are expressed numerically, whenever the aspects are measurable[4], dynamical systems can be expressed mathematically or computationally.

In this case, the elements of $\mathscr{A}$ become vectors $\mathbf{q}$ of quantities that conform to (37.1). Furthermore, if time and space are not entangled, that is, their dimensions being independent, relation (37.2) becomes additive[5] at each point $\mathbf{p}$ of the event-space:

$$\mathsf{Var}_t(\mathscr{A})|_{\mathbf{p}} = (\mathsf{Var}_{\mathbf{x}}(\mathscr{A}) + \mathsf{Var}_{\wp(\{\mathbf{x}\})}(\mathscr{A}) + \mathscr{I}_{\boldsymbol{who}}(\mathscr{A}))|_{\mathbf{p}}, \tag{37.3}$$

where $\mathbf{p}$ is a point in $dom(\mathscr{R})$. Usually $\mathbf{p} = (t, \mathbf{x}; \boldsymbol{\pi})$, embracing localization in space-time and environmental conditions. Observe that relation (37.3) does not apply for aspects like information, that may not be conserved or unambiguously

[4] Aspects may be made numerical by reckoning the number of objects presenting a certain aspect. Measuring, however, requires a comparison with a standard. In general, the first method results in discrete values for variables, while the second in continuum ones.

[5] Additiveness, at this point, has nothing to do with reduction or linearity. It stems from properties of the quantities used in counting, independence among observables and the balance of aspects implied by conservation.

attached to material things. Moreover, vectors $\mathbf{q}$ of quantified aspects do encompass environmental (inputs and output) as well as internal (state) variables and need to conform to observability and consistency requirements [42].

Dynamical systems are sound mathematical objects from many perspectives. From the systems theory perspective, its definition makes explicit the possible interactions with other systems and the environment, the form they are observed, and the possible changes in state, binding to the more general relation (37.2) and do not requiring additiveness (37.3). From this perspective, dynamical systems have the following mathematical structure [26, 27, 42]:

Definition 37.1 (Dynamical System). An octuple $\Sigma = \{T, U, \mathscr{U}, X, Y, \mathscr{Y}, \phi, \eta\}$ is a dynamical state-space system, with time-domain T, inputs $u \in U$, outputs $y \in Y$ and (internal) states $x \in X$, where $\mathscr{U} = \{u_I : T \to U\}$ is a set of admissible input functions, $\mathscr{Y} = \{y_O : T \to Y\}$ a set of output functions, $\phi : T \times T \times X \times \mathscr{U} \to X$ a state-transition function and $\eta : T \times X \to Y$ the observation function, if:

1. $T \subset \mathrm{I\!R}$ is an ordered set,
2. $\mathscr{U} \neq \emptyset$ and is closed under concatenation,
3. $x(t) = \phi(t; \tau, x, u_I)$, the state of Σ at time t resulting from the *initial-state* $x(\tau)$ at initial time $\tau \in T$ under the action of input $u_I \in \mathscr{U}$, satisfies:
   (a) (Direction of Time) ϕ is defined for all $t > \tau$,
   (b) (Consistency) $\phi(t; t, x, u_I) = x, \forall t \in T, \forall x \in X$ and $\forall u_I \in \mathscr{U}$,
   (c) (Composition) $(\forall t_1 < t_2 < t_3)\, \phi(t_3; t_2, \phi(t_2; t_1, x, u_I), u_I)$, for all $x \in X$ and $u_I \in \mathscr{U}$,
   (d) (Causality) For any $t > \tau$, if $u, u' \in \mathscr{U}$ and $u(\tau, t] = u'(\tau, t]$, then
   $$\phi(t; \tau, x, u) = \phi(t; \tau, x, u')$$
4. $y(t) = \eta(t, x(t))$ is the observed output.

A careful choice of elements and properties in Σ reduces the systems of Definition (37.1) to several well known classes of systems: time invariant, finite dimensional, finite state, input–output etc, where time may be continuous, discrete, or both. Considerations about observable inputs and outputs will be recalled in Sect. 37.6. The arguments in the sequel are centred solely in the dynamics of state transitions ϕ. Descriptions of inputs and outputs will be then disregarded by considering time-invariant, autonomous, perfectly observable systems (that is, systems such that $\eta = I$ or $y(t) = x(t), \forall t$).

Observe that the conception of system as a device for distinguishing what pertains to a phenomenon and what does not naturally divides the set $\mathscr{A}$ into three groups: aspects describing observations that strictly pertain to the phenomenon ($\mathscr{A}_S$), aspects depicting what is transferred from the environment to the system ($\mathscr{A}_I$), and aspects depicting what is transferred from the system to the environment ($\mathscr{A}_O$). If the number of aspects in $\mathscr{A}_S$ is greater than 1, the state-set X (of system S) is multi-dimensional and $\mathscr{I}_{\boldsymbol{who}}(\mathscr{A}_S) = \mathscr{I}_{\{\mathrm{x}\}}$ is a relation which domain is X.

Note that Definition (37.1) is a step towards the common view proposed at the end of Sect. 37.2. In this definition, $\mathscr{A}_I$ is made explicit by $\mathscr{U}$ and $\mathscr{A}_O$ by $\mathscr{Y}$ and η. The domain suggested by Fig. 37.1a is part of the specification of the state-set X and the aspects and their interaction exemplified in Fig. 37.1b appear as $\mathbf{x}$ and ϕ. However, Definition (37.1) leaves the specification of dynamics open; that is, the specification of ϕ and of how variations should be tracked (computed).

For systems with only a few quantifiable aspects, $\mathscr{A} = \{1, \ldots, n\}$, the state becomes a vector of values $\mathbf{x} = \{x_1, \ldots, x_n\}$, the interactions of aspects ($\mathscr{I}_{who}$) become simply a finite set of relations in $\mathbf{x}$. Under these constrains, and when time is continuous and aspects are conserved and smooth enough, the transition function ϕ frequently find expression as the flow of a system of differential equations:

$$\frac{d\mathbf{x}}{dt}(t) \overset{def}{=} \dot{\mathbf{x}}(t) = \mathbf{I}(\mathbf{x}(t)), \tag{37.4}$$

where $\mathbf{x}(0) = \mathbf{x_0} \in \boldsymbol{M}$ is a given vector, $\mathbf{I} : \boldsymbol{M} \longrightarrow \boldsymbol{T(M)}$ is a continuous regular enough function, $\boldsymbol{M}$ is the aspect space (variety) and $\boldsymbol{T(M)}$ its tangent bundle. Or, if time is discrete, as the iterated application of a mapping from the aspect space $\boldsymbol{M}$ into a copy of it:

$$\mathbf{x}(k+1) = \mathbf{I}'(\mathbf{x}(k)), k = 1, 2, \ldots \tag{37.5}$$

where $\mathbf{x}(0) = \mathbf{x_0} \in \boldsymbol{M}$ is given.

The latter system may result from a discretisation of equations (37.4) or be the direct result of modelling a phenomenon when observations are only feasible at discrete time intervals τ [20]. For both kinds of dynamics though, degeneracies apart, we have that $dim(\boldsymbol{M}) = dim(\boldsymbol{T(M)})$; even if the system's flow remains confined to sub-varieties of $\boldsymbol{M}$.

Such systems are studied within the mathematical field of Dynamical Systems; a well established discipline counting with more than one century of active developments. Their study focus on global properties of solutions of equations of type (37.4) or (37.5) employing geometrical and topological arguments, having stemmed from work by Henri Poincare [20]. Dynamical systems may also be directly expressed in computational form. There is, though, a close relationship between their mathematical and computational expressions [27] up to the algorithmic form, that ingenious speciations of Definition (37.1) reveal to be dynamical systems as well.

The second movement relates to discrete and relational mathematics, particularly Graph Theory. Graphs are mathematical objects as old as dynamical systems and have been used since long as models or realisations of polyhedra, algebraic structures, molecular bindings, reaction chains, flux networks, and object-to-object relations at large. Graph theory is a rich subject. It embraces and requires both theoretical and algorithmic-computational investigations; being unable to prescind from either [47]. Despite a first result by Euler, Graph Theory blossomed in the second half of the nineteenth century, after seminal work done by Kuratowsky and Cayley. Mathematical relations and graphs are close concepts [61] and are at the root of the (general) system concept when dynamics is not being considered [32, 42]. They reflect the connections established among entities by interactions or exchange-channels [3] (see Fig. 37.1b).

Disguised under different names, like "circuit language, block diagrams, bond graphs, causal networks, dominance maps, flowcharts, Forrester diagrams tendency matrices, transition diagrams, trophic webs, wiring diagrams," to name a few, graphs have been used as a tool to inspect and organise observations and knowledge about systems, as well as, to straightforwardly build models for the underlying phenomenon [21, 39, 40, 48, 54, 62].

From cooperative-competitive relations in socio-economic systems [54], to mass and energy flows in ecosystems [72], to the intricate connectivity of possible reactions in metabolic networks [48], to biological systems at large [21, 39, 40, 54] and beyond into cognitive maps, they support the identification of mutual influences, the tracking of paths and cycles of matter and energy fluxes and the identification of dynamic dependencies and relationships at large[6]. Not just this, graphs forged a large class of models of variable success in many disciplines [21, 36, 39, 44, 54], conveniently representing different sorts of relationships in a variety of phenomena.

Graphs model relations and are therefore universal and ubiquitously applicable [19, 28, 47, 54]. Graphs are just a pair of sets. Notwithstanding being mathematical objects simpler than the dynamical systems of Definition (37.1), they are often defined in slightly different forms resulting in not completely equivalent classes of objects. However, variants and generalisations of graphs are obtainable by properly choosing how to define these two sets. Those relevant to the following discussion are described in the definitions below.

Definition 37.2 (Undirected Graphs). A *graph* G is a pair of sets $\{V, E\}$, which elements are respectively the vertices and edges (unordered pairs) of G, where E is a family of subsets of V of cardinality two. That is, $e \in E \Rightarrow e = \{v_1, v_2\}, v_i \in V$.

Proposition 37.1. *This definition precludes multi-edges and loops.*

Proof. Immediate. The elements of a set are, by definition, distinct and unique. □

Definition 37.3. The variants of interest are:

1. A *directed* or *oriented* graph has arcs (ordered pairs) instead of edges. That is, $G = \{N, A \subset N \times N\}$. In this case the vertices are usually called nodes[7].
2. An *edge-labeled* graph (respectively, arc-labeled oriented graph) is a triple $\{V, E, \lambda : E \to \Lambda\}$ ($\{N, A, \lambda : A \to \Lambda\}$), where Λ is the set of labels and $dom(\lambda) = E$ ($dom(\lambda) = A$).

[6] Cognitive maps appear everywhere: in the early stages of modelling [21] or explanations about mathematical properties [42, Fig. 4.1]. Strictly speaking, all these versions of interaction graphs are cognitive maps, since interactions are indeed a form of relationship.

[7] Although the distinction between edges and arcs will be maintained in the text to distinguish when talking about undirected or directed graphs, vertices V and nodes N will be used interchangeably.

3. A *bipartite* graph (respectively, a bipartite oriented graph) has a partitioned node set $N = N_1 \cup N_2,\ N_1 \cap N_2 = \emptyset$, and edges (respectively, arcs) from one set to the other. That is, $e \in E \Rightarrow [e \cap N_1 \neq \emptyset \wedge e \cap N_2 \neq \emptyset]$ (respectively, $A \subset N_1 \times N_2 \cup N_2 \times N_1$).
4. A hyper-graph is a pair of sets $\{N, E \subset \wp(N)\}$, where $E = \{e_1, \dots, e_p\}$ is a family of non-empty subsets of N such that $\cup_i e_i = N$ [6].

Note 37.2. The following facts about graphs will not be proven. Their proof may be found in textbooks or the references cited.

1. Loops are allowed by Definition 37.3.1.
2. Vertex-labeled graphs may be defined analogously to edge-labeled ones. Of course, a graph may have both its edges and vertices (or its arcs and nodes) labeled.
3. A bipartite graph may have isolated nodes and nodes occurring in just one edge (arc) but no loops.
4. A hyper-graph may be made directed by a collection of procedures and can have, as well, edges and vertices labeled. Hyper-graphs have no isolated nodes [6].
5. Hyper-graphs may be re-written as bipartite graphs [35, 61], but not all bipartite graphs represent hyper-graphs. □

It is a marvel how rich the theory of these simple objects is. Moreover, it is at the same time theoretical (mostly, combinatorial) and algorithmic; constructive proofs providing immediate algorithms and algorithms being needed in proofs [47]. The structure of individual graphs provide a plethora of properties and classes of them are studied from topological and other stands [6, 7, 19, 44, 54]. Within mathematics itself, graphs are models or representations of relations and algebras [61]. Their connection to dynamical systems will be addressed below.

The third movement is about the inherent association between graphs and dynamical systems. Given a dynamical system, in the form of equations (37.4) or (37.5), we define its interaction graph in the sequel. There exists many ways to associate graphs and dynamical systems but the procedure provided is general, based on the interactions among aspects of a phenomenon. This association further partitions the class of dynamical systems, as characterised by vector the fields $\mathbf{I}$ or $\mathbf{I}'$.

Remind that, observable aspects of a system may be associated with inputs, states, responses, environmental factors etc. The sequel refers, though, essentially to state variables. Let $J_n = \{1, 2, \dots, n-1, n\}$ be labels (names) that denote either the state variables $\{x_1, \dots, x_n\}$ of a system or their rates of change $\{\dot{x}_1, \dots, \dot{x}_n\}$. Any dynamical system is associated with a graph. Under form (37.4), a dynamical system is identified by its vector field $\mathbf{I}$. Therefore, the following definition of interaction graphs also give a procedure to associate dynamical systems to them:

Definition 37.4. An *Interaction Graph* of a dynamical system $\mathbf{I}$ is a directed graph $\{N, A\}$ such that:

1. $N = J_n$;
2. Its arcs represent dependencies of components of **I** on components of **x**, that is:

$$\forall(1 \leq i, j \leq n),\ [(i,j) \in A] \qquad \textit{if and only if}$$
$$[I_i(x_1, \ldots, x_j, \ldots, x_n) - I_i(x_1, \ldots, x'_j, \ldots, x_n) \not\equiv 0], \quad \forall\, \mathbf{x}, \mathbf{x}'_{(j)} \in dom(\mathbf{I}), \tag{37.6}$$

where $\mathbf{x}'_{(j)} = (x_1, \ldots, x_{j-1}, x'_j, x_{j+1}, \ldots, x_n)$.

This same definition is valid in the case of discrete systems ($\mathbf{I}'$) as well.

Interaction graphs are symbolic devices that help registering dependencies among the various observables. Besides dependencies they also register tendencies imposed in the variation of an observable by another. Interaction Graphs have been used to help establishing and improving models in the life and social sciences (where gathering observations about changes along time, particularly quantitative data, is often problematic [1, 40, 48, 62]) either as graph diagrams in strict sense [21, 36, 39, 54] or by means of their adjacency matrices [48].

It is nevertheless important to note that labels J_n of Definition 37.4 stand for both states and their rates of variation, and that interaction graphs are indeed bipartite graphs where $N = \dot{J}_n \cup J_n$ and the arcs of (37.6) belong to $\dot{J}_n \times J_n$ and $J_n \times \dot{J}_n$, those in $J_n \times \dot{J}_n$ remaining implicit. The dotted arcs in Fig. 37.2 mean that $x_i(t + dt) = \dot{x}_i dt$ and that $x_i^t \leftarrow x_i^{t+1}$ (x_i^{t+1} replaces x_i^t), at each advance of time. Moreover, nodes representing $\dot{x}_i$ and x_i^{t+1} are connected to exactly one node of the other partition. Consequently, it is common to identify respective nodes in the two partitions reducing the graph to a non-bipartite digraph while using interaction graphs in analyses and modelling.

Remark 37.2. While depicting interaction graphs as digraphs, it is important to note that objects represented by the nodes at the source and sink of arcs belong to different mathematical spaces. □

The bridge between dynamical systems and interaction graphs may be summarised by the following mapping. Let $\mathbf{G}_n$ denote the class of all graphs G which node set $N(G)$ has n elements, and let $\mathscr{C}^r(\mathbb{R}^n; \boldsymbol{T}(\mathbb{R}^n)) = \mathscr{C}^r(\mathbb{R}^n)$ be the class of all vector-fields $\mathbf{F} : \mathbb{R}^n \longrightarrow \boldsymbol{T}(\mathbb{R}^n) = \mathbb{R}^n$ that are $r \in \mathbb{Z}, r \geq 0$ times differentiable. Each dynamical system is attached to a graph by the mapping:

$$\begin{aligned} \mathscr{G} : \mathscr{C}^r(\mathbb{R}^n) &\longrightarrow \mathbf{G}_n \\ \mathbf{F} &\mapsto G, \end{aligned} \tag{37.7}$$

where $G = \mathscr{G}(\mathbf{F})$ is given by Definiton 37.4. The mapping $\mathscr{G}$ is a well defined function and $dom(\mathscr{G}) = \mathscr{C}^r(\mathbb{R}^n)$. Notwithstanding, $img(\mathscr{G}) \neq \mathbf{G}_n$, if $\mathscr{C}^r(\mathbb{R}^n)$ stands for irreducible n-dimensional dynamical systems (see Sect. 37.6).

Interaction graphs are often labeled with information related to dynamics, particularly rates and fluxes. When $r \geq 1$ and $\mathbf{F}$ and $\mathscr{G}(\mathbf{F})$ are known, even partially,

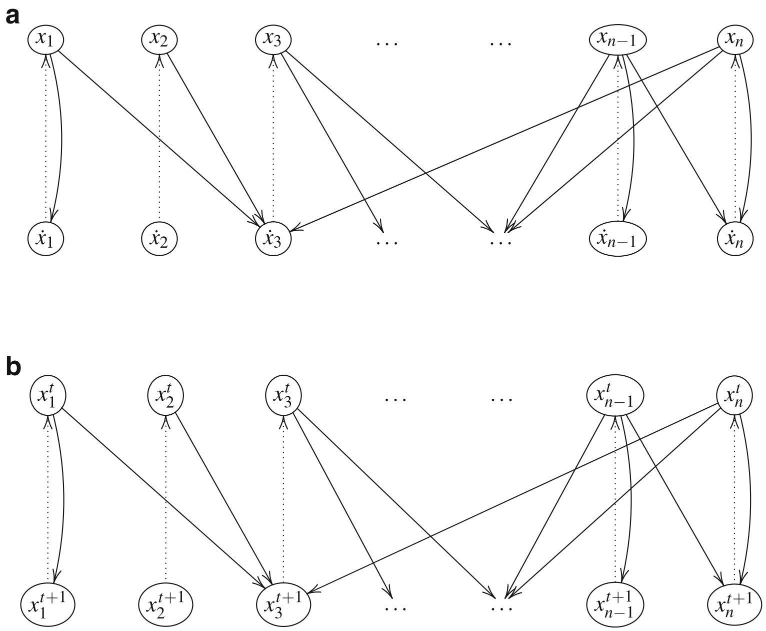


Fig. 37.2 Interaction Graphs: (**a**) continuous time; (**b**) discrete time

there is a simple and frequently used way for labeling arcs, where labels depend on position and parameters of the vector field *dom*(**F**). This labeling attributes to arcs values related to transfers or fluxes among the aspects of a system. That is, the (numerical) label of an arc $a_{i,j} = (i, j)$ is given by:

$$\lambda(a_{i,j}) = \frac{\partial}{\partial x_j}(\frac{dx_i}{dt}) = \frac{\partial}{\partial x_j}(\dot{x}_i) = \frac{\partial}{\partial x_j}F_i, \tag{37.8}$$

where $\frac{\partial}{\partial x_j}F_i$ is the $\mathbf{J}_{ij}$ entry of the Jacobian matrix of **F** and depends on parameters and the state variables **x**.

Labels like these reflect tendencies in variation rates of the aspects that describe a system. Both topics will be addressed further on. They may be made independent of states and parameters by many procedures: taking averages, considering minimum or maximum values, or using other statistics. They may also be made less variable at each point by considering only discrete values, like the signal of the Jacobian matrix entries:

$$\lambda(a_{i,j}) = sign\left(\frac{\partial}{\partial x_j}F_i\right). \tag{37.9}$$

Labels as defined in (37.9) contain information about how an aspect inhibits or enhances the *rate of change* of another. This sort of relationship is known as activation-inhibition switching in biology [57] and socio-economic disciplines [54, Chap. 9].

37.4 Interplay Between Dynamical Systems and Interaction Graphs

Lately, the most conspicuous and apparently widespread use of graphs to investigate natural or artificial ([63]) phenomena wander around network topology. These results centre on a dynamics over sets of graphs characterised by the addition/deletion of vertices and edges [24, 25, 67]. Despite of this, other often more straightforward forms of using graphs to model and investigate natural phenomena have existed since long [40,44,54,62]. Even if not as prominently as Network Theory, the interplay between graphs and dynamical systems suggested by associated interaction graphs (37.7) have been investigated for quite a while now, perhaps not as systematically as networks and dynamical systems per se. A partial overview of these investigations is presented in this section.

Explorations of interaction graphs originate mostly on applications and proceed along two main avenues. One is concerned with the development or improvement of algorithms for computing solutions of dynamical systems, particularly with respect to their control and observability. The other strive to deepen our understanding about relationships between the system's structure brought afloat by interaction graphs and its dynamical behaviour.

Along this reasoning, theoretical and algorithmic properties of graphs — which reflect the interacting possibilities between the aspects represented in a system — are used to unveil facts and properties about the system's dynamics. Due to time constrains and the disperse situation of these accomplishments over several fields, the account provided below is far from complete and only illustrative of possibilities relevant to arguments in the sequel; it fails to do justice to many interesting lines of enquiry. Alongside dynamics, interaction graphs were used to also investigate structural solvability, controllability and observability of systems (see [27] for more information on these concepts).

Since the seventies, stimulated by applications concerning systems science, graph properties have been used to understand dynamics[8] and to develop or simplify algorithms for computing solutions of dynamical systems expressed in terms of differential equations, as exemplified by [28,45]. The very first applications refer largely to linear systems, represented in either form (37.4) or (37.5), requiring **I** to be a linear mapping. Interaction graphs were highly successful in both unveiling dynamic properties and reducing the computational complexity of algorithms

[8] See [54, Chap. 10], [39, Part II] and [45], for information on early achievements.

in these applications. From our present knowledge about deep connections existent between matrices and graphs this success should not be surprising. These early achievements are nevertheless impressive [39,45], [54, Chap. 10 and 11].

Interaction graphs were further used to support investigations about global dynamical properties of non-linear systems and about relations between dynamical behaviour of systems and their solvability, observability and controllability. In particular, interaction graphs, also called representation graphs, are useful for investigating the decomposition or decoupling of dynamical systems into simpler ones [28, 45] by identifying strongly connected components and interactions of lesser strength, leading often to systems of lower dimension. This approach greatly improved algorithms for computing orbits and flows of dynamical systems.

At least since the beginning of the eighties, graph properties and interaction graphs are being more systematically used to infer and study properties of associated dynamical systems. These investigations, as well as earlier ones on linear systems, rely on the same concepts — signed labels and signed circuits — although grounded on different ontologies; circuits being another name for cycles in a graph.

The sign of a circuit is defined as follows. Given an interaction graph G_λ with label defined as in (37.9) and a circuit $\mathtt{c} = \ <n_{i_1}, n_{i_2}, \ldots, n_{i_{k-1}}>$ in G_λ (see definitions further on in Sect. 37.5), the sign of $\mathtt{c}$ is given by:

$$sign(\mathtt{c}) = \prod_{m=1}^{k-1} \lambda(\{n_{i_m}, n_{i_{m+1}}\}). \tag{37.10}$$

To strength the discrete nature of its labels, graphs with signed labels will be henceforth denoted by $G_\pm$.

Of notice here is the work originated by conjectures rosen by R. Thomas since early eighties [70]. His conjectures all stem from biological considerations and the first one reads as follows:

Conjecture 37.1 (R. Thomas, 1981). The presence of a positive circuit (anywhere in the phase space) is a necessary condition for multi-stationarity.

As stated in [29], this conjecture was proved by C. Soulé [65], after several partial results [8, 17, 64]. Other conjectures under the same lines of inquiry exist but their statement require the introduction of other concepts originated from interaction graphs [29], which is outside the scope of this review. Partitioning the phase-space (aspect-space) based on properties of interaction graphs' circuits (or cycles) is possible too [29], leading to decompositions similar to basins of attraction. Partitions grounded on graph-decompositions are also feasible as will be indicated in Sect. 37.5.

Notwithstanding the close relationship between discrete and continuous versions of a system revealed by discretisation procedures and their analysis, connections between results about graphs and dynamical systems seem to be tighter or easier to prove among discrete systems. Recently, several results on these connections were obtained with respect to discrete dynamical systems. Some carved by conjectures along lines similar to those of Thomas [51–53]; other under different approaches and concepts for system and dynamics [13, 30]. Yet, another line of enquiry look after revealing dynamical properties directly from observed interaction graphs [9, 10], focusing on properties observed at the biochemical scale and without considering properties of living systems valid more generally.

The interplay between graph and dynamical system theories have been profitable since long. Still, the employ of graph algorithms to investigate characteristics of dynamical systems and their phase-spaces have not been, to the best of our knowledge, as widely used. The next section addresses this point.

37.5 Graph Decomposition and Lorenz System: An Example

Surveying literature about rapports between graphs and dynamical systems is a lengthy task due multiple nuances concerning their objectives and the distinct meaning of several concepts. Instead, we shall work out an example, which shall illustrate the relations between the characteristics of flows and the structure of interaction graphs and their possible uses. The example hereafter applies a graph decomposition algorithm, first developed for ecological systems, to the well known Lorenz dynamical system.

In the eighties, Ulanowicz [71,72] developed an algorithm to investigate the cycle structure of food-webs and their function as components of ecological systems. The algorithm decomposes a numerically labeled graph of measured fluxes into a collection of cycles, associating a flux-value to each cycle, and leaving a residual acyclic graph with remaining eventually null fluxes as labels. In this way, each cycle can be associated to an ecosystem's 'function,' the cycle flux-value indicating in what proportion the population represented by each node is engaged in that function. The residual acyclic graph depicts whatever aspects flow through the food-web without remaining longer in the ecosystem.

We have recently generalised this algorithm to operate on bipartite graphs and employed it on graphs of biochemical reactions labeled with steady-state fluxes, that arise from chemical considerations and observations in metabolic networks [35]. The interpretation is not as straightforward as in the original application, but this investigation leads to interesting results about metabolism that provide information towards the organisation of metabolic networks. Although there is little reference to the dynamical aspects of metabolism due to missing data, this application conforms to and illustrates the actual stage of biological modelling.

Before introducing the algorithm, a few more elements related to graphs need to be defined. The definitions and the decomposition algorithm are valid for either directed or undirected graphs and shall be presented below without any explicit note but using their notation consistently. Let $G = \{V, E\}$ ($G = \{N, A\}$) be a graph. Two nodes $n_1, n_2 \in N(G)$ are *adjacent* if $\{n_1, n_2\} \in E(G)$ ($(n_1, n_2) \in A(G)$). A sequence $\mathsf{p} = (n_{i_1}, n_{i_2}, \ldots, n_{i_k})$ of adjacent nodes in G such that $n_{i_j} \neq n_{i_l}$, whenever $j \neq l$, $\forall 1 \leq j, l < k$, is called an *elementary path*. The edges of p are then the sets $e_l(\mathsf{p}) = \{n_{i_l}, n_{i_{l+1}}\}$, $\forall 1 \leq l < k$. An *elementary cycle* in G is an elementary path in G such that $n_{i_1} = n_{i_k}$. Elementary cycles will be denoted by $<n_{i_1}, n_{i_2}, \ldots, n_{i_{k-1}}>$, not including the recurring node n_{i_k}. The cardinality of a set S will be denoted by $|S|$. Note that paths and cycles are *sub-graphs* of G. Two other definitions are yet needed for the decomposition algorithm: *critical arc* and *nexus*.

Definition 37.5. Given a set of cycles $\mathbf{C} = \{\mathtt{c}^1, \ldots, \mathtt{c}^p\}$ of a numerically labeled graph $G_\lambda = \{N, A, \lambda : A \to L \subset \mathbb{R}\}$, we define:

1. A *critical arc* of a cycle $\mathtt{c}$ is an arc $a^\circ(\mathtt{c})$ (or edge $e^\circ(\mathtt{c})$) such that

$$\lambda(a^\circ(\mathtt{c})) = min_{a \in \mathtt{c}}\, \lambda(a). \tag{37.11}$$

2. A *critical arc* of a set $\mathbf{C}$ of cycles is an arc $a^\circ(\mathbf{C})$ (or edge $e^\circ(\mathbf{C})$) such that

$$\lambda(a^\circ(\mathbf{C})) = min_{\mathtt{c} \in \mathbf{C}}\, \lambda(a^\circ(\mathtt{c})). \tag{37.12}$$

3. The *nexus*, $\mathbf{N}(a)$, of an arc a (or edge e) with respect to a set of cycles $\mathbf{C}$, is the set of all cycles in $\mathbf{C}$ sharing the arc a (or edge e). That is,

$$\mathbf{N}(a) = \{\mathtt{c} \in \mathbf{C} \mid a \in A(\mathtt{c})\}. \tag{37.13}$$

It is important to remark that the *critical arc* may be defined differently, depending on ontological considerations and on what is being investigated [71].

The decomposition algorithm presupposes that all elementary cycles are known. Their identification can be accomplished by several algorithms, Tarjan's [68] being one of the most efficient. Strictly speaking, Tarjan's algorithm presupposes a directed graph, but this is not a restriction since most applications present directed graphs. Furthermore, any undirected graph may be associated to a directed graph with the same nodes. The algorithm that enumerates all cycles of a graph will be denoted by *CyclesOf*. The nexus of an arc, $\mathbf{N}(a)$, is usually found by and exhaustive search. The algorithm that constructs a nexus given an arc a and a set of cycles $\mathbf{C}$ will be denoted by *NexusOf*$(a, \mathbf{C})$.

Being given a labeled network $G_\lambda = \{N, A, \lambda : A \to \mathbb{R}\}$, such that $\lambda > 0$, the decomposition algorithm builds a flux-label ϕ on a copy $\mathbf{C}_\phi$ of the set of cycles $\mathbf{C}$ and produces a residual graph G_{res}. A sketch of this algorithm is in the table Algorithm 1.

In Algorithm 1, ← denotes assignment and $[b - expression] \Rightarrow$ `statement` denotes Dijkstra's guarded commands [18]: `statement` is only executed if the boolean expression $[b - expression]$ is **true**. Furthermore, boolean expressions are surrounded by brackets, $[B]$, *AnsatzGeneration*$(\mathbf{N}(a^\circ))$ means that the weights w_j are generated from ontological considerations, or ontologically grounded hypotheses, that take into account the structure of $\mathbf{N}(a^\circ)$ and $\phi(\mathtt{c}_j) = \kappa$ means that $\phi(a) = \kappa$, $\forall a \in A(\mathtt{c}_j)$. The following facts about Algorithm 1 hold.

Remark 37.3. Algorithm 1 is a pretty rich object. The following observations reveal some of its broadness and possibilities of application.

Algorithm 1 Cycle Decomposition Algorithm (Ulanowicz)

```
Require: A numerically labeled (direct) graph $G_\lambda$ .
Input: $G_\lambda$
Require: An algorithm for finding a complete cycles enumeration of $G_\lambda$.
Initialisation::
    $\mathbf{C}_\phi, \mathbf{C} \leftarrow \mathit{CyclesOf}(G_\lambda)$
    $img(\phi) \leftarrow \emptyset, \quad G_{res} \leftarrow G_\lambda$
Ensure: $[\mathbf{C}_\phi, \mathbf{C} \neq \emptyset]$
Require: Algorithms for finding critical arc (equations (37.11) and (37.12)) and
         assembling the nexus $\mathbf{N}(a)$ of an arc $a$ in a set of cycles $\mathbf{C}$ (equation (37.13)).
while $\mathbf{C}$ is non empty do
    $a^\circ \leftarrow \mathit{criticalArc}(\mathbf{C})$
    $\mathbf{N}(a^\circ) \leftarrow \mathit{NexusOf}(a^\circ, \mathbf{C})$
    Heuristics::
    $\{w_j, 1 \leq j \leq |\mathbf{N}(a^\circ)| = n^\circ\} \leftarrow \mathit{AnsatzGeneration}(\mathbf{N}(a^\circ))$
Ensure: $\forall(1 \leq j \leq n^\circ)[0 \leq w_j \leq 1] \wedge [\sum_{j=1}^{n^\circ} w_j = 1]$
    $\phi$–Update:: $\forall(\mathsf{c_j} \in \mathbf{N}(a^\circ))$
    $img(\phi) \leftarrow img(\phi) \cup \{w_j \lambda(a^\circ)\}$;
    $\phi(\mathsf{c}_j) \leftarrow w_j \lambda(a^\circ)$
    $\lambda$–Update:: $\forall(\mathsf{c}' \in \mathbf{N}(a^\circ))$
    $\lambda(a) \leftarrow \lambda(a) - \phi(\mathsf{c}') \qquad \forall(a \in A(\mathsf{c}'))$
Ensure: $[\lambda(a^\circ) = 0]$
    $G_{res}$–Update:: $\forall(\mathsf{c}' \in \mathbf{N}(a^\circ))$
    $[\lambda(a') = 0] \quad \Rightarrow \quad A(G_{res}) \leftarrow A(G_{res}) \backslash \{a'\} \qquad \forall(a' \in A(\mathsf{c}'))$
    C–Update::
    $\mathbf{C} \leftarrow \mathbf{C} \backslash \mathbf{N}(a^\circ)$
end while
Output: $G_{res}$ and $\phi : \mathbf{C}_\phi \longrightarrow \mathbb{R}$
```

Loops Cycle enumeration algorithms usually require loopless graphs. Loops, $\mathsf{c}^n = \{\{n\}, a^n = (n,n)\}, n \in N(G_\lambda)$, if added to $\mathbf{C}$, would be singled out by Algorithm 1 as degenerate cases ($a^\circ = a^n$, $\mathbf{N}(a^\circ) = \{a^n\}$ and $\phi(a^\circ) = \lambda(a^n)$) and directly “transferred” to $\mathbf{C}_\phi$ with $\phi(\mathsf{c}^n) = \lambda(a^n)$.

Heuristics The heuristics supporting the *AnsatzGeneration*$(\cdot)$ procedure steams from ontological considerations and may have distinct interpretations and justifications. Ulanowicz [71, 72] computes probabilities that give the chance of an element of a population being involved in the ecological processes represented by each cycle in $\mathbf{N}(a^\circ)$. Here and in [35] the same computations are employed but interpreted as proportions of mass or energy flowing among processes present in the cycles. Different heuristics may be used for the same system to investigate distinct aspects of its behaviour.

Non-determinism Algorithm 1 is non-deterministic in principle. When different arcs of G_λ have the same label, it may occur that $|\{\mathit{criticalArc}(\mathbf{C})\}| > 1$. In this case, *criticalArc*$(\cdot)$ returns more that one value, or must chose arc one from the set $\{\mathit{criticalArc}(\mathbf{C})\}$. □

The following theorem asserts the correctness of Algorithm 1, clarifying its execution.

Theorem 37.1. *Algorithm 1 stops independently of the chosen heuristics as long as* $0 \leq w_j \leq 1$ *and* $\sum_{j=1}^{n^\circ} w_j = 1$. *Furthermore,* G_{res} *has no cycles and* ϕ *attributes one value to each* $\mathsf{c} \in \mathbf{C}$.

Proof. Assuming that *CyclesOf* and any other undocumented pre-processing are correct, $\mathbf{C}$ contains all cycles of G_λ possibly excluding loops (see Remark 37.3: Loops). Let us examine the two possibilities: $\mathbf{C} = \emptyset$ and $\mathbf{C} \neq \emptyset$.

If $\mathbf{C} = \emptyset$, commands between lines 3 and 11 are never executed and nothing changes in the algorithm inputs, but the algorithm **stops**. Notwithstanding, $G_{res} = G_\lambda$ and the theorem's assertions are true, since $dom(\phi) = \emptyset$.

If $\mathbf{C} \neq \emptyset$, the critical arc a° of each cycle $\mathsf{c} \in \mathbf{C}$ exists due to the finiteness of $|A(\mathsf{c})|$ and so does a° of $\mathbf{C}$ due to the finiteness of $\mathbf{C}$. Moreover, there is at least one cycle $\mathsf{c} \in \mathbf{C}$ such that $[(a^\circ \in \mathsf{c}) \wedge (a^\circ = criticalArc(\mathbf{C}))]$. Hence, $\mathbf{N}(a^\circ) \subset \mathbf{C}$ exists and $|\mathbf{N}(a^\circ)| > 1$. Both, a° and $\mathbf{N}(a^\circ)$ are found by exhaustive search on $\mathbf{C}$ or an equivalent substitute. Since $[0 \leq w_j \leq 1]$ and $[\sum_{j=1}^{n^\circ} w_j = 1]$, after completion of line 7 we have that:

$$\{w_1\lambda(a^\circ), \ldots, w_{n^\circ}\lambda(a^\circ)\} \subset img(\phi) \tag{37.14}$$

and that

$$\sum_{j=1}^{n^\circ} \phi(\mathsf{c}_j) = \lambda(a^\circ). \tag{37.15}$$

Therefore, the execution of line 8 at the arc a° for all $\mathsf{c}' \in \mathbf{N}(a^\circ)$ will result in:

$$\lambda(a^\circ) \leftarrow \lambda(a^\circ) - \sum_{j=1}^{n^\circ} \phi(\mathsf{c}_j) = 0, \tag{37.16}$$

due to (37.15). Furthermore, a° will be removed from G_{res} in line 9. This disrupts all cycles of G_{res} recorded in $\mathbf{N}(a^\circ)$. These cycles are also removed from $\mathbf{C}$ in line 10, and thus

$$|\mathbf{C}| \leftarrow |\mathbf{C}| - |\mathbf{N}(a^\circ)|. \tag{37.17}$$

Finally, observe that at any point in the algorithm, except possibly between lines 9 and 10, $\mathbf{C}$ is a collection of sub-graphs of G_{res}, hence finite. Consequently, since $|\mathbf{N}(a^\circ)| \geq 1$, $|\mathbf{C}|$ will diminish at each execution of line 10, becoming 0 ($\mathbf{C} = \emptyset$) after a finite number of loop executions. The while-loop (lines 3–11) thus *stops*. Moreover, as $\mathbf{C}$ is a snapshot of the cycles yet in G_{res} at any time, there will be no more cycles left in G_{res} when the algorithm stops. □

Note that the points highlighted in the algorithm under the **Ensure** tag are fundamental in proving its correctness. Thus, special care while programming the algorithm must be taken at these points, for instance, for handling approximations.

Theorem 37.2. *Algorithm 1 works for any label* $\lambda : A(G_\lambda) \to \mathbb{R}$ *(not just if* $\lambda > 0$*), as long as the critical arc is defined by (37.11) and (37.12).*

Proof. Let λ be a non-positive label for G. Since $|A(G_\lambda)| < \infty$ for any G_λ, there exists

$$\lambda^{\bullet}(G_\lambda) = min_{a \in A(G_\lambda)} \lambda(a).$$

Consider an invertible transformation $\lambda' : A(G_\lambda) \longrightarrow \mathbb{R}$ of λ, such that

$$\lambda'(a) = \lambda(a) + abs(\lambda^{\bullet}(G_\lambda)) + \epsilon,$$

where ϵ is 'just big enough' to prevent rounding errors, and the same graph G with label λ' instead of λ, that is, $G_{\lambda'}$. The following assertions hold:

1. Since the cycles of a graph do not depend on labels, $A(G_{\lambda'}) = A(G_\lambda)$, we have that $\mathbf{C}(G_{\lambda'}) = \mathbf{C}(G_\lambda)$.
2. $\lambda'(a) > 0, \ \forall a \in A(G)$, by construction.
3. $\lambda'(a_1) \geq \lambda'(a_2) \Rightarrow \lambda(a_1) \geq \lambda(a_2), \ \forall a_1, a_2 \in A(G_\lambda)$.

The above inequalities imply that $a^{\circ}_{\lambda'} = a^{\circ}_{\lambda}$, for any set of cycles $\mathbf{C}$ in both G_λ and $G_{\lambda'}$, where $a^{\circ}_{\lambda'} = criticalArc_{\lambda'}(\mathbf{C})$ and $a^{\circ}_{\lambda} = criticalArc_{\lambda}(\mathbf{C})$. Moreover, $\mathbf{N}(a^{\circ}_{\lambda'}) = \mathbf{N}(a^{\circ}_{\lambda})$, for any critical arc $a^{\circ}_{\lambda'} = a^{\circ}_{\lambda}$. Therefore, since the algorithm stops for λ', it will also stop for λ. Besides this, label $\phi : \mathbf{C} \longrightarrow \mathbb{R}$ is well defined because its construction does not assumes that $\lambda > 0$. □

To illustrate the possibility of using mathematical and algorithmic graph theory to forth a better understanding of dynamical behaviour, we now apply this cycle decomposition algorithm to a well known dynamical system with low dimension, namely the one given by Lorenz equations. The decomposition is performed at each fixed point and around them and the results compared among themselves and with known facts about its dynamical behaviour.

Lorenz equations were derived in 1963 as a model of a forced-dissipative hydrodynamic system, heated from below and cooled from above [66]. Lorenz was mainly interested in convective motion in the atmosphere, and in applying conclusions drawn from the model to weather forecasting. The importance of this seminal work relies on results obtained for bounded non-periodic orbits, which are extremely unstable with respect to small perturbations of initial conditions. Lorenz equations represent the dynamics of three modes of the Oberbeck–Boussinesq equations for fluid convection and accurately represent properties of these modes for $r \approx 1$ (see [20]). They are:

$$\left.\begin{aligned} \dot{x} &= I_1(x, y) = \sigma(y - x) \\ \dot{y} &= I_2(x, y, z) = -xz + rx - y \\ \dot{z} &= I_3(x, y, z) = xy - bz \end{aligned}\right\}, \tag{37.18}$$

where $\mathbf{x} = (x, y, z)$ are state variables, belonging to the aspect (phase) space, and $\sigma, r, b > 0$ are parameters characterising the fluid and flow, that is, the object causing the phenomenon (see Sects. 37.2 and 37.3).

The state variables are associated to quantities describing convective flow, relevant to atmospheric phenomena [66]:

- x is proportional to the intensity of convective motion;
- y is proportional to the temperature difference between ascending and descending convective currents; and
- z is proportional to the distortion of vertical temperature profile from linearity.

Note that none of the state variables is directly observable. The Lorenz system parameters satisfy $\sigma, r, b > 0$, due to physical constrains, and depend on the acceleration of gravity g; on the coefficient of thermal expansion α; on the kinematic viscosity ν; and on the thermal conductivity κ.

The equilibrium solutions ($\dot{\mathbf{x}} = 0$) of system (37.18) are, straightforwardly:

$$\left.\begin{aligned} \mathbf{x}_0^{\mathbf{e}} &= (0,0,0), \\ \mathbf{x}_1^{\mathbf{e}} &= (-\sqrt{b(r-1)}, -\sqrt{b(r-1)}, r-1), \text{ and} \\ \mathbf{x}_2^{\mathbf{e}} &= (\sqrt{b(r-1)}, \sqrt{b(r-1)}, r-1), \end{aligned}\right\} \tag{37.19}$$

where $\mathbf{x}_1^{\mathbf{e}}$ and $\mathbf{x}_2^{\mathbf{e}}$ exist and are nontrivial only when $r > 1$. The Jacobian matrix $\nabla\mathbf{I}$ of the interaction term in (37.18) (see (37.8) and the end of Sect. 37.3) reads:

$$\mathbf{J} = \begin{bmatrix} -\sigma & \sigma & 0 \\ (r-z) & -1 & -x \\ y & x & -b \end{bmatrix}. \tag{37.20}$$

The interaction graph G_L (Definition 37.4) of the Lorenz system is easily constructed either from (37.18) or using its Jacobian matrix. The condition established by equation (37.6) is satisfied whenever $\mathbf{J}_{ij} \not\equiv 0$. Hence, there are three arcs leading to y and z, from the three state variables, and only two leading to x, one from x itself and another from y, as shown in Fig. 37.3. The arcs of G_L are labeled by the Jacobian of $\mathbf{I}$, as in (37.8).

Remark 37.4. The following observations are not restricted to the Lorenz system and apply to other systems as well.

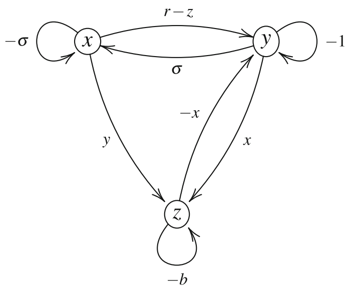


Fig. 37.3 Interaction Graph G_L of the Lorenz system.

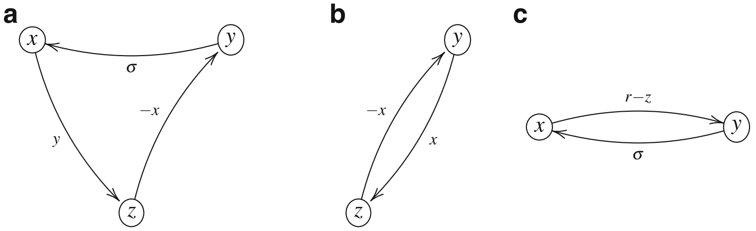


Fig. 37.4 The three cycles of Lorenz's system interaction graph G_L

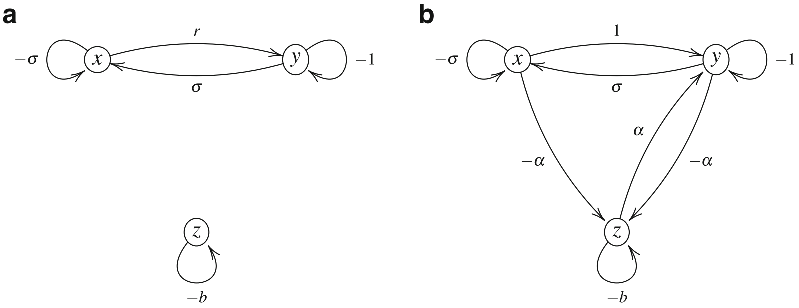


Fig. 37.5 G_L evaluated at each fixed point: (**a**) $\mathbf{x}_0^{\mathbf{e}}$, (**b**) $\mathbf{x}_1^{\mathbf{e}}$

1. Labels $\lambda(a_{ij})$ do depend on the state variables (x, y, z), but not necessarily for all indices i and j, and parameter values.
2. The graph G_L is the *same*, no matter how labels are defined over $A(G_L)$.
3. When labels depend on states or parameters, arcs disappear in the regions where the associated label vanishes, if $\lambda(a_{ij}) = 0$ implies that $\dot{x_i}$ is independent of x_j.

□

In the case of Lorenz system, a complete enumeration of the cycles in G_L may be found by direct inspection. The only three cycles of G_L, $\mathsf{c}_1, \mathsf{c}_2, \mathsf{c}_3$, are displayed in Fig. 37.4, together with the labels induced by (37.8). They correspond respectively to Fig. 37.4a, b, c.

At any point of subspace $\mathbb{R}_z = \{(0, 0, z),\ z \in \mathbb{R}\}$ the graph G_L splits, becoming disconnected whatever values the parameters $\sigma, r, b > 0$ assume, as shown in Fig. 37.5a. Node z remains isolated under these conditions and only the cycle between x and y remains (Fig. 37.4c). There is no exchange of energy between aspects z and x, y and only the trivial solution is admissible for the xy-cycle. Therefore, should an orbit cross the z-axis, it would abruptly become trapped there. From a physics stand, this means that the interaction of distortion from linearity in the temperature profile with the other variables is a key issue for the behaviour captured by Lorenz system. The loop in node z is a feedback relation with a negative coefficient, reverting the sign of the influence of z on itself.

Otherwise, G_L does not change or become disconnected by changing the parameters σ, b, r, as long as they remain positive. This is true for all $\mathbf{x} \in \mathbb{R}^3$ except in its sub-space given by $r = z$, where the xy-cycle is disrupted for any feasible value of σ, b, r. It is important to note that graph-cycles allow for the existence of periodic and quasi-periodic orbits restricted to the variables in its nodes but do not enforce their existence. Otherwise, the *absence* of arcs and graph-cycles in G_L provide information about orbits that cannot exist and regions of aspect-spaces that cannot be visited by any orbit of a system, being thus of great relevance.

Algorithm 1 cannot be used for G_L with its functional labels. But, in consequence of Theorem 37.2, it can be applied to instances of G_L where λ is evaluated at particular states $\mathbf{x}^\star$ of system (37.18). Evaluating, for instance, the labels of G_L at each equilibrium point (37.19) results in graphs with constant coefficients, as depicted in Fig. 37.5, where $\alpha = \sqrt{b(r-1)}$. The interaction graph evaluated at $\mathbf{x}_2^\mathbf{e}$ is not shown. It is the same as the one in Fig. 37.5b with labels α exchanged for $-\alpha$, and vice-versa.

Note that, in the parameter sub-space given by $r = 1$, the interaction graphs at both equilibrium points $\mathbf{x}_1^\mathbf{e}$ and $\mathbf{x}_2^\mathbf{e}$ become equal to the disconnected interaction graph at equilibrium point $\mathbf{x}_0^\mathbf{e}$ for any state $\mathbf{x}$, since $(r = 1) \Rightarrow (\alpha = 0)$, which is consistent with the existence of just one equilibrium point, $\mathbf{x}_0^\mathbf{e}$. This corresponds to the onset of convective motion [20, 37].

The results of applying Algorithm 1 to the interaction graph with labels evaluated at distinct points do depend on relations between parameters σ, b, r and state values. The following conditions hold for the results presented below:

- $r < \sigma$, for λ evaluated at $\mathbf{x}_0^\mathbf{e}$.
- $1 < \sigma + \alpha$, for λ evaluated at $\mathbf{x}_1^\mathbf{e}$.
- $r < \sigma + w_1\alpha$, for λ evaluated at $\mathbf{x}_2^\mathbf{e}$.

At each equilibrium point, Algorithm 1 outputs fluxes ($\phi(\mathtt{c})$) for cycles $\mathtt{c}_1, \mathtt{c}_2, \mathtt{c}_3$ and one acyclic graph. The fluxes are:

1. for $\mathbf{x}_0^\mathbf{e}$: $\phi(\mathtt{c}_3) = r$,
2. for $\mathbf{x}_1^\mathbf{e}$, $\phi(\mathtt{c}_1) = \phi(\mathtt{c}_2) = -\alpha$, $\phi(\mathtt{c}_3) = 1$,
3. for $\mathbf{x}_2^\mathbf{e}$, $\phi(\mathtt{c}_2) = -\alpha$, $\phi(\mathtt{c}_3) = 1$.

The acyclic sub-graphs resulting from applying the decomposition algorithm at each fixed point are shown in Fig. 37.6. At $\mathbf{x}_0^\mathbf{e}$ the acyclic sub-graph of Fig. 37.6a remains. For G_L evaluated at $\mathbf{x}_1^\mathbf{e}$ the residual acyclic graph is depicted in Fig. 37.6b and in Fig. 37.6c when evaluated at $\mathbf{x}_2^\mathbf{e}$. In Fig. 37.6, α is the same as before, $w_1 = \frac{\sigma}{\sigma+\alpha+1}$ and $w_2 = \frac{1+\alpha}{\sigma+\alpha+1}$.

If the above conditions are violated, the cycle fluxes become:

1. for $\mathbf{x}_0^\mathbf{e}$: $\phi(\mathtt{c}_3) = \sigma$,
2. for $\mathbf{x}_1^\mathbf{e}$, $\phi(\mathtt{c}_1) = \phi(\mathtt{c}_2) = -\alpha$, $\phi(\mathtt{c}_3) = \sigma + \alpha$,
3. for $\mathbf{x}_2^\mathbf{e}$, $\phi(\mathtt{c}_2) = -\alpha$, $\phi(\mathtt{c}_3) = \sigma + w_1\alpha$,

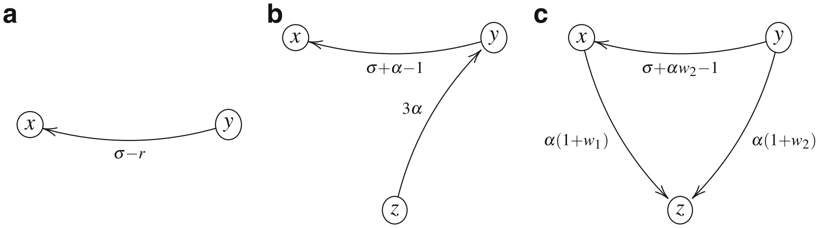


Fig. 37.6 Acyclic graphs resulting from Algorithm 1 at each equilibria: (**a**) $\mathbf{x}_0^{\mathbf{e}}$, (**b**) $\mathbf{x}_1^{\mathbf{e}}$ and (**c**) $\mathbf{x}_2^{\mathbf{e}}$

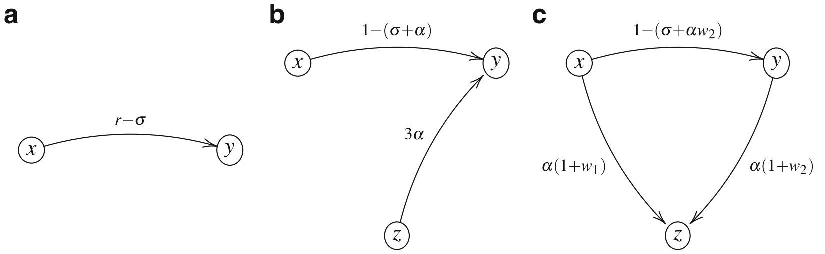


Fig. 37.7 Acyclic graphs resulting from Algorithm 1 at each equilibria: (**a**) $\mathbf{x}_0^{\mathbf{e}}$, (**b**) $\mathbf{x}_1^{\mathbf{e}}$ and (**c**) $\mathbf{x}_2^{\mathbf{e}}$, when the conditions are violated

while the resulting acyclic graphs become, for each equilibrium point, those displayed in Fig. 37.7. Although the residual graph at $\mathbf{x}_0^{\mathbf{e}}$ may seem uninteresting, since *a fortiori* $x = y = 0$, it may give information about tendencies in nearby orbits when compared to decompositions in its neighbourhood.

Let us now investigate what happens to G_L and its decomposition in a neighbourhood of each equilibrium point. With this purpose, consider an arbitrary small displacement $\mathbf{h} = (h_x, h_y, h_z)$ and perform the cycle decomposition at an arbitrarily small vicinity of the equilibrium points $\mathbf{x}_i + \mathbf{h}$.

Since $\mathbf{h} \neq \mathbf{0}$ no label vanishes and the interaction graph at $\mathbf{x}_i^{\mathbf{e}} + \mathbf{h}$, $\forall(1 \leq i \leq 3)$ is the same as G_L in Fig. 37.3 with particular labels, shown in Fig. 37.8. Clearly, the cycles remain the same, while the fluxes assigned to each cycle change according to $\mathbf{h}$.

We first inspect the residual acyclic graphs around equilibria $\mathbf{x}_1^{\mathbf{e}}$ and $\mathbf{x}_2^{\mathbf{e}}$. It is easy to see that in both cases the residual acyclic sub-graphs remain the same as before, whenever $\|\mathbf{h}\| \ll min(\sigma, b, r)$. However, there are important changes at $\mathbf{x}_0^{\mathbf{e}}$.

At $\mathbf{x}_0^{\mathbf{e}}$, any small displacement $\mathbf{h}$ completely changes the interaction graph, that remains connected. The graph in Fig. 37.8 is the same as that of Fig. 37.3. All three variables interact and convective motion is possible. Notwithstanding, $\mathbf{h}$ affects the decomposition as much as parameters σ, r and b. Therefore, it is possible to get different decompositions depending on the sign of $\mathbf{h}$ components, as exemplified above in Figs. 37.6 and 37.7. Moreover, inspecting $\mathbf{h}$ points to regions of the state-space where the decomposition changes or where cycles are disrupted ($r = z_i^{\mathbf{e}} + h_z$), implying changes in the flow regime.

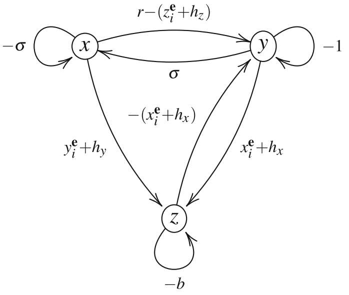


Fig. 37.8 Lorenz system interaction graph around its equilibrium points

37.6 Inverse Problems in Systems Biology and Ecosystems

In classical physics and chemistry we can observe aspects at several points in time and even on whole time intervals obtaining gorgeous series of observations. In contrast, with present-day observation methods used in the life sciences at large, more often than not only snapshots about the constitution of a phenomenon components and about interactions among them can be observed; observation of interactions providing but meagre information in the majority of cases. Particularly in biology and ecology, observations convey information mainly about the nodes of interaction graphs (aspects of entities) and partially about the strength of interactions and exchanges at certain points in time. Hence, although it is quite straightforward to construct representative models for the dynamics and flows of aspects in physics and chemistry directly from observations, even if spatially dependent, it is rather the contrary when dealing with biological and ecological systems.

In life sciences, from the biochemistry of intracellular systems up to organisms, populations and ecological systems, we need to content ourselves with observations about what constitutes the various components of a phenomenon and about the interactions between them; sometimes appended with observations concerning interdependencies among these interactions. In special cases, it is possible to further gather information about the intensity and type of interactions (what is exchanged, their structure and organisation). Except for intensities, that emanate from dynamics, observations about aspects of the living, particularly those relative to interactions and relations, tend to vary slightly and slowly. They issue from previous knowledge about organisations and interrelationships in a phenomenon.

They are grounded on chemical affinities in the case of cellular networks and on direct observation concerning affinities and exchanges between components in the case of larger organic entities, way up to macroscopic ecosystems; although these observations may require the registration and handling of events at various time and space scales. Nevertheless, with just a handful of exceptions, observation of life science subjects can only be performed at certain points in time often destroying the organisation of the living; at least part of it.

Furthermore, at any life scale, the identification of components and their interactions is difficult due to several reasons. Crucial data for understanding their structure and functioning are routinely missing, not to say the data needed to understand the most distinguishing characteristic of life phenomena — the organisation of intervening entities and processes [3,33,41,43,55,63,75].

Therefore, while modelling life phenomena we customarily face the following inverse problem. The statement below assumes that the most relevant observation is arranged and represented as an interaction graph G_λ [1,11,48,52], except possibly for information about qualitative changes in dynamical behaviour.

Problem 37.1. Given a labeled interaction graph G_λ, where λ assumes real numerical values, or a tendency interaction graph $G_\pm$, where λ assumes $+$ or $-$ values, find or, more properly, *guess*:

A. what are key characteristics of the phenomenon and what is the underlying dynamics subjacent to observations about it, and
B. what relations and interactions are being observed and what should be a dynamical system to represent a possible dynamics, given by a vector field **I** whenever feasible. □

This is a general scheme for the problem which assumes a plethora of distinct forms and is stated in the literature in widely diverse terms [1,10–12,14,21,39,48,54,71]. Steps A. and B. distinguish two important modelling moods: the organisation of observations and concept formation, and the expression of them in symbolic or mathematical terms. Both go hand in hand, subject to the scientific questions (see Sect. 37.2), step A. being more intuitive and synthetic whereas step B. more analytical and deductive. They cannot be separated, though, and are recurrent: step A. re-intensifying after step B. increases knowledge about the phenomenon.

To better appreciate this problem, yet without considering hierarchies and organisation, let us rewrite it with the aid of function $\mathscr{G}$, given by (37.7). More precisely, if a labeled graph with n nodes $G_\lambda \in \mathbf{G}_n$ is built out of observations, we need to find a dynamical system $\mathbf{I}$ such that $G_\lambda = \mathscr{G}(\mathbf{I})$ and λ is associated to $\mathbf{I}$ on sensible grounds. Moreover, it is expected that $\mathbf{I}$ qualitatively reproduces conspicuous dynamical features of the phenomenon or at least the more relevant in a global manner. As examples of such features we list: saturation, diffusion, switching, explosion, degradation etc. Stated as this, inventing adequate morphisms $G_\lambda \longrightarrow \mathbf{I}$ is a general description of modelling in the life sciences .

Notwithstanding, the set

$$\mathscr{G}^{-1}(G_\lambda) = \{\mathbf{I} \in \mathscr{C}^r(\mathbb{R}^n; \boldsymbol{T}(\mathbb{R}^n)) \mid G_\lambda = \mathscr{G}(\mathbf{I})\}, \tag{37.21}$$

that is, the set of dynamical systems having the *same* interaction graph, is a pretty large set to choose from as it is non-denumerable.

Letting $G_\circ \in \mathbf{G}_n$ without restriction while considering inverse images of $\mathscr{G}$, where $G_\circ$ denotes a graph irrespective of its labels, will bring several uninteresting or duplicated systems into account. For instance, if the adjacency matriz of $G_\circ$ is the identity, $A(G_\circ) = \{(i,i), i = 1,\ldots,n\}$. That is, $G_\circ$ is a set of looped nodes otherwise unconnected. This means that $\mathbf{I}$ is a set of n one-dimensional flows and we will be dealing with n instances of systems in $\mathscr{C}^r(\mathbb{R}^1; \boldsymbol{T}(\mathbb{R}^1))$. Similarly if the adjacency matrix can be rearranged into a non-overlapping block-diagonal matrix. To prevent be dealing with less dimensional systems, better handled in $\mathscr{C}^r(\mathbb{R}^l; \boldsymbol{T}(\mathbb{R}^l))$, $1 \leq l < n$, we consider only strongly connected interaction graphs $G_\circ \in \mathbf{G}_n^{sc} \subset \mathbf{G}_n$. Strongly connected graphs, are graphs where there is a path $\mathsf{p}(i,j)$ from i to j and another $\mathsf{p}(j,i)$ from j to i, for any two nodes $i, j \in N(G)$. Hence, changes in a variable are (indirectly) influenced by all other variables. Dynamical systems $\mathbf{I}$ associated to this type of interaction graphs are non-decomposable [45,54] into lower dimension ones.

Having such a large set from which to draw models shouldn't be an issue, if we could guarantee that any two dynamical systems drawn from $\mathscr{G}^{-1}(G_\lambda)$ have the right dynamical characteristics for the modelled phenomenon. For instance, that they have the right number of equilibrium points with proper types and that they transit from one basin of attraction to another in similar ways. To have this confidence, a classification of $\mathscr{G}^{-1}(G_\lambda)$ along the lines of dynamical systems theory, so well represented [20] by Peixoto's theorem [49], would be mostly welcome. With this in mind, note that $\mathscr{G}^{-1}$ partitions $\mathscr{C}^r(\mathbb{R}^n; \boldsymbol{T}(\mathbb{R}^n))$ into classes which contain dynamical systems sharing the *same* interaction graph. Moreover, other partitions of $\mathscr{C}^r(\mathbb{R}^n; \boldsymbol{T}(\mathbb{R}^n))$ may be found by joining the classes $\mathscr{G}^{-1}(G_\circ)$ for $G_\circ$ possessing particular graph properties.

The first thing to note is that the number of inverse images in (37.21) is finite, as both $|\mathbf{G}_n|$ and $|\mathbf{G}_n^{sc}|$ are finite. Moreover, even $\mathbf{G}_n^{sc}$ can be further subdivided sole in terms of connection topology, completely disregarding any label λ defined on $G \in \mathbf{G}_n^{sc}$. One way of performing this is to inspect the cycles of G. For instance, if $A(G) = \{(1,2), (2,3), \ldots, (n-1,n), (n,1)\}$, $G \in \mathbf{G}_n^{sc}$ and it has only one cycle. Whichever the graph G is, the number of cycles in G depend only on the set of arcs $A(G)$ and not on λ. Hence, we can define classes of graphs $\mathbf{G}_q^c \subset \mathbf{G}_n^{sc}$ such that $G \in \mathbf{G}_q^c \Leftrightarrow |\mathbf{C}(G)| = q$, where $\mathbf{C}(G)$ is the set of all cycles of G. We have that:

Proposition 37.2. $G \in \mathbf{G}_n^{sc} \Rightarrow \mathbf{C}(G) \neq \emptyset$

Proof. $\mathsf{p}(i,j)$ concatenated with $\mathsf{p}(j,i)$ is a cycle, $\forall i, j \in N(G)$. □

Therefore, the collection of all $\mathbf{G}_q^c$ form a partition of $\mathbf{G}_n^{sc}$ since $\bigcup_q \mathbf{G}_q^c = \mathbf{G}_n^{sc}$ and, consequently, the collection $\{\mathscr{G}^{-1}(\mathbf{G}_q^c)\}$ is a partition of $\mathscr{C}^r(\mathbb{R}^n; \boldsymbol{T}(\mathbb{R}^n))$.

The conjectures and results discussed in [51–53] for discrete systems and in [1, 9, 10, 13, 29, 30] for other systems, indicate that each $\mathbf{G}_q^c$ can be further partitioned according to dynamical properties, by resorting to label characteristics related to dynamics. The suggested classification is thus feasible.

Biological dynamics, nevertheless, is plenty of swift changes from one steady-state to another — which may be homeostatic, cyclic, circa-cyclic, etc — that are associated with states moving from the influence region of one point of equilibrium to another [1, 11, 69]. To cope with that, models should admit large (but controlled) changes or moves in orbits to be caused by small changes in values of parameters, state-variables or labels. Biological dynamics thus intensifies and reinforces challenges about the relevance of structural stability [14, 21], already arising from other fields [20] and reflected on applications of chaotic dynamics and catastrophe theory within biological disciplines.

In the present setting several approaches suggest themselves as instruments to closer inspect the tessiture of each partition member of $\mathscr{G}^{-1}(\mathbf{G}_n^{sc})$ and relations of its elements to properties of graphs, that do not require the well developed analytical and geometric tools of dynamical systems theory. These instruments can thus be employed in a manner complementary to the traditional ones. Labels may be refined and enriched (they may be vectors or more complex structures), while the decomposition resulting from Algorithm 1 may change suddenly with respect to changes in parameters or state-variables, as shown in Sect. 37.5, providing information. From another stand, there is great flexibility in Algorithm 1, particularly with respect to the heuristics employed to distribute fluxes among cycles (line 6) and even in the definition of a critical arc (equations (37.11) and (37.12)). This concerns mathematics.

The arguments of Sects. 37.2 and 37.3 show that whatever we do while studying life phenomena, we will end up with a relational description and a dynamics to build. Disregarding dynamics in favour of purely relational descriptions or vice-versa, does not seem the wisest thing to do. Dealing with both approaches together may, on the contrary, open new horizons. From the biological point of view, not only properties of cycle-labeling procedures or those relative to dynamics have ontological interpretations and meaning [1, 11]. The cycle decomposition too has a meaning in ecological systems [71, 72] and is related to the behaviour and organisation of biochemical networks [11, 35, 69], and other biological systems as well.

The possibility of using operations on graphs associated with graph properties to reveal the dynamics of biological phenomena rises the following possibly overlapping questions, some of which are already addressed in the existing literature.

- Is the number of cycles associated to the number of equilibrium points? Or is label information needed to illuminate this?
- How tight is the relation between cycles and equilibria?
- What dynamical properties come solely from graph properties and which require more information?
- Do the missing arcs of an interaction graph reflect regions of the aspect-space forbidden to orbits?
- How are changes in graph structure, caused by variations in parameters and variables, reflected in dynamics?
- How far do the cycle-decomposition provide clues about dynamics?
- How are variations in cycle-decomposition along the state and parameter spaces related to changes in dynamical behaviour?
- Do these changes furnish biologically relevant information?
- Does the presence of residual graphs, for any conceivable decomposition of fluxes among the cycles of an interaction graph, relate to chaotic behaviour?

These questions are far from being an exhaustive. Many other questions on this subject arise from biological puzzles. The last question above seems to be purely mathematical. Notwithstanding, since acyclic residual graphs have fluxes associated to them, they reveal transfers of energy that must accumulate in internal aspects when the system is closed. This accumulation may be a source of disorganisation, as suggested by the Lorenz system (Sect. 37.5). Answering this sort of questions will be a direct enhancement of biological knowledge. This becomes specially conspicuous when the number of nodes (i.e., aspects and equations) is large.

37.7 Conclusions and Prospects

> I think there are two acts in mathematics. There is the ability to prove and the ability to understand.
>
> Marc Kac, 1982 [15]

Cross-fertilisation among mathematical disciplines have always enriched and enlarged the mathematical adventure [2,38]. Cross-fertilisation among mathematics and other sciences has always been of great value to both partners. The work and the problems above described illuminate a bridge between methods of dynamical and relational approaches in mathematics. In elaborating this text we opted to sacrifice the completeness of the overview in favour of binding to the spirit of Mac Lane's quotation at the beginning.

This paper conveys the idea that properties of interaction graphs are useful in prospecting the nature of dynamical systems associated to them. It argues also that a more systematic investigation about the behaviour of dynamical systems associated to particular sub-graphs of the interaction graph would enrich their knowledge from a mathematical perspective and be of great value to the life sciences.

Properties of dynamical systems linked to properties of graphs are shared by all dynamical systems having the same interaction graph and equivalent labels. A deeper knowledge concerning the dynamics of systems that share the same interaction graph, about which properties they have in common or not, will reduce difficulties in modelling and understanding life processes brought in by the scarcity of data about them. Furthermore, this knowledge can suggest new observation methods.

A clearer panorama concerning the possibilities of dynamical behaviour for a given observation, will enhance our understanding with respect to what is feasible and what is not feasible in the design and organisation of living systems. This is a case where bringing together two mathematical areas would greatly enrich another science. The standpoint advanced is a partial answer to a question posed by Michael Grinfeld during the workshop Emerging Modelling Methodologies in Medicine and Biology—EM3B[9]; namely, what is the usefulness of mathematics in building theories for the life sciences. Why is it so, and possibilities for strengthening this fundamental role, is addressed in the sequel.

Abduction [21, 22], and the consistent modelling and hypothesis construction from partial data emanating from it, is a mandatory procedure while prospecting the unknown in the empirical sciences. A point that is often overlooked, though, is the

[9] IMCS, July 20-24, 2009, http://www.icms.org.uk/workshops/modellingmethodologies

effect of the available knowledge in this process. Knowledge provides the spectacles for observing and the tools for treating data. Therefore, it partially shapes what can be observed, how it can be observed as well as the means of recording observations. Mathematics, being the science of reasoning and a pillar of our cognitive abilities, provides a mostly needed skeleton for structuring our inquiring and thinking about life. Thus, enlarging mathematics in directions suggested by empirical investigations in the life-sciences, even in the absence of experimental data, will be always of enormous value.

Two of the most conspicuous characteristics of life phenomena are their organisation and their dynamical character [4, 14, 55]. Thus, there is a call for the joint treatment of organisations and dynamical behaviour to viabilise theories in the life sciences [3, 56, 74]. Although there is yet no consensus about models and meaning of organisation, a model for organisation was proposed in [33] that includes some conspicuous characteristics of organisation, such as whole-part hierarchy and associativeness. According to it, graphs are the simplest instances of organisation, immediately followed by hyper-graphs. Moreover, the properties and behaviour of organisations at a higher level (a less detailed level) in the whole-part hierarchy of organisations depend only partially on the properties of their parts [34]. That is, parts which properties remain within appropriate bounds are equally adequate for building organisations at a higher organisational level.

What is currently called reductionistic approach or, less controversially, bottom-up approach requires the full description of parts or components to study a system. Among life phenomena, this results on a pile of information difficult to handle [23]. Reverting this requires the ability to confidently describe phenomena without going to the very bottom level of organisations. Seeing living entities as lively organisations [33, 34], and confidently relinquishing the necessity of thoroughly enrolling every detail, requires understanding what sort of components could give rise to the same organisation at a higher level. The suggested relationship between properties of graphs (organisations) and their dynamic behaviour is of utmost relevance to further investigate organisations, and to find good observables to unveil them in life phenomena [55, 58]. Knowledge about graph related partitions of dynamical spaces and their properties shall bring understanding concerning which dynamics is feasible for an observed organisation. Furthermore, this knowledge will support many ongoing efforts to organise biological observations of relational-dynamical character at all scales, see [35] and references therein.

Besides, the matching of dynamics to the next complex instance of organisations, hyper-graphs, leads directly into games, delayed and distributed signals, and differential games as modes of interaction.

Acknowledgements The maturing of ideas exposed in this text as well as their expression owes much to our collaboration and discussion with colleagues in the UK. These collaborations have been made possible by funds from PCI/LNCC/MCT program under contracts number 170089/2008-8 and 170114/2008-2, and we gratefully acknowledge it. Special thanks to Sean McKee, Michael Grinfeld, Ian Stewart and Steven Webb, from University of Strathclyde, and Jean-Marc Schwartz, from University of Manchester. One of us wish also to acknowledge the opportunity to participate in the EM3B workshop (footnote 9). Though uncommon, we wish to thank the authors and curators of LaTeX system TUG distribution and Rogério Brito (IME/USP) for the algorithm/algorithmic package without which this paper would not be so easily composed. Last, but not least, the authors would like to thank the editors of this volume, particularly prof. M.M. Peixoto, for the opportunity to present this paper.

References

1. Alon, U.: An introduction to systems biology: design principles of biological circuits. Mathematical and Computational Biology. Chapman & Hall/CRC, London, (2007)
2. Arnol'd, V.I.: Polymathematics: is mathematics a single science or a set of arts? Available online at http://www.pdmi.ras.ru/arnsem/Arnold/arn-papers.html, on server since 10-Mar-1999 (1999)
3. Ashby, W.R.: Principles of the self-organizing system. In: von Foster, H., Zopf, G.W. Jr. (eds.) Principles Of Self-Organization: Transactions of the University of Illinois Symposium, pp. 255–278, London, UK. University of Illinois, Pergamon Press, New York (1962)
4. Auger, P.: The methods and limits of scientific knowledge. In: Heisenberg, W., et al. (eds.) On Modern Physics. Clarkson N. Potter, Inc. Publisher, New York, NY (1961)
5. Bellman, R.E., Smith, C.P.: Simulation in Human Systems: Decision-Making in Psychotherapy, volume 22 of Publications in Operations Research. Wiley-Interscience, Wiley, New York, NY (1973)
6. Berge, C.: Graphs and Hypergraphs. North-Holland, Amsterdam (1973)
7. Bollobás, B.: Random Graphs. Cambridge studies in advanced mathematics, 2nd edn. Cambridge University Press, Cambridge (2001)
8. Cinquin, O., Demongeuot, J.: Positive and negative feedback: Striking a balance between necessary antagonists. J. Theor. Biol. **216**, 229–241 (2002)
9. Cracium, G., Feinberg, M.: Multiple equilibria in complex chemical reaction networks: extensions to entrapped species models. IEE Proc. Syst. Biol. **153**(4), 179–186 (2006)
10. Cracium, G., Feinberg, M.: Multiple equilibria in complex chemical reaction networks: Ii. the species-reaction graph. SIAM J. Appl. Math. **66**(4), 1321–1338 (2006)
11. de Jong, H. Modeling and simulating genetic regulatory systems: a literature review. J. Comput. Biol. **9**, 67–103 (2002)
12. Domijan, M., Kirkilionis, M.: Bistability and oscillation in chemical reaction networks. Math. Biol. **59**, 467–501 (2009)
13. Domijan, M., Kirkilionis, M.: Graph theory and qualitative analysis of reaction networks. Netw. Heterog. Media **3**(2), 295–322 (2009)
14. Ellner, S.P., Guckenheimer, J.: Dynamic Models in Biology. Princeton University Press, Princeton and Oxford (2006)
15. Feigenbaum, M.: Reflections of the polish masters. Los Alamos Science, Fall (1982)
16. Goldstein, H.: Classical Mechanics. Addison-Wesley Physics Books. Addison-Wesley Publishing Company, Inc., Reading, MA (1950) (2nd Printing, 1964)
17. Gouzé, J.-L.: Positive and negative circuits in dynamical systems. J. Biol. Syst. **6**(1), 11–15 (1998)
18. Gries, D.: The Science of Programming. Texts and Monographs in Computer Science. Springer, New York, NY (1981)
19. Gross, J.L., Tucker, T.W.: Topological Graph Theory. Wiley-Interscience Series in Discrete Mathematics and Optimization. A Wiley-Interscience Publication, Wiley, New York, NY (1987)
20. Guckenheimer, J., Holmes, P.: Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields, volume 42 of Applied Mathematical Sciences. Springer, New York, NY, 2nd printing, revised and corrected edition (1983)
21. Haefner, J.W.: Modeling Biological Systems: Principles and Applications, 2nd edn. Springer, New York, NY (2005)

22. Hanson, N.R.: Patterns of Discovery: An Inquiry into the Conceptual Foundations of Science. Cambridge University Press, Cambridge (1958)
23. Harris-Warrick, R.M., Marder, E., Selverston, A.I., Moulins, M. (eds.): Dynamic Biological Networks: The Stomatogastric Nervous System. Computational Neuroscience. A Bradford Book – The MIT Press, Cambridge, MA (1992)
24. Hayes, B.: Graph theory in practice: Part i. Am. Sci. **88**(1), 9–13 (2000)
25. Hayes, B.: Graph theory in practice: Part ii. Am. Sci. **88**(2), 104–109 (2000)
26. Hinrichsen, D., Pritchard, A.J.: Mathematical Systems Theory I: Modelling, State Space Analysis, Stability and Robustness, volume 48 of Texts in Applied Mathematics. Springer, Berlin (2005)
27. Kalman, R.E., Falb, P.L., Arbib, M.A.: Topics in Mathematical System Theory. McGraw-Hill Book Co., Inc., New York, NY (1969)
28. Kasinski, A., Levine, J.: A fast graph theoretic algorithm for the feedback decoupling problem of nonlinear systems. In: Fuhrmann, P.A. (ed.) Mathematical Theory of Networks and Systems. Proceedings of the MTNS-83 International Symposium, Beer Sheva, Israel, June 20–24, 1983, volume 58 of Lecture Notes in Control and Information Systems, pp. 550–562. Springer, Berlin (1984)
29. Kaufman, M., Soulé, C., Thomas, R.: A new necessary condition on interaction graphs for multistationarity. J. Theor. Biol. **248**, 675–685 (2007)
30. Kirkilionis, M., Sbano, L.: An averaging principle for combined interaction graphs, part i: Connectivity and applications to genetic switches. arXiv:0803.0635v2 [q-bio.MN] (2008)
31. Kitto, K.: High end complexity. Int. J. Gen. Syst. **37**(6), 689–714 (2008)
32. Klir, G.J.: Facets of Systems Science 2nd edn. Plenum Press, New York, NY, 2001 (1991)
33. Kritz, M.V.: Biological organizations. In: Mondaini, R. (ed.) Proceedings of the IV Brazilian Symposium on Mathematical and Computational Biology — BIOMAT IV, Rio de Janeiro, (2005). e-papers Editora
34. M. V. Kritz. Biological information and knowledge. Relatório de P&D 23/2009, LNCC/MCT, Petrópolis, December 2009
35. M. V. Kritz, M. T. dos Santos, S. Urrutia, and J.-M. Schwartz. Organizing metabolic networks: Cycles in flux distribution. J. Theor. Biol. **265**(3), 250–260, August 2010
36. Ljung, L., Glad, T.: Modeling of Dynmical Systems. Prentice Hall Information and System Sciences Series. PTR Prentice Hall, Upper Sadle River, NJ (1994)
37. Lorenz, E.N.: Deterministic non-periodic flows. J. Atmos. Sci. **20**, 130–141 (1963)
38. Mac Lane, S.: Mathematics Form and Function. Springer, New York, NY (1986)
39. MacDonald, N.: Trees and Networks in Biological Models. Wiley-Interscience/Wiley, Chichester (1983)
40. Maki, D.P., Thompson, M.: Mathematical Models and Applications: with Emphasis on the Social, Life and Management Sciences. Prentice-Hall, Inc., Englewood Cliffs, NJ (1973)
41. Mayr, E.: This is Biology: The Science of the Living World. Belknap Press/Havard University Press, Cambridge, MA (1997)
42. Mesarovic, M.D., Takahara, Y.: General Systems Theory: Mathematical Foundations, volume 113 of Mathematics in Science and Engineering. Academic, New York, NY (1975)
43. Miller, J.G.: Living Systems. McGraw-Hill Book Co., Inc., New York, NY (1978)
44. Mirkin, B.G., Rodin, S.N.: Graphs and Genes, volume 11 of Lecture Notes in Biomathematics. Springer, Berlin (1984)
45. Murota, K.: Systems Analysis by Graphs and Matroids: Structural Solvability and Controlability, volume 3 of Algorithms and Combinatorics. Springer, London, UK (1987)
46. Nagel, E.: The Structure of Science: Problems in the Logic of Scientific Explanation, 2nd edn. Hackett Publishing Company, Indianapolis, IN (1979)
47. Nishizeki, T., Chiba, N.: Planar Graphs: Theory and Algorithms, volume 32 of Annals of Discrete Mathematics. North-Holland, Amsterdam (1988)
48. Palsson, B.: Systems Biology: Properties of Reconstructed Networks. Cambridge University Press, Cambridge (2006)
49. Peixoto, M.M.: Structural stability on two-dimensional manifolds. Topology **1**, 101–120 (1962)

50. Penrose, R.: The Road to Reality: A Complete Guide to the Laws of the Universe. Alfred A. Knopf, N. York, 1st american edition (2005)
51. Remy, E., Ruet, P.: From minimal signed circuits to the dynamics of boolean regulatory networks. Bioinformatics **24**, i220–i226 (2008)
52. Richard, A.: Positive circuits and maximal number of fixed points in discrete dynamical systems. Discrete Appl. Math. **157**(15), 3281–3288 (2009)
53. Richard, A., Comet, J.-P.: Necessary conditions for multistationarity in discrete dynamical systems. Discrete Appl. Math. **155**(18), 2403–2413 (2007)
54. Roberts, F.S.: Graph Theory and Its Application to Problems of Society, volume 29 of Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania (1978)
55. Rosen, R.: Biological systems as organizational paradigms. Int. J. Gen. Syst. **1**(3), 165–174 (1974)
56. Rosen, R.: Observation and biological systems. Bull. Math. Biol. **39**, 663–678 (1977)
57. Rosen, R.: Some comments on activation and inhibition. Bull. Math. Biol. **41**(3), 427–445 (1979)
58. Rosen, R.: Anticipatory Systems. Pergamon Press, New York, NY (1985)
59. Rosenblueth, A., Wiener, N.: The role of models in science. Philos. Sci. 1946:316–321, 1946.
60. Rosenblueth, A., Wiener, N., Bigelow, J.: Behavior, purpose and teleology. Philos. Sci. **10**(1), 316 (1943)
61. Schmidt, G., Ströhlein, T.: Relations and Graphs: Discrete Mathematics for Computer Scientists. EACTS Monagrphs on Theoretical Computer Science. Springer, Berlin (1993)
62. Shannon, R.E.: Systems Simulation: the Art and Science. Prentice-Hall, Inc., Englewood Cliffs, NJ (1975)
63. Simon, H.A.: The Sciences of the Artificial, 3rd edn. MIT, Cambridge, MA (1996)
64. Snoussi, E.H.: Necessary conditions for multistationarity and stable periodicity. J. Biol. Syst. **6**(1), 3–9 (1998)
65. Soulé, C.: Graphic requirements for multistationarity. Complexus **1**(3), 123–133 (2003)
66. Sparrow, C.: The Lorenz Equations: Bifurcations, Chaos, and Strange Attractors, volume 41 of Applied Mathematical Sciences. Springerg, New York, NY (1982)
67. Strogatz, S.H.: Exploring complex networks. Nature **410**, 268–276 (2001)
68. Tarjan, R.E.: Enumeration of the elementary circuits of a directed graph. SIAM J. Comput. **2**, 211–216 (1973)
69. Thieffry, D.: Dynamical roles of biological regulatory circuits. Brief. Bioinform. **8**(4), 220–225 (2007)
70. Thomas, R.: On the relation of logical structure of systems and their ability to generate multiple steady states and sustained oscillations. In: Dela-Dora, J., Demongeuot, J., Lacolle, B. (eds.) Numerical Methods in the Study of Critical Phenomena, volume 9 of Springer Series of Sinergetics, pp. 180–193. Springer, Berlin (1981)
71. Ulanowicz, R.E.: Identifying the structure of cycling in ecosystems. Math. Biosci. **65**, 219–237 (1983)
72. Ulanowicz, R.E.: Growth and Development: Ecosystems Phenomenology. Springer, New York, NY (1986)
73. von Bertalanffy, L.: General Systems Theory. Allen Lane The Penguin Press, London (1971)
74. Waddington, C.H. (ed.): Towards a Theoretical Biology, vol. 1–4. Edinburgh University Press, Edinburgh (1968–1972)
75. Weaver, W.: Science and complexity. Am. Sci. **36**, 536–544 (1948)
76. Weinberg, G.M.: An Introduction to General Systems Thinking. Dorset Hause Publishing, New York, NY, silver anniversary edition (2001) (1st ed. 1975).
77. Wigner, E.P.: Events, laws of nature, and invariance principles. Science **145**(3636), 995–999 (1964)

Chapter 38
The Dynamics of Scalar Fields in Cosmology

José Pedro Mimoso

Abstract Scalar field models have been a focal point in cosmology during the last two decades or so. They play a central role in inflationary models, they arise in modified gravity theories that extend Einstein's General Relativity (GR) which are, often, quantum motivated, and, recently, they have been put forward as a dark component of the universe. Here we analyse their dynamics in the framework of isotropic cosmologies presenting an unified approach that encompasses models both in Einstein's GR and more general metric gravity theories. We perform a qualitative analysis of the major dynamical features of these models, discussing the existence of asymptotic regimes and their connection to a classification of the scalar fields potentials and couplings. A special interest is devoted to the interplay between scalar fields and matter which gives rise to scaling behaviour.

