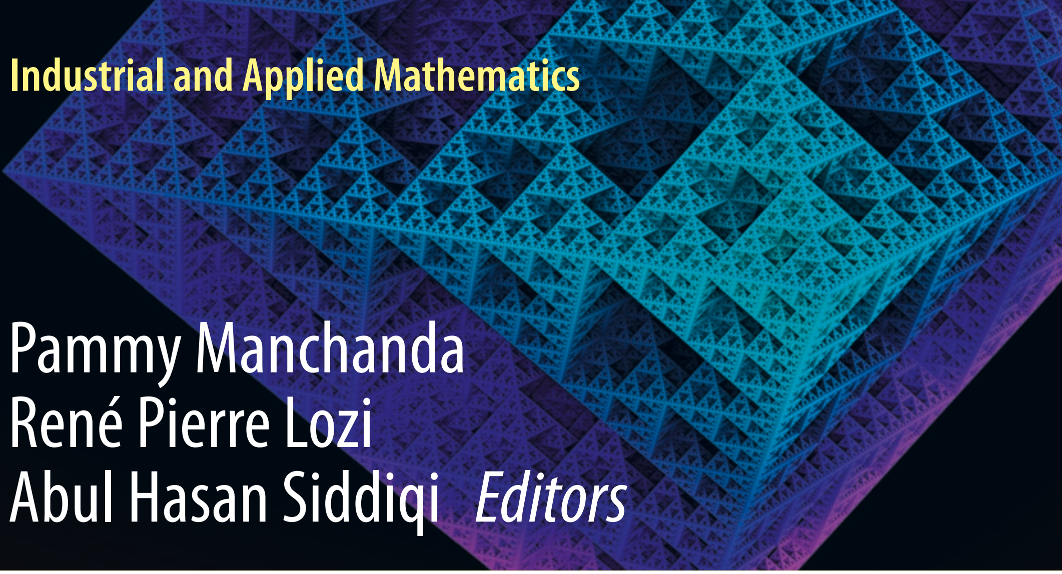
Industrial and Applied Mathematics



Pammy Manchanda
René Pierre Lozi
Abul Hasan Siddiqi *Editors*

Mathematical Modelling, Optimization, Analytic and Numerical Solutions

Springer

Industrial and Applied Mathematics

The *Industrial and Applied Mathematics* series publishes high-quality research-level monographs, lecture notes and contributed volumes focusing on areas where mathematics is used in a fundamental way, such as industrial mathematics, bio-mathematics, financial mathematics, applied statistics, operations research and computer science.

More information about this series at http://www.springer.com/series/13577

Pammy Manchanda · René Pierre Lozi ·
Abul Hasan Siddiqi
Editors

Mathematical Modelling, Optimization, Analytic and Numerical Solutions

Springer

Editors
Pammy Manchanda
Department of Mathematics
Guru Nanak Dev University
Amritsar, Punjab, India

René Pierre Lozi
Laboratory Math, J.A. Dieudonné
University Côte d'Azur
Nice, France

Abul Hasan Siddiqi
School of Basic Sciences and Research
Sharda University
Greater Noida, Uttar Pradesh, India

ISSN 2364-6837 ISSN 2364-6845 (electronic)
Industrial and Applied Mathematics
ISBN 978-981-15-0927-8 ISBN 978-981-15-0928-5 (eBook)
https://doi.org/10.1007/978-981-15-0928-5

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Dedicated to

Professor Abul Hasan Siddiqi (1943–2020)

Preface

The edited volume is based on 12 invited talks, presidential address, and selected nine peer-reviewed papers during an international conference at GNDU, Amritsar, India, in February 2018. The volume covers a variety of themes of industrial and applied mathematics. The main attraction of the conference was two symposiums, one organized by Prof. Guenter Leugering, Vice President, FAU, Erlangen–Nuremberg, Germany, and the other organized by Prof. GDV Gowda, Dean, TIFR Center of Applicable Mathematics, Bangalore, in which speakers were from reputed institutions of Germany and India. All talks were devoted to the role of partial differential equations in understanding the real-world problems. Leugering and his group focused on the system of hyperbolic equations on networks such as nonlinear elastic strings and beams, potentially being coupled via masses and viscoelastic springs, or networks of pipes conveying fluids or gas. Chapter 1 contains a resume of work of Prof. A. H. Siddiqi and his co-workers. Professor Pavel Exner, Current President of European Mathematical Society, has discussed the effect of an abrupt spectral change for some classes of Schrodinger operators depending on the value of the coupling constant, from below bounded and partly or fully discrete, to the continuous one covering the whole real axis in Chap. 2. In Chap. 3, Prof. Hans G. Feichtinger and Mads S. Jakobsen have presented an interesting account of distribution theory by Riemann integrals. In this chapter, they have outlined a syllabus for a course that can be given to engineers looking for an understandable mathematical description of the foundation of distribution theory and the necessary functional analytic methods. Guenter Leugering has presented in Chap. 4 the study of partial differential equations on metric graphs: a survey of results on optimization, control and stabilizability with special focus on shape and topological sensitivity problems. His research collaborators Martin Gugat and Michael Herty have presented a new model for transient flow in gas transportation networks in Chap. 6. In Chap. 5, D. Provitolo, R Lozi and E. Tric have presented their study regarding a new model of weighted human behavior in the context of urban terrorist attacks. It may be remarked that in the context of disasters, and in order to better protect the population, one of the major challenges today is to better understand and anticipate both individual and collective human behavior, and the dynamics of the displacements associated with these behaviors. In Chap. 7, Falk M. Hante has presented his research on “Mixed-Integer Optimal Control for PDES: Relaxation via Differential Inclusions and Applications to Gas Network Optimization.” In Chap. 8, Prof. Mukhayo Rasulova has presented her studies “Application of Solution of the Quantum Kinetic Equations for Information Technology and Renewable Energy Problem.” Professor Taufiquar Rahman Khan, Clemson University, USA, has introduced the general idea of inverse problems particularly with applications to imaging in Chap. 9. Professor K. Sreenadh, IIT Delhi, and T. Mukherjee, TIFR Bangalore, have written a survey on critical growth elliptic problems with Choquard-type nonlinearity in Chap. 10. Professor Ratish Kumar, IIT Kanpur, along with his research collaborator Gopal Priyadashi has presented work on Wavelet Galerkin Methods for Higher-Order Partial Differential Equations in Chap. 11. In Chap. 12, Samares Pal and Joydeb Bhattacharyya have studied resilience and dynamics of coral reefs impacted by chemically rich seaweeds and unsustainable fishing. Interesting mathematical models are discussed. Chapter 13—Multigrid Methods for the Simulations of Surfactant Spreading on a Thin Liquid Film—by Satyananda Panda and Aleksander Grm contains a multigrid approach for the simulations of surfactant spreading on a thin liquid film. Chapter 14 by Eenezer Bonyah, Fahad Al Basir and Santanu Ray is devoted to Hopf bifurcation in a mathematical model of tuberculosis with delay. In Chap. 15, Amit Kumar Roy and Priti Kumar Roy have presented their studies related to Treatment of Psoriasis by Interleukin-10 through Impulsive Control Strategy. A. Akhilandeeswari, K. Balachandran and N. Annapoorani have investigated the existence of solution of the fractional partial differential equations of diffusion type with integral kernel in Chap. 16. In Chap. 17 Mathematical Study on Human Cells Interaction Dynamics for HIV-TV Co-infection by Suman Dolai, Amit Kumar Roy and Priti Kumar Roy is presented. In Chap. 18, P. Suresh Kumar presents his work on Relative Controllability of Nonlinear Fractional Damped Delay Systems with Multiple Delays in Control. Chapter 19 by Rohit Khokher and Ram Chandra Singh is devoted to A Graphical User Interface-Based Fingerprint Recognition. In Chap. 20, P. Umamaheswari, K. Balachandran and N. Annapoorani have investigated the existence and stability results of stochastic fractional delay differential equations with Gaussian noise. Chapter 21 by Kausika Chellamuthu deals with “Asymptotic Stability of Implicit Fractional Volterra Integrodifferential Equations.” Few invited and contributory talks of the international conference at GNDU by Prof. Stephane Jaffard, Prof. M. Shah Jahan, Prof. G. Fairweather, Prof. A. K. Pani, Prof. S. Dharam Raja, Prof. Rashmi Bhardwaj and Prof. T. D. Narang could not be included in this volume due to unavoidable reasons. We thank all of them for sparing time to deliver lectures and for encouraging young researchers.

We take this opportunity to thank Ms. Pooja and Ms. Mamta, UGC Research Fellows at Guru Nanak Dev University, Amritsar, for their contribution toward compiling the manuscript.

Amritsar, India	Pammy Manchanda
Nice, France	René Pierre Lozi
Greater Noida, India	Abul Hasan Siddiqi

Acknowledgements

Editors and organizers gratefully acknowledge the financial support for the Department of Science and Technology, New Delhi: ICIAM and Duty Society, Aligarh Muslim University (AMU). The Organizing Committee and editors also express their gratitude to the vice-chancellor and other functionaries of Guru Nanak Dev University, Amritsar, for generous support without which the conference could not have been organized at this large scale. We also take this opportunity to thank Mr. Shamim Ahmad of Springer India for his valuable support.

Contents

About the Editors

Pammy Manchanda is Professor at the Department of Mathematics Guru Nanak Dev University, Amristar. Professor Manchanda has published more than 50 research papers in several international journals of repute, edited 4 proceedings volumes for international conferences of the Indian Society of Industrial and Applied Mathematics (ISIAM) and co-authored 4 books. She has attended, delivered talks and chaired sessions at reputed academic conferences and workshops across the world, including International Council of Industrial and Applied Mathematics (ICIAM) from 1999 to 2019 and the International Congress of Mathematicians (ICM) since 2002. She is the managing editor of the *Indian Journal of Industrial and Applied Mathematics* and a member of the editorial board of the Springer book series, *Industrial and Applied Mathematics*. She was invited twice to the Industrial Mathematics Group of Prof. Helmut Neunzert at Kaiserslautern University, Germany, and has visited the International Center for Theoretical Physics (UNESCO institution) at Trieste, Italy, several times to carry out research activities. She has been joint secretary of ISIAM from 1999–2016, and is currently secretary of the society. She is ISIAM representative in the International Council of Industrial and Applied Mathematics (ICIAM). She has been actively engaged in the organization of international conferences by the ISIAM and was joint convener, Satellite Conference of ICM, during 15–17 August 2010, New Delhi, India.

René Pierre Lozi is Professor at the Dieudonné Center of Mathematics, University Côte d'Azur, France. In 1991, he became Full Professor at the University Côte d'Azur and the Institute for Teacher Trainees (IUFM), France. He served as Director of IUFM during 2001–2006 and as Vice-Chairman of the French Board of Directors of IUFM during 2004–2006. He completed his French state thesis on chaotic dynamical systems under the supervision of Prof. René Thom, a Fields Medalist, in 1983. Professor Lozi currently serves on the editorial boards of respected international journals. In 1977, he discovered a particular mapping of the plane having a strange attractor (now classically known as “Lozi map”). Today, his research areas include complexity and emergence theory, dynamical systems, bifurcation, control of chaos, cryptography-based chaos, and memristors (a physical device for neuro-computing).

Abul Hasan Siddiqi is Professor Emeritus and a distinguished scientist at the School of Basic Sciences and Research, Sharda University, Greater Noida, India. Professor Siddiqi has held several important administrative positions, such as Chairman of the Department of Mathematics, Dean of the Faculty of Science, and Pro-Vice-Chancellor of Aligarh Muslim University, India. He has been actively associated with the International Centre for Theoretical Physics, Trieste, Italy (UNESCO organization), in different capacities for more than 20 years. He served as Professor of Mathematics at the King Fahd University of Petroleum and Minerals (KFUPM), Saudi Arabia, for 10 years; and a consultant to Sultan Qaboos University, Oman; Istanbul Aydin University, Turkey; and the Institute of Microelectronics, Malaysia. He was awarded the German Academic Exchange Service (DAAD) Fellowship thrice to pursue mathematical research in Germany. He has jointly published more than 100 research papers with his research collaborators, 5 books, and edited proceedings of 9 international conferences. He is Founder Secretary of the Indian Society of Industrial and Applied Mathematics (ISIAM) and Editor-in-Chief of the *Indian Journal of Industrial and Applied Mathematics*, published by the ISIAM, and the Springer book series, *Industrial and Applied Mathematics*. He has recently been elected as President of ISIAM, which represents India at the apex forum of industrial and applied mathematics—ICIAM.

Chapter 1
Certain Areas of Industrial and Applied Mathematics

Abul Hasan Siddiqi

Abstract This chapter is based on Presidential address at the International Conference and 14th Biennial Conference of Indian Society of Industrial and Applied Mathematics, GNDU, Amritsar, Feb 2–4, 2018.

Keywords Walsh and Haar systems · 2-dimensional analogues of Banach space · Variational inequality · Wavelet theory · Fractals · Shearlets

1.1 Introduction

Here we introduce our contribution in certain areas of industrial and applied mathematics such as dyadic harmonic analysis, variational inequalities, wavelet theory, fractals, financial mathematics. People interested in these areas may go through books authored and edited by myself and colleagues and research papers written jointly by research scholars and colleagues. See selected research papers, books authored and edited volumes.

1.2 Areas Pursued By Our Research Group

i. Fourier and Dyadic Harmonic Analysis: I tried to popularize certain areas with which researchers in our country were not familiar. In the 60s, most of the researchers in northern India were engaged in Summability theory of sequences and special functions. I tried to introduce to Indian researchers, new areas such as classes of Fourier coefficients, approximation by Fourier series, and other orthonormal systems such as Walsh and Haar Systems. I pursued vigorously Walsh Fourier analysis and collaborated with Indian engineers (professors of electrical engineers IIT Kanpur) and Hungarian Mathematicians like F. Moricz, S. Fridli and F. Schipp. We learnt from Hungarian Mathematician's approximation by Walsh functions and Haar–Vilenkin

A. H. Siddiqi (✉)
Sharda University, Greater Noida, India

P. Manchanda et al. (eds.), *Mathematical Modelling, Optimization, Analytic and Numerical Solutions*, Industrial and Applied Mathematics,
https://doi.org/10.1007/978-981-15-0928-5_1

system. I also benefited from Prof. William Wade, USA who is a well-known expert in this field. References of my joint work with Moricz and Fridli [14, 39] are given at the end. One can find fairly complete research work up to 1977 in my book of 1978 on Walsh functions, AMU Press. Our book to be published soon by Springer Nature named "Construction of Wavelets through Walsh Functions" provides very recent work by Farkov and research group of Manchanda, especially Meenakshi. These are the relevant Refs. [40, 56, 59, 72].

ii. Two-Dimensional Analogues of Banach Spaces: The concept of two-dimensional analogues of metric normed and Banach spaces were introduced and studied by Siegfried Gaehler of the German Academy of Science, Berlin around 1963. I introduced this topic to researchers in India, Iran, and Algeria. I also spent quite some time with Dr. Gaehler in Berlin. I wrote several research papers on non-Archimedean (ultra) 2-metric and 2-normed spaces quasi-normed spaces, 2-semi-inner product spaces, orthogonally in 2-normed spaces, ultra m-metric, and non-Archimedean m-normed spaces. Some typical results obtained by me and my collaborators will be given in the next section. However a list of papers on theme is given in the bibliography, see for example [20, 72].

iii. Non-Archimedean functional analysis: Besides the concept of compact operators and a fixed point in non-Archimedean functional analysis, we studied the concept of invariant means in non-Archimedean functional analysis which is published in Springer lecture notes in Math vol. 399, Springer, Berlin, 1974. Prof. Grande Kimpe of Belgium, a well-known expert of the field visited AMU, Constantine University, Algeria and also invited me to interact on research problems in non-Archimedean functional analysis. Several Ph.D. students of AMU worked under supervision of Prof. Siddiqi on different topics of this field.

iv. Variational Inequalities: Variational inequalities introduced and studied by celebrated mathematicians J. L. Lions, G. Stampacchia, and G. Fichera in the early 60s were not known in the Indian subcontinent till the 80s. I tried to popularize the subject in the Indian subcontinent and guided successfully 10 research scholars in this field. Some of them are well-known mathematicians of the present time. Four of these research scholars are full professors in the Aligarh Muslim University (AMU); eight of their research scholars are attending this conference. I studied applications of variational inequalities in areas like superconductivity, elasticity (rigid punch problem), and American option pricing references of our research papers in these areas which are given at the end, see for example [3–7, 11, 17, 18, 25, 47–49, 52, 61, 62, 69–71, 84–90].

v. Wavelet Theory: You will be surprised to know that wavelet theory was introduced to me by Helmut Neunzert while driving me from Kaiserslautern to Darmstadt in June 1990. He tried to introduce this subject to his Ph.D. students in the early 90s and other Indian researchers particularly, Prof. Manchanda. I joined KFUPM, Saudi Arabia in 1998 and devoted my full time to promote the study of wavelet theory and its applications. Some of my efforts yielded in publications of two special volumes of Arabian Journal of Science and Engineering, see references of edited volumes [58]. I published several research papers with my research collaborators and completed five research projects related to this theme. See Refs. [2, 31, 94] and edited volumes [35, 83, 91] and authored books [54, 64, 96] for details. The following are additional Refs. [12, 13, 15, 28–30, 38, 73, 80–82, 93].

vi. Fractals: I was motivated by a distinguished electrical engineering professor (Prof. M. N. Faruqi, Former Deputy Director, IIT, Kharagpur and Vice Chancellor, AMU) in 1993 to take up the study of fractals and their applications in image processing. I guided research in this area and one of my Ph.D. students (Mr. Aiman Mukheimar) worked in this area got a Ph.D. degree and is now HoD of Applied Sciences in the Prince Sultan University, Riyadh, Saudi Arabia. I have published a good number of papers given in the bibliography, for example, application of fractal methods in metrological studies some leading scientists have cited this work [42, 43]. see also Refs. [32, 58].

vii. Industrial and Financial Mathematics: Writings of Prof. H. Neunzert inspired me to take up the study of Industrial and applied mathematics. I made serious efforts to initiate teaching and research on industrial mathematics in India. We established the Indian Society of Industrial and Applied Mathematics (ISIAM) with the help of senior academicians of our country such as Prof. J. N Kapur, Prof. H. P. Dixit, Prof. U. P. Singh, Prof. D. Sinha, Prof. V. P. Saxena, Prof. G. C. Sharma, Prof. N. K. Gupta, Prof. Bhola Ishwar, Prof. Karmeshu, Prof. O. P. Bhutani, and Prof. Rudraiah. We made serious efforts for organizing Biennial Conferences of the society and International Conference since 1992. The proceedings of these conferences are published by reputed publishers. The society is publishing its journal named Indian Journal of Industrial and Applied Mathematics. Researchers from all over the world are associated with this journal. Famous publisher Springer, now Springer Nature has started publishing a series called Industrial and Applied Mathematics Series jointly with the society. Some useful references are [9, 10, 19, 33, 34, 37, 55, 57, 60, 66, 95].

viii. Oil Exploration and Wavelets: I worked as a consultant in research projects of the largest oil companies in the world named ARMACO. I also completed research projects on applications of wavelets to meteorological data of the kingdom of Saudi Arabia. See [42, 43] and edited volumes [53, 58].

ix. Wavelets Inverse Problems and Medical Signals: I have worked on the following themes with my research scholars in Sharda University, Greater Noida, for example, Ruchira Aneja, Nagma Irfan, Noor-E- Zahra have already got their Ph.D. degree in 2016. Other researchers with whom I have collaborated are Vivek Singh Bhadouria, Shafali Pande, Amita Garg, Padmesh Tripathi, and Nitender Kumar Shukla. Themes pursued are scalar tomography, vector value tomography, seismic tomography ANFIS, SVM and neuro-fuzzy methods, MRI, inverse problems related to heat equation, etc., see Refs. [35, 96] and research papers [21–24, 26, 27, 44–46, 75, 76, 92, 97].

1.3 Ruchira Aneja—Variants of Wavelet in Medical Imaging

This section is based on [46]: Ruchira Aneja and A. H. Siddiqi, **A Hybrid Shearlet-based compression coefficients and ROI Detection, Journal of Medical Imaging and Health Informatics, USA**.

Need For Geometric Transformations

The need to understand geometric structures arises since it is essential to efficiently analyze and process the data. Data are highly correlated and it is essential to extract the relevant information. This relevant information can be extracted and can be grouped into a certain class if we have an understanding of its dominant features, which are associated with their geometric properties. For instance, edges in natural images. One major goal of applied harmonic analysis is constructing classes of analyzing elements that capture the most relevant information in a certain data class.

Properties of shearlets Shearlets are well localized; they exhibit high directional sensitivity and satisfy parabolic scaling; they are spatially localized and optimally sparse.

Shearlet system is a special case of Composite Wavelet systems which provide optimally sparse representation for a large class of bivariate functions. A Composite Wavelet system in dimension $n = 2$ is

$$\Psi_{j,k} = |det A|^{i/2}\Psi(B^j A^i - k) : i, j \in Z, K \in Z^2$$

$$\text{where } A = \begin{pmatrix} 2 & 0 \\ 0 & \sqrt{2} \end{pmatrix} \text{ and } B = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

$$G = (M, t) : M \in D_{a,t} \in R^2$$

where for each $0 < \alpha < 1$, $D_\alpha < L_2(R)$ is a set of matrices

$$D_\alpha - \{M - M_{\alpha s} - \begin{pmatrix} a & -a^\alpha s \\ 0 & a^\alpha \end{pmatrix}\}\text{where } a > 0, s \in R.$$

Continuous Shearlet Transform For $\Psi \in L^2(R^2)$, the continuous shearlet system $SH(\Psi)$ is define by

$$SH(\Psi) = \Psi_{a,s,t} = T_t D_{A_\alpha} D_{S_s}\Psi : a > 0, s \in R, t \in R^2.$$

$$A_a = \begin{pmatrix} a & 0 \\ 0 & a^{1/2} \end{pmatrix}.$$

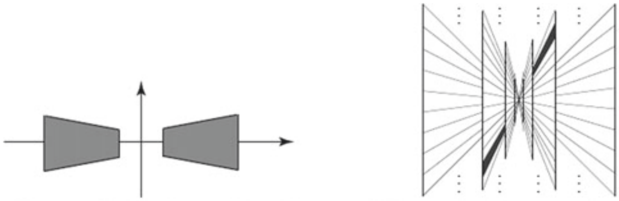
Shearing operator is D_{S_s}, $s \in R$, where the shearing matrix S_s is given by parabolic scaling

$$S_s = \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix}.$$

T_t is the translation operator on $L^2(R^d)$, defined by

$$T_t\Psi(x) = \Psi(x - t), \text{for } t \in R^d.$$

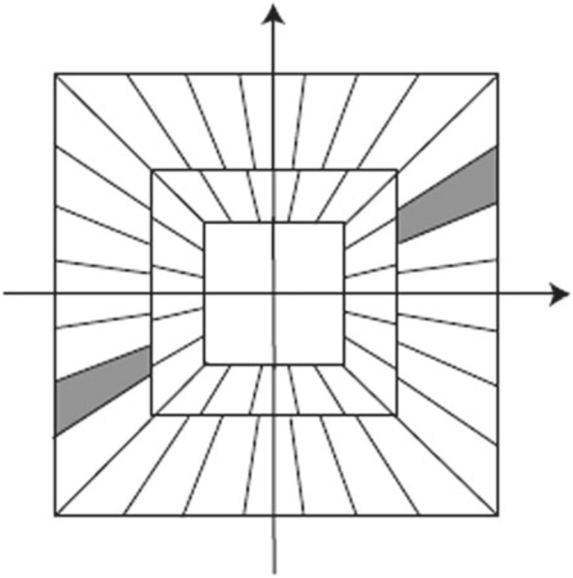
Discrete Shearlet Transform Discrete versions of shearlet systems can be constructed by appropriate sampling of the continuous parameter set S or Scone. Various approaches have been suggested, aiming for discrete shearlet systems which preferably form an orthonormal basis or a tight frame for $L^2(R^2)$ A (regular) discrete shearlet system associated with Ψ, denoted by $SH(\Psi)$, which is defined by



(a) Support of the Fourier transform of a classical shearlet.

(b) Fourier domain support of several elements of the shearlet system, for different values of a and s.

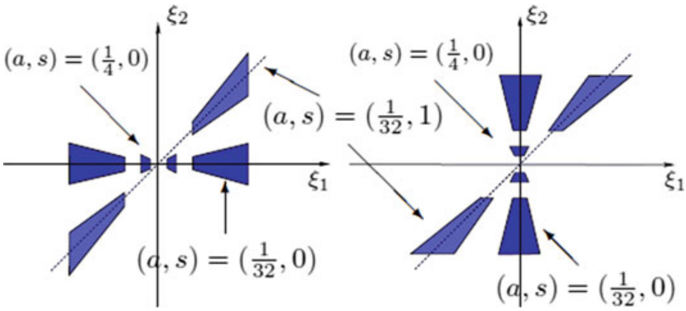
$$SH(\Psi) = \Psi_{j,k,m} = 2^{\frac{3j}{4}}\Psi(S_k A_{2^j} - m) : j, k \in Z, m \in Z^2.$$



Shearlet systems are derived from a single or finite set of generators. It also ensures a unified treatment of the continuum and digital world due to the fact that the shear matrix preserves the integer lattice.

Shearlet and Curvelet in Image Processing

Digital signal and image processing is the most important technique to analyze, manipulate, and process real-world data and images. These types of signals and images may be time series, collection of numbers or measured values. These include audio signals, video images, real-world data like seismic data, rainfall data, biomedical data, and images. Edges are prominent features in images and their analysis and detection are an essential goal in computer vision and image processing. Indeed, identifying and localizing edges are a low-level task in a variety of applications such as 3D reconstruction, shape recognition, image compression, enhancement, and restoration.


ξ_2
$(a, s) = (\frac{1}{4}, 0)$
$(a, s) = (\frac{1}{32}, 1)$
ξ_1
$(a, s) = (\frac{1}{32}, 0)$
ξ_2
$(a, s) = (\frac{1}{4}, 0)$
ξ_1
$(a, s) = (\frac{1}{32}, 0)$

Shearlet and curvelet are a novel directional multiscale mathematical framework which is especially adapted for identification and analysis of distributed discontinuities such as edges occurring in natural images. Multiscale methods based on wavelets have been successfully applied to the analysis and detection of edges. Despite their success, wavelets are however known to have a limited capability in dealing with directional information. The shearlet and curvelet approach is particularly designed to deal with directional and anisotropic features typically present in natural images, and has the ability to accurately and efficiently capture the geometric information on edges. As a result, the shearlet framework provides highly competitive algorithms, for detecting both the location and orientation of edges, and for extracting and classifying basic edge features such as corners and junctions. The shearlet framework provides a unique combination of mathematical rigidness and computational efficiency when addressing edges, optimal efficiency in dealing with edges, and computational efficiency.

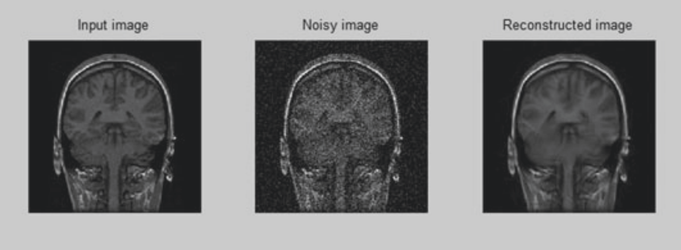
Shearlet in MRI of Brain

Image denoising is a process of recovering the original image from the image corrupted with various types of noise such as Gaussian, speckle, salt and pepper, impulse, etc. Shearlets can be used effectively for image denoising by using various shrinkage rules. The main steps of image denoising are

Table 1.1 Results cont...

MRI brain image	Gaussian noise		
Sigma	Noisy PSNR	PSNR	MSE
10	28.11	34.82	21.45
15	24.61	32.98	32.71
20	22.10	31.72	43.75
25	20.21	30.69	55.51

1. Compute shearlet transform of the noisy image.
2. Apply hard/ soft threshold to the obtained shearlet coefficients.
3. The thresholded shearlet coefficients are subjected to reconstruction to recover the original image (Table 1.1).



1.4 Wavelet Cross-Correlation

Two signals are said to be correlated if they are linearly associated, i.e., if their wavelet spectrum at a certain scale or wavelength is linearly associated.

Broadly speaking, graphs of a vs E(a) for two signals are similar (increase or decrease together).

Wavelet Spectrum Wavelet spectrum E(a) is defined as

$$E(a) = \frac{1}{C_g} \int_{-\infty}^{\infty} W(a,b)^2 db.$$

Wavelet spectrum defines the energy of wavelet coefficients for scale $'a'$.

Peaks in $E(a)$ highlight the dominant energetic scales within the signal.

For details see edited volume [53]. See also authored books [1, 8, 41, 50, 51, 54, 63, 64, 77, 79, 96] and edited volumes [16, 35, 36, 53, 65, 67, 74, 78, 78, 83, 91] for more recent developments in areas mentioned above.

1.5 Fractal Dimension and Predictability

$$D = 2 - H.$$

If the fractal dimension D for the time series is 1.5, there is no correlation between amplitude changes corresponding to two successive time intervals. Therefore, no trend in amplitude can be discerned from the time series and hence the process is unpredictable. However, as the fractal dimension decreases, to 1, the process becomes more and more predictable as it exhibits “persistence”.

Predictably indices (denoted by PI_T, PI_P, and PI_R respectively) for temperature, pressure, and precipitation are defined as follows.

$$PI_T = 2|D_T - 1.5|;\ PI_P = 2|D_P - 1.5|,\ PI_R = 2|D_R - 1.5|.$$

Concepts of fractal and multifractal and their relevance to real-world systems were introduced by Benoit B. Mandeldort, for updated references and interesting introduction of the theme we refer to Benoit B. Mandelbort and Richard L. Hudson, 2004.

In many real-world systems, represented by time series, understanding of the pattern of singularities that is a graph of points at which time-series changes abruptly is quite a challenging task. The time series of rainfall data are usually fractal or multifractal. For details see Refs. [42, 43] .

References

1. K. Adzievski, A.H. Siddiqi, *Introduction to Partial Differential Equations for Scientist and Engineers using Mathematica* (Chapman and Hall/CRC Press, Taylor and Francis Group, U.S., 2013). ISBN 9781466510562
2. M.K. Ahmad, A.H. Siddiqi, Image classification and comparative study of compression techniques. J. Sampl. Theory Signal Image Process. 152–180 (2002)
3. R. Ahmad, A.H. Siddiqi, S. Irfan, J. Inequalities Appl. 515–524 (2007)
4. M.K. Ahmad, A.H. Siddiqi, New distortion measure in image processing. J. Sampl. Theory Signal Image Process. **4**, 151–167 (2005)
5. R. Ahmad, A.H. Siddiqi, Z. Khan, Proximal point algorithm for genralization multivalued nonlinear quasi-variational like inclusions in Banach Spaces. Appl. Math. Comput. **163**, 295–308 (2005)
6. Q.H. Ansari, A.H. Siddiqi, S.Y. Wu, Existence and duality of generalized vector equilibrium problems. J. Math. Anal. Appl. **259**, 115–126 (2001)
7. Q.H. Ansari, Z. Khan, A.H. Siddiqi, Weighted variational inequalities. J. Optim. Theory Appl. **127**, 263–283 (2005)
8. M. Brokate, P. Manchanda, A.H. Siddiqi, *Calculus with Applications* (Springer Nature, Berlin, 2019). (In Press)
9. M. Brokate, A.H. Siddiqi (eds.), *Functional Analysis with Current Applications in Science, Technology and Industry*. Pitman Research Note in Mathematics series, vol. 377 (Addison Wesley Longman, U.K. 1997). 2nd edn. 2015, ISBN 978-0582312609
10. M. Brokate, A.H. Siddiqi, *Functional Analysis with Current Applications in Science, Technology and Industry*. Pitman Research Note in Mathematics series, vol. 377 (Addison Wesley Longman, U.K, 1997). 2nd edn. 2015, ISBN 978-0582312609
11. M. Brokate, A.H. Siddiqi, Sensitivity in the rigid punch problem. Adv. Math. Sci. Appl. **2**, 445–456 (1993)
12. Z. Can, Z. Aslan, O. Oguz, A.H. Siddiqi, Wavelet transform of metrological parameters and gravity waves. Ann. Geophys. **23**, 659–663 (2005)

13. M. El Gebeily, A.H. Siddiqi, Wavelet packets for saudi meteorological time series analysis, in *Proceedings of the International Workshop III: on Applications of Wavelets to Real World Problems*, pp. 1–21 (2008)
14. S. Fridli, P. Manchanda, A.H. Siddiqi, Approximation by walsh-norlud means. Acta Sci. Math. (Szeged) **74**, 593–608(2008)
15. K.M. Furati, P. Manchanda, A.H. Siddiqi, Wavelet methods in oil industry, in *Proceedings of the International Workshop III on Applications of Wavelets to Real World Problems*, pp. 26–36 (2008)
16. K.M. Furati, Z. Nashed, A.H. Siddiqi, *Mathematical Models and Method for Real Worlds System* (Marcel-Dekkar/Taylor and San/ Francis/ Chapman/CRC, New-York, 2005). ISBN 9780849337437
17. K.M. Furati, A.H. Siddiqi, Quazi-variational inequality modeling problems in superconductivity. Numer. Funct. Anal. Optim. **26**, 193–204 (2005)
18. K.M. Furati, A.H. Siddiqi, Fast algorithm for the bean critical model for superconductivity. Numer. Funct. Anal. Optim. **26**, 177–192 (2005)
19. K.M. Furati, H.Z. Tawfiq, A.H. Siddiqi, Simulation and Visulization of safing sensor by fastflo. Am. J. Appl. Sci. **2**, 1261–1265 (2005)
20. S. Gaehler, A.H. Siddiqi, S.C. Gupta, Contribution to non-archimedean functional analysis. Math. Nachr **69**, 163–171 (1975)
21. N. Irfan, A.H. Siddiqi, A novel computational and stable hybrid approach in solving Hankel transform. Appl. Math. Comput. **281**, 121–129 (2016)
22. N. Irfan, A.H. Siddiqi, A wavelet algorithm for Fourier-Bessel transform arising in optics. Int. J. Eng. Math. **789675**, 1–10 (2015)
23. N. Irfan, A.H. Siddiqi, Application of CAS wavelets in Numerical Evaluation of Hankel transform occurring in seismology. In: Mathematical Models, Methods and Applications. Industrial and Applied Mathematics series of Springer (2015), pp. 285–298
24. N. Irfan, A.H. Siddiqi, Sine-cosine wavelets approach in numerical evaluation of Hankel transform. Appl. Math. Model. **40**, 4900–4907 (2016)
25. A. Khaliq, A.H. Siddiqi, S. Krishnan, Some existence result for generalized vector quazi-variational on equalities. Nonlinear Funct. Anal. Appl. **10**, 415–425 (2005)
26. N. Kumar, K. Alam, A. Hasan Siddiqi, Savitzky-golay and wavelet transform based rama spectroscopic data denoising. IJCTA, Int. Sci. Press **9**, 297–305 (2016)
27. N. Kumar, K. Alam, A.H. Siddiqi, Wavelet transform for classification of EEG signal using SVM and ANN. Biomed. Pharmacol. J. [manuscript no. is BPJ-1279] Accepted (2016)
28. P. Manchanda, Meenakshi, A.H Siddiqi, Haar-Vilenkin wavelet, Bull. Ali **27**, 59–73 (2008)
29. P. Manchanda, Meenakshi, A.H. Siddiqi, Construction of vector-valued nonuniform wavelets and wavelet packets. Indian J. Ind. Appl. Math. **4**, 105–125 (2013)
30. P. Manchanda, Meenakshi, A.H. Siddiqi, Wavelet associated with non uniform multi resolution analysis, on positive half-line. Int. J. Wavelets Multi Resolut. Anal. Inf. Process. (Word Scientific publishing company, Singapore) **10**, 1–27 (2012)
31. P. Manchanda, Mukheimer, A.S. Aiman, A.H. Siddiqi, Point-wise convergence of wavelet expansions associated with dilation matrix. Appl. Anal. **76**, 301–308 (2000)
32. P. Manchanda, Mukheimer, A.S. Aimen, A.H. Siddiqi, Certain Results concerning the Iterative function system. Numer. Funct. Anal. Optim. **21**, 217–225 (2000)
33. P. Manchanda, K. Ahmad, A.H. Siddiqi, *Recent Trends in Industrial and Applied Mathematics* (Anamaya Publishers, New-Delhi/London, 2002). ISBN 13: 978-1-4613-7967-6
34. P. Manchanda, R. Lozi, A.H. Siddiqi (eds.), *Mathematical Modelling, Optimization, Analysis and Numerical solutions* (Springer Nature, Berlin, 2019). (In Press)
35. P. Manchanda, R. Lozi, A.H. Siddiqi, *Industrial Mathematics and Complex Systems Emerging Mathematical Models, Methods and Algorithms* (Springer Nature, Berlin, 2017). ISBN 978-981-10-3758-0
36. P. Manchanda, R. Lozi, A.H. Siddiqi, *Mathematical Modeling, Optimization, Analytic and Numerical Solutions* (Springer Nature, Berlin, 2019). (In Process)

37. P. Manchanda, J. Kumar, A.H. Siddiqi, Mathematical methods for modelling price fluctuations of financial time series. J. Frankl. Inst. **344**, 613–636 (2007)
38. Meenakshi, P. Manchanda, A.H. Siddiqi, Wavelets associated with vector-valued nonuniform multiresolution analysis. Appl. Anal. **93**, 84–104 (2014)
39. F. Moricz, A.H. Siddiqi, Approximation by Norlund means of Walsh-Fourier series. J. Appr. 2 Theory **70**, 375–389 (1992)
40. F. Moricz, A.H. Siddiqi, A quantified version of the Dirichlet-Jordan test in L'-norm. Rend. Circ. Mat. Palermo **45**, 19–24 (1996)
41. H. Neunzert, A.H. Siddiqi, *Topics in Industrial Mathematics, Case Studies and Related Mathematical Methods* (Kluwer Academic Publishers, Boston, 2000). (Now Springer), PAPER BAK 2010 electronic version also available, ISBN 978-1-4757-3222-1
42. S. Rehman, A.H. Siddiqi, Wavelet based correlation coefficient of time series of Saudi Meteorological data. Chaos Solitions Fractals (Elsevier) **39**, 1764–1789 (2009)
43. S. Rehman, A.H. Siddiqi, Wavelet based Hurst exponent and fractal dimensional analysis of Saudi climatic dynamics. Chaos Solitions Fractals (Elsevier) **40**, 1081–1090 (2009)
44. A. Ruchira, A.H. Siddiqi, Comparative analysis of image denoising of biomedical images using modified fast shearlet transform and discrete shearlet transform. Int. J. Comput. Sci. Inf. Technol. Res. Excel. (IJCSITRE) **4**, 1–6 (2014)
45. A. Ruchira, A.H. Siddiqi, Comparative analysis of image Denoising of biomedical images using wavelet, curvelet and shearlet transform. Int. J. Appl. Eng. Res. **10**, 36881–26890 (2015)
46. A. Ruchira, A.H. Siddiqi, Hybrid image compression using shearlet coefficients and region of interest detection. J. Med. Imaging Health Inform. **6**(2), 506–517 (2016)
47. M.K. Salahuddin, Ahmad, A.H. Siddiqi, Parametric problem of completely generalized quasi-variational inequalities. J. Inequalities 1–12 (2006)
48. Salahuddin, M.K. Ahmad, A.H. Siddiqi, Existence results for generalized nonlinear variational inclusion. Appl. Math. Lett. **18**, 859–864 (2005)
49. A.H. Siddiqi, *Applicable Mathematics* (MACMILLAN India Limited, 1994)
50. A.H. Siddiqi, *Applied Functional Analysis* (Anamaya Publisher, 2010). ISBN 978-981-10-3725-2
51. A.H. Siddiqi, *Applied Functional Analysis* (Marcel Dekker, New York, 2004). ISBN 9780203913017
52. A.H. Siddiqi, Certain current developments in Variational inequalities, Topological vector spaces, algebras and related areas. Pitman Res. Notes Math. Ser., 316 Longman Sci. The Harlow 219–238 (1994)
53. A.H. Siddiqi, *Emerging Application of Wavelet Methods*, vol. 1463 (American Institute of Physics (AIP), USA, 2012). ISBN 978-0-7354-1067-1
54. A.H. Siddiqi, *Functional Analysis and Applications* (Springer Nature, Berlin, 2018). ISBN 978-981-10-3725-2
55. A.H. Siddiqi, *Introduction and Functional Analysis with Applications* (Tata Mc Graw Hill, New York, 1987). Under UGC book writing project. ISBN 9781904798910, 1904798918
56. A.H. Siddiqi, On the summability of a sequence of Walsh functions. J. Austral. Math. Soc. **10**, 385–394(1969)
57. A.H. Siddiqi, Role of mathematics for economics and scientific development, in *Proceedings of the 2nd Conference on Planning and Development of Education and Scientific Research in the Arab States*, 24–27 Feb 2008 at KFUM, Dhahran, Saudi Arabia (Invited Paper), pp. 987–992 (peer Reviewed Paper)
58. A.H. Siddiqi, Theme issues on "Wavelet and Fractal Methods for Science and Engineering". Arab. J. Sci. Eng. **28**(1C) (2003) and **29**(2C) 2004 (KFUPM)
59. A.H. Siddiqi, *Walsh Function* (AMU Press, Aligarh, 1978). (UGC grant publication)
60. A.H. Siddiqi, K. Ahmad, *Mathematics and Its Applications in Industry and Business* (Narosa, New Delhi and London, 2000)
61. A.H. Siddiqi, R. Ahmad, Mixed variational like inclusion and proximal operator equation in Banach spaces. J. Math. Anal. Appl. **265**, 515–524 (2007)

62. A.H. Siddiqi, R. Ahmad, S.S. Irfan, Set-valud variational inclusions with fuzzy mappings in Banach Spaces. J. Concr. Appl. Math. **4**, 171–181 (2006)
63. A.H. Siddiqi, K. Ahmad, P. Manchanda, *Introduction to Functional Analysis* (Anamaya/Anshan, New Delhi/London, 2006). ISBN 9781904798910
64. A.H. Siddiqi, M. Al-Lawati, M. Boulbrachene, *Modern Engineering Mathematics*. (CRC Press Taylor and Francis Group A Chapman and Hall Book, 2018). ISBN 9781498712057
65. A.H. Siddiqi, I. Duff, O. Christensen (eds.), *Modern Mathematical Methods* (Model and Algorithms, Anamaya/Anshan, New Delhi, London, 2007). ISBN 978-1905740123
66. A.H. Siddiqi, G.D.V. Gowda, R.C. Singh (eds.), *Recent Developments in Computational Science* (Taylor and Francis, Milton Park, 2019). (In Process)
67. A.H. Siddiqi, G.D.V. Gowda, R.C. Singh, *Computational Science and Applications* (CRC Press Taylor and Francis Group, Boca Raton, 2019) (in Press)
68. A.H. Siddiqi, S.C. Gupta, A. Siddiqi, On semi m-normed spaces. Indian J. Math. **30**, 49–55 (1988)
69. A.H. Siddiqi, S. Hussain, K.R. Kazmi, Generalized mixed nonlinear quasi-variational inequalities in Hilbert spaces. Mem. Fac. Sci, Kochi Univ. Ser. A **16**, 7–12 (1995)
70. A.H. Siddiqi, S. Irfan, Completely generalized Co-complementary problems Involving P-Related Accretive operators with Fuzzy Mappings. Accepted for Publication in Professor George Isac (Canada) Memorial Volume, Springer, New York. P. Pardalos (USA), T.M. Rassias (Greece), A. Khan (USA) (eds.) (2009)
71. A.H. Siddiqi, A. Khaliq, Q.H. Ansari, On variational -like-inequalities. Ann. Sci. Math. Quebec **18**, 95–104
72. A.H. Siddiqi, H.H. Khan, J. Ahmed Siddiqi and his contribution, The world of mathematics. Aligarh Bull. Math. **27**, 1–6 (2008)
73. A.H. Siddiqi, S. Khan, S. Rehman, Wind speed simulation using wavelet. Am. J. Appl. Sci. **2**, 557–564 (2004)
74. A.H. Siddiqi, M. Kocvara, *Trends in Industrial and Applied Mathematics, International Conference Proceedings* (Kluwer Academic Publishers (Now Springer), Boston, 2002). ISBN 1-4020-0751-5
75. A.H. Siddiqi, N. Kumar, K. Alam, Raman spectra denoising using wavelet and wavelet packet transform. Indian J. Ind. Appl. Math. (ISIAM) **7**, 1–10 (2016)
76. A.H. Siddiqi, N. Kumar, K. Alam, Raman spectra de-noising using wavelet packet thresholding. IJCTA, Int. Sci. Press **9**, 325–334 (2016)
77. A.H. Siddiqi, P. Manchanda, *Introduction to Differential Equations* (Macmillan India, New-Delhi/Banglore, 2006). ISBN 9781403930187
78. A.H. Siddiqi, P. Manchanda, R. Bhardwaj, *Mathematical Methods, Models and Applications* (Springer, Berlin, 2015). ISBN 978-981-287-973-8
79. A.H. Siddiqi, P. Manchanda, M. Brokate, *Calculus with Applications Written Under a Project of ICTP* (Italy, IK International publisher, Trieste, 2011). ISBN 978-93-81141-32-38
80. A.H. Siddiqi, P. Manchanda, M. Kocvara, Fast wavelet-based algorithms for option pricing, in *Proceedings the 6th Multi Conference on Systematic, Cybernetics and Informatics*, vol. 13 (2002), pp. 140–146
81. A.H. Siddiqi, H. Sevindir, Z. Aslam, A wavelet based energetic approach for the analysis of electroencephalogram, Special Volume of Sultan Qaboos University General of Science, 2012 (In Press, SQUJS-22SVR)
82. A.H. Siddiqi, F.A. shah, Construction of wavelet packets with multi resolution analysis. Aligarh Bull. Math. **27**, 13–22(2008)
83. A.H. Siddiqi, R.C. Singh, P. Manchanda, *Proceedings of Satellite Conference ICM 2010 on Mathematics in Science and Technology* (World Scientific Publisher, Singapore, 2011). ISBN 13 978-981-4338-81-3
84. A.H. Siddiqi, Rais Ahmad, Ishikawa type iterative algorithm for a generalized nonlinear quasi-variational like inclusions in Banach spaces. Math. Comp. Model. **45**, 594–605 (2007)
85. A.H. Siddiqi, Q.H. Ansari, General strongly nonlinear variational inequalities. J. Math. Anal. Appl **166**, 386–392 (1992)

86. A.H. Siddiqi, P. Manchanda, Certain remarks on a class of evolution quasi variational inequalities. Int. J. Math. Math. Sci. **24**, 851–855 (2000)
87. A.H. Siddiqi, Q.H. Ansari, A. Khaliq, on vector variational inequalities. J. Optim. Theory Appl. **84**, 171–180 (1995)
88. A.H. Siddiqi, M.F. Khan, R. Ahmad, Wiener-Hopf equations and general mildly nonlinear variational inequality. Indian J. Pure Appl. Math. **28**, 1317–1325 (1997)
89. A.H. Siddiqi, P. Manchanda, M. Kocvara, An Iterative two-step algorithm for American option pricing. IMA J. Math. Appl. **11**, 71–84 (2001)
90. A.H. Siddiqi, P. Manchanda, M. Brokate, On some recent developments concerning moreaus sweeping process. Trends Ind. Appl. Math. Appl. Optim. Ser. **72**, 339–354 (2002)
91. A.H. Siddiqi, A.K. Gupta, M. Brokate, *Models of Engineering and Technological Problems* (American Institute of Physics (AIP), USA, 2009). ISBN 978-0-7354-0683-4
92. V. Singh Bhadouria, D. Ghoshal, A.H. Siddiqi, A new approach for high density saturated impulse noise removal using decision-based coupled window median filter. Signal Image Video Process. Springer **21** (2014)
93. M.A. Sohail, A.H. Siddiqi, S. Ispon, A new fast DCT based watermarking technique. Trends Ind. Appl. Math. Appl. Optim. Ser. **72**, 117–117 (2002)
94. M.A. Suhail, A.H. Siddiqi, Parameterization of discrete wavelet transform. an image processing application for multimedia copyright security enhancement using wavelet marking. J. Sampl. Theory Signal Image Process. **4**, 57–72 (2004)
95. M. Yu Rasulova, A.H. Siddiqi, Relationship between the solution of BBGK-Hierachy of kinetic equations and the particle solution of Vlasov equation. J. Dyn. Syst. Geom. Theor. **2**, 17–22 (2004)
96. A. Yu Farkov, P. Manchanda, A.H. Siddiqi, *Construction of Wavelets Through Walsh Function*. (Springer, Berlin, 2019). ISBN 978-981-13-6370-2
97. N.E. Zahra, A.H. Siddiqi, Interaction of vector valued wavelet and vector field tomography. Indian J. Ind. Appl. Math. ISIAM **6**, 120–138 (2015)

Chapter 2
Schrödinger Operators with a Switching Effect

Pavel Exner

Abstract This paper summarizes the contents of a plenary talk given at the 14th Biennial Conference of Indian SIAM in Amritsar in February 2018. We discuss here the effect of an abrupt spectral change for some classes of Schrödinger operators depending on the value of the coupling constant, from below bounded and partly or fully discrete, to the continuous one covering the whole real axis. A prototype of such a behavior can be found in the Smilansky–Solomyak model devised to illustrate that an irreversible behavior is possible even if the heat bath to which the systems are coupled has a finite number of degrees of freedom and analyze several modifications of this model, with regular potentials or a magnetic field, as well as another system in which $x^p y^p$ potential is amended by a negative radially symmetric term. Finally, we also discuss resonance effects in such models.

Keywords Smilansky model · Switching effect · Asymptotic expansions · Magnetic field · Resonances

2.1 Introduction

The class of problems we are going to discuss here has a twofold motivation. Let us start with physics. It is well-known that while the equations of motion governing quantum dynamics are invariant with respect to time reversal; we often encounter quantum systems behaving in an irreversible way, for instance, spontaneous decays of particles and nuclei, inelastic scattering processes in nuclear, atomic or molecular systems, or the current passing through a microscopic element attached to poles of a battery. Furthermore, an irreversible process *par excellence* is, of course, the *wave packet reduction* which is the core of Copenhagen description of a measuring process performed on a quantum system. The description of such a process is typically associated with enlarging the state Hilbert space, conventionally referred to as cou-

P. Exner (✉)
Nuclear Physics Institute, Czech Academy of Sciences, Hlavní 130, 25068 Řež, Prague, Czechia
e-mail: exner@ujf.cas.cz

P. Manchanda et al. (eds.), *Mathematical Modelling, Optimization, Analytic and Numerical Solutions*, Industrial and Applied Mathematics,
https://doi.org/10.1007/978-981-15-0928-5_2

pling the system to a *heat bath*. It is generally accepted that to obtain an irreversible behavior through such a coupling, the model has to exhibit typical properties, in particular

- the bath is a system with an infinite number of degrees of freedom,
- the bath Hamiltonian has a continuous spectrum, and
- the presence (or absence) of irreversible modes is determined by the energies involved rather than the coupling strength.

While this all is without any doubt true in many cases, one of our aims here is to show that *neither of the above need not be true in general*. To make this point, Uzy Smilansky constructed a simple model which will be our starting point here. In a sense he did a similar thing as Agatha Christie: when some people tried to introduce in the 1920s mystery rules saying, in particular, that in any such book there must be a single murderer, she wrote *Murder on the Orient Express* in which everyone is a killer, except for Hercule Poirot, of course.

On the other hand, as a mathematician one may ask whether, in contrast to the usual perturbation theory results [16], *small change of the coupling constant* can have a profound influence on the spectrum. Posed like that the answer is trivial: consider the one-dimensional Schrödinger operator

$$H_\lambda = -\frac{\mathrm{d}^2}{\mathrm{d}x^2} + \lambda x^2\,;$$

it is obvious that for all $\lambda = \omega^2 > 0$, the spectrum of such an operator is *purely discrete*, $\sigma(H_\lambda) = \{(2n+1)\omega : n = 0, 1, \dots\}$ while for $\lambda = 0$ and $\lambda < 0$, we have $\sigma(H_\lambda) = [0, \infty)$ and $\sigma(H_\lambda) = \mathbb{R}$, respectively. A much more subtle question is whether similar things could happen if the potential modification concerns a *small part* of the configuration space, or even a *"set of zero measure"*. Smilansky model and its various modifications we are going to discuss provide an *affirmative answer*.

Let us describe briefly the contents of the paper. In the next section, we summarize the known results about the Smilansky–Solomyak model and present a numerical method to analyze its discrete spectrum. Section 2.3 is devoted to discussion of various modifications of the model consisting, in particular, of replacing the δ interaction "channel" by a regular potential one or, on the contrary, by the more singular δ' interaction, or adding a homogeneous magnetic field. In Sect. 2.4, we discuss another model exhibiting similar behavior, a two-dimensional Schrödinger operator with the potential consisting of the $x^p y^p$ part amended by a negative radially symmetric term. In Sect. 2.5, we return to the original Smilansky–Solomyak model and show that it also has a rich resonance structure. Finally, in conclusion we will mention several open questions.

2.2 Smilansky–Solomyak Model

Let us first describe the model proposed by Uzy Smilansky in [23] which in its simplest form describes a one-dimensional system interacting with a caricature heat bath represented by a single harmonic oscillator. Its mathematical properties and various extensions were subsequently analyzed by Michael Solomyak—let us pay memory to this great mathematician who left us 2 years ago—and coauthors in [6, 7, 19, 21, 24–26] from the spectral point view, the corresponding time evolution was discussed in [13, 14].

