Springer Proceedings in Mathematics & Statistics

Bourama Toni *Editor*

Advances in Interdisciplinary Mathematical Research

Applications to Engineering, Physical and Life Sciences

Springer

Springer Proceedings in Mathematics & Statistics

Volume 37

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

Springer Proceedings in Mathematics & Statistics

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

Bourama Toni
Editor

Advances in Interdisciplinary Mathematical Research

Applications to Engineering, Physical and Life Sciences

Springer

Editor
Bourama Toni
Department of Mathematics
and Computer Sciences
Virginia State University
Petersburg, Virginia, USA

ISSN 2194-1009 ISSN 2194-1017 (electronic)
ISBN 978-1-4614-6344-3 ISBN 978-1-4614-6345-0 (eBook)
DOI 10.1007/978-1-4614-6345-0
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2013935590

Mathematics Subject Classification (2010): 97M10, 70F10, 51M04, 74R99, 97R99, 97R30, 82D80, 91A80, 93C95

Printed on acid-free paper

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

To

My wife Emi

To

My children Emma Akira and Andy Shingo

To

Dr. Huiqing Helen Yang

You have honored us with an everlasting friendship You have inspired all during you, tenure at Virginia State University with your devotion to student success, with the highest ethical, moral and professional standards, and with your remarkable joie de vivre.
May your memories live on and continue to inspire through this publication

Preface

Mathematics is playing an ever more important role in the physical and life sciences, engineering and technology, blurring the boundaries between scientific disciplines. This is evidenced in this volume which contains cutting-edge contributions to mathematical sciences and to their applications to the STEAM-H disciplines, that is, science, technology, engineering, agriculture, mathematics, and health. This volume is a written and published thematic continuation of the seminar series at Virginia State University during the academic year 2011–2012. Contributors in this volume, as leading researchers, present their own work in the perspective to advance their specific fields in a way to generate a genuine interdisciplinary interaction. All articles therein were carefully edited and peer-reviewed; they are reasonably self-contained and pedagogically exposed.

This volume features new advances in mathematical research represented here on anti-periodicity, almost stochastic difference equations, absolute and conditional stability in delayed equations, gamma-convergence, the dynamics of collision and near-collision in celestial mechanics, and almost and pseudo-almost limit cycles. It includes current advances in the applications to "rainbows" in spheres with connections to ray, wave, and potential scattering theory; null-controllability of the heat equation with constraints; optimal control for systems subjected to null-controllability; the Galerkin method for fluid flow analysis; wavelet transforms for real-time noise cancellation; signal, image processing, and machine learning in medicine and biology; durability, reliability, and damage tolerance of aerospace materials and structures at NASA Langley Research Center; and Γ-convergence for block copolymer morphology.

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

This volume as a whole will enhance the overall objective of the seminar, that is, to foster student interest in the STEAM-H disciplines and stimulate graduate and undergraduate research and collaboration among researchers on a genuine interdisciplinary basis.

Chapter "An Overview of Durability and Damage Tolerance Methodology at NASA Langley Research Center", by Jonathan Ransom, Edward Glaessgen, and James Ratcliffe, is about damage science at NASA Langley Research Center (LaRC), that is, the design and implementation of computational, analytical, and experimental strategies and methodologies to simulate and assess damage growth and to characterize the damage tolerance of aerospace materials and structures, including many fracture mechanics methods to predict and characterize damage in both metallic and composite materials. This is at the core of the research portfolio of the branch, the Durability, Reliability and Damage Tolerance Branch (DDTRB), which the author is heading. This chapter presents a selection of such strategies and methodologies in a self-contained and streamlined form, accessible to even the classically trained mathematicians. It discusses new methodologies in continuum mechanics, damage tolerance capabilities for composite structures, and related activities for prediction and verification methods for delamination, debonding, and identification of failure mechanisms. To illustrate the applicability of the expertise the authors include a fractographic analysis in the case of AA 587 accident investigation. Finally this chapter advocates a multidisciplinary knowledge base that efficiently combines, for instance, multi-scale simulation capability, optical microscopy, physical metallurgy, organic chemistry, finite elements analysis, molecular dynamics, optimization, and high-performance computing.

Chapter "On the Γ-convergence Theory and its Application to Block Copolymer Morphology", by Xiaofeng Ren, draws from the expanding field of calculus of variations the tool of gamma-convergence theory to develop a rigorous notion for a family of functionals to converge to a functional of a seemingly different type, while still retaining vital properties in the limiting functional. Using this theory the author reduces the Ohta–Kawasaki density theory for block copolymers to a geometric problem containing perimeter minimization and nonlocal interaction. Block copolymers are soft materials characterized by fluidlike disorder on the molecular scale and a high degree of order at a longer length scale. In the process global and local minimizers for the Ohta–Kawasaki theory are characterized. This chapter also discusses the issue of non-locality which often forces a periodic repetition in condensed materials such as charged Langmuir monolayers, chiral liquid crystals, and block copolymers.

Chapter "'Rainbows' in Homogeneous and Radially Inhomogeneous Spheres: Connections with Ray, Wave and Potential Scattering Theory", by John Adam, starts with an introduction to the scientific and mathematical beauty of rainbows, which, according to Sassen, in reference, "have long been a source of inspiration both for those who would prefer to treat them impressionistically or mathematically. The attraction to this phenomenon of Descartes, Newton, and Young, among others, has resulted in the formulation and testing of some of the most fundamental principles of mathematical physics." Follow other interesting descriptions by respectively

Nussenzveig and Lock. In this work, the author elegantly discusses some direct and indirect connections between ray theory, wave theory, and potential scattering theory, providing several perspectives to understand the mathematical nature of the rainbow. In addition, the pedagogical exposition of the profound and complex mathematics and physics behind rainbows and the mechanism of their formation enhance one's appreciation of same: scattering by transparent spheres and the correspondent scattering matrix; ray path integral; morphology-dependent resonances; complex angular momentum; Regge poles and Regge trajectories; and Mie solutions of electromagnetic scattering theory. In the end, using the universal attraction to the beauty and "mystery" of rainbows, the author successfully unifies the treatment of the theories of ray, wave, and potential scattering.

Chapter "Understanding the Dynamics of Collision and Near-Collision Motion in the N-Body Problem", by Lennard Bakker, is on celestial mechanics where stands unsolved the N-body problem since Newton's time. The author makes the point of the necessity of understanding first the nature and dynamics of collisions and near-collisions as an indispensable step towards a complete understanding of the N-body problem, including in a regularized setting which removes the collision singularities. This chapter also discusses the probabilities of collisions, in fact rare as opposed to near-collision motions. It also features some historical remarks interwoven throughout as well as in the footnotes.

Chapter "Absolute Stability and Conditional Stability in General Delayed Differential Equations", by Junping Shi, is concerned with the stability of equilibrium of delay differential equations. In the case of absolute and conditional stability, it provides explicit criteria for one or two equations in general form. The dependence of stability on both the instantaneous feedback and the delayed feedback is also derived from the results. The chapter closes with an interesting open question on stability for distributed delay.

Chapter "Existence of Antiperiodic Solutions to Semilinear Evolution Equations in Intermediate Banach Spaces", by Gaston N'Guerekata and Gisele Mophou, considers a class of semilinear evolution equations with an unbounded sectorial operator not necessarily densely defined in a Banach space, together with an intermediate Banach space. Using an approach based on Banach's fixed-point theorem, the authors proved the existence and uniqueness of an antiperiodic mild solution.

Chapter "Signal, Image Processing and Machine Learning: The Key to Complex Problems in Medicine and Biology", by Mahsa Zahery and Kayvan Najarian, discusses some signal processing and machine learning methods used in biomedical applications and emphasizes their importance on addressing complex problems in medicine and biology, that is, the computer-aided decision-making procedure aiming at producing accurate and timely diagnosis and prognosis to improve the overall service and reduce the cost of health care. This chapter describes applications such as hemorrhage detection involving error correcting output codes (ECOC) and attention detection using dual-tree complex wavelet transform.

Chapter "Real-Time Noise Cancellation Using Wavelet Transforms", by Eshan Sheybani, is about the use of wavelet transforms to develop computationally

low-power, low-bandwidth, and low-cost filters that will remove the noise acquired in datasets, effectively in real time for a decision to be made at the node level. The performance and merit of the approach are profusely illustrated with some experimental results in a series of expressive figures.

Chapter "Null Controllability of the Heat Equation with Two Constraints on the Control: Application to a Discriminating Sentinel with Given Sensitivity", by Ousseinou Nakoulima and Sadou Tao, studies the null-controllability problem with two constraints on a pair of controls. The results are then applied to a discriminating sentinel with given sensitivity to detect some parameters in a pollution problem, governed by a semilinear parabolic equation with Dirichlet boundary condition.

In chapter "A Galerkin Method Solution of Heat Transfer Problems in Closed Channels: Fluid Flow Analysis", by Nasser Ghariban, a fluid flow inside closed channels is analyzed with a heat transfer model built from momentum and energy equations and a Galerkin-based method. The results are validated by comparison with the results from numerical methods and experimental data.

Chapter "Optimal Control for Distributed Linear Systems subjected to Null-Controllability with Constraints on the State", by Michelle Mercan, applies the notion of hierarchical control on a distributed system in which the state is governed by a parabolic equation, assuming two controls: the Leader supposed to bring the solution of the parabolic equation subjected to finite number of constraints to rest at time T, while the second, the Follower, expresses that the state does not move too far from a given state. The results are achieved by means of an observability inequality of Carleman adapted to the constraint.

Chapter "Almost and Pseudo-Almost Limit Cycles with Applications to Quasiperiodic Solitary Waves", by Melissa Watts and Bourama Toni, extends the theories of limit cycles, quasi-periodicity, and the related isochrons to the new concepts of almost and pseudo-almost limit cycles. It addresses the usual questions of conditions of existence, uniqueness, stability, and bifurcation. Several illustrative examples are presented, including some almost and pseudo-almost periodic perturbations of the harmonic oscillator and the renowned Liénard systems. The existence of almost and pseudo-almost periodic waves is also derived. The chapter concludes with many interesting open problems, in particular the question of transitioning an almost or pseudo-almost periodic behavior to a chaotic one by coupling and synchronization.

The concluding chapter "On Almost Periodic Stochastic Difference Equations", by Paul Bezandry, investigates almost periodic random sequence in mean and derives the existence and uniqueness of almost periodic solution of a semilinear system of stochastic difference equations using exponential dichotomy.

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

Accelerator Facility, and the Homeland Security Institute. The seminar, through its written thematic continuation published by a world-renowned publisher, Springer, is expected to become a national and international reference in STEAM-H education and research.

Acknowledgements

We would like to express our most sincere appreciation of the President of Virginia State University Dr. Keith T. Miller's encouragement and support of the seminar series in line with his vision of "Building a Better World".

We gratefully acknowledge the following supports: the Office of the Provost, Dr. W. Weldon Hill, Mr. Daniel Roberts, Ms Dorothy Yancey and Ms Marie Singfield; the Office of the Dean, School of Engineering, Science and Technology, Professor Keith M. Williamson, Professor Dawit Haile, Professor Larry C. Brown, Ms Victoria Perkins and Mrs Bonnie Grant; the Department of Mathematics and Computer Science, the Chair Professor Kenneth Bernard and Administrative Assistants Ms Caroline Price and Ms Vickie Crowder; the Department of Education HBCU MS Program and Professor Pamela Leigh-Mack; the NSF/HBCU-UP, Professor Ali Ansari and Ms Amber Dollete; the NIH/RIMI program and Dr Omar Faison.

We would like to thank very much Mr. Leroy Lane, Ms Melissa Watts, Dr. Giti Javidi, Dr Tony Bryant, Dr. Yahya Njai, Ms Eleanor Poarch-Wall, Mr. Daniel Fritz, Mr. Andrew Wynn, Ms Owens Azzala, Mr. Calvin Smith, Ms Jazarai Studivant, Ms Jewel Booker, Mr. Taurus Richardson and Mr. Joshua Silve, whose tireless efforts have contributed to the smooth organization, presentations, recording and attendance. We sincerely appreciate the promotional support by the Office of Students Activities and its Director Ms Martin Menjiwe and VSU Radio Station WVST 91.3 and Station Manager Ms Jennifer Williamson and her assistants Ms Melony Negron, Ms Melissa Thornton and Mr. Jermane O'Neal.

Special thanks are extended to all the contributors, the faculty and student participants, in particular to Dr. Zhifu Xie, Dr. Ju Wang, Dr. Kostadin Damevski, Dr. Hui Chen, Dr. David Walter, Professor Oliver Hill, Professor Rana Singh, Professor Emeritus Walter Elias, Dr. Brian Sayre, Dr. Laban Rutto, Dr. Wanda D. Gay, Dr. Eshan Sheybani, Dr. Nasser Ghariban, Professor Ephrem Eyob, Professor Toka Diagana, Dr. Ahmed Mohamed, Professor Gaston N'Guérékata, Professor Godwin Mbagwu, Dr. Grace Ndip, Dr. Jonathan Ransom, all great supporters and/or frequent presenters at the seminar since its inception. Their genuine efforts to engage VSU faculty in interdisciplinary research and to encourage the students into the STEAM-H disciplines are very much appreciated.

We would like toexpress our sincere thanks to all the anonymous referees for their professionalism. They all made the seminar and its published thematic continuation a reality for the greater benefice of the community of science, technology, engineering, agriculture, mathematics and health.

Petersburg, VA, USA Bourama Toni

Contents

Contributors

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

Lennard Bakker
Department of Mathematics, Brigham Young University, Provo, UT, USA

Paul Bezandry
Department of Mathematics, Howard University, Washington, DC, USA

Nasser Ghariban
Department of Engineering, Virginia State University, Petersburg, VA, USA

Edward Glaessgen
Head, Durability, Damage Tolerance & Reliability Branch, NASA Langley Research Center, Hampton, VA, USA

Michelle Mercan
Laboratoire CEREGMA, Université des Antilles et de la Guyane, Guadeloupe, France

Gisele Mophou
Département de Mathématiques et Informatique, Université des Antilles et de la Guyane, Guadeloupe, France

Kayvan Najarian
Computer Science & Emergency Medicine, Reanimation Engineering Science Center Director, Biomedical Signal and Image Processing Group, Virginia Commonwealth University, Richmond, VA, USA

Ousseynou Nakoulima
Laboratoire CEREGMA, Université des Antilles et de la Guyane, Guadeloupe, France

Gaston N'Guerekata
Department of Mathematics, Morgan State University, Baltimore, MD, USA

Jonathan Ransom
Durability, Damage Tolerance & Reliability Branch, NASA Langley Research Center, Hampton, VA, USA

James Ratcliffe
Durability, Damage Tolerance & Reliability Branch, NASA Langley Research Center, Hampton, VA, USA

Xiaofeng Ren
Department of Mathematics, George Washington University, Washington, DC, USA

Eshan Sheybani
Department of Computer Engineering, Virginia State University, Petersburg, VA, USA

Junping Shi
Department of Mathematics, College of William & Mary, Williamsburg, VA, USA

Sadou Tao
Laboratoire d'Analyse numérique, d'Informatique et de Biomathématique, Université de Ouagadougou, Ouagadougou, Burkina Faso

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

Melissa Watts
Department of Mathematics and Computer Sciences, Virginia State University, Petersburg, VA, USA

Masha Zahery
Computer Science & Emergency Medicine, Reanimation Engineering Science Center, Virginia Commonwealth University, Richmond, VA, USA

Chapter 1
An Overview of Durability and Damage Tolerance Methodology at NASA Langley Research Center

Jonathan B. Ransom, Edwards H. Glaessgen, and James G. Ratcliffe

Introduction

Engineering fracture mechanics, in particular linear elastic fracture mechanics (LEFM), has played a vital role in the development and certification of virtually every aerospace vehicle that has been developed since the mid-twentieth century. Often, LEFM is associated with a damage tolerance design philosophy where a critical flaw size must be significantly larger than the minimum detectable flaw (e.g., crack) to insure safety. Here, the critical flaw is assumed to exist in a location and under a loading where fracture occurs. In this philosophy, analysis or testing, or a combination of both must show that the detectable crack will not reach a critical length before a subsequent inspection.

Traditional engineering fracture mechanics is a continuum mechanics construct that is based on the premise that crack growth will occur when a computed fracture parameter reaches its empirically determined critical value. For example, brittle fracture in metals (in-plane strain) will occur when $K_I > K_{IC}$, i.e., when the computed value of the stress intensity factor, K_I, is greater than or equal to the experimentally obtained fracture toughness, K_{IC}. In engineering fracture mechanics, fracture toughness is considered to be a property of the material, and the plane strain fracture toughness, K_{IC}, is the lowest value of material toughness. Similarly in laminated composite materials that are susceptible to failure mechanisms such as delamination, the strain energy release rate, G, has been traditionally used as a

J.B. Ransom (✉) • E.H. Glaessgen • J.G. Ratcliffe
Durability, Damage Tolerance and Reliability Branch, NASA Langley Research Center, Hampton, Virginia, 23681, USA
e-mail: Jonathan.B.Ransom@nasa.gov; Edward.H.Glaessgen@nasa.go; James.Ratcliffe@nianet.orgx

B. Toni (ed.), *Advances in Interdisciplinary Mathematical Research*, Springer Proceedings in Mathematics & Statistics 37, DOI 10.1007/978-1-4614-6345-0_1,

measure of the driving force for delamination growth. In a manner similar to that employed for brittle fracture in metals, onset of delamination growth is expected to take place when G becomes equal to or greater than a critical value G_c. As a consequence of delamination growth being constrained by the bounding plies, a mixed-mode loading condition can be imparted along the delamination front, involving opening and shear components of G. In such cases, the total strain energy release rate must be decomposed into its individual components, and a mixed-mode delamination growth criterion must be used to determine the onset of growth.

There are numerous examples of fracture-related mechanisms exhibited by both metallic and composite structure where LEFM fails to provide a sufficient representation of the failure mechanism in question. Examples include fracture in metals involving significant levels of metal plasticity and delamination growth in composite laminates that is accompanied by additional energy dissipating mechanisms such as fiber bridging or the failure of through-the-thickness reinforcement. In these circumstances, approaches are required that either involve alternative approaches to LEFM or employ LEFM-based approaches that are amended to account for the additional failure mechanisms.

Furthermore, with the increasing use of composite materials in airframe primary structure, there is motivation to improve the efficiency of the certification of these relatively new materials. The typical response to this situation has been the attempt to establish damage tolerance analysis methods for replacing, and thus reducing, the amount of testing involved in certifying a composite structure.

Consequently, the Durability, Damage Tolerance and Reliability Branch (DDTRB) at NASA Langley Research Center (LaRC) continues to develop a broad portfolio of fracture mechanics methods aimed at understanding damage in both metallic and composite aerospace structures. Additionally, the branch continues to develop methods with the aim of decreasing the time required for certifying aerospace structures. The latter aim has motivated the development of analysis methods for new forms of metallic structure and has led to a continued effort that is geared towards the development of standardized testing practices for characterizing various fracture mechanisms in composite materials.

This chapter presents an overview of the computational, analytical, and experimental strategies for fracture mechanics for fatigue, fracture, and damage tolerance of metallic and composite aerospace structures at NASA Langley Research Center (LaRC). Methodologies for simulating and characterizing fatigue and fracture of metallic materials are presented. This discussion includes new methodologies in continuum mechanics as well a new paradigm in damage mechanics, referred to herein as damage science. Damage tolerance capabilities for composite structures, including sandwich construction, are then presented. A selection of activities associated with composite materials is presented, including those involved with the development of test standards, prediction and verification methods for delamination and debonding of composite laminates, and the identification of failure mechanisms for a recent failure investigation.

Fracture Mechanics of Metallic Materials

Although methodologies for characterization of the fatigue and fracture of metallic materials have been extensively developed over the past several decades, work in this area remains an active topic of research. The focus of much of the work in the branch in metallic materials has centered on development of methods to predict crack growth in new material forms (e.g., friction stir weld panels) and to improve our understanding of the fundamental mechanisms of deformation and fracture. The effort on predicting crack growth in new material forms builds upon well-established methods in continuum fracture mechanics for predicting fracture in built-up structures. Conversely, a relatively new and largely unproven effort has also been undertaken that offers the promise of changing the fundamental paradigm of fracture mechanics (and greatly extending the length scales for which it is valid) by examining damage processes at the micro- and even the nanoscale. This section will discuss work in the broad range from continuum fracture mechanics to atomistic simulation of fundamental damage processes.

Residual Strength Predictions for Friction Stir Weld Panels

Friction stir welding (FSW) is a new solid-state joining technology that is being considered by many airframe manufacturers as a replacement for traditional joining methods. As is common with other welding methods, FSW results in a residual stress state that may affect crack growth rates. Thus, determination of the fatigue life of friction-stir-welded structure requires the ability to predict the residual stress intensity, $K_{residual}$.

A new method being developed to predict $K_{residual}$ is based on determination of equivalent thermal loads [1]. Equivalent thermal loads are calculated by defining initial strain due to welding along the length and width of the weld region. The method determines the equivalent thermal loads that produce the residual stress field using the elastic modulus, E; coefficient of thermal expansion, α; and change in temperature, ΔT. The methodology always satisfies self-equilibrium and allows rapid convergence. Temperature change, ΔT, is calibrated by comparing the predicted residual stress field to that measured for coupon test data. Similitude can be assumed such that the same ΔT may be used to generate residual stress fields for any other configuration (specimen type, component, or panel of any size and shape) as long as the same welding parameters are used.

The analysis results are compared with experimental data obtained from both cut-compliance and crack-compliance tests. Compact tension C(T) specimens that are 4 inches wide and 0.25 inches thick representing two welding configurations (tensile-dominated and compression-dominated) were modeled using isoparametric eight-node brick elements in the ZIP3D finite element code as shown in Fig. 1.1 (half the specimens were modeled on the assumption of symmetry about the x-axis). Aluminum alloy 2024-T3 was considered for all analyses with modulus,

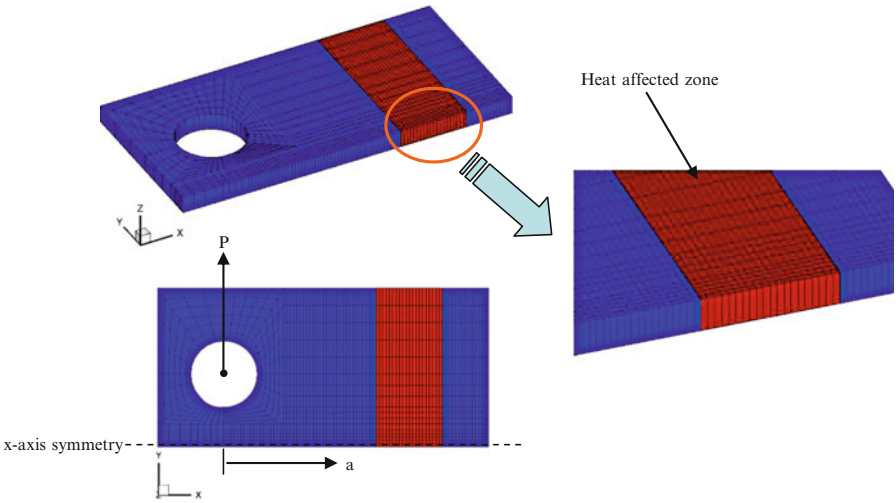

Fig. 1.1 A typical finite element mesh for 4-inch wide C(T) specimen

E=10 Msi, and coefficient of thermal expansion, $\alpha = 13.0 \times 10^{-6}$in/in°F. By comparing the predicted residual stress intensity factor (SIF) distribution to the experimentally measured values in the weld zone, the change in temperature was empirically determined to be −200°F for the tensile-dominated configuration. This same value of ΔT was then applied to the finite element analysis of the compression-dominated specimen. Stress intensity factor solutions were generated using virtual crack closure technique (VCCT) and the J-integral technique and are shown in Figs. 1.2 and 1.3. Both crack-compliance (symbols) and cut-compliance (lines) experimental data are shown for comparison. The SIF solution compares well with experimental data.

Once the residual stress distributions were determined, their effects on residual strength could be determined. Because the 4-inch wide C(T) specimen was manufactured using procedures that mimicked those used on production panels, similitude between the coupon tests and complex FSW panels could be assumed. Hence, the thermal parameters obtained from the analyses of the C(T) specimens could be used to account for residual stress effects in the FSW panel. After an equilibrium solution was obtained in the finite element analysis, the panel was analyzed under tensile loading (illustrated in Fig. 1.4), and a residual strength prediction was carried out using crack-tip opening angle (CTOA) fracture criteria. A typical 3D finite element model of the 24-inch wide FSW panel containing a through-thickness center crack is shown in Fig. 1.4 with the heat-affected FSW zone colored red. The analysis accounted for crack branching, plasticity, variation in panel thickness, residual stress, and the presence of multiple materials.

The corresponding load-crack extension data are shown in Fig. 1.5. In the figure, the open and filled symbols correspond to the test data. The black line represents an analysis carried out without considering the effects of residual stress and is shown

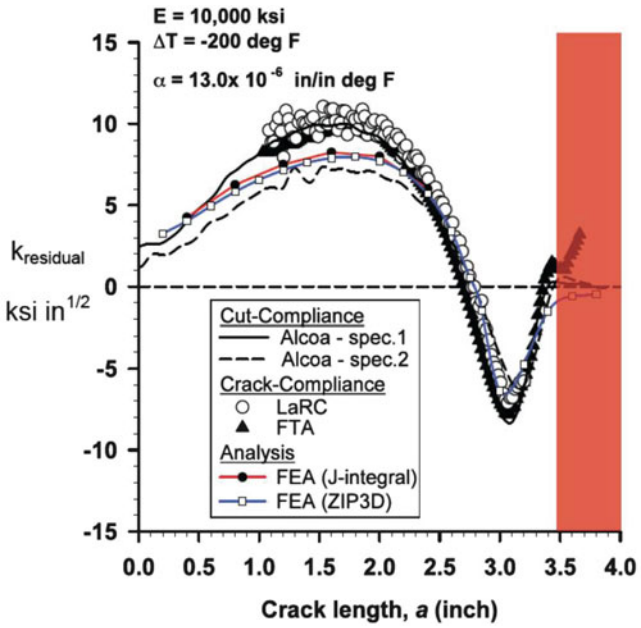

Fig. 1.2 Residual stress variation for tension-dominated specimen. Experimental data are shown for comparison

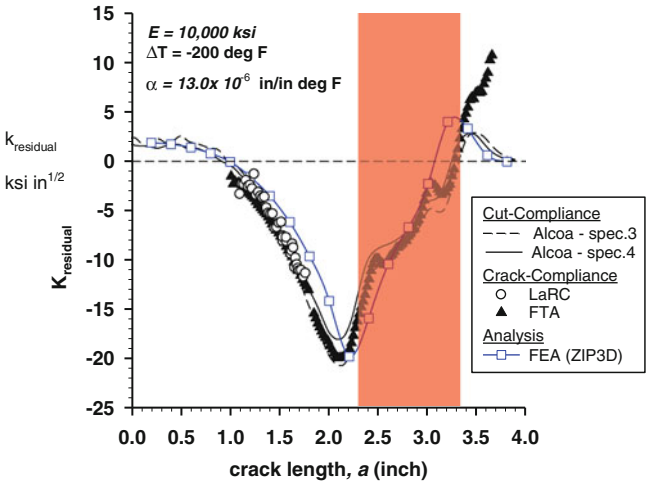

Fig. 1.3 Residual stress variation for compression-dominated specimen. Experimental data are shown for comparison

to overpredict the test results. However, by including the residual stress field and change in material properties in the heat-affected zone, the analysis prediction represented by the red line is much better and well within the test scatter.

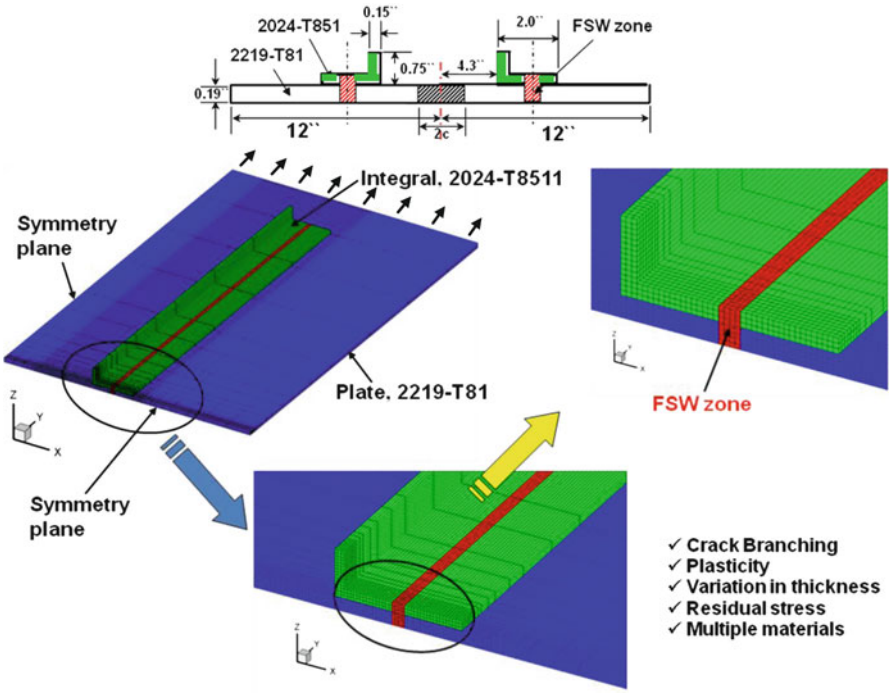

Fig. 1.4 A typical finite element mesh for 24-inch wide FSW integral panel

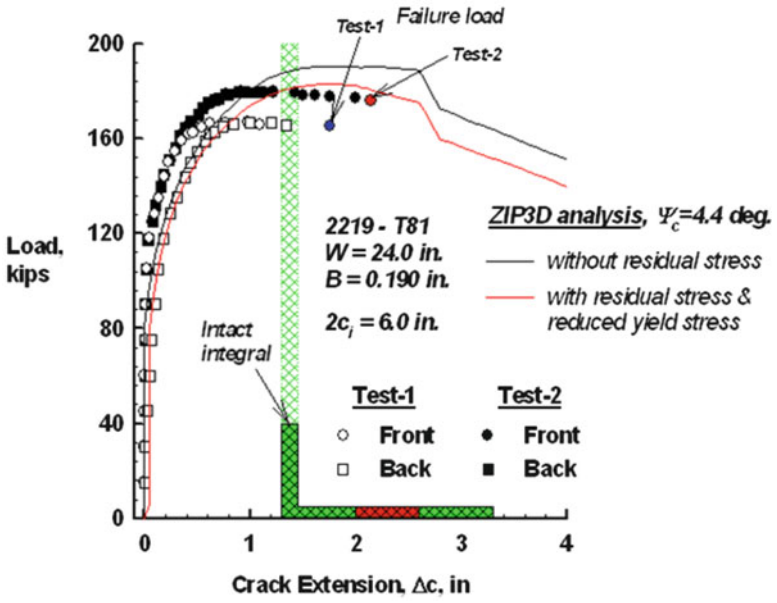

Fig. 1.5 Load-crack extension data for 24-inch wide FSW panel

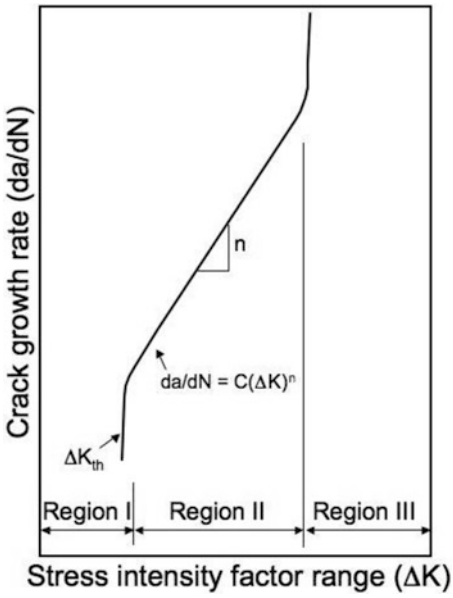

Fig. 1.6 Typical fatigue crack growth behavior

Fatigue Crack Growth and Crack Closure

The most ubiquitous damage in metallic aerospace structures is the slow development and propagation of fatigue cracks during the service life of the aircraft. This process, known as fatigue crack growth, is usually characterized using standardized coupon tests [2]. Fatigue crack growth (FCG) test data are most commonly presented in plots of FCG rate, d*a*/dN, (amount of crack growth per number of cycles) as a function of the cyclic crack-tip stress intensity factor, ΔK, as shown in Fig. 1.6. Three regions are shown, with Region I (the near-threshold region) being of primary concern to the present work. The near-threshold region corresponds to very low values of cyclic crack-tip stress intensity factor and is characterized by slow crack growth rates. It is of practical importance for life prediction because the majority of fatigue life for many aircraft components is consumed in this regime.

Naturally occurring cracks typically initiate under near-threshold loading conditions and propagate under increasing ΔK conditions. However, because of the time required to propagate cracks at low values of ΔK, laboratory tests are typically started in the Paris regime (Region II), and the applied loads are gradually reduced such that ΔK values decrease as the crack propagates. The procedure requires that care be taken to ensure that this artificial loading sequence does not affect the fatigue crack growth rate data. As a result, ASTM standard E647 ("Standard Test Method for Measurement of Fatigue Crack Growth Rates") was developed to ensure that satisfactory test results are obtained [2].

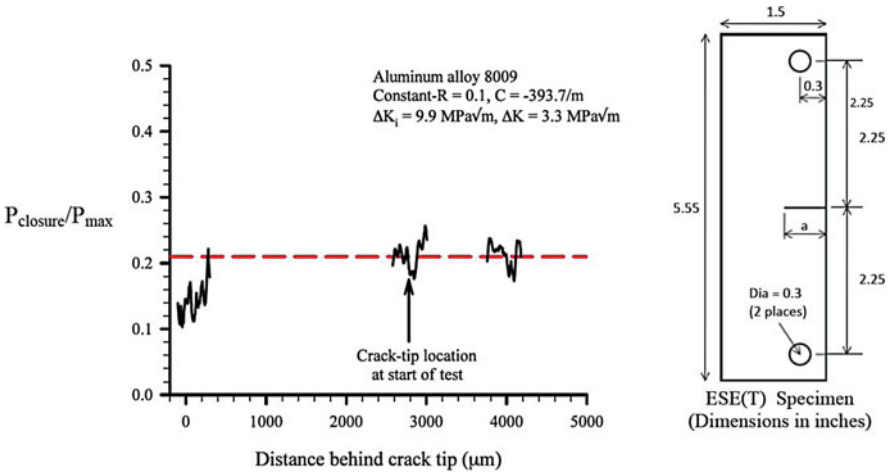

Fig. 1.7 Closure profile for a specimen of aluminum alloy 8009 with initial $K_{max} = 11$ MPa$\sqrt{}$m ($\Delta K_i = 9.9$MPa$\sqrt{}$m) and C= −393.7/m. Remote closure occurs

Recent research suggests that performing crack growth tests under ΔK-reduction conditions can adversely affect the FCG data [3, 4] due to a test-history-induced crack closure phenomenon [5–8]. Fatigue crack closure may result because of crack face contact near the crack tip during decreasing load but before the minimum value is reached. Although crack-tip closure is a naturally occurring phenomenon, the prescribed load reduction method can induce an artificial "remote closure" that occurs away from the crack tip and can artificially affect the FCG data. Thus, the test data may be affected by the crack-tip plasticity created at relatively high ΔK near the start of the FCG test and may not be an accurate indication of the mechanical performance of the material.

To assess the effects of testing procedure and the resulting remote closure on FCG data, fatigue crack growth tests were performed using closed-loop servo-hydraulic test machines with constant amplitude sinusoidal loading. Testing was conducted in accordance with ASTM standard E647 using eccentrically loaded single-edge notch tension (ESE(T)) specimens [9] having width, W, and thickness, B, of 38.1 mm and 2.3 mm, respectively. A schematic of the specimen is shown in the insert in Fig. 1.7. A computer-controlled system [10] was used to continuously monitor crack length during testing using the back-face compliance technique [11]. This system automatically adjusts the applied loads as the crack grows to ensure that programmed stress intensity factors are applied throughout the tests.

Crack closure data were obtained by analyzing a series of high-magnification (300–700X) digital images of the crack obtained during cyclic loading. A random pattern of 4-mm speckles was deposited on specimen surfaces in the region of crack growth to provide features whose motion could be tracked as a function of load using the VIC-2D [12] software program. The presence of crack closure was determined by tracking the relative displacement of speckles on features on either side of the

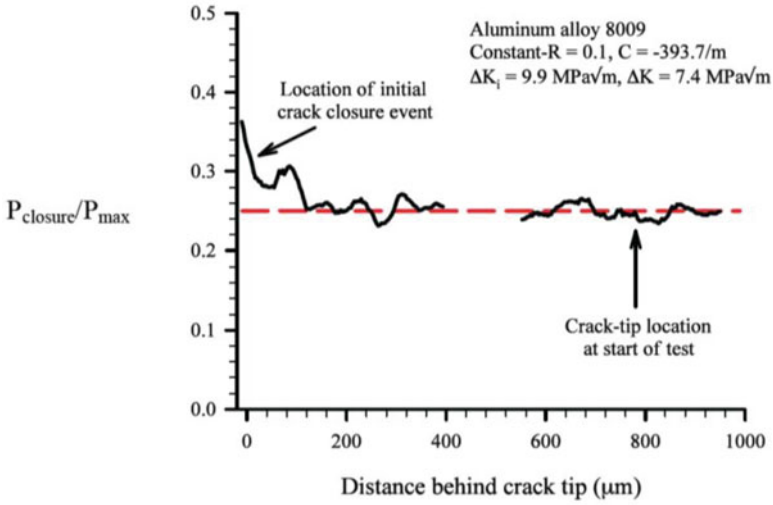

Fig. 1.8 Closure profile for a specimen of aluminum alloy 8009 with initial $K_{max} = 11\text{MPa}\sqrt{\text{m}}(\Delta K_i = 9.9\text{MPa}\sqrt{\text{m}})$ and C= −393.7/m. No remote closure

crack. After the digital images were analyzed using VIC-2D, the displacement data was analyzed using the Elber method to determine the load at which the crack closed [13]. In this method, the crack-opening displacement at a point along the crack wake is plotted against load resulting in a compliance plot for that specific crack wake location. Deviations in linearity on the plot indicate crack closure.

Preliminary results of this study [14] demonstrated that remote closure can occur when the testing parameters recommended by ASTM E647 are greatly exceeded. Figure 1.7 shows the closure profile of an aluminum alloy 8009 specimen tested at constant load ratio, $R = 0.1$, conditions (initial $K_{max} = 11\text{MPa}\sqrt{\text{m}}$ and C= −393.7/m) obtained using VIC-2D. The parameter, C, is the K-gradient and is evaluated as described in [2]. The experimental results of Fig. 1.7 were taken after approximately 3mm of crack growth ($\Delta K = 3.3\text{MPa}\sqrt{\text{m}}$), nearly the $R = 0.1$ FCG threshold value for this alloy. High-magnification images were taken at three locations: near the crack tip, the location corresponding to the start of the ΔK-reduction test, and further behind the crack tip (corresponding to steady-state pre-cracking at $\Delta K = 9.9\text{MPa}\sqrt{\text{m}}$, $R = 0.1$). This test exceeds the ASTM standard E647 limits on the K-gradient, C, by a factor of 5. Here, remote closure is shown to occur because crack closure occurs in the crack wake before occurring at the crack tip. The horizontal red dashed line in the figure corresponds to the mean normalized closure load ($P_{closure}/P_{max} = 0.21$) of the crack wake. In comparison, the crack tip closes later (at a lower load $P_{closure}/P_{max} = 0.12$), corresponding to remote closure, and is assumed to be an artifact of the test procedure.

Experimental results also show that the guidelines of ASTM standard E647 are overly conservative for some load scenarios. Figure 1.8 shows the closure profile (crack closure loads as a function of distance behind the crack tip) of a specimen

of an aluminum alloy 8009 specimen tested at an initial $K_{max} = 11$ MPa$\sqrt{m}$ and C= −393.7/m. The data presented in Fig. 1.8 correspond to a $\Delta K = 7.4$ MPa$\sqrt{m}$, after 800 μm of crack growth from the start of the test. Here, crack closure occurs in the crack wake at approximately $P_{closure}/P_{max} = 0.25$, as indicated by the horizontal red dashed line. Closure loads ($P_{closure}/P_{max}$) increase closer to the crack tip (within 100 μm of the crack tip), increasing to approximately $P_{closure}/P_{max} = 0.36$ at the crack tip. In this case, crack closure occurs first at the crack tip with closure occurring in the crack wake at lower loads, in a manner characteristic of steady-state crack closure in the absence of load history effects [15].

Damage Accumulation in Aluminum Microstructures

The continuum-level behavior discussed previously has its underpinnings at the micro- and nanoscales, so an understanding of the myriad of microscale and nanoscale mechanisms is needed to fully understand the mechanics of fracture. Thus, *Damage Science* methodologies are being developed. The work studies phenomena that occur at the scale of grains, dislocations, and atoms using novel computational and experimental methodologies.

At the microscale, near-crack-tip plasticity is dominated by the presence of large plastic strain gradients and the corresponding geometrically necessary dislocations (GNDs). The effect of GNDs on conventional plasticity formulations is overviewed by Hutchinson [16]. In this case, a critical issue is the underestimated work hardening during plastic deformation within the strain gradient-dominated field. This gives rise to strain gradient crystal plasticity formulations in which GNDs are assumed to dominate micron-scale plastic strain and to be associated with an internal length scale parameter. These gradient formulations are conceptually related to dislocation dynamics (as discussed in an upcoming section), thereby providing a natural linkage to simulations at submicron length scales.

Alternatively, conventional crystal plasticity (CCP) formulations can be employed to study material state fields within a microstructure even though they do not accurately capture some aspects of plastic deformation at or below the micron scale. As with all continuum plasticity formulations, CCP formulations must be subjected to initial calibration to the particular material at hand. Calibration typically consists of generating a polycrystal model—consisting of a representative population of grain size, aspect ratio, and texture—and matching a simulated response to an observed response by varying several material parameters. The result is a CCP model that is calibrated to incorporate micron-scale mechanisms in a homogenized sense.

Even with their quantitative limitations, CCP formulations are being used in conjunction with precise geometrical representations of metallic microstructures to develop a dramatically improved understanding of the sequence of plastic dissipation preceding crack growth at the micron scale. By incorporating models for slip accumulation, a relationship between plastic exhaustion and crack growth can be computed [17].

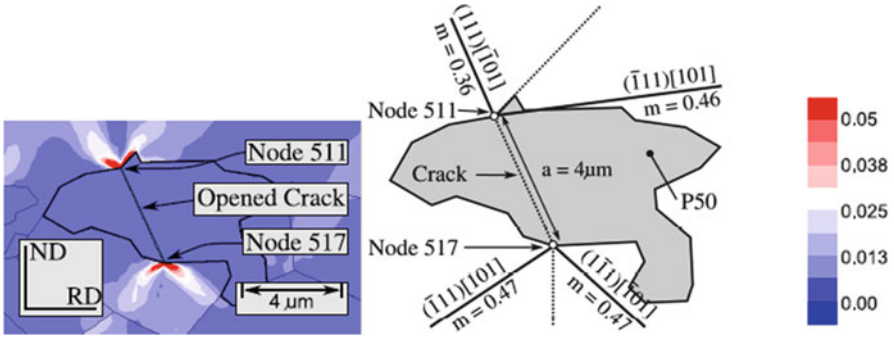

Fig. 1.9 Computed slip fields near a cracked constituent particle that was observed to nucleate a crack into the surrounding grains

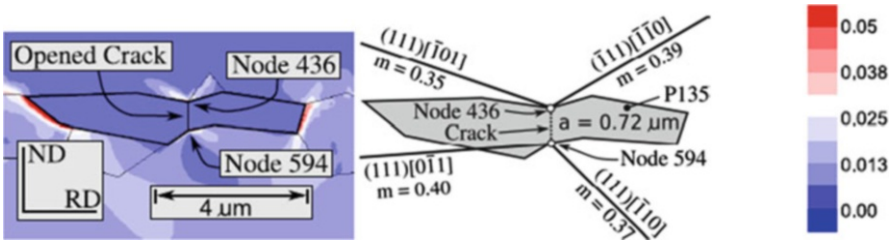

Fig. 1.10 Computed slip fields near a cracked constituent particle that was observed NOT to nucleate a crack into the surrounding grains

To better understand the plastic dissipation during cyclic loading that precedes nucleation events, finite element models were generated using observed microstructural data—constituent particles and grain texture and geometry—and slip localization and accumulation was computed near cracked particles [18]. Figures 1.9 and 1.10 illustrate the computed slip localization near a cracked constituent particle in aluminum AA 7075-T651 for two such models. The contoured fields in both figures are the maximum value of slip on any one of the twelve-face-centered cubic (FCC) slip systems; the corresponding values given by the contour bars are the magnitude of slip on the dominant system. The particle shown in Fig. 1.9, P50, was observed to nucleate a crack into the surrounding grains, while the particle in Fig. 1.10, P135, did not. It is apparent from these results that slip localization and accumulation (i.e., plastic dissipation) plays a governing role in crack nucleation at this scale; see [17] for further discussion. Figures 1.9 and 1.10 also show the correspondence between computed slip localization and dominant slip system directions, as measured via electron backscatter diffraction (EBSD). However, the directions of slip localization did not correspond with the nucleation direction, given by the dotted line in Fig. 1.9. This observation leads to a hypothesis for crack trajectory based on alternating shear or maximum tangential stress with neighboring grains. More simulations are currently underway to investigate these hypotheses.

Experimental Investigations at the Microscale

An environmental scanning electron microscope (ESEM) equipped with in situ loading frame and EBSD system has been developed to characterize damage processes in single crystals of pure aluminum and polycrystalline aluminum alloys (Fig. 1.11). The EBSD orientation mapping tools can be used to measure the extent of high plastic deformation near the fatigue crack tip and crack-tip wake. Plasticity near the crack tip is related to the plastic strain gradients and thus the geometrically necessary dislocation density.

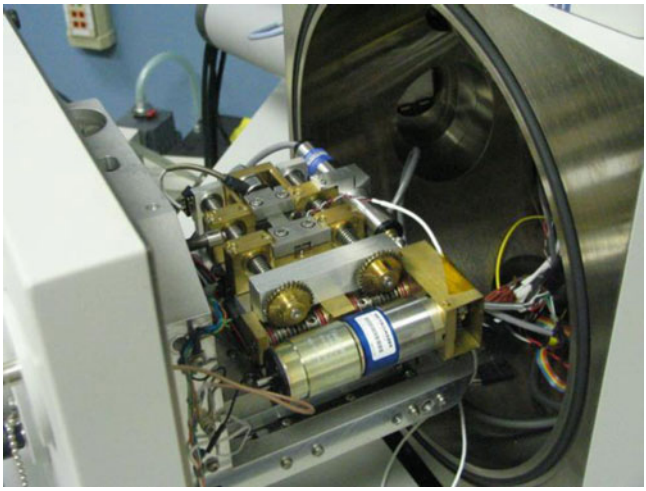

Fig. 1.11 Mechanical loading frame in environmental scanning electron microscope (ESEM) for the examination of damage propagation at high resolution

These dislocations result in bending of the lattice and may be detected as an orientation gradient within a single grain. Additionally, a zone of "significant plastic strain" about a fatigue crack tip and crack-tip wake can be determined by measuring the width of the highly defected region (e.g., green-to-red rainbow color scheme on misorientation maps). Experimentally determined locations of orientation discontinuities, e.g., at sector boundaries, slip bands, near the crack tip, and GND densities estimated from local lattice rotations can be compared with model predictions to enable the *physics-based* models to include correct input parameters, such as source and obstacle densities.

Recent studies [19, 20] of single crystals and bicrystals have shown that it is possible to extract some of the components of the Nye dislocation density tensor [21] using orientation data obtained by EBSD mapping, provided that the crystal orientation and deformation conditions are carefully controlled to constrain the number of independent components. The present work follows [19] and [20] and considers a connection between the GND content and the lattice curvature tensor through spatially resolved local orientation measurements using EBSD.

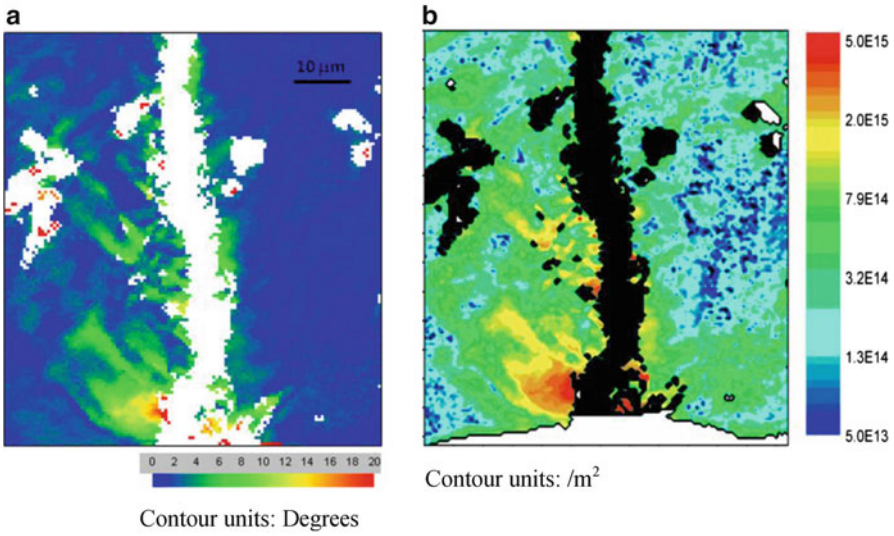

Fig. 1.12 Maps of misorientation and geometrically necessary dislocation density. (**a**) EBSD misorientation map (**b**) Enhanced dislocation density map

For the purpose of illustration, these approaches have been applied to the EBSD orientation data obtained from the vicinity of a fatigue crack in precipitation-hardened aluminum alloy Al-Cu-Mg 2024-T351. The intra-grain misorientation map (Fig. 1.12a) displays changes in the local orientation, along with large amounts of intragranular misorientation associated with the large plastic deformation in the vicinity of a crack-tip wake [22]. White regions in Fig. 1.12a correspond to pixels that were not indexed. The misorientation map reveals distinctions in the morphology of plastic damage, e.g., the presence of slipbands near the crack-tip wake. These maps suggest the presence of a high dislocation content resulting in extensive disorientation.

Figure 1.12b shows the estimated distribution of GND density within the scanned area. The regions of lower dislocation density (i.e., base material, $\sim 0.5\text{–}1 \times 10^{14}/\text{m}^2$) are separated by regions of higher dislocation density (i.e., *plastically deformed* crack wake, $\geq 10^{15}/\text{m}^2$ and higher) and can be identified by marked orientation change (Fig. 1.12a) or by the enhanced dislocation density (Fig. 1.12b) [22]. The boundaries of these banded structures (dislocation patterning) contain a high GND density, and regions within the bands are relatively free of dislocations that contribute to lattice curvature. An inhomogeneous distribution of the dislocation density becomes obvious for such cases.

The measurements of local orientation changes and estimates of GND content near the crack tips and wakes of fatigue cracks can be qualitatively compared with those predicted by computational models developed with the aid of molecular dynamics and finite element simulations. This experimental effort will contribute a

significant quantitative and physical understanding of damage mechanisms that will enable next-generation damage models to progress beyond the current empirical models.

In support of experimental studies at the microscale, a comprehensive metallic materials processing facility has been established. This facility enables unique heat treatments of commercially available alloys as well as for the production of idealized microstructures to study specific damage processes. Single and bi-crystalline pure and simple alloy materials are manufactured using either the Czochralski or Bridgman crystal growth methods. The materials are characterized for crystallographic orientation, and grain structure and mechanical test specimens are machined to study specific damage processes in specific crystallographic orientations.

Discrete Dislocation Simulation

Unlike continuum plasticity formulations wherein the elastic-plastic constitutive behavior is assumed, discrete dislocation plasticity approaches have been developed to predict both the plastic stress-strain response and the corresponding evolution of the dislocation structure as part of the solution [23]. Dislocation dynamics (DD) simulation methods have been developed to represent large numbers of dislocations at relatively large length scales compared to atomic dimensions. In these approaches, dislocations are represented as lines of displacement discontinuity where the magnitude of the discontinuity is equal to the Burgers vector. Away from the core region, the displacement, stress, and strain fields may be suitably represented by analytic elasticity solutions. Thus, the displacements and velocities of individual atoms are not computed. Simulations can involve infinite domains that are modeled using periodic boundary conditions or as finite domains with various prescribed boundary conditions.

In discrete dislocation plasticity, the goal of simulation is often to determine the amount of plastic strain exhibited by a material due to the generation and interaction of the dislocations. For example, Fig. 1.13 depicts the stress-strain response of a single-crystal aluminum loaded in tension using the in-house developed two-dimensional dislocation dynamics code DD-SIM. As shown, both the inelastic plastic yielding together with hardening can be obtained directly from the simulation [24].

Atomistic Simulation of Crack Growth

Atomistic simulation of fracture has been a topic of considerable study during the past two decades. Early studies were focused on idealized perfect or non-defect structures, but improved methodologies and increases in computing power are making the study of deformation and fracture in structural materials attainable. Although only very small volumes of material can be studied using atomistic simulation, the studies employ interatomic potentials that are grounded in the results of ab initio

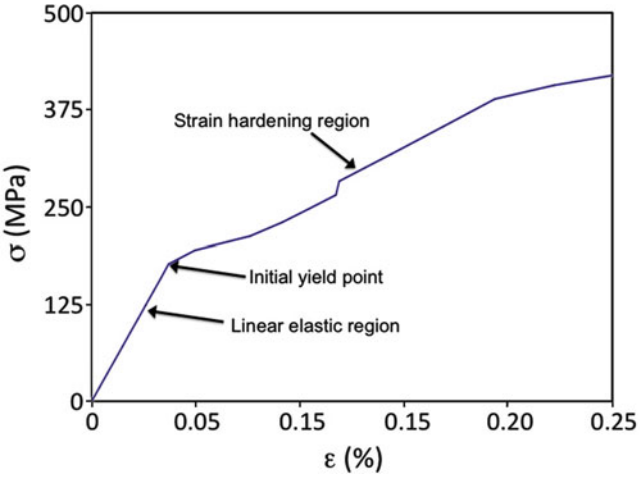

Fig. 1.13 Stress-strain behavior of an aluminum material domain under normal tensile loading

calculations and hence give the promise of understanding damage processes at a truly fundamental level (for an overview of atomistic simulation, see Allen and Tildesley [25]).

Atomistic simulations are used to determine the fundamental processes of crack initiation and growth including plastic mechanisms (e.g., twinning, dislocations, stacking faults) and the creation of free surfaces (i.e., crack propagation). Because of the extreme computational cost of interrogating large volumes of materials with atomistic simulation, both concurrent and sequential multiscale methods are being developed. The concurrent multiscale methods are developed to dramatically improve computational efficiency by virtually embedding a small (several million atom) atomistic simulation within a large finite element model [26], whereas the sequential multiscale methods recast the results of the atomistic simulations for use in continuum-based methods [27].

Fracture processes in aluminum and aluminum-based alloys are of particular interest. Recently, a number of atomistic simulation studies on intergranular and transgranular crack propagation in aluminum have been published [27–31]. The results of these investigations show that two main mechanisms of crack propagation and plasticity operate at the nanoscale. These mechanisms include propagation through deformation twinning and propagation through the emission of full dislocations from the crack tip (see Fig. 1.14). One major finding of these and other atomistic simulations disagrees with experiment: most atomistic simulations predict deformation twinning as the dominant deformation mechanism, whereas experimental observations show that dislocation slip is dominant in aluminum [28].

The discrepancy between simulations and experiments has attracted considerable attention among researchers because it prevents the reliable and accurate modeling of fracture in particular and puts doubt on the reliability of the atomistic simulations in general [17, 31]. Most likely, the source of this discrepancy is related to the very

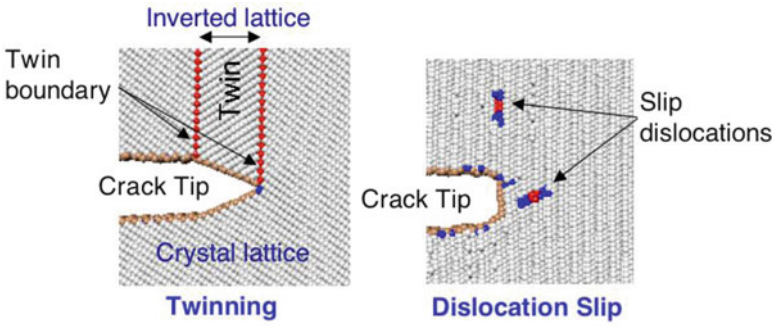

Fig. 1.14 Twinning and slip near a crack tip [27]

different length (nanometers vs. millimeters) and time (nanoseconds vs. seconds) scales at which simulations and experiments are usually performed. Nonetheless, the exact mechanism of how these length and time scales affect the propagation process remains unclear and is a very active topic of research.

To improve the understanding of the sources of the discrepancy between simulation and experiment, a detailed study has been undertaken to determine the conditions under which twinning or dislocation emission occur at a crack tip under Mode I loading [32]. The recently developed embedded statistical coupling method (ESCM) [26] for concurrent multiscale modeling was used. Studying the crack tip nucleation process at different crack orientations and loads revealed the existence of a transition stress intensity, K_{IT}, below which the crack emits full dislocations and above which deformation twinning becomes dominant. The transition stress intensity was found to depend on the crystallographic orientation and temperature. This understanding of the competition between the two mechanisms under the conditions of an atomistic simulation will enable determination of the regimes that are most suitable for study using these methods.

Fracture Mechanics of Composite Materials

The focus of much of the work conducted in the branch on composite materials has centered on the investigation of the damage tolerance capabilities of laminated composites and sandwich composites. A long-established theme of this work has been the development of standardized testing practices for characterizing failure modes of composite laminates, such as a double cantilever beam test [33] for measuring mode I delamination resistance and a curved beam test [34] for measuring the interlaminar strength of composite laminates. Efforts have also focused on computational methods for simulating failure mechanisms, geared towards improving the design and certification of structure manufactured from composite materials. The work was performed in support of a number of NASA aeronautics-related programs and has included collaborative efforts with major domestic airframe manufacturers.

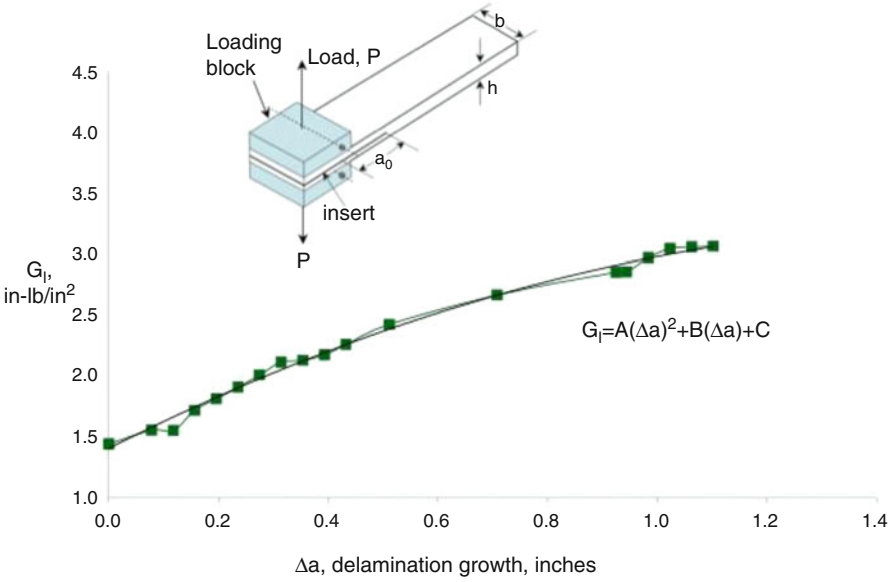

Fig. 1.15 Delamination resistance curve (R-curve) and DCB specimen

Mode I Fatigue Delamination Round Robin

An ASTM International Round Robin test exercise [35] is being conducted to develop a standard test method for mode I fatigue delamination propagation in unidirectional fiber-reinforced polymer matrix composites. Round robin participants include six different laboratories in three countries. The goal of this round robin is to develop a standard test method for determining delamination growth rate under constant amplitude fatigue loading as a function of the cyclic strain energy release rate, G_{Imax}.

The round robin uses the double cantilever beam (DCB) specimen, shown in the inset in Fig. 1.15, to determine the delamination growth rate, da/dN, of three different laminated composite materials [35]. Currently, standards exist for using the DCB specimen to determine mode I fracture toughness, G_{Ic}, (ASTM International Standard D5528) [33] and G_{Imax} for delamination onset under cyclic loading (ASTM International Standard D6115) [36].

Prior to fatigue testing, static DCB tests were conducted on the test materials, using ASTM Standard D5528, to determine critical displacement levels corresponding to delamination onset and data reduction constants for the fatigue tests. There is an artificial increase in the fracture toughness because the DCB specimen experiences fiber bridging as the delamination grows. The resulting curve of toughness vs. crack length, known as an R-curve, is shown in Fig. 1.15, where a polynomial

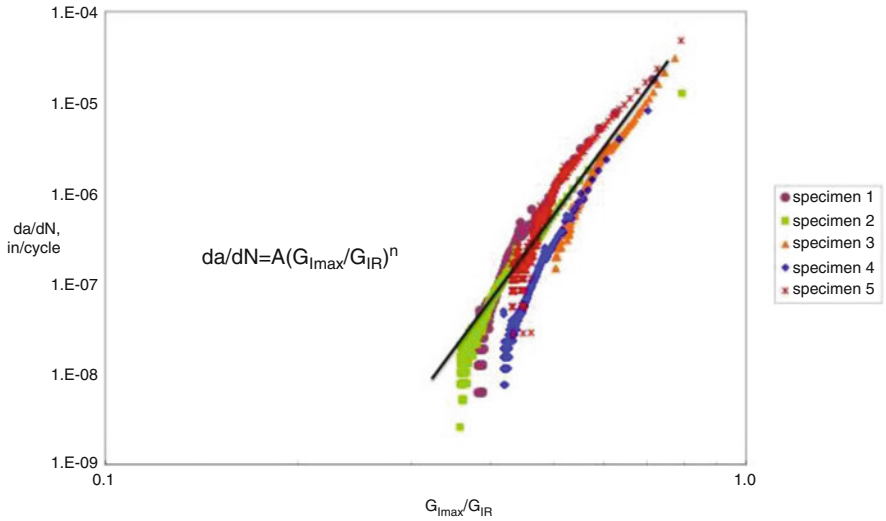

Fig. 1.16 Delamination growth rate data normalized by R-curve

expression has been fit to the static data. In order to account for the effect of fiber bridging in the fatigue tests, the fatigue data are normalized by this R-curve expression [35].

Fatigue tests were conducted in displacement control using an R-ratio of 0.1 and a frequency of 10Hz. For DCB specimens under displacement control, da/dN decreases as the delamination grows. To obtain the complete da/dN curve, tests were run at a G_{Imax} level just below G_{Ic} and allowed to continue until delamination arrested or until da/dN was 10^{-6}mm/cycle or less. Figure 1.16 shows results of the fatigue testing of five specimens of one material where the applied G_{Imax} has been normalized by the plateau value of the R-curve (G_{IR}). Delamination growth rates were consistent for all five specimens. Typically, a power-law expression of the form $da/dN = A(G_{Imax}/G_{IR})^n$ is fit to this data plot, as shown in the figure. The purpose of the round robin exercise is to streamline the test protocol to ensure that the test yields reliable fatigue delamination growth data [35].

Analysis Benchmarking

Over the past two decades, the use of fracture mechanics has become common practice for characterization of the onset and growth of delaminations. In order to predict delamination onset or growth, the calculated strain energy release rate components are compared to interlaminar fracture toughness properties measured over a range from pure mode I loading to pure mode II loading, using delamination growth characterization tests similar to the specimen discussed in the previous section.

The virtual crack closure technique (VCCT) is widely used for computing energy release rates based on results from 2D and 3D finite element analyses and for

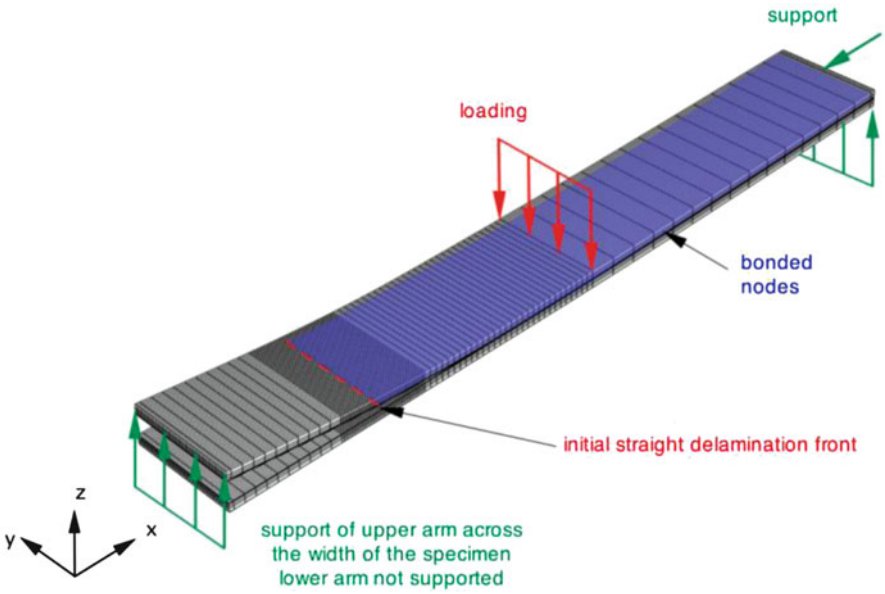

Fig. 1.17 Finite element mesh of a single-leg bending (SLB) specimen

supplying the mode separation required when using a mixed-mode fracture criterion [37]. The VCCT procedure was recently implemented in the commercial finite element codes ABAQUS®, MSC Nastran®, and Marc™ These implementations must be benchmarked to ensure that the method reproduces accurately reference solutions.

An approach for assessing the delamination propagation capabilities in commercial finite element codes under static loading was demonstrated for VCCT for ABAQUS® [38]. First, full three-dimensional finite element models of the single-leg bending (SLB) specimen shown in Fig. 1.17 were developed. Second, starting from an initially straight front, a benchmark solution that involved manual nodal release and computation of fracture parameters using VCCT was developed. Third, the commercial implementation was executed on an identical configuration. Comparison of the load-displacement relationship and the total strain energy release rates obtained from the commercial implementation and the benchmark solution showed that good agreement could be achieved by selecting the appropriate input parameters as shown in Fig. 1.18. Selecting the appropriate input parameters, however, was not straightforward and often required an iterative procedure. Overall, the results are encouraging but further assessment on a structural level is required.

Ongoing efforts include the application of the recently developed benchmark examples to the commercial finite element codes MSC Nastran™ and Marc™.

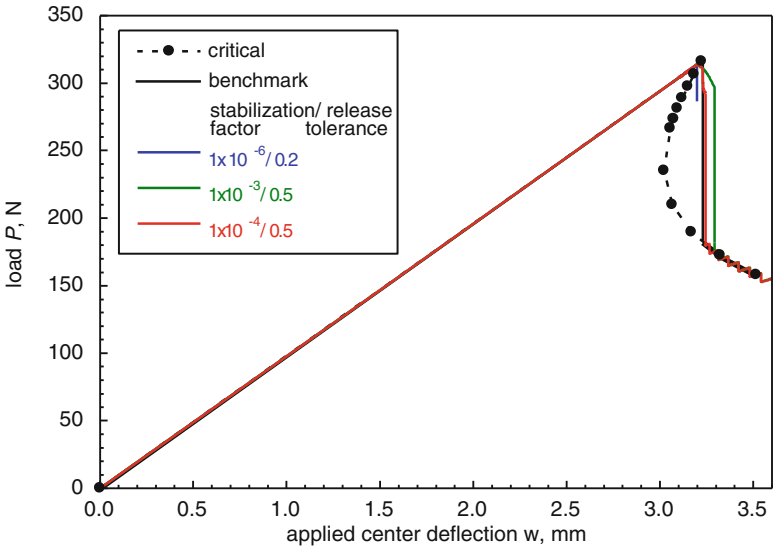

Fig. 1.18 Force-displacement response used in benchmarking procedure

Additionally, new benchmark examples are being created. The focus is on the assessment of the delamination growth prediction capabilities in commercial finite element codes when the specimen is subjected to cyclic loading.

Predicting Delamination Growth in Z-Pin-Reinforced Laminates

The previous two activities focused on methods for characterizing and analyzing delamination growth in composite laminates. Work has also been conducted to evaluate new methods for enhancement of the delamination resistance of a laminate. A number of techniques have been developed previously to achieve this enhancement, including stitching [39] and a process known as z-pinning [40]. Both methods involve the placement of fibers through the thickness of a laminate for the purpose of providing closure tractions to delaminations. In the case of z-pinning, pultruded carbon rods, available in diameters ranging from 0.25 to 0.5 mm, are placed into an uncured component using an ultrasonic hammer, as illustrated in Fig. 1.19a. An example of z-pins bridging the delamination in a DCB specimen is pictured in Fig. 1.19b.

As a delamination proceeds through a z-pin reinforced laminate, the pins first provide elastic closure tractions that oppose the delamination process until the bond between the pins and surrounding laminate begins to fail, after which the pins pull out from one side of the delaminating sections. A laminate showing debonded and partially debonded z-pins is shown in Fig. 1.19b. Subsequently, any analysis that

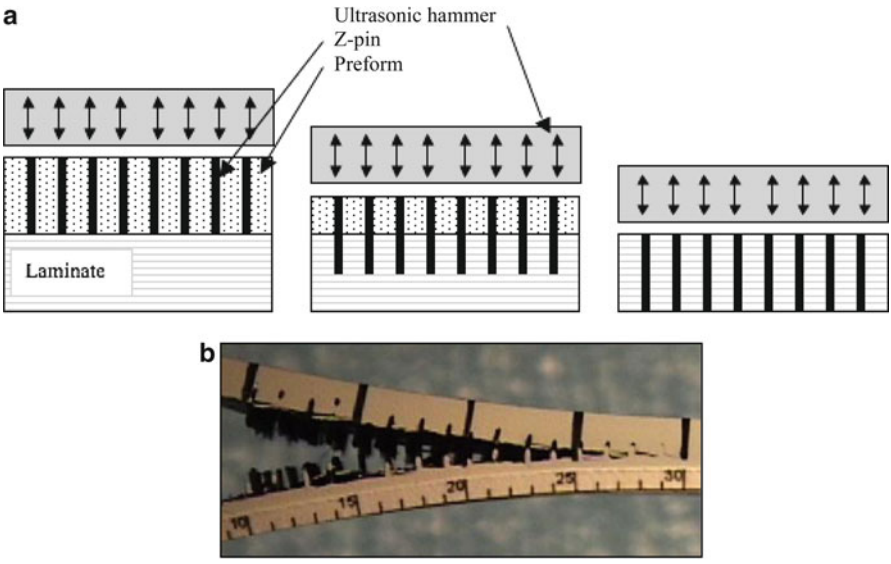

Fig. 1.19 Z-pin insertion process and example of z-pins bridging a delamination. (**a**) Z-pin insertion process (**b**) Z-pins bridging a delamination

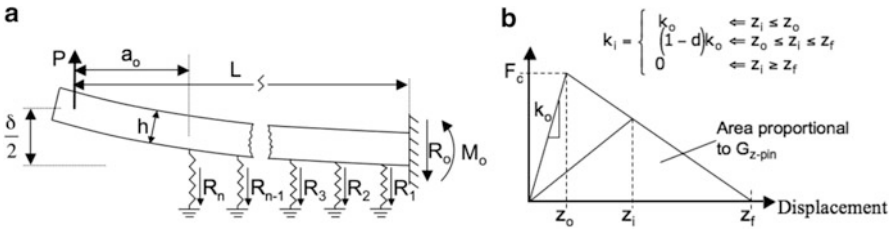

Fig. 1.20 Model of DCB specimen reinforced with z-pins. (**a**) Model of DCB specimen. (**b**) Constitutive z-pin failure law

aims to predict delamination growth under such circumstances must include the z-pin failure mechanisms. A recent analysis method [41] that modeled the DCB specimen as a cantilever beam and the bridging z-pins as a series of springs is illustrated in Fig. 1.20a.

The analysis results in a closed-form solution of the specimen stiffness vs. delamination length relationship, expressed generally as [41]:

$$C_n = \frac{\delta_n}{P} = 2\left[\frac{a_\circ^3}{3EI} + \frac{L^3 - a_\circ^3}{3E_{zp}I_{zp}}\right] + \frac{1}{3PE_{zp}I_{zp}}\left[\sum_{i=1}^{n} k_i z_i a_i^2 \left(a_i - 3L\right)\right] \tag{1}$$

where EI is the flexural rigidity of the beam. The other terms in Eq. (1) are illustrated in Fig. 1.20. The solution to Eq. (1) requires an iterative procedure in order to account for the bilinear constitutive law that is employed to represent the z-pin

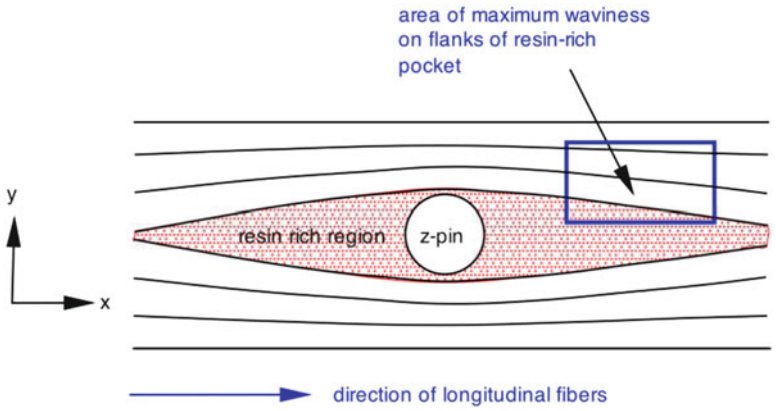

Fig. 1.21 Disruption of fiber alignment due to presence of a z-pin

failure mechanism and is illustrated in Fig. 1.20b. Once Eq. (1) is solved for a range of delamination lengths, the apparent debond toughness is computed using linear elastic fracture mechanics. The analysis was used to study the effect of z-pin spacing on delamination growth in the DCB specimens and also to estimate the enhancement in fracture toughness relative to the parent material. This and other studies [42, 43] indicated that small areal densities of z-pins, on the order of 1.5%, can increase delamination resistance by an order of magnitude. It is noted however that the inclusion of z-pins in a laminate can negatively affect some in-plane properties, which must therefore be evaluated when z-pins are being considered for use. An example of work performed in the branch on this topic follows in the proceeding section.

Influence of Fiber Misalignment Due to Z-Pins on the Compressive Response of Composite Laminates

Although the toughening properties of stitches, z-pins, and similar structures have been studied extensively, only a few investigations have focused on the effect of z-pins on the in-plane properties of laminates. Steeves demonstrated that disruption in the alignment of the fibers in the composite leads to a significant reduction in the in-plane compressive strength [44]. The z-pins may cause significant misalignment of the fibers (see Fig. 1.21) of the composite because the diameter of the z-pins ($\sim$ 280–510 μm) is large relative to the diameter of the fibers ($\sim$7μm). Previously, Sun and coworkers studied the influence of shear loads on the uniaxial compression strength of composites by testing an off-axis unidirectional lamina and extrapolating the compression strength [45, 46]. They found that the addition of small shear loads significantly reduce the compression strength of unidirectional composite lamina.

Therefore, the influence of compression and shear loads on the strength of composite laminates with z-pins was evaluated parametrically using a 2D finite element

code (FLASH) [47] based on Cosserat couple stress theory [48, 49]. Meshes of unit cells were generated for three unique combinations of z-pin diameter and density [50]. First, a laminated plate theory analysis was performed on several layups to determine the biaxial stresses in the zero-degree plies. Second, these stresses were used to determine the magnitude of the relative load steps prescribed in the FLASH analyses. Results indicated that increasing pin density was more detrimental to in-plane compression strength than increasing pin diameter. Compression strengths of lamina without z-pins agreed well with a closed-form expression derived by Budiansky and Fleck [51]. FLASH results for lamina with z-pins were consistent with the closed-form results, and FLASH results without z-pins, if the initial fiber waviness due to z-pin insertion was added to the fiber waviness in the material to yield a total misalignment. The addition of 10% shear to the compression loading significantly reduced the lamina strength compared to pure compression loading. The addition of 50% shear to the compression indicated shear yielding rather than kink-band formation as the likely failure mode. Two different stiffener reinforced skin configurations with z-pins, one quasi-isotropic and one orthotropic, were also analyzed. Six unique loading cases ranging from pure compression to compression plus 50% shear were analyzed assuming material fiber waviness misalignment angles of 0, 1, and 2 degrees. Compression strength decreased with increased shear loading for both configurations, with the quasi-isotropic configuration yielding lower strengths than the orthotropic configuration [50].

Designing Specimens for Characterizing Facesheet-Core Debonding in Sandwich Structure

Other activities at NASA Langley examine the damage tolerance capabilities of sandwich structure, with particular attention paid to sandwich employed in rotorcraft, as discussed in the following two sections. A recent activity, in support of NASA's Subsonic Rotary Wing Program and in collaboration with researchers from the University of Utah, is focused on the development of a standardized testing protocol for characterizing facesheet-core peel debonding in sandwich structure. The purpose of the test is to determine the critical strain energy release rate, G_c, associated with the facesheet-core debonding process. Following the recent identification of an appropriate test specimen [52], namely, the single cantilever beam (SCB) illustrated in Fig. 1.22, a procedure was developed for determining appropriate dimensions of the specimen [53].

The specimen sizing method is based on the beam-on-elastic-foundation model of the SCB specimen, depicted in Fig. 1.22. Subsequent analysis yields a closed-form solution of the stiffness-debond length relationship of the SCB specimen, which is used in the computation of G_c, and is expressed as [54]

$$C_{SCB} = \frac{\delta}{P} = \frac{4\lambda}{k}\left[\frac{\lambda^3 a^3}{3} + \lambda^2 a^2 F_1 + \lambda a F_2 + \frac{3ak}{10\lambda G_{xz,f} t_f b} + \frac{F_3}{2}\right] \tag{2}$$

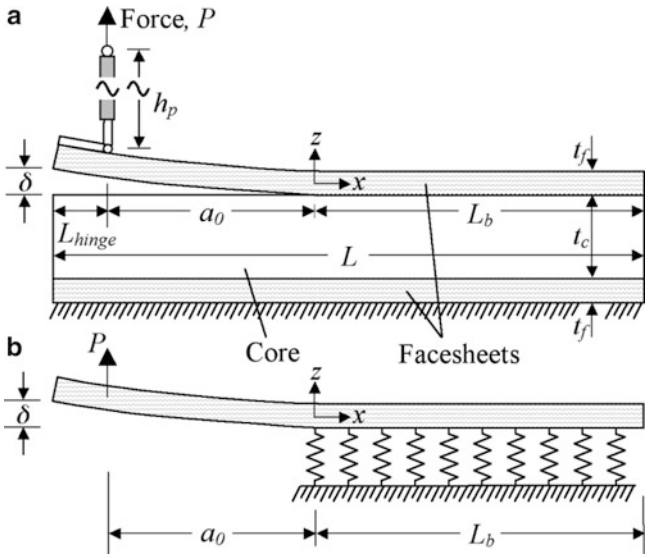

Fig. 1.22 Schematic and beam-on-elastic-foundation model of the SCB specimen

See [53] for a complete description of Eq. 2. This relationship is simplified to a format that is acceptable for use in a standardized testing protocol by imposing limitations on the SCB specimen dimensions, thereby resulting in the specimen design method (details of the limitations imposed on the SCB specimen can be found in [53]). With these limitations imposed, the stiffness-debond length relationship simplifies to [53]

$$C_{SCB} = \frac{4}{E_f b t_f^3}\left[a + \frac{1}{\lambda}\right]^3 \tag{3}$$

In addition to obtaining the desired stiffness-debond length relationship, the sizing method is also geared to result in specimens that behave in a linear elastic manner, as required by the procedures used for computing G_c from the test data. The sizing method will form part of the standardized testing protocol, developed under ASTM International's committee on composite materials, D30.

Predicting the Residual Compressive Strength of Impact-Damaged Sandwich Panels

In addition to being susceptible to delamination, sandwich composite materials are very susceptible to damage from out-of-plane loading, including low-velocity impact. These structures must be designed to sustain ultimate load with barely visible impact (BVID) damage. BVID can result in a compressive strength reduction

Fig. 1.23 Edgewise compression test on an impact-damaged sandwich panel

of 50% or more relative to an undamaged structure [55]. Subsequently, a series of edgewise compression tests were conducted to identify mechanisms involved in the compressive failure of impact-damaged sandwich panels, and analysis methods were developed to predict residual compressive strength of impact-damaged sandwich panels, as detailed in this section.

In order to measure the residual compressive strength of impact-damaged sandwich panels, the specimens are subjected to an axial compressive load using a test configuration similar to that shown in Fig. 1.23. Specimens subjected to this test have been observed to fail via one of two distinct failure modes [55], namely, kink-band propagation or indentation growth. Figure 1.24 shows shadow moiré images of failure sequences from the two failure modes. With kink-band propagation (Fig. 1.24a), the damage acts as a stress concentration similar to an open hole. As a compressive load is applied, the tows or fibers in the loading direction buckle and break normal to the plane of the facesheet, creating a band of broken fibers (on both sides of the impact damage) that propagates perpendicular to the loading direction. This kink band continues to stably propagate away from the damaged region with increasing load until a critical length is reached where the kink band becomes unstable, resulting in panel failure. For the indentation-growth failure mode (Fig. 1.24b), the residual indentation from the impact buckles inward and expands as the compressive load increases. The local buckle in the facesheet applies compressive loads to the core, causing additional crushing as well as elastic deflections. When a critical compressive force is reached, the facesheet rapidly buckles across the width and fails.

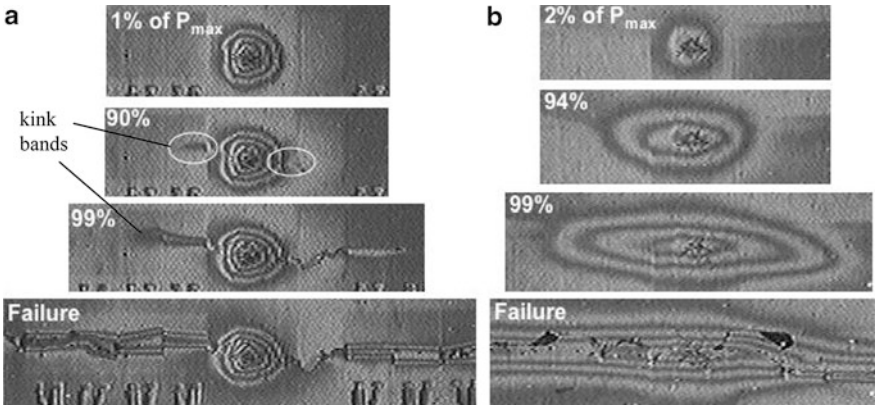

Fig. 1.24 Examples of compressive failure modes (P_{max} denotes force at panel failure). (**a**) Kink-band propagation (**b**) Indentation growth

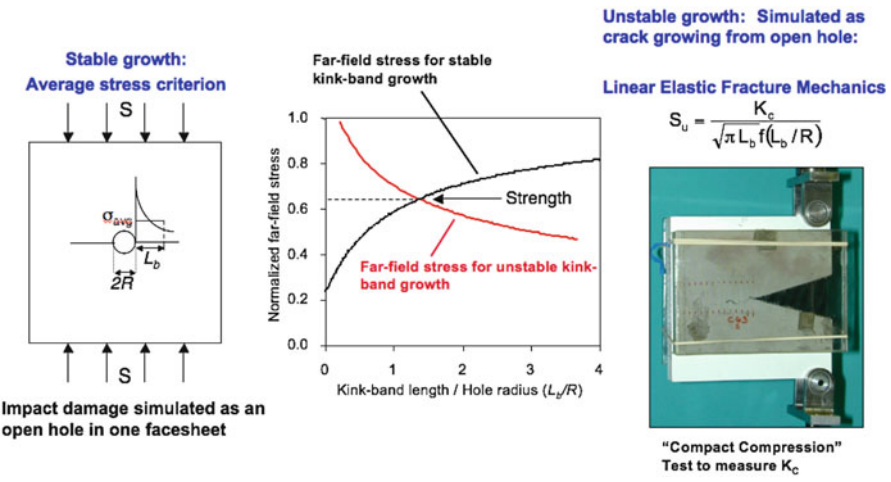

Fig. 1.25 Computing residual compressive strength associated with kink-band growth failure mode

Two separate analysis methods have been developed to predict the residual compressive strength of panels that exhibit either the kink-band or indentation-growth failure mode. The method tailored towards the former failure mode proceeds in two parts [56] and is a modification to a previous analysis [57] designed for predicting the residual compressive strength of impact-damaged, monolithic laminates. In the first part, the far-field stress required for stable kink-band growth is computed by modeling the damaged facesheet as an orthotropic plate with an open hole (Fig. 1.25). This computation is repeated for a range of kink-band lengths and plotted as illustrated in Fig. 1.25. In the second part, the far-field stress required for unstable kink-band growth is computed using linear elastic fracture mechanics,

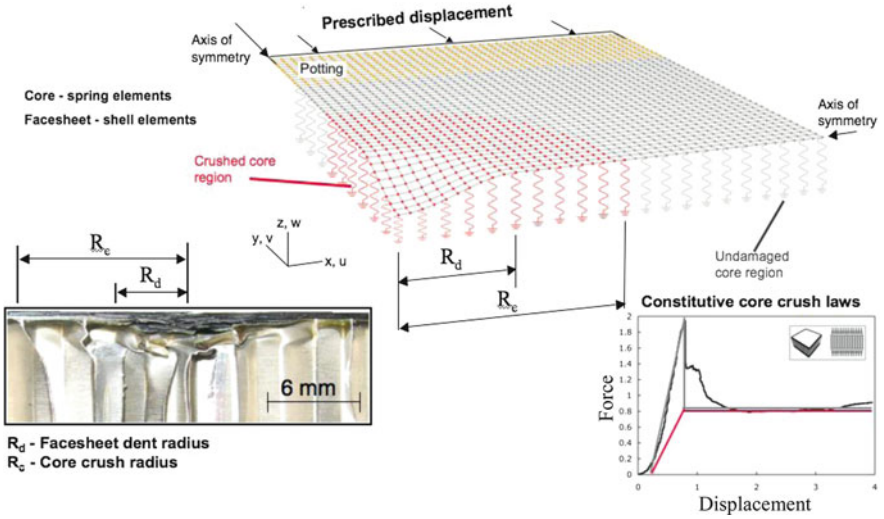

Fig. 1.26 Finite element model for computing residual strength associated with indentation growth

where the apparent fracture toughness associated with kink-band growth is measured using the compact-compression test pictured in Fig. 1.25. This computation is repeated for a range of kink-band lengths and superimposed onto the plot of far-field stress for stable kink-band growth. Given that panel failure is actually observed when kink-band growth transitions from stable to unstable conditions, the intersection of the plots in Fig. 1.25 is deemed to correspond to the residual strength of the panel.

The method for predicting residual compressive strength of sandwich panels exhibiting the indentation-growth failure mode involves a finite element model of the impact-damaged facesheet and core material [58]. The facesheet is modeled using shell elements, with the impact damage represented as a residual indentation, as depicted in the sample mesh shown in Fig. 1.26. The core material is represented using 2-node spring elements. Idealized constitutive traction laws shown in Fig. 1.26 are assigned to the spring elements to represent the crushing response of the core material during the compression test. These traction laws are based on the crush response of honeycomb structure subjected to flatwise compressive loading. The finite element analysis is executed using the loading and boundary conditions illustrated in Fig. 1.26. The global force-displacement response is computed at the end of each increment of the analysis. An example of a typical force-displacement response is presented in Fig. 1.27, where the maximum force is deemed to correspond to panel failure (on the basis of experimentally observed force-displacement response), and is thus used to compute residual compressive strength.

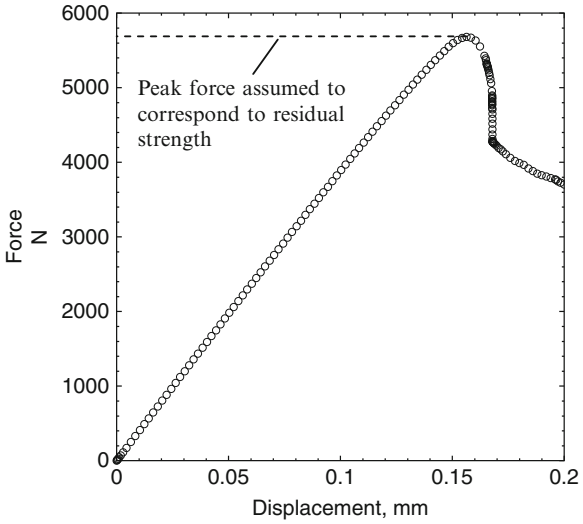

Fig. 1.27 Computed sandwich panel force-displacement response

Fractographic Analysis

Expertise developed within the branch in the area of damage tolerance of composite laminates has lead to requests for participation in aviation accident investigations, as highlighted in the following discussion of a fractographic analysis conducted in the branch in support of the American Airlines 587 accident investigation.

The accident investigation concerned the American Airlines Airbus A300–600R aircraft that crashed shortly after takeoff from Kennedy International Airport in November, 2001. The National Transportation Safety Board (NTSB) determined that the likely cause of the crash was the in-flight separation of the vertical stabilizer, arising from loads beyond ultimate design that were applied to the stabilizer. It was found that the stabilizer had separated from the main fuselage of the aircraft via failures of the stabilizer's six main lug attachment points. Researchers from the branch were requested by the NTSB to conduct fractographic analysis on samples removed from the aircraft debris as part of an effort to determine the cause of the accident [59].

Prior to the main fractography activities, a series of studies were conducted to evaluate the general configuration of the laminated structure, including stacking sequence, void content, chemical composition, and glass transition temperature. Specimens were removed from recovered debris, edges polished and examined optically to determine stacking sequence and void content. Infrared spectroscopy (IR) was used to determine chemical composition, and differential scanning calorimetry (DSC), among other methods, was used to determine glass transition temperature, T_g. These investigations indicated that the general state of the laminate was consistent with that prescribed by the manufacturers specifications.

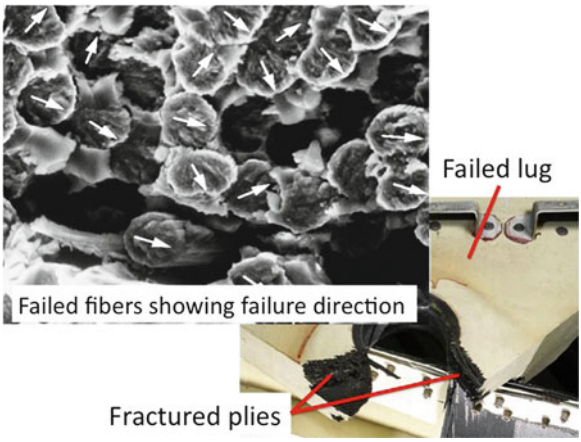

Fig. 1.28 Micrograph of failed fibers in lug attachment

An initial optical microscopy investigation was performed on recovered debris in order to establish a preliminary determination of the sequence of events and failure mechanisms involved in the failure of the composite lugs during detachment of the vertical stabilizer. This investigation determined that the three right-hand lugs failed in tension due to the bending moment-induced overloading, followed by a combination of tensile and compressive failures of the three left-hand lugs, caused by the continued bending deformation of the stabilizer after the right-hand lugs had failed.

The remaining challenge for the investigation was to determine whether any fracture events, such as delamination fatigue, had occurred prior to the accident. To address this, an extensive examination of recovered debris was conducted using scanning electron microscopy (SEM), involving magnifications ranging from 250X to over 2000X. This series of inspections revealed that no fracture events occurred prior to the accident, supporting the hypothesis that the failure was associated with unanticipated, in-flight loading, rather than damage that may have accumulated during the service of the aircraft. Data from the inspections also revealed the fracture modes that took place in each lug, and the direction of fracture propagation. For instance, the micrograph shown in Fig. 1.28 was used to reveal fracture surface features of broken fibers in a lug that had failed under tension [59]. In this instance, the crack growth direction of individual broken fibers (indicated by white arrows in Fig. 1.28) was averaged to determine an overall direction of fracture in this section of the lug [59].

Outlook

An overview of computational, analytical, and experimental strategies for fracture mechanics and its application to understanding damage tolerance of aerospace structures made of metallic and composite materials has been presented. Method-

ologies for simulating and characterizing damage growth in metallic materials under monotonic and cyclic loading were presented in this chapter, including continuum-based mechanics as well as a new paradigm in damage mechanics, damage science. In addition, damage tolerance capabilities for composite materials including current computational and experimental methods for composite structures were discussed.

A multiscale view of fracture mechanics for metallic materials is being developed with the aim of better understanding the fundamental mechanisms of damage progression at each relevant length scale. At the continuum scale, methods for predicting residual strength and fatigue crack growth in friction-stir-welded aluminum panels were presented. New investigations into the mechanisms of fatigue crack closure demonstrated that remote closure can occur when the testing parameters recommended by the testing standard, ASTM E647, are greatly exceeded. These continuum-level behaviors in metallic materials have underpinnings at the micro- and nanoscales, so an understanding of the myriad of microscale and nanoscale mechanisms is needed to fully understand continuum fracture mechanics. Thus, damage science methodologies are being developed to facilitate the understanding of durability and damage tolerance at a very fundamental level.

Recent developments for composite materials include the development of standardized test methods for delamination growth, prediction and verification methods for characterization of delamination and debonding, and fractographic analysis for determining underlying mechanisms of damage. The development of standardized test methods for composite materials is focused on developing a means for generating reliable data for characterizing delamination growth rate under constant amplitude fatigue loading. In addition, a test method for characterizing facesheet-core peel debonding in sandwich structure is being developed based on a specimen sizing method acceptable for use in a standardized test protocol. Other developments for composite materials include methods for predicting delamination growth in z-pin-reinforced laminates, determining the effect of the z-pins on laminate compressive strength, and predicting the residual compressive strength of impact-damaged sandwich panels. Finally, fractographic analysis has been used to determine mechanisms of delamination growth and laminate failure during the American Airlines Flight 587 accident investigation.

The outlook for durability and damage tolerance or, more generally, structures and materials includes a new paradigm in which models and experiments will overwhelm data management, storage, and, more importantly, analysis capabilities. Predictive capabilities will require multiscale and multi-physics software algorithms to be compatible with modern computing platforms having tens of thousands, or millions, of processors, including revolutionary computational paradigms (e.g., quantum computers). However, even with the most optimistic forecast of computing power, breakthroughs in computational methodologies are required to enable the bridging of the vast length and time scales discussed previously. Moreover, breakthroughs in experimental capabilities at all length scales are required to validate the computational methods and to facilitate developing fundamental knowledge in the assessment of structural and material response. This fundamental knowledge

enables development of technologies to support the design, development, processing, qualification, and sustainment of structural materials that are multifunctional, lightweight, durable, and have optimized performance characteristics. These technologies optimize material development and reduce material insertion time while enabling new and more aggressive structural designs for aerospace applications and for more demanding space exploration missions.