38.1 Scalar Field Models

Scalar field cosmological models have been a focal point in cosmology during the last two decades or so. They play a central role in inflationary models, they arise in modified gravity theories [1] that extend Einstein's General Relativity (GR), and, recently, they have been put forward as a dark component of the universe [2].

We consider scalar field cosmologies in a unified representation which includes both General Relativity and non-minimal coupling theories. In this framework the basic action takes the form

$$S = \int \mathrm{d}^4x \, \sqrt{-g} \, \Big[(R - 2V(\varphi)) - g^{ab} \varphi_{,a} \varphi^{,b} + 16\pi G_* L_m(\psi^m, m(\varphi) g_{ab}) \Big], \tag{38.1}$$

J.P. Mimoso
Centro de Astronomia e Astrofísica da Universidade de Lisboa & Departamento de Física, Faculdade de Ciências, Universidade de Lisboa, Ed. C8 Campo Grande, 1749-016 Lisboa, Portugal

e-mail: jpmimoso@cii.fc.ul.pt

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_38,

where g_{ab} represents the space-time metric, R the corresponding Ricci curvature scalar, G_* is the gravitational constant, and L_m is the lagrangean describing the matter fields ψ^m. The dependence of L_m on the metric multiplied by the factor $m(\varphi)$ implies an effective coupling between the matter fields and the φ scalar field. In the so-called minimal coupling case this factor takes a constant value, and we recover GR. In generalized metric theories of gravity, where m is a function of φ, the interaction between the usual matter fields and φ implies that test particles do not satisfy the equivalence principle. However the consistency of Einstein's equations is preserved as it is the combination of the scalar and matter fields that should obey the principle. We have $\nabla_b(T^{ab}_{(\varphi)} + T^{ab}_{(m)}) = 0$, where $T^{ab}_{(\varphi)}$ and $T^{ab}_{(m)}$ are the energy-momentum tensors associated with the scalar field and the matter fields, respectively.

Here we briefly perform a unified qualitative analysis of the major dynamical features of these scalar field cosmological models [3–7]. In what follows we shall restrict to the homogeneous and isotropic universes given by the Friedmann–Robertson–Walker (FRW) metric

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + a^2(t)\left[\frac{\mathrm{d}r^2}{1-k\,r^2} + r^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)\right], \qquad (38.2)$$

where $k = 0, \pm 1$ distinguishes the curvature of the spatial hypersurfaces. We assume that the matter sources are a perfect fluid characterized by the equation of state $p = (\gamma - 1)\,\rho$, where $0 \leq \gamma \leq 2$ is a constant, and a self-interacting scalar field φ, with the potential $V(\varphi)$, which is coupled to the perfect fluid if $m(\varphi) \neq const$ (NB: In what follows we shall adopt units that set $8\pi G_* = 1$).

Introducing the new time variable $N = \ln a$ and the dimensionless density parameters $x^2 = \dfrac{\dot{\varphi}^2}{6H^2}$ and 0 $y^2 = \dfrac{V(\varphi)}{3H^2}$, as well as the expansion normalized curvature term $K = k/(aH)^2$, the Einstein field equations become a fourth order, autonomous dynamical system

$$\begin{aligned} x' = {} & Kx - 3x - \sqrt{\frac{3}{2}}\left(\frac{\partial_\varphi V}{V}\right) y^2 + \frac{3}{2}x\left[2x^2 + \gamma\,(1 - x^2 - y^2 + K)\right] \\ & -\sqrt{\frac{3}{2}}\left(\frac{\partial_\varphi m}{m}\right)(1 - x^2 - y^2 + K) \end{aligned} \qquad (38.3)$$

$$y' = \sqrt{\frac{3}{2}}\left(\frac{\partial_\varphi V}{V}\right) xy + \frac{3}{2}y\left[2x^2 + \gamma\,(1 - x^2 - y^2 + K)\right] - yK\,, \qquad (38.4)$$

$$K' = -2K\left(1 - \frac{3}{2}\left[2x^2 + \gamma\,(1 - x^2 - y^2 + K)\right]\right) - K^2 \qquad (38.5)$$

$$\varphi' = \sqrt{6}\,x\,, \qquad (38.6)$$

where we have used $\rho/3H^2 = \Omega_m = 1 - x^2 - y^2 + K^2$.

GR models correspond to the case where $\partial_\varphi \ln m(\varphi) = 0$, and Brans–Dicke models are characterised by an exponential coupling m, i.e., by $\partial_\varphi \ln m(\varphi) = \alpha_0$

(as well as by $V(\varphi) = 0$ in BD original version, but we do not require it here). Notice also that the dynamical system is akin to that of a decaying scalar field [7]. The crucial point regarding the qualitative study of general models with scalar fields lies in the φ'-equation, since it allows the consideration of arbitrary choices of $V(\varphi)$ and of $m(\varphi)$ [3]. We compactify the phase space of (38.3)–(38.5) by considering $\varphi \in \Re \cup \{\infty\}$, and distinguish the fixed points arising at finite values of φ, which require $x = 0$, from those at $\varphi = \infty$ (which we shall denote φ_∞) that require either $x = 0$ or $\psi = \varphi^{-1} = 0$. The $K = 0$ and the $y = 0$ subspaces are invariant manifolds.

For $K = 0$ the system reduces to the (38.3), (38.4), (38.6) and we find the following fixed points. At finite φ

- $x = 0$, $y = 1$, φ_0 where $\partial_\varphi V(\varphi_0) = 0$. This case corresponds to de Sitter solutions, dominated by the scalar field, and arise at maxima or minima of the potential V. Notice that $\partial_\varphi m$ may be different from 0. These solutions are attractors at minima of V and repellors at maxima of V.
- $x = y = 0$, φ_0 where $\partial_\varphi m = 0$. The second case corresponds to matter dominated solutions with $V = 0$ at maxima or minima of m. Note however that the system (38.3)–(38.6) is singular when $V = 0$, translating the fact that the variables in use are not regular. In the original variables, this second class of solutions exists provided that V has also an extremum, or is identically zero.

At φ_∞, the finite φ solutions may exist also at $\varphi = \infty$ provided that V or m are asymptotically flat. However, other solutions may show up at $\varphi = \infty$, if both V and m are asymptotically exponential. In fact in this case the equilibrium in $\psi = 1/\varphi$ is trivially satisfied at $\psi = 0$ [3]. Defining $W = \sqrt{\frac{3}{2}} \lim_{\varphi\to\infty} \frac{\partial_\varphi V}{V}$ and $Z = \sqrt{\frac{3}{2}} \lim_{\varphi\to\infty} \frac{\partial_\varphi m}{m}$. The exponential asymptotic behaviour of $m(\varphi)$ amounts to having a late-time Brans–Dicke behaviour. One finds the following fixed points (see Fig. 38.1) [8–10]

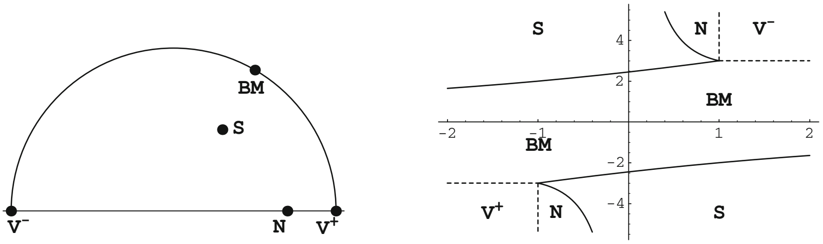


Fig. 38.1 In the left part of this figure we represent the fixed points at φ_∞ phase-plane for V and m asymptotically exponential. In the right part of the figure, we exhibit the regions in parameter space that correspond to the asymptotic Brans–Dicke behaviour. The horizontal axis is the Z-axis and the vertical axis is the W-axis. The regions S and BM are separated by the line $9\gamma/2 - W^2 + WZ = 0$, and the separations between the S and N regions is given by the line $2Z^2 - 2WZ + 9\gamma(2-\gamma)/2 = 0$

- $x_\infty^{V\pm} = \pm 1$, $y^{V\pm} = 0$, that lie on the intersection of the invariant lines $x^2+y^2 = 1$ and $y = 0$. These points correspond to the vacuum solutions of Brans–Dicke theory found by [11, 12].
- $x_\infty^{BM} = -W/3$, which lies on the invariant line $x^2 + y^2 = 1$, provided $|W| < 3$. These solutions (BM) correspond to those found in [13].
- $x_\infty^{N} = -2Z/(3(2-\gamma))$, which lies on the invariant line $y = 0$ and exists if $|Z| < 3(2-\gamma)/2$. These solutions (N) correspond to the matter dominated solutions found by [12, 14].
- $x_\infty^{S} = \frac{3\gamma/2}{Z-W}$, $(x_\infty^S)^2 + (y_\infty^S)^2 = \frac{3x_\infty^S + Z}{Z-W}$, that lies in the interior of the phase space domain, provided that

$$0 < (x_\infty^S)^2 + (y_\infty^S)^2 = \frac{9\gamma/2 + Z(Z-W)}{(Z-W)^2} < 1.$$

These are scaling solutions (S) (see [5, 8–10] and references therein).

For the $K \neq 0$ case, as $K' = 0$ for $K = 0$, we see from the linearization of the system that

$$\frac{\partial K'}{\partial K} = -2 + 2\,Q_{K=0}\,, \quad Q_{K=0} \equiv \frac{3}{2}\left[2x^2 + \gamma\,(1 - x^2 - y^2)\right], \tag{38.7}$$

which shows that the $K = 0$ subspace is stable whenever $Q < 1$ which is precisely the condition for inflationary behaviour. Thus inflation is required to guarantee both the stability of the $K = 0$ asymptotic solutions and the attraction to GR, whenever the latter applies.

Summing up the qualitative analysis of the dynamical system (38.3)–(38.5) associates exact solutions to a classification of the fixed points, and also allows the consideration of how extended gravity theories dynamically relate to GR [5].

Acknowledgements The author is grateful to his collaborators Ana Nunes, Diego Pavón and Tiago Charters for allowing him to draw freely from their works with him, and also wishes to thank Filipe Mena for his kind hospitality. This work was partially supported by the grant PTDC/FIS/102742/2008 (Portugal).

References

1. Will, C.M.: Theory and Experiment in Gravitational Physics. Cambridge University Press, Cambridge, England (1993)
2. Copeland, E.J., Sami, M., Tsujikawa, S.: Int. J. Mod. Phys. D **15**, 1753 (2006). arXiv:hep-th/0603057
3. Nunes, A., Mimoso, J.P.: Phys. Lett. B **488**, 423 (2000). arXiv:gr-qc/0008003
4. Mimoso, J.P., Nunes, A.M.: Phys. Lett. A **248**, 325 (1998)
5. Mimoso, J.P., Nunes, A.: Astrophys. Space Sci. **283**, 661 (2003)
6. T.C. Charters, A. Nunes and J. P. Mimoso, Class. Quant. Grav. **18**, 1703 (2001). arXiv:gr-qc/0103060

7. Mimoso, J.P., Nunes, A., Pavon, D.: Phys. Rev. D **73**, 023502 (2006). arXiv:gr-qc/0512057
8. Holden, D., Wands, D.: Phys. Rev. D **61**, 043506 (2000)
9. Uzan, J.P.: Phys. Rev. D **59**, 123510 (1999)
10. Amendola, L.: Phys. Rev. D **60**, 043501 (1999)
11. O'Hanlon, J., Tupper, B.O.J.: Nuovo Cimento **7**, 305 (1972)
12. Barrow, J.D., Mimoso, J.P.: Phys. Rev. D **50**, 3746 (1994)
13. Barrow, J.D., Maeda, K.: Nucl. Phys. **B341**, 294 (1990)
14. Nariai, H.: Progr. Theor. Phys. **40**, 49 (1968)

Chapter 39
The Dynamics of the Nash Map for 2 By 2 Games

R.A. Becker, S.K. Chakrabarti, W. Geller, B. Kitchens, and M. Misiurewicz

Abstract We describe the dynamics of a better response map from the space of mixed strategy profiles to itself, used by Nash to prove the existence of equilibrium points for finite games. We do it for the case of 2 players and 2 pure strategies. The maps are classified, according to the dynamics, as dominant strategy, elliptic or hyperbolic.

39.1 Introduction

In one of J. Nash's proofs of the existence of equilibrium points for finite games [5] he defined a *better response* map from the space of mixed strategy profiles to itself. The better response map changes each player's strategy in a way that improves the player's payoff with respect to the other players' strategies at the present time. We will call it the *Nash map*. He observed that a point is an equilibrium point for the game if and only if it is a fixed point for the better response map. He then applied the Brouwer Fixed Point Theorem to show a Nash equilibrium exists for any finite game.

While Nash did not introduce his map in order to investigate its dynamics, the idea of studying its dynamics appears for example in [6].

R.A. Becker
Department of Economics, Indiana University, Bloomington, IN 47405, USA
e-mail: becker@indiana.edu

S.K. Chakrabarti
Department of Economics, Indiana University – Purdue University Indianapolis, 425 University Blvd., Indianapolis, IN 46202, USA
e-mail: imxl100@iupui.edu

W. Geller, B. Kitchens, and M. Misiurewicz (✉)
Department of Mathematical Sciences, Indiana University – Purdue University Indianapolis, 402 N. Blackford Street, Indianapolis, IN 46202, USA
e-mail: wgeller@math.iupui.edu, kitchens@math.iupui.edu, mmisiure@math.iupui.edu

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_39,

The iterates of the better response map produce a dynamical system on the space of strategy profiles. We investigate this dynamical system in the case of 2 by 2 games (i.e., 2 players and 2 pure strategies). In papers [1, 2], and [4] we studied particular cases of these maps. Here we summarize those results, present an overview and classify all 2 by 2 games with respect to the Nash map.

We divide the nondegenerate 2 by 2 games into three classes. In the first and simplest class the game has a unique globally attracting fixed point and at least one player has a dominant strategy. This class contains the game of Prisoner's Dilemma. The second class consists of the games having elliptic dynamics, where there is a unique repelling fixed point and the trajectories of all other points move around the fixed point. This class contains the game of Matching Pennies [1, 4]. The third class consists of the games where the dynamics is hyperbolic, in this case meaning that there are three fixed points, two of which are attracting and the third is either repelling or a saddle point. In these games there may be regions consisting of points that are attracted to other periodic orbits. This class contains the Coordination game and the game of Chicken [2].

39.2 2 By 2 Games

Let a matrix

$$\begin{bmatrix} (a, a') & (b, b') \\ (c, c') & (d, d') \end{bmatrix}$$

with ordered pairs of real numbers as entries define a two person, two strategy (2 by 2) game with X the row player and Y the column player. The two players have strategies given by the probability vectors $\boldsymbol{x} = (x, 1 - x)$ and $\boldsymbol{y} = (y, 1 - y)$, respectively. The payoff matrices for X and Y are

$$R_x = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \qquad \text{and} \qquad R_y = \begin{bmatrix} a' & b' \\ c' & d' \end{bmatrix}.$$

The expected payoff for X is $\boldsymbol{x} R_x \boldsymbol{y}^T$ and the expected payoff for Y is $\boldsymbol{x} R_y \boldsymbol{y}^T$. The idea of a better response map is that each player adds weight to a pure strategy if it will improve his payoff against the opposing players' present strategy.

Thus, we define

$$\begin{aligned} t_x &= \quad x \quad + \max\{0, (e_1 - \boldsymbol{x}) R_x \boldsymbol{y}^T\}, \\ t_{1-x} &= (1 - x) + \max\{0, (e_2 - \boldsymbol{x}) R_x \boldsymbol{y}^T\}, \\ t_y &= \quad y \quad + \max\{0, \boldsymbol{x} R_y (e_1 - \boldsymbol{y})^T\}, \\ t_{1-y} &= (1 - y) + \max\{0, \boldsymbol{x} R_y (e_2 - \boldsymbol{y})^T\}, \end{aligned}$$

with e_1, e_2 the standard basis vectors. The vectors (t_x, t_{1-x}) and (t_y, t_{1-y}) must be normalized to produce the new strategies. Therefore, we define the *Nash map* by the formula

$$(\boldsymbol{x}, \boldsymbol{y}) \mapsto \left(\left(\frac{t_x}{t_x + t_{1-x}}, \frac{t_{1-x}}{t_x + t_{1-x}}\right), \left(\frac{t_y}{t_y + t_{1-y}}, \frac{t_{1-y}}{t_y + t_{1-y}}\right)\right).$$

All information is contained in the *essential Nash map* on the unit square

$$n = (n_1, n_2) : [0,1]^2 \to [0,1]^2,$$

defined by

$$n_1(x, y) = \frac{t_x}{t_x + t_{1-x}} = \frac{x + \max\{0, (e_1 - \boldsymbol{x})R_x \boldsymbol{y}^T\}}{1 + \max\{0, (e_1 - \boldsymbol{x})R_x \boldsymbol{y}^T\} + \max\{0, (e_2 - \boldsymbol{x})R_x \boldsymbol{y}^T\}},$$

$$n_2(x, y) = \frac{t_y}{t_y + t_{1-y}} = \frac{y + \max\{0, \boldsymbol{x} R_y (e_1 - \boldsymbol{y})^T\}}{1 + \max\{0, \boldsymbol{x} R_x (e_1 - \boldsymbol{y})^T\} + \max\{0, \boldsymbol{x} R_y (e_2 - \boldsymbol{y})^T\}}$$

with $\boldsymbol{x}$ and $\boldsymbol{y}$ as before. It is clear from the definition that in this setting the essential Nash map is a continuous map of the unit square into itself.

If we let $[r]^+ = \max\{0, r\}$, $[r]^- = \max\{0, -r\}$ and $\alpha = a - c$, $\beta = b - d$, $\gamma = a' - b'$, $\delta = c' - d'$, the essential Nash map reduces to

$$n_1(x, y) = \frac{x + (1 - x)[\alpha y + \beta(1 - y)]^+}{1 + (1 - x)[\alpha y + \beta(1 - y)]^+ + x[\alpha y + \beta(1 - y)]^-},$$

$$n_2(x, y) = \frac{y + (1 - y)[\gamma x + \delta(1 - x)]^+}{1 + (1 - y)[\gamma x + \delta(1 - x)]^+ + y[\gamma x + \delta(1 - x)]^-}.$$

Consequently, essential Nash maps arising from 2 by 2 games form a four parameter family of piecewise rational, continuous functions of the unit square to itself. In what follows we assume a nondegeneracy condition which is that $\alpha, \beta, \gamma, \delta \neq 0$.

We prove that the games fall into three classes determined by the type of dynamics that occur for the essential Nash map. In the first class, a player has a dominant strategy and there is a single pure strategy equilibrium. A *dominant strategy* occurs for a player when one strategy has a higher expected payoff than the other without regard to the other player's strategy. Player X has a dominant strategy if α and β have the same sign and player Y has a dominant strategy if γ and δ have the same sign. This is the case for the Prisoner's Dilemma game which is defined by the matrix

$$\begin{bmatrix} (-3,-3) & (-1,-4) \\ (-4,-1) & (-2,-2) \end{bmatrix}.$$

In this case $\alpha = \beta = \gamma = \delta = 1$ and both X and Y have a dominant strategy which is the first strategy and corresponds to "talking".

In the second class, which contains the game of Matching Pennies, there is a single repelling mixed strategy Nash equilibrium.

For the third class, which includes the games of Coordination and Chicken, there are two attracting pure strategy Nash equilibria and one mixed strategy equilibrium which may have several different types of behavior but is never attracting.

Theorem 39.2.1. *There are three classes of nondegenerate 2 by 2 games. They are:*

1. *Dominant strategy. When either α and β or γ and δ have the same sign, there is a single pure strategy Nash equilibrium which is a globally attracting fixed point for the essential Nash map. When α and β have the same sign, player X has a dominant strategy and when γ and δ have the same sign, player Y has a dominant strategy.*
2. *Elliptic dynamics. When α and δ share one sign while β and γ have the other sign, the essential Nash map is one-to-one, and there is one mixed strategy Nash equilibrium which is a repelling fixed point for the essential Nash map.*
3. *Hyperbolic dynamics. When α and γ share the same sign while β and δ have the other sign, there are two attracting pure strategy and one mixed strategy Nash equilibrium points. The mixed strategy equilibrium is never attracting.*

In Example 39.5.1 we define a one parameter family of maps whose members fall into the third case. In Theorem 39.5.1 we show that the set of nonwandering points of each map in this family consists of a finite number of periodic points. Except for three special values of the parameter and except for the points $(0, 0)$ and $(1, 1)$, all nonwandering points are hyperbolic periodic points. For this reason we refer to the dynamics as *hyperbolic*.

If we interchange the order of X's strategies, the values of α, β, γ and δ in the essential Nash map are changed. The signs of α and β are switched and the roles of γ and δ are interchanged. Thus, we can always assume $\alpha > 0$.

We prove the statements made in Theorem 39.2.1 in the next sections.

39.3 Dominant Strategy

This class contains the game of Prisoner's Dilemma.

Proof (of Theorem 39.2.1 (1)). Here one or both players have a dominant strategy. Assume that α, β and γ are positive. Then $\alpha y + \beta(1 - y)$ is positive and the map n strictly increases the x-coordinate of a point (x, y) unless $x = 1$, in which case the x coordinate is unchanged. When x is sufficiently close to 1, the term $\gamma x + \delta(1 - x)$ is positive and the map n strictly increases the y-coordinate of a point (x, y) unless $y = 1$, in which case the y coordinate is unchanged. The point $(1, 1)$ is a fixed point which attracts every point in the unit square. The same reasoning applies to the other possibilities. □

In the next two classes of maps α and β have opposite signs and γ and δ have opposite signs.

An examination of the essential Nash map when $x, y \in (0,1)$ shows the following.

$$\begin{aligned}
&n_1(x,y) > x \text{ if } \alpha y + \beta(1-y) > 0,\\
&n_1(x,y) = x \text{ if } y = \tfrac{\beta}{\beta-\alpha},\\
&n_1(x,y) < x \text{ if } \alpha y + \beta(1-y) < 0,\\[1ex]
&n_2(x,y) > y \text{ if } \gamma x + \delta(1-x) > 0,\\
&n_2(x,y) = y \text{ if } x = \tfrac{\delta}{\delta-\gamma},\\
&n_2(x,y) < y \text{ if } \gamma x + \delta(1-x) < 0.
\end{aligned}$$

Observe that $0 < \frac{\beta}{\beta-\alpha}, \frac{\delta}{\delta-\gamma} < 1$. The two lines $x = \frac{\delta}{\delta-\gamma}$ and $y = \frac{\beta}{\beta-\alpha}$ divide the square into four quadrants. We refer to the quadrants by compass points, *NE*, *SE*, *SW* and *NW*. We will refer to the lines separating the quadrants as borders. The intersection of the two lines is the point $(\frac{\delta}{\delta-\gamma}, \frac{\beta}{\beta-\alpha})$, which is the unique fixed point of the map in the interior of the square.

By the preceding observation the formula for n_1 depends on whether the term $\alpha y + \beta(1-y)$ is positive or negative and the formula for n_2 depends on whether $\gamma x + \delta(1-x)$ is positive or negative.

When $\alpha y + \beta(1-y) \geq 0$ the formula for n_1 becomes

$$n_1^+(x,y) = \frac{x + (1-x)[\alpha y + \beta(1-y)]}{1 + (1-x)[\alpha y + \beta(1-y)]}$$

and when $\alpha y + \beta(1-y) \leq 0$ the formula for n_1 becomes

$$n_1^-(x,y) = \frac{x}{1 - x[\alpha y + \beta(1-y)]}.$$

Similarly, when $\gamma x + \delta(1-x) \geq 0$ the formula for n_2 becomes

$$n_2^+(x,y) = \frac{y + (1-y)[\gamma x + \delta(1-x)]}{1 + (1-y)[\gamma x + \delta(1-x)]}$$

and when $\gamma x + \delta(1-x) \leq 0$ the formula for n_2 becomes

$$n_2^-(x,y) = \frac{y}{1 - y[\gamma x + \delta(1-x)]}.$$

We need all eight partial derivatives of these four functions.

$$\frac{\partial n_1^+}{\partial x} = \frac{1}{(1 + (1-x)[\alpha y + \beta(1-y)])^2},$$

$$\begin{aligned}
\frac{\partial n_1^+}{\partial y} &= \frac{(\alpha-\beta)(1-x)^2}{(1+(1-x)[\alpha y+\beta(1-y)])^2},\\
\frac{\partial n_1^-}{\partial x} &= \frac{1}{(1-x[\alpha y+\beta(1-y)])^2},\\
\frac{\partial n_1^-}{\partial y} &= \frac{(\alpha-\beta)x^2}{(1-x[\alpha y+\beta(1-y)])^2},\\
\frac{\partial n_2^+}{\partial x} &= \frac{(\gamma-\delta)(1-y)^2}{(1+(1-y)[\gamma x+\delta(1-x)])^2},\\
\frac{\partial n_2^+}{\partial y} &= \frac{1}{(1+(1-y)[\gamma x+\delta(1-x)])^2},\\
\frac{\partial n_2^-}{\partial x} &= \frac{(\gamma-\delta)y^2}{(1-y[\gamma x+\delta(1-x)])^2},\\
\frac{\partial n_2^-}{\partial y} &= \frac{1}{(1-y[\gamma x+\delta(1-x)])^2}.
\end{aligned}$$

From this point we assume $\alpha > 0$, $\beta < 0$, while γ and δ have opposite signs.

39.4 Elliptic Dynamics

This class contains the game of Matching Pennies.

This is the case where $\alpha, \delta > 0$ and $\beta, \gamma < 0$. The essential Nash map is a rational map in each quadrant. The four maps are defined as follows: $NEn = (n_1^+, n_2^-)$, $SEn = (n_1^-, n_2^-)$, $SWn = (n_1^-, n_2^+)$ and $NWn = (n_1^+, n_2^+)$.

The eight partial derivatives have the following signs.

$$\frac{\partial n_1^+}{\partial x}, \frac{\partial n_1^-}{\partial x}, \frac{\partial n_2^+}{\partial y}, \frac{\partial n_2^-}{\partial y} > 0,$$

$$\begin{aligned}
\frac{\partial n_1^+}{\partial y} &> 0, \text{ except when } x = 1,\\
\frac{\partial n_1^-}{\partial y} &> 0, \text{ except when } x = 0,\\
\frac{\partial n_2^+}{\partial x} &< 0, \text{ except when } y = 1,\\
\frac{\partial n_2^-}{\partial x} &< 0, \text{ except when } x = 0.
\end{aligned}$$

Lemmas 39.4.1 and 39.4.4 constitute Theorem 39.2.1 (2).

Lemma 39.4.1. *In the class of elliptic dynamics the essential Nash map is a homeomorphism from the unit square to its image.*

Before proving Lemma 39.4.1, we state two useful lemmas.

Lemma 39.4.2. *Suppose that C is a Jordan curve in the plane and D is the closure of the region bounded by C. Let f be a map from D into $\mathbb{R}^2$, which is a local homeomorphism on the interior of D and a homeomorphism from C onto its image. Then:*

1. *$f(C)$ is a Jordan curve and the image of the interior of D is the region bounded by $f(C)$.*
2. *f is a homeomorphism of D onto its image.*

Proof. The first statement is an elementary exercise in point set topology. The second statement was proved in [3]. □

Lemma 39.4.3. *Let Γ_1 and Γ_2 be two smooth curves in $\mathbb{R}^2$, intersecting transversally at p. Let D be a disk centered at p, such that $\Gamma_1 \cup \Gamma_2$ divides it into the four connected sets A_1, A_2, A_3, A_4, counting in the counterclockwise direction. Let F_1, F_2, F_3, F_4 be smooth maps from D to $\mathbb{R}^2$ with positive Jacobians. Assume that for each pair $i, j \in \{1, 2, 3, 4\}$ the maps F_i and F_j agree on the intersection of the closures of A_i and A_j, and define a map F as F_i on the closure of A_i, $i = 1, 2, 3, 4$. Then there is a neighborhood of p on which F is a homeomorphism onto its image.*

Proof. Choose a disk E, centered at p, contained in D, and so small that the sets $A_i \cap E$ are connected and each F_i restricted to E is a homeomorphism onto its image. Let B be a disk centered at $F(p)$ and contained in the intersection of those four images. Fix $\varepsilon > 0$.

Take a small piece of Γ_1 or Γ_2 with one endpoint p which is mapped by F to a curve $\gamma \subset B$ joining $F(p)$ with the boundary of B. If B is sufficiently small, then by smoothness of Γ_i and F_j, the curve γ is contained in a cone with vertex at $F(p)$ and angular width less than ε. Inside this cone there is a half-line tangent to γ at $F(p)$.

Let us perform this operation with all four germs of $\Gamma_1 \cup \Gamma_2$ at p. As we take two consecutive germs (in the counterclockwise direction), the oriented angle between them is less than π, so the oriented angle between their images under F (which are their images under one of the maps F_j) is less than π. As we go around, the sum of the image angles is less than 4π, and therefore, since we have to end up where we started, it has to be 2π. Since those angles are positive, if ε is sufficiently small, the four cones from the preceding paragraph are pairwise disjoint (except for their common vertex). Thus, the sets $F(A_j \cap E) \cap B = F_j(A_j \cap E) \cap B$ are four disjoint sector-like regions. Therefore, since additionally we know that each $F_j(A_j \cap E)$ is a homeomorphism onto its image, we see that F restricted to $E \cap F^{-1}(B)$ is a homeomorphism onto its image. □

Proof (of Lemma 39.4.1). First we show the image of the boundary of the square is a Jordan curve. There are eight "break" points on the boundary. Starting at the top and going clockwise around the boundary of square we examine their images. They are $n(\frac{\delta}{\delta-\gamma}, 1) = (\frac{\delta-\alpha\gamma}{\delta-(1+\alpha)\gamma}, 1)$, $n(1, 1) = (1, \frac{1}{1-\gamma})$, $n(1, \frac{\beta}{\beta-\alpha}) = (1, \frac{-\beta}{\alpha-(1-\gamma)\beta})$,

$n(1,0) = (\frac{1}{1-\beta}, 0)$, $n(\frac{\delta}{\delta-\gamma}, 0) = (\frac{\delta}{-\gamma+(1-\beta)\delta}, 0)$, $n(0,0) = (0, \frac{\delta}{1+\delta})$, $n(0, \frac{\beta}{\beta-\alpha}) = (0, \frac{\beta}{\beta-(1+\delta)\alpha})$ and $n(0,1) = (\frac{\alpha}{1+\alpha}, 1)$.

First examine the image of the boundary segment $\{(x,1) : \frac{\delta}{\delta-\gamma} \leq x \leq 1\}$. It is a curve connecting the points $(\frac{\delta-\alpha\gamma}{\delta-(1+\alpha)\gamma}, 1)$ and $(1, \frac{1}{1-\gamma})$. Since $\frac{\partial n_1^+}{\partial x} > 0$ and $\frac{\partial n_2^-}{\partial x} < 0$, except when $x = 0$, the curve minus the endpoints is contained in the rectangle where $\frac{\delta-\alpha\gamma}{\delta-(1+\alpha)\gamma} < x < 1$ and $\frac{1}{1-\gamma} < y < 1$. The essential Nash map is a homeomorphism of the curve onto its image because the partial derivatives are not 0.

Next consider the image of the boundary segment $\{(1,y) : \frac{\beta}{\beta-\alpha} \leq y \leq 1\}$. The image is the boundary segment $\{(1,y) : \frac{-\beta}{\alpha-(1-\gamma)\beta} \leq y \leq \frac{1}{1-\gamma}\}$. Since $\frac{\partial n_2^-}{\partial y} > 0$ the essential Nash map restricted to the boundary segment is a homeomorphism onto its image.

Continuing in this way we examine the image of each boundary segment and observe that the essential Nash map is a homeomorphism of each onto its image and that the images do not intersect except at the endpoints. Consequently, the essential Nash map restricted to the boundary of the square is a homeomorphism onto its image which is a Jordan curve.

The essential Nash map n is a local homeomorphism on the interior of each quadrant since the Jacobian is always positive. At the point $(\frac{\delta}{\delta-\gamma}, \frac{\beta}{\beta-\alpha})$ apply Lemma 39.4.3 to see that n is a local homeomorphism. If p is a point on a border between two quadrants we can still apply Lemma 39.4.3 to see that n map is a local homeomorphism; just introduce a spurious curve Γ' and use F_1 and F_2 again in the new regions.

Now apply Lemma 39.4.2 to prove that n is a homeomorphism from the square onto its image. □

Lemma 39.4.4. *In the class of elliptic dynamics the point* $(\frac{\delta}{\delta-\gamma}, \frac{\beta}{\beta-\alpha})$ *is the unique fixed point for the essential Nash map and it is exponentially repelling.*

Proof. Consider the northern boundary of the northeast quadrant. There $n(\frac{\delta}{\delta-\gamma}, 1) = (\frac{\delta-\alpha\gamma}{\delta-(1+\alpha)\gamma}, 1)$ and $n_2^-(x,1) < 1$ when $x > \frac{\delta}{\delta-\gamma}$. There are no fixed points for n on the northern boundary.

On the eastern boundary $n(1,y) = (1, \frac{y}{1-\gamma y})$ but $\frac{y}{1-\gamma y} < y$. Hence, there are no fixed points for n on the eastern boundary of the northeast quadrant. The same reasoning applies to the other quadrants, so there are no fixed points for the essential Nash map on the boundary of the square. The interior fixed point is the only fixed point for the map.

The derivative of *NEn* at the fixed point is

$$D_{2NE} = \begin{bmatrix} 1 & \frac{(\alpha-\beta)\gamma^2}{(\gamma-\delta)^2} \\ \frac{(\gamma-\delta)\beta^2}{(\alpha-\beta)^2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & G \\ -B & 1 \end{bmatrix},$$

where

$$G = \frac{(\alpha-\beta)\gamma^2}{(\gamma-\delta)^2} \qquad \text{and} \qquad B = -\frac{(\gamma-\delta)\beta^2}{(\alpha-\beta)^2} = \frac{(\delta-\gamma)\beta^2}{(\alpha-\beta)^2}.$$

In this case $G, B > 0$.

The determinant is $1 + BG > 1$, the characteristic polynomial is $c(\lambda) = \lambda^2 - 2\lambda + (1 + BG)$ and the eigenvalues are $\lambda = 1 \pm \sqrt{-BG}$. The eigenvalues are complex so the matrix is conjugate to a rotation through $-\theta$, where $\theta = \tan^{-1}\sqrt{BG}$, followed by an expansion of $\varepsilon = \sqrt{1+BG}$.

The derivative maps the ellipse $Bx^2 + Gy^2 = c^2$ to the ellipse $Bx^2 + Gy^2 = (1 + BG)c^2 = (\varepsilon c)^2$. These form a family of ellipses that is invariant under the derivative map. If the matrix is divided by ε the new map is just a rotation on each ellipse.

Let

$$D = \frac{(\alpha-\beta)\delta^2}{(\gamma-\delta)^2} > 0 \qquad \text{and} \qquad A = \frac{(\delta-\gamma)\alpha^2}{(\alpha-\beta)^2} > 0.$$

Then the derivative of NWn at the fixed point is

$$D_{2NW} = \begin{bmatrix} 1 & \frac{(\alpha-\beta)\gamma^2}{(\gamma-\delta)^2} \\ \frac{(\gamma-\delta)\alpha^2}{(\alpha-\beta)^2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & G \\ -A & 1 \end{bmatrix}$$

with ellipses $Ax^2 + Gy^2 = c^2$.

The derivative of SWn at the fixed point is

$$D_{2SW} = \begin{bmatrix} 1 & \frac{(\alpha-\beta)\delta^2}{(\gamma-\delta)^2} \\ \frac{(\gamma-\delta)\alpha^2}{(\alpha-\beta)^2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & D \\ -A & 1 \end{bmatrix}$$

with ellipses $Ax^2 + Dy^2 = c^2$.

Finally, the derivative of SEn at the fixed point is

$$D_{2SE} = \begin{bmatrix} 1 & \frac{(\alpha-\beta)\delta^2}{(\gamma-\delta)^2} \\ \frac{(\gamma-\delta)\beta^2}{(\alpha-\beta)^2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & D \\ -B & 1 \end{bmatrix}$$

with ellipses $Bx^2 + Dy^2 = c^2$.

For $c \geq 0$ let E_c be the piecewise ellipse in $\mathbb{R}^2$ defined as follows. In the NE quadrant it is $Bx^2 + Gy^2 = c^2$ which goes from $(c/\sqrt{B}, 0)$ to $(0, c/\sqrt{G})$. In the NW quadrant it is $Ax^2 + Gy^2 = c^2$ which goes from $(0, c/\sqrt{G})$ to $(-c/\sqrt{A}, 0)$. In the SW quadrant it is $Ax^2 + Dy^2 = c^2$ which goes from $(-c/\sqrt{A}, 0)$ to $(0, -c/\sqrt{D})$. Then in the SE quadrant it is $Bx^2 + Dy^2 = c^2$ which goes from $(0, -c/\sqrt{D})$ back to $(c/\sqrt{B}, 0)$. In the NW quadrant the graph of E_c has as its derivative $\frac{dy}{dx} = -\frac{Ax}{Gy}$ and in the NE quadrant $\frac{dy}{dx} = -\frac{Bx}{Gy}$. At the point $(0, c/\sqrt{G})$

both derivatives are 0. The derivatives also agree at the other three points on E_c that are on the borders of the quadrants. The piecewise ellipse E_c is a compact, strictly convex, differentiable, closed curve.

Define a continuous, piecewise linear map J from $\mathbb{R}^2$ to itself by letting D_{2NE} act on the first quadrant, D_{2NW} act on the second quadrant, D_{2SW} on the third and D_{2SE} on the fourth.

Consider J on the first quadrant. For a point (x, y) on E_c the vector $(J - I)(x, y) = (Gy, -Bx)$ at (x, y) is orthogonal to the gradient of the function $Bx^2 + Gy^2$. This means it is a tangent vector to E_c pointing in the clockwise direction. Consequently, J maps the point outside of E_c. The same is true for every point on E_c. So $J(E_c)$ is a compact, closed curve that lies strictly outside of E_c. There is a positive distance from E_c to $J(E_c)$.

The family of piecewise ellipses $\{E_c : c \geq 0\}$ fills out $\mathbb{R}^2$ so that for $(x, y) \in \mathbb{R}^2$ there is a unique $c \in [0, +\infty)$ with E_c containing (x, y). Use this family to define a size function s on $\mathbb{R}^2$ by $s((x, y)) = c$ where E_c contains (x, y). This is a measure of distance from (x, y) to the origin. It may not be a norm because for $t < 0$ it may be that $s((tx, ty))$ and $|t|s((x, y))$ are not equal. However, when $t \geq 0$ we do have $s((tx, ty)) = ts((x, y))$. This size function is uniformly equivalent to the Euclidean norm in the sense that there are constants $c_1, c_2 > 0$ such that $c_1 s(x, y) \leq ||(x, y)|| \leq c_2 s(x, y)$. Let $\lambda = \min\{s((x, y)) : (x, y) \in J(E_1)\}$. Then $\lambda > 1$ and for all $(x, y) \in \mathbb{R}^2$, $s(J((x, y))) \geq \lambda s((x, y))$. This means the origin is an exponentially repelling fixed point for J.

The map J is the piecewise linear approximation to the essential Nash map at the fixed point $(\frac{\delta}{\delta-\gamma}, \frac{\beta}{\beta-\alpha})$. It follows that the fixed point is exponentially repelling for the essential Nash map. □

Example 39.4.1. We consider a one parameter family of maps where $\delta = \alpha$ and $\beta = \gamma = -\alpha$. The game of Matching Pennies is defined by the matrix

$$\begin{bmatrix} (1,-1) & (-1,1) \\ (-1,1) & (1,-1) \end{bmatrix}.$$

The essential Nash map for Matching Pennies occurs in this family when $\alpha = 2$. It is investigated in detail in [1].

For this family the four terms in the essential Nash map simplify to

$$\begin{aligned} n_1^+(x, y) &= \frac{x + \alpha(1-x)(2y-1)}{1 + \alpha(1-x)(2y-1)}, \\ n_1^-(x, y) &= \frac{x}{1 - \alpha x(2y-1)}, \\ n_2^+(x, y) &= \frac{y + \alpha(1-y)(1-2x)}{1 + \alpha(1-y)(1-2x)}, \\ n_2^-(x, y) &= \frac{y}{1 - \alpha y(1-2x)]}. \end{aligned}$$

The essential Nash map commutes with rotation by $\pi/2$, meaning that if R_θ is rotation in $\mathbb{R}^2$ by the angle θ then $n = R_{-\pi/2} \circ n \circ R_{\pi/2}$. The fixed point is always $(1/2, 1/2)$ and it is repelling. All four derivatives at the fixed point are

$$\begin{bmatrix} 1 & \frac{\alpha}{2} \\ -\frac{\alpha}{2} & 1 \end{bmatrix}.$$

The characteristic polynomial is $c(\lambda) = \lambda^2 - 2\lambda + (1 + \alpha^2/4)$. The eigenvalues are $1 \pm (\alpha/2)i$ so all four derivatives are conjugate to rotation by $\theta = -\tan^{-1}(\alpha/2)$ followed with multiplication by $\sqrt{1 + \alpha^2/4}$.

There is another interpretation of this family as using an *index of caution* for the Nash map in the game of Matching Pennies. By putting a parameter in the definition of the Nash map it is possible to adjust the caution with which each player responds. In fact, any game gives rise to a family of Nash maps using an index of caution. This is equivalent to varying the payoff matrices and applying the original Nash map. Define

$$\begin{aligned} t_{x,\mu} &= \quad x \quad + \mu \max\{0, (e_1 - \boldsymbol{x}) R_x \boldsymbol{y}^T\}, \\ t_{1-x,\mu} &= (1 - x) + \mu \max\{0, (e_2 - \boldsymbol{x}) R_x \boldsymbol{y}^T\}, \\ t_{y,\mu} &= \quad y \quad + \mu \max\{0, \boldsymbol{x} R_y (e_1 - \boldsymbol{y})^T\}, \\ t_{1-y,\mu} &= (1 - y) + \mu \max\{0, \boldsymbol{x} R_y (e_2 - \boldsymbol{y})^T\}, \end{aligned}$$

The parameterized Nash map becomes

$$(\boldsymbol{x}, \boldsymbol{y}) \mapsto \left(\left(\frac{t_{x,\mu}}{t_{x,\mu} + t_{1-x,\mu}}, \frac{t_{1-x,\mu}}{t_{x,\mu} + t_{1-x,\mu}} \right), \left(\frac{t_{y,\mu}}{t_y(\mu) + t_{1-y,\mu}}, \frac{t_{1-y,\mu}}{t_{y,\mu} + t_{1-y,\mu}} \right) \right).$$

This is the usual Nash map for the game with $\alpha = 2\mu$.

The essential Nash map for the game of Matching Pennies has an invariant topological circle. On this circle there is an orbit of period 8 which attracts all points in the square except the fixed point [1]. This is illustrated in Fig. 39.1. In the figure

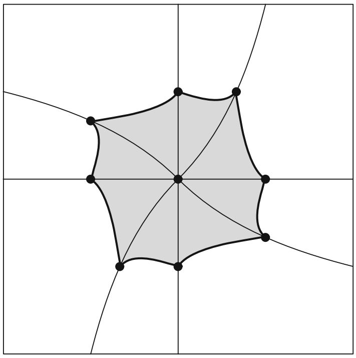


Fig. 39.1 The essential Nash map for the game of Matching Pennies

the shaded region is the intersection of all images of the square. The boundary of the shaded region is the invariant circle and the marked points make up the orbit of period 8.

This leads to the following two conjectures. In what follows, "circle" means a topological circle, that is, a Jordan curve.

Conjecture 39.4.1. In the elliptic case the essential Nash map has an invariant circle that attracts every point in the square except the unique fixed point. If the rotation number of the map restricted to its invariant circle is irrational, then this map is conjugate to an irrational rotation. If the rotation number is rational, there is one attracting periodic orbit.

Figures 39.2 and 39.3 are computer simulations of the Matching Pennies family and show what appear to be invariant attracting circles (if the rotation number is rational, we see only an attracting periodic orbit). In [4] we proved that in this family the conjecture is true when μ is sufficiently small. Moreover, the size of the invariant circle is of order μ, after rescaling the circles by $1/\mu$ the circles converge to a geometric circle of radius $3\pi/32$, and the rotation number of the map restricted to the circle is of order μ.

The second weaker conjecture would follow from the first.

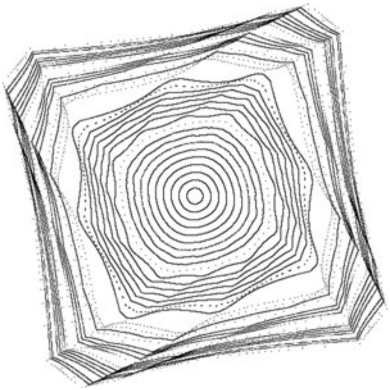


Fig. 39.2 Attractors for various values of μ for the Matching Pennies family; the phase space

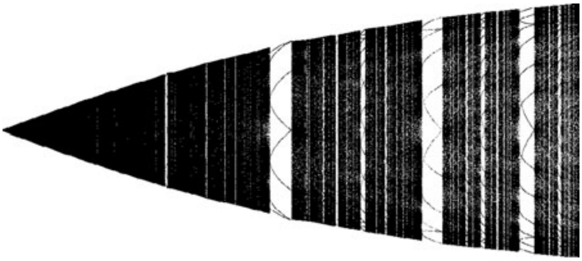


Fig. 39.3 Dependence of the attractor on μ for the Matching Pennies family. The horizontal axis is μ and the vertical axis x

Conjecture 39.4.2. In the elliptic case there is no chaotic behavior in the essential Nash map. In particular, the topological entropy of the map is zero.

39.5 Hyperbolic Dynamics

This class contains the games of Coordination and Chicken.

This is the case where $\alpha, \gamma > 0$ and $\beta, \delta < 0$. The essential Nash map is a rational map in each quadrant. The four maps are defined as follows: $NEn = (n_1^+, n_2^+)$, $SEn = (n_1^-, n_2^+)$, $SWn = (n_1^-, n_2^-)$ and $NWn = (n_1^+, n_2^-)$.

Lemmas 39.5.1 and 39.5.2 constitute Theorem 39.2.1 (3).

Lemma 39.5.1. *In the class of hyperbolic dynamics the points* $(0, 0)$ *and* $(1, 1)$ *are attracting fixed points for the essential Nash map and they are the only fixed points on the boundary of the square. Except for the point* $(\frac{\delta}{\delta-\gamma}, \frac{\beta}{\beta-\alpha})$ *every point in the southwest quadrant is attracted to* $(0, 0)$ *and except for the point* $(\frac{\delta}{\delta-\gamma}, \frac{\beta}{\beta-\alpha})$ *every point in the northeast quadrant is attracted to* $(1, 1)$.

Proof. Consider the northeast quadrant and observe that $n_1^+(1, y) = 1$ and $n_2^+(x, 1) = 1$. When $x \neq 1$ and $y > \frac{\beta}{\beta-\delta}$, n_1^+ increases the x-coordinate of the image point and when $y \neq 1$ and $x > \frac{\delta}{\delta-\gamma}$, n_2^+ increases the y-coordinate of the image point. Consequently, $(1, 1)$ is the only fixed point of NEn and every point except $(\frac{\delta}{\delta-\gamma}, \frac{\beta}{\beta-\alpha})$ in the northeast quadrant is attracted to it. The same statement holds for $(0, 0)$ and the southwest quadrant.

Now consider the boundary of the southeast quadrant. On the southern boundary n_1^- decreases the x coordinate of the image point. On the eastern boundary n_2^+ increases the y-coordinate of the image point. There are no fixed points on the boundary of the southeast quadrant and the same reasoning applies to the northwest quadrant. □

We compute the derivatives at the mixed strategy fixed point as we did in the class of elliptic dynamics. The derivative of NEn at the fixed point $(\frac{\delta}{\delta-\gamma}, \frac{\beta}{\beta-\alpha})$ is

$$D_{3NE} = \begin{bmatrix} 1 & \frac{(\alpha-\beta)\gamma^2}{(\gamma-\delta)^2} \\ \frac{(\gamma-\delta)\alpha^2}{(\alpha-\beta)^2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & G \\ A' & 1 \end{bmatrix},$$

where

$$G = \frac{(\alpha-\beta)\gamma^2}{(\gamma-\delta)^2} > 0 \qquad \text{and} \qquad A' = \frac{(\gamma-\delta)\alpha^2}{(\alpha-\beta)^2} > 0.$$

Let

$$D = \frac{(\alpha-\beta)\delta^2}{(\gamma-\delta)^2} > 0 \qquad \text{and} \qquad B' = \frac{(\gamma-\delta)\beta^2}{(\alpha-\beta)^2} > 0.$$

Then the derivative of *NWn* at the fixed point is

$$D_{3NW} = \begin{bmatrix} 1 & \frac{(\alpha-\beta)\gamma^2}{(\gamma-\delta)^2} \\ \frac{(\gamma-\delta)\beta^2}{(\alpha-\beta)^2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & G \\ B' & 1 \end{bmatrix}.$$

The derivative of *SWn* at the fixed point is

$$D_{3SW} = \begin{bmatrix} 1 & \frac{(\alpha-\beta)\delta^2}{(\gamma-\delta)^2} \\ \frac{(\gamma-\delta)\beta^2}{(\alpha-\beta)^2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & D \\ B' & 1 \end{bmatrix}.$$

Finally, the derivative of *SEn* at the fixed point is

$$D_{3SE} = \begin{bmatrix} 1 & \frac{(\alpha-\beta)\delta^2}{(\gamma-\delta)^2} \\ \frac{(\gamma-\delta)\alpha^2}{(\alpha-\beta)^2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & D \\ A' & 1 \end{bmatrix}.$$

Note that all four matrices are strictly positive so that the Perron–Frobenius Theorem applies to each. In two dimensions the Perron–Frobenius Theorem says that a strictly positive matrix has two real eigenvalues, with one, the *Perron eigenvalue*, strictly larger in absolute value than the other and positive. The eigendirection for the Perron eigenvalue lies in the interior of the first and third quadrants. The eigendirection for the other eigenvalue lies in the interior of the second and fourth quadrants. The theorem also states that the Perron eigenvalue is at least as large as the smallest row sum. In our case each row sum is greater than one so the Perron eigenvalue is greater than 1. Of course, this also can be directly computed.

Lemma 39.5.2. *In the class of hyperbolic dynamics the point* $(\frac{\delta}{\delta-\gamma}, \frac{\beta}{\beta-\alpha})$ *is the unique mixed strategy fixed point for the essential Nash map and it is never attracting.*

Proof. Lemma 39.5.1 shows that there are no mixed strategy fixed points on the boundary of the square and we previously observed that the point $(\frac{\delta}{\delta-\gamma}, \frac{\beta}{\beta-\alpha})$ is the unique mixed strategy fixed point in the interior of the square. The derivatives of *NEn* and *SWn* at the fixed point obey the Perron–Frobenius Theorem as noted. This implies that the fixed point has an expanding direction in the northeast quadrant and an expanding direction in the southwest quadrant. □

The smaller eigenvalue for the derivative of *SEn* is $1-\sqrt{A'D}$ and the smaller eigenvalue for the derivative of *NWn* is $1-\sqrt{B'G}$. These eigenvalues are less than 1 but can be positive, zero or negative in any combination.

Example 39.5.1. We consider a one parameter family of maps where $\gamma = \alpha$ and $\beta = \delta = -\alpha$. The game of Coordination is defined by the matrix

$$\begin{bmatrix} (1,1) & (-1,-1) \\ (-1,-1) & (1,1) \end{bmatrix}.$$

The essential Nash map for Coordination occurs in this family when $\alpha = 2$. The game of Chicken is defined by the matrix

$$\begin{bmatrix} (-10,-10) & (5,-5) \\ (-5,5) & (0,0) \end{bmatrix}.$$

Interchanging the ordering for X's choices one observes that the essential Nash map for Chicken occurs in this family when $\alpha = 5$.

The four terms that occur in the essential Nash map simplify to the following:

$$\begin{aligned} n_1^+(x,y) &= \frac{x + \alpha(1-x)(2y-1)}{1+\alpha(1-x)(2y-1)}, \\ n_1^-(x,y) &= \frac{x}{1-\alpha x(2y-1)}, \\ n_2^+(x,y) &= \frac{y + \alpha(1-y)(2x-1)}{1+\alpha(1-y)(2x-1)}, \\ n_2^-(x,y) &= \frac{y}{1-\alpha y(2x-1)]}. \end{aligned}$$

The three fixed points for the essential Nash map are $(0,0)$, $(1,1)$ and $(1/2,1/2)$. The essential Nash map is symmetric about the main diagonal and the anti-diagonal, meaning that if $r_d(x,y) = (y,x)$ and $r_a(x,y) = (1-y, 1-x)$ then $n = r_d \circ n \circ r_d$ and $n = r_a \circ n \circ r_a$. All four derivatives at $(1/2,1/2)$ are

$$\begin{bmatrix} 1 & \frac{\alpha}{2} \\ \frac{\alpha}{2} & 1 \end{bmatrix}.$$

The derivative has two eigenvalues. The larger in modulus is $1 + \frac{\alpha}{2}$ with eigenvector $(1,1)^T$ and the smaller in modulus is $1 - \frac{\alpha}{2}$ with eigenvector $(1,-1)^T$. In [2] we prove the following theorem.

Theorem 39.5.1. *Let n be the essential Nash map described above.*

1. *For all α, $(0,0)$ and $(1,1)$ are orientation preserving, topologically attracting fixed points and $(1/2,1/2)$ is a fixed point. There are no other fixed points.*
2. *When $0 < \alpha \leq 1/2$, n is a homeomorphism onto its image. When $1/2 < \alpha$, n is not one-to-one.*
3. *For $0 < \alpha < 2$, $(1/2,1/2)$ is an orientation preserving fixed saddle point whose stable manifold is the anti-diagonal and unstable manifold is the diagonal (without $(0,0)$ and $(1,1)$).*
4. *For $2 < \alpha < 4$, $(1/2,1/2)$ is an orientation reversing fixed saddle point whose stable manifold is the anti-diagonal and unstable manifold is the diagonal (without $(0,0)$ and $(1,1)$).*
5. *For $\alpha \leq 4$, all points below the anti-diagonal are attracted to $(0,0)$ and all points above the anti-diagonal are attracted to $(1,1)$.*

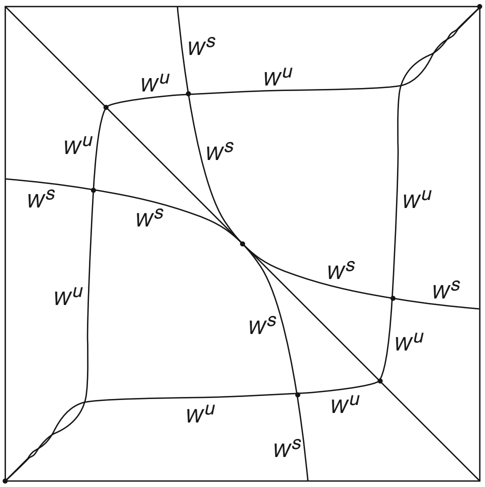


Fig. 39.4 Periodic points for $\alpha = 6$ for the family from Theorem 39.5.1

6. *For* $4 < \alpha < 2(1 + \sqrt{2})$*,* $(1/2, 1/2)$ *is an orientation reversing repelling fixed point. There is a saddle period two orbit on the anti-diagonal. One point of this orbit lies below* $(1/2, 1/2)$ *and its stable manifold is the part of the anti-diagonal with* $x < 1/2$*. The other point lies above* $(1/2, 1/2)$ *and its stable manifold is the part of the anti-diagonal with* $x > 1/2$*. All points below the anti-diagonal are attracted to* $(0, 0)$ *and all points above the anti-diagonal are attracted to* $(1, 1)$.
7. *For* $\alpha > 2(1 + \sqrt{2})$*,* $(1/2, 1/2)$ *is an orientation reversing repelling fixed point. There is an attracting orbit of period two on the anti-diagonal. There are two saddle period two orbits that follow the orbit of period two on the anti-diagonal. One saddle orbit lies below the anti-diagonal and one above. There are no other periodic points. Every other point is attracted to one of the periodic orbits mentioned (including* $(0, 0)$ *and* $(1, 1)$*). For illustration, see Fig. 39.4.*

Conjecture 39.5.1. In the hyperbolic case any periodic behavior that may occur already occurs in the family described in Example 39.5.1. This periodic behavior is explained by Theorem 39.5.1.

Conjecture 39.5.2. In the hyperbolic case there is no chaotic behavior in the essential Nash map. In particular, the topological entropy of the map is zero.

Acknowledgements M. Misiurewicz's research was partially supported by NSF grant DMS 0456526.

References

1. Becker, R.A., Chakrabarti, S.K., Geller, W., Kitchens, B., Misiurewicz, M.: Dynamics of the Nash map in the game of matching Pennies. J. Differ. Equ. Appl. **13**, 223–235 (2007)

2. Becker, R.A., Chakrabarti, S.K., Geller, W., Kitchens, B., Misiurewicz, M.: Hyperbolic Dynamics in Nash maps, to appear in Proceedings of ICDEA 2007
3. Derrick, W.R.: A condition under which a map is a homeomorphism. Am. Math. Mon. **80**, 554–555 (1973)
4. Geller, W., Kitchens, B., Misiurewicz, M.: Microdynamics for Nash maps, to appear in Discrete Contin. Dyn. Syst.
5. Nash, J.F.: Non-cooperative games. Ann. Math. **54**, 286–295 (1951)
6. Skyrms, B.: The Dynamics of Rational Deliberation. Harvard University Press, Cambridge, MA (1990)

Chapter 40
Resort Pricing and Bankruptcy

Alberto A. Pinto, Marta Faias, and Abdelrahim S. Mousa

Abstract We introduce a resort pricing model, where different types of tourists choose between different resorts. We study the influence of the resort prices on the choices of the different types of tourists. We characterize the coherent strategies of the tourists that are Nash equilibria. We find the prices that lead to the bankruptcy of the resorts and, in particular, their dependence on the characteristics of the tourists.

40.1 Introduction

The distribution of different types of tourists reaching a destination affects both the demand and supply side of the tourism industry. From the demand perspective, the choice of a particular destination will depend greatly on the beliefs of the agent about which kind of tourists will share the resort with him/her (see [4, 5]). On the supply side, resorts try to sell their destination based on reputation, and a large factor that determines the character and reputation of a resort is the type of tourists who frequent that resort (see [6]). Brida et al. [1] presented a tourism model where the choice of a resort by a tourist depends not only on the product offered by the resort,

A.A. Pinto (✉)
LIAAD-INESC Porto LA e Departamento de Matemática, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, 687, 4169-007, Portugal
and
Centro de Matemática e Departamento de Matemática e Aplicações, Escola de Ciências, Universidade do Minho, Campus de Gualtar, 4710-057 Braga, Portugal
e-mail: aapinto@fc.up.pt

M. Faias
Department of Mathematics, Universidade Nova de Lisboa, Lisbon, Portugal
e-mail: mcm@fct.unl.pt

A.S. Mousa
LIAAD-INESC Porto LA e Departamento de Matemática, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, 687, 4169-007, Portugal
e-mail: abed11@ritaj.ps

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_40,

but also depends on the characteristics of the other tourists present in the resort. In order to explore the effect the type of resident tourist has on other potential tourists selecting the same resort, they introduced a game theoretical model and described some relevant Nash equilibria. We add to the previous models the influence of resort prices on the tourist's choice of a resort (see [7]). We characterize the prices that lead to the bankruptcy of the resorts and, in particular, their dependence on the characteristics of the tourists.

40.2 Resort Pricing Model

The *resort pricing model* has two types $\mathbf{T} = \{t_1, t_2\}$ of tourists $i \in \mathbf{I}$ that have to choose between two goods or services. For instance, the tourists have to choose between spending their holidays in a beach resort B or in a mountain resort M, i.e. $r \in \mathbf{R} = \{B, M\}$. Let $n_q \geq 1$ be the number of tourists with type t_q. Let $\mathscr{P}$ be the *price vector* whose *coordinates* p^r indicates the *standard price* of the resort r for each tourist, independently of its type,

$$\mathscr{P} = (p^B, p^M).$$

Let $\mathscr{L}$ be the *preference location matrix* whose *coordinates* ω_q^r indicate how much the tourist, with type t_q, likes, or dislikes, to choose resort r

$$\mathscr{L} = \begin{pmatrix} \omega_1^B & \omega_1^M \\ \omega_2^B & \omega_2^M \end{pmatrix}.$$

The preference location matrix indicates, for each type, the resort that the tourists prefer, i.e. the tourists taste type (see [1, 8]).

Let $\mathscr{N}_r$ be the *preference neighbors matrix* whose *coordinates* $\alpha_{qq'}^r$ indicate how much the tourist, with type t_q, likes, or dislikes, that tourist, with type $t_{q'}$, chooses resort r

$$\mathscr{N}_r = \begin{pmatrix} \alpha_{11}^r & \alpha_{12}^r \\ \alpha_{21}^r & \alpha_{22}^r \end{pmatrix}.$$

The preference neighbors matrix indicates, for each type of tourists, whom they prefer to be with or to not be with at each resort, i.e. the tourists crowding type (see [1, 8]).

We describe the tourists' location by a *strategy map* $S : \mathbf{I} \to \mathbf{R}$ that associates to each tourist $i \in \mathbf{I}$ its location $S(i) \in \mathbf{R}$. Let $\mathbf{S}$ be the space of all strategies S. Given a strategy S, let $\mathscr{O}_S$ be the *strategic occupation matrix*, whose coordinates $l_q^r = l_q^r(S)$ indicate the number of tourists, with type t_q, that choose resort r

$$\mathscr{O}_S = \begin{pmatrix} l_1^B & l_1^M \\ l_2^B & l_2^M \end{pmatrix}.$$

The *strategic occupation vector* $\mathcal{V}_S$, associated to a strategy S, is the vector $(l_1, l_2) = (l_1^B(S), l_2^B(S))$. Hence, l_1 (resp. $n_1 - l_1$) is the number of tourists, with type t_1, who choose the resort B (resp. M). Similarly, l_2 (resp. $n_2 - l_2$) is the number of tourists, with type t_2, who choose the resort B (resp. M). The set $\mathbf{O}$ of all possible *occupation vectors* is

$$\mathbf{O} = \{(l_1, l_2) : 0 \le l_1 \le n_1 \quad \text{and} \quad 0 \le l_2 \le n_2\}.$$

Let $U_1 : \mathbf{R} \times \mathbf{O} \to \mathbb{R}$ the *utility function*, of the tourist with type t_1, be given by

$$\begin{aligned} U_1(B; l_1, l_2) &= -p^B + \omega_1^B + \alpha_{11}^B(l_1 - 1) + \alpha_{12}^B l_2 \\ U_1(M; l_1, l_2) &= -p^M + \omega_1^M + \alpha_{11}^M(n_1 - l_1 - 1) + \alpha_{12}^M(n_2 - l_2). \end{aligned}$$

Let $U_2 : \mathbf{R} \times \mathbf{O} \to \mathbb{R}$ the *utility function*, of the tourists with type t_2, be given by

$$\begin{aligned} U_2(B; l_1, l_2) &= -p^B + \omega_2^B + \alpha_{22}^B(l_2 - 1) + \alpha_{21}^B l_1 \\ U_2(M; l_1, l_2) &= -p^M + \omega_2^M + \alpha_{22}^M(n_2 - l_2 - 1) + \alpha_{21}^M(n_1 - l_1). \end{aligned}$$

Given a strategy $S \in \mathbf{S}$, the *utility* $U_i(S)$, of the tourist i with type $t_{p(i)}$, is given by $U_{p(i)}(S(i); l_1^B(S), l_2^B(S))$.

We note that, if the price can depend on the tourist type, then the prices can be encoded in the preference decision matrix and, therefore, the model can be studied as the yes–no decision model presented in [7].

Definition 40.1. A strategy $S^* : \mathbf{I} \to \mathbf{R}$ is a *Nash equilibrium* if, for every tourist $i \in \mathbf{I}$ and for every strategy S, with the property that $S^*(j) = S(j)$ for every tourist $j \in I \setminus \{i\}$, we have

$$U_i(S^*) \ge U_i(S).$$

A *coherent strategy*[1] is a strategy in which all tourists, with the same type, prefer to choose the same resort. A *coherent strategy* is described by a map $C : \mathbf{T} \to \mathbf{R}$ where for every tourist i, with type $t_{q(i)}$, $C(q(i))$ indicates its location. Hence, a coherent strategy $C : \mathbf{T} \to \mathbf{R}$ determines an unique strategy $S : \mathbf{I} \to \mathbf{R}$ given by $S(i) = C(q(i))$.

Let $x = \omega_1^B - \omega_1^M$ be the *horizontal relative location preference* of the tourists with type t_1 and let $y = \omega_2^B - \omega_2^M$ be the *vertical relative location preference* of the tourists with type t_2. Let $p = p^B - p^M$ be the *relative price*. Given a pair (x, y) of relative location preferences, the *Nash equilibrium prices interval* $P(R_1, R_2) = P(x, y; R_1, R_2)$ of a coherent strategy (R_1, R_2) is the set of all relative prices p for which the strategy (R_1, R_2) is a Nash equilibrium. Our aim is to determine and characterize all Nash equilibrium prices intervals.

[1] or equivalently, *no-split strategy* or *heard strategy*.

40.3 Nash Equilibrium Prices

We observe that there are four distinct *coherent* strategies:

- (B, B) *strategy* – all tourists choose the resort B
- (B, M) *strategy* – all tourists, with type t_1, choose the resort B, and all tourists, with type t_2, choose the resort M
- (M, B) *strategy* – all tourists, with type t_1, choose the resort M and all tourists, with type t_2, choose the resort B
- (M, M) *strategy* – all tourists choose the resort M

The *horizontal* $H(B, B)$ and *vertical* $V(B, B)$ *strategic thresholds* of the (B, B) strategy are given by

$$H(B, B) = -\alpha^B_{11}(n_1 - 1) - \alpha^B_{12} n_2 \quad \text{and} \quad V(B, B) = -\alpha^B_{22}(n_2 - 1) - \alpha^B_{21} n_1.$$

The *(B,B) Nash equilibrium prices interval* $P(B, B)$ is the semi-line

$$P(B, B) = \{p \in \mathbb{R} : p \le x - H(B, B) \quad \text{and} \quad p \le y - V(B, B)\}.$$

In the red half-plane of the upper left section of Fig. 40.1, for a given relative preferences pair (x, y), the first coordinate of the blue vector, i.e. the yellow horizontal projection, represents the maximum price in $P(B, B)$; in the green half-plane, for a given relative preferences pair (x, y), the second coordinate of the blue vector, i.e. the orange vertical projection, represents the maximum price in $P(B, B)$.

The horizontal $H(B, M)$ and vertical $V(B, M)$ *strategic thresholds* of the (B, M) strategy are given by

$$H(B, M) = -\alpha^B_{11}(n_1 - 1) + \alpha^M_{12} n_2 \quad \text{and} \quad V(B, M) = \alpha^M_{22}(n_2 - 1) - \alpha^B_{21} n_1.$$

The *(B,M) Nash equilibrium prices interval* $P(B, M)$ is the segment line (that can be empty)

$$P(B, M) = \{p \in \mathbb{R} : p \le x - H(B, M) \quad \text{and} \quad p \ge y - V(B, M)\}.$$

In the blue half-plane of the upper right section of Fig. 40.1, for a given relative preferences pair (x, y), the second coordinate of the blue vector, i.e. the orange vertical projection, represents the minimum price in $P(B, M)$ and the first coordinate of the blue vector, i.e. the yellow horizontal projection, represents the maximum price in $P(B, M)$; in the purple half-plane, there are no Nash equilibrium prices.

The horizontal $H(M, B)$ and vertical $V(M, B)$ *strategic thresholds* of the (M, B) strategy are given by

$$H(M, B) = \alpha^M_{11}(n_1 - 1) - \alpha^B_{12} n_2 \quad \text{and} \quad V(M, B) = -\alpha^B_{22}(n_2 - 1) + \alpha^M_{21} n_1.$$

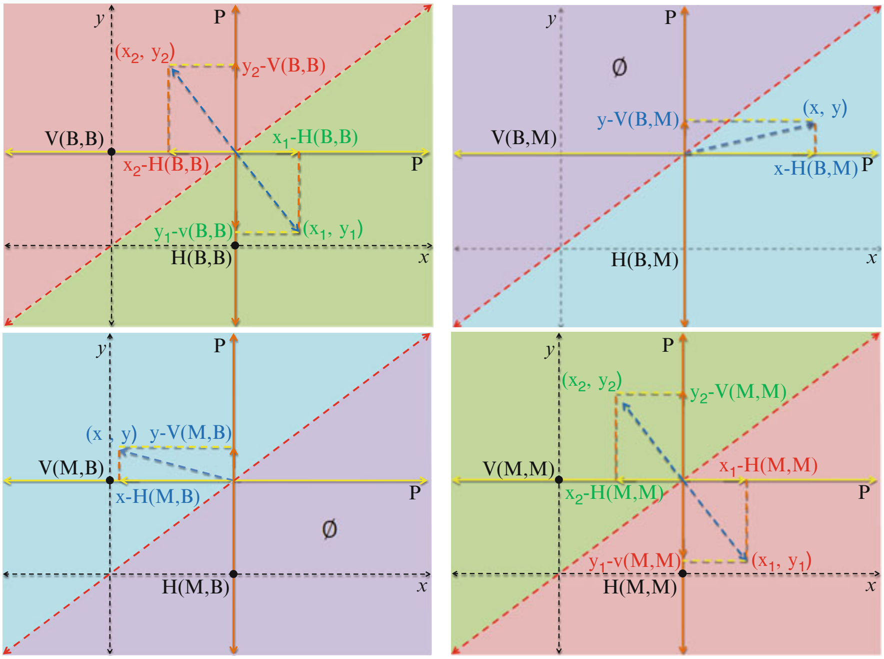


Fig. 40.1 Nash equilibrium prices

The *(M,B) Nash equilibrium prices interval* $P(M, B)$ is the segment line (that can be empty)

$$P(M, B) = \{p \in \mathbb{R} : p \geq x - H(M, B) \quad \text{and} \quad p \leq y - V(M, B)\}.$$

In the blue half-plane of the lower left section of Fig. 40.1, for a given relative preferences pair (x, y), the first coordinate of the blue vector, i.e. the yellow horizontal projection, represents the minimum price in $P(M, B)$ and the second coordinate of the blue vector, i.e. the orange vertical projection, represents the maximum price in $P(M, B)$; in the purple half-plane, there are no Nash equilibrium prices.

The horizontal $H(M, M)$ and vertical $V(M, M)$ strategic thresholds of the (M, M) strategy are given by

$$H(M, M) = \alpha_{11}^{M}(n_1 - 1) + \alpha_{12}^{M} n_2 \quad \text{and} \quad V(M, M) = \alpha_{22}^{M}(n_2 - 1) + \alpha_{21}^{M} n_1.$$

The *(M,M) Nash equilibrium prices interval* $P(M, M)$ is the semi-line

$$P(M, M) = \{p \in \mathbb{R} : p \geq x - H(M, M) \quad \text{and} \quad p \geq y - V(M, M)\}.$$

In the red half-plane of the lower right section of Fig. 40.1, for a given relative preferences pair (x, y), the first coordinate of the blue vector, i.e. the yellow horizontal

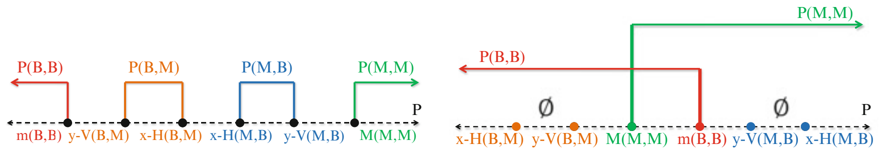


Fig. 40.2 Bankruptcy and competitive business Nash equilibria prices, where $M(M,M) = \max\{x - H(M,M), y - V(M,M)\}$ and $m(B,B) = \min\{x - H(B,B), y - V(B,B)\}$

projection, represents the minimum price in $P(M,M)$; in the green half-plane of the lower right section of Fig. 40.1, for a given relative preferences pair (x,y) the second coordinate of the blue vector, i.e. the orange vertical projection, represents the minimum price in $P(M,M)$.

40.4 Bankruptcy Nash Equilibrium Prices

Let the *coherent uniqueness Nash equilibria prices* be the regions $U(B,B) \subset P(B,B)$, $U(B,M) \subset P(B,M)$, $U(M,B) \subset P(M,B)$ and $U(M,M) \subset P(M,M)$, where for every point in these regions, there is a unique coherent Nash equilibrium. We call the prices in $U(B,B)$ the *bankruptcy* Nash equilibrium prices of the mountain resort M, because, for every price in $U(B,B)$, there are no tourists choosing the mountain resort M. Similarly, we call the prices in $U(M,M)$ the *bankruptcy* Nash equilibrium prices of the beach resort B, because, for every price in $U(M,M)$ there are no tourists choosing the beach resort B. We call the prices in $U(B,M)$ and $U(M,B)$ the *competitive business* Nash equilibrium prices, because, for every price in $U(B,M)$ and in $U(M,B)$, one type of tourist chooses the beach resort B and the other type of tourist chooses the mountain resort M. We note that, the bankruptcy Nash equilibria prices $U(B,B)$ and $U(M,M)$ are non-empty, but the competitive business Nash equilibrium price can be empty (see Fig. 40.2).

40.5 Conclusions

Small changes in the coordinates of the preference location matrix, which indicates the resort that the tourists prefer, and of the preference neighbors matrix, which indicates who the tourists prefer to be with in each resort, can create and annihilate competitive business Nash equilibrium prices and change the bankruptcy Nash equilibria prices.

Acknowledgements We thank LIAAD-INESC Porto LA, Calouste Gulbenkian Foundation, PRODYN-ESF, FEDER, POFC, POCTI and POSI by FCT and Ministério da Ciência e da

Tecnologia, and the FCT Pluriannual Funding Program of the LIAAD-INESC Porto LA and of the Research Centre of Mathematics of the University of Minho, for their financial support.