With this history of the problem in mind, it is appropriate to speak of the *Smilansky–Solomyak model*. At the same time, it is useful to note that while mathematically it is the same thing, physically there may be two ways in which the system is understood. In the original Smilansky treatment, one considers two one-dimensional systems coupled mutually while Solomyak et al. interpreted it in PDE terms as being described by a two-dimensional Schrödinger operator,

$$H_{\mathrm{Sm}} = -\frac{\partial^2}{\partial x^2} + \frac{1}{2}\left(-\frac{\partial^2}{\partial y^2} + y^2\right) + \lambda y \delta(x), \tag{2.1}$$

on $L^2(\mathbb{R}^2)$; it is easy to see that that one may consider $\lambda \geq 0$ only without loss of generality. We will stick here to the latter interpretation because it opens the way to a wider class of possible generalizations.

Let us summarize the known results about spectral properties of the operator (2.1):

- The existence of a *spectral transition:* if $|\lambda| > \sqrt{2}$ the particle can escape to infinity along the singular "channel" in the y direction. In spectral terms, this corresponds to the switch from a positive spectrum to a below unbounded one at $|\lambda| = \sqrt{2}$. At the heuristic level, the *mechanism* of this spectral transition is easy to understand: we have an effective variable decoupling far from the x-axis and the oscillator potential competes there with the δ interaction eigenvalue $-\frac{1}{4}\lambda^2 y^2$.
- The *eigenvalue absence:* for any $\lambda \geq 0$ there are *no eigenvalues* $\geq \frac{1}{2}$. If $|\lambda| > \sqrt{2}$, the point spectrum of H_{Sm} is *empty*.
- The *existence of eigenvalues:* in the subcritical case, $0 < |\lambda| < \sqrt{2}$, we have $H_{\mathrm{Sm}} \geq 0$. The point spectrum is then non-empty and finite, and

$$N(\tfrac{1}{2}, H_{\mathrm{Sm}}) \sim \tfrac{1}{4\sqrt{2(\mu(\lambda)-1)}} \tag{2.2}$$

  holds as $\lambda \to \sqrt{2}-$, where $\mu(\lambda) := \sqrt{2}/\lambda$.
- The *absolute continuity:* in the supercritical case, $|\lambda| > \sqrt{2}$, we have $\sigma(H_{\mathrm{Sm}}) = \sigma_{\mathrm{ac}}(H_{\mathrm{Sm}}) = \mathbb{R}$.

We are not going to give proofs of these claims referring to the papers quoted above, instead we will show how the discrete spectrum can be found *numerically* following [10] which can provide additional insights. At the time, however, the method we use, rephrasing the task as a *spectral problem for Jacobi matrices*, is the core of the proofs done by Solomyak et al. providing thus a feeling of what is the technique involved.

In the halfplanes $\pm x > 0$ the wave functions can be expanded using the "transverse" base spanned by the functions

$$\psi_n(y) = \frac{1}{\sqrt{2^n n! \sqrt{\pi}}} \, \mathrm{e}^{-y^2/2} H_n(y) \tag{2.3}$$

corresponding to the oscillator eigenvalues $n + \frac{1}{2}$, $n = 0, 1, 2, \dots$. Furthermore, one can make use of the mirror symmetry with respect to $x = 0$ and divide H_λ into the trivial odd part $H_\lambda^{(-)}$ and the even part $H_\lambda^{(+)}$ which is equivalent to the operator on the halfplane, $L^2(\mathbb{R} \times (0, \infty))$, with the same symbol determined by the boundary condition

$$f_x(0+, y) = \frac{1}{2}\, \alpha y f(0+, y). \tag{2.4}$$

We substitute the Ansatz

$$f(x, y) = \sum_{n=0}^{\infty} c_n \, \mathrm{e}^{-\kappa_n x} \psi_n(y) \tag{2.5}$$

with $\kappa_n := \sqrt{n + \frac{1}{2} - \varepsilon}$ into (2.4); this yields for the sought solution with the energy ε the equation

$$B_\lambda c = 0, \tag{2.6}$$

where c is the coefficient vector and B_λ is the operator in ℓ^2 with the representation

$$(B_\lambda)_{m,n} = \kappa_n \delta_{m,n} + \frac{1}{2} \lambda (\psi_m, y\psi_n). \tag{2.7}$$

It is obvious that the matrix is in fact tridiagonal because

$$(\psi_m, y\psi_n) = \frac{1}{\sqrt{2}} \big(\sqrt{n+1}\, \delta_{m,n+1} + \sqrt{n}\, \delta_{m,n-1} \big). \tag{2.8}$$

To solve Eq. (2.6) numerically, one truncates the matrix (2.7). The size depends on λ, the most difficult is the weakly bound state corresponding to small λ where the truncation size should be of the order of 10^4 to achieve a numerically stable solution. The result is plotted in Fig. 2.1. In coincidence with the theoretical result quoted above, the discrete spectrum is non-empty for nonzero λ. It may seem that it consists of a single eigenvalue but a closer look shows that the second one appears at $\lambda \approx 1.387559$; the next thresholds are 1.405798, 1.410138, 1.41181626, 1.41263669, ... To have a better insight, one can plot the discrete spectrum near the critical coupling in the semilogarithmic scale as shown in Fig. 2.2. We see that in this regime, many eigenvalues appear which gradually fill the interval $(0, \frac{1}{2})$ as the critical value $\lambda = \sqrt{2}$ is

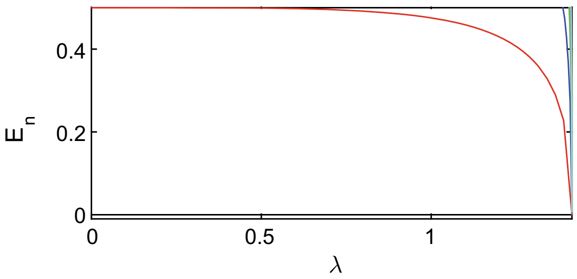


Fig. 2.1 Discrete spectrum of H_{Sm} as a function of the coupling constant λ

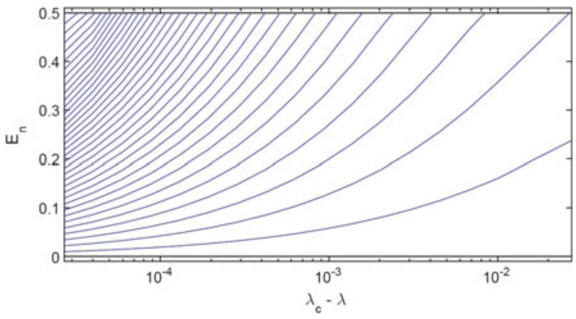


Fig. 2.2 Discrete spectrum of H_{Sm} near the critical value of the coupling constant

approached. Figure 2.3 shows a comparison of their number indicated by dots with the asymptotics (2.2); we see a perfect fit. The numerical solution also indicates other properties. For instance, plotting in Fig. 2.4 the eigenvalue curve for small values of λ in the logarithmic scale, we see that it behaves as $E_1(\lambda) = \frac{1}{2} - c\lambda^4 + o(\lambda^4)$ as $\lambda \to 0$, with $c \approx 0.0156$. In fact, the coefficient value can be found exactly to be $c = 0.015625$. To this aim, we write Eq. (2.6) explicitly in components as

$$\begin{gathered}\sqrt{\mu_\lambda}c_0^\lambda + \frac{\lambda}{2\sqrt{2}}c_1^\lambda = 0, \\ \frac{\sqrt{k}\lambda}{2\sqrt{2}}c_{k-1}^\lambda + \sqrt{k+\mu_\lambda}c_k^\lambda + \frac{\sqrt{k+1}\lambda}{2\sqrt{2}}c_{k+1}^\lambda = 0, \quad k \geq 1\,,\end{gathered} \tag{2.9}$$

where $\mu_\lambda := \frac{1}{2} - E_1(\lambda)$ and $c^\lambda = \{c_0^\lambda, c_1^\lambda, \dots\}$ is the corresponding normalized eigenvector of B_λ. Using the above relations and simple estimates, we get from here

$$\sum_{k=1}^{\infty} |c_k^\lambda|^2 \leq \frac{3}{4}\lambda^2 \quad \text{and} \quad c_0^\lambda = 1 + \mathscr{O}(\lambda^2) \tag{2.10}$$

as $\lambda \to 0+$, hence we have in particular $c_1^\lambda = \frac{\lambda}{2\sqrt{2}} + \mathscr{O}(\lambda^2)$. The first of the above relations then gives $\mu_\lambda = \frac{\lambda^4}{64} + \mathscr{O}(\lambda^5)$ as $\lambda \to 0+$, in other words

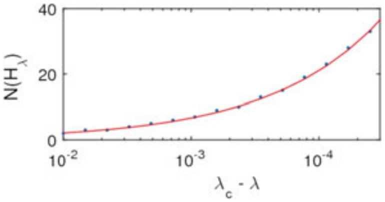


Fig. 2.3 Number of eigenvalues of H_{Sm} versus the asymptotics (2.2)

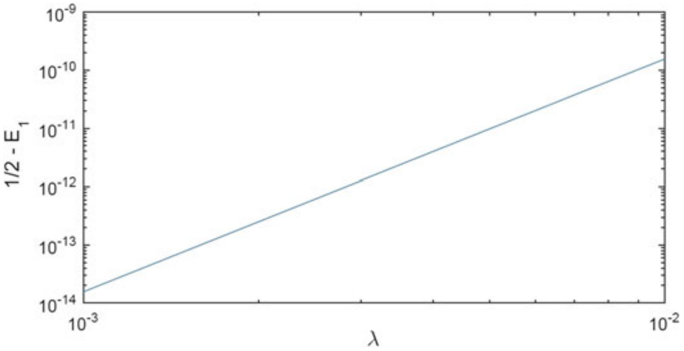


Fig. 2.4 Weak coupling asymptotics of H_{Sm}

$$E_1(\lambda) = \frac{1}{2} - \frac{\lambda^4}{64} + \mathscr{O}(\lambda^5), \tag{2.11}$$

however, the mentioned coefficient 0.015625 is nothing else than $\frac{1}{64}$. Furthermore with the knowledge of the solution to (2.6), we can return to (2.5) and compute the eigenfunctions. In Fig. 2.5, we plot them for a value close to the critical one, namely $\lambda = \sqrt{2} - 0.0086105$. As expected, they are stretched along the axis of the oscillator "channel" and the part of the y-axis where the singular interaction is attractive; the curves on the left side show the y-cuts of the eigenfunctions. The ground state has no zeros and the number of the nodal lines of the nth eigenfunction, $n = 1, 2, \dots,$ is $[\frac{1}{2}n]$, thus only the first excited state is Courant sharp.

2.3 Variations on the Model

We have mentioned in the introduction that various extensions of the Smilansky–Solomyak model described above had been worked out, for instance, using a "heat bath" consisting of more than one oscillator, replacing the line by a loop or a graph, etc. We will not discuss them, instead, we will analyze several other modifications.

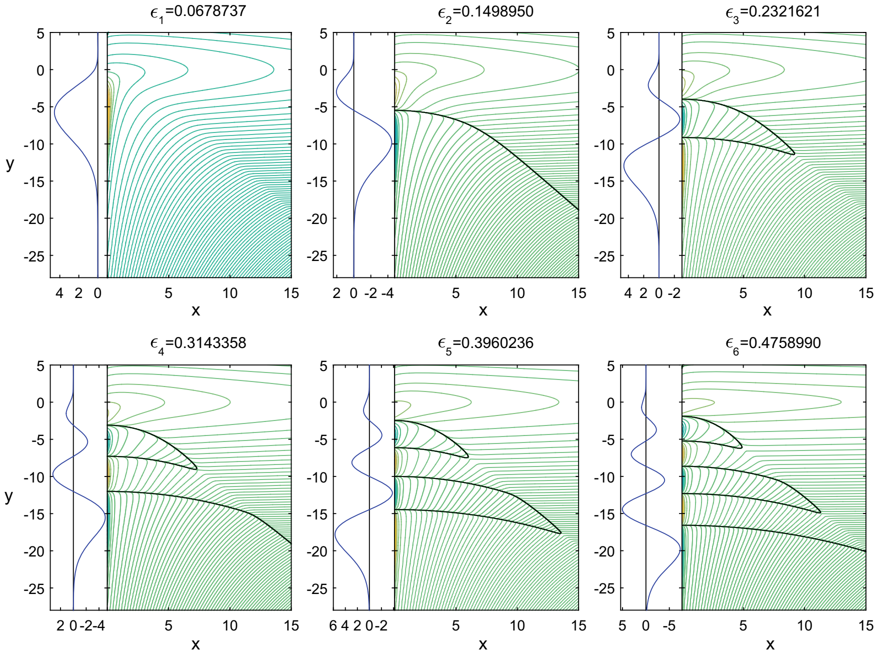


Fig. 2.5 The eigenfunctions of H_{Sm} for $\lambda = 1.4128241$

2.3.1 Regular Version of the Model

The first one is motivated by the question of whether one can observe a similar effect for Schrödinger operators with the δ interaction replaced by a regular potential. It was asked by Italo Guarneri in [13] with a clear motivation: he employed (a modification of) the system described above as a model of the *measuring process* in quantum mechanics in which the supercritical behavior is interpreted as a wave packet reduction. This naturally inspires the question of how the corresponding classical dynamics would look like, and this in turn requires a setting in which the problem can be analyzed in terms of classical mechanics; the first step in this direction has been made in the recent paper [14].

We observe first that the coupling cannot be now linear in y and the profile of the channel has to change with the variable y. We replace the δ by a *family of shrinking potentials*, the mean of which matches the δ coupling constant, $\int U(x, y)\,\mathrm{d}x \sim y$. This can be achieved, e.g., by choosing $U(x, y) = \lambda y^2 V(xy)$ for a fixed function V. This motivates us to investigate the following operator on $L^2(\mathbb{R}^2)$,

$$H = -\frac{\partial^2}{\partial x^2} - \frac{\partial^2}{\partial y^2} + \omega^2 y^2 - \lambda y^2 V(xy), \tag{2.12}$$

where ω, a are positive constants and the potential V is a nonnegative function with bounded first derivative and $\operatorname{supp} V \subset [-a, a]$. By Faris–Lavine theorem H is essentially self-adjoint (e.s.a. in the following) on $C_0^\infty(\mathbb{R}^2)$—cf. [20, Thms. X.28 and X.38]—and the same argument can be applied to various generalizations of the operator (2.12), with more than one "decay" channel, periodicity in the variable x, etc.

To state the result we need a one-dimensional comparison operator $L = L_V$, namely

$$L = -\frac{\mathrm{d}^2}{\mathrm{d}x^2} + \omega^2 - \lambda V(x) \tag{2.13}$$

on $L^2(\mathbb{R})$ with the domain $H^2(\mathbb{R})$. What matters is the sign of its spectral threshold; since $V \geq 0$, the latter is a monotonous function of λ and there is a $\lambda_{\text{crit}} > 0$ at which the sign changes. First of all, we have the following result [2].

Theorem 2.1 *Under the stated assumption, the spectrum of the spectrum of H is bounded from below provided the operator L is positive.*

Sketch of the proof. The claim can be proved using *Neumann bracketing*, imposing additional boundary conditions at the lines $y = \pm \ln n$, $n = 2, 3, \ldots$, and showing that the components of H in these strips have a uniform lower bound by an operator unitarily equivalent to L, cf. [2] for details. □

One the other hand, in the supercritical case when the transverse channel principal eigenvalue dominates over the harmonic oscillator contribution, the spectral behavior changes [2].

Theorem 2.2 *Under our hypotheses, $\sigma(H) = \mathbb{R}$ holds if* $\inf \sigma(L) < 0$.

Sketch of the proof. The argument relies on a choice of an appropriate Weyl sequence: we have to find $\{\psi_k\}_{k=1}^\infty \subset D(H)$ such that $\|\psi_k\| = 1$ which contains no convergent subsequence, and at the same time

$$\|H\psi_k - \mu\psi_k\| \to 0 \quad \text{as} \quad k \to \infty. \tag{2.14}$$

Specifically, we choose

$$\psi_k(x, y) = h(xy)\,\mathrm{e}^{i\varepsilon_\mu(y)}\chi_k\left(\frac{y}{n_k}\right) + \frac{f(xy)}{y^2}\,\mathrm{e}^{i\varepsilon_\mu(y)}\chi_k\left(\frac{y}{n_k}\right),$$

where $\varepsilon_\mu(y) := \int_{\sqrt{|\mu|}}^{y} \sqrt{t^2 + \mu}\,\mathrm{d}t$, h is the normalized ground-state eigenfunction of L, furthermore $f(t) := -\frac{i}{2}\,t^2 h(t)$, and finally, χ_k are suitable, compactly supported mollifier functions, cf. [2] for details. □

The regular version shares also other properties with the original Smilansky–Solomyak model, namely [3]

- in the subcritical case, $\inf\sigma(L) > 0$, we have $\sigma_{\text{ess}}(H) = [\omega, \infty)$ and a non-empty $\sigma_{\text{disc}}(H) \subset [0, \omega)$ and
- in the critical case, $\inf\sigma(L) = 0$, we have $\sigma(H) = \sigma_{\text{ess}}(H) = [0, \infty)$.

2.3.2 Magnetic Version of the Model

One can also consider another modification of the Smilansky–Solomyak model in which the system is placed into a *homogeneous magnetic field* perpendicular to the plane representing the configuration space, described thus by the Hamiltonian

$$H = (i\nabla + A)^2 + \omega^2 y^2 + \lambda y \delta(x), \tag{2.15}$$

where A is a suitable vector potential; note that in this case the original Smilansky interpretation is lost. The spectral properties are similar, in the subcritical case we now have $\sigma_{\text{ess}}(H(A)) = [\sqrt{B^2 + \omega^2}, \infty)$ but again a sufficiently small nonzero λ gives rise to a discrete spectrum which fills the interval $[0, \sqrt{B^2 + \omega^2})$ as $|\lambda|$ approaches the critical coupling 2ω, and above this value the spectrum fills the whole real line. The effect of the magnetic field on the regular version of the model is similar, cf. [4] for details.

2.3.3 The δ' Version of the Model

One can also say that the spectral transition effect is robust. To illustrate this claim, let us consider the version of the model in which the interaction is replaced by a *more singular* one, specifically the one known as δ' [1]. The Hamiltonian then corresponds to the differential expression

$$H_\beta \psi(x, y) = -\frac{\partial^2 \psi}{\partial x^2}(x, y) + \frac{1}{2}\left(-\frac{\partial^2 \psi}{\partial y^2}(x, y) + y^2 \psi(x, y)\right) \tag{2.16}$$

with the domain consisting of $\psi \in H^2((0, \infty) \times \mathbb{R}) \oplus H^2((-\infty, 0) \times \mathbb{R})$ such that

$$\psi(0+, y) - \psi(0-, y) = \frac{\beta}{y}\frac{\partial \psi}{\partial x}(0+, y)\,, \quad \frac{\partial \psi}{\partial x}(0+, y) = \frac{\partial \psi}{\partial x}(0-, y). \tag{2.17}$$

The problem can be treated by a modification of the methods employed in [6, 7, 19, 21, 24–26] leading to the following results which we present referring to [9] for the proofs. Let $\mathfrak{m}_{\text{ac}}$ be the multiplicity of the absolutely continuous spectrum.

Theorem 2.3 *The spectrum of operator H_0 is purely* ac, $\sigma(H_0) = [\frac{1}{2}, \infty)$ *with the multiplicity* $\mathfrak{m}_{\text{ac}}(E, H_0) = 2n$ *for* $E \in (n - \frac{1}{2}, n + \frac{1}{2})$, $n \in \mathbb{N}$. *For* $\beta > 2\sqrt{2}$ *the* ac *spectrum of H_β coincides with $\sigma(H_0)$. For $\beta \le 2\sqrt{2}$ there is a new branch of continuous spectrum added to the spectrum; for $\beta = 2\sqrt{2}$ we have $\sigma(H_\beta) = [0, \infty)$ and for $\beta < 2\sqrt{2}$ the spectrum* covers the whole real line.

We note in passing that with the standard convention used here, *small* values of the parameter β represent a *strong* coupling.

Theorem 2.4 *Assume $\beta \in (2\sqrt{2}, \infty)$, then the discrete spectrum of H_β is* non-empty *and lies in the interval $(0, \frac{1}{2})$. The number of eigenvalues is approximately given by*

$$\frac{1}{4\sqrt{2\left(\frac{\beta}{2\sqrt{2}} - 1\right)}} \quad \text{as} \quad \beta \to 2\sqrt{2}+ .$$

Theorem 2.5 *For large enough β there is a single eigenvalue which asymptotically behaves as*

$$E_1(\beta) = \frac{1}{2} - \frac{4}{\beta^4} + \mathcal{O}\left(\beta^{-5}\right).$$

2.4 Another Model

The Smilansky–Solomyak model is not the only system in which the effect of an abrupt spectral transition can be observed. Now we are going to describe another model in which the transition is even more dramatic as a switch from a *purely discrete spectrum* in the subcritical case to the whole real line in the supercritical one. To begin, recall that there are situations where *Weyl's law fails* and the spectrum is discrete even if the classically allowed phase-space volume is infinite. A classical example of such a situation is due to [22] a two-dimensional Schrödinger operator with the potential $V(x, y) = x^2y^2$, or more generally, $V(x, y) = |xy|^p$ with $p \ge 1$. Similar behavior can also be observed for Dirichlet Laplacians in *regions with hyperbolic cusps*—see [12] for more recent results and a survey; recall also that using the *dimensional-reduction technique* of Laptev and Weidl [17], one can prove tight spectral estimates for such operators.

A common feature of these models is that the particle motion is confined into *channels narrowing toward infinity*; the increasing “steepness” of those “walleys” is responsible for the discreteness of the spectrum. This may remain true even for Schrödinger operators whose potential are *unbounded from below* in which a classical particle can escape to infinity with an increasing velocity. The situation changes, however, if the *attraction is strong enough*; recall that a similar behavior was noted already in [27]. As an illustration, let us thus analyze the following class of operators on $L^2(\mathbb{R}^2)$,

$$L_p(\lambda) \,:\, L_p(\lambda)\psi = -\Delta\psi + \left(|xy|^p - \lambda(x^2 + y^2)^{p/(p+2)}\right)\psi, \quad p \ge 1, \qquad (2.18)$$

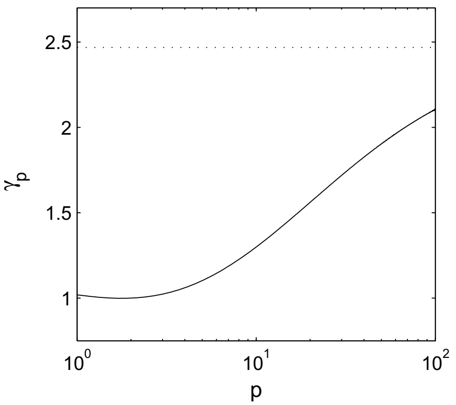


Fig. 2.6 γ_p as a function of p in the semilogarithmic scale

where (x, y) are the standard Cartesian coordinates in $\mathbb{R}^2$ and the parameter λ in the second term of the potential is nonnegative; unless its value is important we write it simply as L_p. Note that $\frac{2p}{p+2} < 2$ so the operator is e.s.a. on $C_0^\infty(\mathbb{R}^2)$ by Faris–Lavine theorem mentioned above; the symbol L_p or $L_p(\lambda)$ will always mean its closure. Needless to say, the power in the last term is chosen in a way that makes it possible to play with the balance between the repulsion coming from the narrowing channels and attraction coming from the negative potential part.

Let us start with the *subcritical case* which occurs for sufficiently small values of λ. To characterize the smallness quantitatively, we need an auxiliary operator which will be an (an)harmonic oscillator Hamiltonian on line,

$$\tilde{H}_p \, : \, \tilde{H}_p u = -u'' + |t|^p u \tag{2.19}$$

on $L^2(\mathbb{R})$ with the standard domain. The principal eigenvalue $\gamma_p = \inf \sigma(H_p)$ equals one for $p = 2$; for $p \to \infty$ it becomes $\gamma_\infty = \frac{1}{4}\pi^2$; it smoothly interpolates between the two values; a numerical solution gives true minimum $\gamma_p \approx 0.998995$ attained at $p \approx 1.788$; in the semilogarithmic scale the plot of γ_p looks as shown in Fig. 2.6

As we have said, the spectrum is bounded from below and discrete if $\lambda = 0$; our first claim [8] is that this remains to be the case provided λ is small enough.

Theorem 2.6 *For any $\lambda \in [0, \lambda_{\text{crit}}]$, where $\lambda_{\text{crit}} := \gamma_p$, the operator $L_p(\lambda)$ is bounded from below for any* $p \geq 1$*; if $\lambda < \gamma_p$ its spectrum is purely discrete.*

Sketch of the proof. Let $\lambda < \gamma_p$. By the minimax principle [20, Sect. XIII.1], we need to estimate L_p from below by a self-adjoint operator with a purely discrete spectrum. To construct it, we employ bracketing imposing additional Neumann conditions at concentric circles of radii $n = 1, 2, \ldots$. In the estimating operators, the variables decouple asymptotically and the spectral behavior is determined by their angular parts; to prove the discreteness one has to check that the lowest eigenvalues in the annuli tend to infinity as $n \to \infty$. For $\lambda = \gamma_p$ this is no longer true but the sequence remains bounded from below. □

A similar argument can be used in the *supercritical case* with a few differences:

- now we seek an *upper* bound to $L_p(\lambda)$ by a below unbounded operator, hence we impose *Dirichlet* conditions on concentric circles;
- the estimating operators have now nonzero contributions from the radial part, however, those are bounded by π^2 independently of n; and
- the negative λ-dependent term now outweights the anharmonic oscillator part so that for the annuli operators $L^{\mathrm{D}}_{n,p}$, we have $\inf \sigma(L^{\mathrm{D}}_{n,p}) \to -\infty$ as $n \to \infty$.

This yields the following conclusion [8]:

Proposition 2.1 *The spectrum of $L_p(\lambda)$, $p \geq 1$, is unbounded below from if $\lambda > \lambda_{\text{crit}}$.*

One can prove a stronger result, however, using a suitable Weyl sequence constructed in a way similar to that employed in the proof of Theorem 2.1, it is possible to make the following conclusion [5].

Theorem 2.7 *$\sigma(L_p(\lambda)) = \mathbb{R}$ holds for any $\lambda > \gamma_p$ and $p > 1$.*

In the subcritical case, one can derive various results concerning properties of the discrete spectrum. Let us first mention an inequality obtained in a variational way for the proof of which we refer to [8]. We define $\alpha := \frac{1}{2}\left(1+\sqrt{5}\right)^2 \approx 5.236 > \gamma_p^{-1}$ and denote by $\{\lambda_{j,p}\}_{j=1}^{\infty}$ the eigenvalues of $L_p(\lambda)$ arranged in the ascending order, then we can make the following claim.

Proposition 2.2 *To any nonnegative $\lambda < \alpha^{-1} \approx 0.19$, there exists a positive constant C_p depending on p only such that the following estimate is valid,*

$$\sum_{j=1}^{N} \lambda_{j,p} \geq C_p(1-\alpha\lambda)\frac{N^{(2p+1)/(p+1)}}{(\ln^p N+1)^{1/(p+1)}} - c\lambda N, \quad N = 1, 2, \ldots, \tag{2.20}$$

where $c = 2\left(\frac{\alpha^2}{4}+1\right) \approx 15.7$.

A similar, and even simpler result can be derived for regions with four hyperbolic "horns" such as $D = \{(x,y) \in \mathbb{R}^2 : |xy| \leq 1\}$ which can be formally viewed as the limit of $p \to \infty$ of our model, and more rigorously they are described by the Schrödinger operator

$$H_D(\lambda) : H_D(\lambda)\psi = -\Delta\psi - \lambda(x^2+y^2)\psi \tag{2.21}$$

with a parameter $\lambda \geq 0$ and Dirichlet condition on the boundary ∂D. Following [8], one can make the following claim.

Theorem 2.8 *The spectrum of $H_D(\lambda)$ is discrete for any $\lambda \in [0, 1)$ and the spectral estimate*

$$\sum_{j=1}^{N} \lambda_j \geq C(1-\lambda)\frac{N^2}{1+\ln N} \quad N = 1, 2, \dots, \tag{2.22}$$

holds true with a positive constant C.

Sketch of the proof. One can check that for any $u \in H^1$ satisfying the condition $u|_{\partial D} = 0$ the inequality

$$\int_D (x^2+y^2)u^2(x,y)\,\mathrm{d}x\,\mathrm{d}y \leq \int_D |(\nabla u)(x,y)|^2\,\mathrm{d}x\,\mathrm{d}y \tag{2.23}$$

is valid which in turn implies $H_D(\lambda) \geq -(1-\lambda)\Delta_D$ where Δ_D is the Dirichlet Laplacian on the region D. The result then follows from the eigenvalue estimates on Δ_D known from [15, 22]. □

We will not sketch the proof of Proposition 2.2 because we are able to demonstrate a substantially stronger result of Lieb–Thirring type [5].

Theorem 2.9 *Given $\lambda < \gamma_p$, let $\lambda_1 < \lambda_2 \leq \lambda_3 \leq \cdots$ be eigenvalues of $L_p(\lambda)$. Then for $\Lambda \geq 0$ and $\sigma \geq 3/2$ the following inequality is valid,*

$$\begin{aligned} &\operatorname{tr}\ \big(\Lambda - L_p(\lambda)\big)_+^{\sigma} \\ &\leq C_{p,\sigma}\left(\frac{\Lambda^{\sigma+(p+1)/p}}{(\gamma_p-\lambda)^{\sigma+(p+1)/p}}\ln\left(\frac{\Lambda}{\gamma_p-\lambda}\right) + C_\lambda^2\left(\Lambda + C_\lambda^{2p/(p+2)}\right)^{\sigma+1}\right), \end{aligned} \tag{2.24}$$

where the constant $C_{p,\sigma}$ depends on p and σ only and

$$C_\lambda := \max\left\{\frac{1}{(\gamma_p-\lambda)^{(p+2)/(p(p+1))}}, \frac{1}{(\gamma_p-\lambda)^{(p+2)^2/(4p(p+1))}}\right\}. \tag{2.25}$$

Sketch of the proof. By minimax principle, we can estimate $L_p(\lambda)$ from below by a self-adjoint operator with a purely discrete negative spectrum and derive a bound to the momenta of the latter. We split the plane $\mathbb{R}^2$ again, now in what one could call a "lego" fashion, cf. Fig. 2.8, using a monotone sequence $\{\alpha_n\}_{n=1}^{\infty}$ such that $\alpha_n \to \infty$ and $\alpha_{n+1} - \alpha_n \to 0$ holds as $n \to \infty$. Estimating the "transverse" variables by their extremal values, we reduce the problem essentially to assessment of the spectral threshold of the anharmonic oscillator with Neumann cuts. We derive easily the following asymptotic result.

Lemma 2.1 *Let $l_{k,p} = -\frac{\mathrm{d}^2}{\mathrm{d}x^2} + |x|^p$ be the Neumann operator on $[-k,k]$, $:\ k > 0$. Then*

$$\inf \sigma\left(l_{k,p}\right) \geq \gamma_p + o\big(k^{-p/2}\big) \quad \textit{as} \quad k \to \infty.$$

In fact the error is exponentially small, but the above relation is sufficient for our purposes. Combining it with the "transverse" eigenvalues $\left\{\frac{\pi^2 k^2}{(\alpha_{n+1}-\alpha_n)^2}\right\}_{k=0}^{\infty}$, using

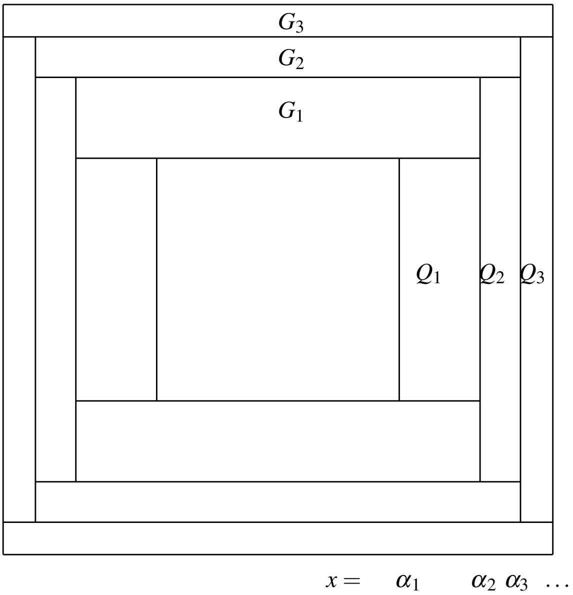


Fig. 2.7 Bracketing in the proof of Theorem 2.9

Lieb–Thirring inequality for this situation [18], and choosing properly the sequence $\{\alpha_n\}_{n=1}^{\infty}$, cf. [5], we are able to prove the claim. □

Let us finally look at the *critical case*, $L := -\Delta + |xy|^p - \gamma_p(x^2 + y^2)^{p/(p+2)}$. The essential spectrum is as expected [5] as one can check easily using properly chosen Weyl sequences.

Theorem 2.10 *We have* $\sigma_{\mathrm{ess}}(L) \supset [0, \infty)$.

The question about the negative spectrum is more involved. First of all, we have the following result [5] which can be proved by the same technique as Theorem 2.9 using another "lego bracketing" estimate.

Theorem 2.11 *The negative spectrum of L is discrete.*

For the moment, however, we are unable to prove that $\sigma_{\mathrm{disc}}(L)$ is non-empty. We conjecture that it is the case having a *strong numerical evidence* for that. For simplicity consider the case $p = 2$. We restrict the operator to a circle of radius R with Dirichlet or Neumann boundary and compute the first two eigenvalues in both cases; they are plotted in Fig. 2.8 as functions of the cut-off radius. By [20, Sect. XIII.15], the possible negative eigenvalues are squeezed between those curves. We see that the bounds become very tight for $R \gtrsim 7$ and indicate that the critical problem has for $p = 2$ an eigenvalue $E_1 \approx -0.18365$. Furthermore, $\sigma_{\mathrm{disc}}(L)$ consists of a single point because the second lower (Neumann) estimate is in positive values for R large enough. A similar numerical analysis also suggests the ground state existence for *other values of* p but it becomes unreliable for $p \gtrsim 20$. We *conjecture* that the *discrete spectrum is nonvoid for all* $p > 1$ *but empty for hyperbolic regions,* $p = \infty$.

Fig. 2.8 The Dirichlet–Neumann estimate of the spectrum in the critical case for $p = 2$

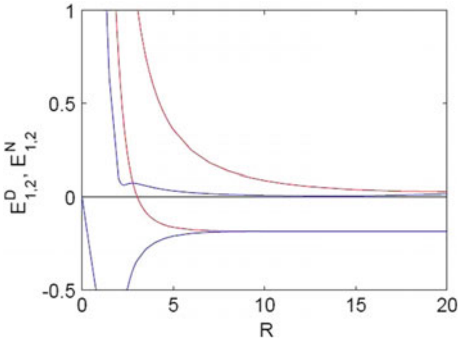
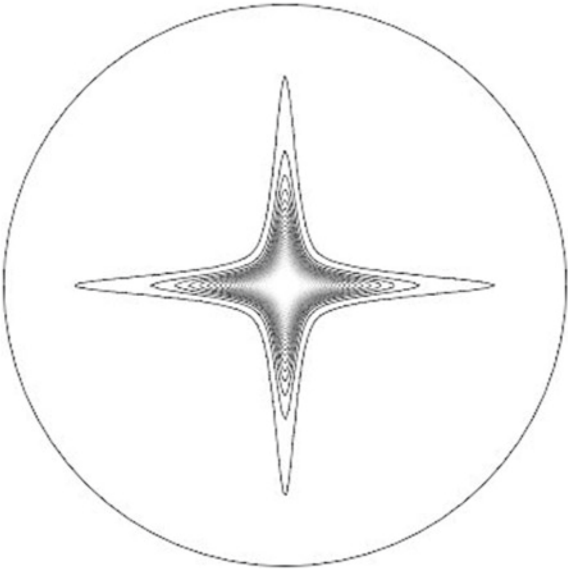


Fig. 2.9 The ground state eigenfunction in the critical case for $p = 2$



We are able to get in a numerical way an idea about the ground state eigenfunction, again in the case $p = 2$, as plotted in Fig. 2.9 based on the solution in the circle with either boundary condition; we note that with the $R = 20$ cutoff, the Dirichlet and Neumann ones are practically identical which is not surprising since one expects a superexponential decay along the axes. The outer level in the plot marks the 10^{-3} value.

2.5 Resonances in Smilansky–Solomyak Model

The models we consider have other interesting properties. Let us return to the setting of Sect. 2.2 and show that the system exhibits a rich *resonance structure*; we refer to

[10, 11] for a detailed discussion of these phenomena. To begin with, we have to say *which resonances* we speak about. There are *resolvent resonances* associated with poles in the analytic continuation of the resolvent over the cut(s) corresponding to the continuous spectrum, *scattering resonances* identified with complex singularities of the scattering matrix.

The former are found using the same Jacobi matrix problem as before, of course, this time with a "complex energy". Let us look at the latter. Suppose the incident wave comes in the mth channel from the left. We use the Ansatz

$$f(x,y)=\begin{cases}\sum_{n=0}^{\infty}\Big(\delta_{mn}\mathrm{e}^{-ipx}\psi_n(y)+r_{mn}\,\mathrm{e}^{ix\sqrt{p^2+\varepsilon_m-\varepsilon_n}}\psi_n(y)\Big)\\ \sum_{n=0}^{\infty}t_{mn}\,\mathrm{e}^{-ix\sqrt{p^2+\varepsilon_m-\varepsilon_n}}\psi_n(y)\end{cases}\tag{2.26}$$

for $\mp x>0$, respectively, where $\varepsilon_n=n+\frac{1}{2}$ and the incident wave energy is assumed to be $p^2+\varepsilon_m=:k^2$. It is straightforward to compute from here the boundary values $f(0\pm,y)$ and $f'(0\pm,y)$. The continuity requirement at $x=0$ together with the orthonormality of the basis $\{\psi_n\}$ yields

$$t_{mn}=\delta_{mn}+r_{mn}.\tag{2.27}$$

Furthermore, we substitute the boundary values coming from the Ansatz (2.26) into

$$f'(0+,y)-f'(0-,y)-\lambda yf(0,y)=0\tag{2.28}$$

and integrate the obtained expression with $\int\mathrm{d}y\;\psi_l(y)$. This yields

$$\sum_{n=0}^{\infty}\Big(2p_n\delta_{ln}-i\lambda(\psi_l,y\psi_n)\Big)r_{mn}=i\lambda(\psi_l,y\psi_m),\tag{2.29}$$

where we have denoted $p_n=p_n(k):=\sqrt{k^2-\varepsilon_n}$. In particular, poles of the scattering matrix are associated with the kernel of the ℓ^2 operator on the left-hand side. In particular, putting $l=m$ we obtain essentially the same condition we had before, cf. (2.6) and (2.7), thus we arrive at the following conclusion.

Proposition 2.3 *The resolvent and scattering resonances coincide in the Smilansky–Solomyak model.*

Let us add a few comments:

- The on-shell scattering matrix for the initial momentum k is a $\nu\times\nu$ matrix where $\nu:=\left[k^2-\frac{1}{2}\right]$ whose elements are the transmission and reflection amplitudes; they have common singularities.
- The resonance condition may have (and in fact it has) numerous solutions, but only those "not far from the physical sheet" are of interest.

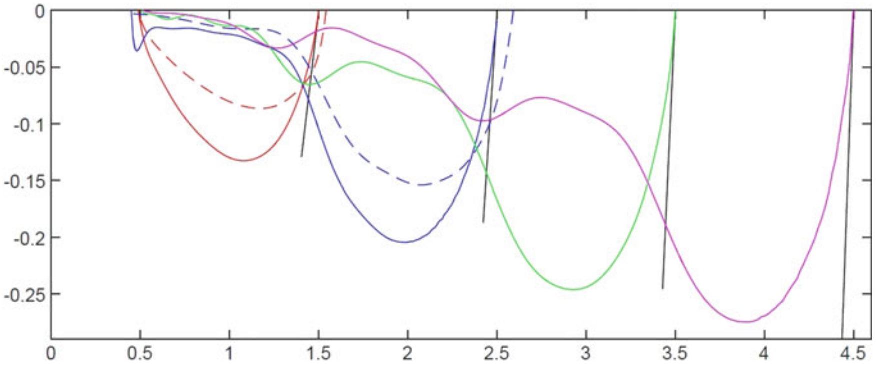


Fig. 2.10 Resonance trajectories as the coupling constant λ varies from zero to $\sqrt{2}$

- The Riemann surface of energy has an infinite number of sheets determined by the choices *branches of the square roots*. The interesting resonances on the nth sheet are obtained by *flipping sign of the first $n - 1$ of them*.

The *weak-coupling analysis* follows the route as for the discrete spectrum, cf. (2.9)–(2.11) above; in fact it includes the eigenvalue case if we stay on the “first” sheet. It shows that for small λ, a resonance pole splits of each threshold according to the asymptotic expansion

$$\mu_n(\lambda) = -\frac{\lambda^4}{64}\big(2n + 1 + 2in(n+1)\big) + o(\lambda^4). \tag{2.30}$$

Hence the distance for the corresponding threshold is proportional to λ^4 and the trajectory asymptote is the “steeper” the larger n is. However, one can solve the condition (2.29) numerically [10]. This allows us to go beyond the weak coupling regime and *the picture becomes more intriguing* as shown in Fig. 2.10. The picture shows clearly the asymptotes of the resonance trajectories for small values λ when the poles split from the channel threshold given by the oscillator eigenvalues. For stronger coupling the behavior changes and eventually the poles return to the real axis as λ approaches the critical value. What is even more interesting, the numerical solutions reveals other, “non-threshold” resonances at the second and third Riemann sheet, indicated by dotted lines that appear at $\lambda = 1.287$ and $\lambda = 1.19$, respectively.

2.6 Concluding Remarks

While we have been able to demonstrate many properties of the models under consideration, various mathematical questions remain open, for instance,

- in the original Smilansky–Solomyak model and its δ' modification of Sect. 2.3.3, we know that the essential spectrum is *absolutely continuous*. We expect that this will also be the case for the models with regular potential channels but this remains to be demonstrated.
- in the regular Smilansky–Solomyak model, the "escape channel" may have more than one mode provided $\#\sigma_{\text{disc}}(L) > 1$ holds for the operator (2.13). In this situation, it is natural to ask how the *spectral multiplicity* changes with λ.
- many questions concern *resonances* in the Smilansky–Solomyak model. One would like to know, *inter alia*, what is their number in a given part of the complex plane, whether there are resonance-free zones for a fixed λ, or whether all the poles will eventually return to the real axis as λ increases. Furthermore, we are interested in the mechanism which produces the "non-threshold" resonances and the coupling constant values at which they appear. Finally, resonance effects are also expected to occur in the regular version of the model.

From the physical point of view the most interesting question concerns the classical motion in the regular model, magnetic and nonmagnetic, as well as in the model of Sect. 2.4. We have mentioned in the opening of Sect. 2.3.1 that a step in this direction was made in [14], however, the importance of the question goes beyond the motivation of that paper dealing with modeling quantum measurements as it may offer a new and interesting insight into the quantum-classical correspondence in unusual situations we have discussed here.

Acknowledgements Our recent results discussed in this survey are the result of a common work with a number of colleagues, in the first place Diana Barseghyan, Andrii Khrabustovskyi, Vladimir Lotoreichik, and Miloš Tater whom I am grateful for the pleasure of collaboration. The research was supported by the Czech Science Foundation (GAČR) within the project 17-01706S.

References

1. S. Albeverio, F. Gesztesy, R. Høegh-Krohn, H. Holden, *Solvable Models in Quantum Mechanics*, 2nd edn. (AMS Chelsea Publishing, Providence, 2005)
2. D. Barseghyan, P. Exner, A regular version of Smilansky model. J. Math. Phys. **55**, 042104 (2014)
3. D. Barseghyan, P. Exner, A regular analogue of the Smilansky model: spectral properties. Rep. Math. Phys. **80**, 177–192 (2017)
4. D. Barseghyan, P. Exner, A magnetic version of the Smilansky-Solomyak model. J. Phys. A: Math. Theor. **50**, 485203 (24pp) (2017)
5. D. Barseghyan, P. Exner, A. Khrabustovskyi, M. Tater, Spectral analysis of a class of Schrödinger operators exhibiting a parameter-dependent spectral transition. J. Phys. A: Math. Theor. **49**, 165302 (2016)
6. W.D. Evans, M. Solomyak, Smilansky's model of irreversible quantum graphs. I: the absolutely continuous spectrum. J. Phys. A: Math. Gen. **38**, 4611–4627 (2005)
7. W.D. Evans, M. Solomyak, Smilansky's model of irreversible quantum graphs. II: the point spectrum. J. Phys. A: Math. Gen. **38**, 7661–7675 (2005)

8. P. Exner, D. Barseghyan, Spectral estimates for a class of Schrödinger operators with infinite phase space and potential unbounded from below. J. Phys. A: Math. Theor. **45**, 075204 (14pp) (2012)
9. P. Exner, J. Lipovský, Smilansky-Solomyak model with a δ'-interaction. Phys. Lett. A **382**, 1207–1213 (2018)
10. P. Exner, V. Lotoreichik, M. Tater, Spectral and resonance properties of Smilansky Hamiltonian. Phys. Lett. A **381**, 756–761 (2017)
11. P. Exner, V. Lotoreichik, M. Tater, On resonances and bound states of Smilansky Hamiltonian. Nanosyst.: Phys. Chem. Math. **7**, 789–802 (2016)
12. L. Geisinger, T. Weidl, Sharp spectral estimates in domains of infinite volume. Rev. Math. Phys. **23**, 615–641 (2011)
13. I. Guarneri, Irreversible behaviour and collapse of wave packets in a quantum system with point interactions. J. Phys. A: Math. Theor. **44**, 485304 (22 pp) (2011)
14. I. Guarneri, A model with chaotic scattering and reduction of wave packets. J. Phys. A: Math. Theor. **51**, 095304 (16 pp) (2018)
15. V. Jakšić, S. Molchanov, B. Simon, Eigenvalue asymptotics of the Neumann Laplacian of regions and manifolds with cusps. J. Funct. Anal. **106**, 59–79 (1992)
16. T. Kato, *Perturbation Theory for Linear Operators* (Springer, Berlin, 1995)
17. A. Laptev, T. Weidl, Sharp Lieb-Thirring inequalities in high dimensions. Acta Math. **184**, 87–111 (2000)
18. O. Mickelin, Lieb-Thirring inequalities for generalized magnetic fields. Bull. Math. Sci. **6**, 1–14 (2016)
19. S. Naboko, M. Solomyak, On the absolutely continuous spectrum in a model of an irreversible quantum graph. Proc. Lond. Math. Soc. **92**(3), 251–272 (2006)
20. M. Reed, B. Simon, *Methods of Modern Mathematical Physics. II. Fourier Analysis, Self-Adjointness, IV. Analysis of Operators* (Academic Press, New York, 1975, 1978)
21. G. Rozenblum, M. Solomyak, On a family of differential operators with the coupling parameter in the boundary condition. J. Comput. Appl. Math. **208**, 57–71 (2007)
22. B. Simon, Some quantum operators with discrete spectrum but classically continuous spectrum. Ann. Phys. **146**, 209–220 (1983)
23. U. Smilansky, Irreversible quantum graphs. Waves Random Media **14**, S143–S153 (2004)
24. M. Solomyak, On the discrete spectrum of a family of differential operators. Funct. Anal. Appl. **38**, 217–223 (2004)
25. M. Solomyak, On a mathematical model of irreversible quantum graphs. St. Petersbg. Math. J. **17**, 835–864 (2006)
26. M. Solomyak, On the limiting behaviour of the spectra of a family of differential operators. J. Phys. A: Math. Gen. **39**, 10477–10489 (2006)
27. M. Znojil, Quantum exotic: a repulsive and bottomless confining potential. J. Phys. A: Math. Gen. **31**, 3349–3355 (1998)

Chapter 3
Distribution Theory by Riemann Integrals

Hans G. Feichtinger and Mads S. Jakobsen

Abstract It is the purpose of this article to outline a syllabus for a course that can be given to engineers looking for an understandable mathematical description of the foundations of distribution theory and the necessary functional analytic methods. Arguably, these are needed for a deeper understanding of basic questions in signal analysis. Objects such as the Dirac delta and the Dirac comb should have a proper definition, and it should be possible to explain how one can reconstruct a band-limited function from its samples by means of simple series expansions. It should also be useful for graduate mathematics students who want to see how functional analysis can help to understand fairly practical problems, or teachers who want to offer a course related to the "Mathematical Foundations of Signal Processing" at their institutions. The course requires only an understanding of the basic terms from linear functional analysis, namely Banach spaces and their duals, bounded linear operators, and a simple version of w^*-convergence. As a matter of fact, we use a set of function spaces which is quite different from the collection of Lebesgue spaces $(L^p(\mathbb{R}_d), \|.\|_p)$ used normally. We thus avoid the use of the Lebesgue integration theory. Furthermore, we avoid topological vector spaces in the form of the Schwartz space. Although practically all the tools developed and presented can be realized in the context of LCA (locally compact abelian) groups, i.e., in the most general setting where a (commutative) Fourier transform makes sense, we restrict our attention in the current presentation to the Euclidean setting, where we have (generalized) functions over $\mathbb{R}^d$. This allows us to make use of simple BUPUs (bounded, uniform partitions of unity), to apply dilation operators and occasionally to make use of concrete special functions such as the (Fourier invariant) standard Gaussian, given by $g_0(t) = \exp(-\pi|t|^2)$. The problems of the overall current situation, with the separation of theoretical Fourier analysis as carried out by (pure) mathematicians

H. G. Feichtinger (✉)
Faculty of Mathematics, University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria
e-mail: hans.feichtinger@univie.ac.at

Charles University, Prague, Czechia

M. S. Jakobsen
Weibel Scientific A/S, Solvang 30, Alleroed, Denmark
e-mail: mja@weibel.dk

P. Manchanda et al. (eds.), *Mathematical Modelling, Optimization, Analytic and Numerical Solutions*, Industrial and Applied Mathematics,
https://doi.org/10.1007/978-981-15-0928-5_3

and Applied Fourier analysis (as used in engineering applications) are getting bigger and bigger and therefore courses filling the gap are in strong need. This note provides an outline and may serve as a guideline. The first author has given similar courses over the past years at different schools (ETH Zürich, DTU Lyngby, TU Munich, and currently Charles University Prague) and so one can claim that the outline is not just another theoretical contribution to the field.

Keywords Distributions · Banach Gelfand Triple · Kernel theorem · Wiener's algebra · Feichtinger's algebra · Distributional convergence · Kohn–Nirenberg symbol · Spreading function · Impulse response · Transfer function · Digital signal processing · Fast Fourier transform · Riemann integral

3.1 Overall Motivation

3.1.1 Psychological Aspects

It is not a secret that the way how engineers or physicists are describing "realities" is quite different from the way mathematicians want to describe the same thing. The usual agreement is that applied scientists are motivated by the concrete applications and therefore do not need to be so pedantic in the description, because they have a "better feeling" about what is true and what is not true. After all, it does not pay to be too pedantic if one wants to make progress.

On the other hand, mathematicians have a tendency to be too formal, to consider formal correctness of a statement as more important than the possible usefulness of a statement, simply because usefulness is not a category in mathematical sciences. Applicability by itself is not a criterion for important mathematical results which often go for the details of a structure without taking care of its relevance for applications. Sometimes this "abstract viewpoint" is very helpful, because it reveals important, underlying structures or allows finding connections between fields which appear to have very little in common at first sight. However, in the right (abstract) mathematical model, they appear to be almost identical. Such observations allow to sometimes transfer information and insight, or computational rules established in one area to another area, which certainly is not possible if only one single application is in the focus.

There are different ways to view these discrepancies. What we could call the *negative attitude* is to say as a mathematician: *You know, engineers and physicists are extremely sloppy, you never can trust their formulas. They claim to derive mathematical identities by using divergent integrals and so on, so one has to be careful in taking over what they "prove"*. In the same way, the engineer might say: *You know, mathematicians are pedantic people who care only about technical details and not for the content of a formula. Whenever they claim that our formulas are not correct,* *they find after some while a way to produce more theory in order to then prove that our formulas have been correct after all.*

A more positive and ambitious approach would be to agree from both sides on a few facts which are on average quite valid:

- Any mathematical statement should, at least at the end, have a proper mathematical justification;
- Formulas developed from applied scientists may, at least at the beginning, come from intuition or experiments, so they might be valid under particular conditions or under implicit assumptions (which are often clear from the physical context, e.g., positivity assumptions, etc.);
- For the progress new formulas might be more important than a refined analysis of established formulas, but the goal is to have *useful formulas whose range of applications* (the relevant assumptions) are well understood; it is important to know when there is a guarantee that the formula can be applied (because there is proof), and when one might be at risk of getting a wrong result (even if it is with low probability);
- This goal requires cooperation between applied scientists and mathematicians; usually, the first group is better trained in establishing unexplored problems while the second is expected to provide a theoretical setup which ensures that things are under control, in terms of correctness of assumptions and conclusions. Obviously, in an ideal world one group, can and should learn a lot from the other.

So in the cooperation between the two communities, mathematicians should learn more about *the goals and the motivation* and, for example, *engineers and physicists* might learn that it is also beneficial to cooperate with mathematicians and to have clear guidelines concerning the correct use of formulas and mathematical identities and where perhaps caution is in place.

3.1.2 The Search for a Banach Space of Test Functions

The overall goal of this paper is to propose a path that allows us to introduce a family of *generalized functions* which is large enough to contain most of those generalized functions which are relevant in the context of (abstract or applied) Fourier analysis and for engineering applications. Specifically Dirac measures and Dirac combs. We will demonstrate that this is possible using modest tools from functional analysis.