A well-supported infrastructure must include evolutionary and revolutionary computational and experimental facilities and the supporting personnel to operate them. Additionally, an environment that integrates analysis and experiment is required to provide seamless interactions between test and analysis. Most importantly, personnel with expertise in computer science and algorithm development are required to exploit the new computational architectures, develop new theories to better understand physical phenomena, and develop experimental capabilities that are needed to discover new phenomena, test hypotheses, and validate analyses.

Moreover, the future of structures and materials is based on knowledge beyond traditional structural mechanics and materials science disciplines and must integrate the previously disjoint testing and analysis elements of high-performance computing, solid mechanics, and manufacturing. This multidisciplinary knowledge base should address these disciplines including experimental methods development, analytical model development and characterization, verification and validation, data standards and structure, and manufacturing considerations. Skills that are critical to structures and materials include atomistic and multiscale simulation and experimental capabilities, and optical microscopy, scanning electron microscope and transmission electron microscope imaging, physical metallurgy, and organic chemistry. Relevant skills in mathematics and computer science (e.g., finite element analysis, molecular dynamics, multidisciplinary analysis and optimization, probability theory, and algorithm development for high-performance computing) are critical to multiscale, multidisciplinary simulation capabilities. Hence, the aforementioned personnel skills and facilities that integrate both experimental and computational capabilities offer the promise of dramatically increasing the insight of the research engineers and facilitating their understanding of fundamental physical processes in structures and materials.

Acknowledgements The authors would like to thank the members of the Durability, Damage Tolerance, and Reliability Branch and for their contributions to this chapter. In particular, the authors are grateful to Dr. Ronald Kruger of the Durability, Damage Tolerance, and Reliability Branch and Dr. Bourama Toni from Virginia State University for their thorough review of this chapter.

References

1. Smith, S.W., Newman, J.A., James, M.A., Donald, J.K., Brazill, R.L., Schultz, R.W., Blair, A., Seshadri, B.R.: "An On-line Methodology for Measuring Residual Stress and Producing Reliable Fatigue Life Assessments," Presented at the 9th International ASTM/ESIS Symposium on Fatigue and Fracture Mechanics. (37th ASTM National Symposium on Fatigue and Fracture Mechanics), May 20–22, 2009, Vancouver, BC

2. ASTM International Standard E647–08: “Standard Test Method for Measurement of Fatigue Crack Growth Rates” 2008 ASTM International Annual Book of Standards, Vol. 03.01, ASTM International, West Conshohocken, PA
3. Newman, Jr. J.C., Yamada, Y.: Compression precracking methods to generate near-threshold fatigue-crack growth-rate data. Int. J. Fatig. **32**, 879–885 (2010)
4. Forth, S.C., Newman, Jr. J.C., Forman, R.G.: On generating fatigue crack growth thresholds. Int. J. Fatig. **25**, 9–15 (2003)
5. Elber, W.: Fatigue crack closure under cyclic tension. Eng. Fract. Mech. **2**, 37–45 (1970)
6. Newman, J.A., Piascik, R.S.: Plasticity and roughness closure interactions near the fatigue crack growth threshold. In: Reuter, W.G., Piascik, R.S. (eds.) Fatigue and Fracture Mechanics: 33rd volume, ASTM STP 1417. ASTM International, West Conshohocken, PA (2002)
7. Suresh, S.: Fatigue of Materials. Cambridge University Press, Cambridge (1991)
8. Newman, J.A.: The effects of load ratio on threshold fatigue crack growth of aluminum alloys. Ph.D. dissertation, Virginia Polytechnic Institute and State University (2000)
9. Piascik, R.S., Newman, Jr. J.C, Underwood, J.H.: The extended compact tension specimen. Fatig. Fract. Eng. Mater. Struct. **20**, 559–563 (1997)
10. Fracture Technology Associates: User’s Reference Manual for Automated Fatigue Crack Growth (Compliance), Version 2.43, Fracture Technology Associates, Bethlehem, PA
11. Deans, W.F., Jolly, C.B., Poyton, W.A., Watson, W.: A strain gauging technique for monitoring fracture specimens during environmental testing. Strain **13**, 152–154 (1977)
12. Sutton, M.A., Orteu, J.-J., Schreier, H.W.: Image Correlation for Shape, Motion and Deformation Measurements. Springer Science Business Media, New York, NY (2009)
13. Elber, W.: Crack Closure and Crack Growth Measurements in Surface-Flawed Titanium Alloy Ti-6Al-4V, NASA TN-D-8010 (1975)
14. Leser, W.P., Newman, J.A., Johnston, W.M.: Fatigue Crack Closure Analysis Using Digital Imaging Correlation, NASA TM-2010–216695 (2010)
15. Riddell, W.T., Piascik, R.S.: Stress Ratio Effects on Crack Opening Loads and Crack Growth Rates in Aluminum Alloy 2024, NASA/TM-1998–206929
16. Hutchinson, J.W.: Plasticity at the micron scale. Int. J. Solid Struct.**37**, 225–238 (2000)
17. Hochhalter, J.D., Littlewood, D.J., Christ, R.J., Veilleux, M.G., Bozek, J.E., Ingraffea, A., Maniatty, A.M.: A geometric approach to modeling microstructurally small fatigue crack formation: II. Physically-based modeling of microstructure-dependent slip localization and actuation of the crack nucleation mechanism in AA 7075-T651. Model. Simulat. Mater. Sci. Eng. **18** (2010)
18. Bozek, J.E., Hochhalter, J.D., Veilleux, M.G., Liu, M., Heber, G., Sintay, S.D., Rollett, A.D., Littlewood, D.J., Maniatty, A.M., Weiland, H., Christ Jr., R.J., Payne, J., Welsh, G., Harlow, D.G., Wawrzynek, P.A., Ingraffea, A.R.: A geometric approach to modeling microstructurally small fatigue crack formation: I. Probabilistic simulation of constituent particle cracking in AA 7075-T651. Model. Simulat. Mater. Sci. Eng. **16** (2008)
19. Sun, S., Adams, B.L., King, W.E.: Observations of lattice curvature near the interface of a deformed aluminum crystal. Phil. Mag. A **80**(1), 9–25 (2000)
20. Kysar, J.W., Briant, C.L.: Crack tip deformation fields in ductile single crystals. Acta Mater. **50**, 2367–2380 (2002)
21. Nye, J.F.: Some geometrical relations in dislocated crystals. Acta Metall. **1**, 153–162 (1953)
22. Gupta, V.K.: Ph.D. Dissertation, University of Virginia, Charlottesville (2009)
23. Arsenlis, A., Cai, W., Tang, M., Rhee, M., Oppelstrup, T., Hommes, G., Pierce, T.G., Buylatov, V.V.: Enabling strain hardening simulations with dislocation dynamics. Model. Simulat. Mater. Sci. Eng. **15**, 553–595 (2007)
24. Glaessgen, E.H., Saether, E., Hochhalter, J.D., Yamakov, V.: Modeling near-crack-tip plasticity at nano to micro scales. To be presented at the 51st AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference and Exhibit, Orlando, FL, April 12–15, 2010
25. Allen M.P., Tildesley D.J.: Computer Simulation of Liquids. Oxford science publications, Oxford (1987)

26. Saether, E., Yamakov, V., Glaessgen, E.H.: An embedded statistical method for coupling molecular dynamics and finite element analyses. Int. J. Numer. Meth. Eng. **78**, 1292–1319 (2009)
27. Yamakov, V., Saether, E., Phillips, D.R., Glaessgen, E.H.: Molecular-dynamics simulation-based cohesive zone representation of intergranular fracture processes in aluminum. J. Mech. Phys. Solid **54**, 1899–1928 (2006)
28. Farkas, D., Duranduru, M., Curtin, W.A., Ribbens, C.: Multiple-dislocation emission from the crack tip in the ductile fracture of Al. Phil. Mag. A **81**, 1241–1255 (2001)
29. Hai, S., Tadmor, E.B.: Deformation twinning at aluminum crack tips. Acta Mater. **51**, 117–131 (2003)
30. Tadmor, E. B., Hai, S.: A peierls criterion for the onset of deformation twinning at a crack tip. J. Mech. Phys. Solid **51**, 765–793 (2003)
31. Warner, D.H., Curtin, W.A., Qu, S.: Rate dependence of crack-tip processes predicts twinning trends in f.c.c. metals. Nat. Mater. **6**, 876–880 (2007)
32. Yamakov, V., Saether, E., Glaessgen, E.: A continuum-atomistic analysis of transgranular crack propagation in aluminum. In: 50th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference and Exhibit, Palm Springs, CA, May 4–7, 2009
33. ASTM International Standard D5528–01: Standard test method for Mode I interlaminar fracture toughness of unidirectional fiber-reinforced polymer matrix composites. 2008 ASTM International Annual Book of Standards, vol. 15.03. ASTM International, West Conshohocken, PA (2007)
34. ASTM International Standard D6415/D 6415M-06a: Standard test method for measuring the curved beam strength of fiber-reinforced polymer-matrix composites. 2008 ASTM International Annual Book of Standards, vol. 15.03. ASTM International, West Conshohocken, PA
35. Paris, I.: Mode I fatigue delamination propagation of unidirectional fiber-reinforced polymer matrix composites (DCB). ASTM International Committee D30 Inter-laboratory Study, ILS # 0189 (2009)
36. ASTM International Standard D6115–97: Standard test method for Mode I fatigue delamination growth onset of unidirectional fiber-reinforced polymer matric composites. 2008 ASTM International Annual Book of Standards, vol. 15.03. ASTM International, West Conshohocken, PA (2004)
37. Rybicki, E.F., Kanninen, M.F.: A finite element calculation of stress intensity factors by a modified crack closure integral. Eng. Fract. Mech. **9**, 931–938 (1977)
38. Krueger, R.: An Approach to Assess Delamination Propagation Simulation Capabilities in Commercial Finite Element Codes, NASA/TM-2008–215123, 2008
39. Poe, Jr., C.C., Harris, C.E.: Mechanics of Textile Composites Conference, NASA Contractor Report, NASA/CP-3311, Parts 1 and 2, 1995
40. Freitas, G., Magee, C., Boyce, J., Bott, R.: Service tough composite structures using the Z-direction reinforcement process. In: Proceedings of the 9th DoD/NASA/FAA Conference on Fibrous Composites in Structural Design, Lake Tahoe, Nevada, USA, November 1991, NASA-CR-198718
41. Ratcliffe, J., O'Brien, T.K.: Discrete spring model for predicting delamination growth in Z-fiber reinforced DCB specimens. NASA Technical Memorandum, NASA/TM-2004–213019, 2004
42. Cartié, D.D.R., Partridge, I.K.: A finite element tool for parametric studies of delamination In Z-pinned laminates. In: Proceedings of the Sixth International Conference on Deformation and Fracture of Composites, pp. 49–55, Manchester, UK, April 2001
43. Robinson, P., Das, S.: Mode I DCB testing of Z-fiber reinforced laminates: a simple model for the investigation of data reduction strategies. J. Eng. Fract. Mech. 71(3), 345–364 (2004)
44. Steeves, C.A.: Mechanics of failure in composite structures. Ph.D. Dissertation, Department of Engineering, University of Cambridge, Cambridge (2001)
45. Sun, C.T.: Novel methods for testing and modelling composite materials and laminates. In: Proceedings of the Second International Conference on Composites Testing and Model Identification, CompTest 2004, Bristol, England, September, 2004

46. Sun, C.T., Jun, A.W.: Compressive strength of unidirectional composites with matrix nonlinearity. Compos. Sci. Tech. **52**(4), 577–587 (1994)
47. Fleck, N.A., Shu, J.Y.: Microbuckle initiation in fibre composites: a finite element study. J. Mech. Phys. Solid **43**(2), 1887–1918 (1995)
48. Shu, J.Y., Fleck, N.A.: User's manual for finite element code for fibre microbuckling. Cambridge University Engineering Department C-MATS Technical Report 224 (ISSN 0309–6505), May, 1995
49. Liu, D., Fleck, N.A.: User's manual II for finite element code FLASH for fibre microbuckling. Cambridge University Engineering Department C-MICROMECH Technical Report 29 (ISSN 0309–7420), November, 1999
50. O'Brien, T.K., Krueger, R.: Influence of compression and shear on the strength of composite laminates with Z-pinned reinforcement. Appl. Compos. Mater. **13**, 173–189 (2006)
51. Budiansky, B., Fleck, N.A.: Compressive failure of fibre composites. J. Mech. Phys. Solid **41**(1), 183–211 (1993)
52. Weaver, C.: Evaluation of Mode I fracture mechanics test methods for sandwich composites. M.Sc Thesis, University of Utah, Salt Lake City, UT (2009)
53. Ratcliffe, J.: Sizing single cantilever beam specimens for characterizing facesheet/core peel debonding in sandwich structure, NASA Technical Publication, NASA/TP-2010–216169, 2010
54. Li, X., Carlsson, L.A.: Elastic foundation analysis of tilted sandwich debond (TSD) specimen. J. Sandwich Struct. Mater. **2**, 3–32 (2000)
55. Cvitkovich, M.K., Jackson, W.C.: Compressive failure mechanisms in composite sandwich structures. J. Am. Helicopter Soc. **44**(4), 260–268 (1999)
56. Ratcliffe J., Jackson, W.C., Schaff, J.: Compression strength prediction of impact-damaged composite sandwich panels. In: Proceedings of the American Helicopter Society 60th Annual Forum, Baltimore, MD, June 7–10, 2004
57. Soutis, C., Fleck, N.A.: Static compression failure of carbon fibre T800/924C composite plate with a single hole. J. Compos. Mater. **24**, 536–558 (1990)
58. Ratcliffe, J., Jackson, W.C.: A Finite Element Analysis for Predicting the Residual Compressive Strength of Impact-Damaged Sandwich Panels. NASA Technical Memorandum, NASA/TM-2008–215341, 2008
59. Fox, R.F., Schulteisz, C.R., Reeder, J.R., Jensen, B.J.: Materials examination of the vertical stabilizer from American Airlines Flight 587. In: Proceedings of Materials Science and Technology, vol. 2, pp. 171–185, 2005

Chapter 2
On the Γ-Convergence Theory and Its Application to Block Copolymer Morphology

Xiaofeng Ren

Introduction

In this chapter we discuss one major phenomenon that has generated much interest among analysts and applied mathematicians working in the area of the calculus of variations. It is the rigorous study of singularly perturbed variational problems and their associated Euler–Lagrange equations. Usually in such a problem a small, positive parameter ε appears in front of the highest order term, often the gradient, of the energy functional. Examples include the Allen–Cahn problem in the phase transition theory, the Cahn–Hilliard problem for binary alloys, and the Ginzburg–Landau problem for superconductors.

Singular perturbation is often accompanied by a concentration phenomenon. Energy minimizers, local minimizers, and saddle points form localized structures such as interfaces, droplets, spikes, and vortices.

One ingredient in many currently studied variational problems is non-locality. Examples in condensed materials include charged Langmuir monolayers, chiral liquid crystals, and most famously block copolymers. In the Ohta–Kawasaki density functional theory of diblock copolymers there is a nonlocal quadratic term in the integrand of the free energy functional. Non-locality may also be introduced by an additional variable, like the magnetic field in the Ginzburg–Landau problem, that mediates the primary variable nonlocally. Such non-locality often forces the above-mentioned structures, like interfaces and droplets, to periodically repeat themselves.

We will discuss these issues with the diblock copolymer problem as the main example, although the techniques one learns here may be used in many other problems.

The tool in our studies is the Γ-convergence theory developed by De Giorgi [5] and later applied to some of the above-mentioned problems by Modica [9],

Xiaofeng Ren (✉)
Department of Mathematics, The George Washington University
Washington, DC 20052, USA
e-mail: ren@gwu.edu

B. Toni (ed.), *Advances in Interdisciplinary Mathematical Research*, Springer Proceedings in Mathematics & Statistics 37, DOI 10.1007/978-1-4614-6345-0_2,

Modica and Mortola [10], Kohn and Sternberg [8], Ren and Wei [15], and others. This theory produces a singular limit, known as the Γ-limit, as the singular perturbation parameter ε tends to zero. In the case of diblock copolymers the Γ-limit is a free boundary problem which is simpler than the original integrodifferential equation.

The Γ-limit retains many vital properties of the original problem. Solving the Γ-limit yields much information about the original problem. We will solve the Γ-limit of the diblock copolymers in one dimension to obtain the so-called lamellar solutions that are used to model one of the fundamental phases in the theory of block copolymer morphology: the lamellar phase (see Fig. 2.1).

The Ohta–Kawasaki Theory of Diblock Copolymers

The main example in these lectures is the Ohta–Kawasaki density functional theory of diblock copolymer morphology. A diblock copolymer melt is a soft material, characterized by fluid-like disorder on the molecular scale and a high degree of order at a longer length scale. A molecule in a diblock copolymer is a linear sub-chain of A-monomers grafted covalently to another sub-chain of B-monomers. Because of the repulsion between the unlike monomers, the different type sub-chains tend to segregate, but as they are chemically bonded in chain molecules, segregation of sub-chains cannot lead to a macroscopic phase separation. Only a local micro-phase separation occurs: micro-domains rich in A monomers and micro-domains rich in B monomers emerge as a result. These micro-domains form patterns that are known as morphology phases. Various phases, including lamellar, cylindrical, spherical, and gyroid, have been observed in experiments. See Fig. 2.1.

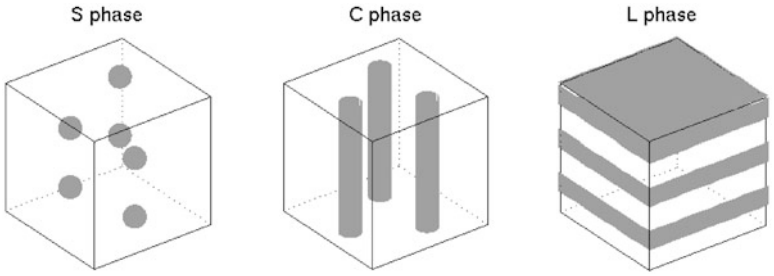

Fig. 2.1 The spherical, cylindrical, and lamellar morphology phases commonly observed in diblock copolymer melts. The *dark color* indicates the concentration of type A monomers, and the *white color* indicates the concentration of type B monomers

The free energy of a diblock copolymer melt proposed by Ohta and Kawasaki [13] takes the form

$$I(u) = \int_D [\frac{\varepsilon^2}{2}|\nabla u|^2 + W(u) + \frac{\sigma}{2}|(-\Delta)^{-1/2}(u-a)|^2]\,dx. \tag{1}$$

Here the sample occupies a bounded and smooth open set D in R^n. The physically relevant dimensions are $n = 1,2,3$. A derivation of this model written for mathematicians may be found in Choksi and Ren [3].

This is a rather complex variational problem. It was first introduced to mathematicians by Nishiura and Ohnishi [12]. The function u is the density field. $u(x)$ is the relative density of A monomers at the place x in a sample. If $u(x)$ is close to 1, then A monomers occupy x; if $u(x)$ is close to 0, then B monomers occupy x; if $u(x)$ lies between 0 and 1, then a mixture of A monomers and B monomers occupy x.

There are three parameters in this problem: ε, σ, and a. They are all positive. ε must be small, and a must be strictly between 0 and 1, i.e.,

$$a \in (0,1). \tag{2}$$

We will discuss σ a little later.

The function W is a balanced double well function. For simplicity we can take it to be

$$W(t) = \frac{1}{4}t^2(1-t)^2. \tag{3}$$

In general, W is a smooth function. $W(t) \geq 0$ for all $t \in (-\infty,\infty)$ and $W(t) = 0$ if and only if $t = 0$ or $t = 1$. A technical condition is that W has a certain growth rate which is quadratic, i.e., there exist $t_0 > 0$, $C_1 > 0$ and $C_2 > 0$ such that for all t with $|t| > t_0$,

$$C_1|t|^2 \leq W(t) \leq C_2|t|^2. \tag{4}$$

Note that W in (3) does not have a quadratic growth rate. We can modify this W so that for $t \in [-1,2]$ it is given by (3) and for $t \in (-\infty,\infty)\backslash[-1,2]$, it is given by a positive function which grows quadratically as $|t| \to \infty$.

The nonlocal operator $(-\Delta)^{-1/2}$ is defined from the $-\Delta$ operator by first solving

$$-\Delta v = u - a \text{ in } D, \quad \partial_\nu v = 0 \text{ on } \partial D, \quad \int_D v(x)\,dx = 0. \tag{5}$$

This defines the operator $(-\Delta)^{-1}$ by $(-\Delta)^{-1}(u-a)=v$. On $\{w \in L^2(D) : \int_D w(x)\,dx = 0\}$, $(-\Delta)^{-1}$ is a bounded and positive operator, for which one can define its positive square root $(-\Delta)^{-1/2}$ via the spectrum of $(-\Delta)^{-1}$. Without the nonlocal term or in other words if $\sigma = 0$ in (1), the functional I becomes the Cahn–Hilliard functional [1] which is used to study binary alloys.

Of course for (5) to be solvable one must assume that $\int_D (u(x) - a)\,dx = 0$. We use

$$\overline{u} = \frac{1}{|D|}\int_D u(x)\,dx \tag{6}$$

to denote the average of u. Then we impose the condition $\overline{u} = a$ on u. Physically this means that the average A monomer density in a sample is fixed at a and the average B monomer density at $1 - a$. This happens if all the chain molecules in a diblock copolymer are the same. Moreover to make the first term in (1) meaningful, we need $u \in W^{1,2}(D)$.

The Euler–Lagrange equation of I is an integrodifferential equation

$$-\varepsilon^2\Delta u + W'(u) + \sigma(-\Delta)^{-1}(u-a) = \lambda, \text{ in } D;\ \partial_\nu u = 0, \text{ on } \partial D. \tag{7}$$

The constant λ on the right side of the last equation is unknown. It is the Lagrange multiplier corresponding to the constraint $\overline{u} = a$.

The Γ-Convergence Theory

We now introduce the analytic tool of the Γ-convergence theory. The concept of this convergence looks fairly simple.

Definition 1. Let F_ε, $\varepsilon > 0$, be a family of functionals all defined on an admissible set $\mathscr{A}$ that takes values in $[-\infty, \infty]$. Here $\mathscr{A}$ is a complete metric space with the distance function d. Let F be another functional also from $\mathscr{A}$ to $[-\infty, \infty]$. We say that F_ε Γ-converges to F if the following two statements hold:

1. For every family $\{u_\varepsilon\} \subset \mathscr{A}$ with $\lim_{\varepsilon\to 0} d(u_\varepsilon, u) = 0$, $\liminf_{\varepsilon\to 0} F_\varepsilon(u_\varepsilon) \geq F(u)$.
2. For every $u \in \mathscr{A}$, there exists a family $\{u_\varepsilon\} \subset \mathscr{A}$ such that $\lim_{\varepsilon\to 0} d(u_\varepsilon, u) = 0$ and $\limsup_{\varepsilon\to 0} F_\varepsilon(u_\varepsilon) \leq F(u)$.

Note that we allow F_ε to be $\pm\infty$. An obvious property of Γ-convergence is that it is robust under continuous perturbation.

Proposition 2. *Suppose F_ε: $\mathscr{A} \to [-\infty, \infty]$ Γ-converges to F. If G: $\mathscr{A} \to (-\infty, \infty)$ is a continuous functional with respect to the metric d on $\mathscr{A}$, then $F_\varepsilon + G$ Γ-converges to $F + G$.*

To apply this theory to the Ohta–Kawasaki functional I we need to choose the parameter σ properly. The correct choice is that σ has the same order as ε. We assume that there exists a fixed γ independent of ε such that

$$\sigma = \varepsilon\gamma. \tag{8}$$

Rewrite the functional I as

$$I_\varepsilon(u) = \int_D [\frac{\varepsilon^2}{2}|\nabla u|^2 + W(u) + \frac{\varepsilon\gamma}{2}|(-\Delta)^{-1/2}(u-a)|^2]\,dx. \tag{9}$$

Here we emphasize that in the notation I_ε, ε is the singular perturbation parameter. In the Γ-convergence theory, ε tends to 0. The problem has two other parameters: a and γ which are held fixed in the Γ-convergence process. We want to be a bit more flexible about the admissible set of I_ε. We define I_ε on

$$\mathscr{A}_a = \{u \in L^2(D) :\ \overline{u} = a\}. \tag{10}$$

If u happens to be in $W^{1,2}(D)$, then $I_\varepsilon(u)$ is given by (9); otherwise we set $I_\varepsilon(u) = \infty$. It turns out that in this setting $\varepsilon^{-1} I_\varepsilon$ has a Γ-limit.

Let us define this limit functional now and prove that it is indeed the Γ-limit later. The functional is denoted by J. We first define it for subsets E of D whose Lebesgue measure is $a|D|$, i.e.,

$$|E| = a|D|, \tag{11}$$

and whose characteristic function

$$\chi_E(x) = 1 \text{ if } x \in E, \quad \chi_E(x) = 0 \text{ if } x \notin E \tag{12}$$

is in $BV(D)$. Here $BV(D)$ is the space of all functions of bounded variation on D. For such an E we set

$$J(E) = \tau \|D\chi_E\|(D) + \frac{\gamma}{2} \int_D |(-\Delta)^{-1/2}(\chi_E - a)|^2 \, dx. \tag{13}$$

The constant τ in (13) depends on the double well potential W:

$$\tau = \int_0^1 \sqrt{2W(t)} \, dt. \tag{14}$$

This constant is called the surface tension. See the appendix for more thorough discussion on this number.

The nonlocal term in (13) is similar to that in (1). The only difference is that here χ_E is the characteristic function of a set while in (1) there is a more general function u. The first term $\|D\chi_E\|(D)$ needs some explanation. Since $\chi_E \in BV(D)$, we view $D\chi_E$ as a vector valued, signed measure, and let $\|D\chi_E\|$ be the positive total variation measure of $D\chi_E$. The first term in (13), $\|D\chi_E\|(D)$, is the $\|D\chi_E\|$ measure of the entire domain D. When $\partial_D E$ is a smooth surface or a union of smooth surfaces, $\|D\chi_E\|(D)$ is just the area of $\partial_D E$. For this reason $\|D\chi_E\|(D)$ is called the perimeter of E in D and is sometimes denoted by $P_D(E)$.

Now as in the definition of I_ε we extend the admissible set of J to the same $\mathscr{A}_a$. Namely, if an element u in $\mathscr{A}_a$ is the characteristic function χ_E of a set E and χ_E is in $BV(D)$, then $J(u)$ is given by (13); otherwise we set $J(u) = \infty$.

On $\mathscr{A}_a$ we use the L^2-norm $\|\cdot\|_2$ to define the metric, i.e., for $u_1, u_2 \in \mathscr{A}_a$,

$$d(u_1, u_2) = \|u_1 - u_2\|_2. \tag{15}$$

A few more words on $\|D\chi_E\|(D)$. This quantity can be alternatively given by the formula

$$\|D\chi_E\|(D) = \sup\{\int_D \chi_E(x) \operatorname{div} \sigma(x) \, dx : \ \sigma \in C_0^\infty(D; R^n), \ |\sigma| \le 1\}. \tag{16}$$

There is a structure theorem which says that one can define a measure theoretic boundary of E, denoted by $\partial^* E$ called the reduced boundary, such that the measure $\|D\chi_E\|$ is exactly the $n-1$ dimensional Hausdorff measure H^{n-1} restricted to $\partial^* E$.

The reduced boundary $\partial^* E$ coincides with the topological boundary $\partial_D E$ when E is smooth. In general $\partial^* E \subset \partial_D E$. Note that being measure theoretic, one may add to or delete from E any set of Lebesgue measure 0 without changing $\partial^* E$. See [6, Sect. 5.7] for more discussion on this subject.

The functional J has an elegant Euler–Lagrange equation. If E is a critical point of J and $\partial_D E$, the boundary of E in D, is smooth, then on every point of $\partial_D E$

$$H(\partial_D E) + \gamma(-\Delta)(\chi_E - a) = \lambda. \tag{17}$$

Here H is the curvature of $\partial_D E$ if $E \subset D \subset R^2$, and H is the mean curvature of $\partial_D E$ if $E \subset D \subset R^3$. If $E \subset D \subset R^1$, then $H(\partial_D E) = 0$. The constant λ is again a Lagrange multiplier corresponding to the condition that $|E| = a|D|$. Equation (17) is a free boundary problem. It states that the sum of the curvature of $\partial_D E$ and $\gamma(-\Delta)^{-1}(\chi_E - a)$ at every $x \in \partial_D E$ is constant.

The main result in this section is the following lemma:

Lemma 3. *$\varepsilon^{-1} I_\varepsilon$ Γ-converges to J.*

Proof. Let us write

$$\varepsilon^{-1} I_\varepsilon = L_\varepsilon + N \tag{18}$$

where

$$L_\varepsilon(u) = \int_D [\frac{\varepsilon}{2}|\nabla u|^2 + \frac{1}{\varepsilon} W(u)]\, dx \tag{19}$$

if $u \in W^{1,2}(D)$ and $L_\varepsilon(u) = \infty$ if $u \in \mathscr{A}_a \backslash W^{1,2}(D)$, and

$$N(u) = \frac{\gamma}{2} \int_D |(-\Delta)^{-1/2}(u - a)|^2\, dx, \quad u \in \mathscr{A}_a. \tag{20}$$

Here L_ε is a local functional and N a nonlocal functional. The functional N is continuous on $\mathscr{A}_a$ with respect to its metric. Therefore by Proposition 2 it suffices to prove that L_ε Γ-converges to L where L is

$$L(E) = \tau \|D\chi_E\|(D), \tag{21}$$

if $u = \chi_E$ for some Lebesgue measurable set E and $\chi_E \in BV(D)$; otherwise $L(u) = \infty$. With these notations, $J = L + N$.

The Γ-convergence of L_ε to L was studied by Modica [9]. We only prove the first property in Definition 1. The proof of the second property is more involved, and we refer to [9, Lemma 2, Proposition 2]. Suppose $u_\varepsilon \to u$ in $L^2(D)$. Let

$$\phi(t) = \int_0^t W^{1/2}(s)\, ds. \tag{22}$$

We first claim that $\phi(u_\varepsilon)$ converges to $\phi(u)$ in $L^1(D)$. For this we need Vitali's convergence theorem [7, p. 203].

Vitali's Convergence Theorem. *Let $\{f_n\}$ be a sequence in $L^p(D,\mu)$, $1 \le p < \infty$, and f be an μ-measurable function such that $f_n \to f$ μ-a.e. Then $f \in L^p(D,\mu)$ and $\|f_n - f\|_p \to 0$ if and only if:*

1. *For each $\varepsilon > 0$, there is a μ-measurable set $A_\varepsilon \subset D$ such that $\mu(A_\varepsilon) < \infty$ and $\int_{D\setminus A_\varepsilon} |f_n|^p \, d\mu < \varepsilon$ for all n.*
2. *For each $\varepsilon > 0$ there is $\delta > 0$ such that for every μ-measurable set S, $\mu(S) < \delta$ implies $\int_S |f_n|^p \, d\mu < \varepsilon$ for all n.*

Part 1 of Vitali's convergence theorem is not needed here because D itself has finite Lebesgue measure. Let $\{u_{\varepsilon_n}\} = \{u_n\}$ be any sequence in $\{u_\varepsilon\}$, $\varepsilon_n \to 0$ as $n \to \infty$, $u_n \to u$ in L^2, and $u_n \to u$ a.e. By Vitali's convergence theorem, for every $\varepsilon > 0$ there exists $\delta > 0$ such that for every S, $|S| < \delta$, one has $\int_S u_n^2 \, dx < \varepsilon$ for all n. Then (4) implies that

$$|\phi(t)| \le C + Ct^2, \tag{23}$$

and consequently

$$\int_S |\phi(u_n)| \, dx \le C\delta + C\varepsilon.$$

The Vitali's theorem applied to $\phi(u_n)$ implies that $\phi(u_n)$ converges to $\phi(u)$ in $L^1(D)$.

If $\liminf_{n\to\infty} L_{\varepsilon_n}(u_n) = \infty$, then the first property of Definition 1 holds trivially. Thus we assume that $L_{\varepsilon_n}(u_n)$ is bounded in n. The Fatou's lemma now implies that

$$0 \le \int_D W(u) \le \liminf_{n\to\infty} \int_D W(u_n) \, dx \le \liminf_{n\to\infty} \varepsilon_n L_{\varepsilon_n}(u_n) = 0.$$

Then for a.e. $x \in D$, $u(x) = 0$ or 1. We can write $u = \chi_E$ where $E = \{x \in D : u(x) = 1\}$ is measurable.

Simple estimation shows that

$$\begin{aligned} L_{\varepsilon_n}(u_n) &= \int_D [\frac{\varepsilon_n}{2} |\nabla u_n|^2 + \frac{1}{\varepsilon_n} W(u_n)] \, dx \\ &\ge \sqrt{2} \int_D \sqrt{W(u_n)} |\nabla u_n| \, dx \\ &= \sqrt{2} \int_D |\nabla \phi(u_n)| \, dx \end{aligned}$$

The lower semi-continuity of the BV norm [6, Theorem 1, page 172] asserts that, since $\phi(u_n) \to \phi(u)$ in L^1,

$$\liminf_{n\to\infty} \int_D |\nabla \phi(u_n)| \, dx \ge \|D\phi(u)\|(D).$$

Hence

$$\liminf_{n\to\infty} L_{\varepsilon_n}(u_n) \ge \sqrt{2} \|D\phi(u)\|(D).$$

Finally we note that since $u = \chi_E$, $\phi(u) = \phi(1)\chi_E$. Then

$$\sqrt{2}\|D\phi(u)\|(D) = \sqrt{2}\phi(1)\|D\chi_E\|(D) = \tau\|D\chi_E\|(D).$$

Therefore

$$\liminf_{n\to\infty} L_{\varepsilon_n}(u_n) \geq L(E).$$

This, together with [9, Lemma 2, Proposition 2], proves the lemma. □

Global and Local Minimizers

Let us see how the theory of Γ-convergence helps us understand the minimizers, local and global, of F_ε and F. The following proposition is easy to prove.

Proposition 1. *Suppose that F_ε Γ-converges to F and F_ε has a global minimizer u_ε. If u_ε converges, possibly along a subsequence $\{u_{\varepsilon_n}\}$, to u_0 in $\mathscr{A}$, then u_0 is a global minimizer of F and $\lim_{n\to\infty} F_{\varepsilon_n}(u_{\varepsilon_n}) = F(u_0)$.*

Proof. The first property in Definition 1 says that

$$\liminf_{n\to\infty} F_{\varepsilon_n}(u_{\varepsilon_n}) \geq F(u_0). \tag{24}$$

Let $u \in \mathscr{A}$ be an arbitrary element. The second property in Definition 1 says that there exists $v_{\varepsilon_n} \in \mathscr{A}$ such that $v_{\varepsilon_n} \to u$ in $\mathscr{A}$ and

$$\limsup_{n\to\infty} F_{\varepsilon_n}(v_{\varepsilon_n}) \leq F(u). \tag{25}$$

Since u_{ε_n} minimizes F_{ε_n}, $F_{\varepsilon_n}(u_{\varepsilon_n}) \leq F_{\varepsilon_n}(v_{\varepsilon_n})$. Then (24) and (25) imply that

$$F(u_0) \leq F(u), \tag{26}$$

i.e., u_0 is a global minimizer of F.

If we take u to be u_0 in (25), then

$$\limsup_{n\to\infty} F_{\varepsilon_n}(v_{\varepsilon_n}) \leq F(u_0). \tag{27}$$

Now $F_{\varepsilon_n}(u_{\varepsilon_n}) \leq F_{\varepsilon_n}(v_{\varepsilon_n})$ and (24) imply that

$$F(u_0) \leq \liminf_{n\to\infty} F_{\varepsilon_n}(u_{\varepsilon_n}) \leq \limsup_{n\to\infty} F_{\varepsilon_n}(u_{\varepsilon_n}) \leq F(u_0).$$

Therefore $\lim_{n\to\infty} F_{\varepsilon_n}(u_{\varepsilon_n}) = F(u_0)$. □

In the last proposition we need an extra assumption that u_{ε_n} converges to u_0 in $\mathscr{A}$. In general the two properties in the Definition 1 are often insufficient when one studies the relationships between the minimizers of F_ε and those of F. We now add another property, which is a kind of uniform coercivity, to the Γ-convergence theory.

Definition 2. Let F_ε be a family of functionals from a complete metric space $\mathscr{A}$ to $[-\infty,\infty]$. Then F_ε is said to be uniformly coercive if for every sequence $\{u_{\varepsilon_n}\}\subset\mathscr{A}$, $\varepsilon_n\to 0$ as $n\to\infty$, with $F_{\varepsilon_n}(u_{\varepsilon_n})$ bounded with respect to n, there is a subsequence of $\{u_{\varepsilon_n}\}$ which converges in $\mathscr{A}$.

Lemma 3. *The family of the functionals $\varepsilon^{-1}I_\varepsilon$ is uniformly coercive, i.e., every sequence $\{u_{\varepsilon_n}\}$ with the property that $\varepsilon_n^{-1}I_{\varepsilon_n}(u_{\varepsilon_n})$ is bounded has a convergent subsequence $\{u_{\varepsilon_{n_l}}\}$ whose limit u is χ_E for some measurable set E and $\chi_E\in BV(D)$.*

Proof. Let us denote u_{ε_n} by u_n for simplicity. Set

$$\phi(t)=\int_0^t W^{1/2}(s)ds. \tag{28}$$

Then (4) implies

$$|\phi(t)|\le C+C|t|^2.$$

Set $w_n=\phi(u_n)$. We claim that w_n is bounded in $W^{1,1}(D)$. For by (4), we find

$$|w_n|\le C+C|u_n|^2\le C+CW(u_n).$$

Therefore since $\varepsilon_n^{-1}I_{\varepsilon_n}(u_n)$ is bounded, $\{w_n\}$ is bounded in $L^1(D)$. On the other hand,

$$\begin{aligned}\int_D|\nabla w_n|\,dx&=\int_D W^{1/2}(u_n)|\nabla u_n|\,dx\\&\le\frac{\sqrt{2}}{2}\Big(\int_D[\frac{\varepsilon_n}{2}|\nabla u_n|^2+\frac{1}{\varepsilon_n}W(u_n)]\,dx\Big)\\&\le\frac{\sqrt{2}}{2}L_{\varepsilon_n}(u_n).\end{aligned}$$

So $\{w_n\}$ is bounded in $W^{1,1}(D)$. The Sobolev embedding theorem asserts that $\{w_n\}$ is relatively compact in $L^1(D)$.

Now consider $u_n=\phi^{-1}(w_n)$. Equations (4) and (28) imply that

$$|\phi^{-1}(t)|\le C+C|t|^{1/2},\ |\phi^{-1}(t)|^2\le C+C|t|. \tag{29}$$

To prove that $\{u_n\}$ is relatively compact we show that every subsequence of $\{u_n\}$ has a L^2-convergent further subsequence. Let $\{u_{n_l}\}$ be a subsequence of $\{u_n\}$. Then there is a subsequence of $\{w_{n_l}=\phi(u_{n_l})\}$, denoted by $\{w_{n_{l_m}}\}$, and a function $w\in L^1(D)$ such that $w_{n_{l_m}}\to w$ in L^1 and $w_{n_{l_m}}\to w$ a.e. Then $u_{n_{l_m}}\to\phi^{-1}(w)$ a.e. Applying Vitali's convergence theorem to $w_{n_{l_m}}$, we find that for every $\varepsilon>0$ there is $\delta>0$ such that for every measurable set S, $|S|<\delta$ implies $\int_S|w_{n_{l_m}}|\,dx<\varepsilon$ for all m. Then (29) implies

$$\int_S|u_{n_{l_m}}|^2\,dx\le\int_S(C+C|w_{n_{l_m}}|)\,dx<C\delta+C\varepsilon.$$

Now Vitali's convergence theorem applied to $\{u_{n_{l_m}}\}$ asserts that $u_{n_{l_m}} \to \phi^{-1}(w)$ in $L^2(D)$.

Let u be the limit of a subsequence $\{u_{n_l}\}$ of $\{u_n\}$. We now show that $u = \chi_E$ for some measurable set E and $\chi_E \in BV(D)$. By passing to a further subsequence if necessary we can assume that u_{n_l} converges to u a.e. The Fatou's lemma and the boundedness of $\varepsilon_n^{-1} I_{\varepsilon_n}(u_n)$ imply that

$$0 \le \int_D W(u) \le \liminf_{l\to\infty} \int_D W(u_{n_l})\,dx \le \liminf_{l\to\infty} I_{\varepsilon_{n_l}}(u_{n_l}) = 0.$$

Then for a.e. $x \in D$, $u(x) = 0$ or 1. We can write $u = \chi_E$ where $E = \{x \in D : u(x) = 1\}$. If we consider $w_{n_l} = \phi(u_{n_l})$, then the boundedness of $\{w_{n_l}\}$ in $W^{1,1}(D)$, proved earlier, implies that $\phi(u)$, the L^1-limit of $\{w_{n_l}\}$, is a BV function. Again we have used the lower semi-continuity of the BV norm. Since $\phi(u) = \phi(1)u$, $u = \chi_E$ is also in $BV(D)$. □

The next important proposition is proved by Kohn and Sternberg [8]. We denote an open ball in $\mathscr{A}$ centered at u_0 of radius δ by $B_\delta(u_0)$, i.e.,

$$B_\delta(u_0) = \{u \in \mathscr{A} :\ d(u, u_0) < \delta\}. \tag{30}$$

Proposition 4. *Suppose that F_ε Γ-converges to F and that F_ε is uniformly coercive. Assume that in every close ball $\overline{B} \subset \mathscr{A}$, F_ε has a minimizer. Let $\delta > 0$ and $u_0 \in \mathscr{A}$ be such that $F(u_0) < F(u)$ for all $u \in B_\delta(u_0)$ with $u \neq u_0$. Then there exists $\varepsilon_0 > 0$ such that for all $\varepsilon < \varepsilon_0$ there exists $u_\varepsilon \in B_{\delta/2}(u_0)$ with $F_\varepsilon(u_\varepsilon) \le F_\varepsilon(u)$ for all $u \in B_{\delta/2}(u_0)$. In addition $\lim_{\varepsilon\to 0} d(u_\varepsilon, u_0) = 0$.*

Proof. Let $u_\varepsilon \in \overline{B}_{\delta/2}(u_0)$ be a minimizer of F_ε in $\overline{B}_{\delta/2}(u_0)$. We claim $u_\varepsilon \in B_{\delta/2}(u_0)$ if ε is small enough, i.e., u_ε is not on the boundary of $B_{\delta/2}(u_0)$. Otherwise there exists a sequence $\varepsilon_l \to 0$, such that $d(u_{\varepsilon_l}, u_0) = \delta/2$ and

$$F_{\varepsilon_l}(u_{\varepsilon_l}) = \min_{u \in \overline{B}_{\delta/2}(u_0)} F_{\varepsilon_l}(u).$$

Property 2 of Definition 1 asserts that there exists a sequence v_{ε_l} in $B_{\delta/2}(u_0)$, if l is large enough, such that

$$\limsup_{l\to\infty} F_{\varepsilon_l}(v_{\varepsilon_l}) \le F(u_0).$$

Therefore,

$$\limsup_{l\to\infty} F_{\varepsilon_l}(u_{\varepsilon_l}) \le \limsup_{l\to\infty} F_{\varepsilon_l}(v_{\varepsilon_l}) \le F(u_0). \tag{31}$$

Definition 2 then asserts that, after passing to a subsequence, again denoted by u_{ε_l}, there exists $\overline{u}_0$ such that $u_{\varepsilon_l} \to \overline{u}_0$ in $\mathscr{A}$ and $d(\overline{u}_0, u_0) = \delta/2$. Part 1 of Definition 1 now implies

$$F(\overline{u}_0) \le \liminf_{l\to\infty} F_{\varepsilon_l}(u_{\varepsilon_l}).$$

By combining this with (31) we find that

$$F(\overline{u}_0) \le \liminf_{l\to\infty} F_{\varepsilon_l}(u_{\varepsilon_l}) \le \limsup_{l\to\infty} F_{\varepsilon_l}(u_{\varepsilon_l}) \le F(u_0).$$

This contradicts the condition that $F(u_0) < F(u)$ for all $u \in B_\delta(u_0)$ with $u \neq u_0$. Therefore u_ε is in the open ball $B_{\delta/2}(u_0)$, i.e., u_ε is a local minimizer of F_ε.

To show $u_\varepsilon \to u_0$ as $\varepsilon \to 0$, we assume that there exists a sequence $\varepsilon_l \to 0$ such that $d(u_{\varepsilon_l}, u_0) \ge \delta_0$ and $\delta_0 < \delta/2$. Then arguing like above, we have $\widetilde{u}_0$ such that, after passing to a subsequence, again denoted by u_{ε_l}, $u_{\varepsilon_l} \to \widetilde{u}_0$ and $d(\widetilde{u}_0, u_0) \ge \delta_0$. By Part 1 of Definition 1 and (31),

$$F(\widetilde{u}_0) \le \liminf_{l\to\infty} F_{\varepsilon_l}(u_{\varepsilon_l}) \le F(u_0),$$

which is again a contradiction. □

In this proposition we have assumed that F_ε has a minimizer in each closed ball $\overline{B}$. This property is satisfied by our Ohta–Kawasaki functional I_ε. The proof of this is standard minimization argument.

The Lamellar Phase of Diblock Copolymers

The lamellar phase of a diblock copolymer is plotted in Fig. 2.1. To study this pattern, we look at a line perpendicular to the parallel layers. Therefore we consider the one-dimension case $D = (0,1)$.

A characteristic function χ_E in $BV(0,1)$, up to a set of Lebesgue measure 0, is a step function. χ_E switches between 0 and 1 at finitely many points $x_1, x_2, \ldots, x_K$, with $0 < x_1 < x_2 < \ldots < x_K < 0$. The set $\{x_1, x_2, \ldots, x_K\}$ is the reduced boundary of E denoted by $\partial^* E$ [6, Sect. 6.7, pages 194–207].

If $\chi_E \in \mathscr{A}_a \cap BV(0,1)$, $\|D\chi_E\|(0,1)$ has to be nonzero. Otherwise χ_E would be a constant. Then $\chi_E(x) = 0$ for a.e. $x \in (0,1)$ or $\chi_E(x) = 1$ for a.e. $x \in (0,1)$. In either case $\int_0^1 \chi_E \neq a \in (0,1)$. So we have the following mutually disjoint decomposition:

$$\begin{aligned} &\mathscr{A}_a \cap BV((0,1),\{0,1\}) = \cup_1^\infty A_K, \text{ where} \\ &A_K = \{\chi_E \in \mathscr{A}_a \cap BV((0,1),\{0,1\}) : \|D\chi_E\|(0,1) = K\}. \end{aligned} \tag{32}$$

Here $BV((0,1),\{0,1\})$ is the set of the BV functions that only take values in $\{0,1\}$.

Two characteristic functions are particularly important in A_K. Let

$$z_1 = \frac{1-a}{K},\ z_3 = z_1 + \frac{2}{K},\ z_5 = z_3 + \frac{2}{K}, \ldots \tag{33}$$

and

$$z_2 = \frac{1+a}{K},\ z_4 = z_2 + \frac{2}{K},\ z_6 = z_4 + \frac{2}{K}, \ldots. \tag{34}$$

Define

$$U_{K,0}(x) = \begin{cases} 0 \text{ if } x \in (0,z_1) \cup (z_2,z_3) \cup (z_4,z_5) \cup \ldots, \\ 1 \text{ if } x \in (z_1,z_2) \cup (z_3,z_4) \cup (z_5,z_6) \cup \ldots \end{cases}. \tag{35}$$

Let

$$z_1 = \frac{a}{K},\ z_3 = z_1 + \frac{2}{K},\ z_5 = z_3 + \frac{2}{K}, \dots \tag{36}$$

and

$$z_2 = \frac{2-a}{K},\ z_4 = z_2 + \frac{2}{K},\ z_6 = z_4 + \frac{2}{K}, \dots . \tag{37}$$

Define

$$U_{K,1}(x) = \begin{cases} 1 \text{ if } x \in (0,z_1) \cup (z_2,z_3) \cup (z_4,z_5) \cup \dots, \\ 0 \text{ if } x \in (z_1,z_2) \cup (z_3,z_4) \cup (z_5,z_6) \cup \dots \end{cases}. \tag{38}$$

Lemma 1. *For every $\chi_E \in A_K$, $\chi_E \neq U_{K,0}$, and $\chi_E \neq U_{K,1}$, we have $J(U_{K,0}) = J(U_{K,1}) < J(E)$.*

Proof. For each $\chi_E \in A_K$, let us denote $\partial^* E$ by $\{x_1, x_2, \dots, x_K\}$, where $0 < x_1 < x_2 \dots < x_K < 1$. Since $\|Du\|(x_i, x_{i+1}) = 0$ for each i and (x_i, x_{i+1}) is connected, $\chi_E = 0$ for a.e. $x \in (x_i, x_{i+1})$ or $\chi_E = 1$ for a.e. $x \in (x_i, x_{i+1})$. And it follows from the definition of reduced boundaries [6, p. 194] that $\chi_E(x)$ must jump from 0 to 1 or 1 to 0 when x moves from (x_{i-1}, x_i) to (x_i, x_{i+1}). We can further decompose A_K into two disjoint sets:

$$\begin{aligned} A_{K,0} &= \{\chi_E \in A_K : u = 0, \text{ for a.e. } x \in (0,x_1)\}, \\ A_{K,1} &= \{\chi_E \in A_K : u = 1, \text{ for a.e. } x \in (0,x_1)\}. \end{aligned} \tag{39}$$

For $\chi_E \in A_{K,0}$ the constraint $\int_0^1 \chi_E = a$ becomes $-x_1 + x_2 - x_3 + x_4 \dots = a$, and for $\chi_E \in A_{K,1}$ the constraint $\int_0^1 \chi_E = a$ becomes $x_1 - x_2 + x_3 - x_4 \dots = a$.

Now $A_{K,0}$ can be identified with the set

$$A_{K,0} = \{(x_1, \dots, x_K) \in R^K : 0 < x_1 < \dots x_K < 1,\ -x_1 + x_2 - x_3 + x_4 \dots = a\}, \tag{40}$$

and $A_{K,1}$ can be identified with the set

$$A_{K,1} = \{(x_1, \dots, x_K) \in R^K : 0 < x_1 < \dots x_K < 1,\ x_1 - x_2 + x_3 - x_4 \dots = a\}. \tag{41}$$

In $A_{K,0}$ and $A_{K,1}$, the functional J becomes a function of $(x_1, x_2, \dots, x_K)$. We want to find all critical points of J in $A_{K,0}$ and $A_{K,1}$. Consider the case $\chi_E \in A_{K,0}$. The case $\chi_E \in A_{K,1}$ is similar. First consider $N(E)$, the nonlocal part of $J(E)$. Let v be the solution of

$$-v'' = \chi_E - a,\ v'(0) = v'(1) = 0,\ \int_0^1 v = 0.$$

Sometimes we write $v(x) = v(x; x_1, x_2, \dots, x_K)$ since v depends on $x_1, x_2, \dots, x_K$. Denote the Green's function of this equation by $G(x,y)$. Then

$$\begin{aligned} N(E) &= \frac{\gamma}{2} \int_0^1 [(-\Delta)^{-1/2}(\chi_E - a)]^2 dx \\ &= \frac{\gamma}{2} \int_0^1 (\chi_E - a) v\, dx \end{aligned}$$

$$= \frac{\gamma}{2}\int_0^1 \chi_E v\,dx$$
$$= \frac{\gamma}{2}[\int_{x_1}^{x_2} v\,dx + \int_{x_3}^{x_4} v\,dx + \ldots].$$

Treating N as a function of $(x_1, x_2, \ldots, x_K)$ in $A_{K,0}$, we calculate

$$\frac{\partial N(E)}{\partial x_1} = \frac{\gamma}{2}[-v(x_1; x_1, \ldots, x_K) + \int_0^1 \chi_E \frac{\partial v}{\partial x_1}\,dx].$$

Since

$$\begin{aligned}\frac{\partial v(x; x_1, \ldots, x_K)}{\partial x_1} &= \frac{\partial}{\partial x_1}\int_0^1 (\chi_E - a)G(x,y)\,dy \\ &= \frac{\partial}{\partial x_1}\int_E G(x,y)\,dy \\ &= \frac{\partial}{\partial x_1}[\int_{x_1}^{x_2} G(x,y)\,dy + \int_{x_3}^{x_4} G(x,y)\,dy + \ldots] \\ &= -G(x, x_1),\end{aligned}$$

we deduce

$$\begin{aligned}\frac{\partial N(E)}{\partial x_1} &= \frac{\gamma}{2}[-v(x_1; x_1, \ldots, x_K) - \int_0^1 \chi_E(x) G(x, x_1)\,dx] \\ &= -\gamma v(x_1; x_1, \ldots, x_K)\end{aligned}$$

The same argument applied to differentiations with respect to the other x_i's yields

$$\begin{aligned}\nabla N(x_1, \ldots, x_K) = \gamma(&-v(x_1; x_1, \ldots, x_K), v(x_2; x_1, \ldots, x_K), \ldots, \\ &(-1)^K v(x_K; x_1, \ldots, x_K)).\end{aligned} \tag{42}$$

Since $\int_0^1 \chi_E = a$, or $-x_1 + x_2 - \ldots = a$, the Lagrange multiplier method asserts that if $(x_1, x_2, \ldots, x_K)$ is a critical point of N in $A_{K,1}$, there exists λ such that

$$\nabla N(x_1, \ldots, x_K) = \lambda(-1, 1, -1, \ldots, (-1)^K).$$

Then (42) implies that

$$v(x_1; x_1, \ldots, x_K) = v(x_2; x_1, \ldots, x_K) = \ldots = v(x_K; x_1, \ldots, x_K). \tag{43}$$

On (x_1, x_2), v solves the linear equation $-v'' = 1 - a$. Then $v(x_1) = v(x_2)$ implies that v is symmetric about $(x_1 + x_2)/2$, and hence $v'(x_1) = -v'(x_2)$. On intervals $(0, x_1)$ and (x_2, x_3), v satisfies the linear equation $-v'' = -a$. Since v also satisfies the conditions $v(x_1) = v(x_2)$, $v'(x_1) = -v'(x_2)$, $v(x_2) = v(x_3)$, and $v'(0) = 0$, we conclude by solving the equation on $(0, x_1)$ and (x_2, x_3) that v on $(0, x_1)$ is a reflection of v on $(x_2, (x_2 + x_3)/2)$. Hence the length of $(0, x_1)$ is half of that of (x_2, x_3). Next we compare intervals (x_2, x_3) and (x_4, x_5) and similarly find that they have the same length. By repeating this argument we conclude that the intervals where $\chi_E = 0$ all

have the same length with the exception of $(0,x_1)$ and possibly $(x_K,1)$ if $\chi_E=0$ there, whose length is half. The same can be said for the intervals where $\chi_E=1$. Taking $-x_1+x_2-x_3+x_4\ldots=a$ into consideration, we find that

$$x_1=\frac{1-a}{K},\ x_3=x_1+\frac{2}{K},\ x_5=x_3+\frac{2}{K},\ldots,\quad x_2=\frac{1+a}{K},\ x_4=x_2+\frac{2}{K},\ x_6=x_4+\frac{2}{K},\ldots.$$

This means that $\chi_E=U_{K,0}$.

We have proved that N has a unique critical point $U_{K,0}$ in $A_{K,0}$. Similarly $U_{K,1}$ is the only critical point of N in $A_{K,1}$. We proceed to prove that $U_{K,0}$ minimizes N in $A_{K,0}$. We first compute $N(U_{K,0})$. Let v be the solution of

$$-v''=U_{K,0}-a,\ v'(0)=v'(1)=0,\ \int_0^1 v=0.$$

Then

$$N(U_{K,0})=\frac{\gamma}{2}\int_0^1(U_{K,0}-a)v=\frac{\gamma}{2}\int_0^1|v'|^2$$

On $(0,x_1)$ $v'(x)=ax+v'(0)=ax$. Then

$$\frac{\gamma}{2}\int_0^{x_1}|v'|^2=\frac{\gamma a^2x_1^3}{6}=\frac{\gamma a^2(1-a)^3}{6K^3}.$$

On $(x_1,(x_1+x_2)/2)$, $v'(x)=-(1-a)(x-(x_1+x_2)/2)+v'((x_2+x_2)/2)=-(1-a)(x-(x_1+x_2)/2)$. Then

$$\frac{\gamma}{2}\int_{x_1}^{(x_1+x_2)/2}|v'|^2=\frac{\gamma(1-a)^2}{6}(\frac{x_2-x_1}{2})^3=\frac{\gamma(1-a)^2a^3}{6K^3}.$$

On the whole interval $(0,1)$, we deduce

$$\frac{\gamma}{2}\int_0^1|v'|^2=K\frac{\gamma}{2}\int_0^{(x_1+x_2)/2}|v'|^2=K[\frac{\gamma a^2(1-a)^3}{6K^3}+\frac{\gamma(1-a)^2a^3}{6K^3}]=\frac{\gamma a^2(1-a)^2}{6K^2}.$$

Hence

$$N(U_{K,0})=\frac{\gamma a^2(1-a)^2}{6K^2}. \tag{44}$$

Similar argument shows that

$$N(U_{K,1})=\frac{\gamma a^2(1-a)^2}{6K^2}. \tag{45}$$

We now show that $N(\chi_E)>N(U_{K,0})$ for every $\chi_E\in A_{K,0}$, $\chi_E\neq U_{K,0}$. If this is not the case, since there is only one critical point, $U_{K,0}$, in $A_{K,0}$, there must be a sequence $\{(x_{n,1},x_{n,2},\ldots,x_{n,K})\}$ converging to a point $(y_1,y_2,\ldots,y_K)$ on the boundary of $A_{K,0}$ such that

$$\lim_{n\to\infty}N(x_{n,1},x_{n,2},\ldots,x_{n,K})\le N(U_{K,0}).$$

For the point $(y_1, y_2, \ldots, y_K)$ to be on the boundary of $A_{K,0}$, at least two of $0, y_1, \ldots, y_K, 1$ must be identical. Then $(y_1, y_2, \ldots, y_K)$ is identified as a point in $A_{K',0}$ or $A_{K',1}$ for some $K' < K$. Let us denote this point by $(z_1, z_2, \ldots, z_{K'})$ and assume, without the loss of generality, $(z_1, z_2, \ldots, z_{K'}) \in A_{K',0}$. We ask whether $U_{K',0}$ is the strict minimum of N in $A_{K',0}$. If so,

$$N(U_{K',0}) \le N(z_1, z_2, \ldots, z_{K'}) = \lim_{n\to\infty} N(x_{n,1}, x_{n,2}, \ldots, x_{n,K}) \le N(U_{K,0}),$$

which, since $K' < K$, is inconsistent with (44), where $N(U_{K,0}) = N(U_{K,1})$ decreases in K. If $U_{K',0}$ is not the strict minimum of N in $A_{K',0}$, we use the same argument on $U_{K',0}$ and end up in a $A_{K'',0}$ or $A_{K'',1}$ with $K'' < K'$. This process stops at $K = 1$, and there since $A_{1,0}$ has only one element $U_{1,0}$ and $A_{1,1}$ has only one element $U_{1,1}$, each is trivially regarded as the strict minimum in its class. Thus we find $N(U_{1,0}) = N(U_{1,1}) \le N(U_{K,0})$, inconsistent with (44) or (45).

So we have proved that $U_{K,0}$ is the strict minimum of N in $A_{K,0}$. And since for $\chi_E \in A_{K,0}$, $J(E) = \tau K + N(E)$, Lemma 1 is proved. □

We now show that the $U_{K,0}$'s and the $U_{K,1}$'s are strict local minimizers of J in $\mathscr{A}_a$ under the L^2-norm.

Lemma 2. *Given any positive integer K, one can find $\delta > 0$ such that for all $u \in B_\delta(U_{K,0})$ with $u \neq U_{K,0}$, $J(U_{K,0}) < J(u)$ and for all $u \in B_\delta(U_{K,1})$ with $u \neq U_{K,1}$, $J(U_{K,1}) < J(u)$.*

Proof. Let us only consider $U_{K,0}$. The study of $U_{K,1}$ is the same. Take δ to be a positive number to be specified later. Let $u \in B_\delta(U_{K,0})$ and $u \neq U_{K,0}$.

If $u \in A_{K,0}$, then Lemma 1 implies Lemma 2.

If $u \in A_{K,1}$, then we choose δ small enough so that $U_{K,1} \notin B_\delta(U_{K,0})$. Then $u \neq U_{K,1}$ and Lemma 1 again implies Lemma 2.

Now we consider $u \in \mathscr{A}_a \backslash A_K$. If $u \in (\mathscr{A}_a \backslash A_K) \backslash BV((0,1), \{-1, 1\})$, then $J(U_{K,0}) < J(u) = \infty$.

So we need only to consider $u \in (\mathscr{A}_a \backslash A_K) \cap BV((0,1), \{-1, 1\})$. In this case $u = \chi_E$ for some measurable set E, $\|D\chi_E\|(0,1) < \infty$ and $\|D\chi_E\|(0,1) \neq K$. There are two cases $\|D\chi_E\|(0,1)$ is either $\le K-1$ or $\ge K+1$. We study them separately.

First we prove that the case $\|D\chi_E\|(0,1) \le K-1$ does not happen if δ is small enough. We claim that there is $\delta > 0$ such that for all $\chi_E \in B_\delta(U_{K,0}) \cap BV((0,1), \{0,1\})$, $\|D\chi_E\|(0,1) \ge K$. Otherwise there exist $\delta_n \to 0$ and $\chi_{E_n} \in B_{\delta_n}(U_{K,0}) \cap BV((0,1), \{0,1\})$ such that $\|D\chi_{E_n}\|(0,1) \le K-1$. Then $\chi_{E_n} \to U_{K,0}$ in $L^2(0,1)$ implies, by the lower semi-continuity of the BV norm (see [6, Theorem 1, page 172]), that

$$K = \|DU_{K,0}\|(0,1) \le \liminf_{n\to\infty} \|D\chi_{E_n}\|(0,1) \le K-1,$$

a contradiction.

Second we consider the case $\|D\chi_E\|(0,1) \ge K+1$. Here

$$\begin{aligned} J(E) &\ge \tau(K+1) + N(E) \\ &= J(U_{K,0}) + \tau + N(E) - N(U_{K,0}). \end{aligned}$$

Because $N : \mathscr{A}_a \to (-\infty, \infty)$ is continuous with respect to the L^2-norm, by making δ small, we have

$$|N(E) - N(U_{K,0})| < \frac{\tau}{2}.$$

Then

$$J(E) \geq J(U_{K,0}) + \tau - \frac{\tau}{2} = J(U_{K,0}) + \frac{\tau}{2} > J(U_{K,0}).$$

This proves the lemma. □

Now we can apply Proposition 4 to $\varepsilon^{-1} I_\varepsilon$ and J to obtain the following theorem:

Theorem 3. *Let $D = (0,1)$. For each positive integer K there are $\delta > 0$ and $\varepsilon_0 > 0$ such that for all $\varepsilon < \varepsilon_0$, there exist a local minimizer $u_{\varepsilon,K,0}$ of I_ε in $B_{\delta/2}(U_{K,0})$ and a local minimizer $u_{\varepsilon,K,1} \in B_{\delta/2}(U_{K,1})$. As $\varepsilon \to 0$, $\|u_{\varepsilon,K,0} - U_{K,0}\|_2 \to 0$ and $\|u_{\varepsilon,K,1} - U_{K,1}\|_2 \to 0$.*

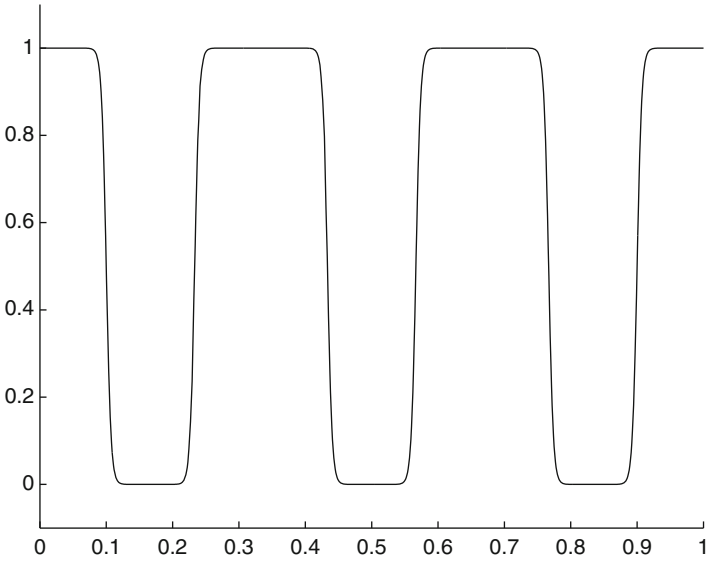

Fig. 2.2 A local minimizer of I_ε with 6 interfaces

Figure 2.2 depicts a $u_{\varepsilon,6,1}$. Now we turn our attention to the global minimizers of I_ε. It is easy to show, by the standard minimization argument, that for each $\varepsilon > 0$, there exists a global minimizer u_ε of I_ε. If we take any set E so that χ_E is in $\mathscr{A}_a \cap BV(D)$, then there exists $v_\varepsilon \in \mathscr{A}_a$ such that $\lim_{\varepsilon \to 0} \|v_\varepsilon - \chi_E\|_2 = 0$ and $\limsup_{\varepsilon \to 0} \varepsilon^{-1} I_\varepsilon(v_\varepsilon) \leq J(E)$, by Lemma 3. Consequently $\limsup_{\varepsilon \to 0} \varepsilon^{-1} I_\varepsilon(u_\varepsilon) \leq J(E) < \infty$. We deduce from Lemma 3 that along every sequence $\{u_{\varepsilon_n}\}$ of $\{u_\varepsilon\}$, there exist a subsequence $\{u_{\varepsilon_{n_l}}\}$ and $u_0 \in \mathscr{A}_a$ such that $u_{\varepsilon_{n_l}} \to u_0$ as $l \to \infty$. Proposition 1 implies that u_0 is a global minimizer of J.

Since $J(u_0) < \infty$, u_0 must be in one of the A_K's. By Lemma 1, we deduce that $u_0 = U_{K,0}$ or $u_0 = U_{K,1}$ for some K.

To identify this K, we consider

$$J(U_{K,0}) = J(U_{K,1}) = \tau K + \frac{\gamma a^2(1-a)^2}{6K^2} \tag{46}$$

as a function of K. Since u_0 is a global minimizer of J, $J(u_0) \leq J(U_{K,0}) = J(U_{K,1})$ for all positive integers K, i.e., u_0 must minimize (46). Although (46) is convex in K, the fact that K only takes positive integer values gives rise to two possibilities:

1. For most values of γ, a, and τ, the quantity (46) is minimized by a unique K, denoted by K_{opt}.
2. For some exceptional values of γ, a, and τ, the quantity (46) is minimized by two consecutive positive integers, denoted by K_{opt} and $K_{opt}+1$.

If the first case occurs, we find that $u_0 = U_{K_{opt},0}$ or $U_{K_{opt},1}$. If the second case occurs, we deduce that $u_0 = U_{K_{opt},0}$, $U_{K_{opt},1}$, $U_{K_{opt}+1,0}$, or $U_{K_{opt}+1,1}$. We have proved the following theorem:

Theorem 4. *Let $D = (0,1)$ and u_ε be a global minimum of I_ε. Along any sequence $\{u_{\varepsilon_n}\}$ of $\{u_\varepsilon\}$ ($\varepsilon_n \to 0$ as $n \to \infty$), there exists a subsequence $\{u_{\varepsilon_{n_l}}\}$ such that as $l \to \infty$, $u_{\varepsilon_{n_l}}$ converges in L^2 to a global minimizer u_0 of J.*

1. *For most values of γ, a, and τ, $u_0 = U_{K_{opt},0}$ or $U_{K_{opt},1}$.*
2. *For some exceptional values of γ, a, and τ, $u_0 = U_{K_{opt},0}$, $U_{K_{opt},1}$, $U_{K_{opt}+1,0}$, or $U_{K_{opt}+1,1}$.*

Discussion

Obviously most of the ideas presented here may be used to study radial solutions of (7). One can find solutions with ring patterns. See Ren and Wei [16, 23, 25] for more in this direction. Several other problems involving non-locality have also been successfully studied by the Γ-convergence approach outlined here [4, 14, 19–22].

Although the Γ-convergence is elegant and powerful, it does not answer all questions.

The description of the global minimizers in Theorem 4 is not entirely clear. If (46) is minimized by a unique K_{opt}, the first statement of the theorem does not say exactly to which of $U_{K_{opt},0}$ and $U_{K_{opt},1}$ u_ε converges. A much more involved analysis by Ren and Wei [17] shows that there are exactly two global minimizers of I_ε when ε is sufficiently small. One converges to $U_{K_{opt},0}$ and the other converges to $U_{K_{opt},1}$. A related result was proved earlier by Müller [11] for the functional I in the case $a = 1/2$ and $\sigma = 1$ in (1). Note that $\sigma = 1$ is a different parameter range from our assumption $\sigma = \varepsilon\gamma$ for I_ε. Müller actually studied a different looking problem which can be changed to (1) with $a = 1/2$. See also Chen and Oshita [2].

If K_{opt} is odd, one global minimizer is the reflection of the other global minimizer with respect to the $x = 1/2$ vertical line. If K_{opt} is even, one can obtain one

global minimizer by reflecting the other minimizer with respect to the $x = 0$ axis and shift the part of the graph on $(-1/2, 1/2)$ to $(0, 1)$. In the odd K_{opt} case, the simple reflection about $x = 1/2$ clearly turns a global minimizer to a global minimizer and $U_{K_{opt},0}$ to $U_{K_{opt},1}$ or vice versa. In the even K_{opt} case to show that the above mentioned operation turns a global minimizer to another global minimizer, one needs the following periodicity property for all the local minimizers of I_ε constructed in Theorem 3.

It was shown in [17] that the local minimizer $u_{\varepsilon,K,0}$ and $u_{\varepsilon,K,1}$ found in Theorem 3 must satisfy

$$u_{\varepsilon_n}(x+\frac{2m}{K}) = u_{\varepsilon_n}(x), \;\; x \in (0, \frac{1}{K}); \;\; u_{\varepsilon_n}(x+\frac{2m-1}{K}) = u_{\varepsilon_n}(\frac{1}{K}-x), \;\; x\in(0, \frac{1}{K}) \tag{47}$$

for $m = 1, 2, 3, \ldots$.

If (46) is minimized by two integers, K_{opt} and $K_{opt} + 1$, the story is more complex. It is shown in the same paper [17] that there are three possibilities:

1. Along some sequences $\{\varepsilon_n\}$, I_{ε_n} has exactly two global minimizers, one converging to $U_{K_{opt},0}$ and the other to $U_{K_{opt},1}$.
2. Along some sequences $\{\varepsilon_n\}$, I_{ε_n} has exactly two global minimizers, one converging to $U_{K_{opt}+1,0}$ and the other to $U_{K_{opt}+1,1}$.
3. Along some sequences $\{\varepsilon_n\}$, I_{ε_n} has exactly four global minimizers. They converge to $U_{K_{opt},0}$, $U_{K_{opt},1}$, $U_{K_{opt}+1,0}$, and $U_{K_{opt}+1,1}$, respectively.

The local minimizers $u_{\varepsilon,K,0}$ and $u_{\varepsilon,K,1}$ found in Theorem 3 may be extended trivially to a box in R^3, i.e., define

$$\tilde{u}_{\varepsilon,K,0}(x_1, x_2, x_3) = u_{\varepsilon,K,0}(x_1), \;\; \tilde{u}_{\varepsilon,K,1}(x_1, x_2, x_3) = u_{\varepsilon,K,1}(x_1)$$

for $(x_1, x_2, x_3) \in (0, 1) \times (0, 1) \times (0, 1)$. These extended functions are still critical points of I_ε on $(0, 1) \times (0, 1) \times (0, 1)$. One wishes to use them to model the lamellar phase of diblock copolymers depicted in Fig. 2.1. However this is not so trivial. To claim that $\tilde{u}_{\varepsilon,K,0}$ and $\tilde{u}_{\varepsilon,K,1}$ model the lamellar phase, one must show that they are still stable in three dimensions, i.e., they are local minimizers of I_ε now defined on $(0, 1) \times (0, 1) \times (0, 1)$. Surprisingly it turns out that they are stable in three dimensions if K is large and unstable if K is small [18]. The number K_{opt} given after (46) is almost the stability borderline for K.

The constraint value a in (10) must be held fixed as $\varepsilon \to 0$ in our Γ-convergence setting. Sometimes it is necessary to study the case that a is small, i.e., $a \to 0$ in a certain way as $\varepsilon \to 0$. This situation seems to be outside the scope of the Γ-convergence theory. However, using another singular perturbation technique, Ren and Wei [24] were able to study a case of I that

$$a = a_0 \varepsilon^{1/2}, \;\; \sigma \sim 1 \tag{48}$$

where $a_0 > 0$ is fixed.

Also outside the scope of the Γ-convergence theory is the study of unstable critical points of I_ε. Due to the existence of a large number of local minimizers of I_ε when ε is small, one expect that I_ε has many saddle points. However in J the $U_{K,0}$'s and the $U_{K,1}$'s are the only critical points of J. So it appears that saddle points are "lost," after we pass to the Γ-limit. Actually even if J had a saddle point, we could not claim from the Γ-convergence theory that there existed a saddle point of I_ε. The Γ-convergence theory only deals with minimizers.

The study of the Ohta–Kawasaki theory in higher dimensions is far from complete. Solutions of the free boundary problem (17) of the Γ-limit problem J which match the cylindrical and spherical phases in Fig. 2.1 have been constructed by Ren and Wei [26–28].

Acknowledgements Supported in part by NSF grant DMS-0907777.

Appendix: Interface Profile and Surface Tension

Suppose that $u_{\varepsilon_n,K,0}$ is a local minimizer found in Theorem 3 which converges to $U_{K,0}$ as $n \to \infty$. Since $u_{\varepsilon_n,K,0}$ is smooth but $U_{K,0}$ is discontinuous, one may be interested in the behavior of $u_{\varepsilon_n,K,0}$ near a discontinuous point of $U_{K,0}$.

Let z_1 be the first discontinuous point and z_2 the second discontinuous point of $U_{K,0}$ given in (33) and (34) so that $U_{K,0}(x) = 0$ if $x \in (0, z_1)$ and $U_{K,0}(x) = 1$ if $x \in (z_1, z_2)$. One can show (see [17]) that there exists $x_{\varepsilon_n} \to z_1$ as $n \to \infty$ where $u_{\varepsilon_n,K,0}(x_{\varepsilon_n}) = 1/2$. If we stretch the variable x, i.e., let $\varepsilon_n y + x_{\varepsilon_n} = x$, then $u_{\varepsilon_n}(\varepsilon_n y + x_{\varepsilon_n})$ as a function of y converges in $C^2([-M,M])$ to a function H for every given $M > 0$. The function $H(y)$ satisfies

$$-H''(y)+W'(H(y)) = 0,\ y \in (-\infty,\infty);\ \lim_{y\to-\infty} H(y) = 0,\ \lim_{y\to\infty} H(y) = 1;\ H(0) = \frac{1}{2}. \tag{A.1}$$

This $H(y)$ is called a heteroclinic orbit of the ODE in (A.1). It describes the asymptotic profile of $u_{\varepsilon,K,0}$ near the discontinuous point z_1. Note that near z_2, $u_{\varepsilon,K,0}$ is approximately $1 - H(y)$.

If W is given by (3), for $t \in (-1, 2)$, one can solve for H explicitly. In this case,

$$W'(t) = t(t - \frac{1}{2})(t - 1). \tag{A.2}$$

Let $Q = H'$. Then

$$\frac{dQ}{dH} = \frac{W'(H)}{Q},$$

from which we deduce that

$$\frac{Q^2}{2} = W(H) + C_1$$

for some constant C_1. The decay conditions on $H(y)$ as $|y| \to \infty$ implies that $C_1 = 0$ and hence

$$Q^2 = \frac{H^2(1-H)^2}{2}.$$

We look for an H whose values are between 0 and 1 and $Q \geq 0$. Then

$$H' = Q = \frac{1}{\sqrt{2}} H(1-H).$$

Separating variables shows that

$$\frac{y}{\sqrt{2}} = \int \frac{dH}{H(1-H)} = \int \frac{dH}{H} + \int \frac{dH}{1-H} = \log H - \log(1-H) + C_2$$

for some constant C_2. The condition $H(0) = 1/2$ implies that $C_2 = 0$, and hence

$$H(y) = \frac{e^{y/\sqrt{2}}}{1+e^{y/\sqrt{2}}} = \frac{1}{2}(1 + \tanh \frac{y}{2\sqrt{2}}). \tag{A.3}$$

The surface tension τ given in (14) can also be expressed in terms of H. For a general W, we multiply the equation in (A.1) by H' to obtain a first integral

$$-\frac{(H'(y))^2}{2} + W(H(y)) = C_3.$$

The decay condition of $H(y)$ as $|y| \to \infty$ implies that $C_3 = 0$ and hence

$$H'(y) = \sqrt{2W(H(y))}.$$

Now we deduce that

$$\begin{aligned} \tau &= \int_0^1 \sqrt{2W(t)}\, dt \\ &= \int_{-\infty}^{\infty} \sqrt{2W(H(y))}\, H'(y)\, dy \\ &= \int_{-\infty}^{\infty} (H'(y))^2\, dy. \end{aligned}$$

Thus we have another formula for the surface tension τ:

$$\tau = \int_{-\infty}^{\infty} (H'(y))^2\, dy. \tag{A.4}$$

References

1. Cahn, J.W., Hilliard, J.E.: Free energy of a nonuniform system. I. Interfacial free energy. J. Chem. Phys. **28**(2), 258–267 (1958)
2. Chen, X., Oshita, Y.: Periodicity and uniqueness of global minimizers of an energy functional containing a long-range interaction. SIAM J. Math. Anal. **37**(4), 1299–1332 (2005)
3. Choksi, R., Ren, X.: On the derivation of a density functional theory for microphase separation of diblock copolymers. J. Stat. Phys. **113**(1–2), 151–176 (2003)
4. Choksi, R., Ren, X.: Diblock copolymer - homopolymer blends: derivation of a density functional theory. Phys. D **203**(1–2), 100–119 (2005)
5. De Giorgi, E.: Sulla convergenza di alcune successioni d'integrali del tipo dell'area. Rend. Mat. **8**(6), 277–294 (1975)
6. Evans, L.C., Gariepy, R.F.: Measure Theory and Fine Properties of Functions. CRC Press, Boca Raton (1992)
7. Hewitt, E., Stromberg, K.: Real and Abstract Analysis. Springer, New York (1965)
8. Kohn, R., Sternberg, P.: Local minimisers and singular perturbations. Proc. Roy. Soc. Edinburgh Sect. A **111**(1–2), 69–84 (1989)
9. Modica, L.: The gradient theory of phase transitions and the minimal interface criterion. Arch. Rat. Mech. Anal. **98**(2), 123–142 (1987)
10. Modica, L., Mortola, S.: Un esempio di Γ^--convergenza. Boll. Un. Mat. Ital. B (5) **14**(1), 285–299 (1977)
11. Müller, S.: Singular perturbations as a selection criterion for periodic minimizing sequences. Calc. Var. Part. Differ. Equat. **1**(2), 169–204 (1993)
12. Nishiura, Y., Ohnishi, I.: Some mathematical aspects of the microphase separation in diblock copolymers. Phys. D **84**(1–2), 31–39 (1995)
13. Ohta, T., Kawasaki, K.: Equilibrium morphology of block copolymer melts. Macromolecules **19**(10), 2621–2632 (1986)
14. Ren, X., Truskinovsky, L.: Finite scale microstructures in nonlocal elasticity. J. Elasticity **59**(1–3), 319–355 (2000)
15. Ren, X., Wei, J.: On the multiplicity of solutions of two nonlocal variational problems. SIAM J. Math. Anal. **31**(4), 909–924 (2000)
16. Ren, X., Wei, J.: Concentrically layered energy equilibria of the di-block copolymer problem. Eur. J. Appl. Math. **13**(5), 479–496 (2002)
17. Ren, X., Wei, J.: On energy minimizers of the di-block copolymer problem. Interfac. Free Boundaries **5**(2), 193–238 (2003)
18. Ren, X., Wei, J.: On the spectra of 3-D lamellar solutions of the diblock copolymer problem. SIAM J. Math. Anal. **35**(1), 1–32 (2003)
19. Ren, X., Wei, J.: Soliton-stripe patterns in charged Langmuir monolayers. J. Nonlinear Sci. **13**(6), 603–624 (2003)
20. Ren, X., Wei, J.: Triblock copolymer theory: ordered ABC lamellar phase. J. Nonlinear Sci. **13**(2), 175–208 (2003)
21. Ren, X., Wei, J.: Chiral symmetry breaking and the soliton-stripe pattern in Langmuir monolayers and smectic films. Nonlinearity **17**(2), 617–632 (2004)
22. Ren, X., Wei, J.:. The soliton-stripe pattern in the Seul-Andelman membrane. Phys. D **188** (3–4), 277–291 (2004)
23. Ren, X., Wei, J.: Stability of spot and ring solutions of the diblock copolymer equation. J. Math. Phys. **45**(11), 4106–4133 (2004)
24. Ren, X., Wei, J.: Droplet solutions in the diblock copolymer problem with skewed monomer composition. Calc. Var. Part. Differ. Equat. **25**(3), 333–359 (2006)
25. Ren, X., Wei, J.: Existence and stability of spherically layered solutions of the diblock copolymer equation. SIAM J. Appl. Math. **66**(3), 1080–1099 (2006)

26. Ren, X., Wei, J.: Many droplet pattern in the cylindrical phase of diblock copolymer morphology. Rev. Math. Phys. **19**(8), 879–921 (2007)
27. Ren, X., Wei, J.: Single droplet pattern in the cylindrical phase of diblock copolymer morphology. J. Nonlinear Sci. **17**(5), 471–503 (2007)
28. Ren, X., Wei, J.: Spherical solutions to a nonlocal free boundary problem from diblock copolymer morphology. SIAM J. Math. Anal. **39**(5), 1497–1535 (2008)

Chapter 3
"Rainbows" in Homogeneous and Radially Inhomogeneous Spheres: Connections with Ray, Wave, and Potential Scattering Theory

John A. Adam

Introduction: The Rainbow, Its Scientific and Mathematical Beauty

"Rainbows have long been a source of inspiration both for those who would prefer to treat them impressionistically or mathematically. The attraction to this phenomenon of Descartes, Newton, and Young, among others, has resulted in the formulation and testing of some of the most fundamental principles of mathematical physics."

K. Sassen [1]

"The rainbow is a bridge between two cultures: poets and scientists alike have long been challenged to describe it... Some of the most powerful tools of mathematical physics were devised explicitly to deal with the problem of the rainbow and with closely related problems. Indeed, the rainbow has served as a touchstone for testing theories of optics. With the more successful of those theories it is now possible to describe the rainbow mathematically, that is, to predict the distribution of light in the sky. The same methods can also be applied to related phenomena, such as the bright ring of color called the glory, and even to other kinds of rainbows, such as atomic and nuclear ones."

H.M. Nussenzveig [2]

"The theory of the rainbow has been formulated at many levels of sophistication. In the geometrical-optics theory of Descartes, a rainbow occurs when the angle of the light rays emerging from a water droplet after a number of internal reflections reaches an extremum. In Airy's wave-optics theory, the distortion of the wave front of the incident light produced by the internal reflections describes the production of

John A. Adam (✉)
Department of Mathematics & Statistics, Old Dominion University, Norfolk, VA 23508, USA
e-mail: jadam@odu.edu

B. Toni (ed.), *Advances in Interdisciplinary Mathematical Research*, Springer Proceedings in Mathematics & Statistics 37, DOI 10.1007/978-1-4614-6345-0_3,

the supernumerary bows and predicts a shift of a few tenths of a degree in the angular position of the rainbow from its geometrical-optics location. In Mie theory, the rainbow appears as a strong enhancement in the electric field scattered by the water droplet. Although the Mie electric field is the exact solution to the light-scattering problem, it takes the form of an infinite series of partial-wave contributions that is slowly convergent and whose terms have a mathematically complicated form. In the complex angular momentum theory, the sum over partial waves is replaced by an integral, and the rainbow appears as a confluence of saddle-point contributions in the portion of the integral that describes light rays that have undergone m^* internal reflections within the water droplet."

J. A. Lock [3]

*In this chapter, $p-1$ will replace m, where $p \geq 1$.

Complementary Domains of Description

This chapter addresses three related topics: the existence of direct transmission (or zero-order) bows in radially inhomogeneous spheres, the Mie solution of electromagnetic scattering, and the associated wave-theoretic/potential scattering connection, to be discussed in detail below. This connection is well illustrated in a series of recent papers by Lock [4–6] (see section "Analysis of Specific Profiles").

Geometrical optics and wave (or physical) optics are two very different but complementary approaches to describing many optical phenomena and here, specifically, the rainbow. However, there is a broad "middle ground," the *semiclassical* régime. Thus, there are essentially three domains within which scattering phenomena may be described: the scattering of waves by objects which in size are (i) small, (ii) comparable with, and (iii) large, compared to the wavelength of the incident (plane wave) radiation. There may be considerable overlap of region (ii) with the others, depending on the problem of interest, but basically, the wave-theoretic principles in region (i) tell us why the sky is blue (amongst many other things!). At the other extreme, the "classical" domain (iii) enables us in particular to be able to describe the basic features of the rainbow in terms of ray optics. The wave-particle duality so fundamental in quantum mechanics is relevant to region (ii) because the more subtle features exhibited by such phenomena involve both these aspects of description and explanation. Indeed, it is useful to relate (somewhat loosely) the régimes (i)–(iii) above to three domains, as stated by Grandy [7]:

(a) *The classical domain*: geometrical optics and particle and particle/raylike trajectories
(b) *The wave domain*: physical optics, acoustic and electromagnetic waves, and quantum mechanics
(c) *The semiclassical domain*: "the vast intermediate region between the above two, containing many interesting physical phenomena"

Geometrical optics is associated with "real" rays, but their analytic continuation to complex values of some associated parameters enables the concept of "complex rays" to be used, often in connection with surface or "evanescent" rays travelling along a boundary while penetrating the less dense medium in an exponentially damped manner. However, complex rays can also be used to describe the phenomenon of *diffraction*: the penetration of light into regions that are forbidden to the real rays of geometrical optics [8], so there are several different contexts in which this term can be used. In fact, the primary bow light/shadow transition region is associated physically with the confluence of a pair of geometrical rays and their transformation into complex rays; mathematically this corresponds to a pair of real saddle points merging into a complex saddle point. For the primary bow then, the two (supernumerary) rays coalesce when they are incident on the sphere surface at the Descartes angle, and the subsequent vanishing of these rays is associated with the complex ray on the shadow side of the rainbow. This does not involve "grazing incidence" at all. On the other hand, rays that graze the sphere *and* just miss grazing it may "tunnel" into the interior, or more accurately, *both* of these regions together form an "edge region" that gives rise to the tunneling ray. This phenomenon is well known in quantum mechanics, specifically tunneling through a classically forbidden potential barrier. Because it occurs in the edge region of semiclassical scattering, it permits grazing rays (and those just outside the sphere) to interact with it (and contribute to the radiation field) [8–10]. As shown by Nussenzveig in a series of very elegant but technical papers [9–12], scattering of scalar waves by a transparent sphere is in many respects isomorphic to the problem of scattering of particles by a spherical potential well. In quantum mechanics, as will be shown later in this chapter, the bound states of a potential well correspond to poles in the elements of a certain matrix, the *scattering matrix*, on the negative real energy axis, whereas *resonances* of the well (as we shall see) correspond to poles that are just below the positive real energy axis of the second Riemann sheet associated with those matrix elements. The closer these poles are to the real axis, the more the resonances behave like very long-lived bound states or "almost bound" states of the system. In very simplistic terms, if a particle with a resonance energy is "shot" at the well from far enough away, it is captured by the well for a considerable time and acts like a bound particle, but eventually it escapes from the well (this, e.g., is a crude description of the mechanism of α-decay from a nucleus, though that is a decay phenomenon, not a scattering one). The reciprocal of the half-width of the resonance is a measure of the lifetime of the resonance particle in the well.

In view of all this then, mathematically at least, a primary "rainbow" is, amongst other things [13, 14]:

(1) a concentration of light rays corresponding to an extremum of the deviation or scattering angle (this extremum is identified as the Descartes' or rainbow ray); (2) a caustic, separating a two-ray region from a 0-ray (or shadow) region; (3) an integral superposition of waves over a (locally) cubic wave front (the *Airy approximation*); (4) a coalescence of two real saddle points; (5) a result of scattering by a square well potential; (6) an example of "Regge-pole dominance"; and (7) a *fold diffraction catastrophe*. Most of these complementary descriptions will not be discussed here; instead the reader is referred to [13, 14] for further details.

Scattering by a Transparent Sphere: Ray Description

In the following discussion, i refers to the angle of incidence for the incoming ray, r is the radial distance within a sphere of radius a (which may be taken to be unity), and $D(i)$ is the deviation undergone by the ray from its original direction. Below, the subscripts 0 and 1 will be used to distinguish the respective deviations of the exiting ray for the direct transmission (or zero order) and the primary bow. For $p-1$ internal reflections in a spherical droplet of *constant* refractive index $n>1$, straightforward geometrical optics reveals that the deviation from its original direction of a ray incident from infinity upon the sphere at angle of incidence i is in radians ($i \in [0, \pi/2]$)

$$D_{p-1}(i) = (p-1)\pi + 2i - 2p \arcsin\left(\frac{\sin i}{n}\right). \tag{1}$$

In general, an extremum of this angle exists at $i = i_c$, where

$$i_c = \arccos\left[\frac{n^2-1}{p^2-1}\right]^{1/2}, p > 1. \tag{2}$$

Naturally, for real optical phenomena such as rainbows, n is such that i_c exists. A primary bow corresponds to $p=2$, a secondary bow to $p=3$, and so forth. That a zero order (or direct transmission bow) corresponding to $p=1$ cannot exist for constant n is readily shown from Eq. (1). Nevertheless, it has been established that such relative extrema (for zero- and higher-order bows) can exist for radially inhomogeneous spheres (see [15, 16] for more details). In fact, multiple zero-order and primary bows may exist depending on the refractive index profile. A well-known result is that the curvature of the ray path is towards regions of higher refractive index n. This is a consequence of Snel's law of refraction generalized to continuously varying media. Thus within the sphere, if $dn(r)/dr \equiv n'(r) < 0$, an incoming ray bends towards the origin; if $n'(r) > 0$, it bends away from it. From Fig. 3.1 it can be seen that for direct transmission in the former case,

$$i + 2\Theta(i) + (i - |D_0(i)|) = \pi \Rightarrow |D_0(i)| = 2i - \pi + 2\Theta(i). \tag{3}$$

In this equation, $2\Theta(i)$ is the angle through which the radius vector turns from the point at which the ray enters the sphere to its point of exit. It is readily noted that for one internal reflection (corresponding to a primary bow)

$$|D_1(i)| = 2i - \pi + 4\Theta(i). \tag{4}$$

In what follows the absolute value notation will be dropped. The deviation formulae can be extended to higher-order bows in an obvious fashion. The quantity $\Theta(i)$ is an improper definite integral to be defined in section "The Ray Path Integral". Analytic expressions for $\Theta(i)$ are difficult to obtain except for a few specific $n(r)$ profiles; several examples are indicated below. For a constant refractive index, $\Theta(i)$

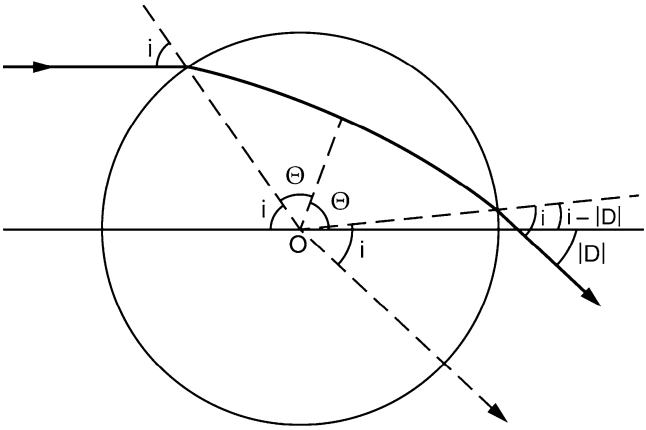

Fig. 3.1 The ray path for direct transmission through a radially inhomogeneous sphere for $n'(r) < 0$

is a standard integral resulting in the inverse secant function and can be readily evaluated. Specifically,

$$D_0(i) = 2i - 2\tilde{r}(i) \quad \text{and} \quad D_1(i) = 2i + \pi - 4\tilde{r}(i), \tag{5a, b}$$

where $\tilde{r}(i)$ is the angle of refraction inside the sphere. Of course, these results are readily determined from elementary geometry and are the $p = 1$ and $p = 2$ cases referred to earlier. As already noted, there can be no "zero-order rainbow" for the direct transmission of sunlight in uniform spheres, only primary and secondary bows (ignoring theoretically possible but practically almost unobservable higher-order bows).

In Fig. 3.2 the dashed curve D_h represents the deviation $D_1(i)$ through a homogeneous sphere of *constant* refractive index $n = 4/3$. The other graphs represent the deviations corresponding to a zero bow and a primary bow for the particular (but arbitrary) choice of refractive index

$$n_1(r) = 1.3 - 0.2\cos\left\{[1.9(r - 0.85)]^2\right\}. \tag{6}$$

Note that both $D_0(i)$ and $D_1(i)$ exhibit fairly broad double extrema in this case. It is interesting to note that the relative maximum for D_1 is much less pronounced than that for D_0. Further discussion of such extrema can be found in [16].