References

1. Brida, J., Defesa, M., Faias, M., Pinto, A.: A Tourist's Choice Model. Dynamics, Games and Science I. In: Peixoto, M., Pinto, A.A., Rand, D. (eds.) Proceedings in Mathematics series, Springer-Verlag, Chapter 10, 159–167 (2011)
2. Brida, J., Defesa, M., Faias, M., Pinto, A.A.: Strategic choice in tourism with differentiated crowding types. Econ. Bull. **30**(2), 1509–1515 (2010)
3. Conley, J.P., Wooders, M.H.: Tiebout economies with differential genetic types and endogenously chosen crowding characteristics. J. Econ. Theory **98**, 261–294 (2001)
4. Ferreira, F.A., Ferreira, F., Pinto, A.A.: 'Own' price influences in a Stackelberg leadership with demand uncertainty. Braz. J. Bus. Econ. **8**(1), 29–38 (2008)
5. Ferreira, F., Ferreira, F.A., Pinto, A.A.: Price-setting dynamical duopoly with incomplete information. In: Machado, J.A.T., et al. (eds.) Nonlinear Science and Complexity, Springer, Berlin, 1–7 (2009)
6. Liu, Z., Siguaw, J.A., Enz, C.A.: Using Tourist Travel Habits and Preferences to assess Strategic Destination Positioning: The case of Costa Rica. Cornell Hosp. Q. **49**, 3 (2008)
7. Pinto, A.A.: Game Theory and Duopoly Models. Interdisciplinary Applied Mathematics Series. Springer, New York (2011)
8. Pinto, A.A., Mousa, A.S., Mousa, M.S., Samarah, R.M.: Tilings and Bussola for Making Decisions. Dynamics, Games and Science I. In: Peixoto, M., Pinto, A.A., Rand, D. (eds.) Proceedings in Mathematics series, Springer-Verlag, Chapter 44, 689–710 (2011)

Chapter 41
On the Hannay–Ozorio De Almeida Sum Formula

M. Pollicott and R. Sharp

Abstract In this note we consider the well known Hannay–Ozorio de Almeida sum formula from a mathematically rigorous viewpoint. In particular, we discuss situations where we can obtain the Sinai–Ruelle–Bowen measure as a limit taken over periodic orbits with periods in an interval which shrinks as it moves to infinity.

41.1 Introduction

The Hannay–Ozorio de Almeida sum formula is a useful principle in the study of the distribution of closed orbits for Hamiltonian flows [7]. Roughly speaking, it asserts that an appropriately weighted sum of measures supported on periodic orbits converges to the physical measure as the periods become large. This formula was originally introduced and used in the study of Quantum Chaos. In particular, Berry used the so-called diagonal approximation and the Hannay–Ozorio de Almeida sum rule to determine the asymptotics of the spectral form factor, which is the Fourier transform of the two-point correlation function for the eigenvalues of the Laplacian [1,6,8]. The traditional setting is in the context of Hamiltonian flows, which include the canonical example of geodesic flows on negatively curved manifolds.

Let us now give a brief description of the formula in the context of a C^2 attracting hyperbolic flows $\phi_t : \Lambda \to \Lambda$, where the attractor Λ is contained in a Riemannian manifold M. Let τ denote a (prime) periodic orbit and let $\lambda(\tau)$ denote its least period. Let $f : \Lambda \to \mathbf{R}$ be a continuous function, then we can introduce a weighted period $\lambda_f(\tau) = \int_0^{\lambda(\tau)} f(\phi_t x_\tau)dt$, where $x_\tau \in \tau$. In particular, if we define the

M. Pollicott (✉)
Department of Mathematics, Warwick University, Coventry CV4 7AL, UK
e-mail: masdbl@warwick.ac.uk

R. Sharp
Manchester University, Oxford Road, Manchester M13 9PL, UK
e-mail: sharp@ma.man.ac.uk

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_41,

expansion coefficient $E : \Lambda \to \mathbf{R}$ by

$$E(x) := \lim_{t\to 0} \frac{1}{t} \log |\mathrm{Jac}(D\phi_t | E^u(x))|$$

then we shall write $\lambda^u(\tau) = \lambda_E(\tau)$. In this setting, one version of the Hannay–Ozorio de Almeida sum formula takes the following form:

$$\lim_{T\to+\infty} \frac{1}{\delta} \sum_{T-\frac{\delta}{2}\le\lambda(\tau)\le T+\frac{\delta}{2}} \lambda_f(\tau)e^{-\lambda^u(\tau)} = \int f\, d\mu, \qquad (41.1)$$

where μ is the SRB (Sinai–Ruelle–Bowen) measure, i.e., the unique ϕ_t-invariant probability measure which is absolutely continuous with respect to the volume on M. William Parry was one of the first people to make a mathematically rigorous study of such results. In particular, he gave a completely rigorous proof of (41.1) in the very general setting of weak mixing Axiom A flows and a general class of Hölder weights [11, 12].

In this note we want to address the question of whether $\delta = \delta(T)$ can be allowed to shrink to zero as T increases and, if so, at what rate. This seems a natural question from both a mathematical and physical perspective, given that there is no natural choice of scale for δ.

Our main results are the following theorems which strengthen (41.1), in the appropriate settings. The first theorem is in the special case of geodesic flows.

Theorem 41.1. *Let $\phi_t : M \to M$ be the geodesic flow on the unit-tangent bundle over a compact negatively curved surface. There exists $\epsilon > 0$ such that if $\delta(T)^{-1} = O(e^{\epsilon T})$ then, for Hölder continuous functions $f : M \to \mathbf{R}$,*

There exists $\epsilon > 0$ such that if $\delta(T)^{-1} = 0(e^{\epsilon T})$

$$\lim_{T\to+\infty} \frac{1}{\delta(T)} \sum_{T-\frac{\delta(T)}{2}\le\lambda(\tau)\le T+\frac{\delta(T)}{2}} \lambda_f(\tau)e^{-\lambda^u(\tau)} = \int f\, d\mu \qquad (41.2)$$

The proof of Theorem 41.1 is based on estimates of Dolgopyat originally used in the proof of exponential mixing of geodesic flows [4]. In fact, the conclusion actually holds for any contact Anosov flows for which the stable and unstable foliations which are non-jointly integrable. In particular, it holds for the geodesic flow on the unit tangent bundle of a compact manifold with negative sectional curvatures, provided these curvatures are pinched between -1 and $-\frac{1}{4}$.

Definition 41.1. We say that β is Diophantine if there exist $\alpha > 2$ and $C > 0$ for which there are no rationals p/q satisfying $|\beta - p/q| \le C/q^\alpha$.

Our second theorem is the following.

Theorem 41.2. *Let $\phi_t : \Lambda \to \Lambda$ be a weak mixing C^2 hyperbolic attracting flow. Assume that we can chose two distinct closed orbits τ_1 and τ_2 such that*

$\beta = \lambda(\tau_1)/\lambda(\tau_2)$ is Diophantine. There exists $\gamma > 0$ such that if $\delta(T)^{-1} = O(T^\gamma)$ then, for Hölder continuous functions $f : \Lambda \to \mathbf{R}$,

There exists $\gamma > 0$ such that if $\delta(T)^{-1} = 0(T^\gamma)$

$$\lim_{T\to+\infty} \frac{1}{\delta(T)} \sum_{T-\frac{\delta(T)}{2}\le\lambda(\tau)\le T+\frac{\delta(T)}{2}} \lambda_f(\tau)e^{-\lambda^u(\tau)} = \int f\,d\mu \qquad (41.3)$$

The proof of Theorem 41.2 is based on estimates of Dolgopyat used to establish polynomial rates of mixing in a wider setting [5]. In particular, the conclusion holds for any weak mixing C^2 Anosov flow.

Remark 41.1. Complementary results to Theorems 41.1 and 41.2 are obtained by fixing $\delta > 0$ and asking about the rate of convergence in (41.1). However, this follows easily using the ideas in [14, 15]. The results are the following.

1. Let ϕ_t be the geodesic flow on the unit-tangent bundle of a compact negatively curved surface and let $f : M \to \mathbf{R}$ be a Hölder continuous function. Then there exists $\epsilon > 0$ such that we have that

$$\frac{1}{\delta} \sum_{T-\frac{\delta}{2}\le\lambda(\tau)\le T+\frac{\delta}{2}} \lambda_f(\tau)e^{-\lambda^u(\tau)} = \int f\,d\mu + O(e^{-\epsilon T}), \text{ as } T \to +\infty.$$

2. Let ϕ_t be a weak mixing hyperbolic attracting flow and let $f : M \to \mathbf{R}$ be a Hölder continuous function. Assume that we can find two distinct closed orbits τ_1 and τ_2 such that $\beta = \lambda(\tau_1)/\lambda(\tau_2)$ is Diophantine. Then there exists $\eta > 0$ such that we have that

$$\frac{1}{\delta} \sum_{T-\frac{\delta(T)}{2}\le\lambda(\tau)\le T+\frac{\delta(T)}{2}} \lambda_f(\tau)e^{-\lambda^u(\tau)} = \int f\,d\mu + O(T^{-\eta}), \text{ as } T \to +\infty$$

Throughout the paper, we use the standard Landau big O and little o notation, i.e, we write $A(T) = O(B(T))$ if there exists $D > 0$ such that $|A(T)| \le DB(T)$ and $A(T) = o(B(T))$ if $|A(T)|/B(T) \to 0$, as $T \to +\infty$.

41.2 Hyperbolic Flows and Symbolic Dynamics

Let $\phi_t : M \to M$ be a C^∞ flow on a compact manifold. Let Λ be a closed ϕ-invariant subset. We call the set Λ *hyperbolic* if:

1. There exists a $D\phi$-invariant splitting $T_\Lambda M = E^0 \oplus E^s \oplus E^u$ and constants $C > 0$ and $\lambda > 0$ such that

   (a) E^0 is tangent to the direction of the flow
   (b) $\|D\phi_t|E^u\| \le Ce^{-\lambda t}$, for $t \ge 0$
   (c) $\|D\phi_{-t}|E^s\| \le Ce^{-\lambda t}$, for $t \ge 0$

2. The periodic orbits in Λ are dense
3. The flow restricted to Λ has a dense orbit
4. There exists an open set $U \supset \Lambda$ such that $\Lambda = \cap_{t \in \mathbf{R}} \phi_t U$

We call the restriction of the flow $\phi_t : \Lambda \to \Lambda$ a *hyperbolic* flow. If $\Lambda = \cap_{t>0} \phi_t U$ then we say that ϕ_t is an attracting hyperbolic flow or, more succinctly, a *hyperbolic attractor*. For any $x \in \Lambda$ we denote the associated unstable manifold by

$$W^u(x) = \{y \in M : \lim_{t \to \infty} d(\phi_t x, \phi_t y) = 0\},$$

and if Λ is an attractor then $W^u(x) \subset \Lambda$.

If a hyperbolic attractor is C^2 then it supports a unique probability measure which is both invariant and absolutely continuous with respect to the natural volume induced on each unstable manifold by the ambient Riemannian volume $\overline{m}$. This measure, which we denote by μ, is called the Sinai–Ruelle–Bowen measure and describes the behaviour of $\overline{m}$-almost every point in a neighbourhood of the attractor [3].

Example 41.1 (Geodesic flow). Let M be the unit tangent bundle of a compact C^∞ surface V, i.e., the tangent vectors to V of unit length. The geodesic flow $\phi_t : M \to M$ is defined as follows. Given a unit tangent vector v we consider the unit speed geodesic $\gamma_v : M \to M$ such that $\dot{\gamma}_v(0) = v$. We then define $\phi_t(v) = \dot{\gamma}_v(t)$. If V has negative curvature then the associated geodesic flow is a hyperbolic attractor with $\Lambda = M$. Here μ is the Liouville measure.

Example 41.2 (Suspension flow). Let $T : \Omega \to \Omega$ be a solenoid. Let $r : \Omega \to \mathbf{R}^+$ be a strictly positive Hölder continuous function. We define the flow space by

$$\Lambda = \{(x, u) \in \Omega \times \mathbf{R} : 0 \leq u \leq r(x)\}$$

where we identify $(x, r(x))$ and $(T(x), 0)$. We define a flow by $\phi_t(x, u) = (x, u+t)$, subject to the identifications.

We shall prove our results via the symbolic description of a hyperbolic flow as a suspended flow over a subshift of finite type. We begin by recalling a few basic definitions and results. Let A be a $k \times k$ aperiodic matrix. We shall then let X be the space

$$X = \{x = (x_n)_{n=-\infty}^{\infty} \in \{1, \ldots, k\}^{\mathbf{Z}} : A(x_n, x_{n+1}) = 1 \text{ for all } n \in \mathbf{Z}\}$$

and define a metric on X by

$$d(x, y) = \sum_{n=-\infty}^{\infty} \frac{1 - \delta(x_n, y_n)}{2^{|n|}},$$

where $\delta(i, j) = 0$ if $i \neq j$ and $\delta(i, i) = 1$. The subshift of finite type $\sigma : X \to X$, defined by $(\sigma x)_n = x_{n+1}$, $n \in \mathbf{Z}$, is a homeomorphism. Given a strictly positive

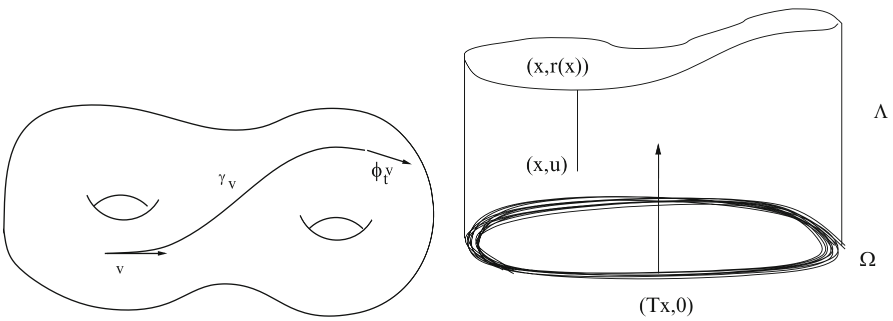


Fig. 41.1 (**a**) The geodesic flow moves the unit tangent vector v along the geodesic γ_v to $\phi_t v$; (**b**) The suspension flow is defined on the area under the graph of $r : \Omega \to \mathbf{R}$

Hölder continuous function $r : X \to \mathbf{R}^+$, let us denote

$$X^r = \{(x,u) \in X \times \mathbf{R} : 0 \le u \le r(x)\},$$

where $(x, r(x))$ and $(\sigma x, 0)$ are identified. We can define the suspended flow $\sigma_t^r : X^r \to X^r$ by $\sigma_t^r(x,u) = (x, u+t)$, subject to the identifications (Fig. 41.1).

To proceed, we state the following, now classical, result.

Lemma 41.1. *Given a hyperbolic flow $\phi_t : \Lambda \to \Lambda$, there exists a subshift of finite type $\sigma : X \to X$, a strictly positive Hölder continuous function $r : X \to \mathbf{R}^+$ and a Hölder continuous semi-conjugacy $\pi : X \to \Lambda$ such that:*

1. *π is one-to-one on a residual set.*
2. *A closed σ-orbit $\{x, \sigma x, \dots, \sigma^{n-1}x\}$ projects to a closed orbit τ of period $\lambda(\tau) = r^n(x) := r(x) + r(\sigma x) + \dots + r(\sigma^{n-1}x)$. Moreover, if we define $g : X \to \mathbf{R}$ by*
$$g(x) = -\int_0^{r(x)} E(\pi(x,u))du$$
then g is Hölder continuous and $-\lambda^u(\tau) = g^n(x)$.

Proof. This follows from the work of Bowen [2] and Bowen–Ruelle [3]. □

41.3 Dirichlet Series

In order to understand the limits in Theorems 41.1 and 41.2, we shall need to study the analytic properties of certain complex functions. We start with the following definition.

Definition 41.2. Given a non-negative Hölder continuous function $f : \Lambda \to \mathbf{R}$, we formally define an η-function for ϕ_t to be the Dirichlet series

$$\eta(s) = \sum_{\tau} \sum_{m=1}^{\infty} \lambda_f(\tau) e^{m(-\lambda^u(\tau)-(s-1)\lambda(\tau))}, \ s \in \mathbf{C},$$

where the sum is taken over all prime periodic orbits of ϕ_t.

It is not difficult to show that $\eta(s)$ converges to an analytic function on the half-plane $\mathrm{Re}(s) > 1$, and thus the definition makes sense on this domain. (We refer the reader to [13] for the general theory.) The proofs of Theorems 41.1 and 41.2 require showing that $\eta(s)$ has an analytic extension to a larger domain. To achieve this we need to relate $\eta(s)$ to a complex function defined in terms of X and functions thereon.

Given $f : \Lambda \to \mathbf{R}$ we define $f_0 : X \to \mathbf{R}$ by $f_0(x) = \int_0^{r(x)} f(\pi(x,u))du$.

Definition 41.3. We define a symbolic η-function by

$$\eta_0(s) = \sum_{n=1}^{\infty} \frac{1}{n} \sum_{\sigma^n x = x} f_0^n(x) e^{g^n(x)-(s-1)r^n(x)}.$$

For a continuous function $w : X \to \mathbf{R}$, we define its pressure $P(w)$ by

$$P(w) = \sup \left\{ h_\nu(\sigma) + \int w \, d\nu : \nu \text{ is a } \sigma\text{-invariant probability measure} \right\},$$

where $h_\nu(\sigma)$ denotes the entropy of σ with respect to ν. If w is Hölder continuous then there is a unique measure, called the equilibrium state for w, for which the supremum is attained.

It is a standard result that $\eta_0(s)$ converges to an analytic function for $P(g - \mathrm{Re}(s-1)r) < 0$ [13]. Since $P(g) = 0$, this holds for $\mathrm{Re}(s) > 1$.

The following lemma relates $\eta(s)$ and $\eta_0(s)$.

Lemma 41.2. *There exists* $\epsilon > 0$ *such that* $\eta_0(s) - \eta(s)$ *is analytic for* $\mathrm{Re}(s) > 1 - \epsilon$.

Proof. The functions $\eta(s)$ and $\eta_0(s)$ agree up to a small discrepancy (due to over-counting caused by orbits passing through the boundaries of the cross sections used to construct the symbolic dynamics in Lemma 41.1). This can be easily accounted for using the a construction of Bowen [2] (following [10]): the difference of the two functions can be written in terms of functions associated to a finite number of auxiliary subshifts of finite type. There are Hölder continuous maps from each of these to Λ but, crucially, they are not surjective. This forces a strict inequality of pressure functions which implies that the difference $\eta_0(s) - \eta(s)$ is analytic in a strictly larger half-plane than $\mathrm{Re}(s) > 1$. □

One of the interesting features of the present problem is the need to extend the region for which certain functions of two variable are bi-analytic. To address this problem, it is convenient to use some classical results in the theory of several complex variables [9]. We recall that a complex function of two variables is bi-analytic at a point $(z, s) \in \mathbf{C}^2$ if it has a uniformly convergent power series expansion (in two variables) in a neighbourhood of the point. Let

$$D(r_1, r_2) = \{(z, w) \in \mathbf{C}^2 : |z| < r_1, |w| < r_2\}$$

denote a polydisc in $\mathbf{C}^2$, where $r_1, r_2 > 0$.

Lemma 41.3 (Hartog's Theorem). *Let $F : D(r_1, r_2) \to \mathbf{C}$ be a function such that*

(i) $F(z, w)$ is bi-analytic on the smaller polydisc $D(r, r_2)$ $(0 < r < r_1)$.
(ii) For each $|w| < r_2$ the functions $f(\cdot, w) : \{z \in \mathbf{C} : |z| < r_1\} \to \mathbf{C}$ are analytic.

Then $F : D(r_1, r_2) \to \mathbf{C}$ is bi-analytic.

To prove Theorem 41.1 we shall require the following result on $\eta_0(s)$.

Lemma 41.4. *Let ϕ_t be a geodesic flow on a surface of negative curvature. We can write*

$$\eta_0(s) = \frac{\int f \, d\mu}{s - 1} + A(s),$$

where $A(s)$ is analytic for $Re(s) > 1 - \epsilon$, for some $\epsilon > 0$. Furthermore,

$$|\eta(s)| = O(\max\{|\mathrm{Im}(s)|^{\rho}, 1\})$$

for some $0 < \rho < 1$.

Proof. Let us define

$$L(s, z) = \exp\left(\sum_{n=1}^{\infty} \frac{1}{n} \sum_{\sigma^n x = x} e^{g^n(x) - (s-1)r^n(x) + z f_0^n(x)}\right)$$

with $s, z \in \mathbf{C}$ where $g : X \to \mathbf{R}$ and $f_0 : X \to \mathbf{R}$ are as defined above. It is easy to see that function $L(s, z)$ converges to a non-zero and bi-analytic function in (s, z) provided $\mathrm{Re}(s) > 0$ and $|z|$ sufficiently small [9]. Moreover, it follows from the approach in Dolgopyat's paper [4] (explicitly in the case $z = 0$, or by a simple modification for any fixed z) that we have analyticity of $L(s, z)$ in s for $\mathrm{Re}(s) > 1 - \epsilon$, where $\epsilon > 0$ can be chosen independently of z. The key ingredients in this approach are estimates on the transfer operator $L_{g - z f_0 - (s-1)r} : C^{\alpha}(X^+) \to C^{\alpha}(X^+)$ defined by

$$L_{g - z f_0 - (s-1)r} w(x) = \sum_{\sigma y = x} e^{(g - z f_0 - (s-1)r)(y)} w(y)$$

on a suitable family $C^{\alpha}(X^+)$ of Hölder continuous functions on the corresponding one-sided shift $\sigma : X^+ \to X^+$, where

$$X^+ = \left\{x = (x_n)_{n=0}^{\infty} \in \{1, \ldots, k\}^{\mathbf{Z}^+} : A(x_n, x_{n+1}) = 1 \text{ for all } n \in \mathbf{Z}^+\right\}.$$

(Here we assume that g, f_0 and r have been replaced by functions in $C^\alpha(X^+)$, chosen so that their sums around periodic orbits remain unchanged. We refer the reader to [13] for more details. To avoid complicating the exposition, we do not change the notation.)

The domain of analyticity corresponds to those (z, s) for which 1 is not in the spectrum of $L_{g-zf_0-(s-1)r}$. Moreover, one can also show that $L(s, z)$ is analytic in a neighbourhood of $s = 1$, provided $s \neq s(z)$, where $s(z)$ is an analytic function with $s(0) = 1$ satisfying $P(g - zf_0 - (s(z) - 1)r) = 0$, for $|z|$ sufficiently small, where $P(\cdot)$ is the analytic extension of the pressure function (i.e., the logarithm of the maximal eigenvalue of the associated transfer operator) [13]. We claim that $L(s, z)^{-1}$ can be differentiated in the second variable at $z = 0$. This is the point in the proof where it is convenient to use the Hartog's Theorem (Lemma 41.3). We have already observed that $L(s, z)^{-1}$ is bi-analytic in the pair of variables (s, z) for $\mathrm{Re}(s) > 0$ and $|z|$ then chosen sufficiently small. We can apply Hartog's Theorem to extend the domain of analyticity to $\mathrm{Re}(s) > 1 - \epsilon/2$, say, and $|z|$ sufficiently small (independent of s). It is now routine to show that $s = 0$ in a simple pole for $\eta_0(s)$ with the claimed residue. Briefly, for s in a sufficiently small neighbourhood of 0 we can write

$$\begin{aligned}\eta_0(s) &= \frac{\partial \log L(s,z)}{\partial z}|_{z=0} \\ &= -\frac{\partial \log(1 - e^{P(g-zf_0-(s-1)r)})}{\partial z}|_{z=0} + A_0(s) \\ &= \frac{1}{s-1}\int f d\mu + A_1(s),\end{aligned}$$

where $A_0(s)$, $A_1(s)$ are analytic functions in a neighbourhood of $s = 1$ and $\int f d\mu = \int f_0 dm / \int r dm$, where m is the equilibrium state on X^+ for g [13].

To complete the proof, we need bounds on $\eta_0(s)$. There exists $\rho_0 > 0$ such that in the same region we have a bound $L(s, z) = O(|\mathrm{Im}(s)|^{\rho_0})$ for $|\mathrm{Im}(s)| \geq 1$. Moreover, the implied constants are uniform in z (in a small neighbourhood of 0). This is implicit in the details of the proof of the result cf. [4, 14]. Thus, we can use Cauchy's Theorem to obtain

$$\eta_0(s) = \frac{\partial}{\partial z} L(s,z)|_{z=0} = \frac{1}{2\pi i \delta}\int_{|\xi|=\delta} \frac{L(s,\xi)}{\xi^2} d\xi = O(|\mathrm{Im}(s)|^\rho),$$

where $\delta > 0$ is chosen sufficiently small.

Finally, as in [14], we may use an argument based on the Phragmén–Lindelöf Theorem, to show that, decreasing ϵ if necessary, ρ may be chosen to be less than 1. □

To prove Theorem 41.2 we shall require the following, somewhat weaker, result on $\eta(s)$.

Lemma 41.5. *Let ϕ be a hyperbolic flow satisfying the hypotheses of Theorem 41.2. We can write*

$$\eta(s) = \frac{\int f d\mu}{s-1} + A(s),$$

where $A(s)$ is analytic for $\mathrm{Re}(s) > 1 - \epsilon \min\{|\mathrm{Im}(s)|^{-\alpha}, 1\}$, for some $\epsilon, \alpha > 0$. Furthermore,

$$|A(s)| = O(\max\{|\mathrm{Im}(s)|^{\rho}, 1\}),$$

for some $\rho > 0$.

Proof. The proof is similar to that of Lemma 41.4. Again the function $L(s,z)$ is bi-analytic in (s,z) for $\mathrm{Re}(s) > 1$ and $|z|$ sufficiently small (cf. [13]). This time we apply the approach in Dolgopyat's paper [5] and for fixed z (with $|z|$ sufficiently small) we have analyticity in s for $\mathrm{Re}(s) > 1 - \epsilon \min\{|\mathrm{Im}(s)|^{-\alpha}, 1\}$, for some uniform (in z) choice of $\epsilon > 0$. The uniformity of the implied constants for small $|z|$ is implicit in the proofs. We can again apply Hartog's theorem for functions of two variables to deduce that $L(s,z)$ is bi-analytic in (s,z) for $\mathrm{Re}(s) > 1 - \frac{\epsilon}{2} \min\{|\mathrm{Im}(s)|^{-\alpha}, 1\}$, say, and $|z|$ sufficiently small. The pole free region for $\eta(s)$, the bounds on modulus $|\eta(s)|$ and the form of the pole and residue at $s = 1$ follow by arguments analogous to those in the previous case. □

41.4 Proof of Theorem 41.1

Given Lemma 41.4, the proof of Theorem 41.1 now follows fairly traditional lines. We recall the following standard identity [16].

Lemma 41.6. *Let $c > 0$ and $k \geq 1$. Then*

$$\frac{1}{2\pi i} \int_{c+i\infty}^{c+i\infty} \frac{T^{s+k}}{s(s+1)\cdots(s+k)} ds = \frac{1}{k!} \left(1 - \frac{1}{T}\right)^k \delta_{[1,+\infty)}(T)$$

For $T > 0$, we shall write

$$\psi_0(T) = \sum_{e^{m\lambda(\tau)} \leq T} \lambda_f(\tau) e^{m(\lambda(\tau) - \lambda^u(\tau))},$$

where the summation is taken over all prime periodic orbits τ and all $m \geq 1$ satisfying $e^{m\lambda(\tau)} \leq T$.

Lemma 41.7. *Under the hypotheses of Theorem 41.1, there exists $\epsilon' > 0$ such that*

$$\psi_0(T) = \left(\int f \, d\mu\right) T + O(T^{1-\epsilon'}).$$

Proof. We introduce an auxiliary function $\psi_1(T) = \int_1^T \psi_0(u)du$. Using Lemma 3.1, with $k = 1$, we can write

$$\psi_1(T) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} \eta(s)\frac{T^{s+1}}{s(s+1)}ds,$$

for any $c > 1$. We want to move the curve of integration to $d = 1 - \epsilon'$, say, where $0 < \epsilon' < \epsilon$, with ϵ as in Lemma 41.4. Since $\eta(s)$ has a simple pole at $s = 1$, we may use the Residue Theorem and the bound $|\eta(s)| = O(|\mathrm{Im}(s)|^\rho)$, for $\rho < 1$, to obtain

$$\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} \eta(s)\frac{T^{s+1}}{s(s+1)}ds = \left(\int f d\mu\right)\frac{T^2}{2} + \frac{1}{2\pi i}\int_{d-i\infty}^{d+i\infty} \eta(s)\frac{T^{s+1}}{s(s+1)}ds. \tag{41.4}$$

Again using the bound on $|\eta(s)|$, the second term on the Right Hand Side of (41.4) can be estimated by

$$\left|\frac{1}{2\pi i}\int_{d-i\infty}^{d+i\infty} \eta(s)\frac{T^{s+1}}{s(s+1)}ds\right| = O\left(T^{d+1}\int_1^\infty \frac{t^\rho}{t(t+1)}dt\right) = O(T^{2-\epsilon'}),$$

To finish the proof, we need to replace the estimate on $\psi_1(T)$ with one on $\psi_0(T)$. Since $\psi_0(T)$ and $\psi_1(T)$ are both monotone increasing, we may write

$$\begin{aligned}\psi_0(T) \le \frac{\psi_1(T) - \psi_1(T-\Delta)}{\Delta} &= \left(\int f\, d\mu\right)\left(\frac{T^2 - (T-\Delta)^2}{2\Delta}\right) + O\left(\frac{T^{2-\epsilon'}}{\Delta}\right)\\ &= \left(\int f\, d\mu\right)T + O\left(\frac{T^{2-\epsilon'}}{\Delta}, \Delta\right).\end{aligned}$$

If we choose $\Delta = T^{1-\epsilon'/2}$ then we have that

$$\psi_0(T) \le \left(\int f\, d\mu\right)T + O(T^{1-\epsilon'/2})$$

Similarly, we can show that

$$\psi_0(T) \ge \left(\int f\, d\mu\right)T + O(T^{1-\epsilon'/2})$$

and the result follows. □

Lemma 41.8. *For some* $\epsilon' > 0$, *we have*

$$\sum_{\lambda(\tau)\le T} \lambda_f(\tau)e^{\lambda(\tau)-\lambda^u(\tau)} = \left(\int f\, d\mu\right)e^T + O(e^{(1-\epsilon')T}).$$

Proof. By Lemma 3.2, we have

$$\sum_{m\lambda(\tau)\le T} \lambda_f(\tau)e^{m(\lambda(\tau)-\lambda^u(\tau))} = \left(\int f\,d\mu\right)e^T + O(e^{(1-\epsilon')T}),$$

where m runs over all $m \ge 1$. We need to show that the terms with $m \ge 2$ make a contribution of smaller order. By simple estimates

$$\limsup_{T\to+\infty} \frac{1}{T}\log\left(\sum_{m\ge 2\,:\,m\lambda(\tau)\le T} \lambda_f(\tau)e^{m(\lambda(\tau)-\lambda^u(\tau))}\right) \le 1 + \sup_{m\ge 2}\frac{P(-mE)}{m},$$

where E is the function defined in the introduction and P denotes pressure. It follows from standard properties of pressure that

(a) $P(-mE) < 0$, for all $m \ge 2$

(b)
$$\lim_{m\to+\infty}\frac{P(-mE)}{m} = e_- := \inf_\nu \int -E\,d\nu < 0,$$
where the infimum is taken over all ϕ_t-invariant probability measures.

In particular, there exists $N \ge 1$ such that

$$\frac{P(-mE)}{m} \le \frac{e_-}{2}$$

for $m > N$ and so

$$1 + \sup_{m\ge 2}\frac{P(-mE)}{m} \le 1 + \max\left\{\frac{P(-2E)}{2},\ldots,\frac{P(-NE)}{N},\frac{e_-}{2}\right\} < 1.$$

Decreasing ϵ' if necessary, this gives the required result. □

Proof (Proof of Theorem 41.1). Lemma 41.8 shows that

$$\pi_f(T) := \sum_{\lambda(\tau)\le T}\lambda_f(\tau)e^{\lambda(\tau)-\lambda^u(\tau)} = \left(\int f\,d\mu\right)e^T + O\big(e^{(1-\epsilon')T}\big), \text{ as } T \to +\infty.$$

Thus, for $\delta = \delta(T)$, we can write that

$$\begin{aligned}\sum_{T-\delta/2\le\lambda(\tau)\le T+\delta/2}\lambda_f(\tau)e^{\lambda(\tau)-\lambda^u(\tau)} &= \pi_f\left(T+\frac{\delta}{2}\right) - \pi_f\left(T-\frac{\delta}{2}\right)\\ &= \left(\int f\,d\mu\right)\left(e^{(T+\delta/2)} - e^{(T-\delta/2)}\right)\\ &\quad +O\left(e^{(1-\epsilon')T}\right)\end{aligned}$$

$$= \left(\int f \, d\mu\right) e^T \delta + O\left(e^{(1-\epsilon')T}, \delta^2 e^T\right).$$

We then have the asymptotic upper bound

$$\begin{aligned}
\sum_{T-\delta/2 \le \lambda(\tau) \le T+\delta/2} \lambda_f(\tau) e^{-\lambda^u(\tau)} &\le \exp\left(-T + \frac{\delta}{2}\right) \\
&\qquad \sum_{T-\delta/2 \le \lambda(\tau) \le T+\delta/2} \lambda_f(\tau) e^{\lambda(\tau)-\lambda^u(\tau)} \\
&= \left(\int f \, d\mu\right) \delta \exp\left(\frac{\delta}{2}\right) + O\left(e^{-\epsilon' T}, \delta^2\right) \\
&= \left(\int f \, d\mu\right) \left(\delta + \frac{\delta^2}{2}\right) + O\left(e^{-\epsilon' T}, \delta^2\right) \\
&= \left(\int f \, d\mu\right) \delta + O\left(e^{-\epsilon' T}, \delta^2\right).
\end{aligned}$$

Similarly, we have an asymptotic lower bound

$$\begin{aligned}
\sum_{T-\delta/2 \le \lambda(\tau) \le T+\delta/2} \lambda_f(\tau) e^{-\lambda^u(\tau)} &\ge \exp\left(-T - \frac{\delta}{2}\right) \\
&\qquad \sum_{T-\delta/2 \le \lambda(\tau) \le T+\delta/2} \lambda_f(\tau) e^{\lambda(\tau)-\lambda^u(\tau)} \\
&= \left(\int f \, d\mu\right) \delta \exp\left(-\frac{\delta}{2}\right) + O\left(e^{-\epsilon' T}, \delta^2\right) \\
&= \left(\int f \, d\mu\right) \left(\delta - \frac{\delta^2}{2}\right) + O\left(e^{-\epsilon' T}, \delta^2\right) \\
&= \left(\int f \, d\mu\right) \delta + O\left(e^{-\epsilon' T}, \delta^2\right).
\end{aligned}$$

Comparing these estimates, we see that

$$\frac{1}{\delta} \sum_{T-\delta/2 \le \lambda(\tau) \le T+\delta/2} \lambda_f(\tau) e^{-\lambda^u(\tau)} = \int f \, d\mu + O\left(\frac{e^{-\epsilon' T}}{\delta}, \delta\right).$$

In particular, providing $\delta(T) \to 0$ as $T \to +\infty$ with $\delta(T)^{-1} = o(e^{\epsilon' T})$ then the estimate (0.2) holds, provided f is non-negative. The result for general f follows from considering positive and negative parts. □

41.5 Proof of Theorem 41.2

We again write $\psi_0(T) = \sum_{e^{m\lambda(\tau)} \le T} \lambda_f(\tau) e^{m(\lambda(\tau)-\lambda^u(\tau))}$ and $\psi_1(T) = \int_1^T \psi_0(u) du$ (Fig. 41.2).

Lemma 41.9. *There exists $a > 0$ such that*

$$\psi_0(T) = \left(\int f \, d\mu\right) T + O\left(\frac{T}{(\log T)^a}\right).$$

Proof. First, let us suppose the exponent $\rho > 0$ in Lemma 41.5 satisfies $0 < \rho < 1$. For $c > 1$ we can again write

$$\psi_1(T) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} \eta(s) \frac{T^{s+1}}{s(s+1)} ds. \tag{41.5}$$

As before we want to move the line of integration to left, however, this time to a curve $\Gamma = \Gamma(T)$ depending on T . More precisely, Γ is the union of the arcs:

1. $\Gamma_0 = [1 + iR, 1 + i\infty]$
2. $\Gamma_1 = [d + iR, 1 + iR]$
3. $\Gamma_2 = [d - iR, d + iR]$
4. $\Gamma_3 = [1 - iR, d - iR]$
5. $\Gamma_4 = [1 - i\infty, 1 - iR]$

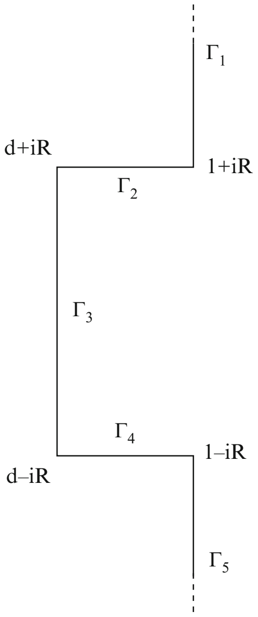


Fig. 41.2 The curve of integration

where $R = R(T) = (\log T)^{\epsilon}$, with $0 < \epsilon < \min\{\frac{\alpha}{2}, \frac{1}{\rho}\}$ and $d = d(T) = 1 - (\log T)^{-1/2}$.

By the Residue Theorem, we can write

$$\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} \eta(s)\frac{T^{s+1}}{s(s+1)}ds = \left(\int f\, d\mu\right)\frac{T^2}{2} + \frac{1}{2\pi i}\int_{\Gamma} \eta(s)\frac{T^{s+1}}{s(s+1)}ds. \tag{41.6}$$

Moreover, we can bound

$$\left|\frac{1}{2\pi i}\int_{\Gamma_1\cup\Gamma_3} \frac{T^{s+1}}{s(s+1)}ds\right| = O(R^{\rho-2}T^2) = O\left(\frac{T^2}{(\log T)^{\epsilon(2-\rho)}}\right), \tag{41.7}$$

$$\left|\frac{1}{2\pi i}\int_{\Gamma_0\cup\Gamma_4} \frac{T^{s+1}}{s(s+1)}ds\right| = O\left(\frac{T^2}{R^{1-\rho}}\right) = O\left(\frac{T^2}{(\log T)^{\epsilon(1-\rho)}}\right), \tag{41.8}$$

$$\left|\frac{1}{2\pi i}\int_{\Gamma_2} \frac{T^{s+1}}{s(s+1)}ds\right| = O\left(\frac{T^{2-(\log T)^{-1/2}}}{R^{1-\rho}}\right) = O\left(\frac{T^2 e^{-(\log T)^{\frac{3}{2}}}}{(\log T)^{\epsilon(1-\rho)}}\right). \tag{41.9}$$

We can then estimate

$$\psi_1(T) = \left(\int f\, d\mu\right)\frac{T^2}{2} + O\left(\frac{T^2}{(\log T)^{-a}}\right),$$

for $a > 0$ chosen sufficiently small.

Using the same method as in the proof of Lemma 3.2 we can write

$$\begin{aligned}\psi_0(T-\Delta) &\le \frac{\psi_1(T) - \psi_1(T-\Delta)}{\Delta}\\ &= \int f d\mu \left(\frac{T^2 - (T-\Delta)^2}{2\Delta}\right) + O\left(\frac{T^2}{\Delta(\log T)^a}\right)\\ &= T\int f d\mu + O\left(\frac{T^2}{\Delta(\log T)^a}, \Delta\right).\end{aligned}$$

If we choose $\Delta = T(\log T)^{-a/2}$ then we have that

$$\psi_0(T-\Delta) \le \left(\int f\, d\mu\right)T + O\left(\frac{T}{(\log T)^{a/2}}\right)$$

and thus

$$\psi_0(T) \le \left(\int f\, d\mu\right)T + O\left(\frac{T}{(\log T)^{a/2}}\right).$$

Modifying the proof of Lemma 3.2, we can also show that

$$\psi_0(T) \geq \left(\int f\, d\mu\right) T + O\left(\frac{T}{(\log T)^{a/2}}\right)$$

and the result follows.

More generally, if $k-1 \leq \rho < k$ then we can inductively define a sequence of functions

$$\psi_2(T) = \int_1^T \psi_1(u)\, du, \ldots, \psi_k(T) = \int_1^T \psi_{k-1}(u)\, du.$$

By repeatedly using the above arguments we reach the same conclusion. □

Proof (Proof of Theorem 41.2). Lemma 4.1 and the arguments in Lemma 41.8 show that

$$\pi_f(x) := \sum_{\lambda(\tau)\leq T} \lambda_f(\tau) e^{\lambda(\tau)-\lambda^u(\tau)} = \left(\int f\, d\mu\right) e^T + O\left(\frac{e^T}{T^a}\right), \text{ as } T \to +\infty,$$

for some choice of $a > 0$, and thus we can write that

$$\begin{aligned}\sum_{T-\delta/2\leq\lambda(\tau)\leq T+\delta/2} \lambda_f(\tau) e^{\lambda(\tau)-\lambda^u(\tau)} &= \pi_f\left(T+\frac{\delta}{2}\right) - \pi_f\left(T-\frac{\delta}{2}\right)\\ &= \left(\int f\, d\mu\right)\left(e^{(T+\delta/2)} - e^{(T-\delta/2)}\right)\\ &\quad +O\left(\frac{e^T}{T^a}\right)\\ &= \left(\int f\, d\mu\right) e^T\delta + O\left(\frac{e^T}{T^a}, \delta^2 e^T\right).\end{aligned}$$

We can then write that

$$\begin{aligned}\sum_{T-\delta/2\leq\lambda(\tau)\leq T+\delta/2} \lambda_f(\tau) e^{-\lambda^u(\tau)} &\leq \exp\left(-T+\frac{\delta}{2}\right)\\ &\quad \sum_{T-\delta/2\leq\lambda(\tau)\leq T+\delta/2} \lambda_f(\tau) e^{\lambda(\tau)-\lambda^u(\tau)}\\ &= \left(\int f\, d\mu\right)\delta \exp\left(\frac{\delta}{2}\right) + O\left(\frac{1}{T^a}, \delta^2\right)\\ &= \left(\int f\, d\mu\right)\left(\delta + \frac{\delta^2}{2}\right) + O\left(\frac{1}{T^a}, \delta^2\right)\\ &= \left(\int f\, d\mu\right)\delta + O\left(\frac{1}{T^a}, \delta^2\right),\end{aligned}$$

with a similar lower bound.

Comparing these estimates, we see that

$$\frac{1}{\delta}\sum_{T-\delta/2\leq\lambda(\tau)\leq T+\delta/2}\lambda_f(\tau)e^{-\lambda^u(\tau)} = \int f\,d\mu + +O\left(\frac{1}{\delta T^a},\delta\right).$$

In particular, providing $\delta(T) \to 0$ as $T \to +\infty$ with $\delta(T)^{-1} = o(T^{-a})$, then the estimate (0.3) holds, provided f is non-negative. The result for general f follows from considering positive and negative parts. □

References

1. Berry, M.: Semiclassical theory of spectral rigidity. Proc. R. Soc. London Ser. A **400**, 229–251 (1985)
2. Bowen, R.: Symbolic dynamics for hyperbolic flows. Am. J. Math. **95**, 429–460 (1973)
3. Bowen, R., Ruelle, D.: The ergodic theory of Axiom A flows. Invent. Math. **29**, 181–202 (1975)
4. Dolgopyat, D.: On decay of correlations in Anosov flows. Ann. Math. **147**, 357–390 (1998)
5. Dolgopyat, D.: Prevalence of rapid mixing in hyperbolic flows. Ergod. Theory Dyn. Syst. **18**, 1097–1114 (1998)
6. Elton, J., Lakshminarayan, A., Tomsovic, S.: Fluctuations in classical sum rules. Phys. Rev. E82, 046223(2010) preprint (2009)
7. Hannay, J.H., Ozorio de Almeida, A.M.: Periodic orbits and a correlation function for the semiclassical density of states. J. Phys. A **17**, 3429–3440 (1984)
8. Keppeler, S.: Classical sum rules. Springer Tracts in Modern Physics, vol. 193, pp. 111–125. Springer, Berlin (2003)
9. Krantz, S.: Function theory of several complex variables. American Mathematical Society, Providence, RI (1992)
10. Manning, A.: Axiom A diffeomorphisms have rational zeta functions. Bull. Lond. Math. Soc. **3**, 215–220 (1971)
11. Parry, W.: Synchronisation of canonical measures for hyperbolic attractors. Comm. Math. Phys. **106**, 267–275 (1986)
12. Parry, W.: Equilibrium states and weighted uniform distribution of closed orbits. Dynamical systems, Proc. Spec. Year, College Park/Maryland Lect. Notes Math., vol. 1342, pp. 617–625. Springer, Berlin (1988)
13. Parry, W., Pollicott, M.: Zeta functions and the periodic orbit structure of hyperbolic dynamics. Astrisque **187–188**, 1–268 (1990)
14. Pollicott, M., Sharp, R.: Exponential error terms for growth functions on negatively curved surfaces. Am. J. Math. **120**, 1019–1042 (1998)
15. Pollicott, M., Sharp, R.: Error terms for closed orbits of hyperbolic flows, Ergod. Theory Dyn. Syst. **21**, 545–562 (2001)
16. Rademacher, H.: Topics in Analytic Number Theory. Springer, Berlin (1973)

Chapter 42
A Fourier Transform Method for Relaxation of Kinetic Equations

Manuel Portilheiro

Abstract A Fourier transform method is used to analyze the relaxation limit of the linear Boltzman–Poisson system for electron density. Two scalings of interest are analyzed, the *low field scaling* and the *drif-colision balance scaling*, corresponding two different regimes of the equation. The limits are obtained in the sense of 'dissipative' solutions.

42.1 Introduction

The relaxation of general kinetic equations can be seen as natural way to understand macroscopic phenomena of large systems of interacting particles. In analogy with the hydrodynamic and diffusive limits for the Boltzmann equation, where one obtains the Euler system and the Navier–Stokes equations, respectively, similar limits can be carried out for other kinetic equations.

The notion of dissipative solution introduced in [2] is particularly suitable to carry out such relaxation limits. Some discrete velocity models were analized in [3], in particular the two-velocity Carleman model. In [4] mixed and continuous velocity L^1 models were studied, and an abstract framework for kinetic equations with L^1 collision terms was proposed.

In this work we look at the semiconductor equation, with the extra electric field term, coupled with its respective Poisson equation. Without analizing compactness, the limits corresponding to hydrodynamic and diffusive scalings yield the expect equations. Related work, for the Vlasov–Poisson–Fokker–Planck system, was done by Poupaud, Soler, Nieto, and Goudon (see [1] and references therein). We start by introducing the Boltzmann–Poisson system, the two scalings we are interested and their respective limits. We also mention the "dissipative solutions" we will work with.

M. Portilheiro
Universidad Autónoma de Madrid, Campus de Cantoblanco, 28036 Madrid, Spain
e-mail: manuel.portilheiro@uam.es

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_42,

42.1.1 Boltzmann–Poisson System

We consider the semi classical Boltzmann–Poisson system in the form

$$\partial_t f + v \cdot \nabla_x f - \frac{e}{m} E(x,t) \nabla_v f = \frac{1}{\tau}(M_{\Theta_0}\, \rho(f) - f), \tag{42.1}$$

where $f = f(t,x,v)$ denotes the density function of an electron at time $t > 0$ and position $x \in D$ and with velocity $v \in \mathbb{R}^3$ while e and m denote the unit charge and effective electron mass. The charge density is given by

$$\rho(f)(t,x) = \int_{\mathbb{R}^3} f(t,x,v)\, dv.$$

The electric field $E(x,t)$ is determined by the Poisson equation for the potential

$$\begin{aligned} &\varepsilon_0 \Delta \Phi(x,t) = e(\rho(f)(x,t) - C(x)) \\ &E(x,t) = -\nabla_x \Phi. \end{aligned} \tag{42.2}$$

Here $C(x)$ is the doping profile. The relaxation time τ is such that the mobility $\mu = (e/m)\tau E$ is linear for small values of $|E|$, with slope μ_0, and has a horizontal asymptote as $|E|$ increases. Finally the absolute Maxwellian is given by

$$M_{\Theta_0} = (2\pi\Theta_0)^{-3/2} \exp\left(-\frac{|v|^2}{2\Theta_0}\right),$$

where Θ_0 is the lattice temperature, $\Theta_0 = (k_B/m)T_0$, k_B the Boltzmann. We will write simply M for M_1.

We consider two different scalings of (42.1), the *low field scaling* (LFS) and the *drift-collision balance scaling* (DCBS). After writing the equation in dimensionless form these correspond respectively to

$$\begin{cases} \varepsilon \partial_t f + v \cdot \nabla_x f - E(t,x) \cdot \nabla_v f = \frac{1}{\varepsilon}\frac{1}{\tau}(M\rho(f) - f) \\ \Delta_x \Phi = \gamma(\rho(f) - C(x)), \end{cases} \tag{LFS}$$

and

$$\begin{cases} \eta \partial_t f + v \cdot \nabla_x f - \frac{\eta}{\varepsilon} E(t,x) \cdot \nabla_v f = \frac{1}{\varepsilon}\frac{1}{\tau}(M\rho(f) - f), \\ \Delta_x \Phi = \gamma(\eta^{-1}\rho(f) - C(x)), \end{cases} \tag{DCBS}$$

where τ, η and γ are dimensionless parameters.

The limits of these systems are

$$\begin{cases} \partial_t u - \mathrm{div}_x \left[\nabla_x u + uE\right] = 0, \\ \varDelta_x \varPhi = \gamma \left(u - C(x)\right), \end{cases} \tag{42.3}$$

for the (LFS) system, and

$$\begin{cases} u_t - \mathrm{div}_x \left(uE\right) = 0, \\ \varDelta_x \varPhi = \gamma \left(\frac{1}{\eta} u - C\right), \end{cases} \tag{42.4}$$

for the (DCBS) system.

Related systems have been studied in [4] using the same "perturbed test function" method. These systems have the extra transport term $E \cdot \nabla_v f$ coupled with the Poisson equation, which is the term that requires the use of the Fourier transform to find the steady state and carry the relaxation limit with the same technique as in [4]. We will not deal here with compactness issues and rather assume that $f^\varepsilon \to f$ pointwise and $u^\varepsilon = \int f^\varepsilon \, dv$ converges in L^1 (see [4] for the compactness arguments of simpler systems).

42.1.2 Dissipative Solutions

The notion of weak solution that we will make use of is the notion of "dissipative solution" introduced in [2] and [3]. We refer to this papers for a motivation.

For a nonlinear accretive (differential) operator A in $L^1(\mu)$ we say that u is a dissipative solution of $Au = f$ if

$$\int sgn(u - \phi)\,(f - A\phi)\; d\mu \geq 0$$

for every $\phi \in C_0^\infty + k, k \in \mathbb{R}$.

For the equations we consider we get the following. $(f^\varepsilon, \varPhi^\varepsilon)$ is a dissipative solution of (LFS) if

$$\begin{aligned} 0 \leq \iiint & sgn(f^\varepsilon - \phi^\varepsilon)\Big[-\partial_t \phi^\varepsilon - \frac{1}{\varepsilon}\left(v \cdot \nabla_x \phi^\varepsilon + \nabla_x \theta^\varepsilon \cdot \nabla_v \phi^\varepsilon\right) \\ & + \frac{1}{\tau\varepsilon^2}\left(M(v)\rho(\phi^\varepsilon) - \phi^\varepsilon\right)\Big]\, dt\, dx\, dv \\ & + \iint sgn(\varPhi^\varepsilon - \theta^\varepsilon)\left[-\varDelta_x \theta^\varepsilon + \gamma(\rho(\phi^\varepsilon) - C(x))\right]\, dt\, dx, \end{aligned} \tag{42.5}$$

for arbitrary ϕ^ε and θ^ε, and u solves the limit (42.3) if

$$\begin{aligned} 0 \le & \iint sgn(u-\psi)\left[-\partial_t\psi + \tau\,\mathrm{div}_x\left(\nabla_x\psi - \psi\nabla_x\theta\right)\right] dt\,dx \\ & + \iint sgn(\Phi-\theta)\left[-\Delta_x\theta + \gamma\left(\psi - C(x)\right)\right] dt\,dx. \end{aligned} \tag{42.6}$$

for arbitrary ψ. Similarly, $(f^\varepsilon, \Phi^\varepsilon)$ is dissipative solution of (DCBS) if

$$\begin{aligned} 0 \le & \iiint sgn(f^\varepsilon - \phi^\varepsilon)\Big[-\eta\partial_t\phi^\varepsilon - v\cdot\nabla_x\phi^\varepsilon - \frac{1}{\varepsilon}\big(\eta\nabla_x\theta\cdot\nabla_v\phi^\varepsilon \\ & - \frac{1}{\tau}\left(M\rho(\phi^\varepsilon) - \phi^\varepsilon\right)\big)\Big]\,dx\,dt\,dv \\ & + \iint sgn(\Phi^\varepsilon - \theta^\varepsilon)\left[-\Delta\theta + \gamma\left(\frac{1}{\eta}\rho(\phi^\varepsilon) - C\right)\right] dx\,dt, \end{aligned} \tag{42.7}$$

for every par $(\phi^\varepsilon, \theta^\varepsilon)$, and u solves (42.4) when

$$\begin{aligned} 0 \le & \iint sgn(u-\psi)\big[-\partial_t\psi - \tau\,\mathrm{div}_x(\psi\nabla_x\theta)\big]\,dx\,dt \\ & + \iint sgn(\Phi-\theta)\left[-\frac{1}{\eta}\Delta\theta + \frac{\gamma}{\eta}\left(\frac{1}{\eta}\psi - C\right)\right] dx\,dt, \end{aligned} \tag{42.8}$$

42.2 Low Field Scaling

We make the following choice for the test functions ϕ^ε in (42.5), $\phi^\varepsilon = \phi_0 + \varepsilon\phi_1$, where

$$\begin{aligned} \phi_0(x,t,v) &= M(v)\psi(x,t), \\ \phi_1(x,t,v) &= -\tau M(v)v\cdot[\nabla_x\psi(x,t) - \psi(x,t)\nabla_x\theta(x,t)], \end{aligned}$$

$\psi(x,t)$ is an arbitrary test function in (x,t) and we take $\theta^\varepsilon = \theta(x,t)$ independent of ε, to be chosen later. Notice that $\rho(\phi_1) = 0$, therefore $\rho(\phi^\varepsilon) = \psi$ and

$$\frac{1}{\tau\varepsilon^2}(M(v)\rho(\phi^\varepsilon) - \phi^\varepsilon) = \frac{1}{\varepsilon}M(v)v\cdot(\nabla_x\psi - \psi\nabla_x\theta).$$

Furthermore

$$\begin{aligned} \frac{1}{\varepsilon}(v\cdot\nabla_x\phi^\varepsilon + \nabla_x\theta\cdot\nabla_v\phi^\varepsilon) = & \frac{1}{\varepsilon}(M(v)v\cdot\nabla_x\psi + \nabla_x\theta\cdot\nabla_v M\psi) \\ & - \tau\left[v\cdot\nabla_x A + \nabla_x\theta\cdot\nabla_v A\right], \end{aligned}$$

where $A{:=}M(v)\,v\cdot(\nabla_x\psi - \psi\nabla_x\theta)$, hence

$$\begin{aligned}&\frac{1}{\varepsilon}(v\cdot\nabla_x\phi^\varepsilon+\nabla_x\theta\cdot\nabla_v\phi^\varepsilon)\\&\quad=\frac{1}{\tau\varepsilon^2}(M(v)\rho(\phi^\varepsilon)-\phi^\varepsilon)-\tau\left[v\cdot\nabla_x A+\nabla_x\theta\cdot\nabla_v A\right].\end{aligned}\tag{42.9}$$

Using this in (42.5), we find

$$\begin{aligned}0\le&\iiint sgn(f^\varepsilon-\phi^\varepsilon)\Big[-M(v)\partial_t\psi+\tau\,(v\cdot\nabla_x A+\nabla_x\theta\cdot\nabla_v A)\Big]\,dt\,dx\,dv\\&+\iint sgn(\Phi^\varepsilon-\theta)\left[-\Delta_x\theta+\gamma\,(\psi-C(x))\right]\,dt\,dx.\end{aligned}$$

Notice that

$$v\cdot\nabla_x A=M(v)\sum_{i,j=1}^{3}v_iv_j\left(\psi_{x_j}-\psi\theta_{x_j}\right)_{x_i}.$$

Therefore, letting $\varepsilon\to 0$, and since $f^\varepsilon\to u(x,t)M(v)$, $\Phi^\varepsilon\to\Phi$, we get

$$\begin{aligned}0\le&\iiint sgn(u-\psi)\Big[-M(v)\partial_t\psi+\tau M(v)\sum_{i,j=1}^{3}v_iv_j\left(\psi_{x_j}-\psi\theta_{x_j}\right)_{x_i}\\&+\tau\nabla_x\theta\cdot\nabla_v A\Big]\,dt\,dx\,dv\\&+\iint sgn(\Phi-\theta)\left[-\Delta_x\theta+\gamma\,(\psi-C(x))\right]\,dt\,dx\end{aligned}$$

or

$$\begin{aligned}0\le&\iint sgn(u-\psi)\left[-\partial_t\psi+\tau\,\mathrm{div}_x\,(\nabla_x\psi-\psi\nabla_x\theta)\right]\,dt\,dx\\&+\iint sgn(\Phi-\theta)\left[-\Delta_x\theta+\gamma\,(\psi-C(x))\right]\,dt\,dx,\end{aligned}$$

that is, (u,Φ) is a dissipative solution of (42.6).

42.3 Drift-Colision Balance Scaling

In this case we choose $\phi^\varepsilon=\phi$ so that

$$\eta\nabla_x\theta\cdot\nabla_v\phi-\frac{1}{\tau}\left(M\rho(\phi)-\phi\right)=0.\tag{42.10}$$

We denote by ψ the first moment of ϕ,

$$\psi:=\rho(\phi)=\int\phi\,dv,$$

and let

$$\boldsymbol{\alpha}{:=}\tau\eta\nabla_x\theta.$$

We want to take Fourier transforms above, in the variable v. Recall that

$$\widehat{M}(\xi) = \exp\left(-\frac{|\xi|^2}{2}\right),$$

where we are using the Fourier transform in the form

$$\widehat{f}(\xi) = \frac{1}{(2\pi)^{3/2}}\int_{\mathbb{R}^3} f(v)\exp\left(-iv\cdot\xi\right)\,dv.$$

We then have

$$\widehat{\phi}(\xi) + i\boldsymbol{\alpha}\cdot\xi\widehat{\phi}(\xi) - \psi\widehat{M}(\xi) = 0,$$

hence

$$\widehat{\phi}(\xi) = \psi\frac{1}{1+i\boldsymbol{\alpha}\cdot\xi}\widehat{M}(\xi) = \psi\widehat{M}(\xi) - \psi\frac{i\boldsymbol{\alpha}\cdot\xi-(i\boldsymbol{\alpha}\cdot\xi)^2}{1+|\boldsymbol{\alpha}\cdot\xi|^2}\widehat{M}(\xi)$$

and

$$\phi(v) = \psi M(v) - \psi(\boldsymbol{\alpha}\cdot\nabla_v)(1-\boldsymbol{\alpha}\cdot\nabla_v)(h_\alpha(v)*M(v)), \tag{42.11}$$

where h_α is such that

$$\widehat{h}_\alpha = \frac{1}{1+|\boldsymbol{\alpha}\cdot\xi|^2} \in L^\infty(\mathbb{R}^3_\xi),$$

(thus $h_\alpha \in \mathscr{M}$) and $\phi = \psi(M - h_\alpha * (\boldsymbol{\alpha}\cdot v - (\boldsymbol{\alpha}\cdot v)^2)M) \in \mathscr{S}_v$, the Schwartz space (the case $\boldsymbol{\alpha} = 0$ is also clear). Therefore, for any given $\psi,\theta \in C_c^\infty(\mathbb{R}^3\times[0,\infty))$, with ϕ given by (42.11), it is easy to check that indeed ϕ satisfies (42.10) and $\rho(\phi){=}\psi$.

Now we can let $\varepsilon\to 0$ and we obtain (42.4). Since the derivatives of ϕ in v are in the direction $\boldsymbol{\alpha}$, it is convenient to perform the change of variables $w = O_\alpha v$, where O_α is the pure rotation which takes $\boldsymbol{\alpha}/|\boldsymbol{\alpha}|$ to, say, $|\boldsymbol{\alpha}|\mathbf{e}_1 = (|\boldsymbol{\alpha}|,0,0)$,

$$O_\alpha\cdot\boldsymbol{\alpha} = |\boldsymbol{\alpha}|\mathbf{e}_1.$$

Then with

$$\lambda(w){:=}\frac{1}{\psi}\phi(x,t,v) = \frac{1}{\psi}\phi(x,t,O_\alpha^{\mathrm{T}}w),$$

we have

$$\begin{aligned}\widehat{\lambda}(\zeta) &= \frac{1}{1+i\boldsymbol{\alpha}\cdot O^{\mathrm{T}}{}_\alpha\zeta}\widehat{M}(O^{\mathrm{T}}{}_\alpha\zeta) = \frac{1}{1+i|\boldsymbol{\alpha}|\zeta_1}\widehat{M}(\zeta)\\ &= \widehat{M^2}(\zeta_2,\zeta_3)\frac{1}{1+i|\boldsymbol{\alpha}|\zeta_1}\widehat{M^1}(\zeta_1)\end{aligned}$$

(where M^d is the gaussian in $\mathbb{R}^d$). Going back to the variables w,

$$\begin{aligned}\lambda(w) &= M^2(w_2,w_3)\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{iw_1\zeta_1}\frac{1-i|\boldsymbol{\alpha}|\zeta_1}{1+|\boldsymbol{\alpha}|^2\zeta_1^2}\widehat{M^1}(\zeta_1)\,d\zeta_1\\ &= M^2(w_2,w_3)(1-|\boldsymbol{\alpha}|\partial_{w_1})(h_\alpha^1 * M^1)(w_1),\end{aligned}$$

where

$$\widehat{h_\alpha^1}(\zeta_1){:=}\frac{1}{1+|\boldsymbol{\alpha}|^2\zeta_1^2},$$

that is,

$$h_\alpha^1(w_1) = \frac{1}{2|\boldsymbol{\alpha}|}e^{-|w_1|/|\boldsymbol{\alpha}|}.$$

Since

$$(1-|\boldsymbol{\alpha}|\partial_{w_1})h_\alpha^1(w_1) = \begin{cases}0, & w_1<0\\ \frac{1}{|\boldsymbol{\alpha}|}e^{-|w_1|/|\boldsymbol{\alpha}|}, & w_1>0,\end{cases}$$

we get

$$\begin{aligned}\lambda(w) &= M^2(w_2,w_3)\frac{1}{\sqrt{2\pi}}\frac{1}{|\boldsymbol{\alpha}|}\int_0^{+\infty} e^{-\frac{y}{|\boldsymbol{\alpha}|}}e^{-\frac{(w_1-y)^2}{2}}\,dy\\ &= M^2(w_2,w_3)\frac{1}{\sqrt{2\pi}}e^{-\frac{w_1^2}{2}}\frac{1}{|\boldsymbol{\alpha}|}e^{\frac{1}{2}(w_1-\frac{1}{|\boldsymbol{\alpha}|})^2}\int_{-(w_1-\frac{1}{|\boldsymbol{\alpha}|})}^{+\infty} e^{-\frac{y^2}{2}}\,dy\\ &= M(w)\frac{1}{|\boldsymbol{\alpha}|}\frac{F(w_1-\frac{1}{|\boldsymbol{\alpha}|})}{M^1(w_1-\frac{1}{|\boldsymbol{\alpha}|})}.\end{aligned}$$

where F is the antiderivative of M^1 with $F(-\infty)=0$ (the Error Function of the Gaussian). We have

$$\begin{aligned}\phi(x,t,v) = \psi(x,t)\lambda(w) &= M(w)\frac{1}{|\boldsymbol{\alpha}|}\frac{F}{M^1}(w_1-\frac{1}{|\boldsymbol{\alpha}|})\\ &= \psi(x,t)M(v)\frac{1}{|\boldsymbol{\alpha}|}\frac{F((\boldsymbol{\alpha}\cdot v-1)/|\boldsymbol{\alpha}|)}{M^1((\boldsymbol{\alpha}\cdot v-1)/|\boldsymbol{\alpha}|)}.\end{aligned}$$

Since we have a similar formula for f with u in place of ψ, it is clear that $sgn(f-\phi) = sgn(u-\psi)$ and so, using (42.11), we obtain

$$I_1 = \iint sgn(u-\psi)\Big[-\eta\partial_t\psi - \mathrm{div}_x\int \psi v M(v)\,dv + \mathrm{div}_x\int \psi v K\,dv\Big]\,dx\,dt,$$

with

$$K{:=}(\boldsymbol{\alpha}\cdot\nabla_v)(1-\boldsymbol{\alpha}\cdot\nabla_v)(h_\alpha * M).$$

Integrating by parts twice, we find that

$$\begin{aligned}\int vK\,dv &= -\boldsymbol{\alpha}\int(1-\boldsymbol{\alpha}\cdot\nabla_v)(h_\alpha * M)\,dv\\ &= -\boldsymbol{\alpha}\int h_\alpha * M\,dv = -\boldsymbol{\alpha}\,\widehat{h_\alpha}(0)\,\widehat{M}(0) = -\boldsymbol{\alpha},\end{aligned}$$

Noting also that $\int vM(v)\,dv = 0$, we get

$$\begin{aligned}I_1 &= \iint sgn(u-\psi)\big[-\eta\partial_t\psi - \mathrm{div}(\psi\alpha)\big]\,dx\,dt\\ &= \eta\iint sgn(u-\psi)\big[-\partial_t\psi - \tau\,\mathrm{div}_x(\psi\nabla_x\theta)\big]\,dx\,dt.\end{aligned}$$

From this (42.8) follows.

References

1. Goudon, T., Nieto, J., Poupaud, F., Soler, J.: Multidimensional high-field limit of the electrostatic Vlasov-Poisson-Fokker-Planck system, J. Diff. Eqs. **213**, 418–442 (2005)
2. Portilheiro, M.: Weak solutions for equations defined by accretive operators I. Proc. R. Soc. Edinb. Sec. A **133A**, 1193–1207 (2003)
3. Portilheiro, M.: Weak solutions for equations defined by accretive operators II: relaxation limits. J. Diff. Eqs. **195**, 66–81 (2003)
4. Portilheiro, M., Tzavaras, A.: Hydrodynamic limits for kinetic equations and the diffusive approximation of radiative transport for acoustic waves. Trans. Am. Math. Soc. **359**, 529–565 (2007)

Chapter 43
Bethe Ansatz Solution of the Finite Bernoulli Matching Model of Sequence Alignment

V.B. Priezzhev and G.M. Schütz

Abstract We map the Bernoulli matching model of sequence alignment to the discrete-time totally asymmetric exclusion process with backward sequential update and step function initial condition. The Bethe ansatz allows for deriving the exact distribution of the length of the longest common subsequence of two sequences of finite lengths X, Y in the Bernoulli mean field approximation.

43.1 Bernoulli Matching Model of Sequence Alignment

Sequence alignment refers to a special kind of pattern recognition problem in which one wishes to quantify the degree of similarity between two sequences of letters taken from some (usually) finite alphabet. Applications range from molecular biology where one wishes to compare DNA or RNA strands or proteins [1]. To computer science where e.g. one wishes to compare to versions of a data file [2]. One important measure for the similarity of two sequences is the length of the longest common subsequence (LCS for short). This problem has a long history of study. There are explicit results from combinatorics where one is interested in the LCS between two random sequences of letters [3].

Given a pair of fixed sequences of c letters of lengths X and Y, the length of their LCS is defined by the recursion [2, 4]

$$L_{X,Y} = \max[L_{X-1,Y}, L_{X,Y-1}, L_{X-1,Y-1} + \eta_{X,Y}] \tag{43.1}$$

with the boundary conditions $L_{i,0} = L_{0,j} = L_{0,0} = 0$ for all $i, j \geq 0$. The variable $\eta_{X,Y}$ is 1 if the letters at the positions X and Y match each other, and 0 if

G.M. Schütz (✉)
Institut für Festkörperforschung, Forschungszentrum Jülich, 52425 Jülich, Germany
e-mail: g.schuetz@fz-juelich.de

V.B. Priezzhev
Laboratory of Theoretical Physics, Joint Institute for Nuclear Research, 141980 Dubna, Russia
e-mail: priezzvb@theor.jinr.ru

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_43,

they do not. The set of variables $\eta_{X,Y}$ form the scoring matrix. One should note that even if the underlying sequences are random, the elements of the scoring matrix are correlated. If one ignores these correlations and takes them as i.i.d. random variables from the bimodal distribution $F(\eta) = p\delta_{\eta,1} + (1-p)\delta_{\eta,0}$, one gets the Bernoulli matching (BM) model of sequence alignment [5]. Here is p is the empirical mean of the scores of the scoring matrix and the quantity of interest is the distribution of the LCS as a function of p.

The hope that gives rise to making such a fairly crude mean field approximation is twofold. Very long sequences may exhibit universal properties that could then be captured in the simplified BM model. Moreover, for finite sequences an exact result for the distribution of the LCS allows for benchmarking, i.e., a quantitative statement about the strength of correlations in the scoring matrix of a real sequence.

In the thermodynamic limit of infinitely long sequences this problem has been studied in some detail. With $X = xN$, $Y = yN$, Seppäläinen derived rigorously the law of large numbers limit. Asymptotically the quantity $L_{X,Y}/N$ is a random variable converging a.s. to a function of p, x, y which he computed explicitly [6]. Using an exact mapping to a directed polymer problem, complemented with scaling arguments, it was shown more recently [7] that asymptotically the quantity $L_{X,Y}$ is a random variable of the form

$$L_{X,Y} \overset{N\to\infty}{\longrightarrow} \gamma_p(x,y)N + \delta_p(x,y)N^{1/3}\chi \tag{43.2}$$

where $\gamma_p(x,y), \delta_p(x,y)$ are known scale factors and χ is a random variable drawn from the Tracy–Widom distribution of the largest eigenvalue of GUE random matrices [8]. A mapping of the sequence alignment problem onto the asymmetric simple exclusion process with sublattice-parallel update has been proposed in [9]. This admits a transfer-matrix formulation and direct hence allows for diagonalization of the transfer matrix to obtain numerically the distribution of the LCS or finite X and Y.

Since our interest lies in an analytical solution for fixed, but arbitrary X and Y, we choose a mapping onto a discrete-time fragmentation process which is equivalent to a totally asymmetric simple exclusion process with backward sequential update [10, 11]. The initial distribution of this TASEP that is obtained by the mapping turns out to be a step configuration with a given number N of particles on sites $-N+1$, $-N+2, \ldots, 0$. This allows us to use earlier results obtained directly from Bethe ansatz [12] for this stochastic lattice gas model. For this model we define $P(M,N,t)$ to be the probability that the Nth particle hops at least M times up to time t.

Specifically, we then express the probability that the length of the LCS is at most Q by the probability that the number of jumps of a selected particle in the exclusion process up to time Y is at least $X - Q$. We define the cumulative distribution

$$\Xi^{Q}_{X,Y} := \mathrm{Prob}[L_{X,Y} \le Q] = \sum_{M=0}^{Q} \Lambda^{M}_{X,Y} \tag{43.3}$$

Our main result is the following theorem [13].

Theorem 43.1.

$$\Xi^{Q}_{X,Y} = P(Y-Q, X-Q, Y) \tag{43.4}$$

with the natural convention that $P(Y, 0, Y) = 1$.

The result (43.4) provides a simple relation between the cumulative distribution of the length of the LCS in the BM model and the distribution of the time-integrated current in the backward-sequential TASEP for the step function initial condition. The probability that the length of the LCS is at most Q is given by the probability that the number of jumps across bond $Y - X$ up to time Y is at least $X - Q$. The proof of the theorem consists of the steps described in the following sections. An explicit expression for $P(Y - Q, X - Q, Y)$ is given below in Theorem 2.

43.2 Mapping to a Fragmentation Process

The mapping of the LCS problem to a one-dimensional discrete-time fragmentation process is described in detail [13]. Here we only summarize the main steps. The scoring matrix generates a rectangular grid on which the recursion (43.2) induces a terrace structure where the value of the LCS for a given pair of sequences is the height of the terrace. It is useful to view the grid that defines the matching matrix as a square lattice with $X \times Y$ bulk sites, embedded in the rectangle of size $(X + 1) \times (Y + 1)$. Each square (defining the dual square lattice) is labelled (i, j) with $0 \leq i \leq X$ and $0 \leq j \leq Y$. Numbers at left corners of terraces appear when $\eta_{X,Y} = 1$ and have weight p. All the rest of numbers at edges of terraces do not depend on $\eta_{X,Y}$ and have therefore weight 1. All remaining numbers appear when $\eta_{X,Y} = 0$ having weight $(1 - p)$. Due to the terraces, each site can take one of five different states. It may be (a) traversed horizontally or vertically by a terrace line (b) represent a left or right corner of a terrace, or (c) be empty. By construction, in the BM model each empty site has weight $q = 1 - p$, each left corner of each line has weight p, and all remaining sites have weight 1. This property allows for a mapping to a five-vertex model. The resulting pattern of intersecting lines then becomes isomorphic to the pattern of in- and outging arrows in the five-vertex model with vertex weights given by the weights of the BM model. One simply identifies black (red) horizontal lines with right-pointing (left-pointing) arrows and black (red) vertical lines with up-pointing (down-pointing) arrows, see Fig. 43.1.

To obtain the fragmentation process, we first turn the vertex lines with arrows pointing left or down into non-intersecting particle world lines by replacing a right-left turn with a diagonal "shortcut". After a space reflection $i \to i' = 1 - i$ this yields a non-intersecting line ensemble. A final mapping is aimed to obtain the line ensemble of particle world lines of the discrete-time totally asymmetric exclusion process (TASEP) with the backward sequential update, introduced in [10] and solved in [11]. This moel is identical tothe fragmentation process studied and solved in [12].

To this end, we consider each trajectory and replace each move upward by a diagonal move right and each diagonal move left by a move upward. In a more formal way, we consider a new square lattice $(i' + j + 1/2, j)$ and draw new trajectories using the correspondence $(i' + 1/2, j) \to (i' + j + 1/2, j)$. The sites of the new lattice are denoted by coordinates (k, t) numbered by integers $k = i' + j$ and $t = j$. By construction the red lines move upward or diagonally and define the world lines of exclusion particles which jump only to the right. The vertex weights assign the appropriate probability to each path ensemble.

The backward sequential dynamics encoded in the vertex weights may be described as follows. In each time particle position are updated sequentially from right to left, starting from the rightmost particle. Each step of a particle by one lattice unit in positive direction has probability $1 - p$, provided the neighbouring target site is empty. If the target site is occupied, the jump attempt is rejected with probability 1. No backward moves are allowed, making the exclusion process totally asymmetric. The horizontal boundary condition of the original sequence matching problem maps into an initial condition where at time $t = 0$ particles occupy consecutive dual lattice points $-X + 1 \leq k \leq 0$. Since the motion of a particle is not influenced by any particles to its left, we may extend the lattice to minus infinity. The vertical boundary condition is equivalent to extending the lattice to plus infinity, such that at time $t = 0$ all sites $k > 0$ are vacant. Thus one has a TASEP on an initially half-filled infinite lattice with step initial state. However, only the first X particles contribute to the statistical properties of the BM model.

In the exclusion picture the terrace height has a simple probabilistic interpretation. It counts the number of world lines that intersects with a diagonal in the square lattice starting from the point $(k, t) = (-x, 0)$ (the left dotted line in Fig. 43.1.

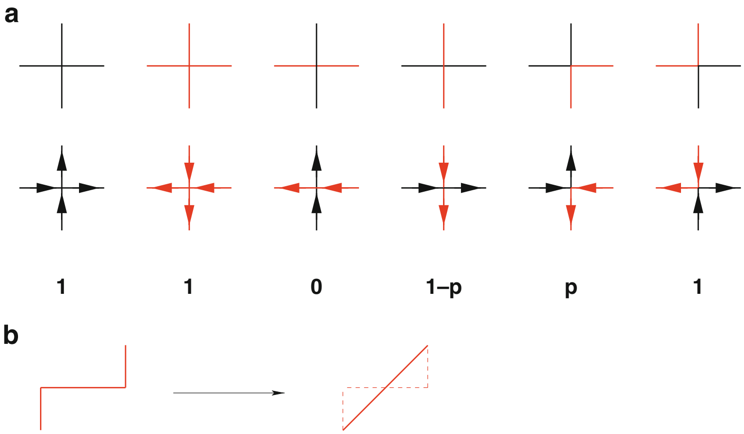


Fig. 43.1 (**a**) Mapping of line intersection to vertices of the six-vertex model which is effectively a five-vertex model since one of the vertex weights is zero. (**b**) A way to avoid line intersections in the five vertex model: if a horizontal line has a left adjacent vertical line below and a right vertical line above, it is replaced by the diagonal shortcut. The figure is taken from [13]

Hence, at any given time step, the terrace increases at each site from right to left by one unit, unless a world line has been crossed when going from right to left. Therefore the number n of trajectories ending at time $t = Y$ and the length L_{XY} of the LCS of the BM model on the rectangle $(X + 1) \times (Y + 1)$ are related by $L_{X,Y} = X - n$.

43.3 Current Distribution in the Fragmentation Process

In what seems to be a departure of the topic of this paper we recall the following theorems for the fragmentation process, proved in [12]. Consider the fragmentation process defind above with an initial state with a particle on a site k_{max} such that for all $k > k_{max}$ the lattice is empty. Here $k_{max} < \infty$ is arbitrary and will without loss of generality taken to be 0. The particle on site k_{max} is called the rightmost particle. Define furthermore the polynomials

Definition 43.1.

$$\begin{aligned} D_q(n,t) &= \frac{1}{2\pi}\int_0^{2\pi} dk \left(1 - p + pe^{-ik}\right)^t \left(1 - e^{i(k+i0)}\right)^{-p} e^{ikn} \\ &= \frac{1}{2\pi i}\oint_{|z|=1-0} dz \left(1 - p + \frac{p}{z}\right)^t (1-z)^{-q} z^{n-1} \end{aligned} \tag{43.5}$$

One has [12]:

Theorem 43.2. *Let A_N be the set of sites on which the N rightmost particles are located at time $t = 0$ and B_N be another set of sites. Let $Q(A_N, B_N; t)$ be the conditional probability to find the N rightmost particles on B_N at time step t, given that they started on A_N at time step* 0*. Then*

$$Q(A_N, B_N; t) = \begin{vmatrix} D_0(k_1 - l_1; t) & D_{-1}(k_1 - l_2; t) & \cdots & D_{-N+1}(k_1 - l_N; t) \\ D_1(k_2 - l_1; t) & D_0(k_2 - l_2; t) & \cdots & D_{-N+2}(k_2 - l_N; t) \\ \vdots & \vdots & & \vdots \\ D_{N-1}(k_N - l_1; t) & D_{N-2}(k_N - l_2; t) & \cdots & D_0(k_N - l_N; t) \end{vmatrix}. \tag{43.6}$$

Proof. The proof uses the evolution equation of the conditional probability $Q(A_N, B_N; t)$ together with fact that for the totally asymmetric dynamics the motion of the first N particles is unaffected by all particles to their left. One proves by induction in N that $Q(A_N, B_N; t)$ follows the evolution equation with the prescribed initial condition, using a series of elementarr determinant manipulations and recursion relations for the functions $D_q(n, t)$, see [12] for details. The proof follows along the lines of a similar result for the continuous-time totally asymmetric simple exclusion process [14]. □

We remark that this proof is simple, but non-constructive. In actual fact, the determinant was determined without paying attention to mathematical rigour by Bethe ansatz. In this sense this is a solution of the fragmentation process by Bethe ansatz.

This theorem is the decisive ingredient in the proof of the following theorem [12] for the step function initial condition where all sites with $k \leq 0$ are occupied.

Theorem 43.3. *Let $A_N = 0, -1, -2, \ldots, -N+1$ and let $P(M,N,t)$ be the probability that the Nth particle (which was on the $(1-N)$th site of the infinite cluster at $t = 0$) hops at least M times up to time t. Then*

$$P(M,N,t) = Z(M,N)^{-1} \sum_{t_1,t_2,\cdots,t_N=0}^{t-1-M+N} \prod_{j=1}^{N} \big(t_j + M - NM \\ -N(1-p)^{t_j}\big) \prod_{i<j} (t_i - t_j)^2. \tag{43.7}$$

Proof. The proof for this theorem is much more involved than the proof of Theorem 2. It involves a summation over the determinantal transition probabilities (43.6). Somewhat miraculously this sum over determinants is again a determinant. A series further determinant manipulations then proves the result. For details see [12]. □

Remark: A very interesting extension to other initial conditions was obtained by Nagao and Sasamoto [15].