Before going to the technical side of the exposition, let us motivate the use of dual spaces and functional analytic methods, and shed some light on the idea of *distributions*. Let us start with some observations:

- First of all, it is clear that *generalized functions* should form a linear space so that linear combinations of those objects (sometimes called signals) can be formed, and under certain conditions, even limits, and hence infinite series;

- Second, we would like to have "ordinary functions" included in a natural way within the world of generalized functions, so we need a *natural embedding* of as many linear spaces of ordinary functions as possible;
- As a third variant we can think of generalized functions as a kind of "limits" of ordinary functions, but in a specific sense (and ideally the convergence should also be allowed to be applied to the generalized functions);
- Finally, there are many operations that can be carried out for (certain) functions, such as translation, convolution, dilation, Fourier transform, and we will go for a setting where the approximation properties of the previous item allow extending these operations to the linear space of generalized functions.

In order to explain our understanding of "distribution theory", let us first formulate again some general thoughts. In fact it is not surprising that we have to use functional analytic methods in this context because after all at least for continuous variables, signal spaces tend to be *not finite dimensional* anymore[1] and so we have to resort to methods that allow us to describe the convergence of infinite series. The simplest way to do this is to assume that one has a linear space and a normed space, $(B, \| \cdot \|_B)$. If one has, in addition, a kind of multiplication $(a, b) \mapsto a \bullet b$ (with the usual rules), one speaks of *normed algebras*, if

$$\|b_1 \bullet b_2\|_B \leq \|b_1\|_B \cdot \|b_2\|_B \text{ for all } b_1, b_2 \in B.$$

Among the normed spaces those which are *complete*, the *Banach spaces* are the most important ones, because like $\mathbb{R}$ itself with the mapping $x \mapsto |x|$ one has (by definition) *completeness*, meaning that every *Cauchy sequence* is convergent. This is known to be equivalent to the fact that every *absolutely convergent sequence* with $\sum_{k=1}^{\infty} \|b_k\|_B < \infty$, is convergent so that the partial sums $\sum_{k=1}^{n} b_k$ have a limit (in $(B, \| \cdot \|_B)$). Therefore the infinite sum is (unconditionally, or independent of the order) well- defined, and thus the symbol $\sum_{k=1}^{\infty} b_k$ is meaningful in this situation.

The most important tool within the linear functional analysis is the *linear functionals*, or bounded linear mappings from B into $\mathbb{C}$ (or into $\mathbb{R}$ for the case of real vector spaces). Such a functional σ has to satisfy two properties:

1. Linearity: $\sigma(\alpha \mathbf{b}_1 + \beta \mathbf{b}_2) = \alpha\sigma(\mathbf{b}_1) + \beta\sigma(\mathbf{b}_2), \quad \mathbf{b}_1, \mathbf{b}_2 \in \mathrm{B}, \alpha, \beta \in \mathbb{C}$.
2. Boundedness: There exists $c > 0$ such that $|\sigma(\mathbf{b})| \leq c\|\mathbf{b}\|_B, \ \forall \mathbf{b} \in \mathrm{B}$.

For any given normed space $(B, \| \cdot \|_B)$, the collection of all such bounded linear functionals constitutes the *dual space*, denoted by B'. It carries a norm, given by

$$\|\sigma\|_{B'} := \sup_{\|\mathbf{b}\|_B \leq 1} |\sigma(\mathbf{b})|.$$

[1] Commonly the term "infinite dimensional" is used, and we will also use it later on, but this expression wrongly suggests that instead of a *finite basis* one just has an infinite basis and this is not what we should expect or use!

With this norm, B' turns out to be a Banach space.[2] One can think of the dual space as the collection of all coordinate functionals (describing the contribution of a fixed element in a basis) over all finite dimensional subspaces of B, thus capturing all the information about the underlying normed space.

In addition to norm convergence on B', we will use what is called the w^*-convergence. It can be described for sequences as *convergence in action*.

For all practical purposes,[3] the following definition is a simple way of describing what is called w^*-convergence.

Definition 3.1 A sequence of linear functionals $(\sigma_n)_{n\geq 1}$ *converges in action* or *in the weak*-sense* to some $\sigma_0 \in B'$ if we have

$$\lim_{n\to\infty} \sigma_n(\mathbf{b}) = \sigma_0(\mathbf{b}) \text{ for all } \mathbf{b} \in B. \tag{3.1}$$

By the Banach–Steinhaus Theorem, the convergence for *all* $\mathbf{b} \in B$ implies boundedness, i.e., $\sup_{n\geq 1} \|\sigma\|_{B'} < \infty$, and that conversely it is (under this condition!) enough to claim that the limits on the left-hand side exist for any $\mathbf{b} \in B$, thus defining the functional σ_0. In fact, it would be even enough (given the boundedness condition) to know that one has a limit for all b from a dense subspace of $(B, \|\cdot\|_B)$.

Infinite dimensional Banach spaces $(B, \|\cdot\|_B)$ do not satisfy the Heine–Borel property. A bounded sequence may fail to have a (norm) convergent subsequence. But the *Banach–Alaoglu Theorem* (see [8]) ensures that any bounded sequence (σ_k) in $(B', \|\cdot\|_{B'})$ has a subsequence $(\sigma_{n_k})_{k\geq 1}$ which is w^*-convergent to some $\sigma_0 \in B'$, i.e.,

$$\lim_{k\to\infty} \sigma_{n_k}(b) = \sigma_0(b) \text{ for all } \mathbf{b} \in B.$$

In a similar way, the set of all bounded and linear operators between two normed spaces is defined; we denote it by $\mathscr{L}(B^1, B^2)$. It is always a normed space with respect to the operator norm

$$|||T||| := \sup_{\|b_1\|_{B^1}\leq 1} \|T(b_1)\|_{B^2}$$

and if $(B^2, \|\cdot\|_{(2)})$ is a Banach space, the space of operators is complete as well. In particular, for the choice $B^2 = \mathbb{C}$ the space reduces to the dual space.

For the case $B^1 = B = B^2$ these operators form a normed algebra, and in fact a *Banach algebra* if $(B, \|\cdot\|_B)$ is a Banach space.

Since many sequences of functions which do not have a reasonable pointwise limit, such as a sequence of compressed box functions which converge to the so-called *Dirac delta*, often denoted by $\delta(t)$ in the engineering literature, are in fact

[2] Even if $(B, \|\cdot\|_B)$ is just a normed space.

[3] Technically speaking, for *separable* Banach spaces $(B, \|\cdot\|_B)$ which are those that contain a countable, dense subset. This will be the case for all the situations where we make use of this concept.

limits in this sense, it is at least plausible to work with dual spaces in order to capture these limits.

Without going too much into the psychological and didactical side of this issue, let us just state here that indeed, it is *meaningful* to model generalized functions as what we will call *distributions*, namely elements of dual spaces for suitable chosen Banach spaces $(B, \| \cdot \|_B)$ of integrable and bounded, continuous functions.

We admit that of course this terminology is influenced by the existing traditional way of introducing generalized functions, for example, by using the *tempered distributions* developed by Laurent Schwartz [45] using the (nuclear Frechet) space $\mathscr{S}(\mathbb{R}^d)$ of *rapidly decreasing functions*. While differentiability is in the focus of attention there, we leave this aspect aside and allow ourselves to call an algebra (with respect to pointwise multiplication and/or convolution) of continuous functions a *space of test functions* and the dual space a space of *distributions*. This will be the setting we choose for our approach. Thus from now on, we will mostly talk about test functions and distributions, but we will still have to explain in which sense distributions are generalized functions in the spirit of the above description.

One can also motivate the use of dual spaces for the description of linear spaces of signals by the following argument:

A signal is something that can be measured!

Just thinking of an audio signal which we can record using a microphone, we can compress using MP3 coding based on the FFT, and we can transmit it. All this is on the basis of *linear measurements* which are of course continuous in some sense, meaning that quite similar signals (whatever they are) will provide similar measurements. But is the audio signal a pointwise almost everywhere defined function in $L^2(\mathbb{R})$ in the mathematical sense? Of course we can take pictures of a natural scene and enjoy the quality of a color picture taken by a 16-million pixel camera, but does that device really sample (in the mathematical sense) a continuous, $2D$-function describing the analog picture which we use in a conversational situation?

The situation is really much more like an abstract probability distribution, say a normal distribution with some expectation value and some variance. We will never be able (except through indirect mathematical description) to provide a pointwise description of such a "distribution" (a different but related use of this word), so normally one resorts to the use of *histograms*. Given the bins used for the histogram, one can describe the height of the bars simply as the value obtained by applying the (nonnegative) measure (via integration) to the indicator function of the corresponding interval (bin), making sure that the union of the bins is the whole real line or at least the range of the random variable respectively, the support of the corresponding measure.

What we are doing here is essentially replace those (finer and finer) bins by BUPUs (uniform partitions of unity), with the extra demand for assuming that they are continuous and not just step functions. The reader should see this as a minor and just technical modification (which is avoiding the distinction between step functions and continuous functions, and is also much more convenient for the setting of LCA groups).

The (abstract) viewpoint of considering signals as something that can be measured also suggests very naturally a measure of similarity of signals. If for a given (potentially comprehensive) set of measurements only very small deviations are observed, then we think of those signals as "quite similar", and a sequence of signals may converge in this way to a limit signal (e.g., coarse approximations to the continuous limit). But this kind of convergence is encapsulated mathematically in the concept of w^*-convergence described above, that will be used intensively in this text.

3.2 Notations and Preliminaries

Although the approach described below can be used to develop Harmonic Analysis in the context of locally compact abelian (LCA) groups, we restrict our attention to the setting of Euclidean spaces $\mathbb{R}^d$. This is the framework relevant for most engineering work and physics.

Let us fix some notation. It all starts with the most simple vector space of functions on $\mathbb{R}^d$, namely $C_c(\mathbb{R}^d)$, the space of continuous, complex-valued and compactly supported functions on $\mathbb{R}^d$, i.e., with $\text{supp}(k) \subset B_R(0) := \{x : |x| \le R\}$ for some $R > 0$. For such a function $f \in C_c(\mathbb{R}^d)$ the notion of an integral, $\int_{\mathbb{R}_d} f(t)\,dt$, is well-defined by Riemann integration, and thus this (infinite dimensional) linear space of functions can be endowed with many different norms, such as the maximum-norm or uniform-norm, $\|k\|_\infty = \sup_{t\in\mathbb{R}^d} |f(t)|$ and the p-norms $\|k\|_p = (\int_{\mathbb{R}_d} |k(t)|^p\,dt)^{1/p}$ for $1 \le p < \infty$. The *completion* of $C_c(\mathbb{R})$ with respect to the p-norm yields the Lebesgue spaces, $(L^p(\mathbb{R}^d), \|.\|_p)$. Most notably are $L^1(\mathbb{R}^d)$ and $L^2(\mathbb{R}^d)$. The latter being a Hilbert space with respect to the inner product $\langle f, g\rangle = \int_{\mathbb{R}_d} f(t)\,\overline{g(t)}\,dt$.

For complex-valued functions f, g on $\mathbb{R}^d$, we define the following operations,

pointwise multiplication, $(f \cdot g)(t) = f(t) \cdot g(t), t \in \mathbb{R}^d$,
flip operation, $f^{\checkmark}(t) = f(-t)$,
complex conjugation, $\overline{f}(t) = \overline{f(t)}$,
translation by $x \in \mathbb{R}^d$, $T_x f(t) = f(t - x)$,
modulation by $\omega \in \mathbb{R}^d$, $E_\omega f(t) = e^{2\pi i\omega\cdot t}\, f(t)$,
dilation by an invertible $d \times d$ matrix A, $\alpha_A f(t) = |\det(A)|^{1/2}\, f(At)$,
specifically homogeneous dilations for $\rho > 0$,
$[\text{St}_\rho f](t) = \rho^{-d} f(t/\rho)$, and $[\text{D}_\rho h](t) = h(\rho t)$
with $\|\text{St}_\rho f\|_1 = \|f\|_1$ and $\|\text{D}_\rho f\|_\infty = \|f\|_\infty$.

Let $\varDelta$ be the *tent* function given by

$$\varDelta(t) = \prod_{j=1}^{d} \max\big(1 - 2|t^{(j)}|, 0\big), \quad t = (t^{(1)}, t^{(2)}, \dots, t^{(d)}) \in \mathbb{R}^d.$$

Observe that $\text{supp}\,\varDelta = [-1/2, 1/2]^d$. We define the family of functions $(\psi_n)_{n\in\mathbb{Z}^d}$ to be the collection of half-integer translates of $\varDelta$ so that

$$\psi_n(t) = \Delta\big(t - \tfrac{1}{2}n\big),\ t \in \mathbb{R}^d,\ n \in \mathbb{Z}^d. \tag{3.2}$$

The crucial properties of the functions (ψ_n) are for us that they satisfy the general assumptions of a BUPU (*bounded uniform partition of unity*), of which we give the definition below. Throughout this work (ψ_n) will always refer to the functions in (3.2). However, any other BUPU can also be used, which entails only minor modifications to our proofs.

For most applications, *regular BUPUs* will be sufficient (and easier to handle), which are obtained as translates of one (smooth) function with compact support along some lattice in $\mathbb{R}^d$. In this setting it is natural to use smooth BUPUs with respect to some lattice $\Lambda = \mathbf{A}\mathbb{Z}^d$, for some non-singular $d \times d$ matrix $\mathbf{A}$. For convenience of notation we use mostly lattices of the form $\gamma\mathbb{Z}^d$, for some $\gamma > 0$.

Definition 3.2 A family $\Psi = (\psi_k)_{k\in\mathbb{Z}^d} = (T_{\gamma k}\psi_0)_{k\in\mathbb{Z}^d}$ in $C_c(\mathbb{R})$ (for some $\gamma > 0$) is called a *regular, uniform partition of unity* on $\mathbb{R}^d$ of size R, (we write $|\Psi| \leq R$ or $\operatorname{diam}\Psi \leq R$) if

1. ψ_0 is compactly supported in $B_R(0)$.[4]
2. $\sum_{k\in\mathbb{Z}^d} \psi_k(x) = \sum_{k\in\mathbb{Z}^d} \psi_0(x - \gamma k) \equiv 1$ on $\mathbb{R}^d$.

Usually it is assumed that $\psi_0(x) \geq 0$.

3.3 Continuous Functions That Vanish at Infinity

The uniform or sup norm of functions on $\mathbb{R}^d$ is defined by $\|f\|_\infty = \sup_{t\in\mathbb{R}^d} |f(t)|$.

Observe that $C_b(\mathbb{R}^d)$, the space of all bounded, continuous, complex-valued functions on $\mathbb{R}^d$ is a Banach algebra with respect to this norm and pointwise multiplication. It is easy to show that $(C_c(\mathbb{R}^d), \|\cdot\|_\infty)$ is not complete. Its completion in $(C_b(\mathbb{R}^d), \|\cdot\|_\infty)$, which is the same as the closure within $(C_b(\mathbb{R}^d), \|\cdot\|_\infty)$, is just the space of continuous functions that vanish at infinity. We denote this space by $(C_0(\mathbb{R}^d), \|\cdot\|_\infty)$. For $f \in C_0(\mathbb{R}^d)$ and $h \in C_b(\mathbb{R}^d)$ the pointwise product $f \cdot h$ is again in $C_0(\mathbb{R}^d)$. In particular, $(C_0(\mathbb{R}^d), \|\cdot\|_\infty)$ is itself a (commutative) Banach algebra with respect to pointwise multiplication, with

$$\|f \cdot h\|_\infty \leq \|f\|_\infty \, \|h\|_\infty. \tag{3.3}$$

We *define* the space of *bounded measures* $M_b(\mathbb{R}^d)$ to be the continuous (Banach space) dual of $(C_0(\mathbb{R}^d), \|\cdot\|_\infty)$. That is, $M_b(\mathbb{R}^d) = C_0'(\mathbb{R}^d)$ consists of all linear and continuous functionals $\mu : C_0(\mathbb{R}^d) \to \mathbb{C}$. We write the action of a functional $\mu \in M_b(\mathbb{R}^d)$ on a function $f \in C_0(\mathbb{R}^d)$ as $\mu(f)$. Naturally, $M_b(\mathbb{R}^d)$ is a Banach space with respect to the operator norm,

[4] $B_R(0)$ is the ball of radius $R > 0$ around zero in $\mathbb{R}^d$.

$$\|\mu\|_{M_b} = \sup_{f\in C_0(\mathbb{R}^d),\ \|f\|_\infty\le 1} \big|\mu(f)\big|. \tag{3.4}$$

There are two simple and natural examples of bounded measures. First of all the Dirac measure (or Dirac delta) of the form $\delta_x : f \mapsto f(x),\ x \in \mathbb{R}^d$.[5] Their finite linear combinations are called *finite discrete measures* and belong also to $M_b(\mathbb{R}^d)$.

Second, any function $g \in C_c(\mathbb{R})$ defines a bounded measure μ_g by

$$\mu_g : C_0(\mathbb{R}^d) \to \mathbb{C}, \quad \mu_g(f) = \int_{\mathbb{R}_d} f(t)\, g(t)\, dt, \quad f \in C_0(\mathbb{R}^d). \tag{3.5}$$

This integral is well-defined as $f \cdot g \in C_c(\mathbb{R}^d)$.

We mention the following operations that one can do with bounded measures: we define the product of a bounded measure $\mu \in M_b(\mathbb{R}^d)$ with a function $h \in C_b(\mathbb{R}^d)$ to be the bounded measure given by

$$\big(\mu \cdot h\big)(f) := \mu(h \cdot f) \ \text{ for all } \ f \in C_0(\mathbb{R}^d). \tag{3.6}$$

Observe that $\|\mu \cdot h\|_{M_b} \le \|h\|_\infty\, \|\mu\|_{M_b}$, and of course associativity.

Furthermore, we define the complex conjugation of a bounded measure, its flip, translation, modulation, and dilation to be, for any $\mu \in M_b(\mathbb{R}^d)$ and $f \in C_0(\mathbb{R}^d)$,

$$\begin{aligned}
\overline{\mu}(f) &= \overline{\mu(\overline{f})},\\
\mu^{\checkmark}(f) &= \mu(f^{\checkmark}),\\
\big(T_x\mu\big)(f) &= \mu(T_{-x}f),\ x \in \mathbb{R}^d,\\
\big(E_\omega\mu\big)(f) &= \mu(E_\omega f),\ \omega \in \mathbb{R}^d,\\
\big(\alpha_A\mu)(f) &= \mu(\alpha_{A^{-1}} f),\ A \in \mathrm{GL}_{\mathbb{R}}(d).
\end{aligned}$$

The reader may verify consistency with the corresponding operators defined on ordinary functions, i.e., that for any $g \in C_c(\mathbb{R})$

$$\overline{\mu_g} = \mu_{\overline{g}},\ \ (\mu_g)^{\checkmark} = \mu_{g^{\checkmark}},\ \ T_x\mu_g = \mu_{T_x g},\ \ E_\omega\mu_g = \mu_{E_\omega g},\ \ \alpha_A\mu_g = \mu_{\alpha_A g}.$$

Furthermore, one has the following rather natural rules:

$$T_y\delta_x = \delta_{x+y},\ \ \delta_x^{\checkmark} = \delta_{-x},\ \ \overline{\delta_x} = \delta_x,\ \ \delta_x \cdot h = h(x)\cdot\delta_x.$$

Finally, we define $\mu * f$ to be the convolution of a function $f \in C_0(\mathbb{R}^d)$ with a measure $\mu \in M_b(\mathbb{R}^d)$. It is a new function on $\mathbb{R}^d$ given pointwise by

[5] What we denote by δ_x is often called the Dirac delta *function* and denoted by $\delta_x(t)$ or $\delta(t-x)$ (the argument indicating that it is a "function" of, e.g., a time variable t). We do not view the Dirac delta in this way.

$$\big(\mu * f\big)(x) = \mu(T_x[f^{\checkmark}]) = \big(T_{-x}\mu\big)(f^{\checkmark}), \quad x \in \mathbb{R}^d. \tag{3.7}$$

Observe that $\delta_x * f = T_x f$. This correspondence is in fact the reason why the "moving average" described in (3.7) makes use of the flip operator.

Theorem 3.1 *For any $\mu \in M_b(\mathbb{R}^d)$ and any $f \in C_0(\mathbb{R}^d)$, the convolution product $\mu * f$ is a function in $C_0(\mathbb{R}^d)$. Moreover, $C_\mu : f \mapsto \mu * f$ is a bounded operator*

$$\|\mu * f\|_\infty \leq \|\mu\|_{M_b} \|f\|_\infty, \quad f \in C_0(\mathbb{R}^d),$$

which commutes with translations, i.e., $\mu * (T_x f) = T_x(\mu * f)$ for all $x \in \mathbb{R}^d$. Moreover, the operator norm of C_μ equals the functional norm of μ.

One can, in fact, show that every continuous operator $T : C_0(\mathbb{R}^d) \to C_0(\mathbb{R}^d)$ that satisfied the commutation relation $T \circ T_x = T_x \circ T$ for all $x \in \mathbb{R}^d$ is given by an operator that convolves with some uniquely determined measure $\mu \in C_b(\mathbb{R}^d)$. A proof of this statement and Theorem 3.1 can be found in the first author's lecture notes.[6] Such an operator is also called TILS (*translation invariant linear system*). For more on this, see Sect. 3.11.

Definition 3.3 Given $f \in C_b(\mathbb{R}^d)$ and $\delta > 0$, we define the oscillation function

$$\mathrm{osc}_\delta(f)(x) := \max_{|y|\leq\delta} |f(x) - f(x+y)|. \tag{3.8}$$

We also define the *local maximal function* for any $f \in C_b(\mathbb{R}^d)$,

$$f^{\#}(x) = \max_{|y|\leq 1} |f(x+y)|, \quad x \in \mathbb{R}^d. \tag{3.9}$$

There are a couple of harmless but useful pointwise estimates.

Lemma 3.1 *For any two functions $f, f_1, f_2 \in C_b(\mathbb{R}^d)$ one has that*

(i) $osc_\delta(f) \leq 2f^{\#}$;
(ii) $osc_\delta(f_1 + f_2) \leq osc_\delta(f_1) + osc_\delta(f_2)$;
(iii) $|f| \leq |g| \Rightarrow f^{\#} \leq g^{\#}$;
(iv) $(f_1 + f_2)^{\#} \leq f_1^{\#} + f_2^{\#}$;
(v) $osc_\delta(T_x f) = T_x osc_\delta(f)$;
(vi) $(T_x f)^{\#} = T_x(f^{\#})$.

Proof The proof is left as an exercise to the reader.

Using these relations, the following is a simple observation.

[6] See the lectures notes on "Harmonic and Functional Analysis" at https://www.univie.ac.at/nuhag-php/home/skripten.php.

Lemma 3.2 *A function $f \in C_b(\mathbb{R}^d)$ is uniformly continuous if and only if*

$$\|osc_\delta(f)\| \to 0 \ for \ \delta \to 0.$$

For every BUPU Ψ, we define the spline-type *quasi- interpolation operator*

$$f \mapsto \mathrm{Sp}_\Psi f : \quad \mathrm{Sp}_\Psi f(t) = \sum_{n \in \mathbb{Z}^d} f(t_n)\psi_n(t), \ \ t \in \mathbb{R}^d. \tag{3.10}$$

Lemma 3.3 *For any regular BUPU Ψ, the operator Sp_Ψ maps $C_0(\mathbb{R}^d)$ and $C_b(\mathbb{R}^d)$ onto itself, respectively, with $\|Sp_\Psi f\|_\infty \leq \|f\|_\infty$. One has $\|Sp_\Psi f - f\|_\infty \to 0$ as* $\mathrm{diam}(\Psi) \to 0$ *if and only if f is uniformly continuous (e.g., $f \in C_0(\mathbb{R}^d)$).*

Proof The first statement follows easily from the fact that all ψ_n are continuous and compactly supported together with the assumed properties of the function f. For the second statement, note that we only have to do a pointwise estimate between $f(t)$ and $\mathrm{Sp}_\Psi f(t) = \sum_{n \in \mathbb{Z}^d} \psi_n(t_n) f(t)$, where $I \subset \mathbb{Z}^d$ is such that $\mathrm{supp}\,\psi_n \cap B_\delta(t) \neq \emptyset$ for all $n \in \mathbb{Z}^d$. Using the fact that the (ψ_n) form a partition of unity, we establish that

$$|\mathrm{Sp}_\Psi f(t) - f(t)| \leq \sum_{n \in \mathbb{Z}^d} |f(t_n) - f(t)| \cdot \psi_n(t).$$

If Ψ is a BUPU such that $|t - t_n| \leq \delta$ for all $t \in \mathrm{supp}(\psi_n)$, then we find that

$$|\mathrm{Sp}_\Psi f(t) - f(t)| \leq \mathrm{osc}_\delta(f)(t).$$

As the support of the functions in the BUPU Ψ is made smaller, we write $|\Psi| \to 0$, δ goes to zero. By Lemma 3.2 we conclude that $\|\mathrm{Sp}_\Psi f - f\|_\infty \to 0$ as $|K| \to 0$.

One important result that we need for later is the following one. We give a proof of Theorem 3.2 at the end of this section.

Theorem 3.2 *Let $\Psi = (\psi_n)_{n \in \mathbb{Z}^d}$ be the BUPU as in (3.2). Every $\mu \in M_b(\mathbb{R}^d)$ can be represented by the absolutely norm convergent series $\mu = \sum_{n \in \mathbb{Z}^d} \mu \cdot \psi_n$. Moreover,*

$$\|\mu\|_{M_b} = \sum_{n \in \mathbb{Z}^d} \|\mu \cdot \psi_n\|_{M_b}. \tag{3.11}$$

Corollary 3.1 *For any $\mu \in M_b(\mathbb{R}^d)$ and any $\varepsilon > 0$, there exists a finite subset $F_0 \subset \mathbb{Z}^d$ such that $\|\mu - \sum_{n \in F} \mu \cdot \psi_n\|_{M_b} < \varepsilon$ for any finite subset of $\mathbb{Z}^d$ with $F \supseteq F_0$. One can think of $p = \sum_{n \in F} \psi_n \in C_c(\mathbb{R}^d)$ as a plateau-type function with $\|\mu - \mu \cdot p\|_{M_b} < \varepsilon$.*

Proof of Theorem 3.2. For any given $\varepsilon > 0$, let $\varepsilon_n > 0$, $n \in \mathbb{Z}^d$ be such that $\sum_{n \in \mathbb{Z}^d} \varepsilon_n < \varepsilon$. By the definition of $\|\mu \cdot \psi_n\|_{M_b}$, we can find $f_n \in C_0(\mathbb{R}^d)$, $\|f_n\|_\infty \leq 1$ such that

$$|\big(\mu \cdot \psi_n\big)(f_n)| > \|\mu \cdot \psi_n\|_{M_b} - \varepsilon_n.$$

Without loss of generality, we can assume that $\big(\mu \cdot \psi_n\big)(f_n)$ is real-valued and non-negative. For any finite set $F \subset \mathbb{Z}^d$, we define $f \in C_c(\mathbb{R}^d)$ by $f = \sum_{n \in F} f_n \cdot \psi_n$. We now observe that

$$\begin{aligned}\mu(f) &= \sum_{n \in F} \mu(f_n \cdot \psi_n) = \sum_{n \in F} \big(\mu \cdot \psi_n\big)(f_n)\\ &> \sum_{n \in F} \big(\|\mu \cdot \psi_n\|_{M_b} - \varepsilon_n\big) > \Big(\sum_{n \in F} \|\mu \cdot \psi_n\|_{M_b}\Big) - \varepsilon.\end{aligned}$$

By a simple pointwise estimate, we find that $\|f\|_\infty \leq 1$. Thus that for every $\varepsilon > 0$ and any finite set $F \subset \mathbb{Z}^d$, there is a function $f \in C_c(\mathbb{R}^d)$, $\|f\|_\infty \leq 1$ such that

$$\sum_{n \in F} \|\mu \cdot \psi_n\|_{M_b} \leq \mu(f) + \varepsilon.$$

This being true for any $\varepsilon > 0$ and any finite set we conclude that

$$\sum_{n \in \mathbb{Z}^d} \|\mu \cdot \psi_n\|_{M_b} \leq \|\mu\|_{M_b}.$$

Hence $\sum_{n \in \mathbb{Z}^d} \mu \cdot \psi_n$ is absolutely convergent in $M_b(\mathbb{R}^d)$. Finally, we show that $\mu = \sum_{n \in \mathbb{Z}^d} \mu \cdot \psi_n$. For any $f \in C_c(\mathbb{R}^d)$ we clearly have

$$\Big(\sum_{n \in \mathbb{Z}^d} \mu \cdot \psi_n\Big)(f) = \sum_{n \in F} \big(\mu \cdot \psi_n)(f) = \mu\Big(\sum_{n \in F} \psi_n \cdot f\Big) = \mu(f),$$

where F is some finite subset of $\mathbb{Z}^d$ that depends on the support of f. Since this equality holds for all $C_c(\mathbb{R})$ which is dense in $C_0(\mathbb{R}^d)$, we get $\mu = \sum_{n \in \mathbb{Z}^d} \mu \cdot \psi_n$. The opposite estimate, namely $\|\mu\|_{M_b} \leq \sum_{n \in \mathbb{Z}^d} \|\mu \cdot \psi_n\|_{M_b}$ is clear by the triangle inequality and the completeness of $M_b(\mathbb{R}^d)N$.

3.4 The Wiener Algebra on $\mathbb{R}^d$

At this point, we are in a situation where we can define pointwise multiplication within the Banach algebra $(C_0(\mathbb{R}^d), \|\cdot\|_\infty)$ and we can convolve a measure with a function $C_0(\mathbb{R}^d)$. Furthermore, we can multiply any measure with a function in $C_b(\mathbb{R}^d)$, always together with the corresponding norm estimates.

But not every function $f \in C_0(\mathbb{R}^d)$ defines a measure and it is not possible to define the convolution product of two arbitrary functions $f_1, f_2 \in C_0(\mathbb{R}^d)$. Hence it is desirable to reduce the reservoir of "test functions" from $(C_0(\mathbb{R}^d), \|\cdot\|_\infty)$ to a smaller one. The first step into this direction will be the introduction of "our new space of test functions", the Wiener algebra. It is defined as follows.

Definition 3.4 Given the BUPU $\Psi = (\psi_n)_{n\in\mathbb{Z}^d}$ in (3.2) the *Wiener algebra* $W(\mathbb{R}^d)$ consists of all continuous functions $f \in C_b(\mathbb{R}^d)$ for which the following norm is finite:

$$\|f\|_W := \sum_{n\in\mathbb{Z}^d} \|f \cdot \psi_n\|_\infty < \infty. \tag{3.12}$$

One can show that the definition does *not* depend on the particular choice of the BUPU, i.e., different BUPUs Ψ^1 or Ψ^2 define the same space. Also, $(W(\mathbb{R}^d), \|\cdot\|_W)$ is a Banach space. We mention that an equivalent norm on $W(\mathbb{R}^d)$ is given by

$$\|f\|_{W,\sqcap} = \sum_{n\in\mathbb{Z}^d} \|f \cdot T_n \mathbb{1}_{[0,1]^d}\|_\infty,$$

where $\mathbb{1}_{[0,1]^d}$ is the characteristic function on the set $[0, 1]^d$. This is the norm still widely used in the literature, and used in H. Reiter's book [38] as an example of an interesting Segal algebra (and even going back to N. Wiener's work on Tauberian theorems). Convolution relations for this (and more general Wiener amalgam spaces) are given in [6, 16, 29].

Observe that for any $f \in W(\mathbb{R}^d)$ and $x \in \mathbb{R}^d$ we have, in general, that $\|T_x f\|_W \neq \|f\|_W$. We will not need a norm that is strictly isometric with respect to translation. One way to do this is to introduce the *continuous description* of amalgam norms, which has been given already in [11].

The Wiener algebra relates to the previously considered function spaces as follows: all functions in $W(\mathbb{R}^d)$ belong to $C_0(\mathbb{R}^d)$. The space $C_c(\mathbb{R}^d)$ is contained in $(W(\mathbb{R}^d), \|\cdot\|_W)$ and $W(\mathbb{R}^d)$ is contained in $(C_0(\mathbb{R}^d), \|\cdot\|_\infty)$, both as dense subspaces. All the inclusions are in fact continuous embeddings. Furthermore, just as $C_0(\mathbb{R}^d)$ and $C_b(\mathbb{R}^d)$, the Wiener algebra behaves well with respect to multiplication.

Lemma 3.4 *(i) The Wiener algebra $W(\mathbb{R}^d)$ is continuously embedded into $C_b(\mathbb{R}^d)$ and $C_0(\mathbb{R}^d)$. Specifically, one has that*

$$\|f\|_\infty \leq \|f\|_W \ \textit{for all} \ f \in W(\mathbb{R}^d).$$

(ii) The Wiener algebra is an ideal of $C_b(\mathbb{R}^d)$ with respect to pointwise multiplication. In fact, for any $h \in C_b(\mathbb{R}^d)$ and $f \in W(\mathbb{R}^d)$ one has that

$$\|h \cdot f\|_W \leq \|h\|_\infty \|f\|_W.$$

(iii) The Wiener algebra is a Banach algebra with respect to pointwise multiplication. For any $f, h \in W(\mathbb{R}^d)$ we have that $\|h \cdot f\|_W \leq \|h\|_W \|f\|_W$.

Proof (i). By assumption we have $1 = \sum_{n\in\mathbb{Z}^d} \psi_n(x)$ for all $x \in \mathbb{R}^d$. Hence

$$\sup_{x\in\mathbb{R}^d} |f(x)| = \sup_{x\in\mathbb{R}^d} |\sum_{n\in\mathbb{Z}^d} f(x)\,\psi_n(x)| \le \sum_{n\in\mathbb{Z}^n} \|f\cdot\psi_n\|_\infty = \|f\|_W < \infty, \quad \forall f\in W(\mathbb{R}^d).$$

(ii). Let h and f be as in the statement. It follows from the easy estimate

$$\sum_{n\in\mathbb{Z}^d} \|h\cdot f\cdot\psi_n\|_\infty \le \|h\|_\infty \sum_{n\in\mathbb{Z}^d} \|f\cdot\psi_n\|_\infty = \|h\|_\infty\,\|f\|_W.$$

(iii). This follows by (i) and (ii).

Lemma 3.5 *The translation and the modulation operator are continuous on the Wiener algebra* $(W(\mathbb{R}^d), \|\cdot\|_W)$. *In fact,*

$$\|T_x f\|_W \le 4^d\,\|f\|_W \;\; \textit{and} \;\; \|E_\omega f\|_W = \|f\|_W \;\; \textit{for all} \;\; x,\omega\in\mathbb{R}^d, \;\; f\in W(\mathbb{R}^d).$$

Moreover, the dilation by an invertible $d\times d$ *matrix* A, $\alpha_A f(t) = |\det(A)|^{1/2} f(At)$ *is a continuous operator on* $W(\mathbb{R}^d)$ *for each such* A.

Proof The relation for the modulation operator is trivial. For the translation operator we have to work a bit harder. First, observe that for any $t, x\in\mathbb{R}^d$ we have

$$\wedge(t+x) = \wedge(t+x)\cdot 1 = \wedge(t+x)\cdot \sum_{k\in F} \wedge\left(t-\tfrac{k}{2}\right),$$

where F is a finite subset of $\mathbb{Z}^d$. In fact, it can be taken to have 4^d summands. It is helpful to make a sketch of the situation in the 1D and 2D setting. With this equality, we achieve the desired result as follows

$$\begin{aligned}
\|T_x f\|_W &= \sum_{n\in\mathbb{Z}^d} \|T_x f\cdot\psi_n\|_\infty = \sum_{n\in\mathbb{Z}^d} \sup_{t\in\mathbb{R}^d} \left|f(t)\;\wedge\left(t+x-\tfrac{n}{2}\right)\right| \\
&= \sum_{n\in\mathbb{Z}^d} \sup_{t\in\mathbb{R}^d} \left|f(t) \sum_{k\in F} \wedge\left(t+x-\tfrac{n}{2}\right)\;\wedge\left(t-\tfrac{n-k}{2}\right)\right| \\
&= \sum_{n\in\mathbb{Z}^d} \sup_{t\in\mathbb{R}^d} \left|f(t) \sum_{k\in F} \wedge\left(t+x-\tfrac{n+k}{2}\right)\;\wedge\left(t-\tfrac{n}{2}\right)\right| \\
&\le \#F\,\|\wedge\|_\infty \sum_{n\in\mathbb{Z}^d} \sup_{t\in\mathbb{R}^d} \left|f(t)\;\wedge\left(t-\tfrac{n}{2}\right)\right| = 4^d\,\|f\|_W.
\end{aligned}$$

The argument for the continuity of the dilation operator is equivalent to the fact that different BUPUs define equivalent norms on the Wiener algebra. We omit the proof.

The reader may verify the following statement.

Lemma 3.6 *If* f *is a function in* $W(\mathbb{R}^d)$ *and* $h\in C_b(\mathbb{R}^d)$ *is such that* $|h(t)|\le|f(t)|$ *for all* $t\in\mathbb{R}^d$, *then* $h\in W(\mathbb{R}^d)$ *and* $\|h\|_W\le\|f\|_W$.

From Lemma 3.6, it is easy to prove the following implications: if f belongs to the Wiener algebra, then so does its absolute value, $|f|$, its real and imaginary part

$\Re(f)$ and $\Im(f)$, and in case f is real valued, also its positive and negative part f^+ and f^-,

$$|f|: t \mapsto |f(t)|, \;\; \Re(f): t \mapsto \Re(f(t)), \;\; \Im(f): t \mapsto \Im(f(t)),$$
$$f^+: t \mapsto \tfrac{1}{2}\big(|f(t)| + f(t)\big) \;\text{ and }\; f^-: t \mapsto \tfrac{1}{2}\big(|f(t)| - f(t)\big), \;\; t \in \mathbb{R}^d.$$

Let us turn to the obstacle that we encountered with the function space $C_0(\mathbb{R}^d)$: not every $f \in C_0(\mathbb{R}^d)$ can be embedded into $M_b(\mathbb{R}^d)$ and we could not define the convolution between arbitary functions in $C_0(\mathbb{R}^d)$. The function space $W(\mathbb{R}^d)$ can be completely embedded into $M_b(\mathbb{R}^d)$. Essential in this embedding is the key property of a function in the Wiener algebra to be integrable. The Riemann integral can be extended from $C_c(\mathbb{R})$ to a linear and continuous functional on $W(\mathbb{R}^d)$. That is,

$$I: W(\mathbb{R}^d) \to \mathbb{C}, \; I(f) = \int_{\mathbb{R}^d} f(t)\, dt, \; f \in W(\mathbb{R}^d), \tag{3.13}$$

is a well-defined linear functional satisfying $I(f) = I(T_x f)$, $x \in \mathbb{R}^d$. Actually,

$$\Big| \int_{\mathbb{R}^d} f(t)\, dt \Big| = |I(f)| \le I(|f|) \le \|f\|_W \;\text{ for all }\; f \in W(\mathbb{R}^d). \tag{3.14}$$

Proof of (3.14). Indeed, if we use the specific BUPU in (3.2), then we find

$$\Big| \int_{\mathbb{R}_d} f(t)\, dt \Big| = \Big| \int_{\mathbb{R}_d} \sum_{n \in \mathbb{Z}^d} f(t)\, \psi_n(t)\, dt \Big| \le \sum_{n \in \mathbb{Z}^d} \int_{\mathbb{R}_d} \big| f(t)\, \psi_n(t) \big|\, dt$$
$$= \sum_{n \in \mathbb{Z}^d} \int_{n + \left[-\frac{1}{2}, \frac{1}{2}\right]^d} |f(t)\, \psi_n(t)|\, dt \le \sum_{n \in \mathbb{Z}^d} \|f \psi_n\|_\infty = \|f\|_W.$$

For functions in the Wiener algebra, we define the L^1-norm to be

$$\|f\|_1 : W(\mathbb{R}^d) \to \mathbb{R}_0^+, \; \|f\|_1 = \int_{\mathbb{R}_d} |f(t)|\, dt.$$

The Riemann integral allows to embed the Wiener algebra $W(\mathbb{R}^d)$ into $M_b(\mathbb{R}^d)$:

$$\mu_k(f) = \int_{\mathbb{R}_d} f(t)\, k(t)\, dt, \quad f \in C_0(\mathbb{R}^d), k \in W(\mathbb{R}^d). \tag{3.15}$$

It is easy to show that $\|\mu\|_{M_b} \le \|k\|_W$ for all $k \in W(\mathbb{R}^d)$ (combine (3.14) and Lemma 3.4) and that the mapping $k \mapsto \mu_k$ from $W(\mathbb{R}^d)$ into $M_b(\mathbb{R}^d)$ is injective.

With this embedding, we define the convolution of two functions in the Wiener algebra: if $f, k \in W(\mathbb{R}^d)$, then their convolution product is defined to be

$$\big(k * f\big)(t) = \big(\mu_k * f\big)(t) = \int_{\mathbb{R}_d} f(t-s)\, k(s)\, ds,\ t \in \mathbb{R}^d. \tag{3.16}$$

Lemma 3.7 *The convolution defined in* (3.16) *turns* $(W(\mathbb{R}^d), \|\cdot\|_W)$ *into a commutative Banach algebra with respect to convolution. In fact,*

$$\|k * f\|_W \le 4^d\, \|k\|_W\, \|f\|_W \ \textit{for all}\ \ k, f \in W(\mathbb{R}^d). \tag{3.17}$$

Proof That the function $k * f$ is continuous follows from the fact that for any $f \in W(\mathbb{R}^d)$, the mapping $t \mapsto T_t f$ is continuous from $\mathbb{R}^d$ to $W(\mathbb{R}^d)$. We can easily establish that the Wiener algebra norm is finite: for all $f, k \in W(\mathbb{R}^d)$

$$\begin{aligned}
\sum_{n\in\mathbb{Z}^d} \|(k * f)\cdot \psi_n\|_\infty &= \sum_{n\in\mathbb{Z}^d} \sup_{t\in\mathbb{R}^d} \Big| \int_{\mathbb{R}_d} f(t-s)\, k(s)\, ds\, \psi_n(t)\Big| \\
&\le \int_{\mathbb{R}_d} |k(s)| \sum_{n\in\mathbb{Z}^d} \sup_{t\in\mathbb{R}^d} \big| f(t-s)\, \psi_n(t)\big|\, ds \\
&= \int_{\mathbb{R}_d} |k(s)|\, \|T_s f\|_W\, dt \le 4^d\, \|k\|_W\, \|g\|_W < \infty.
\end{aligned}$$

It is an easy application of Fubini's theorem that establishes the well-known inequality for the convolution in relation to the L^1-norm,

$$\|k * f\|_1 \le \|k\|_1\, \|f\|_1 \ \text{ for all }\ k, f \in W(\mathbb{R}^d). \tag{3.18}$$

Remark 3.1 This observation opens up the possibility to *define* $(L^1(\mathbb{R}^d), \|\cdot\|_1)$ within $(M_b(\mathbb{R}^d), \|\cdot\|_{M_b})$ as the closure of (the copy of) $C_c(\mathbb{R}^d)$ within $(M_b(\mathbb{R}^d), \|\cdot\|_{M_b})$, avoiding measure theory and Lebesgue integration completely. Even the Riemann–Lebesgue Theorem can be derived in this way. We do not pursue this idea further.

Remark 3.2 As every function in the Wiener algebra is integrable and uniformly bounded, it follows that $W(\mathbb{R}^d) \subset L^1(\mathbb{R}^d)$ and $W(\mathbb{R}^d) \subset C_b(\mathbb{R}^d) \subset L^\infty(\mathbb{R}^d)$. This implies that $W(\mathbb{R}^d)$ is a subspace of all the $L^p(\mathbb{R}^d)$ spaces for $p \in [1, \infty]$. Moreover, $\|f\|_p \le \|f\|_W$ for all $f \in W(\mathbb{R}^d)$ and all $p \in [1, \infty]$. Observe that $L^1(\mathbb{R}^d)$, just as $W(\mathbb{R}^d)$, is a Banach algebra with respect to convolution. Unlike $W(\mathbb{R}^d)$ however, $L^1(\mathbb{R}^d)$ is *not* a Banach algebra with respect to pointwise multiplication.

Lemma 3.8 *A function* $f \in C_b(\mathbb{R}^d)$ *belongs to* $W(\mathbb{R}^d)$ *if and only if* $f^\# \in W(\mathbb{R}^d)$ *and*

$$\|f\|_W \le \|f^\#\|_W \le 8^d\, \|f\|_W \ \textit{for all}\ \ f \in W(\mathbb{R}^d). \tag{3.19}$$

Proof The upper inequality follows by applying the same method as in the proof of Lemma 3.5 where we show that the translation operator is bounded on $W(\mathbb{R}^d)$. As $|f(t)| \le f^\#(t)$ for all $t \in \mathbb{R}^d$, the lower inequality follows by Lemma 3.6.

Lemma 3.9 *If $f \in W(\mathbb{R}^d)$, then $osc_\delta(f) \in W(\mathbb{R}^d)$ and $\lim_{\delta\to 0} \|osc_\delta(f)\|_{W(\mathbb{R}^d)} = 0$.*

Proof By Lemma 3.1 we have the inequality $\mathrm{osc}_\delta(f) \leq 2f^\#$. In Lemma 3.8, we established that $f \in W(\mathbb{R}^d)$ implies that also $f^\# \in W(\mathbb{R}^d)$. It follows from Lemma 3.6 that $\mathrm{osc}_\delta(f) \in W(\mathbb{R}^d)$. We leave the second statement as an exercise for the reader.

$W(\mathbb{R}^d) \subset C_0(\mathbb{R}^d)$ implies that existence of the usual convolution, given by

$$\big(\mu * f\big)(x) = \mu(T_x[f^{\checkmark}]), \quad \mu \in M_b(\mathbb{R}^d), f \in W(\mathbb{R}^d). \tag{3.20}$$

Clearly $\mu * f \in C_0(\mathbb{R}^d)$. For the claim $M_b(\mathbb{R}^d) * W(\mathbb{R}^d) \subset W(\mathbb{R}^d)$, we need a lemma.

Lemma 3.10 *For every compact set K, there exists a constant $c_K > 0$ such that for every function $f \in C_c(\mathbb{R}^d)$ with supp(f) $\subseteq K + x$, $x \in \mathbb{R}^d$ one has*

$$\|f\|_W \leq c_K \|f\|_\infty. \tag{3.21}$$

Proof From the definition of a BUPU, it follows that for any given compact set K there is a uniform bounded finite number of functions such that for all $x \in \mathbb{R}^d$ $\operatorname{supp} \psi_n \cap K \neq \emptyset$. Therefore, for any $f \in W(\mathbb{R}^d)$ with $\operatorname{supp} f \subset K + x$

$$\begin{aligned}\|f\|_W = \sum_{n\in\mathbb{Z}^d} \|f \cdot \psi_n\|_\infty &= \sum_{n\in\mathbb{Z}^d} \big(\sup_{t\in K+x} |f(t) \cdot \psi_n(t)|\big)\\ &\leq \Big(\sum_{n\in F_x} \|\psi_n\|_\infty\Big) \|f\|_\infty = c_K \|f\|_\infty,\end{aligned}$$

where c_K is this uniform bound in the number of elements in F_x.

Proposition 3.1 *We have $M_b(\mathbb{R}^d) * W(\mathbb{R}^d) \subset W(\mathbb{R}^d)$ and moreover there is a constant $c > 0$ such that*

$$\|\mu * f\|_W \leq c\, \|\mu\|_{M_b} \|f\|_W \ \textit{for all}\ \ \mu \in M_b(\mathbb{R}^d),\ \ f \in W(\mathbb{R}^d). \tag{3.22}$$

Proof We use the fact that both $\mu \in M_b(\mathbb{R}^d)$ and $f \in W(\mathbb{R}^d)$ have an absolutely convergent series representation if one applies a BUPU to each of them, i.e., $\mu = \sum_{n\in\mathbb{Z}^d} \mu \cdot \psi_n$ with $\|\mu\|_M = \sum_{n\in\mathbb{Z}^d} \|\mu \cdot \psi_n\|_M$ and $f = \sum_{k\in\mathbb{Z}^d} f \cdot \psi_k$ with $\|f\|_W = \sum_{k\in\mathbb{Z}^d} \|f \cdot \psi_k\|_\infty$. Observe that for each $k, n \in \mathbb{Z}^d$ the function

$$x \mapsto \big(\mu\psi_n * f\psi_k\big)(x) = \mu\psi_n([T_x f \psi_k]^{\checkmark})$$

is continuous and compactly supported, hence an element in $W(\mathbb{R}^d)$. Furthermore, due to the uniform size of the support of the BUPU (ψ_n) the functions, $\mu\psi_n * f\psi_k$, $k, n \in \mathbb{Z}^d$ all have support within $K + x$, where K is a fixed compact set and x

depends on k and n. With the BUPU as in (3.2) $K = [0,1]^d$. By Lemma 3.10 we have

$$\|\mu\psi_n * f\psi_k\|_W \leq c_K \, \|\mu\psi_n * f\psi_k\|_\infty \leq c_K \, \|\mu\psi_n\|_M \, \|f\psi_k\|_\infty.$$

Combining these inequalities allows us to deduce the desired estimate:

$$\begin{aligned}\|\mu * f\|_W &= \Big\| \big(\sum_{n\in\mathbb{Z}^d} \mu \cdot \psi_n \big) * \big(\sum_{k\in\mathbb{Z}^d} f \cdot \psi_k \big) \Big\|_W \\ &\leq \sum_{k,n\in\mathbb{Z}^d} \|(\mu \cdot \psi_n) * (f \cdot \psi_k)\|_W \\ &\leq c_K \sum_{k,n\in\mathbb{Z}^d} \|\mu\psi_n\|_M \, \|f\psi_k\|_\infty \\ &= c_K \, \|\mu\|_M \, \|f\|_W < \infty.\end{aligned}$$

For later use, we note the following result.

Lemma 3.11 *For any* $d, m \in \mathbb{N}$ *such that* $0 < m < d$*, the operator*

$$\mathscr{R}_m : W(\mathbb{R}^d) \to W(\mathbb{R}^m), \; \mathscr{R}_m f(x^{(1)}, \ldots, x^{(m)}) = f(x^{(1)}, \ldots, x^{(m)}, 0, \ldots, 0), \; x^{(i)} \in \mathbb{R}$$

is continuous. In fact, $\|\mathscr{R}_m f\|_{W(\mathbb{R}^m)} \leq \|f\|_{W(\mathbb{R}^d)}$ *for all* $f \in W(\mathbb{R}^d)$.

Proof The desired inequality is achieved as follows:

$$\begin{aligned}\|\mathscr{R}_m f\|_{W(\mathbb{R}^m)} &= \sum_{n\in\mathbb{Z}^m} \|\mathscr{R}_m f \cdot \psi_n^{(m)}\|_\infty \\ &= \sum_{n\in\mathbb{Z}^m} \sup_{t\in\mathbb{R}^m} |f(t,0) \cdot \psi_n^{(m)}(t)| \qquad (0 \in \mathbb{R}^{d-m}) \\ &\leq \sum_{n\in\mathbb{Z}^d} \sup_{t\in\mathbb{R}^d} |f(t) \cdot \psi_n^{(d)}(t)| = \|f\|_{W(\mathbb{R}^d)}.\end{aligned}$$

3.5 The Fourier Transform

As functions in the Wiener algebra are integrable (in the sense of *Riemann!*), we can use $W(\mathbb{R}^d)$ as the domain of the Fourier transform.

Definition 3.5 For $f \in W(\mathbb{R}^d)$ we define the *Fourier transform*,

$$\mathscr{F} f(s) = \hat{f}(s) = \int_{\mathbb{R}^d} f(t) \, e^{-2\pi i s \cdot t} \, dt, \quad s \in \mathbb{R}^d. \tag{3.23}$$

We mention the following classical result.

Lemma 3.12 (Riemann–Lebesgue Lemma) *The Fourier transform is a non-expansive and injective linear operator from* $(W(\mathbb{R}^d), \|\cdot\|_W)$ *into* $(C_0(\mathbb{R}^d), \|\cdot\|_\infty)$, *i.e.,*

$$\|\hat{f}\|_\infty \le \|f\|_1 \le \|f\|_W \text{ for all } f \in W(\mathbb{R}^d). \tag{3.24}$$

A cornerstone of our approach will be the following formula, which has been called *fundamental identity for the Fourier transform* by H. Reiter.

Theorem 3.3

$$\int_{\mathbb{R}_d} f(t)\,\hat{g}(t)\,dt = \int_{\mathbb{R}_d} \hat{f}(x)\,g(x)\,dx \text{ for all } f, g \in W(\mathbb{R}^d). \tag{3.25}$$

Equally important is the convolution theorem for the Fourier transform

$$\widehat{f * g} = \hat{f} \cdot \hat{g} \text{ for all } f, g \in W(\mathbb{R}^d), \tag{3.26}$$

Proof of (3.25) *and* (3.26). The Fourier transforms $\hat{f}$ and $\hat{g}$ are bounded and continuous. By Lemma 3.4 both integrands are in $W(\mathbb{R}^d)$ and thus integrable. The relation (3.25) follows via Fubini's theorem (which is easy to prove for Riemann integrals):

$$\begin{aligned} \int_{\mathbb{R}_d} f(t)\hat{g}(t)\,dt &= \int_{\mathbb{R}_d} f(t) \left(\int_{\mathbb{R}_d} e^{-2\pi i x\cdot t} g(x)\,dx \right) dt \\ &= \int_{\mathbb{R}_d} g(x) \left(\int_{\mathbb{R}_d} e^{-2\pi i x\cdot t} f(t)\,dt \right) dx \\ &= \int_{\mathbb{R}_d} \hat{f}(x)g(x)\,dx. \end{aligned} \tag{3.27}$$

The convolution theorem (3.26) is shown in a similar fashion, making use of the exponential law via the identity $e^{2\pi i s\cdot t} = e^{2\pi i s\cdot(t-y)} e^{2\pi x\cdot y}$.