The Ray Path Integral

In a spherically symmetric medium with refractive index $n(r)$ each ray path satisfies the following equation [17]:

$$rn(r)\sin\phi = \text{constant}, \tag{7}$$

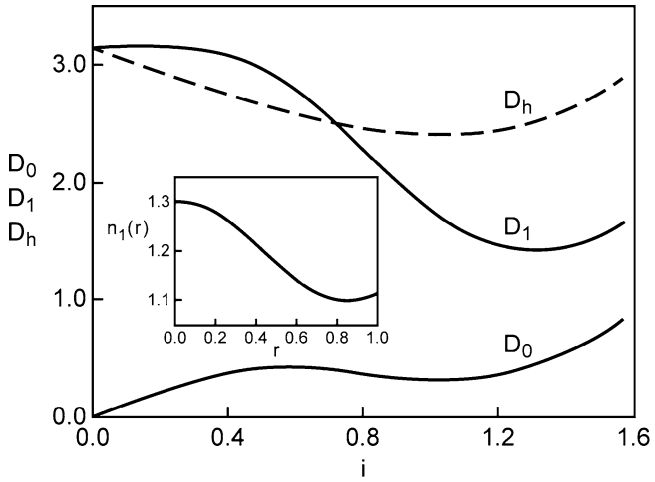

Fig. 3.2 Deviation functions for both a homogeneous (D_h) and inhomogeneous spheres (D_0 and D_1) for the profile $n_1(r)$ [inset]

where ϕ is the angle between the radius vector $\boldsymbol{r}$ and the tangent to the ray at that point (note that $r = |\boldsymbol{r}|$). This expression may be thought of as the optical analogue of the conservation of angular momentum for a particle moving under the action of a central force. The result, known as *Bouguer's formula* (for Pierre Bouguer, 1698–1758), implies that all the ray paths $r(\theta)$ are curves lying in planes through the origin (θ is the polar angle). Elementary differential geometry establishes that

$$\sin\phi = \frac{r(\theta)}{\sqrt{r^2(\theta) + (dr/d\theta)^2}}. \tag{8}$$

From this the angular deviation of a ray, $\Theta(i)$ within the sphere can be determined and subsequently the total angle of deviation $D(i)$ through which an incoming ray at angle of incidence i is rotated. From this the formula for $\Theta(i)$ is found to be

$$\Theta(i) = \sin i \int_{r_c(i)}^{1} \frac{dr}{r\sqrt{r^2 n^2(r) - \sin^2 i}}. \tag{9}$$

The lower limit $r_c(i)$ is the point at which the integrand is singular and is therefore the solution of Eq. (10) below in which (for a *unit* sphere) $\sin i$ is the *impact parameter*. The quantity $r_c(i)$ is the radial point of closest approach to the center of the sphere, sometimes called the *turning point*. The value of $r_c(i)$ is determined implicitly from the following expression:

$$\eta(r_c(i)) \equiv r_c(i)\, n(r_c(i)) = \sin i. \tag{10}$$

The nature of $\eta(r) = rn(r)$ will be very significant in what follows; in particular, $r_c(i)$ will have only one value if $\eta(r)$ is a monotonic function. The integral in Eq. (9) can be evaluated analytically in certain special cases. Consider first the (somewhat unphysical and singular) power-law profile $n(r) = n(R)(r/R)^m$ where m can be of either sign [18]. By a judicious change of variable, this can be reduced to the standard result for a constant refractive index. For the choice of a "shifted hyperbolic" profile of the form $n(r) = (ar+b)^{-1}$, the integral (9) can be evaluated in terms of elementary transcendental functions [15]. The complexity of these integrals increases rapidly with even relatively simple expressions for $n(r)$. In the case of a linear profile, Eq. (3) can be evaluated in terms of incomplete elliptic integrals of the first and third kinds [19, 20]. A parabolic profile of the form $n(r) = a - br^2$ also yields a result also in terms of a purely imaginary elliptic integral of the third kind [20].

Whether the ray path integral is evaluated analytically or numerically, it contributes to the *direct problem* of geometrical optics, namely, (for direct transmission) the total angular deviation $2\Theta(i)$ of the ray inside the sphere for a given profile $n(r)$. Coupled with the refraction at the (in general discontinuous) boundary entrance and exit points, this naturally yields the total deviation of an incoming ray as a function of its angle of incidence. The corresponding *inverse problem* is to determine the profile $n(r)$ from knowledge of the observable deflection function $D(i)$ (note that $D(i) = D(\Theta(i))$). This is generally more difficult to accomplish. Another reason for pursuing the inverse problem is that it would be valuable to find at least some sufficient conditions under which inhomogeneous spheres can exhibit bows of any order but especially of zero order (particularly with regard to industrial techniques such as rainbow refractometry, e.g., see references in [16]). By choosing a generic profile for $D_0(i)$ or $D_1(i)$, for example, it should be possible in principle to examine the implications on $n(r)$ for such profiles. From a strict mathematical point of view, inverse problems in general are notorious for their lack of solution uniqueness. In practical terms it is not significant in this context, and we shall address the topic no further here.

Properties of $\eta(r)$ and Interpretation of the Ray Path Integral

A careful analysis of the integral (9) for $\Theta(i)$ in the neighborhood of the singularity yields two possibilities depending on whether or not $\eta(r)$ is a monotone increasing function:

(i) *Monotonic case*. If $\eta'(r_c) \neq 0$, then in the neighborhood of $r = r_c$, the integral for Θ has the dominant behavior $(r - r_c)^{1/2}$ which tends to zero as $r \to r_c^+$.
(ii) *Non-monotonic case*. If $\eta'(r_c) = 0$, then in the neighborhood of $r = r_c$, the integral for Θ has the dominant behavior $\ln|r - r_c|$ which tends to $-\infty$ as $r \to r_c^+$.

To see this, we expand the quantity $r^2n^2(r)$ about the point $r = r_c$. The radicand then takes the form

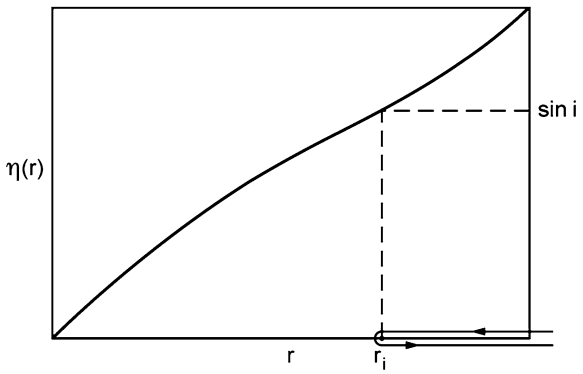

Fig. 3.3 $\eta(r) = rn(r)$ for the monotonic case. The point of closest approach is $r = r_c$

$$r^2n^2(r) - K^2 = r_c^2n^2(r_c) - K^2 + \frac{d}{dr}\left[r^2n^2(r)\right]_{r_c}(r - r_c)$$
$$+\frac{1}{2}\frac{d^2}{dr^2}\left[r^2n^2(r)\right]_{r_c}(r - r_c)^2 + O\left((r - r_c)^3\right). \tag{11}$$

Simplifying (and neglecting extraneous multiplicative and additive constants), we find that, as indicated in Fig. 3.3, if $\left(d\left[r^2n^2(r)\right]/dr\right)_{r_c} > 0$, then the integral in Eq. (9) has the functional form [16]

$$I \propto \int (r - r_c)^{-1/2} dr \propto (r - r_c)^{1/2} \to 0 \tag{12}$$

as $r \to r_c^+$. If on the other hand, $\left(d\left[r^2n^2(r)\right]/dr\right)_{r_c} = 0$, then

$$I \propto \int |r - r_c|^{-1} dr \propto \ln|r - r_c| \to -\infty \tag{13}$$

as $r \to r_c^+$.

Generic $\eta(r)$ profiles for these two cases are illustrated schematically in Figs. 3.3 and 3.4. In the monotonic case, the radius of closest approach for a given angle of incidence is denoted by r_i in Fig. 3.3; the distance of the ray trajectory from the center of the sphere is indicated on the r-axis. This is also indicated in the non-monotonic case in Fig. 3.4. To interpret this figure, it is best to consider rays with angles of incidence increasing away from zero. The radius (point) of closest approach increases in a continuous manner until $i = i_2$ as shown. At that stage the point of closest approach increases discontinuously by an amount Δr to $r = r_c$, thereafter increasing continuously once again. This behavior corresponds to a spherical "zone" of thickness Δr into which *no rays* can penetrate. The situation is reversible: starting with $i = \pi/2$ and reducing, it yields the same zonal gap.

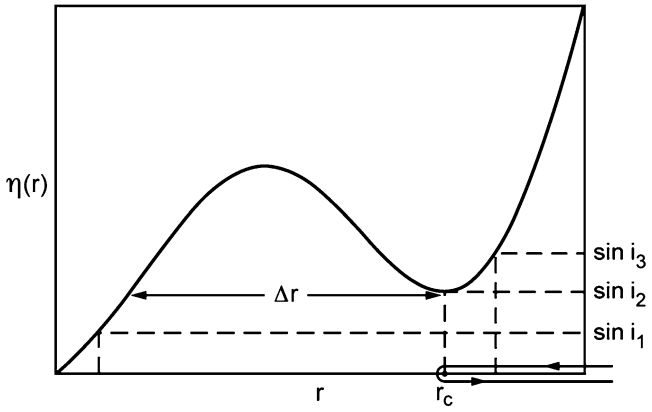

Fig. 3.4 $\eta(r) = rn(r)$ for the non-monotonic case. The point of closest approach for $i > i_2$ is $r = r_c^+$, and a zone of width Δr exists into which no ray penetrates

In scattering theory, the logarithmic singularity (ii) above is associated with the phenomenon of *orbiting*. An extremum of $\eta(r)$ arises at $r = r_c$ when

$$n'(r_c) = -\frac{n(r_c)}{r_c} < 0, \tag{14}$$

meaning that the refractive index profile $n(r)$ either possesses a local minimum at $r = r_m > r_c$, or it tends monotonically to a constant value as r increases to one (see Fig. 3.5). Of course, unlike the case of classical and/or atomic or molecular scattering, $n(r)$ and its corresponding potential $V(r)$ is in general piecewise continuous. The orbiting behavior illustrated in Fig. 3.5 (lower figures) can be thought of as a type of "mechanical" version of a limit cycle in a dynamical system. The connection between the two cases of "classical" and "potential" scattering is illustrated in Appendix 3.

Analysis of Specific Profiles

We now examine two specific (and possibly singular) refractive index profiles for the *unit*sphere, generalizing somewhat that considered in [21]. Before so doing, we introduce some new notation. Electromagnetic waves possess two different polarizations: the transverse electric (*TE*) and transverse magnetic (*TM*) modes. Spherical *TE* modes have a magnetic field component in the direction of propagation, in this case that is in the radial direction, and spherical *TM* modes have an electric field component in the radial direction.

The first profile to be considered is

$$n(r) = n_1 r^{1/b-1}\left(2 - r^{2/b}\right)^{1/2}, n_1 = n(1) > 1. \tag{15}$$

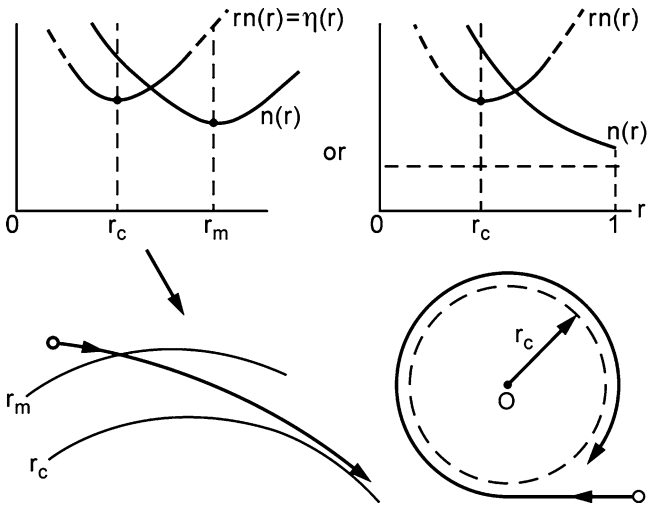

Fig. 3.5 The phenomenon of orbiting illustrated schematically associated with a zero of $\eta'(r)$ showing the $\eta(r)$ and $n(r)$ profiles associated with the existence of a "critical" ray separating two types of ray behavior (upper diagrams). The lower diagrams illustrate two different ways in which rays can approach the critical radius r_c. (See Eq. (14) and the associated discussion in section "Properties of $\eta(r)$ and Interpretation of the Ray Path Integral")

Note that if $b = 1$ and $n_1 = 1$, this profile corresponds to the classic *Luneberg lens* [22]. Using the result (3.5) $D_0(i) = 2i - \pi + 2\Theta$, and substituting for $n(r)$ in the Θ-integral, after some algebra the deviation angle can be shown to be

$$D_0(i) = \pi(b-1) + 2i - b\arcsin\left(\frac{\sin i}{n_1}\right). \tag{16}$$

For a zero-order bow to exist for some critical angle of incidence $i_c \in [0, \pi/2]$, it is necessary and sufficient that $D_0'(i_c) = 0$. This is the case if

$$\cos i_c = 2\left(\frac{n_1^2 - 1}{b^2 - 4}\right)^{1/2}, \tag{17}$$

which implies that $b \geq 2n_1$ if we restrict ourselves to the least potentially singular case of $b > 0$. We have therefore established that a zero bow can exist, unless $n_1 = 1$, whence Eq. (16) is a linear function of incidence angle i. It is interesting to note that the *TE* wave equation (see Appendix 2) has an exact solution for this choice of profile, finite for $0 \leq r \leq 1$, namely,

$$\mathscr{S}_l(r) = r^{l+1}\exp\left(-\frac{bkr^{2/b}}{2}\right) \times {}_1F_1\left(\frac{1}{2} + \frac{b}{2}\left(l + \frac{1}{2} - k\right); 1 + b\left(l + \frac{1}{2}\right); bkr^{2/b}\right). \tag{18}$$

Here ${}_1F_1$ refers to the confluent hypergeometric function. The *TM* equation cannot be expressed in terms of well-known functions, though it can be written in terms of generalized hypergeometric functions and solved by power series expansions in special cases. In a recent series of papers, Lock [4–6] analyzed the scattering of plane electromagnetic waves by a modified Luneberg lens. This "lens" is a dielectric sphere of radius a with a radially varying refractive index [22], specifically

$$n(r) = \frac{1}{f}\left[1 + f^2 - \left(\frac{r}{a}\right)^2\right]^{1/2}. \tag{19}$$

Here f is a parameter determining the focal length of the lens. If $0 < f < 1$, the focus is inside the sphere (i.e., the focal length $< a$); for $f = 1$ it is on the surface, and for $f > 1$ the focal point is outside the sphere. Note that, in contrast to the refractive index profiles (15) and (20), for the profile (19), $n(a) = 1$. Lock also found the existence of a transmission bow for this profile; indeed, this will occur for $f > 1$,whereas for $f = 1$ this bow evolves into an orbiting ray, and if $0 < f < 1$, this ray in turn evolves into a family of morphology-dependent resonances. In a wave-theoretic approach to this problem [5], Lock studied the related radial "Schrödinger" equation for the *TE* mode using the effective potential approach, discussed in section "Morphology-Dependent Resonances: The Effective Potential $U_l(r)$ (Constant n)" below.

When a family of rays has a near-grazing incidence on a dielectric sphere, the so-called far zone consists of (i) an illuminated region containing rays refracted into the sphere and making $p-1$ internal reflections (where $p \geq 1$) before exiting the sphere and (ii) a shadow zone into which no rays enter. (On a related topic, Lock showed that the asymptotic form of the Airy theory bow far into the illuminated region becomes the interference pattern of two supernumerary rays (with slightly different optical path lengths through the sphere.) In an earlier paper [23] he showed that the zero ray/one ray transition for direct transmission is really a regular zero ray/two ray transition (as for a primary bow), with the second ray being a "tunneling ray"; such tunneling will be discussed in section "Morphology-Dependent Resonances: The Effective Potential $U_l(r)$ (Constant n)".)

The other choice for refractive index profile discussed here is

$$n(r) = \frac{2n_1 r^{1/c-1}}{1 + r^{2/c}}, n_1 = n(1). \tag{20}$$

Detailed algebraic manipulation indicates that in this case,

$$D_0(i) = \pi(c-1) + 2i. \tag{21}$$

Obviously, $D_0'(i) \neq 0$ for any value of i, i.e., there is no zero-order bow for this profile. Both *TE* and *TM* modes have finite solutions for $0 \leq r \leq 1$, expressible in terms of the hypergeometric functions ${}_2F_1$, but we do not state them here. For the special case of $c = 1$ and $n_1 = 1$, this profile corresponds to the classic *Maxwell fish-eye* lens [24]. Other analytic solutions for the *TE*/*TM* modes will be discussed elsewhere [19].

Scattering by a Transparent Sphere: Scalar Wave Description

The essential mathematical problem for scalar waves can be thought of either in terms of classical mathematical physics, e.g., the scattering of sound waves, or in quantum mechanical terms, e.g., the nonrelativistic scattering of particles by a square potential well (or barrier) of radius a and depth (or height) V_0 [7, 8]. In either case we can consider a scalar plane wave impinging in the direction $\theta = 0$ on a sphere of radius a. In what follows, a boldface letter refers to a vector quantity, thus here, $\boldsymbol{r} = \langle |\boldsymbol{r}|, \theta, \phi \rangle$ (or $\langle r, \theta, \phi \rangle$) denotes a position vector in space (using a spherical coordinate system). Suppose that we had started with the "classical wave equation" with dependent variable $\tilde{\psi}(\boldsymbol{r}, t) = \psi(\boldsymbol{r}) e^{-i\omega t}$. For the scalar electromagnetic problem, the angular frequency ω, wave number k, and (constant) refractive index n are related by $\omega = kc/n$, c being the speed of light in vacuo. Then for a penetrable (="transparent") sphere, the spatial part of the wave function $\psi(\boldsymbol{r})$ satisfies the scalar Helmholtz equation

$$\nabla^2 \psi + k^2 n^2 \psi = 0, r < a, \tag{22a}$$

$$\nabla^2 \psi + k^2 \psi = 0, r > a. \tag{22b}$$

Again, k is the wave number and $n > 1$ is the (for now, constant) refractive index of the sphere. We can expand the wave function $\psi(\boldsymbol{r})$ as

$$\psi(\boldsymbol{r}) = \sum_{l=0}^{\infty} B_l(k) u_l(r) r^{-1} Y_l^m(\theta, \phi) \equiv \sum_{l=0}^{\infty} A_l(k) u_l(r) r^{-1} P_l(\cos\theta), \tag{23}$$

where $r = |\boldsymbol{r}|$ as noted above and the coefficients $A_l(k)$ will be "unfolded" below. (The coefficients A_l and B_l are related by a multiplicative normalization constant that need not concern us here.) The reason that the spherical harmonics $Y_l^m(\theta, \phi)$ reduce to the Legendre polynomials in the above expression is because the cylindrical symmetry imposed on the system by the incident radiation renders it axially symmetric (i.e., independent of the azimuthal angle ϕ). The equation satisfied by $u_l(r)$ is

$$\frac{d^2 u_l(r)}{dr^2} + \left[k^2 - V(r) - \frac{l(l+1)}{r^2} \right] u_l(r) = 0, \tag{24}$$

where the potential $V(r)$ is now k-dependent, i.e.,

$$V(r) = k^2 \left(1 - n^2\right), r < a$$

$$V(r) = 0, r > a. \tag{25a, b}$$

Since $n > 1$ within the sphere, this potential corresponds to that of a spherical potential well of depth $V_0 = k^2 \left(n^2 - 1\right)$. This leads very naturally to a discussion of the effective potential, wherein the potential $V(r)$ is combined with the "centrifugal barrier" term $l(l+1)/r^2$.

Morphology-Dependent Resonances: The Effective Potential $U_l(r)$ (Constant n)

A rather detailed study of the radial wave equations was carried out by Johnson [25], specifically for the *Mie solution* of electromagnetic theory (see section "The Vector Problem: The Mie Solution of Electromagnetic Scattering Theory"). A crucial part of his analysis was the use of the effective potential for the *TE* mode of the Mie solution, but without any loss of generality, we may still refer to the scalar problem here. This potential is defined as

$$U_l(r) = V(r) + \frac{l(l+1)}{r^2} = k^2(1-n^2) + \frac{l(l+1)}{r^2}, r \le a, \qquad (26a, b)$$

$$= \frac{l(l+1)}{r^2} \approx \frac{\lambda^2}{r^2},\ r > a.$$

It should be noted here that λ as defined here is *not* the wavelength of the incident radiation. For large enough values of $l, [l(l+1)]^{1/2} \approx l + 1/2$. It is clear that $U_l(r)$ has a discontinuity at $r = a$ because of the "addition" of a potential well to the centrifugal barrier. Thus, there arises a tall and thin enhancement corresponding to a barrier surrounding a well (see Fig. 3.6), and this suggests the possible existence of resonances, particularly between the top of the former and bottom of the latter, where there are three turning points (where the energy k^2 is equal to $U_l(r)$). Such resonances are called "shape resonances" (or sometimes "morphology-dependent resonances"); they are quasi-bound states in the potential well that escape by tunneling through the centrifugal barrier. The widths of these resonances depend on where they are located; the smaller the number of nodes of the radial wave function within the well, the deeper that state lies in the well. This in turn determines the width (and lifetime) of the state, because the tunneling amplitude is "exponentially sensitive" to the barrier height and width [13]. Since the latter decreases rapidly with the depth of the well, the smaller is the barrier transmissivity, and the lowest-node resonances become very narrow for large values of $\beta = ka$. The lifetime of the resonance (determined by the rate of tunneling through the barrier) is inversely proportional to the width of the resonance, so these deep states have the longest lifetimes. (To avoid confusion of the node number n with the refractive index in Fig. 3.6, the latter has temporarily been written as N.)

Note that as k^2 is reduced, the bottom B of the potential rises (and for some value of k the energy will coincide with the bottom of the well [25]); however, at the top of the well, $U_l(a) = \lambda^2/a^2$ is independent of k^2, but if k^2 is increased, it will eventually coincide with the top of the well (T). Consider a value of k^2 between the top and the bottom of the well: within this range there will be three radial turning points, the middle one obviously occurring at $r = a$ and the largest at $r = b$ for which $U_l(b) = \lambda^2/b^2$. The smallest of the three ($r_{\min}$) is found by solving the equation

$$k^2 = \frac{\lambda^2}{r_{\min}^2} - \left(n^2 - 1\right)k^2 \qquad (27)$$

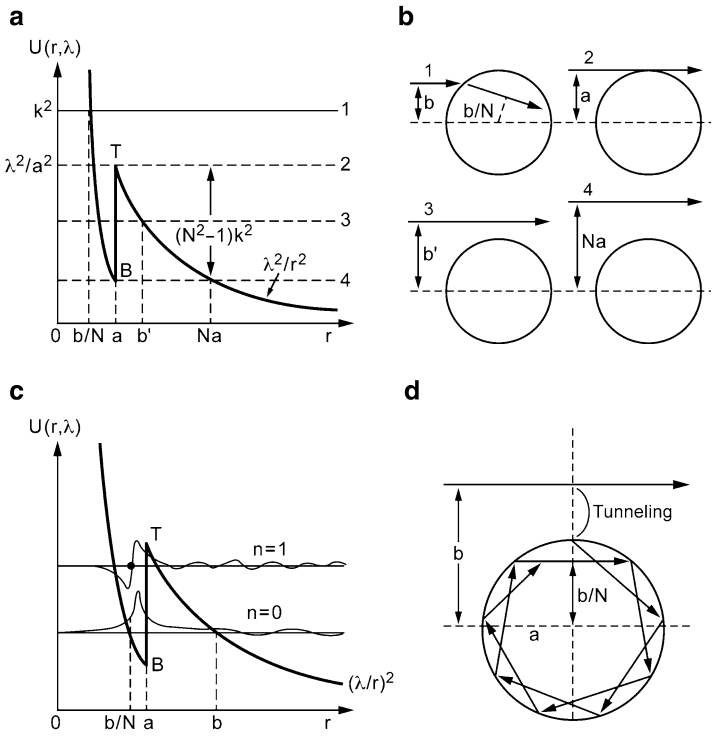

Fig. 3.6 (**a**–**d**) (Redrawn from [8]): (**a**) The effective potential $U(r)$ for a transparent sphere of radius a showing four "energy levels," respectively, above the top of the potential well, at the top, in the middle, and at the bottom of the well. Note that the constant refractive n has temporarily been replaced by N to distinguish it from the node number n in (**c**). (**b**) The corresponding incident rays and impact parameters. Case 2 shows a tangentially incident ray; note that in case 1 the refracted ray is shown. It passes the center at a distance of $l = b/N$; the case is readily shown from simple geometry: from Snel's law of refraction $\sin i = N \sin r = b/a$, and since $l = a \sin r$, the result follows directly. (**c**) Similar to (**a**) but with resonant wave functions shown, corresponding to node numbers $n = 0$ and $n = 1$ (the latter possessing a single node). (**d**) The "tunneling" phenomenon illustrated for an impact parameter $b > a$, being multiply reflected after tunneling, between the surface $r = a$ and the caustic surface $r = b/N$ (the inner turning point)

to obtain, in terms of the impact parameter $b(\lambda) = \lambda/k$,

$$r_{\min} = \frac{\lambda}{nk} \equiv \frac{b}{n}, \tag{28}$$

By applying Snel's law for given b, it is readily shown that the distance of nearest approach of the equivalent ray to the center of the sphere is just $r_{\min}$; indeed, there are in general many nearly total internal reflections (because of internal incidence beyond the critical angle for total internal reflection) within the sphere between $r = b/n$ and $r = a$. This is analogous to orbiting in a ray picture; on returning to its original location after one circumnavigation just below the sphere surface, a ray

must do so with constructive interference. The very low leakage of these states allows the resonance amplitude and energy to build up significantly during a large resonance lifetime which in turn can lead to nonlinear optical effects. In acoustics these are called "whispering gallery modes."

The energy at the bottom of the well (i.e., $\lim_{r\to a^-} U_l(r)$) corresponding to the turning point at $r=a$ is determined by the impact parameter inequalities $a<b<na$ or in terms of $\lambda = kb$:

$$U_l\left(a^-\right) = \left(\frac{\lambda}{na}\right)^2 < k^2 < \left(\frac{\lambda}{a}\right)^2 = U_l\left(a^+\right). \tag{29}$$

This is the energy range between the top and bottom of the well (and in which the resonances occur). To cross the "forbidden region," $a<r<b$ requires tunneling through the centrifugal barrier, and near the resonance energies, the usual oscillatory/exponential matching procedures can lead to very large ratios of internal to external amplitudes (see Fig. 3.6c); these resonances correspond to "quasi-bound" states of electromagnetic radiation (that would be bound in the limit of zero leakage).

We now make a transition to discuss some of the related mathematical properties associated with resonances. In so doing, the reader should be alerted to a somewhat flexible notation used in connection with the scattering function (or S-matrix element to be discussed in section "Introduction to the Scattering Matrix"). This is variously denoted by $\mathscr{S}_l(\lambda,k)$ or $\mathscr{S}_l(\beta)$, where $\beta = ka$, depending on the context. Mathematically, the resonances are complex eigenfrequencies associated with the poles λ_n of the scattering function $\mathscr{S}_l(\lambda,k)$ in the first quadrant of the complex λ-plane; these are known as *Regge poles* (for real k). Corresponding to the energy interval $[U_l(a^-), U_l(a^+)]$, the real parts of these poles lie in the interval $(\beta, n\beta)$ (or equivalently, (ka, nka)); this corresponds to the tunneling region. The imaginary parts of the poles are directly related to resonance widths (and therefore lifetimes). As the node number n decreases, $\mathrm{Re}\lambda_n$ increases and $\mathrm{Im}\lambda_n$ decreases very rapidly (reflecting the exponential behavior of the barrier transmissivity). As β increases, the poles λ_n trace out Regge trajectories, and $\mathrm{Im}\lambda_n$ tend exponentially to zero. When $\mathrm{Re}\lambda_n$ passes close to a "physical" value, $\lambda = l+1/2$, it is associated with a resonance in the lth partial wave; the larger the value of β, the sharper the resonance becomes for a given node number n.

Introduction to the Scattering Matrix

The scattering matrix describes the relationship between the initial and final states of the "system," whatever that may be. In fact it is very useful to relate these states at '$t=-\infty$' and '$t=\infty$' by means of the scattering operator S acting on the wave function ψ, such that $\psi(\infty) = S\psi(-\infty)$. The matrix elements of the operator S form the scattering matrix itself, not surprisingly.

Consider first, for simplicity, a scalar plane wave incident upon an *impenetrable* sphere of radius a. The solution of the Helmholtz equation (22) (outside the sphere) is [7]

$$\psi_k(r,\theta) = \frac{1}{2}\sum_{l=0}^{\infty}(2l+1)\,i^l\left[h_l^{(2)}(kr) + \mathscr{S}_l(\beta)\,h_l^{(1)}(kr)\right]P_l(\cos\theta), \tag{30}$$

where $h_l^{(1)}(kr)$ and $h_l^{(2)}(kr)$ are spherical Hankel functions of the first and second kind, respectively, and

$$\mathscr{S}_l(\beta) = -\frac{h_l^{(2)}(\beta)}{h_l^{(1)}(\beta)}; \beta \equiv ka = \frac{2\pi a}{\lambda}. \tag{31}$$

The quantity $\mathscr{S}_l(\beta)$ is the element (for a given l-value) of the scattering or S-matrix. For "elastic" (or nonabsorptive) scattering, $\mathscr{S}_l(\beta)$ is a phase factor and a very important one—it completely determines the nature of scattering in a potential field. As $|\boldsymbol{r}| = r \to \infty$,

$$h_l^{(1)}(kr) \sim (-i)^{l+1}\frac{e^{ikr}}{kr}; h_l^{(2)}(kr) \sim i^{l+1}\frac{e^{-ikr}}{kr}. \tag{32a, b}$$

Hence inside the summation we have the term

$$\frac{(-1)^{l+1}}{kr}\mathscr{S}_l(\beta)\left[e^{ikr} + \frac{(-1)^{l+1}e^{-ikr}}{\mathscr{S}_l(\beta)}\right]. \tag{33}$$

Again, the reader should note that several possible contexts can be considered here. The modified partial wave number $\lambda = l + 1/2$ is in general considered to be complex, with k being a real quantity, but here we consider k to be a complex quantity also. Thus, so-called bound states (of interest in quantum mechanics) are characterized by a pure imaginary wave number $k = ik_i$, $k_i > 0$ corresponding to energy $E = k^2 < 0$. In order for such a solution to be square integrable in (a,∞), it is necessary that the second term vanish in Eq. (33) above. Formally, this will be the case if $\beta = ka$ is a pole of $\mathscr{S}_l(\beta)$. This is the essential significance of the poles of the S-matrix in what follows.

For a spherical square well or barrier, corresponding to a transparent sphere with constant refractive index n, the form of the scattering matrix elements for scalar waves is more complicated than (31). In fact [8]; see also [26] in terms of spherical Bessel functions (j_l) and spherical Hankel functions, the S-matrix is

$$\mathscr{S}_l(\beta) = -\frac{\beta\,j_l(\alpha)\,h_l'^{(2)}(\beta) - \alpha\,j_l'(\alpha)\,h_l^{(2)}(\beta)}{\beta\,j_l(\alpha)\,h_l'^{(1)}(\beta) - \alpha\,j_l'(\alpha)\,h_l^{(1)}(\beta)}. \tag{34}$$

Equation (34) is an expression of the matching at the finite boundary of the potential of the regular internal solution with the appropriate external solution of the

Schrödinger equation. Using the notation of Nussenzveig [8], the expression (34) is equivalent to

$$\mathscr{S}_l(\beta) = -\frac{h_l^{(2)}(\beta)}{h_l^{(1)}(\beta)}\left[\frac{\ln' h_l^{(2)}(\beta) - n\ln' j_l(\alpha)}{\ln' h_l^{(1)}(\beta) - n\ln' j_l(\alpha)}\right] \tag{35}$$

where $\ln'$ represents the logarithmic derivative operator, j_l is a spherical Bessel function. The "size parameter" $\beta = ka$ plays the role of a dimensionless external wave number, and $\alpha = n\beta$ is the corresponding *internal*wave number. Not surprisingly, $\mathscr{S}_l(\beta)$ may be equivalently expressed in terms of cylindrical Bessel and Hankel functions of half-integer order (see Eq. (39)). Note that for $l = 0$ the S-matrix element takes the simpler form [27]

$$\mathscr{S}_0(\beta) = e^{-2i\beta}\frac{\alpha\cot\alpha + i\beta}{\alpha\cot\alpha - i\beta}. \tag{36}$$

The lth "partial wave" in the series solution (23) (or (30)) is associated with an *impact parameter* $b(l) = (l+1/2)/k$, i.e., only rays "hitting" the sphere $(b \le a)$ are significantly scattered, and the number of terms that must be retained in the series to get an accurate result is slightly larger than β. Unfortunately, for visible light scattered by water droplets in the atmosphere, β is approximately several thousand, and the partial-wave series converges very slowly. This is certainly a nontrivial problem! In the next section, we examine the resolution of this difficulty for both the scalar and the vector wave problem.

Introduction to Complex Angular Momentum (CAM) Theory: The Watson Transform

In the early twentieth century there was a significant mathematical development that eventually had a profound impact on the study of scalar and vector scattering, and the present problem in particular. The *Watson transform*, originally introduced in 1918 by Watson in connection with the diffraction of radio waves around the earth, is a method for transforming the slowly converging partial-wave series (e.g., (30)) into a rapidly convergent expression involving an integral in the complex angular momentum plane. This allows the above transformation to effectively "redistribute" the contributions to the partial-wave series into a few points in the complex plane—specifically the Regge poles and saddle points. Such decomposition means that instead of identifying angular momentum with certain discrete real numbers, it is now permitted to vary continuously through complex values. However, despite this modification, the poles and saddle points have profound physical interpretations in the rainbow problem.

The Watson transform was subsequently modified by several mathematical physicists, including Nussenzveig [10, 12], in studies of the rainbow problem. It is intimately related to the *Poisson sum formula*

$$\sum_{l=0}^{\infty} g\left(l+\frac{1}{2},x\right) = \sum_{m=-\infty}^{\infty} e^{-im\pi} \int_0^{\infty} g(\lambda,x) e^{2\pi i m\lambda} d\lambda, \tag{37}$$

given an "interpolating function" $g(\lambda,x)$, where x denotes a set of parameters and $\lambda = l+1/2$ is again considered to be the complex angular momentum variable. The function g is introduced to generate poles at the "physical" values of λ (or l) so that the corresponding residues account for the original partial-wave series. By means of this conversion of a series to an integral in the complex plane, one is free to deform the path appropriately. The path can be chosen in such a way that the dominant high-frequency contributions to the radiation field come from a small number of "critical points" (such as saddle points or complex poles). This avoids the complexity of summing these contributions over $\beta(=ka)$ partial waves (where $\beta \gg 1$).

It transpires that certain poles in the complex λ-plane are associated with surface waves (Regge poles; see below) and others are associated with morphology-dependent resonances in a particular partial wave. The latter are determined by the poles of the S-function in Eq. (34). But why is *angular momentum* the relevant parameter? A little physics helps us here. Although they possess zero rest mass, in terms of their associated de Broglie wavelength $\hat{\lambda}$, photons have energy $E = hc/\hat{\lambda}$ and momentum $E/c = h/\hat{\lambda}$, where h is Planck's constant and c is the speed of light in vacuo. (Note that the standard notation for wavelength is of course the Greek letter λ; here $\hat{\lambda}$ is used instead to avoid confusion with the complex angular momentum variable.) Thus, for a nonzero impact parameter b_i, a photon will carry an angular momentum $b_i h/\hat{\lambda}$ (b_i being the perpendicular distance of the incident ray from the axis of symmetry of the sun-raindrop system). Each of these discrete values can be identified with a term in the partial-wave series expansion. Furthermore, as the photon undergoes repeated internal reflections, it can be thought of as orbiting the center of the raindrop. As will be reemphasized below, the complex (Regge) poles mentioned above are associated with so-called creeping rays, generated by tangential incidence and propagating around the surface, shedding energy exponentially in a tangential direction. The damping is a result of the increasingly large imaginary part of these poles, leading to a rapidly convergent residue series in the shadow region (inhabited, not by real rays, but by diffracted rays). This approach works well for the impenetrable sphere discussed earlier. In the illuminated region, the primary contributions come, not surprisingly, from real rays—stationary optical paths determined by Fermat's principle of least time. These rays are associated with stationary phase points on the real λ-axis (real saddle points).

Unfortunately, for a penetrable (or transparent, or dielectric) sphere, the Regge poles are situated much closer to the real λ-axis, and the convergence is compromised. To remedy this, the solution must be "unfolded" in terms of surface-to-center reflections (and vice versa)—resulting in the so-called *Debye series* (see Appendix 1). The scattering amplitudes can then be expanded in a series, each term of which represents a surface interaction. When the modified Watson transform is applied to each term, one set of the resulting Regge–Debye poles, as they are called, are associated with rapidly damped surface waves (see below), and rapidly convergent asymptotic expansions are obtained for each term in the Debye series. In this

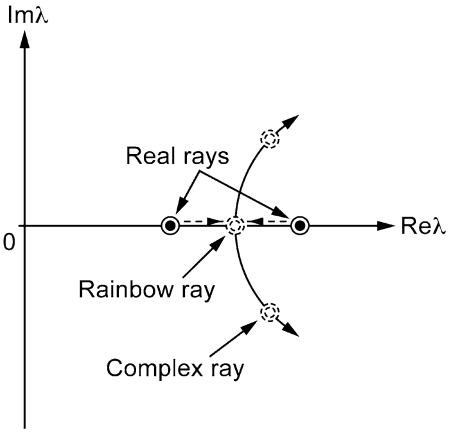

Fig. 3.7 (Redrawn from [11]): The 'collision' of two real saddle points in the complex λ-plane as the rainbow angle (θ_R) is approached from below (i.e., from the illuminated side). At θ_R the points collide and subsequently move away from each other along complex conjugate directions as θ increases away from θ_R into the shadow region. It is the lower complex saddle point that contributes to the wave field in this region

case, the critical points in the λ-plane are exactly those poles and (possibly complex) saddle points. There is a significant difference between the surface waves in this case and the case for the impenetrable sphere; however, they can also take a shortcut through the sphere (critical refraction) and reemerge tangentially as surface waves.

For a Debye term of a given order, p (where $p-1$, $p \geq 1$) is the number of internal reflections at the surface, and a primary rainbow (in particular) is associated in the λ-plane with the existence of two real saddle points that move towards each other as the "rainbow scattering angle" is approached (see Fig. 3.7), merging together at this angle and beyond which (i.e., in the shadow region) the saddle points become complex and move away from the real axis in complex conjugate directions. Thus, as described in [7, 8, 10], from a mathematical point of view, *a rainbow can be defined as a collision between two saddle points in the complex angular momentum plane*.

As will be shown in section "The Partial-Wave Scattering Phase Shift $\delta_l(k)$", *the scattering amplitude* $f(k,\theta)$ is a quantity of fundamental importance in scattering theory; see section "The Partial-Wave Scattering Phase Shift $\delta_l(k)$" (see Eqs. (50) and (51)). It is defined in terms of the scattering matrix elements $\mathscr{S}_l(k)$, and using the Poisson summation formula it may be recast as

$$f(k,\theta) = \frac{i}{ka} \sum_{m=-\infty}^{\infty} (-1)^m \int_0^{\infty} (1 - \mathscr{S}_l(\lambda,k)) P_{\lambda-1/2}(\cos\theta) e^{2\pi i m\lambda} \lambda d\lambda. \qquad (38)$$

For fixed β, $\mathscr{S}_l(\lambda,\beta)$ is a meromorphic function of the complex variable $\lambda = l+1/2$, and again it is the poles of this function that are of interest. In terms of

cylindrical Bessel and Hankel functions, they are defined by the condition

$$\ln' H_{\lambda}^{(1)}(\beta) = n \ln' J_{\lambda}(\alpha). \tag{39}$$

As already noted, they are called Regge poles in the scattering theory literature [7, 8]. For the transparent sphere, *two* types of Regge poles arise. Nussenzveig's *class I poles* [9], located near the real λ-axis, are associated with resonances, via the internal structure of the potential, which is now of course accessible. These are characterized by an effective radial wave number within the potential well. Typically, *class II poles* are associated with surface waves for the impenetrable sphere problem mentioned above—and lead to a rapidly convergent residue series, representing the surface wave (or diffracted or creeping ray) contributions to the scattering amplitude. Seeking poles of the S-matrix in the complex angular momentum plane and their Regge trajectories as the energy E (or wave number k) is varied is in fact equivalent to analyzing these singularities and their trajectories in the complex k-plane as the angular momentum l is varied continuously through real values. In [26] it is pointed out that these two approaches—Regge trajectories and k-trajectories—are two different but complementary mathematical descriptions of the same physical phenomena, and that each one can provide insight into the other.

In the next section we examine another fundamental concept in scattering theory: the *phase shift*. This will prove to be crucial to understanding the changes induced on an incident wave on encountering a potential, be it of finite range or not.

The Partial-Wave Scattering Phase Shift $\delta_l(k)$

We return to the radial equation (24) in order to introduce this fundamental entity. The boundary conditions are that $u_l(r)$ and $u_l'(r)$ are continuous at the surface. We seek a solution satisfying the boundary condition at the origin

$$u_l(r)_{r\to 0} \sim r^{l+1}. \tag{40}$$

In the absence of a potential, the solutions $u_l(r)$ can be expressed in terms of Riccati–Bessel functions of the first and second kind (which are in turn related to the spherical Bessel functions of the first and second kind, $j_l(kr)$ and $y_l(kr)$, respectively):

$$\psi_l(kr) = kr j_l(kr) = \left(\frac{\pi kr}{2}\right)^{1/2} J_{l+1/2}(kr) \sim \sin(kr - l\pi/2) \text{ as } r \to \infty, \text{ and} \tag{41}$$

$$\xi_l(kr) = kr y_l(kr) = (-1)^{l-1} \left(\frac{\pi kr}{2}\right)^{1/2} Y_{(l+1/2)}(kr) \sim \cos(kr - l\pi/2) \text{ as } r \to \infty. \tag{42}$$

(Note that some definitions of $\xi_l(kr)$ use the negative of the above expression, although $\chi_l(kr)$ is commonly used in the literature instead of $\xi_l(kr)$.) Based on the

asymptotic forms of the Riccati–Bessel functions, we expect the solution of (24) to have the following property involving a k- and l-dependent phase shift:

$$u_l(r)_{r\to\infty} \sim \sin(kr - l\pi/2 + \delta_l(k)). \tag{43}$$

In fact, if $V(r)$ can be neglected for $r > r_0$, say, the solution of Eq.(24) can be written in terms of the phase shift $\delta_l(k)$ as [28–30]

$$u_l(r) = kr[j_l(kr)\cos\delta_l(k) - y_l(kr)\sin\delta_l(k)]. \tag{44}$$

In particular, for a spherical well or barrier of radius a, the potential is zero for $r > a$. The k- or energy-dependent partial-wave phase shifts $\delta_l(k)$ represent the effect the potential $V(r)$ on the partial waves comprising the incident plane wave. The quantities $\delta_l(k)$ are real functions of the wave number k when the potential $V(r)$, energy $E(= k^2)$, and angular momentum l are all real. Shortly we shall reintroduce the S-matrix, this time with matrix elements defined in terms of the phase shifts $\delta_l(k)$. Particle scattering in a potential field is completely determined by these elements. The physical interpretation of the phase shifts can be understood as follows. The incoming plane wave is broken up into an infinite number of parts of differing angular momentum (these are the partial waves). Each partial wave interacts individually with the potential to produce a scattered outgoing partial wave. The phase of the outgoing wave is "pushed out" by an amount delta by a repulsive potential, and the phase is "pulled in" by an amount delta for an attractive potential. In optical terms for a sphere of refractive index $n > 1$, it is the latter case that applies: the potential is attractive.

Although it is the poles of the S-matrix that are of interest in this chapter, it is valuable to reflect on the significance of several other concepts introduced here and below. As noted earlier, the phase shift is a measure of the departure of the radial wave function from the form it has when the potential $V(r)$ is zero. It follows from the definition below of the *K-matrix* that this too is a related measure of the distortion induced by a nonzero potential. The K-matrix is especially useful if the interaction is in some sense "weak." The *differential cross section* (Eq. (52b)) is useful because it is the quantity that is directly measured in scattering experiments. The *Jost functions* are useful because they help express the pole structure and associated zero structure of the S-matrix in a very straightforward way.

Returning to the asymptotic result (43), it is also of interest to note that it can be expressed in two other equivalent ways. They are

$$(i)u_l(r)_{r\to\infty} \sim \cos\delta_l[\sin(kr - l\pi/2) + K_l\cos(kr - l\pi/2)], \tag{45}$$

$$\text{and}(ii)u_l(r)_{r\to\infty} \sim \frac{e^{-i\delta_l}}{2i}\left[e^{-i(kr-l\pi/2)} - \mathscr{S}_l(k)\,e^{i(kr-l\pi/2)}\right]. \tag{46}$$

The first of these equations defines the elements of the K-matrix, i.e., $K_l = \tan\delta_l$, and the second (re)defines the S-matrix elements, i.e., $\mathscr{S}_l(k) = e^{2i\delta_l}$. In fact,

$$\mathscr{S}_l(k) = \exp[2i\delta_l(k)] = \frac{1 + i\tan\delta_l(k)}{1 - i\tan\delta_l(k)} \equiv \frac{1 + iK_l(k)}{1 - iK_l(k)}. \tag{47}$$

The integral equation satisfied by the radial wave function $u_l(r)$ can also be written in terms of the Riccati–Bessel functions as follows:

$$u_l(r) = \psi_l(kr) - k^{-1}\int_0^r \left[\psi_l(kr)\,\xi_l(kr') - \psi_l(kr')\,\xi_l(kr)\right] V(r')u_l(r')\,dr'. \tag{48}$$

This may be verified by direct substitution into Eq. (24), where now

$$\lim_{r\to 0} u_l(r) = \lim_{r\to 0}\psi_l(kr) \to \frac{(kr)^{l+1}}{(2l+1)!!}. \tag{49}$$

At large distances from the sphere ($r >> a$) the complete wave field $\psi(\boldsymbol{r})$ can be decomposed into an (axially symmetric) incident wave + scattered field, i.e.,

$$\psi(r,\theta) \sim e^{ikr\cos\theta} + \frac{f(k,\theta)}{r}e^{ikr}. \tag{50}$$

In terms of the scattering matrix element for a given l, and therefore $\mathscr{S}_l(k)$, the *scattering amplitude* is defined as

$$f(k,\theta) = (2ik)^{-1}\sum_{l=0}^{\infty}(2l+1)\left(\mathscr{S}_l(k) - 1\right)P_l(\cos\theta). \tag{51}$$

$P_l(\cos\theta)$ is a Legendre polynomial of degree l. In terms of the phase shift δ_l, the scattering amplitude can be written as

$$f(k,\theta) = k^{-1}\sum_{l=0}^{\infty}(2l+1)e^{i\delta_l}\sin\delta_l P_l(\cos\theta); \tag{52a}$$

For completeness, in the scattering literature, the *differential scattering cross section* is defined by

$$\frac{d\sigma}{d\Omega} = \frac{\text{scattered flux/unit solid angle}}{\text{incident flux/unit area}} = |f(\theta)|^2, \tag{52b}$$

and the *total* (*elastic*) *cross section* σ is obtained by integrating the differential cross section over all scattering angles, i.e.,

$$\sigma = \int_0^{2\pi} d\phi \int_0^{\pi} |f(\theta)|^2 \sin\theta\, d\theta = 2\pi\int_0^{\pi} |f(\theta)|^2 \sin\theta\, d\theta. \tag{52c}$$

The quantity

$$p_l\left(k^2\right) = k^{-1}e^{i\delta_l}\sin\delta_l = (2ik)^{-1}\left(e^{2i\delta_l} - 1\right) \tag{53}$$

is often referred to as the *partial-wave scattering amplitude*.

Analytic Properties of the S-Matrix: The Jost Functions

We now consider in more detail the analytic properties of the partial-wave S-matrix, with elements defined by Eqs. (34) or (35), for example, in the complex momentum plane. We can show that the poles of the S-matrix lying on the positive imaginary k-axis correspond to bound states, while poles lying in the lower half k-plane close to the positive real k-axis correspond to the resonances discussed above (see Appendix 4). We may also derive an expression for the behavior of the phase shift and the cross section when the energy of the scattered particle is in the neighborhood of these poles. Consider again the solution $u_l(r)$ of the radial Schrödinger equation (24) describing the scattering of a particle by a spherically symmetric potential $V(r)$. Implicit in the results to be stated here are certain requirements on the potential $V(r)$. It must be a real, almost everywhere continuous function vanishing at infinity. Furthermore [31, 32], it must be the case that

$$(i)\ \int_c^\infty |V(r)|dr = M(c) < \infty \quad \text{and}$$
$$(ii)\ \int_0^{c'} r|V(r)|dr = N(c') < \infty,$$

where c and c' are positive constants (but otherwise arbitrary). The first of these conditions is equivalent to $V \sim r^{-(1+\varepsilon)}$ as $r \to \infty$, $\varepsilon > 0$ (i.e., $rV(r) \to 0$ as $r \to \infty$), and the second implies that $V \sim r^{-(2+\varepsilon')}$ as $r \to 0, \varepsilon' > 0$ (i.e., $r^2 V(r) \to 0$ as $r \to 0$). (Note that Burke [28] places more stringent conditions on the potential for the existence of bound states; instead of (i) he requires that $\int_0^\infty r^2 |V(r)|dr < \infty$.) We also introduce two (normalized) *Jost solutions* $f_l(\pm k, r)$ of (24), defined by the relations

$$\lim_{r\to\infty} f_l(\pm k, r)\, e^{\pm i(kr \mp l\pi/2)} = 1. \tag{54}$$

This condition at infinity defines $f_l(k, r)$ uniquely in the lower half k-plane, where it is analytic. In the upper half plane, $f_l(k, r)$ is no longer unique because it is always possible to add to it a term proportional to the other Jost solution $f_l(-k, r)$. If the potential vanishes identically beyond a certain distance a then $f_l(\pm k, r)$ are analytic functions of k in the open k-plane for all fixed values of r, that is, they are entire functions of k. We can express the physical solution of (24), defined by the boundary conditions as a linear combination of $f_l(\pm k, r)$, in keeping with the form (44). Thus,

$$u_l(r) \propto \left[f_l(k, r) + (-1)^{l+1} f_l(-k, r)\, \mathscr{S}_l(k) \right]. \tag{55}$$

From a theorem proved by Poincaré, the absence of a k-dependence in this boundary condition implies that this solution is an entire function of k. The *Jost functions* are then defined by

$$\tilde{f}_l(\pm k) = W\left[f_l(\pm k, r), u_l(r) \right], \tag{56}$$

where the Wronskian W is independent of r. It is also convenient to introduce a *normalized* Jost function $f_l(\pm k)$ by

$$f_l(\pm k) = \frac{k^l \exp(\pm il\pi/2)}{(2l+1)!!}\tilde{f}_l(\pm k). \tag{57}$$

(Note that the notation for these functions should not be confused with the definition of the scattering amplitude in Eqs. (51) and (52a).) The functions $f_l(+k)$ and $f_l(-k)$ are continuous at $k=0$ and approach unity at large $|k|$ for Im $k \leq 0$ and Im $k \geq 0$, respectively.

Since

$$W[f_l(\pm k, r), f_l(\mp k, r)] = \pm 2ik, \tag{58}$$

$u_l(r)$ may be written in the form

$$u_l(r) = \frac{1}{2ik}[\tilde{f}_l(k) f_l(-k,r) - \tilde{f}_l(-k) f_l(k,r)]. \tag{59}$$

Comparing this equation with the asymptotic form (44) and using (54) then yields the following expression for the S-matrix elements:

$$\mathscr{S}_l(k) = e^{i\pi l}\frac{\tilde{f}_l(k)}{\tilde{f}_l(-k)} = \frac{f_l(k)}{f_l(-k)}. \tag{60}$$

This equation relates the analytic properties of the S-matrix with the simpler analytic properties of the Jost functions [29]. Since, in particular, $f_l(-k,r)$ satisfies Eq. (24), i.e.,

$$\left(\frac{d^2}{dr^2} + k^2 - V(r) - \frac{l(l+1)}{r^2}\right) f_l(-k,r) = 0. \tag{61}$$

It follows that if we now take the complex conjugate of this equation, we obtain (for real l and $V(r)$)

$$\left(\frac{d^2}{dr^2} + \bar{k}^2 - V(r) - \frac{l(l+1)}{r^2}\right) \bar{f}_l(-k,r) = 0. \tag{62}$$

If we also let $k \to -\bar{k}$ in (61), we also have that

$$\left(\frac{d^2}{dr^2} + \bar{k}^2 - V(r) - \frac{l(l+1)}{r^2}\right) f_l(\bar{k},r) = 0. \tag{63}$$

Furthermore,

$$\bar{f}_l(-k,r)_{r\to\infty} \sim \exp(-i\bar{k}r) \text{ and } f_l(\bar{k},r)_{r\to\infty} \sim \exp(-i\bar{k}r), \tag{64a, b}$$

i.e., they satisfy the same boundary conditions at infinity. Since these functions also satisfy the same differential equation, namely, (62) and (63), respectively, they are

equal for all r for all points in the upper half k-plane and for all other points which admit an analytic continuation from the upper half k-plane. Hence in this region $\bar{f}_l(-k,r) = f_l(\bar{k},r)$, and hence, from (56), $\bar{\tilde{f}}_l(-k) = \tilde{f}_l(\bar{k})$. Therefore, from (60) we find that

$$\mathscr{S}_l(k)\mathscr{S}_l(-k) = e^{2\pi i l}\frac{\tilde{f}_l(k)}{\tilde{f}_l(-k)}\frac{\tilde{f}_l(-k)}{\tilde{f}_l(k)} = 1. \tag{65}$$

We also have the unitarity condition

$$\mathscr{S}_l(k)\bar{\mathscr{S}}_l(\bar{k}) = \frac{\tilde{f}_l(k)}{\tilde{f}_l(-k)}\frac{\bar{\tilde{f}}_l(\bar{k})}{\bar{\tilde{f}}_l(-k)} = 1. \tag{66}$$

These relations give in turn the reflection property

$$\mathscr{S}_l(k) = e^{2\pi i l}\bar{\mathscr{S}}_l(-\bar{k}). \tag{67}$$

It follows from (66) that if k is real then $|\mathscr{S}_l(k)| = 1$ and in terms of the real phase shift $\delta_l(k)$,

$$\mathscr{S}_l(k) = \exp[2i\delta_l(k)]. \tag{68}$$

This is a result already noted above. The poles and zeros of the S-matrix are symmetrically situated with respect to the imaginary k-axis, because it follows from (67) that if the S-matrix has a pole at the point k, then it also has a pole at the point $-\bar{k}$, and from (65) and (6) it has zeros at the points $-k$ and $\bar{k}$. For potentials satisfying the conditions stated at the beginning of this section, only a finite number of bound states can be supported, and these give rise to the poles lying on the positive imaginary axis in Fig. 3.8. However, an infinite number of poles can occur in the lower half k-plane. If they do not lie on the negative imaginary k-axis, they occur in pairs symmetric with respect to this axis, as discussed above. If they lie on the negative imaginary k-axis, they are often referred to as *virtual state* poles; the wave functions corresponding to these states cannot be normalized. Poles lying in the lower half k-plane and close to the real positive k-axis give rise to resonance effects in the cross section equation (52c). Poles lying in the lower half k-plane and far away from the real positive k-axis contribute to the smooth "background" or "nonresonant" scattering. The distribution of poles in the complex k-plane has been discussed in detail in a few cases, (see, e.g., [27]) for scattering by a square well potential.

The Breit-Wigner Form

Consider an isolated pole in the S-matrix which lies in the lower half k-plane close to the positive real k-axis. This pole gives rise to resonance scattering at the nearby real energy. We note (by virtue of Appendix 5) that the pole occurs at the complex energy

$$E = E_r - \frac{i}{2}\Gamma, \tag{69}$$

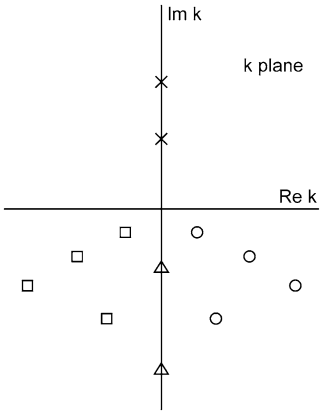

Fig. 3.8 (Redrawn from [28]): A generic distribution of poles for the S-matrix. Crosses correspond to bound-state poles, circles to resonance poles, squares to their conjugate poles, and triangles to virtual states

where E_r is the resonance position and Γ is the resonance width and both are real positive numbers. From the unitarity relation (66), we see that corresponding to this pole there is a *zero* in the S-matrix (at a complex energy given by $E = E_r + i\Gamma/2$) in the upper half k-plane. For energies E on the real axis in the neighborhood of this pole, the S-matrix can be written in a form which is both unitary and explicitly contains the pole and zero:

$$\mathscr{S}_l(k) = \exp\left[2i\delta_l^0(k)\right] \frac{E - E_r - i\Gamma/2}{E - E_r + i\Gamma/2}. \tag{70}$$

The quantity $\delta_l^0(k)$ in this equation is called the "background" or "nonresonant" phase shift. Provided that the energy E_r is not close to threshold, $E = 0$, nor to another resonance, then the background phase shift is slowly varying with energy. Comparing (68) and (70), we obtain the following expression for the phase shift:

$$\delta_l(k) = \delta_l^0(k) + \delta_l^r(k). \tag{71}$$

The quantity

$$\delta_l^r(k) = \arctan\left(\frac{\Gamma/2}{E_r - E}\right) \tag{72}$$

is called the "resonant" phase shift which is seen to increase through π radians as the energy E increases from well below to well above the resonance position E_r.

Further Comments on Jost Functions and Bound States

It can be seen from Eq. (46) that $\mathscr{S}_l(\beta)$ is proportional to the ratio of the coefficients of the outgoing and incoming waves (recall that the harmonic factor $e^{-i\omega t}$ has

been suppressed). According to the theorem of Poincaré mentioned earlier, if the boundary conditions on a differential equation are independent of the parameters in the equation, the solutions will be analytic functions of those parameters. Therefore, the solutions $u_l(r)$ of Eq. (24) will be analytic functions of energy $E = k^2$ if the normalization condition on the behavior of $r^{-(l+1)}u_l(r)$ as $r \to 0$ is also independent of k^2 [33, 34]. For small values of k it can be shown that $\tan\delta_l \sim k^{2l+1} = (k^2)^{l+1/2}$; in other words, δ_l is an analytic function of k (as opposed to k^2) near zero energy. Since $\exp(2i\delta_l)$ is an analytic function of δ_l, the Jost functions will share the branch points of δ_l. As noted earlier, it is customary to divide the k^2-plane into two Riemann sheets by requiring that the "physical" sheet corresponds to $\mathrm{Im}\ k = \mathrm{Im}(k^2)^{1/2} > 0$ on that sheet. The positive k^2-axis is a branch cut [35].

From Eq. (60), poles of $\mathscr{S}_l(k)$ occur when $f_l(-k) = 0$. In the neighborhood of such a zero, we see from Eqs. (54) and (59) that asymptotically

$$u_l(r) \propto f_l(k)\, e^{ikr}. \tag{73}$$

Recalling that on the physical sheet $\mathrm{Im}\ k = k_i > 0$, it follows that $u_l(r) \propto e^{ik_r r}e^{-k_i r}$ so that it is a square integrable and hence normalizable solution; this means it represents a bound state. But such a state for an attractive potential (such as a spherical square well) implies that $k^2 < 0$, that is, $k = ik_i$. Poles on the physical sheet produce an exponentially decaying wave function, so the zeros of $f_l(-k)$ for $k_i > 0$ are bound states. In particular, for the case $l = 0$ it can be shown that $\mathscr{S}_l(k)$ can have poles only where either $\mathrm{Re}(k) = 0$ or $\mathrm{Im}(k) < 0$ [27, 36, 37] (this is proved in Appendix 4). Furthermore, since the partial-wave amplitude $p_l(k^2)$ can be expressed in terms of the Jost functions, this means that poles of $p_l(k^2)$ (Eq. (53)) on the physical sheet are also associated with bound states.

In summary at this point, the scattering matrix elements $\mathscr{S}_l(k)$, regarded as functions of the complex variable k, have several valuable physical interpretations. If k is real, the scattering is defined in terms of real phase shifts δ_l which in turn determine the scattering cross section. Poles of the elements which are pure imaginary with (i) $\mathrm{Im}(k) > 0$ correspond to bound states of the potential, those with (ii) $\mathrm{Im}(k) < 0$ correspond to "virtual" or non-normalizable states (or "antibound" states [27]). If the poles are complex with $\mathrm{Im}(k) < 0$, they are sometimes referred to as "quasi-stationary states," and if $\mathrm{Re}(k) > 0$ and $|\mathrm{Im}(k)| \ll 1$, they are called resonance poles. In the complex E-plane, poles associated with quasi-stationary states are on the second sheet of the Riemann energy surface.

Regge Poles and Regge Trajectories

Following directly from the previous sentence, the "unphysical" Riemann sheet (but close to the branch cut) and poles of the S-matrix elements (now written as $\mathscr{S}_l(k^2)$), i.e., at $E = k^2 = k_r^2 - i\Gamma/2$ (where Γ is "small" and positive), give rise to the familiar Breit–Wigner expression examined above for the phase shift δ_l. Each such pole on

the unphysical sheet corresponds to a *resonance* with energy k_r^2 and "half-width" $\Gamma/2$. What happens as l varies in the radial Schrödinger equation? Again, from Poincaré's theorem, the Jost functions will be analytic functions of l as well as k^2, and we know that the bound states of $V(r)$ are found as the zeros of $f_l(-k)$. This criterion can be regarded as an implicit function in l and k (or indeed, l and $E = k^2$), i.e., $l = g(E)$ (this is a generic function, not the same one as in Eq. (37)). Again, Eqs. (54) and (59) imply that for $k^2 < 0$,

$$u_l(r)_{r\to\infty} \propto \left(f_l(k)\, e^{-k_i r} - f_l(-k)\, e^{k_i r}\right). \tag{74}$$

Since the radial Schrödinger equation is expressed in terms of real quantities only, the solution $u_l(r)$ is real and so are the Jost functions by virtue of (74); therefore, (in particular) $f_l(-k)$ is also real. Hence the zeros $l = g(E)$ of this function must also be real functions. On the other hand, if $k^2 > 0$, $u_l(r)$ is still real, but the complex exponential factors imply that $f_l(-k)$ will be a complex function, whence in general, the Regge pole trajectories $l = G(E)$ (say) will be complex. However, bound states of angular momentum l exist when a trajectory intersects the line $l = m, m = 0, 1, 2, \ldots$, with corresponding energy $k^2 = k^2(l)$.

By contrast, in the complex k-plane for real and positive values of $\lambda = l + 1/2$, all poles in the upper half plane must lie on the imaginary axis. Both complex and pure imaginary poles can be present in the lower half plane [27], and for physical (half-integer) values of λ, the symmetry of these poles with respect to the imaginary axis is established from the following property for the generalized S-matrix element $\mathscr{S}(\lambda, k)$, namely, that $\mathscr{S}(\lambda, k) = \bar{\mathscr{S}}\left(\bar{\lambda}, -\bar{k}\right)$. Note that, according to [26], this relation is no longer valid for unphysical values of λ. In summary, there are two infinite families of "k-poles" (corresponding to the two classes of Regge poles discussed by Nussenzveig [9, 10]; see also section "Introduction to Complex Angular Momentum (CAM) Theory: The Watson Transform" above). Class I poles, we recall, are determined by the interior of the potential and are located in the fourth quadrant near the positive real semiaxis. By contrast, class II poles correspond to surface modes on the "spherical potential" and are located in the third and fourth quadrants. More details can be found in [26].

The Vector Problem: The Mie Solution of Electromagnetic Scattering Theory

The quantum mechanical scalar analysis in previous sections is appropriate primarily for nonrelativistic scattering of a projectile "particle" of mass m. In this section a very different phenomenon is discussed: scattering of zero rest-mass photons. The crucial point to note here is that both of these very different physical systems share the same mathematical structure, namely, the properties of the scalar wave equation.

So having made considerable reference to the scalar problem and its connection with the potential scattering theory, we now turn to the vector problem which for

electromagnetic waves possesses two polarizations (the *TE* and *TM* modes); each radial equation can be examined in turn as a scalar problem. Mie theory is based on the solution of Maxwell's equations of electromagnetic theory for a monochromatic plane wave from infinity incident upon a homogeneous isotropic sphere of radius a. The surrounding medium is transparent (as the sphere may be), homogeneous, and isotropic. The incident wave induces forced oscillations of both free and bound charges in synchrony with the applied field, and this induces a secondary electric and magnetic field, each of which has components inside and outside the sphere [17].

In this section reference will be made to the intensity functions i_1, i_2, the Mie coefficients a_l, b_l, and the angular functions π_l, τ_l. The intensity functions are proportional to the square of the magnitude of two incoherent, plane-polarized components scattered by a single particle; they are related to the scattering amplitudes S_1 and S_2 in the notation of Nussenzveig [11]. The function $i_1(\beta, n, \theta)$ is associated with the electric oscillations perpendicular to the plane of scattering (sometimes called horizontally polarized), and $i_2(\beta, n, \theta)$ is associated with the electric oscillations parallel to the plane of scattering (vertically polarized). The scattered spherical wave is composed of an infinite number of partial waves, the amplitudes of which depend on $a_l(\beta, n)$ and $b_l(\beta, n)$. In physical terms, these may be interpreted as the l^{th} electrical and magnetic multipole waves, respectively. The first set is that part of the solution for which the radial component of the magnetic vector in the incident wave is zero; in the second set the corresponding radial component of the electric vector is zero. A given partial wave can be thought of as coming from an electric or a magnetic multipole field, the first wave coming from a dipole field, the second from a quadrupole, and so on [17]. The angular functions $\pi_l(\cos\theta)$ and $\tau_l(\cos\theta)$ are, as their name implies, independent of size (β) and refractive index (n).

For a point P located at distance r from the origin of coordinates, at polar angle θ and azimuthal angle ϕ, the scattered intensities I_θ and I_ϕ are, respectively,

$$I_\theta = i_2 \left(\frac{1}{kr}\right)^2 \cos^2\phi \text{ and } I_\phi = i_1 \left(\frac{1}{kr}\right)^2 \sin^2\phi, \tag{75a, b}$$

where $i_j = \left|S_j\right|^2, j = 1, 2$ and the amplitude functions S_j are given by

$$S_1 = \sum_{l=1}^{\infty} \frac{2l+1}{l(l+1)} \left[a_l \pi_l(\cos\theta) + b_l \tau_l(\cos\theta)\right], \text{and}$$

$$S_2 = \sum_{l=1}^{\infty} \frac{2l+1}{l(l+1)} \left[a_l \tau_l(\cos\theta) + b_l \pi_l(\cos\theta)\right]. \tag{76a, b}$$

l is the order of the induced electric or magnetic multipole. The Mie angular functions $\pi_l(\cos\theta)$ and $\tau_l(\cos\theta)$ are defined in terms of the associated Legendre functions of the first kind, $P_l^1(\cos\theta)$ as

$$\pi_l(\cos\theta) = \frac{P_l^1(\cos\theta)}{\sin\theta} \text{ and } \tau_l(\cos\theta) = \frac{d}{d\theta} P_l^1(\cos\theta). \tag{77a, b}$$

The scattering coefficients a_l and b_l are defined in terms of the previously encountered Riccati-Bessel functions of the first and second kinds, respectively. a_l and b_l can be written in terms of the Riccati–Hankel function of the first kind, $\zeta_l^{(1)}(z) = zh_l^{(1)}(z) = \psi_l(z) + i\xi_l(z)$, i.e.,

$$a_l = \frac{\psi_l(\beta)\,\psi_l'(\alpha) - n\psi_l(\alpha)\,\psi_l'(\beta)}{\zeta_l^{(1)}(\beta)\,\psi_l'(\alpha) - n\psi_l(\alpha)\,\zeta_l^{(1)\prime}(\beta)} \text{and}$$

$$b_l = \frac{\psi_l(\alpha)\,\psi_l'(\beta) - n\psi_l(\beta)\,\psi_l'(\alpha)}{\zeta_l^{(1)\prime}(\beta)\,\psi_l(\alpha) - n\psi_l'(\alpha)\,\zeta_l^{(1)}(\beta)}. \tag{78a, b}$$

For future reference, the Riccati–Hankel function of the second kind is defined by $\zeta_l^{(2)}(z) = zh_l^{(2)}(z) = \psi_l(z) - i\xi_l(z)$. The dimensionless size parameters $\beta = ka$ and $\alpha = n\beta$ are again used in Eqs. (78a, b). These expressions can be simplified by the introduction of phase shift angles and result in considerable simplification if the refractive index is real [38]. In [38] it is demonstrated that the Mie formulae lead, for large values of β, to a principle for localizing rays and separating diffracted, refracted, and reflected light (in the sense of geometrical optics). The principle asserts that the term of order l in the partial-wave expansion corresponds approximately to a ray of distance $(l+1/2)/k$ from the center of the particle (this is just the impact parameter). When $\beta >> 1$, the expansions for the $S_j\,(j = 1,2)$ may be truncated at $l + 1/2 \approx \beta$ (in practice, $l_{\max} \sim \beta + 4\beta^{1/3} + 2$; see [8, 9, 39]), and the remaining sum is separated into two parts: a diffracted light field component independent of the nature of the particle and reflected and refracted rays dependent on the particle (see also [40]).

From (78a, ba, b) above, we can define the new quantities [7]

$$\begin{aligned} P_l^e &\equiv \psi_l(\beta)\,\psi_l'(\alpha) - n\psi_l(\alpha)\,\psi_l'(\beta), \\ Q_l^e &\equiv \xi_l(\beta)\,\psi_l'(\alpha) - n\psi_l(\alpha)\,\xi_l'(\beta), \\ P_l^m &\equiv \psi_l(\alpha)\,\psi_l'(\beta) - n\psi_l(\beta)\,\psi_l'(\alpha), \\ Q_l^m &\equiv \xi_l'(\beta)\,\psi_l(\alpha) - n\psi_l'(\alpha)\,\xi_l(\beta). \end{aligned} \tag{79a–d}$$

The notation of Grandy [7] is followed here (but a common alternative notation is N/D rather than P/Q). These quantities are real if n is real. Then the external coefficients (in particular) may be written as

$$a_l = \frac{P_l^e}{P_l^e + iQ_l^e}, b_l = \frac{P_l^m}{P_l^m + iQ_l^m}. \tag{80a, b}$$

Furthermore, we may define (for real n) the real phase shifts δ_l in terms of the K-matrix elements

$$\tan\delta_l^e \equiv \frac{P_l^e}{Q_l^e} \text{ and } \tan\delta_l^m \equiv \frac{P_l^m}{Q_l^m}. \tag{81a, b}$$

Hence,

$$a_l = \frac{1}{2}\left[1 - \exp\left(2i\delta_l^e\right)\right], b_l = \frac{1}{2}\left[1 - \exp\left(2i\delta_l^m\right)\right]. \tag{82a, b}$$

Also it is readily shown that

$$a_l = \frac{\left(P_l^e\right)^2}{\left(P_l^e\right)^2 + \left(Q_l^e\right)^2} - i\frac{P_l^e Q_l^e}{\left(P_l^e\right)^2 + \left(Q_l^e\right)^2}, \tag{83}$$

from which it follows that for no absorption (i.e., elastic scattering),

$$\mathrm{Re}\left(a_l\right) = \left|a_l\right|^2 = \sin^2\delta_l^e \in [0,1], \quad \text{and} \quad \mathrm{Im}\left(a_l\right) = \frac{1}{2}\sin 2\delta_l^e \in \left[-\frac{1}{2}, \frac{1}{2}\right]. \tag{84a, b}$$

A similar set of equations can be deduced for b_l. It is interesting to note that the locus of a_l and b_l in the complex δ_l-plane is a circle of radius $1/2$ with center at $(1/2, 0)$. The scalar *partial* scattering amplitudes $f_l(k)$ can be defined using Eq. (51) as

$$f_l(k) = \frac{e^{-il\pi/2}}{2ik}\left[\mathscr{S}_l(k) - 1\right], \tag{85}$$

(on reverting to the former notation for $\mathscr{S}_l(k)$), where $\mathscr{S}_l(k) = \exp\left(2i\delta_l\right)$, the vector problem can be characterized by (for real n) the unitary matrix

$$\mathscr{S}_l = \begin{pmatrix} \mathscr{S}_l^e & 0 \\ 0 & \mathscr{S}_l^m \end{pmatrix}. \tag{86}$$

If we now write

$$a_l = \frac{1}{2}\left[1 - \mathscr{S}_l^e(k)\right], b_l = \frac{1}{2}\left[1 - \mathscr{S}_l^m(k)\right]. \tag{87a, b}$$

Substitution into (82a, b) yields the expressions in terms of α and β

$$\mathscr{S}_l^e(k) = -\frac{\zeta_l^{(2)}(\beta)}{\zeta_l^{(1)}(\beta)}\left[\frac{\ln'\zeta_l^{(2)}(\beta) - n^{-1}\ln'\psi_l(\alpha)}{\ln'\zeta_l^{(1)}(\beta) - n^{-1}\ln'\psi_l(\alpha)}\right],$$

$$\mathscr{S}_l^m(k) = -\frac{\zeta_l^{(2)}(\beta)}{\zeta_l^{(1)}(\beta)}\left[\frac{\ln'\zeta_l^{(2)}(\beta) - n\ln'\psi_l(\alpha)}{\ln'\zeta_l^{(1)}(\beta) - n\ln'\psi_l(\alpha)}\right]. \tag{88a, b}$$

In these expressions, the notation $\ln' f(z) = d\left(\ln f(z)\right)/dz$ has been used. As we have seen, $\mathrm{Re}\left(a_l\right)$ reaches its maximum value (unity) when $Q_l^e = 0$ (for the *TM* modes), and similarly, a maximum occurs for $\mathrm{Re}\left(b_l\right)$ when $Q_l^m = 0$ (*TE* modes). These conditions correspond to Johnson's condition for resonance [25], and as Grandy [7] shows in some detail, they are also equivalent to the poles of the Mie coefficients a_l and b_l in the complex β-plane, which are *in turn* equivalent to the poles of the scattering matrix elements $\mathscr{S}_l^m(\lambda, \beta)$ and $\mathscr{S}_l^e(\lambda, \beta)$ in the complex λ-plane. A valuable examination of the formal analogies between Mie theory and time-independent quantum scattering by a radial potential for both transparent and absorbing "particles" has been carried out in [41].

Solutions of the radial (Debye) equation (24) are linear combinations of the Riccati–Bessel functions $\psi_l(kr)$ and $\xi_l(kr)$ which vanish at the origin and match appropriately at $r = a$, i.e.,

$$u_l^{\nu}(r) \propto \psi_l(nkr), 0 \le r \le a, \text{and}$$

$$u_l^{\nu}(r) \propto \left(\xi_l(kr) - \frac{Q_l^{\nu}}{P_l^{\nu}}\psi_l(kr)\right), r \ge a. \tag{89a, b}$$

The superscript $\nu = e$ or m refers to the electric or magnetic multipole modes, respectively. Within the barrier, the solution $u_l^{\nu}(r)$ must be exponentially increasing with r, from which we infer that $Q_l^e = 0$ for the *TM* modes and $Q_l^m = 0$ for the *TE* modes. As pointed out in [7], these conditions determining the discrete "energy levels" of a resonance are precisely the conditions mentioned above.

Conclusion

This article attempts to categorize and summarize some of the many and various connections that exist between ray theory, wave theory, and potential scattering theory. By "meandering" through these related areas in the broader field of mathematical physics, it is hoped that the reader will recognize how each of the levels of description can inform the others, resulting (it is to be hoped) in a greater appreciation for the whole. More specifically, the mechanism of rainbow formation by the scattering of light from a transparent sphere is examined from a ray-theoretic viewpoint, for both homogeneous and radially inhomogeneous spheres. By examining the complementary approach of wave scattering theory, the resulting radial equations (for scalar and vector wave equations) can be regarded as time-independent Schrödinger-like equations. Consequently it is possible to exploit some of the mathematical techniques in potential scattering theory because every refractive index profile $n(r)$ defines a (wave number-dependent) scattering potential $V(k;r)$ for the problem. This is significantly different from the case of time-independent potential scattering in quantum mechanics because it ensures that there are no bound states of the system (this result is established in Appendix 2). The close correspondence between the resonant modes in scattering by a potential of the "well-barrier" type and the behavior of electromagnetic "rays" in a transparent (or dielectric) sphere is discussed in some detail.

Acknowledgements I have been heavily influenced by the work of Professors H. M. Nussenzveig and J. A. Lock in the preparation of this chapter. I would particularly like to thank Professor Lock for his generous advice, detailed and constructive suggestions on this material (also pointing out an error in Appendix 5), and permission to use the quotation from his paper [3]. The comments of an anonymous reviewer also contributed significantly to the improvement of this chapter and are gratefully acknowledged.

Appendix 1: The Debye Series

In [8, 13]; see references therein it is shown that, in terms of cylindrical Hankel functions of the first and second kinds,

$$\mathscr{S}_l(\lambda,\beta) = \frac{H_\lambda^{(2)}(\beta)}{H_\lambda^{(1)}(\beta)} R_{22}(\lambda,\beta) + T_{21}(\lambda,\beta)\, T_{12}(\lambda,\beta) \frac{H_\lambda^{(1)}(\alpha)}{H_\lambda^{(2)}(\alpha)} \sum_{p=1}^{\infty} [\rho(\lambda,\beta)]^{p-1} \tag{A1}$$

where

$$\rho(\lambda,\beta) = R_{11}(\lambda,\beta) \frac{H_\lambda^{(1)}(\alpha)}{H_\lambda^{(2)}(\alpha)}. \tag{A2}$$

This is the *Debye expansion*, arrived at by expanding the expression $[1-\rho(\lambda,\beta)]^{-1}$ as an infinite geometric series. The quantities R_{22}, R_{11}, T_{21}, and T_{12} are, respectively, the external/internal reflection and internal/external transmission coefficients for the problem. This procedure transforms the interaction of "wave + sphere" into a series of surface interactions. In so doing it "unfolds" the stationary points of the integrand so that a given integral in the Poisson summation contains a few stationary points. This permits a ready identification of the many terms in accordance with ray theory. The first term represents direct reflection from the surface. The term $p = 1$ has one such point (the transmitted ray), whereas $p = 2$ has either two or zero stationary points (the former corresponding to the two supernumerary rays of the first-order rainbow). The pth term in the summation represents transmission into the sphere, via the term T_{21} subsequently "bouncing" back and forth between $r = a$ and $r = 0$ a total of p times with $p-1$ internal reflections at the surface (this time via the R_{11} term in ρ). The final factor in the second term, T_{12}, corresponds to transmission to the outside medium. In general, therefore, the pth term of the Debye expansion represents the effect of $p+1$ surface interactions. Now $f(\beta,\theta)$ can be expressed as

$$f(\beta,\theta) = f_0(\beta,\theta) + \sum_{p=1}^{\infty} f_p(\beta,\theta), \tag{A3}$$

where

$$f_0(\beta,\theta) = \frac{i}{\beta} \sum_{m=-\infty}^{\infty} (-1)^m \int_0^{\infty} \left(1 - \frac{H_\lambda^{(2)}(\beta)}{H_\lambda^{(1)}(\beta)} R_{22}\right) P_{\lambda-1/2}(\cos\theta) e^{2\pi i m\lambda} \lambda\, d\lambda. \tag{A4}$$

This is the direct reflection term. The expression for $f_p(\beta,\theta)$ involves a similar type of integral for $p \geq 1$. The direct transmission term is the one of interest for zero-order bows, but the analysis of Nussenveig and coworkers deals with constant n, for which no such bow exists. As noted earlier, Lock [4] identified the existence of a zero bow for a Luneberg lens with focal length exceeding its radius. In general

however, further work is necessary to determine the nature of direct transmission bows in other radially inhomogeneous transparent (or dielectric) spheres [19].

Returning to the constant n case, the application of the modified Watson transform to the third term ($p = 2$) in the Debye expansion of the scattering amplitude shows that it is *this term* which is associated with the phenomena of the primary rainbow. More generally, for a Debye term of given order p, a rainbow is characterized in the λ-plane by the occurrence of two real saddle points λ and λ' between 0 and β in some domain of scattering angles θ, corresponding to the two scattered rays on the illuminated side. As $\theta \to \theta_R^+$ (θ_R being the *rainbow angle*), the two saddle points move towards each other along the real axis (Fig. 3.7), merging together at $\theta = \theta_R$. As θ moves into the dark side, the two saddle points become complex, moving away from the real axis in complex conjugate directions. Therefore, as noted earlier, from a mathematical point of view, a rainbow can be defined as a collision of two saddle points in the complex angular momentum plane. The primary bow light/shadow transition region is thus associated physically with the confluence of a pair of geometrical rays and their transformation into "complex rays."

Appendix 2: Radially Inhomogeneous Media

In electromagnetic scattering, for radially symmetric media, the electric field vector $\boldsymbol{E}$ must satisfy the scattering boundary conditions and the vector wave equation

$$\nabla \times \nabla \times \boldsymbol{E} - k^2 n^2(r)\boldsymbol{E} = \boldsymbol{0}. \tag{A5}$$

By expanding $\boldsymbol{E}$ in terms of vector spherical harmonics, the following radial equations are obtained for the *transverse electric* (*TE*) and *transverse magnetic* (*TM*) modes, respectively [25]:

$$\frac{d^2\mathscr{S}_l(r)}{dr^2} + \left[k^2 n^2(r) - \frac{l(l+1)}{r^2}\right]\mathscr{S}_l(r) = 0; \tag{A6}$$

$$\frac{d^2 T_l(r)}{dr^2} - \frac{2n'(r)}{n(r)}\frac{dT_l(r)}{dr} + \left[k^2 n^2(r) - \frac{l(l+1)}{r^2}\right]T_l(r) = 0. \tag{A7}$$

Each of these equations can be reworked into a time-independent Schrödinger equation form, with $\psi(r)$ now being a generic-dependent variable for the two modes above. Thus,

$$\frac{d^2\psi(r)}{dr^2} + \left[k^2 - V(r) - \frac{l(l+1)}{r^2}\right]\psi(r) = 0, \tag{A8a}$$

or equivalently, as indicated earlier,

$$\frac{d^2\psi(r)}{dr^2} + \left[k^2 - V(r) - \frac{\lambda^2 - 1/4}{r^2}\right]\psi(r) = 0, \tag{A8b}$$

where $k^2 = E$ is the energy of the 'particle', $\lambda = l + 1/2$. The "scattering potential" is now

$$V(r) = k^2 \left[1 - n^2(r)\right] \tag{A9}$$

for the *TE* mode and (by eliminating the first derivative term in (A7); see (A13) below)

$$V(r) = k^2 \left[1 - n^2(r) + k^{-2} n(r) \frac{d^2}{dr^2} (n(r))^{-1}\right] \tag{A10}$$

for the *TM* mode. Thus, for scattering by a dielectric sphere, the corresponding potential has finite range. Note that for constant refractive index, these two equations are identical in form. We examine one property of the Eqs. (A8a, b) above in more detail. Although they are formally identical to the radial Schrödinger equation, there are important differences for both the scalar and the vector problems. Pure "bound-state" solutions, that is, real, regular, and square-integrable solutions, corresponding to $k^2 < 0 (\mathrm{Im}\, k > 0)$ do not in general exist in the "non-QM case." To see this, assume that $\mathscr{S}_l(r)$ is a square-integrable solution of Eq. (A6). On multiplying by $\bar{S}_l(r)$ (the complex conjugate of $\mathscr{S}_l(r)$) and integrating by parts, we obtain

$$\bar{S}_l(r)\mathscr{S}_l'(r)\big|_0^\infty - \int_0^\infty \left[\left|\mathscr{S}_l'(r)\right|^2 + \left\{\frac{l(l+1)}{r^2} - k^2 n^2(r)\right\} \left|\mathscr{S}_l(r)\right|^2 dr\right] = 0. \tag{A11}$$

The integrated term vanishes because to be square integrable, $S(r)$ must vanish at infinity, and we have noted already that near the origin, $\mathscr{S}_l(r) \sim r^{l+1}$. Hence,

$$\int_0^\infty \left[\left|\mathscr{S}_l'(r)\right|^2 + \frac{l(l+1)}{r^2} \left|\mathscr{S}_l(r)\right|^2\right] dr = \int_0^\infty k^2 n^2(r) \left|\mathscr{S}_l(r)\right|^2 dr. \tag{A12}$$

Clearly, this cannot be satisfied for $k^2 < 0$ unless $n^2(r) < 0$ in some interval or set of intervals. This actually "opens the door" for some insight into properties of "meta-materials" for which the refractive index may be pure imaginary [42]. Regarding the second of the two potentials (A10), if we write $T_l(r) = U_l(r) n(r)$, then from (A7), $U_l(r)$ satisfies the equation

$$\frac{d^2 U_l(r)}{dr^2} + \left[k^2 n^2(r) - n(r) \frac{d^2}{dr^2}\left[\frac{1}{n(r)}\right] - \frac{l(l+1)}{r^2}\right] U_l(r) = 0. \tag{A13}$$

A similar procedure to that above yields the less useful form:

$$\int_0^\infty \left[\left|U_l'(r)\right|^2 + \left\{\frac{l(l+1)}{r^2} + n(r) \frac{d^2}{dr^2}\left(\frac{1}{n(r)}\right)\right\} \left|\mathscr{S}_l(r)\right|^2\right] dr$$
$$= \int_0^\infty k^2 n^2(r) \left|U_l(r)\right|^2 dr. \tag{A14}$$

Clearly, this expression places some conditions on the concavity of $n^{-1}(r)$, but with the Liouville transformation [43] $r \mapsto s : s = \int_0^r n^2(t)\,dt$, and $U_l \mapsto W_l : W_l(s) = m(s)W_l(s)$, where $m(s) = n(r(s))$, it follows that

$$\frac{d^2U_l(r)}{dr^2} = m^2(s)\left[m(s)\frac{d^2W_l(s)}{ds^2} - W_l(s)\frac{d^2m(s)}{ds^2}\right], \tag{A15}$$

and

$$\text{and}\frac{d^2}{dr^2}\left(\frac{1}{n(r)}\right) = -m^2(s)\frac{d^2m(s)}{ds^2}. \tag{A16}$$

Therefore, Eq. (A13) simplifies to the form

$$\frac{d^2W_l(s)}{ds^2} + \left[\frac{k^2}{m^2(s)} - \frac{l(l+1)}{m^4(s)r^2(s)}\right]W_l(s) = 0. \tag{A17}$$

The transformation $r \mapsto s$ is monotonic (and linear for $r > 1$), and $s \sim r$ in the neighborhood of the origin, so the previous analysis carries over, and we can conclude that for $n^2 > 0$, no bound states are possible.

Appendix 3: Connection with Classical Scattering

In the theory of classical scattering of a nonrelativistic projectile particle of mass m by a central force with potential $V(r)$, the total deflection angle θ is given by [8, 44]

$$\theta = \pi - 2b\int_a^\infty \frac{dr}{r^2\left[1 - b^2/r^2 - V(r)/E\right]^{1/2}}, \tag{A18}$$

where b is the impact parameter, a is the distance of closest approach, and E is the particle energy. The integral can be recast to the optical case (using Eq. (9)) by setting $b = \sin i$ and

$$n(r) = \left[1 - \frac{V(r)}{E}\right]^{1/2}, \tag{A19}$$

with $V(r) < 0$ corresponding to an attractive potential with refractive index $n > 1$. This justifies the notion of a refracting sphere having the characteristics of a potential well, with implications, as we have noted, for morphology-dependent resonances.

Appendix 4: The Location of the *S*-Matrix Poles

From Eqs. (23) and (46), noting the implicit time-dependence $\exp(-i\omega t)$, we may write the asymptotic form of the solution for $\psi_l(r,t)$ as

$$\psi_l(r,t) = O\left(\frac{1}{r}\left\{e^{-ikr} - \mathscr{S}_l(k)\, e^{ikr}\right\} e^{-i\omega t}\right). \tag{A20}$$

The scattering matrix elements $\mathscr{S}_l(k)$ are given in terms of the Jost functions by Eq. (60), and since both functions $\tilde{f}_l(\pm k)$ are defined for complex values of k, (60) defines $\mathscr{S}_l(k)$ throughout the complex k-plane [37]. Using the probability conservation law (derived from the time-dependent Schrödinger equation)

$$\frac{\partial}{\partial t}\int_V |\psi|^2 dV = -\int_S \boldsymbol{j}\cdot d\boldsymbol{S}, \tag{A21}$$

where $\boldsymbol{j}$ is the probability flux density (in units for which $m = \hbar = 1$),

$$\boldsymbol{j} = \frac{i}{2}\left(\psi\nabla\bar{\psi} - \bar{\psi}\nabla\psi\right). \tag{A22}$$

The integration in (A21) is carried out on the surface of a large sphere of radius R such that the asymptotic solution (A20) may be used. Furthermore, if $\mathscr{S}_l(k)$ has a pole at the complex k-value $k = k_r + ik_i$, then the first term in (A20) may be neglected in the neighborhood of this point, and we may write

$$\psi_l(r,t) = \frac{u_l(r)}{r}e^{-i\omega t} = O\left(-\frac{\mathscr{S}_l(k)}{r}e^{i(kr-\omega t)}\right), r\to\infty \tag{A23}$$

near the pole k. From (A22) we then find that

$$k_r k_i \int_0^R u_l^2(r)\, dr = -\frac{k_r}{2}\left|\mathscr{S}_l(k)\right|^2 e^{-2k_i R} < 0. \tag{A24}$$

Therefore, it follows that either $k_r = 0$ (the poles of $\mathscr{S}_l(k)$ lie on the imaginary axis) or if $k_r \neq 0$, then $k_i < 0$ (i.e., the poles of $\mathscr{S}_l(k)$ lie in the lower half plane). Equivalently, the only poles in the upper half plane must lie on the imaginary axis.

Note that in the above discussion, we have tacitly assumed that the angular frequency can be identified with the energy of the "particle." This is justified by virtue of the famous relation $E = h\nu \propto \omega$. Without loss of generality here we make set the constant of proportionality to be unity, whence $\omega = E = k^2$, so that

$$\omega = (k_r + ik_i)^2 = \left(k_r^2 - k_i^2\right) + 2ik_r k_i \equiv E_r - \frac{i\Gamma}{2}, \Gamma = -4k_r k_i. \tag{A25}$$

Appendix 5: Poles and Resonances on the k-Plane and E-Plane

For algebraic simplicity, we consider the (simple) poles of the S-matrix for the one-dimensional scalar problem [30, 45]. In this approach, the analysis is based on a slightly different formulation of the governing time-independent "Schrödinger" equation, namely,

$$\frac{1}{2}\frac{d^2u(x)}{dx^2} + \left[k^2 - V(x)\right]u(x) = 0. \tag{A26}$$

For a square well of depth $V_0 > 0$ (i.e., $V(r) = -V_0, |x| < a/2$ and is zero elsewhere), the incident "wave" is represented by

$$u(x) = Ae^{ikx}, x < -a/2, \tag{A27}$$

and a transmitted wave

$$u(x) = Ae^{ik(x-a)}S(E), x > a/2. \tag{A28}$$

The transmission coefficient $S(E)$ is the one-dimensional scattering matrix in this problem. It can be shown that [45]

$$S(E) = \left\{\cos Ka - \frac{i}{2}\left(\frac{k}{K} + \frac{K}{k}\right)\sin Ka\right\}^{-1}, \tag{A29}$$

where now $k = \sqrt{2E}$ and $K = \sqrt{2(E+V_0)}$. Note the similarity of this expression with the denominator of the S-matrix in Eq. (36). The transmissivity of the well is defined as

$$T(E) = |S(E)|^2 = \left\{1 + \frac{V_0^2 \sin^2 Ka}{4E(E+V_0)}\right\}^{-1}. \tag{A30}$$

This expression has maxima equal to one whenever $\sin Ka = 0$, i.e., when $Ka = n\pi, n = 1, 2, 3, \ldots$ Equivalently, $E = n^2\pi^2/2a^2 - V_0 > 0$. These maxima correspond to resonances—perfect transmission—in this system. The well contains an integral number of half wavelengths when this condition is satisfied.

We examine $S(E)$ as an analytic function of the energy E in what follows. For $E > 0, 0 < T(E) \leq 1$. Therefore, poles of $T(E)$ (and $S(E)$) can only occur when $-V_0 < E < 0$. In fact $S(E)$ has a pole whenever

$$\cos Ka - \frac{i}{2}\left(\frac{k}{K} + \frac{K}{k}\right)\sin Ka = 0, \tag{A31}$$

i.e., when

$$\cot Ka = \frac{1}{2}\left(\frac{K}{k} - \frac{k}{K}\right). \tag{A32}$$

Furthermore, from the identity $2\cot 2\theta = (\cot\theta - \tan\theta)$, the solutions of (A32) can be recast in terms of odd and even parity bound-state solutions, i.e.,

$$K\cot\left(\frac{Ka}{2}\right) = ik, \quad \text{and} \quad K\tan\left(\frac{Ka}{2}\right) = -ik. \tag{A.33a, b}$$

(Again, notice the similarity of (A.33a, b) with $\alpha\cot\alpha = i\beta$ from Eq. (36).) Suppose now that a resonance occurs at $E = E_r \equiv k_r^2/2 > 0$. In the *vicinity* of such value of the resonance energy, we may expand the expression $\left(\frac{k}{K} + \frac{K}{k}\right)\tan Ka$ as

$$\left(\frac{k}{K}+\frac{K}{k}\right)\tan Ka=\frac{d}{dE}\left[\left(\frac{k}{K}+\frac{K}{k}\right)\tan Ka\right]_{E_r}(E-E_r)+O(E-E_r)^2. \quad \text{(A33)}$$

To first order in $(E-E_r)$, on simplifying, we find that

$$\left(\frac{k}{K}+\frac{K}{k}\right)\tan Ka\approx a\left[\frac{dK}{dE}\left(\frac{k}{K}+\frac{K}{k}\right)\right]_{E_r}(E-E_r)\equiv\frac{4}{\Gamma}(E-E_r). \quad \text{(A34)}$$

We can rewrite Eq. (A29) as

$$S(E)=\sec Ka\left\{1-\frac{i}{2}\left(\frac{k}{K}+\frac{K}{k}\right)\tan Ka\right\}^{-1}\approx\sec Ka\left\{1-i\frac{2}{\Gamma}(E-E_r)\right\}^{-1}$$
$$=\sec Ka\left(\frac{i\Gamma/2}{E-E_r+i\Gamma/2}\right)\approx\left(\frac{i\Gamma/2}{E-E_r+i\Gamma/2}\right). \quad \text{(A35)}$$

To this order of approximation, then, the pole of $S(E)$ lies in the fourth quadrant of the complex E-plane. There is a branch cut along the real axis, $E>0$ since if $E=|E|e^{i\theta}$, and $E^{1/2}=|E|^{1/2}e^{i\theta/2}$, in the limit $\theta\to 2\pi^-$, $\sqrt{E}=-|E|^{1/2}$, and for $E<0$, $k=i|2E|^{1/2}$. As can be seen from the term $\exp(ikx)$ in Eq. (A28), therefore, $E<0$ corresponds to a decaying transmitted wave, and (A26) then defines the conditions for the bound states to exist within the potential well. These conditions are exactly the Eqs. (A.33a, b) above.

Similarly, for the more general three-dimensional case we would expect that, near a resonance, $\mathscr{S}_l(E)$ also has a pole in the fourth quadrant. This pole is in the analytic continuation of $\mathscr{S}_l(E)$ from above to below the positive real axis and lies on the second Riemann sheet of $\mathscr{S}_l(E)$. The bound states of the well correspond to poles of $\mathscr{S}_l(E)$ on the negative real energy axis. The closer the resonances are to the real axis, the "stronger" they become, that is, the more they behave like very long-lived bound states [45].

Finally, a nice connection can be made to the phase shift from Eq. (A30). Retaining E as the independent variable, we can write

$$S(E)=e^{i\delta(E)}|T(E)|^{1/2}. \quad \text{(A36)}$$

For notational convenience, we write Eq. (A29) as $S(E)=[A(E)-iB(E)]^{-1}$, with obvious choices for A and B. Then it follows that

$$\tan\delta(E)=\frac{B(E)}{A(E)}=\frac{1}{2}\left(\frac{k}{K}+\frac{K}{k}\right)\tan Ka\approx\frac{2}{\Gamma}(E-E_r) \quad \text{(A37)}$$

on using Eq. (A34). Hence,

$$\delta(E)\approx\arctan\left[\frac{2}{\Gamma}(E-E_r)\right]. \quad \text{(A38)}$$

Note also that

$$\frac{d\delta(E)}{dE} = \frac{2\Gamma}{\Gamma^2 + 4(E - E_r)^2}. \tag{A39}$$

And this derivative has a maximum value when $E = E_r$, that is, at a resonance, so $\delta(E)$ varies rapidly there.