43.4 Proof of the Distribution of the LCS in the BM Model

With the results of the preceding sections we arrive at the main conclusion of this work. Our aim is the evaluation of the probability distribution $\Lambda^{Q}_{XY} = \mathrm{Prob}[L_{X,Y} = Q]$ of the Bernoulli model. Having the TASEP interpretation of the original model, we need to evaluate an appropriate sum over end points of trajectories of particles. To do this, we select the first trajectory (counted from the right) which does not end at time Y in the target range of the dual lattice given by the top row $(i, Y$ with $1 \leq i \leq X$ (the green line in Fig. 43.1). An important observation is that the sum of weights of all trajectories ending at times $T_1, T_2, \ldots T_k < Y$ (all lines to the left of the green line) is 1 for the conservation of probabilities in the TASEP. Then, the distribution Λ^{Q}_{XY} is the sum over the probabilities of all trajectories with end points right of the green line and over end points of the green line itself. Hence of all X particles only the rightmost $n + 1 = X - Q + 1$ particles are relevant for the computation of Λ^{Q}_{XY}. The initial positions of these particles are $k_1 = -X + Q, k_2 = k_1 + 1, k_3 = k_2 + 1, \ldots, k_{n+1} = 0$.

Following the relation between terrace height Q and particle trajectories as discussed above, we may consider the final positions $x_1, x_2, \ldots, x_{n+1}$ at the moment of time Y. By the construction, we have $Y - X + 1 \leq x_2 < x_3 < \cdots < x_{n+1} \leq Y$

and $x_1 \leq Y - X$. We first consider $Q = X$. In this case no particle has reached site $Y - X + 1$. In particular, this implies that the first particle (initially at site 0) has not reached site $Y - X + 1$. The complementary probability for this event is the probability $P(Y - X + 1, 1, Y)$ that the first particle has jumped at least $Y - X + 1$ times up to time Y. Hence

$$\Lambda^{X}_{X,Y} = 1 - P(Y - X + 1, 1, Y). \tag{43.8}$$

Now consider $Q < X$. Then Λ^{Q}_{XY} is the joint probability that after Y time steps all rightmost $X - Q$ particles (located initially on $(-X + Q + 1, \ldots, 0)$) have reached sites $\geq Y - X + 1$ and the next particle (located initially on $-X + Q$) has not reached site $Y - X + 1$. This is equivalent to the joint probability that the particle originally at $-X + Q + 1$ has jumped at least $Y - Q$ times and the particle originally at $-X + Q$ has jumped not more than $Y - Q$ times. By construction of the process this joint probability may be expressed as the statistical weight of all paths where the particle initially at $-X + Q + 1$ jumps at least $Y - Q$ times minus the statistical weight of all paths where the particle initially at $-X + Q$ jumps at least $Y - Q + 1$ times.

We have come to a known problem of the TASEP statistics [12, 16]. Consider an infinite chain, the left half of which is initially occupied by particles while the right half is empty. The problem is to find the probability $P(M, N, t)$ that the Nth particle (counted from the right) of the infinite cluster hops at least M times up to time t. With this quantity, we obtain for the partition function the expression

$$\Lambda^{Q}_{X,Y} = P(Y - Q, X - Q, Y) - P(Y - Q + 1, X - Q + 1, Y) \tag{43.9}$$

for every $Q < X$. For $Q = X$, (43.9) reduces to (43.8). We remark that (43.8) may be viewed as incorporated in (43.9) in agreement with the notion that the transition probability in an exclusion process with no particles (second argument of P for $X + Q$) is equal to 1 (this is the trivial transition from the empty lattice to the empty lattice).

To derive (43.9) more formally, one sums over transitions probabilities of the form (43.6) for the step initial condition to obtain the distribution of the LCS in terms of the hopping distribution $P(M, N, t)$. With (43.9) the main theorem is proved.

We remark that in [12] it was shown explicitly that $P(Y - Q, X - Q, Y) = P(X - Q, Y - Q, X)$ which for the BM model is expected by symmetry. In [12] the hopping distribution of the fragmentation process was also expressed in terms of the analogous quantity for the TASEP with parallel update which computed was computed earlier by by Johansson [16] with combinatorial methods of the theory of symmetric groups. Applying methods from random matrix theory he was able to extract the asymptotic behaviour for large t. From this one recovers after some computation the asymptotics (43.2), see [13].

Acknowledgements We acknowledge financial support by DFG and RFBR. We thank the organizers of the conference Dynamics and Applications for organizing a stimulating meeting. We thank S.N. Majumdar and S. Nechaev for useful discussions.

References

1. Waterman, M.S.: Introduction to Computational Biology. Chapman and Hall, London (1994)
2. Gusfield, D.: Algorithms on Strings, Trees, and Sequences. Cambridge University Press, Cambridge (1997)
3. Sankoff, D., Kruskal, J.: Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Composition. Addison Wesley, Reading, Massachussets (1983)
4. Bundschuh, R., Hwa, T.: Discrete Appl. Math. **104**, 113 (2000)
5. Boutet de Monvel, J.: Eur. Phys. J. B **7**, 293 (1999); Phys. Rev. E **62**, 204 (2000)
6. Seppälainen, T.: Ann. Appl. Probab. **7**(4), 886 (1997)
7. Majumdar, S.N., Nechaev, S.: Phys. Rev. E **72**, 020901(R) (2005)
8. Tracy, C.A., Widom, H.: Commun. Math. Phys. **159**, 151 (1994); **177**, 727 (1996)
9. Bundschuh, R.: Phys. Rev. E **65**, 031911 (2002)
10. Rajewsky, N., Schadschneider, A., Schreckenberg, M.: J. Phys. A **29**, L305 (1996)
11. Priezzhev, V.B.: In: Proceedings Statphys 22 Pramana-J.Phys. **64**, 915 (2005); cond-mat/0211052 (2002)
12. Rákos, A., Schütz, G.M.: J. Stat. Phys. **118**, 511 (2005)
13. Priezzhev, V.B., Schütz, G.M.: J. Stat. Mech. Theor. Exp. P09007 (2008)
14. Schütz, G.M.: J. Stat. Phys. **88**, 427 (1997)
15. Nagao, T., Sasamoto, T.: Nucl. Phys. B **699**, 487 (2004)
16. Johansson, K.: Commun. Math. Phys. **209**, 437 (2000)

Chapter 44
Fitness Function Evaluation Through Fractional Algorithms

Cecília Reis and J.A. Tenreiro Machado

Abstract This paper proposes a Genetic Algorithm (GA) for the design of combinational logic circuits. The fitness function evaluation is calculated using Fractional Calculus. This approach extends the classical fitness function by including a fractional-order dynamical evaluation. The experiments reveal superior results when comparing with the classical method.

44.1 Introduction

In the last decade GAs have been applied in the design of electronic circuits, leading to an area of research called Evolutionary Electronics (EE) [6]. EE considers the concept for automatic design of electronic systems. Instead of using human conceived models, abstractions and techniques, EE employs search algorithms to develop good designs. GAs are adaptive heuristic search algorithms based on the evolutionary ideas of natural selection and genetics [1]. GAs employ a population of individuals that undergo selection in the presence of operators such as mutation and crossover. A fitness function is used to evaluate the individuals. In this study, the evaluation is performed through a fractional-order dynamic fitness function.

The area of Fractional Calculus (FC) deals with the operators of integration and differentiation to an arbitrary order and is as old as the theory of classical differential calculus [3, 4]. FC is a tool well-adapted tool for the modelling of many physical phenomena, taking into account some peculiarities that classical integer-order models neglect.

Bearing these ideas in mind the article is organized as follows. Section 43.2 describes the adopted GA as well as the fractional-order dynamic fitness functions. Section 43.3 presents the experiments and the simulation results. Finally, Sect. 43.4 outlines the main results.

C. Reis (✉) and J.A.T. Machado
Department of Electrotechnical Engineering, Institute of Engineering of Porto, Rua Dr. António Bernardino de Almeida, 431, 4200-072 Porto, Portugal
e-mail: cmr@isep.ipp.pt, jtm@isep.ipp.pt

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_44,

44.2 The Genetic Algorithm

The circuits are specified by a truth table and the goal is to implement a functional circuit with the least possible complexity. Two gate sets were defined: Gset a = {AND,XOR,WIRE} and Gset b = {AND,OR,XOR,NOT,WIRE}. The logic gate denoted WIRE means a logical no-operation. In the presented scheme the circuits are encoded as a rectangular matrix of logic cells and each cell as three genes: *<input1> <input2> <gate type>* [5]. The gate type is one of the elements adopted in the gate set. The chromosome is formed with as many triplets as the matrix size demands. The initial population of circuits is generated at random. The search is then carried out among this population. The three different operators used in the GA are reproduction, crossover and mutation. Single point crossover is performed. The crossover point is only allowed between cells to maintain the chromosome integrity. The mutation operator changes the characteristics of a given cell in the matrix, meaning that a completely new cell can appear in the chromosome. Moreover, an elitist algorithm is applied and, consequently, the best solutions are always kept for the next generation.

To run the GA we have to define the number of individuals to create the initial population P. This population is always the same size across the generations, until the solution is reached. The crossover rate *CR* represents the percentage of the population P that reproduces in each generation. Likewise, the mutation rate *MR* is the percentage of the population P that can mutate in each generation.

The goal is to find new ways of evaluating the individuals of the population in order to achieve better performance GAs. We propose two concepts for the fitness functions, namely the static fitness function F_s and the dynamic fitness function F_d. The calculation of F_s is divided in two parts, f_1 and f_2 such that $f_1 = f_{11} - \delta$ *if* $error_i \neq error_{i-1}$ and $f_2 = f_2 + 1$, *if gate type = wire*, where f_1 measures the functionality and the error discontinuity and f_2 measures the simplicity. In a first phase, we compare the output $\mathbf{Y}$ produced by the GA-generated circuit with the required values $\mathbf{Y}_R$, according with the truth table, on a bit-per-bit basis, namely $f_{11} = f_{11} + 1$, *if* $\{bit\ i\ of\ \mathbf{Y}\} = \{bit\ i\ of\ \mathbf{Y_R}\}$, $i = 1, \ldots, f_{10}$. Therefore, f_{11} is incremented by *one* for each correct bit of the output until f_{11} reaches the maximum value $f_{10} = 2^{ni} \times no$, that occurs, when we have a functional circuit. After this, f_{11} is decremented by δ for each $\mathbf{Y}_R - \mathbf{Y}$ error discontinuity, where discontinuity means passing from $\mathbf{Y}_R - \mathbf{Y} = 0$ to $\mathbf{Y}_R - \mathbf{Y} = 1$ or vice-versa when comparing two consecutive levels of the truth table. Once the circuit is functional, in a second phase, the GA tries to generate circuits with the least number of gates. This means that the resulting circuit must have as much genes *<gate type>* ≡ *<wire>* as possible. Therefore, the index f_2, that measures the simplicity (the number of null operations), is increased by *one* (*zero*) for each *wire* (*gate*) of the generated circuit. The static fitness function yields:

$$F_s = \begin{cases} f_1, & F_s < f_{10} \\ f_1 + f_2, & F_s \geq f_{10} \end{cases} \tag{44.1}$$

where $i = 1, \ldots, f_{10}$, and *ni* and *no* represent the number of inputs and outputs of the circuit. The concept of dynamic fitness function F_d results from an analogy with control systems, where we have a variable to be controlled, similarly with the GA case, where we master the population through the fitness function. The simplest control system is the proportional algorithm; nevertheless, there can be other control algorithms, such as, for example, the proportional and the differential actions. In this line of thought, applying the static fitness function corresponds to using a kind of proportional algorithm. Therefore, to implement a proportional-integral-differential evolution the fitness function needs a scheme of the type: $F_d = F_s + K_I I^{\lambda}[F_s] + K_D D^{\mu}[F_s]$, where $0 \leq \lambda, \mu \leq 1$ is the integral-differential fractional-order and K is the 'gain' of the dynamical term.

The Grünwald–Letnikov formulation [2] inspired a discrete-time calculation algorithm, based on the approximation of the time increment h through the sampling period T and a r-term truncated series yielding the equation:

$$D^{\mu}[x(t)] \approx \frac{1}{T^{\alpha}} \sum_{k=0}^{r} \frac{(-1)^k \Gamma(\mu+1)}{k!\ \Gamma(\mu-k+1)} x(t-kT) \tag{44.2}$$

where Γ is the gamma function.

44.3 Experiments and Simulation Results

A reliable execution and analysis of a GA requires a large number of simulations to provide a reasonable assurance that stochastic effects have been properly considered. Therefore, we developed $n = 1{,}000$ simulations for each case. The experiments consist on running the GA to generate a 2-to-1 multiplexer ($M2-1$), using the fitness scheme described previously. The circuits are generated with the two gate sets presented for $CR = 95\%$, $MR = 20\%$. $P = 100$ and the implementation of the differential/integral fractional order operator adopts (8.2) with a series truncation of $r = 50$ terms. A superior GA performance means achieving solutions with a smaller number N of generations. Due to the huge number of possible combinations of the GA parameters, in the sequel we evaluate only a limited set of cases. Therefore, a priori, other values can lead to different results. Nevertheless, the authors developed an extensive number of experiments and concluded that the following cases are representative.

Figure 44.1 shows the results obtained for the $M2-1$ circuit, in terms of the average number of generations to achieve the solution $AV(N)$, for $PI^{1/4}D^{1/4}$ *versus* $K = K_D = K_I$ and $\delta = \{0.0, 0.25, 0.5, 0.75, 1.0\}$ with Gsets a and b. We verify the superior results for $\delta = 0.25$, $K = 0.01$ and $\lambda = \mu = 0.25$.

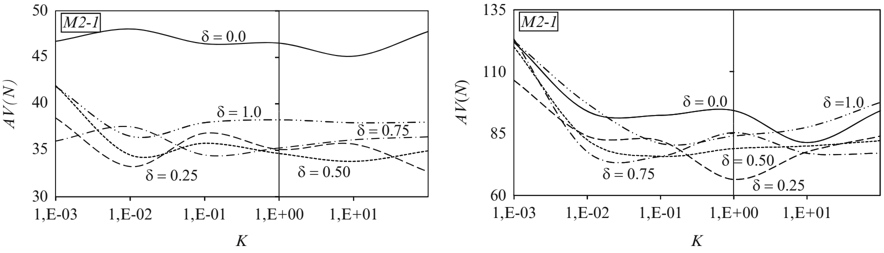


Fig. 44.1 Results obtained for the 2-to-1 multiplexer circuit with Gsets a and b

44.4 Conclusions

This paper presented two techniques for improving the GA performance. Firstly, we concluded that we get superior results by measuring the error discontinuity. Secondly, we verified that, the new concept of fractional-order dynamic fitness function constitutes an important method to outperform the classical static fitness function approach.

References

1. Goldberg, D.E.: Genetic Algorithms in Search Optimization and Machine Learning. Addison-Wesley, Reading, MA (1989)
2. Machado, J.: Analysis and Design of Fractional-Order Digital Control Systems. Systems Analysis-Modelling-Simulation, vol. 27, Issue 2–3, pp. 107–122. Gordon & Breach Science Publishers, Newark, NJ (1997)
3. Miller, K., Ross, B.: An Introduction to the Fractional Calculus and Fractional Differential Equations. Wiley, New York (1993)
4. Oldham, K., Spanier, J.: The Fractional Calculus: Theory and Application of Differentiation and Integration to Arbitrary Order. Academic, New York (1974)
5. Reis, C., Machado, J., Cunha, J.: Evolutionary Design of Combinational Logic Circuits. J. Adv. Comput. Intell. Intell. Informat. 507–513 (2004)
6. Zebulum, R., Pacheco, M., Vellasco, M.: Evolutionary Electronics: Automatic Design of Electronic Circuits and Systems by Genetic Algorithms. CRC, Boca, Raton (2001)

Chapter 45
An Exponential Observer for Systems on SE(3) with Implicit Outputs

Sérgio S. Rodrigues, Naveena Crasta, António Pedro Aguiar, and Fátima Silva Leite

Abstract This paper considers the state estimation problem of a class of systems described by implicit outputs and whose state lives in the special Euclidean group SE(3). This type of systems are motivated by applications in dynamic vision such as the estimation of the motion of a camera from a sequence of images. We propose an observer in the group of motion SE(3) and discuss conditions under which the linearized state estimation error converges exponentially fast. We also analyze the problem when the system is subject to disturbances and noises. We show that the estimate converges to a neighborhood of the real solution. The size of the neighborhood increases/decreases gracefully with the bound of the disturbance and noise.

45.1 Introduction

During the last few decades there has been an extensive study on the design of observers for nonlinear systems. In simple terms, an observer or estimator can be defined as a process that provides in real time the estimate of the state (or some function of it) of the plant from partial and possibly noisy measurements of the inputs and outputs, and inexact knowledge of the initial condition.

N. Crasta and A.P. Aguiar (✉)
Institute for Systems and Robotics, Instituto Superior Técnico, Technical University of Lisbon, Lisbon, Portugal
e-mail: ncrasta@isr.ist.utl.pt, pedro@isr.ist.utl.pt

S.S. Rodrigues
Department of Mathematics, University of Cergy-Pontoise, UMR CNRS 8088, 95000 Cergy-Pontoise, France
e-mail: sesiro@gmail.com

F.S. Leite
Department of Mathematics and Institute for Systems and Robotics, University of Coimbra, Coimbra, Portugal
e-mail: fleite@mat.uc.pt

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_45,

For linear systems evolving on n-dimensional vector spaces, state observer and filter designs employ the traditional Kalman filter [1] and Luenberger type observer [2]. In fact, it is well-known that the Kalman filter [1] is the optimal state estimation algorithm in a well defined sense [27].

For nonlinear systems, the extended Kalman filter is a widely used method for estimating the state. It is obtained by linearizing the nonlinear dynamics and the observation along the trajectory of the estimate. However, if there are substantial nonlinearities or the state lives in some special manifold, there are no guarantees that the state estimate will evolve in the same manifold and even that the estimate will converge to a neighborhood of the true one.

These problems are particularly relevant because they arise in many modern day applications such as the motion control of unmanned aerial vehicles, underwater vehicles, and autonomous robots. See e.g. [9, 17–19] and [4, 6] for other engineering applications such as an exothermic chemical reactor, and a velocity-aided inertial navigation. Typically, these applications require the design of robust nonlinear observers for systems evolving on Lie groups.

Motivated by the above considerations in [3, 13–16, 25] a geometrical framework for the design of symmetric preserving observers on finite-dimensional Lie groups is described. In [5], it is shown that when the output map associated with a left-invariant dynamics on an arbitrary Lie group is right-left equivariant, then it is possible to build non-linear observers such that the error equation is autonomous.

In this paper, we consider left-invariant dynamical systems with implicit outputs, for which the results mentioned above do not apply. Systems of this kind typically arise in mobile robotic applications using dynamic vision such as the estimation of a motion of a camera from a sequence of images. In particular, in [7] and [8], the problem of estimating the position and orientation of a controlled rigid body using measurements from a monocular charged-coupled-device (CCD) camera attached to the vehicle is addressed. The reader is referred to [10–12] for several other examples of implicit output systems in the context of motion and shape estimation.

We propose an observer in the group of motion SE(3) and discuss conditions under which the linearized state estimation error converges exponentially fast. We also analyze the problem when the system is subject to disturbances and noises. We show that the estimate converges to a neighborhood of the real solution. The size of the neighborhood increases/decreases gracefully with the bound of the disturbance and noise.

The outline of the paper is as follows. Section 45.2 introduces the mathematical preliminaries and Sect. 45.3 formulates the state estimation problem. In Sect. 45.4 we propose a left-invariant dynamic observer for estimating the state of systems on SE(3) with implicit outputs, and determine under what conditions the state estimate converges exponentially to the true state. In Sect. 45.5 we analyze the robustness of the proposed observer in the presence of disturbance and noise. Concluding remarks are given in Sect. 45.6.

45.2 Mathematical Preliminaries

In this section we introduce notations and definitions used through out this paper. We denote the Euclidean norm in $\mathbb{R}^n$ by $\|\cdot\|$, and the identity matrix of size n by I_n. Given $A \in \mathbb{R}^{n\times n}$, we let $\det(A)$ and $\mathrm{Tr}(A)$ denote the determinant and the trace of the matrix A, respectively. We consider the scalar product of $A, B \in \mathbb{R}^{n\times n}$ as being defined by $\langle A, B\rangle \overset{\text{def}}{=} \mathrm{Tr}(A^{\mathrm{T}}B)$. The corresponding norm $\|A\| = \sqrt{\langle A, A\rangle}$ is the so-called Frobenius norm. Further, if the entries of $A \in \mathbb{R}^{n\times n}$ depend on t, and $A(t)$ is invertible for all t, from the identity $A^{-1}(t)A(t) = I_n$, one may deduce

$$\frac{\mathrm{d}}{\mathrm{d}t}(A^{-1}(t))A(t) + A^{-1}(t)\frac{\mathrm{d}}{\mathrm{d}t}(A(t)) = 0. \tag{45.1}$$

The cross product of vectors $u, v \in \mathbb{R}^3$ is denoted by $u \times v$. For every $u \in \mathbb{R}^3$,

$$(u\times) = \begin{bmatrix} 0 & -u_3 & u_2 \\ u_3 & 0 & -u_1 \\ -u_2 & u_1 & 0 \end{bmatrix}$$

denotes the matrix representation of the linear map $v \mapsto u \times v$, $v \in \mathbb{R}^3$. It can be easily shown that, for every $u, v \in \mathbb{R}^3$, $\mathrm{Tr}((u\times)^{\mathrm{T}}(v\times)) = 2u^{\mathrm{T}}v$. Given a vector $u \in \mathbb{R}^3$, we denote by $\bar{u} \in \mathbb{R}^4$ its homogeneous coordinates, that is, $\bar{u} = \begin{bmatrix} u \\ 1 \end{bmatrix}$ [26].

The *special orthogonal* group in three-dimensions is denoted by $\mathrm{SO}(3) \overset{\text{def}}{=} \{R \in \mathbb{R}^{3\times 3} \colon R^{\mathrm{T}}R = I_3 \text{ and } \det(R) = +1\}$ and its Lie algebra, that is, the space of all skew-symmetric matrices by $\mathrm{so}(3) \overset{\text{def}}{=} \{(u\times) \in \mathbb{R}^{3\times 3} \colon u \in \mathbb{R}^3\}$.

The *special Euclidean* group is denoted by

$$\mathrm{SE}(3) \overset{\text{def}}{=} \left\{ \begin{bmatrix} g_R & g_T \\ 0 & 1 \end{bmatrix} \in \mathbb{R}^{4\times 4} \colon g_R \in \mathrm{SO}(3) \text{ and } g_T \in \mathbb{R}^3 \right\}$$

and its Lie algebra is defined by $\mathrm{se}(3) \overset{\text{def}}{=} \left\{ \begin{bmatrix} (\omega\times) & v \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{4\times 4} \colon \omega,\ v \in \mathbb{R}^3 \right\}$.

For every $g = \begin{bmatrix} g_R & g_T \\ 0 & 1 \end{bmatrix} \in \mathrm{SE}(3)$, we have $g^{-1} = \begin{bmatrix} g_R^{-1} & -g_R^{-1}g_T \\ 0 & 1 \end{bmatrix}$. Since $g^{-1}g = I_4$, we have $\dot{g} \overset{\text{def}}{=} \frac{\mathrm{d}g}{\mathrm{d}t} = g\left(-\frac{\mathrm{d}}{\mathrm{d}t}g^{-1}\right)g$. Thus, we can rewrite $\dot{g} = g\Omega$ where

$$\Omega \overset{\text{def}}{=} -\left(\frac{\mathrm{d}}{\mathrm{d}t}g^{-1}\right)g \in \mathrm{se}(3).$$

We notice that in order to verify that $\Omega \in \mathrm{se}(3)$ it is sufficient to show the following:

(i) $-\left(\frac{\mathrm{d}}{\mathrm{d}t}g^{-1}\right)g = g^{-1}\dot{g} = \begin{bmatrix} g_R^{-1} & -g_R^{-1}g_T \\ 0 & 1 \end{bmatrix}\begin{bmatrix} \dot{g}_R & \dot{g}_T \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} g_R^{-1}\dot{g}_R & -g_R^{-1}\dot{g}_T \\ 0 & 0 \end{bmatrix}$,

(ii) $(g_R^{-1}\dot{g}_R)^{\mathrm{T}} = (\dot{g}_R)^{\mathrm{T}}g_R = \left(\frac{\mathrm{d}}{\mathrm{d}t}g_R^{\mathrm{T}}\right)g_R = \left(\frac{\mathrm{d}}{\mathrm{d}t}g_R^{-1}\right)g_R = -g_R^{-1}\dot{g}_R.$

We next present a result that will be useful later in the paper.

Lemma 45.1. *Consider* $\xi = \begin{bmatrix} \xi_R & \xi_T \\ 0 & 0 \end{bmatrix} \in \mathrm{se}(3)$*, where* $\xi_R = (\bar{\xi}\times)$ *and* $\bar{\xi}, \xi_T \in \mathbb{R}^3$*. Then* $\|\xi\|^2 = 2\|\bar{\xi}\|^2 + \|\xi_T\|^2$*.*

Proof. A simple computation yields $\xi^{\mathrm{T}}\xi = \begin{bmatrix} \xi_R^{\mathrm{T}}\xi_R & \xi_R^{\mathrm{T}}\xi_T \\ \xi_T^{\mathrm{T}}\xi_R & \xi_T^{\mathrm{T}}\xi_T \end{bmatrix}$. Then by the definition $\|\xi\|^2 = \mathrm{Tr}\left(\xi^{\mathrm{T}}\xi\right) = \mathrm{Tr}(\xi_R^{\mathrm{T}}\xi_R) + \|\xi_T\|^2$. Now the result follows from $\mathrm{Tr}(\xi_R^{\mathrm{T}}\xi_R) = \mathrm{Tr}((\bar{\xi}\times)^{\mathrm{T}}(\bar{\xi}\times)) = 2\|\bar{\xi}\|^2$. □

The Lie bracket of two matrices $A, B \in \mathbb{R}^{n\times n}$ is denoted by $[A, B]$ or, equivalently, $\mathrm{ad}_A B$, and is defined as the commutator $[A, B] = AB - BA$. Given $A, B \in \mathbb{R}^{n\times n}$, we denote $\mathrm{ad}_A^1 B = \mathrm{ad}_A B$ and $\mathrm{ad}_A^{k+1} B = \mathrm{ad}_A \mathrm{ad}_A^k B$ for every $k \in \mathbb{N}$.

45.3 Problem Statement

Consider a left-invariant dynamical system evolving on SE(3), described by

$$\dot{g}(t) = g(t)\mathit{\Omega}(t), \quad g(0) = g_0, \tag{45.2}$$

where $\mathit{\Omega}$ takes values in se(3) and is assumed to be known for all $t \geq 0$.

Consider a set of given points $p_1, \dots, p_N \in \mathbb{R}^3$, and let $y_j = [y_{j_1} \;\; y_{j_2} \;\; 1]^{\mathrm{T}} \in \mathbb{R}^3$, $j \in \mathscr{J}$ be the outputs of the dynamical system (45.2) given implicitly by

$$\alpha_j(t)y_j(t) = F(t)\mathit{\Pi}_0 g(t)\bar{p}_j, \tag{45.3}$$

where $\mathscr{J} \subseteq \{1, 2, \dots, N\}$ is an index set that may depend on time, $\bar{p}_j \in \mathbb{R}^4$ is the homogeneous representation of p_j, the α_j's are unknown scalar continuous function of time satisfying $\alpha_j(t) > 0$ for every $t \geq 0$, $F \in \mathbb{R}^{3\times 3}$ is a known nonsingular matrix, and $\mathit{\Pi}_0 = \left[I_3 \; 0\right] \in \mathbb{R}^{3\times 4}$ is often referred to as the standard (or canonical) projection matrix [26]. We assume that the right-hand-side of (45.3) and $\mathit{\Pi}_0 g(t)\bar{p}_j$ are both bounded below and above, that is, for all $t \geq 0$,

$$m \leq \|F\mathit{\Pi}_0 g\bar{p}_j\|, \|\mathit{\Pi}_0 g\bar{p}_j\| \leq M \quad \text{with} \qquad 0 < m \leq M. \tag{45.4}$$

The problem addressed in this paper can be stated as follows. *Consider the continuous-time left-invariant dynamical system described by* (45.2)–(45.3)*. Let* $\hat{g} \in \mathrm{SE}(3)$ *be the estimate of the state g with a given initial estimate* $\hat{g}(0) = \hat{g}_0$*.*

Design a state observer for (45.2)–(45.3) *that accepts as inputs the measured input* $\Omega(\tau)$ *and the output of the process* $y_j(\tau)$ *for every* $\tau \in [0, t)$, $j \in \mathscr{J}$, *and returns* $\hat{g}(t)$ *at time* t, *for every* $t \geq 0$. *The observer should satisfy some desired performance and robustness properties that will be mentioned later in the paper.*

Remark 45.1. System (45.2)–(45.3) arises for example when one needs to estimate the position and orientation of a robotic vehicle using measurements from an onboard monocular charged-coupled-device (CCD) camera. In that case, adopting the frontal pinhole camera model [26], the scalar α_j captures the unknown depth of a point p_j, and F is a matrix transformation that depends on the parameters of the camera such as the focal length, the scaling factors, and the center offsets. The assumption in (45.4) is very reasonable and only means that the image points are well defined in the sense that they live in some compact set. Notice that if for some point that assumption does not hold, then this only implies to take it out from the index set $\mathscr{J}$.

45.4 Observer Design and Convergence Analysis

Consider the continuous-time left-invariant dynamical system (45.2)–(45.3). We propose the nonlinear observer

$$\dot{\hat{g}}(t) = \hat{g}(t)\Omega(t) + \zeta\, \Theta(\hat{g}(t), y(t))\, \hat{g}(t), \quad \hat{g}(0) = \hat{g}_0, \tag{45.5}$$

where $\hat{g} \in \mathrm{SE}(3)$ is the estimate of the state g, and $\Theta(\hat{g}, y) \in \mathrm{se}(3)$ is given by

$$\Theta(\hat{g}, y) \overset{\text{def}}{=} \begin{bmatrix} \Theta_R(\hat{g}, y) & \Theta_T(\hat{g}, y) \\ 0 & 0 \end{bmatrix}, \tag{45.6}$$

with

$$\Theta_R(\hat{g}, y) = \sum_{j \in \mathscr{J}} \frac{1}{D(\hat{g}\bar{p}_j)} ((((\tilde{y}_j \times \Pi_0 \hat{g}\bar{p}_j) \times \Pi_0 \hat{g}\bar{p}_j) \times \Pi_0 \hat{g}\bar{p}_j)\times), \tag{45.7}$$

$$\Theta_T(\hat{g}, y) = \sum_{j \in \mathscr{J}} \frac{1}{D(\hat{g}\bar{p}_j)} ((-2\tilde{y}_j \times \Pi_0 \hat{g}\bar{p}_j) \times \Pi_0 \hat{g}\bar{p}_j), \tag{45.8}$$

$$\tilde{y}_j = F^{-1} \frac{y_j}{\|y_j\|}, \tag{45.9}$$

where

$$D(\hat{g}\bar{p}_j) \overset{\text{def}}{=} (\#\mathscr{J}) \|\Pi_0 \hat{g}\bar{p}_j\|^2 (1 + \|\Pi_0 \hat{g}\bar{p}_j\|), \tag{45.10}$$

$\#\mathscr{J}$ being the number of elements of $\mathscr{J}$, and $\zeta > 0$ is a tuning constant. Since $\alpha_j > 0$, the expressions (45.3) and (45.9) imply that

$$\tilde{y}_j = \frac{\Pi_0 g \bar{p}_j}{\|F \Pi_0 g \bar{p}_j\|}. \tag{45.11}$$

Remark 45.2. Notice that by defining $\hat{\Theta} \overset{\text{def}}{=} \hat{g}^{-1}\Theta\hat{g}$, system (45.5) can be rewritten as

$$\dot{\hat{g}} = \hat{g}(\Omega + \zeta\hat{\Theta}),$$

and, by a direct computation we can show that $\hat{\Theta} \in \text{se}(3)$. Thus, like the dynamics of g in (45.2), also the dynamics of $\hat{g}$ is *left-invariant*. Moreover, if $\hat{g}(0) = g(0)$, then $\hat{\Theta} = 0$ for every $t \geq 0$, which means that the observer dynamics in that case is exactly the same as the original system.

Using the Lagrange identity for the cross product of vectors together with (45.11), the expression in (45.7) and (45.8) can be simplified respectively as

$$\Theta_R(\hat{g}, y) = \sum_{j\in\mathscr{J}} \frac{1}{D(\hat{g}\bar{p}_j)} \frac{\|\Pi_0\hat{g}\bar{p}_j\|^2}{\|F\Pi_0 g\bar{p}_j\|}((\Pi_0\hat{g}\bar{p}_j \times \Pi_0 g\bar{p}_j)\times), \tag{45.12}$$

$$\Theta_T(\hat{g}, y) = \sum_{j\in\mathscr{J}} \frac{1}{D(\hat{g}\bar{p}_j)} \frac{-2}{\|F\Pi_0 g\bar{p}_j\|}((\Pi_0 g\bar{p}_j \times \Pi_0\hat{g}\bar{p}_j) \times \Pi_0\hat{g}\bar{p}_j). \tag{45.13}$$

Remark 45.3. Note that from (45.4), a lower bound for $D(\hat{g}\bar{p}_j)\|F\Pi_0 g\bar{p}_j\|$ is given by $m^2(m+1)m$, which implies that the observer is well defined.

45.4.1 The Error Dynamics

As in [3], we define the error $\eta(t) \overset{\text{def}}{=} \hat{g}(t)g^{-1}(t)$. Therefore, using (45.1), we may write

$$\dot{\eta} = \dot{\hat{g}}g^{-1} + \hat{g}\dot{g}^{-1} = \zeta\,\Theta(\hat{g}, y)\eta, \quad \eta(0) = \hat{g}_0 g_0^{-1}, \tag{45.14}$$

where, taking into account that $g = \eta^{-1}\hat{g}$, $\Theta(\hat{g}, y)$ can be rewritten as

$$\Theta(\hat{g}, y) = \Theta(\eta) = \begin{bmatrix} \Theta_R(\eta) & \Theta_T(\eta) \\ 0 & 0 \end{bmatrix},$$

with

$$\Theta_R(\eta) = \sum_{j\in\mathscr{J}} \frac{1}{D(\hat{g}\bar{p}_j)} \frac{\|\Pi_0\hat{g}\bar{p}_j\|^2}{\|F\Pi_0 g\bar{p}_j\|}((\Pi_0\hat{g}\bar{p}_j \times \Pi_0\eta^{-1}\hat{g}\bar{p}_j)\times), \tag{45.15}$$

$$\Theta_T(\eta) = \sum_{j\in\mathscr{J}} \frac{1}{D(\hat{g}\bar{p}_j)} \frac{-2}{\|F\Pi_0 g\bar{p}_j\|}((\Pi_0\eta^{-1}\hat{g}\bar{p}_j \times \Pi_0\hat{g}\bar{p}_j)\times \Pi_0\hat{g}\bar{p}_j). \tag{45.16}$$

Since a Lie group is a complex geometric object, it is a standard procedure to estimate results on a matrix Lie group G from results in the vector space which is its Lie algebra, here denoted by $\mathscr{L}$. We will adopt this procedure to analyze the error η and, later, prove convergence results. The Lie algebra $\mathscr{L}$ is the best linear approximation of G in the neighborhood of the identity I, and the exponential map exp, which sends elements in $\mathscr{L}$ to elements in G plays a crucial role in transferring data and results from one structure to the other. The exponential mapping is known to be bijective from a small neighborhood of $0 \in \mathscr{L}$ to a small neighborhood of the identity in G, and its inverse is denoted by log.

If η is sufficiently close to the identity, there is a representation $\eta = \exp(\epsilon\xi)$, where $\epsilon > 0$ and $\xi \in \mathrm{se}(3)$ satisfies $\|\xi\| = 1$. Since, $\exp(\epsilon\xi) = I + \epsilon\xi + O(\epsilon^2)$, where $O(\epsilon^2)$ represents the terms containing ϵ^k, for $k \geq 2$, for small ϵ, $I + \epsilon\xi$ is a good approximation for η. In the rest of the paper, and for the sake of simplicity, we may use the alternative notation e^A instead of $\exp(A)$. We henceforth make the following assumption.

Assumption 1 *We assume that the error η is close enough to I_4, that is, $\eta \in \mathscr{N}_\epsilon \overset{\text{def}}{=} \{v = \exp(\epsilon\xi) \colon \xi \in \mathrm{se}(3)$ and $\|\xi\| = 1\}$, where $0 \leq \epsilon < 1$.*

Remark 45.4. We may, without loss of generality assume that η is close to the identity. This is due to the fact that $x\mathscr{L} \sim \mathscr{L}$, for $x \in G$, is the best linear approximation of G in the neighborhood of x. So, if η is in the neighborhood of $x \in G$, then ηx^{-1} is close to the identity.

Using Lemma 1.7.3 of [20], which can be deduced from Lemma 3.4 in [21], we have

$$\frac{\mathrm{d}}{\mathrm{d}t}(\epsilon\xi) = \frac{u}{e^u - 1}\bigg|_{u\,=\,\mathrm{ad}_{\epsilon\xi}} (\dot{\eta}\eta^{-1}),$$

where $\dfrac{u}{e^u - 1} = \displaystyle\sum_{m=0}^{+\infty} \frac{(-1)^m}{m+1}(e^u - 1)^m$. Using (45.14), we have $\dot{\eta}\eta^{-1} = \zeta\Theta(\hat{g}, y)$ and hence

$$\frac{\mathrm{d}}{\mathrm{d}t}(\epsilon\xi) = \frac{u}{e^u - 1}\bigg|_{u\,=\,\mathrm{ad}_{\epsilon\xi}} (\zeta\Theta)$$

or, equivalently,

$$\frac{\mathrm{d}}{\mathrm{d}t}(\epsilon\xi) = \zeta\Theta - \frac{1}{2}\mathrm{ad}_{\epsilon\xi}\zeta\Theta - \frac{1}{2}\sum_{k=2}^{\infty}\frac{1}{k!}\mathrm{ad}_{\epsilon\xi}^{k}\zeta\Theta + \sum_{m=2}^{+\infty}\frac{(-1)^m}{m+1}(e^u - 1)^m\Big|_{u\,=\,\mathrm{ad}_{\epsilon\xi}}(\zeta\Theta). \tag{45.17}$$

On the other hand

$$\exp(\epsilon\xi) = I_4 + \epsilon\xi + O(\epsilon^2), \quad \exp(-\epsilon\xi) = I_4 - \epsilon\xi + O(\epsilon^2)$$

and, using the fact that $\Theta(g, y) = \Theta(I_4) = 0$ and noticing that Θ is defined in the linear space of 4×4 real matrices, containing both SE(3) and se(3), we have

$$\Theta_R(\eta) = \sum_{j\in\mathscr{J}} -\frac{1}{D(\hat{g}\bar{p}_j)}\frac{\|\Pi_0\hat{g}\bar{p}_j\|^2((\Pi_0\hat{g}\bar{p}_j \times \Pi_0\epsilon\xi\hat{g}\bar{p}_j)\times)}{\|F\Pi_0 g\bar{p}_j\|} + O(\epsilon^2), \tag{45.18}$$

$$\Theta_T(\eta) = \sum_{j\in\mathscr{J}} \frac{1}{D(\hat{g}\bar{p}_j)}\frac{2((\Pi_0\epsilon\xi\hat{g}\bar{p}_j \times \Pi_0\hat{g}\bar{p}_j) \times \Pi_0\hat{g}\bar{p}_j)}{\|F\Pi_0 g\bar{p}_j\|} + O(\epsilon^2). \tag{45.19}$$

From (45.15) to (45.17) with $\Theta(I_4) = 0$, we conclude that

$$\frac{\mathrm{d}}{\mathrm{d}t}(\epsilon\xi) = \zeta\bar{\Theta}(\epsilon\xi) = \zeta\begin{bmatrix}\bar{\Theta}_R(\epsilon\xi) & \bar{\Theta}_T(\epsilon\xi)\\ 0 & 0\end{bmatrix} + O(\epsilon^2),$$

where

$$\bar{\Theta}_R(\epsilon\xi) = \sum_{j\in\mathscr{J}} \frac{1}{D(\hat{g}\bar{p}_j)}\frac{\|\Pi_0\hat{g}\bar{p}_j\|^2}{\|F\Pi_0 g\bar{p}_j\|}((\Pi_0\epsilon\xi\hat{g}\bar{p}_j \times \Pi_0\hat{g}\bar{p}_j)\times), \tag{45.20}$$

$$\bar{\Theta}_T(\epsilon\xi) = \sum_{j\in\mathscr{J}} \frac{1}{D(\hat{g}\bar{p}_j)}\frac{-2}{\|F\Pi_0 g\bar{p}_j\|}((\Pi_0\hat{g}\bar{p}_j \times \Pi_0\epsilon\xi\hat{g}\bar{p}_j) \times \Pi_0\hat{g}\bar{p}_j). \tag{45.21}$$

Up to an approximation of the order ϵ^2, we obtain that $\epsilon\xi$ satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}(\epsilon\xi) = \zeta\begin{bmatrix}\bar{\Theta}_R(\epsilon\xi) & \bar{\Theta}_T(\epsilon\xi)\\ 0 & 0\end{bmatrix}.$$

We have the following result.

Proposition 45.1. *Up to an approximation of the order ϵ^2, the following result holds.*

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\epsilon\xi\|^2 = -4\zeta \sum_{j\in\mathscr{J}} \frac{1}{D(\hat{g}\bar{p}_j)} \frac{1}{\|F\mathit{\Pi}_0 g\bar{p}_j\|} \|\mathit{\Pi}_0\epsilon\xi\hat{g}\bar{p}_j \times \mathit{\Pi}_0\hat{g}\bar{p}_j\|^2. \tag{45.22}$$

Proof. From the fact that

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\epsilon\xi\|^2 = \left\langle \frac{\mathrm{d}}{\mathrm{d}t}(\epsilon\xi), (\epsilon\xi)\right\rangle + \left\langle (\epsilon\xi), \frac{\mathrm{d}}{\mathrm{d}t}(\epsilon\xi)\right\rangle = 2\left\langle \frac{\mathrm{d}}{\mathrm{d}t}(\epsilon\xi), (\epsilon\xi)\right\rangle,$$

and

$$\left\langle \frac{\mathrm{d}}{\mathrm{d}t}(\epsilon\xi), (\epsilon\xi)\right\rangle = \zeta\mathrm{Tr}\left(\left(\bar{\mathit{\Theta}}(\epsilon\xi)\right)^{\mathrm{T}}(\epsilon\xi)\right),$$

it follows that $\frac{\mathrm{d}}{\mathrm{d}t}\|\epsilon\xi\|^2 = 2\zeta\mathrm{Tr}\left(\left(\bar{\mathit{\Theta}}(\epsilon\xi)\right)^{\mathrm{T}}(\epsilon\xi)\right)$. Using the fact that

$$\mathrm{Tr}\left((u_1\times)^{\mathrm{T}}(u_2\times)\right) = 2u_1^{\mathrm{T}}u_2$$

for every $u_1, u_2 \in \mathbb{R}^3$, up to an approximation of order ϵ^3, we have

$$\begin{aligned}
&\mathrm{Tr}\left(\left(\bar{\mathit{\Theta}}(\epsilon\xi)\right)^{\mathrm{T}}(\epsilon\xi)\right) \\
&\quad = 2\sum_{j\in\mathscr{J}} \frac{1}{D(\hat{g}\bar{p}_j)} \frac{1}{\|F\mathit{\Pi}_0 g\bar{p}_j\|} \Big\{ \|\mathit{\Pi}_0\hat{g}\bar{p}_j\|^2 (\mathit{\Pi}_0\epsilon\xi\hat{g}\bar{p}_j \times \mathit{\Pi}_0\hat{g}\bar{p}_j)^{\mathrm{T}}\epsilon\bar{\xi} \\
&\qquad - ((\mathit{\Pi}_0\hat{g}\bar{p}_j \times \mathit{\Pi}_0\epsilon\xi\hat{g}\bar{p}_j) \times \mathit{\Pi}_0\hat{g}\bar{p}_j)^{\mathrm{T}}\epsilon\xi_T \Big\}.
\end{aligned} \tag{45.23}$$

From the relation $(a \times b)^{\mathrm{T}}c = \det[a\ b\ c]$, where $[a\ b\ c]$ stays for the matrix whose first, second, and third columns are respectively the vectors $a, b, c \in \mathbb{R}^3$ and using the skew-symmetry of the determinant function, we obtain

$$\begin{aligned}
&\|\mathit{\Pi}_0\hat{g}\bar{p}_j\|^2 (\mathit{\Pi}_0\epsilon\xi\hat{g}\bar{p}_j \times \mathit{\Pi}_0\hat{g}\bar{p}_j)^{\mathrm{T}}\epsilon\bar{\xi} \\
&\quad = -(((\mathit{\Pi}_0\epsilon\xi\hat{g}\bar{p}_j \times \mathit{\Pi}_0\hat{g}\bar{p}_j) \times \mathit{\Pi}_0 g\bar{p}_j) \times \mathit{\Pi}_0 g\bar{p}_j)^{\mathrm{T}}\epsilon\bar{\xi} \\
&\quad = -((\mathit{\Pi}_0\epsilon\xi\hat{g}\bar{p}_j \times \mathit{\Pi}_0\hat{g}\bar{p}_j) \times \mathit{\Pi}_0 g\bar{p}_j)^{\mathrm{T}}(\mathit{\Pi}_0 g\bar{p}_j \times \epsilon\bar{\xi}) \\
&\quad = -(\mathit{\Pi}_0\epsilon\xi\hat{g}\bar{p}_j \times \mathit{\Pi}_0\hat{g}\bar{p}_j)^{\mathrm{T}}(\mathit{\Pi}_0 g\bar{p}_j \times (\mathit{\Pi}_0 g\bar{p}_j \times \epsilon\bar{\xi})) \\
&\quad = -(\mathit{\Pi}_0\epsilon\xi\hat{g}\bar{p}_j \times \mathit{\Pi}_0\hat{g}\bar{p}_j)^{\mathrm{T}}((\epsilon\bar{\xi} \times \mathit{\Pi}_0 g\bar{p}_j) \times \mathit{\Pi}_0 g\bar{p}_j)
\end{aligned}$$

and

$$\begin{aligned}
&-((\mathit{\Pi}_0\hat{g}\bar{p}_j \times \mathit{\Pi}_0\epsilon\xi\hat{g}\bar{p}_j) \times \mathit{\Pi}_0\hat{g}\bar{p}_j)^{\mathrm{T}}\epsilon\xi_T \\
&\quad = -(\mathit{\Pi}_0\hat{g}\bar{p}_j \times \mathit{\Pi}_0\epsilon\xi\hat{g}\bar{p}_j)^{\mathrm{T}}(\mathit{\Pi}_0\hat{g}\bar{p}_j \times \epsilon\xi_T) \\
&\quad = -(\mathit{\Pi}_0\epsilon\xi\hat{g}\bar{p}_j \times \mathit{\Pi}_0\hat{g}\bar{p}_j)^{\mathrm{T}}(\epsilon\xi_T \times \mathit{\Pi}_0\hat{g}\bar{p}_j).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\|\Pi_0\hat{g}\bar{p}_j\|^2(\Pi_0\epsilon\xi\hat{g}\bar{p}_j \times \Pi_0\hat{g}\bar{p}_j)^{\mathrm{T}}\epsilon\bar{\xi} - ((\Pi_0\hat{g}\bar{p}_j \times \Pi_0\epsilon\xi\hat{g}\bar{p}_j) \times \Pi_0\hat{g}\bar{p}_j)^{\mathrm{T}}\epsilon\xi_T\\
&\quad = -(\Pi_0\epsilon\xi\hat{g}\bar{p}_j \times \Pi_0\hat{g}\bar{p}_j)^{\mathrm{T}}((\epsilon\bar{\xi} \times \Pi_0 g\bar{p}_j) \times \Pi_0 g\bar{p}_j)\\
&\qquad - (\Pi_0\epsilon\xi\hat{g}\bar{p}_j \times \Pi_0\hat{g}\bar{p}_j)^{\mathrm{T}}(\epsilon\xi_T \times \Pi_0\hat{g}\bar{p}_j).
\end{aligned}$$

Note that the right-hand-side is

$$\begin{aligned}
&-(\Pi_0\epsilon\xi\hat{g}\bar{p}_j \times \Pi_0\hat{g}\bar{p}_j)^{\mathrm{T}}\left((\epsilon\bar{\xi} \times \Pi_0 g\bar{p}_j + \epsilon\xi_T) \times \Pi_0\hat{g}\bar{p}_j\right)\\
&\quad = -\|\Pi_0\epsilon\xi\hat{g}\bar{p}_j \times \Pi_0\hat{g}\bar{p}_j\|^2
\end{aligned}$$

by noting that, $\epsilon\bar{\xi} \times \Pi_0\hat{g}\bar{p}_j + \epsilon\xi_T = \Pi_0\epsilon\xi\hat{g}\bar{p}_j$.

Hence (45.23) reduces to

$$\mathrm{Tr}\left(\left(\bar{\Theta}(\epsilon\xi)\right)^{\mathrm{T}}(\epsilon\xi)\right) = -2\sum_{j\in\mathscr{J}}\frac{1}{D(\hat{g}\bar{p}_j)}\frac{1}{\|F\Pi_0 g\bar{p}_j\|}\|\Pi_0\epsilon\xi\hat{g}\bar{p}_j \times \Pi_0\hat{g}\bar{p}_j\|^2,$$

and, consequently

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\epsilon\xi\|^2 = -4\zeta\sum_{j\in\mathscr{J}}\frac{1}{D(\hat{g}\bar{p}_j)}\frac{1}{\|F\Pi_0 g\bar{p}_j\|}\|\Pi_0\epsilon\xi\hat{g}\bar{p}_j \times \Pi_0\hat{g}\bar{p}_j\|^2.$$

□

45.4.2 Exponential Convergence

In this section we show under suitable assumptions that the estimation error converges exponentially to zero as $t \to \infty$. Let M denote the upper bound for $\|F\Pi_0 g\bar{p}_j\|$ for all j.

We will recall Gronwall's Lemma [22, Ch. III, 1.1.3] that is required to prove our next result.

Lemma 45.2 (Gronwall inequality). *Let g, h, y, $\frac{\mathrm{d}y}{\mathrm{d}t}$ be locally integrable functions satisfying*

$$\frac{\mathrm{d}y}{\mathrm{d}t} \leq gy + h \quad \textit{for } t \geq t_0. \tag{45.24}$$

Then, for all $t \geq t_0$,

$$y(t) \leq y(t_0)\exp\left(\int_{t_0}^{t} g(\tau)\,\mathrm{d}\tau\right) + \int_{t_0}^{t} h(s)\exp\left(-\int_{t}^{s} g(\tau)\,\mathrm{d}\tau\right)\mathrm{d}s.$$

Our next result is as follows.

Theorem 45.1. *Let* $\bar{T} \in [0, +\infty]$ *and* $\lambda > 0$ *be such that*

$$\sum_{j\in\mathscr{J}} \frac{\|\Pi_0 \epsilon\xi \hat{g}\bar{p}_j \times \Pi_0 \hat{g}\bar{p}_j\|^2}{(\#\mathscr{J})\|\Pi_0 \hat{g}\bar{p}_j\|^2(1+\|\Pi_0\hat{g}\bar{p}_j\|)} \geq \lambda\|\epsilon\xi\|^2$$

on the time interval $[0, \bar{T}[$. *Then, for every* $t \in [0, \bar{T}[$,

$$\|\epsilon\xi(t)\|^2 \leq \|\epsilon\xi(0)\|^2 e^{-4\zeta\lambda M^{-1}t},$$

where M *is an upper bound for* $\|F\Pi_0 g\bar{p}_j\|$ *for all* j. *In particular, if* $\bar{T} = +\infty$, *then* $\|\epsilon\xi(t)\|^2$ *converges exponentially fast to zero as* $t \to \infty$.

Proof. Under the hypothesis, (45.22) implies that

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\epsilon\xi\|^2 \leq -4\zeta\lambda M^{-1}\|\epsilon\xi\|^2, \tag{45.25}$$

for all $t \in [0, \bar{T}]$. The result follows from Gronwall's inequality (Lemma 45.2). □

Note that the rate of convergence can be improved by tuning $\zeta > 0$, that is, the rate of convergence increases with ζ. Next we prove the following result.

Theorem 45.2. *Let* $\bar{T} \in [0, +\infty]$. *Suppose there exists* $T > 0$ *such that, for every* $t \geq 0$, *with* $t + T \leq \bar{T}$,

$$\frac{1}{T}\int_t^{t+T} \sum_{j\in\mathscr{J}} \frac{\|\Pi_0 \epsilon\xi \hat{g}\bar{p}_j \times \Pi_0 \hat{g}\bar{p}_j\|^2}{(\#\mathscr{J})\|\Pi_0 \hat{g}\bar{p}_j\|^2(1+\|\Pi_0\hat{g}\bar{p}_j\|)} \frac{1}{\|\epsilon\xi\|^2}\mathrm{d}\tau \geq \lambda.$$

Then, for $n \in \mathbb{N}$ *with* $t \geq 0$, $t + nT \leq \bar{T}$,

$$\|\epsilon\xi(t+nT)\|^2 \leq \|\epsilon\xi(t)\|^2 e^{-4\zeta\lambda M^{-1}nT}.$$

In particular, if $\bar{T} = +\infty$, *then* $\|\epsilon\xi(t)\|^2$ *exponentially fast to zero as* $t \to \infty$.

Proof. Multiplying both the sides of (45.22) by $(T\|\epsilon\xi\|^2)^{-1}$, we have

$$\frac{1}{T}\frac{1}{\|\epsilon\xi\|^2}\frac{\mathrm{d}}{\mathrm{d}t}\|\epsilon\xi\|^2 = \frac{1}{T}\frac{1}{\|\epsilon\xi\|^2}\sum_{j\in\mathscr{J}}\frac{-4\zeta}{D(\hat{g}\bar{p}_j)}\frac{1}{\|F\Pi_0 g\bar{p}_j\|}\|\Pi_0\epsilon\xi\hat{g}\bar{p}_j \times \Pi_0\hat{g}\bar{p}_j\|^2,$$

or,

$$\frac{1}{T}\frac{\mathrm{d}}{\mathrm{d}t}\log(\|\epsilon\xi\|^2) \leq \frac{1}{T}\frac{1}{\|\epsilon\xi\|^2}\sum_{j\in\mathscr{J}}\frac{-4\zeta M^{-1}}{D(\hat{g}\bar{p}_j)}\|\Pi_0\epsilon\xi\hat{g}\bar{p}_j \times \Pi_0\hat{g}\bar{p}_j\|^2.$$

Since the statement is trivial for $n = 0$, we consider the case $n \geq 1$. Integrating on the interval $[t + (n-1)T, t + nT]$, we obtain

$$\int_{t+(n-1)T}^{t+nT} \frac{1}{T}\frac{\mathrm{d}}{\mathrm{d}t}\log(\|\epsilon\xi\|^2)\mathrm{d}\tau$$
$$\leq \int_{t+(n-1)T}^{t+nT} \frac{1}{T}\frac{1}{\|\epsilon\xi\|^2}\sum_{j\in\mathscr{J}}\frac{-4\zeta M^{-1}}{D(\hat{g}\bar{p}_j)}\|\Pi_0\epsilon\xi\hat{g}\bar{p}_j\times\Pi_0\hat{g}\bar{p}_j\|^2\mathrm{d}\tau.$$

Note that

$$\int_{t+(n-1)T}^{t+nT} \frac{\mathrm{d}}{\mathrm{d}t}\log(\|\epsilon\xi(\tau)\|^2)\mathrm{d}\tau$$
$$= \log(\|\epsilon\xi(t+nT)\|^2) - \log(\|\epsilon\xi(t+(n-1)T)\|^2),$$

and the properties of logarithm imply that

$$\int_{t+(n-1)T}^{t+nT} \frac{\mathrm{d}}{\mathrm{d}t}\log(\|\epsilon\xi(\tau)\|^2)\mathrm{d}\tau = \log\left(\frac{\|\epsilon\xi(t+nT)\|^2}{\|\epsilon\xi(t+(n-1)T)\|^2}\right).$$

Hence using the assumption, we have $\frac{1}{T}\log\left(\frac{\|\epsilon\xi(t+nT)\|^2}{\|\epsilon\xi(t+(n-1)T)\|^2}\right) \leq -4\zeta M^{-1}\lambda$ or, equivalently,

$$\frac{\|\epsilon\xi(t+nT)\|^2}{\|\epsilon\xi(t+(n-1)T)\|^2} \leq e^{-4\zeta M^{-1}\lambda T},$$

from which we derive $\frac{\|\epsilon\xi(t+nT)\|^2}{\|\epsilon\xi(t)\|^2} \leq e^{-4\zeta M^{-1}\lambda nT}$.

Finally, if $\bar{T} = +\infty$, we have $\|\epsilon\xi(t)\|^2 \leq \max_{s\in[0,T]}\|\epsilon\xi(s)\|^2\, e^{-4\zeta M^{-1}\lambda[t/T]}$, where $[t/T]$ denotes the largest natural number contained in the quotient t/T, that is, $[t/T] \leq t/T < [t/T]+1$. $[t/T] \geq 0$ is a non-negative integer number. □

Remark 45.5. Theorem 45.1 may be seen as the "limit" of Theorem 45.2 when T goes to 0.

45.5 Robustness Analysis of the Observer

In this section we investigate the effect of disturbance and noise on the estimation error. We now consider the process model (45.2)–(45.3) subjected to disturbances and noise as follows:

$$\dot{g}(t) = g(t)\left(\Omega(t) + w(t)\right), \quad g(0) = g_0, \tag{45.26}$$
$$y_j(t) = \tilde{y}_j(t) + v_j(t), \tag{45.27}$$

where $w \in \mathrm{se}(3)$ is the disturbance, $\tilde{y}_j = [\tilde{y}_{j1}\ \tilde{y}_{j2}\ \ 1]^{\mathrm{T}} \in \mathbb{R}^3$ is the real output defined implicitly by $\alpha_j \tilde{y}_j = FHg\bar{p}_j$ with $0 < \kappa \leq \alpha_j$, $y_j = [y_{j1}\ y_{j2}\ \ 1]^{\mathrm{T}} \in \mathbb{R}^3$ is the measured output with noise $v_j = [v_{j1}\ v_{j2}\ \ 0]^{\mathrm{T}} \in \mathbb{R}^3$. Further, the disturbance and noise signals are assumed to be deterministic but unknown. Note that (45.27) is equivalent to $y_j = \alpha_j^{-1}\left(F\Pi_0 g\bar{p}_j + \alpha_j v_j\right)$. Define $M_p \overset{\text{def}}{=} \sup\limits_{\substack{t\in[0,t_1]\\ j\in\mathscr{J}}} \|F\Pi_0 g\bar{p}_j + \alpha_j v_j\|$.

Let $|F^{-1}|$ denotes a bound for the functional norm of $F^{-1}(t)$ defined by

$$|F^{-1}(t)| \overset{\text{def}}{=} \sup\{F^{-1}(t)u \colon u \in \mathbb{R}^3 \text{and } \|u\| = 1\},$$

that is, we assume $F^{-1}(t)$ is bounded in the time interval [0, t_1] we are considering the estimator in. Define $M_v \overset{\text{def}}{=} \sup\limits_{\substack{t\in[0,t_1]\\ j\in\mathscr{J}}} \|v_j(t)\|$ and $M_w \overset{\text{def}}{=} \sup\limits_{t\in[0,t_1]} \|w(t)\|$, that is, M_v and M_w respectively denote the upper bounds for the noise $\|v_j\|$ and disturbance $\|w\|$, we suppose to exist, in the same time interval [0, t_1].

We consider the same observer claimed in (45.5), which can be rewritten as

$$\dot{\hat{g}}(t) = \hat{g}(t)\Omega(t) + \zeta\Theta\left(\hat{g}(t), y(t)\right)\hat{g}(t), \quad \hat{g}(0) = \hat{g}_0, \tag{45.28}$$

where $\Theta(\hat{g}, y)$ is given by

$$\Theta(\hat{g}, y) = \bar{\Theta}(\hat{g}, y) + \tilde{\Theta}(\hat{g}, y),$$

with

$$\bar{\Theta}(\hat{g}, y) = \begin{bmatrix} \bar{\Theta}_R(\hat{g}, y) & \bar{\Theta}_T(\hat{g}, y) \\ 0 & 0 \end{bmatrix} \text{ and } \tilde{\Theta}(\hat{g}, y) = \begin{bmatrix} \tilde{\Theta}_R(\hat{g}, y) & \tilde{\Theta}_T(\hat{g}, y) \\ 0 & 0 \end{bmatrix},$$

where

$$\bar{\Theta}_R(\hat{g}, y) = \sum_{j\in\mathscr{J}} \frac{1}{D(\hat{g}\bar{p}_j)} \frac{\|\Pi_0\hat{g}\bar{p}_j\|^2((\Pi_0\hat{g}\bar{p}_j \times \Pi_0 g\bar{p}_j)\times)}{\|F\Pi_0 g\bar{p}_j + \alpha_j v_j\|},$$

$$\bar{\Theta}_T(\hat{g}, y) = \sum_{j\in\mathscr{J}} \frac{-2}{D(\hat{g}\bar{p}_j)} \frac{((\Pi_0 g\bar{p}_j \times \Pi_0\hat{g}\bar{p}_j) \times \Pi_0\hat{g}\bar{p}_j)}{\|F\Pi_0 g\bar{p}_j + \alpha_j v_j\|},$$

$$\tilde{\Theta}_R(\hat{g}, y) = \sum_{j\in\mathscr{J}} \frac{1}{D(\hat{g}\bar{p}_j)} \frac{\|\Pi_0\hat{g}\bar{p}_j\|^2((\Pi_0\hat{g}\bar{p}_j \times F^{-1}\alpha_j v_j)\times)}{\|F\Pi_0 g\bar{p}_j + \alpha_j v_j\|},$$

$$\tilde{\Theta}_T(\hat{g}, y) = \sum_{j\in\mathscr{J}} \frac{-2}{D(\hat{g}\bar{p}_j)} \frac{((F^{-1}\alpha_j v_j \times \Pi_0\hat{g}\bar{p}_j) \times \Pi_0\hat{g}\bar{p}_j)}{\|F\Pi_0 g\bar{p}_j + \alpha_j v_j\|}.$$

Note that both $\bar{\Theta}$ and $\tilde{\Theta}$ depend on noise v_j. Again, we define the error $\eta(t) \overset{\text{def}}{=} \hat{g}(t)g^{-1}(t)$. Therefore, using (45.1) yields

$$\dot{\eta} = \dot{\hat{g}}g^{-1} + \hat{g}\dot{g}^{-1} = \zeta\Theta(\eta)\eta - \hat{g}w\hat{g}^{-1}\eta, \quad \eta(0) = \hat{g}_0 g_0^{-1}, \tag{45.29}$$

where, by using $g = \eta^{-1}\hat{g}$ we can rewrite $\Theta(\eta)$ as

$$\Theta(\eta) = \bar{\Theta}(\eta) + \tilde{\Theta} = \begin{bmatrix} \bar{\Theta}_R(\eta) & \bar{\Theta}_T(\eta) \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} \tilde{\Theta}_R & \tilde{\Theta}_T \\ 0 & 0 \end{bmatrix},$$

with

$$\bar{\Theta}_R(\eta) = \sum_{j\in\mathscr{J}} \frac{1}{D(\hat{g}\bar{p}_j)} \frac{\|\Pi_0\hat{g}\bar{p}_j\|^2((\Pi_0\hat{g}\bar{p}_j \times \Pi_0\eta^{-1}\hat{g}\bar{p}_j)\times)}{\|F\Pi_0 g\bar{p}_j + \alpha_j v_j\|}, \tag{45.30}$$

$$\bar{\Theta}_T(\eta) = \sum_{j\in\mathscr{J}} \frac{-2}{D(\hat{g}\bar{p}_j)} \frac{((\Pi_0\eta^{-1}\hat{g}\bar{p}_j \times \Pi_0\hat{g}\bar{p}_j) \times \Pi_0\hat{g}\bar{p}_j)}{\|F\Pi_0 g\bar{p}_j + \alpha_j v_j\|}, \tag{45.31}$$

$$\tilde{\Theta}_R = \sum_{j\in\mathscr{J}} \frac{1}{D(\hat{g}\bar{p}_j)} \frac{\|\Pi_0\hat{g}\bar{p}_j\|^2((\Pi_0\hat{g}\bar{p}_j \times F^{-1}\alpha_j v_j)\times)}{\|F\Pi_0 g\bar{p}_j + \alpha_j v_j\|}, \tag{45.32}$$

$$\tilde{\Theta}_T = \sum_{j\in\mathscr{J}} \frac{-2}{D(\hat{g}\bar{p}_j)} \frac{((F^{-1}\alpha_j v_j \times \Pi_0\hat{g}\bar{p}_j) \times \Pi_0\hat{g}\bar{p}_j)}{\|F\Pi_0 g\bar{p}_j + \alpha_j v_j\|}. \tag{45.33}$$

Remark 45.6. Note that $\|F\Pi_0 g\bar{p}_j + \alpha_j v_j\|$ is equal to $\alpha_j\|\alpha_j^{-1}F\Pi_0 g\bar{p}_j + v_j\| = \alpha_j\|y_j\| \geq \alpha_j \geq \kappa$ and from (45.4), it follows that $m^2(m+1)\kappa$ is a lower bound for $D(\hat{g}\bar{p}_j)\|F\Pi_0 g\bar{p}_j + \alpha_j v_j\|$. From this we conclude that the observer is well defined.

Note that $\eta = \exp(\epsilon\xi) = I_4 + \epsilon\xi + O(\epsilon^2)$. Using Lemma 1.7.3 of [20], we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}(\epsilon\xi) = \dot{\eta}\eta^{-1} - \frac{1}{2}[\epsilon\xi, \dot{\eta}\eta^{-1}] + O(\epsilon^2).$$

From (45.29), we have $\dot{\eta}\eta^{-1} = \zeta(\bar{\Theta}(\eta) + \tilde{\Theta}) - \hat{g}w\hat{g}^{-1}$ and hence the above equation becomes

$$\frac{\mathrm{d}}{\mathrm{d}t}(\epsilon\xi) = \zeta\bar{\Theta}(\epsilon\xi) + \zeta\tilde{\Theta} - \hat{g}w\hat{g}^{-1} - \frac{1}{2}[\epsilon\xi, \zeta\tilde{\Theta} - \hat{g}w\hat{g}^{-1}] + O(\epsilon^2).$$

Up to an approximation of the order ϵ^2, we have that $\epsilon\xi$ satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}(\epsilon\xi) = \zeta\bar{\Theta}(\epsilon\xi) + \zeta\tilde{\Theta} - \hat{g}w\hat{g}^{-1} - \frac{1}{2}[\epsilon\xi, \zeta\tilde{\Theta} - \hat{g}w\hat{g}^{-1}], \tag{45.34}$$

and, multiplying by $\epsilon\xi$ yields,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|\epsilon\xi\|^2 &= 2\langle\zeta\bar{\Theta}(\epsilon\xi),\, \epsilon\xi\rangle + 2\langle\zeta\tilde{\Theta},\, \epsilon\xi\rangle - 2\langle\hat{g}w\hat{g}^{-1},\, \epsilon\xi\rangle \\ &\quad - \langle[\epsilon\xi,\, \zeta\tilde{\Theta} - \hat{g}w\hat{g}^{-1}],\, \epsilon\xi\rangle. \end{aligned} \tag{45.35}$$

To estimate a bound for the variation $\frac{\mathrm{d}}{\mathrm{dt}}\|\epsilon\xi\|^2$, we may start by estimate a bound for the individual terms on the right-hand-side of (45.35), which are given by the following result.

Proposition 45.2. *The following statements hold.*

(i) $\langle\bar{\Theta}(\epsilon\xi), \epsilon\xi\rangle = -\sum_{j\in\mathscr{J}} \frac{2}{D(\hat{g}\bar{p}_j)} \frac{\|\Pi_0\epsilon\xi\hat{g}\bar{p}_j \times \Pi_0\hat{g}\bar{p}_j\|^2}{\|F\Pi_0 g\bar{p}_j + \alpha_j v_j\|}.$

(ii) $\left\langle\tilde{\Theta}, \epsilon\xi\right\rangle \leq \sum_{j\in\mathscr{J}} \frac{2|F^{-1}|M_v}{(\#\mathscr{J})}\|\epsilon\xi\|.$

(iii) $\langle\hat{g}w\hat{g}^{-1},\, \epsilon\xi\rangle \leq \|w\|\|\epsilon\xi\|.$

(iv) $\langle[\epsilon\xi,\, \zeta\tilde{\Theta} - \hat{g}w\hat{g}^{-1}],\, \epsilon\xi\rangle \leq \left(2\zeta\sum_{j\in\mathscr{J}} \frac{|F^{-1}|\, M_v}{(\#\mathscr{J})(1 + \|\Pi_0\hat{g}\bar{p}_j\|)} + M_w\right)\|\epsilon\xi\|^2.$

Proof. In the following $\begin{bmatrix}\xi_R & \xi_T \\ 0 & 0\end{bmatrix} \overset{\text{def}}{=} \xi$ with $(\bar{\xi}\times) \overset{\text{def}}{=} \xi_R$.

(i) The result follows from the noise free case, proceeding as in the proof of Proposition 45.1.

(ii) First, note that

$$(\tilde{\Theta})^{\mathrm{T}}\epsilon\xi = \begin{bmatrix}\tilde{\Theta}_R & \tilde{\Theta}_T \\ 0 & 0\end{bmatrix}^{\mathrm{T}} \begin{bmatrix}\epsilon\xi_R & \epsilon\xi_T \\ 0 & 0\end{bmatrix} = \begin{bmatrix}\left(\tilde{\Theta}_R\right)^{\mathrm{T}}\epsilon\xi_R & \left(\tilde{\Theta}_R\right)^{\mathrm{T}}\epsilon\xi_T \\ \left(\tilde{\Theta}_T\right)^{\mathrm{T}}\epsilon\xi_R & \left(\tilde{\Theta}_T\right)^{\mathrm{T}}\epsilon\xi_T\end{bmatrix}.$$

Then $\left\langle\tilde{\Theta}, \epsilon\xi\right\rangle = \mathrm{Tr}\left(\tilde{\Theta}^{\mathrm{T}}\epsilon\xi\right) = \mathrm{Tr}\left(\left(\tilde{\Theta}_R\right)^{\mathrm{T}}\epsilon\xi_R\right) + \left(\tilde{\Theta}_T\right)^{\mathrm{T}}\epsilon\xi_T$. Now, we have

$$\mathrm{Tr}\left(\left(\tilde{\Theta}_R\right)^{\mathrm{T}}\epsilon\xi_R\right) = \sum_{j\in\mathscr{J}} \frac{2}{D(\hat{g}\bar{p}_j)} \frac{\|\Pi_0\hat{g}\bar{p}_j\|^2(\Pi_0\hat{g}\bar{p}_j \times F^{-1}\alpha_j v_j)^{\mathrm{T}}\epsilon\bar{\xi}}{\|F\Pi_0 g\bar{p}_j + \alpha_j v_j\|}, \quad \text{and}$$

$$(\tilde{\Theta}_T)^{\mathrm{T}}\epsilon\xi_T = \sum_{j\in\mathscr{J}} \frac{-2\alpha_j}{D(\hat{g}\bar{p}_j)} \frac{((F^{-1}v_j \times \Pi_0\hat{g}\bar{p}_j) \times \Pi_0\hat{g}\bar{p}_j)^{\mathrm{T}}\epsilon\xi_T}{\|F\Pi_0 g\bar{p}_j + \alpha_j v_j\|}.$$

Proceeding as in the proof of Proposition 45.1, we can arrive to

$$\left\langle \tilde{\Theta}, \epsilon\xi \right\rangle = \sum_{j \in \mathscr{J}} \frac{2\alpha_j}{D(\hat{g}\bar{p}_j)} \frac{(F^{-1}v_j \times \Pi_0 \hat{g}\bar{p}_j)^{\mathrm{T}}(\Pi_0 \epsilon\xi\hat{g}\bar{p}_j \times \Pi_0 \hat{g}\bar{p}_j)}{\|F\Pi_0 g\bar{p}_j + \alpha_j v_j\|},$$

$$\left\langle \tilde{\Theta}, \epsilon\xi \right\rangle \leq \sum_{j \in \mathscr{J}} \frac{2\alpha_j}{D(\hat{g}\bar{p}_j)} \frac{\|F^{-1}v_j \times \Pi_0 \hat{g}\bar{p}_j\| \|\Pi_0 \epsilon\xi\hat{g}\bar{p}_j \times \Pi_0 \hat{g}\bar{p}_j\|}{\|F\Pi_0 g\bar{p}_j + \alpha_j v_j\|}.$$

Since $\|u_1 \times u_2\| \leq \|u_1\|\|u_2\|$ for every $u_1, u_2 \in \mathbb{R}^3$, we have

$$\left\langle \tilde{\Theta}, \epsilon\xi \right\rangle \leq \sum_{j \in \mathscr{J}} \frac{2}{(\#\mathscr{J})(1 + \|\Pi_0 \hat{g}\bar{p}_j\|)} \left(\frac{\alpha_j}{\|F\Pi_0 g\bar{p}_j + \alpha_j v_j\|} \right) \|F^{-1}v_j\| \|\Pi_0 \epsilon\xi\hat{g}\bar{p}_j\|.$$

Further, note that $\|y_j\|^{-1} = \dfrac{\alpha_j}{\|F\Pi_0 g\bar{p}_j + \alpha_j v_j\|} \leq 1$ and $\|F^{-1}v_j\| \leq |F^{-1}|M_v$. Hence

$$\left\langle \tilde{\Theta}, \epsilon\xi \right\rangle \leq \sum_{j \in \mathscr{J}} \frac{2|F^{-1}|M_v}{(\#\mathscr{J})(1 + \|\Pi_0 \hat{g}\bar{p}_j\|)} \|\Pi_0 \epsilon\xi\hat{g}\bar{p}_j\|.$$

Recall that $\Pi_0\epsilon\xi\hat{g}\bar{p}_j = \epsilon\bar{\xi} \times \Pi_0\hat{g}\bar{p}_j + \epsilon\xi_T$ and by triangle inequality it follows that $\|\Pi_0\epsilon\xi\hat{g}\bar{p}_j\| \leq \|\epsilon\bar{\xi} \times \Pi_0\hat{g}\bar{p}_j\| + \|\epsilon\xi_T\|$. In other words, $\|\Pi_0\epsilon\xi\hat{g}\bar{p}_j\| \leq \|\epsilon\xi\|(1 + \|\Pi_0\hat{g}\bar{p}_j\|)$ by noting that

$$\|\epsilon\bar{\xi} \times \Pi_0\hat{g}\bar{p}_j\| \leq \|\epsilon\bar{\xi}\|\|\Pi_0\hat{g}\bar{p}_j\|, \|\epsilon\bar{\xi}\| \leq \|\epsilon\xi\|,$$

and

$$\|\epsilon\xi_T\| \leq \|\epsilon\xi\|.$$

Hence

$$\left\langle \tilde{\Theta}, \epsilon\xi \right\rangle \leq \sum_{j \in \mathscr{J}} \frac{2|F^{-1}|M_v}{(\#\mathscr{J})} \|\epsilon\xi\|.$$

(iii) First note that, $\langle \hat{g}w\hat{g}^{-1}, \epsilon\xi \rangle \leq \|\hat{g}w\hat{g}^{-1}\|\|\epsilon\xi\|$. By the definition, we have

$$\|\hat{g}w\hat{g}^{-1}\| = \sqrt{\mathrm{Tr}(\hat{g}w^{\mathrm{T}}\hat{g}^{-1}\hat{g}w\hat{g}^{-1})}.$$

Since $\hat{g}^{-1}\hat{g} = I_4$, we have $\|\hat{g}w\hat{g}^{-1}\| = \sqrt{\mathrm{Tr}(\hat{g}w^{\mathrm{T}}w\hat{g}^{-1})}$. Recall that, the trace of a matrix is invariant under similarity transformation, that is, $\mathrm{Tr}(BAB^{-1}) = \mathrm{Tr}(A)$ for every $A \in \mathbb{R}^{n \times n}$ [24, Ch. V, 7]. Thus, we conclude that $\|\hat{g}^{-1}w\hat{g}\| = \sqrt{\mathrm{Tr}(w^{\mathrm{T}}w)} = \|w\|$. Hence $\langle \hat{g}w\hat{g}^{-1}, \epsilon\xi \rangle \leq \|w\|\|\epsilon\xi\|$.

(iv) For simplicity, we define $Z \overset{\text{def}}{=} \begin{bmatrix} Z_R & Z_T \\ 0 & 0 \end{bmatrix} \overset{\text{def}}{=} \zeta\tilde{\Theta} - \hat{g}w\hat{g}^{-1}$.

We find $\langle [\epsilon\xi, Z], \epsilon\xi \rangle = \mathrm{Tr}(\epsilon\xi^{\mathrm{T}}(Z\epsilon\xi - \epsilon\xi Z))$. Note that, for every $A, B, C \in \mathrm{se}(3)$, we have

$$A^{\mathrm{T}}BC = \begin{bmatrix} A_R^{\mathrm{T}} & 0 \\ A_T^{\mathrm{T}} & 0 \end{bmatrix} \begin{bmatrix} B_R C_R & B_R C_T \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} A_R^{\mathrm{T}} B_R C_R & A_R^{\mathrm{T}} B_R C_T \\ A_T^{\mathrm{T}} B_R C_R & A_T^{\mathrm{T}} B_R C_T \end{bmatrix}$$

and $\mathrm{Tr}(A^{\mathrm{T}}BC) = -\mathrm{Tr}(A_R B_R C_R) + A_T^{\mathrm{T}} B_R C_T$. Hence

$$\begin{aligned}\langle [\epsilon\xi,\, Z],\, \epsilon\xi\rangle = &\left(-\mathrm{Tr}(\epsilon\xi_R Z_R \epsilon\xi_R) + (\epsilon\xi_T)^{\mathrm{T}} Z_R \epsilon\xi_T\right)\\ &- \left(-\mathrm{Tr}(\epsilon\xi_R \epsilon\xi_R Z_R) + (\epsilon\xi_T)^{\mathrm{T}} \epsilon\xi_R Z_T\right).\end{aligned}$$

It is easy to check that $(\epsilon\xi_R Z_R \epsilon\xi_R)^{\mathrm{T}} = -(\epsilon\xi_R Z_R \epsilon\xi_R)$, and hence $\epsilon\xi_R Z_R \epsilon\xi_R$ is skew-symmetric, which implies that its trace is zero. On the other hand, the term $(\epsilon\xi_T)^{\mathrm{T}} Z_R \epsilon\xi_T = (\epsilon\xi_T)^{\mathrm{T}}(\bar{z} \times \epsilon\xi_T)$ vanishes as well, where $(\bar{z}\times) \overset{\mathrm{def}}{=} Z_R$. The term $\mathrm{Tr}(\epsilon\xi_R \epsilon\xi_R Z_R)$ vanishes because for positive semi-definite matrices A and B, we have $0 \le \mathrm{Tr}(AB) \le \mathrm{Tr}(A)\mathrm{Tr}(B)$ [23, pg. 329], so that

$$0 \le \mathrm{Tr}(-\epsilon\xi_R \epsilon\xi_R Z_R) \le \mathrm{Tr}(-\epsilon\xi_R \epsilon\xi_R)\mathrm{Tr}(Z_R) = 0.$$

Notice that, for a given vector

$$u \in \mathbb{R}^3, u^{\mathrm{T}}(-\epsilon\xi_R)\epsilon\xi_R u = (\epsilon\xi_R u)^{\mathrm{T}}(\epsilon\xi_R u) = \|\epsilon\xi_R u\|^2 \ge 0 \textit{ and } u^{\mathrm{T}} Z_R u = 0.$$

Therefore

$$\langle [\epsilon\xi,\, Z],\, \epsilon\xi\rangle = -(\epsilon\xi_T)^{\mathrm{T}} \epsilon\xi_R Z_T = Z_T^{\mathrm{T}}(\epsilon\bar{\xi} \times \epsilon\xi_T).$$

Note the following facts:

(a) $\|y_j\|^{-1} = \dfrac{\alpha_j}{\|F\Pi_0 g\bar{p}_j + \alpha_j v_j\|} \le 1.$

(b) Using (45.33) together with a), we have

$$\zeta\tilde{\Theta}_T^{\mathrm{T}}(\epsilon\bar{\xi} \times \epsilon\xi_T) \le 2\zeta \sum_{j\in\mathscr{J}} \left| \frac{((F^{-1}v_j \times \Pi_0\hat{g}\bar{p}_j) \times \Pi_0\hat{g}\bar{p}_j)^{\mathrm{T}}(\epsilon\bar{\xi} \times \epsilon\xi_T)}{D(\hat{g}\bar{p}_j)} \right|.$$

From $|((F^{-1}v_j \times \Pi_0\hat{g}\bar{p}_j) \times \Pi_0\hat{g}\bar{p}_j)^{\mathrm{T}}(\epsilon\bar{\xi} \times \epsilon\xi_T)| \le |F^{-1}|\, M_v\, \|\Pi_0\hat{g}\bar{p}_j\|^2\, \|\epsilon\bar{\xi} \times \epsilon\xi_T\|$ we obtain

$$\zeta\tilde{\Theta}_T^{\mathrm{T}}(\epsilon\bar{\xi} \times \epsilon\xi_T) \le 2\zeta \sum_{j\in\mathscr{J}} \frac{|F^{-1}|\, M_v\, \|\epsilon\xi\|^2}{(\#\mathscr{J})\,(1 + \|\Pi_0\hat{g}\bar{p}_j\|)}.$$

(c) It can be easily shown that $(\hat{g}w\hat{g}^{-1})_T^{\mathrm{T}}(\epsilon\bar{\xi} \times \epsilon\xi_T) \le \|w\|\|\epsilon\xi\|^2$ or, equivalently, $(\hat{g}w\hat{g}^{-1})_T^{\mathrm{T}}(\epsilon\bar{\xi} \times \epsilon\xi_T) \le M_w\|\epsilon\xi\|^2$.