The Riemann–Lebesgue lemma tells us that the Fourier transform of a function in the Wiener algebra is a function in $C_0(\mathbb{R}^d)$. As such, they are not necessarily integrable and we have the same issues with it as in Sect. 3.3 (which lead us to the Wiener algebra). Because we cannot guarantee that the Fourier transform of a function in the Wiener algebra is integrable, we cannot always apply the inverse Fourier transform (we also have to show that it is actually a transform which inverts the forward Fourier transform on the given domain),

$$\mathscr{F}^{-1} f(t) = \int_{\mathbb{R}_d} f(s)\, e^{2\pi i s\cdot t}\,dt, \quad t \in \mathbb{R}^d.$$

Therefore, we introduce the following Fourier invariant function space:

$$W_{\mathscr{F}}(\mathbb{R}^d) = \left\{ f \in W(\mathbb{R}^d) \;:\; \hat{f} \in W(\mathbb{R}^d) \right\}. \tag{3.28}$$

This space has been studied by Bürger in [4] (using the symbol $\mathscr{B}_0$). It is a Banach space with respect to the natural norm $\|f\|_{W_{\mathscr{F}}} = \|f\|_W + \|\hat{f}\|_W$. It is nontrivial and in fact dense in $(W(\mathbb{R}^d), \|\cdot\|_W))$ because it contains the Gauss function and all its shifted and modulated versions.

The Banach space $W_{\mathscr{F}}$ is well-suited for the formulation of results in Fourier analysis, such as the Fourier inversion theorem.

Theorem 3.4 *(i) For any* $f \in W_{\mathscr{F}}(\mathbb{R}^d)$ *the* Fourier inversion formula *holds pointwise,*

$$f(t) = \mathscr{F}^{-1}\hat{f}(t) = \int_{\mathbb{R}^d} \hat{f}(s)\, e^{2\pi i s\cdot t}\, ds \;\; \textit{for all} \;\; t \in \mathbb{R}^d. \tag{3.29}$$

(ii) For any pair of functions $f, g \in W_{\mathscr{F}}(\mathbb{R}^d)$ *the* Parseval identity *holds*

$$\int_{\mathbb{R}_d} \hat{f}(t)\, \overline{\hat{g}(t)}\, dt = \int_{\mathbb{R}_d} f(t)\, g(t)\, dt. \tag{3.30}$$

(iii) For any $f, g \in W_{\mathscr{F}}(\mathbb{R}^d)$*, we have the formula* $\widehat{f\cdot h} = \hat{f} * \hat{g}$*.*

(iv) For any $f \in W_{\mathscr{F}}(\mathbb{R}^d)$*, the Poisson formula holds pointwise: given* $m, d \in \mathbb{N}_0$ *with* $0 \le m \le d$ *and any non-singular* $d \times d$ *matrix* A*,*

$$\int_{\mathbb{R}^m} \sum_{k\in\mathbb{Z}^{d-m}} f(A(x,k))\, dx = \frac{1}{\det(A)} \sum_{k\in\mathbb{Z}^{d-m}} \hat{f}(A^{\dagger}(0,k)), \tag{3.31}$$

where $A^{\dagger}$ *is the inverse transpose of the matrix* A*.*

Proof We only prove (i), starting from the fundamental identity of Fourier analysis, (3.25). Denote by g_0 the Gaussian, with $g_0(t) = e^{-\pi t\cdot t}$. It has the remarkable property of *being invariant under the Fourier transform!* Consequently, due to properties of the Fourier transform, we have

$$\mathscr{F}(E_{\omega}\mathrm{D}_{\rho} g_0) = T_x \mathrm{St}_{\rho} g_0, \;\; x \in \mathbb{R}^d, \;\; \rho > 0. \tag{3.32}$$

In (3.25) we choose $g = \mathscr{F}(E_x \mathrm{D}_{\rho} g_0)$, and find that for any $f \in W_{\mathscr{F}}\mathbb{R}^d$,

$$f(x) = \lim_{\rho\to\infty} \int f(t)\, [T_x \mathrm{St}_{\rho} g_0](t)\, dt \overset{(3.25)}{=} \lim_{\rho\to\infty} \int \hat{f}(t)\, [E_x \mathrm{D}_{\rho} g_0](t)\, dt = \int \hat{f}(t)\, e^{2\pi i t x}\, dt. \tag{3.33}$$

The first limit is justified because $\int_{\mathbb{R}_d} h(x)\mathrm{St}_{\rho} g_0 = h(0)$ for any $h \in C_0(\mathbb{R}^d)$. If we apply this to $h = T_{-x} f \in W(\mathbb{R}^d) \subset C_0(\mathbb{R}^d)$, it results in the equality

$$f(x) = f(0+x) = T_{-x} f(0) = \lim_{\rho\to 0} \int_{\mathbb{R}_d} f(t+x) \mathrm{St}_{\rho} g_0(t) dt,$$

which is equal to the expression in the first limit. For the convergence of the second argument, we use the fact that $\hat{f} \in W(\mathbb{R}^d)$ by the density of $C_c(\mathbb{R}^d)$ in

$(W(\mathbb{R}^d), \|\cdot\|_W)$ one can restrict the attention to convergence of $\mathrm{D}_\rho g_0(t) \to 1$ for $\rho \to 0$, uniformly over compact sets. Details are left to the reader. Reading the left-hand side as a function of x, it is easily reinterpreted as $\mathrm{St}_\rho g_0 * f(x)$, which tends to $f(x)$ uniformly for any $f \in C_0(\mathbb{R}^d)$, but also in the Wiener norm for $f \in (W(\mathbb{R}^d), \|\cdot\|_W)$. A detailed proof of the Fourier invariance of the Gauss function can be found in Example 1.3.3 of [1] or in E. Stein's book ([46]).

The Poisson formula (3.31) is often "only" formulated as the Poisson *summation* formula. In this case we set $m = 0$ in (3.31) and obtain

$$\sum_{k\in\mathbb{Z}^d} f(Ak) = \frac{1}{\det(A)} \sum_{k\in\mathbb{Z}^d} \hat{f}(A^\dagger k). \tag{3.34}$$

If we apply (3.34) to the function $E_\omega T_x f$, $f \in W_{\mathscr{F}}(\mathbb{R}^d)$, then we find that

$$\sum_{k\in\mathbb{Z}^d} e^{2\pi i\, Ak\cdot\omega} f(Ak - x) = \frac{e^{2\pi i\, \omega\cdot x}}{\det(A)} \sum_{k\in\mathbb{Z}^d} e^{2\pi i\, A^\dagger k\cdot x} \hat{f}(A^\dagger k - \omega), \tag{3.35}$$

for any invertible $d \times d$ matrix A, any $x, \omega \in \mathbb{R}^d$ and any $f \in W_{\mathscr{F}}(\mathbb{R}^d)$. As a concrete example, we apply (3.35) to the Fourier invariant Gauss function $f(t) = e^{-\pi\, t\cdot t}$, $t \in \mathbb{R}^d$. This yields the equality

$$\sum_{k\in\mathbb{Z}^d} e^{-\pi\,(Ak\cdot Ak - 2\,Ak\cdot(x+i\omega))} = \frac{e^{\pi i\,(x+i\omega)^2}}{\det(A)} \sum_{k\in\mathbb{Z}^d} e^{-\pi\,(A^\dagger k\cdot A^\dagger k - 2\,A^\dagger k\cdot(\omega+ix))}. \tag{3.36}$$

In principle we could already start a "simplified distribution theory" on the basis of the function space $W_{\mathscr{F}}(\mathbb{R}^d)$, by considering its dual space as the reservoir of generalized functions. Indeed, the dual space of $W_{\mathscr{F}}(\mathbb{R}^d)$ already contains Dirac measure (point evaluation functionals) $\delta_{x_0}(f) : f(x_0)$, or integrable as well as bounded or periodic functions, and even objects like Dirac combs.

However, there is one drawback of this space: we cannot prove a *kernel theorem*, which is the "continuous analog" of the matrix representation of a linear mapping from $\mathbb{R}^n$ to $\mathbb{R}^m$ by matrix multiplication with a well-defined $m \times n$-matrix A, see Sect. 3.9. For this we need the *tensor factorization property* of the underlying Banach space of test functions. We will consider this property in the subsequent section by introducing an even smaller space of Banach algebra of test functions, the Segal algebra $(S_0(\mathbb{R}^d), \|\cdot\|_{S_0})$,[7] which satisfies all the properties that we are interested in.

[7] Also called *Feichtinger's algebra* in the literature.

3.6 Tensor Factorization

While the space of functions $W_{\mathscr{F}}$ is convenient for Fourier analysis, it is not suitable enough for our purposes as there is a crucial property we are interested in, namely the tensor factorization property. We explain it here for the space $W_{\mathscr{F}}$. This notion can be defined analogously for the other spaces we have considered so far, and also for the space S_0 that we will define in the next section.

Given two functions, $f^{(1)}, f^{(2)} \in W_{\mathscr{F}}(\mathbb{R}^m)$ their tensor product is

$$\big(f^{(1)} \otimes f^{(2)}\big)(x, y) = f^{(1)}(x) \cdot f^{(2)}(y), \quad (x, y) \in \mathbb{R}^n \times \mathbb{R}^m. \tag{3.37}$$

This function belongs to $W_{\mathscr{F}}(\mathbb{R}^{n+m})$, and there is some constant $c > 0$ such that

$$\|f^{(1)} \otimes f^{(2)}\|_{W_{\mathscr{F}}(\mathbb{R}^{n+m})} \le c \, \|f^{(1)}\|_{W_{\mathscr{F}}(\mathbb{R}^n)} \, \|f^{(2)}\|_{W_{\mathscr{F}}(\mathbb{R}^m)}, \tag{3.38}$$

for all $f^{(1)} \in W_{\mathscr{F}}(\mathbb{R}^n)$ and $f^{(2)} \in W_{\mathscr{F}}(\mathbb{R}^m)$.

With the help of tensor products, we can construct a new Banach space, the *projective tensor product* of $W_{\mathscr{F}}(\mathbb{R}^n)$ and $W_{\mathscr{F}}(\mathbb{R}^m)$,

$$W_{\mathscr{F}}(\mathbb{R}^n) \widehat{\otimes} W_{\mathscr{F}}(\mathbb{R}^m) = \Big\{F \in W_{\mathscr{F}}(\mathbb{R}^{n+m}) : F = \sum_{j=1}^{\infty} f_j^{(1)} \otimes f_j^{(2)}, \text{ and where} \\ \text{furthermore} \sum_{j=1}^{\infty} \|f_j^{(1)}\|_W \, \|f_j^{(2)}\|_W < \infty\Big\}.$$

The norm of a function $F \in W_{\mathscr{F}}(\mathbb{R}^n) \widehat{\otimes} W_{\mathscr{F}}(\mathbb{R}^m)$ is given by

$$\|F\|_{W_{\mathscr{F}}(\mathbb{R}^n) \widehat{\otimes} W_{\mathscr{F}}(\mathbb{R}^m)} = \inf\Big\{\sum_{j=1}^{\infty} \|f_j^{(1)}\|_{W_{\mathscr{F}}(\mathbb{R}^n)} \, \|f_j^{(2)}\|_{W_{\mathscr{F}}(\mathbb{R}^m)}\Big\},$$

where the infimum is taken over all possible representations of F of the type $\sum_{j=1}^{\infty} f_j^{(1)} \otimes f_j^{(2)}$ as described above. One can show that

$$W_{\mathscr{F}}(\mathbb{R}^n) \,\widehat{\otimes}\, W_{\mathscr{F}}(\mathbb{R}^m) \subsetneq W_{\mathscr{F}}(\mathbb{R}^{n+m}). \tag{3.39}$$

That is, the Banach space $W_{\mathscr{F}}$ does *not* have the tensor factorization property. If so, there would be an equal sign in (3.39).

We therefore ask the following: can we find a Banach space of functions that is well-suited for Fourier analysis (such as $W_{\mathscr{F}}$) and which does have the tensor factorization property.

	Banach space	convolution with bounded measures	integration	embedded into its dual space	domain for the Fourier transformation	Fourier inversion theorem	Fourier invariant	Poisson formula	tensor factorization property	kernel theorem
C_0		✓	–	–	–	–	–	–	–	–
W		✓	✓	✓	✓	–	–	–	–	–
$W_{\mathscr{F}}$		✓	✓	✓	✓	✓	✓	✓	–	–
S_0		✓	✓	✓	✓	✓	✓	✓	✓	✓

Fig. 3.1 An overview of some of the properties of the four Banach spaces of functions that we consider, $(C_0(\mathbb{R}^d), \|\cdot\|_\infty)$, $(W(\mathbb{R}), \|\cdot\|_W)$, $(W_{\mathscr{F}}\,\mathbb{R}^n, \|\cdot\|_{W_{\mathscr{F}}})$ and $(S_0(\mathbb{R}^d), \|\cdot\|_{S_0})$

3.7 The Feichtinger Algebra

In this section, we answer the question we posed in the last section. We define a Banach space of functions, to be denoted by $S_0(\mathbb{R}^d)$, that is very well-suited for Fourier analysis; it has the tensor factorization property and consequently allows for the formulation of a kernel theorem. It therefore is the Banach space of test functions that we wish for. Figure 3.1 gives an overview of this and the other spaces that we have considered so far. In relation to the much-used Schwartz space, we mention that it is a dense subspace of S_0. Functions in S_0, however, need not be differentiable.

First, we have to introduce the *Short-Time Fourier Transform* (or STFT) of a function with respect to a window function g. There are various different assumptions that ensure the pointwise existence (and continuity) of the STFT as a function over the *time–frequency plane* or *phase space*. We introduce it as follows.

For a function $g \in W(\mathbb{R}^d)$, the so-called Gabor window, which is typically a nonnegative, even function concentrated near zero, we define the *Short-Time Fourier transform* with respect to g of a function $f \in C_b(\mathbb{R}^d)$ to be the function

$$\begin{aligned}
&\mathscr{V}_g : C_b(\mathbb{R}^d) \to C_b(\mathbb{R}^{2d}),\\
&\mathscr{V}_g f(x,\omega) = \int_{\mathbb{R}^d} f(t)\,\overline{g(t-x)}\,e^{-2\pi i\omega t}\,dt = \mathscr{F}(f\cdot\overline{T_x g})(\omega), \quad x,\omega \in \mathbb{R}^d.
\end{aligned}$$

It is easy to see that the definition makes sense for $g \in W(\mathbb{R}^d),\ f \in C_b(\mathbb{R}^d)$ (still using the Riemann integral).[8] Fix $g_0(t) = e^{-\pi\, t\cdot t}$, $t \in \mathbb{R}^d$ to be the Gaussian.

Definition 3.6 The space $S_0(\mathbb{R}^d)$ consists of all functions $f \in C_b(\mathbb{R}^d)$ for which $\mathscr{V}_{g_0} f$ is a function in $W(\mathbb{R}^{2d})$.[9] It is endowed with the norm

[8] It is also a well-defined function in $C_b(\mathbb{R}^{2d})$, or for $g, f \in L^2(\mathbb{R}^d)$ making use of Lebesgue integration, the usual way of introducing the STFT.

[9] In the book [40], and since then, the space S_0 has been called the Feichtinger algebra.

$$\|\cdot\|_{S_0} : S_0(\mathbb{R}^d) \to \mathbb{R}_0^+, \ \|f\|_{S_0} = \int_{\mathbb{R}^{2d}} \left|\left(\mathscr{V}_{g_0} f\right)(x,\omega)\right| d(x,\omega) = \|\mathscr{V}_{g_0} f\|_1.$$

Observe that this norm is well-defined, as functions in the Wiener algebra are integrable (see Sect. 3.4).

Our goal is to establish the following key result.

Theorem 3.5 *The space $(S_0(\mathbb{R}^d), \|\cdot\|_{S_0})$ is a Banach space, which is isometrically invariant under the Fourier transform and time–frequency shifts, and in fact a Banach algebra under convolution as well as multiplication.*

We start by observing that $S_0(\mathbb{R}^d)$ is a subspace of Wiener's algebra.

Lemma 3.13 *(i) The Feichtinger algebra $S_0(\mathbb{R}^d)$ is a subspace of and continuously embedded into the Wiener algebra $W(\mathbb{R}^d)$.*
(ii) For any $f \in S_0(\mathbb{R}^d)$ it holds that $\|f\|_1 \leq \|f\|_{S_0}$ and $\|f\|_\infty \leq \|f\|_{S_0}$.
(ii) The mapping $S_0(\mathbb{R}^d) \to \mathbb{R}_0^+$, $f \mapsto \|\mathscr{V}_{g_0} f\|_{W(\mathbb{R}^{2d})}$ is an equivalent norm on $S_0(\mathbb{R}^d)$.

Proof Observe that for any $x, s \in \mathbb{R}^d$ we have

$$|f(x)\, g_0(s)| \leq \|f \cdot T_{s-x} g_0\|_\infty.$$

Since $f \in C_b(\mathbb{R}^d)$, $g_0 \in W(\mathbb{R}^d)$ and because the translation operator is continuous on $W(\mathbb{R}^d)$, it follows from Lemma 3.4 that $f \cdot T_{s-x} g_0 \in W(\mathbb{R}^d)$ for any $x, s \in \mathbb{R}^d$. Furthermore, by assumption f is such that

$$(x,\omega) \mapsto \mathscr{V}_{g_0} f(x,\omega) = \int_{\mathbb{R}_d} f(t)\, g_0(t-x)\, e^{-2\pi i x\cdot t}\, dt = \mathscr{F}(f \cdot T_x g_0)(\omega)$$

is a function in $W(\mathbb{R}^{2d})$. This implies, by Lemma 3.11, that for fixed $x \in \mathbb{R}^d$ the function $\omega \mapsto \mathscr{F}(f \cdot T_x g_0)(\omega)$ belongs to $W(\mathbb{R}^d)$ as well. We may therefore apply the Fourier inversion formula, so that, for any $x, s \in \mathbb{R}^d$,

$$\mathscr{F}^{-1}\mathscr{F}(f \cdot T_{s-x} g_0) = f \cdot T_{s-x} g_0.$$

By the Riemann–Lebesgue lemma

$$\|\mathscr{F}^{-1}\mathscr{F}(f \cdot T_{s-x} g_0)\|_\infty \leq \|\mathscr{F}(f \cdot T_{s-x} g_0)\|_W = \sum_{m\in\mathbb{Z}^d} \|\mathscr{F}(f \cdot T_{s-x} g_0) \cdot \psi_m\|_\infty. \tag{3.40}$$

A combination of the observed facts yields the inequality

$$|f(x)\, g_0(s)| \leq \sum_{m\in\mathbb{Z}^d} \|\mathscr{F}(f \cdot T_{s-x} g_0) \cdot \psi_m\|_\infty.$$

Hence

$$\sup_{x\in\mathbb{R}^d} |f(x)\, g_0(s)\psi_n(x)| \le \sum_{m\in\mathbb{Z}^d} \sup_{x,\omega\in\mathbb{R}^d} |\mathscr{F}(f \cdot T_{s-x}g_0)(\omega) \cdot \psi_m(\omega)\psi_n(x)|.$$

Summing over n, and using that the translation operator is continuous on $W(\mathbb{R}^d)$ allows us to deduce that

$$\sum_{n\in\mathbb{Z}^d} \|f \cdot \psi_n\|_\infty \, |g_0(s)| \le 4^d \sum_{n,m\in\mathbb{Z}^d} \left|\mathscr{V}_{g_0} f(x,\omega)\, \psi_n(x)\, \psi_m(\omega)\right| = 4^d \|\mathscr{V}_{g_0} f\|_W,$$

for any $s \in \mathbb{R}^d$ and $f \in S_0(\mathbb{R}^d)$. It follows that

$$\|f\|_W \le 4^d \, \|g_0\|_\infty^{-1} \, \|\mathscr{V}_{g_0} f\|_W = 4^d \, \|\mathscr{V}_{g_0} f\|_W. \tag{3.41}$$

We now show that there exists a constant $c > 0$ such that

$$\|\mathscr{V}_{g_0} f\|_W \le c \, \|f\|_{S_0} \;\text{ for all }\; f \in S_0(\mathbb{R}^d).$$

We first establish the following equality: for any $f \in S_0(\mathbb{R}^d)$ and $x, \omega \in \mathbb{R}^d$

$$\begin{aligned} &\int_{\mathbb{R}^{2d}} \mathscr{V}_{g_0} f(t,\xi)\, \overline{\mathscr{V}_{g_0}[E_\omega T_x g_0](t,\xi)}\, d(t,\xi) \\ &= \int_{\mathbb{R}^{2d}} \mathscr{F}(f \cdot T_t g_0)(\xi)\, \overline{\mathscr{F}([E_\omega T_x g_0] \cdot T_t g_0)(\xi)}\, d(t,\xi) \\ &\overset{(3.30)}{=} \int_{\mathbb{R}^{2d}} (f \cdot T_t g_0)(s)\, \overline{([E_\omega T_x g_0] \cdot T_t g_0)(s)}\, d(t,s) \\ &= \int_{\mathbb{R}_d} f(s)\, \overline{E_\omega T_x g_0(s)} \int_{\mathbb{R}_d} g_0(s-t)\, g_0(s-t)\, dt\, ds = 2^{-d/2}\, \mathscr{V}_{g_0} f(x,\omega). \end{aligned} \tag{3.42}$$

The use of (3.30) is justified as both $f \cdot T_t g_0$ and $\mathscr{F}(f \cdot T_t g_0)$ are functions in the Wiener algebra (as already establish earlier in the proof). We now observe the following:

$$\begin{aligned} \|\mathscr{V}_{g_0} f\|_W &= \sum_{m,n\in\mathbb{Z}^d} \sup_{x,\omega} \left|\mathscr{V}_{g_0} f(x,\omega)\, \psi_n(x)\, \psi_m(x)\,\right| \\ &\overset{(3.42)}{=} 2^{d/2} \sum_{m,n\in\mathbb{Z}^d} \sup_{x,\omega} \left| \int_{\mathbb{R}^{2d}} \mathscr{V}_{g_0} f(t,\xi)\, \overline{\mathscr{V}_{g_0}[E_\omega T_x g_0](t,\xi)}\, d(t,\xi)\, \psi_n(x)\, \psi_m(\omega) \right| \\ &\le 2^{d/2} \int_{\mathbb{R}^{2d}} |\mathscr{V}_{g_0} f(t,\xi)\, \|T_{t,\xi} \mathscr{V}_{g_0} g_0\|_W \, d(t,\xi) \\ &\le 2^{9d/2} \|\mathscr{V}_{g_0} g_0\|_W \int_{\mathbb{R}^{2d}} |\mathscr{V}_{g_0} f(t,\xi)|\, d(t,\xi) = 2^{9d/2}\, \|\mathscr{V}_{g_0} g_0\|_W \, \|f\|_{S_0}. \end{aligned}$$

The second equality follows by the boundedness of the translation operator on the Wiener algebra. Combining the just established inequality with (3.41) yields

$$\|f\|_W \leq 2^{13d/2}\,\|\mathscr{V}_{g_0}g_0\|_W\,\|f\|_{S_0} \text{ for all } f \in S_0(\mathbb{R}^d).$$

Furthermore, we have just established that

$$\|\mathscr{V}_{g_0}f\|_W \leq 2^{9d/2}\,\|\mathscr{V}_{g_0}g_0\|_W\,\|f\|_{S_0} \text{ for all } f \in S_0(\mathbb{R}^d).$$

The inequality $\|f\|_{S_0} \leq \|\mathscr{V}_{g_0}f\|_W$ is clear from (3.14). We have thus shown (i) and (iii). In order to show (ii) we replace (3.40) with the inequality

$$\|\mathscr{F}^{-1}\mathscr{F}(f\cdot T_{s-x}g_0)\|_\infty \leq \|\mathscr{F}(f\cdot T_{s-x}g_0)\|_1,$$

and make similar steps as before. We then obtain the estimate

$$|f(x)g_0(s)| \leq \int_{\mathbb{R}^d}\big|\mathscr{V}_{g_0}f(s-x,\omega)\big|\,d\omega \text{ for all } x,s\in\mathbb{R}^d.$$

An integration over $x\in\mathbb{R}^d$ and taking the supremum over s yields

$$\|f\|_1\,\|g_0\|_\infty \leq \int_{\mathbb{R}^{2d}}\big|\mathscr{V}_{g_0}f(s-x,\omega)\big|\,d(x,\omega) = \|f\|_{S_0}.$$

Switching the role of x and s implies the inequality $\|f\|_\infty \leq \|f\|_{S_0}$. This shows (ii).

As every function in $S_0(\mathbb{R}^d)$ belongs to $W(\mathbb{R}^d)$, we can apply the Fourier transform to the space $S_0(\mathbb{R}^d)$. It turns out that $S_0(\mathbb{R}^d)$ is invariant under the Fourier transform.

Proposition 3.2 *The Fourier transform is an isometric bijection from $S_0(\mathbb{R}^d)$ onto itself, i.e., $\|\mathscr{F}f\|_{S_0} = \|f\|_{S_0}$ for all $f\in S_0(\mathbb{R}^d)$.*

Corollary 3.2 *$S_0(\mathbb{R}^d)$ is continuously embedded into $W_{\mathscr{F}}(\mathbb{R}^d)$.*

That $S_0(\mathbb{R}^d)$ is a proper subspace of $W_{\mathscr{F}}(\mathbb{R}^d)$ was shown by Losert [35, Theorem 2]. Observe that the inclusion $S_0(\mathbb{R}^d)\subset W_{\mathscr{F}}(\mathbb{R}^d)$ implies that all the statements in relation to the Fourier transform in Sect. 3.5 also hold for all functions in $S_0(\mathbb{R}^d)$.
Proof of Proposition 3.2. First of all $S_0(\mathbb{R}^d)\subset W(\mathbb{R}^d)$ so that $\mathscr{F}f$ is a well-defined function in $C_0(\mathbb{R}^d)$. Since $g_0\in W(\mathbb{R}^d)$ and $S_0(\mathbb{R}^d)\subset W(\mathbb{R}^d)$, we can use the fundamental identity of Fourier analysis to establish the following:

$$\begin{aligned}\mathscr{V}_{g_0}\hat{f}(x,\omega) &= \int_{\mathbb{R}^d}\hat{f}(t)\,\overline{g_0(t-x)}e^{-2\pi i\omega t}\,dt \overset{(3.25)}{=} \int_{\mathbb{R}^d} f(t)\,\mathscr{F}\big(\overline{E_\omega T_x g_0}\big)(t)\,dt\\ &= e^{-2\pi i x\cdot\omega}\,\mathscr{V}_{g_0}f(-\omega,x).\end{aligned}$$

Observe that the phase factor $e^{2\pi i x\cdot\omega}$ and also the change of variable $(x,\omega)\mapsto(-\omega,x)$ are continuous operators on the Wiener algebra so that also $\mathscr{V}_{g_0}\hat{f}$ belongs to $W(\mathbb{R}^{2d})$. Moreover, the operations leave the S_0-norm invariant. Indeed,

$$\begin{aligned}\|\hat{f}\|_{S_0} &= \int_{\mathbb{R}^{2d}} |\mathscr{V}_{g_0}\hat{f}(x,\omega)|\,d(x,\omega) = \int_{\mathbb{R}^{2d}} |e^{-2\pi i x\cdot\omega}\,\mathscr{V}_{g_0}f(-\omega,x)|\,d(x,\omega)\\ &= \int_{\mathbb{R}^{2d}} |\mathscr{V}_{g_0}f(x,\omega)|\,d(x,\omega) = \|f\|_{S_0}.\end{aligned}$$

The same proof shows that also the inverse Fourier transform maps $S_0(\mathbb{R}^d)$ into itself. It is therefore clear that $\mathscr{F}$ is a continuous bijection on $S_0(\mathbb{R}^d)$.

Concerning the continuity properties of the translation and modulation operator, we easily establish the following.

Lemma 3.14 *(i) Translation and modulation operators are isometries on $S_0(\mathbb{R}^d)$:*

$$\|E_\omega T_x f\|_{S_0} = \|f\|_{S_0} \ \textit{for all}\ \ x,\omega\in\mathbb{R}^d \ \ \textit{and}\ \ f\in S_0(\mathbb{R}^d). \tag{3.43}$$

(ii) If f belongs to $S_0(\mathbb{R}^d)$, then so does $\overline{f}$ and $f^{\checkmark}$ and

$$\|\overline{f}\|_{S_0} = \|f^{\checkmark}\|_{S_0} = \|f\|_{S_0} \ \textit{for all}\ \ f\in S_0(\mathbb{R}^d). \tag{3.44}$$

Proof Observe that

$$\mathscr{V}_{g_0}(E_\omega T_x f)(t,s) = e^{2\pi i\,x\cdot(\omega-s)}\,\mathscr{V}_{g_0}f(t-x,s-\omega). \tag{3.45}$$

Since translation and the phase factor leave the Wiener algebra invariant, it follows that $\mathscr{V}_{g_0}E_\omega T_x f\in W(\mathbb{R}^{2d})$. Hence $E_\omega T_x f\in S_0(\mathbb{R}^d)$ and moreover

$$\begin{aligned}\|E_\omega T_x f\|_{S_0} &= \int_{\mathbb{R}^{2d}} |\mathscr{V}_{g_0}(E_\omega T_x g_0)(t,s)|\,d(x,\omega)\\ &= \int_{\mathbb{R}^{2d}} |e^{2\pi i\,x\cdot(\omega-s)}\,\mathscr{V}_{g_0}f(t-x,s-\omega)|\,d(t,s)\\ &= \int_{\mathbb{R}^{2d}} |\mathscr{V}_{g_0}f(t,s)\,d(t,s) = \|f\|_{S_0}\end{aligned}$$

for any pair $(x,\omega)\in\mathbb{R}^{2d}$. The statement in (ii) is shown in a similar fashion.

Just as the Wiener algebra and $W_{\mathscr{F}}$, also S_0 behaves in a nice way with respect to multiplication and convolution.

Lemma 3.15 *The Banach space $(S_0(\mathbb{R}^d),\|\cdot\|_{S_0})$ is a Banach algebra with respect to pointwise multiplication and convolution. Indeed, for any $f_1, f_2\in S_0(\mathbb{R}^d)$, the functions $f_1\cdot f_2$ and $f_1 * f_2$ also belong to $S_0(\mathbb{R}^d)$ and*

$$\|f_1\cdot f_2\|_{S_0} \le \|f_1\|_{S_0}\,\|f_2\|_{S_0} \ \ \textit{and}\ \ \|f_1 * f_2\|_{S_0} \le \|f_1\|_{S_0}\,\|f_2\|_{S_0}.$$

Proof Let us first establish $f_1\cdot f_2$ belongs to $S_0(\mathbb{R}^d)$.

$$\begin{aligned}
\|\mathscr{V}_{g_0}(f_1\cdot f_2)\|_W &= \sum_{m,n\in\mathbb{Z}^d}\sup_{x,\omega\in\mathbb{R}^d}\big|\mathscr{F}(f_1\cdot f_2\cdot T_x g_0)(\omega)\,\psi_n(x)\,\psi_m(\omega)\big|\\
&= \sum_{m,n\in\mathbb{Z}^d}\sup_{x,\omega\in\mathbb{R}^d}\big|\mathscr{F}(f_1)*\mathscr{F}(f_2\cdot T_x g_0)(\omega)\,\psi_n(x)\,\psi_m(\omega)\big|\\
&= \sum_{m,n\in\mathbb{Z}^d}\sup_{x,\omega\in\mathbb{R}^d}\Big|\int_{\mathbb{R}_d}\mathscr{F}(f_2\cdot T_x g_0)(\omega-t)\,\hat{f}_1(t)\,dt\,\psi_n(x)\,\psi_m(\omega)\Big|\\
&\le \sum_{m,n\in\mathbb{Z}^d}\sup_{x,\omega\in\mathbb{R}^d}\int_{\mathbb{R}_d}|\mathscr{F}(f_2\cdot T_x g_0)(\omega-t)\,\psi_m(\omega)|\,|\hat{f}_1(t)|\,dt\,\psi_n(x)\\
&\le \sum_{m,n\in\mathbb{Z}^d}\sup_{x,\omega\in\mathbb{R}^d}\int_{\mathbb{R}_d}\sup_{\omega\in\mathbb{R}^d}|\mathscr{F}(f_2\cdot T_x g_0)(\omega)\,\psi_m(\omega+t)|\,|\hat{f}_1(t)|\,dt\,\psi_n(x)\\
&\le 4^d\sum_{m,n\in\mathbb{Z}^d}\sup_{x\in\mathbb{R}^d}\int_{\mathbb{R}_d}\sup_{\omega\in\mathbb{R}^d}|\mathscr{F}(f_2\cdot T_x g_0)(\omega)\,\psi_m(\omega)|\,|\hat{f}_1(t)|\,dt\,\psi_n(x)\\
&\le 4^d\|\hat{f}_1\|_W\,\|f_2\|_{S_0}\le 16^d\,\|f_1\|_{S_0}\,\|f_2\|_{S_0}<\infty.
\end{aligned}$$

In the third inequality, we used the same method as in the proof of Lemma 3.5 to get rid of the translation by x. The inequality for the convolution follows by the just established inequality, the equality $\mathscr{F}(f_1\cdot f_2)=\hat{f}_1*\hat{f}_2$, and the fact that the Fourier transform is a bijection on S_0. We have thus established that $\mathscr{V}_{g_0}(f_1\cdot f_2)\in W(\mathbb{R}^{2d})$ and $\mathscr{V}_{g_0}(f_1*f_2)\in W(\mathbb{R}^{2d})$, i.e., the convolution and pointwise product of f_1, f_2 belong to S_0 again. Concerning the desired estimates, we find that

$$\begin{aligned}
\|f_1*f_2\|_{S_0} &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\big|\mathscr{F}([f_1*f_2]\cdot T_x g_0)(\omega)\big|\,dx\,d\omega\\
&= \int_{\mathbb{R}_d}\int_{\mathbb{R}_d}\big|\big(f_1*f_2*E_\omega g_0\big)(x)\big|\,dx\,d\omega\\
&\le \|f_1\|_1\int_{\mathbb{R}_d}\int_{\mathbb{R}_d}|\big(f_2*E_\omega g_0\big)(x)\,dx\,d\omega\\
&= \|f_1\|_1\,\|f_2\|_{S_0}\le\|f_1\|_{S_0}\,\|f_2\|_{S_0}.
\end{aligned}$$

The first inequality is an application of (3.18). The second inequality follows by Lemma 3.13(ii). The inequality for the pointwise product follows by properties of the Fourier transform as mentioned before.

Among other useful properties of $S_0(\mathbb{R}^d)$ are the following ones. In particular, S_0 has the tensor factorization property.

Theorem 3.6 *(i) For any invertible $d\times d$ matrix A, the operator*

$$\alpha_A: S_0(\mathbb{R}^d)\to S_0(\mathbb{R}^d),\ \alpha_A f(x)=|\det(A)|^{1/2}f(Ax),\ \ x\in\mathbb{R}^d$$

is a continuous bijection on $S_0(\mathbb{R}^d)$.

(ii) For any $m\in\mathbb{N}$ such that $0<m<d$, the operator

$$\mathscr{R}_m: S_0(\mathbb{R}^d)\to S_0(\mathbb{R}^m),\ \mathscr{R}_m f(x^{(1)},\dots,x^{(m)})=f(x^{(1)},\dots,x^{(m)},0,\dots,0),\ x^{(i)}\in\mathbb{R}$$

is a continuous surjection.

(iii) The sampling of a function on $\mathbb{R}^d$ *at the integer-lattice points* $\mathbb{Z}^d$

$$\mathscr{R}_{\mathbb{Z}^d} : S_0(\mathbb{R}^d) \to \ell^1(\mathbb{Z}^d),\ \mathscr{R}_{\mathbb{Z}^d} f(k) = f(k),\ \ k \in \mathbb{Z}^d$$

is a continuous and surjective operator from $S_0(\mathbb{R}^d)$ *onto* $\ell^1(\mathbb{Z}^d)$*.*

(iv) For any $m \in \mathbb{N}$ *such that* $0 < m < d$ *the operator*

$$\begin{aligned}&\mathscr{P}_m : S_0(\mathbb{R}^d) \to S_0(\mathbb{R}^m),\\ &\mathscr{P}_m f(x) = \int_{\mathbb{R}^{n-m}} f(x^{(1)}, \ldots, x^{(m)}, x^{(m+1)}, \ldots, x^{(n)})\, dx^{(m+1)} \ldots dx^{(n)},\\ &x = (x^{(1)}, \ldots, x^{(m)}) \in \mathbb{R}^m\end{aligned}$$

is a continuous surjection.

(v) The periodization of functions on $\mathbb{R}^d$ *with respect to the integer lattice* $\mathbb{Z}^n$

$$\mathscr{P}_{\mathbb{Z}^n} : S_0(\mathbb{R}^d) \to A([0,1]^n),\ \mathscr{P} f(x) = \sum_{k \in \mathbb{Z}^n} f(x+k),\ \ x \in [0,1]^n$$

is a continuous and surjective operator from $S_0(\mathbb{R}^d)$ *onto* $A([0,1]^n)$*, the space of all* $\mathbb{Z}^n$*-periodic functions with absolutely summable Fourier coefficients.*

(vi) $S_0(\mathbb{R}^n) \widehat{\otimes} S_0(\mathbb{R}^m) = S_0(\mathbb{R}^{n+m})$ *for any* $n, m \in \mathbb{N}$*.*

Proof We are not in the position to give a proof, as this requires more theory and details about S_0 than we are willing to give here. The statements all follow from [14, Theorem 7].

To highlight the role of $S_0(\mathbb{R}^d)$ among all Banach spaces of functions within $W(\mathbb{R}^d)$, we give the following characterization. It is a direct consequence of [32, Theorem 7.6]

Theorem 3.7 *For each* $d \in \mathbb{N}$ *let* $(B(\mathbb{R}^d), \|\cdot\|_B)$ *be a nontrivial Banach space such that* $B(\mathbb{R}^d) \subseteq W(\mathbb{R}^d)$*. If for each* $d \in \mathbb{N}$ *the Banach space* $B(\mathbb{R}^d)$ *has the properties that*

(i) there is a constant $c > 0$ *such that* $\|f\|_{W(\mathbb{R}^d)} \le c\,\|f\|_{B(\mathbb{R}^d)}$ *for all* $f \in B(\mathbb{R}^d)$*,*

(ii) for all $(x, \omega) \in \mathbb{R}^{2d}$*, the time–frequency shift operators* $E_\omega T_x$ *is bounded on* $B(\mathbb{R}^d)$ *with a uniformly bounded operator norm over all* $(x, \omega) \in \mathbb{R}^{2d}$*,*

(iii) for every invertible $d \times d$ *matrix* A*, the operator* $f \mapsto f \circ A$ *is bounded on* $B(\mathbb{R}^d)$*,*

(iv) the Fourier transform is a bounded operator from $B(\mathbb{R}^d)$ *into* $W(\mathbb{R}^d)$*,*

(v) and $B(\mathbb{R}^n) \widehat{\otimes} B(\mathbb{R}^m) = B(\mathbb{R}^{n+m})$ *for all* $n, m \in \mathbb{N}$*,*

then $(B(\mathbb{R}^d), \|\cdot\|_B) = (S_0(\mathbb{R}^d), \|\cdot\|_{S_0})$ *for all* $d \in \mathbb{N}$*.*

3.8 The Shortcut to Distribution Theory

In the previous sections, we described several Banach spaces of continuous functions on $\mathbb{R}^d$ that have useful properties. Figure 3.1 gives a brief overview. Based on this, we recognize S_0 as a useful space of test functions. It has all the properties that we wish for. We will consider its dual space $(S_0'(\mathbb{R}^d), \| \cdot \|_{S_0'})$ as a suitably large reservoir of "everything else" that is worth to investigate. We call elements in S_O' for *distributions*.

The *shortcut* to distribution theory here is the fact that we have established a useful Banach space as our space of test functions. Hence, we do not require the more technical details that are typically needed to properly understand the Fréchet space formed by the Schwartz functions. Similarly, the dual space, here the Banach space $S_0'(\mathbb{R}^d)$ is also much more convenient that the space of tempered distributions (the dual of the Schwartz space). Ergo, with less mathematical effort we can describe and achieve much of the same types of results that the Schwartz space and the temperate distributions are typically used for.

One of the most important concepts of the dual space is that it is possible to extend operators that act on S_0 to operators that act on S_0'. In particular, the properties of S_0 allow us to define the Fourier transform of elements in S_0' (this is also possible to do with $W_{\mathscr{F}}$ and $W_{\mathscr{F}}'$). Before we get to this, we need to introduce S_0' properly.

The dual space $S_0'(\mathbb{R}^d)$, consists of bounded, linear functionals $\sigma : S_0(\mathbb{R}^d) \to \mathbb{C}$. It is a Banach space with respect to the usual functional norm

$$\|\sigma\|_{S_0'} = \sup_{f \in S_0(\mathbb{R}^d),\ \|f\|_{S_0}=1} |\sigma(f)|. \tag{3.46}$$

This topology is often too strong. Another weaker, yet at least as natural topology on S_0' is the topology it inherits from S_0: we say that a sequence (σ_n) in $S_0'(\mathbb{R}^d)$ converges in the weak*topology toward $\sigma_0 \in S_0'(\mathbb{R}^d)$ exactly if

$$\lim_n \big|\big(\sigma_n - \sigma_0\big)(f)\big| = 0 \ \text{ for all } \ f \in S_0(\mathbb{R}^d). \tag{3.47}$$

Now every $h \in C_b(\mathbb{R}^d)$ (and many more) defines a distribution $\sigma_h \in S_0'(\mathbb{R}^d)$ via the injective embedding operator

$$\iota : C_b(\mathbb{R}^d) \to S_0'(\mathbb{R}^d),\ \iota(k) = \sigma_h = f \mapsto \int_{\mathbb{R}_d} f(t)\, h(t)\, dt. \tag{3.48}$$

Also, any $\mu \in M_b(\mathbb{R}^d)$ defines a distribution $\sigma_\mu \in S_0'(\mathbb{R}^d)$ by the rule

$$\sigma_\mu(f) = \mu(f) \ \text{ for all } \ f \in S_0(\mathbb{R}^d).$$

The mapping $\mu \mapsto \sigma_\mu$ provides a continuous embedding $M_b(\mathbb{R}^d)$ into $S_0'(\mathbb{R}^d)$.

Definition 3.7 Assume T is a continuous operator from $S_0(\mathbb{R}^d)$ into $S_0(\mathbb{R}^d)$. We say that the operator $\widetilde{T} : S_0'(\mathbb{R}^d) \to S_0'(\mathbb{R}^d)$ is an extension of T if the following holds

(i) $\widetilde{T}$ is weak*-weak* continuous and
(ii) $\widetilde{T} \circ \iota(k) = \iota \circ T(k)$ (or, equivalently, $\widetilde{T}\sigma_k = \sigma_{Tk}$) for all $k \in S_0(\mathbb{R}^d)$.

Lemma 3.16 *The Fourier transform $\mathscr{F}$, translation operator T_x, $x \in \mathbb{R}^d$, modulation operator E_ω, $\omega \in \mathbb{R}^d$, and the coordinate transform α_A, $A \in \mathrm{GL}_d(\mathbb{R})$ are extended from operators on $S_0(\mathbb{R}^d)$ to operators on $S_0'(\mathbb{R}^d)$ in the following way: for any $f \in S_0(\mathbb{R}^d)$ and $\sigma \in S_0'(\mathbb{R}^d)$*

$$\begin{aligned}
\widetilde{\mathscr{F}} &: S_0'(\mathbb{R}^d) \to S_0'(\mathbb{R}^d),\ \big(\widetilde{\mathscr{F}}\sigma\big)(f) = \sigma(\mathscr{F} f),\\
\widetilde{T}_x &: S_0'(\mathbb{R}^d) \to S_0'(\mathbb{R}^d),\ \big(\widetilde{T}_x\sigma\big)(f) = \sigma(T_{-x} f),\\
\widetilde{E}_\omega &: S_0'(\mathbb{R}^d) \to S_0'(\mathbb{R}^d),\ \big(\widetilde{E}_\omega\sigma\big)(f) = \sigma(E_\omega f),\\
\widetilde{\alpha}_A &: S_0'(\mathbb{R}^d) \to S_0'(\mathbb{R}^d),\ \big(\widetilde{\alpha}_A\sigma\big)(f) = \sigma(\alpha_{A^{-1}} f).
\end{aligned}$$

Proof We only show the result for the Fourier transform. The statements for the other operators are proven in the same fashion. We have to show that $\widetilde{F}$ satisfies Definition 3.7. In order to show the weak*-weak*continuity, let (σ_n) be a sequence in $S_0'(\mathbb{R}^d)$ that converges in the weak*-sense toward σ_0. We have to show that then also $\widetilde{\mathscr{F}}\sigma_n \xrightarrow{\text{w}^*} \widetilde{\mathscr{F}}\sigma_0$. This follows easily from the definition of $\widetilde{\mathscr{F}}$,

$$\begin{aligned}
&\lim_n \big|\big(\widetilde{\mathscr{F}}\sigma_n - \widetilde{\mathscr{F}}\sigma_0\big)(f)\big| = \lim_n \big|\big(\widetilde{\mathscr{F}}(\sigma_n - \sigma_0)\big)(f)\big|\\
&= \lim_n \big|\big(\sigma_n - \sigma_0\big)(\mathscr{F} f)\big| = 0,
\end{aligned}$$

where the last equality follows by assumption. It remains to show that Definition 3.7(ii) is satisfied. We observe that for all $f, k \in S_0(\mathbb{R}^d)$

$$\begin{aligned}
&\big(\widetilde{\mathscr{F}} \circ \iota(k)\big)(f) = \big(\iota(k)\big)(\mathscr{F} f) = \int_{\mathbb{R}_d} \hat{f}(t)\, k(t)\, dt\\
&\big(\iota \circ \mathscr{F}(k)\big)(f) = \int_{\mathbb{R}_d} f(t)\, \hat{k}(t)\, dt.
\end{aligned}$$

It follows from (3.25) that the latter two integrals are the same so that $\widetilde{\mathscr{F}} \circ \iota(k) = \iota \circ \mathscr{F}(k)$, as desired.

Consider the Dirac delta,

$$\delta_x : S_0(\mathbb{R}^d) \to \mathbb{C}, \quad \delta_x(f) = f(x),\ x \in \mathbb{R}^d.$$

It is easy to show that $\widehat{\delta_x} = \widetilde{\mathscr{F}}\delta_x$ is the distribution given by

$$\widetilde{\mathscr{F}}\delta_x : S_0(\mathbb{R}^d) \to \mathbb{C}, \quad \widetilde{\mathscr{F}}\delta_x(f) = \hat{f}(x) = \int_{\mathbb{R}_d} f(t)\, e^{-2\pi i x\cdot t}\, dt.$$

Or, equivalently, $\widetilde{\mathscr{F}}\delta_x = \iota(e_x)$, where $e_x \in C_b(\mathbb{R}^d)$ is given by $e_x(t) = e^{-2\pi i x \cdot t}$. This can be formulated as to say that "the Fourier transform of the Dirac delta distribution at x, δ_x, is the function $e_x(t) = e^{-2\pi i t \cdot x}$". Or, equivalently, "the Fourier transform of the function $e_x(t) = e^{2\pi i t \cdot x}$, $t \in \mathbb{R}^d$, is the Dirac delta distribution at x, δ_x".

Remark 3.3 This is the characteristic property of the Fourier transform: it maps pure frequencies into Dirac measures and vice versa (see [37], (4.36)).

Consider now the *Dirac comb* or *Shah distribution* for a given invertible $d \times d$ matrix A, it is the element of $S_0'(\mathbb{R}^d)$ defined by

$$\sqcup\!\sqcup_A : S_0(\mathbb{R}^d) \to \mathbb{C}, \ \sqcup\!\sqcup_A(f) = \sum_{k \in \mathbb{Z}^d} f(Ak).$$

By definition of $\widetilde{\mathscr{F}}$ and a use of the Poisson summation formula (3.34), one gets

$$\widetilde{\mathscr{F}}(\sqcup\!\sqcup_A) = |\det(A)|^{-1} \sqcup\!\sqcup_{A^\dagger}.$$

We define *multiplication* and *convolution* of a distribution $\sigma \in S_0'(\mathbb{R}^d)$ *with a test function* $g \in S_0(\mathbb{R}^d)$ to be the distribution $\sigma \in S_0'(\mathbb{R}^d)$ defined as follows.

Definition 3.8

$$\big(\sigma * g\big)(f) = \sigma(g^{\checkmark} * f) \ \text{ and } \ \big(\sigma \cdot g\big)(f) = \sigma(g \cdot f) \qquad f \in S_0(\mathbb{R}^d).$$

The definition of the convolution is consistent with the definition

$$(\sigma * g)(t) = \sigma(T_t g^{\checkmark}), \ \ t \in \mathbb{R}^d.$$

Consequently we have $S_0(\mathbb{R}^d) * S_0'(\mathbb{R}^d) \subset C_b(\mathbb{R}^d)$, viewed as a subspace of $S_0'(\mathbb{R}^d)$. Observe that $\sqcup\!\sqcup_A * g$ equals the A-period function in $C_b(\mathbb{R}^d)$ given by

$$\big(\sqcup\!\sqcup_A * g\big)(t) = \sum_{k \in \mathbb{Z}^d} g(t + Ak), \ \ t \in \mathbb{R}^d,$$

where the convergence of the series is uniform and absolute within $(C_b(\mathbb{R}^d), \|\cdot\|_\infty)$. Furthermore, one can show that

$$\widetilde{\mathscr{F}}(\sigma * g) = (\widetilde{\mathscr{F}}\sigma) \cdot (\mathscr{F}g), \ \ \widetilde{\mathscr{F}}(\sigma \cdot g) = \widetilde{\mathscr{F}}\sigma * \mathscr{F}g. \tag{3.49}$$

We shall use these relations in Sect. 3.10, where we take a look at the Shannon sampling theorem.

Proof of (3.49). This follows by the definition of the extended Fourier transform and the convolution theorem: for any $\sigma \in S_0'(\mathbb{R}^d)$ and $g, f \in S_0(\mathbb{R}^d)$

$$\begin{aligned}\big(\widetilde{\mathscr{F}}[\sigma * g]\big)(f) &= (\sigma * g)(\mathscr{F} f) = \sigma(g^{\checkmark} * \mathscr{F} f)\\ &= \sigma\big([\mathscr{F}\mathscr{F}^{-1} g^{\checkmark}] * \mathscr{F} f\big) = \sigma\big(\mathscr{F}[\mathscr{F}^{-1} g^{\checkmark} \cdot f]\big)\\ &= \widetilde{\mathscr{F}}\sigma(\mathscr{F} g \cdot f) = \big(\widetilde{\mathscr{F}}\sigma \cdot \mathscr{F} g\big)(f).\end{aligned}$$

The proof of the other equality is done in the same spirit.

3.9 The Kernel Theorem

The reason why $W_{\mathscr{F}}(\mathbb{R}^d)$ is not quite good enough to be our Banach space of test functions is that it does not allow for the formulation of a kernel theorem. For this we have to turn to $S_0(\mathbb{R}^d)$.

The kernel theorem is the continuous analog of the matrix representation for linear mappings from $\mathbb{R}^n$ to $\mathbb{R}^m$, showing that they are represented in a unique way through matrix multiplication. Recalling that such a linear mapping T takes the form $T(\mathbf{x}) = \mathbf{A} \cdot \mathbf{x}$ for a column vector $\mathbf{x} \in \mathbb{R}^{\mathbf{n}}$ (matrix-vector multiplication), where the columns $(a_k)_{k=1}^n$ are just the images of the unit vectors $(\mathbf{e_k})_{\mathbf{k=1}}^{\mathbf{n}}$ in $\mathbb{R}^n$, we find that with the usual convention of using indices describing row and column positions of the entries of a matrix we have $a_{j,k} = \langle T(\mathbf{e_j}), \mathbf{e_k}\rangle_{\mathbb{R}^{\mathbf{m}}}$, with $1 \le j \le n$ and $1 \le k \le m$.

Even by replacing the unit vectors by Dirac measures, one cannot hope to get a "continuous matrix representation", respectively, a description of any given operator (say on $(L^2(\mathbb{R}^d), \|\cdot\|_2)$) as an integral operator, because for example multiplication operators cannot have nonzero contributions outside the main diagonal. But we can formulate (in analogy with the Schwartz Kernel Theorem for tempered distributions) a kernel theorem for S_0.

Theorem 3.8 *(i) The Banach space of operators $\mathscr{L}(S_0(\mathbb{R}^d), S_0'(\mathbb{R}^d))$ can be identified with the space $S_0'(\mathbb{R}^{2d})$. Specifically, to each operator T, there corresponds a unique distribution $K \in S_0'(\mathbb{R}^{2d})$ such that*

$$\big(Tf\big)(g) = K(f \otimes g) \ \textit{for all} \ f, g \in S_0(\mathbb{R}^d). \tag{3.50}$$

(ii) The Banach space of operators $\mathscr{L}_{w^*}(S_0'(\mathbb{R}^d), S_0(\mathbb{R}^d))$ that map weak convergent sequences in $S_0'(\mathbb{R}^d)$ into norm convergent sequences in $S_0(\mathbb{R}^d)$ can be identified with the space $S_0(\mathbb{R}^{2d})$. Specifically, to each operator T there corresponds a unique function $K \in S_0(\mathbb{R}^{2d})$ such that*

$$\big(T\sigma\big)(x) = \int_{\mathbb{R}_d} K(x, y)\, dy \ \textit{for all} \ \sigma \in S_0'(\mathbb{R}^d), \ x \in \mathbb{R}^d. \tag{3.51}$$

Moreover, one has $K(x, y) = (T\delta_y)(x) = \delta_x(T(\delta_y))$ for all $x, y \in \mathbb{R}^d$.