References

1. Sassen, K.: J. Opt. Soc. Am. **69**, 1083–1089 (1979)
2. Nussenzveig, H.M.: Sci. Am. **236**(4), 116–127 (1977)
3. Lock, J.A.: J. Opt. Soc. Am. **A5**, 2032–2044 (1988)
4. Lock, J.A.: J. Opt. Soc. Am. **A25**, 2971–2979 (2008)
5. Lock, J.A.: J. Opt. Soc. Am. **A25**, 2980–2990 (2008)
6. Lock, J.A.: J. Opt. Soc. Am. **A25**, 2991–3000 (2008)
7. Grandy, W.T., Jr.: Scattering of Waves from Large Spheres. Cambridge University Press, Cambridge (2000)
8. Nussenzveig, H.M.: Diffraction Effects in Semiclassical Scattering. Cambridge University Press, Cambridge (1992)
9. Nussenzveig, H.M.: J. Math. Phys. **10**, 82–125 (1969)
10. Nussenzveig, H.M.: J. Math. Phys. **10**, 126–178 (1969)
11. Nussenzveig, H.M.: J. Opt. Soc. Am. **69**, 1068–1079 (1979)
12. Nussenzveig, H.M.: Ann. Phys. **34**, 23–95 (1965)
13. Adam, J.A.: Phys. Reports **356**, 229–365 (2002)
14. Adam, J.A.: Not AMS **49**, 1360–1371 (2002)
15. Adam, J.A., Laven, P.: Appl. Opt. **46**, 922–929 (2007)
16. Adam, J.A.: Appl. Opt. **50**, F50–F59 (2011)
17. Born, M., Wolf, E.: Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light. Cambridge University Press, Cambridge (1999)
18. Brockman, C.L., Alexopoulos, N.G.: Appl. Opt. **16**, 166–174 (1977)
19. Adam, J.A., Pohrivchak, M., Nuntaplook, U.: to be published
20. Vetrano, M.R., van Beeck, J.P., Riethmuller, M.: Opt. Lett. **30**, 658–660 (2005)
21. Uslenghi, P.L.E.: IEEE Trans. Ant. Prop. **17**, 235–236 (1969)
22. Luneberg, R.K.: The Mathematical Theory of Optics. University of California Press, Berkeley and Los Angeles (1964)
23. Lock, J.A.: J. Opt. Soc. Am. **A20**, 499–507 (2003)
24. Leonhardt, U., Philbin, T.: Geometry and Light: The Science of Invisibility. Dover Publications, New York (2010)
25. Johnson, B.R.: J. Opt. Soc. Am. **A10**, 343–352 (1993)
26. Sanz, P., Sanudo, J., Sesma, J.: J. Math. Phys. **22**, 2594–2597 (1981)
27. Nussenzveig, H.M.: Nucl. Phys. **11**, 499–521 (1959)
28. Burke, P.G.: Potential Scattering in Atomic Physics. Plenum Press, New York & London (1977)
29. Burke, P.G., Berrington, K.A. (eds.): Atomic and Molecular Processes: an R-Matrix Approach. Taylor & Francis Group (1993)
30. Schiff, L.I.: Quantum Mechanics, 3rd edn. McGraw-Hill, New York (1968)
31. de Alfaro, V., Regge, T.: Potential Scattering. North-Holland Publishing, Amsterdam (1965)
32. Frisk, G.V., DeSanto, J.A.: J. Acoust. Soc. Am. **47**, 172–180 (1970)
33. Kronenfeld, R.: Am. J. Phys. **39**, 1056–1068 (1971)
34. Omnes, R., Froissart, M.: Mandelstam Theory and Regge Poles. W.A. Benjamin, New York (1963)

35. Newton, R.G.: The Complex j-Plane. W.A. Benjamin, New York (1964)
36. Schutzer, W., Tiomno, J.: Phys. Rev. **83**, 249–251 (1951)
37. Sitenko, A.G.: Scattering Theory. Springer, Berlin (1991)
38. van de Hulst, H.C.: Light Scattering by Small Particles. Dover, New York (1981)
39. Wang, R.T., van de Hulst, H.C.: Appl. Opt. **30**, 106–117 (1991)
40. Ungut, A., Grehan, G., Gouesbet, G.: Appl. Opt. **20**, 2911–2918 (1981)
41. Gouesbet, G.: Opt. Comm. **231**, 9–15 (2004)
42. Pendry, J.B.: Phys. Rev. Lett. **85**, 3966–3969 (2000)
43. Eftimiu, C.: J. Math. Phys. **23**, 2140–2146 (1982)
44. Newton, R.G.: Scattering Theory of Waves and Particles, 2nd edn. Springer, Berlin (1982)
45. Baym, G.: Lectures on Quantum Mechanics. W.A. Benjamin, New York (1969)

Chapter 4
Understanding the Dynamics of Collision and Near-Collision Motions in the *N*-Body Problem

Lennard F. Bakker

Introduction

For ages, humankind has observed the regular and predicable motion of the planets and other bodies in the solar system and asked, will the motion of the bodies in the solar system continue forever as they are currently observed? This philosophical question is the object of the mathematical notion of stability. A difficulty in applying the notion of stability to the motion of the solar system is that of collision and near-collision motions of bodies in the solar system. Collision and near-collision motions do occur in the solar system. Section "Phenomenon" recounts a few of these that have been observed or predicted.

The standard mathematical model for understanding the motion of planets and other bodies in the solar system is the Newtonian *N*-Body problem, presented in section "The *N*-Body Problem". Included here are some of the basic features and mathematical theory of the Newtonian *N*-Body Problem, its integrals or constants of motion, special solutions such as periodic solutions, and the notions of stability and linear stability of periodic solutions and their relationship.

The notions and basic theory of collisions and singularities in the Newtonian *N*-Body Problem is presented in section "Collisions". This includes a discussion of the probabilities of collisions, and the regularization or the lack thereof for collisions. A collision motion is rare in that is has a probability of zero of occurring, whereas a near-collision motion has a positive probability of occurring. Regularization is a mathematical technique that removes the collision singularities from the Newtonian *N*-body problem and enables an analysis of near-collision motions in terms of collision motions through the continuous dependence of motions on initial conditions. This regularization is illustrated in the collinear 2-body problem, the simplest of all the *N*-body problems.

Lennard F. Bakker (✉)
Department of Mathematics, Brigham Young University, Provo, Utah 84602, USA
e-mail: bakker@math.byu.edu

B. Toni (ed.), *Advances in Interdisciplinary Mathematical Research*, Springer Proceedings in Mathematics & Statistics 37, DOI 10.1007/978-1-4614-6345-0_4,

Recent results are presented in section "Results" on the analytic and numerical existence and numerical stability and linear stability of periodic motions with regularizable collisions in various N-body problems with $N = 3$ and $N = 4$. Although fictitious, these periodic motions with regularizable collisions provide a view of their near-collision motions which could be motions of the bodies in the N-body problem that are collision-free and bounded for all time.

Phenomenon

Collisions and near-collisions of two or more solar system bodies are apparent obstacles at which Newton's law of gravity becomes problematic. Velocities of colliding bodies become infinite at the moment of collision, while velocities of near-colliding bodies become very large as they pass by each other. Both of these situations present problems for numerical estimates of the motion of such bodies.

Although collisions are rare, historical evidence of collisions of solar system bodies is viewable on the surface of the Earth and the Moon [8]. Only recently have collisions of solar systems bodies actually been observed. As the comet Shoemaker-Levy 9 approached Jupiter it was torn apart into fragments by tidal forces. In July of 1994, at least 21 discernible fragments of Shoemaker-Levy 9 collided with Jupiter. These were the *first* ever observed collisions of solar system bodies. An animation of some of the fragments of Shoemaker-Levy 9 colliding with Jupiter can be found at www2.jpl.nasa.gov/sl9/anim.html.

Near-collision motion are less rare than collisions. As of March 2012, there are nearly 9,000 known near-Earth asteroids,[1] of which 1,306 are potentially hazardous to Earth.[2] One of these potentially hazardous asteroids, named 2012 DA14, was discovered in 2012. This asteroid will pass by Earth on February 15, 2013, coming closer to the Earth than satellites in geostationary orbit.[3] How close will 2012 DA14 pass by Earth? A mere 17,000 miles (27,000 km).[4] In cosmic terms, this close shave of 2012 DA14 with Earth in 2013 is a near-collision motion.

The N-Body Problem

To model collision and near-collision motions we make some simplifying assumptions and use Newton's inverse square law of gravity. We assume that all the bodies are idealized as particles with zero volume (i.e., as points), that no particle is torn apart by tidal forces, that the mass of each particle never changes, and that besides

[1] See http://neo.jpl.nasa.gov/stats/.

[2] See http://neo.jpl.nasa.gov/neo/groups.html.

[3] See article about 2012 DA14 posted March 6, 2012 on MSNBC.com.

[4] See article about 2012 DA14 posted March 8, 2012 on Earthsky.org.

Newton's law of gravity there are no other forces acting on the bodies. Under these assumptions we would think of Shoemaker-Levy 9 as not being torn apart by tidal forces, but as colliding with Jupiter as a whole.

The Equations

The particles modeling the bodies move in three-dimensional Euclidean space which we denote by $\mathbf{R}^3$. For a positive integer $N \geq 2$, suppose there are N particles with positions $\mathbf{q}_j \in \mathbf{R}^3$ and masses $m_j > 0$, $j = 1, \ldots, N$. The distance between two of the particles is denoted by

$$r_{jk} = |\mathbf{q}_j - \mathbf{q}_k|, \quad j \neq k,$$

which is the standard Euclidean distance between two points in $\mathbf{R}^3$. The Newtonian N-body problem is the system of second-order nonlinear differential equations

$$m_j \mathbf{q}_j'' = \sum_{k \neq j} \frac{G m_j m_k (\mathbf{q}_k - \mathbf{q}_j)}{r_{jk}^3}, \quad j = 1, \ldots, N,$$

where $' = d/dt$ for a time variable t and $G = 6.6732 \times 10^{-11}\ \mathrm{m}^2/\mathrm{s}^2\mathrm{kg}$. By an appropriate choice of units of the $\mathbf{q}_j$, we will assume that $G = 1$ because we are investigating the qualitative or geometric, rather than the quantitative, behavior of collision and near-collision motions.

By the standard existence, uniqueness, and extension theory in differential equations (see [9], for example), the initial value problem

$$m_j \mathbf{q}_j'' = \sum_{k \neq j} \frac{m_j m_k (\mathbf{q}_k - \mathbf{q}_j)}{r_{jk}^3}, \quad \mathbf{q}_j(t_0) = \mathbf{q}_j^0, \quad \mathbf{q}_j'(t_0) = \mathbf{q}_j'^0 \tag{1}$$

has a unique solution

$$\mathbf{q}(t) = (\mathbf{q}_1(t), \ldots, \mathbf{q}_N(t))$$

defined on a maximal interval of definition (t^-, t^+) as long as $r_{jk} \neq 0$ for all $j \neq k$ at $t = t_0$. Such a solution $\mathbf{q}(t)$ describes a motion of the N particles.

Not every initial value problem (1) will have a solution $\mathbf{q}(t)$ with $t^- = -\infty$ and $t^+ = \infty$. A solution with either $t^- > -\infty$ or $t^+ < \infty$ experiences a singularity at the finite endpoint of its maximal interval of definition. The notion of a singularity is addressed in section "Singularities".

Integrals

An integral of motion of the Newtonian N-body problem is a differentiable function F of the position $\mathbf{q}$ and/or the velocity $\mathbf{q}'$ and/or the masses $\mathbf{m} = (m_1, \ldots, m_N)$ such that

$$\frac{d}{dt}F(\mathbf{q}(t),\mathbf{q}'(t),\mathbf{m}) = 0,\ \ t \in (t^-,t^+).$$

Along a solution $\mathbf{q}(t)$, an integral F of motion satisfies

$$F(\mathbf{q}(t),\mathbf{q}'(t),\mathbf{m}) = F(\mathbf{q}(t_0),\mathbf{q}'(t_0),\mathbf{m}),\ \ t^- < t < t^+,$$

i.e., the value of F is constant along the solution. The Newtonian N-body problem has ten known integrals of motion.

The translation invariance of the equations of the Newtonian N-body problem gives rise to 6 integrals of motion. With $M = \sum_{j=1}^{N} m_j$, three of these are given by the components of the center of mass vector

$$\mathbf{C} = \frac{1}{M}\sum_{j=1}^{m} m_j\mathbf{q}_j,$$

and three more are given by the components of the linear momentum vector

$$\mathbf{L} = \frac{1}{M}\sum_{j=1}^{N} m_j\mathbf{q}'_j.$$

Typically, both of these are set to $\mathbf{0}$ so that the *relative motion* of the N particles is emphasized.

The rotational symmetry of the equations of the Newtonian N-body problem gives rise to 3 more integrals of motion. These integrals are given by the components of the angular momentum vector

$$\mathbf{A} = \sum_{j=1}^{N} m_j\mathbf{q}_j \times \mathbf{q}'_j.$$

The angular momentum plays a key role in understanding collisions in the N-body problem, as we will see later on.

There is one more integral of motion of the Newtonian N-body problem. The *self-potential* (or negative of the potential energy) is

$$U = \sum_{j<k} \frac{m_j m_k}{r_{ik}}.$$

The *kinetic energy* is

$$K = \frac{1}{2}\sum_{j=1}^{N} m_j\mathbf{q}'_j \cdot \mathbf{q}'_j.$$

The *total energy*

$$H = K - U$$

is an integral of motion for the Newtonian N-body problem.

In the late 1800s, the mathematical strategy for "solving" the Newtonian N-body problem was to find enough "independent" integrals of motion [25]. This would implicitly give each solution as the curve of intersection of the hypersurfaces corresponding to the integrals of motion. Each solution $\mathbf{q}(t)$ is a curve in $\mathbf{R}^{6N}$. However, the intersection of the hypersurfaces of the ten integrals of motion gives a $6N-10>1$ dimension hypersurface in $\mathbf{R}^{6N}$, which is not a curve! The ten known integrals of motion are independent of each other (one is not a function of the others) and are algebraic functions of positions, velocities, and masses. Are there any more algebraic integrals of motion? This was answered a long time ago in 1887–1888 by Bruns [7].

Theorem 1. *There are no algebraic integrals of motion independent of the ten known integrals of motion.*

Consequently, new integrals of motion, if any, cannot be algebraic! In 1893, Newcomb [20] lamented that no additional integrals had been found to enable the implicit solution of the 3-body problem. It is well-known that the Newtonian 2-body problem can be solved implicitly,[5] but all attempts to solve the N-body problem with $N \geq 3$ have been futile.[6]

Typically then the solution $\mathbf{q}(t)$ of the initial value problem (1) is estimated numerically. From the constant total energy H along a solution $\mathbf{q}(t)$, we observe that if any of the distances r_{jk} get close to 0, i.e., at least two of the particles are near collision, the self-potential becomes large, and the kinetic energy becomes large too. The latter implies that the velocity of at least one of the particles becomes large, and the linear momentum $\mathbf{L}$ along $\mathbf{q}(t)$ implies that the velocity of at least two particles becomes large. In particular, from the equations of the Newtonian N-body problem, the particles that are near collision are the one with the large velocities. These large velocities present problems for the numerical estimates of such a solution.

Special Solutions

Rather than solving the N-body problem for all of its solutions by finding enough independent integrals of motion, it is better to examine special solutions with particular features. The simplest solutions to find are equilibrium solutions, where the position $\mathbf{q}_j(t)$ of each particle is constant for all time. But the Newtonian N-body problem has none of these (see p. 29 in [18]). The next simplest solutions are periodic solutions, i.e., there exist $T>0$ such that $\mathbf{q}(t+T)=\mathbf{q}(t)$ for all $t \in \mathbf{R}$. These are part of the larger collection of solutions $\mathbf{q}(t)$ with $t^- = -\infty$ and $t^+ = \infty$ that are bounded. Such solutions must have a particular total energy (see p. 160 in [25]).

[5] See en.wikipedia.org/wiki/Gravitational_two-body_problem.

[6] Karl Sundman did solve the 3-Body Problem when $\mathbf{A} \neq 0$ by convergent power series defined for all time, but the series converge too slowly to be of any theoretic or numerical use [25].

Theorem 2. *If a solution* $\mathbf{q}(t)$ *of the Newtonian N-body problem exists for all time and is bounded, then the total energy* $H < 0$.

Consequently, any periodic solution $\mathbf{q}(t)$ of the Newtonian N-body problem must have negative total energy. This is why in the search for periodic solutions, the total energy is always assigned a negative value.

Stability

A periodic solution $\mathbf{q}(t)$ of the Newtonian N-body problem gives a predictable future: we know with certainty what the positions of the N particles will be at any time $t > 0$. But what if our measurements of the initial conditions $\mathbf{q}(0)$ and $\mathbf{q}'(0)$ are slightly off? A solution $\tilde{\mathbf{q}}(t)$ with initial conditions near $\mathbf{q}(0)$ and $\mathbf{q}'(0)$ will stay close to $\mathbf{q}(t)$ for a time, by a property of solutions of initial value problems called continuity of solutions with respect to initial conditions (see [9]). But if it stays close for all $t > 0$, we think of $\mathbf{q}(t)$ as being stable.

To *quantify* this notion of stability for a periodic solution, we use a Poincaré section which is a hyperplane S containing the point $(\mathbf{q}(0), \mathbf{q}'(0))$ that is transverse to the curve $(\mathbf{q}(t), \mathbf{q}'(t))$. If $\mathbf{x} = (\tilde{\mathbf{q}}(0), \tilde{\mathbf{q}}'(0))$ is a point on S near the $(\mathbf{q}(0), \mathbf{q}'(0))$, then $P(\mathbf{x})$ is the next point where the curve $(\tilde{\mathbf{q}}(t), \tilde{\mathbf{q}}'(t))$ intersects S,[7] and $P^2(\mathbf{x})$ is the next point, and so on. The initial condition $\mathbf{x}^0 = (\mathbf{q}(0), \mathbf{q}'(0))$ is a *fixed point* of this Poincaré map P from S to S, i.e., $P(\mathbf{x}^0) = \mathbf{x}^0$.

Definition 1. The periodic solution $\mathbf{q}(t)$ is stable if for every real $\varepsilon > 0$, there exist a real $\delta > 0$ such that $|P^k(\mathbf{x}) - \mathbf{x}^0| < \varepsilon$ for all $k = 1, 2, 3, \ldots$, whenever $|\mathbf{x} - \mathbf{x}^0| < \delta$.

When $\mathbf{q}(t)$ is not stable, there are solutions which start nearby but eventually move away from $\mathbf{q}(t)$, and we say that $\mathbf{q}(t)$ is *unstable*.

Showing directly that $\mathbf{q}(t)$ is stable or unstable is very difficult. Instead, the related concept of linearized stability is investigated, at least numerically. The derivative of the Poincaré map at the fixed point $\mathbf{x}^0$ is a square matrix $DP(\mathbf{x}^0)$.

Definition 2. A periodic solution $\mathbf{q}(t)$ is:

1. Spectrally stable[8] if all the eigenvalues of $DP(\mathbf{x}^0)$ have modulus one
2. Linearly unstable if any eigenvalue of $DP(\mathbf{x}^0)$ has modulus bigger than one

In 1907, Liapunov [15] established a connection between the stability of Definition 1 and the linearized stability of Definition 2.

[7] For an illustration of this, see en.wikipedia.org/wiki/Poincaré_map.

[8] There is a more restrictive notion of spectral stability known as linear stability that requires additional technical conditions on the square matrix $DP(\mathbf{x}^0)$.

Theorem 3. *If a periodic solution* $\mathbf{q}(t)$ *is stable, then it is spectrally stable, and if* $\mathbf{q}(t)$ *is linearly unstable, then it is unstable.*

If a periodic solution is shown numerically to be linearly unstable, then by Theorem 3, the periodic solution is unstable. On the other hand, if a periodic solution is shown numerically to be spectrally stable, it may be stable or unstable. Examples exist with spectrally stable fixed points of maps like P that are unstable (see [28]).

The notion of stability for a nonperiodic solution, such as the motion of the sun and planets in the solar system, is harder to grasp. Here is a sampling of the history and opinions on this stability problem. In 1891, Poincaré commented that the stability of the solar system had at that time already preoccupied much time and attention of researchers (see p. 147 in [10]). In 1971, Siegel and Moser lamented that a resolution of the stability problem for the N-body problem would probably be in the distant future (see p. 219 in [28]). In 1978, Moser noted that the answer to the stability of the solar system was still not known (see p. 127 in [10]). In 2005, Saari stated that a still unresolved problem for the N-body problem is that of stability (see p. 132 in [25]). Meyer, Hall, and Offin commented how little is known about the stability problem and how difficult it was to get (see p. 229 in [18]).

In 1996, Diacu and Holmes suggested that the solar system should be considered stable (in a weak sense) if no collisions occur among the sun and the planets, and no planet ever escape from the solar system (see. p.129 in [10]). In this weak sense of stability, the solar system is stable for the next few billion years according to numerical work of Hayes [11] in 2007. Much longer-term numerical studies of the solar system by Batygin and Laughlin [6] in 2008 using small changes in the initial conditions suggest that Mercury could fall into the sun in 1.261 Gyr[9] or that Mercury and Venus could collide in 862 Myr[10] and Mars could escape from the solar system in 822 Myr. The Newtonian N-body problem thus suggests that in the near future, the Solar System should be free of collisions of planets and the Sun, with no planets escaping the solar system. But this still leaves open the possibility that smaller objects, such as asteroids and comets, could collide with any of the planets in the short and long term. Recall that there are nearly 9,000 of those near-Earth asteroids to consider, with 2012 DA14 making its near-collision approach with Earth on February 15 of 2013.

Collisions

Either in the short term or the long term, collisions put a *wrench* into the question of any notion of stability. Why should a solution or any nearby solution of the Newtonian N-body problem be defined for all time? Remember that Shoemaker-Levy 9 has $t^+ < \infty$!

[9] Gyr means giga-year or 1,000,000,000 years.

[10] Myr means mega-year or 1,000,000 years.

Singularities

Collisions are one of the *two* kinds of singularities in the Newtonian N-body problem. The solution $\mathbf{q}(t)$ of initial value problem (1) is real analytic (i.e., a convergent power series) on an interval $(t_0-\delta, t_0+\delta)$ for some $\delta > 0$, as long as $r_{jk} \neq 0$ for all $j \neq k$ at t_0. By a process called analytic continuation (see, e.g., [16]), the interval $(t_0-\delta, t_0+\delta)$ can be extended to the maximal interval (t^-, t^+).

Definition 3. A *singularity* of the Newtonian N-body problem is a time $t = t^+$ or t^- when $t^+ < \infty$ or $t^- > -\infty$.

In 1897, Painlevé [22] characterized a singularity of the Newtonian N-body problem, using the quantity

$$r_{\min}(t) = \min_{j \neq k} r_{jk}(t)$$

determined by a solution $\mathbf{q}(t)$.

Theorem 4. *A singularity for the Newtonian N-body problem occurs at time $t = t^*$ if and only if $r_{\min}(t) \to 0$ as $t \to t^*$.*

An understanding of what this means is obtained by considering the *collision set*

$$\Delta = \bigcup_{j \neq k} \{\mathbf{q} : \mathbf{q}_j = \mathbf{q}_k\} \subset (\mathbf{R}^3)^N,$$

which is the set of points where two or more of the N-particles occupy the same position. Painlevé's characterization means that $\mathbf{q}(t)$ approaches the collison set, i.e.,

$$\mathbf{q}(t) \to \Delta \ \text{ as } \ t \to t^*$$

when t^* is a singularity of the Newtonian N-body problem. Painlevé's characterization introduces two classes of singularities.

Definition 4. A singularity t^* of the Newtonian N-body problem is a collision singularity when $q(t)$ approaches a specific point of Δ as $t \to t^*$. Otherwise the singularity t^* is a non-collision singularity.

Only collision singularities can occur in the Newtonian 2-body problem because it can be implicitly solved. In 1897, Painlevé [22] showed that only one other Newtonian N-body problem has only collision singularities.

Theorem 5. *In the 3-body problem, all singularities are collision singularities.*

Unable to extend his result to more than 3 bodies, Painlevé conjectured that there exist non-collision singularities in the Newtonian 4 or larger body problem. In 1992, Xia [34] mostly confirmed Painlevé's conjecture, giving an example in the Newtonian 5-body problem.

Theorem 6. *There exist non-collision singularities in the N-body problem for $N \geq 5$.*

That leaves unresolved the question of the existence of non-collision singularities in the Newtonian 4-body problem.

An understanding of what a non-collision singular looks like is obtained by considering one-half of the *polar moment of inertia* of the Newtonian N-body problem:

$$I = \frac{1}{2}\sum_{j=1}^{N} m_j \mathbf{q}_j \cdot \mathbf{q}_j.$$

This scalar quantity measures the "diameter" of the N particles in the Newtonian N-body problem. In 1908, von Zeipel [37] characterized a collision singularity in terms of the polar moment of inertia.

Theorem 7. *A singularity of the Newtonian N-body problem at $t = t^*$ is a collision if and only if I is bounded as $t \to t^*$.*

This implies that for a non-collision singularity, at least one of the N-particles has to achieve an infinite distance from the origin in just a finite time. This is a rather strange thing for Newton's law of gravity to predict. On the other hand, by Theorem 7, for a collision singularity, all of the positions of the N particles remain bounded at the moment of the singularity.

A total collapse is an example of a collision singularity in the N-body problem for which all N particles collide at the same point at the singularity t^*. For a solution $\mathbf{q}(t)$, the quantity

$$r_{\max} = \max_{j \neq k} r_{jk}(t)$$

characterizes a total collapse: a total collapse occurs at t^* if and only if

$$r_{\max}(t) \to 0 \text{ as } t \to t^*.$$

There is a relationship between total collapse and the angular momentum that was known by Weierstrass and established by Sundman (see [25]).

Theorem 8. *If $\mathbf{A} \neq 0$, then $r_{\max}(t)$ is bounded away from zero.*

This does not preclude the collision of less than N particles when $\mathbf{A} \neq 0$, as will be illustrated for certain Newtonian N-body problems in section "Results".

Improbability

Recall that there are 1,306 potentially hazardous near-Earth asteroids. What are the chances that Earth will be hit by a near-Earth asteroid or Jupiter will be hit by another comet? Well, it depends on the arrangement of the particles.

Definition 5. A solution $\mathbf{q}(t)$ is called collinear if the N particles always move on the same fixed line in $\mathbf{R}^3$. Otherwise it is called noncollinear.

Every collinear solution has zero angular momentum because $\mathbf{q}_j(t)$ is parallel with $\mathbf{q}_j'(t)$ for all $t \in (t^-, t^+)$. In 1971 and 1973, Saari [23, 24] established the probability of collisions.

Theorem 9. *The probability that a noncollinear solution* $\mathbf{q}(t)$ *will have a collision is zero. Every collinear solution* $\mathbf{q}(t)$ *has a collision.*

With collision singularities being rare for a noncollinear N-body problem, why bother to study them? Diacu and Holmes (see p. 84 and p. 103 in [10]) argue for the study of collision singularities because without such a study, a complete understanding of the Newtonian N-body problem could not be achieved. In particular, solutions near collision singularities could behave strangely, and the probability of a solution coming close to a collision singularity is positive and thus cannot be ignored. Understanding then the collision singularities enables an understanding of the near-collision solutions.

Regularization

Regularization is one method by which we can get an understanding of a collision singularity. To *regularize* a collision means to extend the solution beyond the collision through an elastic bounce without loss or gain of total energy in such a way that all of the solutions nearby have continuity with respect to initial conditions, i.e., they look like the extended collision solution for a time (see p. 104 and p. 107 in [10]). Regularization is typically done by a Levi–Civita-type change of the dependent variables and a Sundman-type change of the independent variable (see [8]), that together *removes* the collision singularity from the equations. We illustrate this regularization in the simplest of the N-body problems.

In the collinear 2-body problem (or Col2BP for short), the positions of the two particles are the scalar quantities q_1 and q_2. If $x = q_2 - q_1$ is the distance between the particle with mass m_1 at q_1 and the particle with mass m_2 at $q_2 > q_1$, then the Col2BP takes the form

$$x'' = -\frac{m_1 + m_2}{x^2},\ x > 0, \tag{2}$$

and the total energy takes the form

$$H = \frac{m_1 m_2}{2(m_1 + m_2)}(x')^2 - \frac{m_1 m_2}{x}. \tag{3}$$

As $x \to 0$ the two particles approach collision, and the total energy implies that the two particles collide with an infinite velocity,

$$(x')^2 \to \infty.$$

To regularize the *binary collision* (or total collapse) in this problem, define a new independent variable s and a new dependent variable w by

$$\frac{ds}{dt} = \frac{1}{x}, \quad w^2 = x,$$

where the former is the Sundman-type change of the independent variable, and the latter is the Levi–Civita-type change of the dependent variable. If $\dot{} = d/ds$, the second-order equation (2) becomes

$$w^2\left[2w\ddot{w} - 2\dot{w}^2 + (m_1 + m_2)\right] = 0, \tag{4}$$

and the total energy (3) becomes

$$Hw^2 = \frac{2m_1m_2}{m_1 + m_2}\dot{w}^2 - m_1m_2. \tag{5}$$

As $w \to 0$, the second-order equation (4) makes sense (no dividing by zero), and the total energy (5) implies that

$$(\dot{w})^2 \to \frac{m_1 + m_2}{2},$$

which is a finite nonzero velocity! The collision singularity has been regularized.

The regularized nonlinear second-order equation (4) can actually be solved! Solving the total energy (5) for $2(\dot{w})^2$ and substituting this into the second-order equation (4) gives

$$2w^3\left[\ddot{w} - \frac{(m_1 + m_2)H}{2m_1m_2}w\right] = 0. \tag{6}$$

This makes sense when $w = 0$, i.e., the moment of collision! For negative H, the linear second-order equation[11] inside the square brackets in (6) solves to give a real analytic stable periodic solution $w(s)$ which experiences a collision every half period in terms of the regularized time variable s. The corresponding solution $x(t)$ is periodic and experiences a collision once a period in terms of the original time variable t. This doubling of the number of collisions per period is because the change of dependent variable $w^2 = x$ has $w(s)$ "doubling" $x(t)$ in that $w(s)$ passes through 0 twice a period, going from positive to negative and then negative to positive, while $x(t)$ is positive except at collision where it is zero.

The binary collision singularity in the Newtonian 2-body problem can be regularized in a similar but more complicated way than what was done above for the Col2BP (see [25]). By Theorem 8, a solution of the 2-body problem with nonzero angular momentum does not experience a collision or total collapse. A nonzero angular momentum near-collision solution looks like the zero angular momentum collision solution.[12] The regularized 2-body problem provides good numerical estimates of the motion because there are no infinite velocities!

[11] This is a simple harmonic oscillator for $H < 0$ whose solutions are in terms of cosine and sine.

[12] Binary star systems are known to exist in the Universe. The Newtonian 2-body problem predicts stability for a binary star system, a collision-free solution that is bounded for all time.

McGehee

What about regularization of a triple collision, when three of the particles meet? In 1974, McGehee [17] showed that regularization of a triple collision is in general not possible.[13] Starting close together, two solutions that approach a near-triple collision can describe *radically different* motions after the near-triple collision. This kind of behavior is known as "sensitive dependence on initial conditions" and is an antithesis of stability. Triple collisions present a numerical nightmare! By extension, collisions with four or more particles present the same nightmare! So the only regularizable collisions are those that are essentially a binary collision.

Results

Spectrally stable periodic solutions have been found in Newtonian N-body problems with regularizable collisions for $N \geq 3$. Three of these situations discussed here are the collinear 3-body problem (or Col3BP), the collinear symmetric 4-body problem (or ColS4BP), and the planar pairwise symmetric 4-body problem (or PPS4BP). There are other Newtonian N-body problems where periodic solutions with regularizable collisions whose existence has been given analytically [27, 35, 36], some of whose stability (in the sense of Definition 1) and linear stability (as defined in Definition 2) has been numerically determined [5, 33, 35, 36].

Col3BP

As a subproblem of the Newtonian 3-body problem, the Col3BP requires that the three particles always lie on the same line through the origin. The positions of the three particles in the Col3BP are the scalars q_1, q_2, and q_3 which can be assumed to satisfy

$$q_1 \leq q_2 \leq q_3.$$

By Theorem 9, collisions always occur in the Col3BP. Because the three particles are collinear for all time, their angular is zero, and by Theorem 8 a total collapse is possible[14] in the Col3BP. In 1974, S.J. Aareth and Zare [1] showed that any two of the three possible binary collisions in the 3-body problem are regularizable.[15]

[13] This is achieved by "blowing-up" the triple collision singularity and slowing down the motion as the particles approach a triple collision. This setting does allow for good numerical estimates of near-triple collisions.

[14] Initial conditions leading to total collapse in the equal mass Col3BP are easy to realize: set $q_1 = -1$, $q_2 = 0$, and $q_3 = 1$ with the initial velocity of each particle set to 0.

[15] A good numerical model for the Sun–Jupiter–Shoemaker-Levy 9 or Earth–Moon-2012DA14 situation is regularized 3-body problem of Aarseth and Zare.

In 1993, Hietarinta and Mikkola [13] used Aarseth and Zare's regularization [1] to regularize the binary collisions $q_1 = q_2$ and $q_2 = q_3$ in the Col3BP.

In 1956, Schubart [26] numerically found a periodic orbit in the equal mass Col3BP of negative total energy in which the inner particle oscillates between binary collisions with the outer particles. In 1977, Hénon [12] numerically extended Schubart's periodic solution to arbitrary masses and investigated their linear stability. In 1993, Hietarinta and Mikkola [13] also numerically investigated the linear stability of Schubart's periodic solution for arbitrary masses. Together they showed that Schubart's periodic solution is spectrally stable for certain masses and linearly unstable for the remaining masses. Hietarinta and Mikkola [13] further numerically investigated the Poincaré section for Schubart's periodic solution for arbitrary masses, showing when there is stability as described in Definition 1. In 2008, Moeckel [19] and Venturelli [32] separately proved the analytic existence of Schubart's solution when $m_1 = m_3$ and m_2 is arbitrary. Only recently, in 2011, did Shibayama [27] analytically prove the existence of Schubart's periodic solution for arbitrary masses in the Col3BP.

Schubart's periodic solution for the Col3BP is also a periodic solution of the 3-body problem, where in the latter the continuity with respect to initial conditions can be seen for near-collision solutions. For example, Schubart's periodic solution for the nearly equal masses

$$m_1 = 0.333333,\ m_2 = 0.333334,\ m_3 = 0.333333$$

is spectrally stable. Considered in 3-body problem, Schubart's periodic solution for these mass values remains spectrally stable [12], and numerically the near-collision solutions in the Newtonian 3-body problem behave like Schubart's periodic solution. It is therefore possible that in the 3-body problem, there are solutions near Schubart's periodic solution that are free of collisions and bounded for all time. Imagine, as did Hénon [12], of Newton's law of gravity predicting a triple star system that is free of collisions and bounded for all time!

ColS4BP

As a subproblem of the Newtonian 4-body problem, the ColS4BP requires that the four particles always lie on the same line through the origin. The positions of the four particles are the scalars q_1, q_2, q_3, and q_4 that satisfy

$$q_4 = -q_1,\ q_3 = -q_2,\ q_1 \geq 0,\ q_2 \geq 0$$

and

$$-q_1 \leq -q_2 \leq 0 \leq q_2 \leq q_1$$

with masses

$$m_1 = 1,\ m_2 = m > 0,\ m_3 = m,\ m_4 = 1.$$

The angular momentum for all solutions of the ColS4BP is zero because of the collinearity, and so by Theorem 8 a total collapse is possible. There are two kinds of non-total collapse collisions in the ColS4BP: the binary collision of the inner pair of particles of mass m each, i.e., $q_2 = 0$, and the *simultaneous* binary collision of the two outer pairs of particles, i.e., $q_1 = q_2 > 0$. In 2002 and 2006, Sweatman [30, 31] showed, by adapting the regularization of Aarseth and Zare [1], that these non-total collapse collisions in the ColS4BP are regularizable.

Sweatman [30, 31] numerically found a Schubart-like periodic solution in the ColS4BP with negative total energy for arbitrary m where the outer pairs collide in a simultaneous binary collision at one moment and then the inner pair collides at another moment. He determined numerically that this Schubart-like periodic solution is spectrally stable when

$$0 < m < 2.83 \text{ and } m > 35.4$$

and is otherwise linearly unstable. In 2010, Bakker et al. [2] verified Sweatman's linear stability for the Schubart-like periodic solution using a different technique. In 2011–2012, Ouyang and Yan [21], Shibayama [27], and Huang [14] proved separately the analytic existence of the Schubart-like periodic solution in the ColS4BP.

PPS4BP

The PPS4BP has two particles of mass 1 located at the planar locations

$$\mathbf{q}_1 \text{ and } \mathbf{q}_3 = -\mathbf{q}_1,$$

and two particles of mass $0 < m \leq 1$ located at the planar locations

$$\mathbf{q}_2 \text{ and } \mathbf{q}_4 = -\mathbf{q}_2.$$

The four particles in the PPS4BP need not be collinear, so that the angular momentum need not be zero. Unlike the ColS4BP, total collapse can be avoided in the PPS4BP by Theorem 8 when the angular momentum is not zero. Like the ColS4BP, there are two kinds of non-total collapse collisions in the PPS4BP: simultaneous binary collisions when $\mathbf{q}_1 = \mathbf{q}_2$ and $\mathbf{q}_3 = \mathbf{q}_4$ or when $\mathbf{q}_1 = \mathbf{q}_4$ and $\mathbf{q}_2 = \mathbf{q}_3$ and binary collisions when $\mathbf{q}_1 = 0$ or when $\mathbf{q}_2 = 0$. In 2010, Sivasankaran, Steves, and Sweatman [29] showed that these non-total collapse collisions in the PPS4BP are regularizable.

The Schubart-like periodic solution in the ColS4BP is also a periodic solution of the PPS4BP, where in the latter the continuity with respect to initial conditions can be observed for near-collision solutions. However, as shown by Sweatman [31], in the PPS4BP the Schubart-like periodic solution of the ColS4BP becomes linearly unstable for

$$0 < m < 0.406 \text{ and } 0.569 < m < 1.02$$

as well as $2.83 < m < 35.4$, while it remains spectrally stable for

$$0.407 < m < 0.567 \text{ and } m > 35.4.$$

By long-term numerical integrations for the Schubart-like periodic solution as a solution of the PPS4BP, Sweatman [31] showed that stability in the sense of Definition 1 is possible when $0.407 < m < 0.567$ and when $m > 35.4$. It is therefore possible for these values of m that near Schubart's periodic solution, there are collision-free solutions of the PPS4BP that are bounded for all time.

In 2011, adapting the regularization of Aarseth and Zare [1] to simultaneous binary collisions, Bakker, Ouyang, Yan, and Simmons [3] proved the analytic existence of a noncollinear periodic solution in the equal mass PPS4BP. This periodic solution has zero angular momentum, negative total energy, and alternates between a simultaneous binary collision of the symmetric pairs in the first and third quadrant where $\mathbf{q}_1 = \mathbf{q}_2$ and $\mathbf{q}_3 = \mathbf{q}_4$ and the simultaneous binary collision of the symmetric pairs in the second and fourth quadrants where $\mathbf{q}_1 = \mathbf{q}_4$ and $\mathbf{q}_2 = \mathbf{q}_3$. Bakker, Ouyang, Yan, and Simmons [3] then numerically extended this noncollinear periodic simultaneous binary collision solution to unequal masses $0 < m < 1$. In 2012, Bakker, Mancuso, and Simmons [4] have numerically determined that the noncollinear periodic simultaneous binary collision solution is spectrally stable when

$$0.199 < m < 0.264 \text{ and } 0.538 < m \leq 1$$

and is linearly unstable for the remaining values of m. Long-term numerical integrations of the regularized equations done by Bakker, Ouyang, Yan, and Simmons [3] suggest instability when $0.199 < m < 0.264$ and stability when $0.538 < m \leq 1$ in the sense of Definition 1. For these latter values of m could the near-collision solutions in the PPS4BP that look like the noncollinear periodic simultaneous binary collision solution be collision-free and bounded for all time?

Future Work

Both the ColS4BP and the PPS4BP are subproblems of the Newtonian 4-body problem, where the non-total collapse collisions in the former two problems are regularizable. What is not known is how to, if possible, regularize binary collisions and simultaneous binary collisions in the Newtonian 4-body problem within one coordinate system.[16] If such a regularization is possible, then all of the periodic solutions

[16] During the special session on Celestial Mechanics at the American Mathematical Society's Sectional Conference in April 2011 at the College of the Holy Cross, Worcester, Massachusetts, Rick Moeckel put forth the problem of finding an elegant coordinate system for the Newtonian 4-body problem in which regularizes binary collisions and simultaneous binary collisions and blows up all triple collisions and total collapse. The regularization of binary collisions and simultaneous binary collisions can be achieved within multiple coordinate systems, with one coordinate system for each regularizable collision.

thus known in the ColS4BP and PPS4BP would also be periodic solutions of the Newtonian 4-body problem, and the investigation of their stability and linear stability in the Newtonian 4-body problem could begin. With more possible perturbations of initial conditions in the Newtonian 4-body problem as compared with the PPS4BP, a loss of spectral stability could indeed happen as it did with going from the ColS4BP to the PPS4BP. But some of the spectral stability might survive passage from the PPS4BP to the Newtonian 4-body problem, giving the possibility of near-collision solutions that are collision-free and bounded for all time.

Acknowledgements The author expresses appreciation for the referee's comments and feedback that improved the quality of this paper. The author also expresses thanks to the organizers of the year-long seminar series held at Virginia State University.

References

1. Aarseth, S.J., Zare, K.: A regularization of the three-body problem. Cel. Mech. **10**, pp. 185–205 (1974)
2. Bakker, L.F., Ouyang, T., Yan, D., Simmons, S.C., Roberts, G.E.: Linear stability for some symmetric periodic simultaneous binary collision orbits in the four-body problem. Celest. Mech. Dynam. Astron. **108**, pp. 147–164 (2010)
3. Bakker, L.F., Ouyang, T., Yan, D., Simmons, S.C.: Existence and stability of symmetric periodic simultaneous binary collision orbits in the planar pairwise symmetric four-body problem. Celest. Mech. Dynam. Astron. **110**, pp. 271–290 (2011)
4. Bakker, L.F., Mancuso, S.C., Simmons, S.C.: Linear stability analysis of symmetric periodic simultaneously binary collision orbits in the planar pairwise symmetric four-body problem. J. Math. Anal. Appl. **392**, pp. 136–147 (2012)
5. Bakker, L.F., Simmons, S.C.: Stability of the rhomboidal symmetric-mass orbit. Submitted to J. Math. Anal. Appl (2012), http://arxiv.org/pdf/1208.3183.pdf
6. Batygin, K., Laughlin, G.: On the dynamical stability of the solar system. Astrophys. J. **683**, pp. 1207–1216 (2008)
7. Bruns, H.: Über die integrale des vielkörper-problems. Acta Math. **11**, pp. 25–96 (1887–1888)
8. Celletti, A.: Singularities, collisions and regularization theory. In: Benest, D., Froeschlé, C. (eds.) Singularities in Gravitational Systems, vol. 590, Lecture Notes in Physics. Springer, New York, pp. 1–24 (2002)
9. Chicone, C.: Ordinary Differential Equations with Applications, Texts in Applied Mathematics, vol. 34. Springer, New York (1999)
10. Diacu. F., Holmes, P.: Celestial Encounters: The Origin of Chaos and Stability. Princeton University Press, Princeton (1996)
11. Hayes, W.: Is the Outer System Chaotic? http://arxiv.org/abs/astro-ph/0702179v1
12. Hénon, M.: Stability of interplay orbits. Cel. Mech. **15**, pp. 243–261 (1977)
13. Hietarinta, J., Mikkola, S.: Chaos in the one-dimensional gravitational three-body problem. Chaos **3**, pp. 183–203 (1993)
14. Huang, H.-Y.: Schubart-like orbits in the Newtonian collinear four-body problem: a variational proof. Dis. Con. Dyn. Sys. **32**, pp. 1763–1774 (2012)
15. Liapunov, A.: Probéme général de la stabilité du mouvement. Ann. Fac. Sci. Toulouse **9**, pp. 203–474 (1907)
16. Marsden, G.E., Hoffman, M.J.: Basic Complex Analysis, 2nd edn. W.H. Freeman and Company, New York (1987)
17. McGehee, R.: Triple collision in the collinear three-body problem. Invent. Math. **27**, pp. 191–227 (1974)

18. Meyer, K.R., Hall, D.R., Offin, D.: Introduction to Hamiltonian Dynamical Systems and the *N*-Body Problem, Second Edition, Springer, New York (2009)
19. Moeckel, R.: A topological existence proof for the schubart orbits in the collinear three-body problem. Dis. Con. Dyn. Syst. Ser. B **10**, pp. 609–620 (2008)
20. Newcomb, S.: Modern mathematical thought. Bull. New York Math. Soc. **4**, pp. 95–107 (1893)
21. Ouyang, T., Yan, D.: Periodic solutions with alternating singularities in the collinear four-body problem. Celest. Mech. Dynam. Astron. **109**, pp. 229–239 (2011)
22. Painlevé, P.: Lecons Sur la Théorie Analytic de Equations Différentielles. Herman, Paris (1897)
23. Saari, D.G.: Improbability of collisions in Newtonian gravitational systems. Trans. Am. Math. Soc. **162**, pp. 267–271 (1971)
24. Saari, D.G.: Improbability of collisions in Newtonian gravitational systems II. Trans. Am. Math. Soc. **181**, pp. 351–368 (1973)
25. Saari, D.G.: Collisions, Rings, and Other Newtonian *N*-Body Problems, CBMS vol. **104**. American Mathematical Society, Providence, Rhode Island (2005)
26. Schubart, J.: Numerische aufsuchung periodischer Lösungen im Dreikörperproblem. Astron. Nachr. **283**, pp. 17–22 (1956)
27. Shibayama, M.: Minimizing periodic orbits with regularizable collisions in the *n*-body problem. Arch. Rational Mech. Anal. **199**, pp. 821–841 (2011)
28. Siegel, C.L., Moser, J.K.: Lectures on Celestial Mechanics. Springer, New York (1971)
29. Sivasankaran, A., Steves, B.A., Sweatman, W.L.: A global regularisation for integrating the Caledonian symmetric four-body problem. Celestial Mech. Dyn. Astron. **107**, pp. 157–168 (2010)
30. Sweatman, W.L.: Symmetrical one-dimensional four-body problem. Celestial Mech. Dyn. Astron. **82**, pp. 179–201 (2002)
31. Sweatman, W.L.: A family of symmetrical schubart-like interplay orbits and their stability in the one-dimensional four-body problem. Celestial Mech. Dyn. Astron. **94**, pp. 37–65 (2006)
32. Venturelli, A.: A variational proof of the existence of von Schubart's orbit. Discrete Contin. Dyn. Syst. Ser. B **10**, pp. 699–717 (2008)
33. Waldvogel, J.: The rhomboidal symmetric four-body problem. Celestial Mech. Dyn. Astron. **113**, pp. 113–123 (2012)
34. Xia, Z.: The existence of noncollision singularities in Newtonian systems. Ann. Math. **135**, pp. 411–468 (1992)
35. Yan, D.: Existence and linear stability of the rhomboidal periodic orbit in the planar equal mass four-body problem. J. Math. Anal. Appl. **388**, pp. 942–951 (2012)
36. Yan, D.: Existence of the Broucke orbit and its linear stability. J. Math. Anal. Appl. **389**, pp. 656–664 (2012)
37. von Zeipel, E.H.: Sur les singularités du problème des n corps. Ark. Mat. Astron. Pys. **4**, pp. 1–4 (1908)

Chapter 5
Absolute Stability and Conditional Stability in General Delayed Differential Equations

Junping Shi

Introduction

Delay differential equations are a class of mathematical models describing various natural and engineered phenomena with delayed feedbacks in the system. Mathematical theory of delay differential equations or functional-differential equations have been developed in the second half of twentieth century to study mathematical questions from models of population biology, biochemical reactions, neural conduction, and other applications [4, 6, 10, 17, 20].

A basic delay differential equation was proposed by renowned biologist George Evelyn Hutchinson in 1948 (see [8]):

$$\frac{du(t)}{dt} = ru(t)\left(1 - u(t-\tau)\right), \tag{1}$$

where $u(t)$ is the population as a function of time t, r is growth rate per capita parameter, and the system carrying capacity is assumed to be rescaled to 1. When $\tau = 0$, the Eq. (1) is reduced to the classical logistic equation, and it is well-known that the equilibrium $u = 1$ is globally asymptotically stable for all positive initial values. On the other hand, when τ is larger, then $u = 1$ becomes unstable, and there exists a periodic orbit of (1) which attracts all positive initial values except $u = 1$. To illustrate the cause of instability, we linearize the Eq. (1) at $u = 1$ to obtain

$$v'(t) = -rv(t-\tau). \tag{2}$$

Junping Shi (✉)
Department of Mathematics, College of William and Mary,
Williamsburg, Virginia, 23187–8795, USA
e-mail: jxshix@wm.edu

B. Toni (ed.), *Advances in Interdisciplinary Mathematical Research*, Springer Proceedings in Mathematics & Statistics 37, DOI 10.1007/978-1-4614-6345-0_5,

If an exponential function $v(t) = \exp(\lambda t)$ is a solution of (2), then the exponent λ satisfies a characteristic equation in form

$$\lambda + re^{-\lambda\tau} = 0. \tag{3}$$

While the exponent λ in the Eq. (3) cannot be explicitly solved, one can observe that $\lambda = 0$ is not a root of (3), and also the root λ of (3) varies continuously with respect to parameters r and τ. Since when $\tau = 0$, the only root of (3) is $\lambda = -r < 0$, then (3) can only have a root with positive real part if $\lambda = \omega i$ is a root of (3) for some (r, τ). Thus one can assume the "neutral stability" condition $\lambda = \omega i$ for some $\omega > 0$ (as $\lambda = -\omega i$ is also a root), which implies

$$\omega i + re^{-\omega\tau i} = 0$$

and

$$\cos(\omega\tau) = 0, \ \ r\sin(\omega\tau) = \omega. \tag{4}$$

Solving (4) we obtain that only when

$$\tau_n = \frac{(2n+1)\pi}{2r}, \ \ n \in \mathbb{N} \cup \{0\}, \tag{5}$$

the neutral stability condition holds with $\omega = r$. This simple example demonstrates that an equilibrium in delay differential equation can lose the stability with a larger delay value $\tau > 0$. In this case, we call the equilibrium $u = 1$ conditionally stable for the delay differential equation (1).

In general, for a delay differential equation with k different delays and variable $x \in \mathbb{R}^n$:

$$\dot{x}(t) = f(x(t), x(t-\tau_1), \cdots, x(t-\tau_k)), \tag{6}$$

A steady state $x = x_*$ of system (6) is said to be *absolutely stable* (i.e., asymptotically stable independent of the delays) if it is locally asymptotically stable for all delays $\tau_j \geq 0$ $(1 \leq j \leq k)$, and $x = x_*$ is said to be *conditionally stable* (i.e., asymptotically stable depending on the delays) if it is locally asymptotically stable for τ_j $(1 \leq j \leq k)$ in some intervals, but not necessarily for all delays (see [13]).

A variation of (1) can demonstrate the absolute stability of an equilibrium. Consider

$$\frac{du}{dt} = ru(t)[1 - au(t) - bu(t-\tau)]. \tag{7}$$

Here a and b represent the portions of instantaneous and delayed dependence of the growth rate on the population, respectively, and we assume that $a, b \in (0,1)$ and $a + b = 1$ (see [14]). Then $u_* = 1$ is an equilibrium. Following [14], we use the same procedure as above, then the linearized equation is now:

$$v'(t) = -arv(t) - brv(t-\tau), \tag{8}$$

and the characteristic equation becomes

$$\lambda + ar + bre^{-\lambda\tau} = 0. \tag{9}$$

By substituting the neutral stability condition $\lambda = \omega i$ into (9) and separating the real and imaginary parts, we obtain

$$\cos(\omega\tau) = -\frac{a}{b}, \ \sin(\omega\tau) = \frac{\omega}{br}. \tag{10}$$

If $a < b$, then one can find that the neutral stability condition $\lambda = \omega i$ can be achieved when $\tau = \tau_n$ as defined by

$$\tau_n = \frac{1}{r\sqrt{b^2 - a^2}} \left(\arccos\left(-\frac{a}{b}\right) + 2n\pi \right), \tag{11}$$

with

$$\omega = r\sqrt{b^2 - a^2}.$$

In this case, similar to (1), the equilibrium $u_* = 1$ is conditionally stable. However, if $a \geq b$, then the neutral stability condition cannot be achieved for any $\tau > 0$; hence it is absolutely stable, that is, the equilibrium $u_* = 1$ is locally asymptotically stable for any $\tau \geq 0$. Indeed one can prove that $u_* = 1$ is globally asymptotically stable by using a Lyapunov function argument (see [9, 11, 14]).

Biologically the phenomenon described above has the following meaning: if the instantaneous feedback of the population dominates the delayed feedback, then the system has a globally asymptotically stable equilibrium; but if the delayed feedback is more dominant, then the equilibrium is conditionally stable, and it loses the stability for a larger value of delay. It is the aim of this notes to show that this phenomenon occurs for a wider class of delayed differential equations, including some systems from biology or physics. Some recent results by the author and his collaborators in this direction will be reviewed in section "Main Results", while the proof of these results can be found in references given below. In section "Concluding Remarks" some concluding remarks and open questions will be given.

Main Results

Scalar Equations

First we state a result for scalar equation which generalizes the example of instantaneous and delayed feedback given in the Introduction. Consider a general delayed differential equation:

$$\frac{du}{dt} = f(u(t), u(t-\tau)). \tag{12}$$

Here $f = f(u, w)$ is a smooth function, and we assume that $u = u_*$ is an equilibrium. Then the linearization of (12) at $u = u_*$ is

$$v'(t) = f_u(u_*, u_*)v(t) + f_w(u_*, u_*)v(t-\tau), \tag{13}$$

where $f_u(u_*, u_*)$ and $f_w(u_*, u_*)$ are the partial derivatives of f with respect to the variables u and w, respectively. In the following when there is no confusion, we will simply write f_u and f_w, with the understanding of evaluation at (u_*, u_*). Then the corresponding characteristic equation is

$$\lambda - f_u - f_w e^{-\lambda\tau} = 0. \tag{14}$$

We assume that when $\tau = 0$, the equilibrium $u = u_*$ is stable; hence the following condition is satisfied:

$$f_u(u_*, u_*) + f_w(u_*, u_*) < 0. \tag{15}$$

Substituting the neutral stability condition $\lambda = \omega i$ into (14), we get

$$\cos(\omega\tau) = -\frac{f_u}{f_w}, \ \sin(\omega\tau) = -\frac{\omega}{f_w}. \tag{16}$$

Squaring each equation in (16) and taking the sum, we obtain

$$\omega^2 = f_w^2 - f_u^2. \tag{17}$$

By using the well-known stability result, we obtain the following general criterion.

Theorem 1. *Suppose that $u = u_*$ is an equilibrium of (12), and (15) is satisfied.*

1. *If $|f_u(u_*, u_*)| \geq |f_w(u_*, u_*)|$ (or equivalently $f_u(u_*, u_*) \leq f_w(u_*, u_*)$), then the neutral stability condition cannot be achieved for any $\tau \geq 0$. Hence u_* is absolutely stable.*
2. *If $|f_u(u_*, u_*)| < |f_w(u_*, u_*)|$ (or equivalently $f_u(u_*, u_*) > f_w(u_*, u_*)$), then $u = u_*$ is locally asymptotically stable when $0 \leq \tau < \tau_0$, and it is unstable when $\tau > \tau_0$, where*

$$\tau_0 = \frac{1}{\sqrt{f_w^2(u_*, u_*) - f_u^2(u_*, u_*)}} \arccos\left(-\frac{f_u(u_*, u_*)}{f_w(u_*, u_*)}\right). \tag{18}$$

Moreover the characteristic equation (14) has a pair of purely imaginary root $\lambda = \pm\omega i$ for $\omega > 0$ if and only if $|f_u(u_*, u_*)| < |f_w(u_*, u_*)|$, $\tau = \tau_n$ which is defined by

$$\tau_n = \frac{1}{\sqrt{f_w^2(u_*, u_*) - f_u^2(u_*, u_*)}} \arccos\left(-\frac{f_u(u_*, u_*)}{f_w(u_*, u_*)} + 2n\pi\right) \tag{19}$$

for $n \in \mathbb{N} \cup \{0\}$, and

$$\omega = \sqrt{f_w^2(u_*, u_*) - f_u^2(u_*, u_*)}. \tag{20}$$

An obvious example of Theorem 1 is the instantaneous and delayed feedback given in the Introduction in which $f_u < 0$ and $f_w < 0$. Note that Theorem 1 can also be applied to the case (i) $f_u > 0$ and $f_w < 0$ (conditionally stable), and (ii) $f_u < 0$ and $f_w > 0$ (absolutely stable).

Planar Systems with One Transcendental Term

It is common that in a system of differential equations, there are delayed feedbacks on one of the variables. A general form of such equations can be written as

$$\begin{cases} u_t = f(u,v,u_\tau), & t>0, \\ v_t = g(u,v,u_\tau), & t>0, \\ u(t) = \phi_1(t), & t\in[-\tau,0], \\ v(0) = \phi_2, \end{cases} \tag{21}$$

where $u = u(t)$, $v = v(t)$, and $u_\tau = u(t-\tau)$. The functions $f(u,v,w)$ and $g(u,v,w)$ are continuously differentiable in $\mathbb{R}^3$. We assume that there exist $u^*, v^* \in \mathbb{R}$ such that

$$f(u^*,v^*,u^*) = 0, \quad g(u^*,v^*,u^*) = 0.$$

Then (u^*,v^*) is a constant equilibrium of system (21). Linearizing system (21) at (u^*,v^*), we obtain

$$\frac{d}{dt}\begin{pmatrix}\phi\\ \psi\end{pmatrix} = \begin{pmatrix} f_u & f_v \\ g_u & g_v\end{pmatrix}\begin{pmatrix}\phi\\ \psi\end{pmatrix} + \begin{pmatrix} f_w\phi(t-\tau)\\ g_w\phi(t-\tau)\end{pmatrix}. \tag{22}$$

And the characteristic equation can be derived from

$$Det\begin{pmatrix} \lambda - f_u - f_w e^{-\lambda\tau} & -f_v \\ -g_u - g_w e^{-\lambda\tau} & \lambda - g_v\end{pmatrix} = 0, \tag{23}$$

and it is in a form

$$\lambda^2 + a\lambda + b + (c\lambda + d)e^{-\lambda\tau} = 0, \tag{24}$$

where

$$a = -(f_u + g_v),\ \ b = f_u g_v - f_v g_u,\ \ c = -f_w,\ \ and\ \ d = f_w g_v - f_v g_w. \tag{25}$$

Note that if we define

$$L_1 = \begin{pmatrix} f_u & f_v \\ g_u & g_v\end{pmatrix}, \quad L_2 = \begin{pmatrix} f_w & 0 \\ g_w & 0\end{pmatrix}, \tag{26}$$

then

$$\begin{aligned} &a = -Tr(L_1),\ \ b = Det(L_1),\ \ c = -Tr(L_2), \\ &b+d = Det(L_1+L_2),\ \ b-d = Det(L_1-L_2). \end{aligned} \tag{27}$$

Similar to before, we substitute the neutral stability condition $\lambda = \omega i$ into (24), then we obtain

$$-\omega^2 + a\omega i + b + (c\omega i + d)e^{-i\omega\tau} = 0, \tag{28}$$

or equivalently

$$\begin{aligned} -d\cos(\omega\tau) + c\omega\sin(\omega\tau) &= b - \omega^2, \\ -c\omega\cos(\omega\tau) - d\sin(\omega\tau) &= a\omega. \end{aligned} \tag{29}$$

Squaring each equation in (29) and taking the sum, we obtain an equation of ω^2 in the form

$$\omega^4 - (c^2 - a^2 + 2b)\omega^2 + (b^2 - d^2) = 0. \tag{30}$$

The existence of a positive root ω^2 to (30) determines the stability of equilibrium (u_*, v_*). Again, we assume that the equilibrium (u_*, v_*) is stable when $\tau = 0$; hence the following condition is satisfied (partial derivatives are evaluated at (u_*, v_*, u_*)):

$$\begin{aligned} a + c &= -Tr(L_1 + L_2) = -(f_u + f_w + g_v) > 0, \\ b + d &= Det(L_1 + L_2) = (f_u + f_w)g_v - (g_u + g_w)f_v > 0. \end{aligned} \tag{31}$$

We first state a result for the characteristic equation (24), which was proved in Ruan [13] (see also references therein for earlier results).

Theorem 2. *Suppose that* $a, b, c, d \in \mathbb{R}$ *satisfy*

$$a + c > 0, \;\; b + d > 0. \tag{32}$$

1. *If (i)* $c^2 - a^2 + 2b < 0$ *and* $b - d > 0$ *or (ii)* $(c^2 - a^2 + 2b)^2 - 4(b^2 - d^2) < 0$ *is satisfied, then all roots of* (24) *have negative real parts for any* $\tau \geq 0$.
2. *If (iii)* $b - d < 0$ *or (iv)* $c^2 - a^2 + 2b > 0$, $b - d > 0$, *and* $(c^2 - a^2 + 2b)^2 - 4(b^2 - d^2) \geq 0$ *are satisfied, then* (24) *has purely imaginary roots* $\pm\omega i$ *if and only if* (30) *has a positive root* ω_+ *or* ω_- *where*

$$\omega = \omega_\pm = \sqrt{\frac{c^2 - a^2 + 2b \pm \sqrt{(c^2 - a^2 + 2b)^2 - 4(b^2 - d^2)}}{2}}, \tag{33}$$

and for $n \in \mathbb{N} \cup \{0\}$,

$$\tau = \tau_n = \frac{1}{\omega_\pm}\left(\arccos\left(\frac{(d - ac)\omega_\pm^2 - bd}{d^2 + c^2\omega_\pm^2}\right) + 2n\pi\right). \tag{34}$$

Applying Theorem 2 to the stability of equilibrium of (21), we have

Theorem 3. *Suppose that* $(u, v) = (u_*, v_*)$ *is an equilibrium of* (21), *and* (31) *is satisfied.*

1. *If the matrices* L_1 *and* L_2 *satisfy either*

   (i) $Det(L_1) + Tr(L_1 + L_2)Tr(L_1 - L_2) < 0$ *and* $Det(L_1 - L_2) > 0$ *or*
   (ii) $[Det(L_1) + Tr(L_1 + L_2)Tr(L_1 - L_2)]^2 - 4Det(L_1 + L_2)Det(L_1 + L_2) < 0$,

   then the neutral stability condition cannot be achieved for any $\tau \geq 0$. *Hence* (u_*, v_*) *is absolutely stable.*

2. If the matrices L_1 and L_2 satisfy either

(iii) $Det(L_1 - L_2) < 0$ *or*
(iv) $Det(L_1 - L_2) > 0$ *and*
$Det(L_1) + Tr(L_1 + L_2)Tr(L_1 - L_2) > 2\sqrt{Det(L_1 + L_2)Det(L_1 + L_2)}$,

then there exists $\tau_0 > 0$ *such that* (u_*, v_*) *is locally asymptotically stable when* $0 \le \tau < \tau_0$, *and it is unstable when* $\tau > \tau_0$.

In the second case of Theorem 3, the critical value τ_0 and $(\omega_\pm, \tau_n)$ can all be calculated from the formulas in Theorem 2, and we omit the long formulas due to their long expression. We remark that results in Theorem 3 again demonstrate the phenomenon that if the instantaneous feedback dominates the delayed one, then the equilibrium is absolutely stable; but if the delayed feedback is more dominant, then it is conditionally stable. This is best seen in the scenario (iii) in Theorem 3 as $Det(L_1 - L_2)$ provides a measure of the difference of the two feedbacks.

There are numerous examples from applications in which the above absolute or conditional stability can be determined. Here we show two examples. First one is a Rosenzweig–MacArthur predator–prey model with a delay effect (see [2]):

$$\begin{cases} u'(t) = u(t)\left(1 - \dfrac{u(t)}{k}\right) - \dfrac{mu(t)v(t)}{u(t)+1}, & t > 0, \\ v'(t) = -rv(t) + \dfrac{mu(t-\tau)v(t)}{u(t-\tau)+1}, & t > 0, \\ u(t) = u_0(t) \ge 0, v(t) = v_0(t) \ge 0, & t \in [-\tau, 0], \end{cases} \tag{35}$$

From well-known results (see, e.g., [21]), (35) has a unique positive equilibrium (β, v_β) where $\beta = \dfrac{r}{m-r}$ and $v_\beta = \dfrac{(K-\beta)(1+\beta)}{Km}$. To consider the stability of (β, v_β), we find that

$$L_1 = \begin{pmatrix} \dfrac{\beta(k-1-2\beta)}{k(1+\beta)} & -r \\ 0 & 0 \end{pmatrix}, \; L_2 = \begin{pmatrix} 0 & 0 \\ \dfrac{(k-\beta)}{k(\beta+1)} & 0 \end{pmatrix}. \tag{36}$$

Hence the characteristic equation is in form (24) with

$$a = -\frac{\beta(k-1-2\beta)}{k(1+\beta)}, \; b = c = 0, \; d = \frac{r(k-\beta)}{k(\beta+1)}. \tag{37}$$

Then for $(k-1)/2 < \beta < k$, $a + c > 0$ and $b + d > 0$; hence (31) is satisfied; it is also obvious that $b < d$. Therefore case (iii) in Theorem 3 occurs, and the coexistence equilibrium (β, v_β) is conditionally stable.

The second example is a Leslie–Gower predator–prey system with delay effect [3].

$$\begin{cases} u'(t) = u(t)(p - \alpha u(t) - \beta v(t-\tau_1)), & t > 0, \\ v'(t) = \mu v(t)\left(1 - \dfrac{v(t)}{u(t-\tau_2)}\right), & t > 0, \\ u(t) = u_0(t) \geq 0, & t \in [-\tau_2, 0]. \\ v(t) = v_0(t) \geq 0, & t \in [-\tau_1, 0]. \end{cases} \tag{38}$$

A unique positive equilibrium of (38) is $(u_*, v_*) = \left(\dfrac{p}{\alpha+\beta}, \dfrac{p}{\alpha+\beta}\right)$. Note that (38) is not in the form of (21), but the characteristic equation is still (24) with

$$a = \frac{\alpha p}{\alpha+\beta} + \mu, \ b = \frac{\mu\alpha p}{\alpha+\beta}, \ c = 0, \ d = \frac{\mu\beta p}{\alpha+\beta}, \quad \text{and} \quad \tau = \tau_1 + \tau_2. \tag{39}$$

Thus $a + c > 0$ and $b + d > 0$; hence (31) is satisfied. If $\alpha > \beta$, then

$$\begin{aligned} b - d &= \frac{\mu(\alpha-\beta)p}{\alpha+\beta} > 0, \\ c^2 - a^2 + 2b &= -\left(\frac{\alpha p}{\alpha+\beta}\right)^2 - \mu^2 < 0. \end{aligned} \tag{40}$$

Thus case (i) in Theorem 2 is applicable, and (u_*, v_*) is absolutely stable. Indeed, in [3], it is proved that (u_*, v_*) is globally asymptotically stable for any $\tau_1 \geq 0$, $\tau_2 \geq 0$. On the other hand, if $\alpha < \beta$, then $b - d < 0$, then again case (iii) in Theorem 2 is applicable. Hence there exists $\tau_0 > 0$ such that (u_*, v_*) is stable for $\tau_1 + \tau_2 < \tau_0$, and it is unstable for $\tau_1 + \tau_2 > \tau_0$.

We remark that for many planar systems not in the form of (21), one can still use Theorem 2 to consider the stability of equilibrium in such systems, as long as the characteristic equation is still (24). For example, planar systems in form of

$$\begin{cases} \dot{u}(t) = f(u(t), v(t-\tau_1)), \\ \dot{v}(t) = g(u(t-\tau_2), v(t)), \end{cases} \tag{41}$$

and planar systems in form of

$$\begin{cases} \dot{u}(t) = f(u(t), u(t)) \pm k_1 g(u(t-\tau), v(t-\tau)), \\ \dot{v}(t) = h(u(t), v(t)) \pm k_2 g(u(t-\tau), v(t-\tau)). \end{cases} \tag{42}$$

Notice that Eq. (41) includes the case of Kolmogorov-type predator–prey systems with two delays [13, 15], and Eq. (42) includes the cases of competitive, mutualistic, and predator–prey models with symmetric delayed interaction terms. Yet another example is the second-order delayed feedback system in form

$$u'' + au' + bu = F(u(t-\tau)). \tag{43}$$

A discussion of this system by comparing the delayed feedback and the instantaneous one is given in Smith [17, Sect. 6.4].

General Planar Systems with One Delay

For a general planar system

$$\begin{cases} \dot{x}(t) = f(x(t), y(t), x(t-\tau), y(t-\tau)), \\ \dot{y}(t) = g(x(t), y(t), x(t-\tau), y(t-\tau)), \end{cases} \tag{44}$$

the corresponding characteristic equation is in a form

$$\lambda^2 + a\lambda + b + (c\lambda + d)e^{-\lambda\tau} + he^{-2\lambda\tau} = 0. \tag{45}$$

Here $\tau > 0$ and $a, b, c, d, h \in \mathbb{R}$. Notice that (45) has an additional transcendental term $he^{-2\lambda\tau}$ compared with (24). The characteristic equation (45) was considered recently in [1]. Here we will briefly describe the results in [1] and refer all the details and proofs to [1].

If $\pm i\omega$, $(\omega > 0)$, is a pair of roots of (45), then we have

$$-\omega^2 + a\omega i + b + (c\omega i + d)e^{-i\omega\tau} + he^{-2i\omega\tau} = 0. \tag{46}$$

If $\frac{\omega\tau}{2} \neq \frac{\pi}{2} + j\pi$, $j \in \mathbb{Z}$, then let $\theta = \tan\frac{\omega\tau}{2}$, and we have $e^{-i\omega\tau} = \frac{1 - i\theta}{1 + i\theta}$. Separating the real and imaginary parts, we obtain that θ satisfies

$$\begin{cases} (\omega^2 - b + d - h)\theta^2 - 2a\omega\theta = \omega^2 - b - d - h, \\ (c\omega - a\omega)\theta^2 + (-2\omega^2 + 2b - 2h)\theta = -(a\omega + c\omega). \end{cases} \tag{47}$$

Denote

$$M_1 = \begin{pmatrix} \omega^2 - b + d - h & -2a\omega \\ (c-a)\omega & -2\omega^2 + 2b - 2h \end{pmatrix},$$

$$M_2 = \begin{pmatrix} \omega^2 - b - d - h & -2a\omega \\ -(c+a)\omega & -2\omega^2 + 2b - 2h \end{pmatrix},$$

and

$$M_3 = \begin{pmatrix} \omega^2 - b + d - h & \omega^2 - b - d - h \\ (c-a)\omega & -(c+a)\omega \end{pmatrix}.$$

And define

$$D(\omega) = det(M_1), \ E(\omega) = det(M_2), \quad \text{and} \quad F(\omega) = det(M_3). \tag{48}$$

If $D(\omega) \neq 0$, then we can solve from (47) that

$$\theta^2 = \frac{E(\omega)}{D(\omega)}, \quad \theta = \frac{F(\omega)}{D(\omega)}, \tag{49}$$

and from Eq. (49), we have that ω satisfies

$$D(\omega)E(\omega) = F(\omega)^2, \tag{50}$$

which is a polynomial equation for ω with degree 8:

$$\omega^8 + s_1\omega^6 + s_2\omega^4 + s_3\omega^2 + s_4 = 0, \tag{51}$$

where

$$\begin{aligned} s_1 &= 2a^2 - 4b - c^2, \\ s_2 &= 6b^2 - 2h^2 - 4ba^2 - d^2 + a^4 - a^2c^2 + 2c^2b + 2hc^2, \\ s_3 &= 2d^2b - a^2d^2 - 4b^3 + 2b^2a^2 - c^2b^2 - 2bc^2h \\ &\quad + 4acdh - 2d^2h + 4bh^2 - 2h^2a^2 - c^2h^2, \\ s_4 &= b^4 - d^2b^2 - 2b^2h^2 + 2bd^2h - d^2h^2 + h^4 = (b-h)^2[-d^2 + (b+h)^2], \end{aligned} \tag{52}$$

and ω^2 is a positive root of

$$z^4 + s_1z^3 + s_2z^2 + s_3z + s_4 = 0. \tag{53}$$

The following lemma gives the algorithm of solving the critical delay values for purely imaginary roots of (45).

Lemma 4. *If (53) has a positive root ω_N^2 ($\omega_N > 0$) and $D(\omega_N) \neq 0$, then Eq. (47) has a unique real root $\theta_N = \dfrac{F(\omega_N)}{D(\omega_N)}$ when $\omega = \omega_N$. Hence Eq. (45) has a pair of purely imaginary roots $\pm i\omega_N$ when*

$$\tau = \tau_N^j = \frac{2\arctan\theta_N + 2j\pi}{\omega_N}, \quad j \in \mathbb{Z}. \tag{54}$$

The nondegeneracy condition $D(\omega_N) \neq 0$ can be verified in certain situations, and the tranversality condition for the roots moving across the imaginary axis can also be formulated for (45). The analysis of the quartic polynomial (53) is more complicated than (30), and a complete solution would be cumbersome to present. In [1], several different ways of solving the characteristic equation were presented. Here we only state one of them:

Theorem 5. *Suppose that $a, b, c, d, h \in \mathbb{R}$ satisfy*

(i) $c \neq 0$ *and* $h \neq 0$.
(ii) $b \neq h$ *and* $d^2 > (b+h)^2$.
(iii) $b + h \leq \dfrac{ad}{c}$ *or* $\left(\dfrac{d}{c}\left(2h - \dfrac{ad}{c}\right) - a\left(b + h - \dfrac{ad}{c}\right)\right) \cdot (a - c) \neq 0$.

Recall that $D(\omega)$ and $F(\omega)$ are defined as in (48). Then

1. *The quartic equation (53) has a positive root ω_N^2 for some $\omega_N > 0$ satisfying $D(\omega_N) \neq 0$.*
2. *Let*

$$\theta_N = \frac{F(\omega_N)}{D(\omega_N)} \text{ and } \tau = \tau_N^j = \frac{2\arctan\theta_N + 2j\pi}{\omega_N},$$

where $j \in \mathbb{Z}$. Then the characteristic equation (45) has a pair of purely imaginary eigenvalues $\pm i\omega_N$ when $\tau = \tau_N^j$.

Moreover if $a, b, c, d, h \in \mathbb{R}$ also satisfy

(iv) $a + c > 0$ and $b + d + h > 0$,

then there exists $\tau_* > 0$ such that when $\tau \in [0, \tau_*)$, all the roots of Eq. (45) have negative real parts; if a nondegeneracy condition holds, then when $\tau = \tau_*$, all the roots of Eq. (45) have nonpositive real parts, but Eq. (45) has at least one pair of simple purely imaginary roots $\pm i\omega_0$, and for $\tau \in (\tau_*, \tau_* + \varepsilon)$ with some small $\varepsilon > 0$, Eq. (45) has at least one pair of conjugate complex roots with positive real parts.

We refer to [1] for the detail of the nondegeneracy condition. We remark that all the conditions (i), (ii), and (iii) except $d^2 > (b+h)^2$ hold for all parameter values except a zero measure set. Combining with the condition (iv), we have the following observation for the appearance of roots of Eq. (45) with positive real parts for $\tau > 0$.

Corollary 6. *Define a subset in the parameter space*

$$P = \{(a, b, c, d, h) \in \mathbb{R}^5 : a + c > 0,\ b + d + h > 0,\ b - d + h < 0\}. \tag{55}$$

Then for almost every $(a, b, c, d, h) \in P$, there exists $\tau_* > 0$ such that when $\tau \in [0, \tau_*)$, all the roots of Eq. (45) have negative real parts; when $\tau = \tau_*$, Eq. (45) has at least one pair of simple purely imaginary roots $\pm i\omega_*$, and for $\tau \in (\tau_*, \tau_* + \varepsilon)$ with some small $\varepsilon > 0$, Eq. (45) has at least one pair of conjugate complex roots with positive real parts.

Now we apply these results to (44). We assume that the functions $f(u, v, w, z)$ and $g(u, v, w, z)$ are continuously differentiable in $\mathbb{R}^4$, and there exist $u^*, v^* \in \mathbb{R}$ such that

$$f(u^*, v^*, u^*, v^*) = 0, \quad g(u^*, v^*, u^*, v^*) = 0.$$

The linearized equation is

$$\frac{d}{dt}\begin{pmatrix} \phi(t) \\ \psi(t) \end{pmatrix} = L_1 \begin{pmatrix} \phi(t) \\ \psi(t) \end{pmatrix} + L_2 \begin{pmatrix} \phi(t-\tau) \\ \psi(t-\tau) \end{pmatrix}, \tag{56}$$

where

$$L_1 = \begin{pmatrix} f_u & f_v \\ g_u & g_v \end{pmatrix}, \quad L_2 = \begin{pmatrix} f_w & f_z \\ g_w & g_z \end{pmatrix}, \tag{57}$$

Then the characteristic equation at (u_*, v_*) is

$$Det\begin{pmatrix} \lambda - f_u - f_w e^{-\lambda\tau} & -f_v - f_z e^{-\lambda\tau} \\ -g_u - g_w e^{-\lambda\tau} & \lambda - g_v - g_z e^{-\lambda\tau} \end{pmatrix} = 0, \tag{58}$$

which becomes (45) with

$$\begin{aligned} a &= -(f_u + g_v),\ b = f_u g_v - f_v g_u,\ c = -(f_w + g_z), \\ d &= (f_u g_z - f_z g_u) + (f_w g_v - f_v g_w),\ h = f_w g_z - f_z g_w, \end{aligned} \tag{59}$$

or equivalently

$$\begin{aligned} &a = -Tr(L_1),\ b = Det(L_1),\ c = -Tr(L_2), \\ &d = \frac{1}{2}\left[Det(L_1 + L_2) - Det(L_1 - L_2)\right],\ h = Det(L_2). \end{aligned} \tag{60}$$

Then we can state a general delay-induced instability result based on Theorem 5:

Theorem 7. *Suppose that $f, g \in C^1(\mathbb{R}^4)$, and (u^*, v^*) is an equilibrium of (44). Let L_1 and L_2 be the Jacobian matrices defined as in (57). Assume that*

$$Tr(L_2) \neq 0,\ Tr(L_2) \neq Tr(L_1),\ Det(L_2) \neq 0,\ Det(L_2) \neq Det(L_1), \tag{61}$$

and for a, b, c, d, h defined in (59), we have

$$b + h \leq \frac{ad}{c} \quad or \quad \frac{d}{c}\left(2h - \frac{ad}{c}\right) - a\left(b + h - \frac{ad}{c}\right) \neq 0. \tag{62}$$

If L_1 and L_2 satisfy

$$Tr(L_1 + L_2) < 0,\ Det(L_1 + L_2) > 0,\quad and\quad Det(L_1 - L_2) < 0, \tag{63}$$

then there exists $\tau_0 > 0$, the equilibrium (u^*, v^*) is stable for (44) when $0 \leq \tau < \tau_0$, but it is unstable when $\tau \in (\tau_0, \tau_0 + \varepsilon)$ for $\varepsilon > 0$ and small.

Similarly Corollary 6 implies the following observation:

Corollary 8. *Suppose that f, g, (u^*, v^*), L_1 and L_2 are same as in Theorem 7. Let $M_{2\times 2}$ be the set of all real-valued 2×2 matrices, and let $\mathscr{M}_1$ be a subset of $(M_{2\times 2})^2$ consisting of all matrix pairs (L_1, L_2) satisfying (63). Then for almost every $(L_1, L_2) \in \mathscr{M}_1$, the conclusions in Theorem 7 hold.*

We remark that the results in [1] are mainly about under what conditions, conditional stability is achieved. Only in a very special case, we find a condition for absolute stability. Hence more general condition on a, b, c, d, h for absolute stability is still largely open.

We apply the result above to another Leslie–Gower predator–prey system with delays:

$$\begin{cases} u'(t) = u(t)(p - \alpha u(t) - \beta v(t-\tau)), & t > 0, \\ v'(t) = \mu v(t)\left(1 - \dfrac{v(t-\tau)}{u(t-\tau)}\right), & t > 0, \end{cases} \tag{64}$$

where p, α, β, and μ are positive parameters, and $\tau \geq 0$ is the delay. System (64) has a unique positive equilibrium

$$(u^*, v^*) = \left(\frac{p}{\alpha+\beta}, \frac{p}{\alpha+\beta}\right), \tag{65}$$

and the Jacobian matrices at (u^*, v^*) are

$$L_1 = \begin{pmatrix} -\dfrac{\alpha p}{\alpha+\beta} & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad L_2 = \begin{pmatrix} 0 & -\dfrac{\beta p}{\alpha+\beta} \\ \mu & -\mu \end{pmatrix}.$$

Hence the characteristic equation of system (64) is in the same form as (45) with

$$a = \frac{\alpha p}{\alpha+\beta}, \quad b = 0, \quad c = \mu, \quad d = \frac{\mu\alpha p}{\alpha+\beta}, \text{ and } \quad h = \frac{\mu\beta p}{\alpha+\beta}. \tag{66}$$

Since $a + c > 0$ and $b + d + h > 0$ hold for any parameter $\alpha, \beta, p, \mu > 0$, then (u^*, v^*) is always locally asymptotically stable when $\tau = 0$. If $\alpha > \beta$, then $b - d + h < 0$, and one can apply Theorem 7 to show that there exists a $\tau_0 > 0$, such that (u_*, v_*) is locally asymptotically stable when $0 \leq \tau < \tau_0$, and it is unstable when $\tau \in (\tau_*, \tau_* + \varepsilon)$ for small $\varepsilon > 0$.

Concluding Remarks

For the simplicity of presentation, we only state our results for delayed differential equations without spatial variables. The results in section "Main Results" also hold for the stability of a constant equilibrium of reaction-diffusion systems with Neumann boundary condition; see details in [1] and also [2, 3] for the examples in Section "Planar Systems with One Transcendental Term". For reaction-diffusion systems, the interaction between diffusion and delay can also produce more complex spatiotemporal pattern formation; see [1, 5, 16]. On the other hand, for the reaction-diffusion equation with Dirichlet boundary condition, the positive equilibrium is spatially nonhomogenous, and the corresponding characteristic equation is also nonhomogenous. Thus the stability analysis for Dirichlet boundary PDE models is much more involved. For the instantaneous and delayed feedback model (7), the corresponding PDE model is

$$\begin{cases} u_t(x,t) = d\Delta_x u(x,t) + ru(x,t)(1 - au(x,t) - bu(x,t-\tau)), & x \in \Omega, t > 0, \\ u(x,t) = 0, & x \in \partial\Omega, t > 0. \end{cases} \tag{67}$$

It is known that when $a \geq b$ and $r > d\lambda_1$, then the unique positive equilibrium $u_r(x)$ is globally asymptotically stable for any $\tau \geq 0$ (see [7, 12]); on the other hand, when $a < b$, and assume $r > d\lambda_1$ but $r - d\lambda_1$ is small, then there is a $\tau_0(r) > 0$ satisfying $\lim_{r \to d\lambda_1} (r - d\lambda_1)\tau_0(r) = \frac{1}{r\sqrt{b^2 - a^2}} \arccos\left(-\frac{a}{b}\right)$ such that the unique positive equilibrium $u_r(x)$ is stable when $\tau < \tau_0(r)$, and it is unstable when $\tau > \tau_0(r)$ (see [18]).

For a planar system with two variables and two equations, we have shown here that a general stability/instability criterion can be formulated in terms of Jacobian matrices at the equilibrium point. This delay-induced instability can be compared to the Turing's diffusion-induced instability for planar reaction-diffusion systems [19]. See [1] for more in that direction. It would be interesting to extend such notion for systems with three or more variables. Another interesting question is to prove the stability or instability for distributed delay instead of discrete delays.

Acknowledgements Partially supported by NSF grant DMS-1022648 and Shanxi 100-talent program.

References

1. Chen, S., Shi, J., Wei, J.: Time delay induced instabilities and Hopf bifurcations in general reaction-diffusion systems. J. Nonlinear Sci. **23**(1), 1–38 (2013).
2. Chen, S., Shi, J., Wei, J.: The effect of delay on a diffusive predator–prey system with Holling type-II predator functional response. Comm. Pure Appl. Anal. **12**(1), 481–501 (2013)
3. Chen, S., Wei, J., Shi, J.: Global stability and Hopf bifurcation in a delayed diffusive Leslie-Gower predator–prey system. Int. J. Bifurcat. Chaos **22**(3), 1250061 (2012)
4. Erneux, T.: Applied Delay Differential Equations, vol. 3 of Surveys and Tutorials in the Applied Mathematical Sciences. Springer, New York (2009)
5. Hadeler, K.P., Ruan, S.: Interaction of diffusion and delay. Discrete Contin. Dyn. Syst. Ser. B **8**(1), 95–105 (2007)
6. Hale, J.K., Verduyn Lunel, S.M.: Introduction to Functional-Differential Equations, vol. 99 of Applied Mathematical Sciences. Springer, New York (1993)
7. Huang, W.: Global dynamics for a reaction-diffusion equation with time delay. J. Differ. Equat. **143**(2), 293–326 (1998)
8. Hutchinson, G.E.: Circular causal systems in ecology. Ann. N. Y. Acad. Sci. **50**(4), 221–246 (1948)
9. Kuang, Y., Smith, H.L.: Global stability in diffusive delay Lotka-Volterra systems. Differ. Integr Equat. **4**(1), 117–128 (1991)
10. Kuang, Y.: Delay Differential Equations with Applications in Population Dynamics, vol. 191 of Mathematics in Science and Engineering. Academic Press, Boston, MA (1993)
11. Lenhart, S.M., Travis, C.C.: Global stability of a biological model with time delay. Proc. Am. Math. Soc. **96**(1), 75–78 (1986)
12. Pao, C.V.: Dynamics of nonlinear parabolic systems with time delays. J. Math. Anal. Appl. **198**(3), 751–779 (1996)
13. Ruan, S.: Absolute stability, conditional stability and bifurcation in Kolmogorov-type predator–prey systems with discrete delays. Q. Appl. Math. **59**(1), 159–173 (2001)
14. Ruan, S.: Delay differential equations in single species dynamics. Delay Differential Equations and Applications, vol. 205 of NATO Sci. Ser. II Math. Phys. Chem., pp. 477–517. Springer, Dordrecht (2006)

15. Ruan, S.: On nonlinear dynamics of predator–prey models with discrete delay. Math. Model. Nat. Phenom. **4**(2), 140–188 (2009)
16. Seirin Lee, S., Gaffney, E.A., Monk, N.A.M.: The influence of gene expression time delays on Gierer-Meinhardt pattern formation systems. Bull. Math. Biol. **72**(8), 2139–2160 (2010)
17. Smith, H.: An Introduction to Delay Differential Equations with Applications to the Life Sciences, vol. 57 of Texts in Applied Mathematics. Springer, New York (2011)
18. Su, Y., Wei, J., Shi, J.: Hopf bifurcation in a diffusive logistic equation with mixed delayed and instantaneous density dependence. J. Dyn. Differ. Equat. **24**(4), 897–925 (2012).
19. Turing, A.M.: The chemical basis of morphogenesis. Phil. Trans. R. Soc. Lond. Ser. B Biol. Sci. **237**(641), 37–72 (1952)
20. Wu, J.: Theory and Applications of Partial Functional-Differential Equations, vol. 119 of Applied Mathematical Sciences. Springer, New York (1996)
21. Yi, F., Wei, J., Shi, J.: Bifurcation and spatiotemporal patterns in a homogeneous diffusive predator–prey system. J. Differ. Equat. **246**(5), 1944–1977 (2009)

Chapter 6
Existence of Antiperiodic Solutions to Semilinear Evolution Equations in Intermediate Banach Spaces

Gisèle Mophou and Gaston M. N'Guérékata

Introduction

We consider semilinear evolution equations of the form

$$x'(t) = Ax(t) + f(t, x(t)), \qquad t \in \mathbb{R}, \tag{1}$$

where A is an unbounded sectorial operator with not necessarily dense domain in a Banach space X and $f : \mathbb{R} \times X_\alpha \to X$, where X_α, $\alpha \in (0,1)$, is any intermediate Banach space between $D(A)$ and X. Concrete examples of X_α are the fractional power spaces $D((-A)^\alpha)$, $0 < \alpha < 1$, the real interpolation spaces $D_A(\alpha, \infty)$, introduced by J. L. Lions and J. Peetre, and the Hölder spaces $D_A(\alpha)$ which coincide with the continuous interpolation spaces due to G. Da Prato and P. Grisvard; see section "Preliminaries".

We are concerned in this paper with the existence of antiperiodic mild solutions of the following semilinear integro-differential equation in a Banach space X

$$u'(t) = Au(t) + \int_{-\infty}^{t} a(t-s)Au(s)ds + f(t, Cu(t)), \tag{2}$$

where $C : X \to X$ is a bounded linear operator, A is a closed linear (not necessarily bounded) operator defined in a Banach space X, and $a \in L^1_{loc}(\mathbb{R}^+)$ is a scalar-valued kernel.

Gisèle Mophou
Université des Antilles et de la Guadeloupe, Département de Mathématiques et Informatique, Université des Antilles et de La Guyane, Campus Fouillole 97159 Pointe-à-Pitre Guadeloupe (FWI), France
e-mail: gmophou@univ-ag.fr

Gaston M. N'Guérékata (✉)
Department of Mathematics, Morgan State University, 1700 E. Cold Spring Lane, Baltimore, M.D. 21251, USA
e-mail: Gaston.N'Guerekata@morgan.edu

B. Toni (ed.), *Advances in Interdisciplinary Mathematical Research*, Springer Proceedings in Mathematics & Statistics 37, DOI 10.1007/978-1-4614-6345-0_6,

We are interested in finding conditions under which Eq. (1) has an antiperiodic mild solution.

Our paper is motivated by the recent work [5] where the authors studied the existence of antiperiodic mild solutions of the following semilinear integro-differential equation in a Banach space X

$$u'(t) = Au(t) + \int_{-\infty}^{t} a(t-s)Au(s)ds + f(t, Cu(t)), \tag{3}$$

where $C : X \to X$ is a bounded linear operator, A is a closed linear (not necessarily bounded) operator defined in a Banach space X, and $a \in L^1_{loc}(\mathbb{R}^+)$ is a scalar-valued kernel.

Our main result is Theorem 4.

Preliminaries

In this section we recall some definitions and fix notations which will be used in the sequel. Throughout this paper, X is a Banach space and A is a sectorial operator with not necessarily dense domain, i.e., there are constants $\omega \in \mathbb{R}$, $\theta \in]\frac{\pi}{2}, \pi[$, $M > 0$ such that

$$(i) \qquad \rho(A) \supset S_{\theta,\omega} := \{\lambda \in \mathbb{C} : \lambda \neq \omega, |arg(\lambda - \omega)| < \theta\}. \tag{4}$$

$$(ii) \qquad \|R(\lambda, A)\| \leq \frac{M}{|\lambda - \omega|}, \quad \lambda \in S_{\theta,\omega}. \tag{5}$$

Hence, A generates an analytic semigroup $\mathcal{T} := (T(t))_{t \geq 0}$ on $(0, \infty)$ to $\mathcal{L}(X)$ satisfying

$$\|T(t)\| \leq M_0 e^{\omega t}, \quad t > 0, \tag{6}$$

$$\|t(A - \omega)T(t)\| \leq M_1 e^{\omega t}, \quad t > 0. \tag{7}$$

The semigroup $\mathcal{T}$ is assumed to be hyperbolic, i.e., there exist a projection P and constants $M, \delta > 0$ such that each $T(t)$ commutes with P, $\ker P$ is invariant with respect to $T(t)$, $T(t) : ImQ \longrightarrow ImQ$ is invertible and

$$\|T(t)Px\| \leq Me^{-\delta t}\|x\| \qquad \text{for } t \geq 0, \tag{8}$$

$$\|T(t)Qx\| \leq Me^{\delta t}\|x\| \qquad \text{for } t \leq 0, \tag{9}$$

where $Q := I - P$ and, for $t \leq 0$, $T(t) := (T(-t))^{-1}$.

We recall that if $\mathcal{T}$ is analytic, then $\mathcal{T}$ is hyperbolic if and only if

$$\sigma(A) \cap i\mathbf{R} = \emptyset;$$

see, for instance, [3, Prop 1.15, p.305].

For $\alpha \in (0,1)$, a Banach space X_α with norm $\|\cdot\|_\alpha$ is said to be an intermediate space between $D(A)$ and X, or a space of class $\mathscr{J}_\alpha$, if $D(A) \subset X_\alpha \subset X$ and there is a constant $c > 0$ such that

$$\|x\|_\alpha \leq c\|x\|^{1-\alpha}\|x\|_A^\alpha, \qquad x \in D(A), \tag{10}$$

where $\|\cdot\|_A$ is the graph norm associated to A. Concrete examples of X_α are $D((-A)^\alpha)$, $\alpha \in (0,1)$, the domains of the fractional powers of $-A$, the real interpolation spaces $D_A(\alpha,\infty)$, $\alpha \in (0,1)$, defined as follows:

$$\begin{cases} D_A(\alpha,\infty) := \{x \in X : [x]_\alpha = \sup_{0<t\leq 1} \|t^{1-\alpha}(A-\omega)e^{-\omega t}T(t)x\| < +\infty\} \\ \|x\|_\alpha = \|x\| + [x]_\alpha, \end{cases}$$

and the abstract Hölder spaces $D_A(\alpha) := \overline{D(A)}^{\|\cdot\|_\alpha}$. A very important property of these last two spaces is given by the fact that they depend only on $D(A)$ and X (in contrast with the fractional power spaces of $-A$). That is, for another sectorial operator B with $D(B) = D(A)$, their interpolation and Hölder spaces coincide. For more details about intermediate spaces, see, for instance, [3, Chap. II, Sect. 5.b] and [4].