Using (b) and (c) above, we conclude that

$$\langle [\epsilon\xi,\, \zeta\tilde{\Theta} - \hat{g}w\hat{g}^{-1}],\, \epsilon\xi\rangle \leq \left(2\zeta \sum_{j\in\mathscr{J}} \frac{|F^{-1}|\, M_v}{(\#\mathscr{J})(1 + \|\Pi_0 \hat{g}\bar{p}_j\|)} + M_w\right) \|\epsilon\xi\|^2.$$

□

The following result follows immediately from Proposition 45.2.

Proposition 45.3. *The following statement holds.*

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\epsilon\xi\|^2 \leq -\sum_{j\in\mathscr{J}} \frac{4\zeta}{D(\hat{g}\bar{p}_j)} \frac{\|\Pi_0\epsilon\xi\hat{g}\bar{p}_j \times \Pi_0\hat{g}\bar{p}_j\|^2}{\|F\Pi_0 g\bar{p}_j + \alpha_j v_j\|} + M_w(2 + \|\epsilon\xi\|)\|\epsilon\xi\|$$
$$+ 2\zeta \sum_{j\in\mathscr{J}} \frac{|F^{-1}|\, M_v\|}{(\#\mathscr{J})} \left(2 + \frac{\|\epsilon\xi\|}{1 + \|\Pi_0\hat{g}\bar{p}_j\|}\right) \|\epsilon\xi\|.$$

Proof. The result follows by using the bounds given by Proposition 45.2 in the expression (45.35). □

The following result follows immediately, by recalling that $\epsilon < 1$ from Assumption 1.

Proposition 45.4. *We have the estimate*

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\epsilon\xi\|^2 \leq -\sum_{j\in\mathscr{J}} \frac{4\zeta}{D(\hat{g}\bar{p}_j)} \frac{\|\Pi_0\epsilon\xi\hat{g}\bar{p}_j \times \Pi_0\hat{g}\bar{p}_j\|^2}{\|F\Pi_0 g\bar{p}_j + \alpha_j v_j\|} + 6\zeta|F^{-1}|M_v + 3M_w. \tag{45.36}$$

By (45.4), $\|F\Pi_0 g\bar{p}_j\|$ is bounded above by M and, by the definition $\alpha_j \leq \|F\Pi_0 g\bar{p}_j\|$. Then each $\|F\Pi_0 g\bar{p}_j + \alpha_j v_j\|$ is bounded above by $M_p = M(1+M_v)$.

Next, we derive the noisy version of Theorem 45.1.

Theorem 45.3. *Let* $\bar{T} \in [0, +\infty]$ *and* $\lambda > 0$ *be such that*

(i) $\|\epsilon\xi(0)\| < 1$.

(ii) $\displaystyle\sum_{j\in\mathscr{J}} \frac{\|\Pi_0\epsilon\xi\hat{g}\bar{p}_j \times \Pi_0\hat{g}\bar{p}_j\|^2}{(\#\mathscr{J})\|\Pi_0\hat{g}\bar{p}_j\|^2(1 + \|\Pi_0\hat{g}\bar{p}_j\|)} \geq \lambda\|\epsilon\xi\|^2.$

(iii) $\displaystyle\frac{6\zeta|F^{-1}|M_v + 3M_w}{4\zeta\lambda M_p^{-1}} < 1.$

are satisfied on the time interval $[0, \bar{T})$. *Then, for every* $t \in [0, \bar{T}[$,

$$\|\epsilon\xi(t)\|^2 \leq \|\epsilon\xi(0)\|^2 e^{-4\zeta\lambda M_p^{-1} t} + \frac{6\zeta|F^{-1}|M_v + 3M_w}{4\zeta\lambda M_p^{-1}} \left(1 - e^{-4\zeta\lambda M_p^{-1} t}\right).$$

In particular, if $\bar{T} = +\infty$, then for every constant $\rho > 0$ there exists $t_\rho \geq 0$ such that

$$\|\epsilon\xi(t)\|^2 < \frac{6\zeta|F^{-1}|M_v + 3M_w}{4\zeta\lambda M_p^{-1}} + \rho$$

for all $t \geq t_\rho$.

Proof. Define $M_{v,w} \overset{\text{def}}{=} 6\zeta|F^{-1}|M_v + 3M_w$ and $\tilde{\lambda} \overset{\text{def}}{=} 4\zeta\lambda M_p^{-1}$. Since $\|\epsilon\xi(0)\| < 1$ there exists small enough $t_0 \in (0, \bar{T}]$ such that $\|\epsilon\xi(t)\| < 1$ for all $t \in [0, t_0)$. From Proposition 45.4, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\epsilon\xi\|^2 \leq -\tilde{\lambda}\|\epsilon\xi\|^2 + M_{v,w},$$

for all $t \in [0, t_0]$. From Gronwall's inequality (45.24) we have

$$\|\epsilon\xi(t)\|^2 \leq \|\epsilon\xi(0)\|^2 e^{-\tilde{\lambda}t} + \int_0^t M_{v,w} \exp\left(-\int_t^s -\tilde{\lambda}\,\mathrm{d}\tau\right)\mathrm{d}s. \tag{45.37}$$

Note that $\exp\left(-\int_t^s -\tilde{\lambda}\,\mathrm{d}\tau\right) = e^{\tilde{\lambda}(s-t)}$. Hence

$$\begin{aligned}\|\epsilon\xi(t)\|^2 &\leq \|\epsilon\xi(0)\|^2 e^{-\tilde{\lambda}t} + \int_0^t M_{v,w}e^{\tilde{\lambda}(s-t)}\,\mathrm{d}s\\ &\leq \|\epsilon\xi(0)\|^2 e^{-\tilde{\lambda}t} + M_{v,w}\tilde{\lambda}^{-1}(1 - e^{-\tilde{\lambda}t})\end{aligned} \tag{45.38}$$

for all $t \in [0, t_0)$. Therefore we see that $\|\epsilon\xi\|$ is non-increasing in $[0, t_0)$ if we have that

$$\|\epsilon\xi(0)\|^2 e^{-\tilde{\lambda}t} + M_{v,w}\tilde{\lambda}^{-1}(1 - e^{-\tilde{\lambda}t}) \leq \|\epsilon\xi(0)\|^2,$$

that is, if $\|\epsilon\xi(0)\|^2 \geq M_{v,w}\tilde{\lambda}^{-1}$. Since at time t_0, we have $\|\epsilon\xi(t_0)\|^2 < 1$ we may repeat the argument for a suitable interval $[t_0, t_0 + t_1]$ for some $t_1 > 0$ with $t_0 + t_1 \leq \bar{T}$. Therefore, if $\|\epsilon\xi(t_0)\|^2 \geq M_{v,w}\tilde{\lambda}^{-1}$, then $\|\epsilon\xi(t)\|^2$ decreases in $[t_0, t_0 + t_1]$. Repeating again successively the argument, we see that the sequence of instants of time $s_m \overset{\text{def}}{=} \sum_{i=0}^{m} t_i$ must "reach" the instant $\bar{T}$, otherwise by the definition we must have $\|\epsilon\xi(s)\| = 1$ at $s = \lim_{m\to+\infty} s_m \leq \bar{T}$ that is impossible because $\|\epsilon\xi(s_m)\|^2 \leq \|\epsilon\xi(0)\|^2 < 1$ for all $m \in \mathbb{N}$. So, in particular $\|\epsilon\xi\|^2 \leq \|\epsilon\xi(0)\|^2 < 1$ in $[0, \bar{T}[$. Coming back to the beginning of this proof we may then suppose that $t_0 = \bar{T}$ and so, estimate (45.37) holds for all $t \in [0, \bar{T}[$. In the case $\bar{T} = +\infty$, from (45.37), we conclude that for any given constant $\rho > 0$, we may find $t_\rho \geq 0$ such that $\|\epsilon\xi(t)\|^2 < M_{v,w}\tilde{\lambda}^{-1} + \rho$ for all $t \geq t_\rho$. □

We also have the noisy version of Theorem 45.2.

Theorem 45.4. *Let* $\bar{T} \in]0, +\infty]$. *Suppose there exist positive constants* T, λ *such that,*

(i) $\|\epsilon\xi(0)\|^2 \leq \dfrac{(6\zeta\,|F^{-1}|\,M_v + 3M_w)T}{1 - e^{-4\zeta\lambda M_p^{-1}T}}$

(ii) $\dfrac{1}{T}\displaystyle\int_t^T \sum_{j\in\mathscr{J}} \dfrac{\|\Pi_0\epsilon\xi\hat{g}\bar{p}_j \times \Pi_0\hat{g}\bar{p}_j\|^2}{(\#\mathscr{J})\|\Pi_0\hat{g}\bar{p}_j\|^2(1 + \|\Pi_0\hat{g}\bar{p}_j\|)}\,\mathrm{d}\tau \geq \lambda$ *for every* $0 \leq t, t + T < \bar{T}$

(iii) $\dfrac{(6\zeta\,|F^{-1}|\,M_v + 3M_w)T(2 - e^{-4\zeta\lambda M_p^{-1}T})}{1 - e^{-4\zeta\lambda M_p^{-1}T}} < 1$

are satisfied. Then for all $s \in [0, \bar{T})$,

$$\|\epsilon\xi(s)\|^2 \leq \frac{(6\zeta\,|F^{-1}|\,M_v + 3M_w)T(2 - e^{-4\zeta\lambda M_p^{-1}T})}{1 - e^{-4\zeta\lambda M_p^{-1}T}}.$$

Proof. Define $M_{v,w} \overset{\text{def}}{=} 6\zeta|F^{-1}|M_v + 3M_w$ and $\tilde{\lambda} \overset{\text{def}}{=} 4\zeta\lambda M_p^{-1}$. Suppose that $\|\epsilon\xi(t)\|^2 - M_p(t - s) > 0$ for all $t \in [s, s + T]$. From estimate (45.36), we may derive

$$\frac{1}{T}\frac{1}{\|\epsilon\xi(t)\|^2 - M_{v,w}(t - s)}\frac{\mathrm{d}}{\mathrm{d}t}\left(\|\epsilon\xi(t)\|^2 - M_{v,w}(t - s)\right)$$
$$\leq -\frac{1}{T}\sum_{j\in\mathscr{J}}\frac{4}{D(\hat{g}\bar{p}_j)}\frac{\|\Pi_0\epsilon\xi\hat{g}\bar{p}_j \times \Pi_0\hat{g}\bar{p}_j\|^2}{\|F\Pi_0 g\bar{p}_j + v_j\|(\|\epsilon\xi(t)\|^2 - M_p(t - s))}.$$

Integrating on $[s, s + T]$, we arrive to

$$\log\left(\frac{\|\epsilon\xi(s + T)\|^2 - M_{v,w}T)}{\|\epsilon\xi(s)\|^2}\right) \leq -\tilde{\lambda}T,$$

that is, $\|\epsilon\xi(s + T)\|^2 \leq e^{-\tilde{\lambda}T}\|\epsilon\xi(s)\|^2 + M_{v,w}T$, and so $\|\epsilon\xi(s + T)\|^2 \leq \|\epsilon\xi(s)\|^2$ if

$$\|\epsilon\xi(s)\|^2 \geq \frac{M_{v,w}T}{e^{-\tilde{\lambda}T}}. \tag{45.39}$$

On the other hand notice that from estimate (45.36), we have $\dfrac{\mathrm{d}}{\mathrm{d}t}\|\epsilon\xi\|^2 \leq M_{v,w}$. Since $\|\epsilon\xi(0)\|^2 \leq \dfrac{M_{v,w}T}{1 - e^{-\tilde{\lambda}T}}$, we have that

$$\|\epsilon\xi(t)\|^2 \leq \frac{M_{v,w}T}{1 - e^{-\tilde{\lambda}T}} + M_{v,w}t = \frac{M_{v,w}T + (1 - e^{-\tilde{\lambda}T})M_{v,w}t}{1 - e^{-\tilde{\lambda}T}} \tag{45.40}$$

for all $t \in [0,\ T]$. Now we notice that to have, for some time $\tau \geq T$,

$$\|\epsilon\xi(\tau)\|^2 = \frac{M_{v,w}T(2 - e^{-\tilde{\lambda}T})}{1 - e^{-\tilde{\lambda}T}} \tag{45.41}$$

at some time $\tau > 0$ we need to have $\|\epsilon\xi(t)\|^2 \geq \frac{M_{v,w}T}{1 - e^{-\tilde{\lambda}T}}$ in the interval $[\tau - T,\ \tau]$, because if for some $t \in [\tau - T,\ \tau]$ we have $\|\epsilon\xi(t)\|^2 < \frac{M_{v,w}T}{1 - e^{-\tilde{\lambda}T}}$ then necessarily

$$\begin{aligned}\|\epsilon\xi(\tau)\|^2 <& \frac{M_{v,w}T}{1 - e^{-\tilde{\lambda}T}} + M_{v,w}(\tau - t)\\ \leq& \frac{M_{v,w}T}{1 - e^{-\tilde{\lambda}T}} = \frac{M_{v,w}T(2 - e^{-\tilde{\lambda}T})}{1 - e^{-\tilde{\lambda}T}}\end{aligned}$$

which contradicts (45.41). On the other side, if $\|\epsilon\xi(t)\|^2 \geq \frac{M_{v,w}T}{1 - e^{-\tilde{\lambda}T}}$ in the interval $[\tau - T,\ \tau]$, then since $\frac{M_{v,w}T}{1 - e^{-\tilde{\lambda}T}} \geq M_{v,w}T$ and, using estimate (45.39), we have that $\|\epsilon\xi(\tau)\|^2 \geq \|\epsilon\xi(\tau - T)\|^2$ so, for every $\tau > T$, $\|\epsilon\xi(\tau)\|^2 = \frac{M_{v,w}T(2 - e^{-\tilde{\lambda}T})}{1 - e^{-\tilde{\lambda}T}}$ only if

$$\|\epsilon\xi(\tau - T)\|^2 = \frac{M_{v,w}T(2 - e^{-\tilde{\lambda}T})}{1 - e^{-\tilde{\lambda}T}}. \tag{45.42}$$

Now from estimate (45.40) we have

$$\begin{aligned}\|\epsilon\xi(t)\|^2 \leq& \frac{M_{v,w}T + (1 - e^{-\tilde{\lambda}T})M_{v,w}t}{1 - e^{-\tilde{\lambda}T}}\\ <& \frac{M_{v,w}T(2 - e^{-\tilde{\lambda}T})}{1 - e^{-\tilde{\lambda}T}}\end{aligned}$$

for all $t \in [0,\ T[$. Therefore, from (45.42), we have $\|\epsilon\xi(t)\|^2 < \frac{M_{v,w}T(2 - e^{-\tilde{\lambda}T})}{1 - e^{-\tilde{\lambda}T}}$ for all $t > 0$. □

Remark 45.7. Theorem 45.3 may be "almost" seen as the limit of Theorem 45.4 as T goes to 0. We say almost because we need to impose that the square of the norm of the initial error is smaller than $\frac{M_{v,w}T}{1 - e^{-\tilde{\lambda}T}} < \frac{M_{v,w}T(2 - e^{-\tilde{\lambda}T})}{1 - e^{-\tilde{\lambda}T}}$, that is,

$\|\epsilon\xi(0)\|^2 \leq \dfrac{M_{v,w}T}{1-e^{-\tilde{\lambda}T}} < 1$, while in Theorem 45.3 it was enough to impose that $\|\epsilon\xi(0)\|^2 < 1$. Notice that $\dfrac{M_{v,w}T(2-e^{-\tilde{\lambda}T})}{1-e^{-\tilde{\lambda}T}}$ goes to $\dfrac{M_{v,w}}{\tilde{\lambda}}$ as T goes to 0.

45.6 Conclusions

This paper provides a nonlinear observer design structure for a left-invariant dynamical system evolving on the three-dimensional special Euclidean group with measurements given by implicit functions. Under suitable assumptions, we show that the linearized state estimation error converges exponentially fast to the true state. Furthermore, we show that if the dynamical system is subject to disturbance and noise, the estimator converges to an open neighborhood of the true value of the state. The size of the neighborhood increases/decreases gracefully with the bound of the disturbance and noise.

Acknowledgements This work was supported in part by projects DENO/FCT-PT (PTDC/EEA-ACR/ 67020/2006), Co3-AUVs (EU FP7 under grant agreement No. 231378), FCT (ISR/ IST plurianual funding) through the PIDDAC Program funds and the CMU-Portugal program.

References

1. Kalman, R.: A new approach in linear filtering and prediction problems. Transactions of the American Society of Mechanical Engineers. Journal of Basic Engineering. 82D. (1960) pp. 35–45
2. Luenberger, D.: Observing the state of a linear system with observers of low dynamic order. IEEE Trans. Mil. Electron. 74–80 (1964)
3. Bonnabel, S., Martin, P., Rouchon, P.: Nonlinear symmetry-preserving observers on Lie groups. IEEE Trans. Autom. Control. **5**, 1709–1713 (2009)
4. Bonnabel, S., Martin, P., Rouchon, P.: Symmetry-preserving observers. IEEE Trans. Autom. Control. **53**, 2514–2526 (2008)
5. Bonnabel, S., Martin, P., Rouchon, P.: Non-linear observer on Lie groups for left-invariant dynamics with right-left equivalent output. 17th World Congress The International Federation of Automatic Control, pp. 8594–8598. (2008)
6. Bonnabel, S., Martin, P., Rouchon, P.: A non-linear symmetry-preserving observer for velocity-aided inertial navigation. Proceedings of the 2006 American Control Conference, pp. 2910–2914 (2006)
7. Aguiar, A.P., Hespanha, J.P.: Minimum-energy state estimation for systems with perspective outputs. IEEE Trans. Autom. Control. **51**(2), 226–241 (2006)
8. Aguiar, A.P., Hespanha, J.P.: Robust filtering for deterministic systems with implicit outputs. Syst. Control Lett. **58**(4), 263–270 (2009)
9. Aguiar, A.P., Hespanha, J.P.: Trajectory-tracking and path-following of underactuated autonomous vehicles with parametric modeling uncertainty. IEEE Trans. Autom. Control. **52**(8), 1362–1379 (2007)
10. Ghosh, B.K., Jankovic, M., Wu, Y.T.: Perspective problems in system theory and its application in machine vision. J. Math. Syst. Estimation Control **4**(1), 3–38 (1994)

11. Ghosh, B.K., Loucks, E.P.: A perspective theory for motion and shape estimation in machine vision. SIAM J. Control Optim. **33**(5), 1530–1559 (1995)
12. Takahashi, S., Ghosh, B.K.: Motion and shape parameters identification with vision and range. 2001 American Control Conference, vol. 6, 4626–4631 (2001)
13. Lageman, C., Trumpf, J., Mahony, R.: Gradient-like observers for invariant dynamics on a Lie group. IEEE Trans. Autom. Control (2009)
14. Lageman, C., Trumpf, J., Mahony, R.: State observers for invariant dynamics on a Lie group. 18th International Symposium on Mathematical Theory of Networks and Systems. (2008)
15. Lageman, C., Trumpf, J., Mahony, R.: Observers for systems with invariant outputs. European Control Conference 2009, pp. 4587–4592. Budapest, Hungary (2009)
16. Lageman, C., Trumpf, J., Mahony, R.: Observer design for invariant systems with homogeneous observations. IEEE Trans. Autom. Control (2008)
17. Skjetne, R., Fossen, T.I., Kokotović, P.: Robust output maneuvering for a class of nonlinear systems. Automatica **40**(3), 373–383 (2004)
18. Bullo, F.: Stabilization of relative equilibria for underactuated systems on Riemannian manifolds. Automatica **36**, 1819–1834 (2000)
19. Al-Hiddabi, S., McClamroch, N.: Tracking and maneuver regulation control for nonlinear nonminimum phase systems: Application to flight control. IEEE Trans. Control Syst. Technol. **10**(6), 780–792 (2002)
20. Machado, L.: Least squares problems on Riemannian manifolds. Department of Mathematics, University of Coimbra (2006)
21. Sattinger, D.H., Weaver, O.L.: Lie groups and algebras with applications to physics, geometry, and mechanics. Springer, New York (1980)
22. Temam, R.: Infinite-dimensional dynamical systems in mechanics and physics. Applied Mathematical Sciences, vol. 68, 2nd edn. Springer, New York (1997)
23. Abadir, K.M., Magnus, J.R.: Matrix Algebra. Cambridge University Press, Cambridge (2005)
24. Pease, M.C.: Methods of Matrix Algebra, vol. 16. Elsevier, New York (1965)
25. Bonnabel, S., Rouchon, P.: Control and observer design for nonlinear finite and infinite dimensional systems. LNCIS, vol. 322, pp. 53–67. Springer, Berlin (2005)
26. Ma, Y., Soatto, S., Kosecka, J., Sastry, S.S.: An Invitation to 3-D Vision. Springer, Berlin (2005)
27. Anderson, B.D.O., Moore, J.B.: Optimal Filtering. Prentice-Hall in New Jersey, USA (1979)

Chapter 46
Stability of Polydeuces Orbit

Dulce C. Pinto and Filipe C. Mena

Abstract The stability of the recently discovered Saturn satellite Polydeuces has not been fully studied yet. We use data from the Cassini probe group (NASA and Queen Mary, London) in order to numerically study the stability of the orbit of Polydeuces. We treat the system Saturn-Dione-Polydeuces as a planar, circular, restricted three body problem where Polydeuces is librating around the L_5 Lagrangian point in a tadpole motion. We analyze the eccentricity evolution of Polydeuces trajectory, the Poincaré section and the indicator of its maximum Lyapounov characteristic exponent. Our results suggest that the Polydeuces orbit is stable for at least 10^5 Dione-years.

46.1 Polydeuces and the 3-Body Problem

Polydeuces is a Saturnian moon that was discovered by the Cassini spacecraft [3] on October 2004 (Fig. 46.1a). This moon is about 377,400 km distant from Saturn, has an equatorial diameter of 13 km and is close to Dione, a well known saturn moon. The orbit of Dione has an eccentricity $e \simeq 0.0192$, so its motion is approximately circular. Polydeuces has negligle mass comparatively to Dione and because the inclination $i \simeq 0.1774$ of its orbit is not significant, one treat the system Saturn-Dione-Polydeuces as a planar, circular, restricted three body problem (PCRTBP). More details about Polydeuces can be found in [3, 5].

The equations of motion for the PCRTBP in the synodic frame are well known and can be written as

$$\dot{\mathbf{x}} = \mathbf{F}\left(x, y, \dot{x}, \dot{y}\right), \tag{46.1}$$

D.C. Pinto (✉)
Departamento de Matemática, Universidade do Minho, Campus de Gualtar, 4710 Braga, Portugal
e-mail: dcpinto@iol.pt

F.C. Mena
Centro de Matemática, Universidade do Minho, Campus de Gualtar, 4710-057 Braga, Portugal
e-mail: fmena@math.uminho.pt

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_46,

where

$$\mathbf{F}(\mathbf{x}) = \left(x, y, 2y + \frac{\partial U}{\partial x}, 2x + \frac{\partial U}{\partial y}\right).$$

and U is given by

$$U = U(x, y) = \frac{1}{2}(x^2 + y^2) + \frac{1-\mu}{d_1} + \frac{\mu}{d_2},$$

where $d_1 = \sqrt{(x+\mu)^2 + y^2}$, $d_2 = \sqrt{(x-1+\mu)^2 + y^2}$, (x, y) is the position and $(\dot{x}, \dot{y})$ the velocity of the small body, $\mu = Gm_2$ and $m_1 >> m_2$ are the masses of the two most massive bodies. This system has an integration constant, the so called Jacobi constant, given by

$$\frac{1}{2}C_J = \frac{1}{2}(\dot{x}^2 + \dot{y}^2) - U.$$

It is also well known that the PCRTBP has five equilibrium points in the synodic frame. The Euler Points (L_1, L_2 and L_3) are collinear with two massive bodies and the Lagrangian points (L_4 and L_5) form with those two bodies a double equilateral triangle. The non-linear stability study of the equilibrium points is a problem with a known solution (see e.g. [1,2] or [4] for a review).

46.2 Numerical Stability: Lyapounov Exponents, Poincaré Sections and Eccentricity

In the case of Saturn-Dione-Polydeuces system $\mu = 1.85 \times 10^{-6}$, Polydeuces is librating around the L_5 equilibrium point in a tadpole motion and has a period of about 792 days [5]. General known analytical results allow to conclude the stability of the L_5 point. However, as Polydeuces is not exactly at this point, so we have used numerical analysis to follow its orbit and our results are shown in Fig. 46.1b. This graphic representation was obtained for an integration period of $1,035$ Dione-years.

At each instant of time, we consider an elliptical orbit and calculate its eccentricity e, and our results in Fig. 46.2 presents a pattern characteristic of a regular orbit with T referring to the orbital period of Dione and a to its semi-major axis.

In order to obtain Poincaré sections for Polydeuces orbit we take the fix Jacobi constant $C_J(x, y, \dot{x}, \dot{y}) = c_0$ and express $\dot{y}$ as a function of x, y and $\dot{x}$:

$$\dot{y} = \sqrt{(x^2 + y^2) + 2\left(\frac{1-\mu}{d_1} + \frac{\mu}{d_2}\right) - \dot{x}^2 - c_0}$$

where the dot denotes differentiation with respect to time.

From Cassini's data for the Polydeuces orbit, we have considered the energy level $c_0 = 2.99957$, the position $(x, y) = (0.7831, -0.6519)$ and, the corresponding

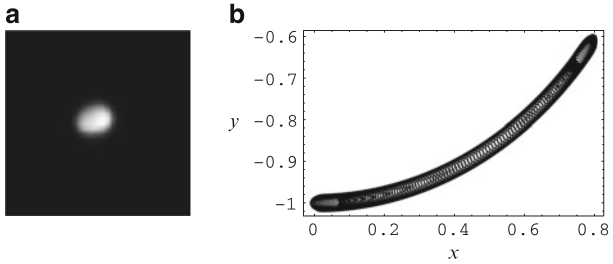


Fig. 46.1 (**a**) Polydeuces image, taken from the site of NASA: http://saturn.jpl.nasa.gov/science/moons/moonDetails.cfm?pageID=19. (**b**) Polydeuces trajectory for an integration period of 1,035 years of Dione

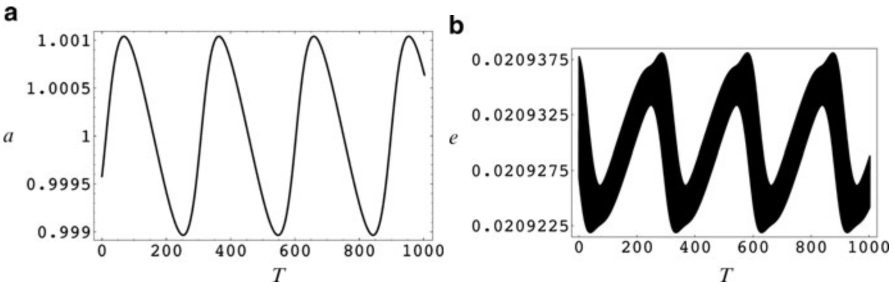


Fig. 46.2 Semi-major axis (**a**) and Eccentricity (**b**) in function of time for the Polydeuces trajectory. The T variable refers to orbital periods of Dione

velocity $(\dot{x}, \dot{y}) = (-0.0181, -0.0341)$. The equations of motion were then integrated for a period of 1.6×10^5 Dione-years. Our results for the Poincaré section of Polydeuces orbit, represented in Fig. 46.3a, suggest quasi-periodicity.

The maximum Lyapounov exponent χ was then computed. In order to do that, we have considered two orbits in the phase space separated by an initial distance $\xi(t_0)$, we found, at each time t, the new distance between both orbits $\xi(t)$. We have then used numerical methods to obtain an indicator of the maximum Lyapounov exponent, represented by χ^*.

In Fig. 46.3b, the points of the ensemble $\{(log\, t,\ log\ \chi^*(t)),\ 0 < t < 2.6{\times}10^6\}$ are plotted, using the equation

$$\chi^*(t) = \frac{1}{t} \log\left(\frac{\xi(t)}{\xi(t_0)}\right),$$

where $\xi(t)$ is, in this case, the norm of the solution of the ODE system

$$\dot{\boldsymbol{\xi}} = \frac{\partial \dot{\mathbf{F}}}{\partial \mathbf{x}}(\mathbf{x}(t))\boldsymbol{\xi} \ \wedge \ \boldsymbol{\xi}(0) = \boldsymbol{\xi}_0$$

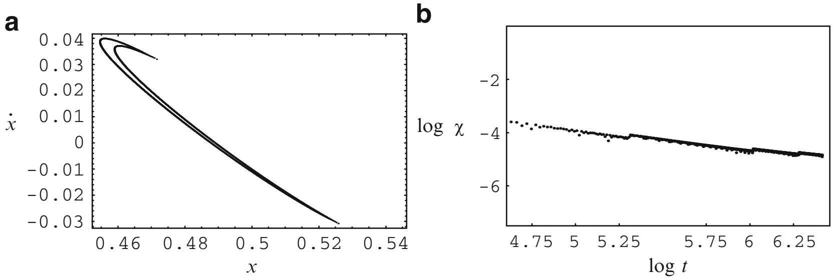


Fig. 46.3 Poincaré section (**a**) and maximum Lyapounov characteristic exponent (**b**) of Polydeuces orbit

and $\boldsymbol{F}$ is the vectorial function of the (46.1). The time value 2.6×10^6 corresponds to approximately 4×10^5 Dione-years. For regular orbits, the initial and final distances should be close and the graphic representation, using a $log - log$ scale, should be approaching a linear function with negative slope. Therefore, our graphic representation above exhibits patterns typical of a regular orbit.

46.3 Conclusion

The results of the eccentricity evolution of Polydeuces trajectory, the Poincaré section and the indicator of its maximum Lyapounov characteristic exponent are compatible and suggest that the Polydeuces orbit is stable for at least 10^5 Dione-years.

Acknowledgements We thank Mike Evans (QM, London and Cassini mission) for providing Polydeuces data. FCM thanks CMAT, Universidade do Minho, for support.

References

1. Arnold, V.: The stability of the equilibrium position of a Hamiltonian system of ordinary differential equations in the general elliptic case. Soviet. Math. Dokl. **2**, 247–249 (1961)
2. Cabral, H., Meyer, K.: Stability of equilibria and fixed points of conservative systems. Nonlinearity **12**, 1351–1352 (1999)
3. Murray, C.D., Cooper, N.J., Evans, M.W., Beurle, K.: S/2004 S 5: A new co-orbital companion for Dione. Icarus **179**, 222–234 (2005)
4. Pinto, D.: Estabilidade do problema de três corpos restrito Master Thesis, Universidade do Minho (2006)
5. Porco, C., et al.: Cassini imaging science: initial results on Saturn's rings and small satellites. Science **307** 1226–1236 (2005)

Chapter 47
Peixoto Classification of 2-Dim Flows Revisited

Antonio R. da Silva

Abstract According to A. Connes attached to a foliation groupoid there is a natural C^*-algebra. Here we present a very brief survey on how this construction together with a glueing method developed by Oshemkov and Sharko allows us to present an alternative formulation of Peixoto classification of Morse–Smale flows on 2-dimensional manifolds. Though all tools involved here are well-known their links seem to be still in progress.

47.1 Connes Foliation Algebra

It is well-known that given a locally compact group G with a left Haar measure dx, the space of the integrable functions $L^1(G, dx)$ can be made into a *-Banach algebra by taking as multiplication

$$(f * g)(x) = \int f(y)\, g(y^{-1}x)\, dy$$

and as involution

$$f^*(x) = \Delta(x)^{-1}\, \overline{f(x^{-1})},$$

where Δ is the modular function of G.

A standard procedure allows us to define the *group C^*-algebra $C^*(G)$* by taking the completion of $L^1(G)$ with respect to the norm

$$||f||_{C^*} = \sup\{||\pi(f)|| : \pi \text{ is a }^*\text{-representation of } G\},$$

where $\pi(f) = \int_G f(x)\pi(x)\, dx$ and $||\cdot||$ is the operator norm for operators acting on the representation space.

A.R. da Silva
Instituto de Matemática, Universidade Federal do Rio de Janeiro, CP 68530, Rio de Janeiro, Brazil
e-mail: ardasilva@ufrj.br

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_47,

By the end of the Seventies Connes showed how to build up a C^*-algebra out of a regular foliation, see e.g. [2–4]. Connes construction is based on the concept of *graph of a foliation*, introduced by Thom in [15], and treated in more detail by Winkelnkemper in [17]. In order to present Connes'construction we first recall some definitions.

By a *groupoid* G *with basis* B we understand a set G endowed with mappings $r: G \to B$, $s: G \to B$ and a partially defined binary operation $(x, y) \mapsto xy$ such that:

(a) $x\,y$ is defined whenever $r(y) = s(x)$
(b) Associativity holds
(c) For each $x \in G$ there is a left neutral element r_x and a right neutral element s_x such that

$$r_x \cdot x = x = x \cdot s_x$$

Thus instead of the unity of the group we have the *unit space*

$$G^o := \{x\,x^{-1} : x \in G\}.$$

Groups are groupoids with $G^o = \{e\}$. The mappings $r, s: G \to G^o$, with $r(x) = x\,x^{-1}$ and $s(x) = x^{-1}\,x$ are naturally associated to G and allow us to identify B with G^o.

A groupoid G endowed with a topology such that the multiplication and inverse mappings, whenever defined, are continuous is called a *topological groupoid.*

Let us assume that the topology on G is second countable, G is locally compact and Hausdorff and that $r: G \to G^o$ is an open mapping.

A *transverse function* on a locally compact *groupoid* is a family of measures $\{\lambda^x : x \in G^o\}$ such that:

(a) The support of λ^x is contained in $G^x := r^{-1}(x)$
(b) For each $f \in C_c(G)$ the mapping

$$\lambda(f): x \to \int f\, d\lambda^x$$

is continuous
(c) For each $\gamma \in G^y_x := r^{-1}(y) \cap s^{-1}(x)$ and each $f \in C_c(G)$ holds

$$\int f(\gamma\,\gamma')d\lambda^x(\gamma') = \int f(\gamma')d\lambda^y(\gamma').$$

The support of the mapping $\lambda: C_c(G) \to C_c(G^o)$ is a subset of G of the form G_F, where F is a closed subset of G^o. The closed subset is the *support of the transverse function*. The transverse function is called a *Haar system* when its support is G^o.

A Haar system for a locally compact groupoid need not exist and if it does, it is usually not unique.

Given a regular foliation $\mathscr{F}$ on a smooth manifold M under the *graph of* $\mathscr{F}$ or the *holonomy groupoid of* $\mathscr{F}$ we understand the groupoid with M as the sets of objects and the set G of morphisms defined by the properties:

(a) there is no morphism $x \to y$ unless x and y lie on the same leaf L of $\mathscr{F}$;
(b) the portion $r^{-1}(L) = s^{-1}(L)$ of G over a leaf L is a quotient of the fundamental groupoid of L, and
(c) two homotopy classes $[\gamma_1]$, $[\gamma_2]$ of paths from x to y in a leaf L are identified in G if and only if they have the same holonomy.

One can give G a natural structure of a manifold (not necessarily Hausdorff). Further, if the leaves of $\mathscr{F}$ are simply connected, or more generally, without holonomy then G reduces to the groupoid of an equivalence relation R on M, R being the relation of "lying on the same leaf". Now let us recall the notion of groupoid equivalence.

Let G be a groupoid. Consider a fiber bundle with base space G^o, total space Z and projection π. Given a pair of points x, y in G^o, each $\gamma \in G_x^y = r^{-1}(y) \cap s^{-1}(x)$ induces a bijection:

$$L_\gamma : \pi^{-1}(x) \to \pi^{-1}(y)$$

which satisfies: $L_{\gamma\gamma'} = L_\gamma \circ L_{\gamma'}$.

A pair (γ, z) is said *compatible* if $\pi(z) = s(\gamma)$ and the *product* of γ by z is defined by:

$$\gamma \cdot z = L_\gamma(z).$$

The set of compatible pairs is denoted by $G * Z$.

If π is open and continuous, and the product map $G * Z \subset G \times Z \to Z$ is continuous then Z is called a *left G-space*.

Similarly one defines *right G-spaces*.

Given two groupoids G and G' we define a *(G, G')-space* as a topological space Z that satisfies:

(a) Z is a left G-space
(b) Z is a right G'-space
(c) the actions of G and G' an Z commute

We say that two groupoids G and G' are *equivalent* whenever there is a (G, G')-space such that:

(a) if $\pi(z') = \pi(z)$ then there is a unique $\gamma \in G$ such that $z' = \gamma z$
(a) if $\pi'(z') = \pi'(z)$ then there is a unique $\gamma' \in G'$ such that $z' = z\gamma'$

Finally, the construction of a C^*-algebra out of a groupoid runs as follows:

Let (G, λ) be a locally compact groupoid with a Haar system. We can define *a* $*$-algebra structure on $C_c(G)$ by taking:

$$(f * g)(\gamma) = \int f(\gamma')g(\gamma'^{-1}\gamma)\, d\lambda^{s(\gamma)}(\gamma');$$
$$f^*(\gamma) = \overline{f(\gamma^{-1})}.$$

The enveloping C^*-algebra of this *-algebra is the *C^*-algebra of the groupoid G*. So given a regular foliation $\mathscr{F}$ on a manifold M on considers its holonomy groupoid $G(\mathscr{F})$ the corresponding C^*-algebra is called the *C^*-foliation algebra of $\mathscr{F}$* denoted here by $C^*(\mathscr{F})$.

47.2 Vector Fields

Let X be a vector field on a manifold M, i.e. $X\colon M \to TM$, where TM is the tangent bundle over M. The integral curves of the system:

$$\begin{cases} \dot{x}(t) = X(x(t)), & t \in \mathbb{R} \\ x(0) = p \end{cases}$$

define a *flow*

$$\begin{aligned} \mathscr{F}_p\colon I \subseteq \mathbb{R} &\to M \\ t &\mapsto \mathscr{F}_p(t) = \mathscr{F}(t,p) \end{aligned}$$

Given $p \in M$ the *orbit* through p is defined by $\mathscr{O}(p) = \{\mathscr{F}_t(p) \in M : t \in I\}$.

If $\mathscr{F}$ is the flow associated to $\dot{x} = X(x)$ then flow $(M, \mathscr{F})$ is identified with the vector field X on M. Further, we recall that

stable manifold $W^s(v) := \{p \in M : \mathscr{F}_t(P) \to v, t \to +\infty\}$;

unstable manifold $W^u(v) := \{p \in M : \mathscr{F}_t(p) \to v, t \to -\infty\}$.

Given two vector fields X and Y on M with corresponding flows $\mathscr{F}$ and $\mathscr{G}$ resp. we say X and Y are *topologically equivalent* whenever there is an homeomorfism $h\colon M \to M$ that maps orbits of X into orbits of Y preserving their orientations. That is, for $p \in M$ and $\delta > 0$ there exists $\varepsilon > 0$ such that:

$$0 < t < \delta \Rightarrow h(\mathscr{F}_t(p)) = \mathscr{G}_{\bar{t}}(h(p)) \text{ for some } 0 < \bar{t} < \varepsilon.$$

Now let M be a n-dimensional connected compact manifold and let $\mathfrak{X}^r(M)$ be the set of all C^r-vector fields on M, $r \geq 1$. Considering

$$X\colon M \to TM = \bigcup_{p \in M} T_pM \equiv \{(p,v) : p \in M, v \in T_pM\}$$

we have that X is a section of the fiber bundle (TM, π, M).

We consider a topology on $\mathfrak{X}^r(M)$, where the open sets are induced by the norm

$$||X||_r = \max_{i=\overline{o,r}} \sup_{u \in B(1)} \{||f^i(u)||, ||df^i(u)||, \ldots, ||d^r f^i(u)||\}.$$

Here we take a finite open cover $V_1, \ldots, V_k$ such that each V_i is contained in the domain of a local chart (x_i, V_i) with $x_i(V_i) = B(1) \equiv$ open unit ball, $f^i \equiv X \circ x_i^{-1} \colon x_i(V_i) \to \mathbb{R}^n$ is of classe C^r.

Definition 47.1. We say that $X \in \mathfrak{X}^r(M)$ is *structurally stable* whenever there is a neighborhood $\mathscr{V}$ of X in $\mathfrak{X}^r(M)$ such that every vector field $Y \in \mathscr{V}$ is topologically equivalent to X. In other words, the topological behaviour of the orbits of X does not change under small perturbations.

Suppose now that M is a 2-dimensional closed manifold (that is, compact without boundary).

Definition 47.2. $X \in \mathfrak{X}^r(M)$ is called *Morse–Smale* if:

(a) The vector field X has a finite number of hyperbolic singular points and a finite number of hyperbolic periodic trajectories.
(b) There is no saddle connection, that is, no separatrix joining one saddle to another or to itself.
(c) Any orbit has a unique α-limit as well as a unique ω-limit.
We have that the α- and ω-limit sets of every trajectory is either a singularity or a closed orbit.

Theorem 47.1 (Peixoto). *Let M be a closed orientable 2-manifold. Then*

(i) A vector field $X \in \mathfrak{X}^r$ is structurally stable iff it is Morse–Smale
(ii) The set $\sum$ of all Morse–Smale vector fields is open and dense in the space $\mathfrak{X}^r$ (approximation theorem!)

As a matter of fact Peixoto proved that the set of Morse–Smale vector fields on a compact two-dimensional manifold is open in $\mathfrak{X}^r(M)$, $r \geq 1$, and it is also dense if M is orientable. It follows from Pugh's closing lemma that this set is also dense in $\mathfrak{X}^1(M)$ for any two-dimensional compact manifold M. A topological equivalence class may have infinitely many connected components.

In [5] Gutiérrez and de Melo gave a classification of the set of connected components of Morse–Smale vector fields on M.

In what concerns the closing lemma we recall the deep contribution of Gutiérrez [6–8], who in particular showed that is not possible to prove the C^2 closing lemma through a local perturbation.

The Morse–Smale vector fields without periodic orbits are called *Morse vector fields*. It is well-known that these vector fields are equivalent to *gradient vector fields* which are vector fields $X \colon M \to TM$ such that $X = \nabla f$, for some $f \in C^{r+1}(M)$.

Peixoto in [14] gave a classification of the 2-dim Morse flows through the so-called *distinguished graphs*, see also [1, 11–13].

The distinguished graphs are defined through the polar flow and the distinguished sets that correspond to the *canonical regions*, that are the open connected components that one gets once the singularities, stable and unstable manifolds are taken out, for details see [14].

47.3 Oshemkov and Sharko Results and the Alternative Version of Peixoto Classification

In [13] Oshemkov and Sharko presented a complete classification of Morse–Smale flows on two-dimensional manifolds. As pointed out in that paper, several previous classifications were either incomplete, by leaving out flows with limit cycles, or hold only in case of oriented manifolds. Among them is a result due to Wang [16] (see also [12, 13]) that holds only for Morse flows on orientable two-manifolds.

We claim that adapting the results of Oshemkov and Sharko to Wang's approach we get the following version of Peixoto classification:

Theorem 47.2. *Let be given two Morse–Smale flows X and X' on closed two-dimensional manifolds M and M' respectively. Then X and X' are topologically equivalent if and only if the C^*-algebras attached to X and X' respectively are isomorphic.*

The C^*-algebras referred to in the above theorem are the ones obtained through Connes' construction and Oshemkov & Sharko glueing method. We now give a brief sketch of the proof. Let us first consider the case where X and X' are Morse flows. In this case one has the following result, see [13].

Lemma A. *Two Morse flows X and X' on closed two-dimensional manifolds M and M' are topologically equivalent if and only if there are, uniquely defined, three-colour graphs attached to them $T(X)$ and $T(X')$ respectively, that are isomorphic.*

For this case (Morse flows) one needs to consider just the so-called *atom technique*. Now as in Wang's paper one can attach to each of these coloured graphs a single C^*-algebra, the main point is to establish a correspondence between each vertex of graph and the C^*-algebra obtained from the foliation of the canonical region described by the flows. Note that the restriction to the orientable case, present in Wang's paper, does not show up here, since the glueing method developed by Oshemkov and Sharko holds in the general case. Further, foliation groupoids corresponding to topologically equivalent foliations are equivalent, see [9] and [10]. The case of Morse–Smale flows is more involved and one needs to use the so-called *molecule technique*, for details see Oshemkov and Sharko's paper. There they prove

Lemma B. *Two Morse–Smale flows X and X' on closed two-dimensional manifolds M and M' are topologically equivalent if and only if the corresponding molecules $W(X)$ and $W(X')$ are isomorphic.*

To these molecules, that again are three-colour graphs, this time obtained through a more complex glueing procedure, can be associated a C^*-algebra as in previous of Morse flows leading to Peixoto's theorem in the C^*-algebra context.

References

1. Bronstein, I., Nikolaev, I.: Peixoto graphs of Morse-Smale foliations on surfaces. Topol. Appl. **77**, 19–36 (1977)
2. Connes, A.: A survey of foliations and operator algebras. Operator Algebras and Applications, Part 1, pp. 521–628. Proc. Sympos. Pure Math. vol. 38. AMS, Providence, RI (1982)
3. Connes, A.: The von Neumann algebra of a foliation. Lecture Notes in Phys., vol. 80, pp. 145–151. Springer, Berlin (1978)
4. Connes, A.: Sur la théorie non commutative de l'integration. Lecture Notes in Math., vol. 725, pp. 19–143, Springer, Berlin (1979)
5. Gutiérrez, C., de Melo, W.: The connected components of Morse-Smale vector fields on two manifolds. Geometry and Topology (Proc. III Latin Amer. School of Math., Rio de Janeiro, 1976), pp. 230–251. Lecture Notes in Math., vol. 597, Springer, Berlin (1977)
6. Gutiérrez, C.: A counter-example to a C^2-closing lemma. Ergod. Theory Dyn. Syst. **7**(4), 509–530 (1987)
7. Gutiérrez, C.: On C^r-closing for flows on 2-manifolds. Nonlinearity **13**(6), 1883–1888 (2000)
8. Gutiérrez, C., Pires, B.: On Peixoto's conjecture for flows on non-orientable 2-manifolds. Proc. Amer. Math. Soc. **133**(4), 1063–1074 (2005)
9. Haefliger, A.: Groupoids and Foliations. Contemp. Math. **282**, 83–100 (2001)
10. Muhly, P., Renault, J., Williams, F.: Equivalence and isomorphism for groupoid C^*-algebras. J. Operat. Theory **17**, 3–22 (1987)
11. Nikolaev, I., Zhuzhoma, E.: Flows on 2-dimensional Manifolds – An Overview. Lecture Notes in Math., vol. 1705. Springer, Berlin (1999)
12. Nikolaev, I.: Foliations on Surfaces. Ergebnisse der Mathematik and ihrer Grenzgebiete, vol. 41. Springer, Berlin (2001)
13. Oshemkov, A.A., Sharko, V.V.: Classification of Morse-Smale flows on two-dimensional manifolds. Sb. Math. **189**(8), 1205–1250 (1998)
14. Peixoto, M.M.: On the classification of flows on 2-manifolds. In: Peixoto, M.M. (ed.) Proceedings Symposium Dynamical Systems, pp. 389–419. Academic, New York (1973)
15. Thom, R.: Généralisation de la théorie de Morse aux variétes feuilletées. Ann. Inst. Fourier **14**, 173–190 (1964)
16. Wang, X.: The C^*-algebras of Morse-Smale flows on two-manifolds. Ergod. Theory Dyn. Syst. **10**, 565–597 (1990)
17. Winkelkemper, H.: The graph of a foliation. Ann. Glob. Anal. Geom. **1**, 51–75 (1983)

Chapter 48
Fractional Control of Legged Robots

Manuel F. Silva and J.A. Tenreiro Machado

Abstract Fractional calculus (FC) is being used in several distinct areas of science and engineering, being recognized its ability to yield a superior modelling and control in many dynamical systems. This article illustrates the application of FC in the area of robot control. A Fractional Order PD$^{\mu}$ controller is proposed for the control of an hexapod robot with 3 dof legs. It is demonstrated the superior performance of the system by using the FC concepts.

48.1 Introduction

Walking machines allow locomotion in terrain inaccessible to other type of vehicles, since they do not need a continuous support surface, but require systems for leg coordination and control [1]. For multi-legged robots, the control at the joint level is usually implemented through a PID scheme with position/velocity feedback. Recently, the application of the theory of FC to robotics revealed promising aspects for future developments [2].

Bearing these ideas in mind, the article presents the application of a FO PD$^{\mu}$ ($0 < \mu \leq 1$) controller in the control of an hexapod robot with 3 dof legs. Section 2 introduces the hexapod robot kinematic and dynamic models and the adopted controller architecture. Section 3 presents some simulation results showing the superior performance of the system under the action of a fractional-order controller. Finally, Sect. 4 addresses the main conclusions.

M.F. Silva (✉) and J.A.T. Machado
Department of Electrotechnical Engineering, Institute of Engineering of Porto, Rua Dr. António Bernardino de Almeida, 431, 4200-072 Porto, Portugal
e-mail: mss@isep.ipp.pt, jtm@isep.ipp.pt

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_48,

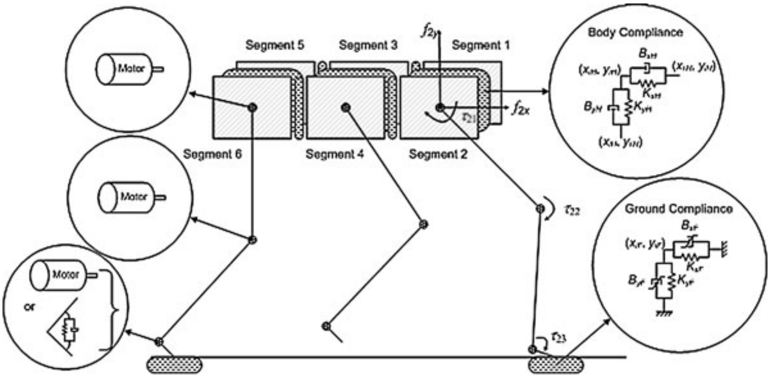


Fig. 48.1 Model of the robot body and foot-ground interaction

48.2 Hexapod Robot Model and Control Architecture

The present study compares the tuning of Fractional Order (FO) algorithms, applied to the joint control of a walking robot with $n = 6$ legs, equally distributed along both sides of the robot body, each with three rotational joints $j = \{1, 2, 3\} \equiv$ {hip, knee, ankle} (Fig. 48.1) [3]. Leg joint $j = 3$ can be either mechanical actuated, or motor actuated. For the mechanical actuated case we suppose that there is a rotational pre-tensioned spring-dashpot system connecting leg links L_{i2} and L_{i3}. This mechanical impedance maintains the angle between the two links while imposing a joint torque [3].

Figure 48.1 presents the dynamic model for the hexapod body and the foot-ground interaction. It is considered the existence of robot intra-body compliance because most walking animals have a spine that allows supporting the locomotion with improved stability. The robot body is divided in n identical segments (each with mass $M_b n^{-1}$) and a linear spring-damper system (with parameters defined so that the body behaviour is similar to the one expected to occur on an animal) is adopted to implement the intra-body compliance [3]. The contact of the i^{th} robot feet with the ground is modelled through a non-linear system, being the values for the parameters based on the studies of soil mechanics [4].

The general control architecture of the hexapod robot is presented in Fig. 48.2. We evaluate the effect of different PD$^{\mu}$ controller implementations for $G_{c1}(s)$, while G_{c2} is a P controller. The PD$^{\mu}$ $0 < \mu_j \leq 1$ $(j = 1, 2, 3)$ algorithm is implemented through a discrete-time 4th-order Padé approximation.

The performance analysis is based on the formulation of two indices measuring the mean absolute density of energy per travelled distance (E_{av}) and the hip trajectory errors (ε_{xyH}) during walking [5]. It is analyzed the system performance of the different PD$^{\mu}$ controller tuning, when adopting a periodic wave gait at a constant forward velocity V_F, for two distinct cases: the hip and knee joints are motor actuated while the ankle joint is mechanically (passively) actuated, and the three leg joints are fully motor actuated [3].

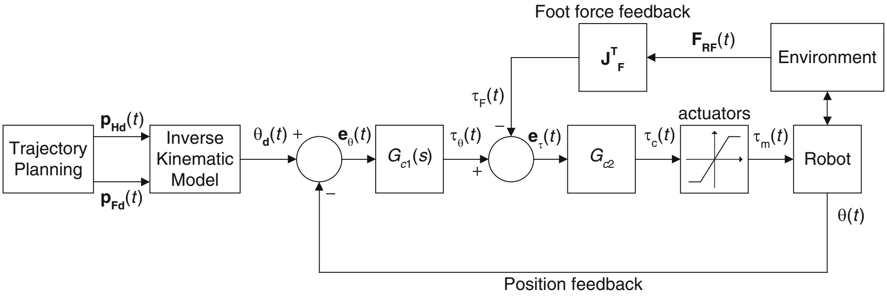


Fig. 48.2 Hexapod robot control architecture

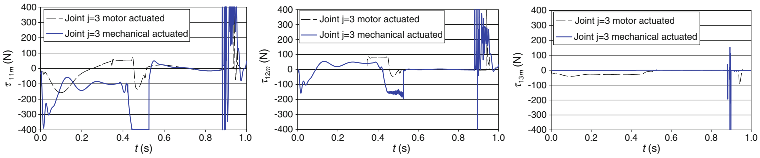


Fig. 48.3 Plots of τ_{1jm} *vs.* t, with joints 1 and 2 motor actuated and joint 3 mechanical actuated and all joints motor actuated, for $\mu_j = 0.5$

48.3 Simulation Results

To tune the different controller implementations we adopt a systematic method, testing and evaluating several possible combinations of parameters, for all controller implementations. We adopt the $G_{c1}(s)$ parameters that establish a compromise in what concerns the simultaneous minimization of E_{av} and ε_{xyH}, and a proportional controller G_{c2} with gain $Kp_j = 0.9$ $(j = 1, 2, 3)$. It is assumed high performance joint actuators, having a maximum actuator torque of $\tau_{ijMax} = 400\,\text{Nm}$. The desired angle between the foot and the ground (assumed horizontal) is established as $\theta_{i3hd} = -15°$. We start by considering that leg joints 1 and 2 are motor actuated and joint 3 has a passive spring-dashpot system. For this case we tune the PD$^\mu$ controllers for values of the fractional order in the interval $0 < \mu_j < 0.9$, establishing $\mu_1 = \mu_2 = \mu_3$. Afterwards, we consider that joint 3 is also motor actuated, and we repeat the controller tuning procedure seeking for the best parameters.

When joint 3 is mechanically actuated, the value of $\mu_j = 0.6$ leads to the best compromise situation in what concerns the simultaneous minimization of ε_{xyH} and E_{av}. When all joints are motor actuated, $\mu_j = 0.5$ leads to the best compromise between ε_{xyH} and E_{av}. Furthermore, the best case corresponds to all leg joints being motor actuated.

In conclusion, the experiments reveal the superior performance of the PD$^\mu$ controller for $\mu_j \approx 0.5$ and a robot with all joints motor actuated. The good performance can be verified in the joint actuation torques τ_{1jm} (Fig. 48.3) and the hip trajectory tracking errors Δ_{1xH} and Δ_{1yH} (Fig. 48.4).

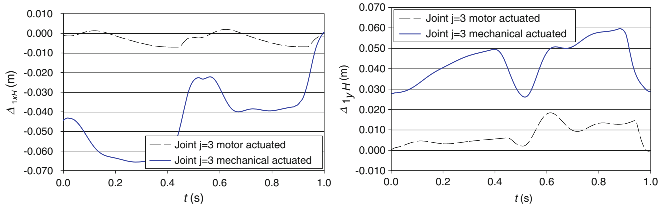


Fig. 48.4 Plots of Δ_{1xH} and Δ_{1yH} *vs.* t, with joints 1 and 2 motor actuated and joint 3 mechanical actuated and all joints motor actuated, for μ_j = 0.5

48.4 Conclusions

This article presented the application of the FC concepts in the area of control systems. A PD$^{\mu}$ controller was used in the control of an hexapod robot with 3 dof legs. It was shown the superior performance of the overall system when adopting a FO PD$^{\mu}$ controller with $\mu_j \approx 0.5$, and a robot having all joints motor actuated.

References

1. Silva, M.F., Machado, J.A.T.: A historical perspective of legged Robots. J. Vib. Control **13**(9–10), 1447–1486 (2007)
2. Silva, M.F., Machado, J.A.T., Lopes, A.M.: Comparison of Fractional and Integer Order Control of an Hexapod Robot. In: Proceedings of the VIB 2003 – ASME International 19th Biennial Conference on Mechanical Vibration and Noise. USA (2003)
3. Silva, M.F., Machado, J.A.T., Jesus, I.S.: Modelling and Simulation of Walking Robots With 3 dof Legs. In: Proceedings of the MIC 2006 – The 25th IASTED International Conference on Modelling, Identification and Control. Lanzarote, Spain (2006)
4. Silva, M.F., Machado, J.A.T.: Position/Force control of a walking Robot. MIROC Mach. Intell. Robot Control **5**, 33–44 (2003)
5. Silva, M.F., Machado, J.A.T.: Kinematic and dynamic performance analysis of artificial legged systems. Robotica **26**(1), 19–39 (2008)

Chapter 49
Evolutionary Dynamics of the Spatial Prisoner's Dilemma with Single and Multi-Behaviors: A Multi-Agent Application

Carla Silva, Welma Pereira, Jan Knotek, and Pedro Campos

Abstract This work explores an application of the spatial prisoner's dilemma in two situations: when all agents use the same type of behavior and when they use a mix of behaviors. Our aim is to explore the evolutionary dynamics of this game to analyze the dominance of one strategy over the other. We also investigate, in some possible scenarios, which behavior has better performance when they all coexist in the same environment.

49.1 Introduction

Game theory is an important way of understanding the dynamics of certain behaviors and to analyze the evolution of the components involved. The prisoners' dilemma is a game that raises the problem of cooperation in a stark form: two strategies are available (cooperate or defect). The payoff to mutual cooperation exceeds the payoff to mutual defection (P). Much of the literature on the evolution of cooperation following Axelrod's seminal contribution [2] has sought to identify the factors that influence the possibility of cooperative behavior emerging in populations of boundedly-rational agents playing the repeated prisoners' dilemma (RPD). Hoffmann and Waring [7] have studied the problem of the localization of the agents in the RPD. One important contribution to this area is due to Nowak and May [17]. The authors study a population of RPD playing cellular automata distributed on squares displayed on a torus which are capable only of the Always Defect (ALL-D) and Always Cooperate (ALL-C) strategies. Individual agents interact with

C. Silva (✉)
Faculty of Economics, University of Porto, Porto, Portugal
e-mail: cmap.silva@ymail.com

W. Pereira, J. Knotek, and P. Campos
LIAAD-INESC Porto and Faculty of Economics, University of Porto, Porto, Portugal
e-mail: welma.pereira@gmail.com, jan.knotek@gmail.com, pcampos@fep.up.pt

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_49,

all neighbors on their eight diagonally and orthogonally adjacent squares and are able to imitate the strategy of any better-performing one among them. Nowak and May have found that the distribution of strategies on the torus depends on the relative size of the RPD-payoffs.

Both the static and dynamic perspectives of evolutionary game theory provide a basis for equilibrium variety as we notice in our multiple simulations. According to Szabó and Fáth [20] there is a static and a dynamic perspective of evolutionary game theory, in the next sections we deeply explore this game dynamics perception. In Zimmermann and Eguluz [25] research the concluding equilibrium solution is composed mostly by cooperative agents, in a prisoner's dilemma with adaptive local interactions, which focus cooperative behavior among a group of agents assuming adaptive interactions. In our case we do not deeply focus on the agent leadership's issue acquire after perhaps existing adaptive interactions or rules. We could analyze this question with specific measures in the link analysis method, but our main goal was to focus on the strategy leadership's issue, view the dominance of one strategy in a spatial environment, focusing assorted issues. In our simulations we had a diversity of situations, as in the more significant payoffs changes occurs when one strategy was trying to be the leader innovation strategy.

In this work, we use Agent-based simulation to analyze the prisoner's dilemma with three types of behavior. We will use three types of behaviors: copy best player (greedy), copy best strategy (conformist) and Pavlovian. After presenting them individually, we join all the three in the same playground. We will explore the game dynamics when the parameters or the initial conditions change. The evaluation is made using statistics and link analysis. Our model is based in physical properties of the automata. Spatial interaction of autonomous actors selects actions from their own logical set based on its own state and on its neighbor's states. We apply this social interaction to 100 agents, each one having one of the three ways of action (behaviors), so that the outcome depends on the choices of all the players based on the Moore Neighborhood, played with eight neighbors (as in [7]). The game dynamics is determined locally since the neighborhood is defined in a finite region. These geographic effects are represented by placing agents in territorial structures and restricting them to interact and learn within certain geographic regions. We consider that in real-world, these agents can be seen as companies. Each agent has an initial strategy which can be either cooperate or defect. Once the dynamics of this game gives rise to clusters of collaborators and/or defectors. We also decided to use social network analysis in order to capture the link relation in the networks of firms that emerged in the different scenarios.

We concluded that Pavlovian behavior scores better than Greedy and Conformist when b (the payoff that corresponds to the defection of a player and the cooperation of the other) is higher. Additionally we can see that when we have more companies with collaborative strategy (small b) we get higher average payoffs in all behaviors. The maximum individual payoff is also higher for small values of b when companies use behavior Greedy or Conformist.

The work is structured as follows: in Sect. 49.2 we define the prisoners' dilemma and give a short overview of game theory. We also define the gain matrix and the strategies involved in the model. Section 49.3 contains the implementation of the model. Section 49.4 describes the experiments. Data analysis is in Sect. 49.5, in Sect. 49.6 we describe possible dynamics GIS application and in Sect. 49.7, we discuss the corresponding results. Finally, in Sect. 49.8, we introduce some future work.

49.2 The Prisoner's Dilemma: A Short Overview of Game Theory and Definition of the Strategies

The prisoner's dilemma is the name given by Albert W. Tucker [6] to the following problem in game theory: Suppose the situation in which there are two suspects (let's say P1 and P2) of a crime that are arrested in separated cells by the police. A prosecutor meets with the prisoners separately and offers the same deal to both of them. They can either testify against the other or to remain silent. If prisoner P1 decides to testify and prisoner P2 remains silent, P1 goes free and P2 receives a 10-years full sentence. Likewise, if prisoner P2 testifies while P1 remains silent, P2 will be the one to be set free while P1 stays in prison for the same 10-years time. If they both testify they both stay in prison for a shorter sentence of 7-years. If in the last case, they both remain in silent, P1 and P2 are sentenced to only 1 year in prison. The problem of the prisoners is to decide weather to remain silent or to testify against the other. Since the prisoners are both isolated, neither of them would know the other's decision. This story is usually generalized to analyze similar situations.

Analyzing the options individually if the other remains in silent, the best is to testify against him (defect) so that you can be set free. But if the other decides to defect the best is also to defect, otherwise you will stay in jail while the other leaves free. In the other hand the best solution for both of them is to remain in silent because in this situation, even though neither of them is set free, they still get a much shorter sentence. The dilemma is that they do not know what the other will do. A common view is that this puzzle illustrates a conflict between individual and group rationality. If each player has chosen a strategy and no player can benefit by changing his or her strategy while the other players keep theirs unchanged, then the current set of strategy choices and the corresponding payoffs constitute a Nash equilibrium (Table 49.1).

Table 49.1 Original payoff matrix [6]

	Cooperate (C)	Defect (D)
Cooperate (C)	7,7	10,0
Defect (D)	0,10	1,1

49.2.1 Companies and Strategies

We decided to apply our model to the situation where there is a bunch of companies in the same market business competing with each other. The companies are randomly disposed in a matrix and they interact with their closest eight neighbors surrounding it. For example, the companies can be restaurants in the city of Porto and their business is to sell “francesinhas”, the typical dish of this Portuguese city. The neighbors can be seen as the nearest restaurants that competes in the same physical area or district. Another possible application is to see the proximity in the matrix as the companies with most similar activities that competes with each other not necessarily in the same region but in the same kind of business.

Each company can choose to cooperate with their neighbor or not. If they cooperate they form a cartel, i.e. to sell the “francesinha” for the same price and therefore their profit will be quite similar. If they decide not to cooperate they can sell it for cheaper price. In this case they can attract more customers and earn more. This can be seen as the situation of defect in the prisoner’s dilemma problem.

The options of collaboration or defect can be seen as two strategies that can be used by the companies. This business market can then be divided into this two types of strategies followed by companies, the ones that follow the strategy of collaboration, let’s call it strategy A, the companies that follow the strategy of defection, that will be called strategy B. In this situation the companies can change the strategies at any time. Our aim is to explore the evolution of this game in different scenarios and analyze if one of the strategies dominates the other.

To make this situation more real and more competitive we also decided to introduce a reward policy that gives to the cooperators some advantage. Each time a company decides to cooperate, it wins a reward based on its own status and on the number of collaborators in its neighborhood.

49.2.2 Spatial-Temporal Definitions

We start by defining our game geometry. Each point of our spatial lattice has a state. The possible states are: Cooperate (C) or Defect (D). The states of the cells are updated after each round according to agents current state, the states of its eight nearest neighbors and the agents behavior. Implemented behaviors are described later. All cells in the lattice are updated synchronously. We can define the neighborhoods depending on the system we pretend to model. Concerning the two dimensional lattice the following definitions are common (Fig. 49.1):

The Von Neumann neighborhood can be called also as 4-neighborhood as it contains four cells: the cell above, below, to the right and to the left. The radius of this definition is one. The Moore neighborhood is an enlargement of the Von Neumann neighborhood containing the diagonal cells too, the radius is also one. The Moore neighborhood is also called 8-neighborhood. The extended Moore neighborhood is equivalent to the description of Moore neighborhood, but it reaches over the

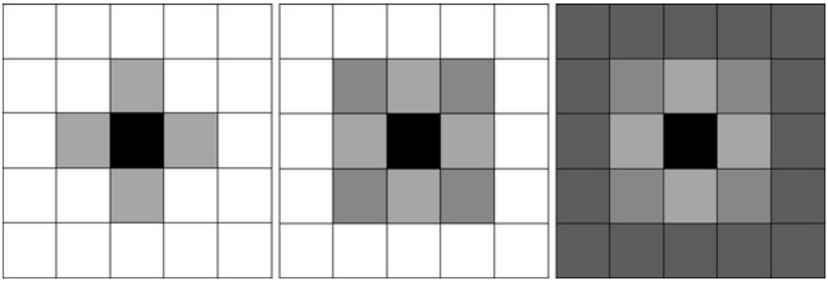
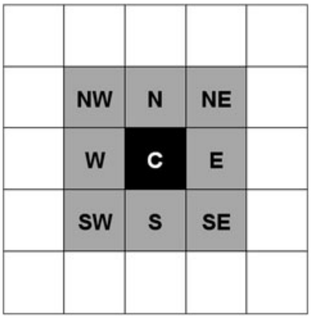


Fig. 49.1 Von Neumann neighborhood, Moore neighborhood and Extended Moore neighborhood

Fig. 49.2 The gray cells are the neighbors of the central one. As previous discussed the states of these cells are used to calculate the subsequently state of the center cell according to the rule defined



distance of the next adjacent cells, in this case the radius is two. Extended Moore Neighborhood can be also called as 25-neighborhood.

In our simulations, we decided to use the technique of Moore neighborhood. Comprising eight cells surrounding a central cell in a square lattice. We could also have used (among others) the technique of Von Neumman neighborhood. The spatial games can vary in many ways its geometry (Fig. 49.2).

Initially random positions are generated. Each cell of the network is occupied by an agent (player, company). To avoid edge effects, the edges of the network (or endpoints of the line) are glued together (Fig. 49.3).

Nowak and May [17] studied a population of agents distributed on cells of a 2-dimensional torus which are capable only of Always – Cooperate (C) or Always – Defect (D) strategies. In our paper we are using the same type of strategies but we rename them as (A) or (B). Nowak and May [17] in their study found out that the distribution of strategies on the torus depends on the relative size of the payoffs. They refer that the future state of each cell depends on the current state of the cell and the states of the cells in the neighborhood, the development of each cell is defined by rules. As we had referred before. In our case rules fall into three basic strategy behaviors. In our game 100 agents play during n time units, which means one or three kind of behaviors in a playground being able to change in each round. The evolution of these rules leaded us to observe dynamical patterns. The survival cooperative and/or defect behavior.

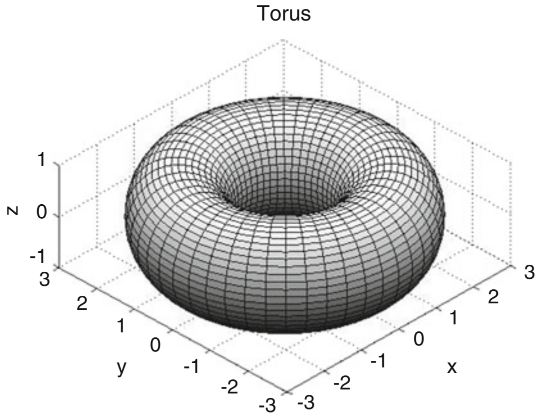


Fig. 49.3 By connecting square lattice from left to right side and from up to bottom side we obtain a torus topology, where all cells are equal (all have the same count of neighbors). Here we have an example image made in *Matlab*, using *surf* command, which draws 3-D shaded surface plot

As Jun and Sethi refer in [9], the survival of cooperative behavior in populations in which each person interacts only with a small set of social neighbors, the individuals adjust their behavior over time by shortsightedly imitating more successful strategies within their own neighborhood. All this process leaded us to a called Demographic Game, as Epstein [5] refer seems an appropriate name for this class of models because they involve spatial, evolutionary, and population dynamics. In Epstein's research [5], each agent is an object whose one main attributes is his vision. He considers vision like the distance an agent can see, looking north, south, east, or west. In our case we assume that all of our agents have a peripheral vision, looking 360° around, so we also include north–west, south–west, north–east and south–east. Our agents move around this space, interacting with Moore Neighbors (all with peripheral vision).

49.2.3 Gains Matrix

In what follows, a player must choose between two strategies. Isaac [8] refers that by altering a single entry of the payoff matrix could demonstrate that payoff cardinality is crucial to prisoner's dilemma outcomes on an evolutionary grid. And also refers that the evolutionary processes are fundamental to cooperation in social situations and have been an enduring theoretical problem in diverse areas, like biological, sociological, and geographical. Using Nowak [16] calculating gains according to his definition, we build for each round a gain matrix, which gives us payoffs values, the ones that determine if the agent will stay with the same strategy or will change according to his neighbors payoffs (Table 49.2).

Table 49.2 Nowak [16] defines the payoffs values according to the above rules

	Cooperate (C)	Defect (D)
Cooperate (C)	1	0
Defect (D)	b	ϵ

Calculation of Gain:

1. If C find D, then D obtains b>1 and C gets 0
2. If D satisfies D, D gets ϵ
3. Cooperate if both C and C, then each gets 1

Where $1<b<2$ and $\epsilon \rightarrow 0$.

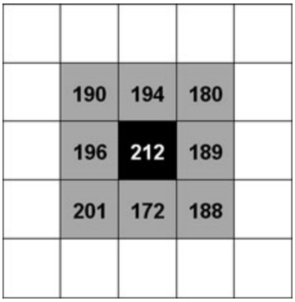
Nowak and May [17] presented the seminal work that showed how the spatial effects of the interactions between simple agents in a cellular automaton model of the iterated prisoner's dilemma was sufficient enough for the evolution of cooperation. A similar cellular automaton model was built that simulated cooperation through the behavioral adaptation of Pavlovian agents as they adjusted their cooperation by mimicking the most successful player in a neighborhood. In this present paper we observe this cooperation or non-cooperation behaviors varying the b value between 1 and 2 and establishing ϵ as 0.01.

49.2.4 Game Dynamics

Cooperation is frequently observed in real-life psycho-economic experiments. This result either suggests that the abstract Prisoner's Dilemma game is not the right model for the situation or that the players do not fulfill all the premises. Indeed, there is good reason to believe that many realistic problems, in which the effect of an agent's action depends on what other agents do are far more complex that perfect rationality of the players could be postulated, Szabó and Fáth [20]. Nevertheless, the standard deductive reasoning loses its appeal when agents have non-negligible cognitive limitations, there is a cost of gathering information about possible outcomes and payoffs. Like Xianyu [23] refer this in his recent paper. We also agree that agents have incomplete information on other agents' strategies, so the agents need to learn and develop their own strategies in this unknown environment. Mind necessarily becomes an endogenous dynamic variable of the model. This kind of bounded rationality may explain that in many situations people respond instinctively, play according to heuristic rules and social norms rather than adopting the strategies indicated by rational game theory. In our simulation we compute three rational behaviors.

49.2.5 Three Behaviors Strategies

According to the payoff and to the strategy agent chooses if he wants to change strategy or not. This means changing from A to B (Cooperate to Defect) or vice-versa,



	190	194	180	
	196	212	189	
	201	172	188	

Fig. 49.4 Each cell of the network is occupied by a player, and each one has one associated payoff value calculated using the gain matrix

or else no change at all. In the following example the player won't change strategy because his payoff is bigger than the payoff of its neighbors (Fig. 49.4).

We chose to implement in R following three behaviors: Copy Best Player (greedy), Copy Best Strategy (conformist) and Pavlovian. Three kinds of social preference theories have been tested. As Oliver Kirchkamp [11] we apply the idea of evolution to a spatial model, were prisoners' dilemmas or coordination games are played repeatedly within neighborhoods where players instead of optimizing in each round, prefer to copy successful strategies. Discriminative behavior of players is introduced representing strategies as small automaton, which can be in different states against different neighbors. These personality types represent certain simple aspects of actual human behavior. Pavlovian agents are the most realistic automaton for the investigation of the evolution of cooperation, because they are simple enough to know nothing about their rational choices but intelligent enough to follow an action that produces a satisfactory state of affairs tends to reinforce the repetition of that particular action.

According to Power [18], greedy is an agent who imitates the neighbor with the highest reward. Then conformist is an agent who imitates the action of the majority in the social unit. And Pavlovian is an agent with a coefficient of learning whose probability of cooperation changes by an amount proportional to the reward/penalty it receives from the environment.

49.2.5.1 Copy Best Player (Oliver Kirchkamp [10])

Greedy, a learning player can simply look around in the neighborhood which he observes and determine the player with the highest payoff. A learning player that uses the rule "copy best player" will pick the strategy of the most successful player. Of course, it could well be that there is more than a single player who has the maximal payoff. Then let players use the following tie breaking rule. Define the set of most successful players in neighborhood N_L^i of player i as

$$M^{i,t} \leftarrow argmax_{j \in N_L^i} \left(\frac{\pi_\epsilon^{j,t}}{n_\epsilon^{i,t}} \right) \tag{49.1}$$

where $\pi_\epsilon^{j,t}$ is player's j payoff in time t and $n_\epsilon^{i,t}$ is a count of players in the neighborhood.

The probability that player i choose strategy s in period $t + 1$ is determined as

$$P(x^{i,t+1} = s) \leftarrow \begin{cases} 1 & \text{if } x^{i,t} \in \left\{x^{j,t} | j \in M^{i,t}\right\} \, \mathit{and} \, s = x^{i,t} \\ 0 & \text{if } x^{i,t} \in \left\{x^{j,t} | j \in M^{i,t}\right\} \, \mathit{and} \, s \neq x^{i,t} \\ \frac{\sum_{j \in M^{i,t} \wedge x^{j,t}=s} n_\epsilon^{j,t}}{\sum_{j \in M^{i,t}} n_\epsilon^{j,t}} & \mathit{otherwise} \end{cases} \tag{49.2}$$

where $M^{i,t}$ is a set of best players in neighborhood of player i in time t and $n_\epsilon^{j,t}$ is a count of neighbors of player j in time t.

Thus the player that is to be copied is chosen randomly with probabilities that are proportional to the number of interactions the respective best players had. In the special case where the player's own strategy is among the best strategies, we assume that the player prefers to keep his own strategy.

In our experiments our greedy agent just changes his strategy based in the highest payoff neighbor.

49.2.5.2 Copy Best Strategy (Oliver Kirchkamp [10])

A learning player is a Conformist when it look at the average payoffs of strategy s at time t in the neighborhood of player i which we designate by $f_s^{i,t}$:

$$f_s^{i,t} \leftarrow \begin{cases} \frac{\sum_{j \in \cup_s^{i,t}} \pi_\epsilon^{j,t}}{\sum_{j \in \cup_s^{i,t}} n_\epsilon^{j,t}} & \text{if } \sum_{j \in \cup_s^{i,t}} n_\epsilon^{j,t} > 0 \\ -\infty & \mathit{otherwise} \end{cases} \tag{49.3}$$

where $\cup_s^{i,t}$ is a set of players in neighborhood of player i with strategy s in time t, $\pi_\epsilon^{j,t}$ is payoff of player j in time t and $n_\epsilon^{j,t}$ is a count of players in neighborhood of player j in time t.

If a strategy is not used in a neighborhood, we define its fitness to be $-\infty$ to make sure that it will be never selected by an evolutionary process. A learning player that uses the rule "copy best strategy" switches to the strategy with the highest average payoff. Again there could be more than one strategy with maximal payoff. Then we use the following tie breaking rule: define the set of most successful strategies as:

$$N^{i,t} \leftarrow argmax_s \left(f_s^{i,t}\right) \tag{49.4}$$

As for the "copy best player", two strategies could achieve exactly the same average payoff. The probability that player i uses strategy s in the next period is then

$$P(x^{i,t+1} = s) \leftarrow \begin{cases} 1 & \text{if } x^{i,t} \in N^{i,t} \ and \ s = x^{i,t} \\ 0 & \text{if } x^{i,t} \in N^{i,t} \ and \ s \neq x^{i,t} \\ \dfrac{\sum_{j \in \cup_s^{i,t}} n_\epsilon^{j,t}}{\sum_{\sigma \in N^{i,t}} \sum_{j \in \cup_\sigma^{i,t}} n_\epsilon^{j,t}} & otherwise \end{cases} \tag{49.5}$$

where $\cup_s^{i,t}$ is a set of players in neighborhood of player i with strategy s in time t, $N^{i,t}$ is a set of strategies with highest mean payoff and $n_\epsilon^{j,t}$ is a number of players in neighborhood of player j in time t.

If a player already uses one of the best strategies, he adopts one of the best strategies randomly with probabilities proportional to the number of interactions the users had with the respective strategies.

In our experiments the conformist agent just changes his strategy to the one, which has higher average payoff in players neighborhood.

49.2.5.3 Pavlovian [18]

Like Power [18] we also define the agents as stochastic learning automata with Pavlovian personalities and attitudes.

By definition Pavlov works according to the following algorithm: Szabó and Fáth [20] "repeat your latest action if that produced one of the two highest possible payoffs, and switch to the other possible action, if your last round payoff was one of the two lowest possible payoffs". As such Pavlov belongs to the more general class of Win-Stay-Lose-Shift strategies, which define a direct payoff aspiration level for strategy change. An alternative definition frequently appearing in the literature is "cooperate if and only if you and your opponent used the same move in the previous round", and this translates into the same rule of the Prisoner's Dilemma.