Note that the Hilbert space $L^2(\mathbb{R}^{2d})$ satisfies $S_0(\mathbb{R}^{2d}) \hookrightarrow L^2(\mathbb{R}^{2d}) \hookrightarrow S_0'(\mathbb{R}^{2d})$ and by the classical characterization of Hilbert–Schmidt operators on $L^2(\mathbb{R})td$, this is an intermediate version of the kernel theorem. Recall that Hilbert–Schmidt operators are compact operators, and form a Hilbert space with respect to the sesquilinear form

$$\langle S, T\rangle_{\mathscr{H}\mathscr{S}} := \mathrm{trace}(S * T^*)$$

and the identification is even unitary at this level. For proof of Theorem 8, we refer to [25].

What we can see from Theorem 3.8(ii), in the case of "regularizing operators", is that they behave very much like matrices, just with continuous entries. This is quite useful for various reasons. It allows assigning (also in the context of S_0 and S_0') to each operator a Kohn–Nirenberg symbol or (via an additional symplectic Fourier transform) a so-called *spreading symbol*. These alternative representations are on $S_0'(\mathbb{R}^d \times \hat{\mathbb{R}}^d)$ or $S_0(\mathbb{R}^d \times \hat{\mathbb{R}}^d)$ respectively if and only if the corresponding kernels are in this space. Again those isomorphisms can be seen as extensions, respectively, restrictions of the Hilbert (Schmidt) case, but we will not have space to discuss this at length here (see [9]).

But we would like to point at least to the natural composition law for regularizing operators. Assume that we have two operators T_1 and T_2 with kernels in $S_0(\mathbb{R}^{2d})$, denoted by K_1 and K_2. Clearly the composition $T_2 \circ T_1$ of these operators belongs again to the operator space $\mathscr{L}_{w^*}(S_0', S_0)$ and therefore has a kernel $K \in S_0(\mathbb{R}^{2d})$. Not very surprising one can show (easily) that one has

$$K(x, z) = \int_{\mathbb{R}_d} K_2(x, y)K_1(y, z)dy, \quad x, z \in \mathbb{R}^d. \tag{3.52}$$

When we want to compose two operators with more general kernels, let us assume that now T_1, T_2 are just bounded operators on $L^2(\mathbb{R}^d)$, so they belong to $\mathscr{L}(L^2, L^2) \subset \mathscr{L}(S_0, S_0')$, then they might not have a representation by kernels in S_0 in general and the question is how to "compose" the kernels. For such cases formula (3.52) above cannot be applied directly, but it is possible to combine this with regularization operators to ensure that the actual composition is performed on "nice kernels". Of course one takes limits after the composition and reaches in this way better and better approximation (in the w^*-sense) to the kernel of the composed mapping.[10]

When applied to the Fourier transform with the continuous, bounded and smooth kernel $K_1(s, y) = e^{-2\pi isy}$ and the inverse Fourier transform with kernel $K_2(s, x) = e^{2\pi ixs}$, one can see that the resulting operator is the identity operator which can be described by the distribution $\delta_\Delta(F) = \int_{\mathbb{R}^d} F(x, x)dx$, for $F \in S_0(\mathbb{R}^{2d})$, which should be seen as the continuous analog of the Kronecker delta symbol. Viewed rowwise (in the continuous sense) the entry is just δ_x at level x, or in other words $T(f)(x) = \delta_x(f) = f(x)$, known as the *sifting property* of the *Dirac delta* (see for example [37], or [2]).

[10] This is comparable with the multiplication of real numbers which is defined as the limit of products of decimal approximations of the involved real numbers, and taking limits afterward!

Taking the naive approach and computing (3.52) for the Fourier kernels and then applying the exponential law results in the (mathematically strange, but often used by engineers) formula

$$\int_{-\infty}^{\infty} e^{-2\pi i s t} ds = \delta(t). \tag{3.53}$$

Such an integral should of course not be viewed as an effective integral, but rather a rule at the level of symbols which is equivalent to the (independently verifiable fact) that $\mathscr{F}^{-1} \circ \mathscr{F} = Id$, e.g., as operators on $S_0(\mathbb{R}^d)$ (using true integrals) or in the spirit of Plancherel's Theorem (by taking limits).

The setting in Theorem 3.8(i) is general enough to be applied to many of the operators arising elsewhere, e.g., bounded on any of the space $(L^p(\mathbb{R}^d), \|.\|_p)$ or even from $(L^p(\mathbb{R}^d), \|.\|_p)$ to some other $(L^q(\mathbb{R}^d), \|.\|_q)$, for $1 \leq p, q \leq \infty$, because one has $S_0(\mathbb{R}^d) \subset L^p(\mathbb{R}^d) \subset S_0'(\mathbb{R}^d)$ (with continuous embeddings), for $p, q \in [1, \infty]$. The book of R. Larsen ([34]) describes such operators as convolution operators by suitable quasi-measures. These *quasi-measures* (introduced by G. Gaudry, [30]) are more general than the elements of $S_0'(\mathbb{R}^d)$ and can only be convolved with compactly supported functions in the Fourier algebra, i.e., the elements of the pre-dual. Moreover, unlike elements of $S_0'(\mathbb{R}^d)$ it is not possible to define a Fourier transform, respectively, a corresponding transfer function in the natural way. Note however that operators with a kernel in S_0' do *not form an algebra*, because the range of the space may be larger than the domain. On the other hand, for operators mapping a given space into itself (e.g., $L^2(\mathbb{R}^d)$, or even $(S_0(\mathbb{R}^d), \| \cdot \|_{S_0})$, etc.) composition is possible and then it should be true (and can be verified) that the convolution of the corresponding kernels "somehow makes sense" (using regularizers) or equivalently, the pointwise product of the associated transfer functions will be also meaningful (e.g., via pointwise multiplication in $L^\infty(\mathbb{R}^d)$ almost everywhere).

The kernel theorem is the starting point for many alternative descriptions of linear operators, more or less by a "change of basis". One can view the space $S_0'(\mathbb{R}^{2d})$ as a (huge) space of operators, which contains a number of interesting operators, such as the collection of all the TF-shifts $\pi(\lambda) = E_s T_x,\ x, s \in \mathbb{R}^d$. The so-called spreading representation of the operators is a kind of "Fourier-like" representation of operators, where these TF-shifts play the role of the Fourier basis for the continuous Fourier transform. This representation will be called the *spreading representation* of operators. For more on this see, e.g., [10, 26].

3.10 Shannon's Sampling Theorem

The claim of the classical Whittaker–Kotelnikov–Shannon Sampling Theorem concerns the recovery of any $L^2(\mathbb{R})$ function whose a Fourier transform whose support is contained in the symmetric interval $I = [-1/2, 1/2]$ around zero (i.e., $\mathrm{supp}(\hat{f}) \subseteq I$) from regular samples of the form $(f(\alpha n))_{n \in \mathbb{Z}}$ as long as $\alpha \leq 1$ (Nyquist rate).

The reconstruction can be achieved using the sinc-function, with $\mathrm{sinc}(t) = sin(\pi t)/\pi t$, the *sinus cardinales*,[11] which can be characterized as the inverse Fourier transform of the box function $\mathbb{1}_I$, the indicator function of I.

It is convenient to apply the following notation:

$$B_I^2 := \{f \,:\, f \in L^2(\mathbb{R}),\ \mathrm{supp}(\hat{f}) \subseteq I\}. \tag{3.54}$$

The Sampling theorem can be deduced as follows: by the usual Fourier series, we know that the functions $(e_k)_{k\in\mathbb{Z}} = (e^{2\pi iks})_{k\in\mathbb{Z}}$ form an complete orthonormal basis in the Hilbert space $L^2([0,1])$, respectively, the space of all functions from $L^2(\mathbb{R})$ with $\mathrm{supp}(\hat{f}) \subseteq I$. Therefore using the standard inner product $\langle\cdot,\cdot\rangle$ on $L^2(I)$ we obtain

$$\hat{f}(s) = \sum_{k\in\mathbb{Z}} \langle \hat{f}, e_k\rangle e_k(s) = \sum_{k\in\mathbb{Z}} \langle \hat{f}, e_k\rangle e^{2\pi iks}\mathbb{1}_I(s).$$

By applying the inverse Fourier transform we obtain

$$f(t) = \sum_{k\in\mathbb{Z}} \langle \hat{f}, e_k\rangle \mathrm{sinc}(t+k), \tag{3.55}$$

$$\text{with} \quad \langle \hat{f}, e_k\rangle = \int_I \hat{f}(s)\, e^{-2\pi iks}\, ds = \int_{\mathbb{R}} \hat{f}(s)\, e^{-2\pi iks}\, ds = f(-k).$$

Plugging this into (3.55) yields the classical version of the Shannon theorem:

$$f(t) = \sum_{k\in\mathbb{Z}} f(k)\,\mathrm{sinc}(t-k) \ \text{ for all } \ t \in \mathbb{R} \ \text{ and } \ f \in B_I^2. \tag{3.56}$$

Thanks to the fact that the sampling values are in $l^2(\mathbb{Z})$ the series is pointwise absolutely convergent, even uniformly, but it is also unconditionally convergent in $(L^2(\mathbb{R}), \|\cdot\|_2)$. Unfortunately, the partial sums are *not well localized* due to the poor decay of the sinc-function (which is in $L^2(\mathbb{R})$, but not in $L^1(\mathbb{R})$ or $S_0(\mathbb{R})$).

Consequently one prefers to make use of alternative building blocks at the cost of working at a slight oversampling rate.[12] Let us formulate this more practical version of the Shannon sampling for bandlimited functions in the Wiener algebra.

For any interval $I \subset \mathbb{R}$ we set $B_I^1 := \{f \in W(\mathbb{R}) \,:\, \mathrm{supp}(\hat{f}) \subset I\}$. One can show that $B_I^1 = \{f \in S_0(\mathbb{R}) \,:\, \mathrm{supp}(\hat{f}) \subset I\} = \{f \in L^1(\mathbb{R}) \,:\, \mathrm{supp}(\hat{f}) \subset I\}$. The more

[11] The word "cardinal" comes into the picture because of the *Lagrange type* interpolation property of the function sinc: $\mathrm{sinc}(k) = \delta_{k,0}$.

[12] Recall that digital audio recordings are meant to capture all the frequencies up to 20 kHz and work with 44100 samples per second although the abstract Nyquist criterion would only ask for $2 * 20000 = 40000$ samples per second (to express the Nyquist criterion in a practical form). Clearly the use of this theorem in a real-time situation requires the reconstruction being well localized in time, in order to cause only minimal delay of the reconstruction process.

practical version of Shannon's Sampling Theorem, now with good localization of the building blocks (rather than the sinc-function) reads as follows.

Theorem 3.9 *Let $\beta > 0$ be such that $I \subset \frac{1}{2}(-\beta, \beta)$ and let $g \in S_0(\mathbb{R})$ be such that $\hat{g}(s) = 1$ for all $s \in I$ and supp $\hat{g} \subset \frac{1}{2}[-\beta, \beta]$ and let $\alpha = \beta^{-1}$. Then we have*

$$f(t) = \alpha \sum_{k \in \mathbb{Z}} f(\alpha k) g(t - \alpha k) \ \text{for all} \ t \in \mathbb{R}, \quad \forall f \in B_I^1, \tag{3.57}$$

with absolute convergence in $(S_0(\mathbb{R}), \| \cdot \|_{S_0})$, $(W(\mathbb{R}), \| \cdot \|_W)$, and $(C_0(\mathbb{R}), \| \cdot \|_\infty)$.

It is even possible to require that g has decay like the inverse of any given polynomial: given $r \in \mathbb{N}$ one can find g such that $|g(t)| \leq C(1 + |t|)^{-r}$ for a suitable constant $C > 0$. The spectrum of g is contained in a small open interval around I.

Proof The assumption about supp$(\hat{f}) \subset I$ implies that the support of all the shifted copies of $\hat{f}$, are disjoint to I and even to the open interval $(-\beta/2, \beta/2)$. Hence for any (ideally smooth) function g as in the theorem satisfies

$$(\sqcup\!\sqcup_\beta * \hat{f}) \cdot \hat{g} = \hat{f}. \tag{3.58}$$

By applying the inverse Fourier transform , we find

$$f = \alpha \cdot (\sqcup\!\sqcup_\alpha \cdot f) * g \tag{3.59}$$

That is, we reach our goal as follows

$$\begin{aligned} f(t) &= \big(\alpha \cdot (\sqcup\!\sqcup_\alpha \cdot f) * g\big)(t) = \alpha \cdot \big(\sqcup\!\sqcup_\alpha \cdot f\big)(T_t g^{\checkmark}) \\ &= \alpha \cdot \big(\sqcup\!\sqcup_\alpha\big)(f \cdot T_t g^{\checkmark}) = \alpha \sum_{k \in \mathbb{Z}} (f \cdot T_t g^{\checkmark})(\alpha\, k) \\ &= \alpha \sum_{k \in \mathbb{Z}} f(\alpha\, k)\, g(t - \alpha k). \end{aligned}$$

3.11 Systems and Convolution Operators

The theory of TILS (*translation invariant linear systems*) is an important subject and most electrical engineering students are exposed to this concept early on in their studies. Unfortunately one must say that—due to the lack of appropriate mathematical descriptions—the way in which the concepts of an impulse response respectively a transfer function is introduced only in a rather vague (but "intuitive") fashion. Furthermore, students who want to dig deeper and understand these concepts in more detail are left alone, because engineering books explaining the relevance of the subject do not provide more details or justifications later on. On the other hand, the mathematical books who talk about convolution do this with a completely different motivation but do not connect to those problems arising in the engineering context.

The article [21] takes the first steps toward a reconciliation of these two approaches[13] by modeling translation invariant systems of what is called BIBOS systems (which means bounded input—bounded output), respectively, as a bounded linear operator from the Banach space $(C_0(\mathbb{R}^d), \|\cdot\|_\infty)$ into itself, commuting with translations.

By choosing as a domain the space $C_0(\mathbb{R}^d)$ and *not* the larger space $C_b(\mathbb{R}^d)$ of all bounded, continuous, complex-valued functions, we avoid indeed the so-called *scandal* in system theory as diagnosed by I. Sandberg in a series of paper (see e.g., [41–44]). Furthermore, we are in fact able to represent every such system as a convolution operator by some *bounded measure*. In order to do so it is not at all required to discuss technical details of measure theory, but one can just *call*[14] the bounded (respectively, continuous) linear functionals on $(C_0(\mathbb{R}^d), \|\cdot\|_\infty)$ bounded measures (as we also did in Sect. 3.3).

Unfortunately, this setting cannot be used to characterize all the TILS which are bounded on $(L^2(\mathbb{R}^d), \|\cdot\|_2)$. It is true that every convolution operator of the form $f \mapsto \mu * f,\ f \in L^2(\mathbb{R}^d)$ with $\mu \in M_b(\mathbb{R}^d)$ extends to all of $L^2(\mathbb{R}^d)$ and satisfies the expected estimate: $\|\mu * f\|_2 \leq \|\mu\|_{M_b(\mathbb{R}^d)}\|f\|_2$, or alternatively can be described on the Fourier transform side as $\hat{f} \mapsto \hat{\mu} \cdot \hat{f}$, where $\hat{\mu} \in C_b(\mathbb{R}^d)$, but not every L^2-TILS can be represented in this form.

It is not so difficult to find out (using Plancherel's Theorem) that the most general TILS on $(L^2(\mathbb{R}^d), \|\cdot\|_2)$ is a pointwise multiplier with an essentially bounded and measurable function, respectively, with some $h \in L^\infty(\mathbb{R}^d)$. So we can write any such operator in the form $f \mapsto T(f) = \mathscr{F}^{-1}(h \cdot \hat{f})$, with transfer "function" $h \in L^\infty(\mathbb{R}^d)$. But then one would expect that we can write $T(f) = \sigma * f$, where $\sigma = \mathscr{F}^{-1}(h)$, but normally no inverse Fourier transform for bounded functions (which are not integrable or at least square integrable) exists. However, this can be made correct by taking the inverse Fourier transform in the sense of $S_0'(\mathbb{R}^d)$ (as defined in Sect. 3.8).

One possible example is the convolution by a chirp signal, which is a bounded, highly oscillating function of the form $ch(t) = e^{i\pi\alpha|t|^2}$. For simplicity we choose the value $\alpha = 1$. The general chirp can be obtained from this one by dilations. This allows us to derive from this also the FT of general chirp signals.

Recall that the chirp $ch(t) = e^{i\pi|t|^2}$ belongs to $S_0'(\mathbb{R}^d)$ and therefore has a Fourier transform in this sense. Moreover, it is in fact Fourier invariant, and consequently convolution by ch corresponds to pointwise multiplication of $\hat{f}$ by $ch(t)$, which is a good operator on $(L^2(\mathbb{R}^d), \|\cdot\|_2)$, because it is continuous and bounded.

On the other hand, one might expect that one can write the convolution for any $f \in L^2(\mathbb{R}^d)$ as an integral, if not as a Riemann integral so at least as a Lebesgue integral, because this is the most general integral (at least for our purposes). Specifically, we would like to convolve ch with the sinc-function. But due to the fact that $|ch(t)| =$

[13] But still much more has to be done!

[14] This is well justified by the Riesz representation theorem.

$1, \forall t \in \mathbb{R}$ and the fact that sinc $\notin L^1(\mathbb{R})$ for no argument, this convolution integral exists in the literal sense. It is however (and of course) possible to approximate $f \in L^2(\mathbb{R})$ by functions $f_n \in S_0(\mathbb{R})$, to perform the convolutions $ch * f_n$ in a classical way, and then take the limit for $n \to \infty$ (with convergence in the L^2 sense).

There are other scenarios, for example, (at least mathematicians) are interested in linear operators from $(L^p(\mathbb{R}_d), \|.\|_p)$ to $(L^q(\mathbb{R}_d), \|.\|_q)$ of a similar nature. All of these cases are covered by the following theorem.

Theorem 3.10 *The Banach space* $H_{L_1}(S_0, S_0')$ *of all bounded linear operators from* $(S_0(\mathbb{R}^d), \|\cdot\|_{S_0})$ *into* $(S_0'(\mathbb{R}^d), \|\cdot\|_{S_0'})$ *which commute with the action of* $W(\mathbb{R}^d)$ *or* $L^1(\mathbb{R}^d)$ *by convolution,*[15] *i.e., which satisfy*

$$T(g * f) = g * T(f), \quad \forall g \in L^1(\mathbb{R}^d),\ f \in S_0(\mathbb{R}^d), \tag{3.60}$$

or equivalently the set of all translation invariant bounded operators

$$T(T_x f) = T_x(T(f)), \quad \forall x \in \mathbb{R}^d,\ f \in S_0(\mathbb{R}^d), \tag{3.61}$$

can be characterized as the set of all convolution operators of the form $T : f \mapsto \sigma * f$ *(given pointwise* $[\sigma * f](x) = \sigma(T_x f^{\checkmark})$*) where* $\sigma \in S_0'(\mathbb{R}^d)$*. In fact, every such operator maps* $(S_0(\mathbb{R}^d), \|\cdot\|_{S_0})$ *into* $(C_b(\mathbb{R}^d), \|\cdot\|_\infty)$*, and the corresponding three norms are equivalent, i.e.,* $\|\sigma\|_{S_0'}$*, or the operator norm of* T *as operator from* $S_0(\mathbb{R}^d)$ *into* $(C_b(\mathbb{R}^d), \|\cdot\|_\infty)$ *or into* $(S_0'(\mathbb{R}^d), \|\cdot\|_{S_0'})$*, respectively. Moreover, any such operator can be described on the Fourier transform side as a Fourier multiplier with the transfer function* $\widehat{\sigma} \in S_0'(\mathbb{R}^d)$*, via*

$$\widehat{T(f)} = \widehat{\sigma} \cdot \hat{f}, \quad f \in S_0(\mathbb{R}^d). \tag{3.62}$$

3.12 Further References

These notes are part of a more comprehensive program running under the title "Conceptual Harmonic Analysis" (see [22]). It aims at providing a more integrative approach to Fourier analysis and its applications, by emphasizing the connections between discrete and continuous Fourier transform. The contribution provided by this article is meant to underline that such a more global approach to Fourier analysis, which certainly requires the use of generalized functions (like Dirac measures, Dirac combs, but also almost periodic function and their Fourier transforms, etc.) does not have to start from the theory of Schwartz functions and Lebesgue integration, or even from the Schwartz–Bruhat distributions (see [3, 36]) and (Haar)-measure theory in

[15] In the terminology of Banach modules, we are talking about the fact that both $S_0(\mathbb{R}^d)$ and $S_0'(\mathbb{R}^d)$ are Banach modules over the Banach convolution algebra $(L^1(\mathbb{R}^d), \|\cdot\|_1)$, and that we are interested in the Banach module homomorphisms.

the case of LCA groups. Instead, at least for the Euclidean case, a simplified approach can be provided on the basis of principles from linear functional analysis and the Riemann integral for continuous and well decaying functions on $\mathbb{R}^d$. Recall that the use of functional analytic methods as such appears unavoidable due to the fact that relevant signal spaces are rarely finite dimensional.

The original paper introducing the Banach space S_0 for general locally compact abelian groups is [15]. At that time it was introduced as a particular Segal algebra in the spirit of H. Reiter [38], in fact the smallest member in the family of all *strongly character invariant* (meaning in modern terminology: *isometrically modulation invariant*) Segal algebras. This minimality property gives a large number of properties of these spaces. It is introduced there in the context of general LCA groups. A comprehensive walkthrough of its important properties (also for general LCA groups) is [32].

It turned out to be the proper domain for the treatment of the metaplectic group by H. Reiter in [39] and even for the treatment of *generalized stochastic processes* (see [24]). Also, it is essential for the development of a general theory of *modulation spaces*, which are nowadays a well established discipline, even with interesting applications in the theory of partial or pseudo-differential operators (see e.g., [18, 19]).

From the point of view of *coorbit theory* as developed in [23] modulation spaces are associated with the STFT, which can be seen as practically equivalent with the matrix coefficients of a pair of vectors f, g in the Hilbert space $(L^2(\mathbb{R}^d), \| \cdot \|_2)$ under the *Schrödinger representation* of the *reduced Heisenberg group*. This makes modulation spaces very suitable for the discussion of operators arising in time–frequency analysis and especially in connection with Gabor Analysis.

It is this area where the usefulness of the spaces $(S_0(\mathbb{R}^d), \| \cdot \|_{S_0})$ and its dual became apparent again and again. Sometimes these two spaces are viewed together as a *Banach Gelfand Triple* denoted by $(S_0, L^2, S_0')(\mathbb{R}^d)$. It has been the experiences especially in this area where the ideas about "well chosen function spaces" became clear. In the spirit of [20], the current article describes the Wiener algebra $W(C_0, l^1)(\mathbb{R}^d)$ and the Segal algebra $(S_0(\mathbb{R}^d), \| \cdot \|_{S_0})$ as the most useful Banach spaces of continuous and integrable functions. It allows using ordinary Riemann integrals in a very natural fashion and also covers more or less all the classical summability kernels. On the way to a distribution theoretical description of the Fourier transform (cf. also the elaborations of J. Fischer in this direction, [27, 28]) the space $W_{\mathscr{F}}\mathbb{R}^d = W(\mathbb{R}^d) \cap \mathscr{F}W(\mathbb{R}^d)$ is a first, intermediate step.

While the concept of modulation spaces was originally to define Wiener amalgam spaces on the Fourier transform side (in the spirit of the Fourier analytic description of the classical smoothness spaces like $(B^s_{p,q}(\mathbb{R}^d), \| \cdot \|_{B^s_{p,q}}$, using dyadic, smooth partitions of unity), also the Wiener algebra is a representative of the equally important class of *Wiener amalgam spaces*. The general theory of Wiener amalgam spaces is described in [29] (Fournier/Stewart) and [5] for the classical case, where the *local component* is $L^p(G)$ and the *global component* is $l^q(\mathbb{Z}^d)$. In [16] much more general ingredients were admitted, which work as long as the local component has a sufficiently rich pointwise multiplier algebra in order to generate BUPUs which are uniformly bounded in that multiplier algebra. For $B = \mathscr{F}L^p$ it is enough to have boundedness in $(\mathscr{F}L^1(\mathbb{R}^d), \|\cdot\|_{\mathscr{F}L^1})$.

The general description of Wiener's algebra (described among others in [38, 40]) is the paper [12]. The minimality of $W(C_0, l^1)(\mathbb{R}^d)$ and then $S_0(\mathbb{R}^d) = W(\mathscr{F}L^1, l^1)(\mathbb{R}^d)$ is studied in [13, 17]. Since the local behavior of $\mathscr{F}_W(\mathbb{R}^d)$ equals that of $\mathscr{F}L^1(\mathbb{R}^d)$ (this is valid for any Segal algebra).

The pair consisting of $(S_0(\mathbb{R}^d), \|\cdot\|_{S_0})$ and its dual space $(S_0'(\mathbb{R}^d), \|\cdot\|_{S_0'})$ can also serve as a basis for the treatment of generalized stochastic processes. This approach is described in [24], based on the Ph.D. thesis [31] of W. Hörmann.

3.13 The Relation to the Schwartz Theory

It is of course legitimate to ask about the relationship of the presented approach to the well-established Schwartz Theory of (tempered) distributions (see [45]) which is widely used for PDE or pseudo-differential operators.

It was first observed by D. Poguntke that $\mathscr{S}(\mathbb{R}^d)$ is continuously and densely embedded into $(S_0(\mathbb{R}^d), \|\cdot\|_{S_0})$ and consequently $(S_0'(\mathbb{R}^d), \|\cdot\|_{S_0'})$ is continuously embedded into $\mathscr{S}'(\mathbb{R}^d)$. It is also clear that the extended Fourier transform for $\mathscr{S}'(\mathbb{R}^d d)$, when restricted to $S_0'(\mathbb{R}^d)$ is just the one defined directly in Lemma 3.16 without the use of tempered distributions. In practice $S_0(\mathbb{R}^d)$ and $\mathscr{S}(\mathbb{R}^d)$ respectively, their duals have very similar properties (except for differentiability issues!), including the existence of a kernel theorem or regularization via smoothing and pointwise multiplication, using the relations

$$\big(S_0'(\mathbb{R}^d) * S_0(\mathbb{R}^d)\big) \cdot S_0(\mathbb{R}^d) \subset S_0(\mathbb{R}^d) \tag{3.63}$$

which resembles the well-known relationship

$$\big(\mathscr{S}'(\mathbb{R}^d) * \mathscr{S}(\mathbb{R}^d)\big) \cdot \mathscr{S}(\mathbb{R}^d) \subset \mathscr{S}(\mathbb{R}^d). \tag{3.64}$$

But there are still various good reasons to consider the approach presented in this note. First of all, as mentioned several times, it is technically much less challenging, and so the hope is that it has better chances to be adopted by engineers or physicists. In particular for courses on signal processing and systems theory, it might be a good way to go. For people interested in either numerical approximation of abstract Harmonic Analysis, the function spaces used should offer good tools for a discussion of the connection between the continuous and the finite discrete setting. Such questions usually do not involve any differentiation.

We also point out that the advantage of a smaller room of distributions is the fact that all the many invariance properties allow showing that one is staying within that smaller area. In [26] it was crucial for the derivation of the Janssen representation of the Gabor frame operator for general lattices to show that the distributional kernel describing the spreading function of that operator is supported by the *adjoint lattice*, i.e., by a discrete set, and that consequently it is a sum of Dirac measures (because there is nothing like a practical derivative of the Dirac delta in $S_0'(\mathbb{R}^d)$!). We could also argue that it is enough to know that for any $p \in [1, \infty]$, all its elements in $L^p(\mathbb{R}^d)$ have a Fourier transform inside of $S_0'(\mathbb{R}^d)$ and not only within some much larger space like $\mathscr{S}'(\mathbb{R}^d)$. Theorem 3.10 is a good example in that direction. Unlike quasi-measures (see [33]) we also find the transfer function inside of the Fourier invariant space $S_0'(\mathbb{R}^d)$, a proper subspace of the space of quasi-distributions.

Acknowledgements The work of M.S.J. was carried out during the tenure of the ERCIM 'Alain Bensoussan' Fellowship Programme at NTNU. This project was written while M.S.J. was visiting NuHAG at the University of Vienna. He is grateful for their hospitality. The senior author was finishing this manuscript while he was holding a guest position at the Mathematical Institute of Charles University in Prague.

References

1. J.J. Benedetto, *Harmonic Analysis and Applications*, Studies in Advanced Mathematics (CRC Press, Boca Raton, 1996)
2. R.N. Bracewell, *The Fourier Transform and Its Applications*, 3rd edn., McGraw-Hill Series in Electrical Engineering. Circuits and Systems (McGraw-Hill Book Co, New York, 1986)
3. F. Bruhat, Distributions sur un groupe localement compact et applications a l'etude des représentations des groupes p-adiques. Bull. Soc. Math. France **89**, 43–75 (1961)
4. R. Bürger, Functions of translation type and Wiener's algebra. Arch. Math. (Basel) **36**, 73–78 (1981)
5. R.C. Busby, H.A. Smith, Product-convolution operators and mixed-norm spaces. Trans. Amer. Math. Soc. **263**, 309–341 (1981)
6. P.L. Butzer, D. Schulz, Limit theorems with O-rates for random sums of dependent Banach-valued random variables. Math. Nachr. **119**, 59–75 (1984)
7. M. Cwikel, A quick description for engineering students of distributions (generalized functions) and their Fourier transforms (2018)
8. J.B. Conway, *A Course in Functional Analysis*, 2nd edn. (Springer, New York, 1990)
9. E. Cordero, H.G. Feichtinger, F. Luef, Banach Gelfand triples for Gabor analysis, *Pseudo-differential Operators*, vol. 1949, Lecture Notes in Mathematics (Springer, Berlin, 2008), pp. 1–33
10. M. Dörfler, B. Torrésani, Spreading function representation of operators and Gabor multiplier approximation, in *Proceedings of SAMPTA07*, Thessaloniki (2007)
11. H.G. Feichtinger, A characterization of Wiener's algebra on locally compact groups. Arch. Math. (Basel) **29**, 136–140 (1977)
12. H.G. Feichtinger, Multipliers from $L^1(G)$ to a homogeneous Banach space. J. Math. Anal. Appl. **61**, 341–356 (1977)
13. H.G. Feichtinger, A characterization of minimal homogeneous Banach spaces. Proc. Amer. Math. Soc. **81**(1), 55–61 (1981)
14. H.G. Feichtinger, Banach spaces of distributions of Wiener's type and interpolation, in *Proceeding of Conference on Functional Analysis and Approximation, Oberwolfach, 1980*, eds. by P. Butzer, S. Nagy, E. Görlich, vol. 69, International Series of Numerical Mathematics (Birkhäuser Boston, Basel, 1981), pp. 153–165
15. H.G. Feichtinger, On a new Segal algebra. Monatsh. Math. **92**, 269–289 (1981)
16. H.G. Feichtinger, Banach convolution algebras of Wiener type, in *Proceeding of Conference on Functions, Series, Operators, Budapest 1980*, eds. by B. Sz.-Nagy, J. Szabados, vol. 35,

Colloquia Mathematica Societatis Janos Bolyai (North-Holland, Amsterdam, 1983), pp. 509–524

17. H.G. Feichtinger, Minimal Banach spaces and atomic representations. Publ. Math. Debrecen **34**(3–4), 231–240 (1987)
18. H.G. Feichtinger, Modulation spaces of locally compact Abelian groups, in *Proceedings of International Conference on Wavelets and Applications*, eds. by R. Radha, M. Krishna, S. Thangavelu (New Delhi Allied Publishers, Chennai, 2002, 2003), pp. 1–56
19. H.G. Feichtinger, Modulation spaces: looking back and ahead. Sampl. Theory Signal Image Process. **5**(2), 109–140 (2006)
20. H.G. Feichtinger, *Choosing Function Spaces in Harmonic Analysis*, vol. 4, The February Fourier Talks at the Norbert Wiener Center; Applied and Numerical Harmonic Analysis (Birkhäuser/Springer, Cham, 2015), pp. 65–101
21. H.G. Feichtinger, A novel mathematical approach to the theory of translation invariant linear systems, in *Novel Methods in Harmonic Analysis with Applications to Numerical Analysis and Data Processing*, eds. by P.J. Bentley, I. Pesenson (2016), pp. 1–32
22. H.G. Feichtinger, Thoughts on numerical and conceptual harmonic analysis, in *New Trends in Applied Harmonic Analysis. Sparse Representations, Compressed Sensing, and Multifractal Analysis*, eds. by A. Aldroubi, C. Cabrelli, S. Jaffard, U. Molter, Applied and Numerical Harmonic Analysis (Birkhäuser, Basel, 2016), pp. 301–329
23. H.G. Feichtinger, K. Gröchenig, Banach spaces related to integrable group representations and their atomic decompositions. I. J. Funct. Anal. **86**(2), 307–340 (1989)
24. H.G. Feichtinger, W. Hörmann, A distributional approach to generalized stochastic processes on locally compact abelian groups, in *New Perspectives on Approximation and Sampling Theory*, eds. by G. Schmeisser, R. Stens, Festschrift in Honor of Paul Butzer's 85th Birthday (Birkhäuser/Springer, Cham, 2014), pp. 423–446
25. H.G. Feichtinger, M.S. Jakobsen, The inner kernel theorem for a certain Segal algebra (2018)
26. H.G. Feichtinger, W. Kozek, Quantization of TF lattice-invariant operators on elementary LCA groups, in *Gabor Analysis and Algorithms*, eds. by H.G. Feichtinger, T. Strohmer, Applied and Numerical Harmonic Analysis (Birkhäuser, Boston, 1998), pp. 233–266
27. J.V. Fischer, On the duality of discrete and periodic functions. Mathematics **3**(2), 299–318 (2015)
28. J.V. Fischer, On the duality of regular and local functions. Mathematics **5**(41), (2017)
29. J.J.F. Fournier, J. Stewart, Amalgams of L^p and ℓ^q. Bull. Amer. Math. Soc. (N.S.) **13**, 1–21 (1985)
30. G.I. Gaudry, Quasimeasures and operators commuting with convolution. Pacific J. Math. **18**, 461–476 (1966)
31. W. Hörmann, Stochastic processes and vector quasi-measures. Master's thesis, University of Vienna, 1987
32. M.S. Jakobsen, On a (no longer) new segal algebra: a review of the Feichtinger algebra. J. Fourier Anal. Appl., 1– 82 (2018)
33. H.-C. Lai, A characterization of the multipliers of Banach algebras. Yokohama Math. J. **20**, 45–50 (1972)
34. R. Larsen, *An Introduction to the Theory of Multipliers* (Springer, New York, 1971)
35. V. Losert, A characterization of the minimal strongly character invariant Segal algebra. Ann. Inst. Fourier (Grenoble) **30**, 129–139 (1980)
36. M.S. Osborne, On the Schwartz-Bruhat space and the Paley-Wiener theorem for locally compact Abelian groups. J. Funct. Anal. **19**, 40–49 (1975)
37. P. Prandoni, M. Vetterli, *Signal Processing for Communications* (CRC Press, 2008)
38. H. Reiter, *Classical Harmonic Analysis and Locally Compact Groups* (Clarendon Press, Oxford, 1968)
39. H. Reiter, *Metaplectic Groups and Segal Algebras*, Lecture Notes in Mathematics (Springer, Berlin, 1989)
40. H. Reiter, J.D. Stegeman, *Classical Harmonic Analysis and Locally Compact Groups*, 2nd edn. (Clarendon Press, Oxford, 2000)

41. I.W. Sandberg, The superposition scandal. Circuits Syst. Signal Process. **17**(6), 733–735 (1998)
42. I.W. Sandberg, A note on the convolution scandal. IEEE Signal Process. Lett. **8**(7), 210–211 (2001)
43. I.W. Sandberg, Continuous multidimensional systems and the impulse response scandal. Multidimens. Syst. Signal Process. **15**(3), 295–299 (2004)
44. I.W. Sandberg, Bounded inputs and the representation of linear system maps. Circuits Syst. Signal Process. **24**(1), 103–115 (2005)
45. L. Schwartz, *Théorie des Distributions. (Distribution Theory). Nouveau Tirage*, vol. 1, xii, 420p (Hermann, Paris, 1957)
46. E.M. Stein, *Singular Integrals and Differentiability Properties of Functions* (Princeton University Press, Princeton, 1970)

Chapter 4
Partial Differential Equations on Metric Graphs: A Survey of Results on Optimization, Control, and Stabilizability Problems with Special Focus on Shape and Topological Sensitivity Problems

Günter Leugering

Abstract We consider ordinary equations on metric graphs. In particular, we consider novelty and review earlier results in the context of shape and topology optimization for second-order equations on such metric graphs. For the sake of brevity, we concentrate on simple topologies, such as star graphs, in order to provide a simple representation of the concepts. In fact, we use the concept of Steklov–Poincaré operators in order to reduce complex graphs to star graphs. As for the differential operators, we also confine ourselves with constant coefficients. In that respect, the current article is the first one, where the results scattered in the literature are put in a unifying framework.

Keywords Problems on metric graphs · Shape and topology optimization · Optimal control

4.1 Introduction

Differential equations on graphs or, more precisely, on metric graphs are ubiquitous. In particular, transportation networks conveying fluids, gas, electrical power or traffic can be modeled by hyperbolic systems on networks. Also mechanical networks involving strings, cables, beams, masses, and springs that are important in civil engineering can be framed within hyperbolic balance laws on graphs. Heat flow in technical devices as well as in buildings can be regarded as parabolic equations on networks. Also flow in crack networks are of importance and lead to uncommon fluid–structure interaction problems in combination with crack-sensitivity theory.

This work has been supported by the grants DFG TRR154 A05, DAAD-UGC.

G. Leugering (✉)
Department of Mathematics, Friedrich-Alexander-University Erlangen-Nürnberg,
Cauerstraße 11, 91058 Erlangen, Germany
e-mail: guenter.leugering@fau.de

P. Manchanda et al. (eds.), *Mathematical Modelling, Optimization, Analytic and Numerical Solutions*, Industrial and Applied Mathematics,
https://doi.org/10.1007/978-981-15-0928-5_4

The underlying quasi-static model or one that is obtained after suitable time discretization is that of an elliptic system on networks. In these notes, we concentrate on such elliptic systems and, as those are edgewise one-dimensional, consequently, on ordinary differential equations on metric graphs, that, in turn, are constituted as assemblies of Jordan curves in three-spaces or more simply by straight edges in the plane. Such metric graphs and elliptic problems "carried" by the graph have been the subject of many publications. In particular in physics, such graphs together with Sturm–Liouville problems have come to be known as quantum graphs (see Exner [9, 10] and Kuchment [34–36] for typical references). Articles in this area are typically concerned with well-posedness questions and analytical properties, in particular with respect to the spectrum. Another branch of literature is concerned with wave equations and heat equations defined on such metric graphs. Early contributions are, for example, by von Below [50, 51] on the spectral properties of metric graphs and even nonlinear parabolic problems thereon. At the same time problems of controllability and stabilizability of linear wave equations on metric graphs have been investigated. See Schmidt [15] and Lagnese, Leugering, and Schmidt [37–41]. Later the topic has become very popular and even today it enjoys a lot of attraction. Indeed, a wealth of literature is currently available. As regards the control of wave and heat equations, a fairly recent survey has been given by Dager and Zuazua [5] while nonlinear problems concerning the control of traffic dynamics and supply chains on networks have been discussed in the literature by many authors, culminating in the recent monographs by Garavallo, Hans, and Piccoli [11] and D'Apice, Göttlich, Herty, and Piccoli [6]. Nonlinear and nonlocal hyperbolic balance laws in this context have been investigated in, e.g., [13, 21, 31] and most recently in [32, 33]. Yet another branch of quasilinear problems on metric graphs is given by the transport of gas and water in pipe networks or canal networks. There is virtually no space for an adequate acknowledgement of the literature. Suffice it to refer to a survey article provided in the last proceedings of ISIAM 2016 that was published in [29] and the articles [7, 8, 16–20, 22–24, 27, 28, 44] where the corresponding nonlinear problems of controllability and stablizability are addressed. The literature on mixed-integer nonlinear optimal control of gas flow on networks has been achieved in [12, 25, 26, 45].

Also inverse problems have been studied related to differential equations on graphs. Here the interest is often more focused on the topology of the graph and physical parameters. The interest, in a way, has been motivated by the classical question raised by Kac: Can one hear the shape of a drum? See[30]. The question has been rephrased as Can one hear the shape of a tree [52]? Indeed, in Avdonin, Leugering, and Mikhaylov [2, 3], these questions have been answered positively in the context of planar wave equations on planar graphs. There are other articles in the context of quantum graphs, where similar inverse problems have been solved. See e.g., [1, 49]. As with controllability and stabilizability questions, if the graph contains cycles, no such fully satisfactory answer can be given.

Yet another observation is in order here. When dealing with scalar differential equations on a (metric) graph, only the adjacency structure and the lengths of the edges matters while the actual geometrical objectives, such as the angles between two consecutive edges, do not matter in the class of trees. This is different for planar, that is to say, in-plane problems on planar graphs or for higher dimensional problems on graphs in d-space and those problems where cycles occur.

In these notes, therefore, we address the role of the lengths of the edges and the topology of the graph, thus, we are interested in the shape and the topology optimization of metric graphs upon which differential equations are defined. To the best knowledge of the author, only very few articles are present in the literature, where such problems are addressed. We refer to Leugering and Sokolowski [43, 46] and the dissertation by Ogiermann [48]. In fact, part of this survey is taken from these articles. In these articles, the sensitivity of a graph with respect to exchanging a multiple node by cycle is investigated. This is analogous to what has come to be known as topological sensitivity. Releasing a node to a series of nodes, inserting an edge, deleting an edge or contracting a subgraph to single node, these are questions naturally asked in the context of discrete graphs and integer optimization problems. In the context of continuous problems on metric graphs, many of analogous questions are unresolved as of today. In these notes, we dwell on these problems. For the sake of brevity, we stay with planar graphs. Moreover, in order to keep the material compact, we dwell on a few special problems in order to provide the scope of different directions.

4.2 Planar Graphs

We consider a simple planar graph $(V, E) = G$ in $\mathbf{R}^2$, with vertices $V = \{v_J | J \in \mathscr{J}\}$ and edges $E = \{e_i | i \in \mathscr{I}\}$. Let $m = |\mathscr{J}|, n = \|\mathscr{I}\|$ be the numbers of vertices and edges, respectively. In general the edge set may be a collection of smooth curves in $\mathbf{R}^2$, parametrized by their arc lengths. The restriction to planar graphs and *straight edges* is for the sake of simplicity only. The more general case, which is of course also interesting in the combination of shape and topology optimization, can also be handled. However, this is beyond these notes.

We associate to the edge e_i the unit vector $\underline{e}_i$ aligned along the edge. $\underline{e}_i^\perp$ denotes the orthogonal unit vector. Given a node v_J, we define

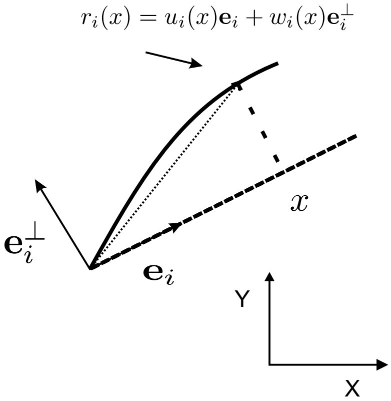
$$\mathscr{I}_J := \{i \in \mathscr{I} | e_i \text{ is incident at } v_J\}$$

as the incidence set, and $d_J = |\mathscr{I}_J|$ as the edge degree of v_J. The set of nodes splits into simple nodes $\mathscr{J}_S$ and multiple nodes $\mathscr{J}_M$ according to $d_J = 1$ and $d_J > 1$, respectively. On G, we consider a vector-valued function y representative of the displacement of the network (see Fig. 4.1)

$$y : G \to \mathbf{R}^{np} := \Pi_{i \in \mathscr{I}}^{p_i} \mathbf{R},\ \ p_i \geq 1\, \forall i. \tag{4.1}$$

The numbers p_i represents the degrees of freedom of the physical model used to describe the behavior of the edge with number i. For instance, $p = 1$ is representative of a heat problem and out-of-the-plane models for elastic strings, whereas $p = 2, 3$

Fig. 4.1 Representation of planar displacement



is used in an elasticity context on graphs in 2 or 3 dimensions. The $p_i's$ may change in the network in principle. However, for the sake of brevity, in this article we insist on $p_i = p = 1$, or $2, \forall i$. See Lagnese, Leugering, and Schmidt [41] and Lagnese and Leugering [42] for details on the modeling.

Once the function y is understood as being representative of, say, a deformation of the graph, we may localize it to the edges

$$y^i := y|_{e_i} : [\alpha_i, \beta_i] \to \mathbf{R}^p,\ i \in \mathscr{I}, \tag{4.2}$$

where e_i is parametrized by $x \in [\alpha_i, \beta_i] =: I_i,\ 0 \le \alpha_i < \beta_i,\ \ell_i := \beta_i - \alpha_i$. See Fig. 4.1.

We introduce the incidence relation

$$d_{iJ} := \begin{cases} 1 & \text{if } e_i \text{ ends at } v_J \\ -1 & \text{if } e_i \text{ starts at } v_J \end{cases}.$$

Accordingly, we define

$$x_{iJ} := \begin{cases} 0 & \text{if } d_{iJ} = -1 \\ \ell_i & \text{if } d_{iJ} = 1 \end{cases}.$$

We will use the notation $y^i(v_J)$ instead of $y^i(x_{iJ})$. In order to represent the material considered on the graph, we introduce stiffness matrices

$$K_i := h_i\left[\left(1 - \frac{1}{s_i}\right) I + \frac{1}{s_i}\underline{e}_i\underline{e}_i^T\right]. \tag{4.3}$$

Obviously, the longitudinal stiffness is given by h_i, whereas the transverse stiffness is given by $h_i(1 - \frac{1}{s_i})$. This can be related to 1-d analoga of the Lamé parameters. We introduce Dirichlet and Neumann simple nodes as follows. We define

$$\mathscr{J}_D^t := \{J \in \mathscr{J}_S | y^i(v_J) \cdot \underline{e}_i = 0\},$$

$$\mathscr{J}_D^n := \{J \in \mathscr{J}_S | y^i(v_J) \cdot \underline{e}_i^\perp = 0\},$$

$$\mathscr{J}_N^t := \{J \in \mathscr{J}_S | d_{iJ} K_i \frac{d}{dx} y^i(v_J) \cdot \underline{e}_i = 0\},$$

$$\mathscr{J}_N^n := \{J \in \mathscr{J}_S | d_{iJ} K_i \frac{d}{dx} y^i(v_J) \cdot \underline{e}_i^\perp = 0\}.$$

Notice that these sets are not necessarily disjoint. Obviously, the set of completely clamped vertices can be expressed as

$$\mathscr{J}_D^0 := \mathscr{J}_D^t \cap \mathscr{J}_D^n. \tag{4.4}$$

Similarly, a vertex with completely homogenous Neumann conditions is expressed as $\mathscr{J}_N^n \cap \mathscr{J}_N^t$. At tangential Dirichlet nodes in $\mathscr{J}_D^t$ we may, however, consider normal Neumann conditions as in $\mathscr{J}_N^n$ and so on.

In this article, we restrict ourselves to "rigid" joints in the sense that the angles between edges in their reference configuration remain fixed. The continuity is expressed simply as

$$y^i(v_J) = y^j(v_J), \quad i, j \in \mathscr{I}_J, \; J \in \mathscr{J}_M.$$

We consider the energy of the system

$$\mathscr{E}_0 := \frac{1}{2} \sum_{i \in \mathscr{I}} \int_0^{\ell_i} K_i \frac{d}{dx} y^i \cdot \frac{d}{dx} y^i + c_i y^i \cdot y^i dx, \tag{4.5}$$

where the primes denote the derivative with respect to the running variable x_i, c_i represents an additional spring stiffness term or an elastic support.

In order to analyze the problem, we need to introduce a proper energy space

$$\mathscr{V} := \{y : G \to \mathbf{R}^{2n} | y^i \in H^1(I_i) \tag{4.6}$$

$$y^i(v_D) = 0, \; i \in \mathscr{I}_D, \; D \in \mathscr{J}_D^0 \tag{4.7}$$

$$y^i(v_J) = y^j(v_J), \; \forall i, j \in \mathscr{I}_J, \; J \in \mathscr{J}_M\}. \tag{4.8}$$

$\mathscr{V}$ is clearly a Hilbert space in

$$\mathscr{H} := L^2(0, \ell_i)^{2n}. \tag{4.9}$$

We introduce the bilinear form on $\mathscr{V} \times \mathscr{V}$

$$a(r,\phi) := \sum_{i\in\mathscr{I}} \int_0^{\ell_i} \left[K_i \frac{d}{dx} y^i \cdot \frac{d}{dx}\phi^i + c_i y^i \cdot \phi^i \right] dx. \tag{4.10}$$

Let now distributed and boundary forces, $f^i,\ g_J$ be given along the edge e_i and at the node v_J, respectively, which define a continuous linear functional in $\mathscr{V}$

$$L(\phi) := \sum_{i\in\mathscr{I}^f} \int_0^{\ell_i} f^i \cdot \phi^i dx + \sum_{J\in\mathscr{J}_N^g} g_J \cdot \phi^{\hat{i}}(v_J), \tag{4.11}$$

where $\hat{i}$ indicates that the simple nodes have just one incident edge, and where $f^i \in H^1(0,\ell_i)^*$. We now consider minimizing the energy over the set of constrained displacements. The Ritz approach to deriving the problem now can be stated as follows

$$\min_{y\in\mathscr{V}} \frac{1}{2} a(y,y) - L(y). \tag{4.12}$$

The fact that this convex optimization problem admits a unique solution is then proved by standard arguments. The classical first-order necessary optimality conditions in their strong formulation then read as follows.

$$\begin{cases} -K_i \dfrac{d^2}{dx^2} y^i + c_i y^i = f^i,\ \forall i \in \mathscr{I} \\ y^i(v_D) = 0,\ i \in \mathscr{I}_D,\ D \in \mathscr{J}_D \\ d_{iN} K_i \dfrac{d}{dx} y^i(v_N) = g_J,\ i \in \mathscr{I}_N,\ N \in \mathscr{J}_N\ , \\ y^i(v_J) = y^j(v_J),\ \forall i,j \in \mathscr{I}_J,\ J \in \mathscr{J}_M \\ \displaystyle\sum_{i\in\mathscr{I}_J} d_{iJ} K_i \frac{d}{dx} y^i(v_J) = 0,\ J \in \mathscr{J}_M \end{cases} \tag{4.13}$$

where $f^i = 0,\ i \in \mathscr{I} \setminus \mathscr{I}^f$, $g_N = 0$, and $J \in \mathscr{J}_N \setminus \mathscr{J}_N^g$. Notice that (4.13) line 6 is an example of the classical Kirchhoff condition known from electrostatics. See Lagnese, Leugering, and Schmidt [41, 42] f. (4.13) can be seen as an example of a general second-order planar elliptic problem on a metric graph with mixed boundary conditions. A general shape and topological sensitivity analysis even in this locally 1-D problem is not available in the literature.

In order to follow this cycle of ideas in a nutshell, we consider very simple such equations on networks, in fact on star graphs with $p = 1$. We provide some first-hand information before we embark on shape optimization problems related to problems on metric graphs. We consider the following star graph with m edges, m external nodes at $x = \ell_i$ and one multiple node at $x = 0$. On this graph we consider the Laplacean with Dirichlet conditions (Fig. 4.2).