For the hyperbolic analytic semigroup $\mathscr{T}$, we can easily check that estimations similar to (8) and (9) hold also with norms $\|\cdot\|_\alpha$. In fact, as the part of A in ImQ is bounded, it follows from the inequality (9) that

$$\|AT(t)Qx\| \leq c'e^{\delta t}\|x\| \qquad \text{for } t \leq 0.$$

Hence, from (10) there exists a constant $c(\alpha) > 0$ such that

$$\|T(t)Qx\|_\alpha \leq c(\alpha)e^{\delta t}\|x\| \qquad \text{for } t \leq 0. \tag{11}$$

We have also

$$\|T(t)Px\|_\alpha \leq \|T(1)\|_{\mathscr{L}(X,X_\alpha)}\|T(t-1)Px\| \qquad \text{for } t \geq 1,$$

and then from (8), we obtain

$$\|T(t)Px\|_\alpha \leq M'e^{-\delta t}\|x\|, \qquad t \geq 1,$$

where M' depends on α. For $t \in (0,1]$, by (7) and (10)

$$\|T(t)Px\|_\alpha \leq M''t^{-\alpha}\|x\|.$$

Hence, there exist constants $M(\alpha) > 0$ and $\gamma > 0$ such that

$$\|T(t)Px\|_\alpha \leq M(\alpha)t^{-\alpha}e^{-\gamma}\|x\| \qquad \text{for } t > 0. \tag{12}$$

Antiperiodic Functions

Definition 1. A function $f \in BC(\mathbb{R},X)$ is said to be ω-antiperiodic (or simply antiperiodic) if there exists $\omega > 0$ such that $f(t+\omega) = -f(t)$ for all $t \in \mathbb{R}$. The least such ω will be called the antiperiod of f.

We will denote by $P_{\omega ap}(X)$ the space of all ω-antiperiodic functions $\mathbb{R} \to X$.

Theorem 2. *[5] Let $f, f_1, f_2 \in P_{\omega ap}(X)$. Then the following also are in $P_{\omega ap}(X)$:*

- *$f_1 + f_2, cf$, c is an arbitrary real number.*
- *$g(t) := (\frac{1}{f})(t)$, provided $f \neq 0$ on $\mathbb{R}$. (Here $X = \mathbb{R}$).*
- *$f_a(t) := f(t+a)$ a is an arbitrary real number.*

Remark 3. It is clear that every ω-antiperiodic function is 2ω-periodic.

Remark 4. If $A \in \mathscr{B}(X)$, the space of all bounded linear operators $X \to X$, and f is an ω-antiperiodic X-valued function, then Af is also ω-antiperiodic.

A classical example of such function is

$$f(t) = \sum_{n=1}^{\infty} \frac{cos[(2n+1)t]}{n^2}, \ t \in \mathbb{R}$$

which is π-antiperiodic. See also [2, 5] for more examples.

Theorem 5 ([5]). *Let $f_n \in P_{\omega ap}(X)$, such that $f_n \to f$ uniformly on $\mathbb{R}$. Then $f \in P_{\omega ap}(X)$. Thus, $P_{\omega ap}(X)$ is a Banach space equipped with the supnorm.*

Now let X, Y be Banach spaces. Then we have the following which is slightly more general than Definition 2.13 [5].

Definition 6. A function $F \in BC(\mathbb{R} \times Y, X)$ is said to be ω-antiperiodic with antiperiod ω if $F(t+\omega,x) = -F(t,x)$ for all $t \in \mathbb{R}$ uniformly in $x \in Y$.

Define the Nemytskii's superposition operator

$$\mathscr{N}(\varphi)(\cdot) := F(\cdot, \varphi(\cdot)), \ \ \varphi \in P_{\omega ap}(X_\alpha).$$

We state here a slight generalization of Theorem 2.16 [5].

Theorem 7. *Let $F \in BC(\mathbf{R} \times X_\alpha, X)$. The following properties are equivalent:*

i) *For every $\varphi \in P_{\omega ap}(X_\alpha)$, $\mathscr{N}(\varphi) \in P_{\omega ap}(X)$.*
ii) $\forall (t,x) \in \mathbf{R} \times X_\alpha, F(t+\omega,-x) = -F(t,x)$.

Proof. The proof is similar to the one of Theorem 2.16 [5]. □

Existence of Anti-periodic Solutions

Let's first consider the following inhomogeneous problem:

$$\frac{d}{dt}x(t) = Ax(t) + h(t), \qquad t \in \mathbb{R}. \tag{13}$$

Definition 1. A mild solution of (13) is a continuous function $x : \mathbb{R} \to X_\alpha$ satisfying

$$x(t) = T(t-s)x(s) + \int_s^t T(t-\sigma)h(\sigma,)\,d\sigma \tag{14}$$

for all $t \geq s$ and all $s \in \mathbb{R}$.

Remark 2. [1] If $h \in BC(\mathbb{R}, X)$, then there is a unique mild solution $x(\cdot)$ of Eq. (13) in $BC(\mathbb{R}, X_\alpha)$ given by

$$x(t) = \int_{-\infty}^t T(t-s)Ph(s)ds - \int_t^{+\infty} T(t-s)Qh(s)ds, \quad t \in \mathbb{R}. \tag{15}$$

Theorem 3. *If $h \in P_{\omega ap}(X)$, then the unique bounded and continuous mild solution of Eq. (13) is also in $P_{\omega ap}(X)$.*

Proof. It is well-known that such a mild solution $x(t)$ of Eq. (13) is represented by

$$x(t) = \int_{-\infty}^t T(t-s)Ph(s)ds - \int_t^{+\infty} T(t-s)Qh(s)ds, \quad t \in \mathbb{R}.$$

So

$$x(t+\omega) = \int_{-\infty}^{t+\omega} T(t+\omega-s)Ph(s)ds - \int_{t+\omega}^{+\infty} T(t+\omega-s)Qh(s)ds.$$

Letting $s - \omega = \sigma$, we obtain

$$\begin{aligned} x(t+\omega) &= \int_{-\infty}^t T(t-\sigma)Ph(\sigma+\omega)d\sigma - \int_t^{+\infty} T(t-\sigma)Qh(\sigma+\omega)d\sigma \\ &= -\int_{-\infty}^t T(t-\sigma)Ph(\sigma)d\sigma + \int_t^{+\infty} T(t-\sigma)Qh(\sigma+)d\sigma \\ &= -x(t). \end{aligned}$$

The proof is complete. □

Theorem 4. *Suppose that*

- *H1.* $\forall (t,x) \in \mathbb{R} \times X_\alpha, f(t+\omega, -x) = -f(t,x)$.
- *H2. f satisfies the condition*

$$\|f(t,x) - f(t,y)\| \leq k(t)\|x-y\|_\alpha$$

for every $t \in \mathbb{R}$ and $x, y \in X_\alpha$ and some function $k \in L^p(\mathbb{R}, \mathbb{R}^+)$ with $p \in (\frac{1}{1-\alpha}; \infty]$, such that

$$\left[M(\alpha)(\gamma q)^\alpha (\Gamma(1-\alpha q))^{1/q} + \frac{c(\alpha)}{(\gamma q)^{1/q}}\right] \|k\|_p < 1 \tag{16}$$

where q is the conjugate of p (note that $1 - q\alpha > 0$ since $p > \frac{1}{1-\alpha}$).

Then Eq. (1) has a unique mild solution in $P_{\omega ap}(X)$.

Proof. Note that mild solutions of Eq. (1) are of the form

$$x(t) = \int_{-\infty}^{t} T(t-s)Pf(s,x(s))ds - \int_{t}^{+\infty} T(t-s)Qf(s,x(s))ds, \quad t \in \mathbb{R}.$$

By Theorem 3 above, $f \in P_{\omega ap}(X_\alpha)$. Then by Theorem 7, we deduce that $x(t) \in P_{\omega ap}(X)$.

So the mapping $\mathcal{G} : P_{\omega ap}(X_\alpha) \to P_{\omega ap}(X_\alpha)$ given by

$$(\mathcal{G}x)(t) := \int_{-\infty}^{t} T(t-s)Pf(s,x(s))ds - \int_{t}^{+\infty} T(t-s)Qf(s,x(s))ds, \quad t \in \mathbb{R}$$

is well-defined.

Now let $u, v \in P_{\omega ap}(X_\alpha)$. Then we have

$$\begin{aligned}
\|(\mathcal{G}u)(t) - (\mathcal{G}v)(t)\|_\alpha &\le \int_{-\infty}^{t} \|T(t-s)P[f(s,u(s)) - f(s,v(s))]\|_\alpha ds \\
&\quad + \int_{t}^{+\infty} \|T(t-s)Q[f(s,u(s)) - f(s,v(s))]\|_\alpha ds \\
&\le M(\alpha) \int_{-\infty}^{t} (t-s)^{-\alpha} e^{-\gamma(t-s)} \|[f(s,u(s)) - f(s,v(s))]\| ds \\
&\quad + c(\alpha) \int_{t}^{+\infty} e^{\delta(t-s)} \|[f(s,u(s)) - f(s,v(s))]\| ds \\
&\le M(\alpha) \int_{-\infty}^{t} (t-s)^{-\alpha} e^{-\gamma(t-s)} k(s) \ \|u(s) - v(s)]\|_\alpha ds \\
&\quad + c(\alpha) \int_{t}^{+\infty} e^{\delta(t-s)} k(s) \|u(s) - v(s)]\|_\alpha ds \\
&\le \left[M(\alpha) \int_{-\infty}^{t} (t-s)^{-\alpha} e^{-\gamma(t-s)} k(s) ds + c(\alpha) \int_{t}^{+\infty} e^{\delta(t-s)} k(s) ds\right] \sup_t \|u(t) - v(t)]\|_\alpha.
\end{aligned}$$

Now we use Hölder's inequality. Assume first that p is finite. We can write

$$\|(\mathscr{G}u)(t)-(\mathscr{G}v)(t)\|_\alpha \le \left[M(\alpha)\left(\int_{-\infty}^{t}(t-s)^{-q\alpha}e^{-q\gamma(t-s)}ds\right)^{\frac{1}{q}}\left(\int_{-\infty}^{t}(k(s))^p ds\right)^{\frac{1}{p}}\right.$$
$$\left.+c(\alpha)\left(\int_{t}^{+\infty}e^{q\delta(t-s)}ds\right)^{\frac{1}{q}}\left(\int_{t}^{\infty}(k(s))^p ds\right)^{\frac{1}{p}}ds\right]\sup_t\|u(t)-v(t)]\|_\alpha$$
$$\le\left[M(\alpha)(\gamma q)^\alpha(\Gamma(1-\alpha q))^{1/q}+\frac{c(\alpha)}{(\gamma q)^{1/q}}\right]\|k\|_p\sup_t\|u(t)-v(t)]\|_\alpha.$$

When $p=\infty$, we obtain directly the same result (with $q=1$). So, it is true for any p. And so

$$\sup_t\|(\mathscr{G}u)(t)-(\mathscr{G}v)(t)\|_\alpha$$
$$\le\left[M(\alpha)(\gamma q)^\alpha(\Gamma(1-\alpha q))^{1/q}+\frac{c(\alpha)}{(\gamma q)^{1/q}}\right]\|k\|_p\sup_t\|u(t)-v(t)]\|_\alpha.$$

The proof is completed, by using Banach's fixed-point theorem. □

Acknowledgements We are grateful to the referee for his/her valuable suggestions and corrections.

References

1. Boulite, S., Maniar, L., N'Guérékata, G.M.: Almost automorphic solutions for hyperbolic semilinear evolution equations. Semigroup Forum **71**, 231–240 (2005)
2. Chen, H.L.: Antiperiodic functions. J. Comput. Math **14**(1), 32–39 (1996)
3. Engel, K.J., Nagel, R.: One Parameter Semigroups for Linear Evolution Equations, Graduate texts in Mathematics. Springer, New York (1999)
4. Lunardi, A.: Analytic Semigroups and Optimal Regularity in Parabolic Problems. Birkhäuser, Basel (1995)
5. N'Guérékata, G.M., Valmorin, V.: Antiperiodic solutions of semilinear integrodifferential equations in Banach spaces. App. Math. Comput. **218**(22), 11118–11124 (2012)

Chapter 7
Signal, Image Processing, and Machine Learning: The Key to Complex Problems in Medicine and Biology

Mahsa Zahery and Kayvan Najarian

Introduction

Computer-aided decision-making systems have been introduced into many fields, such as economics, medicine, architecture, and agriculture. The increasing demand and rapid pace of development of such computer-aided decision-making systems displays their popularity and success in aiding and enhancing various fields. In the field of medicine, the advantage of having such systems is in the expense, labor, energy, and budget savings they provide to the health care environments. In the following sections, a brief description of the application of such systems in hemorrhagic shock, attention detection, traumatic brain injuries, and pelvic fracture detection has been provided. A flowchart of the procedure of developing such systems is represented in Fig. 7.1.

Hemorrhage Detection

An example of using a computer-aided decision-making system is in dealing with traumatic injuries (injuries caused by an accident, a battle, or an illness) wherein the effective decisions produced by a computer-aided system can be very handy in controlling the situation and quickly assessing the patient's health condition. Such systems are typically designed to detect the type of illness, assess the severity, and thereby help in the allocation of resources.

In [11], a computational system is proposed, which is designed to estimate the severity of blood volume loss. Severe hemorrhage is the event of losing large volumes of blood which leads to reduced blood and oxygen perfusion to vital

Mahsa Zahery (✉) • Kayvan Najarian
Department of Computer Science, Virginia Commonwealth University, Richmond, VA, USA
e-mail: zaherym@vcu.edu; knajarian@vcu.edu

B. Toni (ed.), *Advances in Interdisciplinary Mathematical Research*, Springer Proceedings in Mathematics & Statistics 37, DOI 10.1007/978-1-4614-6345-0_7,

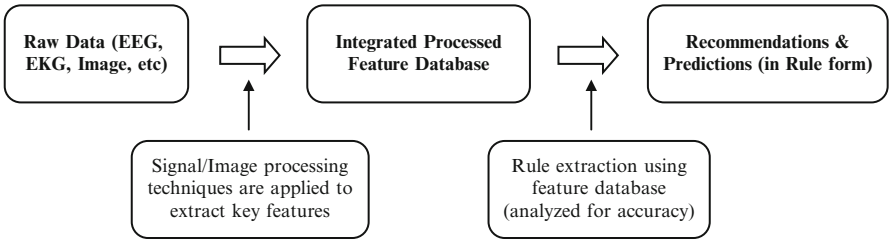

Fig. 7.1 Procedure of developing a computer-assisted system using signal processing and machine learning [3]

organs. This can be life-threatening and hence requires immediate care and attention. Depending on the severity, hemorrhage can be categorized into different levels, such as mild, moderate, and severe. The proposed system is capable of estimating and incorporating the severity of blood volume loss and hemorrhage knowledge. This is very efficient not only in saving lives but also in reducing the cost of treatments.

The methodology can be introduced in three steps, each of which contains novel and transformative concepts. Preprocessing of the raw signals is the first step which includes algorithms to detect QRS complex and systolic/diastolic waveforms along with the usage of an adaptive filtering method to filter the noise in the signals. The second step involves combining the features which are extracted from time domain, frequency domain, nonlinear analysis, and multi-model analysis (feature extraction step). This way, a better representation of the hemorrhage patterns is provided. The last step uses a machine learning algorithm for high-accuracy and real-time decision-making. At this stage, a new version of error-correcting output code (ECOC) has been developed. Accuracy obtained by the proposed system is much higher compared to the accuracy from the United States Army Institute of Surgical Research lower body negative pressure (USAISR LBNP) dataset thereby justifying the reliability of the proposed system (an accuracy of 99.89% in case of QRS detection and 99.95% in case of systole and diastole detection). In the following section, the conventional ECOC algorithm, as well as the properties of its improved version, is explained briefly.

Error-Correcting Output Codes (ECOC)

Combining the output of binary classifiers, ECOC [8] solves multiclass learning problems by using an error-correcting output code matrix. The framework of this method is given in Fig. 7.2.

A matrix of k rows and n columns is generated, with k representing the number of classes and n not being limited to any value as long as it satisfies the $n > \log_2 k$. Matrix elements are either 1 or -1 which are the binary codes adjusted to each class label.

The training stage of an ECOC classifier builds the coding matrix of size k by $n$following the steps provided below:

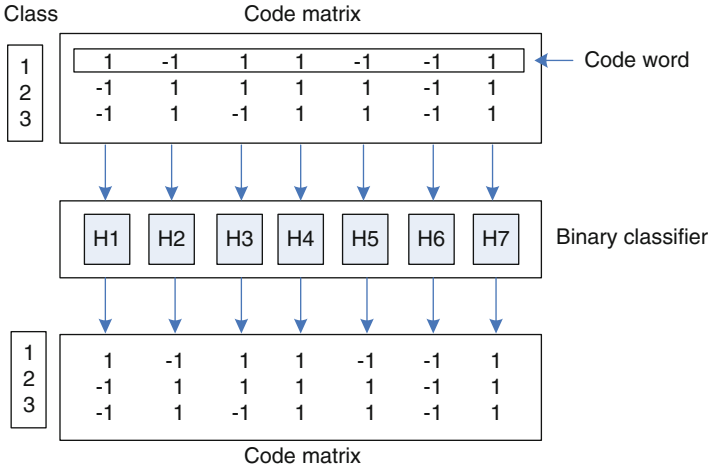

Fig. 7.2 Framework of ECOC algorithm [10]

- A coding matrix of values 1/−1 is generated.
- For each column j of this matrix:
  - Two superclasses are made. One has all the labels i(the row numbers) for which the (ij)th element of the matrix is 1, and the other consists of the labels for −1 elements of the matrix.
  - A binary classifier is generated to differentiate between the superclasses made in the previous step.

The testing stage of the algorithm classifies a new example, given the matrix constructed at the training stage, using the following steps:

- For each column j of the matrix:
  - The probability with which the binary classifier for column j allocates the new example to label 1 superclass is calculated.
- The proximity (according to Hamming distance) of the vector containing the probabilities from all the binary classifiers (each column has one binary classifier) to each row of the matrix is calculated.
- The row giving the minimum value is the class label for the new example.

The proposed framework for ECOC in [11] improves the conventional ECOC algorithm in the following areas:

- It changes the code matrix to BCH [15] which is a specific type of error-correcting output code and is one of the most common used approaches. The choice of the code matrix is avoided in BCH, but this does not affect its error-correcting capability.
- It makes use of support vector machine (SVM) as binary classifier. SVM can handle data/sample imbalance and small sample size problems more effectively.

– It decomposes the learning problem, in the sense that by assuming a normal distribution for the data, data points in one class form several normal distributions with different expectations and standard deviations, distinguishing them from each other, while these different subclasses are still under the coverage of the same class. Afterwards, the data is allocated to different layers in each of which the region is decomposed to two subregions: one with high confidence on the prediction result and the other with low confidence. Next, the dataset is sent to the next layer for classification.

For the details on the framework of the improved ECOC algorithm, the reader is encouraged to refer to [11].

Attention Detection

With the stressful environments, extended work hours, and high workload, the high-paced life of most people has made sleep disorders a more commonplace in societies. In addition, there are certain daily tasks which are repetitive and tedious in nature, thus leading to fluctuations in people's attention spans and capacity. This is a very critical problem since losing attention during certain activities or profession can be very dangerous and deadly.

Using ECG (electrocardiograph) which is a fundamental physiological signal, a real-time monitoring system has been developed in [3] to predict whether an individual is paying attention during a task execution or not. The aim of this study is to find the effect of the body's physiological parameters on the individuals' attention level. Using noninvasive portable monitors, these signals are collected to predict an individual's inclination to sleep or loss of attention well ahead of time. With advanced signal processing techniques, informative features are extracted. Specific features related to heart's rhythm are extracted using a QRS complex detection algorithm. Next, using dual-tree complex wavelet transform (DT-CWT) and Stockwell transform, the ECG signal is decomposed to extract more features which are informative in differentiating the subtle changes in the acquired ECG signal [2] and [4]. The next step involves using machine learning algorithms to categorize these extracted features between cases of attention and non-attention. Finally, EEG (electroencephalograph) signals are analyzed and classified to act as a benchmark for comparison with ECG classification, since EEG signals are fundamentally more informative in providing information regarding human cognitive activity [5]. However, typically EEG signal collection devices are cumbersome and many times non-portable thereby limiting its usability in real-world scenarios.

ECG and EEG signals of around 15 subjects are collected. The volunteers are asked to view videos for 40 min, consisting of 20 min of interesting clips and 20 min of clips that are not interesting. After decomposing the acquired data and analyzing

it for feature extraction and classification, a fairly reasonable accuracy of 78.27% shows that with only the ECG signals of the volunteers, it is possible to distinguish between the presence and lack of attention in the subjects [5].

Dual-Tree Complex Wavelet Transform

Wavelet transform was developed to overcome the deficiencies of short-time Fourier transform (STFT). Regular Fourier transform is not always able to represent a signal's time-dependent nature. The problem is that Fourier transform does not reflect the time at which a frequency exists. This is not a problem for stationary signals. However, for nonstationary signals, STFT was introduced. STFT moves a window throughout the signal. Fourier transform is then applied to each window to obtain the frequency information of each window. The problem with STFT is that it considers the same resolution for all frequencies.

To tackle this problem, wavelet transform was developed. Wavelet transform considers different resolutions for different frequencies. Discrete wavelet transform (DWT) substitutes the infinitely fluctuating sinusoidal basis functions of Fourier transform with locally fluctuating basis functions referred to as wavelets. Wavelets are basis functions with concentrated energy using which the signals are decomposed.

Dual-tree complex wavelet transform (DT-CWT) was proposed by [9] to come up with solutions to the constraints of DWT. DWT operates a decimation task while transforming a signal. This makes DWT a shift variant transformation which creates various output wavelet coefficients in response to a small shift in the analyzed signal. The other deficiencies of DWT are susceptibility to aliasing, oscillations, and lack of directionality [7, 9], and [12].

Providing directional wavelets, shift-invariant property, as well as amending angular resolution, DT-CWT uses a dual tree of real filters to achieve the real and imaginary parts of the generated complex coefficients [13]. Figures 7.3 and 7.4 illustrate the analysis and synthesis filter banks, respectively.

Consisting of two parallel wavelet transforms, DT-CWT calculates the wavelet coefficients and scaling coefficients of the first tree in the following manner:

$$d_l^{Re}(k) = 2^{l/2} \int_{-\infty}^{+\infty} x(t)\psi_h(2^l t - k)dt,\ l = 1, \ldots, j \tag{1}$$

$$c_j^{Re}(k) = 2^{j/2} \int_{-\infty}^{+\infty} x(t)\phi_h(2^j t - k)dt \tag{2}$$

where l is the scaling factor and $j$is the maximum scale. The coefficients of the second tree are calculated in similar manner:

$$d_l^{Im}(k) = 2^{l/2} \int_{-\infty}^{+\infty} x(t)\psi_g(2^l t - k)dt,\ l = 1, \ldots, j \tag{3}$$

$$c_j^{Im}(k) = 2^{j/2} \int_{-\infty}^{+\infty} x(t)\phi_g(2^j t - k)dt \tag{4}$$

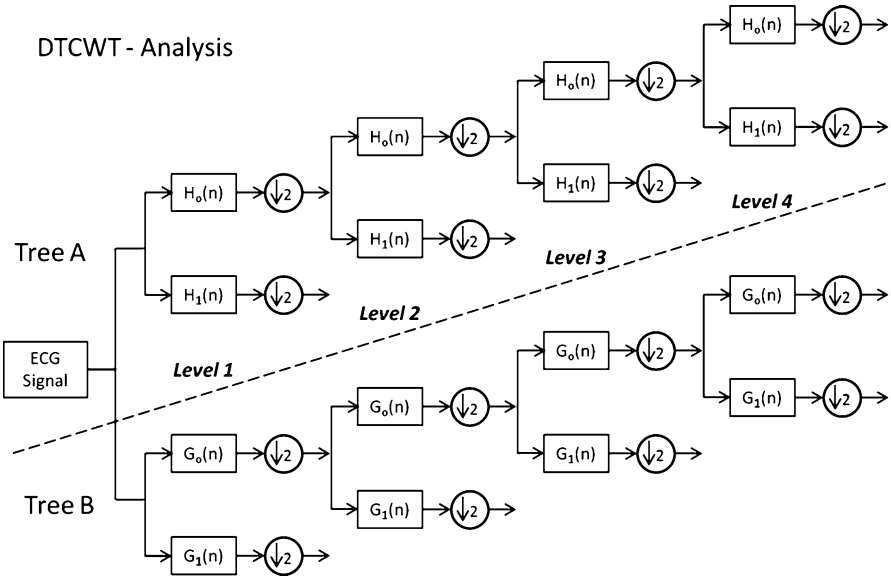

Fig. 7.3 $H_0(n)$ and $H_1(n)$ are, respectively, representatives of high-pass and low-pass filters for tree A. $G_0(n)$ and $G_1(n)$ are high- and low-pass filters for tree B, similarly. The input signal is down sampled to approximately half its original size at each level of decomposition. The output of each level is the detailed and approximation coefficient of the input signal [1]

The coefficients of DT-CWT are calculated as provided below:

$$d_l^C(k) = d_l^{Re}(k) + jd_l^{Im}(k),\ l = 1, \ldots, j \tag{5}$$

$$c_j^C(k) = c_j^{Re}(k) + jc_j^{Im}(k) \tag{6}$$

Feature Extraction in DT-CWT

The real and imaginary coefficients from the DT-CWT decomposition and the real part of the approximation coefficient are used to extract features. Five levels of DT-CWT are performed on the ECG signal for windows of length 10 s. For each level, real and complex detailed coefficients as well as the real parts of the level 5 approximate coefficient are considered.

Having $x_1, x_2, \cdots, x_n$ as the values of each coefficient obtained from each 10 s window, several statistical features such as standard deviation, median, minimum and maximum, energy, power, entropy, skewness, kurtosis, range, signal complexity, signal mobility, log of variance, mean of frequencies, variance of probability distribution, sum of autocorrelation, mean of auto-covariance, and entropy of frequency are calculated. Below a brief description of skewness and kurtosis, two statistical features affecting the shape of a signal, is provided.

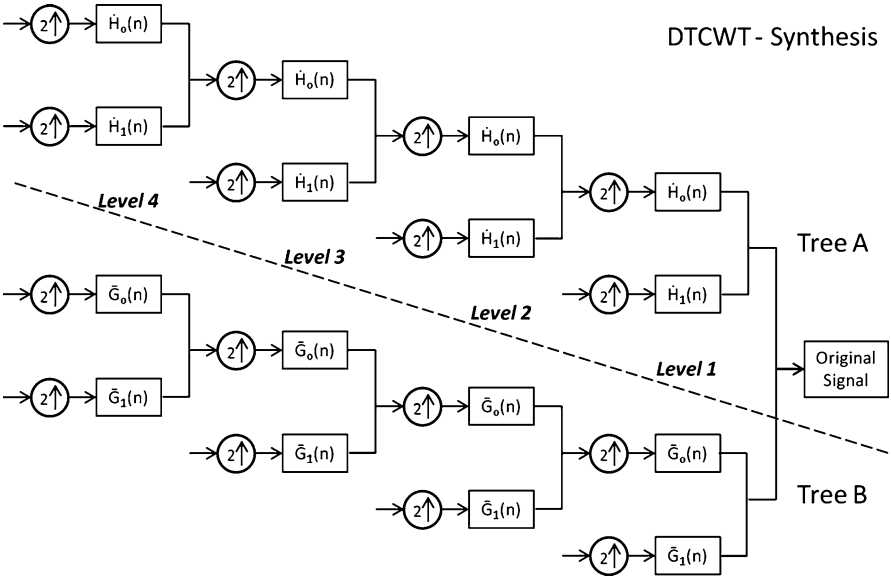

Fig. 7.4 $\dot{\mathbf{H}}_{0(n)}$ and $\dot{\mathbf{H}}_{1(n)}$ are, respectively, representatives of high-pass and low-pass filters for tree A. $\bar{\mathbf{G}}_{0(n)}$ and $\bar{\mathbf{G}}_{1(n)}$ are high- and low-pass filters for tree B. The input signal is down sampled to roughly half its original size at each level of decomposition [1]

Skewness and Kurtosis

Shape parameters are considered as parameters affecting the shape of a distribution as opposed to its location and scale. Skewness and kurtosis are shape parameters measuring the degree of asymmetry and peakedness of the probability distributions, respectively.

Mathematically speaking, skewness gives the third moment of a random variable as shown in (7):

$$Skewness = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left[\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]^3} \tag{7}$$

Negative value for skewness represents a probability distribution skewed to the left tail (left of the mean), and positive value demonstrates a distribution skewed more to the right tail (right of the mean).

With kurtosis, the fourth moment of a random variable is provided, as given in (8):

$$Kurtosis = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2} \tag{8}$$

A distribution concentrated around the mean has a high, sharp peak with a kurtosis value of greater than 3. In contrast, a kurtosis value of less than 3 portrays a flat distribution with low, less obvious peak. Value of 3 is the kurtosis of a normal distribution used as a reference standard.

In the attention detection decision-making computer system by [3], the feature selection process resulted in selecting skewness for all levels of the imaginary parts of the detailed coefficients and levels 1, 3, and 5 of real parts of the detailed coefficients. For kurtosis, level 5 of the approximate coefficient along with levels 4 and 5 of the imaginary part of the detailed coefficient was selected.

Traumatic Brain Injury

Each year, over 1.4 million people in the United States suffer from traumatic brain injury (TBI) [6]. Over 50,000 of these victims do not survive, out of which 50% die in the first two hours after the injury. Hence, it is vital to be able to diagnose the injury quickly. Here, computer tomography (CT) imaging, a fast and economical medical scan, comes handy as the gold standard for initial TBI assessment. Another advantage of CT is that it is capable of uncovering fractures or hematomas.

Increased intracranial pressure (ICP) is a common cause of death and a major complication of TBI which results in deformation of brain tissue. Cranial trepanation is the current method of ICP assessment which can result in patients' bleeding and infection due to its highly invasive nature. Therefore, a noninvasive approach as a preliminary step to perform trepanation would be more preferred.

Changes in the location and size of the ventricles can be helpful in deciding whether to perform cranial trepanation, and these changes are detectable by CT scan. The solution proposed by [6] focuses on automatic processing of CT images of brain for segmentation and identification of the ventricular systems. Segmentation of the ventricles helps in providing vital diagnosis knowledge through measuring the changes in the location and size of the ventricles.

Key features are extracted from CT ventricular images via image processing. The features include the extent of midline shift (a normal midline shift is defined according to the skull symmetry and anatomical features of a normal subject) in the brain and the size of the lateral ventricles.

An ideal midline detection algorithm proposed in [6] consists of the following three steps:

- Approximate midline detection based on the symmetry of the skull
- Falx cerebri and anterior bone protrusion detection
- Midline position refinement using these features

This algorithm resulted in 90% accuracy in midline detection.

A two-step approach is taken for segmentation of the ventricles. The first step is a low-level segmentation on each pixel of the CT images. Iterated conditional mode (ICM) and maximum A posteriori spatial probability (MASP) are the two

algorithms used in this step. Comparing the results of these algorithms, ICM with K-means as the initial segmentation method resulted in smoother segmentation, although some small parts were missing. MASP resulted in more noise than ICM. A modified version of MASP is used which speeds up the segmentation process for each slice by dealing only with each pixel's current estimated neighborhood. The comparison is illustrated in Fig. 7.5. In the next step, template matching algorithm is used to isolate the ventricles. The CT dataset used for ventricle segmentation contains mild and severe TBI subjects. In all the cases, the ventricles are detected successfully in all CT slices (100% accuracy).

Pelvic Fracture Detection

One of the most severe types of injuries suffered by trauma patients is traumatic pelvic injuries. Traumatic pelvic injuries along with the associated complications, specifically, hemorrhage and infected hematomas account for 8.6% to 50% of the mortalities. A computer-aided system is required to accelerate the decision-making procedure by analyzing large datasets of patients' information.

The computer-assisted decision-making system proposed in [14] allows for detection of fracture and hemorrhage through processing of CT images to assess the severity of a pelvic injury. A hierarchal procedure, capable of combining image enhancement and segmentation approaches, is proposed which results in accurate bone segmentation.

The procedure starts with detecting the bone regions and applying histogram equalization to the area for achieving a better contrast. To enhance the favorable features in the area, speckle reducing anisotropic diffusion (SRAD) is performed. Finally, with the aid of automated seeded region growing, the initial bone segmentation is refined. In 83% of the cases, the detected contours were acceptably accurate. Figure 7.6 represents the performance of the proposed approach compared to the actual image.

Conclusion

Nowadays, signal processing and machine learning techniques play a major role in dealing with biomedical problems. The noninvasive and computerized solutions these techniques provide to health care environments have increased their popularity and made them reliable tools for addressing medical issues.

In this chapter, two medical problems, hemorrhage detection in traumatic injuries and attention detection using ECG and EEG analysis, have been discussed. Error-correcting output codes (ECOC) and dual-tree complex wavelet transform (DT-CWT) are explained as two machine learning and signal processing techniques,

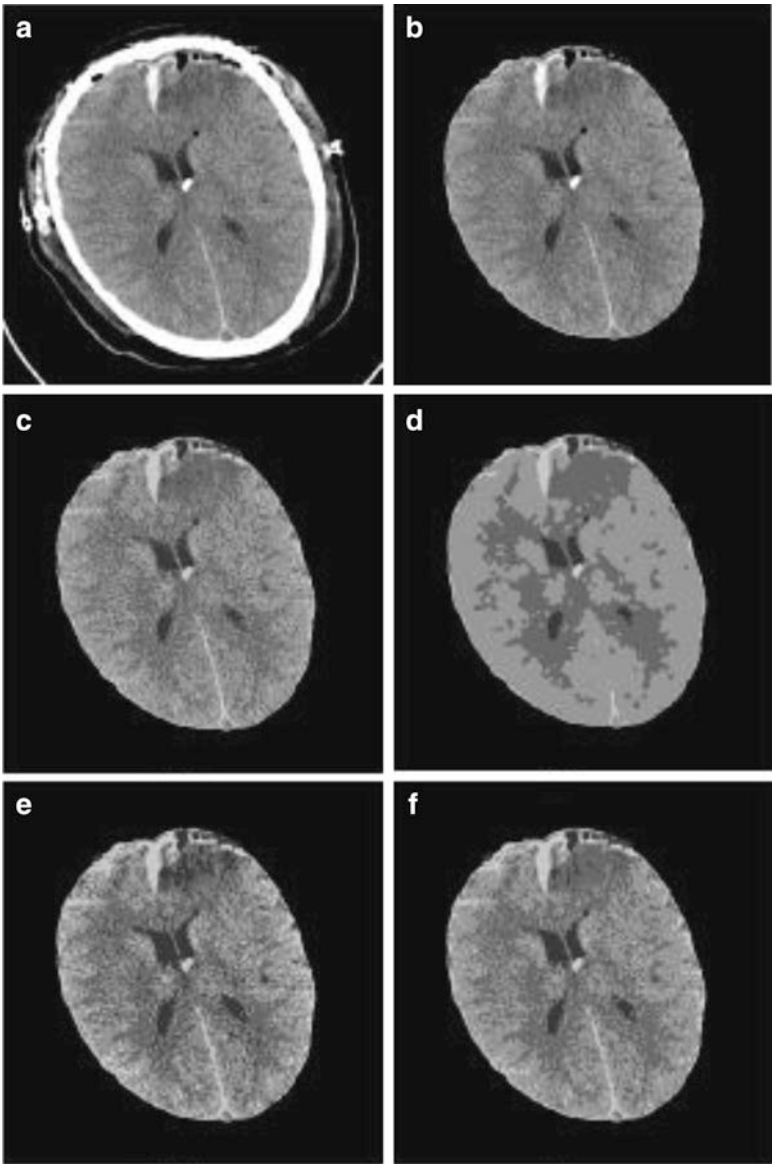

Fig. 7.5 Comparing segmentation methods. (**a**) Actual CT image. (**b**) CT image without the skull. (**c**) Result of K-means algorithm. Four clusters are identified with initial seeds. Noise and uneven intensity distribution have created some holes in the segmentation. (**d**) ICM segmentation with K-means as initial result. ICM gives smoother segmentation, although some small parts are missing. (**e**) MASP segmentation. MASP has resulted in more noise compared to ICM (*the right upper corner* is labeled wrongly as part of ventricles). (**f**) Modified MASP segmentation which is not as smooth as ICP, but resulted in less noise than MASP [6]

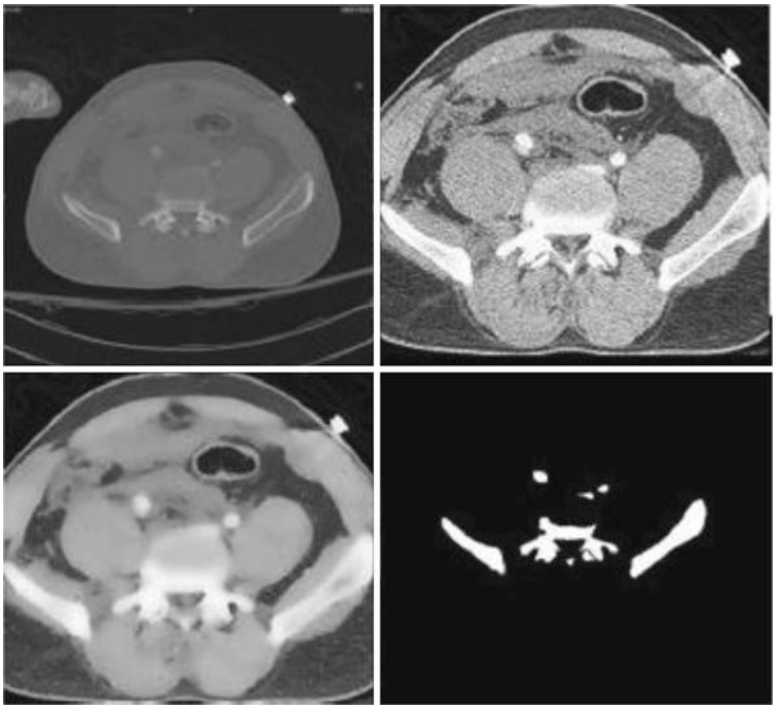

Fig. 7.6 The actual image can be seen in the *upper left corner*. The *upper right corner* represents the region where bone is detected after histogram equalization. The *lower left corner* image shows the image after SRAD filtering. In the *lower right corner*, the segmentation results are provided. The bone contour and shape of the segmented image match with those of the actual image [14]

respectively. Finally, the role of skewness and kurtosis as two statistical parameters for extracting features from DT-CWT has also been discussed.

Traumatic brain and pelvic injuries are investigated using image processing techniques on CT images. Segmentation methods such as ICP and MASP are used to address low-level segmentation on each pixel of the brain CT slices. In the next step, ventricles are detected using template matching algorithm. For pelvic images, a hierarchical procedure merging filtering and histogram equalization is proposed to enhance segmentation quality of the CT images.

Acknowledgement The authors would like to acknowledge Dr. Ashwin Belle, Dr. Yurong Lue, Dr. Simina Vascilache, and Dr. Wenan Chen for contributing their research to this chapter.

References

1. Belle, A.: A Physiological Signal Processing System for Optimal Engagement and Attention Detection (Unpublished doctoral dissertation). Virginia Commonwealth University, Richmond, VA (2012)
2. Belle, A., Hobson Hargraves, R., Najarian, K.: A physiologica signal processing system for optimal engagement and attention detection. In: Proceedings of the IEEE International Conference on Bioinformatics & Biomedicine Workshops, Atlanta, GA., Nov. 12–15, 2011, pp. 555–561
3. Belle, A., Hobson Hargraves, R., Najarian, K.: An automated optimal engagement and attention detection system using electrocardiogram. J. Comput. Math. Meth. Med. (2012). doi:10.1155/2012/528781
4. Belle, A., Ji, S.Y., Ansari, S., Hakimzadeh, R., Ward, K.R., Najarian, K.: Frustration detection with electrocardiogram signal using wavelet transform. In: Proceedings of the International Conference on Biosciences, Cancun, Mexico, Mar. 7–13, 2010, pp. 91–94
5. Belle, A., Pfaffenberger, M., Hobson Hargraves, R., Najarian, K.: An automated decision making system for detecting loss of attention in individuals using real time processing of electroencephalogram. In: Biosignal Interpretation- 7th International Workshop (2012)
6. Chen, W., Smith, R., Ji, S.-Y., Ward, K.R., Najarian, K.: Automated ventricular systems segmentation in brain CT images by combining low-level segmentation and high level template matching. BMC Med. Informat. Decis. Making 9, S4 (2009)
7. Choi, H., Romberg, J., Baraniuk, R., Kingsbury, N.: Hidden markov tree modeling of complex wavelet transforms. In: Proceedings of the IEEE International Conference of Acoustics, Speech, Signal Process, Istanbul, Turkey, June 2000, pp. 133–136
8. Dietterich, T.G., Bakiri, G.: Solving multiclass learning problems via error-correcting output codes. J. Artif. Intell. Res. **2**, 263–286 (1995)
9. Kingsbury, N.G.: The dual-tree complex wavelet transform: A new technique for shift invariance and directional filters. In: Proceedings 8th IEEE DSP Workshop, Utah, Aug. 9–12, 1998
10. Luo, Y.: The Severity of Stages Estimation During Hemorrhage Using Error Correcting Output Codes Method (Unpublished doctoral dissertation). Virginia Commonwealth University, Richmond, VA (2012)
11. Luo, Y., Najarian, K.: Employing decoding of specific error correcting codes as a new classification criterion in multiclass learning problems. In: 2010 International Conference on Pattern Recognition, 2010, pp. 4238–4241
12. Romberg, J., Choi, H., Baraniuk, R.: Multiscale classification using complex wavelets and hidden markov tree models. In: Proceedings of IEEE International Conference on Image Processing, Vancouver, Canada, September 2000. pp. 371–374
13. Selesnick, I., Baraniuk, R., Kingsbury, N.: The dual-tree complex wavelet transform. IEEE Signal Process. Mag. **22**, 123–151 (2005)
14. Vasilache, S., Ward, K., Najarian, K.: Unified wavelet and Gaussian filtering for segmentation of CT images; application in segmentation of bone in pelvic CT images. BMC Med. Informat. Decis. Making **9**, S8 (2009)
15. Wicker, S.B.: Error Control Systems for Digital Communication and Storage. Prentice-Hall, Englewood Cliffs, NJ (1995)

Chapter 8
Real-Time Noise Cancellation Using Wavelet Transforms

Ehsan Sheybani

Introduction

Noise from different sources can have dramatic effects on the performance and decision-making process of the systems. As such, total elimination of the noise could also be damaging to the final outcome, as it may result in removing useful information that can benefit the decision-making process. Several efforts have been made to find the optimal balance between noise and data parameters. For the most part, experts in the field agree that it is more beneficial to remove noise at the node level where data is collected [1–3]. This is mainly stressed so that the low-power, low bandwidth, and low computational overhead constraints are met while fused datasets can still be used to make reliable decisions [4–6].

Digital signal processing algorithms, based on advanced mathematical concepts, have long served to manipulate data to be a good fit for analysis and synthesis of any kind. For the noise removal application, a special wavelet-based approach has been considered to suppress the effect of noise in data. The proposed technique uses the orthogonality properties of wavelets to decompose the dataset into spaces of coarse and detailed signals. With the filter banks being designed from special bases for this specific application, the output signal in this case would be components of the original signal represented at different time and frequency scales and translations. A detailed description of the techniques follows in the next section.

Ehsan Sheybani (✉)
Virginia State University, Petersburg, VA, USA
e-mail: esheybani@vsu.edu

B. Toni (ed.), *Advances in Interdisciplinary Mathematical Research*, Springer Proceedings in Mathematics & Statistics 37, DOI 10.1007/978-1-4614-6345-0_8,

Wavelet-Based Transforms

Traditionally, Fourier transform (FT) has been applied to time-domain signals for signal processing tasks such as noise removal. The shortcoming of the FT is in its dependence on time averaging over entire duration of the signal. Due to its short time span, analysis of dataset requires resolution in particular time and frequency rather than frequency alone. Wavelets are the result of translation and scaling of a finite-length waveform known as mother wavelet. A wavelet divides a function into its frequency components such that its resolution matches the frequency scale and translation. To represent a signal in this fashion, it would have to go through a wavelet transform. Application of the wavelet transform to a function results in a set of orthogonal basis functions which are the time-frequency components of the signal. Due to its resolution in both time and frequency, wavelet transform is the best tool for detection and classification of signals that are nonstationary or have discontinuities and sharp peaks. Depending on whether a given function is analyzed in all scales and translations or a subset of them, the continuous (CWT), discrete (DWT), or multi-resolution wavelet transform (MWT) can be applied.

An example of the generating function (mother wavelet) based on the sinc function for the CWT is

$$\psi(t) = 2Sinc(2t) - Sinc(t) = \frac{Sin(2\pi t) - Sin(\pi t)}{\pi t} \tag{1}$$

normalized with scale one frequency band [1, 2]. The subspaces of this function are generated by translation and scaling. For instance, the subspace of scale (dilation) a and translation (shift) b of the above function is

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right) \tag{2}$$

$a > 0$ defines the scale and frequency band [1/a, 2/a], whereas b ϵ R is any real number defining the shift. When a function x is projected into this subspace, an integral would have to be evaluated to calculate the wavelet coefficients in that scale:

$$WT_{\psi}\{x\}(a,b) = \langle x, \psi_{a,b}\rangle = \int_R x(t)\overline{\psi_{a,b}(t)}dt \tag{3}$$

where $\overline{\psi_{a,b}(t)}$ indicates the conjugate of function $\psi_{a,b}$ and $\langle x, \psi_{a,b}\rangle$ is the inner product of $L_2(R)$, the space of square-integrable function over R. And therefore, the function x can be shown in term of its components:

$$x_a(t) = \int_R WT_{\psi}\{x\}(a,b).\psi_{a,b}(t)db \tag{4}$$

projection of x onto the subspace of scale a. Due to computational and time constraints, it is impossible to analyze a function using all wavelet coefficients. Therefore, usually a subset of the discrete coefficients is used to reconstruct the best approximation of the signal. This subset is generated from the discrete version of the generating function with the corresponding wavelet coefficients:

$$\psi_{m,n}(t) = a^{-m/2}\psi\left(a^{-m}t - nb\right). \tag{5}$$

with integers m, n ε Z. Applying this subset to a function x representing a signal with finite energy will result in DWT coefficients from which one can closely approximate (reconstruct) x using the coarse coefficients of this sequence:

$$x(t) = \sum_{m\in Z}\sum_{n\in Z}\langle x, \psi_{m,n}\rangle.\psi_{m,n}(t). \tag{6}$$

The MWT is obtained by choosing a finite number of wavelet coefficients from a set of DWT coefficients. However, to avoid computational complexity, two generating functions ϕ and ψ are used to create the subspaces restricting a to a = 2 and b to b = 1. As a result, we have the subspace $V_m = \text{span}(\phi_{m,n}, n \in Z)$ generated by the coefficients

$$\phi_{m,n}(t) = 2^{-m/2}\phi\left(2^{-m}t - n\right) \tag{7}$$

and the subspace $W_m = \text{span}(\psi_{m,n}, n \in Z)$ generated by the coefficients

$$\psi_{m,n}(t) = 2^{-m/2}\psi\left(2^{-m}t - n\right). \tag{8}$$

The subspace V_m forms a decreasing sequence in $L^2(R)$, with W_m its orthogonal complement from which the two (fast) wavelet transform pairs (MWT) can be generated:

$$\phi(t) = \sqrt{2}\sum_{n\in Z} h_n\phi(2t - n) \tag{9}$$

and

$$\psi(t) = \sqrt{2}\sum_{n\in Z} g_n\phi(2t - n) \tag{10}$$

with $h_n =< \phi_{0,0}, \phi_{-1,n} >$, and $g_n =< \psi_{0,0}, \psi_{-1,n} >$.

In this paper the DWT has been used to suppress noise in a dataset. Due to its ability to extract information in both time and frequency domain, DWT is considered a very powerful tool. The approach consists of decomposing the signal of interest into its detailed and smoothed components (high- and low-frequency). The detailed components of the signal at different levels of resolution localize the time and frequency of the event. Therefore, the DWT can extract the coarse features of the signal (compression) and filter out details at high frequency (noise). DWT has been successfully applied to system analysis for removal of noise [7, 8]. In this paper we present how

DWT can be applied to detect and filter out noise. A detailed discussion of theory and design methodology for the special-purpose filters for this application follows.

Theory of DWT-Based Filters for Noise Suppression

DWT-based filters can be used to localize abrupt changes in signals in time and frequency. Creative techniques have been implemented to suppress noise in datasets using this approach [7–12]. These techniques range in their approach from calculating the wavelet transforms for all circular shifts and selecting the "best" one that minimizes a cost function [9] to using the entropy criterion [10] and adaptively decomposing a signal in a tree structure so as to minimize the entropy of the representation. In this paper a new approach to cancellation of noise in data has been proposed. The discrete Meyer adaptive wavelet (DMAW) is both translation- and scale-invariant and can represent a signal in a multi-scale format. While DMAW is not the best fit for entropy criterion, it is well suited for the proposed noise cancellation purposes [12].

The process to implement DMAW filters starts with discretizing the Meyer wavelets defined by wavelet and scaling functions as

$$\phi(t) = \sqrt{2} \sum_{n \in Z} h_n \phi(2t - n) \tag{11}$$

and

$$\psi(t) = \sqrt{2} \sum_{n \in Z} g_n \phi(2t - n). \tag{12}$$

The masks for these functions are obtained as

$$\left\{ \phi(0), \phi\left(\frac{1}{2^m}\right), \cdots, \phi\left(\frac{M-1}{2^m}\right) \right\} \tag{13}$$

and

$$\left\{ 0, 0, \ldots, 0, \psi(0), \psi\left(\frac{1}{\sigma}\right), \cdots, \psi\left(\frac{N}{\sigma}\right) \right\} \tag{14}$$

As these two masks are convolved, the generating function (mother wavelet) mask (F) can be obtained as

$$F\left(\frac{k}{2^m}\right) (-M \le k \le N) \tag{15}$$

where for every integer k, integers $n_1^k, n_2^k, \ldots.., n_q^k$ can be found to satisfy the inequality

$$-3 < \mu - n_i^k + \frac{k\sigma}{2^m} < \frac{3\sigma}{2^m} (1 \le i \le q). \tag{16}$$

The corresponding values from mother wavelet mask can then be taken to calculate

$$\alpha_i^k = \frac{2^{m/2}}{\sigma} F\left(\frac{\rho_i^k}{2^m}\right),$$

where $\rho_i^k = \left[(\mu - n_i^k)2^m + k\sigma\right]$ $(1 \le i \le q)$ and

$$\frac{c_{-m,k}}{\sqrt{\alpha}} - \sum_{i=1}^{q} c_{ni}\alpha_i^k. \tag{17}$$

Decomposing the re-normalized signal $\frac{c_{-m,k}}{\sqrt{\alpha}}(k \in Z)$ according to the conventional DWT will result in the entire DMAW filter basis for different scales:

$$\frac{c_{-m+1,k}}{\sqrt{\alpha}}, \frac{d_{-m+1,k}}{\sqrt{\alpha}}, \frac{c_{-m+2,k}}{\sqrt{\alpha}}, \frac{d_{-m+2,k}}{\sqrt{\alpha}}, \cdots, \frac{c_{0,k}}{\sqrt{\alpha}}, \frac{d_{0,k}}{\sqrt{\alpha}}. \tag{18}$$

Experimental Results

Noisy Sinusoidal Signal

Figures 8.1, 8.2, and 8.3 show the experimental results for the application of the proposed filter banks to a noisy sinusoidal signal. As is evident from these figures, a signal can be decomposed in as many levels as desired by the application and allowed by the computational constraints. Levels shown from top to bottom represent the coarse to detailed components of the original signal. Once the signal is decomposed to its components, it is easy to do away with pieces that are not needed. For instance, noise, which is the lowermost signal in Fig. 8.1, can be totally discarded. The reconstructed signal is a fairly good approximation of the original signal. Figure 8.2 shows the thresholds and coefficients of the signal being filtered. Figure 8.3 shows the histogram (frequency of components distribution) of the signal.

Comparison to Other Noisy Signals

For comparison purposes, the same filter banks have also been applied to a quad-chirp signal with noise, and the results are shown in Figs. 8.4–8.9. The versions of the signal have been computed and plotted. In each case the coefficients that have remained intact have also been displayed. Finally, in Figs. 8.7–8.16, the histogram for the denoised quad-chirp, auto-regressive, and white noise has been compared to the original signal. The effectiveness of the proposed filter banks and their capability to maintain the important components of the original signal is evident in these figures.

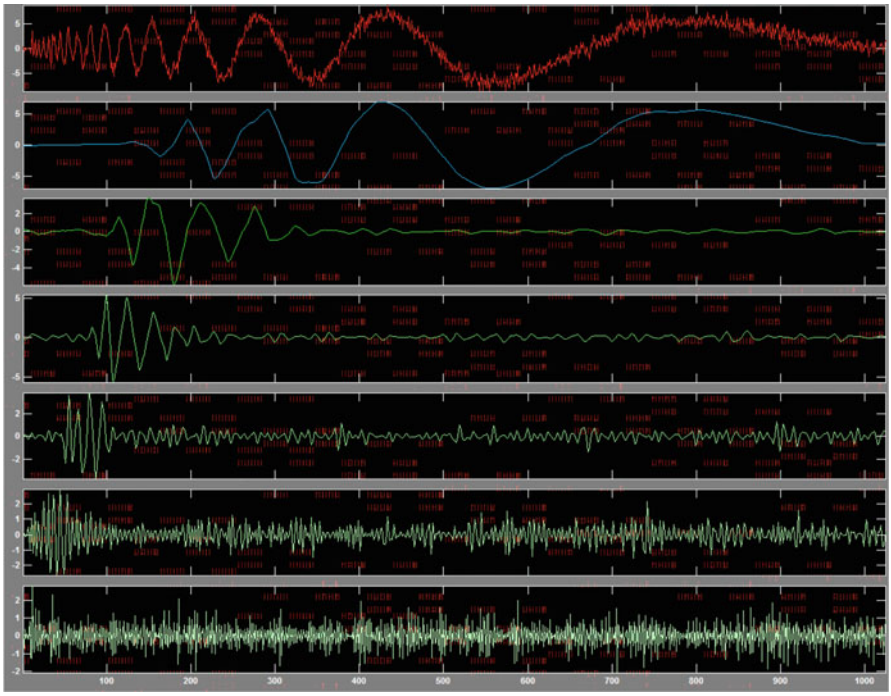

Fig. 8.1 Decomposed signal showing all the components of a mixed sine wave with noise

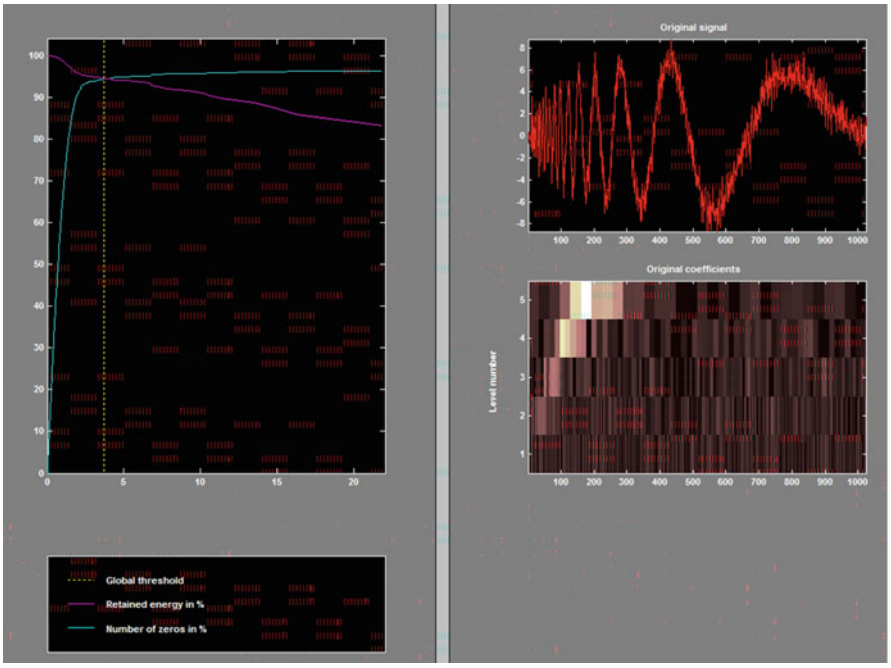

Fig. 8.2 Threshold and coefficients of the decomposed signal

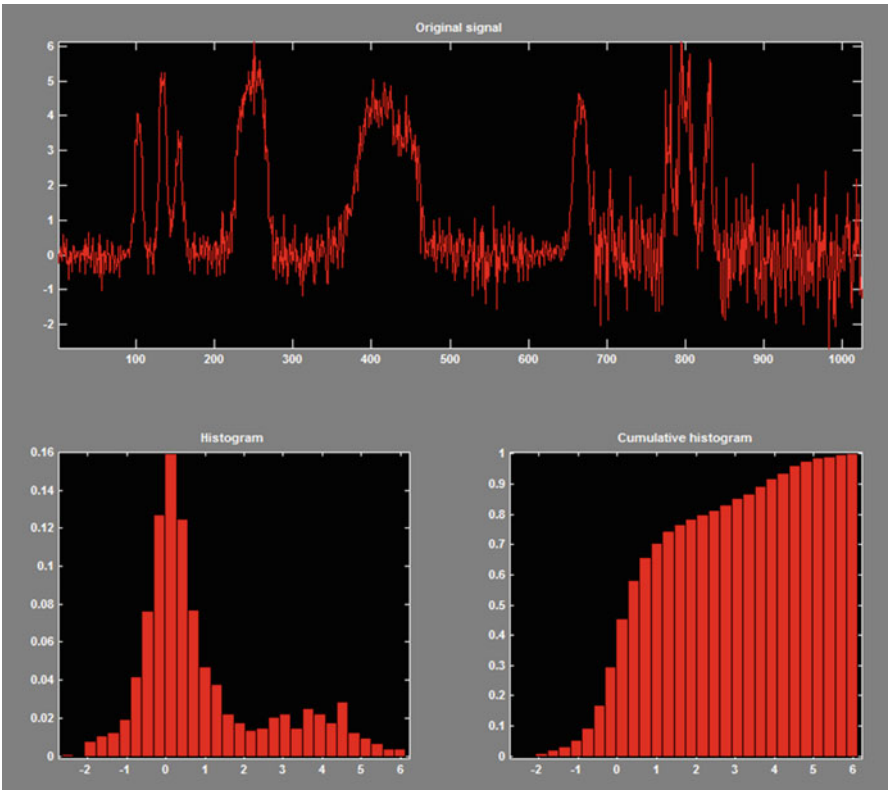

Fig. 8.3 Histogram and cumulative histogram of the signal

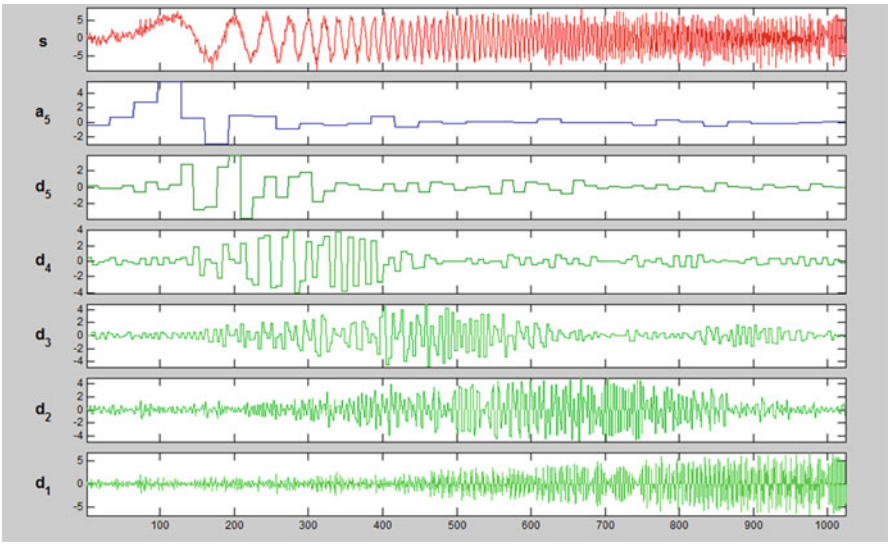

Fig. 8.4 Decomposed signal showing all the components of a quad-chirp wave with noise

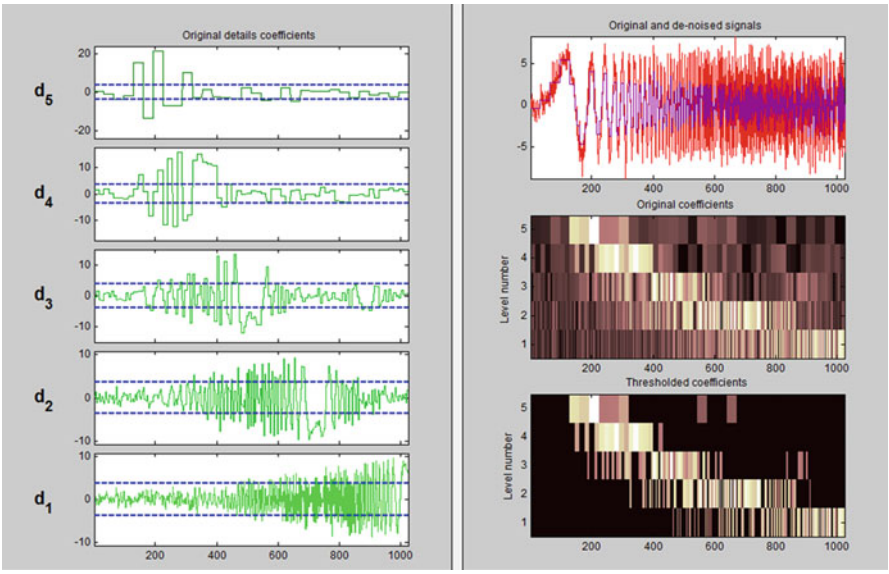

Fig. 8.5 Original and denoised signal with original and thresholded coefficients

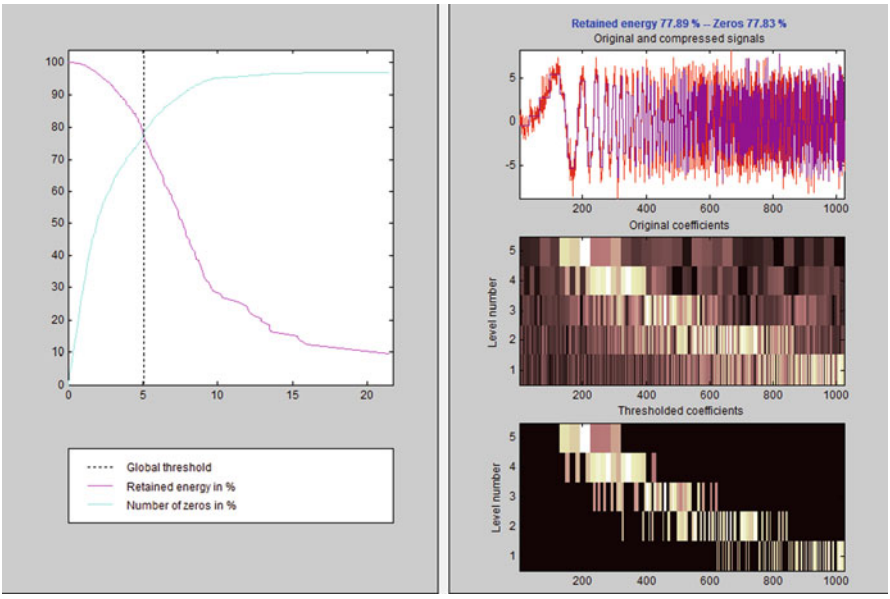

Fig. 8.6 Threshold and coefficients of the decomposed signal showing retained energy and number of zeros

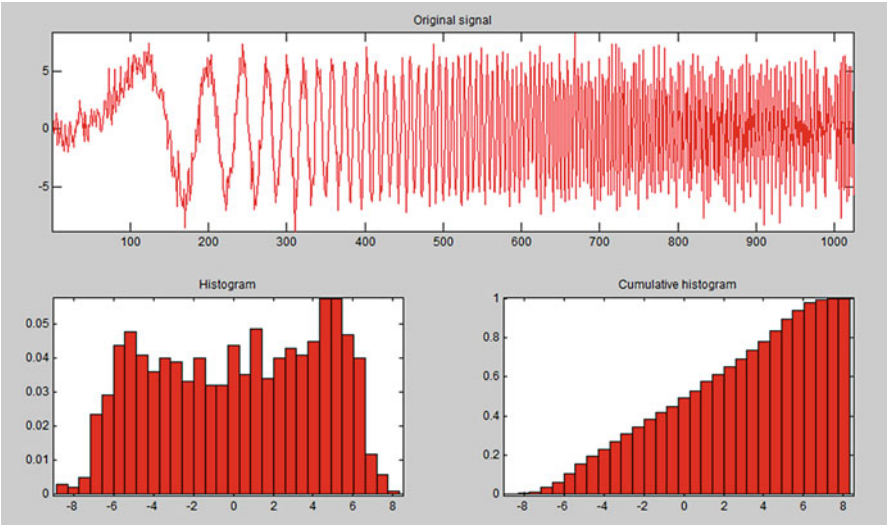

Fig. 8.7 Histogram and cumulative histogram of the original quad-chirp signal

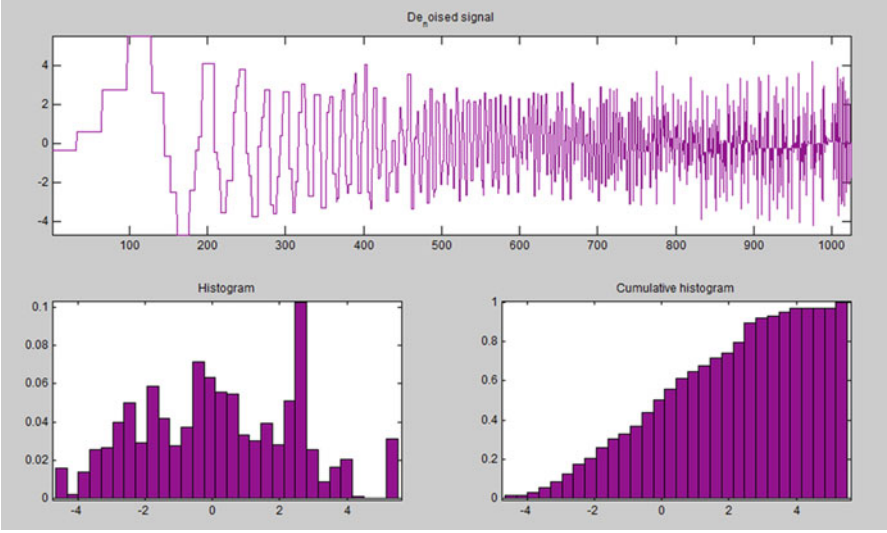

Fig. 8.8 Histogram and cumulative histogram of the denoised quad-chirp signal

Conclusions and Future Work

As expected from the theory, the DMAW filters performed well under noisy conditions. The decomposed signal could be easily freed up from noise. Future plans include the application of these filters to fused datasets and comparison between the

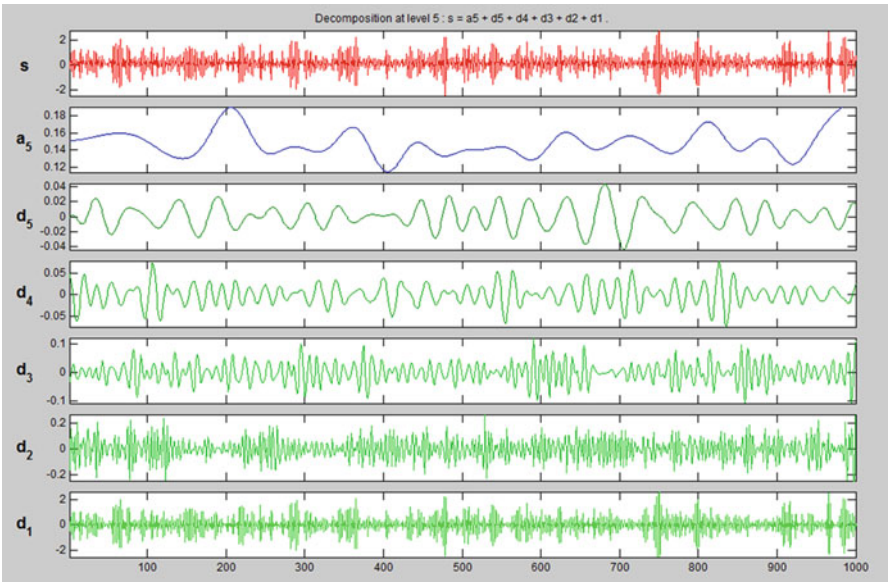

Fig. 8.9 Decomposed signal showing all the components of an auto-regressive wave with noise

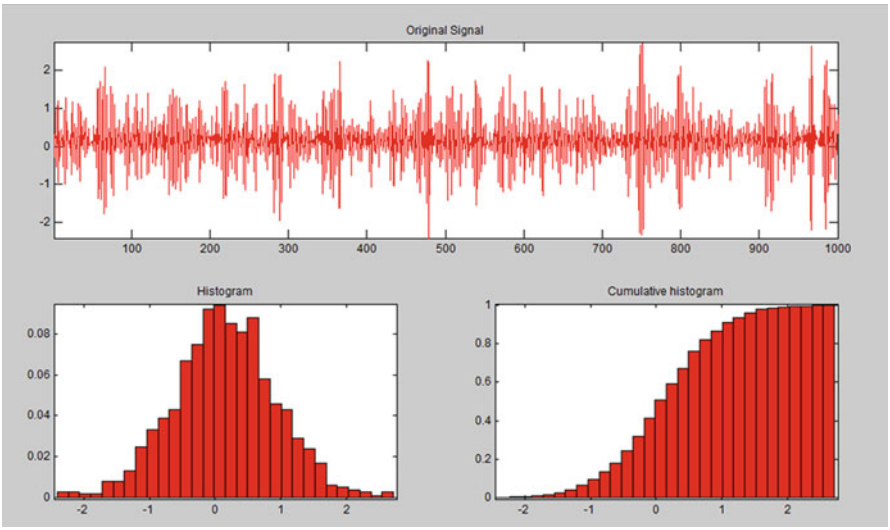

Fig. 8.10 Histogram and cumulative histogram of the original auto-regressive signal

two approaches. Additionally, the results of this study can be used in the decision-making stage to realize the difference this approach can make in accuracy of this process.

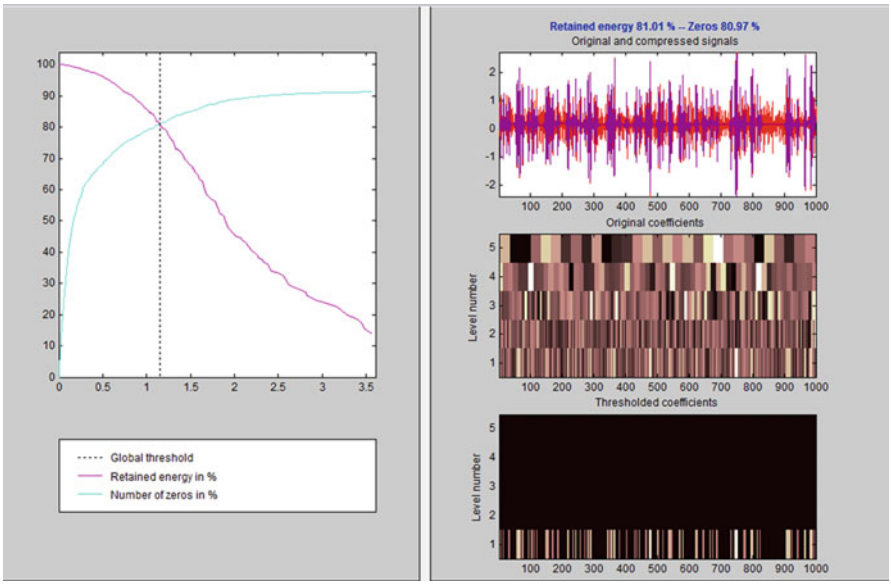

Fig. 8.11 Threshold and coefficients of the decomposed signal showing retained energy and number of zeros

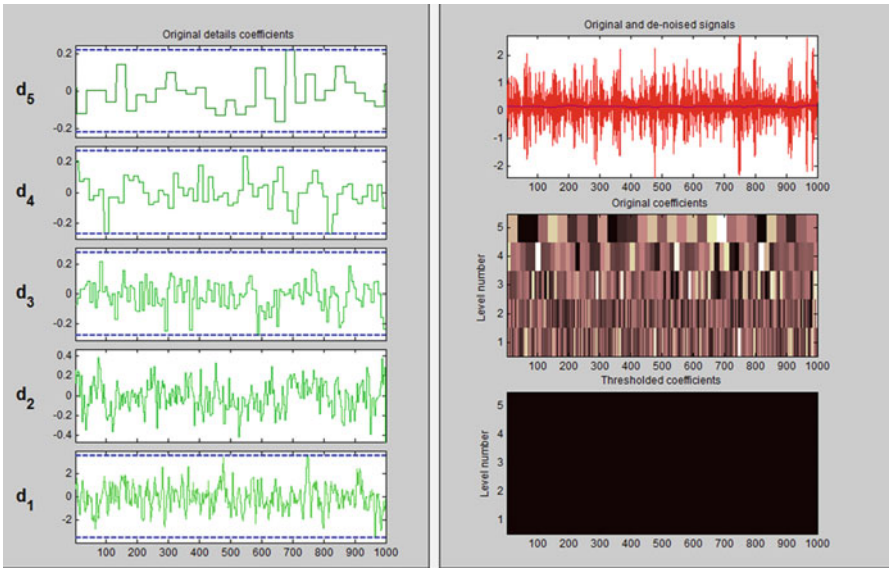

Fig. 8.12 Original and denoised signal with original and thresholded coefficients

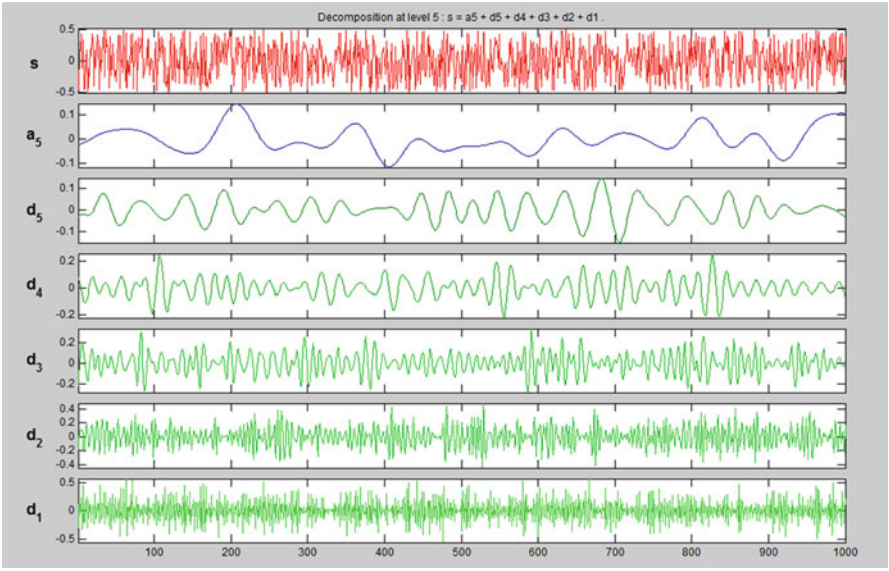

Fig. 8.13 Decomposed signal showing all the components of white noise

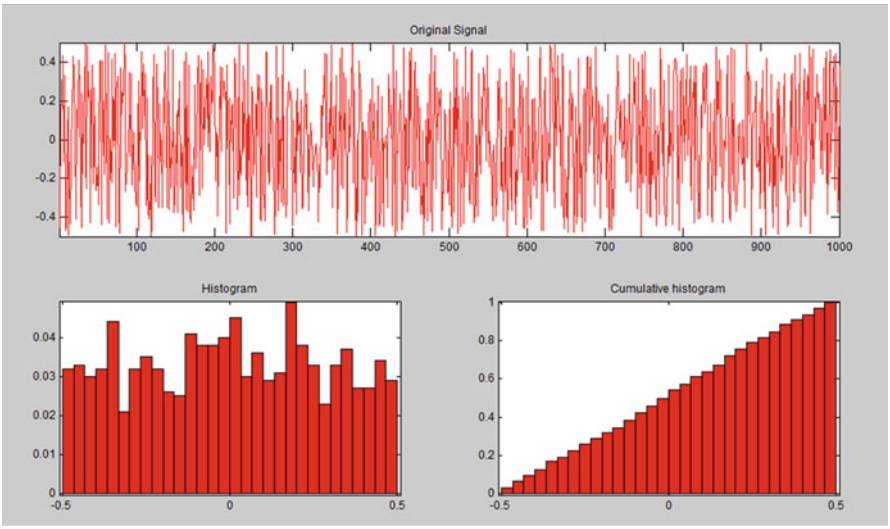

Fig. 8.14 Histogram and cumulative histogram of the original white noise signal

Future work will address issues such as characterizing the parameters for simulation and modeling of the proposed filter for wireless sensor networks, showing how complex examples with correlated sensor data will be filtered for redundancy,

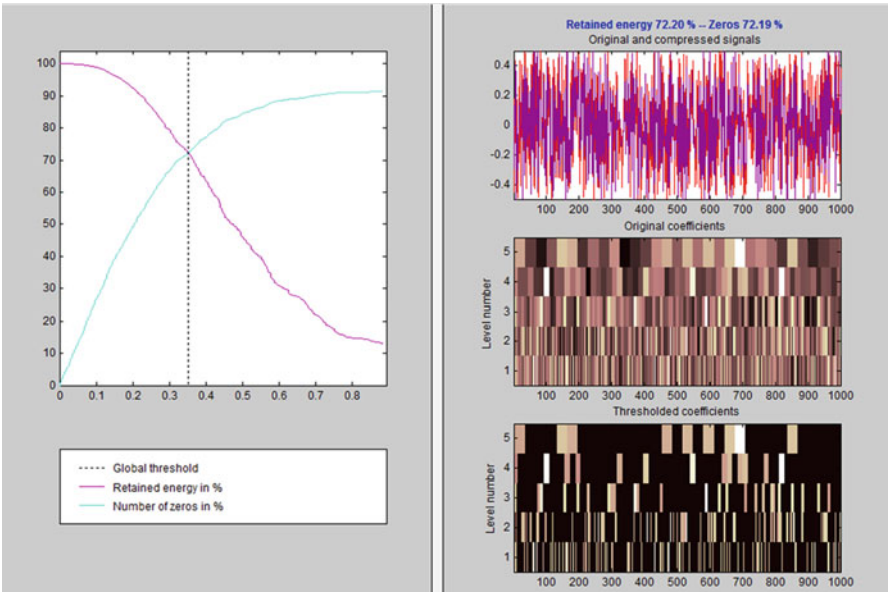

Fig. 8.15 Threshold and coefficients of the decomposed signal showing retained energy and number of zeros

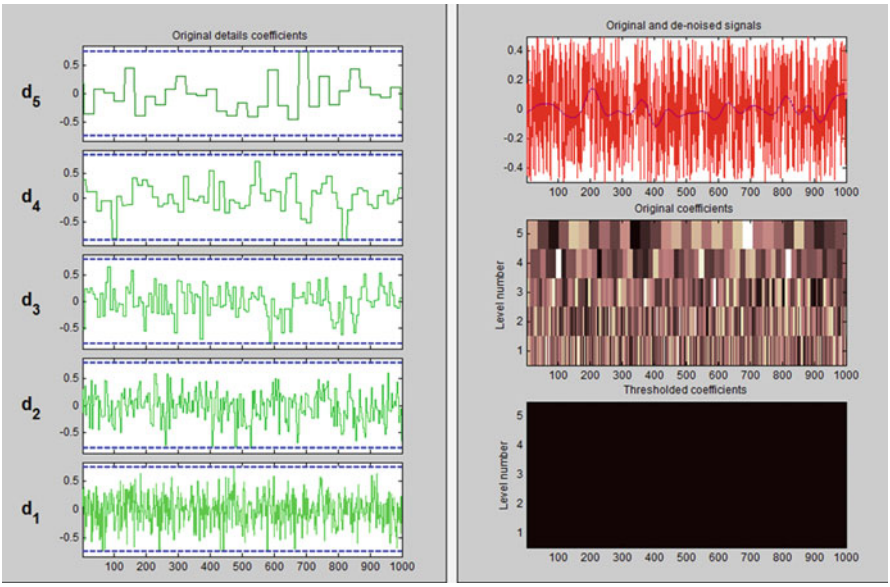

Fig. 8.16 Original and denoised signal with original and thresholded coefficients

and comparing the proposed approach with other similar approaches and giving comparative results to support the claimed advantages, both theoretically and experimentally.

Acknowledgements This is to thank all of the anonymous reviewers and referees who with their constructive comments made this a better chapter for publication.

References

1. Closas, P., Calvo, E., Fernandez-Rubio, J.A., Pages-Zamora, A.: Coupling noise effect in self-synchronizing wireless sensor networks. In: IEEE 8th Workshop on Signal Processing Advances in Wireless Communications, 2007, SPAWC 2007, 17–20 June 2007, pp. 1–5
2. Yamamoto, H., Ohtsuki, T.: Wireless sensor networks with local fusion. In: IEEE Global Telecommunications Conference, 2005, vol. 1, GLOBECOM '05, 28 Nov.–2 Dec. 2005, p. 5
3. Son, S.-H., Kulkarni, S.R., Schwartz, S.C., Roan, M.: Communication-estimation tradeoffs in wireless sensor networks. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings, 2005, (ICASSP apos;05), vol. 5, 18–23 March 2005, pp. 1065–1068
4. Abdallah, A., Wolf, W.: Analysis of distributed noise canceling. In: The 2nd International Conference on Distributed Frameworks for Multimedia Applications, 2006, May 2006, pp. 1–7
5. Schizas, I.D., Giannakis, G.B.: Zhi-Quan Luo optimal dimensionality reduction for multisensor fusion in the presence of fading and noise. In: IEEE International Conference on Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006, 14–19 May 2006, vol. 4, pp. IV–IV
6. Pescosolido, L., Barbarossa, S., Scutari, G.: Average consensus algorithms robust against channel noise. In: IEEE 9th Workshop on Signal Processing Advances in Wireless Communications, 2008. SPAWC 2008, 6–9 July 2008, pp. 261–265
7. Cohen, I., Raz, S., Malah, D.: Shift invariant wavelet packet bases. In:Proceedings 20th IEEE Int. Conf. Acoustics, Speech, Signal Processing, Detroit, MI, May 8–12, 1995, pp. 1081–1084
8. Daubechies, I.: Ten lecture on wavelet. In: CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia, PA (1992)
9. Liang, J., Parks, T.W.: A translation invariant wavelet representation algorithm with applications. IEEE Trans. Signal Process. **44**, 225–232 (1996)
10. Coifman, R.R., Wickhauser, M.V.: Entropy-based algorithms for best basis selection. IEEE Trans. Inform. Theory **38**, 713–718 (1992)
11. Mallat, S.: Zero-crossings of a wavelet transform. IEEE Trans. Inform.Theory **37**, 1019–1033 (1991)
12. Mallat, S., Zhang, S.: Characterization of signals from multiscale edges. IEEE Trans. Pattern Anal. Mach. Intell. **14**, 710–732 (1992)

Chapter 9
Null Controllability of the Heat Equation with Two Constraints on the Control: Application to a Discriminating Sentinel with Given Sensitivity

Sadou Tao and Ousseynou Nakoulima

Introduction

For $N \in \mathbb{N}^*$, let Ω be an open bounded subset of $\mathbf{R}^N$ of boundary Γ of class $\mathscr{C}^2$. Let also ω be an nonempty open bounded subset of Ω. For a time $T > 0$, we set $Q = \Omega \times (0,T)$, $\Sigma = \Gamma \times (0,T)$ and $U = \omega \times (0,T)$. We consider the following parabolic equation:

$$\begin{cases} -\dfrac{\partial q}{\partial t} - \Delta q + a_0 q = h + v\chi_\omega + w\chi_\omega & \text{in } Q \\ q = 0 & \text{on } \Sigma \\ q(T) = 0 & \text{in } \Omega \end{cases} \tag{1}$$

where $a_0 \in L^\infty(Q)$, $h \in L^2(Q)$. The controls v and w belong to $L^2(U)$; χ_ω denotes the characteristic function of ω. In referring to [15, 21], the problem (1) admits a unique solution in the space

$$H^{2,1}(Q) = \{\varphi,\ \frac{\partial \varphi}{\partial x_i},\ \frac{\partial^2 \varphi}{\partial x_i \partial x_j},\ \frac{\partial \varphi}{\partial t} \in L^2(Q)\}$$

equipped with the norm

Sadou Tao
Laboratoire d'Analyse numérique, d'Informatique et de Biomathématiques, Université de Ouagadougou, Ouagadougou, Burkina Faso
e-mail: sadoutao@yahoo.fr

Ousseynou Nakoulima (✉)
Centre d'Etude et de Recherche en Economie, Gestion, Modélisation et Informatique Appliquée (CEREGMIA), Université des Antilles et de la Guyane, Campus de Fouillole, 97159 Pointe à Pitre Guadeloupe (France)
e-mail: onakouli@univ-ag.fr

B. Toni (ed.), *Advances in Interdisciplinary Mathematical Research*, Springer Proceedings in Mathematics & Statistics 37, DOI 10.1007/978-1-4614-6345-0_9,

$$\|\varphi\|_{H^{2,1}(Q)} = \left\{ \int_Q \left[|\varphi|^2 + \sum_{1\le i,j\le N} \left| \frac{\partial \varphi}{\partial x_i} \right|^2 + \sum_{1\le i,j\le N} \left| \frac{\partial^2 \varphi}{\partial x_i \partial x_j} \right|^2 + \left| \frac{\partial \varphi}{\partial t} \right|^2 \right] dxdt \right\}^{\frac{1}{2}}.$$

Let

$$Y_\lambda \text{ and } \mathscr{M} \text{ be two real closed vector subspaces of } L^2(U). \tag{2}$$

$Y_\lambda^\perp$ and $\mathscr{M}^\perp$ represent, respectively, the orthogonal subspaces of Y_λ and $\mathscr{M}$ in $L^2(U)$.

We use the following notation:

$$q = q(x,\ t;\ (v,w)).$$

This means that the solution q of (1) depends on two controls v and w.

The problem is as follows

Let a function h be in $L^2(Q)$, $a_0 \in L^\infty(Q)$; find two controls $\widehat{v}$ and $\widehat{w}$ in $L^2(U)$ with

$$\widehat{v} \in Y_\lambda^\perp \text{ and } \widehat{w} \in \mathscr{M}^\perp \tag{3}$$

such that if $q = q(x,t;(\widehat{v},\widehat{w})) \in H^{2,1}(Q)$ is the unique solution of

$$\begin{cases} -\dfrac{\partial q}{\partial t} - \Delta q + a_0 q = h + \widehat{v}\chi_\omega + \widehat{w}\chi_\omega & \text{in } Q \\ q = 0 & \text{on } \Sigma \\ q(T) = 0 & \text{in } \Omega \end{cases} \tag{4}$$

then

$$q(.,\ 0;\ (\widehat{v},\widehat{w})) = 0 \text{ in } Q \tag{5}$$

and the couple $(\widehat{v},\widehat{w})$ is the minimal norm in $L^2(U)$, i.e.

$$||(\widehat{v},\widehat{w})|| = \min_{(v,w)\in\mathscr{E}} ||(v,w)|| \tag{6}$$

where

$$||(v,w)|| = (||v||^2_{L^2(U)} + ||w||^2_{L^2(U)})^{\frac{1}{2}}$$

and

$$\mathscr{E} = \left\{ \begin{array}{c} (v,w) \in Y_\lambda^\perp \times \mathscr{M}^\perp,\ \textit{such that the pair } ((v,w),\ q(x,t,(v,w))) \\ \text{satisfies (3), (4) } \textit{and} \text{ (5)} \end{array} \right\}.$$

Problems (3)–(6) are null-controllability problems with constraint on the control.

– If $Y_\lambda^\perp = \{0\}$ and $\mathscr{M}^\perp = \{0\}$, this problem becomes a null-controllability problem without constraints on the control. There exists a large literature on this problem. In [13] G. Lebeau and L. Robbiano solved this problem for the heat equation. E. Fernandez-Cara, M. González-Burgos, S. Guerrero et al. and J.P. Puel in [7] used the Carleman estimate for the weak solution of heat equation with nonhomogeneous Neumann boundary conditions to prove the null controllability

of the heat equation. In the nonlinear case, A. Fursikov and O. Yu. Imanuvilov in [11] showed using a Carleman's estimate that, when the control acts on the boundary, null-controllability holds for bounded continuous and sufficiently small initial data. One can also see [3, 6, 8, 9, 20] and the references therein.

– If $Y_\lambda^\perp = \{0\}$ and $\mathscr{M}^\perp \neq \{0\}$ or if $\mathscr{M}^\perp = \{0\}$ and $Y_\lambda^\perp \neq \{0\}$, this problem becomes a null-controllability problem with one constraint on the control; one can see [17, 18].

In this paper we are interested in the case $Y_\lambda^\perp \neq \{0\}$ and $\mathscr{M}^\perp \neq \{0\}$, in other words, the case where two constraints act on the state q. The main result of this paper is the following:

We assume that the subspaces

$$Y_\lambda^\perp \text{ and } \mathscr{M}^\perp \text{ are finite dimensional} \tag{7}$$

and

$$\begin{cases} (\forall \rho \in Y_\lambda), (\dfrac{\partial \rho}{\partial t} - \Delta \rho + a_0 = 0 \text{ in } \omega \times (0,T) \Longrightarrow \rho = 0 \text{ in } \omega \times (0,T)) \\ (\forall \rho \in \mathscr{M}), (\dfrac{\partial \rho}{\partial t} - \Delta \rho + a_0 = 0 \text{ in } \omega \times (0,T) \Longrightarrow \rho = 0 \text{ in } \omega \times (0,T)). \end{cases} \tag{8}$$

Theorem 1. *We assume that (7), (8) are satisfied. Then there exists a weight function θ (which does not vanish in θ its domain with $\dfrac{1}{\theta}$ bounded) such that for any function $h \in L^2(Q)$ with $\theta h \in L^2(Q)$, the null-controllability problems (3)–(6) admit a solution $(\widehat{v}_\theta, \widehat{w}_\theta)$. Moreover, the pair $(\widehat{v}_\theta, \widehat{w}_\theta)$ is such that*

$$\begin{aligned} \widehat{v}_\theta &= -(\widehat{\rho}_\theta \chi_\omega - P_1 \widehat{\rho}_\theta) \text{ in } \omega \times (0,T) \\ \widehat{w}_\theta &= -(\widehat{\rho}_\theta \chi_\omega - P_2 \widehat{\rho}_\theta) \text{ in } \omega \times (0,T), \end{aligned} \tag{9}$$

where P_1 and P_2 are, respectively, the orthogonal projection operator from $L^2(\omega \times (0,T))$ into Y_λ and into $\mathscr{M}$ and $\widehat{\rho}_\theta$ satisfies

$$\begin{cases} \dfrac{\partial \widehat{\rho}_\theta}{\partial t} - \Delta \widehat{\rho}_\theta + a_0 \widehat{\rho}_\theta = 0 \text{ in } Q \\ \widehat{\rho}_\theta = 0 \text{ on } \Sigma. \end{cases} \tag{10}$$

The rest of this article is organized as follows. In section "Preliminaries" we prove Theorem 1 after establishing a Carleman inequality adapted to (3). In section "Antiperiodic Functions", we show by the penalization method an approximation of solution of the controllability problems (3)–(5) of Theorem 1. In the end, in section "Existence of Anti-periodic Solutions" we give an application of these results at a sentinel that will be defined.

Null Controllability with Two Constraints on the Control

Adapted Carleman Inequality

It is well known that the analysis of the null-controllability problem is associated to the Carleman appropriate inequality. The main contributions are due to O. Yu. Emanuvilov, who developed the use of Carleman inequality to the null-controllability problem in [12].

To establish the Carleman inequality we adopt the following notations:

$$\begin{cases} L_0 = \dfrac{\partial}{\partial t} - \Delta \\ L = \dfrac{\partial}{\partial t} - \Delta + a_0 I \\ \mathscr{V} = \left\{ \rho \in C^{\infty}(\overline{Q}) / \ \rho = 0 \ on \ \Sigma \right\} \end{cases} \tag{11}$$

where $a_0 \in L^{\infty}(Q)$ is defined in (1). The classical Carleman inequality can be formulated by the following.

Proposition 1.

There exists a weight function θ, $\theta \in C^2(Q)$, and $\dfrac{1}{\theta}$ is bounded; there exists a constant $c = c(\Omega, a_0, T)$ such as for any $\rho \in \mathscr{V}$, the following inequality holds:

$$\int_0^T \int_{\Omega} \frac{1}{\theta^2} |\rho|^2 \, dxdt \leq C \left(\int_0^T \int_{\Omega} |L\rho|^2 \, dxdt + \int_0^T \int_{\omega} |\rho|^2 \, dxdt \right). \tag{12}$$

All these results are well known. We refer to E. Fernández-Cara and E. Zuazua in [10] and O. Nakoulima in [18].

To handle the constraint (3), we use the Carleman inequality adapted to the spaces Y_λ and $\mathscr{M}$ defined in (2). The following lemma is key to our results.

Lemma 1. *Assume that (7) and (8) hold, then there exists a positive constant* $C=C(\Omega, \omega, a_0)$ *such that for any* $\rho \in \mathscr{V}$

$$\begin{aligned} &\int_0^T \int_{\Omega} \frac{1}{\theta^2} |\rho|^2 \, dxdt \\ &\leq C \left(\int_0^T \int_{\Omega} |L\rho|^2 \, dxdt + \int_0^T \int_{\omega} |\rho - P_1\rho|^2 \, dxdt + \int_0^T \int_{\omega} |\rho - P_2\rho|^2 \, dxdt \right) \end{aligned} \tag{13}$$

where P_1 *and* P_2 *are, respectively, the orthogonal projections from* $L^2(\omega \times (0,T))$ *into* Y_λ *and into* $\mathscr{M}$.

Proof. The proof uses a well-known compactness-uniqueness argument and the inequality (12). Indeed, suppose that (13) does not hold;
then

$$\begin{cases} \forall n \in \mathbb{N}^*,\ \exists \rho_n \in \mathscr{V},\ \int_0^T \int_\Omega \dfrac{1}{\theta^2} |\rho_n|^2\, dxdt = 1, \\ \int_0^T \int_\Omega |L\rho_n|^2\, dxdt \le \dfrac{1}{n},\ \int_0^T \int_\omega |\rho_n - P_1\rho_n|^2\, dxdt \le \dfrac{1}{n} \\ \text{and } \int_0^T \int_\omega |\rho_n - P_2\rho_n|^2\, dxdt \le \dfrac{1}{n}. \end{cases} \tag{14}$$

The proof consists in showing that (14) yields a contradiction. We proceed in four steps:

Step 1. We have

$$\int_0^T \int_\omega \frac{1}{\theta^2} |P_1\rho_n|^2\, dxdt \le 2\Big(\int_0^T \int_\omega \frac{1}{\theta^2} |\rho_n|^2\, dxdt + \int_0^T \int_\omega \frac{1}{\theta^2} |\rho_n - P_1\rho_n|^2\, dxdt\Big)$$
$$\int_0^T \int_\omega \frac{1}{\theta^2} |P_2\rho_n|^2\, dxdt \le 2\Big(\int_0^T \int_\omega \frac{1}{\theta^2} |\rho_n|^2\, dxdt + \int_0^T \int_\omega \frac{1}{\theta^2} |\rho_n - P_2\rho_n|^2\, dxdt\Big).$$

Since $\dfrac{1}{\theta^2}$ is bounded, it follows from (14) that

$$\begin{aligned} \int_0^T \int_\omega \frac{1}{\theta^2} |P_1\rho_n|^2\, dxdt &\le C \\ \int_0^T \int_\omega \frac{1}{\theta^2} |P_2\rho_n|^2\, dxdt &\le C. \end{aligned} \tag{15}$$

Since $P_1\rho_n \in Y_\lambda$ and $P_2\rho_n \in \mathscr{M}$ and that Y_λ and $\mathscr{M}$ are finite dimensional, $(P_1\rho_n)_n$, $(P_2\rho_n)_n$ are bounded in $L^2(\omega \times (0,T))$ and so $(\rho_n)_n$ because $\int_0^T \int_\omega \dfrac{1}{\theta^2} |\rho_n|^2\, dxdt \le \int_0^T \int_\omega \frac{1}{\theta^2} |\rho_n - P_i\rho_n|^2\, dxdt + \int_0^T \int_\omega \dfrac{1}{\theta^2} |P_i\rho_n|^2\, dxdt$, $i = 1,2$.