Pavlovian strategies are formulated as a weighted payoff, an average production function, and a three-step memory coefficient of learning. Given an agent, the weighted payoff is defined as:

$$RPwt = \sum_{i=1}^{3} Mc_i \cdot w_i \tag{49.6}$$

$RPwt \leftarrow$ weighted payoff of agent in last three rounds.

W_i is a weighting parameter such that all weights sum to one, and Mc_i is the history payoff. Assuming that the effects of memory decrease with time, $w_1 \geq w_2 \geq w_3$ and $w_1 + w_2 + w_3 = 1$. Let $S(t)$ be the strategy of the player in time t. Then parameter α–learning rate is after each round for each Pavlovian player set in a following manner:

$$\alpha_i(t+1) \leftarrow \begin{cases} \alpha_i(t) + 0.15 & , \text{if } (S(t) = S(t-1)) \wedge (S(t-1) = S(t-2)) \\ \alpha_i(t) + 0.10 & , \text{if } (S(t) = S(t-1)) \wedge (S(t-1) \neq S(t-2)) \\ \alpha_i(t) - 0.10 & , \text{if } (S(t) \neq S(t-1)) \end{cases} \tag{49.7}$$

The probability of cooperation for agent i at time $t + 1$ is:

$$p(t+1) \leftarrow \begin{cases} p(t) + (1 - p(t)) \cdot \alpha_i & \text{, for } S(t) = C \text{ and } RPwt > pf_{avg} \\ (1 - \alpha_i) \cdot p(t) & \text{, for } S(t) = C \text{ and } RPwt \leq pf_{avg} \end{cases} \tag{49.8}$$

For every t there is $q(t) = 1 - p(t)$. If previous action is D:

$$q(t+1) \leftarrow \begin{cases} q(t) + (1 - q(t)) \cdot \alpha_i & \text{, for } S(t) = D \text{ and } RPwt > pf_{avg} \\ (1 - \alpha_i) \cdot q(t) & \text{, for } S(t) = D \text{ and } RPwt \leq pf_{avg} \end{cases} \tag{49.9}$$

The state of agent i is updated contingent on its previous state, the average neighborhood production function, and the probabilities for both C and D. The neighborhood production function for time t is the cooperation payoff for the group following:

$$pf(t) = \frac{\sum C_j}{N} \tag{49.10}$$

C_j is the payoff value for agent j and N is the total number of agents in the neighborhood. The average neighborhood function for three memory events is given by:

$$pf_{avg} = \frac{\sum_i^3 pf_i}{3} \tag{49.11}$$

pf_{avg} – average payoff in neighborhood in last three rounds.

Thus, the state of agent i at time $t + 1$ with $S(t)$, where $Ru \in [0, 1]$ is a uniform random value:

For $S(t) = C$:

$$S(t+1) \leftarrow \begin{cases} D & \text{if } RPwt \text{ for agent } i < pf_{avg} \text{ and } p(t+1) < q(t+1) \\ & \text{and } q(t+1) > Ru \\ C & \text{if conditions for } D \text{ are not satisfied} \end{cases} \tag{49.12}$$

For $S(t) = D$:

$$S(t+1) \leftarrow \begin{cases} C & \text{if } RPwt \text{ for agent } i < pf_{avg} \text{ and } q(t+1) < p(t+1) \\ & \text{and } p(t+1) > Ru \\ D & \text{if conditions for } C \text{ are not satisfied} \end{cases} \tag{49.13}$$

Pavlovian strategies according to Kraines and Kraines [12], are quite stable even in a noisy environment. Although this strategy cooperates and retaliates, as does Tit-For-Tat [2], it is not tolerant. The Pavlovian behavior will exploit altruistic strategies until he is punished by mutual defection. Pavlovian strategies are natural models for many real life conflict-of-interest.

49.3 Implementation

The experiment was implemented in R [19]. Its main advantage is that there is a large library of packages to compute many statistics, draw graphs etc. R is an open source software and its distributed for free.

49.3.1 Definition of the Game

In our experiments, we used two dimensional arrays of agents as it is done in [17]. Our implementation involves the iterated n-person prisoner's dilemma.

For most of the statistics and simulations, a ten by ten square lattice is used. Boundary conditions are solved periodically – square lattice is bended by two sides into a torus. Periodic conditions simplifies the problem, because each cell (agent) has exactly the same conditions – eight neighbors. Behaviors and rewards are then influenced only by the configuration of strategies around the agent.

In each round, each agent is playing with all agents in a Moore eight-neighborhood and also with himself. Rewards are driven by a simplified evaluation system:

Table 49.3 shows the rewards for each combination of strategies, that can be played by two agents. When Cooperator meets Defector, Defector receives a payoff equal to constant b from interval $\langle 1, 2 \rangle$ and Cooperator receives 0. Cooperators meeting each other receive reward 1 and finally two Defectors playing together receive $\epsilon \rightarrow 0$.

In our experiments, ϵ is set to 0.01 and b is the variable which we are modifying to see the changes in process. These rules of the game are implemented in the function *play*:

```
play<-function(a, b,payoffB, payoffEps){
 if(a== 0 && b== 0) {return (1)}
 if(a== 0 && b== 1) {return (0)}
 if(a== 1 && b== 0) {return (payoffB)}
 if(a== 1 && b== 1) {return (payoffEps)}
}
```

49.3.2 Initialization

The whole simulation process is started by function *PlayDilemma*, which takes as parameters initial strategy matrix, behavior matrix, Payoff B, Payoff ϵ, number of rounds and size of playground.

Table 49.3 Payoff matrix of the two agents game Nowak [16]

	Cooperate (C)	Defect (D)
Cooperate(C)	1	0
Defect (D)	b	ϵ

Initial strategy matrix contains the first strategy, which agents will play. "Zeros" mean Cooperate, "ones" are interpreted as Defect.

Behavior matrix contains 1 for Greedy behavior, Conformist behavior and Pavlovian agent.

49.3.3 One Round

A typical round works as follows:

1. Calculate payoffs for all agents according to their current strategies (Cooperate/Defect)
2. Evaluate last rounds payoffs, compute learning rate α and probability of cooperation (Pavlovian agent)
3. Change strategies by last rounds payoffs and strategies in agents neighborhood using agent's predefined behavior

As the calculation of payoffs for each behavior is the same, the only aspect which one implemented differently for each behavior is the function *changeStrategies*, which is called after each round. For Pavlovian agents, there are also two matrices to be evaluated – Learning rate α matrix and probability matrix. This is described in detail in Pavlovian behavior Sect. 49.3.4.3.

49.3.4 Implemented Behaviors

After each round, agents are evaluating the payoffs and if they need, they change strategy – cooperate or defect, for the next round. Evaluations for changing the strategy are implemented as three types of behavior. In this paper, we are also evaluating how these behaviors are influencing the distribution of complete payoffs among the agents after 50 or more rounds.

49.3.4.1 Copy Best Player

Copy Best Player is the simplest behavior we have used. Agent compares its payoff with payoffs of players in neighborhood and adopts the strategy of the most successful player in previous round.

49.3.4.2 Copy Best Strategy

The second behavior we used is Copy Best Strategy. To compare the strategies in neighborhood, agent uses mean of payoffs gained by players using strategy

Cooperate and mean of payoffs gained by players using strategy Defect. Player compares these means and picks the strategy with higher mean. If the player himself has higher payoff than these means, he keeps its last strategy.

49.3.4.3 Pavlovian

Pavlovian behavior is described in [24] and also in [18]. This behavior is much more complex than previous two, and its results are not always the best.

Pavlovian agent is defined as an agent with a coefficient of learning whose probability of cooperation changes by an amount proportional to the reward/penalty it receives from the environment.

We had to make some adaptation for the Pavlovian agent defined in previous papers, because our environment is not giving any penalties. Our Pavlovian agent is comparing his reward to the mean of rewards of the neighborhood.

Each agent has learning rate α and probability of cooperation p. These two variables are adjusted in each round looking three rounds back on a payoff response from the environment.

49.3.4.4 Three Rounds Payoff

The agent uses weights $w_1 \geq w_2 \geq w_3$ and $w_1 + w_2 + w_3 = 1$. Weight w_1 is the largest in order to make last round more important than previous rounds.

The three rounds payoff is weighted for each agent and computed using this equation:

$$RPwt = \sum_{i=1}^{3} Mc_i \cdot w_i \tag{49.14}$$

where Mc_i is a payoff in one round and w_i is the corresponding weight.

49.3.4.5 Neighborhood Three Rounds Payoff

pf_{avg} is the mean of payoffs in agents neighborhood for last three rounds. It is computed as follows:

$$pf(t) = \frac{\sum Mc_j}{n} \tag{49.15}$$

$$pf_{avg} = \frac{\sum_{i=0}^{2} pf(t-i)}{3} \tag{49.16}$$

Mc_j are the payoffs of neighborhood agents and n is the complete count of agents in neighborhood. In our case it is always set to eight because we used Moore neighborhood. From the first equation we get the average production in the neighborhood in one round. To get the three rounds mean, we compute the mean of $pf(t-2), pf(t-1), pf(t)$ using the second equation.

49.3.4.6 Learning Rate α

To know if the player was changing the strategy a lot or he wasn't changing at all, we are introducing parameter α. This parameter allows the Pavlovian agent to start changing the strategy more often, if the constant behavior has low payoff compared to mean payoff of the neighbors or in other situation, agent can remain in similar changing rate as he had in last rounds. Learning rate α is a bounded variable always set to a value from the interval [0, 1]. If α is close to zero, it means, that agent was changing strategies often in the past rounds. If α is close to one, the agent was stable in last rounds. In each round α is adjusted according to the following (49.17):

$$\alpha_i(t+1) \leftarrow \begin{cases} \alpha_i(t) + 0.15 & \text{, if } (S(t) = S(t-1)) \wedge (S(t-1) = S(t-2)) \\ \alpha_i(t) + 0.10 & \text{, if } (S(t) = S(t-1)) \wedge (S(t-1) \neq S(t-2)) \\ \alpha_i(t) - 0.10 & \text{, if } (S(t) \neq S(t-1)) \end{cases} \tag{49.17}$$

Where t means the time of the round, and $S(t)$ is the agent's strategy in round t.

49.3.4.7 Probability of Cooperation

p is the probability of using a cooperate strategy in a given round. Notation $p(t)$ means probability of cooperation in round t. Probability is adjusted in each round by the payoff response of the environment. If previous strategy is C, then probability of Cooperation is computed as:

$$p(t+1) \leftarrow \begin{cases} p(t) + (1 - p(t)) \cdot \alpha_i & \text{, for } S(t) = C \text{ and } RPwt > pf_{avg} \\ (1 - \alpha_i) \cdot p(t) & \text{, for } S(t) = C \text{ and } RPwt \leq pf_{avg} \end{cases} \tag{49.18}$$

Note that for every t there must be $q(t) = 1 - p(t)$. This is used in the implementation to have only one matrix for probabilities.

The same set of equations is used for updating the action probabilities when the previous action is D. Probability q of defect is computed as:

$$q(t+1) \leftarrow \begin{cases} q(t) + (1 - q(t)) \cdot \alpha_i & \text{, for } S(t) = D \text{ and } RPwt > pf_{avg} \\ (1 - \alpha_i) \cdot q(t) & \text{, for } S(t) = D \text{ and } RPwt \leq pf_{avg} \end{cases} \tag{49.19}$$

Final strategy is chosen by probability of cooperation and defection and also by the last round's payoff. Conditions we used are slightly modified from conditions of previous paper [18]. To make the Pavlovian agent a non-random agent, in our (49.18) and (49.19) we removed the condition containing random number Ru, which was compared with cooperation/defection probability in equations in Sect. 49.2.5.3 from [18]. This was needed to have the chance to repeat each play in the same way.

For $S(t) = C$:

$$S(t+1) \leftarrow \begin{cases} D & \text{if } RPwt \text{ for agent } i < pf_{avg} \text{ and } p(t+1) < q(t+1) \\ C & \text{if conditions for } D \text{ are not satisfied} \end{cases} \tag{49.20}$$

For $S(t) = D$:

$$S(t+1) \leftarrow \begin{cases} C & \text{if } RPwt \text{ for agent } i < pf_{avg} \text{ and } q(t+1) < p(t+1) \\ D & \text{if conditions for } C \text{ are not satisfied} \end{cases} \tag{49.21}$$

Pavlovian agent is the most sophisticated agent we used, because it takes longer to compute each round. This higher demand for time should have been rewarded by better results in experiments, which were not so good as we expected.

49.3.5 Cooperation Reward Policy

We decided to use some reward policy to give some advantage to the cooperators. We only use it though in our multi-behaviors situation. Here is how it works: every time a player decides to cooperate it will receive a reward that will be added to his payoff. The reward is defined as follows:

$$0.1 \times Size.Group \tag{49.22}$$

where the Size.Group represents the number of cooperators among its eight neighbors. The idea behind is that group of cooperators will get benefited proportionally to its group size.

49.4 Experiments

This section analyzes the dynamics of the spatial prisoner's dilemma that we have implemented. The game is played repeatedly for twenty rounds for each behavior described before. The evolution of the strategies and their respective payoffs are displayed through graphics.

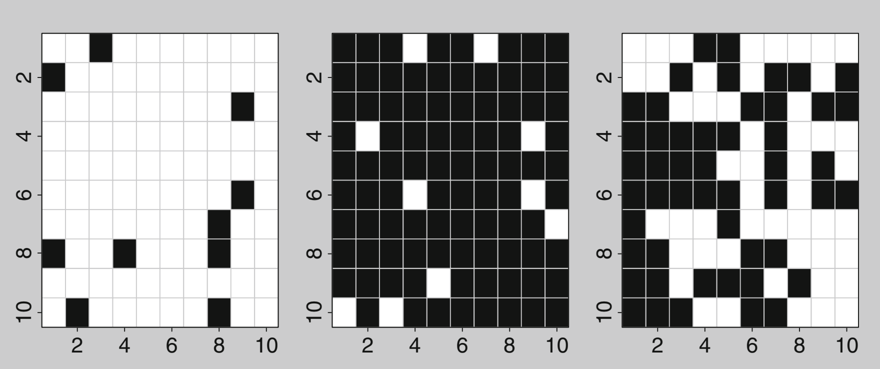


Fig. 49.5 The initial matrices (M1, M2 and M3)

We investigate our application in the situation where the companies have all the same behavior (Single behavior situation) and when all three behaviors coexist in the same playground (Multi-behaviors situation).

The initial matrix was M3 (Fig. 49.5). M3 was generated randomly with 50–50% of the companies using strategies A and B. In the single behavior situation, we always used the same initial matrix so that allows us to compare our results. In the multi-behaviors situation, the experiments were performed in different scenarios with the 3 types of initial matrices.

49.4.1 Single Behavior Situation

The experiments performed in this section analyze the dynamics of a situation where all the companies have the same type of behavior.

Table 49.4 Mean Payoffs of the three behaviors

Payoffs (bs)	1.1	1.5	1.9
Greedy Behavior	169.3	130.5	8.7
Conformist Behavior	162.9	84.8	50.5
Pavlovian Behavior	99.9	87.9	90.2

Figure 49.6 shows the evolution of strategies A and B when all the companies use the behavior greedy the b is equal to 1.1. After stage 4 it doesn't change anymore. We represent the companies with strategy A by black squares and the companies with strategy B by white squares. We can see that the number of blacks increases until only one white company survives in the end.

For $b = 1.5$, Fig. 49.7 shows that even though more blacks can be seen in stage 10, the dynamics of the game for this value of b is more intense and therefore the companies change the strategies more often. If we increase the value of b to 1.9 (very close to the limit of 2), all the companies choose strategy B just after the second interaction.

With the conformist behavior, a similar situation happens. We can see in Fig. 49.9 that after five interactions two companies with strategy B survives in the world of A's with a low b. For b = 1.5 the instability occurs once again and for a high b, the "whites" are the majority. But there are a few "blacks" that survive, more than in the greedy behavior situation.

As for the Pavlovian behavior Figs. 49.12–49.14 show that there are even more changing for all values of b. Figure 49.15 shows how the frequencies of A's (solid line) and B's (dashed line) change. The frequencies start at the same point (0.5) as we start with an initial matrix with a 50–50% of A's and B's. As time passes, the solid line goes up and the dashed line goes down, showing that there are more A's and B's for a b of 1.1. As the value of the b increases, the dashed line exceeds the solid line for all behaviors. We notice though that the competitions between A's and B's are harder for the smarter companies. This can seen as looking on how close (and how much they move up and down) the solid and dashed lines are for the situation where companies use the Pavlovian or the conformist behavior in relation to the situation where they use the greedy behavior (our less intelligent behavior).

Concerning the payoffs we notice from the Table 49.4 that in general the average payoff of all companies decreases as we increase the value of b. We can also notice that our smarter behavior (Pavlovian) scores better in situations with higher values of b. For b = 1.9, Pavlovians have an average payoff of 90.2 while the conformists have 50.5 and the greedy agents have only 8.7. Additionally we can see that when we have more companies with strategy A (small b), ie. selling a product for the same price, we get higher average payoff for all behaviors. The maximum individual payoff is also higher for small values of b when the companies use behavior greedy or conformist but when they use the Pavlovian the maximum is reached when b is high (1.9).

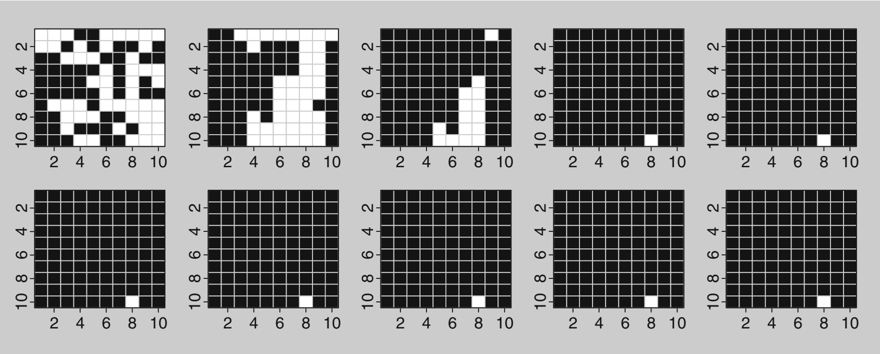


Fig. 49.6 The evolution of strategies A (*black*) and B (*white*) using the Greedy behavior and $b = 1.1$

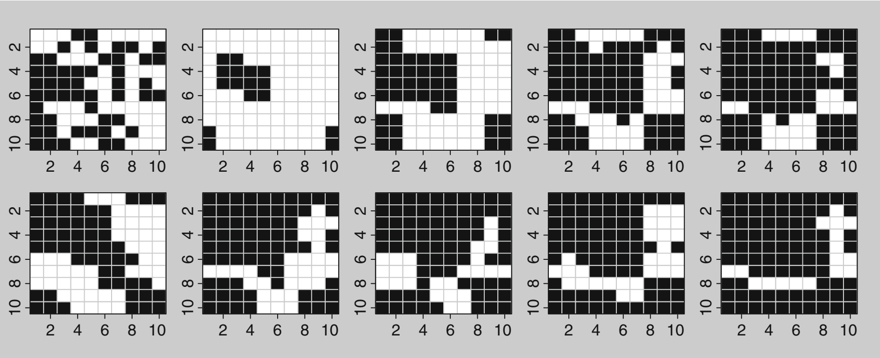


Fig. 49.7 The evolution of strategies A (*black*) and B (*white*) using the Greedy behavior and $b = 1.5$

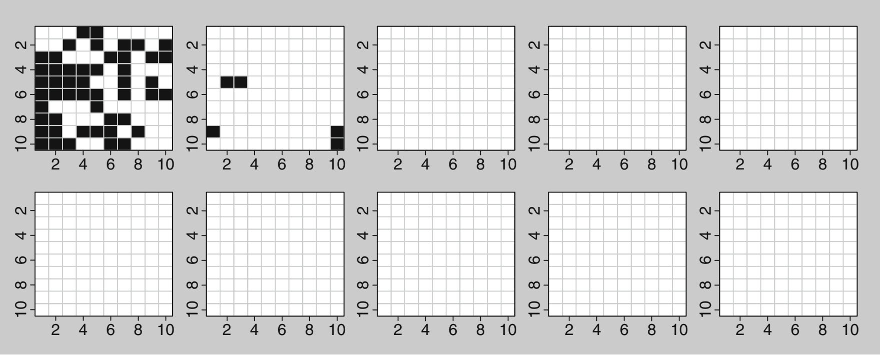


Fig. 49.8 The evolution of strategies A (*black*) and B (*white*) using the Greedy behavior and $b = 1.9$

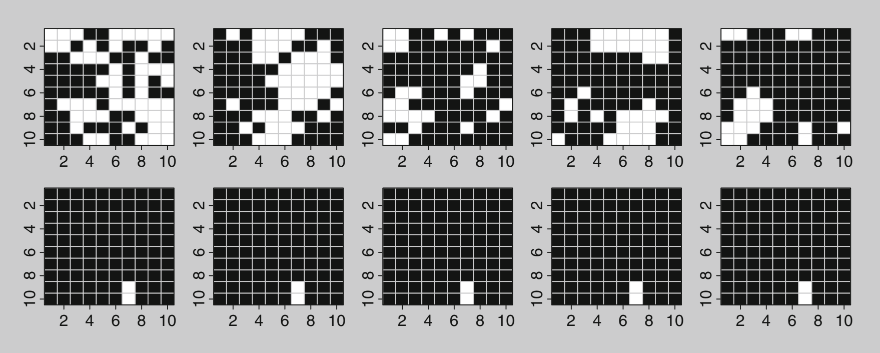


Fig. 49.9 The evolution of strategies A (*black*) and B (*white*) using the Conformist behavior and $b = 1.1$

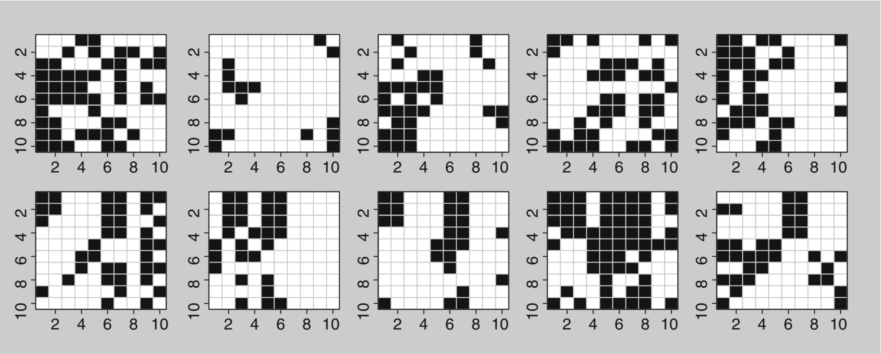


Fig. 49.10 The evolution of strategies A (*black*) and B (*white*) using the Conformist behavior and $b = 1.5$

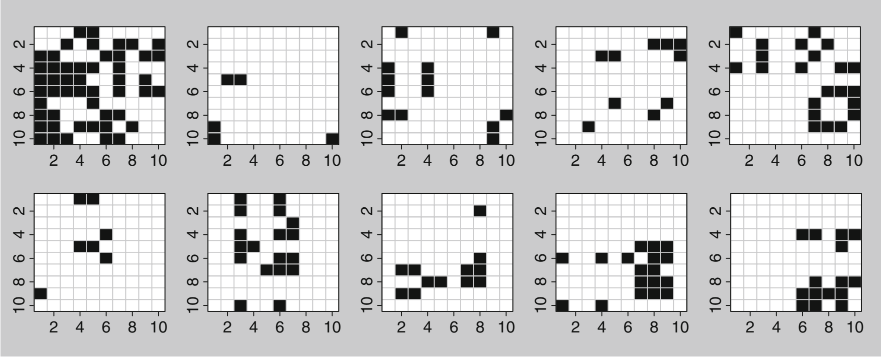


Fig. 49.11 The evolution of strategies A (*black*) and B (*white*) using the Conformist behavior and $b = 1.9$

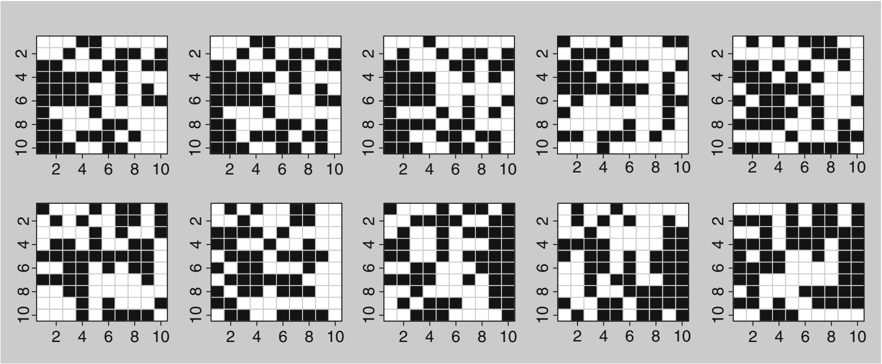


Fig. 49.12 The evolution of strategies A (*black*) and B (*white*) using the Pavlovian behavior and $b = 1.1$

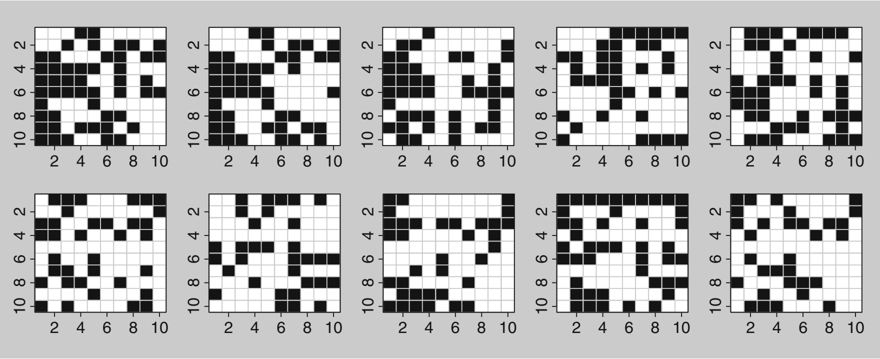


Fig. 49.13 The evolution of strategies A (*black*) and B (*white*) using the Pavlovian behavior and $b = 1.5$

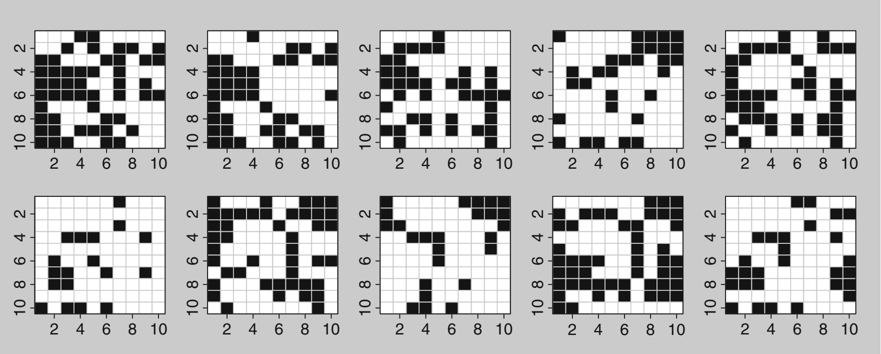


Fig. 49.14 The evolution of strategies A (*black*) and B (*white*) using the Pavlovian behavior and $b = 1.9$

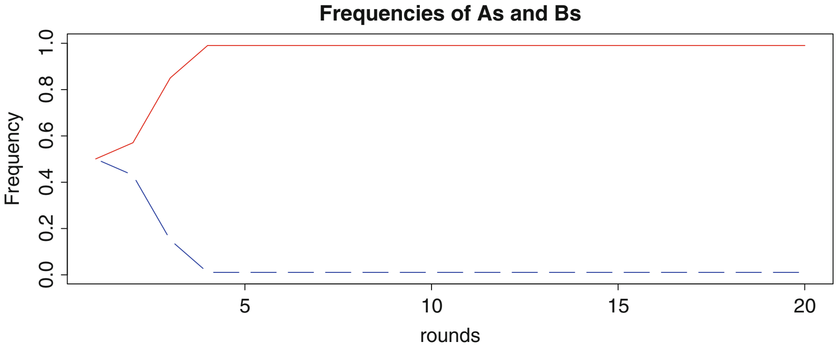


Fig. 49.15 The frequencies of strategies A and B for the Greedy behavior ($b = 1.1$)

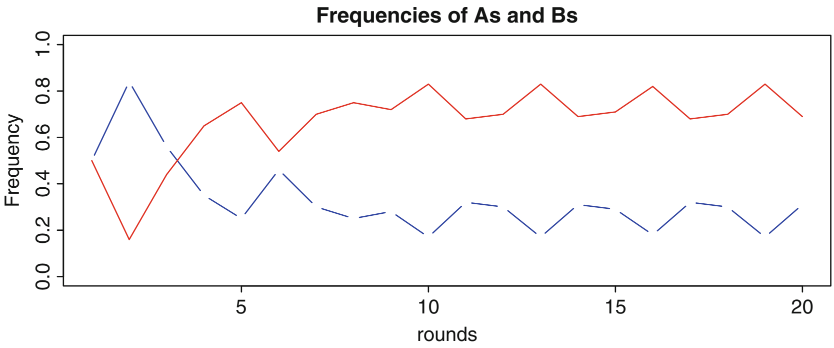


Fig. 49.16 The frequencies of strategies A and B for the Greedy behavior ($b = 1.5$)

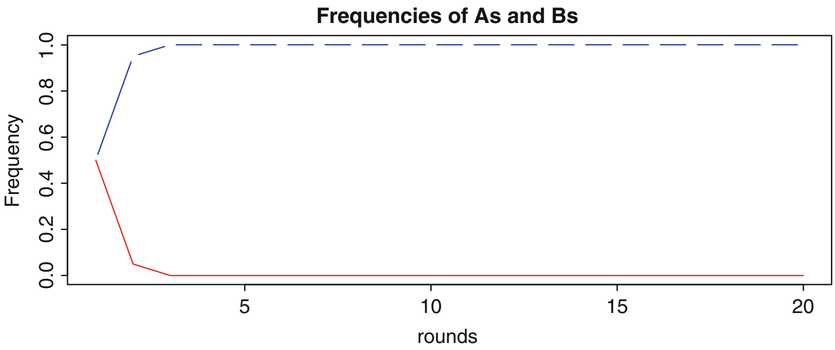


Fig. 49.17 The frequencies of strategies A and B for the Greedy behavior ($b = 1.9$)

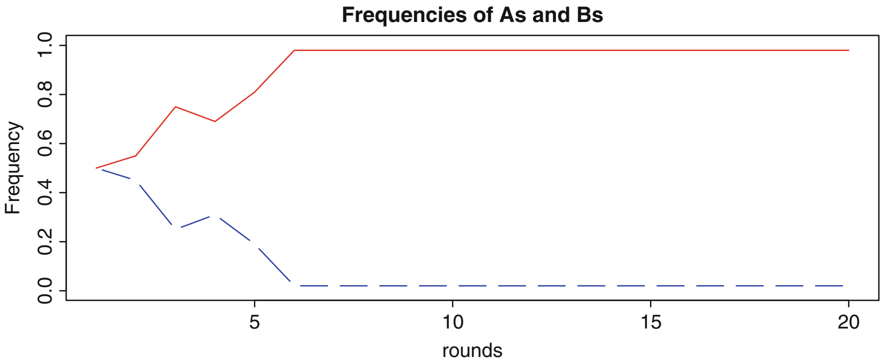


Fig. 49.18 The frequencies of strategies A and B for the Conformist behavior ($b = 1.1$)

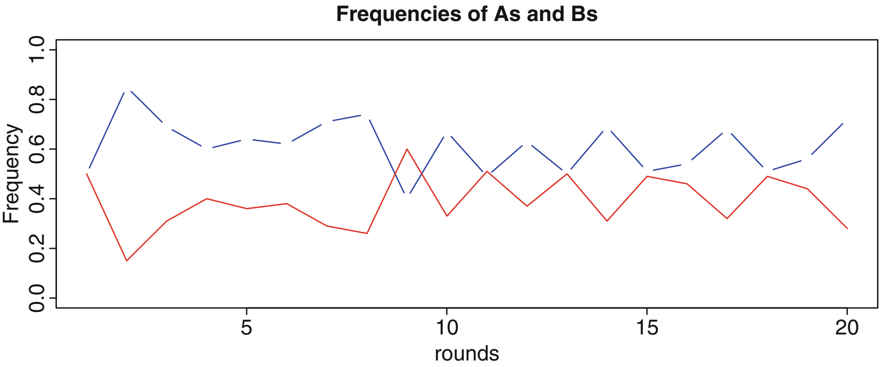


Fig. 49.19 The frequencies of strategies A and B for the Conformist behavior ($b = 1.5$)

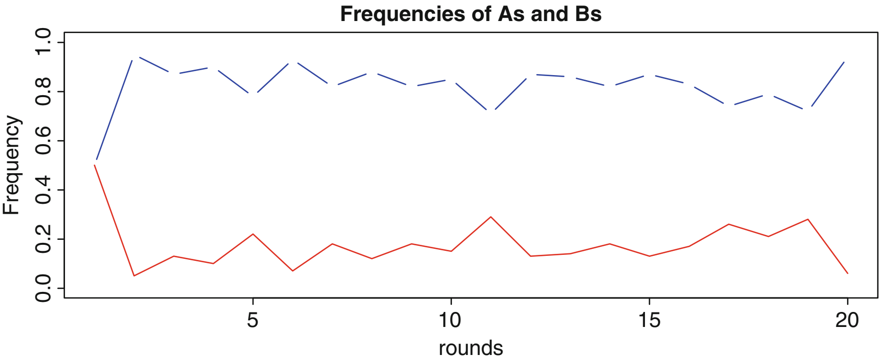


Fig. 49.20 The frequencies of strategies A and B for the Conformist behavior ($b = 1.9$)

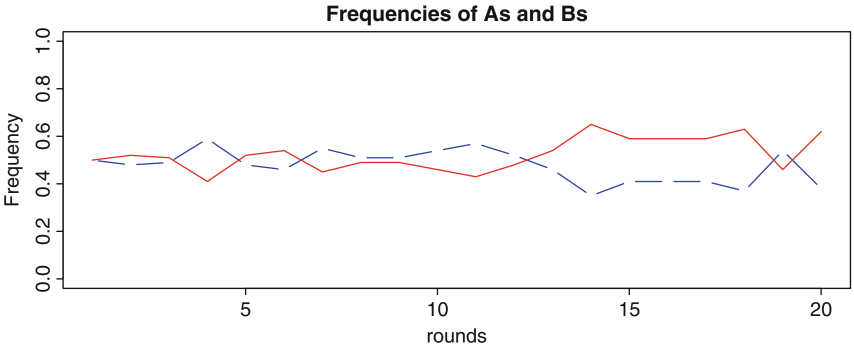


Fig. 49.21 The frequencies of strategies A and B for the Pavlovian behavior ($b = 1.1$)

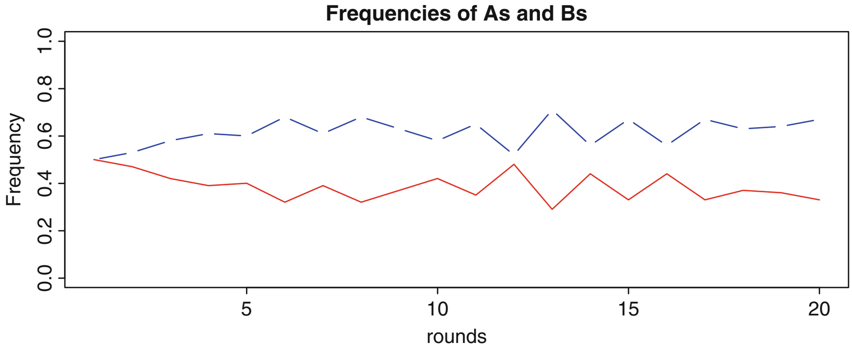


Fig. 49.22 The frequencies of strategies A and B for the Pavlovian behavior ($b = 1.5$)

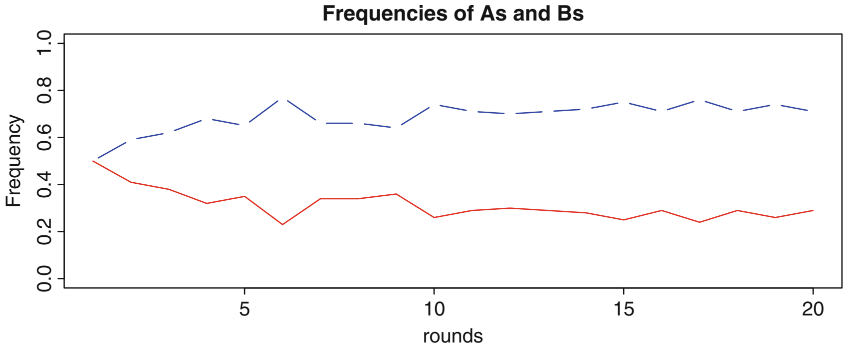


Fig. 49.23 The frequencies of strategies A and B for the Pavlovian behavior ($b = 1.9$)

49.4.2 Multi-Behaviors

In this section we investigate the situation where there are groups of companies with different behaviors and we check which group of behavior gets the highest payoff (wins the game). Our motivation comes from the fact that not all companies have the same behaviors, in reality. Even though, this is still a simplification of the reality that can give some insight into the competition between firms.

We created and tested the following scenarios:

- Scenario 1: The initial matrix is M1: there are 90% of companies using strategy B and 10% using strategy A
- Scenario 2: The initial matrix is M2: there are 90% of companies using strategy A and 10% using strategy B
- Scenario 3: The initial matrix is M3: there are 50/50% of companies using strategies A/B

For each scenario we change the value of b (1.2, 1.5 and 1.9) and create a game with three possible percentages of behaviors (Fig. 49.5):

- Balanced: similar percentage of behaviors (33% of greedy, 33% of conformist and 34% of Pavlovian)
- More greedy (90% of greedy, 5% of conformist and 5% of Pavlovian)
- More conformist (90% of conformist, 5% of greedy and 5% of Pavlovian)
- More Pavlovian (90% of Pavlovian, 5% of conformist and 5% of greedy)

The results can be seen in Table 49.5. This table shows the average payoffs of the sets of companies with one of the three behaviors. In scenario 1 and more greedy

Table 49.5 Average payoffs of the multi-behaviors situation

	Balanced	More Greedy	More Conformist	More Pavlovian
Scenario 1	($b = 1.2$)			
Pavlovian	127.259	13.054	135.182	91.978
Conformist	132.244	7.714	148.697	89.706
Greedy	126.233	7.979	157.482	91.162
	($b = 1.5$)			
Pavlovian	73.747	12.486	92.894	79.452
Conformist	77.639	8.558	90.121	90.752
Greedy	79.195	9.003	96.132	88.482
	($b = 1.9$)			
Pavlovian	50.476	12.806	12.630	81.573
Conformist	50.706	10.318	10.875	93.422
Greedy	60.863	10.896	10.872	92.566
Scenario 2	($b = 1.2$)			
Pavlovian	164.789	194.738	182.158	120.832
Conformist	171.726	192.414	183.389	121.764
Greedy	170.108	192.699	184.828	124.860
	($b = 1.5$)			
Pavlovian	79.289	139.104	112.756	111.431
Conformist	81.822	154.916	110.912	120.320
Greedy	82.152	150.755	110.558	115.842
	($b = 1.9$)			
Pavlovian	65.849	89.724	61.702	99.994
Conformist	71.692	127.056	69.413	122.212
Greedy	79.091	112.997	70.626	123.532
Scenario 3	($b = 1.2$)			
Pavlovian	125.418	175.840	159.234	113.844
Conformist	137.533	185.380	171.578	115.448
Greedy	130.555	173.732	169.214	118.768
	($b = 1.5$)			
Pavlovian	81.914	133.956	104.500	88.550
Conformist	86.944	143.686	104.207	100.012
Greedy	87.807	129.863	98.998	105.756
	($b = 1.9$)			
Pavlovian	57.749	17.238	51.898	88.642
Conformist	58.934	17.038	58.291	86.646
Greedy	68.102	16.101	60.890	116.872

environment we can see the advantage of the group of Pavlovians as their mean payoffs are always higher than for the other behaviors for all values of b. For the other possible combinations it is harder to make conclusions because the mean values fluctuates fairly often.

49.5 Data Analysis

According to Epstein [4], simulation is a particularly tool, when the aim is to establish that some set of micro assumptions is sufficient to generate a macro phenomenon of awareness. The behavior emerged from the complexity generated by the moving agents in our spatial geometry, is analyzed by implementing experiments and link analysis methods. Masuda and Aihara [15] found that for intermediate values of b, small-world architecture realizes a quasi-optimal behavior in the sense of rapid convergence to a good equilibrium. Here it is implicitly measured by a hierarchy of states.

49.5.1 Link Analysis

As Luo, Chakraborty, Sycara [14] in their research about the Prisoner's Dilemma game in a graph, we also use multiple types of agents. As them, we assume there are different types of agents forming the nodes of the graph. Prisoner's Dilemmas game in graphs with synchronized strategy update, is a game where the graph topology is assumed to be essential to analyze. So we decide to construct the network and observe the results using a special library present in R a package called "sna" – Social Network Analysis tool. We use an undirected graph $G = (V, E)$ to represent the agents of groups and their connections, where $V = v_i | i = 1, ..., n$ is a set of n nodes representing the set of n agents, and $E = (v_i, v_j) | i \neq j, i, j \in 1, ..., n$ represents a set of edges so that $(v_i, v_j) \in E$ if v_i and v_j are connected to each other (Fig. 49.21).

Szabó and Fáth [20] in evolutionary games on graphs define its evolutionary form when the interacting agents are linked in a specific social network, the core solution concepts and methods are very similar to those applied in non-equilibrium statistical physics. To also evaluate this situation, we produced simulations to compare the three b values (payoffs). In this section we build an adjacency matrix. Our aim is being able to according the strategy matrix, for each round, connect cooperating strategy agents. These connected agents form graph components. For this link analysis, we play 20 rounds, with 100 agents using the innovations strategies A and B. It is worth mentioning that in the evolution of spatial games, social preference has received negligible attention, according to Xianyu [23] although it has been accepted that the structure of agent interaction indeed plays an important role in a significant number of spatial games. In our study we are reverting this situation giving absolutely attention to social behaviors (Fig. 49.22).

49.5.2 The Graphs Payoff-Based Link Analysis

Jun and Sethi [9] visualize an interesting implication where there is a sense in which dense networks are more conducive to the evolution of cooperation than sparse networks. In our study we explore the graph density and according to the results for b = 1.1 (density = 0.06909091), for b = 1.5 (density = 0.01151515) and finally for b = 1.9 (density = 0.001616162). We also verify that the graph for b = 1.1 has higher density as cooperation strategy is used much more (Fig. 49.23).

We observe that for b = 1.9 we almost did not pick significant information because link analysis is stronger when cooperators are winning, which means lowest payoffs. We observe centrality positions in lowest b values. The stronger behavior from each strategy was furthermore reviewed finding strong components, seeking who is in what component, what are the component sizes, and which is the largest component. We observe that for higher b values (highest payoffs), most of all agents are in diverse components. When the payoffs are smaller (lower b values), the agents are in less components. We finally verify that the number of agents getting into the largest component increases with the decreasing of b value. That's because number of defecting agents is negatively affected by lowering the b value. (Figs. 49.24–49.30).

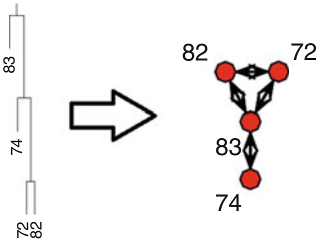


Fig. 49.24 Data example: Cluster Dendrogram and Graph for corresponding cooperator cluster, following Innovation Strategy A

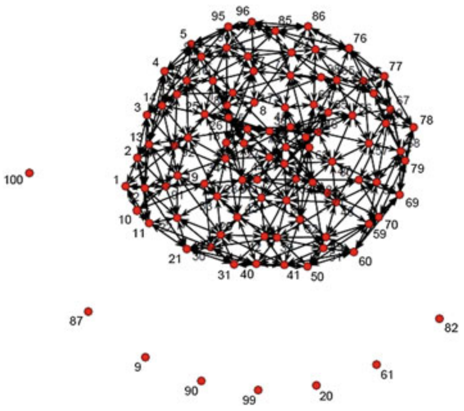


Fig. 49.25 Graph plot for $b = 1.1$ – network composed by 92 Cooperate Agents (Strategy A) and 8 Defect Agents (Strategy B)

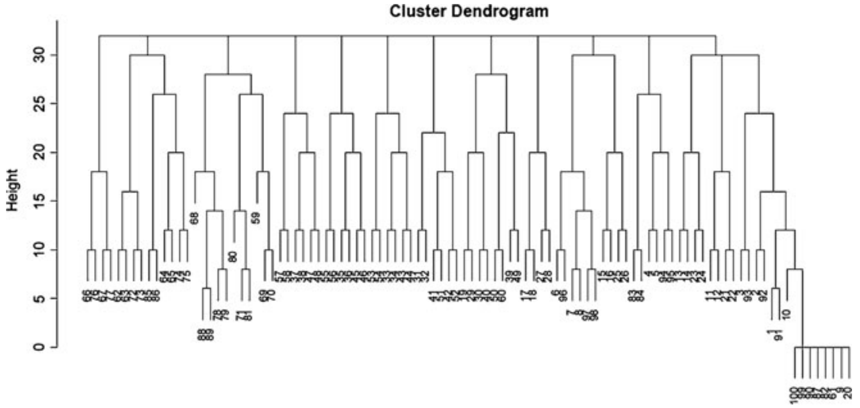


Fig. 49.26 Cluster Dendrogram for $b = 1.1$ – network composed by 92 Cooperate Agents (Strategy A) and 8 Defect Agents (Strategy B)

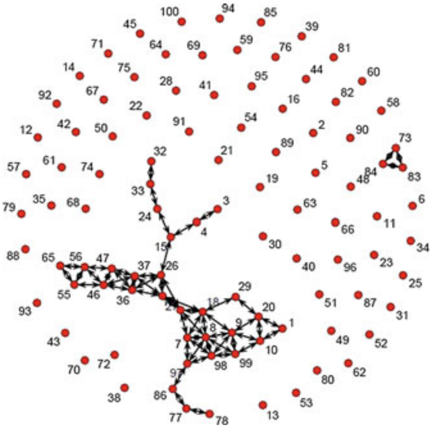


Fig. 49.27 Graph plot for $b = 1.5$ – network composed by 35 Cooperate Agents (Strategy A) and 65 Defect Agents (Strategy B)

Through link analysis we observe, like we did with statistics analysis, that the innovation Strategy A (cooperate) is dominant for lowest payoffs, and Strategy B (defect) dominant for high payoffs.

In Figs. 49.31–49.33 we see three different scenarios, each one having three different payoffs, where we observe (by link analysis) graphs representing different strategy matrices. We change the initial strategy A and strategy B portion.

Then in Fig. 49.34 we use the same initial strategy matrix, but agents use different behaviors all of them coexisting in the same playground. We change Greedy (G), Conformist (C) and Pavlovian (P) behaviors portion this time. The down-right plot shows the more balanced situation. The up-right filled with almost all cooperators, shows the more Conformist, the up-left more Greedy and the down-left more Pavlovian situations.

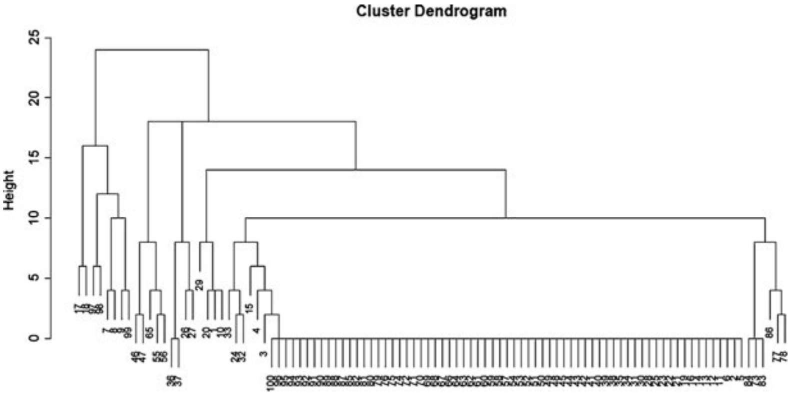
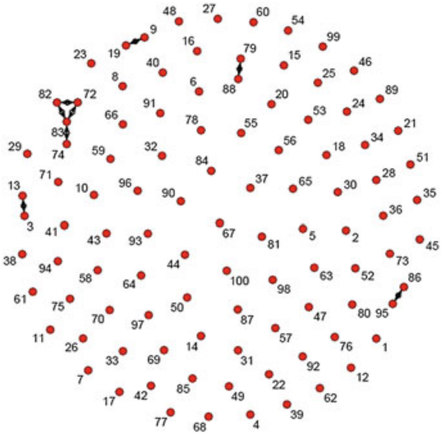


Fig. 49.28 Cluster Dendrogram for $b = 1.5$ – network composed by 35 Cooperate Agents (Strategy A) and 65 Defect Agents (Strategy B)

Fig. 49.29 Graph plot for $b = 1.9$ – network composed by 12 Cooperate Agents (Strategy A) and 88 Defect Agents (Strategy B)



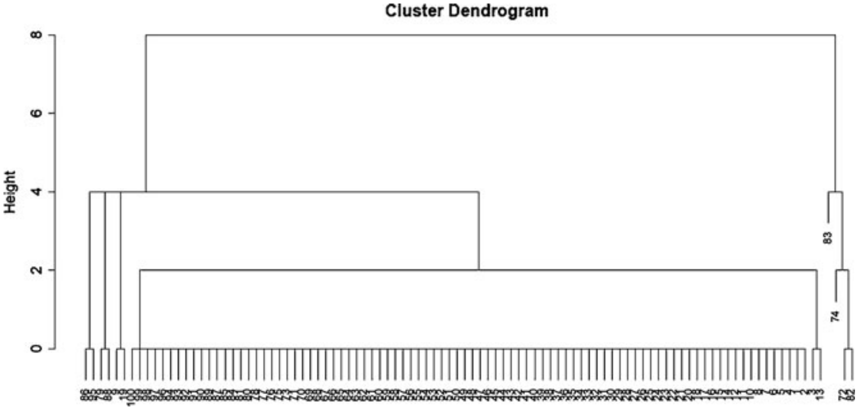


Fig. 49.30 Cluster Dendrogram for $b = 1.9$ – network composed by 12 Cooperate Agents (Strategy A) and 88 Defect Agents (Strategy B)

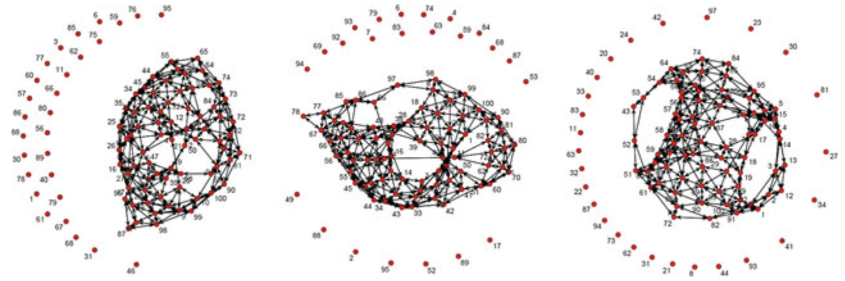


Fig. 49.31 Same behaviors, 20 rounds, payoff = 1.2. Different initial strategy matrices (90% - A, 10% - B); (10% - A, 90% - B); (50% - A, 50% - B)

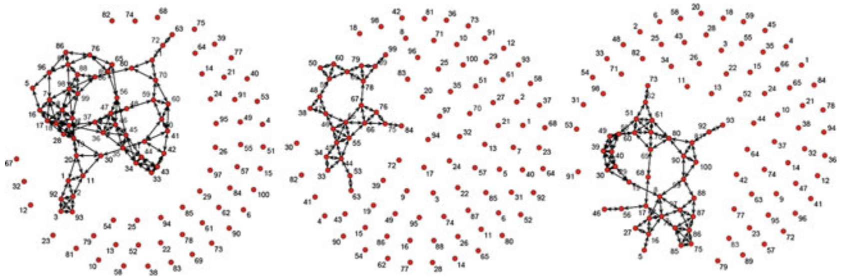


Fig. 49.32 Same behaviors, 20 rounds, payoff = 1.5. Different initial strategy matrices (90% - A, 10% - B); (10% - A, 90% - B); (50% - A, 50% - B)

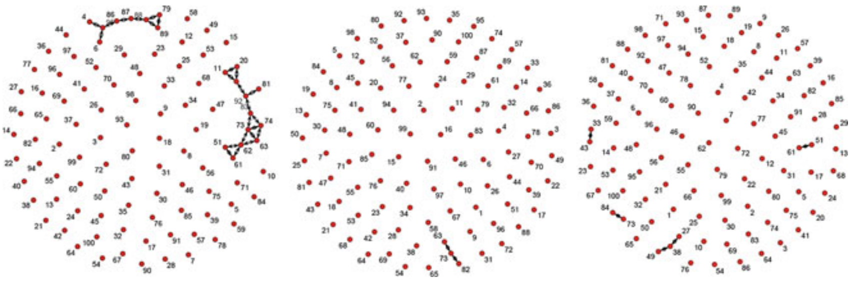


Fig. 49.33 Same behaviors, 20 rounds, payoff = 1.9. Different initial strategy matrices (90% - A, 10% - B); (10% - A, 90% - B); (50% - A, 50% - B)

In this paper our main concern is to analyze how the graph structure of interactions can modify and enrich the representation of behavioral patterns emerging in evolutionary games. One of our innovation strategies leading this research was based on using social network analysis, to evaluate b values, not just statistically, as it usually appears in most of the Prisoner's Dilemma studies.

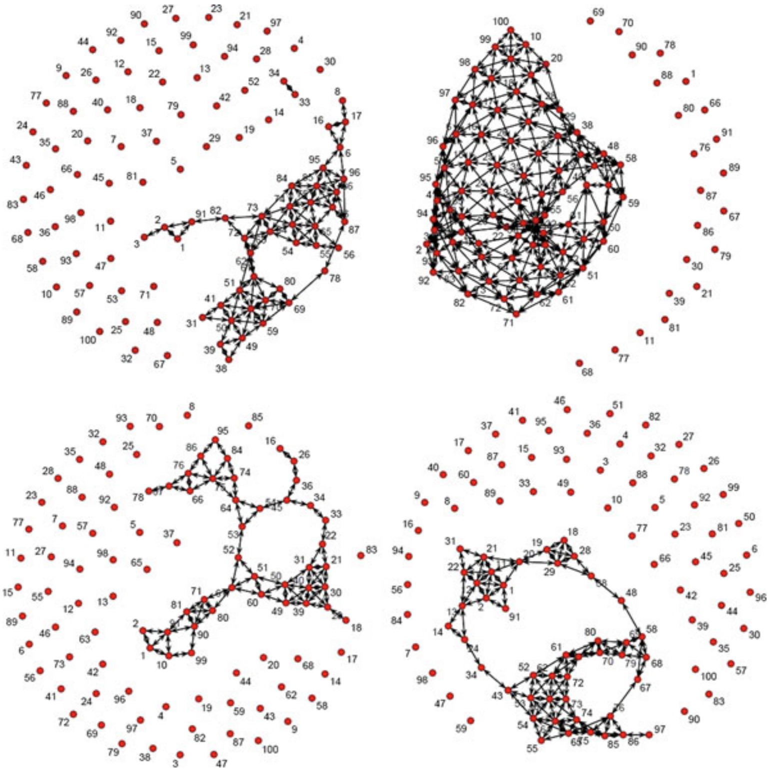


Fig. 49.34 Same initial strategy matrix, 20 rounds, payoff = 1.5. Different behaviors: (90% - G, 5% - C, 5% - P); (5% - G, 90% - C, 5% - P); (5% - G, 5% - C, 90% - P);(33% - G, 33% - C, 34% -P)

49.6 Dynamics GIS

Following the new Essays on Geography and GIS on GIS Best Practices, ESRI article [1], we felt the need to develop a GIS tool which allows us to acquire information from a base map of real-world locations. Bringing us the possibility to combine different data sets performed dynamically, we thought about a space-time combination which could provide a spatiotemporal understand and the possibility to predict reality. Dynamics GIS described by May Yuan [1] seemed the perfect tool to achieve our aim once we need to perform knowledgeable decisions, and analyze adaptation strategies for this dynamic world multi-agent-based.

Using especially agent-based modeling, with spatiotemporal analysis we became able to extract spatiotemporal information, survey clusters and change detection. Knowing that the complexity of reality can be understood by analyzing spatial and social organizations, by representing an dynamics GIS we can understand the structure connecting subsystems and supersystems, according to May Yuan [1] and also detecting existent hierarchies, like we did in Sect. 49.5.1 with link analysis methods, and now with a spatiotemporal representation.

We build a system, so we can be able to obtain georeferenced information about innovation strategies A and B. To do so, we developed a methodology using a cartographic projection designed to create a coordinate system, since we need to represent objects in space. We start modeling our data because we want to link different datasets together by the fact they relate to specific and also different geographic locations. We set one of the flat representation system used in Portuguese mapping, Hayford-Gauss System. We use the datum for the planimetric geodetic network based on the International Hayford ellipsoid parameters: $a = 6378388$ (semi-major axis), $f = 1/297$ (flattening), $e^2 = 0.00672267002233$ (squared eccentricity). The ellipsoid is positioned at the Central Meridian, $\lambda = 9°07'54.862''W$ (geodetic longitude). We use a file containing the geodetic latitude and longitude of the geodetic border points in Portugal. The longitude of the points was adjusted to the Central Meridian. We use Transverse Mercator Projection (also named as Gauss Projection), is a conformal projection, where the lengths remains along a meridian called the Central Meridian of projection, the origin of the ordinate is at the equator, the origin of abscises is at Central Meridian. It can be defined by the following equations:

In the Central Meridian,

$$\begin{cases} x = 0 \\ y = S\varphi = \int\limits_0^{\varphi} \rho d\varphi \end{cases} \tag{49.23}$$

$$\rho = \frac{a(1-e^2)}{\sqrt{(1-e^2 sin^2\varphi)^3}} \tag{49.24}$$

Formulas for direct transformation of Transverse Mercator projection,

$$S\varphi = a[A_0\varphi - A_2 \sin 2\varphi + A_4 \sin 4\varphi + ...] \tag{49.25}$$

Obtaining the different A values by the following equation,

$$\begin{aligned} A_0 &= 1 - \frac{1}{4}e^2 - \frac{3}{64}e^4 + ... \\ A_2 &= \frac{3}{8}(e^2 + \frac{1}{4}e^4 + ...) \\ A_4 &= \frac{15}{256}(e^4 + ... \end{aligned} \tag{49.26}$$

$$\begin{cases} x = N \cos\varphi * \lambda \\ y = S\varphi + \frac{1}{2}N \cos\varphi * \sin\varphi * \lambda^2 \end{cases} \tag{49.27}$$

Where φ is the Geodetic Latitude, λ Geodetic Longitude, a semi-major axis of the ellipsoid, e eccentricity of the ellipsoid, N 1st Vertical Curvature Radius, ρ Meridian Curvature Radius and finally x and y our agents position according to all this parameters previously defined.

49.6.1 Dynamics Business Parks

Using the equations defined in the Sect. 49.6 we build the algorithm in *Matlab* to perform our dynamics GIS. Modeling geographic dynamics, we obtain each agent or company georeferenced information in dynamic reality, we now know, who is playing, which strategy behavior, where and when.

We set three different Business Parks: North Park, Centre Park and South Park. They are fictitious (not existent business parks) but referred to existent (real-world) positions. Three Business Parks, with 100 agents each, each agent with a calculated (referenced-based) position, based on the parameters establish on the previous section (Fig. 49.35).

In our research the agent or company is located always in the same position, as we established on Sect. 49.2.2 but each agent changes his states according to game dynamics related to psychological geography (and geometry, in the context of the discrete model cellular automata), deeply well-marked in this single and multi-behavior spatial prisoner's dilemma dynamics (Fig. 49.36).

This technique allows us to survey in a long run which is the most used innovation strategy per company, per business park and per country. Scanning in any direction, performing zoom in and out, agents or companies positions can be observed looking into the axis, or by numbers in output coordinates (Fig. 49.37).

With this dynamics GIS tool and sample analysis we are suggesting a spatiotemporal Business Park classification, this temporal classification plus temporal agents or companies gains, would provide significant information to the construction of some kind of Business Park Dynamics Rating System. If globally applied

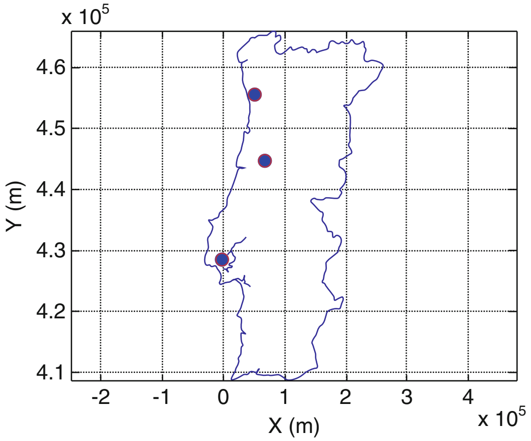


Fig. 49.35 Portugal Transverse Mercator Projection. Three Dynamics Business Parks playing spatial prisoner's dilemma innovation strategies A and B, with single and multi-behaviors. North Business Park, Centre Business Park, South Business Park

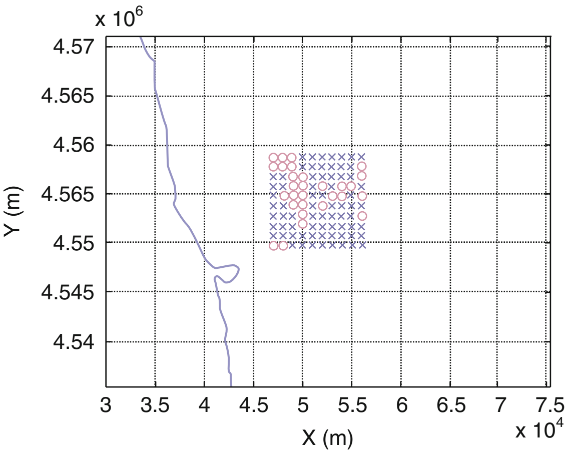


Fig. 49.36 North Business Park, playing spatial prisoner's dilemma innovation strategies A (x) and B (o), with more Pavlovian behavior

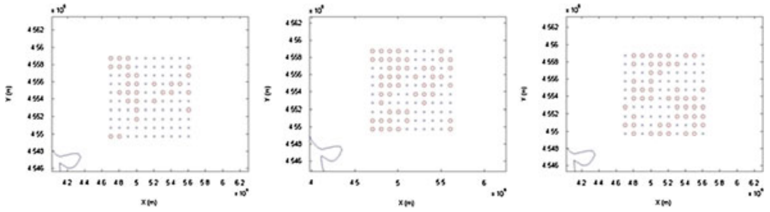


Fig. 49.37 North Business Park playing innovation strategies A (x) and B (o) with more Pavlovian behavior. At time units, *left*: $t = 1$, (strategy A wins), *middle*: $t = 2$ (strategy B wins), *right*: $t = 3$ (no winning strategy, it's balanced). During this three time units analysis we can classify North Business Park as following innovation strategy A and B, a balanced strategy behavior

innovation strategy A or B analyzing who is obtaining bigger gains, not just a payoff-based analysis but also a social analysis we can rate the country following the winning strategy with higher global results, basing our full-model analysis in human growth contribution, welfare and development. Obviously much more parameters enter in this kind of rating; we are just contributing with a small step. Since change and movement are two essential elements in temporal GIS, in this research we didn't define the time topology, as Giacomo Bonanno [3] did, when describing perfect information games time, agents predictions in branching time (also referred by May Yuan [1] on GIS Best Practices), perhaps when we build the future work, in Sect. 49.8, we can explore temporal definitions, associating to each agent or company prediction strategies linked to their spatiotemporal definitions (Fig. 49.38).

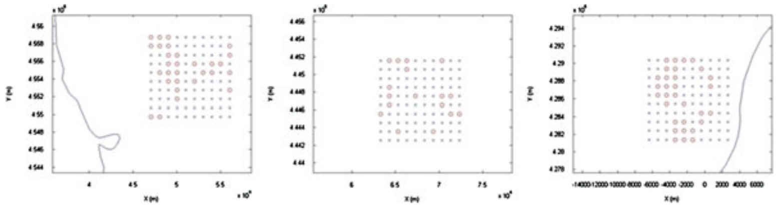


Fig. 49.38 At time unit $t = 1$, each business park is using one kind of behavior when playing innovation strategies A (x) and B (o). *left*: North Business Park (more Pavlovian behavior – strategy A wins), middle: Centre Business Park (more greedy behavior – strategy A wins), right: South Business Park (more conformist behavior – strategy A wins). During this one time unit analysis we can classify Portuguese Business Parks as following innovation strategy A.

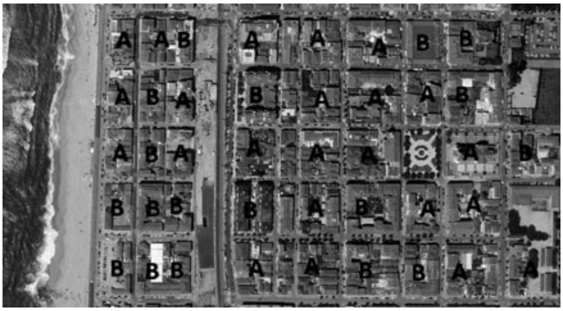


Fig. 49.39 Illustrative fictitious Business Park playing Innovation Strategies A and B. Image font: Google Earth, Espinho, Portugal.

49.6.2 Another Real: World Application

Robert Axelrod and John Holland, mentioned by Hoffmann and Waring [7] as responsible for important discoveries such as the Tit-For-Tat strategy in Prisoners' dilemma (Axelrod) or the genetic algorithms (Holland), consider social tagging as a means of influencing game outcome. In the context of the social sciences, the interactions and learning simulations take place within a neighborhood of players. Neighborhoods can be formed geographically, or in a more abstract sense, such as in the range of partners available for trading.

So we thought in a Real – World Application for our study, based in many few companies clustered in parks, business parks. These parks for many reasons need to be rated and we thought that some kind of rating system should exist. Each company belonging to a park, plays Spatial Prisoner's Dilemma, which means: we could classify the parks with the classification following innovation Strategy A or Strategy B, like we did to each company or agent in the paper.

There are a lot of parameters entering in parks classification based in some theories; we believe our research may well contribute in some way to a Real-World Business Park Dynamics Rating System.

49.7 Discussion

In this work, the experiments are presented in two different ways depending on the type of behavior: single behavior and multi-behavior. In the single behavior situation, we noticed that in general, the average payoff of all companies decreases as we increase the value of b. It is also possible to conclude that Pavlovian behavior scores better then Greedy and Conformist when b is higher. Additionally we can see that when we have more companies with strategy A (small b) we get higher average payoffs in all behaviors. The maximum individual payoff is also higher for small values of b when companies use behavior Greedy or Conformist. In the multi-behavior situation (different percentages of Pavlovian, Greedy and Conformists are defined), and in the scenario 1 (90% of companies use strategy B and 10% use strategy A) in the more greedy environment, we state the Pavlovian behavior dominates for all values of b.

Results of link analysis show that for greater values of b, the density of the network decreases. In addition, it is possible to see that Strategy A is dominant for lower payoffs (defined by the b values) and strategy is dominant for higher payoffs.

In some researches, players in prisoner's dilemma are modeled as learning automaton; such models can be viewed as complicated variants of reinforcement learners. However, following Wakano and Yamamura [21], in our research we were also interested in the behavior of the simplest form of reinforcement learner. During the study we start asking ourselves if Pavlovian would be more efficient than the Conformist, or the simplest one, the Greedy. Our result shows that agents with a simple and instinctively familiar learning rule put up a very efficient and tough cooperation in some situations. At the same time, Kraines and Kraines [13] prove that a society of fast adapting agents may experience conflict and disagreement while another society with slower learning agents will benefit in cooperation. As Alexander Pope was mentioned in Kraines and Kraines [13] leaving the following thought "a little learning is a valuable thing and it is too much learning that is dangerous"–we feel like being in a great adventure, so we will just keep walking into knowledge flow.

49.8 Future Work

Collect geo-spatial data in Prisoner's Dilemma theory game and build real models based on geo-referenced images representing real world situations, using Geographic Information Systems tools. Once the dynamics is implemented, the major concern is the long run behavior of the system: fixed points, cycles, and their stability, chaos, other parameters, and the connection between static concept as Nash equilibrium or evolutionary stability, and dynamic predictions. According to Wolfram [22] even when the general problem is undecidable, the emergence of particular finite sequences in the limit set for a cellular automaton may be decidable. We need revolutionary new concepts and methods for the categorization of the rising complex and self-organizing patterns. One could be the development of dynamics GIS using Multi-Agent Systems with dynamics predictions.

References

1. Yuan. M., Chrisman, N., "Jack" Owens, J.B., Dobson, J.E., Goodchild, M.F., Moll, G., Gallis, M., Millar, H., Getis, A., Wrigth, D.J.: GIS Best Practices. Essays on Geography and GIS. ESRI, Sept (2008)
2. Axelrod, R.: The Evolution of Cooperation. Basic Books, New York (1984)
3. Bonanno, G.: Branching time, perfect information games and backward induction. Department of Economics, University of California, Nov (1999)
4. Epstein, J.M.: Zones of cooperation in demographic prisoner's dilemma. Complexity **4**, 36–48 (1996)
5. Epstein, J.M.: Zones of cooperation in demographic prisoner's dilemma. Working Papers 97-12-094, Santa Fe Institute, Dec (1997)
6. Tucker, A.W., Kuhn, H.W.: Contributions to the Theory of Games I. Annals of Mathematics Studies, no. 24. Princeton University Press, Princeton (1950)
7. Hoffmann, R., Waring, N.: The localisation of interaction and learning in the repeated prisoner's dilemma. Working Papers 96-08-064, Santa Fe Institute, Aug (1996)
8. Isaac, A.G.: Simulating evolutionary games: a python-based introduction. J. Artif. Soc. Soc. Simul. **11**(3), 8 (2008)
9. Jun, T., Sethi, R.: Neighborhood structure and the evolution of cooperation. J. Evol. Econ. **17**, 623–646 (2007)
10. Kirchkamp, O.: Spatial evolution of automata in the prisoner's dilemma. In: Social Science Microsimulation [Dagstuhl Seminar, May, 1995], pp. 307–358, London, UK. Springer, Berlin (1996)
11. Kirchkamp, O.: Spatial evolution of automata in the prisoners' dilemma. J. Econ. Behav. Organ. **43**(2), 239–262 (2000)
12. Kraines, D., Kraines, V.: Learning to cooperate with pavlov an adaptive strategy for the iterated prisoner's dilemma with noise. J. Theory Decis. **35**(2), 107–150 (1993)
13. Kraines, D., Kraines, V.: The threshold of cooperation among adaptive agents: Pavlov and the stag hunt. In: ECAI '96: Proceedings of the Workshop on Intelligent Agents III, Agent Theories, Architectures, and Languages, pp. 219–231, London, UK. Springer, Berlin (1997)
14. Sycara, K., Luo, L., Chakraborty, N.: Prisoners dilemma on graphs with heterogeneous agents (2009)
15. Masuda, N., Aihara, K.: Spatial prisoner's dilemma optimally played in small-world networks. Phys. Lett. **13**, 55–61 (2003)
16. Nowak, M.A.: Evolutionary Dynamics: Exploring the Equations of Life. Harvard University Press, Cambridge, MA (2006)
17. Nowak, M.A., May, R.M.: The Spatial Dilemmas of Evolution. University of Oxford, Oxford (1992)
18. Power, C.: A spatial agent-based model of n-person prisoner's dilemma cooperation in a socio-geographic community. J. Artif. Soc. Soc. Simul. **12**, 1 (2009)
19. R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, (2009). ISBN 3-900051-07-0.
20. Szabó, G., Fáth, G.: Evolutionary Games on Graphs. Physics Reports (2007)
21. Wakano, J.Y., Yamamura, N.: A simple learning strategy that realizes robust cooperation better than pavlov in iterated prisoners' dilemma. J. Ethol. **19**, 1–8 (2001)
22. Wolfram, S.: Computation Theory of Cellular Automata. Princeton University Press, Princeton (1984)
23. Xianyu, B.: Social preference, incomplete information, and the evolution of ultimatum game in the small world networks: An agent-based approach. J. Artif. Soc. Soc. Simul. **13**, 2 (2010)

24. Wang, X.-L., Sheng, Z.-H., Hou, Y.-Z., Du, J.-G.: The Evolution of Cooperation with Memory, Learning, and Dynamic Preferential Selection in Spatial Prisoner's Dilemma Game. International Symposium on Nonlinear Dynamics. J. Phys. Conf. Ser. **96**, 1–6 (2007)
25. Zimmermann, M.G., Eguluz, V.M.: Cooperation, social networks, and the emergence of leadership in a prisoner's dilemma with adaptive local interactions. Phys. Rev. E **72**(5), 056118 (2005)

Chapter 50
Euclidean Jordan Algebras and Strongly Regular Graphs

Luís Vieira

Abstract We analyze the spectra of strongly regular graphs in the environment of Euclidean Jordan algebras. In particular we obtain the spectra of the strongly regular graphs constructed in the Euclidean Jordan algebra studied in Cardoso and Vieira (J Math Sci 120:881–894, 2004) recurring to homogeneous linear difference equations of second order with constant coefficients. Next, we associate a three dimensional Euclidean Jordan algebra V to the adjacency matrix of a strongly regular graph τ with three distinct eigenvalues and we define the generalized Krein parameters of τ. Finally, we establish necessary conditions for the existence of a strongly regular graph.

50.1 Euclidean Jordan Algebras

Euclidean Jordan algebras are more and more used in the various branches of Mathematics. For instance, we may cite the application of this theory to interior point methods [3, 4], to statistics [7] and to combinatorics [1].

A deep study on Euclidean Jordan algebras can be found in Koecher's lecture notes [6] and in the monograph by Faraut and Korányi [2]. Herein, we only present the results of Euclidean Jordan algebras more often used in this work.

Consider a finite dimensional real algebra $\mathscr{V}$ with the bilinear mapping $(x, y) \mapsto x \cdot y$. $\mathscr{V}$ is a real power associative algebra if for all x in $\mathscr{V}$, $(x \cdot x) \cdot x = x \cdot (x \cdot x)$. Let $\mathscr{V}$ be a $n-$dimensional real power associative algebra with unit $\mathbf{e}$. For x in $\mathscr{V}$ the rank of x is the least natural number k such that $\{\mathbf{e}, x, \ldots, x^k\}$ is linearly dependent and we write $\mathrm{rank}(x) = k$. Since $\mathrm{rank}(x) \leq n$ we define the rank of $\mathscr{V}$ as being the natural number $\mathrm{rank}(\mathscr{V}) = \max\{\mathrm{rank}(x) : x \in \mathscr{V}\}$. An element x in $\mathscr{V}$ is regular if $\mathrm{rank}(x) = \mathrm{rank}(\mathscr{V})$. The set of regular elements of $\mathscr{V}$ is open and dense in $\mathscr{V}$.

L. Vieira
Center of Mathematics of University of Porto, Faculty of Engineering, University of Porto, Portugal
e-mail: lvieira@fe.up.pt

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_50,

Let x be a regular element of $\mathscr{V}$ and $r = \operatorname{rank}(x)$. Then, there exist real numbers $a_1(x), a_2(x), \ldots, a_{r-1}(x)$ and $a_r(x)$ such that

$$x^r - a_1(x)x^{r-1} + \cdots + (-1)^r a_r(x)\mathbf{e} = 0, \tag{50.1}$$

where 0 is the null vector of $\mathscr{V}$. Taking in account (50.1) the polynomial (50.2)

$$p(x, \lambda) = \lambda^r - a_1(x)\lambda^{r-1} + \cdots + (-1)^r a_r(x) \tag{50.2}$$

is called the characteristic polynomial of x. Each coefficient a_i of the characteristic polynomial of x is a homogeneous polynomial of degree i in the coordinates of x in a fixed basis of $\mathscr{V}$. Since the set of regular elements of $\mathscr{V}$ is dense in $\mathscr{V}$ and since each polynomial a_i is a homogeneous polynomial of degree i then the definition of characteristic polynomial is extensible to every element in $\mathscr{V}$. We call the roots of the characteristic polynomial of x the eigenvalues of x. Consider x in $\mathscr{V}$. The coefficients $a_1(x)$ and $a_r(x)$ of the polynomial (50.2) are called the trace of x and the determinant of x respectively. The notation for the determinant and for the trace of x is $\operatorname{tr}(x)$ and $\det(x)$ respectively.

Example 50.1. $E = \mathbb{R}^n$ endowed with the binary application

$$(x_1, x_2, \ldots, x_n) \cdot (y_1, y_2, \ldots, y_n) = (x_1y_1, x_2y_2, \ldots, x_ny_n)$$

is a n-dimensional real power associative algebra with the unit element $\overline{\mathbb{e}} = (1, 1, \ldots, 1)$. Let $\overline{x} = (x_1, x_2, \ldots, x_n)$ be an element of E. Then, $\overline{x}^k = (x_1^k, x_2^k, \ldots, x_n^k)$. Since

$$\triangle_n(\overline{x}) = \begin{vmatrix} 1 & x_1 & \cdots & x_1^{n-1} \\ 1 & x_2 & \cdots & x_2^{n-1} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_n & \cdots & x_n^{n-1} \end{vmatrix} = \Pi_{i>j}(x_i - x_j)$$

then the set $\{\overline{e}, \overline{x}, \ldots, \overline{x}^{n-1}\}$ is linear independent if and only if the x_i's for $i = 1, \ldots, n$ are distinct. So $\overline{x}$ is a regular element of E if and only if the x_i's are all distinct and therefore we can conclude that $\operatorname{rank}(E) = n$. Consider a regular element $\overline{x} = (x_1, x_2, \ldots, x_n)$ of E and p the characteristic polynomial of $\overline{x}$.

$$p(\overline{x}, \lambda) = \lambda^n - a_1(\overline{x})\lambda^{n-1} + \cdots + (-1)^n a_n(\overline{x})$$

Since $p(\overline{x}, \overline{x}) = \overline{0}$ then $\overline{x}^n - a_1(\overline{x})\overline{x}^{n-1} + \cdots + (-1)^n a_n(\overline{x})\overline{e} = \overline{0}$, where $\overline{0} = (0, 0, \ldots, 0)$. Therefore, $p(\overline{x}, x_i) = 0$ for $i = 1, \cdots, n$. Because $\overline{x}$ is a regular element of E then $\triangle_n(\overline{x}) \neq 0$ and so $p(\overline{x}, \lambda) = \prod_{i=1}^n(\lambda - x_i)$. That is,

$$p(\overline{x},\lambda) = \lambda^n - \sum_{i=1}^{n} x_i \lambda^{n-1} + \cdots + (-1)^n \prod_{i=1}^{n} x_i. \tag{50.3}$$

So taking in account (50.3) we conclude that $\mathrm{tr}(\overline{x}) = \sum_{i=1}^{n} x_i$ and $\det(\overline{x}) = \prod_{i=1}^{n} x_i$.

Example 50.2. The real vector space of real symmetric matrices of order n, $E = \mathrm{Sym}(n,\mathbb{R})$, equipped with the bilinear map $x \circ y = (xy + yx)/2$ is a $(n^2+n)/2$-dimensional real power associative algebra whose unit is $\mathbf{e} = I_n = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$.

An element x of E is regular if and only if x has n distinct eigenvalues $\lambda_1, \ldots, \lambda_{n-1}$ and λ_n and the characteristic polynomial of x is $p(x,\lambda) = \prod_{i=1}^{n}(\lambda - \lambda_i)$. That is

$$p(x,\lambda) = \lambda^n - \sum_{i=1}^{n} \lambda_i \lambda^{n-1} + \cdots + (-1)^n \prod_{i=1}^{n} \lambda_i. \tag{50.4}$$

Then, analyzing the expression of the characteristic polynomial (50.4), it follows that $\mathrm{tr}(x) = \sum_{i=1}^{n} \lambda_i$ and $\det(x) = \prod_{i=1}^{n} \lambda_i$.