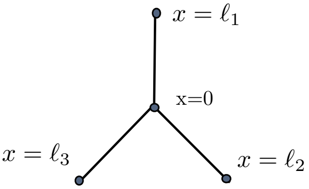


Fig. 4.2 The three-star

$$\begin{aligned} -\frac{d^2}{dx^2}y^i + c_i y^i &= f^i \text{ in } (0,\ell_i) \\ y^i(\ell_i) &= u^i \\ y^i(0) &= y^j(0), i \neq j = 1,\dots,m \\ \sum_{i=1}^{m} \frac{d}{dx} y^i(0) &= 0. \end{aligned} \tag{4.14}$$

Example 4.2.1 Let $f^i = 0, c^i = 0, i = 1,\dots,m$, then the solution is given by

$$y^i(x) = \frac{1}{\ell_i}\left(u^i - \frac{1}{\sum\limits_{j=1}^m \frac{1}{\ell_j}} \sum_{j=1}^m \frac{1}{\ell_j} u^j\right) x + \frac{1}{\sum\limits_{j=1}^m \frac{1}{\ell_j}} \sum_{j=1}^m \frac{1}{\ell_j} u^j. \tag{4.15}$$

Example 4.2.2 Let $f^i = 0, c_i > 0,\ i = 1,\dots,m$. Then the solution is given by

$$\begin{aligned} y^i(x) =& \frac{u^i}{\sinh(\sqrt{c_i}\ell_i)} \sinh(\sqrt{c_i}x) \qquad (4.16) \\ &- \left(\frac{1}{\sum\limits_{j=1}^m \sqrt{c_j}\coth(\sqrt{c_j}\ell_j)} \sum_{j=1}^m \frac{u^j\sqrt{c_j}}{\sinh(\sqrt{c_j}\ell_j)}\right) \coth(\sqrt{c_i}\ell_i)\sinh(\sqrt{c_i}x) \\ &+ \frac{1}{\sum\limits_{i=1}^m \sqrt{c_i}\coth(\sqrt{c_i}\ell_i)} \sum_{i=1}^m \frac{u^i\sqrt{c_i}}{\sinh(\sqrt{c_i}\ell_i)} \cosh(\sqrt{c_i}x),\ i = 1,\dots,m. \end{aligned}$$

Example 4.2.3 A Helmholtz-type example is given by $f^i = 0,\ c_i = -\omega^2 k_i < 0$, where ω denotes a frequency. In this case, the solutions have the following form

$$y^i(x) = a_i \sin(\omega\sqrt{k_i}x) + b_i \cos(\omega\sqrt{k_i}x),\ i = 1,\dots,m. \tag{4.17}$$

The same procedure as in the previous examples provides

$$b = \frac{1}{\sum_{i=1}^{m} \omega\sqrt{k_i}\cot(\omega\sqrt{k_i}\ell_i)} \sum_{i=1}^{m} \frac{u^i\sqrt{k_i}}{\sin(\omega\sqrt{k_i}\ell_i)} \tag{4.18}$$

$$\begin{aligned} a_i = &\frac{1}{\sin(\omega\sqrt{k_i}\ell_i)} \Bigg(u^i \\ &- \frac{1}{\sum_{j=1}^{m} \omega\sqrt{k_j}\cot(\omega\sqrt{k_i}\ell_j)} \sum_{j=1}^{m} \frac{u^j\sqrt{k_j}}{\sin(\omega\sqrt{k_i}\ell_j)} \cos(\omega\sqrt{k_i}\ell_i)\Bigg). \end{aligned} \tag{4.19}$$

We now look at the same problem, however, with different nodal conditions at the multiple center node.

$$\begin{aligned} -\frac{d^2}{dx^2}y^i + c_i y^i &= f^i \text{ in } (0,\ell_i) \\ y^i(\ell_i) &= u^i \\ \frac{d}{dx}y^i(0) &= \frac{d}{dx}y^j(0), i \neq j = 1,\dots,m \\ \sum_{i=1}^{m} y^i(0) &= 0. \end{aligned} \tag{4.20}$$

Remark 1 If we introduce the vectors

$$Y(0) = \begin{pmatrix} y^1 \\ y^2 \\ \ddots \\ y^m \end{pmatrix}(0),\ \frac{d}{dx}Y(0) = \begin{pmatrix} \frac{d}{dx}y^1 \\ \frac{d}{dx}y^2 \\ \ddots \\ \frac{d}{dx}y^m \end{pmatrix}(0)$$

and the matrices

$$A_0 = \begin{pmatrix} 1 & 0 & 0 & \dots & -1 \\ 0 & 1 & 0 & \dots & -1 \\ 0 & 0 & 1 & \dots & -1 \\ \dots & \dots & \dots & \dots & -1 \\ 0 & \dots & \dots & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad A_1 = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ 0 & \dots & \dots & \dots & 0 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}, \tag{4.21}$$

then we can write down the nodal conditions for (4.14) as follows

$$A_0 Y(0) + A_1 \frac{d}{dx}Y(0) = 0.$$

Accordingly, for (4.20), we have

$$A_0 \frac{d}{dx} Y(0) + A_1 Y(0) = 0.$$

We can, thus, identify the nodal conditions as vector-valued Robin-type conditions for the vectorial problem. In particular, if one normalizes the lengths to 1, which can always be achieved by introducing a diagonal matrix in the equations (see below), one can rewrite the star-graph problems as two-point boundary value problems for a vectorial elliptic equation or an elliptic system as follows

$$AY - \frac{d^2}{dx^2} Y = F \tag{4.22}$$

$$A_0 Y(0) + A_1 \frac{d}{dx} Y(0) = 0, \; Y(1) = U, \tag{4.23}$$

where A can be regarded as yet another coupling matrix; this time the coupling occurs in the domains, rather than on the boundary. In a sense, this system can be taken as 1-D prototype of a two-point boundary value problem for elliptic systems. This remark can be extended to general Sturm–Liouville type problems on star graphs and, in fact, on general metric graphs.

Remark 2 If the edges have individual lengths ℓ_i, we can use $w^i(x) = y^i(x\ell_i)$, $x \in (0, 1)$. We define diagonal matrices $\mathscr{L} := (\ell_i a_{ik})_{ik}$ where a_{ik} is the adjacency matrix: $a_{ik} = 1$ if edge i is connected to edge k ($\exists j \in \mathscr{J}^S : d_{ij} \neq 0, d_{kj} \neq 0$) and 0 else. In the same way $\mathscr{L}^2 := (\ell_i^2 a_{ik})_{ik}$. Then the system (4.22) is replaced with

$$\begin{aligned} &AY - \mathscr{L}^{-2} \frac{d^2}{dx^2} Y = F \\ &A_0 Y(0) + A_1 \mathscr{L}^{-1} \frac{d}{dx} Y(0) = 0, \; Y(1) = U. \end{aligned}$$

Thus, a problem with varying lengths can be reinterpreted as a problem with normalized lengths but with varying coefficients. In a more general context, this is precisely the basis for the method of the transformation on a fixed domain which is the basic tool in this book.

4.2.1 Steklov–Poincaré Map for Metric Graphs

Steklov–Poincaré maps provide a useful tool in nonoverlapping domain decomposition methods; see e.g., [42]. In the context of metric graphs, they can be computed rather explicitly. We consider a simple situation in order to show the concept. We consider again the general star graph. We also assume that the data $f^i, \; i = 1, \ldots, m$ are constant in a (small) neighborhood of the central node which we assume to be located at $x = 0$. That is to say, f^i constant for $i = 1, \dots, m$ for $x \in [0, \ell_i]$. We now take as the first m edges, the ones emerging from $x = 0$ with individual lengths ℓ_i. We introduce new serial nodal points n_i at $x = \ell_i$ for each edge e_i, $i = 1, \dots, m$. Emerging from the nodes n_i, $i = 1, \dots, m$ we introduce m new edges e_{i+m}, $i = 1, \dots, m$ starting at n_i with $x = 0$. Thus, we have m serial nodes for the newly arranged network. Our aim is to replace the inner star graph by the Steklov–Poincaré map and find new transmission conditions at the nodes n_i, $i = 1, .., m$. This is analogous to what has come to be known as domain decomposition by Steklov–Poincaré maps. In order to proceed, we first show the explicit solution on the inner star graph. In the case of constant coefficients and loads, the solution on each edge is clearly quadratic. In fact, simple calculations show that the solutions y^i look as follows

$$\begin{aligned} y^i(x) &= -\frac{1}{2}x^2 f^i \\ &+ \frac{1}{\ell_i}\left(u^i - \frac{1}{\sum\limits_{j=1}^m \frac{1}{\ell_j}} \sum_{j=1}^m \left(\frac{u^j}{\ell_j} + \frac{1}{2}\ell_j f^j \right) + \frac{1}{2}\ell_i^2 f^i \right) x \\ &+ \frac{1}{\sum\limits_{j=1}^m \frac{1}{\ell_j}} \left(\sum_{j=1}^m \left(\frac{u^j}{\ell_j} \right) + \frac{1}{2}\ell_i f^i \right), \; i = 1, \dots, m. \end{aligned}$$

The corresponding Steklov–Poincaré map, or Dirichlet–Neumann map is then given by

$$\begin{aligned} \mathscr{S}_i(u) &= \frac{d}{dx} y^i(\ell_i) \\ &= \frac{1}{\ell_i}u^i - \frac{1}{\ell_i}\frac{1}{\sum\limits_{j=1}^m \frac{1}{\ell_j}} \left(\sum_{j=1}^m \frac{u^j}{\ell_j} + \frac{1}{2}\ell_i f^i \right) - \frac{1}{2}\ell_i f^i \\ &= \frac{1}{\ell_i}u^i - \frac{1}{\ell_i}\frac{1}{\sum\limits_{j=1}^m \frac{1}{\ell_j}} \sum_{j=1}^m \frac{u^j}{\ell_j} - \frac{1}{\ell_i}\frac{1}{\sum\limits_{j=1}^m \frac{1}{\ell_j}} \sum_{j=1}^m \frac{1}{2}\ell_j f^j - \frac{1}{2}\ell_i f^i. \end{aligned} \tag{4.24}$$

As mentioned above, we now assume that the connecting edges e_{i+m}, $i = 1, \dots, m$ start at the nodes n_i, $i = 1, \dots, m$, then the Steklov–Poincaré-based boundary conditions there read as follows

$$\frac{d}{dx}y^{i+m}(0) - \frac{1}{\ell_i}y^{i+m}(0) + \frac{1}{\ell_i}\frac{1}{\sum\limits_{j=1}^m \frac{1}{\ell_j}} \sum_{j=1}^m \frac{y^{j+m}(0)}{\ell_j} \tag{4.25}$$

$$= -\frac{1}{\ell_i}\frac{1}{\sum\limits_{j=1}^{m}\frac{1}{\ell_j}}\sum_{j=1}^{m}\frac{1}{2}\ell_j f^j - \frac{1}{2}\ell_i f^i,\ i = 1, \ldots, m.$$

Remark 3 The Robin boundary conditions (4.25) are still coupled over the endpoints of the star graph that has been cut out using the Steklov–Poncaré map (4.24). It is interesting to note that these coupling conditions are self-adjoint. Indeed, if, for the sake of simplicity, we put all lengths equal to 1 and neglect the loads f^i, $i = 1, .., 3$ then, upon introducing the vectors

$$Y(0) = \begin{pmatrix} y^1 \\ y^2 \\ \ddots \\ y^m \end{pmatrix}(0),\ \frac{d}{dx}Y(0) = \begin{pmatrix} \frac{d}{dx}y^1 \\ \frac{d}{dx}y^2 \\ \ddots \\ \frac{d}{dx}y^m \end{pmatrix}(0)$$

and the matrices

$$A_0 = \begin{pmatrix} 1-\frac{1}{m} & -\frac{1}{m} & -\frac{1}{m} & \ldots & -\frac{1}{m} \\ -\frac{1}{m} & 1-\frac{1}{m} & -\frac{1}{m} & \ldots & -\frac{1}{m} \\ -\frac{1}{m} & -\frac{1}{m} & 1-\frac{1}{m} & \ldots & -\frac{1}{m} \\ \ldots & \ldots & \ldots & \ldots & -\frac{1}{m} \\ -\frac{1}{m} & \ldots & \ldots & -\frac{1}{m} & 1-\frac{1}{m} \end{pmatrix},\quad A_1 = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ 0 & \ldots & \ldots & \ldots & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \tag{4.26}$$

we can write down the nodal conditions for (4.14) as follows

$$A_0 Y(0) + A_1\frac{d}{dx}Y(0) = 0.$$

We have $A_0 A_1^T$ is symmetric and, hence, the boundary conditions are self-adjoint.

For the second transmission condition according to (4.20), we obtain

$$y^i(x) = -\frac{1}{2}x^2 f^i + ax + b_i,\ i = 1, \ldots, m \tag{4.27}$$

$$a = \frac{1}{\sum\limits_{j=1}^{m}\ell_j}\left(\sum_{j=1}^{m}\left(u^j + \frac{1}{2}\ell_j^2 f^j\right)\right)$$

$$b_i = u^i - \frac{1}{\sum\limits_{j=1}^{m}\ell_j}\left(\sum_{j=1}^{m}\left(u^j + \frac{1}{2}\ell_j^2 f^j\right)\right)\ell_i + \frac{1}{2}\ell_i^2 f^i,\ i = 1, \ldots, m. \tag{4.28}$$

The corresponding Steklov–Poincaré map is then given by

$$\begin{aligned}\mathscr{S}_i(u) &= \frac{d}{dx}y^i(\ell_i) \\ &= \frac{1}{\sum\limits_{j=1}^{m}\ell_j}\left(\sum_{j=1}^{m}u^j + \frac{1}{2}\sum_{j=1}^{m}\ell_j^2 f^j\right) - \ell_i f^i.\end{aligned} \tag{4.29}$$

If we now assume that the connecting edges $e_{i+m},\ i = 1, \dots, m$ start at the nodes $n_i,\ i = 1, \dots, m$, then the Steklov–Poincaré-based boundary conditions there read as follows

$$\frac{d}{dx}y^{i+m}(0) - \frac{1}{\sum\limits_{j=1}^{m}\ell_j}\left(\sum_{j=1}^{m}y^{j+m}(0)\right) \tag{4.30}$$

$$= \frac{1}{\sum\limits_{j=1}^{m}\ell_j}\left(\frac{1}{2}\sum_{j=1}^{m}\ell_j^2 f^j\right) - \ell_i f^i,\ i = 1, \dots, m. \tag{4.31}$$

Similar calculations can be done for reverse Steklov–Poincaré problem, namely, where one starts with Neumann data and solves for the Dirichlet data at the same nodes.

We continue our discussions in the next section, where we resort again to the general graphs.

4.2.2 Problems on General Metric Graphs

The Steklov–Poincaré or Dirichlet–Neumann maps, together with the corresponding Neumann–Dirichlet maps, can be used to decompose any given problem on a metric graph into star graphs. Let us briefly consider this method. We use the standard notation introduce at the beginning of this section and consider the problem on a general network.

The complete system of type (4.14) reads as follows

$$\begin{aligned} y^i - \gamma_i \frac{d^2}{dx^2}y^i &= f^i, i \in \mathscr{I}, x \in (0,1) \\ y^i(n_j) &= y^k(n_j), \forall i,k \in \mathscr{I}_j, j \in \mathscr{J}^M \\ \sum_{i \in \mathscr{I}_j} d_{ij}\gamma_i \frac{d}{dx}y^i(n_j) &= 0, j \in \mathscr{J}^M \end{aligned} \tag{4.32}$$

$$
\begin{aligned}
y^i(n_j) &= u^j, i \in \mathscr{I}_j, j \in \mathscr{J}_D^S \\
\frac{d}{dx} y^i(n_j) &= u^j, i \in \mathscr{I}_j, j \in \mathscr{J}_N^S,
\end{aligned}
$$

whereas the complete elliptic network problem of type (4.20) is written as

$$
\begin{aligned}
y^i - \gamma_i \frac{d^2}{dx^2} y^i &= f^i, i \in \mathscr{I}, x \in (0,1) \\
\gamma_i \frac{d}{dx} y^i(n_j) &= \gamma_k \frac{d}{dx} y^k(n_j), \forall i, k \in \mathscr{I}_j, j \in \mathscr{J}^M \\
\sum_{i \in \mathscr{I}_j} d_{ij} y^i(n_j) &= 0, j \in \mathscr{J}^M \\
y^i(n_j) &= u^j, i \in \mathscr{I}_j, j \in \mathscr{J}_D^S \\
\frac{d}{dx} y^i(n_j) &= u^j, i \in \mathscr{I}_j, j \in \mathscr{J}_N^S.
\end{aligned}
\tag{4.33}
$$

The problem of well-posedness (4.33) is considered next. The analysis is similar for the even more classical system (4.32) which has been analyzed in the literature. See e.g., [41]. It turns out that, in this case, for any graph G, such that each node is connected to a Dirichlet node by a simple path, the underlying elliptic operator is self-adjoint and positive definite, thus, the problem admits a unique solution. Indeed, the Lagrange identity

$$
\begin{aligned}
\langle Ay, \phi \rangle &:= \sum_{i \in \mathscr{I}} \int_0^1 \left(y^i - \gamma_i \frac{d^2}{dx^2} y^i - f^i \right) \phi^i dx \qquad (4.34) \\
&= -\sum_{j \in \mathscr{J}} \sum_{i \in \mathscr{I}_j} d_{ij} \gamma_i \frac{d}{dx} y^i(n_j) \phi^i(n_j) + \sum_{i \in \mathscr{I}} \int_0^1 \left(y^i \phi^i + \gamma_i \frac{d}{dx} y^i \frac{d}{dx} \phi^i - f^i \phi^i \right) dx \\
&= -\sum_{j \in \mathscr{J}} \sum_{i \in \mathscr{I}_j} d_{ij} \gamma_i \frac{d}{dx} y^i(n_j) \phi^i(n_j) + \sum_{j \in \mathscr{J}} \sum_{i \in \mathscr{I}_j} d_{ij} \gamma_i y^i(n_j) \frac{d}{dx} \phi^i(n_j) + \langle y, A\phi \rangle \\
&= \langle y, A\phi \rangle, \ \forall y, \phi \in D(A),
\end{aligned}
$$

where the domain $D(A)$ is given by

$$
\begin{aligned}
D(A) := \Bigg\{ & y = (y^i)|_{i \in \mathscr{I}} \in \Pi_{i \in \mathscr{I}} H^2(0,1) | y^i(n_j) = 0, i \in \mathscr{I}_j, j \in \mathscr{J}_D^S \\
& \frac{d}{dx} y^i(n_j) = 0, i \in \mathscr{I}_j, j \in \mathscr{J}_N^S \\
& \gamma_i \frac{d}{dx} q(n_j) = \gamma_k \frac{d}{dx} y^k(n_j), \forall i, k \in \mathscr{I}_j, j \in \mathscr{J}^M \\
& \sum_{i \in \mathscr{I}_j} d_{ij} y^i(n_j) = 0, j \in \mathscr{J}^M \Bigg\}.
\end{aligned}
\tag{4.35}
$$

Thus, we can define the unbounded, self-adjoint operator

$$Aq := \left(y^i - \gamma_i \frac{d^2}{dx^2} y^i \right)_{i \in \mathscr{I}} \tag{4.36}$$

with domain $D(A)$ in the Hilbert space $\mathscr{H} := \Pi_{i \in \mathscr{I}} L^2(0, 1)$, which is positive definite. We may also introduce the energy space

$$\begin{aligned} V := \Bigg\{ & y = (y^i)|_{i \in \mathscr{I}} \in \Pi_{i \in \mathscr{I}} H^1(0, 1) | y^i(n_j) = 0, i \in \mathscr{I}_j, j \in \mathscr{J}_D^S \\ & \frac{d}{dx} y^i(n_j) = 0, i \in \mathscr{I}_j, j \in \mathscr{J}_N^S \\ & \sum_{i \in \mathscr{I}_j} d_{ij} y^i(n_j) = 0, j \in \mathscr{J}^M \Bigg\}. \end{aligned} \tag{4.37}$$

Theorem 4.2.4 *Let a network $G = (V, E)$ be given such that each node is connected to a Dirichlet node by a simple path. Moreover, let controls u_j, $j \in \mathscr{J}^S$ and right-hand sides $f \in \mathscr{H}$ be given. Then there exists a unique solution $y \in D(A)$ to $Ay = f$* (4.33). *If $f \in V^*$, then there is a unique solution $y \in V$.*

We continue the discussion for the use of the Dirichlet–Neumann (Steklov–Poincaré) and Neumann–Dirichlet maps for the purpose of decomposing problems on a general graph G into star-graph problems. Assume we are given a multiple node n_j, $j \in \mathscr{J}^M$ with edge degree d_j. Then d_j edges emanate from n_j. Denote these edges $i_1^0, \dots i_{d_j}^0$. If we cut these edges in the middle, then we obtain d_j serial nodes $n_{j_1}, \dots, n_{j_{d_j}}$, where we apply the canonical transmission conditions and another set of d_j edges with labels $i_1^1, \dots i_{d_j}^1$. Then we impose

$$y^{i_1^0}(n_{j_1}) = y^{i_1^1}(n_{j_1}), \dots, y^{i_{d_j}^0}(n_{j_{d_j}}) = y^{i_{d_j}^1}(n_{j_{d_j}}),$$

and

$$\frac{d}{dx} y^{i_1^0}(n_{j_1}) = \frac{d}{dx} y^{i_1^1}(n_{j_1}), \dots, \frac{d}{dx} y^{i_{d_j}^0}(n_{j_{d_j}}) = \frac{d}{dx} y^{i_{d_j}^1}(n_{j_{d_j}}).$$

Then we use the Dirichlet–Neumann (D2N) or the Neumann–Dirichlet (N2D) mappings in order to cut out the star graph with center node n_j and the edges with indices $i_1^0, \dots i_{d_j}^0$ as indicated above. Therefore, for a first approach, it is sufficient to consider star graphs as they contain the problem of multiple nodes. We do not dwell further on problems on general graphs here and instead refer to [41, 42].

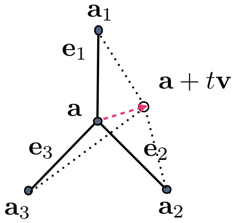


Fig. 4.3 The deformed three-star

4.2.3 Shape Variations for Graph Problems

We are now interested in moving the center node of the graph, Fig. 4.3, and to calculate the sensitivities with respect to that movement. We introduce the following notation. Let $\mathbf{a_i}$, $i = 1, \dots, m$ and $\mathbf{a}$ be vectors in the plane, where the $\mathbf{a}_i$, $i = 1, \dots, m$ denote the endpoints of the edges labeled $1, \dots, m$ and $\mathbf{a}$ denotes the center node in the reference configuration. We are going to move $\mathbf{a}$ in the direction $\mathbf{v}$. To this end, we introduce the normalized edge vectors

$$\mathbf{e}_i := \mathbf{e}_i(0, \mathbf{0}) := \frac{\mathbf{a} - \mathbf{a_i}}{\|\mathbf{a} - \mathbf{a}_i\|}, \quad \mathbf{e}_i(t, \mathbf{v}) := \frac{\mathbf{a} + t\mathbf{v} - \mathbf{a_i}}{\|\mathbf{a} + t\mathbf{v} - \mathbf{a}_i\|}, \quad i = 1, \dots, m. \tag{4.38}$$

See Fig. 4.3. The edges have lengths $\ell_i := \|\mathbf{a} - \mathbf{a}_i\|$, $i = 1, \dots, m$.

There are several ways to introduce a vector field that transports the point $\mathbf{a}$ to the point $\mathbf{a} + t\mathbf{v}$, $t > 0$. One natural choice is

$$\begin{aligned} T_t(\mathbf{v}) &: \mathbb{R}^2 \to \mathbb{R}^2 \\ T_t(\mathbf{v})(\mathbf{a}_i) = \mathbf{a}_i,\ & T_t(\mathbf{v})(\mathbf{a}) = \mathbf{a} + t\mathbf{v} \end{aligned} \tag{4.39}$$

We can realize this field by introducing a velocity $V(\mathbf{a}_i) = 0$, $i = 1, \dots, m$ and $V(\mathbf{a}) = \mathbf{v}$, hence,

$$V(\mathbf{x}) = \frac{\|\mathbf{x} - \mathbf{a}_1\| \|\mathbf{x} - \mathbf{a}_2\| \dots \|\mathbf{x} - \mathbf{a}_m\|}{\|\mathbf{a} - \mathbf{a}_1\| \|\mathbf{a} - \mathbf{a}_2\| \dots \|\mathbf{a} - \mathbf{a}_m\|} \mathbf{v}. \tag{4.40}$$

Note, however, that V deforms the graph in that the images of the edges $T_t(\mathbf{v})(\mathbf{a}_i + s\mathbf{e}_1) = \mathbf{a}_i + s\mathbf{e}_i + tV(\mathbf{a}_1 + s\mathbf{e}_i)$ are no longer straight lines. Alternatively, we can define m linear fields as follows.

$$T_t^i(\mathbf{v})(\mathbf{a}_i + s\mathbf{e}_i) = \mathbf{a}_i + s\mathbf{e}_i + t\frac{s\mathbf{v}}{\ell_i} \tag{4.41}$$

which map the straight edges $\mathbf{a}_i + s\mathbf{e}_i$, $s \in [0, \ell_i]$ onto the straight edges $\mathbf{a}_i + \left(s\mathbf{e}_i + t\frac{\mathbf{v}}{\ell_i}\right)$, $s \in [0, \ell_i]$. Let us introduce the lengths of the edges in the moving domain $\ell_i(t) := \|\mathbf{a} + t\mathbf{v} - \mathbf{a}_i\|$, $i = 1, \dots, m$. We also denote by $\ell_i(\hat{\mathbf{a}})$, the individual length with a given center multiple node $\hat{\mathbf{a}}$. In case of $\hat{\mathbf{a}} = \mathbf{a} + t\mathbf{v} - \mathbf{a}_i$ we have $\ell_i(\hat{\mathbf{a}}) = \ell_i(t)$.We can now pose the following optimization problem, where we look for the optimal position in an ϵ neighborhood of the given positon of the multiple node.

$$\begin{aligned} \min_{\hat{\mathbf{a}} \in B_{\epsilon(\mathbf{a})}} \frac{1}{2} \sum_{i=1}^{m} \int\limits_{0}^{\ell_i(\mathbf{a})} |y^i(x) - y_d(x)|^2 dx \text{ s.t.} \\ -\frac{d^2}{dx^2} y^i + c_i y^i = f^i \text{ in } (0, \ell_i(\hat{\mathbf{a}})) \\ y^i(\ell_i(\hat{\mathbf{a}})) = u^i \\ y^i(0) = y^j(0),\ i \neq j = 1, \ldots, m \\ \sum_{i=1}^{m} \frac{d}{dx} y^i(0) = 0. \end{aligned} \tag{4.42}$$

Example 4.2.5 We let $c_i = 0,\ i = 1, \ldots, m$ and $y_d(x) = 0,\ x \in [0, \ell_i],\ i = 1, \ldots, m$. According to Fig. 4.3, we consider the following transformations. See (4.39).

$$T_t^i(\mathbf{v})(x) := \mathbf{a}_i + x\mathbf{e}_i + t\frac{\mathbf{v}}{\ell_i}x,\ i = 1, \ldots, m. \tag{4.43}$$

Then, clearly, $T_t^i(\mathbf{v})(0) = \mathbf{a}_i,\ i = 1, \ldots, m$ and $T_t^i(\mathbf{v})(\ell_i) = \mathbf{a} + t\mathbf{v}$. We can reformulate problem (4.42) as follows.

$$\begin{aligned} -\frac{d^2}{dx^2} y^i = 0 \text{ in } (0, \ell_i(\hat{\mathbf{a}})) \\ y^i(0) = u^i \\ y^i(\ell_i(t)) = y^j(\ell_j(t)),\ i \neq j = 1, \ldots, m \\ \sum_{i=1}^{m} \frac{d}{dx} y^i(\ell_i(t)) = 0. \end{aligned} \tag{4.44}$$

Let us observe that the unique solution is given by

$$y_t^i(x) = \frac{1}{\ell_i(t)} \left(u^i - \frac{1}{\sum\limits_{j=1}^{m} \frac{1}{\ell_j(t)}} \sum_{j=1}^{m} \frac{u^j}{\ell_j(t)} \right) (\ell_i(t) - x) + \frac{1}{\sum\limits_{j=1} m \frac{1}{\ell_j(t)}} \sum_{j=1} m \frac{u^j}{\ell_j(t)}. \tag{4.45}$$

On the fixed graph, we have

$$
\begin{aligned}
(y^i)^t(x) &= y^i_t(T_t(\mathbf{v})(x) \qquad (4.46)\\
&= \frac{1}{\ell_i(t)}\left(u^i - \frac{1}{\sum\limits_{j=1}^m \frac{1}{\ell_j(t)}}\sum_{j=1}^m \frac{u^j}{\ell_j(t)}\right)\left(\ell_i(t) - \frac{\ell_i(t)}{\ell_i}x\right)\\
&\quad + \frac{1}{\sum\limits_{j=1}^m \frac{1}{\ell_j(t)}}\sum_{j=1}^m \frac{u^j}{\ell_j(t)}.
\end{aligned}
$$

From this it is possible to derive the material and shape derivative, $\dot{y}^i$, $(y^i)'$, directly. We obtain for the material derivatives

$$\dot{y}^i(0) = 0,\ i = 1, \ldots, m, \qquad (4.47a)$$

$$\dot{y}^i(\ell_i) = \frac{1}{\left(\sum\limits_{j=1}^m \frac{1}{\ell_j}\right)^2}\sum_{j=1}^m \frac{\ell_j'(0)}{\ell_j^2}\sum_{i=1}^m \frac{u^j}{\ell_j} - \frac{1}{\sum\limits_{j=1}^m \frac{1}{\ell_j}}\sum_{j=1}^m \frac{u^j \ell_j'(0)}{\ell_j^2} \qquad (4.47b)$$

which is independent of i and

$$\sum_{i=1}^m \frac{d}{dx}\dot{y}^i(\ell_i) = -\sum_{i=1}^m \frac{\ell_i'(0)}{\ell_i}\frac{d}{dx}y^i(\ell_i). \qquad (4.47c)$$

For the shape derivatives, we get

$$
\begin{aligned}
&\frac{d^2}{dx^2}(y^i)' = 0, \quad \text{in } (0, \ell_i(\hat{\mathbf{a}}))\\
&(y^i)'(0) = 0,\ i = 1, \ldots, m, \qquad (4.48a)\\
&(y^i)'(\ell_i) - (y^j)'(\ell_j) = \ell_j'(0)\frac{d}{dx}y^j(\ell_j) - \ell_i'(0)\frac{d}{dx}y^i(\ell_i) \qquad (4.48b)\\
&\sum_{i=1}^m \frac{d}{dx}(y^i)'(\ell_i) = 0.
\end{aligned}
$$

These relations can be verified using the system (4.44) directly, without using the explicit solution (4.45) (Fig. 4.4).

Example 4.2.6 Let $f^i = c_i = 0,\ i = 1, \ldots, m$ and $y_d(x) = 0,\ x \in [0, \ell(\hat{\mathbf{a}})]$. Then the solution on the moving graph is given by

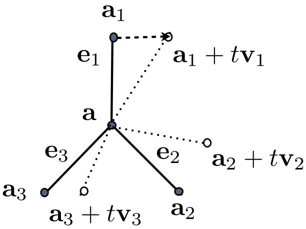


Fig. 4.4 The three-star with transformed edges

$$y_t^i(x) = \frac{1}{\ell_i(t)}\left(u^i - \frac{1}{\sum\limits_{j=1}^m \frac{1}{\ell_j(t)}}\sum_{j=1}^m \frac{u^j}{\ell_j(t)}\right)x + \frac{1}{\sum\limits_{j=1}^m \frac{1}{\ell_j(t)}}\sum_{j=1}^m \frac{u^j}{\ell_j(t)}.$$

We can equally well keep the center and move the three edges. It turns out, in fact, that for such scalar problems, the solution does not depend on the angles between the edges. In scalar problems the variable $y^i(x)$ is seen as either a temperature, in case we are dealing with a heat transport problem on the graph, or it is viewed as an out-of-the-plane displacement, in case the problem is considered as representing a network of strings with displacements pointing out of the plane. For problems on graphs in the plane, where the variables $y^i(x) \in \mathbb{R}^2$ are vectors in the plane, the situation is different. This case will be treated later. It is, thus, sufficient to treat the case, where

$$T_t^i(\mathbf{v})(s) := \frac{\ell_i(t)}{\ell_i}s, \; s \in [0, \ell_i], \tag{4.49}$$

where

$$\ell_i(t) = \|\mathbf{a}_i - \mathbf{a} + t\mathbf{v}_i\|. \tag{4.50}$$

Clearly,

$$\ell_i'(t) = \frac{(\mathbf{a}_i - \mathbf{a} + t\mathbf{v}_i, \mathbf{v}_i)}{\|\mathbf{a}_i - \mathbf{a} + t\mathbf{v}_i\|}, \; \ell_i'(0) = \frac{(\mathbf{a}_i - \mathbf{a}, \mathbf{v}_i)}{\|\mathbf{a}_i - \mathbf{a}\|}. \tag{4.51}$$

We first transform the cost functional onto the fixed graph

$$\sum_0^m \int_0^{\ell_i(t)} |y_t^i(x)|^2 dx = \sum_0^m \frac{\ell_i(t)}{\ell_i}\int_0^{\ell_i} |y_t^i\left(\frac{\ell_i(t)}{\ell_i}s\right)|^2 ds.$$

We differentiate with respect to t and obtain

$$\begin{aligned}\frac{d}{dt}\sum_{0}^{m}\int_{0}^{\ell_i(t)}|y_t^i(x)|^2dx &= \frac{d}{dt}\sum_{0}^{m}\frac{\ell_i(t)}{\ell_i}\int_{0}^{\ell_i}\left|y_t^i\left(\frac{\ell_i(t)}{\ell_i}s\right)\right|^2 ds\\ &= \sum_{i=1}^{m}\frac{\ell_i'(t)}{\ell_i}\int_{0}^{\ell_i}\left|y_t^i\left(\frac{\ell_i(t)}{\ell_i}s\right)\right|^2 ds+\\ \sum_{0}^{m}\frac{\ell_i(t)}{\ell_i}\int_{0}^{\ell_i}2y_t^i&\left(\frac{\ell_i(t)s}{\ell_i}\right)\frac{\partial}{\partial t}y_t^i\left(\frac{\ell_i(t)s}{\ell_i}\right)\frac{\ell_i'(t)}{\ell_i}sds.\end{aligned} \tag{4.52}$$

We evaluate (4.52) at $t = 0$ and obtain

$$\frac{d}{dt}\sum_{0}^{m}\int_{0}^{\ell_i(t)}|y_t^i(x)|^2dx|_{t=0} = \sum_{i=1}^{m}\frac{\ell_i'(0)}{\ell_i}\int_{0}^{\ell_i}|y^i(s)|^2ds + 2\sum_{0}^{m}\int_{0}^{\ell_i}\dot{y^i}(s)y^i(s)\frac{\ell_i'(0)}{\ell_i}sds, \tag{4.53}$$

where now $\dot{y^i}(s)$ is the material derivative of $y^i(s)$. We use the relation between the shape derivative $(y^i)'$ and the material derivative $\dot{y}^i$, namely,

$$\dot{y}^i(x) = (y^i)'(x) + \frac{d}{dx}y^i(x)\frac{\ell_i'(0)}{\ell_i}x.$$

Then (4.53) takes the form

$$\begin{aligned}\frac{d}{dt}\sum_{0}^{m}\int_{0}^{\ell_i(t)}|y_t^i(x)|^2dx|_{t=0} &= \sum_{i=1}^{m}\frac{\ell_i'(0)}{\ell_i}\int_{0}^{\ell_i}|y^i(s)|^2ds\\ +2\sum_{0}^{m}\int_{0}^{\ell_i}(y^i)'(s) &+ \frac{d}{dx}y^i(s)\frac{\ell_i'(0)}{\ell_i}sy^i(s)\frac{\ell_i'(0)}{\ell_i}sds\\ &= \sum_{i=1}^{m}\ell_i'(0)y^i(\ell_i)^2 + 2\sum_{i=1}^{m}\int_{0}^{\ell_i}(y^i)'(x)y^i(x)dx\\ &= \sum_{i=1}^{m}(\mathbf{e}_i,\mathbf{v}_i)y^i(\ell_i)^2 + 2\sum_{i=1}^{m}\int_{0}^{\ell_i}(y^i)'(x)y^i(x)dx.\end{aligned} \tag{4.54}$$

We consider the problem that the shape derivatives $(y^i)'(\cdot),\ i = 1, \dots, m$ have to solve. This problem can be derived by direct differentiation.

$$\begin{aligned}
-\frac{d^2}{dx^2}(y^i)' &= 0 \text{ in } (0,\ell_i)\\
(y^i)'(\ell_i) &= -\frac{d}{dx}y^i(\ell_i)\ell_i'(0)\\
(y^i)'(0) &= (y^j)'(0), i \neq j = 1,\dots,m\\
\sum_{i=1}^m \frac{d}{dx}(y^i)'(0) &= 0.
\end{aligned}\tag{4.55}$$

Furthermore, we introduce the adjoint problem for the adjoint variable $p^i(\cdot)$ as follows.

$$\begin{aligned}
-\frac{d^2}{dx^2}p^i &= y^i \text{ in } (0,\ell_i)\\
p^i(\ell_i) &= 0\\
p^i(0) &= p^j(0), i \neq j = 1,\dots,m\\
\sum_{i=1}^m \frac{d}{dx}p^i(0) &= 0.
\end{aligned}\tag{4.56}$$

We use the adjoint in (4.55) and integrate by parts and use the boundary conditions for p^i in (4.56). We have

$$\begin{aligned}
-\sum_{i=1}^m \int_0^{\ell_i}(y^i)'\frac{d^2}{dx^2}p^i dx &= -\sum_{i=1}^m (y^i)'\frac{d}{dx}p^i|_0^{\ell_i} + \sum_{i=1}^m \frac{d}{dx}(y^i)'p^i|_0^{\ell_i}\\
&= \sum_{i=1}^m \frac{d}{dx}y^i(\ell_i)\frac{d}{dx}p^i(\ell_i)\ell_i'(0).
\end{aligned}$$

This gives the shape derivative of the cost functional as follows.

$$dJ(y,v) = \sum_{i=1}^m\left(\frac{1}{2}y^i(\ell_i)^2 + \frac{d}{dx}y^i(\ell_i)\frac{d}{dx}p^i(\ell_i)\right)(\mathbf{e}_i,\mathbf{v}_i)\tag{4.57}$$

$$= \sum_{i=1}^m\left(\frac{1}{2}(u^i)^2 + \frac{d}{dx}y^i(\ell_i)\frac{d}{dx}p^i(\ell_i)\right)(\mathbf{e}_i,\mathbf{v}_i).\tag{4.58}$$

We can, for instance, consider the following scenario: let $u^i{=}u$, $\ell_i{=}\ell$ $i{=}1,\dots,m$. It is clear from (4.49) evaluated at $t = 0$ that $y_0^i(\ell_i) = u \neq 0$, $i = 1,\dots,m$ and $\frac{d}{dx}y^i(\ell_i) = 0$ $i = 1,\dots,m$. Thus,

$$dJ(y,v) = \frac{1}{2}(u)^2\sum_{i=1}^m(\mathbf{e}_i,\mathbf{v}_i).$$

If we take the velocity vectors equal $\mathbf{v}_i = \mathbf{v},\ i = 1, \dots, m$, then $\sum\limits_{i=1}^{m} \mathbf{e}_i = 0$ is a condition which leads to stationarity. This is the configuration where all edges have the same angle of $\frac{2\pi}{m}$.

4.2.4 Shape Sensitivity for the First Eigenvalue

We consider the following system on the star graph

$$\begin{aligned} -\frac{d^2}{dx^2} z^i &= \lambda z^i, i \in \mathscr{I}, x \in (0, \ell_i) \\ z^i(0) &= z^k(0), \forall i, k = 1, \dots, m \\ \sum_{i=1}^{m} \frac{d}{dx} z^i(0) &= 0, \qquad (4.59) \\ z^i(\ell_i) &= 0, i = 1, \dots, m. \qquad (4.60) \end{aligned}$$

We consider the smallest eigenvalue λ. To this end, we define

$$a(z, \phi) := \sum_{i=1}^{m} \int_0^{\ell_i} \frac{d}{dx} y^i \frac{d}{dx} \phi^i dx, \qquad (z, \phi) := \sum_{i=1}^{m} \int_0^{\ell_i} y^i \phi^i dx \qquad (4.61)$$

$$V := \{y \in \Pi_{i=1}^{m} H^1(0, \ell_i) | y^i(0) = y^j(0), i, j = 1, \dots, m, y^i(\ell_i) = 0\}.$$

We pose the eigenvalue problem in variational form as follows.

$$a(z, \phi) = \lambda(z, \phi),\ \forall \phi \in V. \qquad (4.62)$$

The smallest eigenvalue satisfies

$$\lambda(G) = \min_{\phi \in V} \{a(\phi, \phi) | \|\phi\| = 1\}. \qquad (4.63)$$

Clearly, $\lambda(G) > 0$. We now consider G_t, where the edges are dependent on t: $[0, \ell_i(t)]$. Consequently, we look for the smallest eigenvalue $\lambda(G_t)$:

$$\lambda(G_t) = \min_{\phi \in V_t} \{a_t(\phi, \phi) | \|\phi\| = 1\} = \min_{\phi \in V_t \setminus \{0\}} \left\{ \frac{a_t(\phi, \phi)^2}{\|\phi\|} \right\}, \qquad (4.64)$$

where

$$V_t := \{y \in \Pi_{i=1}^m H^1(0, \ell_i(t)) | y^i(0) = y^j(0), i, j = 1, \dots, m, y^i(\ell_i(t) = 0\}. \tag{4.65}$$

We introduce

$$F(t, \psi) := \frac{a^t(\psi, \psi)}{\|\sqrt{\gamma(t, x)}\psi\|^2},$$

with $\gamma^i(t, x) := \partial_x T_t^i(x)$. We obtain

$$\frac{\partial}{\partial t} F(0, \psi) = \frac{a'(\psi, \psi)\|\psi\|^2 - a(\psi, \psi)\|\sqrt{\gamma'(0)}\psi\|^2}{\|\sqrt{\gamma(0)}\psi\|^4} \tag{4.66}$$

$$= a'(\psi, \psi) - \lambda \sum_{i=1}^m \int_0^{\ell_i} |\psi^i|^2 \partial_x V^i(0, x) dx, \tag{4.67}$$

where we used that $\|\psi\| = 1, \gamma(0) = 1$. On the other side, by direct computation

$$-\frac{d^2}{dx^2}(z^i)' = \lambda (z^i)' + \lambda' z^i, i \in \mathscr{I}, x \in (0, \ell_i)$$

$$(z^i)'(0) = (z^k)'(0), \forall i, k = 1, \dots, m$$

$$\sum_{i=1}^m \frac{d}{dx}(z^i)'(0) = 0, \tag{4.68}$$

$$(z^i)'(\ell_i) = -\frac{d}{dx} z^i V^i(0, \ell_i), i = 1, \dots, m. \tag{4.69}$$

We multiply by the eigenelement z^i and obtain

$$-\sum_{i=1}^m \int_0^{\ell_i} \frac{d^2}{dx^2}(z^i)' z^i dx = \lambda \sum_{i=1}^m \int_0^{\ell_i} (z^i)' z^i dx + \lambda'. \tag{4.70}$$

We come back to to the constrained optimization problem and introduce the Lagrangean:

$$\mathscr{L}(G; z, \beta) := \sum_{i=1}^m \int_0^{\ell_i} \left| \frac{d^2}{dx^2}(z^i) \right|^2 dx + \beta \left(\sum_{i=1}^m \int_0^{\ell_i} |z^i|^2 dx - 1 \right). \tag{4.71}$$

Shape variation in the direction of V yields at the optimum

$$\begin{aligned} 0 = d\mathscr{L}(G; z, \beta; V) &= \sum_{i=1}^{m} \int\limits_{0}^{\ell_i} \left(2\frac{d}{dx}(z^i)' \frac{d}{dx} z^i + \frac{d}{dx}(\left(\frac{d}{dx} z^i \right)^2 V^i(0, x) \right) dx \\ &\quad + \beta \sum_{i=1}^{m} \int\limits_{0}^{\ell_i} \left(2(z^i)' z^i + \frac{d}{dx}((z^i)^2 V^i(0, x) \right) dx \\ &= \sum_{i=1}^{m} \int\limits_{0}^{\ell_i} 2\frac{d}{dx}(z^i)' \frac{d}{dx} z^i dx + \sum_{i=1}^{m} \left(\frac{d}{dx} z^i \right)^2 (\ell_i) V^i(0, \ell_i) + 2\beta \sum_{i=1}^{m} \int\limits_{0}^{\ell_i} (z^i)' z^i dx. \end{aligned}$$

This implies

$$\sum_{i=1}^{m} \int\limits_{0}^{\ell_i} \frac{d}{dx}(z^i)' \frac{d}{dx} z^i dx + \frac{1}{2} \sum_{i=1}^{m} \left(\frac{d}{dx} z^i \right)^2 (\ell_i) V^i(0, \ell_i) = -\beta \sum_{i=1}^{m} \int\limits_{0}^{\ell_i} (z^i)' z^i dx. \tag{4.72}$$

Together with (4.70), (4.72) implies

$$\lambda \sum_{i=1}^{m} \int\limits_{0}^{\ell_i} (z^i)' z^i dx + \lambda' + \frac{1}{2} \sum_{i=1}^{m} \left(\frac{d}{dx} z^i \right)^2 (\ell_i) V^i(0, \ell_i) = -\beta \sum_{i=1}^{m} \int\limits_{0}^{\ell_i} (z^i)' z^i dx. \tag{4.73}$$

We conclude $\beta = -\lambda$ and

$$d\lambda(G; V) = -\frac{1}{2} \sum_{i=1}^{m} \left(\frac{d}{dx} z^i \right)^2 (\ell_i) V^i(0, \ell_i) = -\frac{1}{2} \sum_{i=1}^{m} \left(\frac{d}{dx} z^i \right)^2 (\ell_i) \ell_i'(0). \tag{4.74}$$

We may now add a volume constraint

$$\sum_{i=1}^{m} \int\limits_{0}^{\ell_i} 1 dx = M$$

together with a new Lagrange multiplier μ and rewrite the Lagrangean form as

$$\mathscr{L}(G; z, \mu) := \lambda(G) + \mu \left(\sum_{i=1}^{m} \int\limits_{0}^{\ell_i} 1 dx - M \right). \tag{4.75}$$

The shape derivative finally amounts to

$$d\lambda(G; V) + \mu \sum_{i=1}^{m} V^i(0, \ell_i), \tag{4.76}$$

which gives the first-order optimalitiy condition

$$\sum_{i=1}^{m} \left(\mu - \frac{1}{2} \left(\frac{d}{dx} z^i \right)^2 (\ell_i) \right) V^i(0, \ell_i) = 0, \ \forall V^i. \tag{4.77}$$

This clearly shows, that the boundary conditions for $\frac{d}{dx} z^i(\ell_i)$ are independent of i. But, as $\sqrt{\lambda}\ell_i \neq k\pi$, we have $\frac{d}{dx} z^i(\ell_i) = -a \frac{\sqrt{\lambda}}{\sin(\sqrt{\lambda}\ell_i)}$. But this implies that the lengths are all equal: $\ell_i = \frac{M}{m}$. This result is analogous to the 2-D problem, where it is well-known that the optimal domain for the first eigenvalue is a ball.

4.2.5 Transmission Problem

We again consider the star-graph problem, however, for the sake of simplicity, with uniform lengths of the edges, i.e., $\ell_i = \ell, i \in \mathscr{I}$. We introduce $0 < r < \ell$ and a stiffness parameter $\delta > 0$. We consider uniform distributed forces $f^i = 1, i \in \mathscr{I}, \mathscr{I} = \{i = 1, \ldots, m\}$. Let $\chi^{\delta}_{(0,r)}$ be the function

$$\chi^{\delta}_{(0,r)}(x) := \begin{cases} \delta & x \in [0, r) \\ 1 & x \in [r, \ell]. \end{cases}$$

The transmission problem on the star graph then reads as follows.

$$\begin{aligned} -\chi^{\delta}_{(0,r)} \frac{d^2}{dx^2} y^i &= 1, \ x \in (0, \ell_i), \ i \in \mathscr{I} \\ y^i(r^-) &= y^i(r^+), \ i \in \mathscr{I} \\ \delta \frac{d}{dx} y^i(r^-) &= \frac{d}{dx} y^i(r^+), \ i \in \mathscr{I} \\ y^i(0) = y^j(0), \quad \sum_{i \in \mathscr{I}} & \frac{d}{dx} y^i(0) = 0 \\ y^i(\ell) &= 0, \ i \in \mathscr{I}. \end{aligned} \tag{4.78}$$

We consider the optimization problem

$$\min \mathscr{J}(G_\delta, y) := \frac{1}{2}\sum_{i=1}^{m}\int_0^r |y^i - y^i_{d1}|^2 dx + \frac{1}{2}\sum_{i=1}^{m}\int_r^\ell |y^i - y^i_{d2}|^2 dx$$

$$\text{such that } (y,\delta) \text{ satisfies (4.78).} \tag{4.79}$$

We will specify the target functions $y^i_{dk}, k = 1, 2$ in the example following the analysis. As of now, we assume $y^i_{d1}(r^-) = y^i_{d2}(r^+), i = 1, \dots, m$. We first consider the shape derivative $(y^i)'(x)$ of y^i at x.

$$\begin{aligned}
-\frac{d^2}{dx^2}(y^i)' &= 0,\ x \in (0,\ell_i),\ i \in \mathscr{I}\\
(y^i)'(r^-) &= (y^i)'(r^+),\ i \in \mathscr{I}\\
\delta\frac{d}{dx}(y^i)'(r^-) - \frac{d}{dx}(y^i)'(r^+) &= (\delta - 1)\frac{d}{dx}y^i(r^-)r'(0),\ i \in \mathscr{I}\\
y^i(0) = y^j(0),\quad \sum_{i\in\mathscr{I}}\frac{d}{dx}y^i(0) &= 0\\
y^i(\ell) &= 0,\ i \in \mathscr{I}.
\end{aligned} \tag{4.80}$$

In order to establish the gradient of the cost function, we note

$$d\mathscr{J}(r,y) = \sum_{i=1}^{m}\int_0^r (y^i - y^i_{d1})(y^i)'dx + \int_r^\ell (y^i - y^i_{d2})(y^i)'dx. \tag{4.81}$$

In order to establish the shape gradient, we can, of course, either solve (4.80), or, which is more common and most often more convenient, introduce the adjoint problem

$$\begin{aligned}
-\chi^\delta_{(0,r)}\frac{d^2}{dx^2}p^i &= y^i - y^i_{d1},\ x \in (0,r),\ i \in \mathscr{I}\\
-\frac{d^2}{dx^2}p^i &= y^i - y^i_{d2},\ x \in (r,\ell),\ i \in \mathscr{I}\\
\delta\frac{d}{dx}p^i(r^-) &= \frac{d}{dx}p^i(r^+),\ i \in \mathscr{I}\\
p^i(0) = p^j(0),\quad \sum_{i\in\mathscr{I}}\frac{d}{dx}p^i(0) &= 0\\
p^i(\ell) &= 0,\ i \in \mathscr{I}.
\end{aligned} \tag{4.82}$$

After some calculus, we arrive at

$$d\mathscr{J}(r,y) = (1-\delta)r'(0)\sum_{i=1}^{m}\frac{d}{dx}p^i(r^-)\frac{d}{dx}y^i(r^-). \tag{4.83}$$

Then, depending of the sign of the sum, $r'(0)$ needs to be positive or negative in order to decrease the cost function.

Example 4.2.7 We compute the solutions of (4.78). We obtain

$$y^i(x) = \begin{cases} \frac{1}{2}\left(\frac{1}{\delta}-1\right)r^2 + \frac{1}{2}\ell^2 - \frac{1}{2\delta}x^2 & x \in [0,r) \\ \frac{1}{2}(\ell^2 - x^2) & x \in [r,\ell]. \end{cases}$$

We now specify the target functions. In order to provide simple evidence, we introduce $0 < r_0 < r$ and take

$$y^i_{d1}(x) = \begin{cases} \frac{1}{2}\left(\frac{1}{\delta}-1\right)r^2 + \frac{1}{2}\ell^2 - \frac{1}{2\delta}x^2 & x \in [0,r_0) \\ \frac{1}{2}(\ell^2 - x^2) & x \in [r_0,r), \quad y^i_{d2} = \frac{1}{2}(\ell^2 - x^2), \; x \in [r,\ell]. \end{cases}$$

Clearly, in order to reduce the gradient to zero, r has to tend to r_0 and, therefore, $r'(0) < 0$. In order to derive this from (4.83), we calculate the adjoint explicitly. We obtain

$$p^i(x) = \tag{4.84}$$

$$-\frac{\delta-1}{2\delta^2}\begin{cases} \frac{1}{2}r^2r_0^2 - \frac{1}{4}r_0^4 - \frac{1}{4}r^4 + (r^2r_0 - \frac{1}{3}r_0^3 - \frac{2}{3}r^3)\ell & x \in [0,r_0) \\ (r^2r_0 - \frac{1}{3}r_0^3 - \frac{2}{3}r^3)(x-\ell) + \frac{2}{3}r^3x - \frac{1}{2}r^2x^2 + \frac{1}{12}x^4 - \frac{1}{4}r^4 & x \in [r_0,r) \\ (r^2r_0 - \frac{1}{3}r_0^3 - \frac{2}{3}r^3)(x-\ell) & x \in [r,\ell]. \end{cases} \tag{4.85}$$

This implies

$$d\mathscr{J}(r,y) = \frac{(1-\delta)^2}{2\delta^2}mr\left(\frac{2}{3}r^3 + \frac{1}{3}r_0^3 - r^2r_0\right)r'(0). \tag{4.86}$$

As the factor in front of $r'(0)$ is positive, in order to have decent, we have to take $r'(0) < 0$. Clearly, the gradient is zero for $y = r_0$.