Step 2. We can extract a subsequence, still denoted $(\rho_n)_n$, such that on the one hand,

$$\rho_n \rightharpoonup g \text{ weakly in } L^2(\omega \times (0,T)), \tag{16}$$

and on the other hand,

$$\begin{aligned} &\rho_n - P_1\rho_n \to 0 \text{ strongly in } L^2(\omega \times (0,T)) \\ &\qquad\qquad \text{and} \\ &\rho_n - P_2\rho_n \to 0 \text{ strongly in } L^2(\omega \times (0,T)). \end{aligned} \tag{17}$$

Next we deduce from the compactness of P_1 and of P_2 (because Y_λ and $\mathscr{M}$ are of finite dimensional) that there exist $\sigma \in Y_\lambda$ and $\delta \in \mathscr{M}$ such that

$$\begin{aligned} &P_1\rho_n \to \sigma \text{ strongly in } L^2(\omega \times (0,T)) \\ &\qquad\qquad \text{and} \\ &P_2\rho_n \to \delta \text{ strongly in } L^2(\omega \times (0,T)). \end{aligned} \tag{18}$$

We deduce from (17) and (18) that $\rho_n \to g = \sigma$ and $\rho_n \to g = \delta$ strongly in $L^2(\omega \times (0,T))$. Due to the continuity of P_1 and of P_2, we have $P_1\rho_n \to P_1 g$ and $P_2\rho_n \to P_2 g$ strongly in $L^2(\omega \times (0,T))$. Therefore, $P_1 g = g$, $P_2 g = g$ and $g \in Y_\lambda \cap \mathscr{M}$.
Step 3. In fact, we have $g = 0$. Indeed, from (14), we also have $L\rho_n \to 0$ strongly in $L^2(Q)$. Thus, $L\rho_n \to 0$ strongly in $L^2(\omega \times (0,T))$. We deduce $L\rho_n \rightharpoonup 0$ weakly in $D'(\omega \times (0,T))$ and so $Lg = 0$. The assumption (8) implies that $g = 0$ on $\omega \times (0,T)$. Finally, $\rho_n \to 0$ strongly in $L^2(\omega \times (0,T))$.
Step 4. Since $\rho_n \in \mathscr{V}$, it follows from the observability inequality (12) that

$$\int_0^T \int_\Omega \frac{1}{\theta^2} |\rho_n|^2 \, dxdt \leq C \left(\int_0^T \int_\Omega |L\rho_n|^2 \, dxdt + \int_0^T \int_\omega |\rho_n|^2 \, dxdt \right).$$

Then, from the conclusions in the third step, we deduce that
$\int_0^T \int_\Omega \frac{1}{\theta^2} |\rho_n|^2 \, dxdt \to 0$ when $n \to +\infty$. The contradiction occurs because of the first condition in (14), where $\int_0^T \int_\Omega \frac{1}{\theta^2} |\rho_n|^2 \, dxdt = 1$. The proof of (13) is complete.

Proof of Theorem 1

The main tool is the observability inequality (13) adapted to the constraints.
Consider now the following symmetric bilinear form:

$$\begin{aligned} a(\rho, \widehat{\rho}) = & \int_Q L\rho L\widehat{\rho} dxdt + \int_{L^2(\omega \times (0,T))} (\rho - P_1\rho)(\widehat{\rho} - P_1\widehat{\rho}) dxdt \\ & + \int_{L^2(\omega \times (0,T))} (\rho - P_2\rho)(\widehat{\rho} - P_2\widehat{\rho}) dxdt. \end{aligned} \tag{19}$$

Due to Lemma 1, this bilinear form is a scalar product on $\mathscr{V}$. Let V be the Hilbert space obtained from taking the closure of $\mathscr{V}$ under the norm

$$\rho \longmapsto ||\rho||_V = \sqrt{a(\rho, \rho)}. \tag{20}$$

Remark 1. Observe that the norm $\|.\|_V$ is related to the right-hand side of inequality (13). Similarly, the left-hand side of (13) leads to the norm

$$||\rho||_\theta = \left(\int_Q \frac{1}{\theta^2} |\rho|^2 dxdt \right)^{\frac{1}{2}}. \tag{21}$$

The completion of $\mathscr{V}$ is the weighted Hilbert space usually denoted by $L^2_{\frac{1}{\theta}}(Q)$.
The inequality (13) shows that

$$||\rho||_\theta \leq C||\rho||_V. \tag{22}$$

This inequality extends to $\rho \in V$. This shows that V is continuously imbedded in $L^2_{\frac{1}{\theta}}(Q)$. Let us now consider $h \in L^2(Q)$ such that $\theta h \in L^2(Q)$. Then due to (13) and the Cauchy–Schwartz inequality, we deduce that the linear form defined on V by

$$\rho \longmapsto \int_Q h\rho dxdt$$

is continuous. By the Lax–Milgram theorem, for any $h \in L^2(Q)$ such that $\theta h \in L^2(Q)$, there exists one and only one solution $\rho_\theta \in V$ to the variational problem:

$$\forall \rho \in V \;\; a(\rho_\theta, \rho) = \int_Q h\rho dxdt. \tag{23}$$

Proposition 2. *Assume that (7) and (8) hold. Let h be in $L^2(Q)$ such that $\theta h \in L^2(Q)$. Let ρ_θ be the unique solution of (23), $P_1\rho_\theta$ the projection from $\rho_\theta\chi_\omega$ into Y_λ and $P_2\rho_\theta$ the projection from $\rho_\theta\chi_\omega$ into $\mathscr{M}$. Set*

$$v_\theta = -(\rho_\theta\chi_\omega - P_1\rho_\theta) \tag{24}$$

$$w_\theta = -(\rho_\theta\chi_\omega - P_2\rho_\theta) \tag{25}$$

and

$$q_\theta = L\rho_\theta. \tag{26}$$

The pair $((v_\theta, w_\theta); q_\theta)$ is such that (3)–(5) hold. Moreover, we have

$$||\rho_\theta||_V \leq C||\theta h||_{L^2(Q)}, \tag{27}$$

$$||v_\theta||_{L^2(\omega\times(0,T))} \leq C||\theta h||_{L^2(Q)}, \tag{28}$$

$$||w_\theta||_{L^2(\omega\times(0,T))} \leq C||\theta h||_{L^2(Q)}, \tag{29}$$

$$||q_\theta||_{H^{2,1}(Q)} \leq C||\theta h||_{L^2(Q)}, \tag{30}$$

where C is a positive constant depending only for Ω, ω, a_0, T, Y_λ and $\mathscr{M}$.

Proof. Since $\rho_\theta \in V$ it follows that $v_\theta = -(\rho_\theta\chi_\omega - P_1\rho_\theta) \in L^2(\omega\times]0,T[)$, $w_\theta = -(\rho_\theta\chi_\omega - P_2\rho_\theta) \in L^2(\omega\times]0,T[)$ and $q_\theta \in L^2(Q)$. Since $P_1\rho_\theta \in Y_\lambda$ and $P_2\rho_\theta \in \mathscr{M}$, we have $v_\theta = -(\rho_\theta\chi_\omega - P\rho_\theta) \in Y_\lambda^\perp$ and $w_\theta = -(\rho_\theta\chi_\omega - P_2\rho_\theta) \in \mathscr{M}^\perp$. Substitute $-(\rho_\theta\chi_\omega - P_1\rho_\theta)$, $-(\rho_\theta\chi_\omega - P_2\rho_\theta)$ and $L\rho_\theta$, respectively, by v_θ, w_θ and q_θ in the formula (23), it follows that

$$\int_Q q_\theta L\rho dxdt - \int_0^T \int_\omega v_\theta(\rho - P_1\rho)dxdt$$

$$-\int_0^T \int_\omega w_\theta(\rho - P_2\rho)dxdt = \int_Q h\rho dxdt, \;\; \forall \rho \in V.$$

Taking into account that $P_1\rho \in Y_\lambda$ and $P_2\rho \in \mathscr{M}$, the above identity reduces to

$$\int_Q q_\theta L\rho dxdt = \int_Q h\rho dxdt + \int_0^T \int_\omega v_\theta \rho dxdt + \int_0^T \int_\omega w_\theta \rho dxdt \ \ \forall \rho \in V. \quad (31)$$

We show now that q_θ is in fact the weak solution by transposition of a backward heat problem. More precisely, if $f \in L^2(Q)$, let z be the solution of

$$\begin{cases} \dfrac{\partial z}{\partial t} - \Delta z + a_0 = f \ \text{ in } \ Q \\ z = 0 \ \text{ on } \Sigma \\ z(0) = 0 \ \text{ in } \ \Omega. \end{cases} \quad (32)$$

Then $z \in V$, and so

$$\int_Q q_\theta f dxdt = \int_Q h\rho dxdt + \int_0^T \int_\omega v_\theta \rho dxdt \quad (33)$$
$$+ \int_0^T \int_\omega w_\theta \rho dxdt \ \ \forall \rho \in V.$$

Thus, q_θ is the weak solution by transposition of problem (4) with $v = v_\theta$ and $w = w_\theta$ (see J. L. Lions [15]). We know that the solution of this equation is in $H^{2,1}(Q)$.

In other words q_θ is the solution of the following problem:

$$\begin{cases} -\dfrac{\partial q_\theta}{\partial t} - \Delta q_\theta + a_0 q_\theta = h_0 + v_\theta \chi_\omega + w_\theta \chi_\omega \ \text{ in } \ Q, \\ q_\theta = 0 \qquad\qquad\qquad\qquad\quad \text{on } \Sigma, \\ q_\theta(T) = 0 \qquad\qquad\qquad\qquad\quad \text{in } \ \Omega. \end{cases} \quad (34)$$

Multiplying the first equation of (34) by $\rho \in \mathscr{V}$ and integrating by parts over Q, it follows that

$$-\int_\Omega q_\theta(T)\rho(T)dx + \int_\Omega q_\theta(0)\rho(0)dx + \int_Q q_\theta L\rho dxdt$$
$$= \int_Q h\rho dxdt + \int_0^T \int_\omega v_\theta \rho dxdt + \int_0^T \int_\omega w_\theta \rho dxdt \ \ \forall \rho \in \mathscr{V}.$$

Since $\rho \in \mathscr{V}$ we deduce from (31) that

$$\int_\Omega q_\theta(0)\rho(0)dx = 0 \ \ \rho \in \mathscr{V}\ .$$

Therefore, $q_\theta(0) = 0$ in Ω. Hence the first statement of Proposition 2 is proved. It remains to prove the estimates (27)–(30). We set $\rho = \rho_\theta$ in (23). It follows from (23) that

$$\begin{aligned} a(\rho_\theta,\rho_\theta) &= ||q_\theta||^2_{L^2(Q)} + ||v_\theta||^2_{L^2(\omega\times(0,T))} + ||w_\theta||^2_{L^2(\omega\times(0,T))} \\ &\leq ||\theta h||_{L^2(Q)}||\rho_\theta||_\theta \\ &\leq C||\theta h||_{L^2(Q)}||\rho_\theta||_V. \end{aligned} \tag{35}$$

From (20) we obtain (27) and thus (28) and (29). Finally, (30) is a consequence of (28), (29) and classical properties of the heat equation.

The adapted observability inequality (13) shows that the choice of the scalar product on $\mathscr{V}$ is not unique. Thus there exist infinitely many control functions (v_θ, w_θ) such that (3)–(5) hold.

Let consider the set of control variables

$$(v_\theta, w_\theta) \in L^2(\omega\times(0,T)) \times L^2(\omega\times(0,T))$$

such that (3)–(5) hold. By Proposition 2, this set is nonempty, and it is clearly convex and closed in $L^2(\omega\times(0,T)) \times L^2(\omega\times(0,T))$. Therefore, there exists a unique *pair* of variable *controls* (v_θ, w_θ) of minimal norm in $L^2(\omega\times(0,T))$ such that $(\widehat{v}_\theta, \widehat{w}_\theta,\ \widehat{q}_\theta = \widehat{q}_\theta(x,t;(\widehat{v}_\theta,\widehat{w}_\theta)))$ is the solution of (3)–(6).

Optimality System for the Optimal Solution

Penalization

The optimal solution $(\widehat{v}_\theta, w_\theta, \widehat{q}_\theta)$ can be approximated considering the penalization method by J. L. Lions [16]. Let $\varepsilon > 0$. Define the functional

$$\begin{aligned} J_\varepsilon(v,w,q) = &\frac{1}{2}\|v\|^2_{L^2(\omega\times(0,T))} + \frac{1}{2}\|w\|^2_{L^2(\omega\times(0,T))} \\ &+\frac{1}{2\varepsilon}\| - \frac{\partial q}{\partial t} - \Delta q + a_0 q - h_0 - v\chi_\omega - w\chi_\omega\|^2_{L^2(Q),} \end{aligned} \tag{36}$$

for any (v,w,q) such that

$$\begin{cases} v \in Y_\lambda^\perp,\ w \in \mathscr{M}^\perp \\ -\dfrac{\partial q}{\partial t} - \Delta q + a_0 q \in L^2(Q) \\ q = 0 \text{ on } \Sigma,\ q(0) = q(T) = 0 \text{ in } \Omega. \end{cases} \tag{37}$$

Consider the minimization problem

$$\min J_\varepsilon(v,w,q), (v,w,q) \ \text{ subject to } (37). \tag{38}$$

We show the following result:

Proposition 3. *Under the assumptions of Proposition 2, the minimization problem (38) has an optimal solution. In other words, there exists* $(v_\varepsilon, w_\varepsilon, q_\varepsilon)$ *such that*

$$J_\varepsilon(v_\varepsilon, w_\varepsilon, q_\varepsilon) = \min\{J_\varepsilon(v,w,q)|(v,w,q)\text{subject to } 37\}. \tag{39}$$

Proposition 4. *Under the assumptions of Proposition 2, the triplet* $(v_\varepsilon, w_\varepsilon, q_\varepsilon)$ *is the optimal solution (39) if and only if there exists a function* ρ_ε *such that triplet* $((v_\varepsilon, w_\varepsilon), q_\varepsilon, \rho_\varepsilon)$ *satisfies the following approximate optimality condition:*

$$\begin{cases} -\dfrac{\partial q_\varepsilon}{\partial t} - \Delta q_\varepsilon + a_0 q_\varepsilon &= h\chi_O + v_\varepsilon\chi_\omega + w_\varepsilon\chi_\omega + \varepsilon\rho_\varepsilon \text{ in } Q \\ q_\varepsilon &= 0 \quad \text{on} \quad \Sigma \\ q_\varepsilon(T) &= 0 \quad \text{in} \quad \Omega \end{cases} \tag{40}$$

$$q_\varepsilon(0) = 0 \text{ in } \Omega \tag{41}$$

$$\begin{cases} \dfrac{\partial \rho_\varepsilon}{\partial t} - \Delta\rho_\varepsilon + a_0\rho_\varepsilon &= 0 \text{ in } Q \\ \rho_\varepsilon &= 0 \text{ on } \Sigma \end{cases} \tag{42}$$

$$v_\varepsilon = -(\rho_\varepsilon\chi_\omega - P_1\rho_\varepsilon) \text{ in } \omega\times]0,T[, \tag{43}$$

$$w_\varepsilon = -(\rho_\varepsilon\chi_\omega - P_2\rho_\varepsilon) \text{ in } \omega\times]0,T[. \tag{44}$$

Proof. Express the Euler–Lagrange optimality conditions which characterize $(v_\varepsilon, w_\varepsilon, q_\varepsilon)$:

$$\frac{d}{d\lambda}J_\varepsilon(v_\varepsilon, w_\varepsilon, q_\varepsilon + \lambda\varphi)|_{\lambda=0} = 0$$

for any $\varphi \in C^\infty(\overline{Q})$ such that $\varphi = 0$ on Σ, $\varphi(0) = \varphi(T) = 0$ in Ω,

$$\frac{d}{d\lambda}J_\varepsilon(v_\varepsilon + \lambda v, w_\varepsilon, q_\varepsilon)|_{\lambda=0} = 0 \ \ \forall\, v \in Y_\lambda^\perp \ ,$$

$$\frac{d}{d\lambda}J_\varepsilon(v_\varepsilon, w_\varepsilon + \lambda w, q_\varepsilon)|_{\lambda=0} = 0 \ \ \forall w \in \mathscr{M}^\perp.$$

After some calculations, we have

$$\int_Q \frac{1}{\varepsilon}(-\frac{\partial q_\varepsilon}{\partial t} - \Delta q_\varepsilon + a_0 q_\varepsilon - h_0 - v_\varepsilon\chi_\omega - w_\varepsilon\chi_\omega)(-\frac{\partial\varphi}{\partial t} - \Delta\varphi + a_0\varphi)dxdt = 0 \tag{45}$$

for any $\varphi \in C^\infty(\overline{Q})$ such that $\varphi = 0$ on Σ, $\varphi(0) = \varphi(T) = 0$ in Ω,

$$\int_0^T \int_\omega v_\varepsilon v dxdt - \int_Q \frac{1}{\varepsilon}(-\frac{\partial q_\varepsilon}{\partial t} - \Delta q_\varepsilon + a_0 q_\varepsilon - h_0 - v_\varepsilon\chi_\omega - w_\varepsilon\chi_\omega)v dxdt = 0 \tag{46}$$

$\forall v \in Y_\lambda^\perp$ and

$$\int_0^T \int_\omega w_\varepsilon w dxdt - \int_Q \frac{1}{\varepsilon}(-\frac{\partial q_\varepsilon}{\partial t} - \Delta q_\varepsilon + a_0 q_\varepsilon - h_0 - v_\varepsilon \chi_\omega - w_\varepsilon \chi_\omega) w dxdt = 0 \quad (47)$$

$\forall w \in \mathscr{M}^\perp$.

We set

$$\rho_\varepsilon = -\frac{1}{\varepsilon}(-\frac{\partial q_\varepsilon}{\partial t} - \Delta q_\varepsilon + a_0 q_\varepsilon - h_0 - v_\varepsilon \chi_\omega - w_\varepsilon \chi_\omega).$$

Then (45), (46), and (47) become, respectively,

$$\int_Q \rho_\varepsilon(-\frac{\partial \varphi}{\partial t} - \Delta \varphi + a_0 \varphi) dxdt = 0 \quad (48)$$

for any $\varphi \in C^\infty(\overline{Q})$ such that $\varphi = 0$ on Σ, $\varphi(0) = \varphi(T) = 0$ in Ω,

$$\int_0^T \int_\omega v_\varepsilon v dxdt + \int_Q \rho_\varepsilon v dxdt = 0, \ \forall v \in Y_\lambda^\perp \quad (49)$$

and

$$\int_0^T \int_\omega w_\varepsilon w dxdt + \int_Q \rho_\varepsilon w dxdt = 0, \ \ \forall w \in \mathscr{M}^\perp. \quad (50)$$

Consider the first part of (48), we deduce that

$$\frac{\partial \rho_\varepsilon}{\partial t} - \Delta \rho_\varepsilon + a_0 \rho_\varepsilon = 0 \ in \ Q$$

so $\rho_\varepsilon \in L^2(Q)$ with $L\rho_\varepsilon \in L^2(Q)$. Then we can define on the one hand ρ *on* Γ and on the other hand we prove that $\rho_\varepsilon = 0$ on Σ.

Now, we consider (49)

$$\int_0^T \int_\omega (v_\varepsilon + \rho_\varepsilon) v dxdt = 0 \ \ \forall v \in Y_\lambda^\perp.$$

Thus $v_\varepsilon + \rho_\varepsilon \chi_\omega \in Y_\lambda$. Since $v_\varepsilon \in Y_\lambda^\perp$, we have $v_\varepsilon + \rho_\varepsilon \chi_\omega = P_1(v_\varepsilon + \rho_\varepsilon \chi_\omega) = P_1 \rho_\varepsilon$ and therefore $v_\varepsilon = -(\rho_\varepsilon \chi_\omega - P_1 \rho_\varepsilon)$.

Finally we consider (50)

$$\int_0^T \int_\omega (w_\varepsilon + \rho_\varepsilon) w dxdt = 0 \ \forall w \in \mathscr{M}^\perp;$$

we have therefore $v_\varepsilon + \rho_\varepsilon \chi_\omega \in \mathscr{M}^\perp$. Since $w_\varepsilon \in \mathscr{M}^\perp$, we have $w_\varepsilon + \rho_\varepsilon \chi_\omega = P_2(w_\varepsilon + \rho_\varepsilon \chi_\omega) = P_2 \rho_\varepsilon$ and thus $w_\varepsilon = -(\rho_\varepsilon \chi_\omega - P_2 \rho_\varepsilon)$. The proposition is proved.

Remark 2. There is no information available for $\rho_\varepsilon(0)$ and $\rho_\varepsilon(T)$.

Proposition 5. *Let $v_\varepsilon, w_\varepsilon, q_\varepsilon$ and ρ_ε be the functions defined in Proposition 4. Then for $\varepsilon \to 0$, we have the following limits:*

$$v_\varepsilon \rightharpoonup \widehat{v}_\theta \text{ weakly in } L^2(\omega \times (0,T)), \tag{51}$$
$$w_\varepsilon \rightharpoonup \widehat{w}_\theta \text{ weakly in } L^2(\omega \times (0,T)) \tag{52}$$
$$q_\varepsilon \rightharpoonup \widehat{q}_\theta \text{ weakly in } H^{2,1}(Q), \tag{53}$$
$$\rho_\varepsilon \rightharpoonup \widehat{\rho}_\theta \text{ weakly in } V. \tag{54}$$

Proof. We show the proposition in three steps.

Step 1. We now look for a priori estimates for the approximate controls and state v_ε, w_ε, q_ε and ρ_ε.

From Proposition 3 we have

$$||-\frac{\partial q_\varepsilon}{\partial t} - \Delta q_\varepsilon + a_0 q_\varepsilon - h - v_\varepsilon \chi_\omega||_{L^2(Q)} \leq C\sqrt{\varepsilon}, \tag{55}$$
$$||v_\varepsilon||_{L^2(\omega\times(0,T))} \leq C, \tag{56}$$
$$||w_\varepsilon||_{L^2(\omega\times(0,T))} \leq C. \tag{57}$$

Since q_ε satisfies (37), we obtain from (55), (56) and (57) the following estimate:

$$||q_\varepsilon||_{H^{2,1}(Q)} \leq C. \tag{58}$$

From (43) and (56), we obtain

$$||-\rho_\varepsilon + P_1\rho_\varepsilon||_{L^2(\omega\times(0,T))} \leq C. \tag{59}$$

In the same way, from (44) and (57), we have

$$||-\rho_\varepsilon + P_2\rho_\varepsilon||_{L^2(\omega\times(0,T))} \leq C. \tag{60}$$

Since $L\rho_\varepsilon = 0$, using the definition of the norm on V given by (20), we have

$$||\rho_\varepsilon||_V \leq C. \tag{61}$$

Since $\rho_\varepsilon \in \mathscr{V}$, applying the observability inequality (13) to ρ_ε, we have $\left\|\frac{1}{\theta}\rho_\varepsilon\right\|_{L^2(\omega\times(0,T))} \leq C$. Then, using (59), (60) and the fact that $\frac{1}{\theta}$ is in $L^\infty(Q)$, we deduce that

$$\left\|\frac{1}{\theta}P_1\rho_\varepsilon\right\|_{L^2(\omega\times(0,T))} \leq C \text{ and } \left\|\frac{1}{\theta}P_2\rho_\varepsilon\right\|_{L^2(\omega\times(0,T))} \leq C.$$

Since $P_1\rho_\varepsilon \in Y_\lambda$ and $P_2\rho_\varepsilon \in \mathscr{M}$ with Y_λ and $\mathscr{M}$ finite dimensional, we have $\|P_1\rho_\varepsilon\|_{L^2(\omega\times(0,T))} \leq C$, $\|P_2\rho_\varepsilon\|_{L^2(\omega\times(0,T))} \leq C$. Thus using again (59) et (60), we obtain

$$||\rho_\varepsilon||_{L^2(\omega\times(0,T))} \leq C, \tag{62}$$

Step 2. We study the convergence of $(v_\varepsilon, w_\varepsilon, q_\varepsilon)$.
From (56), (57) and (58) we can extract some subsequences still denoted $(q_\varepsilon)_\varepsilon$, $(v_\varepsilon)_\varepsilon$ and $(w_\varepsilon)_\varepsilon$ such that

$$v_\varepsilon \rightharpoonup v_0 \text{ weakly in } L^2(\omega \times (0,T)), \tag{63}$$
$$w_\varepsilon \rightharpoonup w_0 \text{ weakly in } L^2(\omega \times (0,T)), \tag{64}$$
$$q_\varepsilon \rightharpoonup q_0 \text{ weakly in } H^{2,1}(Q). \tag{65}$$

Since $v_\varepsilon \in Y_\lambda^\perp$ and $w_\varepsilon \in \mathscr{M}^\perp$ with Y_λ and $\mathscr{M}$ the real closed subspaces of $L^2(\omega \times (0,T))$, we have

$$v_0 \in Y_\lambda^\perp \text{ and } w_0 \in \mathscr{M}^\perp. \tag{66}$$

Since the injection from $H^{2,1}(Q)$ into $L^2(Q)$ is compact, the pair (v_0, w_0, q_0) is such that

$$\begin{cases} -\dfrac{\partial q_0}{\partial t} - \Delta q_0 + a_0 q_0 = h_0 + v_0 \chi_\omega + w_0 \chi_\omega & \text{in } Q, \\ q_0 = 0 & \text{on } \Sigma, \\ q_0(T) = 0 & \text{in } \Omega, \end{cases} \tag{67}$$

$$q_0(0) = 0 \text{ in } \Omega.$$

Step 3. We show that $(v_0, w_0, q_0 = q_0(x,t;(v_0,w_0))) = (\widehat{v}_\theta, w_\theta, \widehat{q}_\theta = \widehat{q}_\theta(x,t;(\widehat{v}_\theta, \widehat{w}_\theta)))$.
From the expression of J_ε given by (36), we can write

$$\frac{1}{2}||v_\varepsilon||^2_{L^2(\omega\times(0,T))} + \frac{1}{2}||w_\varepsilon||^2_{L^2(\omega\times(0,T))} \leq J_\varepsilon(v_\varepsilon, q_\varepsilon).$$

Since $(\widehat{v}_\theta, \widehat{w}_\theta, \widehat{q}_\theta)$ satisfies (3)–(5), we have

$$\begin{aligned} \frac{1}{2}||v_\varepsilon||^2_{L^2(\omega\times(0,T))} + \frac{1}{2}||w_\varepsilon||^2_{L^2(\omega\times(0,T))} &\leq J_\varepsilon(v_\varepsilon, q_\varepsilon) \\ &\leq \frac{1}{2}\|\widehat{v}_\theta\|^2_{L^2(\omega\times(0,T))} + \frac{1}{2}\|\widehat{w}_\theta\|^2_{L^2(\omega\times(0,T))}. \end{aligned} \tag{68}$$

Thus using (63) and (64) and taking the limit in (68), we obtain

$$\begin{aligned} \frac{1}{2}||v_0||^2_{L^2(\omega\times(0,T))} + \frac{1}{2}||w_0||^2_{L^2(\omega\times(0,T))} &\leq \liminf_{\varepsilon\to 0} J_\varepsilon(v_\varepsilon, q_\varepsilon) \\ &\leq \frac{1}{2}\|\widehat{v}_\theta\|^2_{L^2(\omega\times(0,T))} + \frac{1}{2}\|\widehat{w}_\theta\|^2_{L^2(\omega\times(0,T))}, \end{aligned}$$

consequently,

$$||v_0||^2_{L^2(\omega\times(0,T))} + ||w_0||^2_{L^2(\omega\times(0,T))} \leq \|\widehat{v}_\theta\|^2_{L^2(\omega\times(0,T))} + \|\widehat{w}_\theta\|^2_{L^2(\omega\times(0,T))}$$

since the triplet $(\widehat{v}_\theta, w_\theta, \widehat{q}_\theta)$ is the optimal solution, we have

$$\begin{aligned} ||v_0||^2_{L^2(\omega\times(0,T))} + ||w_0||^2_{L^2(\omega\times(0,T))} &= \|\widehat{v}_\theta\|^2_{L^2(\omega\times(0,T))} + \|\widehat{w}_\theta\|^2_{L^2(\omega\times(0,T))}, \\ ||(v_0,w_0)||^2_{L^2(\omega\times(0,T))} &= ||(\widehat{v}_\theta,\widehat{w}_\theta)||^2_{L^2(\omega\times(0,T))}, \\ ||(v_0,w_0)||_{L^2(\omega\times(0,T))} &= ||(\widehat{v}_\theta,\widehat{w}_\theta)||_{L^2(\omega\times(0,T))}. \end{aligned}$$

From the uniqueness of $\widehat{v}_\theta$ and $\widehat{w}_\theta$, we have $v_0 = \widehat{v}_\theta$ and $w = \widehat{w}_\theta$.
Step 4. From (62) et (61), there exists a subsequence, still denoted $(\rho_\varepsilon)_\varepsilon$ and $\widehat{\rho}_\theta \in V$ such that

$$\rho_\varepsilon \rightharpoonup \widehat{\rho}_\theta \text{ weakly in } V \tag{69}$$
$$\rho_\varepsilon \rightharpoonup \widehat{\rho}_\theta \text{ weakly in } L^2(\omega\times(0,T)). \tag{70}$$

Since P is the compact operator, we deduce from (70) that

$$\begin{aligned} P_1\rho_\varepsilon &\to P_1\widehat{\rho}_\theta \text{ weakly in } L^2(\omega\times(0,T)), \\ P_2\rho_\varepsilon &\to P_2\widehat{\rho}_\theta \text{ weakly in } L^2(\omega\times(0,T)). \end{aligned} \tag{71}$$

Combining (70) and (71), we obtain

$$\begin{aligned} v_\varepsilon = \rho_\varepsilon\chi_\omega - P\rho_\varepsilon &\rightharpoonup \widehat{v}_\theta = \widehat{\rho}_\theta\chi_\omega - P_1\widehat{\rho}_\theta \text{ weakly in } L^2(\omega\times(0,T)), \\ w_\varepsilon = \rho_\varepsilon\chi_\omega - P_2\rho_\varepsilon &\rightharpoonup \widehat{w}_\theta = \widehat{\rho}_\theta\chi_\omega - P_2\widehat{\rho}_\theta \text{ weakly in } L^2(\omega\times(0,T)). \end{aligned}$$

This achieves the proof of existence.

Discriminating Sentinels with Given Sensitivity

In this section, we use the previous results to identify some pollution parameters in a problem governed by the semi-linear parabolic equation.

More precisely, let N, $M_1 \in \mathbb{N}\backslash\{0\}$ and Ω be a bounded subset of $\mathbb{R}^N$ with boundary Γ of class C^2. For any $T > 0$, we set $Q = \Omega\times(0,T)$ and $\Sigma = \partial\Omega\times]0,T[= \Gamma\times]0,T)$. We consider now the system modelling the following pollution problem:

$$\begin{cases} \dfrac{\partial y}{\partial t} - \Delta y + f(y) = \xi + \displaystyle\sum_{i=1}^{M_1} \lambda_i\hat{\xi}_i \quad \text{in } Q \\ y = 0 \text{ on } \Sigma \\ y(0) = y^0 + \tau\,\hat{y}^0 \text{ in } \Omega \end{cases} \tag{72}$$

where

- $y : Q \longrightarrow \mathbb{R}$ is an unknown function which represents, for example, the pollutant concentration.
- f is a real-valued given function of class C^1, globally Lipschitz.

– The source term is unknown and represents pollution source of the form $\xi + \sum_{i=1}^{M_1} \lambda_i \hat{\xi}_i$. The functions ξ and $\{\hat{\xi}_i\}_{i=1}^{M_1}$ are known whereas the real coefficients $\{\lambda_i\}_{i=1}^{M_1}$ are unknown.
– The initial condition is of the form $y^0 + \tau\, \hat{y}^0$ where the function y^0 is known while τ real is unknown.

We assume that:

– y^0 and $\hat{y}^0$ belong to $L^2(\Omega)$; ξ and ξ_i belong to $L^2(Q)$.
– The functions ξ_i, $1 \le i \le M_1$ are linearly independent.
– The real τ is sufficiently small.
– The function f verifies

$$f(0) = 0 \tag{73}$$

and satisfies the growth condition:

$$\begin{aligned} &|f(s_1) - f(s_2) - f'(0)(s_1 - s_2)\,| \\ &\le c(\,|s_1|^{P-1} + |s_2|^{P-1})|s_1 - s_2|\ \ \forall s_1, s_2 \in \mathbb{R} \end{aligned} \tag{74}$$

for some $c > 0$ and $p > 1$ such that $p < \dfrac{N+4}{N}$.

We assume without loss generality of the problem that

$$\xi = 0 \ \text{ in } Q \ \text{ and } y^0 = 0 \text{ in } \Omega. \tag{75}$$

Under the above conditions on the data, it is proved in [5, 20] that there exists $\alpha > 0$ such that if

$$||\tau \hat{y}^0||_{L^2(\Omega)} + ||\sum_{i=1}^{M_1} \lambda_i \hat{\xi}_i\,||_{L^2(Q)} \le \alpha$$

the problem (72) admits a unique solution in $C([0,T];\, L^2(\Omega)) \cap L^2(0,T; H_0^1(\Omega))$. Moreover, if we denote by I a neighbourhood of 0, the applications

$$\begin{aligned} &\tau \to y(\lambda_i, \tau) \ \text{ and } \ \lambda_i \to y(\lambda_i, \tau),\ (1 \le i \le M_1) \\ &\qquad \text{are in } \mathscr{C}\left(I, L^2(0,T; L^2(\Omega))\right). \end{aligned} \tag{76}$$

for simplicity, we denote

$$y(x,t;\lambda,\tau) = y(\lambda,\tau).$$

the unique solution to (72).

We need to obtain by observation of (72) some information on the pollution term $\sum_{i=1}^{M_1} \lambda_i \hat{\xi}_i$ without calculating the missing term $\tau \hat{y}^0$.

In the model (72), we are interested in identifying the parameters λ_i without any attempt of computing the missing term $\tau \hat{y}^0$. To identify these parameters, we use the method of sentinels due to J.L. Lions dans [14]. It was revisited by O. Nakoulima in [18].

The theory of sentinels introduced by J.L. Lions relies on three considerations:

1. A state equation represented here by (72) whose solution $y = y(x,t,\lambda,\tau) = y(\lambda,\tau)$ depends on two families of parameters $\lambda = \{\lambda_1, \ldots \lambda_{M_1}\}$ and τ.
2. An observation

$$y_{obs} = m_0 + \sum_{i=1}^{M_2} \beta_i m_i, \tag{77}$$

where $M_2 \in \mathbb{N}$; $m_0, m_1, m_2, \cdots, m_{M_2} \in L^2(O \times (O,T))$ are given by measurement of y in $L^2(O \times (O,T))$. The real-valued β_i are unknown, and we suppose that they are small enough. The terms $\beta_i m_i$ are the interference terms. m_0 is a measurement of y if there is no noise. The functions m_i are linearly independent on $O \times (0,T)$. O is nonempty open subset of Ω it's called observatory.

3. A function $S = S(\lambda,\tau)$ called "sentinel" defined for $h \in L^2(O\times]0,T[)$ and a nonempty open subset ω of O ($\omega \subset \overline{\omega} \subset O$) by

$$S(\lambda,\tau)(u) = \int_0^T \int_O hy(\lambda,\tau)dxdt + \int_0^T \int_\omega uy(\lambda,\tau)dxdt \tag{78}$$

where the control function u is to be found with minimal norm in $L^2(\omega\times]0,T[)$ among functions S defined in (78) and satisfying the following conditions:

– S is stationary to the first order with respect to the missing term $\tau\, \hat{y}^0$, i.e.

$$\frac{\partial S}{\partial \tau}(0,0) = 0 \ \ \forall \hat{y}^0. \tag{79}$$

– S is stationary to the first order with respect to the interference terms $\beta_i m_i$, i.e.

$$\int_0^T \int_O h_0 m_i dxdt + \int_0^T \int_\omega u m_i dxdt = 0,\ 1 \le i \le M_2. \tag{80}$$

– S is sensitive to the first order with respect to the pollution terms $\lambda_i \hat{\xi}_i$:

$$\frac{\partial S}{\partial \lambda_i}(0,0) = c_i,\ 1 \le i \le M_1 \tag{81}$$

where c_i, $(1 \le i \le M_1)$, are some given constants not all identically null.

At this step, several remarks are indispensable.

Remark 3. One of the purposes of this work is to show that the set of u satisfying (78)–(81) is nonempty and has a infinitely many solutions. The problem is then to propose a criterion which permits to select one of them.

Remark 4. For a sentinel without noise and without a given sensitivity, one chooses the control of minimal norm in $L^2(U)$; see J. L. Lions [14]. One proceeds similarly for a discriminating sentinel studied by O. Nakoulima [18] or a discriminating sentinel studied by G. Massengo Mophou and O. Nakoulima in [17].

In the case of a discriminating sentinel with given sensitivity, it seems natural to look for a criterion in the product space. We proceed then as follows:

– We look for the control u in the form:

$$\begin{aligned} u &= u_0 + \widehat{v} + \widehat{w} \\ &\text{with} \\ u_0 &\in Y_\lambda \ + \mathscr{M}, \widehat{v} \in Y_\lambda^\perp \ \text{and} \ \widehat{w} \in \mathscr{M}^\perp. \end{aligned} \tag{82}$$

where Y_λ and $\mathscr{M}$ are defined in (2). We then choose u among the controls called "admissibles", i.e. the controls for which the function S defined in (78) satisfying (79)–(81), with

$$||(\widehat{v},\widehat{w})|| = \min_{(v,w)\in E} \{||(v,w)||\} \tag{83}$$

where

$$E = \{(v,w) \in Y_\lambda \ \times \mathscr{M}, \ \text{such that} \ (u,S(u)) \ \text{satisfies} \ (79)\text{–}(81)\}.$$

Using the adjoint problem, one shows that the existence of these sentinels is reduced to the solution of null-controllability problem with two constraints on the control. These types of controllability problems were the subject of many numerical methods which in fact reduce them to an approximate controllability problem with constraints on the state. It is in this context, for instance, that J.P. Kernevez et al. use these sentinels in [1, 2] to identify parameters of pollution in a river. O. Bodart applies them in [4] to identify an unknown boundary.

O. Nakoulima in [18] studied the null-controllability problem with Dirichlet condition using a discriminating sentinel. G.Massengo Mophou and O. Nakoulima in [17] studied this problem given a sensitivity sentinel. In this paper we propose to study a null-controllability problem for Dirichlet boundary condition using a discriminating sentinel with given sensitivity.

Remark 5. To estimate the parameter λ_i, one proceeds as follows.

Assume that the solution of the state equation (72) when $\lambda = 0$ and $\tau = 0$ is known. Then one has the following information:

$$S(\lambda,\tau) - s(0,0) \approx \sum_{i=1}^{M_1} \lambda_i \frac{\partial S}{\partial \lambda_i}(0,0).$$

Therefore, fixing $i_0 \in \{1,\ldots,M_1\}$ and choosing j such that

$$\frac{\partial S}{\partial \lambda_j}(0,0) = 0 \quad \text{for} \ \ i_0 \neq j \ \text{and} \ \frac{\partial S}{\partial \lambda_{i_0}}(0,0) = 1,$$

one obtains the following estimate of the parameter λ_{i_0}:

$$\lambda_{i_0} \approx S(\lambda,\tau) - s(0,0).$$

Remark 6. Notice that for the J.L. Lions's sentinel, one has $\omega = O$.

Remark 7. Since $y \in L^2(0,T;H^1(\Omega))$, $h \in L^2(O \times (0,T))$ and $u \in L^2(\omega \times (0,T))$, the relation (78) is well defined. Furthermore, in view of (76), the derivatives of y with respect to τ denoted by

$$y_\tau = \frac{d}{d\tau} y(\lambda, \tau)|_{\tau=0} \tag{84}$$

and with respect to λ_i denoted by

$$y_{\lambda_i} = \frac{d}{d\lambda_i} y(\lambda, \tau)|_{\lambda_i=0} \tag{85}$$

exist. Thus the conditions (79)–(81) are well defined.

Remark 8. In the sensitivity condition (81), the c_i are chosen according to the importance associated with the component $\hat{\xi}_i$ of the pollution source.

Remark 9. If the function S defined by (78)–(81) exists, then it is unique since u verifies (83). In this case, proceeding as in Remark 5, we get

$$\lambda_i \approx \frac{1}{c_i}(S(\lambda, \tau) - S(0,0)).$$

Definition 1. We will refer to the function S given by (78)–(81) as a discriminating sentinel with given $\{c_i\}$ sensitivity.

Let y_0 be the solution of (72) when $\lambda = 0$ and $\tau = 0$. Then, in view of (75), we have

$$y_0 = 0 \text{ in } Q. \tag{86}$$

From (84) and (85), y_τ and y_{λ_i} are, respectively, solution of

$$\begin{cases} \dfrac{\partial y_\tau}{\partial t} - \Delta y_\tau + f'(0) y_\tau = 0 \text{ in } Q, \\ y_\tau = 0 \text{ on } \Sigma, \\ y_\tau(0) = \hat{y}^0 \text{ in } \Omega \end{cases} \tag{87}$$

and

$$\begin{cases} \dfrac{\partial y_{\lambda_i}}{\partial t} - \Delta y_{\lambda_i} + f'(0) y_{\lambda_i} = \hat{\xi}_i \text{ in } Q, \\ y_{\lambda_i} = 0 \text{ on } \Sigma, \\ y_{\lambda_i}(0) = 0 \text{ in } \Omega \end{cases} \tag{88}$$

where $f'(0)$ denotes the derivative of f at $y_0 = 0$. From condition (74), the problems (87) and (88) admit, respectively, unique solutions $y_\tau \in C([0,T]; L^2(\Omega)) \cap L^2(0,T;H_0^1(\Omega))$ and $y_{\lambda_i} \in H^{2,1}(Q) = L^2(0,T;H^2(\Omega)) \cap H^1(0,T;L^2(\Omega))$.

Let χ_ω be the characteristic function of the set ω. We set

$$Y_\lambda = vect(y_{\lambda_1}\chi_\omega, \cdots, y_{\lambda_{M_1}}\chi_\omega) \tag{89}$$

$$\mathscr{M} = vect(m_1\chi_\omega, \cdots, m_{M_2}\chi_\omega) \tag{90}$$

the vector subspace of $L^2(\omega \times (0,T))$ generated by the M_1-independent functions $y_{\lambda_i}\chi_\omega$ $(1 \le i \le M_1)$ and the M_2-independent functions $m_i\chi_\omega$ $(1 \le i \le M_2)$.

We also set

$$a_0 = f'(0). \tag{91}$$

Equivalence to the Null-Controllability Problem

Since y_τ and y_{λ_i} are, respectively, solutions of (87) and (88), the stationary conditions (79)–(80) and, respectively, the sensitivity conditions (81) hold if and only if

$$\int_0^T \int_O h_0 y_\tau dxdt + \int_0^T \int_\omega u y_\tau dxdt = 0 \ \ \hat{y}^0 \in L^2(\Omega), \tag{92}$$

$$\int_0^T \int_O h m_i dxdt + \int_0^T \int_\omega u m_i dxdt = 0, \ 1 \le i \le M_2, \tag{93}$$

and

$$\int_0^T \int_O h_0 y_{\lambda_i} dxdt + \int_0^T \int_\omega u y_{\lambda_i} dxdt = c_i, \ 1 \le i \le M_1. \tag{94}$$

In order to transform equation (92) we introduce the classical adjoint state.

More precisely, we consider the solution $q = q(x,t,u)$ of the linear problem

$$\begin{cases} -\dfrac{\partial q}{\partial t} - \Delta q + a_0 q = h\chi_O + u\chi_\omega & \text{in } Q \\ q = 0 = 0 & \Sigma \\ q(T) = 0 & \Omega \end{cases} \tag{95}$$

where χ_O the characteristic function of the open set O. It is well known that the problem (95) admits a unique solution q in $H^{2,1}(Q)$ (see [5]).

First, multiplying both sides of the differential equation in (95) by $y_\tau \in C([0,T], L^2(\Omega)) \cap L^2(0,T;H_0^1(\Omega))$ which is solution of (87) and integrating by parts in Q, we get

$$\int_0^T \int_O h_0 y_\tau dxdt + \int_0^T \int_\omega u y_\tau dxdt = \int_\Omega q(0)\hat{y}^0 dx \ \ \forall \hat{y}^0 \in L^2(\Omega).$$

Thus, condition (79) or (92) holds if and only if

$$q(0) = 0 \ \text{ in } \Omega. \tag{96}$$

Then, multiplying both sides of the differential equation in (95) by $y_{\lambda_i} \in H^{2,1}(Q)$ which is solution of (88) and integrating by parts in Q, we have

$$\int_0^T \int_{\Omega} q\widehat{\xi}_i dxdt = \int_0^T \int_O h_0 y_{\lambda_i} dxdt + \int_0^T \int_{\omega} u y_{\lambda_i} dxdt \quad 1 \leq i \leq M_1.$$

Thus, condition (81) or (94) is equivalent to

$$\int_0^T \int_{\Omega} q\widehat{\xi}_i dxdt = c_i \quad 1 \leq i \leq M_1. \tag{97}$$

Let Y_λ be a real closed vector subspace defined in (89). Since Y_λ is a vector subspace $L^2(\omega \times (0,T))$ generated by the independent functions and Y_λ is finite dimensional, there exists a unique $u_1 \in Y_\lambda$ such that

$$c_i - \int_0^T \int_O h_0 y_{\lambda_i} dxdt = \int_0^T \int_{\omega} u_1 y_{\lambda_i} dxdt \quad 1 \leq i \leq M_2.$$

Therefore, condition (94) or (97) holds if and only if

$$u - u_1 \in Y_\lambda^{\perp}.$$

Therefore there exists $k_1 \in Y_\lambda^{\perp}$ such that

$$u = u_1 + k_1. \tag{98}$$

We consider now (93). Let $\mathscr{M}$ be a real closed vector subspace defined in (90). Since $\mathscr{M}$ is a vector subspace $L^2(\omega \times (0,T))$ generated by the independent functions and $\mathscr{M}$ is finite dimensional, there exists a unique $u_2 \in \mathscr{M}$ such that

$$\int_0^T \int_O h_0 m_i dxdt = -\int_0^T \int_{\omega} u_2 m_i dxdt, \ 1 \leq i \leq M_1.$$

Therefore, condition (93) holds if and only if

$$u - u_2 \in \mathscr{M}^{\perp};$$

therefore, there exists $k_2 \in \mathscr{M}^{\perp}$ such that

$$u = u_2 + k_2. \tag{99}$$

In (98) and (99) we do the sum member to member

$$2u = (u_1 + k_1) + (u_2 + k_2)$$

which gives

$$u = \frac{1}{2}(u_1 + u_2) + \frac{1}{2}k_1 + \frac{1}{2}k_2.$$

Setting $u_0 = \frac{1}{2}(u_1 + u_2)$, $v_1 = \frac{1}{2}k_1$, $v_2 = \frac{1}{2}k_2$, we have

$$u = u_0 + v_1 + v_2; \tag{100}$$

we replace u by $u_0 + v_1 + v_2$ in $(95)_1$, and we set

$$h = h_0\chi_O + u_0\chi_\omega \in L^2(Q). \tag{101}$$

One can state the following proposition:

Proposition 6. *Let u be in $L^2(\omega\times]0,T[)$ and $u_0 \in Y_\lambda \ + \mathscr{M}$*

$$S(\lambda,\tau)(u) = \int_0^T \int_O hy(\lambda,\tau)dxdt + \int_0^T \int_\omega uy(\lambda,\tau)dxdt;$$

then the following are equivalent:

1. *Find u such that S satisfies (79)–(81).*
2. *Find $(v_1,v_2) \in Y_\lambda^\perp \times \mathscr{M}^\perp$ such that if $u = u_0 + v_1 + v_2$ and if q is the solution of (95), then q satisfies (96).*

Detection of Parameters

We are now able to give the expression of the sentinel S defined by (78)–(81) and identify the parameter λ_i.

Expression of the Discriminating Sentinel with Given Sensitivity

We consider the results obtained in section "Antiperiodic Functions" and we assume that h given by (101) and θ given by Theorem 1 are such that $\theta h \in L^2(O\times(0,T))$. Let $(\widehat{v}_\theta,\widehat{w}_\theta,\widehat{q}_\theta)$ as in Theorem 2. Since $(\widehat{v}_\theta,\widehat{w}_\theta) = (-(\widehat{\rho}_\theta\chi_\omega - P_1\widehat{\rho}_\theta),-(\widehat{\rho}_\theta\chi_\omega - P_2\widehat{\rho}_\theta))$ realizes the minimum in $L^2(O\times(0,T))$ among all controls (v,w) such that the pair $((v,w);q)$ satisfies (3)–(6), using (100), we deduce that $u = u_0 + \widehat{v}_\theta + \widehat{w}_\theta = u_0 - (\widehat{\rho}_\theta\chi_\omega - P_1\widehat{\rho}_\theta) - (\widehat{\rho}_\theta\chi_\omega - P_2\widehat{\rho}_\theta)$. Consequently, replacing u by its expression in (78), the function S becomes:

$$\begin{aligned} S(\lambda,\tau)(u) = &\int_0^T \int_O h_0 y(\lambda,\tau)dxdt \\ &+ \int_0^T \int_\omega (u_0 - (\widehat{\rho}_\theta\chi_\omega - P_1\widehat{\rho}_\theta) - (\widehat{\rho}_\theta\chi_\omega - P_2\widehat{\rho}_\theta))y(\lambda,\tau)dxdt \end{aligned} \tag{102}$$

and $(u,S(u))$ is such that (78)–(81) holds.

Identification of Parameter λ_i

Let us show now how the sentinel permits to detect the pollution λ_i. Let y_0 be the solution of (72) for $\lambda = 0, \tau = 0$. One obtains from (75) that $y_0 = 0$ in Q. Then taking $\lambda = 0, \tau = 0$ in (102), we can calculate

$$S(0,0)(u) = \int_0^T \int_O h_0 y_0 dxdt + \int_0^T \int_\omega (u_0 + \widehat{v}_\theta + \widehat{w}_\theta) y_0 dxdt = 0. \tag{103}$$

From (79), we have

$$S(\lambda,\tau)(u) \simeq S(0,0) + \sum_{i=1}^{M_2} \lambda_i \frac{S}{\partial \lambda_i}(0,0),\ for\ \lambda_i\ et\ \tau\ petit. \tag{104}$$

If y_{obs} is known, its follows from (77) and (80) that

$$S(\lambda,\tau)(u) = \int_0^T \int_O h_0 m_0 dxdt + \int_0^T \int_\omega (u_0 + \widehat{v}_\theta + \widehat{w}_\theta) m_0 dxdt. \tag{105}$$

Thus,

$$S(\lambda,\tau) - S(0,0) = \int_0^T \int_O h_0 (m_0 - y_0) dxdt + \int_0^T \int_\omega (u_0 + \widehat{v}_\theta + \widehat{w}_\theta)(m_0 - y_0) dxdt.$$

We have also the following information:

$$\sum_{i=1}^{M_1} \lambda_i \frac{S}{\partial \lambda_i}(0,0) = \int_0^T \int_O h_0 (m_0 - y_0) dxdt + \int_0^T \int_\omega (u_0 + \widehat{v}_\theta + \widehat{w}_\theta)(m_0 - y_0) dxdt$$

which gives (80) using

$$\sum_{i=1}^{M_1} \lambda_i c_i \approx \int_0^T \int_O h_0 (m_0 - y_0) dxdt + \int_0^T \int_\omega (u_0 + \widehat{v}_\theta + \widehat{w}_\theta)(m_0 - y_0) dxdt.$$

Now, fixing $i_0 \in \{1, M_1\}$ and choosing $c_{i_0} \neq 0$ and $c_j = 0$, for all $j \in \{1, M_1\}$ with $j \neq i_0$, we obtain of this estimation the parameter λ_{i_0}

$$\begin{aligned} \lambda_{i_0} &\approx \frac{1}{c_{i_0}} \{ \int_0^T \int_O h_0 (m_0 - y_0) dxdt + \int_0^T \int_\omega (u_0 + \widehat{v}_\theta + \widehat{w}_\theta)(m_0 - y_0) dxdt \} \\ &\approx \frac{1}{c_{i_0}} \{ \int_0^T \int_O h_0 m_0 dxdt + \int_0^T \int_\omega (u_0 + \widehat{v}_\theta + \widehat{w}_\theta) m_0 dxdt \} \end{aligned}$$

because $y_0 = 0$ in Q.

References

1. Ainseba, B.E., Kernevez, J.P., Luce, R.: Application des sentinelles à l'identification des pollution dans une rivière. M2AN Math. Model. Numer. Anal. **28**(3), 297–312 (1994)
2. Ainseba, B.E., Kernevez, J.P., Luce, R.: Identification de paramètres dans les problèmes non linéaire à donnéés incomplètes. M2AN Math. Model. Numer. Anal. **28**(3), 313–328 (1994)
3. Barbu, V.: Exact controllability of the superlinear heat equation. Appl. Math. Optim. **42**(1), 73–89 (2000)
4. Bodart, O.: Sentinels for the identification of an unknown boundary. Math. Model Method Appl. Sci. **7**(6), 871–885 (1997)
5. Cazenave, Th., Haraux, A.: Introduction aux problèmes d'évolution semi-linéaires. Collection Mathématiques et Applications de la Smai Editions Ellipse, Paris (1991)
6. Doubova, A., Osses, A., Puel, J.P.: Exact controllability to trajectories for semilinear heat equations with discontinuous diffusion coefficients. ESAIM: Contr Optim. Calc. Var. **8**, 621–661 (2002)
7. Fernandez-Cara, E., Gonzalez-Burgos, M., Guerrero et J-P., S.: Puel: Null controllability of the heat equation with boundary Fourier condition: the linear case. ESAIM: COCV **12**(3), 442–465 (2004)
8. Fernández-Cara, E.: Nul controllability of the semilinear heat equation. ESAIM Contr Optim. Calc. Var. **2**, 87–103 (1997)
9. Fernández-Cara, E., Guerrero, S.: Global Carleman inequalities for parabolic systems and applications to controllability. SIAM J. Contr Optim. **45**(4), 1395–1446 (2006)
10. Fernández-Cara, E., Zuazua, E.: The cost of approximate controllability for heat equations: the linear case. Adv. Differ. Equat. **5**, 465–514 (2000)
11. Fursikov, A., Imanuvilov, O.Yu.: Controlability of Evolution Equations, Lecture Notes #34. Seoul National University, Korea (1996)
12. Imanuvilov, O.Yu.: Controllability of parabolic equations. Sbornik Math. **186**(6), 879–900 (1995)
13. Lebeau, G., Robbiano, L.: Contrôle exacte de l'équation de la chaleur. Comm. Part. Differ. Equat. **20**, 335–356 (1995)
14. Lions, J.L.: Sentinelles pour les systèmes distribués à données incomplètes. Recherches en Mathématiques Appliquées 21, Masson, Paris (1992)
15. Lions, J.L., Magenes, M.: Problemes aux limites non homogènes et applications. Dunod, Vol. 1 et 2, Paris (1968). Zb1 0165.10801 MR 1159093
16. Lions, J.L.: Controle optomal de systèmes gouvernés par des équations aux dérivées partielles. Dunod, Gauthier-Villars, Paris (1968). Zb1 017941801 MR 1159093
17. Massengo Mophou, G., Nakoulima, O.: Sentinel with given sensitivity. Eur. J. Appl. Math. **19**(01), 21–40 (2008)
18. Nakoulima, O.: A revision of J.-L. Lions' notion of sentinels. Portugal. Math. **65**(Fasc. 1), 1–22 (2008)
19. Russell, D.L.: A unified boundary controlability theory for hyperbolic and parabolic partial differential equations. **52**, 189–211 (1973)
20. Zuazua, E.: Finite dimentional null controlability for the semi-linear heat equation. J. Math. Pure Appl. **76**(9), 237–264 (1997). Zb1 0872.93014 MR 1441986
21. Lions, J.L.: Contrôle des systèmes distribués singuliers. BORDAS, Gauthier-Villars, Paris (1983). ISBN 2-04-015539-2
22. Lions, J.L.: Quelques méthodes de résolution des problèmes aux limites non linéaires. Dunod, Gauthier-Villars, Paris (1969)

Chapter 10
A Galerkin Method Solution of Heat Transfer Problems in Closed Channels: Fluid Flow Analysis

Nasser Ghariban

Introduction

The study of conventional forced convection in channels is a requirement for well-designed heat transfer equipment. A channel is a configuration for studying internal flows. Being well informed or having knowledge of the structure of flow in a channel is of great engineering interest since it can be applied in many applications. Channels contain flows that are considered Newtonian fluid. The traditional application of this study is in heat transfer equipment, such as heat exchangers; the friction factor and heat transfer coefficient are important parameters for evaluating design performance of these equipment. Recently, the growth in microfluidic systems with needs of transporting of liquids or gases in channels with micro cross-sectional dimensions is of great importance in many applications. These applications in microelectronic cooling, MEMS, fuel cell technology, and medical and biomedical devices motivated researchers to investigate on simple solutions for channel flow. Microchannels have specific characteristics such as high surface area per unit volume and high heat transfer coefficient that will provide further application in the future. A study of velocity and temperature distribution in these channels will help investigators to understand the pressure drop and heat transfer rate at the boundaries. Although the behavior of fluid in microchannels is laminar in larger-scale equipment such as heat exchangers, the designer usually deals with turbulent flow. When flow is turbulent, the computations are difficult. Often, the experimental studies of shear stress and heat transfer in turbulent flow guide researchers toward theoretical predictions. However, the experimental data are not universally available for all possible shapes and flow conditions.

N. Ghariban (✉)
Department of Engineering, Virginia State University, Petersburg Virginia 23806, USA
e-mail: nghariban@vsu.edu

B. Toni (ed.), *Advances in Interdisciplinary Mathematical Research*, Springer Proceedings in Mathematics & Statistics 37, DOI 10.1007/978-1-4614-6345-0_10,

Numerical computation has served engineers well and is a powerful tool. This work seeks a simple mathematical model that can produce relatively accurate results with little computational effort. The Galerkin-based method of solution given by Haji-Sheikh et al. [4] and Beck et al. [1] is modified to solve laminar flow in closed channels as well as study of turbulent flow. This method provided a simple and effective method for calculating laminar flow characteristics in various shape channels. It is also shown that this method can also be used for turbulent flow; however, major modifications are needed. A set of basis functions that are markedly different from the basis functions for laminar flow must be selected. Improved accuracy and rapid convergence are realized when the basis functions include the dependence of turbulent viscosity on the velocity gradient. From several different turbulent viscosity models, the Van Driest model is chosen for this solution method. It was determined that a modified Van Driest model provides computed data that agree well with experimental data of other investigators, e.g., Laufer [7] and Nikuradse [9].

Analysis

The objective of this paper is to develop a simple and computationally efficient method for finding flow properties such as pressure drop. The Galerkin method is selected because it is equally applicable to circular and noncircular ducts. The same method has been extended for soling thermal characteristics of the channels that will be address in second part of this publication.

Governing Equations

The Galerkin-based method is a simple technique for finding the velocity field in ducts with arbitrary cross-section areas. For convenience, the Cartesian coordinates are used to describe the method of solution. The cylindrical coordinates are used for study of pipe flow as demonstrated in example 1 and 3 of this study. The momentum equations for fully developed channel flow are

$$\rho \frac{Du}{Dt} = \rho . f - \nabla P + \mu \nabla^2 u \tag{1}$$

In the absence of external forces and steady-state condition, the equation for a laminar flow will simplify to

$$-\frac{\partial P}{\partial Z} + \mu \left(\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2} \right) = 0 \tag{2a}$$

where w is the velocity of the flow along channel axis (z).

For turbulent flow the equation will have an extra term due to fluctuation of velocity along x, y, and z axis as

$$-\frac{\partial P}{\partial z}+\mu\left(\frac{\partial^2 w}{\partial x^2}+\frac{\partial^2 w}{\partial y^2}\right)-\rho\left(\frac{\overline{\partial u'w'}}{\partial x}+\frac{\overline{\partial v'w'}}{\partial y}\right)=0 \tag{2b}$$

The Boussinesq's eddy-diffusivity coefficients for momentum are defined as

$$\begin{aligned} -\rho\overline{u'w'} &= \mu_t \tfrac{\partial w}{\partial x} \\ -\rho\overline{v'w'} &= \mu_t \tfrac{\partial w}{\partial y} \end{aligned} \tag{3}$$

The following equation is given by substituting the above expressions into Eq. (2b)

$$-\frac{\partial P}{\partial z}+\frac{\partial}{\partial x}\left(\mu_e\frac{\partial w}{\partial x}\right)+\frac{\partial}{\partial y}\left(\mu_e\frac{\partial w}{\partial y}\right)=0 \tag{4}$$

where μ_e is the effective viscosity given as

$$\mu_e=\mu+\mu_t$$

If the vector notation is used, the turbulent momentum equation is shortened to

$$-\frac{\partial P}{\partial z}+\nabla\cdot(\mu_e\nabla w)=0 \tag{5}$$

This equation can be written in nondimensional form as

$$1+\nabla_1\cdot(\mu_e^*\nabla_1 W)=0 \tag{6}$$

where $\mu_e^*=1+(\mu_t/\mu)=1+\mu_t^*$ for laminar flow in absence of eddy-diffusivity $\mu_e^*=1$ and Eq. (6) will be simplified to $1+\nabla_1^2 W=0$

$$W=\frac{\mu w}{-a^2\frac{\partial P}{\partial z}} \text{ where } a \text{ is a characteristic length}$$

and

$$\nabla_1=\frac{\partial}{\partial X}i+\frac{\partial}{\partial Y}j \text{ where } X=x/a \text{ and } Y=y/a$$

Turbulent Viscosity

Equation (6) indicates that the momentum equations contain turbulent viscosity that is a function of the surface sheer stress. Because of the appearance of this term, the mean flow equations are not complete; a turbulent model is necessary to determine the turbulent diffusivity terms before the equations can be solved.

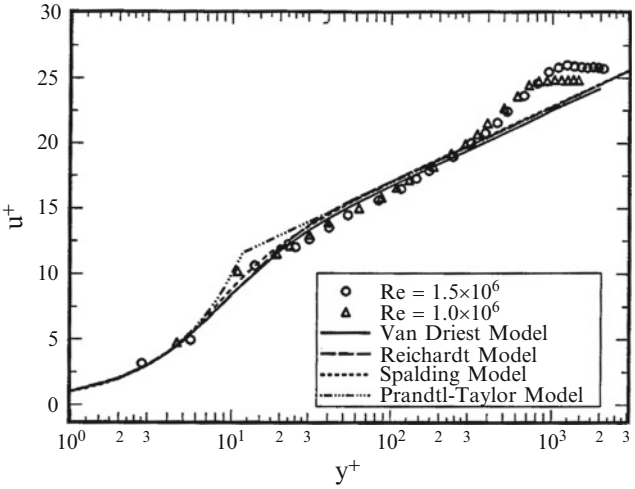

Fig. 10.1 Nondimensional velocity profile for turbulent boundary layer

Kakac et al. [5] has summarized different models of turbulent viscosity for predicting velocity profile inside the turbulent boundary layer. Figure 10.1 shows the velocity distribution from the wall with the nondimensional parameters u^+ and y^+,

$$u^+ = \frac{\overline{u}}{\sqrt{\tau_w/\rho}},\ y^+ = \frac{y\sqrt{\tau_w/\rho}}{\nu}$$

where τ_w is the wall shear stress.

The circular and the triangular symbols are the results of an experimental study for Reynolds number of 1.5×10^6 and 1.0×10^6, respectively, conducted by the author. As Fig. 10.1 illustrates, the Van Driest model is smooth and continuous in the near wall region and follows experimental data with a good agreement. The good prediction of turbulence near the wall by the Van Driest model is the primary reason for its selection. This model is solely a function of y^+; therefore, it would considerably simplify the numerical calculations and increase the accuracy of the results. According to the Van Driest [11] model, the effective viscosity is given as

$$\mu_e = \mu \frac{1 + \{1 + 4\kappa^2 (y^+)^2 [1 - \exp(-y^+/A^+)]^2\}^{1/2}}{2} \tag{7}$$

Researchers who have used the Van Driest model to study turbulent flow inside ducts report that an additional modification is necessary to remove certain inaccuracies. Malhotra and Kang [8] used Eq.(7) with an additional correction factor which

emerged as a result of studies on a two-equation model of turbulence for pipe flow. The turbulence viscosity then becomes

$$\mu_t = \rho\kappa^2 y^2[1-\exp(y^+/A^+)]\left|\frac{\partial u}{\partial y}\right|/(1+\frac{3y}{R}+\frac{3y^2}{R^2})$$

where R is the radius of the pipe and y is the distance from the wall. Richman and Azad [10] assumed a constant turbulence viscosity in the range $0.158 \le y/R \le 1$ as

$$\mu_t = \mu\frac{\{1+4\kappa^2(y^+)^2[1-\exp(-y^+/A^+)]^2\}^{1/2}-1}{2} \tag{8}$$

for $0 \le y/R \le 0.158$ and $\mu_t = (\mu_t)_{y/R=0.158}$ for $0.158 \le y/R \le 1$

In the present study, the second modification yields a closer agreement with the experimental data.

Calculation of Fluid Flow Properties and Pressure Drop

A Galerkin-based integral (GBI) method [1] is used to solve momentum and energy equations. This is based on weighted residual methods. The method can be used for any ordinary differential equation such as $L[y(x)]+f(x)=0$ over interval $a \le x \le b$ where L denotes a linear differential equation.

Multiplying this equation with any arbitrary function w(x) and integrating over the interval [a,b] provide

$$\int_a^b w(x)\{L[y(x)]+f(x)\}dx = 0$$

Weighted residual method provides solution to this equation by introducing a trial solution of u(x) as

$$u(x) = \phi_0(x) + \sum_{j=1}^{n} c_j\phi_j(x)$$

Replacing y(x) with u(x) on the left side of original differential equation, the residual is defined as follows:

$$r(x) = L[u(x)]+f(x)$$

The goal of this method is to construct u(x) so that the integral of the residual will be zero for some choices of weighted functions. This means the following condition, zero residual, must be satisfied for some choices of w(x):

$$\int_a^b w(x)\{L[u(x)]+f(x)\}dx = 0$$

Galerkin Method

Galerkin method is one of the most commonly used weighted residual methods. This method chooses the weight function, w(x), as a function of basis functions, $w(x) \in \phi_i(x)|_{i=1}^{n}$

$$\int_a^b \phi_i(x)\{L[u(x)] + f(x)\}dx = 0 \qquad \text{for} \quad i = 1,2,\ldots\ldots,n$$

Introducing trial function, $u(x) = \phi_0(x) + \sum_{j=1}^{n} c_j\phi_j(x)$, into this equation provides a set of n equations that must be solved to find the coefficients of basis functions C_j:

$$\int_a^b \phi_i(x)\left\{L\left[\sum_{j=1}^{n} c_j\phi_j(x)\right] + L[\phi_o(x)] + f(x)\right\}dx = 0 \quad \text{for} \quad i = 1,2,\ldots\ldots,n \quad (9)$$

Solution of the Momentum Equation

Equation (6) is the momentum equation in nondimensional space with the boundary condition $V = 0$ on the wall. According to the Galerkin-based integral method, the solution is approximated as a linear combination of a set of basis functions

$$V(X,Y) = \sum_{i=1}^{N} c_i f_i(X,Y) \tag{10}$$

The basis functions, f_i, are linearly independent and satisfy the same homogeneous boundary conditions as V; thus, $V(X,Y)$, axial velocity in Z direction, satisfies the given boundary conditions for all choices of the c_i's. Next, if the error or residual is formed and the d_i's are chosen so that the weighted integral of the residual is zero for each $i = 1,\ldots,N$, a linear algebraic system is obtained as

$$\mathbf{A}\cdot\mathbf{c} = \mathbf{b} \tag{11}$$

where **d** is the vector of coefficients which has the elements $d_1, d_2, \ldots, d_N$ and it is the solution of the above system of N linear equations. The vector **g** has elements

$$b_i = -\frac{1}{A}\int_A f_i dA \tag{12}$$

and the matrix **A** has the elements

$$a_{ij} = \frac{1}{A}\int_A f_i \nabla_1 \cdot (\mu_e^* \nabla_1 f_j) \quad \text{for turbulent flow} \tag{13a}$$

$$a_{ij} = \frac{1}{A}\int_A f_i \nabla_1 \cdot (\nabla_1 f_j) \quad \text{for laminar flow} \tag{13b}$$

The solution of Eq. (11) results in the evaluation of coefficients, $c_1, c_2, \ldots, c_N$, and their substitution in Eq. (10) yields the solution for velocity V.

Haji-Sheikh et al. [4] derived the following equations for skin friction and dimensionless velocity by this method:

$$C_f \mathrm{Re} = \frac{2D_e^2}{a^2 V_{av}} = \frac{2\overline{D_e^2}}{V_{av}} \tag{14}$$

$$\frac{V}{V_{av}} = \frac{V}{V_{av}} = \frac{C_f \mathrm{Re}}{2\overline{De^2}} \sum_{i=1}^{N} d_i f_i \tag{15}$$

where $C_f = -(\partial_P/\partial_Z)D_e/(\rho v_{av}/2) = 4\tau_w/(\rho v_{av}/2)$, $\mathrm{Re} = \rho D_e v_{av}/\mu$, and D_e/a is designated as the nondimensional hydraulic diameter.

The solution for laminar flow has simple steps of defining suitable basis functions, solving array $\boldsymbol{b}$ and matrix $\boldsymbol{a}$ Eqs. (12) and (13b), and then solving for array $\boldsymbol{c}$ to find velocity profile from Eq. (10).

The momentum equation for turbulent flow to solve is more complex. The turbulent viscosity (μ_e^*) is a function of wall shear stress (τ_w) that must be determined. An iterative method is used to solve this equation for turbulent flow. Figure 10.2 demonstrates the flow chart for this solution.

Example 1: Laminar Pipe Flow

The momentum equation in cylindrical coordinate can be written as $-\frac{dP}{dz} + \mu\left[\frac{1}{r}\frac{d}{dr}\left(r\frac{dv}{dr}\right)\right] = 0$ with boundary condition of $v = 0$ at $r = R_0$. (R_0 is the radius of the pipe.)

Introducing nondimensional velocity as

$$V = -\frac{\mu v}{R_o^2(dP/dz)} \quad \text{and} \quad R = \frac{r}{R_o}$$

the equation reduces to

$$\left[\frac{1}{R}\frac{d}{dr}\left(R\frac{dV}{dR}\right)\right] + 1 = 0 \text{ with boundary condition of } V = 0 \text{ at } R = 1$$

The exact solution for this equation is

$$V = \frac{1}{4}(1 - R^2)$$

Dividing it by average velocity, it can be normalized to $U^* = \frac{V}{V_{ave}} = 2(1 - R^2)$

And the exact solution for skin friction using Eq. (9) can be evaluated as $C_f \mathrm{Re} = 64$.

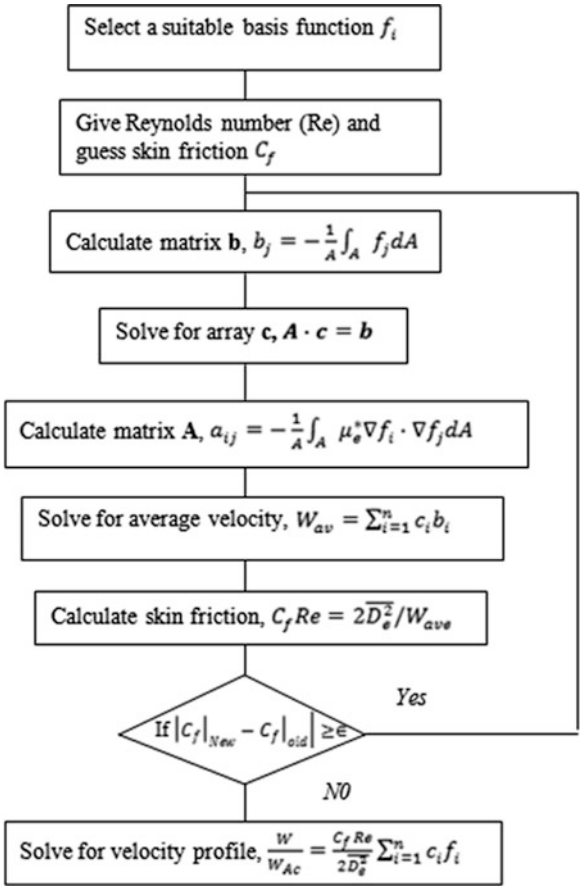

Fig. 10.2 Calculation procedure for turbulent flow

The Galerkin solution can be found by writing Eq. (9) for cylindrical coordinates as Galerkin's solution to this differential equation will be in the form of

$$\int_0^1 \phi_i(R)\left\{\frac{1}{R}\frac{d}{dR}\left(R\frac{d}{dR}\sum_{j=1}^{n} c_j\phi_j(R)\right)+1\right\}dR = 0$$

That can be written in form of a set of linear equations as

$$A_{ij} \cdot C_j = B_i$$

where

$$A_{ij} = \int_0^1 \phi_i(R)\left\{\frac{1}{R}\frac{d}{dR}\left(R\frac{d}{dR}\phi_j(R)\right)\right\}\mathrm{dR} \quad \text{and} \quad B_i = \int_0^1 \phi_i(R)dR$$

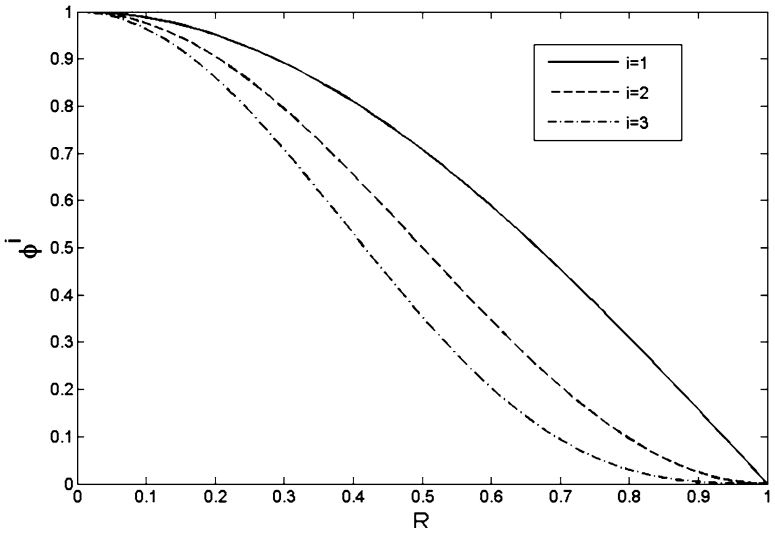

Fig. 10.3 Basis function for a flow in a cylindrical pipe

The Galerkin method starts with selecting the basis functions. The basis functions must satisfy the same boundary condition as governing equation. A suitable basis function can be in the form of $\phi_i(R)| = \left[\cos(\frac{\pi}{2}R)\right]^i$. Figure 10.3 illustrates the basis functions for n = 1,2,3.

A program in MATHLab was developed to evaluate matrics A_{ij}, B_i, C_j, and f for different value of n. Following is calculation for n = 4:

$$A_{ij} = \begin{bmatrix} -2.6882 & -3.3870 & -3.9424 & -4.4201 \\ -2.2171 & -3.2451 & -3.9434 & -4.4945 \\ -1.9310 & -3.0478 & -3.8335 & -4.4488 \\ -1.7334 & -2.8641 & -3.6957 & -4.3559 \end{bmatrix},$$

$$B_i = [-0.6366 \ -0.5000 \ -0.4244 \ -0.3750],$$

$$\text{and} \quad C_j = \begin{bmatrix} 0.3180 \\ -0.0974 \\ 0.0385 \\ -0.0091 \end{bmatrix}$$

The result for skin friction is given in the following table for different value of n:

n	**1**	**2**	**3**	**4**	**5**	**6**
$C_f \cdot R_e$	73.0132	63.7913	64.0286	63.9969	64.0004	63.9999

As the table indicates with only two terms the resistance coefficient can be evaluated with an error of 0.3%.

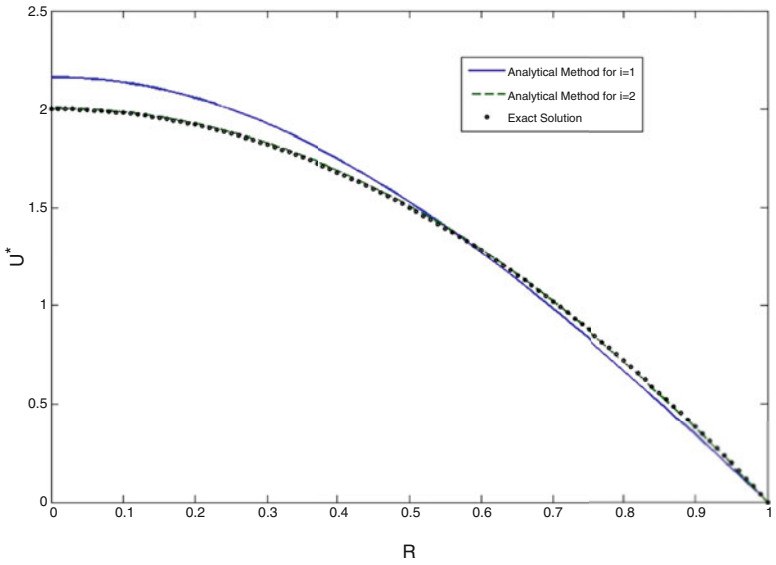

Fig. 10.4 Comparison of analytical and exact solution velocity profile

The velocity profile can be evaluated from

$$v = \sum_{i=1}^{n} C_i \cdot \phi_i$$

For example, for n = 4 velocity profile will be

$$v = 0.3180\cos(\pi/2R) - 0.0974\cos^2(\pi/2R) + 0.0385\cos^3(\pi/2R) - 0.0091\cos^4(\pi/2R)$$

As this figure indicates analytical and exact solutions are in excellent agreement with each other even with two-term solution ($n = 2$) for the velocity (Fig. 10.4). Another basis function that satisfies boundary condition and provides great accuracy is in the form of

$$f_i = (1 - R^2)R^{2(i-1)} \tag{16}$$

The first basis function in this set ($f_1 = 1 - R^2$) matches with the exact solution, and the method and one-term solution will match perfectly with exact solution.

Example 2: Laminar Flow Inside Square Duct

The governing equations in general for pipe flow are Navier–Stokes equations for incompressible laminar steady-state flow and can be written as

$$-\frac{dP}{dz}+\mu\left(\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}\right)=0$$

where u(x, y) is velocity in two dimensions with boundary condition of u(x, y) = 0 at the wall and dP/dz is pressure drop and μ is viscosity.

For a square channel with side "a," assuming $X=\frac{x}{a}$ and $Y=\frac{y}{a}$ the equation can be written in nondimensional form

$$1+\frac{\partial^2 U}{\partial X^2}+\frac{\partial^2 U}{\partial Y^2}=0 \tag{17}$$

where $U=-\frac{\mu/a^2}{dP/dz}u$ with the boundary condition of U = 0 for X = 0, Y = 0, X = 1, and Y = 1.

The following figure illustrates the geometry of such a channel:

Galerkin method starts with assuming a set of basis function then

$$U=\sum_{i=1}^{k} C_i\phi_i$$

where ϕ_i is the basis function that should satisfy the boundary condition (zero velocity on walls). One satisfactory expression for this function can be in the form of

$$\phi_i=\cos^m(\pi X)\cdot\cos^n(\pi Y)$$

where m and n are any set of integer numbers greater than zero. The following table can represent one set of such numbers that has been used in this study:

i	m	n
1	1	1
2	1	2
3	2	1
4	2	2
5	3	1
6	3	2
.	.	.
.	.	.

The following figures illustrate the shape of basis functions (ϕ_1 and ϕ_3) along X axis and Y = 0.3 (Fig. 10.5).

Multiplying both sides of Eq. (17) with the basis function and integrating over the domain, the governing equation will be changed to integral equation as

$$\int_{-1/2}^{1/2}\int_{-1/2}^{1/2}\phi_i(x)\left\{1+\left(\frac{\partial^2}{\partial X^2}+\frac{\partial^2}{\partial Y^2}\right)\left[\sum_{j=1}^{n} c_j\phi_j(x)\right]\right\}dx=0$$

And this integral equation can be converted to a set of linear equation as

$$A_{ij}\cdot C_j=B_i$$

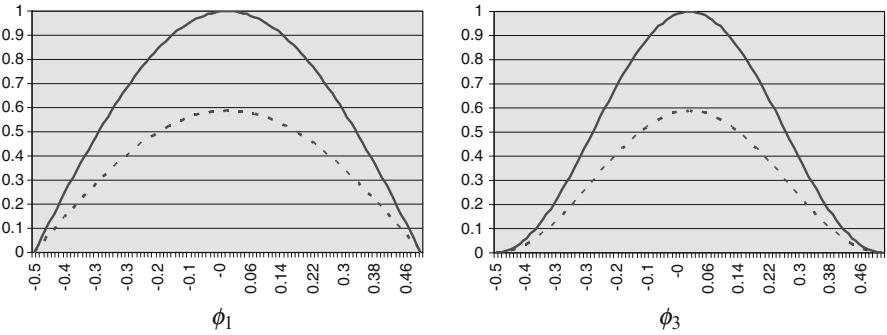

Fig. 10.5 Basis function for a flow in a square closed channel

where

$$A_{ij} = \int_{-1/2}^{1/2}\int_{-1/2}^{1/2} \phi_i(x)\left\{(\frac{\partial^2}{\partial X^2} + \frac{\partial^2}{\partial Y^2})\phi_j(x)\right\} dXdY$$

$$B_i = \int_{-1/2}^{1/2}\int_{-1/2}^{1/2} \phi_i(x) dXdY$$

To solve the problem B_i and A_{ij} must be evaluated. Following is a sample of calculation for $i = j = 1$ where $\phi_1 = \cos(\pi X)\cdot\cos(\pi Y)$:

$$B_1 = \int_{-1/2}^{1/2}\int_{-1/2}^{1/2} \phi_1 dXdY = \int_{-1/2}^{1/2}\int_{-1/2}^{1/2} \cos(\pi X)\cos(\pi Y) dXdY$$

$$= \frac{1}{\pi}\sin(\pi X)|_{-1/2}^{1/2} \cdot \frac{1}{\pi}\sin(\pi Y)|_{-1/2}^{1/2}$$

$$B_1 = \frac{4}{\pi^2}$$

For A_{11} second partial derivative of basis functions must be calculated first before integrating over the area:

$$\frac{\partial\phi_1}{\partial X} = -\pi\sin(\pi X)\cdot\cos(\pi Y) \quad \text{and} \quad \frac{\partial^2\phi_1}{\partial X^2} = -\pi^2\cos(\pi X)\cdot\cos(\pi Y)$$

$$\frac{\partial\phi_1}{\partial X} = -\pi\cos(\pi X)\cdot\sin(\pi X) \quad \text{and} \quad \frac{\partial^2\phi_1}{\partial X^2} = -\pi^2\cos(\pi X)\cdot\cos(\pi X)$$

and

$$\frac{\partial^2\phi_1}{\partial X^2} + \frac{\partial^2\phi_1}{\partial Y^2} = -2\pi^2\cos(\pi X)\cdot\cos(\pi X)$$

then

$$A_{ij} = \int_{-1/2}^{1/2}\int_{-1/2}^{1/2} \cos(\pi X)\cdot\cos(\pi Y)\cdot(-2\pi^2\cos(\pi X)\cdot\cos(\pi Y)) dXdY$$

$$A_{ij} = -2\pi^2 \int_{-1/2}^{1/2} \cos^2(\pi X) dX \int_{-1/2}^{1/2} \cos^2(\pi Y) dY$$

$$A_{ij} = -2\pi^2 \int_{-1/2}^{1/2} \frac{1}{2}(1-\cos(2\pi X)) dX \int_{-1/2}^{1/2} \frac{1}{2}(1-\cos(2\pi Y)) dY$$

$$A_{ij} = -\frac{\pi^2}{2}$$

Velocity coefficient will be calculated from $A_{11} \cdot C_1 = B_1$

$$C_1 = B_1/A_{11} = \frac{8}{\pi^4} = .08213$$

The result for six terms is as follows:

$$B = \begin{bmatrix} -0.4053 \\ -0.3183 \\ -0.3183 \\ -0.2500 \\ -0.2702 \\ -0.2122 \end{bmatrix} \quad A = \begin{bmatrix} -4.9348 & -4.1888 & -4.1888 & -3.5556 & -3.7011 & -3.1416 \\ -4.1888 & -4.3180 & -3.5556 & -3.6652 & -3.1416 & -3.2385 \\ -4.1888 & -3.5556 & -4.3180 & -3.6652 & -4.1888 & -3.5556 \\ -3.5556 & -3.6652 & -3.6652 & -3.7011 & -3.5556 & -3.5605 \\ -3.7011 & -3.1416 & -4.1888 & -3.5556 & -4.3180 & -3.6652 \\ -3.1416 & -3.2385 & -3.5556 & -3.5605 & -3.6652 & -3.6240 \end{bmatrix}$$

$$C = \begin{bmatrix} 0.188037 \\ -0.083799 \\ -0.131458 \\ 0.091192 \\ 0.045730 \\ -0.036434 \end{bmatrix}$$

Figure 10.6 illustrates the velocity profile in different locations from the center of the channel. The maximum velocity will be at location $X = 0$ and $Y = 0$ with a value of $U_{\max} = 0.0735$.

Figure 10.7 illustrates the same data in three-dimensional format.

The normalized velocity can be evaluated from Eq. (15) to be compared with other studies and experimental values. The normalized velocity at the center of the channel with this study is **s**

$$\frac{u_{\max}}{u_{ave}} = 2.093$$

This result was compared to an experimental study by Kakac et al. [5] and a finite difference method by the author. The experimental method is given for rectangular channels as

$$\frac{u}{u_{max}} = \left(\frac{m+1}{m}\right)\left(\frac{n+1}{n}\right)$$

$$m = 1.7 + 0.5\alpha^{*-1.4}$$

$$n = \begin{cases} 2 & for\ \alpha^* \le \frac{1}{3} \\ 2 + 0.3\left(\alpha^* - \frac{1}{3}\right) & for\ \alpha^* \ge \frac{1}{3} \end{cases}$$

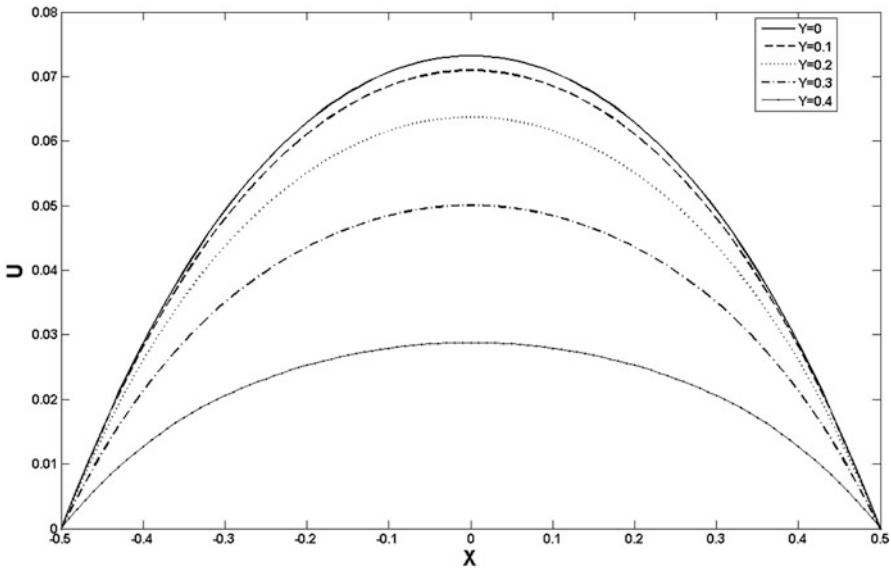

Fig. 10.6 Velocity profile inside a square channel

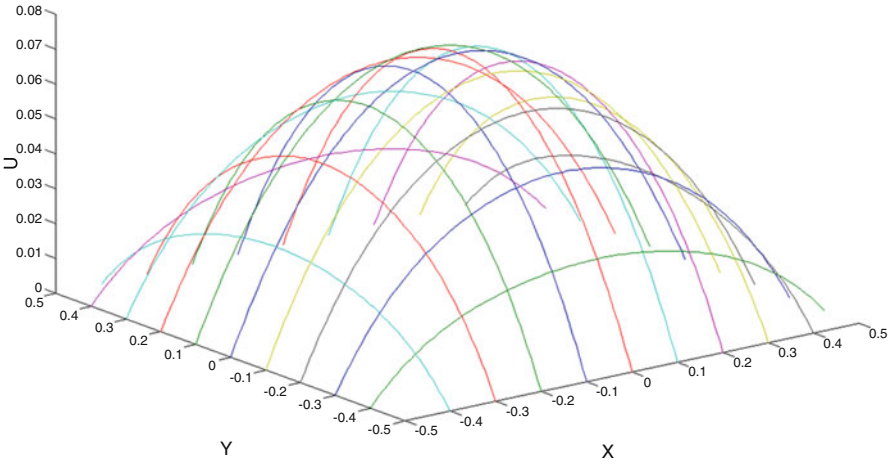

Fig. 10.7 Velocity profile inside a square channel

where $\alpha^* = 1$ for square channel; as a result $m = n = 2.2$ and the maximum velocity will be

$$\frac{u_{\max}}{u_{ave}} = 2.115$$

This is in good agreement with result evaluated by Galerkin-based solution.

The finite difference method yields this value as 2.099 using 101×101 elements. Finite difference method takes significant computational time for converging results.

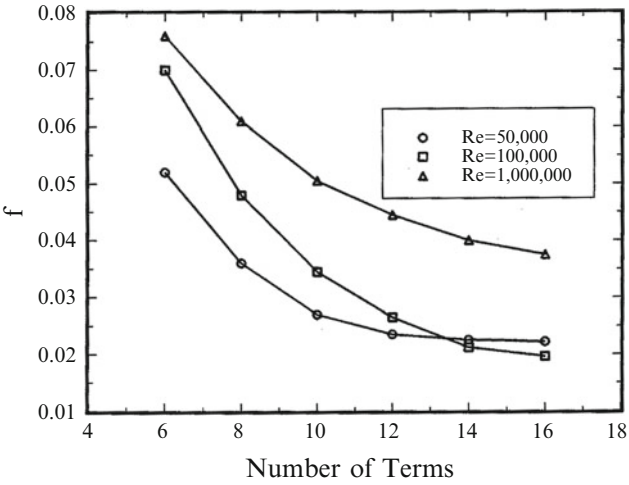

Fig. 10.8 Convergence of skin friction using polynomial basis function

Example 3: Turbulent Pipe Flow

The same procedure as example 1 is used for studying the turbulent flow in a circular pipe. The initial attempt is to use the same basis polynomial functions as introduced in example 1, Eq. (16). Although f_i in Eq. (16) is a simple function of R, the **a** matrix cannot be calculated analytically because of the existence of μ_e^* in Eq. (9) which is a function of r. Numerical integration is used to evaluate the members of the **a** matrix. The array **g**, however, is the same as that given for laminar flow.

Having matrix **a** and array **b** available, the skin friction and the velocity profile are calculated by an iterative procedure as shown in Fig. 10.8. The calculations show that the convergence of the results for skin friction requires a large number of the basis functions. This is the initial difficulty encountered when calculating pressure drop in turbulent flow by the standard GBI method. In contrast, the laminar flow requires as few as two basis functions for an accurate solution.

Figure 10.8 shows the convergence of the results for Reynolds numbers $5 \times 10^4, 10^5$, and 10^6 using a different number of terms. This figure confirms that a higher Reynolds number requires more terms to have convergence. It is also noticed that the upper limit for N is 16. For values of N greater than 16, the matrix inversion routine fails, and convergence never happens. According to Fig. 10.8, solutions for a Reynolds number of 10^6 or higher cannot be obtained.

The velocity profile for Reynolds number 10^5 and for different values of N is given in Fig. 10.9. The computed results are compared with the experimental data given by Nikuradse [9] for the same Reynolds number. This figure shows that, as the number of terms increases, the calculated velocity profile gets closer to the experimental data. For N=16, which satisfies the convergence, the calculated velocity is in agreement with the experimental data within 3%.

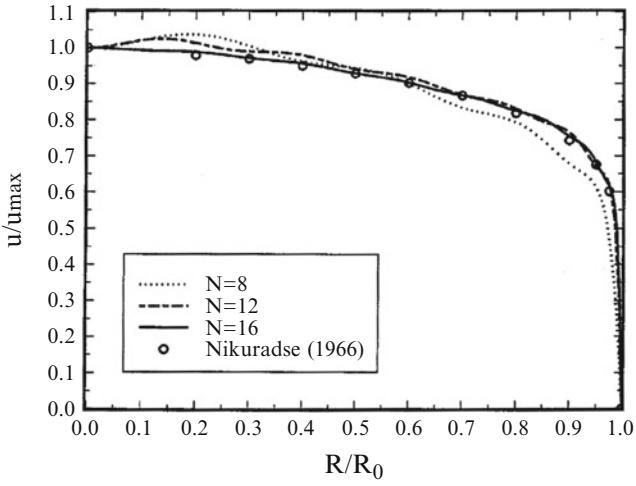

Fig. 10.9 Predicted velocity profile by using polynomial basis functions

The calculated values of skin friction and velocity profile, Figs. 10.8 and 10.9, require a large number of terms to achieve convergence. Increasing the number of terms increases the computer time and reduces the accuracy of the results. For these reasons, selecting a set of basis functions that describes the behavior of the turbulent velocity profile is necessary. The following basis functions have the necessary characteristics to describe the turbulent velocity profile:

$$f_i = \{1 - e^{[-\beta(1-R^2)]}\}R^{2(i-1)} \tag{18}$$

The factor $1 - e^{[-\beta(1-R^2)}$ in Eq. (18) is a turbulence factor that depends on the Reynolds number. It provides a sharp slope for the velocity profile at the wall and disappears far from the wall. The factor B is a constant that depends on the turbulence similar to A in the turbulent viscosity equation of Van Driest. The value of β is arbitrarily selected equal to one for laminar flow.

The computed results for three different Reynolds numbers are given in Fig. 10.10. This figure illustrates that the calculated skin friction decreases by increasing the coefficient β, then begins to increase as β increases. A value of β that makes the skin friction minimum is the proper choice. A justification of selecting β at minimum skin friction is given in the Appendix. The calculated skin friction at the optimum β provides the best agreement with the experimental values.