Let $\mathscr{V}$ be a finite dimensional real power associative algebra with a unit element $\mathbf{e}$, x a regular element of $\mathscr{V}$ and $r = \mathrm{rank}(\mathscr{V})$. $\mathbb{R}[x]$ denotes the subalgebra of $\mathscr{V}$ spanned by $\mathbf{e}$ and x. $\mathscr{L}(x)$ is the linear endomorphism on $\mathscr{V}$ defined by $\mathscr{L}(x)y = x \cdot y$ for every y in $\mathscr{V}$. We call $\mathscr{L}_0(x)$ to the restriction of $\mathscr{L}(x)$ on $\mathbb{R}[x]$. We now show that $p(x,\lambda) = |\lambda I - L_0(x)|$. Consider the basis $\mathscr{B} = \{\mathbf{e}, x, \ldots, x^{r-1}\}$ of $\mathbb{R}[x]$. Since

$$\begin{aligned} L_0(x)\mathbf{e} &= 1x \\ L_0(x)x &= x^2 \\ &\vdots \\ L_0(x)x^{r-2} &= x^{r-1} \\ L_0(x)x^{r-1} &= a_1(x)x^{r-1} + \cdots - (-1)^r a_r(x)\mathbf{e} \\ &= a_1(x)x^{r-1} + \cdots + (-1)^{r-1} a_r(x)\mathbf{e} \end{aligned}$$

then

$$M(\lambda I - L_0(x);\mathscr{B}) = \begin{pmatrix} \lambda & 0 & \cdots & 0 & (-1)^r a_r(x) \\ -1 & \lambda & \cdots & 0 & (-1)^{r-1} a_{r-1}(x) \\ 0 & -1 & \cdots & 0 & (-1)^{r-2} a_{r-2}(x) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & \lambda - a_1(x) \end{pmatrix}. \tag{50.5}$$

So

$$|\lambda I - L_0(x)| = \lambda^r - a_1(x)\lambda^{r-1} + \cdots + (-1)^r a_r(x).$$

Therefore, it is natural to call the polynomial (50.2) the characteristic polynomial of x.

Let $\mathscr{V}$ be a real vector space with the bilinear map $(x, y) \mapsto x \cdot y$. Then $\mathscr{V}$ is a Jordan algebra if for all x and y in $\mathscr{V}$

(a)$x \cdot y = y \cdot x$
(b) $x \cdot (x^2 \cdot y) = x^2 \cdot (x \cdot y)$

where $x^2 = x \cdot x$.

We call $\mathbf{e}$ the unit of the Jordan algebra $\mathscr{V}$ if for all x in $\mathscr{V}$, $x \cdot \mathbf{e} = \mathbf{e} \cdot x = x$.

Remark 50.1. Let E be a finite dimensional real associative algebra. We introduce on E a structure of Jordan algebra by considering a new product $\circ$ defined by $x \circ y = (x \cdot y + y \cdot x)/2$ for all x and y in E.

Example 50.3. The real vector space $\mathrm{Sym}(n, \mathbb{R})$ is a Jordan algebra when endowed with the bilinear map $\circ$ given by $x \circ y = (xy + yx)/2$ for all x and y in E, where xy is the usual matrix multiplication of x and y.

From now on, a Jordan algebra $\mathscr{V}$ is always a finite dimensional real algebra. A Jordan algebra $\mathscr{V}$ with unit is always power associative.

A Euclidean Jordan algebra $\mathscr{V}$ is a Jordan algebra with unit and with an inner product $< \cdot, \cdot >$ such that

$$< x \cdot y, z > = < y, x \cdot z > \tag{50.6}$$

for all x, y and z in $\mathscr{V}$.

Example 50.4. Let $E = \mathbb{R}^n$. Then E is a Euclidean Jordan algebra when we consider on E the bilinear map $\circ$ given by $(x_1, x_2, \ldots, x_n) \circ (y_1, y_2, \ldots, y_n) = (x_1 y_1, x_2 y_2, \ldots, x_n y_n)$ and the inner product $< \cdot, \cdot >$ defined by $< (x_1, x_2, \ldots, x_n), (y_1, y_2, \ldots, y_n) >= \sum_{i=1}^n x_i y_i$.

Example 50.5. The real vector space $\mathrm{Sym}(n, \mathbb{R})$ is a Euclidean Jordan algebra when endowed with the bilinear map $\circ$ defined by $x \circ y = (xy + yx)/2$ and with the inner product $< x, y >= \mathrm{Tr}(xy)$, where Tr denotes the usual trace of matrices.

Let $\mathscr{V}$ be a Euclidean Jordan algebra with a unit element $\mathbf{e}$. An element c in $\mathscr{V}$ is an idempotent if $c^2 = c$. Two idempotents c and d are orthogonal if $c \cdot d = 0$. The set $\{c_1, c_2, \ldots, c_l\}$ is a complete system of orthogonal idempotents if

$$\begin{aligned} c_i^2 &= c_i \text{ for } i = 1, \ldots, l, \\ c_i \cdot c_j &= 0 \text{ if } i \neq j, \\ \sum_{i=1}^{l} c_i &= \mathbf{e}. \end{aligned}$$

An idempotent c is primitive if it is a non-zero idempotent of $\mathcal{V}$ and if it can't be written as a sum of two non-zero idempotents.

We say that $\{c_1, c_2, \ldots, c_k\}$ is a Jordan frame if $\{c_1, c_2, \ldots, c_k\}$ is a complete system of orthogonal idempotents such that each idempotent is primitive.

Example 50.6. Let $E = \mathbb{R}^n$. The set $\{(1, 0, \ldots, 0), (0, 1, \ldots, 0), \ldots, (0, 0, \ldots, 1)\}$ is a Jordan frame of E.

Example 50.7. Let $E = \mathrm{Sym}(n, \mathbb{R})$ and let F_{ii} for $i \in \{1, \ldots, n\}$ be the matrices defined by $(F_{ii})_{pq} = \delta_{ip}\delta_{iq}$ for all p and $q \in \{1, \ldots, n\}$. $\{F_{11}, F_{22}, \ldots, F_{nn}\}$ is a Jordan frame of E.

Theorem 50.1. *([2], p. 43). Let $\mathcal{V}$ be a Euclidean Jordan algebra. Then for x in $\mathcal{V}$ there exist unique real numbers $\lambda_1, \lambda_2, \ldots, \lambda_k$, all distinct, and a unique complete system of orthogonal idempotents $\{c_1, c_2, \ldots, c_k\}$ such that*

$$x = \lambda_1 c_1 + \lambda_2 c_2 + \cdots + \lambda_k c_k. \tag{50.7}$$

Additionally, $c_j \in \mathbb{R}[x]$ for $j = 1, \ldots, k$.

The numbers λ_j's of (50.7) are the eigenvalues of x and the decomposition (50.7) is the first spectral decomposition of x.

Theorem 50.2. *([2], p. 44). Let $\mathcal{V}$ be a Euclidean Jordan algebra with rank$(\mathcal{V}) = r$. Then, for each x in $\mathcal{V}$ there exist a Jordan frame $\{c_1, c_2, \ldots, c_r\}$ and real numbers $\lambda_1, \ldots, \lambda_{r-1}$ and λ_r such that*

$$x = \lambda_1 c_1 + \lambda_2 c_2 + \cdots + \lambda_r c_r. \tag{50.8}$$

The numbers λ_j's (with their multiplicities) are uniquely determined by x.

The decomposition (50.8) is called the second spectral decomposition of x. We must say that the second spectral decomposition of x is not unique.

We may introduce on a Euclidean Jordan algebra $\mathcal{V}$ a new inner product $< \cdot, \cdot >$ defined by $< x, y >= \mathrm{tr}(x \cdot y)$ for all x and y in $\mathcal{V}$. We will also use the notation $< x, y >= \mathrm{tr}_{\mathcal{V}}(x \cdot y)$, where $\mathrm{tr}_{\mathcal{V}}(x \cdot y) = \mathrm{tr}(x \cdot y)$.

Let $\mathcal{V}$ be a m-dimensional Euclidean Jordan algebra. From property (50.6) we conclude that for all x and y in $\mathcal{V}$, $x \cdot y = 0 \Rightarrow < x, y >= 0$. Therefore, if $\mathcal{B}$ is a complete system of orthogonal idempotents of $\mathcal{V}$ then $\mathcal{B}$ is a linearly independent set of $\mathcal{V}$. So, every complete system of orthogonal idempotents of $\mathcal{V}$ with cardinality m is a basis of $\mathcal{V}$. Let x be an element of $\mathcal{V}$. If $\mathcal{L}_0(x)$ has m distinct eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_{m-1}$ and λ_m then $\mathcal{B} = \{c_1, \ldots, c_m\}$ where $c_i = \prod_{i \neq j}(\mathcal{L}_0(x)\mathbf{e} - \lambda_j \mathbf{e})/(\lambda_i - \lambda_j)$ for $i = 1, \ldots, m$ is a Jordan frame that is a basis of $\mathcal{V}$.

We intend, until Sect. 50.6, to complement the work done in [1]. We will obtain the character table of the Euclidean Jordan algebra constructed in [1] relatively to two of its basis, where one of the basis is a complete system of orthogonal primitive idempotents, recurring to homogeneous linear difference equations of second order with constant coefficients of the form $\tilde{\beta}_{j+1,i} - a_i\tilde{\beta}_{j,i} + \tilde{\beta}_{j-1,i} = 0$ where a_i is a real constant.

50.2 A Euclidean Jordan Algebra Spanned by a Family of Matrices

Following [1], let n be a natural number and $m = \lfloor n/2 \rfloor + 1$. For all $i, j \in \{1, \ldots, n\}$ consider the matrices $E_{ij} \in \mathbb{R}^{n\times n}$, defined by $(E_{ij})_{pq} = \delta_{ip}\delta_{jq}$. Let $\mathscr{F} = \{A_i\}_{i\in\{1,\ldots,m\}}$ be the family of matrices such that $A_1 = I_n$,

$$A_r = \sum_{l=r}^{n} E_{l,\,l-r+1} + \sum_{l=r}^{n} E_{l-r+1,\,l} + \sum_{l=1}^{r-1} E_{n-r+1+l,\,l} + \sum_{l=1}^{r-1} E_{l,\,n-r+1+l},$$

for $r = 2, \ldots, m-1$, and

$$A_m = \sum_{l=m}^{n} E_{l,\,l-m+1} + \sum_{l=m}^{n} E_{l-m+1,\,l} + \sum_{l=1}^{m-1} E_{n-m+1+l,\,l} + \sum_{l=1}^{m-1} E_{l,\,n-m+1+l}$$

if n is odd, and

$$A_m = \sum_{l=1}^{m-1} E_{n-m+1+l,\,l} + \sum_{l=1}^{m-1} E_{l,\,n-m+1+l}$$

if n is even.

Let $< \mathscr{V}_n, +, ., < ., . >>$ be the Euclidean space spanned by $\{A_1, A_2, \ldots, A_m\}$ over $\mathbb{R}$, where $+$ and $\cdot$ are the usual sum and product of matrices, respectively, and $< ., . >$ is the inner product defined by $< X, Y >= \mathrm{Tr}(XY)$.

The Theorem 50.3 presents several useful algebraic properties of the family $\mathscr{F}$.

Theorem 50.3. *([1], p. 884). The matrices of the family $\mathscr{F}$ verify*

$$A_iA_j = A_{i+j-1} + A_{j-i+1} \ \textit{if} \begin{cases} 2 \le i < j < m \\ i + j \le m \end{cases},$$

$$A_iA_j = A_{n-(i+j-1)+1} + A_{j-i+1} \ \textit{if} \begin{cases} 2 \le i < j < m \\ i + j \ge m + 2 \end{cases},$$

$$A_i^2 = A_{2i-1} + 2I_n \ \textit{if}\ 2 \le i \le \left\lfloor \frac{m+1}{2} \right\rfloor,$$

$$A_i^2 = A_{n-2(i-1)+1} + 2I_n \ \textit{if}\ i \ge \left\lceil \frac{m}{2} \right\rceil + 1.$$

If n is odd then

$$\begin{aligned} A_i A_j &= A_{i+j-1} + A_{j-i+1} \ \textit{if} \ \begin{cases} 2 \leq i < j < m \\ i + j = m + 1 \end{cases}, \\ A_i A_m &= A_{n-(i+m-1)+1} + A_{m-i+1} \ \textit{if} \ 2 \leq i \leq m - 1, \\ A_m^2 &= A_3 + 2I_n. \end{aligned}$$

If n is even then

$$\begin{aligned} A_i A_j &= 2A_m + A_{j-i+1} \ \textit{if} \begin{cases} 2 \leq i < j < m \\ i + j = m + 1 \end{cases}, \\ A_i A_m &= A_{m-i+1} \ \textit{if} \ i < m, \\ A_m^2 &= I_n. \end{aligned}$$

From the properties of the family $\mathscr{F}$ we conclude that the minimal polynomial of A_2 is the polynomial $p_m \in \mathbb{R}[t]$ given by $p_m(\lambda) = \Pi_{i=0}^{m-1}\Big(\lambda - 2\cos\big((2\pi i)/n\big)\Big)$.

Let $\mathscr{V}_n$ be the algebra spanned by $\mathscr{F}$ over $\mathbb{R}$. Then $\mathscr{V}_n$ is a Euclidean Jordan algebra and $\dim(\mathscr{V}_n) = m$.

Consider $\lambda_{i,2} = 2\cos\big((2\pi(i-1))/n\big)$ for $i = 1, \ldots, m$ and

$$P_i = \prod_{i \neq j} \frac{A_2 - \lambda_{j,2} I_n}{\lambda_{i,2} - \lambda_{j,2}}$$

for $i = 1, \ldots, m$.

Then $\mathscr{B}' = \{P_1, \ldots, P_m\}$ is a basis of $\mathscr{V}_n$. In Sect. 50.3 we obtain a character table relatively to the basis $\mathscr{B} = \{A_1, \ldots, A_m\}$ and $\mathscr{B}'$ of $\mathscr{V}_n$.

50.3 A Character Table Determined by Homogeneous Difference Equations

Let n be a natural even number and $i \in \{1, \ldots, m\}$. Then there exist scalars $\beta'_{k,i}s$ for $k = 1, \ldots, m$ such that $P_i = \sum_{k=1}^m \beta_{k,i} A_k$. After some algebraic manipulation we conclude that:

- If $j \in \{2, \ldots, m-1\}$ then $A_j = \sum_{i=1}^m \lambda_{i,j} P_i = \sum_{i=1}^m 2\frac{\beta_{j,i}}{\beta_{1,i}} P_i$
- $A_m = \sum_{i=1}^m \lambda_{i,m} P_i = \sum_{i=1}^m \frac{\beta_{m,i}}{\beta_{1,i}} P_i$
- $A_1 = I_n = \sum_{i=1}^m P_i$

Consider $i \in \{1, \ldots, m\}$ and $\tilde{\beta}_{k,i} = \beta_{k,i}/\beta_{1,i}$ for $k = 1, \ldots, m$. From Theorem 50.3, see p. 887 of [1], we deduce that the $\tilde{\beta}_{j,i}$'s for $j = 2, \ldots, m$ are determined solving recursively the system (50.9).

$$\begin{cases} 2\tilde{\beta}_{2,i} = \lambda_{i,2}, \\ \tilde{\beta}_{j-1,i} + \tilde{\beta}_{j+1,i} = \lambda_{i,2}\tilde{\beta}_{j,i}, \quad \text{for } j = 2, \ldots, m-1, \\ 2\tilde{\beta}_{m-1,i} = \lambda_{i,2}\tilde{\beta}_{m,i}. \end{cases} \tag{50.9}$$

Since $\lambda_{i,2} = 2\cos\big((2\pi(i-1))/n\big)$ then we conclude that the $\tilde{\beta}_{j,i}$'s for $j = 1, \ldots, m$ are the first m values of the solution of the homogeneous linear difference equation of second order (50.10)

$$\tilde{\beta}_{j-1,i} + \tilde{\beta}_{j+1,i} = 2\cos\left(\frac{2\pi(i-1)}{n}\right)\tilde{\beta}_{j,i} \tag{50.10}$$

with initial conditions $\tilde{\beta}_{1,i} = 1$ and $\tilde{\beta}_{2,i} = \cos\big((2\pi(i-1))/n\big)$. Then

$$\tilde{\beta}_{j,i} = \cos\left(\frac{2\pi(j-1)(i-1)}{n}\right), \forall j = 1, \ldots, m.$$

And so, the character table of $\mathscr{V}_n$ is represented by Table 50.1.

Suppose now that n is odd. For $i \in \{1, \ldots, m\}$ consider the notation $P_i = \sum_{k=1}^{m} \beta_{k,i} A_k$.

Let $i \in \{1, \ldots, m\}$ and $\tilde{\beta}_{k,i} = \beta_{k,i}/\beta_{1,i}$ for $k = 1, \ldots, m$. The $\tilde{\beta}_{j,i}$s for $j = 2, \ldots, m$, see p. 889 of [1], verify the system (50.11).

$$\begin{cases} 2\tilde{\beta}_{2,i} = \lambda_{i,2}, \\ \tilde{\beta}_{j-1,i} + \tilde{\beta}_{j+1,i} = \lambda_{i,2}\tilde{\beta}_{j,i}, \text{for } j = 2, \ldots, m-1, \\ \tilde{\beta}_{m-1,i} + \tilde{\beta}_{m,i} = \lambda_{i,2}\tilde{\beta}_{m,i}. \end{cases} \tag{50.11}$$

We conclude that the $\tilde{\beta}_{j,i}$'s for $j = 1, \ldots, m$ are the first m values of the solution of the homogeneous linear difference equation of second order with constant coefficients (50.12)

$$\tilde{\beta}_{j-1,i} + \tilde{\beta}_{j+1,i} = \lambda_{i,2}\tilde{\beta}_{j,i}, \tag{50.12}$$

with initial conditions $\tilde{\beta}_{1,i} = 1$ and $\tilde{\beta}_{2,i} = \lambda_{i,2}/2$.

Table 50.1 Character table of $\mathscr{V}_n$ when n is even

	A_1	$\cdots$	A_j	$\cdots$	A_m
P_1	1	$\cdots$	2	$\cdots$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
P_i	1	$\cdots$	$2\cos\left(\frac{2\pi(j-1)(i-1)}{n}\right)$	$\cdots$	$\cos\left(\frac{2\pi(m-1)(i-1)}{n}\right)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
P_m	1	$\cdots$	$2(-1)^{j-1}$	$\cdots$	$(-1)^{m-1}$

Table 50.2 Character table of $\mathcal{V}_n$ when n is odd

	A_1	$\cdots$	A_j	$\cdots$	A_m
P_1	1	$\cdots$	2	$\cdots$	2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
P_i	1	$\cdots$	$2\cos\left(\frac{2\pi(j-1)(i-1)}{n}\right)$	$\cdots$	$2\cos\left(\frac{2\pi(m-1)(i-1)}{n}\right)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
P_m	1	$\cdots$	$2\cos\left(\frac{\pi(j-1)(n-1)}{n}\right)$	$\cdots$	$2\cos\left(\frac{\pi(n-1)^2}{2n}\right).$

Since $\lambda_{i,2} = 2\cos\big((2\pi(i-1))/n\big)$ then $\tilde{\beta}_{j,i} = \cos\big((2\pi(j-1)(i-1))/n\big)$ for $i = 1, \ldots, m$ and for $j = 1, \ldots, m$. Finally, we present the character table of $\mathcal{V}_n$ in Table 50.2.

In the next section, we present a brief introduction to strongly regular graphs.

50.4 Strongly Regular Graphs

A graph G consists of a nonempty set $V(G)$ of vertices and a set $E(G)$ of edges. The number of vertices of a graph G is called its order and the number of edges of G its size. An edge whose endpoints are the vertices u and v is denoted by uv. We say that the vertex u is adjacent to the vertex v if uv in $E(G)$. Adjacent vertices are also called neighbors. A simple graph is a graph with neither loops (edges with both ends in the same vertex) nor multiple edges (more than one edge between the same pair of vertices). In what follows, we only deal with simple graphs. A graph of order n in which all pairs of vertices are adjacent is called a complete graph. When there is no pair of adjacent vertices the graph is called an empty graph. Let G be a graph of order n. The adjacency matrix of G is the matrix $A_G = (a_{uv})_n$ such that

$$a_{uv} = \begin{cases} 1, \text{ if } uv \in E(G) \\ 0, \text{ otherwise} \end{cases}.$$

The characteristic polynomial p_G of the adjacency matrix A_G of G is called the characteristic polynomial of G. The eigenvalues of A_G and the spectrum of A_G are also called the eigenvalues and the spectrum of G, respectively. Consider v in $V(G)$. The number of neighbors of v in $V(G)$ is called the degree of v and is denoted by $d_G(v)$. If H is a graph such that for all v in $V(H)$, $d_H(v) = p$ then we say that H is $\mathbf{p}$-regular. A graph is regular if it is $\mathbf{p}$-regular for some $\mathbf{p}$.

A graph H is called strongly regular if it is regular, non-complete, nonempty and if given any two distinct vertices u and v in $V(H)$, the number of vertices which are neighbors of both u and v depends on whether u and v are adjacent or not. H is a $(n, p; a, c)$-strongly regular graph if H is a strongly regular graph of order n which is $\mathbf{p}$-regular and if any pair of adjacent vertices have $\mathbf{a}$ common neighbors and any two distinct non-adjacent vertices have $\mathbf{c}$ common neighbors. graph. In Fig. 50.1

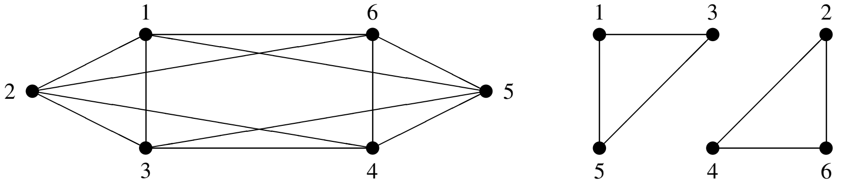


Fig. 50.1 Graphs Γ_1 and Γ_2

we represent the $(6,4;2,4)$-strongly regular graph Γ_1 and the $(6,2;1,0)$-strongly regular graph Γ_2 defined by the adjacency matrices $A_{\Gamma_1} = \begin{pmatrix} 0&1&1&0&1&1\\ 1&0&1&1&0&1\\ 1&1&0&1&1&0\\ 0&1&1&0&1&1\\ 1&0&1&1&0&1\\ 1&1&0&1&1&0 \end{pmatrix}$ and

$$A_{\Gamma_2} = \begin{pmatrix} 0&0&1&0&1&0\\ 0&0&0&1&0&1\\ 1&0&0&0&1&0\\ 0&1&0&0&0&1\\ 1&0&1&0&0&0\\ 0&1&0&1&0&0 \end{pmatrix}.$$

A graph H, see [5], which is non-complete and nonempty is a $(n,p;a,c)$-strongly regular graph if and only if $A_H^2 = pI_n + aA_H + c(J_n - A_H - I_n)$, where J_n denotes the square matrix of order n with only one's. The complement of a graph H, denoted by $\bar{H}$, is such that $V(\bar{H}) = V(H)$ and $E(\bar{H}) = \{uv : u, v \in V(H) \wedge uv \notin E(H)\}$. Therefore, since $A_{\bar{H}} = J_n - A_H - I_n$ we conclude that a graph is a $(n,p;a,c)$−strongly regular graph if and only if its complement is a $(n, n-p-1; n-2p+c-2, n-2p+a)$−strongly regular graph.

50.5 Strongly Regular Graphs in $\mathscr{V}_n$

Let n be an even natural number and let $\mathscr{V}_n$ be the Euclidean Jordan algebra spanned by the family $\mathscr{F}$. The graph τ_1 such that $A_{\tau_1} = \sum_{j=2}^{m-1} A_j$ is a $(n, n-2; n-4, n-2)$−strongly regular graph, see Theorem 4.1 at p. 891 of [1]. The spectra of τ_1 is obtained by summing the columns of Table 50.1 corresponding to $A_2, \ldots, A_{m-2}$ and A_{m-1}. Let $i \in \{2, \ldots, m-1\}$. Since

$$\sum_{j=1}^{p} \cos(px) = \frac{\cos\left(\frac{px}{2}\right)\sin\left(\frac{(p+1)x}{2}\right)}{\sin\left(\frac{x}{2}\right)} - 1 \qquad (50.13)$$

Table 50.3 The spectra of the strongly regular graphs τ_1 and $\overline{\tau}_1$

	A_{τ_1}	$A_{\overline{\tau}_1}$
P_1	$2(m-2)$	1
$\vdots$	$\vdots$	$\vdots$
P_i	$(-1)^i - 1$	$\cos\left(\frac{2\pi(m-1)(i-1)}{n}\right)$
$\vdots$	$\vdots$	$\vdots$
P_m	$(-1)^m - 1$	$(-1)^{m-1}$

Table 50.4 The spectra of the strongly regular graphs τ_2 and $\overline{\tau}_2$

	A_{τ_2}	$A_{\overline{\tau}_2}$
P_1	2	$2(m-3)+1$
$\vdots$	$\vdots$	$\vdots$
P_i	$2\cos\left(\frac{2\pi(i-1)}{3}\right)$	$-1-2\cos\left(\frac{2\pi(i-1)}{3}\right)$
$\vdots$	$\vdots$	$\vdots$
P_m	2	-3

for $p \in \mathbb{N}$ and for $x \in \mathbb{R} \setminus \{2k\pi : k \in \mathbb{Z}\}$ and considering $x = (2\pi(i-1))/n$ and $p = m-2$ on (50.13) we conclude that $\sum_{j=2}^{m-1}\bigl(2\cos(2\pi(i-1)(j-1)/n)\bigr) = (-1)^i - 1$. We present in Table 50.3 the spectra of the strongly regular graphs τ_1 and $\overline{\tau}_1$.

Now we obtain the spectra of another strongly regular graph when $n = 6k$, $k \in \mathbb{N}$. Since $m = \lfloor n/2 \rfloor + 1$ then $m = 3k+1$. We deduce, by Theorem 50.3, that

$$A_{2k+1}^2 = A_{2k+1} + 2I_n$$

and therefore A_{2k+1}^2 is a linear combination of I_n, A_{2k+1} and $J_n - A_{2k+1} - I_n$. We conclude that A_{2k+1} is the adjacency matrix of a $(n, 2; 1, 0)$-strongly regular graph τ_2. Proceeding like in the deduction of the spectra of the strongly regular graph τ_1 we obtain the spectra of the strongly regular graph τ_2. In Table 50.4 we present the spectra of the strongly regular graphs τ_2 and $\overline{\tau}_2$.

50.6 Generalized Krein Parameters of a Strongly Regular Graph

Now we will introduce the generalized Krein parameters of a strongly regular graph like in [8]. Let τ be a $(n, p; a, c)$-strongly regular graph such that $0 < c < p < n-1$ and let A be the adjacency matrix of τ. In what follows, we suppose that A has three distinct eigenvalues. Now, we consider the Euclidean Jordan subalgebra $\mathcal{V}$ of the Euclidean Jordan algebra $\mathrm{Sym}(n, \mathbb{R})$ spanned by I_n and A. Since A has three distinct eigenvalues then $\mathscr{V}$ is a three dimensional Euclidean Jordan algebra and $\operatorname{rank}(\mathscr{V}) = 3$.

Let r and s, see [5], be the eigenvalues of A different from p.

The linear operator $L(A)$ is a symmetric linear endomorphism on $\mathscr{V}$ with only three distinct eigenvalues, namely p, r and s. Let $E_1 = \big((A-rI_n)(A-sI_n)\big)/\big((p-r)(p-s)\big) = J_n/n, E_2 = \big((A-sI_n)(A-pI_n)\big)/\big((r-s)(r-p)\big)$ and $E_3 = \big((A-rI_n)(A-pI_n)\big)/\big((s-r)(s-p)\big)$. We conclude that $S = \{E_1, E_2, E_3\}$ is the unique complete system of orthogonal idempotents of $\mathscr{V}$ associated to A.

Let x be a real number. We define $|A|^x$ by the equality

$$|A|^x = p^x E_1 + r^x E_2 + |s|^x E_3. \tag{50.14}$$

Considering the basis $\mathscr{B} = \{I_n,\ A,\ E_1\}$ of $\mathscr{V}$ and recurring to the scalar product $\operatorname{tr}_{\mathscr{V}}$ we deduce that

$$\begin{aligned}|A|^x = (p-c)\frac{r^{x-1}+|s|^{x-1}}{r-s}I_n - \frac{|s|^x - r^x}{r-s}A\\ + \left(p^x - r^x + (p-r)\frac{|s|^x - r^x}{r-s}\right)E_1.\end{aligned} \tag{50.15}$$

Using the basis of $\mathscr{V}$, $\mathscr{B}' = \{I_n,\ A,\ J_n - A - I_n\}$ and recurring to equalities (50.14) and (50.15) we conclude that

$$\begin{aligned}E_1 &= \frac{1}{n}I_n + \frac{1}{n}A + \frac{1}{n}(J_n - A - I_n),\\ E_2 &= \frac{|s|n + s - p}{n(r-s)}I_n + \frac{n+s-p}{n(r-s)}A + \frac{s-p}{n(r-s)}(J_n - A - I_n),\\ E_3 &= \frac{rn + p - r}{n(r-s)}I_n + \frac{-n+p-r}{n(r-s)}A + \frac{p-r}{n(r-s)}(J_n - A - I_n).\end{aligned}$$

Now, we introduce a compact notation for the Hadamard and for the Kronecker powers of the elements of S. Let j, k, l, m, u and v be natural numbers such that $1 \leq j, u, v \leq 3, k \geq 2$ and $u < v$. For B in $\mathscr{M}_n(\mathbb{R})$, $B^{\circ k}$ and $B^{\otimes k}$ denotes the Hadamard power of order k of B and the Kronecker power of order k of B, respectively and, $B^{\circ 1} = B$ and $B^{\otimes 1} = B$. Consider the following notation: $E^{\circ\, jjk} = (E_j)^{\circ k}$, $E^{\circ uvlm} = (E_u)^{\circ l} \circ (E_v)^{\circ\, m}$, $E^{\otimes\, jjk} = (E_j)^{\otimes k}$ and $E^{\otimes uvlm} = (E_u)^{\otimes l} \otimes (E_v)^{\otimes\, m}$. Since $\mathscr{V}$ is a Euclidean Jordan algebra closed for the Hadamard product and S is a basis of $\mathscr{V}$ then there exist real numbers q^i_{jjk}, q^i_{uvlm} for $1 \leq i \leq 3$ such that

$$\begin{aligned}E^{\circ\, jjk} &= \sum_{i=1}^{3} q^i_{jjk} E_i\ ,\\ E^{\circ\, uvlm} &= \sum_{i=1}^{3} q^i_{uvlm} E_i.\end{aligned} \tag{50.16}$$

We call the parameters q^i_{jjk} and q^i_{uvlm} involved in (50.16) the generalized Krein parameters of the strongly regular graph τ, since q^i_{jj2} and q^i_{uv11} are the Krein parameters of τ.

Lets consider the natural numbers $i,\ j,\ k,\ l,\ m,\ u$ and v such that $u < v$, $1 \leq i, j, u, v \leq 3$, and $k \geq 2$. The matrices $E^{\otimes jjk}$ and $E^{\otimes uvlm}$ are idempotents matrices of $\mathscr{M}_{n^k}(\mathbb{R})$ and of $\mathscr{M}_{n^{l+m}}(\mathbb{R})$, respectively. Since the matrices $E^{\circ jjk}$ and $E^{\circ uvlm}$ are principal submatrices of the matrices $E^{\otimes jjk}$ and of $E^{\otimes uvlm}$ respectively, we conclude that $0 \leq q^i_{jjk},\ q^i_{uvlm} \leq 1$.

We assume now that k, l and m are natural numbers such that k and $l+m$ are odd. From the analysis of the generalized Krein parameters q^1_{33k} and q^1_{32lm}, we establish necessary conditions for the existence of a strongly regular graph in Theorem 50.4.

Theorem 50.4. *Let τ be a $(n, p; a, c)-$strongly regular graph such that $0 < c < p < n-1$, whose adjacency matrix A has the eigenvalues p, r and s. Then*

$$\begin{aligned}&(rn+p-r)^k + (-n+p-r)^k p + (p-r)^k(n-p-1) \geq 0,\ \forall k \in 2\mathbb{N}+1\\&(rn+p-r)^l(|s|n+s-p)^m + (-n+p-r)^l(n+s-p)^m p+\\&+(p-r)^l(s-p)^m(n-p-1) \geq 0, \forall l, m \in \mathbb{N}:\ \ l+m \in 2\mathbb{N}+1.\end{aligned}$$

Acknowledgements This work was supported by the Foundation of Sciences and Technology (FCT) through the Center of Mathematics of the University of Porto. Partial results of the paper [1] and of [8] are presented in this chapter with the kind permission of Springer Science and Business Media and with the permission from AIP Conference Proceedings 1168, 995 (2009), Copyright 2009, American Institute of Physics.

References

1. Cardoso, D.M., Vieira, L.A.: Euclidean Jordan Algebras with strongly regular graphs. J. Math. Sci. **120**, 881–894 (2004)
2. Faraut, J., Korányi, A.: Analysis on Symmetric Cones. Oxford Mathematical Monographs. Clarendon Press, Oxford (1994)
3. Faybusovich, L.: Linear systems in Jordan algebras and primal-dual interior-point methods. J. Comput. Appl. Math. **86**, 149–175 (1997)
4. Faybusovich, L.: Euclidean Jordan Algebras and Interior-point Algorithms. Positivity **1**, 331–357 (1997)
5. Godsil, C.D.: Algebraic Graph Theory. Chapman & Hall, New York (1993)
6. Koecher, M.: The Minnesota Notes on Jordan Algebras and Their Applications. Springer, Berlin (1999)
7. Massam, H., Neher, E.: Estimation and testing for lattice condicional independence models on Euclidean Jordan algebras. Ann. Stat. **26**, 1051–1082 (1998)
8. Vieira, L.A.: Euclidean Jordan Algebras and Inequalities on the Parameters of a Strongly Regular Graph. AIP Conference Proceedings, vol. 1168, pp. 995–998 (2009)

Chapter 51
Parameter Estimation in Stochastic Differential Equations

G.-W. Weber, P. Taylan, Z.-K. Görgülü, H. Abd. Rahman, and A. Bahar

Abstract Financial processes as processes in nature, are subject to stochastic fluctuations. Stochastic differential equations turn out to be an advantageous representation of such noisy, real-world problems, and together with their identification, they play an important role in the sectors of finance, but also in physics and biotechnology. These equations, however, are often hard to represent and to resolve. Thus we express them in a simplified manner of approximation by discretization and additive models based on splines. This defines a trilevel problem consisting of an optimization and a representation problem (portfolio optimization), and a parameter estimation (Weber et al. Financial Regression and Organization. In: Special Issue on Optimization in Finance, *DCDIS-B*, 2010). Two types of parameters dependency, linear and nonlinear, are considered by constructing a penalized residual sum of squares and investigating the related Tikhonov regularization problem for the first one. In the nonlinear case Gauss–Newton's method and Levenberg–Marquardt's method are employed in determining the iteration steps. Both cases are treated using continuous optimization techniques by the elegant framework of conic quadratic programming. These convex problems are well-structured, hence, allowing the use of the efficient interior point methods. Furthermore, we present nonparametric and

G.-W. Weber (✉)
METU, IAM, Ankara, Turkey
e-mail: gweber@metu.edu.tr

P. Taylan
Department of mathematics, Dicle University, Diyarbakır, Turkey
e-mail: ptaylan@dicle.edu.tr

Z.-K. Görgülü
Universität der Bundeswehr München, Munich, Germany
e-mail: das_lemma@gmx.de

H.A. Rahman and A. Bahar
Faculty of Science, UTM, Skudai, Malaysia
e-mail: halizarahman@utm.my, arifah@mel.fs.utm.my

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_51,

related methods, and introduce into research done at the moment in our research groups which ends with a conclusion.

51.1 Introduction

The majority of problems in finance involve either maximization of wealth or minimization of costs. Thus, the modeling of financial processes consists of optimization and optimal control in the organisation of a portfolio apart from security requests or pricing and hedging. The present study focusses on the parameter estimation that is a special step in modeling which serves as a link to the overall context of modeling and decision making.

Real-world data from the financial sector and science are often characterized by their huge quantity and variation, while serving as the basis of future prediction at the same time. Both the real situation and practical requests are difficult to put in equilibrium [23, 41]. In fact, the related mathematical modeling faces a high sensitivity of the model with respect to slightest perturbations of the data and, in the limit, with non-smoothness.

To better understand this sensitivity, we will analyze the corresponding parameter estimation problems by means of Tikhonov regularization, conic quadratic programming and nonlinear regression methods. Herewith, we offer an a priori approach to stochastic differential equations (SDEs) inspired by the martingale method in portfolio optimization. The entire problem is described by three phases, ordered in some sequel: an optimization problem, a representation problem (portfolio optimization) and a parameter estimation which can be done at the first or at the last level.

As a preparation, a brief introduction into our regression method is given from statistical learning called additive models, which we will then be evaluated by continuous optimization. Then, using modern methods of regularization and optimization, we will apply them to SDEs addressing both the linear and nonlinear case of parameter estimation. Various examples are given for the nonlinear case, especially referring to interest rate models. By all of this, our more theoretical paper recalls and improves the scientific results in the pioneering articles [40, 43], and addresses them to a wider community.

This contribution is organized as follows: In Sect. 51.2 we introduce additive models with a classical tool, which we then apply on stochastic differential equations, presented, treated and optimized in Sect. 51.3. Then, in Sect. 51.4, we apply Tikhonov regularization and conic quadratic programming on our equations. The following Sect. 51.5 is devoted nonlinear parameter estimation. Section 51.6 deals with cases of continuous time one-factor interest rate models. Then, further, nonparametric methods are the subject of Sect. 51.7, before Sect. 51.8 offers other recent investigations in our research groups. In Sect. 51.9 we conclude.

51.2 Classical Additive Models

An additive model [10] is a special regression model to estimate an additive approximation of the multivariate regression function.

For N observations on a response (or dependent) variable Y, denoted by $y = (y_1, y_2, \ldots, y_N)^T$ measured at N design vectors $x_i = (x_{i1}, x_{i2}, \ldots, x_{im})^T$, the additive model is defined by

$$Y = \beta_0 + \sum_{j=1}^{m} f_j(X_j) + \varepsilon,$$

with an error (or noise) ε being independent of the factors X_j, $E(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2$ [19]. The functions f_j are arbitrary unknown, univariate functions which are, mostly, considered to be splines. We denote the estimates by $\hat{f}_j$. The standard convention consists in assuming at X_j that $E\left(f_j(X_j)\right) = 0$, since otherwise there will be a free constant in each of the functions [19]; the intercept (bias) β_0 summarizes all those constants.

51.2.1 Estimation Equations for Additive Model

Additive models are mainly used as data analytic tool. Each function is estimated by an algorithm proposed by Friedman and Stuetzle [16], called backfitting (or Gauss–Seidel) algorithm. We estimate β_0 by the mean of the response variable Y, i.e., $\hat{\beta}_0 = E(Y)$. This procedure depends on the partial residual against X_j,

$$r_j = Y - \hat{\beta}_0 - \sum_{k \neq j} \hat{f}_k(X_k),$$

and it consists of estimating each smooth function by holding all the other ones fixed [19, 23]. This yields $E(r_j \,|X_j) = \hat{f}_j(X_j)$, which minimizes $E\big(Y - \hat{\beta}_0 - \sum_{j=1}^{m} \hat{f}_j(X_j)\big)^2$ [16, 18].

51.2.2 On the Theoretical and Practical Background

This study is motivated by the financial mathematics of stochastic differential equations. It is a very hard challenge to find such models of reality as an approximation based on data from the sector of finance. These equations can be, e.g., about processes of prices, interest rates and volatility, about underlyings and derivatives of different kinds as well. The financial data used are usually characterized by noise and a high variation of their values; this fact is a great obstacle for any modelling (data fitting) which should be stable against small perturbations, such as noise, in the data, i.e., it should reveal a very modest complexity [42].

We aim at these goals in a balanced way, by using theories of inverse problems and continuous, actually, convex optimization with a very characteristic global structure which is generated and represented by quadratic cones. This treatment demonstrates and proposes for our problem the use and benefit of a closed mathematical approach which is called model based, in contrast to model free approaches from statistics which are of a more adaptive and heuristic nature [19]. For example, the backfitting (Gauss–Seidel) algorithm and the MARS algorithm from statistical learning in additive, generalized and multiplicative models reveal that nature [19,51].

Our new mathematical approach in the areas of these models proved to be successful and really competitive with traditional statistical methods, as demonstrated for the prediction of credit default and for quality analysis and control in manufacturing [20,51].

51.2.3 Closer Explanations on the Optimization and Computational Methods

Least-squares estimation, i.e., the minimization of the residual sum of squares between the left- and right-hand sides of a time-discretized stochastic differential equation (SDE) is needed to model such an SDE.

In fact, we embed this SDE into a common context together with portfolio optimization. This context turns out to be a multi-stage problem. Our least-squares estimation takes two formes discussed by us, on one hand, a linear regression, after having introduced linear combinations of spline functions to represent model functions in the SDE, and, on the other hand, a nonlinear regression, in the presence of nonlinear parametric dependencies and subtle compositions. Here linear regression techniques will be used in each iteration of the nonlinear method to find the appropriate step (see Sect. 51.5).

Actually, besides of that parameter estimation which aims at accuracy, we also have another target, which consists in a smallest possible complexity of the model or, equivalently, in stability. Both goals are firstly combined in the tradeoff that is given in the form of a penalized problem, i.e., of minimizing the residual sum of squares. Numerical methods which can be applied here are usually called exterior point methods [26]. In fact, since then we unify some discretized complexity (or energy) terms and include them into an inequality constraint, where we impose a bound on it, we obtain a conic quadratic programming problem on which we can apply interior point methods [28]. Our main approach to exterior point methods is given by Tikhonov regularization from the theory of inverse problems. For this kind of stabilization the program package of MATLAB Regularization Toolbox is provided; it strongly uses the techniques of singular value decomposition, including “truncation” and “filtering” [2]. These two approaches both the interior and the exterior one, are equivalent under a suitable constellation of the corresponding two parameters, namely, the penalty (or smoothing) parameter and the upper bound in the constraint [2]. In our study, we prefer to employ interior points methods (IPMs) which we apply on our conic quadratic programming problem, e.g., via the program package MOSEK (for various applications see [51]). Herewith, we exploit the fact that conic programs are well-structured convex problems, which are a very powerful model-based approach that enables us to benefit from the efficiency of IPMs [28].

51.3 Equations and their Optimization

51.3.1 A Short Introduction

The concept of stochastic differential equation (SDE) extends the concept of ordinary differential equation to stochastic processes. Stochastic differential equations come into operation to simulate ordinary, predictable, processes those are additionally affected by outer perturbation (noise).

As in the deterministic case, in the case of stochastic processes we aim at a relationship between value and future trend (derivative) of a function. For, Îto processes are nowhere differentiable, initially, a derivative of a process causes problems.

However, an ordinary differential equation

$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = a\left(t, x_t\right)$$

is equivalent to an integral equation

$$x_t = x_0 + \int_0^t a(\xi, x_\xi)\, \mathrm{d}\xi.$$

The latter formulation does not need the derivative’s term. Thus, in the case of stochastic differential equations

$$X_t = a\left(t, X_t\right) + b(t, X_t)\frac{\mathrm{d}W_t}{\mathrm{d}t},$$

where $a(t, X_t)$ denotes the deterministic, $b(t, X_t) \cdot \mathrm{d}W_t/\mathrm{d}t$ the stochastic influence (noise) and $(W_t)_{t\geq 0}$ a continuous martingale, it makes more sense to define the SDE by reference to the corresponding integral equation

$$X_t = X_0 + \int_0^t a\left(\xi, X_\xi\right)\, \mathrm{d}\xi + \int_0^t b\left(\xi, X_\xi\right)\, \mathrm{d}W_\xi,$$

where the latter expression is an Îto integral.

51.3.2 A Special Approach to Financial Processes

Many phenomena in nature, technology and economy are modelled by means of a deterministic differential equation with initial value $x_0 \in \mathbb{R}$:

$$\begin{cases} \dot{x} := \mathrm{d}x/\mathrm{d}t = a(x,t), \\ x(0) \;=\; x_0. \end{cases}$$

However, this modeling omits stochastic fluctuations and is not appropriate for, e.g., stock prices, population dynamics and biometry to name a few. To consider stochastic movements, stochastic differential equations (SDEs) are used since they arise in modeling many phenomena, such as random dynamics in the physical, biological and social sciences, in engineering and economy. Solutions of these equations are often diffusion processes and, hence, they are connected to the subject of partial differential equations. We try to find a solution for these equations by an additive approximation (cf. Sect. 51.3.3), which is very famous in the statistical area, using spline functions.

Typically, a stochastic differential equation, equipped with an initial value, is given by

$$\begin{cases} \dot{X}(t) = a(X,t) + b(X,t)\delta_t \qquad t \in [0,\infty), \\ X(0) = x_0, \end{cases} \tag{51.1}$$

here, a is the deterministic part, $b\delta_t$ is the stochastic part, and δ_t denotes a generalized stochastic process [22,32]. An example of a generalized stochastic processes is white noise. For a generalized stochastic processes, derivatives of any order can be defined. Suppose that W_t is a generalized version of a Wiener process which is used to model the motion of stock prices, which instantly responds to the numerous arising and, actually, emerging information. A one-dimensional Wiener process (or a Brownian motion) is a time continuous process satisfying the following properties:

(a) $W_0 = 0$, with probability one.
(b) $W_t - W_s \sim N(0, t-s)$ for all s, t with $0 \leq s < t \leq T$. Here, we speak about stationary increments.
(c) All increments $\Delta W_t := W_{t+\Delta t} - W_t$ on nonoverlapping time intervals are independent. That is, the displacements $W_{t_2} - W_{t_1}$ and $W_{t_4} - W_{t_3}$ are independent for all $0 \leq t_1 < t_2 \leq t_3 < t_4$.

From a. we learn that, especially, $W_t \sim N(0,t)$ for all $0 \leq t \leq T$; i.e., for each t the random variable W_t is normally distributed with mean $E(W_t) = 0$ and variance $\mathrm{Var}(W_t) = E\left(W_t^2\right) = t$. A multi-dimensional Wiener process can be similarly defined. Usually a Wiener process is differentiable almost nowhere. To obtain our approximate and, then, smoothened model, we treat W_t as if it was differentiable (a first approach widespread in literature). Then, white noise δ_t is defined as $\delta_t = \dot{W}_t = \mathrm{d}W_t/\mathrm{d}t$ and a Wiener process can be got by smoothing the white noise. As a Wiener process is nowhere differentiable, to obtain an approximation of the

SDE model we replace in (51.1) δ_t with $\mathrm{d}W_t/\mathrm{d}t$ by treating the time interval as continuous, i.e., $\Delta t \to 0$, and thus W_t as differentiable, the following SDE can be rewritten as

$$\mathrm{d}X_t = a(X_t, t)\mathrm{d}t + b(X_t, t)\mathrm{d}W_t, \tag{51.2}$$

here, $a\,(X_t, t)$ and $b\,(X_t, t)$ are drift and diffusion term, respectively, and X_t is a solution. We approximate the solution through discretization of SDE.

Some popular SDE models are listed in Table 51.1. The parameter estimation of the diffusion processes of discrete-time observations should ideally be based on a likelihood function. If the transition densities of X are known, one can use a likelihood function. The transition densities of X are usually unknown and, thus, it has to be approximated, but this is proven to be quite computer intensive. As alternative one can approximate the log-likelihood function based on the continuous observation of X. The maximizer of the approximate log-likelihood function will provide the approximate maximum likelihood estimator (AMLE) [6]. However, the previous works [17, 33] employed methods for maximum likelihood estimation.

In [33] it is explained the oscillations of glycemia in response to hyperinsulinization by extending a system of ordinary differential equation (ODE) to a system of stochastic differential equations (SDE). The parameters estimated for the ODE were based on Iteratively Weighted Least Squares (IRWLS) method. The stochastic model of Euglycemic Hyperinsulinemic Clamp (EHC) was fitted to the data and the system noise was estimated by a simulated maximum likelihood procedure. The system noise estimates were found non-negligible and robust to changes in measurement error values. They concluded that the explicit expression of system noise was physiologically relevant since the glucose uptake rate is affected by a host of additive influences.

In [17] it is examined the main probabilistic characteristics and described an explicit expression of the trends and the stationary distribution of the model based on the homogeneous Rayleigh diffusion process. They estimated the maximum likelihood parameters and simulated the stochastic sample path of the model on the corresponding Îto stochastic differential equation. The model was applied to study the evolution of thermal electricity in Maghribi.

Other methods are methods of moments [30], filtering, e.g., extended Kalman filter [29] and non-linear least squares [24]. One of the drawbacks of the classical least squares and maximum likelihood method is its inadequacy for simultaneous estimation of the drift and diffusion parameters while in methods of moments the sample is not used efficiently.

Many of the analytical solutions of SDE are unknown, we have to approximate the true solution through discretization of SDE.

51.3.3 Discretization of SDE

A number of discretization schemes available for the SDE (51.2) among others are Euler–Maruyama, Milstein and Runge–Kutta [22]. We choose the Milstein scheme

Table 51.1 Popular SDE models

	Constant diffusion term	Linear diffusion term	Nonlinear diffusion term
Constant	Bachelier		
drift term	$\mathrm{d}r_t = \beta \, \mathrm{d}t + \sigma \, \mathrm{d}W_t$		
Linear	Ornstein–Uhlenbeck	Black–Scholes	Feller/Cox–Ingersoll–Ross
drift term	$\mathrm{d}r_t = \beta r_t \, \mathrm{d}t + \sigma \, \mathrm{d}W_t$	$\mathrm{d}X_t = \beta X_t + \sigma X_t \, \mathrm{d}W_t$	$\mathrm{d}X_t = (\alpha - \beta X_t) \, \mathrm{d}t + \sigma \, \mathrm{d}W_t$
			Chan–Karloyi–Logstaff–Sanders
			$\mathrm{d}r_t = \kappa \, (\Theta - r_t) \, \mathrm{d}t + \sigma r_t^{\gamma} \, \mathrm{d}W_t$
			Ornstein–Uhlenbeck
			$\mathrm{d}r_t = (1 + 2\beta r_t) \, \mathrm{d}t + 2\sigma \sqrt{r_t} \, \mathrm{d}W_t$
Nonlinear	Ornstein–Uhlenbeck	Gompertz diffusion	
drift term	$\mathrm{d}X_t = \left(\alpha X_t^{-1} - X_t\right) \, \mathrm{d}t + \sigma \, \mathrm{d}W_t$	$\mathrm{d}r_t = (\alpha r_t - \beta r_t \log r_t) \, \mathrm{d}t + 2\sigma \sqrt{r_t} \, \mathrm{d}W_t$	
	Hyperbolic diffusion	Logistic diffusion	
	$\mathrm{d}r_t = \alpha \frac{r_t}{\sqrt{1+r_t^2}} \, \mathrm{d}t + \sigma \, \mathrm{d}W_t$	$\mathrm{d}r_t = \left(\alpha r_t - \beta X_t^2\right) \, \mathrm{d}t + \sigma r_t \, \mathrm{d}W_t$	
	Kessler–Sørensen		
	$\mathrm{d}r_t = -\Theta \tan r_t \, \mathrm{d}t + \sigma \, \mathrm{d}W_t$		

because it has a strong order of convergence compared with Euler–Maruyama and it is easier compared with Runge–Kutta. Then, we represent an approximation $\hat{X}_{t_j}$, in short: $\hat{X}_j$ $(j \in \mathbb{N})$, of the process X_t by

$$\begin{aligned}\hat{X}_{j+1} = \hat{X}_j + a\left(\hat{X}_j, t_j\right)(t_{j+1} - t_j) + b\left(\hat{X}_j, t_j\right)(W_{j+1} - W_j) \\ + \frac{1}{2}(b'b)\left(\hat{X}_j, t_j\right)\left((W_{j+1} - W_j)^2 - (t_{j+1} - t_j)\right),\end{aligned} \tag{51.3}$$

where the prime "$'$" denotes the derivative with respect to t. Particularly referring to the finitely many sample (data) points $(\overline{X}_j, \overline{t}_j)$ $(j = 1, 2, \ldots, N)$, we obtain

$$\dot{\overline{X}}_j = a\left(\overline{X}_j, \overline{t}_j\right) + b\left(\overline{X}_j, \overline{t}_j\right)\frac{\Delta W_j}{\overline{h}_j} + \frac{1}{2}(b'b)\left(\overline{X}_j, \overline{t}_j\right)\left(\frac{(\Delta W_j)^2}{\overline{h}_j} - 1\right), \tag{51.4}$$

where the value $\dot{\overline{X}}_j$ represents difference quotients raised on the j–th data value $\overline{X}_j$ and on step lengths $\Delta\overline{t}_j = \overline{h}_j := \overline{t}_{j+1} - \overline{t}_j$ between neighboring sampling times:

$$\dot{\overline{X}}_j := \begin{cases} \frac{\overline{X}_{j+1} - \overline{X}_j}{\overline{h}_j}, & \text{if } j = 1, 2, \ldots, N-1, \\ \frac{\overline{X}_N - \overline{X}_{N-1}}{\overline{h}_N}, & \text{if } j = N. \end{cases}$$

The relations (51.4) cannot be expected to hold in an exact sense, since they include real data, but we satisfy them best in the approximate sense of residual sums of squares (**RSS**), also called least squares of errors. For the ease of exposition, we write "=" instead of the approximation symbol "≈". We will study the minimization of **RSS** in Sect. 51.3.4, where we combine it with the need for regularization and speak about **PRSS** ("*P*" abbreviating penalized).

As $W_t \sim N(0, t)$, the increments ΔW_j are independent on non-overlapping intervals and moreover, $\text{Var}(\Delta\overline{W}_j) = \Delta\overline{t}_j$, hence, the increments having normal distribution can be simulated with the help of standard normal distributed random numbers $\overline{Z}_j$. Herewith, we obtain a discrete model for a Wiener process:

$$\Delta\overline{W}_j = \overline{Z}_j\sqrt{\Delta\overline{t}_j}, \quad \overline{Z}_j \sim N(0, 1). \tag{51.5}$$

Inserting this value in our discretized equation, we receive

$$\dot{\overline{X}}_j = a\left(\overline{X}_j, \overline{t}_j\right) + b\left(\overline{X}_j, \overline{t}_j\right)\frac{\overline{Z}_j}{\sqrt{\overline{h}_j}} + \frac{1}{2}(b'b)\left(\overline{X}_j, \overline{t}_j\right)\left(\overline{Z}_j^2 - 1\right), \tag{51.6}$$

what we abbreviate by

$$\dot{\overline{X}}_j = \overline{G}_j + \overline{H}_j c_j + \left(\overline{H}_j{}'\overline{H}_j\right)d_j \tag{51.7}$$

with $c_j := Z_j/\sqrt{\overline{h}_j}$, $d_j := 1/2\left(\overline{Z}_j^2 - 1\right)$, $\overline{G}_j := a\left(\overline{X}_j, \overline{t}_j\right)$ and $\overline{H}_j := b\left(\overline{X}_j, \overline{t}_j\right)$. To determine the unknown values of $\overline{G}_j$ and $\overline{H}_j$, we consider the following optimization problem:

$$\min_{\theta} \sum_{j=1}^{N} \left\| \dot{\overline{X}}_j - \left(\overline{G}_j + \overline{H}_j c_j + \left(\overline{H}_j{}'\overline{H}_j\right) d_j\right) \right\|_2^2, \tag{51.8}$$

where the vector θ comprises all the parameters in the Milstein model. We point out that also vector-valued processes could be studied, referring to sums of terms in the Euclidean norm $\|\cdot\|_2^2$. We note that data from the stock markets, but also from other sources of information or communication, have a high variation; we shall take this carefully into account, subsequently, in terms of regularization.

Now, we have to employ a parameter estimation method which will at the same time control that high variation and give a smoother approximation to the data. Splines are more flexible and they allow us to avoid large oscillation which may be permitted by high-degree polynomial approximation and based on strongly varying data or outliers existing. Here, one sometimes speaks also of overfitting; we want to prevent from that.

Let us call that splines from finite dimensional spline spaces can be described as linear combinations of basis splines and that they approximate the data $(\dot{\overline{X}}_j, \overline{t}_j)$ smoothly. For this reason we approximate each function underlying the numbers $\overline{G}_j = a(\overline{X}_j, \overline{t}_j)$, $\overline{H}_j = b(\overline{X}_j, \overline{t}_j)$ and $\overline{F}_j = \overline{H}_j{}'\overline{H}_j$ in an additive way established on basis splines and then introduce a regularization. This treatment is very useful for the stability of the model in the presence of the many and highly varying data. Let us use basis splines for each function establishing an additive separation of variables (coordinates); e.g., in equation (7):

$$\begin{aligned}
\overline{G}_j &= a\left(\overline{X}_j, \overline{t}_j\right) = \alpha_0 + \sum_{p=1}^{2} f_p\left(\overline{U}_{j,p}\right) = \alpha_0 + \sum_{p=1}^{2}\sum_{l=1}^{d_p^g} \alpha_p^l B_p^l\left(\overline{U}_{j,p}\right), \\
\overline{H}_j c_j &= b\left(\overline{X}_j, \overline{t}_j\right) c_j = \beta_0 + \sum_{r=1}^{2} g_r\left(\overline{U}_{j,r}\right) = \beta_0 + \sum_{r=1}^{2}\sum_{m=1}^{d_r^h} \beta_r^m C_r^m\left(\overline{U}_{j,r}\right), \\
\overline{F}_j d_j &= \left(b'b\right)\left(\overline{X}_j, \overline{t}_j\right) d_j = \varphi_0 + \sum_{s=1}^{2} h_s\left(\overline{U}_{j,s}\right) = \varphi_0 + \sum_{s=1}^{2}\sum_{n=1}^{d_s^f} \varphi_s^n D_s^n\left(\overline{U}_{j,s}\right),
\end{aligned} \tag{51.9}$$

where we used the unifying notation $\overline{U}_j = \left(\overline{U}_{j,1}, \overline{U}_{j,2}\right) := \left(\overline{X}_j, \overline{t}_j\right)$. Let us give an example on how one can gain bases of splines. If we denote the kth order base spline by $B_{\eta,k}$, a polynomial of degree $k-1$, with knots, say x_η, then a great benefit of using the base splines is provided by the following recursive algorithm [15]:

$$B_{\eta,1}(x) = \begin{cases} 1, \text{ if } & x_\eta \leq x < x_{\eta+1}, \\ 0, \text{ otherwise}, \end{cases}$$

$$B_{\eta,k}(x) = \frac{x - x_\eta}{x_{\eta+k-1} - x_\eta} B_{\eta,k-1}(x) + \frac{x_{\eta+k} - x}{x_{\eta+k} - x_{\eta+1}} B_{\eta+1,k-1}(x). \quad (51.10)$$

Before we actually introduce how this regularized parameter estimation becomes modeled and solved, we explain how it is connected with portfolio optimization and introduce into that.

51.3.4 Portfolio Optimization Related

An important scope of application for SDEs in financial mathematics is provided by portfolio optimization and the stochastic control. A thorough investigation of this field is done in [23]. To understand the genesis of a trilevel problem it is indispensable to refresh some elements of portfolio optimization. In portfolio optimization there are basically two approaches in use: the martingale method, and the stochastic control. Let us give a short introduction into both.

51.3.4.1 The Martingale Method

The martingale method describes a bi-level method considering an optimization problem at the lower level and a representation problem at the upper level. This method is mainly based on a separation of the dynamical problem

$$\max_{(\pi,c)\in A'(x)} \quad J(x;\pi,c)$$

into a static optimization problem ("determination of the optimal payoff profile") at the lower level and a representation problem ("compute the portfolio process corresponding to the optimal payoff profile") at the upper level. The definition of the feasible set includes that the expected overall utility, or wealth, should be bounded away from $-\infty$ (therefore, the negative part $(\cdot)^-$ is used):

$$A'(x) := \left\{ (\pi,c) \in A(x) \,\middle|\, E\left(\int_0^T \left(U_1\left(t, c(t)\right)\right)^- \,\mathrm{d}t + \left(U_2\left(x(T)\right)\right)^- \right) < \infty \right\},$$

where we denote a self-financing pair (π, c) consisting of a portfolio process π and a consumption process c that is admissible with initial wealth $x > 0$, by $(\pi, c) \in A(x)$. We note that the functions U_1 and U_2 are utility functions [23].

51.3.4.2 The Stochastic Control

A SDE of the form

$$\mathrm{d}X(t) = \mu\left(t, X(t), u(t)\right)\ \mathrm{d}t + \sigma\left(t, X(t), u(t)\right)\ \mathrm{d}W(t),$$

where $W(t)$ is a m-dimensional Brownian motion, $X(t)$ is a n-dimensional Itô process and where $u(t)$ is an arbitrary d-dimensional stochastic process – the stochastic control – is called a *controlled stochastic differential equation*. The main task in stochastic control consists of determining an optimal control, i.e., a control process $u(t)$ which is optimal with respect to a certain cost functional.

The martingale method bases on the identification and treatment of a bi-level problem, but our entire study here refers to a tri-level problem. There are two ways of how to realize this, recalling that the martingale method consists of two levels. Indeed, an additional parameter estimation can take place (a) at the end (as the third – most lower problem) or, (b) at the beginning (first problem – most upper problem). In case of (a), a parameter estimation is, under some criterion of least squares or maximum likelihood, applied to the solution of the portfolio optimization task which is parametric and can therefore be called a parametric portfolio optimization. However, the case of (b) is the classical one on nonparametric portfolio optimization; here, parameter estimation is taking place in an a priori sense. In the sequel, we shall address the case (b) and investigate it emphasizing the parameter estimation.

Further information about bi- and multi-level problems in optimization and related topics from optimal control we refer to [47, 49].

51.3.5 The Penalized Residual Sum of Squares Problem for SDE

We construct the *penalized residual sum of squares* (**PRSS**) for our SDE in the following form:

$$\begin{aligned}\mathbf{PRSS}(\theta, f, g, h) :=& \sum_{j=1}^{N}\left(\dot{\overline{X}}_j - \left(\overline{G}_j + \overline{H}_j c_j + \overline{F}_j d_j\right)\right)^2 + \sum_{p=1}^{2}\lambda_p \int \left(f_p''(U_p)\right)^2 \mathrm{d}U_p \\ &+ \sum_{r=1}^{2}\mu_r \int \left(g_r''(U_r)\right)^2\ \mathrm{d}U_r + \sum_{s=1}^{2}\tau_s \int \left(h_s''(U_s)\right)^2 \mathrm{d}U_s.\end{aligned} \tag{51.11}$$

Here the integral symbol $\int$ stands for $\int_{[a_\kappa, b_\kappa]}$, i.e., as a dummy variable. Be $[a_\kappa, b_\kappa]$ $(\kappa = p, r, s)$ sufficiently large intervals of integration and be U_p, U_r, U_s the unifying notation of (X_t, t), where $(U_1 = X_t, U_2 = t)$, respectively.

For the ease of exposition, we may always think that they are the same: $a_\kappa = a$ and $b_\kappa = b$ $(\kappa = p, r, s)$.

Furthermore, $\lambda_p, \mu_r, \tau_s \geq 0$ are penalty (or smoothing) parameters, standing for the tradeoff between the first term of "lack-of-fit" and the second terms of complexity (or "energy").

Remark 51.1. Large values of λ_p, μ_r and τ_s enforce "smoother" curves, smaller values can result in more fluctuation.

If we use an additive form based on the basis splines for each function, then **PRSS** becomes

$$\sum_{j=1}^{N}\left(\dot{\overline{X}}_j - \left(\overline{G}_j + \overline{H}_j c_j + \overline{F}_j d_j\right)\right)^2 = \sum_{j=1}^{N}\left(\dot{\overline{X}}_j - \left(\alpha_0 + \sum_{p=1}^{2}\sum_{l=1}^{d_p^g}\alpha_p^l B_p^l\left(\overline{U}_{j,p}\right)\right.\right.$$
$$+\beta_0 + \sum_{r=1}^{2}\sum_{m=1}^{d_r^h}\beta_r^m C_r^m\left(\overline{U}_{j,r}\right)$$
$$\left.\left.+\varphi_0 + \sum_{s=1}^{2}\sum_{n=1}^{d_s^f}\varphi_s^n D_s^n\left(\overline{U}_{j,s}\right)\right)\right)^2. \tag{51.12}$$

The above part

$$\overline{G}_j + \overline{H}_j c_j + \overline{F}_j d_j = \alpha_0 + \sum_{p=1}^{2}\sum_{l=1}^{d_p^g}\alpha_p^l B_p^l\left(\overline{U}_{j,p}\right) + \beta_0 + \sum_{r=1}^{2}\sum_{m=1}^{d_r^h}\beta_r^m C_r^m\left(\overline{U}_{j,r}\right)$$
$$+\varphi_0 + \sum_{s=1}^{2}\sum_{n=1}^{d_s^f}\varphi_s^n D_s^n\left(\overline{U}_{j,s}\right) = \overline{A}_j\theta \tag{51.13}$$

can easily be interpreted as scalar product of two vectors $\overline{A}_j$ and θ, where

$$\overline{A}_j = \begin{pmatrix} 1 \\ B_1^1\left(\overline{U}_{j,1}\right) \\ \vdots \\ B_1^{d_1^g}\left(\overline{U}_{j,1}\right) \\ B_2^1\left(\overline{U}_{j,2}\right) \\ \vdots \\ B_2^{d_2^g}\left(\overline{U}_{j,1}\right) \\ 1 \\ C_1^1\left(\overline{U}_{j,1}\right) \\ C_1^2\left(\overline{U}_{j,1}\right) \\ \vdots \\ C_1^{d_1^h}\left(\overline{U}_{j,1}\right) \\ C_2^1\left(\overline{U}_{j,2}\right) \\ \vdots \\ C_2^{d_2^h}\left(\overline{U}_{j,2}\right) \\ 1 \\ D_1^1\left(\overline{U}_{j,1}\right) \\ \vdots \\ D_1^{d_1^f}\left(\overline{U}_{j,1}\right) \\ D_2^1\left(\overline{U}_{j,2}\right) \\ \vdots \\ D_2^{d_2^f}\left(\overline{U}_{j,2}\right) \end{pmatrix} =: \begin{pmatrix} 1 \\ \left(B_1^l\right)_{1\le l\le d_1^g} \\ \left(B_2^l\right)_{1\le l\le d_2^g} \\ 1 \\ \left(C_1^l\right)_{1\le m\le d_1^h} \\ \left(C_2^l\right)_{1\le m\le d_2^h} \\ 1 \\ \left(D_1^l\right)_{1\le n\le d_1^f} \\ \left(D_2^l\right)_{1\le n\le d_2^f} \end{pmatrix} \quad \text{and} \quad \theta = \begin{pmatrix} \alpha_0 \\ \alpha_1^1 \\ \vdots \\ \alpha_1^{d_1^g} \\ \alpha_2^1 \\ \vdots \\ \alpha_2^{d_2^g} \\ \beta_0 \\ \beta_1^1 \\ \vdots \\ \beta_1^{d_1^h} \\ \beta_2^1 \\ \vdots \\ \beta_2^{d_2^h} \\ \varphi_0 \\ \varphi_1^1 \\ \vdots \\ \varphi_1^{d_1^f} \\ \varphi_2^1 \\ \vdots \\ \varphi_2^{d_2^f} \end{pmatrix} =: \begin{pmatrix} \alpha_0 \\ \left(\alpha_1^l\right)_{1\le l\le d_1^g} \\ \left(\alpha_2^l\right)_{1\le l\le d_2^g} \\ \beta_0 \\ \left(\beta_1^m\right)_{1\le m\le d_1^h} \\ \left(\alpha_2^m\right)_{1\le m\le d_2^h} \\ \varphi_0 \\ \left(\varphi_1^n\right)_{1\le n\le d_1^f} \\ \left(\varphi_2^n\right)_{1\le n\le d_2^f} \end{pmatrix}.$$

Considering the matrix

$$\overline{A} := \begin{pmatrix} \overline{A}_1^T \\ \overline{A}_2^T \\ \vdots \\ \overline{A}_N^T \end{pmatrix}^T \quad \text{and the vector of difference quotients} \quad \dot{\overline{X}} := \begin{pmatrix} \dot{\overline{X}}_1 \\ \dot{\overline{X}}_2 \\ \vdots \\ \dot{\overline{X}}_N \end{pmatrix},$$

that represents the change rates of the given data, we obtain **PRSS** as the squared length of the difference vector between $\dot{\overline{X}}$ and $\overline{A}\theta$:

$$\sum_{j=1}^{N} \left(\dot{\overline{X}}_j - \overline{A}_j\theta\right)^2 = \left\|\dot{\overline{X}} - \overline{A}\theta\right\|_2^2. \tag{51.14}$$

Furthermore, by taking functional values from the left-hand boundaries of the subintervals, we can approximate each integration term using its Riemann sum:

$$\begin{aligned}\int_a^b \left(f_p''\left(U_p\right)\right)^2 \, \mathrm{d}U_p &\approx \sum_{j=1}^{N-1} \left(f_p''\left(U_{j,p}\right)\right)^2 \left(U_{j+1,p} - U_{j,p}\right) \\ &= \sum_{j=1}^{N-1} \left(\sum_{l=1}^{d_p^g} \alpha_p^l B_p^{l''} \left(U_{j,p}\right) u_j \right)^2 . \qquad (51.15)\end{aligned}$$

On the other hand, each integration term can be represented by the squared length of a appropriate vector, i.e.,

$$\begin{aligned}\int_a^b \left(f_p''(U_p)\right)^2 \, \mathrm{d}U_p &\approx \sum_{j=1}^{N-1} \left(B_j^{p''} u_j \alpha_p\right)^2 = \left\| \overline{A}_p^B \alpha_p \right\|_2^2 \qquad (p = 1,2), \\ \int_a^b \left(g_r''(U_r)\right)^2 \, \mathrm{d}U_r &\approx \sum_{j=1}^{N-1} \left(C_j^{r''} v_j \beta_r\right)^2 = \left\| \overline{A}_r^C \beta_r \right\|_2^2 \qquad (r = 1,2), \\ \int_a^b \left(h_s''(U_s)\right)^2 \, \mathrm{d}U_s &\approx \sum_{j=1}^{N-1} \left(D_j^{s''} w_j \varphi_s\right)^2 = \left\| \overline{A}_s^D \varphi_s \right\|_2^2 \qquad (s = 1,2), \quad (51.16)\end{aligned}$$

where

$$\overline{A}_p^B := \begin{pmatrix} \left(B_1^p\right)'' u_1 \\ \left(B_2^p\right)'' u_2 \\ \vdots \\ \left(B_{N-1}^p\right)'' u_{N-1} \end{pmatrix}, \overline{A}_r^C := \begin{pmatrix} \left(C_1^r\right)'' v_1 \\ \left(C_2^r\right)'' v_2 \\ \vdots \\ \left(C_{N-1}^r\right)'' v_{N-1} \end{pmatrix},$$

$$\overline{A}_s^D := \begin{pmatrix} \left(D_1^s\right)'' w_1 \\ \left(D_2^s\right)'' w_2 \\ \vdots \\ \left(D_{N-1}^s\right)'' w_{N-1} \end{pmatrix}$$

and, for $j = 1,2,\ldots,N-1$,

$$u_j := \sqrt{U_{j+1,p} - U_{j,p}}, \quad v_j := \sqrt{U_{j+1,r} - U_{j,r}}, \quad w_j := \sqrt{U_{j+1,s} - U_{j,s}}.$$

Inserting the vector-matrix and approximative forms in (11), **PRSS** turns out to look as follows:

$$\mathbf{PRSS}(\theta, f, g, h) = \left\| \dot{\overline{X}} - \overline{A}\theta \right\|_2^2 + \sum_{p=1}^{2} \lambda_p \left\| \overline{A}_p^B \alpha_p \right\|_2^2 + \sum_{r=1}^{2} \mu_r \left\| \overline{A}_r^C \beta_r \right\|_2^2 + \sum_{s=1}^{2} \tau_s \left\| \overline{A}_s^D \varphi_s \right\|_2^2 . \quad (51.17)$$

However, we obtain a 6-tuple of *penalty* parameters: $\lambda = (\lambda_1, \lambda_2, \mu_1, \mu_2, \tau_1, \tau_2)^T$. Therefore, the minimization of **PRSS** is not yet already a Tikhonov regularization problem with its single such parameter. Thus, let us look at the case which is given by a uniform penalization by taking the same penalty factor $\lambda_p = \mu_r = \tau_s = \lambda =: \delta^2$ for each term. Then, our approximation of **PRSS** can be rearranged as

$$\mathbf{PRSS}(\theta, f, g, h) = \left\| \dot{\overline{X}} - \overline{A}\theta \right\|_2^2 + \delta^2 \left\| \overline{L}\theta \right\|_2^2 , \quad (51.18)$$

with the $(6(N-1) \times m)$-matrix

$$\overline{L} := \begin{pmatrix} 0 & \overline{A}_1^B & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \overline{A}_2^B & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \overline{A}_1^C & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \overline{A}_2^C & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \overline{A}_1^D & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \overline{A}_2^D \end{pmatrix}.$$

Herewith, based on the basis splines, we have identified the minimization of **PRSS** for some stochastic differential equation as a Tikhonov regularization problem [2]:

$$\min_m \; \|Gm - d\|_2^2 + \delta^2 \, \|Lm\|_2^2 \quad (51.19)$$

with penalty parameter $\lambda = \delta^2$. We note that this regularization task is a special kind of multi-objective optimization, and that in statistical learning this method is also known as ridge regression [19]. It is very helpful for problems whose exact solution does not exist or is not unique or not stable under perturbations (noise) of the data. MATLAB Regularization Toolbox can be used for the numerical solution [2].

51.4 Alternative Solution for Tikhonov Regularization Problem with Conic Quadratic Programming

51.4.1 Construction of the Conic Quadratic Programming Problem

As we just mentioned we can solve our Tikhonov regularization problem with MATLAB Regularization Toolbox. In addition, we shall explain how to treat our problem by using continuous optimization techniques which we suppose to become a complementary key technology and alternative to the concept of Tikhonov regularization. In particular, we apply the elegant framework of conic quadratic programming (CQP). Indeed, based on an appropriate, learning based choice of a bound M, we reformulate our Tikhonov regularization as the following optimization problem:

$$\begin{aligned}&\min_{\theta} \left\| \overline{A}\theta - \dot{\overline{X}} \right\|_2^2, \\ &\text{subject to } \left\| \overline{L}\theta \right\|_2^2 \leq M.\end{aligned} \tag{51.20}$$

Here, the objective function in (20) is not linear but quadratic. However, the original objective function can be moved to the list of constraints, and we can write an equivalent problem consisting of minimizing a new "height" variable t over the epigraph in the extended $(t, \theta)-$space as follows:

$$\begin{aligned}&\min_{t,\theta} t, \\ &\text{subject to } \left\| \overline{A}\theta - \dot{\overline{X}} \right\|_2^2 \leq t^2, \quad t \geq 0, \\ &\qquad\qquad \left\| \overline{L}\theta \right\|_2^2 \leq M,\end{aligned} \tag{51.21}$$

or

$$\begin{aligned}&\min_{t,\theta} t, \\ &\text{subject to } \left\| \overline{A}\theta - \dot{\overline{X}} \right\|_2 \leq t, \\ &\qquad\qquad \left\| \overline{L}\theta \right\|_2 \leq \sqrt{M}.\end{aligned} \tag{51.22}$$

Indeed, considering the form of a conic quadratic optimization problem [27]

$$\min_{x} c^T x, \quad \text{subject to} \quad \left\| D_i x - d_i \right\|_2 \leq p_i^T x - q_i \qquad (i = 1, 2, \ldots, k), \tag{51.23}$$

we can identify our optimization problem for parameter estimation in a SDE as a conic quadratic program with

$$c = \left(1\ 0_m^T\right)^T,\ x = \left(t\ \theta^T\right)^T,\ D_1 = \left(0_N, \overline{A}\right),\ d_1 = \dot{\overline{X}},\ p_1 = (1,0,\ldots,0)^T,$$
$$q_1 = 0,\ D_2 = \left(0_{6(N-1)}, L\right),\ d_2 = 0,\ p_1 = 0_{m+1}^T,\ q_1 = -\sqrt{M},$$
$$m = \sum_{p=1}^{2} d_p^g + \sum_{r=1}^{2} d_r^h + \sum_{s=1}^{2} d_s^f + 3.$$

In order to state the optimality conditions, we firstly reformulate our problem as

$$\begin{aligned} &\min_{t,\theta} t, \\ &\text{such that } \chi := \begin{pmatrix} 0_N & \overline{A} \\ 1 & 0_m^T \end{pmatrix} \begin{pmatrix} t \\ \theta \end{pmatrix} + \begin{pmatrix} -\dot{\overline{X}} \\ 0 \end{pmatrix}, \\ &\qquad\qquad \eta := \begin{pmatrix} 0_{6(N-1)} & L \\ 0 & 0_m^T \end{pmatrix} \begin{pmatrix} t \\ \theta \end{pmatrix} + \begin{pmatrix} 0_{6(N-1)} \\ \sqrt{M} \end{pmatrix}. \end{aligned} \tag{51.24}$$

Here, $\chi \in L^{N+1}$ and $\eta \in L^{6(N-1)+1}$, where L^{N+1} and $L^{6(N-1)+1}$ are the $(N+1)$- and $(6(N-1)+1)$- dimensional ice-cream (or second-order or Lorentz) cones, given by

$$L^{\nu} := \left\{ x = (x_1, x_2, \ldots, x_\nu)^T \in \mathbb{R}^\nu \,\middle|\, x_\nu \geq \sqrt{x_1^2 + x_2^2 + \ldots + x_{\nu-1}^2} \right\} \quad (\nu \geq 2).$$

Then, we can also write the dual problem to the latter problem as

$$\begin{aligned} &\max \left(\dot{\overline{X}}^T, 0\right) \kappa_1 + \left(0_{6(N-1)}^T, -\sqrt{M}\right) \kappa_2 \\ &\text{such that} \begin{pmatrix} 0_N^T & 1 \\ \overline{A}^T & 0_m \end{pmatrix} \kappa_1 + \begin{pmatrix} 0_{6(N-1)}^T & 0 \\ L^T & 0_m \end{pmatrix} \kappa_2 = \begin{pmatrix} 1 \\ 0_m \end{pmatrix}, \\ &\kappa_1 \in L^{N+1},\ \kappa_2 \in L^{6(N-1)+1}. \end{aligned} \tag{51.25}$$

Moreover, $(t, \theta, \chi, \eta, \kappa_1, \kappa_2)$ is the primal-dual optimal solution if the following constraints are provided in the corresponding ice-cream cones:

$$\begin{aligned} &\chi := \begin{pmatrix} 0_N & \overline{A} \\ 1 & 0_m^T \end{pmatrix} \begin{pmatrix} t \\ \theta \end{pmatrix} + \begin{pmatrix} -\dot{\overline{X}} \\ 0 \end{pmatrix}, \\ &\eta := \begin{pmatrix} 0_{6(N-1)} & L \\ 0 & 0_m^T \end{pmatrix} \begin{pmatrix} t \\ \theta \end{pmatrix} + \begin{pmatrix} 0_{6(N-1)} \\ \sqrt{M} \end{pmatrix}, \\ &\begin{pmatrix} 0_N^T & 1 \\ \overline{A}^T & 0_m \end{pmatrix} \kappa_1 + \begin{pmatrix} 0_{6(N-1)}^T & 0 \\ L^T & 0_m \end{pmatrix} \kappa_2 = \begin{pmatrix} 1 \\ 0_m \end{pmatrix}, \\ &\kappa_1^T \chi = 0,\ \kappa_2^T \eta = 0, \\ &\kappa_1 \in L^{N+1},\ \kappa_2 \in L^{6(N-1)+1}, \\ &\chi \in L^{N+1},\ \eta \in L^{6(N-1)+1}. \end{aligned} \tag{51.26}$$

Let us note that in the modeling of gene-environment and eco-finance networks, we received versions of CQP which are not formulated in our discrete (Gaussian) sums of squares but in the sum of squared uniform (maximal or Chebychev) errors. Such a kind of problems can be represented as semi-infinite programming (SIP) problems and, if the sets of inequalities even depend on the state variable, as generalized SIP problems [48,49].