4.2.6 *Topological Derivative with Respect to Material Properties*

We now consider the sensitivity of states with respect to the inclusions. That is to say, we introduce a different stiffness parameter close to the origin at the center node and establish a sensitivity result. This time we consider inhomogeneous Dirichlet conditions at the simple nodes.

$$\begin{aligned} -\frac{d^2}{dx^2}y^i &= 0,\ x \in (0,1),\ i \in \mathscr{I} \\ y^i(r^-) &= y^i(r^+),\ i \in \mathscr{I} \\ \delta\frac{d}{dx}y^i(r^-) &= \frac{d}{dx}y^i(r^+),\ i \in \mathscr{I} \\ y^i(0) = y^j(0),\ &\sum_{i\in\mathscr{I}}\frac{d}{dx}y^i(0) = 0 \\ y^i(1) &= g_i,\ i \in \mathscr{I}. \end{aligned} \tag{4.87}$$

We compute the following solution:

$$\begin{aligned} y^i(x) &= \\ &-\frac{1}{\delta(r-1)-r}\left(g_i - \frac{1}{m}\sum_{j=1}^m g_j\right)x + \frac{1}{m}\sum_{j=1}^m g_j,\ x \in [0,r) \\ &-\frac{1}{\delta(r-1)-r}\delta\left(g_i - \frac{1}{m}\sum_{j=1}^m g_j\right)x + \frac{1}{m}\sum_{j=1}^m g_j + g_i \\ &+\frac{\delta}{\delta(r-1)-r}\left(g_i - \frac{1}{m}\sum_{j=1}^m g_j\right),\ x \in [r,1]. \end{aligned} \tag{4.88}$$

With this, it is easy to calculate

$$\frac{1}{r}\left(y^i_r(x) - y^i(x)_0\right) = \frac{1}{\delta}\frac{1-\frac{1}{\delta}}{1-r-\frac{r}{\delta}}\left(g_i - \frac{1}{m}\sum_{j=1}^m g_j\right)x \tag{4.89}$$

and, therefore,

$$\frac{d}{dr}y^i_r(x)|_{r=0} - \frac{1-\delta}{\delta}\left(g_i - \frac{1}{m}\sum_{j=1}^m g_j\right)x. \tag{4.90}$$

4.3 Stars with a Hole

We consider a star graph G_{J^0} with m edges and center at the node v_{J^0}. In particular, we may without loss of generality, assume that the edges e_i stretch from the center to the simple boundary nodes, which we will label from 1 to m. By this assumption, we consider the multiple node at the center as being reached at $x = 0$ for all outgoing edges. Thus, the data u_i are picked up at the ends $x = \ell_i$. As a slight conceptual variation with respect to previous discussions on differential equations on networks, we now consider vectorial states on the graph and, therefore, collect the stiffness information in a symmetric, positive definite constant matrix K_i. By this, the reader should be able to redo the problems discussed in the previous sections in the context of vectorial systems that are more relevant in the applications. Indeed, if one considers elastic strings stretched from node to node in space, the resulting system becomes more relevant and gives rise to still open questions, in particular for nonlinear systems. See [14, 47, 53] and for nonlinear gas transport on networks [24, 27, 44] for further reading. The systems then reads as follows. The material of the next two sections is taken in part from [43, 46].

$$\begin{cases} -K_i \dfrac{d^2}{dx^2} y^i + c_i y^i = f^i, \; i \in \mathscr{I} \\ y^i(\ell_i) = u_i, \; i = 1, \dots, m \\ y^i(0) = y^j(0), \; \forall i, j = 1, \dots, m \\ \displaystyle\sum_{i=1}^m \frac{d}{dx} y^i(0) = 0. \end{cases} \tag{4.91}$$

We are going to cut out the center and connect the corresponding cut nodes via a circuit as seen in Fig. 4.5. In general, we have numbers $\rho_i \in [0, \ell_i), \; i = 1, \dots, m$ which are taken to be the lengths of the edges that are cut out. Thus the remaining edges have lengths $\ell_i - \rho_i$. At $x = \rho_i$ we create a new multiple node v_i. We connect these nodes by edges $e_{m+i}, \; i = 1, \dots, m$ with lengths $\sigma^i(\rho_i)$. After that, these nodes receive a new edge degree. In this section, we assume that all these nodes have the same edge degree of $d_i = 3$. More complicated cutting procedures can be introduced, but obscure the ideas of this section on topological derivatives of graph problems.

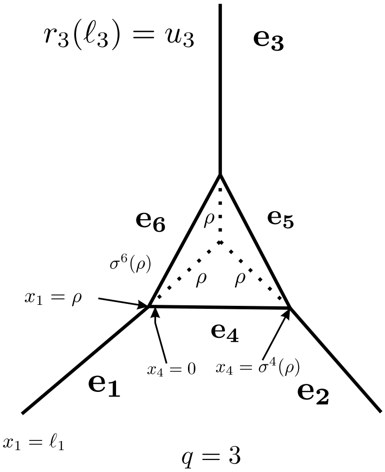


Fig. 4.5 Cutting a hole into star-like subgraph

The problem we have to solve is the following:

$$\left\{\begin{array}{c} -K_i \frac{d^2}{dx^2} y^i + c_i y^i = f^i,\ i \in \mathscr{I} \\ y^i(\ell_i) = u_i,\ \imath = 1, \dots, m \\ y^i(\rho_i) = y^{m+i}(0) = y^{m+i-1}(\sigma^i(\rho_i)),\ \forall i = 2, \dots, m \\ y^1(\rho_1) = y^{m+1}(0) = y^{2m}(\sigma^{2m}(\rho_{2m})), \\ -K_i \frac{d}{dx} y^i(\rho_i) - K_{m+i} \frac{d}{dx} y^{m+i}(0) + K_{m+i-1} \frac{d}{dx} y^{m+i-1}(\sigma^{m+i-1}(\rho_{m+i-1})) = 0,\ i = 2, \dots, m \\ -K_1 \frac{d}{dx} y^1(\rho_1) - K_{m+1} \frac{d}{dx} y^{m+1}(0) + K_{2m} \frac{d}{dx} y^{2m}(\sigma^{2m}(\rho_{2m})) = 0. \end{array}\right. \tag{4.92}$$

4.3.1 Homogeneous Networks

In this subsection, we consider the network under the assumption that all material and geometrical quantities are the same, and a symmetric hole. In fact, for the sake of brevity, we directly move to a tripod. See Fig. 4.5. Here we can solve the resulting system analytically and obtain the coefficients:

$$\begin{aligned} a_i^\rho = & \tfrac{1}{\sinh(\ell)} \left(u_i - \tfrac{1}{3} \sum_{j=1}^{3} u_j \right) \\ & + \rho \tfrac{1}{\cosh(\ell)} \left\{ \left(1 - \tfrac{1}{3}\sigma\right) \coth(\ell)^2 \left(u_i - \tfrac{1}{3} \sum_{j=1}^{3} u_j \right) \right. \\ & \left. + (\sigma - 1) \tfrac{1}{3} \sum_{j=1}^{3} u_j \right\} + O(\rho^2), \end{aligned} \tag{4.93}$$

$$\begin{aligned} b_i^\rho = & \tfrac{1}{\cosh(\ell)} \tfrac{1}{3} \sum_{j=1}^{3} u_j \\ & - \rho \tfrac{\sinh(\ell)}{\cosh(\ell)^2} \left\{ \left((1 - \tfrac{1}{3}\sigma) \coth(\ell)^2 \right) \left(u_i - \tfrac{1}{3} \sum_{j=1}^{3} u_j \right) \right. \\ & \left. + (\sigma - 1) \tfrac{1}{3} \sum_{i=1}^{3} u_i \right\} + O(\rho^2), \end{aligned} \tag{4.94}$$

where $i = 1, 2, 3$.

It is apparent that (4.93) and (4.94) provide the second-order asymptotic expansion we were looking for. We consider the following experiment: we apply longitudinal forces $u_i = ue_i$ with the same magnitude at the simple nodes of the network. The (outer) edges $e_i,\ 1 = 1, 2, 3$ or, respectively the edges of the original star, are given by

$$e_1 = (0, 1),\ e_2 = \left(-\frac{\sqrt{3}}{2}, -\frac{1}{2} \right),\ e_3 = \left(\frac{\sqrt{3}}{2}, -\frac{1}{2} \right)$$

which together with the orthogonal complements

$$e_1^\perp = (-1,0),\ e_2^\perp = \left(\frac{1}{2}, -\frac{\sqrt{3}}{2}\right),\ e_3^\perp = \left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right)$$

form the local coordinate systems of the edges. Obviously $\sum\limits_{i=1}^{3} e_i = 0$. Thus the solution to the unperturbed problem is given by

$$(y^i)(x) = \frac{1}{\sinh(\ell)} u \sinh(x) e_i. \tag{4.95}$$

This is in agreement with the fact that that particular reference configuration is completely symmetric. Now, the solution $(y^i)_\rho$ to the perturbed system and $(\frac{d}{dx}(y^i))_\rho(\ell)$ are then given by

$$\begin{aligned} (y^i)_\rho(x) &= \frac{1}{\sinh(\ell)} \sinh(x) u e_i \\ &\qquad + \rho\left(1 - \frac{\sigma}{3}\right) \frac{1}{\sinh(\ell)} \left(\coth(\ell)\sinh(x) - \cosh(x)\right) u e_i \ + O(\rho^2) \\ \frac{d}{dx}((y^i))_\rho(\ell) &= \coth(\ell) u e_i + \rho(\coth(\ell)^2 - 1)\left(1 - \frac{\sigma}{3}\right) u e_i \ + O(\rho^2) \qquad . \end{aligned} \tag{4.96}$$

The energy of the unperturbed system is given by

$$\mathscr{E}_0 = \frac{1}{2}\sum_{i=1}^{3}\int_0^{\ell} \frac{d}{dx}(y^i)\cdot\frac{d}{dx}(y^i) + (y^i)\cdot(y^i) dx = \frac{3}{2}\coth(\ell) u^2, \tag{4.97}$$

while the energy of the perturbed system is given by

$$\mathscr{E}^\rho = \frac{1}{2}\sum_{i=1}^{3}\int_0^{\ell-\rho}\left[\frac{d}{dx}(y^i)\cdot\frac{d}{dx}(y^i) + (y^i)\cdot(y^i)\right]dx + \frac{1}{2}\sum_{i=4}^{6}\int_0^{\sigma\rho}\left[\frac{d}{dx}(y^i)\cdot\frac{d}{dx}(y^i) + (y^i)\cdot(y^i)\right]dx \tag{4.98}$$

$$= \frac{1}{2}\langle S^\rho u, u\rangle = \frac{1}{2}\langle S^0 u, u\rangle + \rho\frac{1}{2}\left(1 - \frac{\sigma}{3}\right)\left\{((\coth(\ell))^2 - 1)\right\} u^2 \tag{4.99}$$

$$= \frac{1}{2}\langle S^0 u, u\rangle + \rho\frac{\sqrt{3}}{2}(\sqrt{3} - 1)\sinh(\ell)^{-2} u^2. \tag{4.100}$$

From these experiments we may draw the conclusion that nodes of edge degree 3 under symmetric load, where the configuration is at $120A^0$ between the edges (this amounts to $\sigma = \sqrt{3}$) are not going to be replaced by a hole, which would, in turn result in three new multiple nodes of edge degree 3. This seems to support the optimality of such graphs being observed by Buttazzo [4].

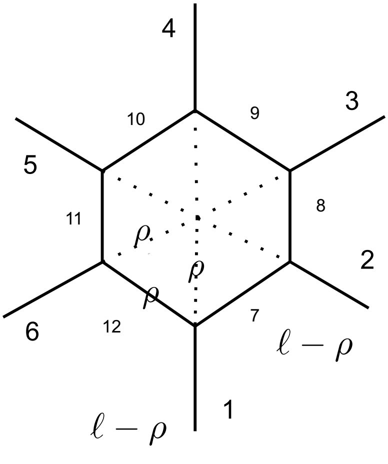


Fig. 4.6 Graph with "critical" edge degree 6

Remark 4 The completely analogous formulae are obtained in the scalar case $((y^i)(x) \in \mathbf{R})$, relevant, for instance, in problems of heat transfer or electrical currents in networks.

If the loads are not symmetric and/or if the geometry of the "hole" is not uniform, the energy may in fact drop. A more detailed analysis can be found in [43]. Suffice it to say here that nodes with higher edge degree, according to our analysis, are "more likely" to be released by a hole, as even in the symmetric case the number $\sigma(\rho)$ which measures the new edge lengths, will be less than 1.

This is true, for example, for a node with edge degree 6 and beyond. Thus, the total length of the new edges is smaller than the total length of the removed edges. This, in turn, is intuitive with respect to the fact that in the higher dimensional problem (in 2- or 3-D, no graphs), digging a hole reduces the amount of mass.

We now consider the homogeneous situation for a star with edge degree 6 at the multiple node. In this case $\sigma = 1$ for the symmetric situation. See Fig. 4.6.

We calculate

$$\begin{aligned} a_1^\rho = &\frac{1}{\sinh(\ell)}\left(u_1 - \frac{1}{6}\sum_{j=1}^{6} u_j\right) \\ &+\rho\frac{\cosh(\ell)}{\cosh^2(\ell)-1}\Bigg\{(-u_5 - u_3 - 4u_2 - 4u_6 + 10u_1) \\ &-7(u_1 - \frac{1}{6}\sum_{j=1}^{6} u_j)\Bigg\}. \end{aligned} \tag{4.101}$$

Notice that the edges 2 and 6 are the "neighboring" edges of edge 1 in the original star graph. The other coefficients a_i^ρ, $1 = 2, \ldots, 6$ are then obvious. For the sake of brevity, we only display, for example, a_{12}^ρ:

$$\begin{aligned} a_{12}^\rho = &\frac{1}{12\sinh(\ell)}[5(u_1-u_6)+3(u_2-u_5)+(u_3-u_4)] \\ &-\rho\frac{\cosh(\ell)}{144(\cosh^2(\ell)-1)}[25(u_1-u_6)-9(u_2-u_5)-7(u_3-u_4)] \\ &\qquad\qquad +O(\rho^2). \end{aligned} \tag{4.102}$$

Again, observe that edge 12, in terms of the edges of the original graph, has direct neighbors 1 and 6, the next level is 2 and 5, and finally we have 3 and 4. One realizes a consequent scaling. Also note that $a_i^\rho = 0$ if u_i are all equal. This shows that the coefficients b_i^ρ, in that case, are independent of ρ and thus the energy will not change for this limiting case.

4.4 The Topological Derivative

We are now in the position to define the topological derivative of an elliptic problem on a graph.

Definition 1 Let G be a graph, and let $v_J \in \mathscr{J}_M$ be a multiple node with edge degree d_J. Let G_ρ be the graph obtained from G by replacing v_J with a cycle of length $\sum\limits_{i=1}^{d_J} c_i\rho$ with vertices $v_J^1, \dots v_J^{d_J}$ of edge degree 3 each, such that the distance from v_J to v_J^i is equal to ρ. Thus, the number n^ρ of edges of G_ρ is $n + d_J$. Let $\mathscr{J} : G \to \mathbf{R}$ be a functional on the edges of G

$$J\left(y, \frac{d}{dx}y, G\right) := \sum_{i=1}^{n}\int_0^{\ell_i} F\left(x, y^i, \frac{d}{dx}y^i\right) \tag{4.103}$$

and let

$$J\left(y_\rho, \frac{d}{dx}y_\rho, G_\rho\right) := \sum_{i=1}^{n+d_J}\int_0^{\ell_i^\rho} F\left(x, y_{\rho_i}, \frac{d}{dx}y_{\rho_i}\right) \tag{4.104}$$

be its extension to $G\rho$. Assume we have an asymptotic expansion as follows

$$J\left(y_\rho, \frac{d}{dx}y_\rho, G_\rho\right) = J\left(r, \frac{d}{dx}y, G\right) + \rho\mathscr{T}(v_J) + O(\rho^2), \tag{4.105}$$

then we define the topological gradient of $J(G_\rho)$ with respect to ρ for $\rho = 0$ at the vertex v_J as follows.

$$\mathscr{T}(v_J) = \lim_{\rho\to 0}\frac{J\left(y_\rho, \frac{d}{dx}y_\rho, G_\rho\right) - J\left(r, \frac{d}{dx}y, G\right)}{\rho}. \tag{4.106}$$

We consider the energy functional or, equivalently, the compliance which is the most natural criterion to begin with. There are five such functionals relevant for the analysis of this section: $E^0(y)$ on the entire graph G, $E^\rho(y_\rho)$ on the entire graph with the hole G^ρ, $E_{CS}(y)$ on the graph $G \setminus \mathscr{S}^{J^0}$, where the star graph without hole $\mathscr{S}^{J^0}$ has been cut out along edges $e_i,\ i \in \mathscr{I}_{J^0}$, $E_S^0(y; v)$ on the star graph without hole, and $E_S^\rho(y; v)$ on the star graph with hole. Obviously

$$E_S^0(y; u) = \frac{1}{2}\langle S^0 u, u\rangle, \tag{4.107}$$

$$E_S^\rho(y; u) = \frac{1}{2}\langle S^\rho u, u\rangle, \tag{4.108}$$

$$E^0(y) = E_{CS}(y) + E_S^0(y, y), \quad E^\rho(y_\rho^0) = E_{CS}(y_\rho) + E_S^\rho(y_\rho, y_\rho), \tag{4.109}$$

where it is understood that in $E_S^\rho(y_\rho, \cdot)$ and $E_S^0(y, \cdot)$, we insert $u_i = y_\rho(\ell_i)$ and $u_i = y_0(\ell_i)$, respectively. Thus

$$E^\rho(y_\rho) - E^0(y) = \frac{1}{2}\langle S^\rho(\tilde{y}), \tilde{y}\rangle - \frac{1}{2}\langle S^0(\tilde{y}), \tilde{y}\rangle, \tag{4.110}$$

where $\tilde{y}$ solves the problem on $G \setminus \mathscr{S}^{J^0}$ and $u_i = \tilde{y}_i(\ell_i), i \in \mathscr{I}_{J^0}$. Thus the asymptotic analysis of the last section carries over to the entire graph. As we have done the complete asymptotic analysis up to order 2 in the homogeneous case only, we consequently dwell on this case now; the more general case will be subject of a forthcoming publication.

4.4.1 Homogeneous Graphs

In order to find an expression of the topological gradient in terms of the solutions y at the node v_{J^0}, the one that is cut out, we need to express the solution in terms of the data u_i.

Example 4.4.1 We consider the star graph as above with three edges. Obviously

$$u_i - \frac{1}{3}\sum_{j=1}^{3} u_j = \sinh(\ell)\frac{d}{dx}y_i(0), \ \frac{1}{3}\sum_{j=1}^{3} u_j = \cosh(\ell)y^i(0). \tag{4.111}$$

Thus using the fact that $\sum\limits_{i=1}^{3} \|u_i - \frac{1}{3}\sum\limits_{j=1}^{3} u_j\|^2 = \sum\limits_{i=1}^{3} \|u_i\|^2 - \frac{1}{3}(\|\sum\limits_{i=1}^{3} u_j\|)^2$, we can express the bilinear expression $\langle \mathscr{S}^\rho(u), u\rangle$ in terms of $\|y_0(0)\|^2$ and $\|(\frac{d}{dx}y_0(0)\|^2$

(where we omit the index 0) as follows

$$\begin{aligned}\langle \mathscr{S}_i^\rho(u), u\rangle &= \langle \mathscr{S}_i^0(u), u\rangle \\ &+\rho\left\{(1-\tfrac{1}{3}\sigma)\sum_{i=1}^{3}\|\tfrac{d}{dx}y^i(0)\|^2 + (\sigma-1)\sum_{i=1}^{3}\|y^i(0)\|^2\right\}.\end{aligned} \tag{4.112}$$

This says that for energy function in the homogeneous case, when cutting out a symmetric hole, e.g., $\sigma^i = \sigma = \sqrt{3},\ i = 1, 2, 3$, we have

$$\mathscr{T}_E(y, v_{J^0}) = \frac{1}{2}\left\{\left(1-\frac{1}{3}\sigma\right)\sum_{i=1}^{3}\left\|\frac{d}{dx}y^i(0)\right\|^2 + (\sigma-1)\sum_{i=1}^{3}\|y^i(0)\|^2\right\}. \tag{4.113}$$

The situation will be different for such vertices having a higher edge degree as 6, and those having nonsymmetric holes. We expect that such networks are more likely to be reduced to edge degree 3 by tearing a hole. But this has to be confirmed by more detailed studies.

4.4.2 Sensitivity with Respect to Edge Inclusion

We now consider a different situation where a node with edge degree $d_J = N$ is released into a node of edge degree 3 and one of degree $N-1$ by introduction of a new edge e_{N+1}. See Fig. 4.7.

We consider this procedure in an explicit example with edge degree 4.

Let, therefore, v_J be a node with edge degree 4. As visualized in Fig. 4.7, we will introduce an additional new edge e_5^ρ of length $\rho > 0$ which together with the two new edges e_1^ρ, e_2^ρ is given by

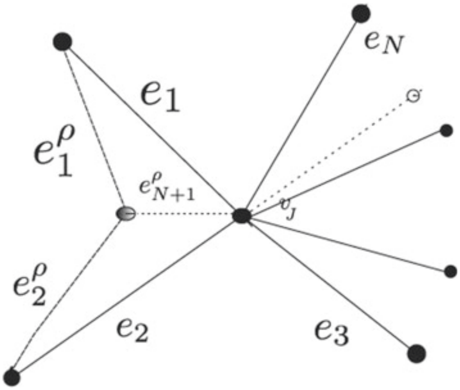


Fig. 4.7 N-node turns into 3-node plus (N+1)-node

$$
\begin{aligned}
e_1^\rho &:= \frac{\ell_1 e_1 - \rho e_{N+1}}{\|\ell_1 e_1 - \rho e_{N+1}\|}\\
e_2^\rho &:= \frac{\ell_2 e_1 - \rho e_{N+1}}{\|\ell_2 e_2 - \rho e_{N+1}\|},
\end{aligned}
\tag{4.114}
$$

where in our case study below $N = 4$.

The new lengths $\ell - \sigma$ of the edges e_1^ρ, e_2^ρ (we consider a symmetric situation where the new additional edge e_{N+1} equally divides the angle between e_1, e_2 with an inclination α toward the corresponding unit vectors) can be computed by elementary trigonometry. The number σ is then found to be

$$
\sigma = \rho\cos\alpha - \rho^2 \frac{1}{2\ell}\left(1 - \frac{1}{\ell}\cos^2\alpha\right) + O(\rho^3). \tag{4.115}
$$

It is interesting to notice that for $\cos\alpha > \frac{1}{2}$, the new graph has actually a smaller total length. This is in contrast to the standard situation where cutting out a hole—which in fact implies introducing the new edges forming that hole—has the opposite effect. For the sake of simplicity, we calculate the sensitivities with respect to introducing the new edge of length ρ for the Laplacian on the graph only. Thus, we do not consider an extra stiffening part due to the presence of a term cy^i.

$$
\begin{cases}
-\dfrac{d^2}{dx^2}y^i = 0 \quad \text{in } I_i,\ i = 1, \dots, 5\\
y^i(\ell) = u_i, i = 1, \dots 4,\\
y^1(\sigma) = y^2(\sigma) = y^5(\rho),\\
\dfrac{d}{dx}y^1(\sigma) + \dfrac{d}{dx}y^2(\sigma) - \dfrac{d}{dx}y^5(\rho) = 0,\\
y^3(0) = y^4(0) = y^5(0),\\
\dfrac{d}{dx}y^3(0) + \dfrac{d}{dx}y^4(0) + \dfrac{d}{dx}y^5(0) = 0
\end{cases}
\tag{4.116}
$$

We perform a similar analysis as in Sect. 4.3 and therefore omit the details. We obtain

$$
\begin{aligned}
y_\rho^1(x) = {} & \frac{1}{\ell}\left(u_1 - \frac{1}{4}\sum_{i=1}^4 u_i\right)x + \frac{1}{4}\sum_{i=1}^4 u_i\\
& -\frac{\rho}{2\ell^2}\left\{\left[\frac{1}{2}\cos\alpha + 1\right](u_2 - u_1) + (2 - \cos\alpha)\left(u_1 - \frac{1}{4}\sum_{i=1}^4 u_i\right)\right\}(x - \ell)
\end{aligned}
$$

$$= y_0^1(x)$$
$$-\frac{\rho}{2\ell^2}\left\{\left[\frac{1}{2}\cos\alpha + 1\right](u_2 - u_1) + (2 - \cos\alpha)\left(u_1 - \frac{1}{4}\sum_{i=1}^{4} u_i\right)\right\}(x - \ell)$$

$$y_\rho^2(x) = y_0^2(x)$$
$$-\frac{\rho}{2\ell^2}\left\{\left[\frac{1}{2}\cos\alpha + 1\right](u_1 - u_2) + (2 - \cos\alpha\left(u_2 - \frac{1}{4}\sum_{i=1}^{4} u_i\right)\right\}(x - \ell)$$

$$y_\rho^3(x) = r_3^0(x)$$
$$-\frac{\rho}{2\ell^2}\left\{\left[1 - \frac{1}{2}\cos\alpha\right](u_2 - u_1) + (2 - \cos\alpha\left(u_2 - \frac{1}{4}\sum_{i=1}^{4} u_i\right)\right\}(x - \ell)$$

$$y_\rho^4(x) = r_4^0(x)$$
$$-\frac{\rho}{2\ell^2}\left\{\left[1 - \frac{1}{2}\cos\alpha\right](u_1 - u_2) + (2 - \cos\alpha)\left(u_1 - \frac{1}{4}\sum_{i=1}^{4} u_i\right)\right\}(x - \ell).$$

In order to calculate the energy, we use the Steklov–Poincaré mapping and multiply by $y^i(\ell)$.

As before, the calculations can be done for scalar problems as well as for vectorial in-plane models. We dispense with the display of the lengthy formulae. Instead, we give two different scenarios for topological derivatives.

Example 4.4.2 In the scalar case, we may set $u_1 = u_2$ and $u_3 = u_4 = 0$, i.e., we apply Dirichlet conditions at the ends of edges 3 and 4 and pull at the end of the edges 1 and 2 by the same amount. This results in

$$\langle S^\rho u, u\rangle = \langle S^0 u, u\rangle - \frac{\rho}{2\ell^2}(2 - \cos\alpha)u^2. \tag{4.117}$$

Obviously, the introduction of a new edge is enhanced. One obtains a decomposition into two multiple nodes with edge degree 3.

Example 4.4.3 In the second example, we take the planar model and set $u_1 = ue_1, u_2 = ue_2$ and again $u_3 = 0 = u_4$. Now we obtain

$$\langle S^\rho u, u\rangle = \langle S^0 u, u\rangle - \frac{3\rho}{4\ell^2}\left[\cos(\alpha)\left(\cos^2\alpha + \frac{2}{3}\cos\alpha - \frac{4}{3}\right)\right]u^2. \tag{4.118}$$

For small enough angles α (e.g., $0 < \alpha < \pi/6$) the expression with ρ, i.e., the topological derivative of the energy becomes negative. This shows that in the planar

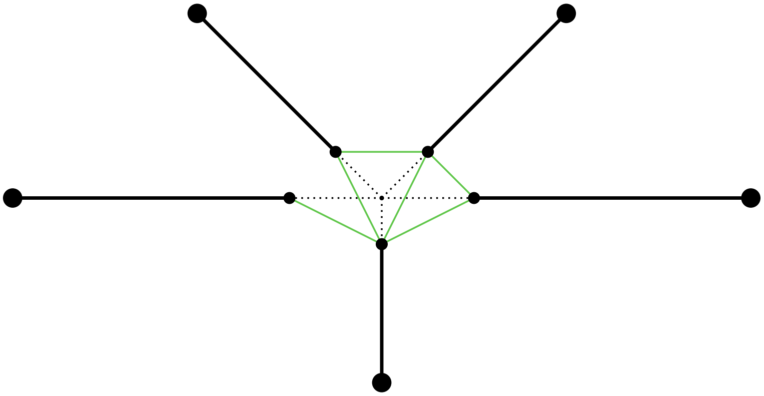


Fig. 4.8 Inserting the interconnections, valid example

situation, the opportunity to create an additional edge depends on the angles between the edges 1 and 2.

Obviously, the examples above can be generalized to more general networks including distributed loads and obstacles. It is also possible to extend this analysis to 3-D networks. As a final remark, one can extend the technique to networks of Timoshenko beams which is much more reasonable due to the stiffness of the resulting structure. See the thesis by Ogiermann [48]. We may then introduce "holes" as in the picture (Fig. 4.8).

References

1. S. Avdonin, P. Kurasov, M. Nowaczyk, Inverse problems for quantum trees II: recovering matching conditions for star graphs. Inverse Probl. Imaging **4**(4), 579–598 (2010)
2. S. Avdonin, G. Leugering, V. Mikhaylov, On an inverse problem for tree-like networks of elastic strings. ZAMM Z. Angew. Math. Mech. **90**(2), 136–150 (2010)
3. S. Avdonin, C. Rivero Abdon, G. Leugering, V. Mikhaylov, On the inverse problem of the two-velocity tree-like graph. ZAMM Z. Angew. Math. Mech. **95**(12), 1490–1500 (2015)
4. G. Buttazzo, B. Ruffini, B. Velichkov, Shape optimization problems for metric graphs. ESAIM Control Optim. Calc. Var. **20**(1), 1–22 (2014)
5. R. Dáger, E. Zuazua, *Wave Propagation, Observation and Control in* $1-d$ *Flexible Multi-structures*, vol. 50. Mathématiques & Applications (Berlin) [Mathematics & Applications] (Springer, Berlin, 2006)
6. C. D'Apice, S. Göttlich, M. Herty, B. Piccoli, *Modeling, Simulation, and Optimization of Supply Chains: A Continuous Approach* (Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 2010)
7. M. Dick, M. Gugat, Günter Leugering, Classical solutions and feedback stabilization for the gas flow in a sequence of pipes. Netw. Heterog. Media **5**(4), 691–709 (2010)
8. M. Dick, M. Gugat, G. Leugering, A strict H^1-Lyapunov function and feedback stabilization for the isothermal Euler equations with friction. Numer. Algebra Control Optim. **1**(2), 225–244 (2011)

9. P. Exner, Introduction to quantum graphs [book review of mr3013208]. Bull. Amer. Math. Soc. (N.S.) **51**(3), 511–514 (2014)
10. P. Exner, O. Turek, Spectrum of a dilated honeycomb network. Integr. Equ. Oper. Theory **81**(4), 535–557 (2015)
11. M. Garavello, K. Han, B. Piccoli, *Models for Vehicular Traffic on Networks*. AIMS Series on Applied Mathematics, vol. 9 (American Institute of Mathematical Sciences (AIMS), Springfield, 2016)
12. B. Geißler, O. Kolb, J. Lang, G. Leugering, A. Martin, A. Morsi, Mixed integer linear models for the optimization of dynamical transport networks. Math. Methods Oper. Res. **73**, 339–362 (2011)
13. M. Gröschel, A. Keimer, G. Leugering, Z. Wang, Regularity theory and adjoint-based optimality conditions for a nonlinear transport equation with nonlocal velocity. SIAM J. Control Optim. **52**(4), 2141–2163 (2014)
14. G. Qilong, G. Leugering, T. Li, Exact boundary controllability on a tree-like network of nonlinear planar Timoshenko beams. Chin. Ann. Math. Ser. B **38**(3), 711–740 (2017)
15. E.J.P. Georg Schmidt, On the modelling and exact controllability of networks of vibrating strings. SIAM J. Control Optim. **30**(1), 229–245 (1992)
16. M. Gugat, G. Leugering, Regularization of L^∞-optimal control problems for distributed parameter systems. Comput. Optim. Appl. **22**(2), 151–192 (2002)
17. M. Gugat, G. Leugering, Global boundary controllability of the de St. Venant equations between steady states. Ann. Inst. H. Poincaré Anal. Non Linéaire **20**(1), 1–11 (2003)
18. M. Gugat, G. Leugering, Global boundary controllability of the Saint-Venant system for sloped canals with friction. Ann. Inst. H. Poincaré Anal. Non Linéaire **26**(1), 257–270 (2009)
19. M. Gugat, M. Dick, G. Leugering, Gas flow in fan-shaped networks: classical solutions and feedback stabilization. SIAM J. Control Optim. **49**(5), 2101–2117 (2011)
20. M. Gugat, F.M. Hante, M. Hirsch-Dick, G. Leugering, Stationary states in gas networks. Netw. Heterog. Media **10**(2), 295–320 (2015)
21. M. Gugat, A. Keimer, G. Leugering, Z. Wang, Analysis of a system of nonlocal conservation laws for multi-commodity flow on networks. Netw. Heterog. Media **10**(4), 749–785 (2015)
22. M. Gugat, G. Leugering, Solutions of L^p-norm-minimal control problems for the wave equation. Comput. Appl. Math. **21**(1), 227–244 (2002). Special issue in memory of Jacques-Louis Lions
23. M. Gugat, G. Leugering, L^∞-norm minimal control of the wave equation: on the weakness of the bang-bang principle. ESAIM Control Optim. Calc. Var. **14**(2), 254–283 (2008)
24. M. Gugat, G. Leugering, E.J.P. Georg Schmidt, Global controllability between steady supercritical flows in channel networks. Math. Methods Appl. Sci. **27**(7), 781–802 (2004)
25. M. Gugat, G. Leugering, A. Martin, M. Schmidt, M. Sirvent, D. Wintergerst, MIP-based instantaneous control of mixed-integer PDE-constrained gas transport problems. Comput. Optim. Appl. **70**(1), 267–294 (2018)
26. M. Gugat, G. Leugering, A. Martin, M. Schmidt, M. Sirvent, D. Wintergerst, Towards simulation based mixed-integer optimization with differential equations. Networks **72**(1), 60–83 (2018)
27. M. Gugat, G. Leugering, K. Schittkowski, E.J.P. Georg Schmidt, Modelling, stabilization, and control of flow in networks of open channels, in *Online Optimization of Large Scale Systems* (Springer, Berlin, 2001), pp. 251–270
28. M. Gugat, G. Leugering, S. Tamasoiu, K. Wang, H^2-stabilization of the isothermal Euler equations: a Lyapunov function approach. Chin. Ann. Math. Ser. B **33**(4), 479–500 (2012)
29. F.M. Hante, G. Leugering, A. Martin, L. Schewe, M. Schmidt, Challenges in optimal control problems for gas and fluid flow in networks of pipes and canals: from modeling to industrial application, in *Industrial Mathematics and Complex Systems*, Ind. Appl. Math. (Springer, Singapore, 2017), pp. 77–122
30. M. Kac, Can one hear the shape of a drum? Amer. Math. Monthly **73**(4, part II), 1–23 (1966)
31. A. Keimer, G. Leugering, T. Sarkar, Analysis of a system of nonlocal balance laws with weighted work in progress. J. Hyperbolic Differ. Equ. **15**(3), 375–406 (2018)

32. A. Keimer, L. Pflug, Existence, uniqueness and regularity results on nonlocal balance laws. J. Differ. Equ. **263**(7), 4023–4069 (2017)
33. A. Keimer, L. Pflug, M. Spinola, Existence, uniqueness and regularity of multi-dimensional nonlocal balance laws with damping. J. Math. Anal. Appl. **466**(1), 18–55 (2018)
34. P. Kuchment, Graph models for waves in thin structures. Waves Random Media **12**(4), R1–R24 (2002)
35. P. Kuchment, Quantum graphs. I. Some basic structures. Waves Random Media **14**(1), S107–S128 (2004). Special section on quantum graphs
36. P. Kuchment, Quantum graphs. II. Some spectral properties of quantum and combinatorial graphs. J. Phys. A **38**(22), 4887–4900 (2005)
37. J.E. Lagnese, G. Leugering, E.J.P.G. Schmidt, Modelling and controllability of networks of thin beams, in *System Modelling and Optimization (Zurich, 1991)*. Lecture Notes in Control and Information Sciences, vol. 180 (Springer, Berlin, 1992), pp. 467–480
38. J.E. Lagnese, G. Leugering, E.J.P.G. Schmidt, Control of planar networks of Timoshenko beams. SIAM J. Control Optim. **31**(3), 780–811 (1993)
39. J.E. Lagnese, G. Leugering, E.J.P.G. Schmidt, Modelling of dynamic networks of thin thermoelastic beams. Math. Methods Appl. Sci. **16**(5), 327–358 (1993)
40. J.E. Lagnese, G. Leugering, E.J.P.G. Schmidt, On the analysis and control of hyperbolic systems associated with vibrating networks. Proc. Roy. Soc. Edinburgh Sect. A **124**(1), 77–104 (1994)
41. J.E. Lagnese, G. Leugering, E.J.P.G. Schmidt, *Modeling, Analysis and Control of Dynamic Elastic Multi-link Structures*. Systems & Control: Foundations & Applications (Birkhäuser Boston, Inc., Boston, 1994)
42. J.E. Lagnese, G. Leugering, *Domain Decomposition Methods in Optimal Control of Partial Differential Equations*, vol. 148. International Series of Numerical Mathematics (Birkhäuser Verlag, Basel, 2004)
43. G. Leugering, J. Sokolowski, Topological derivatives for networks of elastic strings. ZAMM Z. Angew. Math. Mech. **91**(12), 926–943 (2011)
44. G. Leugering, E.J.P. Georg Schmidt, On the modelling and stabilization of flows in networks of open canals. SIAM J. Control Optim. **41**(1), 164–180 (2002)
45. G. Leugering, A. Martin, M. Schmidt, M. Sirvent, Nonoverlapping domain decomposition for optimal control problems governed by semilinear models for gas flow in networks. Control Cybernet. **46**(3), 191–225 (2017)
46. G. Leugering, J. Sokolowski, Topological sensitivity analysis for elliptic problems on graphs. Control Cybernet. **37**(4), 971–997 (2008)
47. G.R. Leugering, E.J.P. Georg Schmidt, On exact controllability of networks of nonlinear elastic strings in 3-dimensional space. Chin. Ann. Math. Ser. B **33**(1), 33–60 (2012)
48. E. Ogiermann, Topological sensitivity analysis for networks of timoshenko beams. Ph.D. Thesis, FAU Department of Mathematics, 2015
49. M. Olivieri, D. Finco, On the inverse spectral problems for quantum graphs, in *Advances in Quantum Mechanics*. Springer INdAM Series, vol. 18 (Springer, Cham, 2017), pp. 267–281
50. J. von Below, A characteristic equation associated to an eigenvalue problem on c^2-networks. Linear Algebra Appl. **71**, 309–325 (1985)
51. J. von Below, Sturm-Liouville eigenvalue problems on networks. Math. Methods Appl. Sci. **10**(4), 383–395 (1988)
52. J. von Below, Can one hear the shape of a network? in *Partial Differential Equations on Multistructures (Luminy, 1999). Lecture Notes in Pure and Appl. Math.*, vol. 219. (Dekker, New York, 2001), pp. 19–36
53. Y. Wang, G. Leugering, T. Li, Exact boundary controllability for 1-D quasilinear wave equations with dynamical boundary conditions. Math. Methods Appl. Sci. **40**(10), 3808–3820 (2017)

Chapter 5
Topological Analysis of a Weighted Human Behaviour Model Coupled on a Street and Place Network in the Context of Urban Terrorist Attacks

D. Provitolo, R. Lozi and E. Tric

Abstract This article introduces a new model of weighted human behaviour in the context of urban terrorist attacks. In this context, one of the major challenges is to improve the protection of the population. In achieving this goal, it is important to better understand and anticipate both individual and collective human behaviour, and the dynamics of the displacements associated with these behaviours. Based on the recently published Panic-Control-Reflex (PCR) model, this new Coupled Weighted PCR model takes into account the role of spatial configurations on behavioural dynamics. It incorporates, via a bottleneck effect, the narrowness and the length of the streets, and thus the pressure and counter pressure of the crowd in dangerous and safe places. The numerical evacuation simulations highlight that, depending on their size, intermediate places or public squares modulate the dynamics and the speed of flow of the crowd as it evacuates to a safe place. This model features a user-friendly graphical representation, which allows planners to accurately decide where to organize host public events in a specific territorial context.

Keywords Human behaviour · Panic · Terrorism · Street and place network · Urban context · Mathematical modelling · Coupled weighted PCR model

D. Provitolo · E. Tric
Université Côte d'Azur, CNRS, Observatoire de la Côte d'Azur, IRD, Géoazur, Nice, France
e-mail: provitol@geoazur.unice.fr
URL: https://www.oca.eu/fr/rech-risques-geoazur

E. Tric
e-mail: emmanuel.tric@univ-cotedazur.fr
URL: https://www.oca.eu/fr/rech-risques-geoazur

R. Lozi (✉)
Université Côte d'Azur, CNRS, LJAD, Nice, France
e-mail: Rene.LOZI@univ-cotedazur.fr
URL: https://math.unice.fr/laboratoire/fiche&id=286

P. Manchanda et al. (eds.), *Mathematical Modelling, Optimization, Analytic and Numerical Solutions*, Industrial and Applied Mathematics,
https://doi.org/10.1007/978-981-15-0928-5_5

5.1 Introduction

In the context of disasters, and in order to better protect the population, one of the major challenges today is to better understand and anticipate both individual and collective human behaviour, and the dynamics of the displacements associated with these behaviours.

Indeed, the impact of a dangerous phenomenon is particularly determined by the behaviour of the affected population. These reactions allow anyone to ensure their own safety and that of their family.

This is especially true in the case of sudden and unpredictable events, such as terrorist attacks. These events require immediate reactions for self-protection and self-evacuation, before the arrival of emergency response services.

In spite of the long history of terrorism, there is currently no uniform definition of this word, because 'depending on the political configuration, this one will be terrorist for some; instead he will be hailed freedom fighters for others' [1]. However, it is possible to identify different forms of terrorist acts, such as attacks, kidnappings, sabotage, bioterrorism, or cyberterrorism. Moreover, terrorist acts can be committed at different levels by individuals, groups or states.

In this article, we focus specifically on the new forms of terrorism, such as the attacks perpetrated by sects or groups which are becoming more and more active. These attacks target places frequented by the public (bus and subway stations, airports, hotels, cafes, etc.), in order to spread panic among the population. The sarin gas attack on the Tokyo subway committed by the Aum sect (1995), the bomb attacks in bus and subway stations (Nigeria 2014; Belgium 2016), at airports (Brussels airport 2016; Atatürk airport 2016) or near the finish line of the Boston Marathon (2013) and acts perpetrated by Islamic organizations (Kashmir 2019), are a few examples of a very long list of terrorist attacks. All these events triggered mass movements of flight and collective panic.

Moreover, in order to cause as many casualties as possible and to complicate the response of the security forces, terrorists have developed a new *modus operandi* by conducting simultaneous attacks in urban areas. In this way, for some coordinated attacks, human losses have been massive, yet constrained in time and space. This was the case for 11 September 2001, in New York and near Washington D.C., where a series of four coordinated terrorist attacks claimed by the Al-Qaeda terrorist group killed nearly 3,000 victims and injured over 6 thousand. India also endured terrifying terrorist attacks in November 2008 when an Islamic terrorist organization based abroad carried out a series of 12 coordinated shooting and bombing attacks lasting 4 days across Mumbai.

This *modus operandi* also makes it possible to prolong the attacks to maximise media coverage and to increase the feeling of terror and panic, as was seen during the Paris and Saint-Denis (France) terrorist attacks of November 2015.

In response to tightened police controls to reduce bombing attacks, new types of ad hoc weapons are now employed, like lorries or cars simply driven directly through a crowd massed for a cultural event or simply strolling peacefully. These attacks are most often led by ‘lone wolves’, who are difficult to detect beforehand and equally difficult to locate because they are ‘nested’ among civilians. On 14 July 2016, 86 people who attended the Bastille Day fireworks on the ‘Promenade des Anglais’ in Nice, France, were killed; 458 more were wounded. Since this day the frequency of such kinds of terrorist attacks have grown rapidly in western countries: in Berlin, Germany, 19 December 2016; in London, England, 22 March 2017; in Stockholm, Sweden, 7 April 2017; and in Barcelona, Spain, 16 August 2017. It is not only cars that have been used in such attacks but also true weapons like military guns (Las Vegas, USA, 2 October 2017) or simple tools like knives (Marseille, France, 1 October 2017), hammers, etc.

Nowadays, in order to protect inhabitants, authorities have developed diversified risk reduction strategies. These are complementary, with some acting directly on the threat, and others targeting the vulnerability of potential victims. In the field of counter-terrorism, the most widespread strategies aim to fight against the threat. Thus, in order to avoid new attacks, actions aiming to dismantle terrorist networks are carried out by the police and intelligence services. Attention is therefore focused on dangerous groups, but it is difficult to detect and neutralize all terrorist threats, especially those perpetrated by such ‘lone wolves’.

Thus, prevention policies also focus on reducing the vulnerability of the population. It is then a question of hindering the terrorist action itself by deploying security measures as close as possible to the potential targets (festivals, concerts, college, train and metro stations, etc.). In many countries, like in India and France, more law enforcement officers and even the army have been deployed for such measures; however this can be done only for a limited period of time.

More recently, in France, other actions, such as the distribution of leaflets and information documents have been deployed, to increase awareness among the population about the responses that could save lives, for example, to flee wherever possible, barricade the entrance, hide behind a solid obstacle and turn off the phone. This effort of dissemination of the best practices in the face of an attack is already a step forward but it remains incomplete.

We must also look at the reality of human behaviour (i.e. what people actually do during a terrorist attack). Their reactions simultaneously depend on their own emotions, the culture of risk and the environmental context (a closed environment like an auditorium or theatre, compared to an open geographical environment, such as a public square, place or networks of streets). When such an attack happens, the topography of the area is very important. The dynamics of human reactions and the associated displacements are guided by the space and the alternatives that it offers, especially in terms of evacuation, flight and accessibility to refuge areas.

Consider, for example, the terrible Jallianwala Bagh massacre, which took place on 13 April 1919 in Amritsar, Punjab, when a crowd of non-violent protesters were fired upon by troops of the British Indian Army.

The Jallianwala Bagh is a public garden with an area of 28,000 m^2, walled on all sides with five entrances (Figs. 5.1 and 5.2). The largest entry point was blocked by a tank and the main exit was locked. The troops fired on the crowd, directing their bullets largely towards the few open gates through which people were trying to flee (Fig. 5.3).

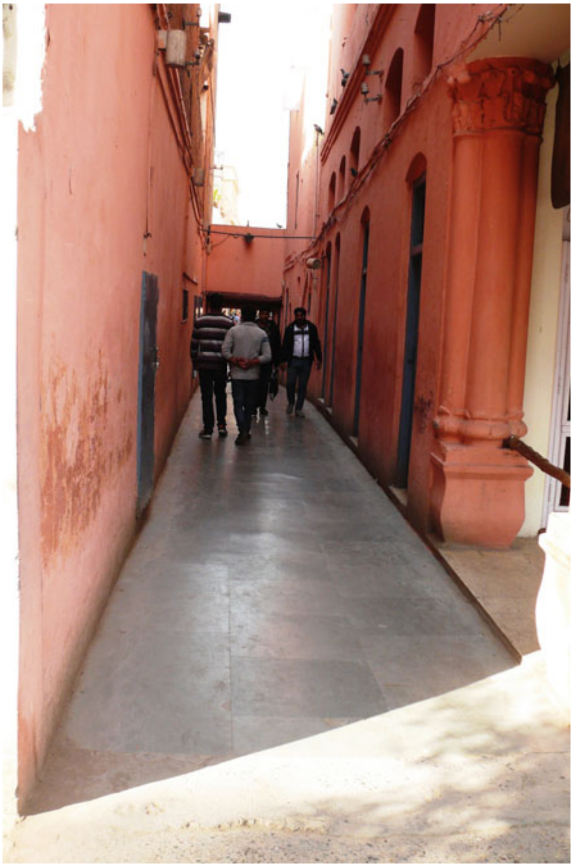


Fig. 5.1 Spatial configuration of Jallianwala Bagh (Amritsar, India) © R. Lozi, Feb. 2018

More recently, two terrorist attacks in Mediterranean cities, in the heart of the historic town centres of Barcelona (2017; Fig. 5.4) and Nice (2016), have shown that the population fled through the labyrinth of alleys to find refuge in urban places, such as public squares, esplanades or parks. Spatial configurations can therefore either increase or mitigate the vulnerability of populations.

In an urban context, it is also important to study the impact of the width of alleys and the shelter capacity of public squares or places so that the population can ensure its own self-evacuation (Fig. 5.4).

Of course, it is difficult to artificially reproduce a disaster, which would otherwise allow us to observe the diversity of human reactions that could occur, to follow the spatio-temporal dynamics and to analyse the impact of territorial configurations on these dynamics. To overcome these limits, it is possible to develop mathematical models from which evolution scenarios are simulated by varying parameters or initial conditions. The computer thus becomes the virtual laboratory and the simulation is understood to be an experiment on a model, a digital experience [2].

Fig. 5.2 Jallianwala Bagh Memorial (Amritsar, India) © D. Provitolo, Feb. 2018

The Com2SiCa research team[1] proposed the Panic-Control-Reflex (PCR) model [3], which is a model that simulates the possible human behaviour that can occur during sudden onset and unpredictable disasters, such as a tsunami, earthquake, or technological disaster. This model is formalized by a system of ordinary differential equations to describe behavioural dynamics over time [4–6].

In this article, we propose an extension of the PCR model in order to take into account the influence of spatial configuration in the mathematical modelling of the dynamics of human reactions in the face of traumatic situations, such as terrorist attacks. This dynamic and the associated displacements are indeed guided by the territorial configurations (networks of streets and places) and the alternatives that they offer in terms of evacuation, flight and accessibility to shelters. We call this extended model the Coupled Weighted PCR (CWPCR).

In Sect. 5.2, we will present the Panic-Control-Reflex model in its graphical and mathematical formalism, as published by the authors cited above. Then, in Sect. 5.3, this model will be improved in order to take into account the role of spatial configurations on behavioural dynamics. The CWPCR model incorporates the pressure and counter pressure of the crowd in each place via a bottleneck effect, which is induced by the narrowness and the length of the streets and the size of places. In Sect. 5.4,

[1] https://geoazur.oca.eu/en/research-geoazur/2158-com2sica-how-to-comprehend-and-simulate-human-behaviors-in-areas-facing-natural-disasters.

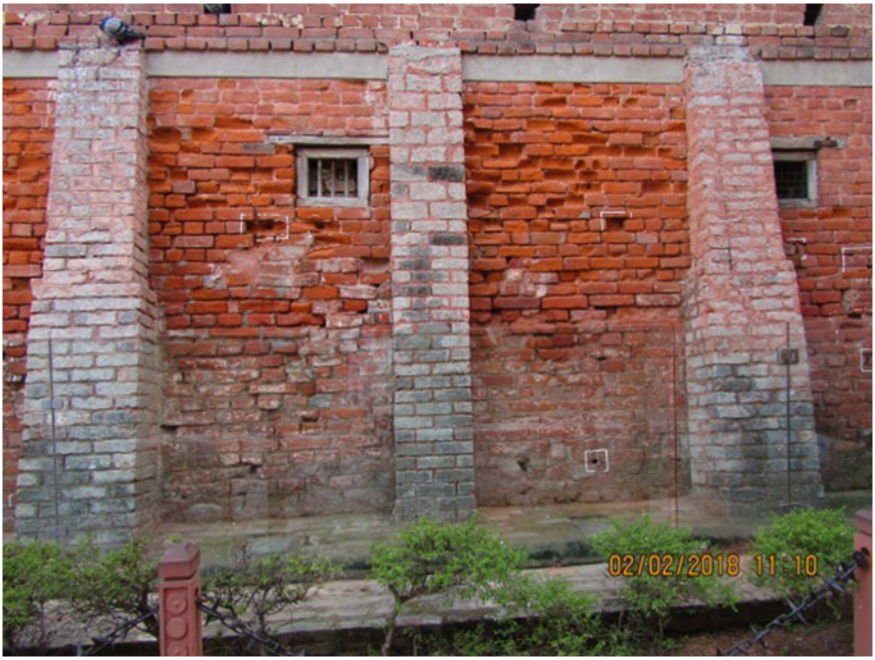


Fig. 5.3 Bullet marks can be easily seen on the wall (Amritsar, India) © D. Provitolo, Feb. 2018

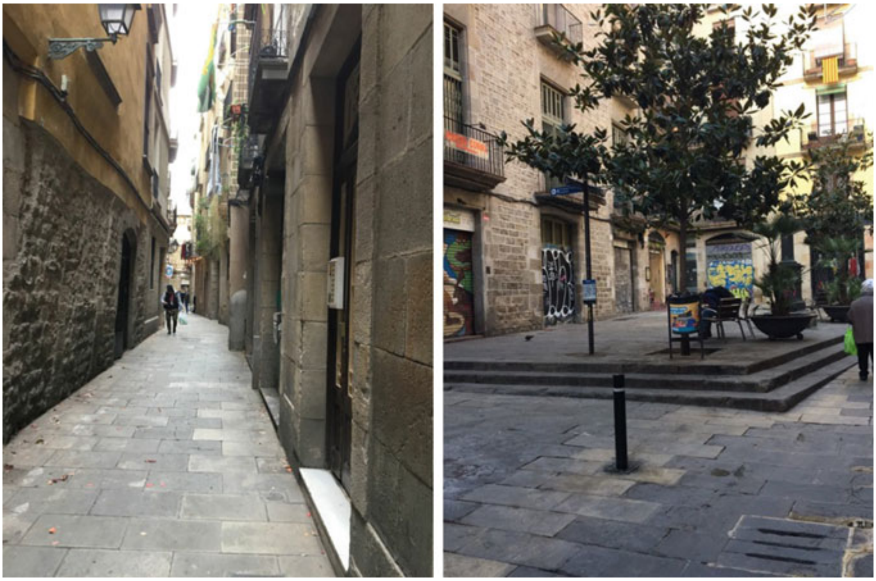


Fig. 5.4 Street and place in Barcelona (Spain) © D. Provitolo, Feb. 2018

we will consider an oriented network with three nodes representing three places or public squares of different sizes, linked by narrow streets. This part is therefore devoted to the analysis of the impact of the parameters (the size of the places, the width of the streets) on the evacuation of the population in the face of a terrorist attack, by means of numerical simulations. The numerical results highlight that, depending on their respective size, intermediate places modulate the dynamics and the speed of flow of the crowd. In this sense, they become strategic places both for the planners who must think about the organization of the area to host public events and festivals, and also for the terrorists who can use these strategic places to multiply the effect of their harmful actions by trapping the flight movements between two areas of action. This model is used with a user-friendly graphical representation, which allows planners to accurately consider where to organize host public events in a specific territorial context. Finally, in Sect. 5.5, a brief conclusion will be drawn.

5.2 The PCR System: An A-Spatial Model for Analysing the Dynamics of Human Behaviour During a Disaster

5.2.1 Neuroscientific Background of the PCR System

The PCR model [3–6] is a simulation model of the dynamics of collective human behaviour during a disaster (Fig. 5.5). It has been developed from the SIR-based models, which are compartmental models that are widely used in epidemiology [7]. In these models, the population can be decomposed into several subpopulations, each of which corresponds to a compartment. The PCR model focuses on the following:

i Different human behavioural states, namely daily behaviour before a disaster occurs, as well as reflex, panic and controlled behaviours that are observed during a disaster.

ii Transitional processes from one reaction to another. Indeed, neuroscience research shows that in disaster situations, humans are rarely stuck in one type of behaviour. The population switches between different behavioural states, some of which are the result of instinctive reactions [8], others of reasoned reactions [9].