The optimum values of β, for different Reynolds numbers, were calculated, and method of least square was used to find correlation for β and skin friction as given in the following equation:

$$B = \frac{C_f Re}{19}$$

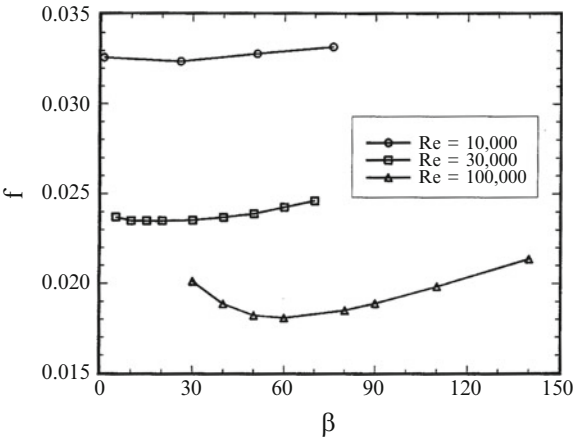

Fig. 10.10 Calculated skin friction for different values of β

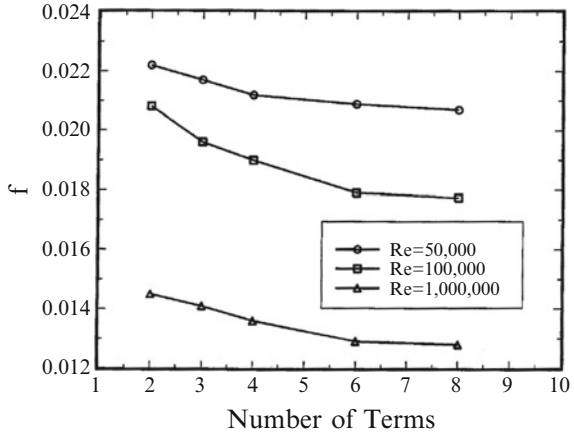

Fig. 10.11 Convergence of the result by using new basis functions

Once a correlation for β is available, one can proceed to solve for velocity profile and skin friction at any Reynolds number. Figure 10.11 shows the convergence of skin friction for three different Reynolds numbers versus the number of terms, N. A comparison between the data in Figs. 10.8 and 10.11shows that the new basis functions provide convergence with fewer terms. In fact, for a small Reynolds number $(\mathrm{Re} = 50,000)$, two terms in the series give the skin friction that has a satisfactory agreement with the experimental values.

The analysis also shows a good agreement with the experimental velocity profile given by Laufer [7]. The experimental data of Laufer [7] are given in Fig. 10.12 and compared with the analytical results for this study when $\mathrm{N} = 2, 4$, and 8. The figure shows the agreement between analytical and experimental velocity profiles to within 6% using as few as 2 terms in the series.

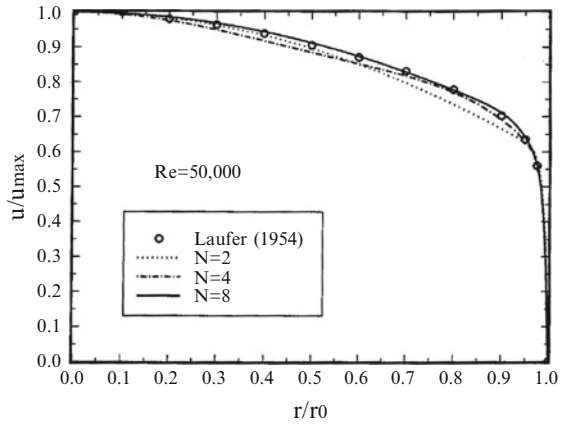

Fig. 10.12 Velocity profiles for 2, 4, and 8 terms in series

The velocity profiles, using 8 terms, and the experimental data of Laufer[7] for $\mathrm{Re} = 5 \times 10^4$ and 5×10^5 and Nikuradse [9] are shown in Fig. 10.13. Both calculated velocity profiles agree with the experimental measurements to within 3%. For more than 8 terms, no significant improvement is observed.

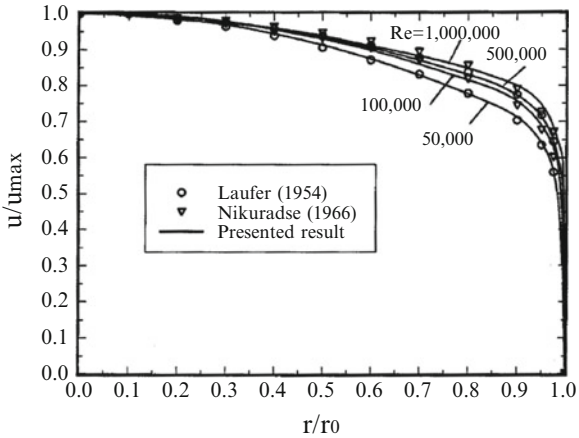

Fig. 10.13 Comparison of predicted velocity profile and experimental data

Conclusions

A simple Galerkin-based solution has been presented that predicts the skin friction and velocity flow field in fully developed duct flow. The analytical steps, described in this study, apply to both laminar and turbulent flow in ducts with various cross-

sectional shapes. A comparison of the calculated values with the experimental data shows satisfactory agreement leading to the following major conclusions.

The solution for turbulent flow shows that special care must be taken in selecting basis functions. A suitable set of basis functions reduces the number of terms in the series, N, and consequently decreases computing time and increases the accuracy of the results. The value of $N = 8$ provides data with good accuracy over a rather large range of the Reynolds number, $4 \times 10^4 < Re < 10^6$. Even for $Re < 10^5$ the method provides an accurate solution with as few as two terms.

It is shown that the modified Van Driest model for effective viscosity given by Richman and Azad [10] is a sufficiently accurate model for predicting turbulence. The calculated Nusselt numbers based on this model show good agreement with the correlation of Gnielinski [2].

Appendix

Highlights of the Variational Steps

The details of the minimization principle are provided by Ghariban [3]; only a brief description is given here. The Galerkin method is based on minimization of the integral

$$I = \int_A \left\{ \frac{1}{2}\mu_e^*(X,Y) \left[\left(\frac{\partial W}{\partial X} \right)^2 + \left(\frac{\partial W}{\partial Y} \right)^2 \right] + W \right\} dA \tag{19}$$

In the minimization of Eq. (19), it is assumed that μ_e^* is a known function of X and Y. The variational steps begin by replacing W by $\overline{W}$ and then setting $\overline{W}(X,Y) = W(X,Y) + \varepsilon\eta(X,Y)$. The calculation of $(\partial I/\partial\varepsilon)$ as $\varepsilon \to 0$ leads to the equation [6],

$$\int_A \eta(X,Y) \lfloor 1 - \nabla_1 \cdot (\mu_e^* \nabla_1 W) \rfloor dA = 0 \tag{20}$$

where $\eta(X,Y)$ is an arbitrarily selected function. The Galerkin method uses Eq. (20) to compute $W(X,Y)$. For example, one substitutes for $W(X,Y)$ in Eq. (20) a quantity

$$W(X,Y) = \sum_{i=1}^{N} d_i f_i(X,Y) \tag{21}$$

and replaces the arbitrary function $\eta(X,Y)$ by $f_J(X,Y)$ for $j = 1,2,\ldots,N$. It is to be noted that μ_e is assumed to be a known function of X and Y. This yields the Galerkin solution described by Eq. (8). For turbulent flow, $\mu_e^*(X,Y)$ is a function of shear stress at the wall, and shear stress is unknown. One must provide a value for the shear stress and then solve for velocity field. The subsequent calculation of shear stress from computed velocity field should be followed by recalculation of the velocity field. The continuation of this iterative procedure leads to a Galerkin-type solution.

For turbulent flow, the numerical studies show that the number of terms using $f_i(X,Y)$ functions with simple forms, Eq. (20), can be prohibitively large. It is suggested, in this paper, to introduce a new form for $f_i(X,Y)$, Eq. (21), that changes as the shear stress changes. The iterative minimization procedure, using this latter form of $f_i(X,Y,B)$ with an additional parameter B, needs some modifications. The first step of iteration is to consider a known value for B and solve for $W(X,Y)$ using the standard Galerkin solution method. A selected value of B influences the value of the calculated average shear stress, and B represents the effect of the turbulence intensity on the velocity profile. The minimization of function I, following some algebra [3], leads to an additional equation

$$\int_A \frac{\partial \mu_e^*(X,Y)}{\partial B} dA \left[\left(\frac{\partial W}{\partial X} \right)^2 + \left(\frac{\partial W}{\partial \mathrm{Y}} \right)^2 \right] dA = 0 \tag{22}$$

Because $(\partial W/\partial X)^2 + (\partial W/\partial Y)^2 > 0$, the integral given by Eq. (22) is zero if $(\partial \mu_e^*/\partial B) = 0$. For turbulent flow, it is assumed that B in Eq. (21) depends on the average shear stress, τ_w . Therefore, the minimization of I requires that $(\partial \mu_e^*/\partial \tau_w)(\partial \tau_w/\partial B) = 0$ in addition to Eq. (20). For turbulent flow, the effective viscosity coefficient, μ_e^*, increases as τ_w increases, indicating $\partial \mu_e^*/\partial \tau_w > 0$; then I is minimum if $\partial \tau_w/\partial B = 0$. The dimensionless form of this condition is used in subsequent calculations; that is, B is computed so that

$$\frac{\partial C_f}{\partial B} = 0$$

Acknowledgment I thank all reviewers and contributors for their valuable suggestions, comments, and correction in the final version of my manuscript. They were able to constructively pinpoint relevant issues that need to be revised in my manuscript and which will strengthen and improve my paper.

References

1. Beck, J.V., Cole, K.D., Haji-Sheikh, A., Litkouhi, B.: Heat Conduction Using Green's Functions, Chapter 11. Hemisphere Publishing, Washington, D.C. (1992)
2. Gnielinski, V.: New equations for heat and mass transfer in turbulent pipe and channel flow. Int. Chem. Eng. **16**, 359–468 (1976)
3. Ghariban, N.: Turbulent flow and heat transfer in ducts. PhD. Thesis, Department of Mechanical Engineering, The University of Texas at Arlington (1993)
4. Haji-Sheikh, A., Mashena, M., Haji-Sheikh, M.J.: Heat transfer coefficient in ducts with constant wall temperature. J. Heat Tran. **105**, 878–883 (1983)
5. Kakac, S., Shah, R., Aung, W.: Handbook of Single-Phase Convective Heat Transfer, Chapter 2. Wiley, New York (1987)
6. Kantorovitch, L.V., Krylov, V.I.: Approximate Methods of Higher Analysis. Wiley, New York (1960)
7. Laufer, J.: The structure of turbulence in fully developed pipe flow. NACA Rept. 1174 (1954)

8. Malhotra, A., Kang, S.S.: Turbulent prandtl number in circular pipes. Int. J. Heat Mass Tran. **27**(8), 2158–2161 (1984)
9. Nikuradse, J.: Gesetzmabigkiten der turbulenten stromung in glatten Rohen. Forsch. Arb. Ing.-Wes. **356** (1932); English transl., NASA TT F-10, 359 (1966)
10. Richman, J.W., Azad, R.S.: Developing turbulent flow in smooth pipes. Appl. Sci. Res. **28**, 419–440 (1973)
11. Van Driest, E.R.: On turbulent flow near a wall. J. Aeronaut. Sci. **23**, 1007–1012 (1956)

Nomenclature

A	Characteristic length
a_{ij},b_{ij},c_{ij}	Element of matrices **A**,**B**,**C**
A	Flow area of duct
A^+	Damping constant
A,**B**,**C**	Matrices
β	Constant in Eq. (12)
c_i	Element of the array **c**
c	Array of velocity coefficient
C_f	Friction factor
C_n	Temperature profile coefficient
D_e	Hydraulic diameter
$\overline{D_e}$	Hydraulic diameter, dimensionless
f_i	Basis function
g_i	Element of the array **g**
g	Auxiliary array
i, j	Indices
N	Number of terms in series
P	Static pressure
r	Cylindrical coordinate, dimensionless
R	Cylindrical coordinate
R_o	Pipe radius
T	Local temperature
T'	Fluctuating temperature
u', v', w'	Fluctuating velocity components
u^+	Velocity parameter
w	Axial velocity
W	Axial velocity, dimensionless
w_{av}	Average velocity
W_{av}	Average velocity, dimensionless
X, Y, Z	Coordinates
X, Y, Z	Coordinates, dimensionless
y^+	Wall distance parameter

κ	Von Karman constant
μ	Molecular viscosity
μ_e	Effective viscosity
μ_e^*	Effective viscosity, dimensionless
μ_t	Turbulent viscosity
μ_t^*	Turbulent viscosity, dimensionless
ρ	Fluid density
τ_w	Wall shear stress

Chapter 11
Optimal Control for Distributed Linear Systems Subjected to Null Controllability with Constraints on the State

Michelle Mercan

Introduction

Let $d \in \mathbb{N}^*$ and Ω be a bounded open subset of $\mathbb{R}^d$ with boundary Γ of class C^2, $T>0$. Let also ω be an open nonempty subset of Ω. Set $Q=\Omega\times(0,T)$, $\Sigma=\Gamma\times(0,T)$, and $G=\omega\times(0,T)$. We consider the parabolic evolution equation

$$\begin{cases} y'-\Delta y+a_0 y = h+k\chi_\omega \text{ in } Q, \\ y=0 \qquad\qquad \text{on } \Sigma, \\ y(0)=0 \qquad\qquad \text{in } \Omega, \end{cases} \tag{1}$$

where $(.)'$ is the partial derivative with respect to time t, $a_0 \in L^\infty(Q)$, $(h,k) \in L^2(Q)\times L^2(G)$, and χ_ω denotes the characteristic function of the control set ω. It is well known that problem (1) admits a unique solution y in the following Hilbert space

$$\Xi^{1,2}(Q)=H^1((0,T);L^2(\Omega))\cap L^2((0,T);H^2(\Omega)\cap H_0^1(\Omega)).$$

Let $\{e_i,\ 1\le i\le M\}$ be a set of functions of $L^2(Q)$ such that

$$e_i\chi_\omega\ 1\le i\le M \text{ are linearly independent.} \tag{2}$$

From now on, we use the notation

$$y=y(h,k)$$

to mean that each source term h and k plays a particular role. More precisely, we would like to choose the control pair (h,k) in order to achieve two objectives that we present under the form (in the cascade sense) of two problems.

Michelle Mercan (✉)
Laboratoire CEREGMIA, Université des Antilles et de La Guyane,
Campus Fouillole 97159 Pointe-à-Pitre Guadeloupe (FWI), France
e-mail: Michelle.Mercan@etu.univ-ag.fr

B. Toni (ed.), *Advances in Interdisciplinary Mathematical Research*, Springer Proceedings in Mathematics & Statistics 37, DOI 10.1007/978-1-4614-6345-0_11,

Problem 1. Let $H \subset L^2(Q)$ be a Hilbert space and $\{e_i,\ 1 \leq i \leq M\}$ be a set of functions of $L^2(Q)$ and assume that (2) holds. Fix $h \in H$. Then the Follower's problem can be stated as follows: Given $a_0 \in L^\infty(Q)$, find a control $k \in L^2(G)$ such that if y is solution of

$$\begin{cases} y' - \Delta y + a_0 y = h + k\chi_\omega \text{ in } Q, \\ y = 0 \qquad\quad \text{on } \Sigma, \\ y(0) = 0 \qquad\quad \text{in } \Omega, \end{cases} \tag{3}$$

then,

$$\int_Q y e_i \, dxdt = 0, \tag{4}$$

and

$$y(T) = 0, \text{ in } \Omega. \tag{5}$$

The role of k is to insure the null-controllability property (5) in the presence of the forcing term h and under the constraint (4).

In the sequel, we introduce a suitable nonnegative weight function θ, which will be defined below, and consider the Hilbert space

$$H = \{h | h \in L^2(Q), \theta h \in L^2(Q)\} \tag{6}$$

endowed with the scalar product and the norm

$$(h,l)_\theta = \int_Q \theta^2 hl \, dxdt, \quad \|h\|_H = \|\theta h\|_{L^2(Q)}.$$

For fixed $h \in H$, we will see that there exists several controls k such that (3), (4), and (5) are satisfied. Thus, we need to add some criteria to select k. More precisely, we will see that k is of the form $k = k_0(h) + v$. We consider then the maps $\mathscr{F}$ and $\mathscr{F}_1$ defined, respectively, by

$$\begin{aligned} \mathscr{F} : H &\to L^2(G) \\ h &\mapsto v = \mathscr{F}(h) \end{aligned} \tag{7}$$

and

$$\begin{aligned} \mathscr{F}_1 : H &\to L^2(G) \\ h &\mapsto \mathscr{F}_1(h) = k_0(h). \end{aligned} \tag{8}$$

We will see below (see section "Optimal Strategy for the Leader") that these maps are linear and continuous from H into $L^2(G)$.

In addition to the null-controllability problem (5) subject to the constraint (4), the second objective is to choose the forcing term h such that

$$y(h,k) \text{ is not too far from } z_d$$

where z_d is given in $L^2(Q)$.

In order to achieve this objective, we introduce the cost function J defined by

$$J(h) = \frac{1}{2}\|y(h,k) - z_d\|^2_{L^2(Q)} + \frac{N}{2}\|h\|^2_H$$

where $z_d \in L^2(Q)$ and $\mathscr{U}_{ad}$ is a nonempty closed convex subset of H. Then, we consider the following minimization problem:

Problem 2. Find $\hat{h} \in \mathscr{U}_{ad}$ such that

$$J(\hat{h}) = \min_{h \in \mathscr{U}_{ad}} J(h). \tag{9}$$

Problem 1 is a null-controllability problem with state constraints. Few results are known for such problem. Indeed, recently O. Nakoulima [6] gave a result of null controllability for the linear heat equation with constraints on a distributed control. His result was based on an observability inequality adapted to the constraint. In [4], G. Mophou and O. Nakoulima proved the existence of sentinels with given sensitivity by solving a problem of null controllability with constraint on the control. In [3], O. Nakoulima and G. Mophou studied a null controllability with constraints on the state for a semilinear heat equation by proving that the considering problem was equivalent to null controllability with constraint on the control. G. Mophou [5] generalizes these results to the case where the nonlinear term contains gradient terms.

Problem 2 is an optimal control problem. Such problem has been widely studied by J.L. Lions [2].

In this paper, we extend the works of G. Mophou and O. Nakoulima [3, 4] to a problem of control with two controls that we have to determine successively under some constraints. This is done by solving the combination of Problems 1 and 2, called Stackelberg problem. In this case, the controls h and k are, respectively, called Leader and Follower.

The main results of this paper are the following theorems.

Theorem 1. *Existence, uniqueness, and characterization of the Follower.*
Let Ω be a bounded open subset of $\mathbb{R}^n$ with boundary Γ of class C^2, and let H be the Hilbert space defined by (6). Then, for every $e_i \in L^2(Q)$, $1 \le i \le M$ verifying (2) and every $h \in H$, there exists a unique control $k = k(h) \in L^2(Q)$ such that the solution $y = y(h,k(h))$ of (3) satisfies (4) and (5). Moreover, the control k can be selected such that

$$\|k\| \le C\|h\|_H \tag{10}$$

where $C = C\left(\Omega, \omega, a_0, T, \sum_{i=1}^{M} \|e_i\|_{L^2(Q)}\right) > 0.$

Theorem 2. *Existence, uniqueness, and characterization of the Leader.*
Let Ω be a bounded open subset of $\mathbb{R}^n$ with boundary Γ of class C^2. Let also θ be defined as previously, and $\mathscr{F}$ and $\mathscr{F}_1$ be the linear and continuous maps, respectively,

defined by (7) *and* (8)*. Then, the minimization problem* (9) *admits a unique solution* $\hat{h}$ *characterized by the following optimality condition*

$$\left(\Lambda^{-1}(\frac{1}{\theta}I+\mathscr{F}_1^*+\mathscr{F})(p)+N\hat{h},h-\hat{h}\right)_H \geq 0,\ \forall h\in\mathscr{U}_{ad} \tag{11}$$

where Λ^{-1} *is the isometric isomorphism from* H' *into* H*,* I *is the identity operator of* $L^2(Q)$*, and* p *is solution of*

$$\begin{cases} -p'-\Delta p+a_0p=y(\hat{h},k)-z_d & \text{in } Q,\\ p=0 & \text{on } \Sigma,\\ p(T)=0 & \text{in } \Omega. \end{cases}$$

The rest of this paper is organized as follows. Section "Equivalence Between the Null-Controllability Problem with Constraints on the State and a Null- Controllability Problem with Constraint on the Control" is devoted to proving the equivalence between the null-controllability problem with constraints on the state and a null-controllability problem with constraint on the control. In section "Optimal Strategy for the Follower", we solve the null-controllability problem with constraint on the control. Finally, in section "Optimal Strategy for the Leader", we solve the Leader's problem.

Equivalence Between the Null-Controllability Problem with Constraints on the State and a Null-Controllability Problem with Constraint on the Control

Proposition 1. *Let* Ω *be a bounded open subset of* $\mathbb{R}^n$ *with boundary* Γ *of class* C^2*. Then, there exists a positive real weight function* θ *(a precise definition of* θ *will be given later on), two finite dimensional subspaces* $\mathscr{M}$ *and* $\mathscr{M}_\theta$ *such that for any* $h\in H$*, there exists* $k_0=k_0(h)\in\mathscr{M}_\theta$ *such that the null-controllability problem with constraints on the state* (3)*,* (4)*, and* (5) *is equivalent to the following null-controllability problem with constraint on the control: Given* $a_0\in L^\infty(Q)$ *and* $k_0\in\mathscr{M}_\theta$*, find* $v\in L^2(G)$ *such that*

$$v\in\mathscr{M}^\perp \tag{12}$$

$$k=k_0+v \tag{13}$$

and the solution y *of*

$$\begin{cases} y'-\Delta y+a_0y=h+(k_0+v)\chi_\omega & \text{in } Q,\\ y=0 & \text{on } \Sigma,\\ y(0)=0 & \text{in } \Omega, \end{cases} \tag{14}$$

satisfies

$$y(T) = 0 \text{ in } Q. \tag{15}$$

Proof. We interpret the constraint (4) by using the adjoint state. More precisely, for any $e_i,\ 1 \le i \le M$, we consider the adjoint system

$$\begin{cases} -q_i' - \Delta q_i + a_0 q_i = e_i \text{ in } Q, \\ q_i = 0 \text{ on } \Sigma, \\ q_i(T) = 0 \text{ in } \Omega. \end{cases} \tag{16}$$

Since $a_0 \in L^\infty(Q)$ and $e_i \in L^2(Q)$, problem (16) admits a unique solution

$$q_i = q_i(z) \in \Xi^{1,2}(Q).$$

We multiply both sides of the differential equation (3) by q_i solution of (16) and we integrate over Q. By applying the Green formula, we obtain

$$\int_Q y e_i \, dxdt = \int_Q (h + k\chi_\omega) q_i \, dxdt.$$

From (4), we have

$$0 = \int_Q (h + k\chi_\omega) q_i \, dxdt.$$

Thus,

$$\int_G k q_i \, dxdt = -\int_Q h q_i \, dxdt. \tag{17}$$

Let

$$\mathscr{M} = Span\{q_i\chi_\omega,\ 1 \le i \le M\}$$

be the vector subspace of $L^2(G)$ generated by the M functions $q_i\chi_\omega,\ 1 \le i \le M$. We denote $\mathscr{M}^\perp$ the orthogonal of $\mathscr{M}$ in $L^2(G)$. We set

$$\mathscr{M}_\theta = \frac{1}{\theta}\mathscr{M}$$

the vector subspace of $L^2(G)$ generated by the M functions $\dfrac{1}{\theta} q_i\chi_\omega,\ 1 \le i \le M$.

Since the matrix $\left(\displaystyle\int_0^T \int_\omega \frac{1}{\theta} q_i q_j \, dxdt\right)_{1 \le i,j \le M}$ is symmetric positive definite (cf. Lemma 3), there exists a unique $k_0 = k_0(h) \in \mathscr{M}_\theta$ such that

$$\int_G k_0 q_i \, dxdt = -\int_Q h q_i \, dxdt,\ 1 \le i \le M. \tag{18}$$

Thus, combining (17) and (18), we deduce that

$$\int_G (k - k_0) q_i \, dxdt = 0 \quad 1 \le i \le M.$$

Consequently

$$k - k_0 \in \mathscr{M}^{\perp}.$$

Then $k = k_0 + v$ with $v \in \mathscr{M}^{\perp}$. Therefore, replacing $k\chi_\omega$ by $(k_0+v)\chi_\omega$ in (3), we obtain (14).

Conversely, fix $h \in L^2(Q)$. For every $e_i \in L^2(Q)$, $1 \le i \le M$, assume that (v,y) is the solution of (12), (13), (14), and (15). Then, by solving (16), we obtain the functions q_i, $1 \le i \le M$. Let $\mathscr{M}$ and $\mathscr{M}_\theta$ be defined as previously. Let also $\mathscr{M}^{\perp}$ be the orthogonal of $\mathscr{M}$ in $L^2(G)$, $v \in \mathscr{M}^{\perp}$ and k_0 verifying (18).

Multiplying both sides of Eq. (14) by q_i and integrating by parts over Q, we obtain

$$\int_Q y' q_i \, dxdt - \int_Q \Delta y q_i \, dxdt + \int_Q a_0 q_i \, dxdt = \int_Q [h + (k_0+v)\chi_\omega] \, q_i \, dxdt,$$

i.e.,

$$-\int_Q h q_i \, dxdt + \int_Q y e_i \, dxdt = \int_Q (k_0+v)\chi_\omega q_i \, dxdt.$$

Since $v \in \mathscr{M}^{\perp}$ and k_0 verifies (18), the previous identity is reduced to (4). Thus, (k,y) is solution of (3), (4), and (5). □

Lemma 1. *Assume that (2) holds. Then, the functions $q_i\chi_\omega$, $1 \le i \le M$ are linearly independent. Moreover, the functions $\frac{1}{\theta} q_i\chi_\omega$, $1 \le i \le M$ are also linearly independent.*

Proof.

For $\gamma_i \in \mathbb{R}$, $1 \le i \le M$, let $\tilde{k} = \sum_{i=1}^{M} \gamma_i q_i$ on Q such that $\tilde{k}_{|_G} = 0$. Since q_i is solution of (16), we have

$$-\frac{\partial \tilde{k}}{\partial t} - \Delta \tilde{k} + a_0 \tilde{k} = \sum_{i=1}^{M} \gamma_i e_i, \text{ in } Q, \tag{19}$$

$$\tilde{k} = 0, \text{ on } \Sigma. \tag{20}$$

Then, $\tilde{k}$ being identically zero on G, we deduce that $\tilde{k} = 0$ in Q. This means that $\sum_{i=1}^{M} \gamma_i e_i = 0$ in Q. Thus,

$$\sum_{i=1}^{M} \gamma_i e_i = 0, \quad \text{in } G.$$

Since the functions $e_i\chi_\omega$, $i \in \{1,\dots,M\}$ satisfy (2), we conclude that $\gamma_i = 0$, $1 \le i \le M$.

The second assertion of the lemma follows immediately. □

In order to obtain a priori estimates on $k_0(h)$, we need the following result which is proved in [3].

Lemma 2. *Let q_i be defined by (16) and θ be a positive function defined below by relation (31). Let also $A_\theta = \left(\int_G \frac{1}{\theta} q_i q_j \, dxdt\right)_{i,j}$, $1 \le i,j \le M$. Then, there exists $\delta > 0$ such that*

$$(A_\theta X, X)_{\mathbb{R}^M} \ge \delta \|X\|^2_{\mathbb{R}^M}$$

where

$$(A_\theta X, X)_{\mathbb{R}^M} = \int_G \frac{1}{\theta} \left(\sum_{i=1}^M X_i p_i\right)\left(\sum_{j=1}^M X_j p_j\right) dxdt$$

and

$$X = (X_1, \ldots, X_M) \in \mathbb{R}^M.$$

Proposition 2. *Let θ be defined below by relation (31) and h be in H. Let also q_i and $k_0(h)$ be defined, respectively, by (16) and (18). Then, there exists $C = C(\Omega, a_0, T, \sum_{i=1}^M \|e_i\|_{L^2(Q)}) > 0$ such that*

$$\|\theta k_0(h)\|_{L^2(G)} \le C\|h\|_H \tag{21}$$

$$\|k_0(h)\|_{L^2(G)} \le C\|h\|_H. \tag{22}$$

Proof. From (18), we have

$$\int_G k_0(h) q_i \, dxdt = -\int_Q h q_i \, dxdt, \, 1 \le i \le M. \tag{23}$$

Since $k_0(h) \in Span\{\frac{1}{\theta} q_1 \chi_\omega, \ldots, \frac{1}{\theta} q_M \chi_\omega\}$, there exists

$$\alpha = (\alpha_1, \ldots, \alpha_M) \in \mathbb{R}^M$$

such that

$$k_0(h) = \sum_{j=1}^M \alpha_j \frac{1}{\theta} q_j \chi_\omega. \tag{24}$$

Thus, replacing $k_0(h)$ by $\sum_{j=1}^M \alpha_j \frac{1}{\theta} q_j \chi_\omega$ in (23), we obtain

$$\int_G \sum_{j=1}^M \alpha_j \frac{1}{\theta} q_j q_i \, dxdt = -\int_Q h q_i \, dxdt, \, 1 \le i \le M$$

and consequently,

$$\int_G \sum_{j=1}^M \alpha_j \frac{1}{\theta} q_j \chi_\omega \sum_{i=1}^M \alpha_i q_i \, dxdt = -\int_Q \theta h \sum_{i=1}^M \alpha_i \frac{1}{\theta} q_i \, dxdt.$$

Applying to this latter identity Lemma 2 with $X = \alpha$ to the left-hand side and to the right-hand side and using Cauchy–Schwartz inequality, we obtain

$$\delta\|\alpha\|^2 \leq \|h\|_H \sum_{i=1}^{M} |\alpha_i| \|q_i\|_{L^2(Q)}. \tag{25}$$

From the energy inequality for q_i solution of (16), it follows that for $1 \leq i \leq M$,

$$\|q_i\|_{L^2(Q)} \leq C(\Omega, a_0, T)\|e_i\|_{L^2(Q)}$$

which, combined with (25) and the fact that $\delta > 0$ gives

$$\|\alpha\|^2 \leq \delta^{-1} C(\Omega, a_0, T)\|h\|_H \|\alpha\|_{\mathbb{R}^M} \sqrt{\sum_{i=1}^{M} \|e_i\|^2_{L^2(Q)}},$$

i.e.,

$$\|\alpha\| \leq \delta^{-1} C(\Omega, a_0, T)\|h\|_H \sqrt{\sum_{i=1}^{M} \|e_i\|^2_{L^2(Q)}}. \tag{26}$$

Finally, from (24), we have

$$\begin{aligned} \|\theta k_0(h)\|_{L^2(G)} &\leq \sum_{j=1}^{M} |\alpha_j| \|q_j\|_{L^2(G)}, \\ &\leq C(\Omega, a_0, T) \sum_{j=1}^{M} |\alpha_j| \|e_j\|_{L^2(Q)}, \\ &\leq C(\Omega, a_0, T)\|\alpha\| \left(\sum_{i=1}^{M} \|e_i\|_{L^2(Q)} \right)^{\frac{1}{2}}, \end{aligned}$$

and

$$\|k_0(h)\|_{L^2(G)} \leq C(\Omega, a_0, T)\|\alpha\| \left(\sum_{i=1}^{M} \|e_i\|_{L^2(Q)} \right)^{\frac{1}{2}}.$$

Hence, using (26) and the fact that $\dfrac{1}{\theta}$ is bounded in $L^\infty(Q)$, and setting

$$C = C(\Omega, a_0, T, \sum_{i=1}^{M} \|e_i\|_{L^2(Q)}) = \delta^{-1} C(\Omega, a_0, T)^2 \sum_{i=1}^{M} \|e_i\|^2_{L^2(Q)},$$

we deduce (21) and (22). □

Optimal Strategy for the Follower

Controllability Problem with Constraint on the Control

We consider a auxiliary function $\psi \in C^2(\overline{\Omega})$ which satisfies the following conditions:

$$\begin{aligned} &\psi(x) > 0 \ \forall x \in \Omega, \\ &\psi(x) = 0 \ \forall x \in \Gamma, \\ &|\psi(x)| \neq 0 \ \forall x \in \overline{\Omega - \omega}. \end{aligned} \tag{27}$$

Such a function exists according to A. Fursikov and O. Yu. Imanuvilov [1]. Then, for any $\lambda > 0$, we define the following weight functions:

$$\varphi(x,t) = \frac{e^{\lambda(\psi(x)+m_1)}}{t(T-t)}, \tag{28}$$

$$\eta(x,t) = \frac{e^{\lambda(|\psi(x)|_\infty+m_2)} + e^{\lambda(\psi(x)+m_1)}}{t(T-t)}, \tag{29}$$

for $(x,t) \in Q$ and $m > 1$ and we adopt the following notations:

$$\begin{aligned} L &= \frac{\partial}{\partial t} - \Delta + a_0 I, \\ L^* &= -\frac{\partial}{\partial t} - \Delta + a_0 I, \\ \mathscr{V} &= \{\rho \in C^\infty(\overline{Q}) \mid \rho = 0 \text{ on } \Sigma\}. \end{aligned}$$

Then, we have the following Carleman inequality (see [1, 3]).

Proposition 3. *Let ψ, φ, and η be defined by (27), (28), and (29). Then, there exists $\lambda_0 = \lambda_0(\Omega, \omega, a_0)$, $s_0 = s_0(\Omega, \omega, a_0, T)$ and $C = C(\Omega, \omega, a_0, T)$ such that for any $\lambda \geq \lambda_0$, any $s \geq s_0$ and any $\rho \in \mathscr{V}$, we have*

$$\int_Q \frac{e^{-2s\eta}}{s\varphi} \left(\left|\frac{\partial \rho}{\partial t}\right|^2 + |\Delta \rho|^2 \right) dxdt + s\lambda^2 \int_Q \varphi e^{-2s\eta} |\nabla \rho|^2 \, dxdt +$$
$$s^3\lambda^4 \int_Q \varphi^3 e^{-2s\eta} |\rho|^2 \, dxdt \leq C \left(\int_Q e^{-2s\eta} |L^*\rho|^2 \, dxdt + s^3\lambda^4 \int_G \varphi^3 e^{-2s\eta} |\rho|^2 \, dxdt \right). \tag{30}$$

As φ does not vanish over Q, we set

$$\theta = \varphi^{-\frac{3}{2}} e^{s\eta}. \tag{31}$$

From the definition of φ and η given by (28) and (29), the function θ is positive and $\dfrac{1}{\theta}$ is bounded. Since $\dfrac{1}{\varphi}$ is also bounded, taking $\lambda \geq \lambda_0 > 1$ and $s \geq s_0 > 1$, we obtain the following observability inequality:

$$\int_Q \frac{1}{\theta^2}|\rho|^2\,dxdt \le C\left(\int_Q |L^*\rho|^2\,dxdt + \int_G |\rho|^2\,dxdt\right),\ \forall \rho \in \mathscr{V}. \tag{32}$$

Denote by:

- P the orthogonal projection operator from $L^2(G)$ into $\mathscr{M}$.
- $P\rho$ the orthogonal projection of $\rho\chi_\omega$ for $\rho \in L^2(Q)$.

From (32), we derive the following adapted Carleman estimate which is proved in [3, 4, 6].

Proposition 4. *Assume that (2) holds. Let θ be defined by (31). Then, there exists $\lambda_0 = \lambda_0(\Omega,\omega,a_0) > 1$, $s_0 = s_0(\Omega,\omega,a_0,T) > 1$ and $C = C(\Omega,\omega,a_0,T) > 0$ such that for any $\lambda \ge \lambda_0$ and $s \ge s_0$ and for any $\rho \in \mathscr{V}$, we have*

$$\int_Q \frac{1}{\theta^2}|\rho|^2\,dxdt \le C\left(\int_Q |L^*\rho|^2\,dxdt + \int_G |\rho - P\rho|^2\,dxdt\right). \tag{33}$$

Now, we consider the following symmetric bilinear form:

$$a(\rho,\hat{\rho}) = \int_Q L^*\rho L^*\hat{\rho}\,dxdt + \int_G (\rho - P\rho)(\hat{\rho} - P\hat{\rho})\,dxdt. \tag{34}$$

According to Proposition 4, this symmetric bilinear form is a scalar product over $\mathscr{V}$. Let $V = \overline{\mathscr{V}}$ the completion of $\mathscr{V}$ with respect to the norm

$$\rho \mapsto \|\rho\|_V = \sqrt{a(\rho,\rho)}. \tag{35}$$

Then, V is a Hilbert space.

Assume that (2) holds. Let H be a Hilbert space defined by (6) and $h \in H$. Let also θ and $k_0(h)$ be, respectively, defined by (31) and (18). Then, thanks to the estimation of $\theta k_0(h)$ given by (21) and the Cauchy–Schwartz inequality, the linear form defined on V by

$$\rho \mapsto \int_Q h\rho\,dxdt + \int_G k_0(h)\rho\,dxdt$$

is continuous on V. Thus, Lax–Milgram theorem allows us to say that for any $h \in H$, there exists a unique $\rho_\theta = \rho_\theta(h) \in V$ solution of the variational equation

$$\begin{aligned} a(\rho_\theta,\rho) &= \int_Q L^*\rho_\theta L^*\rho\,dxdt + \int_G (\rho_\theta - P\rho_\theta)(\rho - P\rho)\,dxdt,\ \forall \rho \in V, \\ a(\rho_\theta,\rho) &= \int_Q (h + k_0(h)\chi_\omega)\rho\,dxdt,\ \forall \rho \in V. \end{aligned} \tag{36}$$

Proposition 5. *Assume that (2) holds. Let $h \in H$, and let ρ_θ be the unique solution of (36). Set*

$$v_\theta = -(\rho_\theta \chi_\omega - P\rho_\theta) \tag{37}$$

and

$$y_\theta = L^*\rho_\theta. \tag{38}$$

Then, the pair (v_θ, y_θ) *is such that* (12)–(15) *hold.*

Moreover, there exists $C = C(\Omega, \omega, a_0, T, \sum_{i=1}^{M} \|e_i\|_{L^2(Q)}) > 0$ *such that*

$$\|\rho_\theta\|_V \leq C\|h\|_H, \tag{39}$$

$$\|v_\theta\|_{L^2(G)} \leq C\|h\|_H, \tag{40}$$

$$\|y_\theta\|_{\Xi^{1,2}(Q)} \leq C\|h\|_H. \tag{41}$$

Proof. We proceed in two steps.

Step 1. We prove that (v_θ, y_θ) is solution of (12)–(15).

Since $\rho_\theta \in V$, then $v_\theta = -(\rho_\theta \chi_\omega - P\rho_\theta) \in L^2(G)$ and $y_\theta \in L^2(Q)$. As $P\rho_\theta \in \mathscr{M}$, the function $v_\theta \in \mathscr{M}^\perp$. Replacing $-(\rho_\theta \chi_\omega - P\rho_\theta)$ by v_θ and $L^*\rho_\theta$ by y_θ in (36), we have

$$\int_Q y_\theta L^*\rho \, dxdt + \int_G (\rho_\theta - P\rho_\theta)(\rho - P\rho) \, dxdt = \int_Q (h + k_0\chi_\omega)\rho \, dxdt.$$

As $P\rho \in \mathscr{M}$, then

$$\int_Q y_\theta L^*\rho \, dxdt \int_G (\rho_\theta - P\rho_\theta)\rho \, dxdt = \int_Q (h + k_0\chi_\omega)\rho \, dxdt \; \forall \rho \in V.$$

This means that

$$\int_Q y_\theta L^*\rho \, dxdt = \int_Q (h + k_0\chi_\omega)\rho \, dxdt + \int_G v_\theta \rho \, dxdt, \; \forall \rho \in V. \tag{42}$$

Actually, y_θ is the weak solution of a heat equation. Indeed, for $\phi \in L^2(Q)$, let $\mathfrak{p}$ be the solution of

$$\begin{cases} -\mathfrak{p}' - \Delta\mathfrak{p} + a_0\mathfrak{p} = \phi \text{ in } Q, \\ \mathfrak{p} = 0 \text{ on } \Sigma, \\ \mathfrak{p}(0) = 0 \text{ in } \Omega. \end{cases}$$

Thus, $\mathfrak{p} \in V$, and replacing ρ in (42) by $\mathfrak{p}$, we obtain

$$\int_Q y_\theta \phi \, dxdt = \int_Q (h + k_0\chi_\omega)\mathfrak{p} \, dxdt + \int_G v_\theta \mathfrak{p} \, dxdt.$$

Consequently, y_θ is the weak solution, by transposition of the system (14) with $k = v_\theta$ (see [2]). And we know that the solution of this equation is in $\Xi^{1,2}(Q)$. Hence, $y_\theta \in C([0,T], L^2(\Omega))$. Then, multiplying the first equation of (3) by $\varphi \in \mathscr{V}$ and integrating by parts over Q, it follows that for any $\varphi \in \mathscr{V}$,

$$\int_\Omega y_\theta(T)\varphi(T)\,dx - \int_\Omega y_\theta(0)\varphi(0)\,dx + \int_Q y_\theta L^*\varphi\,dxdt = \int_Q (h + k_0\chi_\omega)\varphi\,dxdt + \int_G v_\theta\varphi\,dxdt.$$

As $\varphi \in \mathscr{V}$, we deduce from (42) that

$$\int_\Omega y_\theta(T)\varphi(T)\,dx = 0,\ \forall \varphi \in \mathscr{V}.$$

Therefore, $y_\theta(T) = 0$ in Ω. Consequently, the pair (v_θ, y_θ) is solution of the problem (12)–(15).

Step 2. Let us prove the estimates (39)–(41).

Replacing φ by ρ_θ in (36), it follows from (33) and (21) that

$$\begin{aligned} a(\rho_\theta,\rho_\theta) &= \|y_\theta\|^2_{L^2(Q)} + \|v_\theta\|^2_{L^2(G)}, \\ &\le \|\theta(h+k_0)\|_{L^2(Q)} \|\frac{1}{\theta}\rho_\theta\|_{L^2(Q)}, \\ &\le C\|h\|_H \|\rho_\theta\|_V. \end{aligned}$$

From the definition of the norm on V given by (35), we obtain (39) and then (40). Finally, (41) is a consequence of (40) and the classic properties of heat equations.□

Proposition 6. *Assume that the assumptions of Proposition 5 hold. Then there exists a unique control v such that*

$$v = \min_{\tilde{v}\in\mathscr{E}} \|\tilde{v}\| \tag{43}$$

where $\mathscr{E} = \{\tilde{v} \in \mathscr{M}^\perp \mid (\tilde{v},\tilde{y})$ *satisfies* $(12)-(15)\}$.

Furthermore, there exists $C = C(\Omega,\omega,a_0,T,\sum_{i=1}^{M}\|e_i\|_{L^2(Q)}) > 0$ *such that*

$$\|v\|_{L^2(G)} \le C\|h\|_H. \tag{44}$$

Proof. According to Proposition 5, the pair (v_θ, y_θ) satisfies (12)–(15). Consequently, $\mathscr{E}$ is nonempty. Since $\mathscr{E}$ is also a closed convex subset of $L^2(G)$, we deduce that there exists a unique control v of minimal norm in $L^2(G)$. Particularly,

$$\|v\|_{L^2(G)} \le \|v_\theta\|_{L^2(G)}.$$

Hence, using (40), we obtain (44). □

From now on, we denote by $v = \mathscr{F}(h)$ the optimal control verifying (43) and by $y(h,k(h))$ the optimal state with $k(h) = k_0(h) + \mathscr{F}(h)$.

Penalization Method

In this subsection, we characterize the optimal solution. To this end, we use a penalization method of Lions (see [2]).
Let

$$\begin{cases} u \in \mathscr{M}^{\perp},\ z \in L^2(Q), \\ z' - \Delta z \in L^2(Q),\ z = 0 \text{ on } \Sigma, \\ z(0) = 0,\ z(T) = 0. \end{cases} \tag{45}$$

We define for any $h \in H$ and for any (u,z) verifying (45),

$$I_\varepsilon(u,z) = \frac{1}{2}\|u\|^2_{L^2(G)} + \frac{1}{2\varepsilon}\|Lz - h - k_0 - u\chi_\omega\|^2_{L^2(Q)} \tag{46}$$

and we consider the following problem

$$\inf\{I_\varepsilon(u,z),\ (u,z) \text{ verifying (45)}\}. \tag{47}$$

Since I_ε is coercive, continuous, and strictly convex, Problem (47) admits a unique solution $(v_\varepsilon = v_\varepsilon(h), y_\varepsilon = y_\varepsilon(h))$, i.e.,

$$I_\varepsilon(v_\varepsilon, y_\varepsilon) \le I_\varepsilon(u,z).$$

We give now the optimality system verified by $(v_\varepsilon, y_\varepsilon)$.

Proposition 7. *Assume that the assumptions of Proposition 5 hold. Then, the following assertions are equivalent:*

(i) $(v_\varepsilon, y_\varepsilon) \in \mathscr{M}^{\perp} \times \Xi^{1,2}(Q)$ *is an optimal solution of Problem* (47).
(ii) *There exists* $\rho_\varepsilon \in V$ *such that the triplet* $(v_\varepsilon, y_\varepsilon, \rho_\varepsilon)$ *is solution of the following optimality system:*

$$v_\varepsilon = -(\rho_\varepsilon\chi_\omega - P\rho_\varepsilon) \in \mathscr{M}^{\perp} \tag{48}$$

$$\begin{cases} y'_\varepsilon - \Delta y_\varepsilon + a_0 y_\varepsilon = h + k_0\chi_\omega + v_\varepsilon\chi_\omega - \varepsilon\rho_\varepsilon \text{ in } Q, \\ y_\varepsilon = 0 \quad \text{on } \Sigma, \\ y_\varepsilon(0) = 0 \quad \text{on } \Omega, \end{cases} \tag{49}$$

$$y_\varepsilon(T) = 0 \text{ in } \Omega, \tag{50}$$

$$\begin{cases} -\rho'_\varepsilon - \Delta\rho_\varepsilon + a_0\rho_\varepsilon = 0 \text{ in } Q, \\ \rho_\varepsilon = 0 \text{ on } \Sigma. \end{cases} \tag{51}$$

Proof. We express the Euler–Lagrange optimality conditions which characterize $(v_\varepsilon, y_\varepsilon)$.

$$\begin{cases} \dfrac{d}{d\lambda} I_\varepsilon(v_\varepsilon, y_\varepsilon + \lambda\varphi)_{|\lambda=0} = 0,\ \forall \varphi \in C^\infty(\overline{Q}) \text{ such that} \\ \varphi = 0 \text{ on } \Sigma,\ \varphi(0) = \varphi(T) = 0 \text{ in } \Omega, \end{cases}$$

$$\frac{d}{d\lambda} I_\varepsilon(v_\varepsilon + \lambda v, y_\varepsilon)_{|\lambda=0} = 0,\ \forall v \in \mathscr{M}^{\perp}.$$

After calculations, we have

$$\begin{cases} \displaystyle\int_Q \frac{1}{\varepsilon}\,(Ly_\varepsilon - h - k_0\chi_\omega - v_\varepsilon\chi_\omega)\,L\varphi\,dxdt = 0, \\ \forall \varphi \in C^\infty(\overline{Q}) \ \text{ such that }, \ \varphi = 0 \text{ on } \Sigma, \ \varphi(0) = \varphi(T) = 0 \text{ in } \Omega \end{cases} \tag{52}$$

and

$$\int_G v_\varepsilon v\,dxdt - \int_Q \frac{1}{\varepsilon}\,(Ly_\varepsilon - h - k_0\chi_\omega - v_\varepsilon\chi_\omega)\,v\,dxdt = 0, \ \forall v \in \mathscr{M}^\perp. \tag{53}$$

Then we define the adjoint state

$$\rho_\varepsilon = \rho_\varepsilon(h) = -\frac{1}{\varepsilon}(Ly_\varepsilon - h - k_0\chi_\omega - v_\varepsilon\chi_\omega). \tag{54}$$

Hence, we deduce that $Ly_\varepsilon = h + k_0\chi_\omega + v_\varepsilon\chi_\omega - \varepsilon\rho_\varepsilon \in L^2(Q)$. And, since $(v_\varepsilon, y_\varepsilon)$ verifies (45), we have $y_\varepsilon = 0$ on Σ, $y_\varepsilon(0) = 0$ in Ω, and $y_\varepsilon(T) = 0$ in Ω. Thus, $(v_\varepsilon, y_\varepsilon, \rho_\varepsilon)$ is such that (49)–(50) hold. Since $h + k_0\chi_\omega + v_\varepsilon - \varepsilon\rho_\varepsilon \in L^2(Q)$, we obtain that $y_\varepsilon \in \Xi^{1,2}(Q)$. Now, replacing $-\dfrac{1}{\varepsilon}\,(Ly_\varepsilon - h - k_0\chi_\omega - v_\varepsilon\chi_\omega)$ by ρ_ε, in (52) and (53), we, respectively, obtain

$$\begin{cases} \displaystyle\int_Q \rho_\varepsilon L\varphi\,dxdt = 0, \\ \forall \varphi \in C^\infty(\overline{Q}) \ \text{ such that }, \ \varphi = 0 \text{ on } \Sigma, \ \varphi(0) = \varphi(T) = 0 \text{ in } \Omega \end{cases} \tag{55}$$

and

$$\int_G v_\varepsilon v\,dxdt + \int_Q \rho_\varepsilon v\,dxdt = 0, \ \forall v \in \mathscr{M}^\perp. \tag{56}$$

Therefore, from (55), we derive

$$L^*\rho_\varepsilon = -\rho_\varepsilon' - \Delta\rho_\varepsilon + a_0\rho_\varepsilon = 0 \text{ in } Q.$$

Thus, $\rho_\varepsilon \in L^2(Q)$ and $L^*\rho_\varepsilon \in L^2(Q)$. Consequently, we can define ρ_ε on Σ and show that $\rho_\varepsilon = 0$ on Σ.
From (56), we have

$$\int_G (v_\varepsilon + \rho_\varepsilon\chi_\omega)v\,dxdt = 0, \ \forall v \in \mathscr{M}^\perp.$$

Hence, $v_\varepsilon + \rho_\varepsilon\chi_\omega \in \mathscr{M}^\perp$. Since $v_\varepsilon \in \mathscr{M}^\perp$, we have $v_\varepsilon + \rho_\varepsilon\chi_\omega = P(v_\varepsilon + \rho_\varepsilon\chi_\omega) = P\rho_\varepsilon$. Thus, $v_\varepsilon = -(\rho_\varepsilon\chi_\omega - P\rho_\varepsilon) \in \mathscr{M}^\perp$. □

Furthermore, we have the following estimates:

Proposition 8. *Let $(v_\varepsilon, y_\varepsilon, \rho_\varepsilon)$ be defined as in Proposition 7. Then, there exists a positive constant C, independent on ε such that*

$$\|v_\varepsilon\|_{L^2(G)} \leq C, \tag{57}$$

$$\|y_\varepsilon\|_{\Xi^{1,2}(Q)} \leq C, \tag{58}$$
$$\|\rho_\varepsilon \chi_\omega\|_{L^2(G)} \leq C, \tag{59}$$
$$\|\rho_\varepsilon\|_V \leq C. \tag{60}$$

Proof. The structure of I_ε, on the one hand, and the existence of (v_θ, y_θ) on the other hand show that

$$0 \leq I_\varepsilon(v_\varepsilon, y_\varepsilon) \leq I_\varepsilon(v_\theta, y_\theta) = \frac{1}{2}\|v_\theta\|^2_{L^2(Q)} \leq C.$$

Thus, we have (57) and

$$\|Ly_\varepsilon - h - k_0\chi_\omega - v_\varepsilon\chi_\omega\|_{L^2(Q)} \leq C\sqrt{\varepsilon}. \tag{61}$$

Consequently, (54) and (61) give $\|\varepsilon\rho_\varepsilon\|_{L^2(Q)} \leq C\sqrt{\varepsilon}$, and y_ε being solution of (49), we obtain (58), thanks to the regularity properties of heat equations.

Furthermore, since $L^*\rho_\varepsilon = 0$ and (57) holds, using the definition of the norm on V given by (35), we obtain (60).

On the other hand, since $\rho_\varepsilon \in V$, applying the observability inequality (33) to ρ_ε, we have $\|\frac{1}{\theta}\rho_\varepsilon\|_{L^2(G)} \leq C$. Thus, using (48), (57), and the fact that $\frac{1}{\theta} \in L^\infty(Q)$, we deduce that $\|\frac{1}{\theta}P\rho_\varepsilon\|_{L^2(G)} \leq C$. Since $P\rho_\varepsilon \in \mathscr{M}$ which is finite dimensional, we have $\|P\rho_\varepsilon\|_{L^2(G)} \leq C$. Hence, using again (48) and (57), we obtain estimate (59). □

Now, we can pass to the limit when ε tends to zero to obtain the singular optimality system associated to Problem 1.

Proposition 9. *Let $v = \mathscr{F}(h)$ be the unique solution of (43). Let also P be the orthogonal projection operator from $L^2(G)$ into $\mathscr{M}$. Then*

$$\mathscr{F}(h) = -(\rho\chi_\omega - P\rho) \tag{62}$$

where $\rho \in V$ is solution of

$$L^*\rho = 0 \text{ in } Q, \tag{63}$$
$$\rho = 0 \text{ on } \Sigma. \tag{64}$$

Proof. We proceed in three steps.

Step 1. We study the convergence of $(v_\varepsilon, y_\varepsilon)$.

According to (57) and (58), we can extract two subsequences, still denoted $(v_\varepsilon)_\varepsilon$ and $(y_\varepsilon)_\varepsilon$ such that

$$v_\varepsilon \rightharpoonup v_0(h) \quad \text{weakly in} \quad L^2(G), \tag{65}$$
$$y_\varepsilon \rightharpoonup y_0(h) \quad \text{weakly in} \quad \Xi^{1,2}(Q). \tag{66}$$

And, as $v_\varepsilon \in \mathscr{M}^\perp$ which is a closed vector subspace of $L^2(G)$, we have

$$v_0(h) \in \mathscr{M}^\perp. \tag{67}$$

Since the injection of $\Xi^{1,2}(Q)$ into $L^2(Q)$ is compact, the pair $(v_0 = v_0(h), y_0 = y_0(h))$ is such that

$$\begin{cases} y_0' - \Delta y_0 + a_0 y_0 = h + k_0 \chi_\omega + v_0 \chi_\omega & \text{in } Q, \\ y_0 = 0 & \text{on } \Sigma, \\ y_0(0) = 0 & \text{in } \Omega. \end{cases} \tag{68}$$

$$y_0(T) = 0 \text{ in } \Omega. \tag{69}$$

Step 2. We show that $(v_0,\ y_0) = (\mathscr{F}(h), y(h,k(h)))$.
From the expression of I_ε given by (46), we can write

$$\frac{1}{2}\|v_\varepsilon\|^2_{L^2(G)} \leq I_\varepsilon(v_\varepsilon, y_\varepsilon).$$

Since $(\mathscr{F}(h), y(h,k(h)))$ satisfies (12)–(15) and (43), this latter inequality becomes

$$\frac{1}{2}\|v_\varepsilon\|^2_{L^2(G)} \leq I_\varepsilon(v_\varepsilon, y_\varepsilon) \leq \frac{1}{2}\|\mathscr{F}(h)\|^2_{L^2(G)}. \tag{70}$$

Then, using (65) while passing to the limit in (70), we obtain

$$\frac{1}{2}\|v_0\|^2_{L^2(G)} \leq \liminf_{\varepsilon \to 0} I_\varepsilon(v_\varepsilon, y_\varepsilon) \leq \frac{1}{2}\|\mathscr{F}(h)\|^2_{L^2(G)}.$$

Consequently

$$\|v_0\|_{L^2(G)} \leq \|\mathscr{F}(h)\|_{L^2(G)},$$

and thus,

$$\|v_0\|_{L^2(G)} = \|\mathscr{F}(h)\|_{L^2(G)}.$$

Hence,

$$v_0 = \mathscr{F}(h), \tag{71}$$

and since (68) admits a unique solution, it follows that $y_0 = y(h,k(h))$.

Remark 1. *Note that $\|\mathscr{F}(h)\|_{L^2(G)} \leq C\|h\|_H$. Indeed, as (v_θ, y_θ) satisfies (45), we can write*

$$I_\varepsilon(v_\varepsilon, y_\varepsilon) \leq I_\varepsilon(v_\theta, y_\theta) = \frac{1}{2}\|v_\theta\|_{L^2(G)}.$$

Threfore, using the fact that v_θ verifies (40) and the definition of I_ε given by (46), we obtain that $\|v_\varepsilon\|_{L^2(G)} \leq C\|h\|_H$. Hence, in view of (65) and (71), we have $\|\mathscr{F}(h)\|_{L^2(G)} \leq C\|h\|_H$.

Step 3. According to estimates (59) and (60), we can extract a subsequence, still denoted $(\rho_\varepsilon)_\varepsilon$ such that

$$\rho_\varepsilon \chi_\omega \rightharpoonup \rho(h)\chi_\omega \text{ weakly in } L^2(G), \tag{72}$$

$$\rho_\varepsilon \chi_\omega \rightharpoonup \rho(h)\chi_\omega \text{ weakly in } V, \tag{73}$$

and it follows from (51) that $\rho(h)$ is solution of

$$\begin{cases} L^*\rho = 0 \text{ in } Q, \\ \rho = 0 \text{ on } \Sigma. \end{cases}$$

As P is a compact operator, we deduce from (72) that

$$P\rho_\varepsilon \to P\rho(h) \text{ strongly in } L^2(G). \tag{74}$$

Therefore, combining (72) and (74), we obtain

$$v_\varepsilon = -(\rho_\varepsilon \chi_\omega - P\rho_\varepsilon) \rightharpoonup \mathscr{F}(h) = -(\rho(h)\chi_\omega - P\rho(h)) \text{ weakly in } L^2(G).$$

Thus, we have showed that for any $h \in H$, the unique pair $(\mathscr{F}(h), y(h, k(h)))$ satisfies (12)–(15) where $\mathscr{F}(h) = -(\rho(h)\chi_\omega - P\rho(h))$ and $\rho = \rho(h)$ is solution of (63). □

Proof of Theorem 1

We have proven that there exists a unique control $v = v(h) \in \mathscr{M}^\perp$ solution of (43) such that the pair (v, y) verifies (14) and (15). Therefore, Proposition 1 allows us to say that the control $k = k(h) = (k_0(h) + v(h))$ with $k_0 \in \mathscr{M}_\theta$ is such that $(k, y(k))$ satisfies the null-controllability problem with constraints on the state (3), (4), and (5). Therefore, using (22) and (44), we deduce (10).

Optimal Strategy for the Leader

Properties of $\mathscr{F}$

Lemma 3. *For any $h \in H$, let $\rho = \rho(h)$ be the solution of (63). Then, the map $\mathscr{F}$ defined by*

$$\mathscr{F}(h) = -(\rho - P\rho)\chi_\omega \tag{75}$$

is linear and continuous from H into $L^2(G)$.

Proof. Consider the vector subspace V_0 from V defined by

$$V_0 = \{\varphi \in V | L^*\varphi = 0\}.$$

Since $\mathscr{F}(h)$ is solution of problem (12)–(15) and verifies (62), we multiply the first equation of (14) by $\varphi \in V_0$ and we integrate by parts. Then, we obtain

$$\int_Q h\varphi\, dxdt + \int_Q k_0(h)\varphi\, dxdt + \int_Q v\chi_\omega\varphi\, dxdt = 0\ \forall \varphi \in V_0,$$

i.e.,

$$\int_Q h\varphi\, dxdt + \int_Q k_0(h)\varphi\, dxdt - \int_Q (\rho - P\rho)\chi_\omega\varphi\, dxdt = 0\ \forall \varphi \in V_0,$$

or equivalently,

$$\int_Q h\varphi\, dxdt + \int_Q k_0(h)\varphi\, dxdt + \int_Q \mathscr{F}(h)\chi_\omega\varphi\, dxdt = 0\ \forall \varphi \in V_0.$$

Using the fact that the map $\varphi \mapsto \int_Q h\varphi\, dxdt + \int_Q k_0(h)\varphi\, dxdt$ is linear and continuous on V and

$$\begin{aligned} -\int_G \mathscr{F}(h)\varphi\, dxdt &= \int_G (\rho(h) - P\rho(h))\varphi\, dxdt, \\ &= \int_G (\rho(h) - P\rho(h))(\varphi - P\varphi)\, dxdt, \\ &= a(\rho(h), \varphi), \end{aligned}$$

we deduce that $\rho = \rho(h)$ is solution of the variational problem

$$a(\rho, \varphi) = \int_Q h\varphi\, dxdt + \int_Q k_0(h)\varphi\, dxdt\ \forall \varphi \in V_0. \tag{76}$$

Hence, the map $h \mapsto \rho = \rho(h)\chi_\omega$ is linear from H to $L^2(G)$. And since the projection operator $I - P$ which is defined from $L^2(G)$ to $\mathscr{M}^\perp \subset L^2(G)$ is also linear, we deduce that the map $\mathscr{F}$ is linear from H to $L^2(G)$. Hence, it follows from Remark 1 that $\mathscr{F}$ is continuous on H since $\|\mathscr{F}(h)\|_{L^2(G)} \leq C\|h\|_H$. □

Remark 2. *Let k_0 be defined as in (18) . Then*

1. *$k_0 \in H$. Indeed, since $k_0 \in \mathscr{M}_\theta$, we have on the one hand, $k_0 \in L^2(G)$, and on the other hand, $\theta k_0 \in \mathscr{M} \subset L^2(G)$.*
2. *In view of (18), the map $\mathscr{F}_1 : h \mapsto k_0(h)$ is linear, and since (22) holds, this map is continuous on H.*

From now on, we denote $k_0(h) = \mathscr{F}_1(h)$.

Proof of Theorem 2

We consider the cost function J defined by

$$J(h) = \frac{1}{2}\|y(h, k(h)) - z_d\|^2_{L^2(Q)} + \frac{N}{2}\|h\|^2_H \tag{77}$$

from which we associate the minimization problem

$$\inf_{h\in\mathscr{U}_{ad}} J(h) \tag{78}$$

where $\mathscr{U}_{ad}$ is a nonempty closed convex subspace of $L^2(Q)$.
Using the properties of the maps $\mathscr{F}$ given by Lemma 3 and $\mathscr{F}_1$ given by Remark 1, we have that J is strictly convex, continuous, and coercive. Thus, we have the following classic result:

Proposition 10. *Problem* (78) *has a unique control* $\hat{h}\in\mathscr{U}_{ad}$.

Observing that $k(h)=k_0(h)+v(h)=\mathscr{F}_1(h)+\mathscr{F}(h)$, we will denote, now and in the sequel, by $\hat{y}=y(\hat{h},\hat{k}=\hat{k}(\hat{h}))$ the state associated to the optimal control $\hat{h}$. Let us characterize $\hat{h}$.

Writing the Euler–Lagrange condition, we obtain

$$\frac{d}{d\lambda}J(\hat{h}+\lambda(h-\hat{h}))_{|\lambda=0}\geq 0,\ \forall h\in\mathscr{U}_{ad}$$

which after calculations gives

$$\frac{d}{d\lambda}J(\hat{h}+\lambda(h-\hat{h}))_{|\lambda=0}=(\hat{y}-z_d,y(h-\hat{h},k(h-\hat{h}))_{L^2(Q)}+(N\hat{h},h-\hat{h})_H.$$

Thus,

$$(\hat{y}-z_d,y(h-\hat{h},k(h-\hat{h}))_{L^2(Q)}+(N\hat{h},h-\hat{h})_H\geq 0,\ \forall h\in\mathscr{U}_{ad}.$$

We interpret this condition using the adjoint state notion. To make our calculations easier, we set $w=h-\hat{h}$ and we denote $y=y(w,k(w))$. Let p be the solution of the following system:

$$\begin{cases} -p'-\Delta p+a_0p=\hat{y}-z_d & \text{in } Q,\\ p=0 & \text{on } \Sigma,\\ p(T)=0 & \text{in } \Omega. \end{cases} \tag{79}$$

Since $\hat{y}-z_d\in L^2(Q)$, we know that $p\in\Xi^{1,2}(Q)$. Multiply the first equation of (79) by y and integrate by parts over Q, we obtain

$$\int_Q p(w+(\mathscr{F}_1(w)+F(w))\chi_\omega)\,dxdt=\int_Q y(\hat{y}-z_d)\,dxdt.$$

This means that,

$$\int_Q pw\,dxdt+\int_Q p\mathscr{F}_1(w)\chi_\omega\,dxdt+\int_Q p\mathscr{F}(w)\chi_\omega\,dxdt=\int_Q y(\hat{y}-z_d)\,dxdt.$$

Let H' be the dual of the Hilbert space H. Let also Λ^{-1} be the isometric isomorphism from H' to H. Observing on the one hand that $\mathscr{F} = \mathscr{F}^*$ because of the symmetry of the operator $a(.,.)$, and on the other hand that we can write

$$\int_Q pw\,dxdt = \int_Q \frac{1}{\theta} p\theta w = \langle \frac{1}{\theta} p, w\rangle_{H',H},$$

$$\int_G p\mathscr{F}_1(w)\,dxdt = \langle \mathscr{F}_1^*(p), w\rangle_{H',H},$$

and

$$\int_G p\mathscr{F}(w)\,dxdt = \langle \mathscr{F}^*(p), w\rangle_{H',H},$$

we have

$$\int_Q pw\,dxdt = (\Lambda^{-1}(\frac{1}{\theta} p), w)_H,$$

$$\int_G p\mathscr{F}_1(w)\,dxdt = (\Lambda^{-1}\mathscr{F}_1^*(p), w)_H,$$

and

$$\int_G p\mathscr{F}(w)\,dxdt = (\Lambda^{-1}\mathscr{F}(p), w)_H.$$

Therefore, the Euler–Lagrange condition gives

$$\left(\Lambda^{-1}\frac{1}{\theta}(p) + \Lambda^{-1}\mathscr{F}_1^*(p) + \Lambda^{-1}\mathscr{F}(p) + N\hat{h}, h - \hat{h}\right)_H \geq 0,\ \forall h \in \mathscr{U}_{ad}$$

or

$$\left(\Lambda^{-1}(\frac{1}{\theta} I + \mathscr{F}_1^* + \mathscr{F})(p) + N\hat{h}, h - \hat{h}\right)_H \geq 0,\ \forall h \in \mathscr{U}_{ad}$$

where I is the identity operator of $L^2(Q)$.

References

1. Fursikov, A., Imanuvilov, O.Yu.: Controllability of Evolution Equations, Lecture Notes, Research Institute of Mathematics. Seoul National University, Korea (1996)
2. Lions, J.L.: Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles. Dunod, Paris (1968)
3. Massengo-mophou, G., Nakoulima, O.: Null controllability with constraints on the state for the semilinear heat equation. J. Optim. Theory. Appl. **143**(3), 539–565 (2009)
4. Massengo-mophou, G., Nakoulima, O.: Sentinels with given sensitivity. Eur. J. Appl. Math. **19**(1), 21–40 (2008)
5. Mophou, G.: Null controllability with constraints on the state for nonlinear heat equations. Forum Mathematicum (2011)
6. Nakoulima, O.: Contrôlabilité à zéro avec contraintes sur le contrôle. C.R. Acad. Sci. Paris Ser. I **339**, 405–410 (2004)

Chapter 12
Almost and Pseudo-Almost Limit Cycles with Applications to Quasiperiodic Solitary Waves

Bourama Toni and Melissa Watts

Introduction

Periodicity plays an essential role in several natural and man-made systems and is apparent, for example, in simple models of the solar system, in the circadian rhythms by which basic biological functions are regulated, and in electronic devices producing stable periodic signals such as in wireless communications. Periodic trajectories, isolated or otherwise, are crucial in the mathematics of dynamical systems and its applications to science and engineering by virtue of the importance of periodic phenomena as well as by the formidable intellectual challenges in detecting and predicting periodicity.

One important aspect of periodicity is described by the so-called limit cycles, isolated periodic orbits in the phase space, stable or attractive when the neighboring solutions tend to them in an asymptotic sense or unstable if the neighboring solutions unwind from them. As such they can be seen as a set of accumulation points of either the forward or backward trajectory.

Limit cycles, when stable, actually model the dynamical state of self-sustained oscillations found very often in nature, with examples in biology, chemistry, mechanics, electronics, fluid dynamics, etc. See, for example, [3, 4, 8, 19, 21]. They often arise in many physical systems around a state at which energy generation and dissipation balance. One of the most important limit cycles of our lives is the heartbeat. A spectacular example is the Tacoma Narrows Bridge and its 1940 dramatic collapse, where the limit cycle drew its energy from the wind and involved torsional oscillations of the roadbed of about 70°. Dynamic walking in Robotics is another practical example; the stable gait to which the repeated walking pattern converges

Bourama Toni (✉) • Melissa Watts
Department of Mathematics and Computer Sciences, Virginia State University, Petersburg, VA, USA
e-mail: btoni@vsu.edu; mwatts@vsu.edu

B. Toni (ed.), *Advances in Interdisciplinary Mathematical Research*, Springer Proceedings in Mathematics & Statistics 37, DOI 10.1007/978-1-4614-6345-0_12,

is modeled as a stable limit cycle, stability easily lost to even small disturbances, evidence of a narrow basin of attraction of the limit cycle.

Planar limit cycles were defined by Poincaré in the famous paper *Mémoire sur les courbes définies par une équation différentielle* [28], using his so-called Method of Sections, described in section "Overview of Limit Cycles". However, much attention in this century has been drawn to the determination of the number, amplitude, and configuration of limit cycles in a general nonlinear system, which is still an unsolved problem. This is part of the so-called Hilbert's 16th Problem. A weakened version by Arnold called the *tangential Hilbert's problem* concerns the bound on the number of limit cycles which can bifurcate from a first-order perturbation of a Hamiltonian system

$$\dot{x} = -H_y + \varepsilon P(x,y), \quad \dot{y} = H_x + \varepsilon Q(x,y), \tag{1}$$

$P(x,y)$ and $Q(x,y)$ are polynomials of degree $deg(P,Q) \leq n$, and $H(x,y)$ is the Hamiltonian of degree $degH(x,y) = n+1$. The limit cycles appearing in the perturbed system are given by the isolated zeros of the abelian integral (integral of a rational one form along an algebraic oval)

$$I(c) := \oint_{\gamma_c : H=c} P(x,y)dy - Q(x,y)dx. \tag{2}$$

If $I(c_0) = 0, \quad I'(c_0) \neq 0$, then there is a unique hyperbolic (defined below) limit cycle bifurcating from the level set $\gamma_{c_0} : H = c_0$ [4, 11, 15, 16]. Similar analysis was used by Toni for explicitly linearizable polynomial systems [32].

Existence/Nonexistence of Periodicity

The existence or nonexistence of periodic orbits, in particular limit cycles, is investigated in various ways. The possibility of a limit cycle on a plane or a two-dimensional manifold is restricted to nonlinear dynamical systems, due to the fact that, for linear systems, $kx(t)$ is also a solution for any constant k if $x(t)$ is a solution. Therefore, the phase space will contain an infinite number of closed trajectories encircling the origin, with none of them isolated. Conservative and gradient systems do not have limit cycles, though these systems may exhibit almost or pseudo-almost limit cycles [13]. We overview here the most common techniques for predicting the absence or existence of periodicity and limit cycles.

1. Index Theory: The interested reader may find definitions and more details in [4, 12, 15, 21]. The index of a limit cycle is 1. If all equilibria inside the periodic orbit (isolated or not) are hyperbolic, there must be an odd number $2n+1$ of equilibria, n saddle points, and $n+1$ sinks or sources. So if the appropriate equilibria are not present in a region of the phase space, a periodic orbit cannot exist. And if the sum of the indices of the equilibria enclosed in a region does not equal unity,

then a closed path cannot exist in such a region. Moreover, a closed path cannot surround a region containing no equilibrium nor one containing only a saddle point. However, the relationship between equilibria and periodic orbits does not immediately generalize to higher dimensions. The system

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = -x_1, \quad \dot{x}_3 = 1 - (x_1^2 + x_2^2) \tag{3}$$

has no equilibria but has periodic phase paths given by the helices $x_1^2 + x_2^2 = 1$, $x_3 = b$ (b constant). See, for instance, [3].

2. Dulac's Criterion: There are no periodic orbits lying entirely in a simply connected region D where the divergence of $B\mathscr{X}$ is not identically zero and does not change sign, with B a scalar function defined on D and $\mathscr{X}$ the planar vector field. For instance, the system

$$\dot{x} = y, \quad \dot{y} = -x - y + x^2 + y^2 \tag{4}$$

is actually a perturbation of the linear center or linear isochrone, with a continuum of periodic orbits around the origin. But it has no periodic orbits by the Dulac test using $B = e^{-2x}$. All periodic orbits were therefore destroyed by the perturbation. See, for example, [4, 7, 12, 21].

3. Poincaré–Bendixson Test. If a trajectory enters and does not leave a closed and bounded region of phase space with no equilibria, then the trajectory must approach a limit cycle for increasing time. See, for example, [4, 7, 12, 16, 21].

4. Bifurcation theory. A bifurcation, qualitative change in the behavior of the system as the system parameter is varied, could involve a change of stability of the periodic orbit and/or the creation/destruction of periodic orbits. See the above example of Hamiltonian system. For example, it is known that at most k, limit cycles (of small amplitude) bifurcate out of a weak focus of order k under a perturbation of the coefficients [4].
More importantly, Melnikov's theory is a powerful tool for predicting the number, positions, and multiplicities of limit cycles that bifurcate from homoclinic and heteroclinic orbits under perturbations, by associating to a given dynamical system a function whose roots are related to the existence and location of limit cycles. It has been developed for the analysis of planar systems

$$\dot{u}(t) = f(u) + \varepsilon g(u), \tag{5}$$

for $u \in \mathbf{R}^2$, $\varepsilon \ll 1$ and f, and g sufficiently smooth functions, assuming that the unperturbed system at $\varepsilon = 0$ has a one-parameter family of τ_r-periodic solutions γ_r. Then the Melnikov function is given by

$$\mathscr{M}(r) = \int_0^{\tau_r} e^{\int_0^t \nabla f(\gamma_r(s))ds} f \wedge g(\gamma_r(t))dt, \tag{6}$$

where the wedge product of $u = (u_1, u_2)$ and $v = (v_1, v_2)$ in $\mathbf{R}^2$ is $u \wedge v = u_1 v_2 - u_2 v_1$. Therefore, if there exist $r_j \quad j = 1, \ldots, n$ such that $\mathscr{M}(r_j) = 0$, with

$\mathscr{M}'(r_j) \neq 0$, then the system has n hyperbolic limit cycles in an $O(\varepsilon)$ neighborhood of γ_{r_j} that bifurcate from the periodic orbits $\gamma_{r_j}(t)$. And if $\mathscr{M}(r_0) \neq 0$, then the system has no limit cycles in an $O(\varepsilon)$ neighborhood of γ_{r_0}. See, for instance, [11, 15, 16, 32].

5. Configuration of limit cycles. Any configuration C of closed curves, that is, any finite set of mutually disjoint closed curves, is *realizable* as a configuration of limit cycles by a polynomial vector field of degree n, as well as a configuration of algebraic limit cycles by a polynomial vector field of degree $\leq 2(n+r)-1$ where r is the number of its primary curves (containing no other curves). By realizable we mean topologically equivalent with the existence of a homeomorphism between the set of closed curves and the set of limit cycles. An algebraic closed curve is a connected component of the zero set of some polynomial function. See for instance [11, 15].
6. The Toroidal Principle. If a smooth vector field $\mathscr{X}$ leaves a toroidal region (a submanifold M in $\mathbf{R}^n$ diffeomorphic to $\mathbf{D}^{n-1} \times S^1$) positively invariant and has a section S diffeomorphic to the closed unit disk $\mathbf{D}^{n-1}$, then $\mathscr{X}$ has a periodic orbit in M by Brouwer's fixed point theorem. ($\mathbf{D}^n$ is the closed unit disk in $\mathbf{R}^n$.) See [8].

Remarks

The nonlinear character of isolated periodic oscillations renders their detection and construction challenging. In mechanical terms the appraisal of the regions of the phase plane where energy loss and energy gain occur might reveal a limit cycle, for example, in the family of equations of the form

$$\ddot{x} + \varepsilon h(x,\dot{x}) + x = 0, \tag{7}$$

with a small nonlinearity for $\varepsilon \ll 1$. In particular we have the well-known case of $h(x,\dot{x}) = (x^2-1)\dot{x}$ for the Van del Pol equation. In the absence of a forcing term, it has a single, self-excited oscillation approached from all nonzero initial conditions, that is, a stable limit cycle [18, 19, 21].

Let us emphasize that even though in most studies periodicity has been illustrated more frequently, the occurrence of almost and pseudo-almost periodic oscillations or waves is actually much more common than that of periodic ones. For instance, in the simplest model of harmonic oscillator or mathematical pendulum, as well as for the one-dimensional wave equation, diverse kinds of oscillatory trajectories can be displayed, both periodic and more generally nonperiodic.

The theory of almost periodic functions introduced by H. Bohr [6] is connected with problems in differential equations, stability theory, dynamical systems, partial differential equations, or equations in Banach spaces. There are several results concerning the existence and uniqueness of almost periodic solutions for first-order differential equations, for example, in [13, 14, 16, 25, 26, 29]. But in most of these

works the authors derived almost periodic solutions from the existence of bounded solutions.

We extend the theory of limit cycles to that of almost and pseudo-almost limit cycles, isolated almost/pseudo-almost periodic orbits, and we discuss in the current and future work the usual questions of conditions of existence and uniqueness, stability and asymptotic stability, bifurcation and perturbation, the coexistence of limit cycles and almost/pseudo-almost limit cycles, and introduce the idea of *almost isochrons* and *pseudo-almost isochrons*. Section "Overview of Limit Cycles" overviews the theory of limit cycles with some examples and presents the concept of isochrons. Section "Almost Limit Cycles" is devoted to almost limit cycles and includes definition, properties, examples, and the main existence theorem for Liénard systems. In Section "Pseudo-Almost Limit Cycles", we present the concept of pseudo-almost limit cycle, its properties, several illustrative examples including the so-called linear pseudo-center, and existence theorem in the case of Liénard systems. The section shows the applications of the existence theorems for Liénard systems to obtain almost and pseudo-almost periodic waves for some hyperbolic and parabolic partial differential equations. Finally in Section "Almost and Pseudo-Almost Periodic Waves" we discuss some directions for future research, and state several open problems, defining in the process the concept of almost isochrons and pseudo-almost isochrons. One important question is the requirements for transition from almost or pseudo-almost periodic behavior to a chaotic behavior.

Overview of Limit Cycles

Let the multidimensional space $\mathbf{R}^n$ represents all the possible states of a system modeling nonlinear phenomena. The dynamics of the system are determined by the values in $\mathbf{R}^n$ in terms of the time. That is to say we define an *evolution map or flow* Φ, smooth on the smooth manifold $\mathbf{R}^n$:

$$\Phi : \quad \mathbf{R}^n \times \mathbf{R} \longrightarrow \mathbf{R}^n, \tag{8}$$

such that $\Phi(x,t) = y$ indicates that the state $x \in \mathbf{R}^n$ evolved into the state $y \in \mathbf{R}^n$ after t units of time, together with the usual flow properties

$$\Phi(x,0) = x, \quad \Phi(x,t_1+t_2) = \Phi(\Phi(x,t_1),t_2). \tag{9}$$

The flow Φ then determines a *vector field* $\mathscr{X}$ (conversely as well) such that, for $x \in M$

$$\mathscr{X}(x) := \frac{\partial \Phi}{\partial t}(x,0). \tag{10}$$

The orbit or trajectory of the flow through $x \in \mathbf{R}^n$ is given by

$$O(x) := \{\Phi_x(t) := \Phi(x,t) | t \in \mathbf{R}\}. \tag{11}$$

Definition 1. The orbit $\gamma = O(x)$ based at x is called a limit cycle if there is a neighborhood V of γ such that γ is the only periodic orbit contained in V. The limit cycle is stable (unstable) if $\omega(s) = \gamma$ ($\alpha(s) = \gamma$) for any $s \in V$ that is, γ is the ω-limit set (α-limit set) of any point in V.

In other words, a limit cycle is an isolated periodic orbit of some period τ, that is stable (resp. unstable) if it has a neighborhood U such that, for some distance function d on $\mathbf{R}^n$, $d(\Phi(y,t),\gamma) \longrightarrow 0$, as $t \to \infty$ (resp. $t \to -\infty$), for any $y \in U$.

Note that the phase φ of a limit cycle refers to the relative position on the orbit, which is measured by the elapsed time (modulo the period) to go from a reference point to the current position on the limit cycle.

Examples: Linear Center and Its Perturbations

Example 1

The linear center or *linear isochrone*

$$\dot{x} = -y, \quad \dot{y} = x, \tag{12}$$

where the origin of the plane is surrounded by a continuum of periodic orbits (not isolated) given by $x^2 + y^2 = c > 0$, is perturbed into the following system, in polar coordinates (r, θ)

$$\dot{r} = r(1 - r^2), \quad \dot{\theta} = 1. \tag{13}$$

The circle $r = 1$ is a 2π-periodic orbit and is unique. It is therefore a limit cycle. Moreover r is a monotone function on each orbit ($\dot{r} > 0$ inside and < 0 outside) so that all nonconstant orbits tend towards the limit cycle which is therefore stable. This system is the so-called Poincaré Oscillator as in the figure below (Fig. 12.1).

Example 2

The linear center could also be perturbed into a system to generate several limit cycles as in the following example. The C^∞-system

$$\dot{x} = -y + xf(x,y), \quad \dot{y} = x + yf(x,y), \tag{14}$$

where

$$f(x,y) = \sin\left(\frac{1}{x^2+y^2} e^{-\frac{1}{x^2+y^2}}\right),$$

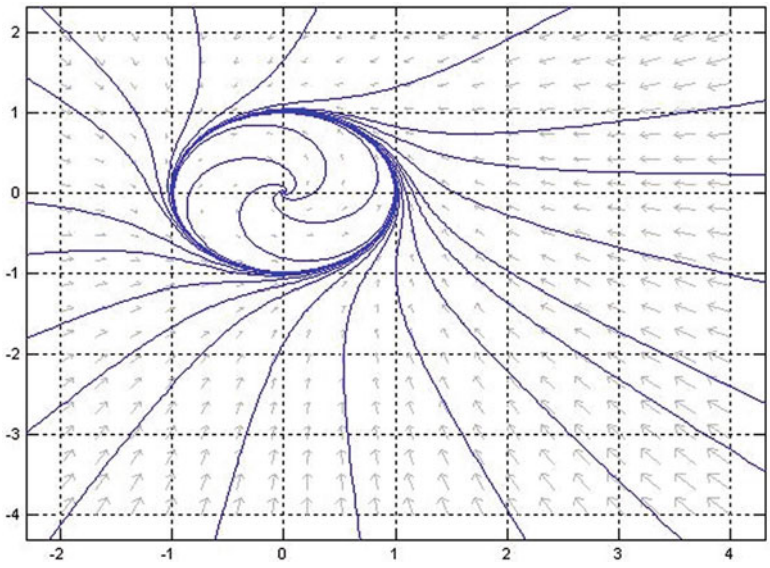

Fig. 12.1 $\dot{r} = r(1-r), \quad \dot{\theta} = 1$

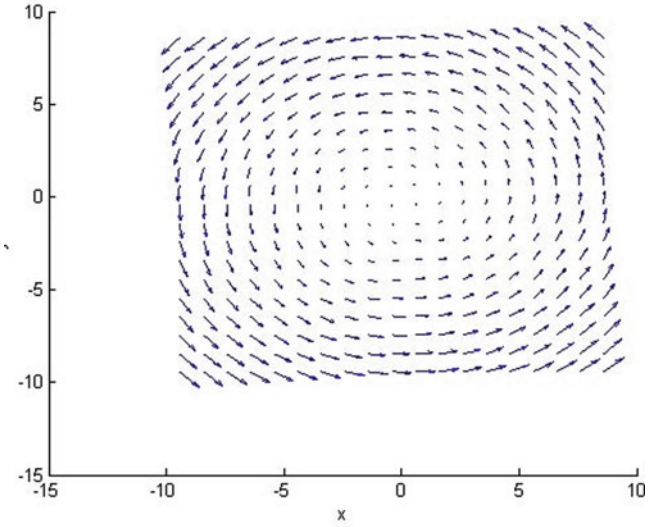

Fig. 12.2 $\dot{x} = -y + xf(x,y), \quad \dot{y} = x + yf(x,y)$ where $f(x,y) = \sin(\frac{1}{x^2+y^2})e^{-\frac{1}{x^2+y^2}}$

has an infinite number of limit cycles

$$\gamma_n : \quad x^2 + y^2 = \frac{1}{n\pi}, \quad n \in \mathbf{Z} \tag{15}$$

accumulating at the origin. The phase portrait appears below in Fig. 12.2.

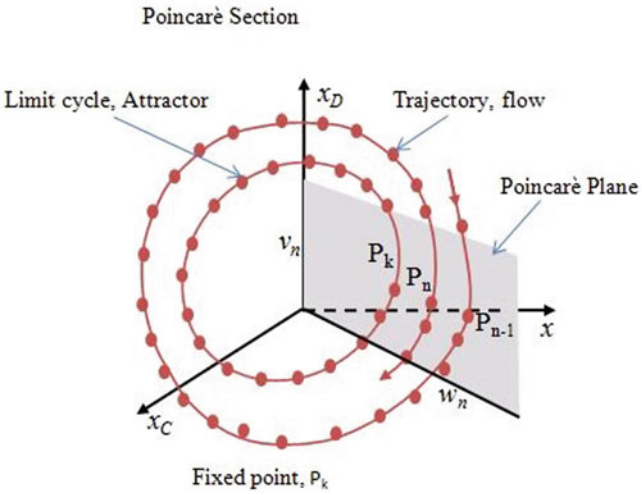

Fig. 12.3 Poincaré section

Poincaré's Method of Sections

Poincaré first observed that if a τ-periodic orbit γ exists for a smooth vector field $\mathscr{X}$, and if $x_0 \in \gamma$, and $\mathscr{H}$ is a hyperplane complementary to the tangent line $T_{x_0}(\gamma)$ to x_0 at γ, then there is a sufficiently small neighborhood, called a *local section* or *cross section,* $\Sigma \subset \mathscr{H}$ on which the implicit function theorem provides for each $x \in \Sigma$ a least positive time t_x for the solution based at x to first return to Σ, defining the so-called smooth Poincaré or "first return" map (monodromy operator) $\mathscr{P}$ on Σ. In other words, we have

1. $x_0 \in \Sigma$, and $\bar{\Sigma} \cap \gamma = \{x_0\}$. ($\bar{\Sigma}$ denotes the closure of Σ.)
2. $T_{x_0}\Sigma + T_{x_0}\gamma = T_{x_0}\mathscr{H}$. ($\Sigma$ is transverse to γ at x_0.)

By continuity of the flow, and the implicit function theorem, the time τ_x of first return exists and is near the period τ for a point x near x_0. Therefore, in practice, $\mathscr{P}(y) = \Phi_{\tau_x}(x)$, where τ_x is the time taken by the orbit $\Phi_x(t)$ to first return to Σ. And $\tau_x \to \tau$ as $x \to x_0$. Of course, $\mathscr{P}(x_0) = x_0$, that is, x_0 is a fixed point for the map $\mathscr{P}$. And the existence of fixed points for $\mathscr{P}$ implies the existence of periodic orbits for the flow, allowing for the use of powerful topological fixed point theorems. But the existence of such a section is itself one of the standard paradigms of the existence of nonlinear oscillations (Fig. 12.3).

Next consider the *monodromy operator* given by the matrix $D_{x_0}\mathscr{P} = [\frac{\partial \mathscr{P}}{\partial x}(x_0)]$ of partial derivatives of $\mathscr{P}$ at x_0. The limit cycle is said to be *hyperbolic or elementary* if $D_{x_0}\mathscr{P}$ has no eigenvalue of modulus one. The eigenvalues are the so-called characteristic (Floquet) multipliers of γ and are independent of the choice of x_0 and Σ. A hyperbolic limit cycle is stable (resp. unstable) if it has all the multipliers

with modulus less than one (resp. greater than one). Each orbit in the neighborhood of γ tends toward (resp. away from) γ exponentially fast. For a planar vector field $\mathscr{X}(x,y) = P(x,y)\partial x + Q(x,y)\partial y$, with P and Q at least C^1, sufficient conditions for stability are given by the following: for τ the period of the limit cycle γ, $I(\tau) = \int_0^\tau (\frac{\partial P(x,y)}{\partial x} + \frac{\partial Q(x,y)}{\partial y})dt$ is negative for stable limit cycle and positive for unstable limit cycle; such limit cycles are said to be *hyperbolic*. A multiple limit cycle is obtained for $I(\tau) = 0$ [11, 15, 28].

The idea of a constant first return time identical to the period of the limit cycle leads to the description of isochrons which we introduce next.

Isochrons

Winfree in [33, 34] introduced the isochrons of limit cycles in biosciences, in particular in relation to biological rhythms. Then Guckenheimer showed that they are in fact the stable manifolds of a point on an attractive hyperbolic limit cycles. Their existence for nonhyperbolic limit cycles was proved by Chicone in [10].

Definition 2. *For a hyperbolic stable limit cycle γ of period τ and for $x_0 \in \gamma$, the isochron at x_0, denoted by $Is(x_0)$, is defined as a cross section of γ at x_0 for which the time of first return is identically the period τ.*

In other words, the isochrons of a limit cycle is the set of points from which state trajectories evolve to the same phase as the limit cycle. That is, a set of initial conditions resulting in oscillations having the same phase. The limit cycle itself, like the unit circle, can be parameterized by one variable called its phase φ.

The existence of isochrons is ensured by the Invariant Manifold Theorem as the leaves of the invariant foliation of the stable manifold of a hyperbolic periodic orbit. In a 2-state system the foliation is visualized as lines traversing the limit cycle. They are used extensively in investigating the dynamics of neural oscillators and to qualitatively illustrate phase resetting in circadian rhythms.

In practice, for a hyperbolic limit cycle γ, there exists a unique $\vartheta(x)$ for any $x \notin \gamma$ such that

$$\lim_{t \leftarrow \infty} |\Phi(t) - \gamma(t + \vartheta(x))| = 0, \tag{16}$$

where $\Phi(t)$ is the trajectory based at x. The value $\vartheta(x)$, bounded by the period T (1 or 2π after normalization), is called the *asymptotic (or latent) phase* of x.

A level set $\vartheta(x) = c$ or $\vartheta^{-1}(c)$ defines an isochron. And it is an $(n-1)$-dimensional hyperplane. In fact all points of an isochron are points of the sequence $\{x(kT)\}_{k \geq 0}$. That is, points on the forward orbit $\Phi(t)$ observed only at times integer multiple of the period of the limit cycle, thereby defined by a Poincaré map. Therefore, an isochron is a special Poincaré section with the time of first return equals the period of the limit cycle. A *phaseless set* is formed by those points where isochrons cannot be defined. See, for example, [3, 19, 33, 34].

Example

Consider the planar differential equations in polar coordinates (r,θ)

$$\dot{r} = (r-1)r^2, \quad \dot{\theta} = r, \tag{17}$$

with a limit cycle $\gamma: \quad r = 1$. Looking for a function f such that the asymptotic phase is defined by $\vartheta(r,\theta) = \theta - f(r)$ leads to each isochron $\vartheta^{-1}(c)$ being defined by $\theta = c + \frac{1}{r} - 1$. Therefore, isochrons exist everywhere in the plane, with the phaseless set reduced to the singleton containing the origin, whose every neighborhood intersects all isochrons. Consequently, using the asymptotic phase as the new phase coordinate allows the dynamics of the phase to be decoupled from the other coordinate, thereby effectively reducing the dimension of the equation in the neighborhood of the limit cycle.

Remarks

Note that the concept of isochrons extends the notion of phase of a periodic orbit to a neighborhood of that orbit. The phase difference between two points in the basin of attraction of a limit cycle can be directly computed as the time difference between the isochrons to which they belong. Computation of isochrons is usually quite difficult, requiring sometimes the coordinate transformation to phase variables, or backward integration of the system from the limit cycle, and collection of points at time interval of the period. The configuration of the isochrons in a given region also determines how fast or slow trajectories are moving in that region. The convergence (resp. divergence) of isochrons indicates a slow (resp. fast) synchronization region. A numerical resolution of isochrons could be found in [3].

Almost Limit Cycles

Definition 3. The orbit $O(x_0)$ based at x_0 as defined above is called an *almost limit cycle* if it is isolated and the function $\Phi(.) := \Phi_{x_0}(.): \quad \mathbf{R} \longrightarrow \mathbf{R}^n$ is almost periodic in the following sense (Bohr): $\forall \varepsilon > 0$, $\exists l_\varepsilon > 0$ such that every interval $(a, a + l_\varepsilon)$ in $\mathbf{R}$ of length l_ε contains a number τ_ε such that

$$||\Phi_{x_0}(t+\tau_\varepsilon) - \Phi_{x_0}(t)|| < \varepsilon. \tag{18}$$

The number τ_ε is called the ε-almost period of $\Phi_{x_0}(.)$, or ε-translation number. The following properties are derived from those of almost periodic functions which could be found for instance in [6, 13, 16]. Denote $AP(\mathbf{R}, \mathbf{R}^n)$ the Banach space of almost periodic functions from $\mathbf{R}$ to $\mathbf{R}^n$.

Properties of Almost Limit Cycles

1. The set $\mathscr{T}_\varepsilon$ of the ε-translation numbers is relatively dense in $\mathbf{R}$.
2. The orbit $O(x_0)$ is bounded, relatively compact in $\mathbf{R}^n$, and the map $\Phi_{x_0}(.)$ is uniformly continuous. Moreover there is a sequence of trigonometric polynomials $P_n(t) = \Sigma_{k=1}^{N} a_k e^{i\lambda_k t}$ converging to $\Phi(t)$ uniformly in $\mathbf{R}$.
3. The family of translates $\mathscr{F} := \{T_\tau \Phi(.) = \Phi(.+\tau); \quad \tau \in \mathbf{R}\}$ is relatively compact in the space of almost periodic functions from $\mathbf{R}$ to $\mathbf{R}^n$.
4. For a sequence of almost periodic solutions $\Phi_k(t), \quad k = 1,\ldots,n$ and $\forall \varepsilon > 0$, there exist common ε-translation numbers.
5. For $\Phi \in AP(\mathbf{R},\mathbf{R}^n)$ the *time mean or mean value* of $\Phi(t)$ exists and is defined by

$$M(\Phi) := \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} f(t)dt. \tag{19}$$

6. The Fourier exponent λ and the related Fourier-Bohr coefficient $c(\lambda)$ of $\Phi \in AP(\mathbf{R},\mathbf{R}^n)$ are defined by $c(\lambda) = M(\Phi(t)e^{-i\lambda t}) \neq 0$. The module $mod(\Phi)$ of Φ is the additive group generated by the set $\Lambda(\Phi) = \{\lambda \in \mathbf{R} | c(\Phi) \neq 0\}$. The almost periodic function is said to be quasi-periodic with frequency $\omega = (\omega_1, \cdots, \omega_m) \in \mathbf{R}^m$ if its module is contained in the additive group generated by ω.
7. Any $\Phi \in AP(\mathbf{R},\mathbf{R}^n)$ satisfies the so-called recurrence property, that is, there exists a real sequence $\{\tau_n\}$ with $\lim_{n\longrightarrow\pm\infty} \tau_n = \pm\infty$ such that $\lim_{n\longrightarrow\infty} ||T_{\tau_n}\Phi - \Phi|| = 0$.