51.4.2 On Solution Methods for Conic Quadratic Programming

For solving "well-structured" convex problems like conic quadratic problems but also linear programs, semidefinite programs and others, there are interior point methods (IPMs) which were firstly introduced by [21]. IPMs are barrier methods. Classically, they base on the interior points of the feasible set of the optimization problem; this set is assumed to be closed and convex. Then, an interior penalty (barrier) function $F(x)$ is chosen, well defined (and smooth and strongly convex) in the interior of the feasible set. This function is "blowing up" as a sequence from the interior approaches a boundary point of the feasible set [28]. Of great importance are primal-dual IPMs which refer to the pair of primal and dual variables.

The *canonical barrier function* for second-order cones $L^{\nu} = \left\{ x = (x_1, x_2, \ldots, x_{\nu})^T \in \mathbb{R}^{\nu} \,\middle|\, x_{\kappa} \geq \sqrt{x_1^2 + x_2^2 + \ldots + x_{\nu-1}^2} \right\}$ $(\nu \geq 2)$ are defined by $L_{\nu}(x) := -\ln\left(x_{\nu}^2 - x_1^2 - \cdots - x_{\nu-1}^2\right) = -\ln\left(x^T J_{\nu} x\right)$, with $J_{\nu} = \begin{pmatrix} -I_{\nu-1} & 0 \\ 0 & 1 \end{pmatrix}$. The parameter of this barrier is $\alpha(L_{\nu}) = 2$. IPMs have the advantage of employing the global structure of the problem, of allowing better complexity bounds and exhibiting a much better practical performance. For closer details we refer to [27,28].

51.4.3 On the Selection of Penalty Parameters and Upper Bounds

The choice of the penalty parameters in the penalized sum of squares (PRSS) and of the parametrical upper bound in a contraint of the conic quadratic problem is not a deterministic issue. The easiest case of minimizing PRSS is given by Tikhonov regularization, where we just have one penalty parameter and the integral terms discretized already; let us refer to that case in the sequel. As indicated above, the two parameters, related with our two approaches, can be chosen in some dependence and compatibility (equivalence) [2].

However, we prefer to draw and look at the so-called efficiency frontier (or efficiency curve) which is given by plotting the optimal solutions according to a larger (finite) number or parameter values, as points in the coordinate scheme with two axes where at one axis the complexity is denoted, whereas the other axis represents the length of the residual vector. We emphasize that each of these points is efficient. In case of Tikhonov regularization, logarithmical scales are employed such that some "kink" kind of a point on the efficiency boundary, called L-curve according to its more pronounced shape now, is caused; this point is considered to be closest to the origin and, together with the corresponding penalty parameter, it is often chosen [2].

For our approach with conic quadratic programming, in [20, 51], a lot of numerical experience is presented, related with varying upper bounds. In fact, numerous computations and efficiency frontiers are presented there, and many comparisons are made among alternative solutions gained by our approach, and with other methods from statistics and statistical learning. Until now, we employed the program packages of MOSEK, Salford MARS and special codes written in MATLAB. With our colleagues we are working on further improvements, e.g., in the selection of the knots of the splines via a preprocessing by clustering and in the selection (and reduction) of input features.

Those studies on credit default and quality management show that our mathematical approach is quite competitive and that it looks promising for the future use indeed, e.g., when becoming applied to real-world financial data from stock markets, but also from insurance companies. In fact, while for the stock exchange, SDEs and portfolio optimization are an important tool since about 30 years and in a process of continuous scientific advance, also in the actuarial sector, SDEs and portfolios are becoming more and more important in these years. Let us not forget about the emerging markets on, e.g., carbon trade, and the fast growing wide field of risk management. We further underline that our studies on quality analysis and control can also be applied to the products of the financial and related sectors, as actually rediscovered and recommended in these days of the financial crises of the years 2008–2009.

These possible applications will benefit from that we also consider nonlinearities in our parametrical dependencies. We approach the problem in a stepwise orientation to overcome nonlinearity which will be the subject of the next section.

51.5 On Nonlinear Dependence on Parameters and their Estimation

Let us return to (51.2) again, with two ways of generalization. (a) The model functions $a(\cdot)$ and $b(\cdot)$ may not only depend on the parameters which appear as coefficients in the linear combination with base splines, but also on really probabilistic (stochastic) parameters. (b) Differently from the earlier linear dependence on the parameters, the dependence on the newly considered parameters may be nonlinear. In that case, we should use any nonlinear parameter estimation methods like, e.g., Gauss–Newton's method or Levenberg–Marquardt's method [26].

Let us look at (a), for example, we consider following the SDE, Black–Scholes model with initial value:

$$\begin{cases} \mathrm{d}X_t = \mu X_t \,\mathrm{d}t + \sigma X_t \,\mathrm{d}W_t, \\ X(0) = x_0, \end{cases}$$

where $X_t = X(t)$ denotes the (random) price of a stock at time $t \geq 0$, and $\mu > 0$ and σ are parameters called the drift and volatility of the stock and x_0 is the starting price, respectively. Then, referring to the finitely many sample (data) points $\left(\overline{X}_\kappa, \overline{t}_\kappa\right)$ $(\kappa = 1, 2, \ldots, N)$ we obtain the Milstein scheme as

$$\dot{\overline{X}}_\kappa = \mu \overline{X}_\kappa + \sigma \overline{X}_\kappa \frac{\Delta W_\kappa}{\overline{h}_\kappa} + \frac{1}{2}\sigma^2 \left(P'P\right)\left(\overline{t}_\kappa\right)\left(\frac{(\Delta W_\kappa)^2}{\overline{h}_\kappa} - 1\right) = g\left(\overline{X}_\kappa, \mu, \sigma\right).$$

To determine the unknown values (μ, σ), we consider following optimization problem:

$$\min_{\beta} f(\beta) := \sum_{\kappa=1}^{N} \left(\dot{\overline{X}}_\kappa - g\left(\overline{X}_\kappa, \mu, \sigma\right)\right)^2 = \sum_{\kappa=1}^{N} f_\kappa^2(\beta) \quad \left(\text{or } \frac{1}{2}\sum_{\kappa=1}^{N} f_\kappa^2(\beta)\right). \tag{51.27}$$

Here, $\theta = (\mu, \sigma)^T$, $P(X) := X$, hence $P'\left(\overline{t}_\kappa\right) := 0$ (since P does not depend on t), and the objective function $f(\theta)$ of parameter estimation is defined linearly in auxiliary functions f_j squared $(j = 1, 2, \ldots, N)$. This problem representation holds true also if the quadratic term $1/2\sigma^2(P'P)\left(\overline{t}_\kappa\right)\left((\Delta W_\kappa)^2/\overline{h}_\kappa - 1\right)^2$ would not vanish and in many further examples where (b) the parametric dependence may be nonlinear indeed.

Nonlinear parametric dependence can occur by the composition of stochastic processes. For example, in financial modelling of the dynamics of wealth from time t to $t + \mathrm{d}t$ or maturity time T, V_t, may be given by

$$\begin{cases} \mathrm{d}V_t = \left(\left(\varpi_t^T (\mu - re) + r\right) V_t\right) \mathrm{d}t - c_t \,\mathrm{d}t + \varpi_t^T \sigma V_t \,\mathrm{d}W_t, \\ V_0 = \nu_0, \end{cases}$$

where ϖ_t is the fraction of wealth invested in the risky asset at time t and c_t is the consumption at time t. We can easily identify both $a\,(t, V_t, c_t, \varpi_t; r, \mu) := \left(\varpi_t^T (\mu - re) + r\right) V_t - c_t$ and $b\,(t, V_t, \varpi_t; \sigma) := \varpi_t^T \sigma V_t$. Here, r is the short-term interest rate, μ denotes the vector of expected rates of return, e is the vector consisting of ones, σ stands for the volatility matrix of the risky assets. The entire parameter $\theta := (r, \mu, \sigma)^T$ (arranged as a column vector) is assumed to be constant through time [1]. Finally, W is a Wiener process with the property that $\mathrm{d}W_t$ is $N(0, \mathrm{d}t)$ distributed. While the dependence of the right-hand side of the stochastic differential equation on θ is linear, nonlinear parametric dependencies can occur via the insertion of the processes c_t and ϖ_t in a and b, but also if r becomes a stochastic process r_t, e.g., in the following way. Namely, as a direct example of nonlinearity, the stochastic interest rate r_t for each $\tau \in \mathbb{R}$ may be given by

$$\mathrm{d}r_t = \alpha \cdot (R - r_t)\ \mathrm{d}t + \sigma_t \cdot r_t^\tau\ \mathrm{d}W_t,$$

where σ_t and W_t are volatility and a Brownian motion, respectively Here, α is a positive constant, and the drift term $\alpha \cdot (R - r_t)$ is positive for $R > r_t$ and negative for $R < r_t$ [38]. We denote $a\,(t, r_t; R) := \alpha \cdot (R - r_t)$ and $b\,(t, r_t, \sigma_t; \tau) := \sigma_t r_t^\tau$. This process on the interest rate can be attached to a price or wealth process. By this interest rate processes and the composition of stochastic processes, further parameters such as (R, τ), can implicitly and in a partially nonlinear way enter the interest rate dynamics r_t and processes beyond of that dynamics. After these first examples, we shall return to further examples and explanations in Sect. 51.6.

In fact, the financial sector with the modeling and prediction of stock prices and interest rate are the most prominent application areas here. Moreover, mixed linear-nonlinear dependencies on the parameters may be possible due to the linearly and the nonlinearly involved parameters of various kinds. This optimization problem (51.27) means a nonlinear least-squares estimation (or nonlinear regression). In the context of data fitting, each of the functions f_j corresponds to a residual in our discrete approximation problem which may arise in a mathematical modelling or in an inverse problem. Let us represent basic ideas of nonlinear regression theory with the help of [26].

Now, (51.27) can be represented in vector notation:

$$\min_\theta f(\theta) := \frac{1}{2} F^T(\theta) F(\theta), \tag{51.28}$$

where F is the vector-valued function $F(\theta) := (f_1(\theta), \ldots, f_N(\theta))^T$ $(\theta \in \mathbb{R}^p)$ and where the factor $1/2$ serves for a convenient normalization of the derivatives. In fact, by the chain rule we obtain

$$\nabla f(\theta) := \nabla F(\theta) F(\theta), \tag{51.29}$$

where $\nabla f(\theta)$ is an $(p \times N)$-matrix-valued function. By row-wise differentiation of $\nabla f(\theta)$ and using this gradient representation, we obtain the Hessian matrix of f:

$$\nabla^2 f(\theta) := \nabla F(\theta) \nabla F^T(\theta) + \sum_{j=1}^N f_j(\theta) \nabla^2 f_j(\theta). \tag{51.30}$$

Let θ^* be a solution of (51.27) and suppose $f\,(\theta^*) = 0$. Then, $f_j\,(\theta^*) = 0$ $(i = 1, 2, \ldots, N)$, i.e., all the residuals r_j vanish and the model fits the data without error. As a result, $F\,(\theta^*) = 0$ and, by (51.29), $\nabla f\,(\theta^*) = 0$, which just confirms our first-order necessary optimality condition. Furthermore, we can obtain the Hessian of f being $\nabla^2 f\left(\theta^*\right) = \nabla F\left(\theta^*\right) \nabla F^T\left(\theta^*\right)$, which is a positive semi-definite matrix, just as we expected by our second-order necessary optimality condition. In case where $\nabla F^T\left(\theta^*\right)$ is a matrix of full rank, i.e., $\operatorname{rank}\left(\nabla F^T\left(\theta^*\right)\right) = p$, then $\nabla f^2\left(\theta^*\right)$ is positive definite, i.e., second-order necessary optimality condition is provided such that θ^* is also a strict local minimizer.

From this basic idea, a number of specialized nonlinear least-squares methods come from. The simplest of this methods, called Gauss–Newton uses this approximative description in an indirect way. It makes a replacement of the Hessian in the formula

$$\nabla^2 f(\theta)\, q = -\nabla f(\theta), \tag{51.31}$$

such that we have relation

$$\nabla F(\theta) \nabla F^T(\theta)\, q = -\nabla F(\theta) F(\theta), \tag{51.32}$$

where q is Gauss–Newton increment $q = \theta_1 - \theta_0$. If $F\left(\theta^*\right) \approx 0$ and rank $\left(\nabla F\left(\theta^*\right)\right) = p(\leq N)$, then, near to a solution θ^*, Gauss–Newton behaves like Newton's method. However, we need not pay the computational cost of calculating second derivatives. Gauss–Newton's method sometimes behaves poor if there is one or a number of outliers, i.e., if the model does not fit the data well, or if rank $\left(\nabla F\left(\theta^*\right)\right)$ is not of full rank p. In these cases, there is a poor approximation of the Hessian.

Many other nonlinear least-squares methods can be interpreted as using an approximation of the second additive form in the formula for the Hessian i.e., of

$$\sum_{j=1}^{N} f_j(\theta) \nabla^2 f_j(\theta). \tag{51.33}$$

Levenberg–Marquardt's method uses the simplest of these approximation:

$$\sum_{j=1}^{N} f_j(\theta) \nabla^2 f_j(\theta) \approx \lambda I_p, \tag{51.34}$$

with some scalar $\lambda \geq 0$. This approximation yields the following linear system:

$$\left(\nabla F(\theta) \nabla^T F(\theta) + \lambda I_p\right) q = -\nabla F(\theta) F(\theta). \tag{51.35}$$

Often, ones can find Levenberg–Marquardt method implemented in the context of a trustregion algorithm. There, q is obtained, e.g., by minimizing a quadratic model of the objective function with Gauss–Newton approximation of the Hessian:

$$\begin{cases} \min\limits_{q} Q(q) := f(\theta) + q^T \nabla F(\theta) F(\theta) + 1/2 q^T \nabla F(\theta) \nabla^T F(\theta) q \\ \text{subject to } \|q\|_2 \leq \Delta. \end{cases} \tag{51.36}$$

Here, λ is indirectly determined by picking a value of $\varDelta$. The scalar $\varDelta$ can be chosen based on the effectiveness of the Gauss–Newton method.

Levenberg–Marquardt method can be interpreted as a mixture between Gauss–Newton method (if $\lambda \approx 0$) and steepest-descent method (if λ is very large) [2, 26]. An adaptive and sequential way of choosing λ and, by this, of the adjustment of mixture between the methods of Gauss–Newton and steepest-descent, is presented in [26]. We note that the term "I_p" can also be regarded as a regularization term that shifts the eigenvalues of $\nabla F(\theta)\nabla^T F(\theta)$ away from 0.

Another way to solve the system (51.35) for given $\theta = \theta_k$, i.e., to find the $(k+1)$-st iterate $q = q_k$, consists in an application of least-squares estimation. If we denote (51.35) by $Gq = d$, where $G = \nabla F(\theta)\nabla^T F(\theta) + \lambda I_p$ and $d = -\nabla F(\theta)F(\theta)$, then we can study the regularized problem by adding to the squared residual norm $\|Gq - d\|_2^2$ a penalty or regularization term of the form $\delta^2 \|Lq\|_2^2$, i.e.,

$$\min_q \left\| \left(\nabla F(\theta)\nabla^T F(\theta) + \lambda I_N\right) q - (-\nabla F(\theta)F(\theta)) \right\|_2^2 + \delta^2 \|Lq\|_2^2, \tag{51.37}$$

where L may be the unit matrix, but it can also represent a discrete differentiation of first or second order. This regularization serves to diminish the complexity of the model. We recall [2] for closer explanation about this Tikhonov regularization. But instead of the penalization approach, we can again bound the regularization term $\|Lq\|_2^2$ by an inequality constraint. What is more, we can turn the optimization problem to a CQP problem in order to find the step q_k and, herewith, the next iterate $\theta_{k+1} := \theta_k + q_k$. By this conic quadratic modelling and solution technique we are back in the methodology that we presented in Sect. 51.4. Indeed, with a suitable and maybe adaptive choice of an upper bound M_1 [19, 39, 41] we can write our problem as

$$\begin{aligned} &\min_q \left\| \left(\nabla F(\theta)\nabla^T F(\theta) + \lambda I_N\right) q - (-\nabla F(\theta)F(\theta)) \right\|_2^2, \\ &\text{subject to } \|Lq\|_2^2 \leq M_1, \end{aligned} \tag{51.38}$$

or we can write an equivalent problem as follows:

$$\begin{aligned} &\min_{t,q} t, \\ &\text{subject to } \left\| \left(\nabla F(\theta)\nabla^T F(\theta) + \lambda I_p\right) q - (-\nabla F(\theta)F(\theta)) \right\|_2^2 \leq t^2, \\ &\qquad\qquad t \geq 0, \ \|Lq\|_2^2 \leq M_1. \end{aligned}$$

If we consider the general problem form [27]

$$\min_x c^T x, \qquad \text{subject to} \quad \|D_i x - d_i\|_2 \leq p_i^T x - q_i \quad (i = 1, 2, \ldots, k),$$

we can identify our optimization problem for determining step length q as a conic quadratic program again, namely, with

$$c = \left(1\ 0_p^T\right)^T, x = \left(t, q^T\right)^T, D_1 = \left(0_A, \overline{A}\right)^T, d_1 = -\nabla F(\beta)F(\beta),$$
$$p_1 = (1, 0, \ldots, 0)^T, q_1 = 0, D_2 = \left(0_p, L_{p\times p}\right)^T, d_2 = 0_p, p_2 = 0_{p+1} \text{ and}$$
$$q_2 = -\sqrt{M_1}.$$

As announced earlier, we are now giving further examples of SDE by different classes of interest rate models.

51.6 Continuous Time One-Factor Interest Rate Models

Various stochastic differential equations used to model a short term interest rate can be depicted in the following general form:

$$\mathrm{d}r_t = (\alpha + \beta r_t)\ \mathrm{d}t + \sigma r_t^{\gamma}\ \mathrm{d}W_t, \tag{51.39}$$

where r_t $(= r(t))$ is a real continuous time process for an interest rate, and α, β, σ and γ are the unknown parameters to be estimated. Some of the models of this form can be summarized as in Table 51.2.

Nowman assumes as an approximation to the true underlying model given by (51.39) that over the interval $[0, T]$, $r(t)$ satisfies the stochastic differential equation [50]

$$\mathrm{d}r_t = (\alpha + \beta r(t))\ \mathrm{d}t + \sigma\left(r\left(t' - 1\right)\right)^{\gamma}\ \mathrm{d}W_t, \tag{51.40}$$

where $t' - 1$ is the largest integer less than t (i.e., t' is the smallest integer greater than or equal to t). He also assumes that, in (51.40), the volatility of the interest rate changes at the beginning of the unit observation period and then remains constant.

Table 51.2 Nonlinear one-factor interest rate models [31]

Merton (1973) [25]	$\mathrm{d}r_t = \alpha\ \mathrm{d}t + \sigma\ \mathrm{d}W_t$
Vasicek (1977) [46]	$\mathrm{d}r_t = (\alpha + \beta r_t)\ \mathrm{d}t + \sigma\ \mathrm{d}W_t$
Cox, Ingersoll and Ross (1985) [13]	$\mathrm{d}r_t = (\alpha + \beta r_t)\ \mathrm{d}t + \sigma r_t^{1/2}\ \mathrm{d}W_t$
Dothan (1978) [14]	$\mathrm{d}r_t = \sigma r_t\ \mathrm{d}W_t$
Black–Scholes model	$\mathrm{d}r_t = \beta r_t\ \mathrm{d}t + \sigma r_t\ \mathrm{d}W_t$
Brennan and Schwartz (1980) [8]	$\mathrm{d}r_t = (\alpha + \beta r_t)\ \mathrm{d}t + \sigma r_t\ \mathrm{d}W_t$
Cox, Ingersoll and Ross (1980) [12]	$\mathrm{d}r_t = \sigma r_t^{3/2}\ \mathrm{d}W_t$
Constant Elasticity of Variance	$\mathrm{d}r_t = \beta r_t\ \mathrm{d}t + \sigma r_t^{\gamma}\ \mathrm{d}W_t$

This is an improvement of Bergstrom's results [3–5], where he assumes that the conditional second moment is constant over time.

Then, (51.40) has the following stochastic integral representation [50]:

$$r(t) - t\left(t' - 1\right) = \int_{t'-1}^{t} (\alpha + \beta r(s))\ \mathrm{d}s \\ + \sigma\left(r\left(t' - 1\right)\right)^{\gamma} \int_{t'-1}^{t} \mathrm{d}W(s) \\ \text{for all} \quad t \in \left[t' - 1, t'\right], \tag{51.41}$$

where $\int_{t'-1}^{t} \mathrm{d}W(s) = W\left[t' - 1, t'\right] := W(t) - W\left(t' - 1\right)$. Following [4], Nowman [31] is able to write the discrete version of (51.41) as

$$r(t) = \mathrm{e}^{\beta} r(t-1) + \frac{\alpha}{\beta}\left(\mathrm{e}^{\beta} - 1\right) + \eta_t, \tag{51.42}$$

where the noise term η_t $(t = 1, 2, \ldots, T)$ satisfies the following moment conditions

$$E\left[\eta_s, \eta_t\right] = 0 \qquad \text{for} \quad s \neq t,$$

$$E\left[\eta_t^2\right] = \int_{t-1}^{t} \mathrm{e}^{2(s-\tau)\beta} \sigma^2 (r(s-1))^{2\gamma}\ \mathrm{d}s \\ = \frac{\sigma^2}{2\beta}\left(\mathrm{e}^{2\beta} - 1\right)(r(t-1))^{2\gamma} = m_{tt}^2.$$

We conclude the excursion of this section by noting that maximum-likelihood (ML) kind of optimization problem on parameter estimation is to maximize the log-likelihood function, which looks as follows:

$$\min_{\alpha,\beta,\sigma,\gamma} L(\alpha, \beta, \sigma, \gamma) := \sum_{t=1}^{T} \left(2\log\left(m_{tt}\right) + \left(r(t) - \mathrm{e}^{\beta} r(t-1) \right.\right. \\ \left.\left. - \frac{\alpha}{\beta}\left(\mathrm{e}^{\beta} - 1\right) \right)^2 / m_{tt}^2 \right)$$

and by recommending the methods presented in our paper for possible application on the great variety of SDEs, independently of any such a special form of ML function – to those mentioned in this section, and much beyond. The study on prediction of credit-default risk [20] already showed the value of our additive model approach. Indeed, further combined applications on real-world data from areas of finance, science and technology may be expected, where our contribution can be utilized.

51.7 Further Nonparametric Methods

The latest work implementing nonparametric approach in estimating the structural parameter of SDE was done in [45]. In [45], an Approximate Maximum Likelihood Estimation (AMLE) is developed for estimating the states and parameters of models described by stochastic differential equations (SDE). Using the initial estimate of Q –the process disturbance intensity and assuming that the σ^2-measurement-noise variance is known– the estimate of $\left(\hat{\Theta}, \hat{\beta}\right) = \arg\min_{(\Theta,\beta)} T$ in the first step, where

$$T = \frac{(y_m - x_{Bm})^T (y_m - x_{Bm})}{2\sigma_m^2} + \frac{1}{2Q}\int_{t_0}^{t_q} (\dot{x}_B(t) - f(x(t), u(t), \Theta))^2 \, \mathrm{d}t \tag{51.43}$$

is the negative natural logarithm of the probability density function.

Here, y_m is the vector of outputs at observation times, x_{Bm} is B-spline expansion of the input variable, $\dot{x}_B$ is the differentiation of the B-spline expansion, $f(x(t), u(t), \Theta)$ is the nonlinear function of the state variable, $\hat{\Theta}$ is the estimated model parameter and $\hat{\beta}$ are the spline coefficients. Later, evaluate $\hat{\sigma}_m^2$ using x_{Bm} in the second step and obtain a new estimate of Q in the final step.

The two step method in estimating the structural parameter of the ordinary differential equation (ODE) is initiated in [44]. It fits the observed data with cubic spline functions in the first step and estimates the parameters in a second step by finding the least squares solution of the differential equation sampled at a set of points.

In [37] it is considered a two-step approach in a functional data analysis (FDA) framework. It is based on the transformation of data into functions with smoothing cubic splines. The paper [35] proposes principal differential analysis (PDA) and using the basis function such as B-splines to estimate the parameters of ODE. The extension of PDA is done by applying it to nonlinear ODE and the iterated PDA (iPDA), thus repeating the two-steps method in [34]. The iPDA has been extended with the introduction of generalized smoothing approach [36] where the smoothing and estimation of ODE parameters are done simultaneously. The paper [36] proposes a generalized profiling procedure which is a variant of the collocation method based on basis function expansion in the form of a penalized log-likelihood criterion

$$J(c \mid \Theta, \sigma, \lambda) = -\sum_{i \in I} \ln\left(g\left(e_i, \Theta, \lambda\right)\right) + \mathrm{PEN}\left(\hat{x} \mid \lambda\right) \tag{51.44}$$

or the least squares criterion

$$J(c \mid \Theta, \sigma, \lambda) = -\sum_{i \in I} w_i \left\| y_i - \hat{x}_i(t) \right\|_2 + \mathrm{PEN}\left(\hat{x} \mid \lambda\right). \tag{51.45}$$

Here, the first term at the RHS of (51.45) is the data fitting criteria and the second term is the equation fidelity criteria with $\mathrm{PEN}\left(\hat{x} \mid \lambda\right) = \int L_{i,\Theta}\left(\hat{x}_i(t)\right)^2 \, \mathrm{d}t$, $L_{i,\Theta}\left(x_i\right) = \hat{x}_i - f_i\left(\hat{x}, \mu, t \mid \Theta\right)$, $\hat{x}_i$ is the spline function, f_i is the model function of the corresponding ODE. The method is applied using noisy measurements on a subset of variables to estimate the parameters defining a system of nonlinear differential equations. For simulated data from models in chemical engineering, the authors derive the point estimates and the confidence interval and show that these have low bias and good coverage properties. The method has also been applied to real data from chemistry and from the progress of the autoimmune disease lupus. Referring to [9,36] proposes a general method of estimating the parameters of ODE from time series data. Brunel [9] uses the nonparametric estimator of the regression function as a first step to construct the M-estimator minimizing

$$R_w^2(\Theta) = \left\| \dot{\hat{x}} - F\left(t, \hat{x}_n, \Theta\right) \right\|_{2,w}, \tag{51.46}$$

where $\dot{\hat{x}}$ is the derivative of the nonparametric estimator of the solution of ODE and $F\left(t, \hat{x}_n, \Theta\right)$ is the ODE. The method is able to alleviate computational difficulties encountered by the classical parametric method. The authors also show the consistency of the derived estimator $\hat{\Theta}$. In the case of spline estimators the authors prove the asymptotic normality and the rate of convergence of the parametric estimators.

51.8 Summary on Studies in Research Groups

In the first part of the research we estimated the drift and diffusion parameters of the stochastic logistic models. The parameters of the drift equation were estimated via fourth-order Runge–Kutta method for deterministic models. The values of the parameter with the least MSE is then utilized to estimate the diffusion parameters in stochastic models employing Milstein discretization. The objective function is then minimized by Levenberg–Marquardt's method. This is considered as parametric method which the result, i.e., the model obtained will be compared with those obtained via nonparametric methods.

In the second part of the research we want to develop a new criterion based on the nonparametric approach of a two step method in estimating the drift and diffusion parameters Θ and σ of stochastic models. We have three alternatives, i.e., by proposing a totally new criterion for simultaneous estimation of both parameters [36], partially estimate of the parameters [9] or, lastly, Varziri's estimation [45] with some modifications.

Another research in this group is by Norhayati Rosli. Part of her research involving the comparison of the performance of numerical methods in SDE is the approximation of the strong solution of SDE. Based on [11] which presented and analyzed Stochastic Runge–Kutta models (SRK), i.e., 2-stage SRK and 4-stage SRK, Norhayati compared the performance of 2-stage SRK and Euler–Maruyama in approximating the strong solutions of the stochastic power law logistic model in describing the growth of C.Acetobutylicum. The performance of both methods were compared, based on their global error. Then, she compared the discretization scheme for delay stochastic differential equations.

Lastly, we mention the work by Mohd. Khairul Bazli using the stochastic power law logistic model to describe the growth of C.Acetobutylicum. The model parameters were estimated by using simulated maximum likelihood and the solution to the power law logistic model was approximated by using Euler–Maruyama.

51.9 Concluding Remarks

Discretization and additive models based on splines which defining a trilevel problem consisting of an optimization and a representation problem (portfolio optimization), and a parameter estimation, are a key approach to approximate stochastic differential equations. Furthermore, continuous optimization techniques make it possible to use highly efficient IPMs.

This paper gave a new contribution to problems related with SDEs using regression under an additive or a nonlinear model, as a preparatory step on the way of organizing assets in terms of portfolios. Indeed, we explained the relations to portfolio optimization, especially, to the martingale method. We made modern methods of inverse problems and continuous optimization, especially, CQP and methods from nonlinear regression, accessible and usable. By this, a bridge has been offered between statistical learning and data mining on the one hand, and the powerful tools prepared for well-structured convex optimization problems [7,27], and Newton- and steepest-descent type regression methods [2,26] on the other hand.

We hope that future research, theoretical and applied achievements on this fruitful interface will be stimulated by our paper.

Acknowledgements The authors express their gratitude to Professor Alberto Pinto for his invitation to prepare this chapter in a book in honour of Prof. Dr. Mauricio Peixoto and Prof. Dr. David Rand.

References

1. Akume, D.: Risk Constrained Dynamic Portfolio Management. PhD Thesis (2007)
2. Aster, A., Borchers, B., Thurber, C.: Parameter Estimation and Inverse Problems. Academic, New York (2004)
3. Bergstrom, A.R.: Gaussian estimation of structural parameters in higher-order continuous time dynamic models. Economica **51**, 117–152 (1983)
4. Bergstrom, A.R.: Continuous time stochastic models and issues of aggregation over time. In: Grilliches, Z., Intrilligator, M.D. (eds.) Handbookof Econometrics, vol. II. Elsevier Science, Amsterdam (1984)
5. Bergstrom, A.R.: The estimation of open higher-order continuous time dynamic models with mixed stock and flow data. Econom. Theory **2**, 350–373 (1986)
6. Bishwal, J.P.N.: Parameter Estimation in Stochastic Differential Equation. Springer, New York (2008)
7. Boyd, S., Vandenberghe, L.: Convex Optmization. Camebridge University Press, Camebridge (2004)

8. Brennan, M.J., Schwartz, E.S.: Conditional predictions of bond prices and returns. J. Finan. **35**, 405–417 (1980)
9. Brunel, N.: Parameter estimation of ODE's via nonparametric estimators. Electronic. J. Stat. **2**, 1242–1267 (2008)
10. Buja, A., Hastie, T., Tibshirani, R.: Linear smoothers and additive models. Ann. Stat. **17**(2), 453–510 (1989)
11. Burrage, K., Burrage, P.M.: High strong order explicit Runge-Kutta methods for stochastic ordinary differential equations. Appl. Numer. Math. **22**, 81–101 (1996)
12. Cox, J., Ingersoll, J., Ross, S.: An analysis of variable rate loan contracts. J. Finan. **35**(2), 389–403 (1980)
13. Cox, J., Ingersoll, J., Ross, S.: A theory of the term structure of interest rates. Econometrica **53**(2), 385–407 (1985)
14. Dothan, L.U.: On the term structure of interest rates. J. Finan. Econom. **6**, 59–69 (1978)
15. De Boor, C.: Practical Guide to Splines. Springer, Berlin (2001)
16. Friedman, J.H., Stuetzle, W.: Projection pursuit regression. J. Am. Stat. Assoc. **76**, 817–823 (1981)
17. Gutiérrez, R., Gutiérrez-Sánchez, R., Nafidi, A.: Trend analysis and computational estimation in a stochastic Rayleigh model: Simulation and application. Math. Comput. Simul. **77**, 209–217 (2008)
18. Hastie, T., Tibshirani: Generalized additive models. Statist. Science **1**, 3, 297–310 (1986)
19. Hastie, T., Tibshirani, Friedman, J.H.: The Elements of Statistical Learning. Springer Verlag, New York (2001)
20. Işcanoğlu Çekiç, A., Weber, G.W., Taylan, P.: Predicting default probabilities with generalized additive models for emerging markets. Invited Lecture, Graduate Summer School on New Advances in Statistics, METU, August 11–24 (2007) available at http://144.122.137.55/gweber/
21. Karmarkar, N.: A new polynomial time algorithm for linear programming. Combinatorica **4**(4), 373–395 (1984)
22. Kloeden, P.E., Platen, E., Schurz, H.: Numerical Solution of SDE Through Computer Experiments. Springer, New York (1994)
23. Korn, R., Korn, E.: Options Pricing and Portfolio Optimization: Modern Methods of Financial Mathematics. Oxford University Press, Oxford (2001)
24. Lo, B.P., Haslam, A.J., Adjiman, C.S.: An algorithm for the estimation of parameters in models with stochastic differential equations. Chem. Eng. Sci. **63**, 4820–4833 (2008)
25. Merton, R.C.: An intertemporal capital asset pricing model. Econometrica **41**(5), 867–887 (1973)
26. Nash, G., Sofer, A.: Linear and Nonlinear Programming. McGraw-Hill, New York (1996)
27. Nemirovski, A.: Lectures on modern convex optimization. Isreal Istitute of Technology (2002). available at http://iew3.technion.ac.il/Labs/Opt/ opt/LN/Final.pdf
28. Nesterov, Y.E., Nemirovskii, A.S.: Interior Point Methods in Convex Programming. SIAM, Philadelphia (1993)
29. Nielsen, J.N., Madsen, H.: Applying the EKF to stochastic differential equations with level effects. Automatica **37**(1), 107–112 (2001)
30. Nielsen, J.N., Madsen, H., Young, P.C.: Parameters Estimation in Stochastic Differential Equations: An Overview. Annu. Rev. Contr. **24**, 83–94 (2000)
31. Nowman, K.B.: Gaussian estimation of single-factor continuous time models of the term structure of interest rates. J. Sci. **52**, 1695–1706 (1997)
32. Øksendal, B.K.: Stochastic Differential Equations: An Introduction with Applications. Springer, Berlin (2003)
33. Picchini,U., Ditlevsen, S., Gaetano, A.D.: Modelling the Euglycemic Hyperinsulinemic Clamp by Stochastic Differential Equations. J. Math. Biol. **53**(5), 771–796 (2006)
34. Poyton, A.A, Varziri, M.S., McAuley, K.B., McLellan, P.J, Ramsay, J.O.: Parameter estimation in continuous-time dynamic models using principal differential analysis. Comput. Chem. Eng. **30**, 698–708 (2006)

35. Ramsay, J.O.: Principal differential analysis: Data reduction by differential operators. J. R. Stat. Soc. Ser. B **58**, 495–508 (1996)
36. Ramsay, J.O., Hooker, G., Campbell, D., Cao, J.: Parameter Estimation for differential equations: A Generalized Smoothing Approach. J. R. Stat. Soc. Ser. B **69**(5), 741–796 (2007)
37. Ramsay, J.O., Silverman, B.W.: Functional Data Analysis. Springer, Berlin (1997)
38. Seydel, R.U.: Tools for Computational Finance. Springer, New York (2003)
39. Taylan, P., Weber, G.W.: New approaches to regression in financial mathematics by additive models. J. Comput. Technol. **12**(2), 3–22 (2007)
40. Taylan, P., Weber, G.W.: Organization in finance prepared by stochastic differential equations with additive and nonlinear models and continuous optimization. Organizacija **41**(5), September-October, 185–193 (2008)
41. Taylan, P., Weber, G.W., Beck, A.: New approaches to regression by generalized additive, models and continuous optimization for modern applications in finance, science and technology. Optimization **56**(5–6), 675–698 (2007)
42. Taylan, P., Weber, G.W., Yerlikaya, F.: Continuous optimization applied in MARS for modern applications in finance, science and technology. ISI Proceedings of the 20th Mini-EURO Conference Continuous Optimization and Knowledge-Based Technologies, pp. 317–322. Neringa, Lithuania, May 20–23 (2008)
43. Taylan, P., Weber, G.W., Kropat, E.: Approximation of stochastic differential equations by additive models using splines and conic programming. In: D.M. Dubois (ed.). Proceedings of CASYS07, Eighth International Conference on Computing Anticipatory Systems, Liege, Belgium, August 6–11 (2007)
44. Varah, J.M.: A Spline Least Squares Method for Numerical Parameter Estimation in Differential Equations. SIAM J. Sci. Stat. Comput. **3**(1), 28–46 (1982)
45. Varziri, M.S., McAuley, K.B, McLellan, P.J.: Parameter and State Estimation in Nonlinear Stochastic Continuous-Time Dynamic Models With Unknown Disturbance Intensity. Can. J. Chem. Eng. **86**, 828–837 (2008)
46. Vasicek, O.: An equilibrium characterisation of the term structure. J. Finan. Econom. **5**, 177–188 (1977)
47. Weber, G.W.: On the topology of parametric optimal control. J. Aust. Math. Soc. Ser. B **39**, 1–35 (1997)
48. Weber, G.W.: On the topology of generalized semi-infinite optimization. In: special issue Optimization of *Journal of Convex Analysis* **9**(2), 665–691 (2002)
49. Weber, G.W.: Generalized Semi-Infinite Optimization and Related Topics. In: Hofmannn, K.H., Wille, R. (eds.) Heldermann publishing house, Research and Exposition in Mathematics, vol. 29. Lemgo (2003)
50. Weber, G.W., Taylan, P., Yıldırak, K., Görgülü, Z.-K.: Financial Regression and Organization. In: Special Issue on Optimization in Finance, *DCDIS-B* (2010)
51. Yerlikaya, F.: A New Contribution to Nonlinear Robust Regression and Classification with MARS and Its Application to Data Mining for Quality Control in Manufacturing. MSc. Thesis, Middle East Technical University, Ankara (2008)

Chapter 52
Stochastic Saddle Paths and Economic Theory

A.N. Yannacopoulos

Abstract Stochastic saddle paths appear in a number of important applications in economic theory, the most prominent of which are economic control theory and rational expectations theory. We provide a formulation of stochastic saddle paths in terms of forward backward stochastic differential equations and through this establish the existence and persistence of saddle paths under the influence of noise, provide qualitative results on the form and structure of the stable saddle path, approximation schemes as well as controllability related results. The general framework is illustrated by examples from economic theory and economic policy.

52.1 Introduction

Saddle points and saddle paths are extremely important concepts for the modern theory of dynamical systems as has been shown in the pioneering works of M. Peixoto [15] and their existence and properties provide interesting qualitative information such as for instance structural stability, occurence of erratic quasi-random behaviour (called chaos) [5] etc. Apart from their importance in theoretical studies they play an important role in applications either in the physical sciences or in economics. From the late 1970s it has been made clear by applied mathematicians such as D. Rand [17] that even simple economic dynamical systems like simple oligopoly models may exhibit highly nontrivial dynamic behaviour. Observations of this type triggered the study of economic systems using the tools of modern dynamical systems theory, in which saddles have a prominent role, and has led to interesting developments in the field (see e.g. [4]).

A.N. Yannacopoulos
Athens University of Economics and Business, 76 Patission Str, Athens 11434, Greece
e-mail: ayannaco@aueb.gr

M.M. Peixoto et al. (eds.), *Dynamics, Games and Science II*, Springer Proceedings in Mathematics 2, DOI 10.1007/978-3-642-14788-3_52,

The aim of the present work is to show a connection of saddle paths with a general class of models in economic theory which incorporate the effect of future expectations on the actions of economic agents today. Such models are very popular in economic theory; as an example one may take rational expectations models which form the basis of modern macroeconomic theory. The effect of expectations concerning future states in economics has been recognised early in its history and their prominent role in the determination of the current state makes economics distinct from the natural sciences. Most of the models currently used in the macroeconomic literature have a general form that consists of two sets of variables, a set of variables what is called fundamentals of the economy and a set of variables that is called assets; denoted in this paper by x and y respectively. The current value of the assets depends on the expectations of the future values of the fundamentals and then through a feedback mechanism consistent with the adopted framework of economic theory, this also has an effect on the current value of the fundamentals themselves. It is desirable, that the structure of the model is such that this feedback mechanism in the long run drives the system to a desired state, the equilibrium state. This is usually the state where the economy functions in the "optimal" way according to the criteria set by the policy makers.

The general mathematical structure of such models, at least in the absence of uncertainty (total forthsight models) usually have a saddle point structure in the appropriate phase space, and the stable saddle path usually corresponds to the path the will lead the economy in the long run to the desired equilibrium state. The determination of this stable saddle path in very important as far as economic policy making is concerned and in many models a passive control rule which is usually expressed as a relationship between different variables of the model is needed to guarantee the existence of such a stable saddlepath. For instance, the Taylor rule used by central banks for the determination of the interest rate in nothing else but such an attempt to endow the commonly accepted dynamic model for the evolution of the state of the system with such a stable saddle path [18].

While, many of the economic models incorporate the effects of uncertainty, the situation becomes less clear from the dynamical systems point of view when this is done, as the concept of phase space and many of the geometric and topological arguments used to understand the dynamics in deterministic systems, are no longer readily applicable. To provide a link with the theory of dynamics as understood by the dynamical systems community with the problem at hand we thus propose an alternative formulation of stochastic economies with expectations feedback, in terms of the concept of forward backward stochastic differential equations (FBSDEs). We show that this alternative formulation allows us to redefine within the stochastic framework many of the desired features of such a model, thus providing us with existence results for the stochastic analogue of the stable saddle paths, information concerning their qualitative properties, uniqueness as well as approximation schemes and control procedures for design of policy to drive the system to a desired state. The general framework is illustrated using examples from economic theory.

52.2 Stochastic Saddle Paths in Continuous Time and Their Connection with FBSDEs

Miller and Weller [11] introduced a linear model that incorporates the effects of expectations of future states in the process of equilibration of a stochastic economy. The model has the basic structure of a rational expectations model, in which future expectations are costantly updated using a common information set available to all agents, in such a way as to direct the economy in the long run to a desired state. The deterministic analogue of the model has the familiar saddle point structure common to many "near equilibrium" models in economic theory.

The nonlinear generalization of the Miller and Weller model can be formulated as follows

$$\begin{aligned} dx(t) &= b(t,x(t),y(t))dt + \sigma(t,x(t),y(t))dW(t) \\ y(t) &= E\left[\int_t^\infty g(s,x(s),y(s))e^{-\int_t^s \delta(r,x(r),y(r))\,dr}\,ds \mid \mathscr{F}_t\right] \end{aligned} \tag{52.1}$$

where $x \in R^n$, $y \in R^m$, $W(t)$ is an $s-$dimensional Wiener process assumed as a model of the stochastic factors driving the economy, and $\mathscr{F}_t = \sigma(W_s, s \le t)$.

We assume without loss of generality that the desired equilibrium state of the economy is $x^* = y^* = 0$. The functions $b(t,x,y) : R^+ \times R^n \times R^m \to R^n$, $g(x) : R^n \to R^m$ are nonlinear functions. In the special case $n = m = 1$, $b(t,x,y) = \alpha x + \beta y$, $g(x) = -\gamma x$ we retrieve the linear model of Miller and Weller [11]. The introduction of the nonlinear function $g(x)$ in the conditional expectation describing the asset dynamics models some saturation effects. The value of the asset is some rational expectation of the deviation of the fundamental from equilibrium but its value may not keep on growing unboundedly as the deviation of the fundamental from equilibrium grows. The above arguments can be formalized by imposing the following standing assumptions on the functions $b(t,x,y)$, $g(t,x,y)$, $\delta(t,x,y)$:

1. $b(t,x,y)$ is globally Lipschitz continuous in (t,x,y) and bounded
2. $g(x)$ is Lipschitz continuous and bounded
3. $\delta \ge C > 0$

One of the major problems when introducing stochastic effects in the model is to define properly the concept of the saddle path. This can be achieved by adopting the proper functional setup for the problem. We define the Banach space

$$\begin{aligned} M_X := \{r(\cdot) \mid\ &\text{stochastic processes with values in } X \\ &\text{s.t. } E\left[\int_0^\infty \| r(s) \|_X^2\ ds\right] < \infty\right\}. \end{aligned}$$

In our case X is a finite dimensional space (either R^n, R^m or $R^{m\times s}$). We will therefore define a saddle path as follows

Definition 52.1. A stable saddle path for system (52.1) is a stochastic process $(x(\cdot), y(\cdot)) \in M_X \times M_Y$, for $X = R^n, Y = R^m$, that satisfies (52.1).

This choice of functional setup guarantees that the state of the system approaches the desired equilibrium state $x^* = 0$, $y^* = 0$ asymptotically in time.

We now present a convenient reformulation of the problem in terms of infinite horizon FBSDEs.

Proposition 52.1. *The Miller and Weller model (52.1) is equivalent to the infinite horizon FBSDE*

$$\begin{aligned} dx(t) &= b(t, x(t), y(t))dt + \sigma(t, x(t), y(t))dW(t) \\ dy(t) &= (-g(t, x(t), y(t)) + \delta(t, x(t), y(t))\, y(t))dt - (z(t), dW(t)) \\ x_0 &= x, \ \ (x(\cdot), y(\cdot), z(\cdot)) \in M_X \times M_Y \times M_Z \end{aligned} \tag{52.2}$$

with $X = R^n$, $Y = R^m$, $Z = R^{m\times s}$, *where* $z(\cdot)$ *is an* $\mathscr{F}_t$*-adapted stochastic process which is to be determined.*

Proof. The proof uses technical tools from stochastic analysis and in particular the martingale representation theorem. We refrain from providing a complete proof which essentially follows the lines of [20]. □

Note that there is only an initial condition for x_t but a condition at infinity for y_t. This turns (52.2) into a stochastic boundary value problems where the fundamentals play the role of the forward variable and the assets play the role of the backward variable. The reformulation of the stochastic saddlepath problem as an infinite horizon FBSDE provides us with a great arsenal of technical tools from stochastic analysis by which we may study the several well-posedeness problems as well as qualitative problems that arise in economic modelling.

52.3 FBSDEs: A Brief Survey

The theory of FBSDEs is a relatively new and exciting field in the theory of stochastic differential equations. An FBSDE is essentially a stochastic boundary value problem for the evolution of a stochastic process $(x(t), y(t))$, the x component of which is defined by its initial condition, whereas the y component of which is defined through a final condition. In other words a general FBSDE is a system of the form

$$\begin{aligned} dx(t) &= b(t, x(t), y(t), z(t))\, dt + \sigma(t, x(t), y(t), z(t))\, dW(t) \\ dy(t) &= h(t, x(t), y(t), z(t))\, dt - z(t)\, dW(t) \\ x(0) &= x, \ \ y(T) = g(x(T)) \end{aligned}$$

for some prescribed function $g(x)$. Eventhough there are only two sets of evolution equations (one for the forward variable x and one for the backward variable y) there are three unknown processes $(x(\cdot), y(\cdot), z(\cdot))$ where $z(\cdot)$ is an auxilliary unknown process whose role is absolutely essential in guaranteeing that the solution of the system is adapted i.e. $x(t), y(t)$ are measurable with respect to the filtration generated by the Wiener process. This condition essentially guarantees that one may completely determine the possible values of the random variables $x(t), y(t)$ by having access only to the history of the economy up to time t. That this condition is not trivially satisfied for stochastic boundary value problems can be seen quite easily by taking the simple form of backward equation $dy(t) = 0$, with final condition $y(T) = \xi$ where ξ is a $\mathscr{F}_T$ measurable random variable. Our immediate candidate for a solution $y(t) = \xi$ for all t, while satisfying the equation and the boundary condition fails to satisfy the adaptability property, and this leads to the need for proper reinterpretation of the equation as a BSDE (see e.g. [21]).

FBSDEs find many applications in a number of fields. They arise quite naturally in stochastic control theory, where the forward part is considered as the state equation, while the backward part is considered as the adjoint equation in the generalization of the Pontryagin maximum principle [21]. In this setting the optimal control is constructed in terms of $(y(\cdot), z(\cdot))$, therefore the third process $z(\cdot)$ has a natural interpretation. Equations of this form arise also in a number of problems in mathematical finance as well as in mathematical economics [3]. For instance the hedging problem for contingent claims [3, 21], the problem of market completeness [21], the famous Black consol rate conjecture [9], portfolio optimization problems (see e.g. [21] and references therein), rational expectations models [20] and many others can be written in terms of an equivalent FBSDE.

The solvability of stochastic boundary value problems of this type has been studied by a number of authors, starting with the pioneering works of Bensoussan and Pardoux and Peng (see e.g. [13]). The basic strands in the literature either employ constructive methods using auxiliary deterministic quasilinear parabolic PDEs within the four step scheme (see e.g. [9]) to decouple the forward equation from the backward by the construction of a mapping f such that $y(t) = f(t, x(t))$, or purely probabilistic arguments that are based on fixed point schemes whose limit is the desired triple of processes $(x(\cdot), y(\cdot), z(\cdot))$ (see e.g. [14, 16]). The convergence of such schemes is guaranteed by monotonicity properties of the functions b, σ, δ, g some of which may be quite general.

52.4 Results Concerning the Infinite Horizon Model

52.4.1 Existence of Saddle Path Solutions

An important first question concerning the equivalent infinite horizon FBSDE form for the rational expectations model (52.1) is its well posedness. By answering the question of solvability in the affirmative, we guarantee the existence of paths that may lead the economy to the desired equilibrium state asymptotically in time ($t \to \infty$). This information is extremely important for an economic model which may be used for policy making.

The following existence result settles this problem.

Proposition 52.2. *Under the stated assumptions on the functions b, σ, δ, g, linear growth conditions on b and g and for large enough values of the lower bound C of the discount factor δ there exists a stable saddle path for model (52.1).*

Proof. We only sketch the proof here which follows [6] with the necessary modifications. The proof utilizes the fixed point scheme $u^{(1)}(t,x) = 0$, $u^{(n)}(t,x) = y_t^{(n)}$ where

$$\begin{aligned} dx^{(n)}(s) &= b(s, x^{(n)}(s), u^{(n)}(s, x^{(n)}(s))\,ds + \sigma(s, x^{(n)}(s), u^{(n)}(s, x^{(n)}(s)))\,dW(s) \\ dy^{(n+1)}(s) &= (\delta(s, x^{(n)}(s), u^{(n)}(s, x^{(n)}(s))y^{(n+1)}(s) - g(s, x^{(n)}(s), u^{n}(s, x^{(n)}(s)))\,ds \\ &\qquad - z^{(n+1)}(s)\,dW(s) \end{aligned}$$

with initial condition $x^{(n)}(t) = x$ and $s \in [t, \infty)$. This scheme essentially decouples the forward with the backward equation in each step, and expresses y_s at step $n+1$ with a function $u^{(n)}(t,x)$ calculated at $x = x^{(n)}(t)$. Furthermore, the backward equation at each step is a linear equation, and the condition the δ is bounded below by a positive number, guarantees that its solution is well behaved for all t. Using Gronwall type inequalities and estimates for the forward equation one may prove the uniform continuity in x of the sequence $u^{(n)}(t,x)$ and then the uniform convergence of this sequence on $R_+ \times R^n$, which in turn implies that the sequences of stochastic processes $x^{(n)}(\cdot)$, $y^{(n+1)}(\cdot) = u^{(n+1)}(\cdot, x^{(n)}(\cdot)), z^{(n)}(\cdot)$ converge in $M_X \times M_Y \times M_Z$ and the limit is the solution of the infinite horizon FBSDE. □

An alternative proof was given in [20], for the case where $\delta(t,x,y) = \delta$, $g(t,x,y) = g(x)$, $b(t,x,y) = b(x,y)$, $\sigma(t,x,y) = \sigma(x,y)$ where the four step scheme [9] was used to obtain a particular class of solutions to this problem, called nodal or Markovian solutions. The solutions of this type as such that $y(t) = f(x(t))$ for some deterministic function f, which is C^2. This function f can be considered as a parametrization of the stable manifold of the stochastic saddle path. Applying Itô's rule on the function f and matching resulting form for the evolution equation for $y(t)$ with the FBSDE we see that the function f must satisfy a system of quasilinear elliptic PDEs of the general form

$$L_i(x,f)f_i = -g_i(x) + \delta f_i(x), \quad i = 1, \cdots, m, \quad x \in R^n \tag{52.3}$$

where we consider $g = (g_1, \cdots, g_m) \in R^m$ and and $L_i(x,f)$ is the quasilinear elliptic operator

$$L_i(x,f) = \sum_{j=1}^{n} b_j(x, f(x)) \frac{\partial f_i}{\partial x_j} + \frac{1}{2} \sum_{j=1}^{n} \sum_{k=1}^{n} (\sigma(x, f(x))\sigma(x, f(x))^T)_{jk} \frac{\partial^2 f_i}{\partial x_j \partial x_k}$$

Therefore, the existence of the solution to (52.1) may be reduced to the existence of solutions for a quasilinear deterministic equation and the to the solvability for the forward SDE

$$dx(t) = b(x(t), f(x(t)))\,dt + \sigma(x(t), f(x(t)))\,dW(t)$$

where of course suitable conditions on b, σ must be imposed so that the solution to the forward problem is in M_X. For details on this approach see [20]. One may consider the quasilinear deterministic (52.3) as the stochastic analogue of the functional equation which is used in deterministic dynamical systems for the construction of the stable manifold.

For the case of linear problems, a more direct approach can be adopted that uses the construction of the stable manifold through the use of the Ricatti equation. According to this approach one may look for solutions using the special ansatz $y(t) = \phi(t)\,x(t) + \psi(t)$, where $\phi(t)$ and $\psi(t)$ are functions to be determined. Applying Itô's lemma on this ansatz and substituting into the equations one obtains that $\phi(t)$ satisfies a Ricatti type equation. Using the resulting Ricatti equation as well as stability results for linear forward SDEs existence results for the stable saddle paths for linear models can be obtained. Here we state for simplicity the results for the case $n = m = 1$.

Proposition 52.3. *Consider the linear two dimensional infinite horizon FBSDE*

$$\begin{aligned} dx_t &= (\alpha x_t + \beta y_t)dt + (\sigma_1 x_t + \sigma_2 y_t)dW_t \\ dy_t &= (\gamma x_t + \delta y_t)dt + z_t dW_t \end{aligned}$$

Let the following two assumptions hold

1. *The deterministic system has a saddle point structure with two real eigenvalues, ordered as follows $\lambda_2 < 0 < \lambda_1$.*
2. *The noise coefficients satisfy*

$$\left(\sigma_1 - \frac{\sigma_2}{\mu_2}\right)^2 <\mid \lambda_2 \mid, \ \ \mu_2 = \frac{(\delta - \alpha) + \sqrt{(\alpha - \delta)^2 + 4\beta\gamma}}{2\gamma}$$

Then the stochastic system has a unique stable saddle path.

52.4.2 Qualitative Properties of Saddle Path Solutions

Another important class of questions is the related to the qualitative properties of the saddle path solutions. Such properties may help in the construction of economic policy etc. We state such a result for the particular case where $n = 1$.

Proposition 52.4. *Assume that $g(x)$ is strictly increasing (decreasing). Then the representing function f is strictly increasing (decreasing).*

The monotonicity result proved in Proposition 52.4 may be used for obtaining information for the location of the stochastic analogue of the stable manifold in the nonlinear case. In the case of $g(x)$ strictly increasing we see by the above proposition that we expect the stable manifold to lie in the first and third quarterplanes where as in the case of $g(x)$ strictly decreasing we expect the stable manifold to lie in the second and fourth quarterplanes. Similar results, however without such explicit geometric interpretations, may be obtained for the case of higher dimensions using natural monotonicity results.

Such qualitative results can be of interest when one wants to gain some insight concerning the basic properties of the stable saddle path, such as the general region in phase space where this is likely to be situated etc.

52.5 Approximation and Construction of Saddle Paths

In many cases one is not interested in the infinite horizon stochastic saddle path model (52.1) but rather in its solutions in finite horizon $[0, T]$ with boundary conditions $(X(T), Y(T))$ such that

$$\mid X(T) \mid < \epsilon, \quad \mid Y(T) \mid < \epsilon,$$

for T arbitrary and possibly large. Such solutions can be considered as approximate saddle paths, which get close to equilibrium at T. While the problem of the asymptotic behaviour of the system which is equivalent to the existence of the stable saddlepath is very important from the dynamical systems point of view, in a number of applications the finite time behaviour of the system is more important than its asymptotic behaviour. One may not have the patience to wait an infinite time until the economy reaches the desired equilibrium state; as John Maynard Keynes has witfully stated we do not always have the luxury of being interested in the long run in the real world.

One way towards the construction of such solutions may be obtained through a stochastic generalization of a shooting method (stochastic control problem). Consider the forward stochastic control problem

$$\begin{aligned} dx(t) &= b(t, x(t), y(t))dt + \sigma(t, x(t), y(t))dW(t) \\ dy(t) &= (-g(t, x(t), y(t)) + \delta y(t))dt - (z(t), dW(t)) \\ x(0) &= x, \ \ y(0) = y \end{aligned} \tag{52.4}$$

where $z(t)$ is now considered as a control process, chosen so as to drive the system to the desired final state. Define the cost functional

$$J(s, x, y; z(t)) := E[\phi(X(T), Y(T))]$$

where $\phi(x, y)$ is a uniformly Lipschitz continuous function such that $\phi(x, y) \geq 0$ for all $(x, y) \in R^n \times R^m$ and $\phi(x, y) = 0$ if and only if $x = y = 0$. One may consider the function ϕ as an alternative to using the norm of the difference. Note that this problem is stated as a controllability problem rather than as a control problem, in the sense that the cost functional only involves $x(T)$, $y(T)$ and not $z(\cdot)$. In this sense, we are not imposing any restrictions on the process $z(\cdot)$ which can be arbitrary as long as it leads the system to the desired state. This is allowed here as $z(\cdot)$ is considered as an auxiliary process, introduced so as to give a differential equation format to the integral equation (52.1). We will return to the problem of control (as opposed to controllability) in Sect. 52.8.

Using techniques for the theory of controlled forward stochastic differential equations we may thus consider the problem of approximate solvability of an FBSDE (see [10]) and reinterpret these approximate solutions properly as paths that approach the equilibrium point close enough for finite time horizon T.

Proposition 52.5. *Let $b(t,x,y)$, $\sigma(t,x,y)$, $\delta(t,x,y)$ and $g(t,x,y)$ be continuous, C^2 with bounded first and second order derivatives and assume that they satisfy the condition*

$$\mid b(t,x,0) \mid + \mid \sigma(t,x,0) \mid + \mid g(t,x,0) \mid < L, \quad \forall (t,x) \in [0,T] \times R^n$$

Then there exists a path approaching the equilibrium state arbitrarily close in finite time T.

Proof. The proof follows by a straightforward application of Theorem 5.1 in [10]. □

The approximate solvability is by no means a property which can be guaranteed for any FBSDE. For instance, one may find explicit examples of equations which are not approximately solvable for given horizons T. This is closely linked with the solvability of related deterministic boundary value problems.

An alternative to the above approach could be to use monotonicity conditions of the general form

$$\begin{aligned} (b(t,x_1,y) - b(t,x_2,y), x_1 - x_2) &\le \lambda_1 \mid x_1 - x_2 \mid^2 \\ (\delta(t,x,y_1)\,y_1 - g(t,x,y_1) - (\delta(t,x,y_2)\,y_2 - g(t,x,y_2))\,,\; y_1 - y_2) &\le \lambda_2 \mid y_1 - y_2 \mid^2 \end{aligned}$$

and then adapt the general arguments of Pardoux and Tang [14] to prove existence of finite horizon solutions that approach equilibrium within a stated accuracy using a fixed point scheme argument. The general type of condition we need in this case is that $\lambda_1 + \lambda_2$ is bounded above by a suitable bound (not necessarily positive). In the particular case where δ is uniformly bounded below the condition on $\lambda_1 + \lambda_2$ essentially reduces to a condition on a δ being large enough compared to the terms b, g describing the feedback effects in the model. This is a very reasonable result as the strength of the discount factor is expected to play a very important role in the evolution of the model towards to equilibrium state. This approach allows us to obtain connections between the growth conditions and the strength of the monotonicity of the b, σ, δ, g with the rate by which the system approaches the equilibrium state, thus providing explicit estimates of the critical time T^* which is needed before the system reaches the desired neighbourghood of the equilibrium.

The approximation of the stable saddle path can be obtained through the approximate solution of the forward controlled problem, using techniques from computational stochastic control theory such as construction of appropriate Markov chains, viscocity solutions etc. (see e.g. [8]. Alternatively, it may be obtained using approximate solutions of the quasilinear system

$$\frac{\partial f_i}{\partial t} + L_i(x,f) f_i = -g_i(x) + \delta f_i(x), \quad i = 1, \cdots, m, \quad x \in R^n \tag{52.5}$$

with a final condition $f_i(T,x) = h(x)$ where h is any function that satisfies the condition $\mid h(x) \mid \leq C \mid x \mid$ for x in a neighbourghood of 0. As has been shown in [20] the particular choice of h is irrelevant as long as certain monotonicity conditions hold.

52.6 Stochastic Saddle Paths in Discrete Time and Backward Stochastic Difference Equations

A large class of models in economic theory are stated in discrete time, either because the actual dynamics happen in discrete time or because of their immediate connection with econometric models. Consider a general class of rational expectation models in the form

$$\begin{aligned} x(t+1) - x(t) &= b(t, x(t), y(t)) + \sigma(t, x(t), y(t))\, \xi(t) \\ y(t) &= E[\sum_{i=t}^{\infty} (1+\delta)^{-i}\, g(i, x(i), y(i)) \mid \mathscr{F}_t] \end{aligned} \tag{52.6}$$

where $\xi(t) = M(t+1) - M(t)$ where $M(t)$ is a martingale process.

The stochastic saddle path for the discrete time model can be expressed as the solution of (52.6) $(x(\cdot), y(\cdot))$ which belong to the sequence spaces $m_X \times m_Y$ where

$$m_X = \{\, \{x(i)\}_{i=1}^{\infty}\ s.t.,\ E[\sum_{i=1}^{\infty} \mid\mid x(i) \mid\mid_X^2] < \infty \}.$$

For the problem in question a possible choice for $X = R^n$ and $Y = R^m$.

Using technical tools from stochastic analysis for discrete time martingales we may show that this general model can be rewritten as a backward stochastic difference equation.

Proposition 52.6. *System (52.6) is equivalent to the backward stochastic difference equation*

$$x(t+1)-x(t) = b(t,x(t),y(t)) + \sigma(t,x(t),y(t))\,\xi(t),$$
$$y(t+1)-y(t) = \delta\, y(t) - (1+\delta)\, g(t,x(t),y(t)) - z(t)\,\xi(t)$$

Furthermore, one may generalize the existence results for the continuous time case, to provide results concerning the existence of stable saddle paths for discrete time models. Also, most of the qualitative results will hold for the discrete time case as well.

52.7 Examples from Economic Theory

52.7.1 The Krugman Model for Target Zones

In the Krugman model for target zones [7], the exchange rate $s(t)$ at any time t is assumed equal to

$$s(t) = m(t) + v(t) + \gamma \frac{E[ds(t)]}{dt}$$

where $s(t)$ is the log of the spot price of foreign exchange, $m(t)$ is the domestic money supply, $v(t)$ is a shift term representing velocity shocks and the last term is the expected rate of depreciation. The term $m(t)$ is considered as a policy variable which is shifted in such a way as to keep $s(t)$ within a specified band, the target zone. The term $v(t)$ is considered as the only source of external noise into the system. The evolution of v is given by the solution of the (forward) SDE

$$dv(t) = a(t,v(t))dt + \sigma(t,v(t))\,dW(t)$$

According to Krugman [7] the basic exchange rate equation can be viewed as arising from a more underlying equation of the form

$$s(t) = \frac{1}{\gamma} E\left[\int_t^{\infty} (m(r)+v(t)) e^{-\frac{1}{\gamma}(r-t)} dr \mid \mathscr{F}_t\right]$$

From the results of Sect. 52.2 it is evident that the Krugman model can be recast in the form of an infinite horizon FBSDE as

$$dx(t) = a(x(t))dt + \sigma(x(t))dW(t)$$
$$dy(t) = -\frac{1}{\gamma}(m(t)+x(t)) + \frac{1}{\gamma}y(t) - z(t)dW(t)$$

where we substituted $x(t) = v(t)$,$y(t) = s(t)$ to be in line with the notation used here. We see that this is a decoupled system of FBSDEs since the forward equation does not depend on the backward equation. As such, this model, can be treated with methods more simple than the ones used here which are taylor made for fully coupled FBSDEs. For instance as long as $a(x)$ and σ satisfy certain monotonicity conditions of the form used often for forward SDEs we may find that $x(\cdot) \in M_X$. Such a monotonicity condition may be for instance $(a(x), x) \leq -\mu \mid x \mid^2$ and $\sigma(\cdot) \in M_X$. We may now turn to the backward SDE for $y(\cdot)$. Using methods similar to those used for the proof of Theorem 4 in [16] we may see that as long as $m(\cdot) \in M_Y$ the backward SDE has a unique solution in M_Y. Therefore, under simple monotonicity conditions for the evolution of the fundamentals and the assumption that $m \in M_Y$ we conclude that the Krugman model has a unique stable saddle path.

52.7.2 The Dornbusch Model

In the Dornbusch overshooting model for exchange rates (see e.g. [2] or [12]) the constituting equations are the following:

$$\begin{aligned} m(t) - p(t) &= k\, y(t) - \lambda\, i(t) \\ y(t) = -\gamma \left(i(t) - \frac{E[dp(t)]}{dt} \right) &+ \eta\, (s(t) - p(t)) \\ \frac{E[ds(t)]}{dt} &= i(t) - i^* \\ dp(t) &= \phi\, (y(t) - \bar{y})dt + \sigma dW(t) \end{aligned}$$

The first equation is the condition for equilibrium of the domestic money market. In this equation, $m(t)$ is the domestic money supply, $p(t)$ is the domestic price level, $y(t)$ is the level of output in the economy and $i(t)$ is the nominal domestic interest rate. The second equation is a goods market equilibrium condition where $s(t)$ is the domestic price of foreign currency ($s(t) - p(t)$ is the real exchange rate). The third equation is an uncovered interest parity condition (the expected rate of depreciation of the domestic currency is set equal to the nominal interest differential) and the fourth equation is a representation of less than instantaneous price adjustment. External shocks in the economy are modeled by the introduction of the Wiener process perturbation $W(t)$.

Using the results of Sect. 52.2 this model may be redressed in the form of an infinite horizon FBSDE as

$$\begin{aligned} dx(t) &= \frac{1}{D}(-\phi(\gamma + \lambda\eta)x(t) + \phi\lambda\eta y(t))dt + \sigma dW(t) \\ dy(t) &= \frac{1}{D}((1 - k\eta - \phi\gamma)x(t) + k\eta y(t))dt - z(t)dW(t) \\ D &= k\gamma + \lambda - \phi\gamma\lambda \end{aligned}$$

where we denoted $x(t) = p(t)$ and $y(t) = s(t)$ the forward and backward variables respectively so as to be in accordance with the notation of this paper, and without loss of generality we set $i^* = \bar{y} = 0$. Then the results on the existence of stable saddle paths for linear systems may be used to provide the range of parameters for this model, for which there exists a stable saddle path leading the system to equilibrium [20].

52.7.3 The Woodford Model

In the Neo-Wicksellian framework the determinants of the equilibrium price level are not monetary factors, but the real factors that determine the equilibrium rate of interest (the natural rate of interest) on the one hand, and the systematic relation between interest rates and prices established by the central banks policy rule on the other Woodford [18] uses this framework to discuss monetary policy and price determination first in an cashless endowment economy without monetary frictions, and therefore without money demand for transaction purposes. the Neo-Wicksellian model for the analysis of monetary policy with nominal rigidities. This model consists of three equations: an IS equation, an AS (the new-keynesian Phillips curve), and an equation expressing the monetary policy rule (MR).

The IS equation is expressed as

$$x(t) = E[x(t+1) \mid \mathscr{F}_t] - \sigma\,[i(t) - E[\pi(t+1) \mid \mathscr{F}_t] - r(t)] \qquad (52.7)$$

where x denotes the output gap , i.e. the difference between the current output, and the equilibrium output, which is defined as the output which is consistent with perfect price flexibility. Thus at equilibrium $x = 0$. The case of positive output gap may be identified as an excess demand for the current output, while the case of a negative output gap may be identified as an excess supply. The market interest rate is denoted by i , the natural interest rate by r, π the inflation rate and $E[\cdot \mid \mathscr{F}_t]$ the expectations conditioned on the state of the economy by time t. Thus the expression $i(t) - E[\pi(t+1) \mid \mathscr{F}_t]$ denotes the real interest rate, and σ the inter-temporal substitution in consumption.

The next step is to provide a link between the output gap, and the inflation rate. This link is provided by the AS equation (New Keynesian Phillips curve) , which is derived from the Calvo staggered price setting model (see e.g. [18]), and has the form:

$$\pi(t) = k\,x(t) + \beta\,E[\pi(t+1) \mid \mathscr{F}_t] + u(t) \qquad (52.8)$$

This equation says that inflation depends positively on both the output gap and on the expected rate of inflation. In the AS equation β is an intertermporal discounting factor, $u(t)$ is a cost push factor and k is related to the price stickiness. It gives the relative change in the rate of inflation when the output gap changes. Therefore in the case in which firms never revise their prices (absolute price rigidity) the value of k

approaches zero. Thus k is decreasing in θ (in Calvo's model) which measures the degree of price rigidity.

The model is closed by an equation describing the monetary rule (MR) which replaces the traditional LM equation. The nominal interest rate is taken as an instrument of the monetary policy. The monetary aggregate is bypassed. According to this rule (originally suggested by Wicksell) nominal interest rate has to rise when inflation is rising, and vice versa when inflation is declining. We will assume that the nominal interest rate responds to inflation according to the rule:

$$i(t) = \gamma\pi(t) + g(t) \tag{52.9}$$

where $\gamma > 1$. This rule is known as the Taylors principle. It says that in order to increase the real interest rate, the nominal interest rate must respond more than one for one to changes in inflation. Thus $\gamma > 1$ expresses the rate of growth on the real interest rate.

Equations (52.7) and (52.8) are supposed to describe the dynamics of the system. Using Proposition 52.6 we may see that this system is equivalent to a backward stochastic difference equation. The using the existence results for the stable saddle path one obtains the parameter ranges for which the system may be driven to equilibrium. This gives us a relationship between γ and the other parameters of the system so that the stable saddlepath exists therefore can be interpreted as a dynamical systems justification of the Taylor's rule. Using this approach, one may further study the effects of a single monetary policy rule in the long term behaviour of monetary unions [1].

52.8 Control of FBSDEs and Applications in Economic Policy

In many problems in economic theory, there may be the posibility that the authorities may actively influence some quantities so as to affect the system's evolution. Such quantities may be considered as control variables. As an example of that one may consider the framework of the Krugman model (see Sect. 52.7.1) where m can now be considered as a quantity properly adjusted by the central bank so as to fulfil some selected targets. Similar interpretations may hold for the interest rate policy in the Woodford model. Therefore, one may rewrite the problem of selection of economic policy in models where future expectations are important, through the use of Proposition 52.1 to a problem of control of a FBSDE with properly selected cost functional.

The general state equation can be written in the form

$$\begin{aligned} dx^u(t) &= b(t, x^u(t), y^u(t); u(t))dt + \sigma(t, x^u(t), y^u(t); u(t))dW(t) \\ dy^u(t) &= (-g(t, x^u(t), y^u(t); u(t)) + \delta(t, x^u(t), y^u(t); u(t))\, y^u(t))dt \\ &\quad - (z^u(t), dW(t)) \\ x^u(0) &= x, \quad y^u(T) = h(x^u(T)), \end{aligned} \tag{52.10}$$

where $u(t)$ denotes a control process and the superscript u is used to denote the dependence of the processes $x(\cdot), y(\cdot), z(\cdot)$ on the choice of the control process. We may also define the cost functional of the general form

$$J(u) = E[\int_0^T \phi(x^u(s), y^u(s), u(s))\, ds]$$

where ϕ is given function. The particular choice of this function allows us to describe our targets. For instance if

$$\phi(x^u(s), y^u(s), u(s)) = C_1\,(x^u(s) - f_1(s))^2\; + C_2\,(x^u(s) - f_1(s))^2 \\ + C_3\,(u(s) - f_3(s))^2$$

then this functional penalizes the deviation of the economy at all times from the prescribed path $(f_1(s), f_2(s))$. By choosing the control in such a way so as to minimize $J(u)$, we obtain the path which is as close as possible to the prescribed path, with the minimum possible intervention cost. Other forms for the cost functional are possible, quantifying different targets for the policy maker.

The Pontryagin maximum principle can be generalized so as to provide solutions to this optimal control problem and specify both the optimal path of the economy as well as the optimal policy $u^*(t)$ needed to drive the economy to the optimal path. Both the optimal path and the optimal policy are obtained through the solution of an augmented system of FBSDEs where now the original system is complemented by a system containing the adjoint variables.

We refrain from giving a full account of this problem here but provide the solution for a simplified case where the state equation consists only of the backward part and the cost functional is a quadratic functional [19]. The relevant economic set up here is the problem of optimal exchange rate control of the Krugman model presented in Sect. 52.7.1 so as to minimize a the deviation of the economy from a desired path with the minimal possible inteventions. This may be stated as the control problem of finding $m^*(t)$ such that

$$J(m^*(t), \xi) = \min_{m(t)} E\left\{\frac{1}{2}Hs^2(0) + \int_0^T [Q(t)(s(t) - c(t))^2 + R(t)m^2(t)]dt\right\}$$

subject to the backward dynamics

$$\begin{aligned} ds(t) &= -\frac{1}{\gamma}(m(t) + F(t)) + \frac{1}{\gamma}s(t) - z(t)dW(t) \\ s(T) &= \xi \end{aligned} \tag{52.11}$$

Proposition 52.7. *The optimal control will be of the form*

$$m^*(t) = \frac{1}{\gamma R(t)}y(t)$$

where $(s(t), y(t), z(t))$ *are the solutions of the following system of FBSDEs*

$$\begin{aligned} ds(t)^* &= -\frac{1}{\gamma}(m^*(t) + F(t)) + \frac{1}{\gamma}s^*(t) - z^*(t)dW(t) \\ dy(t) &= -\left[Q(t)(s^*(t) - c(t)) + \frac{1}{\gamma}y(t)\right]dt + \hat{z}(t)\,dW(t) \\ s(T) &= \xi \\ y(0) &= -H(s^*(0) - c(0)) \end{aligned} \tag{52.12}$$

Without loss of generality we may take $\hat{z} := 0$.

To obtain the optimal policy we have to find a solution of the FBSDE (52.12). This is a linear FBSDE and it is natural to look for solutions which connect the forward and the backward variables in a linear manner. We will thus look for solutions of the form

$$s^*(t) = P(t)y(t) + h(t)$$

where $P(t)$ is a deterministic function of time and $h(t)$ is a stochastic process.

The following Proposition gives the solution of this FBSDE.

Proposition 52.8. *The FBSDE (52.12) admits a solution in the form*

$$s^*(t) = P(t)y(t) + h(t)$$

where $P(t)$ *solves the deterministic Riccati equation*

$$\begin{aligned} &\dot{P}(t) - \frac{2}{\gamma}P(t) - Q(t)P(t)^2 + \frac{1}{\gamma^2 R(t)} = 0 \\ &P(T) = 0 \end{aligned}$$

$(h(t), z^*(t))$ *solve the BSDE*

$$\begin{aligned} dh(t) &= \left\{\left(\frac{1}{\gamma} + P(t)Q(t)\right)h(t) - \left(\frac{1}{\gamma}F(t) + P(t)Q(t)c(t)\right)\right\}dt - z^*(t)dW(t) \\ h(T) &= \xi \end{aligned}$$

and $y(t)$ *solves the forward random ODE*

$$\begin{aligned} dy(t) &= \left\{-\left(Q(t)P(t) + \frac{1}{\gamma}\right)y(t) + Q(t)(c(t) - h(t))\right\}dt \\ y(0) &= -\frac{H(h(0) - c(0))}{1 + HP(0)} \end{aligned}$$

We will conclude this section with the following lemma that provides solutions for the BSDE and the forward random ODE

Lemma 52.1. *(a) The BSDE for $h(t)$ has the solution $(h(t), z^*(t))$*

$$h(t) = \Phi(t)\left\{-\int_0^t \Phi(s)^{-1}\left(\frac{1}{\gamma}F(s) + P(s)Q(s)c(s)\right)ds + \int_0^t \eta(s)dW_s + E[\theta]\right\}$$
$$z^*(t) = \Phi(t)\eta(t)$$

where

$$\Phi(t) = \exp\left(\int_0^t \left(\frac{1}{\gamma} + P(s)Q(s)\right)ds\right)$$
$$\theta = \Phi(T)^{-1}\xi + \int_0^T \Phi(s)^{-1}\left(\frac{1}{\gamma}F(s) + P(s)Q(s)c(s)\right)ds$$

and the stochastic process η is defined by the representation

$$E[\theta \mid \mathscr{F}_t] = E[\theta] + \int_0^t \eta(s)dW(s)$$

(b) The forward random ODE has the solution

$$y(t) = \Psi(t)y(0) + \Psi(t)\int_0^t \Psi(s)^{-1}Q(s)(c(s) - h(s))ds$$
$$\Psi(t) = \exp\left(-\int_0^t \left(Q(s)P(s) + \frac{1}{\gamma}\right)ds\right)$$

Acknowledgements The author would like to thank the organizers of DYNA2008, and in particular Professor A. A. Pinto, for the very interesting and stimulating conference and for the warm hospitality.

References

1. Demopoulos, G.D., Yannacopoulos, N.A., Yannacopoulos, A.N., Warren, S. A.: Neo-Wicksellian monetary theory and monetary unions. In: Brox J.A. and Baltas N.C. (eds.) The Global Economics of a Changing Environment, North Waterloo Academic Press (2009)
2. Dornbusch, R.: Expectations and exchange rate dynamics. J. Polit. Econ. **84**, 1161–1176 (1976)
3. El Karoui, N., Peng, S., Quenez, M.-C.: Backward stochastic differential equations in finance. Math. Financ. **7**, 1–71 (1997)
4. Gandolfo, G.: Economic Dynamics. Springer, New York (1997)
5. Guckenheimer, J., Holmes, P.: Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields. Applied Mathematical Sciences, vol. 42. Springer, New York (1983)
6. Guo, D., Ji, S., Zhao, H.: On the solvability of infinite horizon forwardbackward stochastic differential equations with absorption coefficients. Statist. Probab. Lett. **76**, 1954–1960 (2006)
7. Krugman, P.R.: Target zones and exchange rate dynamics. Q. J. Econom. **106**, 669–682 (1991)

8. Kushner, H.J., Dupuis, P.: Numerical Methods for Stochastic Control Problems in Continuous Time. Applications of Mathematics, 2nd edn., vol.24. Stochastic Modelling and Applied Probability. Springer, New York (2001)
9. Ma, J., Yong, J.: Forward-Backward Stochastic Differential Equations and their Applications. Lecture Notes in Mathematics, vol. 1702, Springer, Berlin (1999)
10. Ma, J., Jong, J.: Approximate solvability of forward-backward stochastic differential equations. Appl. Math. Optim. **45**, 1–22 (2002)
11. Miller, M., Weller P.: Stochastic saddlepoint systems: Stabilization policy and the stock market. J. Econom. Dynam. Control **19**, 279–302 (1995)
12. Neely, C.J., Weller P.A., Corbae, P.D.: Endogenous reallignments and the sustainability of a target zone. Endogenous reallignments in a target zone, Oxford Economic Papers, **55**, 494–511 (2003)
13. Pardoux, E., Peng, S.G.: Adapted solution of a backward stochastic differential equation. Syst. Control Lett. **14**, 55–61 (1990)
14. Pardoux, E., Tang S.: Forward-backward stochastic differential equations and quasilinear parabolic PDEs. Probab. Theory Relat. Fields. **114**, 123–150 (1999)
15. Peixoto, M.M.: Structural stability on two dimensional manifolds. Topology **2**, 101–121 (1962)
16. Peng, S., Shi, Y.: Infinite horizon forward-backward stochastic differential equations. Stochastic Process. Appl. **85**, 75–92 (2000)
17. Rand, D: Exotic phenomena in games and duopoly models. J. Math. Econom. **5**, 173–184 (1978)
18. Woodford, M.: Interest and Prices: Foundations of a Theory of Monetary Policy. Princeton University Press, Princeton and Oxford (2003)
19. Yannacopoulos, A.N.: A novel approach to exchange rate control using controlled backward stochastic differential equations. Ekonomia **8** (2005)
20. Yannacopoulos, A.N.: Rational expectations models: An approach using forwardbackward stochastic differential equations. J. Math. Econom. **44**, 251–276 (2008)
21. Yong. J, Zhou, X. Y.: Stochastic Controls: Hamiltonian Systems and HJB Equations. Applications of Mathematics, vol. 43, Springer, New York (1999)