Example of Linear Almost Center

Let $p(t) \in AP(\mathbf{R},\mathbf{C})$, and consider the differential equation

$$\dot{x}(t) = -\alpha x(t) + p(t), \quad \alpha > 0. \tag{20}$$

Define a kernel

$$K(t) = \left\{0, \quad for \quad t < 0, \quad and \quad e^{-\alpha t}, \quad for \quad t \geq 0\right\}. \tag{21}$$

Therefore, $K \in L^1(\mathbf{R},\mathbf{C})$. Thus, the convolution $x_\alpha(t) = (K * p)(t) = e^{-\alpha t}\int_{-\infty}^{t} e^{\alpha s} p(s)ds$ is also in $AP(\mathbf{R},\mathbf{C})$. Moreover this convolution is an almost periodic solution, not isolated; therefore it is not an almost limit cycle. Indeed the equation being linear, we derive a continuum of parameterized family of almost periodic solutions. Such a continuum is called a *linear almost center.* This example also appears in [13]. We represent below the solution for the almost periodic function $p(t) = \sin t + \sin\sqrt{2}t$ (Fig. 12.4).

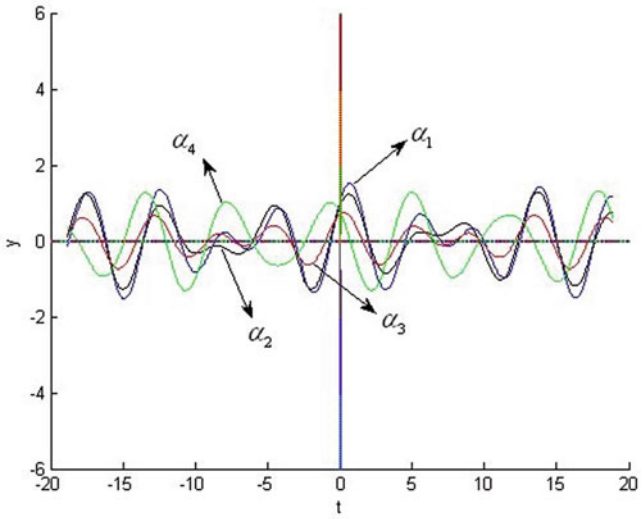

Fig. 12.4 $\dot{x}(t) = \alpha x(t) + (\sin t + \sin\sqrt{2}t), \quad \alpha = 1,2,3,4$

We introduce here an efficient technique of investigating almost periodic solutions and consequently almost limit cycles.

Hull and Method of Auxiliary Systems

Consider the nonlinear system

$$\dot{x}(t) = f(x(t),t), \tag{S}$$

where the function f is continuous on the open set $O = \mathbf{R} \times I, \quad I \subset \mathbf{R}^n$, and almost periodic in t uniformly with respect to $x \in K \subset I$, for K a compact subset of I.

Therefore, $f(\mathbf{R},K)$ is bounded. And the function $f(t,.)$ is uniformly continuous on K. A function g is said to be in the hull $H(f)$ of f if there exists a sequence $\{\tau_n; n \geq 1\}$ in $\mathbf{R}$ with $\lim_{n\to\infty} f(t+\tau_n,x) = g(t,x)$ uniformly on any set $\mathbf{R} \times K, \quad K \subset I$. That is, g is in the closure of the set $\{f(t+\tau,x), \quad \tau \in \mathbf{R}\}$.

Then consider the auxiliary system

$$\dot{x}(t) = g(t,x(t)), \quad g \in H(f). \tag{S_a}$$

Let D be a region of $\mathbf{R} \times \mathbf{R}^n$ given by $D = \mathbf{R} \times K$, K the above compact set. A solution $x(t)$ of the system (S) whose graph is in D is *separated* in D if it is either the only solution with its graph in D or there is a number $\delta > 0$ such that $|x(t) - y(t)| \geq \delta, t \in \mathbf{R}$, where $y(t)$ is another solution with its graph in D. From Amerio [2, 13] we obtain the following two theorems:

Theorem 1. *Consider the systems (S) and (S_a):*

1. *The number of separated solutions with the graph in D is finite.*
2. *If the system (S) has a solution with graph in D, then each of the auxiliary system has also a solution with its graph in D.*
3. *If $x(t), t \geq t_0$, is a solution of the system (S) such that $x(t) \in K$ for $t \geq t_0$, then the auxiliary system (S_a) has a solution defined on* $\mathbf{R}$ *whose graph is in D.*

Consequently it results:

Theorem 2. *Assuming the auxiliary system has its solutions in D separated, then all these solutions are almost periodic.*

Therefore, we are allowed to conclude that bounded separated solutions of system (*S*) are almost periodic. Corduneanu in [13] has effectively used this method to prove the existence of an asymptotically stable almost periodic solution to a Liénard-type second-order differential equation.

Next we illustrate the concept of almost limit cycle with several examples.

Almost Periodic Perturbations of the Harmonic Oscillator

Consider the forced oscillations of the harmonic oscillator given by

$$\ddot{x}(t) + \omega_0 x(t) = f(t) \tag{22a}$$

or equivalently for $\dot{x} = y$

$$\dot{x} = y, \quad \dot{y} = -\omega_0 x + f(t) \tag{22b}$$

where the external forcing term is $f(t) = k \sin \omega_1 t$ with ω_1 such that the ratio $\frac{\omega_1}{\omega_0}$ is irrational. From the Lagrange's method the general solution is computed as

$$x(t) = A \cos(\omega_0 t + \alpha) + k(\omega_0^2 - \omega_1^2)^{-1} \sin \omega_1 t. \tag{22c}$$

This solution certainly represents an oscillatory motion, but due to the fact that the ratio is irrational, the solution $x(t)$ is not periodic in t but is indeed one of the simplest examples of an almost periodic trajectory in an explicit form. The periodic perturbation has indeed destroyed the free harmonic oscillations. Setting the parameters A, α, ω_0 k, and ω_1 to numerical values provides an example of a unique asymptotically almost periodic orbit, thus isolated. It is therefore a unique stable almost limit cycle.

We further illustrate the theory of almost and pseudo-almost limit cycles with the well-known Liénard systems.

Liénard Systems

Why the Liénard Systems?

Liénard equation, which also generalizes the famous Van der Pol oscillator, is ubiquitous in the study of nonlinear systems [1, 2, 7, 12, 22, 23]. We here recall some by now classical results about Liénard-type systems.

Consider the one-parameter family of forced Liénard systems

$$\ddot{x} + f(x)\dot{x} + g(x) = \mu h(t), \tag{23}$$

or equivalently

$$\dot{x} = y - F(x), \dot{y} = -g(x) + \mu h(t), \tag{24}$$

where f, g, and h are continuous functions on $\mathbf{R}$, μ a small real parameter, and $F(x) := \int_0^x f(s)ds$.

Setting the parameter $\mu = 0$, that is, for homogeneous Liénard systems, we obtain the following well-known Liénard theorems. See more details in, for example, [7, 9, 12].

Theorem 3. *Consider the system*

$$\ddot{x}(t) + f(x)\dot{x}(t) + g(x) = 0 \tag{25}$$

where $f(x)$ and $g(x)$ are two functions generally nonlinear, assumed continuous, and differentiable from $\mathbf{R}$ *to* $\mathbf{R}$*, together with the following conditions:*

$(L_1) : xg(x) > 0$, *for* $x \neq 0$.
$(L_2) : \lim_{|x|\to\infty} |F(x)| = \infty$.
(L_3) : *There exist real numbers* α *and* β *such that* $F(x) < 0$, *for* $x < -\alpha$ *or* $0 < x < \beta$, *and* $F(x) > 0$, *for* $-\alpha < x < 0$ *or* $x > \beta$.
$(L_4) : f(x)$ *is symmetric, while* $g(x)$ *is antisymmetric.*

Then there exists a unique nontrivial periodic solution to the equation.

Theorem 4. *If the Liénard's equations satisfies the following conditions:*

1. $f(x)$ *is continuous, even and* $f(0) < 0$.
2. $g(x)$ *is locally Lipschitz, odd, and such that* $xg(x) > 0$ *for* $x \neq 0$.
3. $f(x)$ *has a unique positive zero at* $x = b$, *and it increases at* ∞ *for* $x > b$.

Then there a unique stable limit cycle.

Therefore, these theorems provide conditions under which there exist, for the unperturbed Liénard systems, respectively, a unique periodic solution and a unique limit cycle, isolated periodic orbit controlling the behavior of neighboring trajectories. We next subject some classes of Liénard systems to perturbations that, in fact, destroy the limit cycles to give birth to almost limit cycles or pseudo-almost limit cycles under suitable conditions.

Liénard Almost Limit Cycles

We study system (21) or its equivalent form (23) under the following additional assumptions:

(A_1) $f(x) > 0$, in $\mathbf{R}$, with $F(x)sgnx \to \infty$ as $|x| \to \infty$.
(A_2) $xg(x) > 0$ for $x \neq 0$, $G(x) \to \infty$ as $|x| \to \infty$.
(A_3) $|h(t)| \leq K$, and $|H(t)| \leq K$, with $H(t) = \int_0^t h(s)ds$, $t \in \mathbf{R}$, and K a positive constant.
(A_4) $g'(x) > 0$, and $g''(x)$ exists and is bounded.

It is known that, under such assumptions, for $0 < \mu \ll 1$, there exists in the xy-plane a set E bounded by a regular simple curve (C^1 except possibly at a finite number of points) such that:

1. For every solution $\gamma(t) = (x(t), y(t))$ of system (21), there is a value t_0 such that $\gamma(t_0) \in E$.
2. If, for a value t_0 of t, we have $\gamma(t_0) \in E$, then we have also $\gamma(t) \in E$, for $t \geq t_0$. That is, solutions entering the set cannot leave it for increasing time.

Moreover the set E depends only on the functions $f(x)$, $g(x)$, $h(t)$, the parameter μ, and the constant K. Equivalently, the set E may be described by the inequalities $|x(t)| \leq x_0 \quad |\dot{x}(t)| \leq v_0$, for a solution $x(t)$ of Eq. (20), and where x_0 and v_0 are constants independent of μ. See, for example, [9, 26, 29]. In other words, under the above conditions the solutions ultimately settle in a C^1-bounded set E in $\mathbf{R}^2$.

The main theorem here states:

Theorem 5. *Assume the function $h(t)$ is an almost periodic function, then under the conditions $(A_1), \ldots, (A_4)$, the almost periodically forced Liénard system has a unique stable almost limit cycle.*

This theorem was first presented by the Toni in [31]. We present here an improved and self-contained proof for the sake of clarity.

Proof. Let $\gamma(t) = (x(t), y(t))$ a solution of the system, and $\tilde{\gamma}(t) = (\tilde{x}(t), \tilde{y}(t))$ either another solution of the system or a solution of an associated system with a sufficiently small perturbation $\bar{h}(t)$ of the forcing term $h(t)$. We have then

$$\lim_{t \to \infty} |\tilde{\gamma}(t) - \gamma(t)| = 0,$$

that is,

$$\lim_{t \to \infty} |\tilde{x}(t) - x(t)| = 0 = \lim_{t \to \infty} |\tilde{y}(t) - y(t)|. \tag{26}$$

Indeed, upon the change of variables $u(t) = \tilde{x}(t) - x(t), \quad v(t) = \tilde{x}(t) - y(t)$, we obtain the system

$$\dot{u}(t) = v(t) - \varphi(t)u(t)\dot{v}(t) = -\psi(t)u(t) + \mu \Delta h(t), \tag{27}$$

where

$$\varphi(t) = \frac{F(x_2) - F(x_1)}{x_2 - x_1}, \quad \psi(t) = \frac{g(x_2) - g(x_1)}{x_2 - x_1}. \tag{28}$$

Note that the functions f, g', and g'' are bounded on the compact set E. For sufficiently small values of the parameter $\mu \ll 1$, we can construct the quadratic form

$$Q(t,u,v) = \psi(t)u^2 + v^2 - 2cuv, \tag{29}$$

with $c > 0$ chosen small enough for $Q(t,u,v)$ to be positive definite such that

$$Q(t,u,v) \geq c(u^2 + v^2), \tag{30}$$

c a positive constant, and such that

$$\dot{Q}(t,u,v) + cQ(t,u,v) < 0. \tag{31}$$

Actually we have

$$\frac{dQ}{dt}(t,u,v) = -2(\varphi\psi - \dot{\psi} - 2c\psi)u^2 - 2cv^2 + 2c\varphi uv, \tag{32}$$

yielding

$$\tilde{Q}(t,u,v) := \dot{Q}(t,u,v) + cQ(t,u,v) = -(2\varphi\psi - \dot{\psi} - 3c\psi)u^2 + 2c(\varphi - c)uv - cv^2. \tag{33}$$

The quadratic form $\tilde{Q}(t,u,v)$ can be made negative definite by taking the constant c such that

$$c < \frac{2\varphi\psi - \dot{\psi}}{3\psi}, \quad c(3\psi + (\varphi - c)^2) < 2\varphi\psi - \dot{\psi}, \tag{34}$$

which entails

$$\dot{Q}(t,u,v) < Q(t_0)e^{-c(t-t_0)}. \tag{35}$$

Therefore, $Q(t) \to 0$ as $t \to \infty$, implying that $u \to 0$ and $v \to 0$. The constant c is appropriately chosen so that, when $|\Delta h(t)| = |\tilde{h}(t) - h(t)| \to 0$, we can make $Q(t) \to 0$ for $t \to \infty$. That is, the solutions of the system of the perturbed forcing term ultimately converge to the solutions of the original system.

Next let $\gamma(t) = (x(t), y(t))$ be one of these solutions which settled in E for $t \geq t_0$. We then define the sequence of solutions $\gamma_n(t) = \gamma(t+n) = (x_n(t), y_n(t)), \quad t \geq t_0 - n$. The sequence is therefore equicontinuous and uniformly bounded. Consequently we can extract a subsequence $\gamma_{n_k}(t)$ converging uniformly to a solution $\bar{\gamma}(t) = (\bar{x}(t), \bar{y}(t))$ lying completely in E for all $t \in \mathbf{R}$. $(\lim_{n\to\infty}(t_0 + n, \infty) = (-\infty, \infty))$. And of course $\bar{\gamma}(t)$ is unique. Therefore, the forced Liénard system has a unique solution $\bar{\gamma}(t) = (x(t), y(t))$ defined on the whole real line $\mathbf{R}$ in the set E.

Now $h(t)$ almost periodic implies there is an ε-almost period τ such that

$$\|T_\tau h(t) - h(t)\| < \varepsilon$$

for any arbitrary ε. For such an ε-period consider the function $\bar{\gamma}(t+\tau)=(\bar{x}(t+\tau), \bar{y}(t+\tau))$. It is readily a solution of the following system $(\mathscr{E}_\tau)$:

$$\dot{x} = y - F(x)\dot{y} = -g(x) + \mu h(t+\tau). \tag{36}$$

Take $h(t+\tau)$ as a sufficiently small perturbation of $h(t)$ as above. Therefore, we obtain

$$\|\tilde{\gamma}(t+\tau) - \tilde{\gamma}(t)\| < \varepsilon, \tag{37}$$

which entails that the unique solution $\gamma(t)$ is also almost periodic with the same ε-almost period as the forcing term $h(t)$.

Moreover all other solutions of the system that ultimately settle in E converge to the unique almost periodic solution $\gamma(t) \in E$. Therefore, the system has a unique (isolated) almost periodic solution to which any other solution unwinds in the C^1-bounded set E. It is a stable almost limit cycle as defined above. Hence the claim. □

Remarks

Note that the proof of the theorem actually accomplishes more. That is, under the assumptions above, only one solution of the system settles in the bounded region E for all time; that solution will be of the same nature as the forcing term, almost periodic for an almost periodic forcing in this case. It has been proven also, for example, in [9, 26, 29], that it is periodic under a periodic forcing. In addition, we prove in the next section that this single solution becomes as well pseudo-almost periodic under such a forcing term. Indeed the next section discusses the concept of pseudo-almost limit cycles from the dual concepts of limit cycles and pseudo-almost periodicity.

Pseudo-Almost Limit Cycles

Introductory Concepts

Let $\mathscr{C}(\mathbf{R}\times\Omega,\mathbf{R}^n)$, $\Omega \subset \mathbf{R}^n$ open, be the Banach space of bounded continuous functions $\phi(t,x)$ endowed with the norm $||\phi|| = sup_{t\in\mathbf{R},x\in\Omega}|\phi(t,x)|$. The set $\mathscr{C}(\mathbf{R}\times\Omega,\mathbf{R}^n)$ is a subset of the more general space $\mathbf{L}_b(\mathbf{R}\times\Omega,\mathbf{R}^n)$ of all Lebesgue measurable and bounded functions.

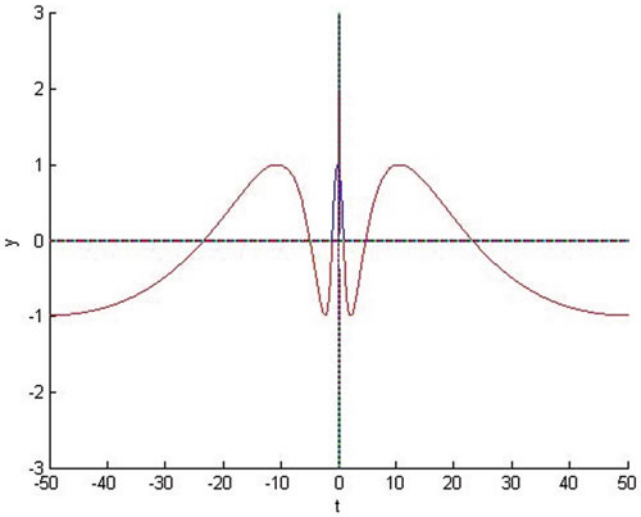

Fig. 12.5 $f(t) = 1 - t^2, \quad for \quad |t| < 1, \text{ and } \quad \sin(log(\frac{1}{t^2})), \quad for \quad |t| \geq 1$

Definition 4. A function f in $\mathscr{L}_b(\mathbf{R} \times \Omega, \mathbf{R}^n)$ is said to be *ergodic* if for every compact subset $K \subset \Omega$, the mean defined by

$$\mathscr{M}(f) := \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} f(t,x)dt \tag{38}$$

exists uniformly for $x \in K$.

We say that the function has a *vanishing mean* if $\mathscr{M}(f) = 0$. Let $\mathscr{E}(\mathbf{R} \times \Omega, \mathbf{R}^n)$ denote the space of all ergodic functions on $\mathbf{R} \times \Omega$. Note in passing that not all uniformly continuous bounded functions on $\mathbf{R}$ are ergodic. For instance the function

$$f(t) = \{1 - t^2, \quad for \quad |t| < 1, \quad and \sin(log(1/t^2)), \quad for \quad |t| \geq 1, \} \tag{39}$$

is uniformly continuous in $\mathbf{R}$, but not ergodic (Fig. 12.5). In the space $\mathscr{L}(\mathbf{R} \times \Omega, \mathbf{R}^n)$ of all Lebesgue measurable functions on $\mathbf{R} \times \Omega$, we consider next the following subspace $\mathscr{L}_0$ of all $\{\phi \in \mathscr{L} : \mathbf{R} \times \Omega \to \mathbf{R}^n$ such that $\forall x \in \Omega$, $\tilde{\phi}(.) := \phi(.,x)$ is Lebesgue measurable on $\mathbf{R}$ with $\mathscr{M}(|\tilde{\phi}|) = 0$, and $\mathscr{M}(|\phi|) = 0$.

For example, the function

$$\phi(t) = t|\sin \pi t|^{t^N}, \quad N > 6, \tag{40}$$

is unbounded, Lebesgue measurable with vanishing mean $\mathscr{M}$.

The unbounded and discontinuous function

$$\phi(t) := \{\sqrt{n}, \quad n \leq t \leq n + 1/n, \quad and \quad 0, \quad otherwise\} \tag{41}$$

is also an element of $\mathscr{L}_0$. Indeed we have $\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}|\phi(t)|dt = \lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}\frac{1}{\sqrt{k}} = 0$.

Definition 5. The orbit $O(x_0)$ based at x_0 as defined above is called a *pseudo-almost limit cycle* if it is isolated, and more importantly if the function $\Phi(.) := \Phi_{x_0}(.) : \mathbf{R} \longrightarrow \mathbf{R}^n$ defining the orbit is *pseudo-almost periodic* in the following sense, $\forall\varepsilon > 0$, $\exists\delta = \delta\varepsilon > 0$, a relatively dense subset $\mathscr{D}_\varepsilon \subset \mathbf{R}$, a subset $C_\varepsilon \subset \mathbf{R}$, such that:

1. For m the Lebesgue measure on $\mathbf{R}$,

$$\lim_{t\to\infty}\frac{m(C_\varepsilon\cap[-t,t])}{2t} = 0, \quad (C_\varepsilon \text{ is called an ergodic zero set}). \tag{42}$$

2. Let $T_\tau\Phi$ denotes the translate of Φ by τ, that is, $(T_\tau\Phi(t)) := \Phi(t+\tau)$. Then

$$||(T_\tau\Phi)(t) - \Phi(t)|| < \varepsilon, \quad \tau \in \mathscr{D}_\varepsilon, \quad t, t+\tau \in \mathbf{R} - C_\varepsilon. \tag{43}$$

3. Finally

$$|t_1 - t_2| < \delta \Longrightarrow ||\Phi(t_1) - \Phi(t_2)|| < \varepsilon, \quad t_1, t_2 \in \mathbf{R} - C_\varepsilon. \tag{44}$$

Denote $\mathscr{PA}$ the space of pseudo-almost periodic functions. These functions satisfy the following properties widely available in the relevant literature [13, 14, 35].

Some Properties of Pseudo-Almost Periodicity

We first give an equivalent definition of a pseudo-almost periodic function, in particular in the space $\mathscr{C}(\mathbf{R}\times\Omega,\mathbf{R}^n)$, with the restriction of $\mathscr{L}_0$ to the space $\mathscr{E}_0$ containing all functions $\phi \in \mathscr{C}(\mathbf{R}\times\Omega)$ such that

$$\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}|\phi(t,x)|dt = 0, \tag{45}$$

uniformly in $x \in \Omega$.

Definition 6. A function $f : \mathbf{R}\times\Omega \longrightarrow \mathbf{R}^n$ is called *pseudo-almost periodic* in t uniformly on compact subsets K of Ω if it has a unique decomposition in the form

$$f(t,x) = a(t,x) + e(t,x), \tag{46}$$

where a is almost periodic and $e \in \mathscr{E} \subset \mathscr{L}_0$. Recall that a is almost periodic if it satisfies the so-called Bohr's property. That is, $\forall\varepsilon > 0\ \exists l = l(\varepsilon)$ such that any interval $(t,t+l) \subset \mathbf{R}$ contains a number τ such that

$$||f(t+\tau,x) - f(t,x)|| < \varepsilon, \quad t \in \mathbf{R}, x \in \Omega. \tag{47}$$

The functions a and e are called, respectively, the *almost periodic component* and the *ergodic perturbation* of f. Moreover we have the following properties [14, 35]:

1. For $f \in \mathscr{PA}$, the set $f(\mathbf{R},K) := \{f(t,x)|t \in \mathbf{R}, x \in K\}$ is bounded for every bounded subset $K \subset \Omega$.
2. The function $f(t,.)$ is uniformly continuous in each bounded subset of Ω uniformly in t.
3. When the ergodic zero set $C_\varepsilon = \emptyset$, the space $\mathscr{PA}$ coincides with the space $\mathscr{AP}$ of almost periodic functions.
4. If both functions f and its derivative f' are pseudo-almost periodic, with $f = a+e$ and $f' = a' + e'$, where a and a' in $\mathscr{PA}$ and e and e' in $\mathscr{L}_0$, then the functions a and e are differentiable with $a' = a$ and $e' = e$.

Some Illustrative Examples of Pseudo-Almost Periodic Functions

We present some by now classic examples of pseudo-almost periodic functions. See also [14, 35]. We include here their graphic requirements.

Example 1

We consider the function

$$\phi_1(t) = \sin t + \sin\sqrt{2}t + \frac{e^{-|t|}}{1+t^2} \tag{48}$$

and represent, respectively,

1. The almost periodic component $a(t) = \sin t + \sin\sqrt{2}t$ and the ergodic perturbation $e(t) = \frac{e^{-|t|}}{1+t^2}$ (Fig. 12.6)
2. The pseudo-almost periodic function $\phi_1(t) = a(t) + e(t)$ (Fig. 12.7)

Example 2

We consider the function

$$\phi_2(t) = \sin t + \sin \pi t + t|\sin \pi t|^{t^N}, \quad N > 6, \tag{49}$$

with the graphic representations given below:

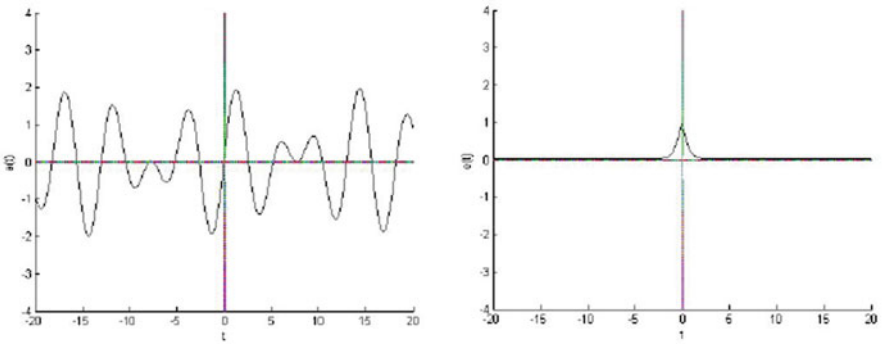

Fig. 12.6 $a(t) = \sin t + \sin\sqrt{2}t$, and $e(t) = \frac{e^{-|t|}}{1+t^2}$

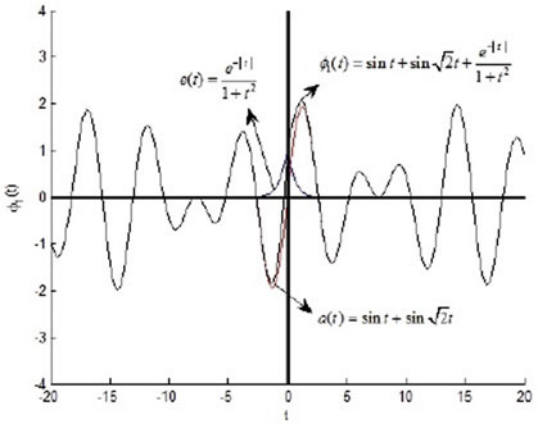

Fig. 12.7 $\phi_1(t) = a(t) + e(t) = \sin t + \sin\sqrt{2}t + \frac{e^{-|t|}}{1+t^2}$

1. The almost periodic component $a(t) = \sin t + \sin \pi t$ and the ergodic perturbation $e(t) = t|\sin \pi t|^{t^N}$ for $N = 8$ (Fig. 12.8)
2. The pseudo-almost periodic function $\phi_2(t) = a(t) + e(t)$ (Fig. 12.9)

Example 3

We finally consider the function

$$\phi_\omega(t) = I_1(t) + I_2(t), \quad \omega \neq 0, \tag{50}$$

where

$$I_1(t) = \int_{-\infty}^{\infty} h(t-s)(\sin s + \sin\sqrt{2}s)ds, \quad h \in L^1(\mathbf{R}) \tag{51}$$

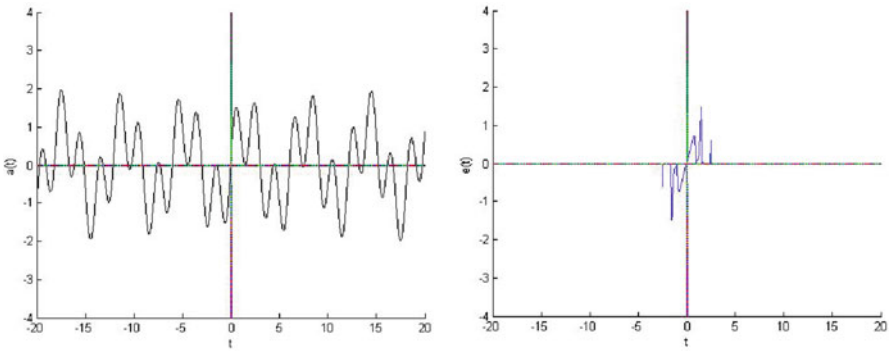

Fig. 12.8 $a(t) = \sin t + \sin \pi t$ and $e(t) = t|\sin \pi t|^{t^8}$

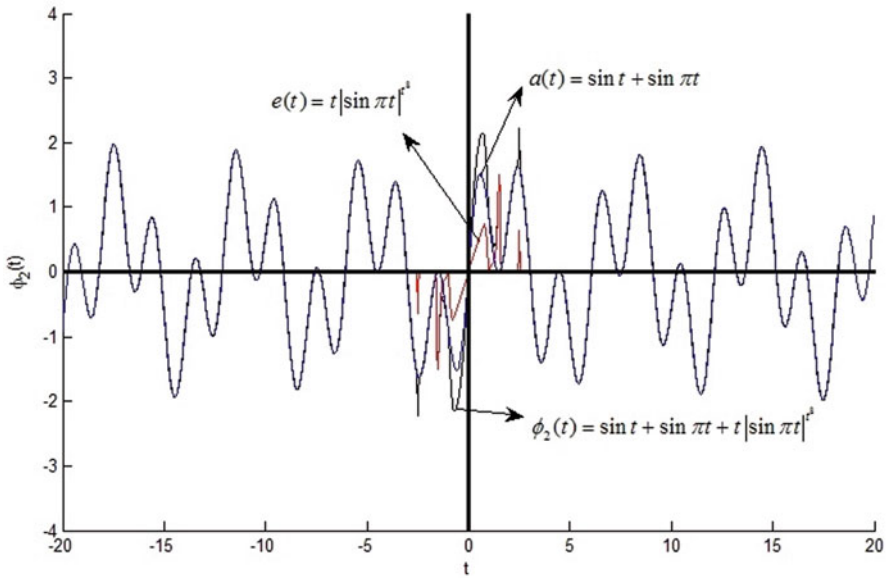

Fig. 12.9 $\phi_2(t) = a(t) + e(t) = \sin t + \sin \sqrt{2}t + \frac{e^{-|t|}}{1+t^2}$

and

$$I_2(t) = \int_{-\infty}^{\infty} \frac{h(t-s)}{s^2+\omega^2} ds. \tag{52}$$

We take $h(t) = t^2$, in $L^1(\mathbf{R})$, $\omega = 1$, and represent in figure below:

1. The almost periodic component $I_1(t)$ and the ergodic perturbation $I_2(t)$ (Fig. 12.10)
2. The pseudo-almost periodic function $\phi_1(t) = I_1(t) + I_2(t)$ (Fig. 12.11)

As in the previous section we now present some examples of existence of pseudo-almost limit cycles. First we mention the case of the *linear pseudo-almost center.*

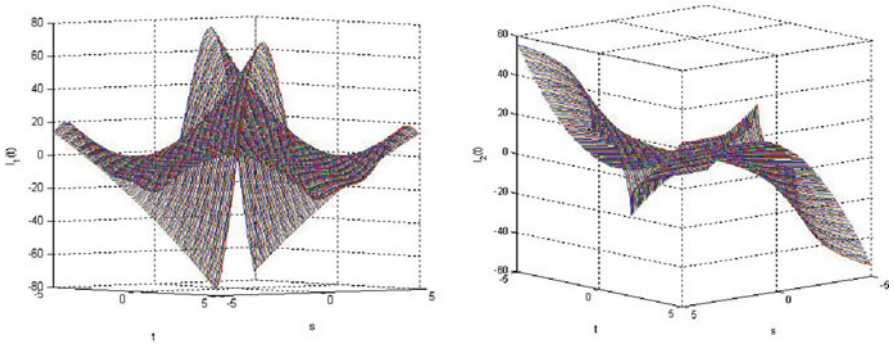

Fig. 12.10 $I_1(t) = \int_{-\infty}^{\infty} h(t-s)(\sin s + \sin\sqrt{2s})ds$ and $I_2(t) = \int_{-\infty}^{\infty} \frac{h(t-s)}{s^2+\omega^2}ds$

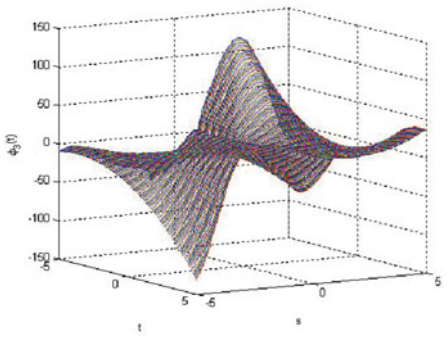

Fig. 12.11 $\phi_\omega(t) = I_1(t) + I_2(t)$

Linear Pseudo-Almost Center: An Example

Let $p(t) \in \mathscr{PA}(\mathbf{R}, \mathbf{C})$, that is, a complex-value pseudo-almost periodic function defined on the real numbers, and consider the differential equation (see also [13])

$$\dot{x}(t) = -\alpha x(t) + p(t), \quad \alpha > 0. \tag{53}$$

Define a kernel

$$K(t) = \{0, \quad t < 0, \quad \textit{and} \quad e^{-\alpha t}, \quad t \geq 0\}. \tag{54}$$

Therefore, $K \in L^1(\mathbf{R}, \mathbf{C})$. Thus, the convolution $x_\alpha = (K * p)(t) = e^{-\alpha t}\int_{-\infty}^{t} e^{\alpha s} p(s)ds$ is also in $\mathscr{PA}(\mathbf{R}, \mathbf{C})$, for every α. Indeed the space $\mathscr{PA}$ is convolution invariant by L^1. The equation being linear, it results in the existence of a continuum of parameterized pseudo-almost periodic solutions which we called *linear pseudo-almost center*. Therefore, these solutions are not isolated and are not pseudo-almost limit cycles.

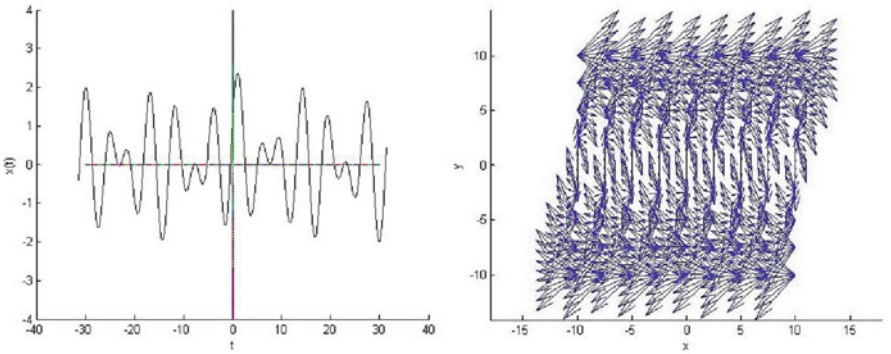

Fig. 12.12 $x(t) = \sin t + \sin\sqrt{2}t + \frac{1}{t^2+1}$ and $\dot{x} = y, \quad \dot{y} = -x + (-\sin\sqrt{2}t + \frac{t^2(t^2+4)}{(t^2+1)^3})$

Pseudo-Almost Periodic Perturbations of the Harmonic Oscillator

Consider the forced oscillations of the harmonic oscillator given by

$$\ddot{x}(t) + x(t) = f(t) \tag{55}$$

where the forcing term is

$$f(t) = -\sin\sqrt{2}t + \frac{t^2(t^2+4)}{(t^2+1)^3} \tag{56}$$

equivalently, for $\dot{x} = y$

$$\dot{x} = y, \quad \dot{y} = -x + f(t). \tag{57}$$

Clearly the function explicitly given by

$$x(t) = \sin t + \sin\sqrt{2}t + \frac{1}{t^2+1} \tag{58}$$

is the unique solution of the equation, and it is one of the classic examples of pseudo-almost periodic function that is not periodic. (See also [11].) Therefore, we obtain an explicit and simple example of pseudo-almost limit cycle. The figure below gives the phase portrait of (57) and the graph of the pseudo-almost periodic function in (58) (Fig. 12.12).

Liénard Pseudo-Almost Limit Cycles

We now reconsider the above Liénard systems (20) and (21) under a forcing term that is now assumed to be a pseudo-almost function. As stated above in the case of

an almost periodic forcing, the assumptions entail the existence of a unique solution that settles in the bounded region E for all time. Moreover the proof of Theorem 5 leads to the following:

Proposition 7. *Assume the conditions $A_1,\ldots,A_4$. Let $\gamma(t) = (x(t), y(t))$ be a solution of the system, and $\tilde{\gamma}(t) = (\tilde{x}(t), \tilde{y}(t))$ either another solution of the system or a solution of an associated system with a sufficiently small perturbation $\bar{h}(t)$ of the forcing term $h(t)$. Then we have*

$$\lim_{t\to\infty} |\tilde{\gamma}(t) - \gamma(t)| = 0,$$

that is,

$$\lim_{t\to\infty} |\tilde{x}(t) - x(t)| = 0 = \lim_{t\to\infty} |\tilde{y}(t) - y(t)|. \tag{59}$$

Proof. It is a direct consequence of the lines of proof for theorem (cite here). That is, the solutions of the system associated to the perturbed forcing term ultimately converge to the solutions of the original system. □

We now state and prove the main result of this section.

Theorem 6. *Assume the forcing term $h(t)$ is a pseudo-almost periodic function. Then under the conditions $(A_1),\ldots,(A_4)$, the pseudo-almost periodically forced Liénard system has a unique stable pseudo-almost limit cycle.*

Proof. The proof is based on the previous proposition, including the existence of a unique solution enclosed in E for all time. First assuming the forcing term $h(t)$ is pseudo-almost periodic entails from the definition above that, for any arbitrary ε, there exists $\delta = \delta(\varepsilon)$, an ε-pseudo-almost period $\tau \in \mathscr{D}_\varepsilon$, a relatively dense set in $\mathbf{R}$ such that

$$\|h(t+\tau) - h(t)\| < \varepsilon, \quad t, t+\tau \in \mathbf{R} - C_\varepsilon \tag{60}$$

and

$$|t_1 - t_2| < \delta \Longrightarrow ||h(t_1) - h(t_2)|| < \varepsilon, \quad t_1, t_2 \in \mathbf{R} - C_\varepsilon, \tag{61}$$

where C_ε is the ergodic zero set defined above. For such an ε-pseudo-almost period, consider the unique solution $\bar{\gamma}(t)$ given in the previous lemma that settles in E for all time $t \in (-\infty, \infty)$, and the associated function $\bar{\gamma}(t+\tau) = (\bar{x}(t+\tau), \bar{y}(t+\tau))$. This function is readily a solution of the following system $(\mathscr{E}_\tau)$

$$\dot{x} = y - F(x, \quad \dot{y} = -g(x) + \mu h(t+\tau). \tag{62}$$

Take $h(t+\tau)$ as a sufficiently small perturbation of $h(t)$ as above. Therefore, according to the previous propositions, the solutions $\bar{\gamma}(t)$ and $\bar{\gamma}(t+\tau)$ converge. Thus, we obtain

$$\|\bar{\gamma}(t+\tau) - \bar{\gamma}(t)\| < \varepsilon, \quad t, t+\tau \in \mathbf{R} - C_\varepsilon. \tag{63}$$

Moreover we also have, for $t_1, t_2 \in \mathbf{R} - C_\varepsilon$,

$$|\bar{\gamma}(t_2) - \bar{\gamma}(t_1)| \leq |t_2 - t_1| sup_E |\dot{\bar{\gamma}}|,$$

which ensures the existence of δ such that

$$|t_1 - t_2| < \delta \Longrightarrow ||\bar{\gamma}(t_1) - \bar{\gamma}(t_2)|| < \varepsilon, \quad t_1, t_2 \in \mathbf{R} - C_\varepsilon. \tag{64}$$

Therefore, we conclude that the unique solution $\bar{\gamma}(t)$ is pseudo-almost periodic.

Moreover, from the previous proposition, all other solutions of the system that ultimately settle in E converge to this unique pseudo-almost periodic solution $\bar{\gamma}(t) \in E$. Therefore, the system has a unique (isolated) almost periodic solution to which any other solution unwinds in the C^1-bounded set E. It is a stable almost limit cycle as defined above. Hence the claim. □

Remarks

For a Liénard system under the assumptions stated above, a forcing term, respectively, periodic, almost periodic, and pseudo-almost periodic leads to the emergence, respectively, of a unique stable limit cycle, stable almost limit cycles, and pseudo-almost limit cycles. Such characteristics, if need be, add to the "mathematical beauty and richness" of the Liénard systems. We derive the following natural question as an open problem.

Open Problem

Re-parameterize the Liénard system if necessary and determine conditions under which the phase space could be partitioned in regions of limit cycles, almost limit cycles, and pseudo-almost limit cycles.

Almost and Pseudo-Almost Periodic Waves

The importance of Liénard systems among nonlinear systems also comes from the fact that several systems can be transformed into Liénard systems and solved [1, 17, 19, 20]. We present next some partial differential equations solvable first by reducing them to some Liénard-type equations, then by applying the previous theorems.

Illustrative Example 1

Consider systems described by the time-perturbed nonlinear hyperbolic equation

$$u_{tt} = u_{xx} + f_0(u)u_x + g_0(u) + p(t) \quad (\mathscr{H}).$$

The search of special solutions of the form

$$u(x,t) = y(x+ct), \quad c \in \mathbf{R} \tag{65}$$

defining the wave with speed $|c|$ yields the Liénard-type equation

$$(1-c^2)\ddot{y} + f_0(y)\dot{y} + g_0(y) = -p(t). \tag{66}$$

Define $f(y) = \frac{f_0(y)}{1-c^2}$, $g(y) = \frac{g_0(y)}{1-c^2)}$, and $h(t) = \frac{-p(t)}{1-c^2}$. The functions f_0 and g_0 are continuously differentiable chosen together with the speed $|c|$ of the waves $u(t,x)$ such that the functions f, g, and h satisfy the assumptions $(A_1),\ldots,(A_4)$. Obviously $p(t)$ almost periodic or pseudo-almost periodic implies $h(t)$, respectively, almost or pseudo-almost periodic. Therefore, we conclude under these assumptions:

Theorem 7. *For an almost periodic perturbation $p(t)$, the nonlinear hyperbolic equation $(\mathscr{H})$ has an almost periodic solitary wave $u(x,t) = y(x+ct)$, where $y(x)$ is an almost limit cycle of the perturbed Liénard-type equation (25).*

Proof. The proof is immediate and is adapted from Theorems 5 and 6. □

In the same lines we prove:

Theorem 8. *For a pseudo-almost periodic perturbation $p(t)$, the nonlinear hyperbolic equation $(\mathscr{H})$ has a pseudo-almost periodic solitary wave $u(x,t) = y(x+ct)$, where $y(x)$ is a pseudo-almost limit cycle of the perturbed Liénard-type equation (25).*

We next consider a parabolic partial differential equation describing a reaction-diffusion equation.

Reaction-Diffusion Model

Consider now the time-perturbed parabolic equation describing a reaction-diffusion model

$$u_t = u_{xx} + f_0(u)u_x + g_0(u) + p(t) \quad (\mathscr{RD}).$$

Looking again for special solutions of the form (24) leads to the Liénard-type equation

$$\ddot{y} + [f_0(y) - c]\dot{y} + g_0(y) = 0. \tag{67}$$

As in the previous case we set $f(y) = f_0(y) - c$, $g(y) = g_0(y)$, and $h(t) = -p(t)$. The functions f_0 and g_0 are continuously differentiable are determined together with the speed $|c|$ of the waves $u(t,x)$ such that the functions f, g, and h satisfy the assumptions $(A_1), \ldots, (A_4)$. Obviously $p(t)$ almost periodic or pseudo-almost periodic implies $h(t)$, respectively, almost or pseudo-almost periodic. We therefore obtain the equivalent theorems of existence of almost and pseudo-almost solitary waves to the reaction-diffusion equation as functions of the corresponding Liénard almost and pseudo-almost limit cycles, as in Theorems 7 and 8.

Outlook and Open Problems

Arnold in [5] states

> Une trajectoire fermée nondégénérée ne disparait pas par une petite déformation du système, mais se déforme légèrement. Donc le système des trajectoires est structurellement stable dans le voisinage de la trajectoire fermée générique.

That is, periodic orbits do not just disappear under small perturbation, but they may be slightly deformed, due to the fact that the system of trajectories is structurally stable in the neighborhood of a periodic orbit.

Many forced systems such as the Liénard ones are actually small perturbations of systems having periodic orbits (limit cycles) in their unperturbed form, and many results such as the above ones are about the existence and uniqueness of almost periodic solution with no mention of the fate of the periodic orbit(s) existing before perturbation. The appearing of almost or pseudo-almost periodic solutions could only results from the bifurcation of the generic orbits for a parameterized system. Therefore, one must investigate the relation between the "new" almost periodic solutions appearing upon perturbation and the periodic orbits of the unperturbed system. For instance, to uncover the existence of the so-called limit periodic almost limit cycles, where a sequence of periodic orbits such as in the linear isochrone $\dot{x} = -y \quad \dot{y} = x$ accumulate on the new almost/pseudo-almost limit cycle.

The following open problems should be of interest to the community of pure and applied mathematicians including graduate students. Note first that a periodic function is also almost periodic and pseudo-almost periodic, as an almost periodic function is also pseudo-almost periodic with a zero ergodic perturbation. Consequently a limit cycle is also an almost or a pseudo-almost limit cycle, but not inversely. To make the distinction, we will call *strictly almost limit cycles* and *strictly pseudo-almost limit cycles*, respectively, those almost or pseudo-almost limit cycles that are not limit cycles.

Open Problem 1

Complete a full study of the bifurcation of strictly almost/pseudo-almost limit cycles in the above forced Liénard systems when the parameter value μ varies in order to investigate conditions on the functions f and g for which the strictly almost/pseudo-almost limit cycles that exist for $\mu \ll 1$ could persist for $\mu = 1$, and eventually accumulate when $\mu \to 1$.

Open Problem 2: Linear Almost and Pseudo-almost Center

Determine the conditions of existence for a continuum of parameterized families of strictly almost and pseudo-almost trajectories possibly surrounding a critical point. Such continuum defines, respectively, the *linear almost center* and the *linear pseudo-almost center*.

Open Problem 3: Multiple Almost and Pseudo-almost Limit Cycles

Find parameterized systems and determine conditions under which exist in the same phase space multiple strictly almost or pseudo-almost limit cycles, similar to several examples in the case of the usual normal limit cycles.

Open Problem 4: Coexistence of Limit Cycles and Almost and/or Pseudo-Almost Limit Cycles

Find parameterized systems and determine conditions under which coexist in the phase space limit cycles and strictly almost or pseudo-almost limit cycles.

Open Problem 5: Isochronous Almost and Pseudo-almost Limit Cycles

Let γ be a strictly almost or pseudo-almost limit cycle of a flow ϕ on $\mathbf{R}^n$ as in Section "Overview of limit cycles". A point x_1 in $\mathbf{R}^n$ has *asymptotic phase* with respect to γ if there is a point $x_0 \in \gamma$ such that $\lim_{t \longrightarrow \pm\infty} |\phi_t(x_1) - \phi_t(x_0)| = 0$. We say that x_1 is *in phase* with x_0.

It is well known that a hyperbolic limit cycle has some neighborhood where every point has asymptotic phase with respect to the limit cycle, due to the existence of

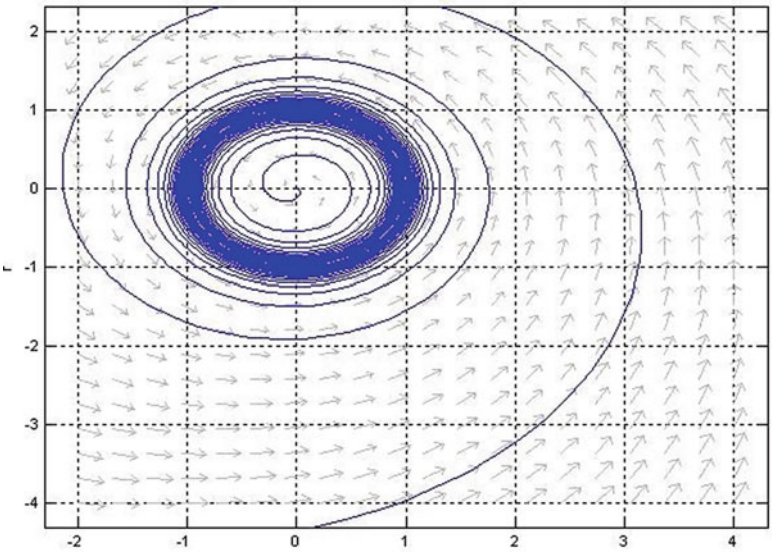

Fig. 12.13 $\dot{r} = -(r-1)^2, \dot{\theta} = 2\pi + (r-1)$

invariant foliation. Similar question needs to be addressed as well in case of strictly almost or pseudo-almost limit cycles.

Definition 7. A strictly almost or pseudo-almost limit cycle is said to be *isochronous* if there is a neighborhood of γ in which every point is in phase with a point on γ.

In the case of limit cycles, we have, for instance, the following examples.

1. The system
$$\dot{r} = -(r-1)^2, \quad \dot{\theta} = 2\pi + (r-1) \tag{68}$$
in polar coordinate (r, θ) has a nonhyperbolic limit cycle γ at $r = 1$, attracting for $r > 1$ and repelling for $r < 1$, but nonisochronous. Indeed no point $(r_0, \theta_0), \quad r_0 > 0$ has asymptotic phase with γ. For more details, see [10]. The nonisochronous limit cycle is represented below (Fig. 12.13).
2. The system
$$\dot{r} = -\frac{1}{3}(r-1)^4 e^{|r-1|^{-3}}, \quad \dot{\theta} = 2\pi \tag{69}$$
has a nonhyperbolic limit cycle at the unit cycle with period 1, attracting for $r>1$. The asymptotic phase of any point (r_0, θ_0) in its neighborhood is $(1, \theta_0)$. The limit cycle is therefore isochronous. For more details, see [10]. The isochronous limit cycle is represented below (Fig. 12.14).

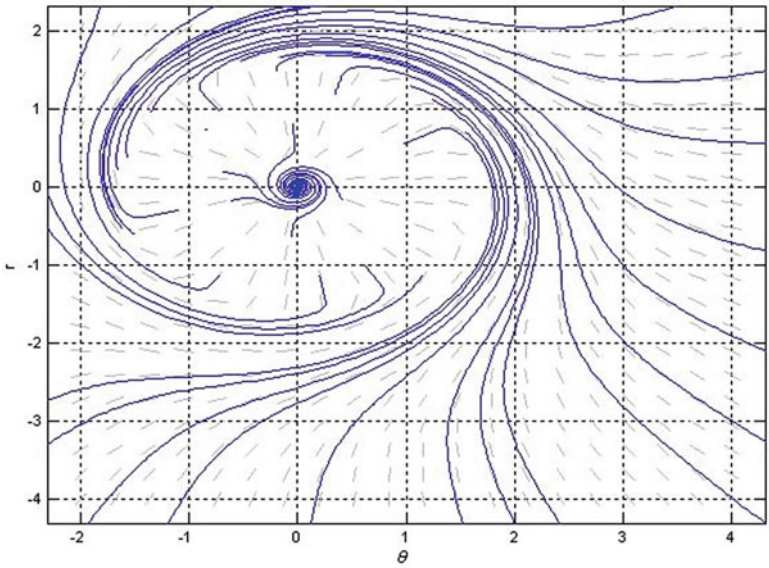

Fig. 12.14 $\dot{r} = -\frac{1}{3}(r-1)^4 e^{|r-1|^{-3}}, \dot{\theta} = 2\pi$

It would be interesting to:

1. Perturb systems (68) and (69), in particular in the angle variable, and study the conditions of appearance of strictly almost and/or pseudo-almost limit cycles
2. Investigate the conditions of existence of isochronous strictly almost or pseudo-almost limit cycles, in particular for the forced Liénard systems

Open Problem 6: Almost and Pseudo-almost Isochrons

Following the previous open problem, we further define:

Definition 8. Given $x_0 \in \gamma$ where γ is a strictly almost or pseudo-almost limit cycle, an *almost or pseudo-almost isochron* $I(x_0)$ based at x_0 is the set of all point $x \in \mathbf{R}^n$ in phase with x_0.

As in the case of limit cycles we conjecture the existence of *almost or pseudo-almost isochrons* and that they will foliate the neighborhood of *almost or pseudo-almost limit cycles.* Their determination is definitely an interesting but difficult question of research. One line of attack might be similar to Guckenheimer and Winfree investigation of isochrons of limit cycles [3, 19, 33, 34].

Open Problem 7: Transition to Chaos

It would be interesting to investigate the possibility for a strictly almost or pseudo-almost behavior to transition to a chaotic behavior. Such study could initiate with coupling of forced Liénard oscillators, as in the following example. Consider two almost or pseudo-almost self-sustained oscillators given by forced Liénard systems under the coupling described as follows:

$$\ddot{x}(t) + f_\alpha(x)\dot{x} + g_\beta(x) = h(t) \tag{70}$$

$$\ddot{y}(t) + f_\alpha(y)\dot{y} + g_\beta(y) = h(t) - K(y-x)H(t-t_0) \tag{71}$$

where $h(t)$ is almost or pseudo-almost periodic, K is the feedback coupling coefficient, t_0 the onset time of synchronization process, and $H(z)$ the Heaviside function defined as

$$H(z) = \{0, \quad for \quad z < 0, \quad 1, \quad for \quad z \geq 0\}. \tag{72}$$

Introduce the new variable $z(t) = y(t) - x(t)$ to measure the closeness between solutions of (70) and (71) and then analyze the resulting second-order equation. The question is to find the appropriate coupling coefficients and conditions on f_α and g_β which enable (70) to adjust its oscillations and to synchronize with (71).

To fix ideas one may start with f_α and g_β such as the systems are two driven chaotic Van der Pol-Duffing systems, paradigm for relaxation oscillations and chaotic behavior in small ranges of control parameter, and also systems well known to be generalized by the Liénard systems. The relevant references include [3, 10, 24, 27, 30].

Acknowledgements The authors express appreciation for the referees' time and efforts and for their valuable suggestions and corrections that help improve the quality of this chapter.

References

1. Albarakati, W.A., Lloyd, N.G., Pearson, J.M.: Transformation to Liénard form. EJDE **2000**(76), 1–11 (2000)
2. Amerio, L.: Soluzioni quasi-periodiche, o limitate, di sistemi differenziali non lineari quasi-periodici, o limitati Annali di Matematica Pura ed Applicata **39**, 97–119 (1955)
3. Amor, H.B., Glade, N., Lobos, C., Demongeot, J.: The isochronal fibration: characterization and implication in biology. Acta Biotheor. **58**(2), 121–142 (2010)
4. Andronov, A.A. et al.: Theory of Oscillators Dover, New York (1989)
5. Arnold, V.: Chapites supplémentaires de la Théorie des équations différentielles ordinaires. Editions MIR, Moscou (1978)
6. Bohr, H.A.: Almost Periodic Functions. Chelsea, New York (1951)
7. Brauer, S.G., Nohel, J.A.: The Qualitative Theory of Ordinary Differential Equations. W.A. Benjamin, New York (1968)
8. Byrnes, C.: Topological methods for nonlinear oscillations. Not. AMS **57**(9), 1080–1091 (2010)

9. Cartwright, M.L., Littlewood, J.E.: On non-linear differential equations of the second order II. Ann.Math. **48**(2), 472–494 (1947)
10. Chicone, C., Liu, W.: Asymptotic phase revisited. J. Differ. Equat. **204**, 227–246 (2004)
11. Christopher, C., Li, C.: Limit Cycles of Differential Equations. Birkhauser Verlag, Basel-Boston-Berlin (2007)
12. Coddington, E.A., Levinson, N.: Theory of Ordinary Differential Equations. Mc-Graw-Hill, New York (1953)
13. Corduneanu, C.: Almost Periodic Oscillations and Waves. Springer, New York (2009)
14. Diagana, T.: Pseudo Almost Periodic Functions in Banach Spaces. Nova Publishers, New York (2007)
15. Dumortier, F.: Qualitative Theory of Planar Differential Systems. Springer, New York (2006)
16. Ecalle, J. et al.: Non-accumulation des cycles limites I-II. C. R. Acad. Sci. Paris **I**(304), 375–431 (1987)
17. Fink, A.M.: Convergence and almost periodicity of solutions of forced Liénard equations. SIAM J. Appl. Math. **26**(1), 6–34 (1974)
18. Guckenheimer, J., Holmes, P.: Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields. Springer, New York (1983)
19. Guckenheimer, J.: Isochrons and phaseless sets. J. Math. Biol. **1**, 259–273 (1975)
20. Hilbert, D.: Mathematical problems. Bull. Amer. Math. Soc. **8**, 437–479 (1902)
21. Jordan, D.W., Smith, P.: Nonlinear Ordinary Differential Equations, 4th edn. Oxford University Press, Oxford (2007)
22. Lloyd, N.G.: Liénard systems with several limit cycles. Math. Proc. Camb. Phil. **102**(03), 565–572 (1987)
23. Loud, W.S.: Boundedness and convergence of solutions of $x'' + cx' + g(x) = e(t)$. Duke Math. J. **24**, 63–72 (1957)
24. Leung, H.K.: Synchronization dynamics of coupled Van der Pol systems. Phys. A **321**, 248–255 (2003)
25. N'Guérékata, G.M.: Almost Automorphic Functions and Almost Periodic Functions in Abstract Spaces. Kluwer Academic/Plenum Publishers, New York-London-Moscow (2001)
26. Opial, Z.: Sur les solutions périodiques et presque-périodiques de l'équations differentielle $x'' + kf(x)x' + g(x) = kp(t)$. Annales Polonici Mathematici **VII**, 309–319 (1960)
27. Pecora, L.M., Caroll, T.L.: Synchronization in chaotic systems. Phys. Rev. Lett. **1990**, 64–821 (1990)
28. Poincaré, H.: Mémoire sur les courbes définies par une équation differentielle. J. Math. Pure Appl. **7**, 375–422 (1881)
29. Reuters, G.E.H.: On certain non-linear differential equaions with almost periodic solutions. J. Lond. Math. Soc. **26**, 215–221 (1951)
30. Szemplinska-Stupnicka, W., Rudonski, J.: The coexistence of periodic, almost periodic, and chaotic attractors in the Van der Pol-Duffing oscillator. J. Sound Vib. **199**, 165 (1997)
31. Toni, B.: Almost and pseudo-almost limit cycles for some forced Liénard systems. Nonlinear Anal. **71**, 4718–4724 (2009)
32. Toni, B.: Upper bounds of limit cycles from isochronous period annulus via birational linearization. Discrete Contin. Syst. **2005**(Supp), 846–853 (2005)
33. Winfree, A.T.: The Geometry of Biological Time. Springer, New York (2001)
34. Winfree, A.T.: Patterns of phase compromise in biological cycles. J. Math. Biol. **1**, 73–95 (1974)
35. Zhang, C.: Almost periodic type functions and ergodicity. Science Press/Kluwer Academic Publishers (2003)

Chapter 13
On Almost Periodic Stochastic Difference Equations

Paul H. Bezandry

Introduction

The study of almost periodicity which generalizes the notion of periodicity is an area of interest in its own right and has sundry applications in fields like Physics. For a study of almost periodic sequences, we refer the reader to Bezandry and Diagana [2], Bezandry et al. [3], Corduneanu [4], Diagana et al. [5], Han and Hong [7], Hong and Nunez [8], and references therein. Almost periodicity is also of importance in the study of stochastic processes.

In Bezandry et al. [3], the notion of almost periodicity in mean was introduced and used to study the existence and uniqueness of almost periodic solutions to the stochastic Beverton–Holt equation.

In this paper, we study the existence and uniqueness of almost periodic solutions to a semi-linear system of stochastic difference equations of the form:

$$X(\omega,n+1) = A(\omega,n)X(\omega,n) + f(n,X(\omega,n)),\ n \in \mathbb{Z}_+,\ \omega \in \Omega, \tag{1}$$

on $\mathbf{R}^k$, where $A(n)$ is an invertible almost periodic $k \times k$ random matrix function defined on $\mathbb{Z}_+$ and $f : \mathbb{Z}_+ \times \mathbf{R}^k \to \mathbf{R}^k$ is a function to be specified later. We assume that the $A(n)$'s are independent and independent of $X(0)$. This assumption together with Eq. (1) imply that the sequence $\{A(n)\}_{n\in\mathbb{Z}_+}$ is independent of the sequences $\{X(n)\}_{n\in\mathbb{Z}_+}$.

The paper is organized as follows. In section "Preliminaries", we recall a basic theory of almost periodic random sequences on $\mathbb{Z}_+$. In section "Almost Periodic

Paul H. Bezandry (✉)
Department of Mathematics, Howard University, Washington, DC 20059, USA
e-mail: pbezandry@howard.edu

B. Toni (ed.), *Advances in Interdisciplinary Mathematical Research*, Springer Proceedings in Mathematics & Statistics 37, DOI 10.1007/978-1-4614-6345-0_13,

Stochastic Difference Equations", we apply the techniques developed in section "Preliminaries" to find some sufficient conditions for the existence and uniqueness of the almost periodic solution to some semi-linear system of stochastic difference equations. In section "Application", we study the stochastic Beverton–Holt difference equation to illustrate our main result.

Preliminaries

In this section we establish a basic theory for almost periodic random sequences. To facilitate our task, we first introduce the notations needed in the sequel.

Let $(\mathbb{B}, \|\cdot\|)$ be a Banach space, and let $(\Omega, \mathscr{F}, \mathbf{P})$ be a complete probability space. Throughout the rest of the paper, $\mathbb{Z}_+$ denotes the set of all nonnegative integers. Define $L^1(\Omega;\mathbb{B})$ to be the space of all $\mathbb{B}$-valued random variables V such that

$$\mathbf{E}\|V\| := \Big(\int_\Omega \|V(\omega)\| d\mathbf{P}(\omega)\Big) < \infty. \tag{2}$$

It is then routine to check that $L^1(\Omega;\mathbb{B})$ is a Banach space when it is equipped with its natural norm $\|\cdot\|_1$ defined by, $\|V\|_1 := \mathbf{E}\|V\|$ for each $V \in L^1(\Omega,\mathbb{B})$.

Let $X = \{X_n\}_{n\in\mathbb{Z}_+}$ be a sequence of $\mathbb{B}$-valued random variables satisfying $\mathbf{E}\|X_n\| < \infty$ for each $n \in \mathbb{Z}_+$. Thus, interchangeably we can, and do, speak of such a sequence as a function, which goes from $\mathbb{Z}_+$ into $L^1(\Omega;\mathbb{B})$.

This setting requires the following preliminary definitions.

Definition 1. An $L^1(\Omega;\mathbb{B})$-valued random sequence $X = \{X(n)\}_{n\in\mathbb{Z}_+}$ is said to be Bohr almost periodic in mean if for each $\varepsilon > 0$ there exists $N_0(\varepsilon) > 0$ such that among any N_0 consecutive integers, there exists at least an integer $p > 0$ for which

$$\mathbf{E}\|X(n+p) - X(n)\| < \varepsilon, \forall n \in \mathbb{Z}_+.$$

An integer $p > 0$ with the above-mentioned property is called an ε-almost period for X. The collection of all $\mathbb{B}$-valued random sequences $X = \{X(n)\}_{n\in\mathbb{Z}_+}$ which are Bohr almost periodic in mean is then denoted by $AP(\mathbb{Z}_+; L^1(\Omega;\mathbb{B}))$.

Similarly, one defines the Bochner almost periodicity in mean as follows:

Definition 2. An $L^1(\Omega;\mathbb{B})$-valued random sequence $X = \{X(n)\}_{n\in\mathbb{Z}_+}$ is called mean Bochner almost periodic if for every sequence $\{m_k\}_{k\in\mathbb{Z}_+} \subset \mathbb{Z}_+$, there exists a subsequence $\{m'_k\}_{k\in\mathbb{Z}_+}$ such that $\{X(n+m'_k))\}_{k\in\mathbb{Z}_+}$ converges (in the mean) uniformly in $n \in \mathbb{Z}_+$.

Following along the same arguments as in the proof of [5, Theorem 2.4, p. 241], one can show that those two notions of almost periodicity coincide.

Theorem 1. *An $L^1(\Omega;\mathbb{B})$-valued random sequence $X = \{X(n)\}_{n\in\mathbb{Z}_+}$ is Bochner almost periodic in mean if and only if it is Bohr almost periodic in mean.*

An important and straightforward consequence of Theorem 1 is the next corollary, which pays a key role in the proof of Theorem 9.

Corollary 1. *If* $X_1 = \{X^1(n)\}_{n\in\mathbb{Z}_+}, X_2 = \{X^1(n)\}_{n\in\mathbb{Z}_+}, \ldots,$ *and* $X_N = \{X^N(n)\}_{n\in\mathbb{Z}_+}$ *are N random sequences, which belong to* $AP(\mathbb{Z}_+;L^1(\Omega,\mathbb{B}))$*, then for each* $\varepsilon > 0$ *there exists* $N_0(\varepsilon) > 0$ *such that among any* $N_0(\varepsilon)$ *consecutive integers, there exists an integer* $p > 0$ *for which*

$$\mathbf{E}\|X^j(n+p) - X(n)\| < \varepsilon$$

for $n \in \mathbb{Z}_+$ *and for* $j = 1,2,\ldots,N$.

Definition 3. A $\mathbb{B}$-valued random sequence $X = \{X(n)\}_{n\in\mathbb{Z}_+}$ is said to be almost periodic in probability if for each $\varepsilon > 0$ and $\eta > 0$, there exists $N_0(\varepsilon,\eta) > 0$ such that among any N_0 consecutive integers, there exists at least an integer $p > 0$ for which

$$\mathbf{P}\big\{\omega \in \Omega : \|X(\omega,n+p) - X(\omega,n)\| > \varepsilon\big\} < \eta, \forall\, n \in \mathbb{Z}_+.$$

Theorem 2. *If X is almost periodic in mean, then it is almost periodic in probability and there also exists a constant* $M > 0$ *such that* $\mathbf{E}\|X(n)\| \le M$ *for all* $n \in \mathbb{Z}_+$. *Conversely, if X is almost periodic in probability and the sequence* $\big\{\|X(n)\|, n \in \mathbb{Z}_+\big\}$ *is uniformly integrable, then X is almost periodic in mean.*

Let $\mathbf{k} = \{k(i)\}_{i\in\mathbb{Z}_+}$, and denote $T_{\mathbf{k}}X(\omega,n) := \lim_{i\to\infty} X(\omega,n+k(i))$ for each $\omega \in \Omega$ and each $n \in \mathbb{Z}_+$ if it exists.

Definition 4. A $\mathbb{B}$-valued random sequence $X = \{X(n)\}_{n\in\mathbb{Z}_+}$ satisfies Bochner's almost sure uniform double sequence criterion if, for every pair of sequences (k'_i) and (l'_i), there exists a measurable subset $\Omega_1 \subset \Omega$ with $\mathbf{P}(\Omega_1) = 1$ and there exist subsequences $\mathbf{k} = (k_i) \subset (k'_i)$ and $\mathbf{l} = (l_i) \subset (l'_i)$, respectively, with the same indexes (independent of ω) such that, for every $n \in \mathbb{Z}_+$,

$$T_{\mathbf{k}}T_{\mathbf{l}}X(\omega,n) = T_{\mathbf{k}+\mathbf{l}}X(\omega,n), \forall\, \omega \in \Omega_1\,.$$

(In this case, Ω_1 depends on the pair of sequences (k'_i) and (l'_i).)

Theorem 3. *The following properties of X are equivalent:*

(i) *X satisfies Bochner's almost sure uniform double sequence criterion.*
(ii) *X is almost periodic in probability.*

The proof of the theorem can be seen in Bedouhene et al. [1], for instance.

Theorem 4. *If X satisfies Bochner's almost sure uniform double sequence criterion and the sequence* $\big\{\|X(n)\|, n \in \mathbb{Z}_+\big\}$ *is uniformly integrable, then X is almost periodic in mean.*

Let $(\mathbb{B}_1, \|\cdot\|_1)$ and $(\mathbb{B}_2, \|\cdot\|_2)$ be Banach spaces, and let $L^1(\Omega;\mathbb{B}_1)$ and $L^1(\Omega;\mathbb{B}_2)$ be their corresponding L^1-spaces, respectively.

Definition 5. A function $F : \mathbb{Z}_+ \times L^1(\Omega;\mathbb{B}_1) \mapsto L^1(\Omega;\mathbb{B}_2)$, $(n,U) \mapsto F(n,U)$ is said to be almost periodic in mean in $n \in \mathbb{Z}_+$ uniformly in $U \in \mathbb{K}$ where $\mathbb{K} \subset L^1(\Omega;\mathbb{B}_1)$ is a compact if for any $\varepsilon > 0$, there exists a positive integer $l(\varepsilon,\mathbb{K})$ such that among any l consecutive integers, there exists at least an integer p with the following property:

$$\mathbf{E}\|F(n+p,U) - F(n,U)\| < \varepsilon$$

for each random variable $U \in L^1(\Omega;\mathbb{B}_1)$ and $n \in \mathbb{Z}^+$.

Here again, the number p will be called an ε-translation of F, and the set of all ε-translations of F is denoted by $\mathscr{E}(\varepsilon,F,\mathbb{K})$.

Let $UB(\mathbb{Z}_+; L^1(\Omega;\mathbb{B}))$ denote the collection of all uniformly bounded $L^1(\Omega;\mathbb{B})$-valued random sequences $X = \{X(n)\}_{n\in\mathbb{Z}_+}$. It is then easy to check that the space $UB(\mathbb{Z}_+; L^1(\Omega;\mathbb{B}))$ is a Banach space when it is equipped with the norm:

$$\|X\|_\infty = \sup_{n\in\mathbb{Z}_+} \mathbf{E}\|X(n)\|.$$

Lemma 1. *$AP(\mathbb{Z}_+;L^1(\Omega;\mathbb{B})) \subset UB(\mathbb{Z}_+;L^1(\Omega;\mathbb{B}))$ is a closed space.*

In view of the above, the space $AP(\mathbb{Z}_+;L^1(\Omega;\mathbb{B}))$ of almost periodic random sequences equipped with the sup norm $\|\cdot\|_\infty$ is also a Banach space.

We now state the following composition result.

Theorem 5. *Let $F : \mathbb{Z}_+ \times L^1(\Omega;\mathbb{B}_1) \mapsto L^1(\Omega;\mathbb{B}_2)$, $(n,U) \mapsto F(n,U)$ be almost periodic in mean in $n \in \mathbb{Z}_+$ uniformly in $U \in L^1(\Omega;\mathbb{B}_1)$. If in addition, F is Lipschitz in $U \in \mathbb{K}$, where $\mathbb{K} \subset L^1(\Omega;\mathbb{B}_1)$ is compact (i.e., there exists $L > 0$ such that*

$$\mathbf{E}\|F(t,U) - F(t,V)\|_2 \leq M\,\mathbf{E}\|U - V\|_1 \ \ \forall U,V \in L^1(\Omega;\mathbb{B}_1), n \in \mathbb{Z}_+),$$

then for any almost periodic random sequence $X = \{X(n)\}_{n\in Z+}$, then the $L^1(\Omega;\mathbb{B}_1)$-valued random sequence $Y(n) = F(n,X(n))$ is almost periodic in mean.

Almost Periodic Stochastic Difference Equations

Let $\left(\mathbb{R}, |\cdot|\right)$, $\left(\mathbb{R}^k, |\cdot|\right)$ be the field of real numbers equipped with its absolute value, the k-dimensional space of real numbers equipped with Euclidean topology, respectively.

Our main objective in this paper is to find sufficient conditions for the existence the existence of an almost periodic solution of the stochastic semi-linear systems of difference equations of type (1).

To study solutions of Eq. (1), we use the fundamental solutions of the system

$$X(\omega,n+1) = A(\omega,n)X(\omega,n),\ n \in \mathbb{Z}_+\,,\ \omega \in \Omega \tag{3}$$

to examine almost periodic solutions of the system of difference equations

$$X(\omega,n+1) = A(\omega,n)X(\omega,n) + g(\omega,n),\ n \in \mathbb{Z}_+\,,\ \omega \in \Omega \tag{4}$$

where $g : \Omega \times \mathbb{Z}_+ \to \mathbb{R}^k$ is almost periodic in mean. Here, we assume that g satisfies the following property: There exists a random variable Y with $\mathbf{E}[Y] < \infty$ such that $|g(n)| \le Y$ for all $n \in \mathbb{Z}_+$.

For $n, m \in \mathbb{Z}_+$, we define the transition matrix

$$\Phi(n,m) = \prod_{r=m}^{n-1} A(r)\,.$$

Definition 6. Equation (3) is said to have a regular exponential dichotomy if there exists $k \times k$ projection matrices $P(n)$ with $n \in \mathbb{Z}_+$ and positive constants M and $\beta \in (0,1)$ such that the following four conditions are satisfied:

(i) $A(n)P(n) = P(n+1)A(n)$.
(ii) The matrix $A(n)|_{R(I-P(n))}$ is an isomorphism from $R(I-P(n))$ onto $R(I-P(n+1)$.
(iii) $\|\Phi(n,m)P(m)X\| \le M\beta^{n-m}\|X\|$, for $0 \le m \le n$, $X \in L^1(\Omega,\mathbb{R}^k)$.
(iv) $\|\Phi(m,n)(I-P(n))X\| \le M\beta^{n-m}\|X\|$, for $0 \le m \le n$, $X \in L^1(\Omega,\mathbb{R}^k)$.

By repeated application of [(i), Definition 6], we obtain

$$P(n)\Phi(n,m) = \Phi(n,m)P(m)\,.$$

Define the hull $H(X)$ of a random sequence X as follows:

Definition 7. The set

$$H(X) = \left\{\tilde{X} \,|\, \text{there exists a sequence } \mathbf{k} \subset \mathbb{Z}_+ \text{with } T_{\mathbf{k}}X = \tilde{X}\right\}.$$

Similarly, for a matrix function $A(n)$, we define

$$H(A) = \left\{\tilde{A} \,|\, \text{there exists a sequence } \mathbf{k} \subset \mathbb{Z}_+ \text{with } T_{\mathbf{k}}A = \tilde{A}\right\},$$

where $T_{\mathbf{k}}A = \tilde{A}$ means that $\lim_{i\to\infty} A(n+l(i)) = \tilde{A}(n)$.

Theorem 6. *Suppose that Eq. (3) has a regular exponential dichotomy and $\tilde{A}(n) \in H(A(n))$. Then the system*

$$X(n+1) = \tilde{A}(n)X(n)$$

satisfies a regular exponential dichotomy with same projections and constants.

Let us now state the main results of this paper. For linear stochastic difference equations, we obtain the following theorem.

Theorem 7. *Let $\{A(n)\}_{n\in\mathbb{Z}_+}$ be a sequence of invertible random matrices satisfying Bochner's almost sure uniform double sequence criterion. Suppose that Eq. (3) has*

a regular exponential dichotomy and $\tilde{A}(n) \in H(A(n))$. Then Eq. (4) has an almost periodic solution given by

$$\bar{X}(n) = \sum_{r=-\infty}^{n-1} \Phi(n, r+1)P(r+1)g(r) - \sum_{r=n}^{\infty} \Phi(n, r+1)(I - P(r+1))g(r), \qquad (5)$$

where $\Phi(n,r)P(r) = 0$ for $r > n$ and $g(r) = 0$ for $r < 0$.

Proof. It is not hard to show that $\bar{X}(n)$ defined by Eq. (5) is a solution of Eq. (3). Moreover,

$$\begin{aligned}
\|\bar{X}(n)\| &\leq \sum_{r=-\infty}^{n-1} \|\Phi(n, r+1)P(r+1)g(r)\| + \sum_{r=n}^{\infty} \|\Phi(n, r+1)(I - P(r+1))g(r)\| \\
&\leq \sum_{r=0}^{n-1} M\beta^{n-r-1}\|g(r)\| + \sum_{r=n}^{\infty} M\beta^{n-r-1}\|g(r)\| \\
&\leq \Big\{ \sum_{r=0}^{n-1} M\beta^{n-r-1} + \sum_{r=n}^{\infty} M\beta^{n-r-1} \Big\} Y \\
&\leq M\frac{1+\beta}{1-\beta} Y .
\end{aligned}$$

This implies that $\{\|\bar{X}(n)\|, n \in \mathbb{Z}_+\}$ is uniformly integrable. Now, to prove the almost periodicity of $\bar{X}(\cdot)$, it suffices by Theorem 4 to show that $\bar{X}(\cdot)$ satisfies Bochner's almost sure uniform double sequence criterion. To this end, let $\mathbf{k}' = (k_i')$ and $\mathbf{l}' = (l_i')$ be arbitrary sequences of nonnegative integers, and then choose a measurable set $\Omega_1 \subset \Omega$ with $\mathbf{P}(\Omega_1) = 1$. Let $(k_i) \subset (k_i')$ and $(l_i') \subset (l_i')$ be their common subsequences such that for each $\omega \in \Omega_1$, $(T_{\mathbf{k}+\mathbf{l}}A(\omega) = (T_{\mathbf{l}}T_{\mathbf{k}}A)(\omega)$ and $(T_{\mathbf{k}+\mathbf{l}}g(\omega) = (T_{\mathbf{l}}T_{\mathbf{k}}g)(\omega)$. For simplicity, we omit ω in what follows. Then we have

$$\begin{aligned}
\bar{X}(n+k_i) &= \sum_{r=-\infty}^{n+k_i-1} \Phi(n+k_i, r+1)P(r+1)g(r) \\
&\quad - \sum_{r=n+k_i}^{\infty} \Phi(n+k_i, r+1)[I - P(r+1)]g(r) \\
&= \sum_{r=-\infty}^{n-1} \Phi(n+k_i, s+k_i+1)P(s+k_i+1)g(s+k_i) \\
&\quad - \sum_{r=n}^{\infty} \Phi(n+k_i, s+k_i+1)[I - P(s+k_i+1)]g(s+k_i) \\
&= \sum_{r=-\infty}^{n-1} A(n+k_i-1)\cdots A(s+k_i+1)P(s+k_i+1)g(s+k_i) \\
&\quad - \sum_{r=n}^{\infty} A(n+k_i-1)\cdots A(s+k_i+1)[I - P(s+k_i+1)]g(s+k_i) .
\end{aligned}$$

Thus, taking the limit of the above expression as $i \to \infty$ and recalling the fact that $\lim_{i\to\infty} \bar{X}(n+k_i) = (T_{\mathbf{k}}\bar{X})(n)$, we can then write

$$\begin{aligned}(T_{\mathbf{k}}\bar{X})(n) &= \sum_{r=-\infty}^{n-1} \tilde{A}(n-1)\cdots\tilde{A}(s+1)\tilde{P}(s+1)\tilde{g}(s) \\ &\quad - \sum_{r=n}^{\infty} \tilde{A}(n-1)\cdots\tilde{A}(s+1)[I-\tilde{P}(s+1)]\tilde{g}(s) \\ &= \sum_{r=-\infty}^{n-1} (T_{\mathbf{k}}A)(n-1)\cdots(T_{\mathbf{k}}A)(s+1)(T_{\mathbf{k}}P)(s+1)(T_{\mathbf{k}}g)(s) \\ &\quad - \sum_{r=n}^{\infty} (T_{\mathbf{k}}A)(n-1)\cdots(T_{\mathbf{k}}A)(s+1)[I-(T_{\mathbf{k}}P)(s+1)](T_{\mathbf{k}}g)(s)\,.\end{aligned}$$

Moreover,

$$\begin{aligned}(T_{\mathbf{l}}T_{\mathbf{k}}\bar{X})(n) &= \sum_{r=-\infty}^{n-1} (T_{\mathbf{l}}T_{\mathbf{k}}A)(n-1)\cdots(T_{\mathbf{l}}T_{\mathbf{k}}A)(s+1)(T_{\mathbf{l}}T_{\mathbf{k}}P)(s+1)(T_{\mathbf{l}}T_{\mathbf{k}}g)(s) \\ &\quad - \sum_{r=n}^{\infty} (T_{\mathbf{l}}T_{\mathbf{k}}A)(n-1)\cdots(T_{\mathbf{l}}T_{\mathbf{k}}A)(s+1)[I-(T_{\mathbf{l}}T_{\mathbf{k}}P)(s+1)](T_{\mathbf{l}}T_{\mathbf{k}}g)(s) \\ &= (T_{\mathbf{l}+\mathbf{k}}\bar{X})(n)\,,\end{aligned}$$

as desired. □

From now on, we assume that the random evolution operator $\Phi(n,m)$ generated by $A(n)$ is uniformly exponentially stable. That is, there exist constants $M > 0$ and $\beta \in (0,1)$ such that $\|\Phi(n,m)\| \le M\beta^{n-m}$ for all $n \ge m$. This implies that the projection matrix $P(n)$ used in the definition of the regular exponential dichotomy is the identity.

In order to state similar results for the nonlinear case, we set $\mathscr{O} = \{y \in \mathbb{R}^k : |y| \le \delta\}$ for a fixed $\delta > 0$ and take a function $f : Z_+ \times L^1(\Omega,\mathscr{O}) \to L^1(\Omega,\mathbb{R}^k)$, $(n,X) \mapsto f(n,X)$ with $f(n,0) = 0$ for which there exists a constant $L > 0$ such that

$$\mathbf{E}\|f(n,U) - f(n,V)\| \le L \cdot \mathbf{E}\|U - V\|,\ \forall U, V \in L^1(\Omega,\mathscr{O}), n \in \mathbb{Z}_+\,.$$

In addition, we assume that there exists a random variable Y with $\mathbf{E}[Y] < \infty$ such that $|f(n,U)| \le Y$ for all $n \in \mathbb{Z}_+$ and $U \in L^1(\Omega,\mathbb{R}^k)$.

Under these conditions on A and f, we have the following theorem.

Theorem 8. *Let $\{A(n)\}_{n\in\mathbb{Z}_+}$ be a sequence of invertible random matrices satisfying Bochner's almost sure uniform double sequence criterion. Suppose that the linear stochastic difference equation Eq. (3) corresponding to Eq. (1) is uniformly exponentially stable and that $f = \{f(n,X)\}_{n\in\mathbb{Z}_+, X\in L^1(\Omega,\mathbb{R}^k)}$ is almost periodic in mean. Then Eq. (1) has a unique almost periodic solution*

$$\bar{X}(n) = \sum_{r=0}^{n-1} \left(\prod_{s=r}^{n-1} A(s) \right) f(r, X(r))$$

provided that

$$\frac{M\beta L}{1-\beta} < 1 . \tag{6}$$

Proof. Note that

$$\begin{aligned} \left\| \bar{X}(n) \right\| &\leq \sum_{r=0}^{n-1} \|\Phi(n,r)\| \, \|f(r, X(r))\| \\ &\leq \left\{ \sum_{r=0}^{n-1} M\beta^{n-r} \right\} Y \leq \frac{M\beta}{1-\beta} Y . \end{aligned}$$

Consider the Banach space $AP(\mathbb{Z}_+; L^1(\Omega, \mathbb{R}^k))$ with the super norm. By Theorem 5, if $\varphi \in AP(\mathbb{Z}_+; L^1(\Omega, \mathbb{R}^k))$, then $f(\cdot, \varphi(\cdot)) \in AP(\mathbb{Z}_+; L^1(\Omega, \mathbb{R}^k))$. Now, define

$$\Gamma : AP(\mathbb{Z}_+; L^1(\Omega, \mathbb{R}^k)) \to AP(\mathbb{Z}_+; L^1(\Omega, \mathbb{R}^k))$$

be the nonlinear operator defined by

$$(\Gamma\varphi)(n) := \sum_{r=0}^{n-1} \Phi(n,r) \, f(r, \varphi(r)) .$$

By Theorem 7, Γ is well defined. Now, let φ, $\psi \in AP(\mathbb{Z}_+; L^1(\Omega, \mathbb{R}^k))$ having the same property as X defined in Eq. (1). We can easily see that

$$\begin{aligned} \mathbf{E}\|(\Gamma\varphi)(n) - (\Gamma\psi)(n)\| &\leq \sum_{r=0}^{n-1} \left(M\beta^{n-r} \mathbf{E}\|f(r, \varphi(r)) - f(r, \psi(r))\| \right) \\ &\leq \frac{M\beta L}{1-\beta} \sup_{r \in Z_+} \mathbf{E}\|\varphi(r) - \psi(r)\| . \end{aligned}$$

Thus,

$$\|\Gamma\varphi - \Gamma\psi\|_\infty \leq \frac{M\beta L}{(1-\beta)} \|\varphi - \psi\|_\infty .$$

Γ is a contraction provided that $\frac{M\beta L}{1-\beta} < 1$. Using the Banach fixed point theorem, we obtain that Γ has a unique fixed point $\bar{X}$, which is the unique almost periodic solution of Eq. (1). □

Application

In constant environments, theoretical discrete-time population models are usually formulated under the assumption that the dynamics of the total population size in generation n, denoted by $X(n)$, are governed by equations of the form

$$X(n+1) = \gamma X(n) + f(X(n)), \tag{7}$$

where $\gamma \in (0,1)$ is the constant "probability" of surviving per generation and $f : \mathbb{R}_+ \to \mathbb{R}_+$ models the recruitment process.

Almost periodic effects can be introduced into Eq. (7) by writing the recruitment function or the survival probability as almost periodic random sequences. This is model with the equation

$$X(n+1) = \gamma_n X(n) + f(n, X(n)), \tag{8}$$

where either $\{\gamma_n\}_{n\in\mathbb{Z}_+}$ or $f(n,X(n)) \in AP(\mathbb{Z}_+; L^1(\Omega,\mathbb{R}))$ and each $\gamma_n \in (0,1)$.

In their paper, Franke and Yakubu [6] studied (8) with the periodic Beverton–Holt recruitment function

$$f(n,X(n)) = \frac{(1-\gamma_n)\mu K_n X(n)}{(1-\gamma_n)K_n + (\mu - 1 + \gamma_n)X(n)}, \tag{9}$$

where the carrying capacity K_n is p-periodic, $K_{n+p} = K_n$ for all $n\mathbb{Z}_+$, and $\mu > 1$.

In this section, we assume that both carrying capacity K_n and the survival rate γ_n are random and that $\{\gamma_n, n \in \mathbb{Z}_+\}$ are independent and independent of the sequence K_n, $n \in \mathbb{Z}_+$.

We have the following theorem:

Theorem 9. *Suppose that both $\{\gamma_n\}_{n\in\mathbb{Z}_+}$ and $\{K_n\}_{n\in Z_+}$ are almost periodic in mean. Then Eqs. (8)–(9) has a unique almost periodic solution whenever*

$$\sup_{n\in Z_+} \mathbf{E}[\gamma_n] < \frac{1}{\mu+1}.$$

Proof. Equation (8) is in the form of Eq. (1), where

$$A(n) = \gamma_n$$

and

$$f(n,X(n)) = \frac{(1-\gamma_n)\mu K_n X(n)}{(1-\gamma_n)K_n + (\mu - 1 + \gamma_n)X(n)}.$$

It is a routine to show that

$$|f(n,U) - f(n,V)| \le \mu |U - V|,$$

and hence

$$\mathbf{E}|f(n,U) - f(n,V)| \le \mu \mathbf{E}|U - V|.$$

To prove the almost periodicity of $n \to f(n,X(n))$, set $A_n = (1-\gamma_n)K_n$ and $B_n = \mu - 1 + \gamma_n$. Then f can be written as follows:

$$f(n,X(n)) = \mu \frac{A_n X(n)}{A_n + B_n X(n)} \quad \text{for each } n \in \mathbb{Z}_+.$$

Using the fact that $\{\gamma_n\}$ and $\{K_n\}$ are almost periodic in mean and making use of, respectively, Theorem 2 and Corollary 1, we can choose a constant $K > 0$ such that $\mathbf{E}|K_n| < K$ for all $n \in \mathbb{Z}_+$ and for each $\varepsilon > 0$ there exists a positive integer $N_0(\varepsilon)$ such that among any $N_0(\varepsilon)$ consecutive integers, there exists an integer $p > 0$, a common ε-almost period for $\{\gamma_n\}$ and $\{K_n\}$ for which

$$\mathbf{E}|\gamma_{n+p} - \gamma_n| \leq \frac{\varepsilon(\mu-1)^2}{2\mu^2 K}, \text{and } \mathbf{E}|K_{n+p} - K_n| \leq \frac{\varepsilon(\mu-1)}{4\mu^2}.$$

Observe that

$$\begin{aligned}\mathbf{E}|f(n+p,U) - f(n,U)| &= \mu\mathbf{E}\left|\frac{A_{n+p}U}{A_{n+p}+B_{n+p}U} - \frac{A_nU}{A_n+B_nU}\right| \\ &\leq \mu\mathbf{E}\left|\frac{(A_{n+p}B_n - A_nB_{n+p})U^2}{B_{n+p}B_nU^2}\right| = \mu\mathbf{E}\left|\frac{A_{n+p}}{B_{n+p}} - \frac{A_n}{B_n}\right|.\end{aligned}$$

To evaluate $\mathbf{E}\left|\frac{A_{n+p}}{B_{n+p}} - \frac{A_n}{B_n}\right|$, we borrow a calculation from [3]. Using the hypothesis of independence of the random sequences $\{\gamma_n\}_{n\in\mathbb{Z}_+}$ and $\{K_n\}_{n\in\mathbb{Z}_+}$, we have

$$\begin{aligned}\mathbf{E}\left|\frac{A_{n+p}}{B_{n+p}} - \frac{A_n}{B_n}\right| &= \mathbf{E}\left|\frac{(1-\gamma_{n+p})K_{n+p}}{\mu-1+\gamma_{n+p}} - \frac{(1-\gamma_n)K_n}{\mu-1+\gamma_n}\right| \\ &= \mathbf{E}\Bigg[\frac{1}{(\mu-1+\gamma_{n+p})(\mu-1+\gamma_n)}\Big|(\mu-1)[K_{n+p}-K_n] - \gamma_n\gamma_{n+p}[K_{n+p}-K_n] \\ &\quad -(\mu-1)[\gamma_{n+p}K_{n+p} - \gamma_nK_n] + [\gamma_nK_{n+p} - \gamma_{n+p}K_n]\Big|\Bigg] \\ &= \mathbf{E}\Bigg[\frac{1}{(\mu-1+\gamma_{n+p})(\mu-1+\gamma_n)}\Big|(\mu-1)[K_{n+p}-K_n] - \gamma_n\gamma_{n+p}[K_{n+p}-K_n] \\ &\quad -(\mu-1)K_{n+p}[\gamma_{n+p}-\gamma_n] + \gamma_n[K_{n+p}-K_n] + \gamma_n[K_{n+p}-K_n] - [\gamma_{n+p}-\gamma_n]\Big|\Bigg] \\ &= \mathbf{E}\Bigg|\frac{\mu-1}{(\mu-1+\gamma_{n+p})(\mu-1+\gamma_n)}[K_{n+p}-K_n] \\ &\quad -\frac{\gamma_n\gamma_{n+p}}{(\mu-1+\gamma_{n+p})(\mu-1+\gamma_n)}[K_{n+p}-K_n] \\ &\quad -\frac{\mu-1}{(\mu-1+\gamma_{n+p})(\mu-1+\gamma_n)}K_{n+p}[\gamma_{n+p}-\gamma_n] \\ &\quad +\frac{(\mu-1)\gamma_n}{(\mu-1+\gamma_{n+p})(\mu-1+\gamma_n)}[K_{n+p}-K_n] \\ &\quad -\frac{\gamma_n}{(\mu-1+\gamma_{n+p})(\mu-1+\gamma_n)}[K_{n+p}-K_n]\end{aligned}$$

$$
\begin{aligned}
&\left. -\frac{1}{(\mu-1+\gamma_{n+p})(\mu-1+\gamma_n)} K_n\left[\gamma_{n+p}-\gamma_n\right] \right| \\
&\leq \frac{1}{\mu-1}\mathbf{E}|K_{n+p}-K_n| + \mathbf{E}|K_{n+p}-K_n| + \frac{1}{\mu-1}\mathbf{E}|K_{n+p}|\,\mathbf{E}|\gamma_{n+p}-\gamma_n| \\
&\quad + \mathbf{E}|K_{n+p}-K_n| + \frac{1}{\mu-1}\mathbf{E}|K_{n+p}-K_n| + \frac{1}{(\mu-1)^2}\mathbf{E}|K_n|\,\mathbf{E}|\gamma_{n+p}-\gamma_n| \\
&\leq \frac{2\mu}{\mu-1}\mathbf{E}|K_{n+p}-K_n|] + \frac{\mu}{(\mu-1)^2}K\cdot\mathbf{E}|\gamma_{n+p}-\gamma_n| \,.
\end{aligned}
$$

Thus, we obtain

$$
\mathbf{E}\,|f(n+p,U)-f(n,U)| \leq \frac{\varepsilon}{2}+\frac{\varepsilon}{2} = \varepsilon\,.
$$

By Theorem 5, we can conclude that $n \to f(n,X(n))$ is almost periodic in mean.

Now, let $M=1$ and $\beta = \sup_{n\in Z_+}\mathbf{E}[\gamma_n]$. Then $\sup_{n\in Z_+}\mathbf{E}[\gamma_n] < \frac{1}{\mu+1}$ implies that

$$
\frac{M\beta L}{1-\beta} = \frac{\mu \sup_{n\in\mathbb{Z}_+}\mathbf{E}[\gamma_n]}{1-\sup_{n\in\mathbb{Z}_+}\mathbf{E}[\gamma_n]} < 1,
$$

and Eq. (6) is satisfied. Applying Theorem 8 yields the result. □

References

1. Bedouhene, F., Mellah, O., Raynaud de Fitte, P.: Bochner-almost periodicity for stochastic processes. Stoch. Anal. Appl. **30**(2), 322–342 (2012)
2. Bezandry, P., Diagana, T.: Almost Periodic Stochastic Processes. Springer, New York (2011)
3. Bezandry, P., Diagana, T., Elaydi, S.: On the stochastic Beverton–Holt equation with survival rates. J. Differ. Eq. Appl. **14**(2), 175–190 (2008)
4. Corduneanu, C.: Almost Periodic Functions, 2nd edn. Chelsea, New York (1989)
5. Diagana, T., Elaydi, S., Yakubu, A.-A.: Population models in almost periodic environments. J. Differ. Equat. Appl. **13**(4), 239–260 (2007)
6. Franke, J.E., Yakubu, A.-A.: Population models with periodic recruitment functions and survival rates. J. Differ. Equat. Appl. **11**(14), 1169–1184 (2005)
7. Han, J., Hong, C.: Almost periodic random sequences in probability. J. Math. Anal. Appl. **336**, 962–974 (2007)
8. Hong, J., Nunez, C.: The almost periodic type difference equations. Math. Comput. Model. **28**(12), 21–31 (1998)

Index

B. Toni (ed.), *Advances in Interdisciplinary Mathematical Research*, Springer Proceedings in Mathematics & Statistics 37, DOI 10.1007/978-1-4614-6345-0,