To take these behavioural sequences into account, the PCR model formalizes human reactions as a chain of behaviours that appear in a certain order. It distinguishes $q(t)$, the daily behaviour before, and $b(t)$, after the disaster; $r(t)$ and $p(t)$, the uncontrolled emotional behaviours which are managed by the reptilian zone of the brain; and $c(t)$, the reasoned behaviours that are controlled by the prefrontal cortex [10–12]. This is represented in Fig. 5.5.

As the brain switches from one behavioural state to another, in the context of terrorist acts and therefore in a situation of sudden and unforeseen threat, the whole impacted population first adopts a behavioural reaction, Reflex $r(t)$, under the influence of surprise and the suddenness of the event, before transiting to the Panic reflex behaviour $p(t)$, or Controlled behaviour $c(t)$.

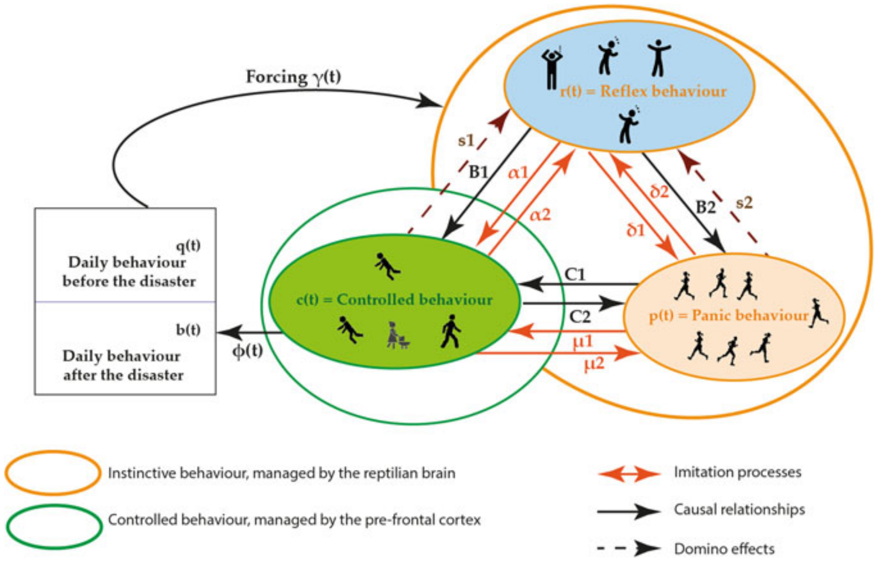

Forcing γ(t)
q(t)
Daily behaviour before the disaster
b(t)
Daily behaviour after the disaster
ϕ(t)
r(t) = Reflex behaviour
c(t) = Controlled behaviour
p(t) = Panic behaviour
s1
B1
α1
α2
δ2
δ1
B2
s2
C1
C2
μ1
μ2
Instinctive behaviour, managed by the reptilian brain
Controlled behaviour, managed by the pre-frontal cortex
Imitation processes
Causal relationships
Domino effects

Fig. 5.5 Graphical representation of the panic, control, reflex behaviour model (PCR) in the exceptional situation of disaster. *Source* from [3]

Reflex $r(t)$ and Panic $p(t)$ behaviour are instinctive, automatic reactions, allowing one to react extremely quickly to the threat, either by being stunned and paralysed $r(t)$, or by fleeing as quickly as possible due to the panic fear $p(t)$. In the context of a dense crowd, the context sought by terrorist groups, panic escape behaviour can worsen the vulnerability of the population because of the risk of crushing and suffocation [13].

Controlled behaviour $c(t)$ concerns reasoned and self-control reactions. They can take different forms during a catastrophe, for example, in the form of evacuation, leak, containment, sheltering, search for help, mutual aid or, on the contrary, looting, etc. Despite their diversity, the PCR model aggregates all of these controlled behaviours.

During the event, the switches from one behavioural state to another are caused by transitional dynamics due to

i Causal relationships (B_1, B_2, C_1, C_2). Once the population is in the reflex behaviour state, a part of it can evolve towards controlled behaviours at the rate B_1 while another part transitions towards panic behaviours at the rate B_2. Likewise, a part of the panicked population may switch to controlled behaviour at the rate C_1. According to the evolution of the situation, individuals who have adopted a controlled behaviour may switch back to panic behaviour at the rate C_2. This process can be iterated many times.

ii Processes of imitation and contagion, which are well known in crowd psychology and have been termed 'emotional contagion' [14]. The imitation processes are modelled identically to epidemiological propagation [15]. The imitation is valid in both directions and is modelled by the function $F(r, c)$ Eq. (5.3) for emotional contagion between reflex and controlled behaviour (using the damping coefficients α_1 and α_2), by the function $G(r, p)$ Eq. (5.4) for emotional contagion between reflex and panic behaviour (δ_1 and δ_2), and by the function $H(c, p)$ Eq. (5.5) for emotional contagion between controlled and panicked behaviour (μ_1 and μ_2).

iii Domino effects, which illustrate a succession of events (s_1 and s_2) correspond, for example, to a new attack in an urban area or to a closed door during an evacuation. In the PCR model, the parameters s_1 and s_2 are either constant or built in a periodic form.

The triggering of the threat is represented by a forcing function $\gamma(t)$, the form of which may vary according to the specificities of the danger (event with fast or slow kinetics, expected or not).

5.2.2 Equations of the PCR System

In [4–6], the authors introduce the Panic-Control-Reflex model (PCR) by using a system of five Ordinary Differential Equations (ODE), which describe the human behaviours in one specific place, during a catastrophic event. We include these ODE Eqs. (5.1)–(5.6) in order to describe the modifications we introduce to them in the next section, where we seek to model such behaviours when a network of places and streets or stairs linking those places in a town is considered. In Sect. 5.3, we will identify this network to a mathematical graph; the places are called vertices or nodes, and the streets and stairs are the edges linking those nodes.

$$\dot{X} = \Phi(t, X) \tag{5.1}$$

with $\dot{X} = \frac{dX}{dt}$, $X = (r, c, p, q, b)^T \in \mathbb{R}^5$ and Φ given by $\Phi(t, X) = (\Phi_i(t, X))^T$, $i = 1, \ldots, 5$, where the functions Φ_i are defined by

$$\begin{cases} \Phi_1(t, X) = \gamma(t)q(t)\left(1 - \frac{r(t)}{r_m}\right) - (B_1 + B_2)\,r(t) + s_1(t)c(t) + s_2(t)p(t) \\ \qquad\qquad\quad +F(r(t), c(t))r(t)c(t) + G(r(t), p(t))r(t)p(t) \\ \Phi_2(t, X) = -\varphi(t)c(t)(1 - b(t)) + B_1 r(t) + C_1 p(t) - C_2 c(t) - s_1(t)c(t) \\ \qquad\qquad\quad -F(r(t), c(t))r(t)c(t) + H(c(t), p(t))c(t)p(t) \\ \Phi_3(t, X) = B_2 r(t) - C_1 p(t) + C_2 c(t) - s_2(t)p(t) - G(r(t), p(t))r(t)p(t) \\ \qquad\qquad\quad -H(c(t), p(t))c(t)p(t) \\ \Phi_4(t, X) = -\gamma(t)q(t)\left(1 - \frac{r(t)}{r_m}\right) \\ \Phi_5(t, X) = \varphi(t)c(t)(1 - b(t)) \end{cases} . \tag{5.2}$$

And the variables $r(t), c(t), p(t), q(t), b(t)$ denote, respectively, the *densities* of people being in a reflex, control, panic, daily or back to daily behaviour [6].

The parameters involved in Eq. (5.2) are real positive coefficients previously defined in Sect. 5.2.1: $r_m > 0$ (reflex behaviour maximum value); $B_i \geq 0$, $C_i \geq 0$, $i = 1, 2$; $\alpha_i \geq 0$, $\delta_i \geq 0$, $\mu_i \geq 0$, $i = 1, 2$; $s_i \geq 0$, $i = 1, 2$.

The imitation functions F, G and H are real-valued functions defined on $\mathbb{R} \times \mathbb{R}$ by

$$F(r(t), c(t)) = -\alpha_1 f_1\left(\frac{r(t)}{c(t)+\varepsilon}\right) + \alpha_2 f_2\left(\frac{c(t)}{r(t)+\varepsilon}\right) \tag{5.3}$$

$$G(r(t), p(t)) = -\delta_1 g_1\left(\frac{r(t)}{p(t)+\varepsilon}\right) + \delta_2 g_2\left(\frac{p(t)}{r(t)+\varepsilon}\right) \tag{5.4}$$

$$H(c(t), p(t)) = \mu_1 h_1\left(\frac{c(t)}{p(t)+\varepsilon}\right) - \mu_2 h_2\left(\frac{p(t)}{c(t)+\varepsilon}\right), \tag{5.5}$$

where ε is a positive number and f_i, g_i, h_i for $i = 1, 2$ are real-valued functions defined on $\mathbb{R}$. They have a decreasing shape indicating that the behaviour imitation is symmetric. Moreover they are normalized,

$$0 \leq f_i(u) \leq 1,\ 0 \leq g_i(u) \leq 1,\ 0 \leq h_i(u) \leq 1, \quad \forall u \in \mathbb{R},\ i = 1, 2. \tag{5.6}$$

Because this model does not take the mortality rate into account, the population remains constant and, in one node, can be normalized to 1. Therefore, Eq. (5.1) is considered when time is proceeding from an initial time $t_0 \geq 0$, with initial condition

$$(r(t_0), c(t_0), p(t_0), q(t_0), b(t_0)) = (r_0, c_0, p_0, q_0, b_0) \tag{5.7}$$

that satisfies the following properties

$$r(t_0) > 0,\ c(t_0) > 0,\ p(t_0) > 0,\ q(t_0) > 0,\ b(t_0) > 0 \tag{5.8}$$

$$r(t_0) + c(t_0) + p(t_0) + q(t_0) + b(t_0) = 1. \tag{5.9}$$

Equation (5.9) remains true throughout all the process because the sum of the five Eqs. (5.2) is null, therefore

$$b(t) = 1 - (r(t) + c(t) + p(t) + q(t)), \quad \forall t \geq t_0 \tag{5.10}$$

which implies that Eq. (5.2) can be reduced to the system of only four ODE

$$\begin{cases} \dot{r}(t) &= \gamma(t)q(t)\left(1-\frac{r(t)}{r_m}\right)-(B_1+B_2)\,r(t)+s_1(t)c(t)+s_2(t)p(t)\\ & \quad +F(r(t),c(t))r(t)c(t)+G(r(t),p(t))r(t)p(t)\\ \dot{c}(t) &= -\varphi(t)c(t)(1-b(t))+B_1r(t)+C_1p(t)-C_2c(t)-s_1(t)c(t)\\ & \quad -F(r(t),c(t))r(t)c(t)+H(c(t),p(t))c(t)p(t)\\ \dot{p}(t) &= B_2r(t)-C_1p(t)+C_2c(t)-s_2(t)p(t)-G(r(t),p(t))r(t)p(t)\\ & \quad -H(c(t),p(t))c(t)p(t)\\ \dot{q}(t) &= -\gamma(t)q(t)\left(1-\frac{r(t)}{r_m}\right) \end{cases}. \tag{5.11}$$

The initial condition of Eq. (5.11) corresponding to $(r_0, c_0, p_0, q_0, b_0)$ for Eq. (5.2) becomes simply

$$(r_0, c_0, p_0, q_0). \tag{5.12}$$

5.2.3 Transitional Dynamics

Both forcing functions, γ and φ, respectively model the beginning of the disaster and the return to a quiescent daily behaviour. Their shape can be adapted to various scenarios. When t is sufficiently large, they satisfy $\gamma(t) = \varphi(t) = 1$. In catastrophic situations it is considered that γ is a stiff function, ranging from 0 to 1 in a very brief interval of time [4–6] because if we consider a bomb attack, all the crowd that is near the explosion passes from daily to reflex behaviour in an instant, and it takes a very long time for people to return to their normal state.

Therefore, one can suppose that a terror attack is shaped by two characteristic times: t_s (for start) and t_e (for end) with $t_0 < t_s < t_e$ for which

$$\begin{cases} \gamma(t) = 1, \ \forall t \geq t_s \\ \varphi(t) = 0, \ \forall t < t_e \end{cases}. \tag{5.13}$$

As an example for $I_{trans} = [2.5, 42.5]$, these functions can be defined by (Fig. 5.6)

$$\varphi(t) = \begin{cases} 0 & \text{if } 0 \leq x < 42.5 \\ \cos^2\left(2\pi\frac{x-2.5}{160}\right) & \text{if } 42.5 \leq x \leq 82.5 \\ 1 & \text{if } x > 82.5 \end{cases} \tag{5.14}$$

$$\gamma(t) = \begin{cases} \cos^2\left(2\pi\frac{x-2.5}{10}\right) & \text{if } 0 \leq x \leq 2.5 \\ 1 & \text{if } x > 2.5 \end{cases}. \tag{5.15}$$

Following [6] we keep the term *transitional dynamics* for the dynamics of the PCR model (likewise for both improved WPCR and CWPCR models presented in Sect. 5.3) in the interval of time $I_{trans} = [t_s, t_e]$ (in terror attacks, this interval of time can last from several minutes up to hours as observed during the 2016 terrorist attack in Nice). Therefore in $\forall t \in I_{trans}$ functions, φ and γ verify

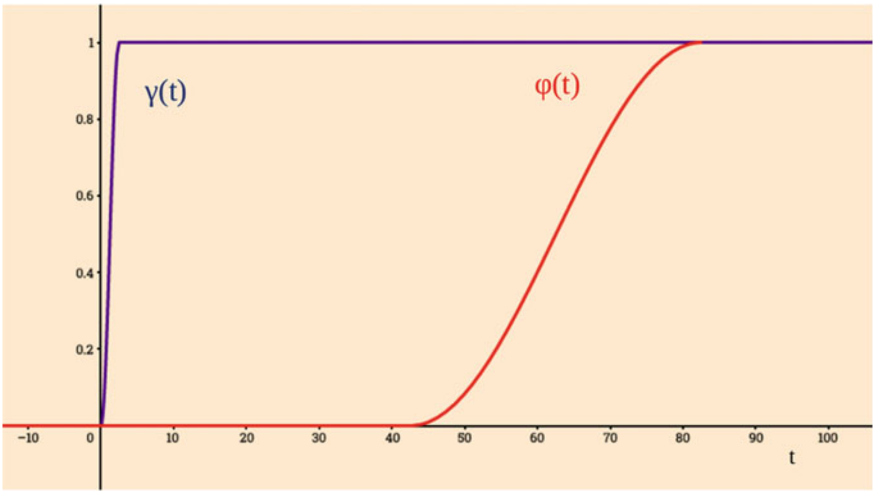


Fig. 5.6 Forcing functions $\gamma(t)$, blue curve, and $\varphi(t)$, red curve, $I_{trans} = [2.5, 42.5]$

$$\begin{cases} \gamma(t) = 1 \\ \varphi(t) = 0 \end{cases} . \tag{5.16}$$

Hence, during the transitional dynamics, the population with daily behaviour collapses and there is not yet a population that is back to daily behaviour (i.e. $q(t) = b(t) = 0$).

System (5.11) is reduced to

$$\begin{cases} \dot{r}(t) &= -(B_1 + B_2)\, r(t) + s_1(t)c(t) + s_2(t)p(t) \\ &\quad +F(r(t), c(t))r(t)c(t) + G(r(t), p(t))r(t)p(t) \\ \dot{c}(t) &= B_1 r(t) + C_1 p(t) - C_2 c(t) - s_1(t)c(t) \\ &\quad -F(r(t), c(t))r(t)c(t) + H(c(t), p(t))c(t)p(t) \\ \dot{p}(t) &= B_2 r(t) - C_1 p(t) + C_2 c(t) - s_2(t)p(t) - G(r(t), p(t))r(t)p(t) \\ &\quad -H(c(t), p(t))c(t)p(t). \end{cases} \tag{5.17}$$

5.3 Mathematical Weighted and Coupled PCR System

5.3.1 The Weighted PCR System

Equations (5.1)–(5.10) model the human behaviours, in one specific place, during a catastrophic event. As explained in the first part of this article, it is based on neuroscience studies.

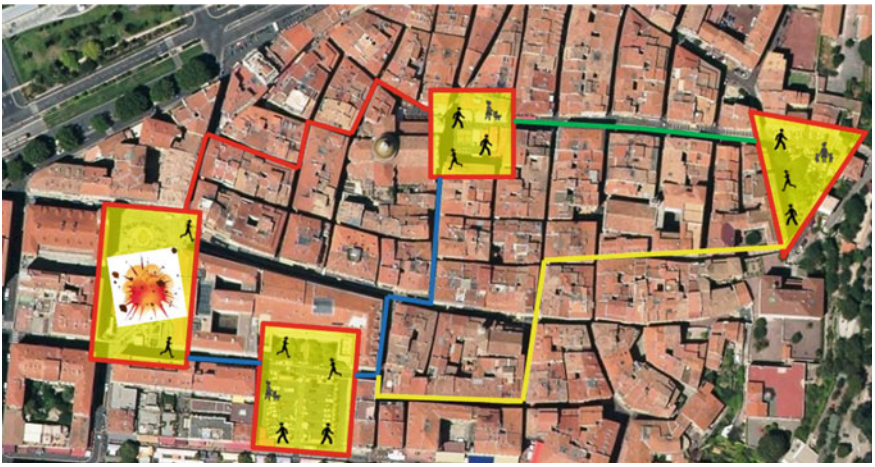


Fig. 5.7 Possible paths of rushing people in a city network (IGN—BD Ortho, 2017, 50 cm resolution)

However just before a dramatic event like a terror attack in a city, the crowd is generally spread across several places, public squares and streets. In the aftermath of the initial shock, people are rushing through the streets to reach what they think will be more secure places (Fig. 5.7).

In this article, we define the city by a mathematical graph, where the places and public squares are called vertices or nodes and are denoted by $N_1, N_2, \ldots, N_p$, and the streets, escalators, doors, and stairs are the oriented edges $(N_i \rightarrow N_j)$ linking these nodes. They are oriented because the flow from one place towards another is not symmetric. Our aim is to model the motion of the crowd through such edges. To achieve this, we must introduce some 'geographical' particularities of the city, like the size of places, the narrowness of streets and the number of people initially present in each place. This is why we need to upgrade the standard PCR model into the Weighted PCR (WPCR) model, by introducing new data, with weight standing for the relative sizes of crowd, places and streets.

First, on each node $N_k, k = 1, p$ we call $r_k(t)$, $c_k(t)$, $p_k(t)$, $q_k(t)$, $b_k(t)$ the *number* of people being in reflex, control, panic, daily and back to daily behaviour and $V_k(t)$ the total number of people present at this node

$$V_k(t) = r_k(t) + c_k(t) + p_k(t) + q_k(t) + b_k(t). \tag{5.18}$$

Since we consider that the nodes are not identical (as is seen in an actual city), they do not generally contain, at each moment, the same number of people. Moreover, this number is varying with time when the crowd is moving through the streets. That is why it is more convenient to consider that the five variables of the WPCR model represent *actual numbers* of people, rather than *densities*, as in the PCR model. Densities can be used only when there is no motion at all between places, and the population in each place is the same. Of course, this number of people can be transformed locally to density whenever it is necessary.

Second, in order to more precisely model the characteristics of the city, we introduce W_k, the maximum capacity of the number of people who can be present in each node N_k (i.e. due to the size of the corresponding place). Each maximum capacity is a constant. At every time one must have

$$V_k(t) \leq W_k. \tag{5.19}$$

The Weighted Panic-Control-Reflex model (WPCR) is then defined on each node k by

$$\dot{X}_k = \Phi(t, X_k) \tag{5.20}$$

with $\dot{X}_k = \frac{dX_k}{dt}$, $X_k = (r_k, c_k, p_k, q_k, b_k)^T \in \mathbb{R}^5$ and Φ given by $\Phi(t, X_k) = (\Phi_i(t, X_k))^T$, $i = 1, \ldots, 5$, where the functions Φ_i are defined by

$$\begin{cases} \Phi_1(t, X_k) = \dot{r}_k(t) = & \gamma(t) q_k(t)\left(W_k - r_k(t)\right) - (B_1 + B_2)\, r_k(t) + s_1(t) c_k(t) \\ & + s_2(t) p_k(t) + F(r_k(t), c_k(t)) r_k(t) c_k(t) \\ & + G(r_k(t), p_k(t)) r_k(t) p_k(t) \\ \Phi_2(t, X_k) = \dot{c}_k(t) = & -\varphi(t) c_k(t)(W_k - b_k(t)) + B_1 r_k(t) + C_1 p_k(t) - C_2 c_k(t) \\ & - s_1(t) c_k(t) - F(r_k(t), c_k(t)) r_k(t) c_k(t) \\ & + H(c_k(t), p_k(t)) c_k(t) p_k(t) \\ \Phi_3(t, X_k) = \dot{p}_k(t) = & B_2 r_k(t) - C_1 p_k(t) + C_2 c_k(t) - s_2(t) p_k(t) \\ & - G(r_k(t), p_k(t)) r_k(t) p_k(t) - H(c_k(t), p_k(t)) c_k(t) p_k(t) \\ \Phi_4(t, X_k) = \dot{q}_k(t) = & -\gamma(t) q_k(t)\left(W_k - r_k(t)\right) \\ \Phi_5(t, X_k) = \dot{b}_k(t) = & \varphi(t) c_k(t)(W_k - b_k(t)) \end{cases} \tag{5.21}$$

the initial condition satisfies

$$\begin{aligned} r_k(t_0) + c_k(t_0) + p_k(t_0) + q_k(t_0) + b_k(t_0) &= V_k(t_0) \\ &= r_{k,0} + c_{k,0} + p_{k,0} + q_{k,0} + b_{k,0} = V_{k,0} \leq W_k^{\cdot} \end{aligned} \tag{5.22}$$

We suppose that the characteristic parameters B_1, B_2, C_1, C_2, of each node have the same value because they depend on cultural and psychological factors specific to each individual rather than to spatial configurations and crowd context. Thus, all the parameters and the functions are the same as those defined in Sect. 5.2.2, which is why we use the function Φ_i instead of function $\Phi_{k,i}$.

5.3.2 Flows and Bottleneck Coupling

As previously defined, nodes are linked by edges. We now aim to model the motion of the crowd through such edges (i.e. streets, stairs, escalators and doors). For the sake

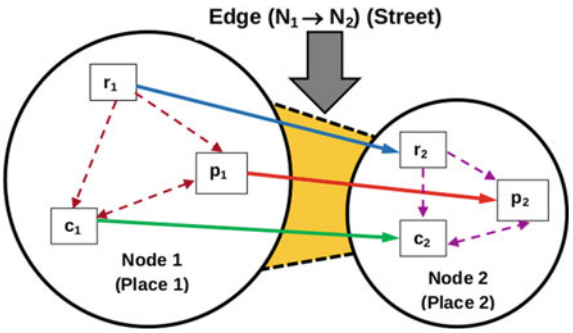


Fig. 5.8 Two nodes network: people in every state are rushing from node 1 towards node 2, maintaining the same behavioural class (this representative colour scheme is used for the remainder of the article.)

of simplicity, we present the coupling on a simplified oriented network with only two nodes (N_1; W_1) and (N_2; W_2), and one edge ($N_1 \rightarrow N_2$). We suppose that during the short interval of time when people are travelling inside one edge, they remain in the same behavioural class (Fig. 5.8.). In this figure, the 'geographical' edge ($N_1 \rightarrow N_2$) (i.e. the street linking node 1 to node 2) is split into three 'behavioural edges' meaning that on the same street, people in reflex, panic or controlled behaviour are escaping from node 1 to node 2, therefore people in each particular behaviour in node 1 are meeting people in the same behaviour in node 2.

To continue focusing on the special coupling that we are introducing here, we suppose that imitation mechanisms are not activated (i.e. $F \equiv H \equiv G \equiv 0$, which is equivalent to $\alpha_i = \delta_i = \mu_i = 0$ for $i = 1, 2$), and, furthermore, there is no domino effect ($s_i = 0, i = 1, 2$) and we also suppose that we are in the interval $I_{trans} = [t_s, t_e]$ where only transitional dynamics are considered. Of course, it is easy to relax such limitations, which are not dependent upon the coupling, by not eliminating the corresponding terms in the equations.

In each node $N_k, k = 1, 2$; such transitional dynamics are the solution of the system

$$\begin{cases} \dot{r}_k(t) &= -(B_1 + B_2)r_k(t) \\ \dot{c}_k(t) &= B_1 r_k(t) + C_1 p_k(t) - C_2 c_k(t) \\ \dot{p}_k(t) &= B_2 r_k(t) - C_1 p_k(t) + C_2 c_k(t) \end{cases} \tag{5.23}$$

which is the reduction of system (5.21) in the transition interval (as Eq. (5.17) is the reduction of Eq. (5.2)).

Note: In the WPCR model, the terms $\dot{r}_k(t)$, $\dot{c}_k(t)$, $\dot{p}_k(t)$, $\dot{q}_k(t)$, $\dot{b}_k(t)$ can be considered as flows, because a flow is a quantity of something divided by a unit of time. There are two kinds of flows. In Eqs. (5.21)–(5.23), flows are 'behavioural', as they represent the quantities of people changing their behaviour per unit of time. Now we consider also 'motion' flows, which are the quantities of people in each behaviour class, moving from one node to another node, per unit of time. Of course, both kinds of flows are combined to produce a global equivalent in the following equations.

In [16] different types of flow situations are considered in pedestrian facilities, such as unidirectional, bidirectional and crossing: it is said that 'Unidirectional and

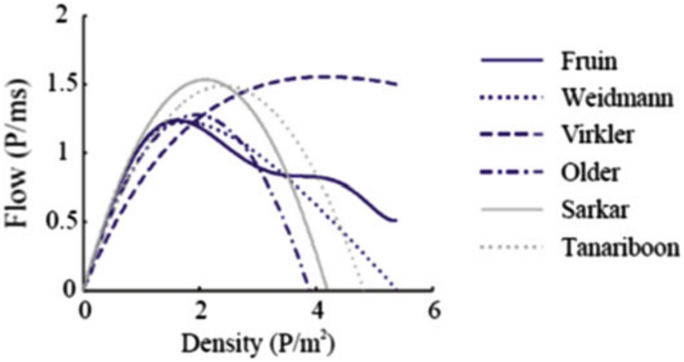


Fig. 5.9 Relationship between flow and density of pedestrians from the literature, from [[17], derived from Fruin, Weidmann, Virkler, Older, Sarkar and Tanariboon]

bidirectional flow conditions can be commonly observed in corridors, stairs and bottlenecks of pedestrian facilities such as transport terminals and shopping malls. An understanding of the fundamental relationship between flow–speed–density is important in the planning, design and operation of pedestrian facilities. Capturing the realistic behaviour of pedestrians in various pedestrian facilities with different geometric elements such as corridors, bottlenecks, stairs and escalators are essential in order to estimate the flow parameters accurately. The important parameters such as the width of the bottleneck and slope of the stairs play a vital role in deciding the capacity of the respective element. The flow density relationship for different geometric elements is important and further analysis like spatial and temporal development of the basic quantities (velocity, density and flow) on different elements like corridors, stairs and bottlenecks should be considered'.

Many studies on pedestrian flows have been published [17–23]. We consider, in particular, the survey of [17], in which the graphs of six different experiments showing the relationship between the flow of pedestrians and their densities are displayed. All these graphs show clearly a non-linear relationship of a logistic type between density and flow (Fig. 5.9).

Moreover, Daamen and co-authors developed a first-order traffic flow theory to describe two-dimensional pedestrian flow operations in the case of an oversaturated bottleneck in front of which a large, high-density region has formed (Fig. 5.10). Such a mathematical model also highlights the logistic relationship for any bottleneck width.

We now introduce our hypothesis for the coupled WPCR (CWPCR), based on this type of logistic relationship. Again, for the sake of simplicity, we consider only the unidirectional motion of the crowd (i.e. motion on an oriented graph) as in Fig. 5.8, because it is supposed that a terrorist attack occurs in node 1 and that people try to escape from this node towards node 2. As an aside, in forthcoming research we will allow bidirectional motion in more complex networks, where the bidirectional motion will be simply obtained by adding symmetric terms in Eq. (5.24).

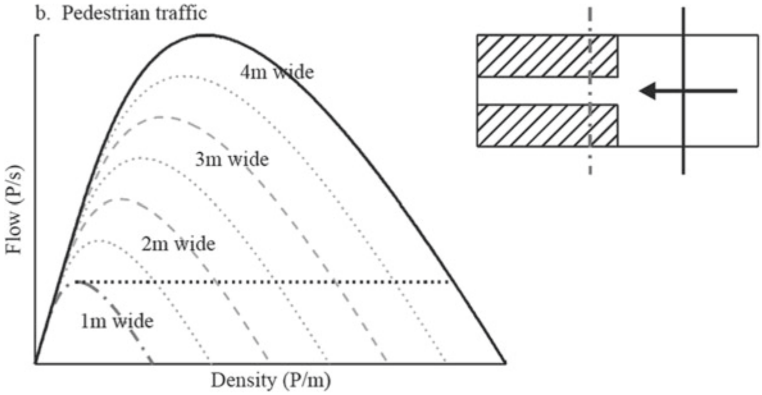


Fig. 5.10 Relationship between flow and density of pedestrians going through an oversaturated bottleneck, from [17]

When the crowd is moving from one node to another, its speed and the corresponding motion flow depends on three factors. The first factor reflects the narrowness and the length of the street. More people can go from one place to the next if the street is large, rather than in the case of a narrow street. This topological characteristic will be modelled by a 'roughness' coefficient $\eta_{1,2}$. In fact, in the considered coupling, we suppose that people cannot change their behaviour when they move from one node to another (e.g. controlled people remain controlled, panicked people remain panicked and so on), and thus we use three such roughness coefficients: $\eta_{c,1,2}$, $\eta_{p,1,2}$, $\eta_{r,1,2}$, to refer to the population in reflex, panic or controlled situations, respectively. Of course, they can have the same value. The second factor is proportional to the number of people present in the first node. This is equivalent to the pressure of the crowd in the first place. The third factor reflects the counter pressure due to the maximal capacity of the second place conjugated with the number of persons already present there. The combination of pressure and counter pressure gives a bottleneck effect.

We propose to model this bottleneck effect by a non-linearity of a logistic type to keep the same philosophy as the authors cited above.

Thus, the bottleneck coupling corresponding to Fig. 5.8 is given by the system

$$\left\{\begin{array}{lcl}
\dot{r}_1(t) & = & -(B_1+B_2)r_1(t) - \eta_{r,1,2}r_1(t)(W_2 - c_2(t) - p_2(t) - r_2(t)) \\
\dot{c}_1(t) & = & B_1 r_1(t) + C_1 p_1(t) - C_2 c_1(t) - \eta_{c,1,2}c_1(t)(W_2 - c_2(t) - p_2(t) - r_2(t)) \\
\dot{p}_1(t) & = & B_2 r_1(t) - C_1 p_1(t) + C_2 c_1(t) - \eta_{p,1,2}p_1(t)(W_2 - c_2(t) - p_2(t) - r_2(t)) \\
\dot{r}_2(t) & = & -(B_1+B_2)r_2(t) + \eta_{r,1,2}r_1(t)(W_2 - c_2(t) - p_2(t) - r_2(t)) \\
\dot{c}_2(t) & = & B_1 r_2(t) + C_1 p_2(t) - C_2 c_2(t) + \eta_{c,1,2}c_1(t)(W_2 - c_2(t) - p_2(t) - r_2(t)) \\
\dot{p}_2(t) & = & B_2 r_2(t) - C_1 p_2(t) + C_2 c_2(t) + \eta_{p,1,2}p_1(t)(W_2 - c_2(t) - p_2(t) - r_2(t))
\end{array}\right. \tag{5.24}$$

with initial conditions satisfying

$$r_{1,0} + c_{1,0} + p_{1,0} = V_{1,0} \leq W_1; \qquad r_{2,0} + c_{2,0} + p_{2,0} = V_{2,0} \leq W_2.$$

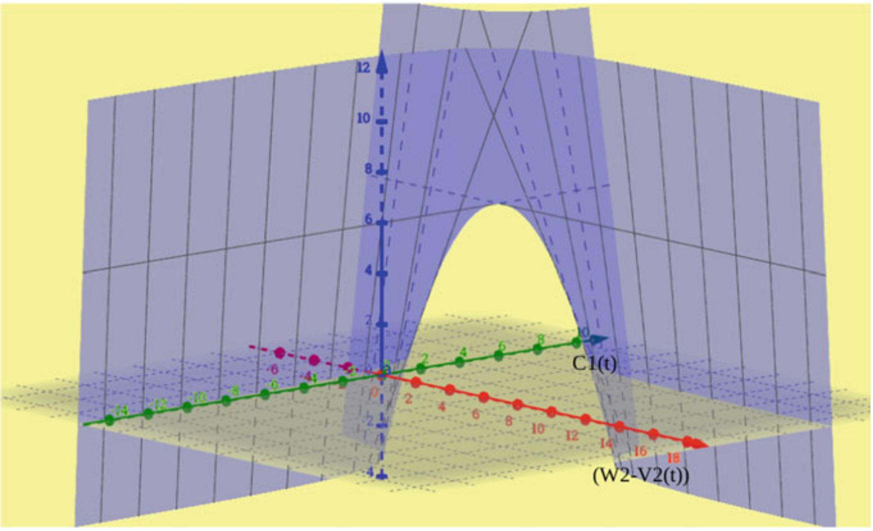


Fig. 5.11 Graph of the bottleneck coupling function

In this system the bottleneck coupling, concerning, for example, the controlled population that is moving from node 1 to node 2 is given by the term:

$$\eta_{c,1,2}c_1(t)(W_2 - c_2(t) - p_2(t) - r_2(t)), \tag{5.25}$$

in the second equation of (5.24)

$$\dot{c}_1(t) = B_1 r_1(t) + C_1 p_1(t) - C_2 c_1(t) - \eta_{c,1,2}c_1(t)(W_2 - c_2(t) - p_2(t) - r_2(t)). \tag{5.26}$$

In this bottleneck coupling (5.25), $\eta_{c,1,2}$ is the parameter which models the topological characteristic of the street linking node 1 to node 2. The second factor $c_1(t)$ of (5.25) reflects the pressure of controlled people in node 1 willing to escape towards node 2 and also the proportionality of people escaping with respect to people staying in node 1. Finally, the factor $(W_2 - c_2(t) - p_2(t) - r_2(t))$ shows the counter-pressure which is maximum (i.e. the term vanishes) when $W_2 = c_2(t) + p_2(t) + r_2(t)$ because, in this case, there is no more room for people coming from node 1. This bottleneck coupling is non-linear as shown in Fig. 5.11.

As we consider only transitional dynamics where $q_k(t) = b_k(t) = 0$, Eq. (5.25) can be written as

$$\eta_{c,1,2}c_1(t)(W_2 - V_2(t)). \tag{5.27}$$

5.3.3 Fixed Points of the Two-Node System

The fixed-point research allows us to identify the point of equilibrium towards which the system tends during the transitional period. This equilibrium point highlights the primordial role of the size of node 2 in the context of evacuation dynamics. It is important to note that the mathematically calculated equilibrium point does not necessarily correspond to the equilibrium situation sought by crisis management personnel. Thus, the equilibrium point obtained in situation 2 below, in which a part of the population cannot escape and remains in a dangerous place, does not correspond to a crisis equilibrium situation. The population stranded in the initial place (node 1) remains very vulnerable to the terrorist threat.

The fixed point $(r_1^*, c_1^*, p_1^*, r_2^*, c_2^*, p_2^*)$ of the system towards which the solution Eq. (5.24) converges is easily computed.

$$\begin{cases}
\begin{cases}
r_1^* &= 0 \\
c_1^* &= 0 \\
p_1^* &= 0 \\
r_2^* &= 0 \\
c_2^* &= \frac{C_1(V_{1,0}+V_{2,0})}{C_1+C_2} \\
p_2^* &= \frac{C_2(V_{1,0}+V_{2,0})}{C_1+C_2}
\end{cases} & \text{if } r_{1,0}+c_{1,0}+p_{1,0}+r_{2,0}+c_{2,0}+p_{2,0}=V_{1,0}+V_{2,0}\le W_2, \\
\begin{cases}
r_1^* &= 0 \\
c_1^* &= \frac{C_1(V_{1,0}+V_{2,0}-W_2)}{C_1+C_2} \\
p_1^* &= \frac{C_2(V_{1,0}+V_{2,0}-W_2)}{C_1+C_2} \\
r_2^* &= 0 \\
c_2^* &= \frac{C_1 W_2}{C_1+C_2} \\
p_2^* &= \frac{C_2 W_2}{C_1+C_2}
\end{cases} & \text{if } r_{1,0}+c_{1,0}+p_{1,0}+r_{2,0}+c_{2,0}+p_{2,0}=V_{1,0}+V_{2,0}> W_2,
\end{cases} . \tag{5.28}$$

The values of this fixed point mean the following.

Situation 1: If the number of people initially staying in both nodes is less than the capacity of refuge in node 2 (i.e. $V_{1,0} + V_{2,0} \le W_2$), after a while, node 1 becomes empty and all the crowd has sought refuge in node 2.

Alternatively, in **Situation 2**: If this number is greater than the capacity (i.e. $V_{1,0} + V_{2,0} > W_2$), then node 2 becomes full and the remaining people $W_2 - (V_{1,0} + V_{2,0})$ are still stranded in node 1.

From Eq. (5.28), it is obvious that only the ratio $\frac{C_1}{C_2}$ is significant for the limit of solutions of Eq. (5.24) because $\frac{c_1^*}{p_1^*} = \frac{c_2^*}{p_2^*} = \frac{C_1}{C_2}$, when defined, instead of parameters B_1 and B_2 becomes important for the pace at which the ‘reservoir’ of people in reflex behaviour is emptied.

Of course, across the world there are different cultures, which lead to different behaviours. These behaviours can be modelled by varying parameters.

For example, if populations are not made aware of major risks and not prepare for them, it causes a panic reaction ($B_1 < B_2$); this behaviour is then regulated by the ratio $\frac{C_1}{C_2}$. The higher this ratio, the more the population remains or transits in the controlled state.

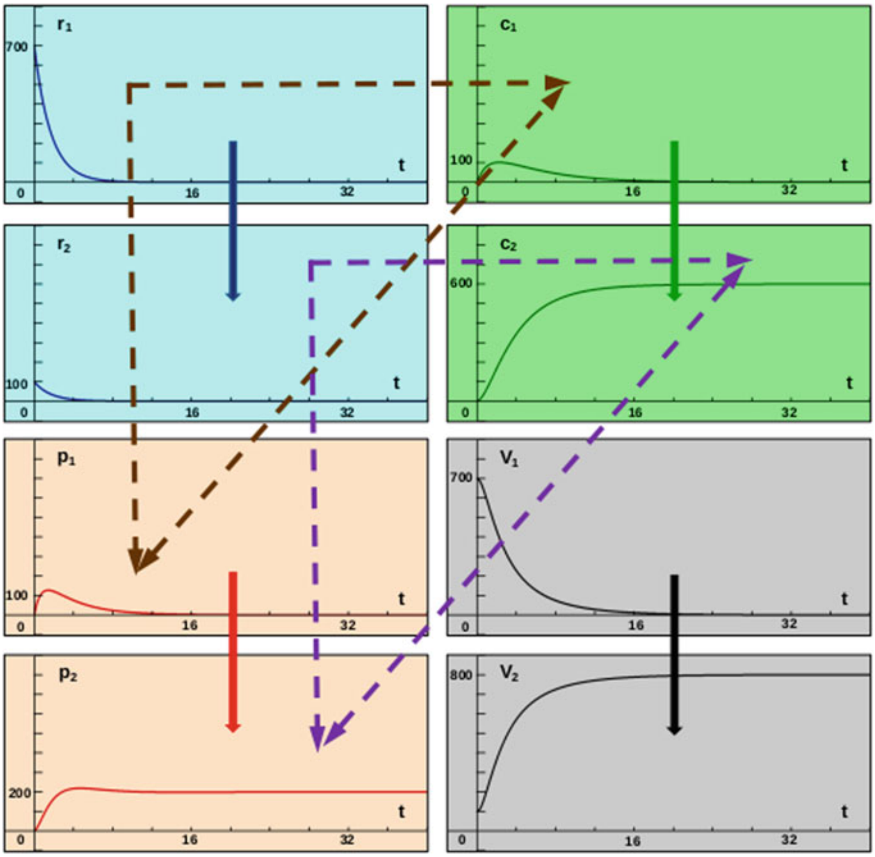


Fig. 5.12 Situation 1: Convergence towards the fixed point (0, 0, 0, 0, 600, 200). In this figure, the change of behavioural states is symbolized by dotted arrows, and the motion between nodes by plain arrows is as displayed in Fig. 5.8

In both simulated situations, we choose a set of parameters that highlight a weak risk culture while favouring the return to a controlled behaviour; instead, the values, $B_1 = 0.2$, $B_2 = 0.4$, mean that there is a weak risk culture and $C_1 = 0.3$, $C_2 = 0.1$ mean that the panic in the crowd context is compensated by controlled reactions for a part of the population who keep self-control, notably because there is no new threat or sudden attack.

Therefore, in **situation 1** (Fig. 5.12) when $V_{1,0} = 700$, $V_{2,0} = 100$, $W_2 = 1000$ and $\eta_{r,1,2} = \eta_{c,1,2} = \eta_{p,1,2} = 0.001$, one obtains the following convergence towards the fixed point: $r_1^* = 0$, $c_1^* = 0$, $p_1^* = 0$, $r_2^* = 0$, $c_2^* = \frac{C_1(V_{1,0}+V_{2,0})}{C_1+C_2} = \frac{0.3(700+100)}{0.3+0.1} = 600$, $p_2^* = \frac{C_2(V_{1,0}+V_{2,0})}{C_1+C_2} = 200$ and $V_1^* = 0$, $V_2^* = 800 < W_2 = 1000$.

In **situation 2** (Fig. 5.13) when $V_{1,0} = 700$, $V_{2,0} = 100$, $W_2 = 500$, one obtains the following convergence towards the fixed point: $r_1^* = 0$, $c_1^* = \frac{C_1(V_{1,0}+V_{2,0}-W_2)}{C_1+C_2} =$

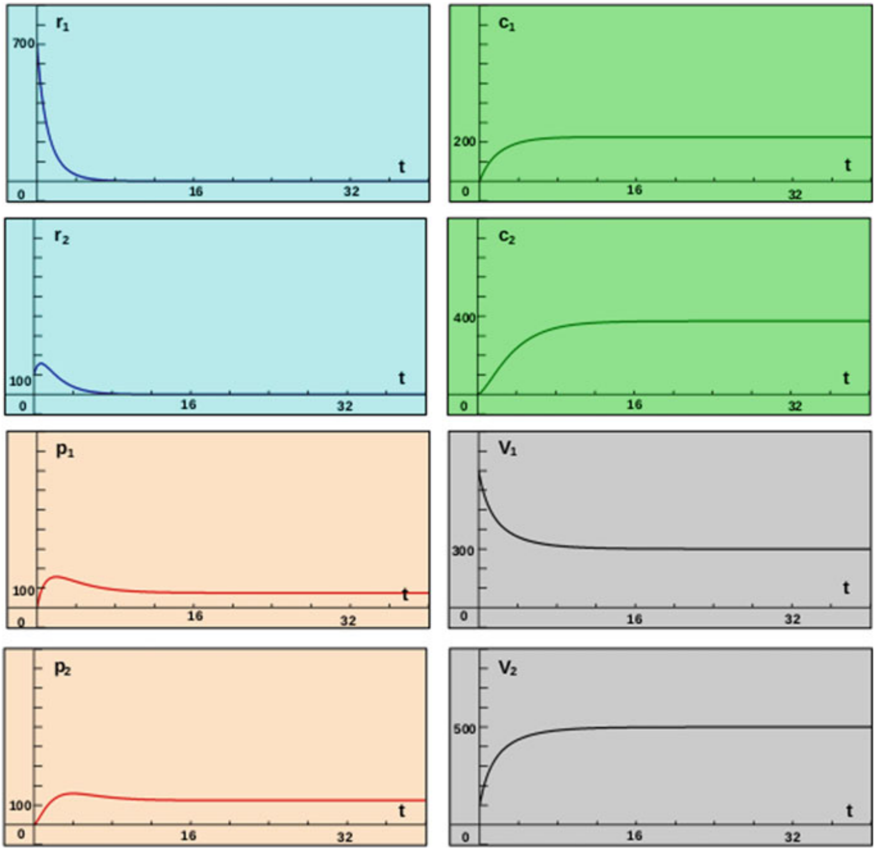


Fig. 5.13 Situation 2: Convergence towards the fixed point (0, 225, 75, 0, 375, 125)

$\frac{0.3(700+100-500)}{0.3+0.1}$=225, $p_1^* = \frac{C_2(V_{1,0}+V_{2,0}-W_2)}{C_1+C_2} = 75$, $r_2^* = 0$, $c_2^* = \frac{C_1 W_2}{C_1+C_2} = 375$, $p_2^* = \frac{C_2 W_2}{C_1+C_2} = 125$ and $V_1^* = 300$, $V_2^* = 500 = W_2$.

Examining both Figs. 5.12 and 5.13, it appears that one of the main trends of the CWPCR model is to take into account the role of the spatial configuration on both behavioural dynamics and on the crowd'ability to escape from a dangerous place towards a place of shelter, when such a place offers sufficient room for the crowd. In effect, in the first situation, everyone can leave place 1, as we can see on Fig. 5.12 where $V_1(t)$ vanishes for $t > 15$ mn, and the entire population escapes to place 2 ($V_2(t) = 800$ for $t > 15$ min). However, in the second situation, a part of the population remains stranded in place 1, because $W_2 = 500$, which is less than the total number of the crowd equal to 800. Therefore, we can see in Fig. 5.13 that $V_1(t)$ tends to $300 = 800 - 500$. These first results lead us to consider more complicated spatial configuration networks in the following section.

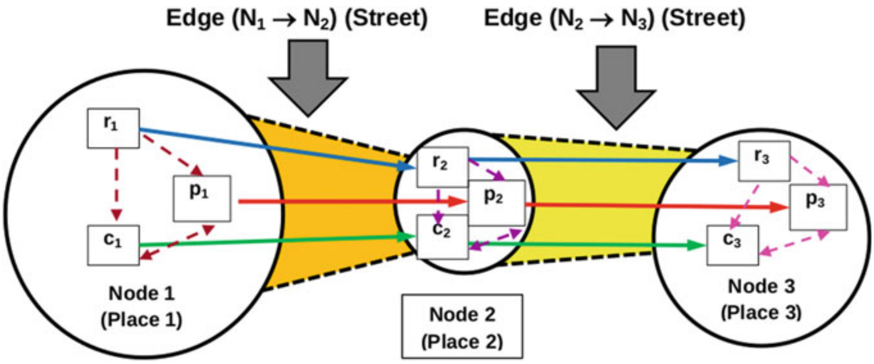


Fig. 5.14 Three-node network: people in every state are rushing from node 1 towards node 2, and from node 2 towards node 3, keeping the same behavioural class

5.4 Influence of the Spatial Configuration on the Pace of Evacuation

We now seek to identify the obstacles that slow the escape of the crowd in the aftermath of the initial shock by analysing the topology of the network of streets and places in a city. This information can potentially be used to improve the design of a city to facilitate the escape of a crowd towards more secure places.

For the sake of simplicity, we consider first a simplified oriented network with only three nodes $(N_1;\ W_1)$; $(N_2;\ W_2)$; $(N_3;\ W_3)$ and two edges $(N_1 \to N_2)$; $(N_2 \to N_3)$ (Fig. 5.14). Such a simplified network can be straightforward complexified by adding as many nodes and edges as necessary, without any difficulty. However, it is better to first focus our attention on the nature of the obstacles in this simplified network.

5.4.1 Equation of the Three-Node Network

The corresponding equation of the three-node network is

$$\left\{\begin{array}{lcl}
\dot{r}_1(t) & = & -(B_1+B_2)r_1(t) - \eta_{r,1,2}r_1(t)(W_2 - c_2(t) - p_2(t) - r_2(t)) \\
\dot{c}_1(t) & = & B_1 r_1(t) + C_1 p_1(t) - C_2 c_1(t) - \eta_{c,1,2}c_1(t)(W_2 - c_2(t) - p_2(t) - r_2(t)) \\
\dot{p}_1(t) & = & B_2 r_1(t) - C_1 p_1(t) + C_2 c_1(t) - \eta_{p,1,2}p_1(t)(W_2 - c_2(t) - p_2(t) - r_2(t)) \\
\dot{r}_2(t) & = & -(B_1+B_2)r_2(t) + \eta_{r,1,2}r_1(t)(W_2 - c_2(t) - p_2(t) - r_2(t)) \\
 & & -\eta_{r,2,3}r_2(t)(W_3 - c_3(t) - p_3(t) - r_3(t)) \\
\dot{c}_2(t) & = & B_1 r_2(t) + C_1 p_2(t) - C_2 c_2(t) + \eta_{c,1,2}c_1(t)(W_2 - c_2(t) - p_2(t) - r_2(t)) \\
 & & -\eta_{c,2,3}c_2(t)(W_3 - c_3(t) - p_3(t) - r_3(t)) \\
\dot{p}_2(t) & = & B_2 r_2(t) - C_1 p_2(t) + C_2 c_2(t) + \eta_{p,1,2}p_1(t)(W_2 - c_2(t) - p_2(t) - r_2(t)) \\
 & & -\eta_{p,2,3}p_2(t)(W_3 - c_3(t) - p_3(t) - r_3(t)) \\
\dot{r}_3(t) & = & -(B_1+B_2)r_3(t) + \eta_{r,2,3}r_2(t)(W_3 - c_3(t) - p_3(t) - r_3(t)) \\
\dot{c}_3(t) & = & B_1 r_3(t) + C_1 p_3(t) - C_2 c_3(t) + \eta_{c,2,3}c_2(t)(W_3 - c_3(t) - p_3(t) - r_3(t)) \\
\dot{p}_3(t) & = & B_2 r_3(t) - C_1 p_3(t) + C_2 c_3(t) + \eta_{p,2,3}p_2(t)(W_3 - c_3(t) - p_3(t) - r_3(t))
\end{array}\right. . \quad (5.29)$$

5.4.2 Scaling the Parameters

Fixing the value of all the parameters in PCR, WPCR or CWPCR is a very complicated task, which has not yet been done on an experimental basis. In the framework of the Com2SiCa project, an experimental protocol is under scrutiny, in order to achieve these results in the near future. However, it is important to note that only relative values between parameters are important, because there is a relationship between the unit of time and the unit used for the parameters. In other words, considering all the parameters of the adimensional Eq. (5.29), (i.e. $B_i \geq 0, C_i \geq 0, \eta_{c,i,j}, \eta_{p,i,j}, \eta_{r,i,j}, i = 1, 2, j = 2, 3$) integrated with respect to the variable time t, it is nearly equivalent to consider such parameters multiplied by the same constant κ and integrated using the time variable $\tau = \frac{t}{\kappa}$ (e.g. t can be considered in seconds, minutes, or hours). There is not, strictly speaking, equivalence because such parameters are linearly used in the WPCR model, however in the CWPCR model the coupling is non-linear and a slight distortion intervenes during a transient short period for some variables.

5.4.3 Speed of Convergence Towards the Fixed Points

To stay closer to reality, we further assume that people in reflex situation are stunned and paralysed. They cannot rush from one node to another (Fig. 5.15), therefore the transitions between r_1, r_2, r_3 are forbidden. People in this stunned state can only change their behaviour (from r to p or c).

To achieve this goal Eq. (5.29) is simply modified to vanish parameters $\eta_{r,i,j}, i = 1, 2, j = 2, 3$.

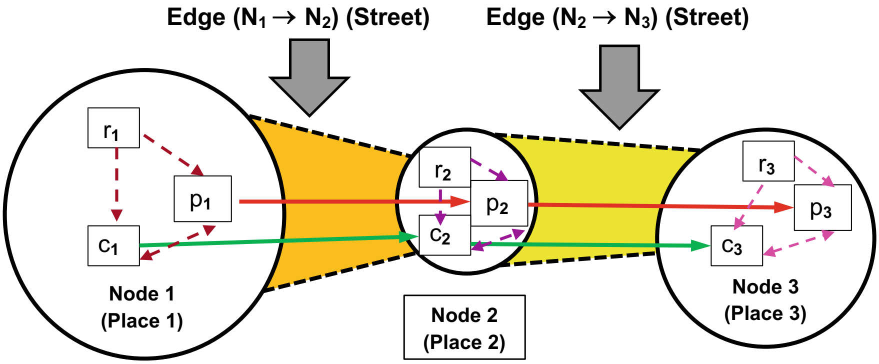


Fig. 5.15 Three-node network: people in the reflex behaviour state are stranded in their original node

As previously described (Sect. 5.3.3), we consider the values of the parameters $B_1 = 0.2$, $B_2 = 0.4$, $C_1 = 0.3$, $C_2 = 0.1$ and we choose $\eta_{c,1,2} = \eta_{c,2,3} = \eta_{p,1,2} = \eta_{p,2,3} = 0.005$.

For the following values $V_{1,0} = 20{,}000$, $V_{2,0} = 0$, $V_{3,0} = 0$, $W_2 > 20{,}000$, $W_1 = 1000$, $W_3 = 20{,}500$, the fixed point $(r_1^*, c_1^*, p_1^*, r_2^*, c_2^*, p_2^*, r_3^*, c_3^*, p_3^*) = (0, 0, 0, 0, 0, 0, 0, 15{,}000, 5000)$ is straightforward to compute the following.

$$\begin{cases} r_1^* &= 0 \\ c_1^* &= 0 \\ p_1^* &= 0 \\ r_2^* &= 0 \\ c_2^* &= 0 \\ p_2^* &= 0 \\ r_3^* &= 0 \\ c_3^* &= \frac{C_1(V_{1,0}+V_{2,0}+V_{3,0})}{C_1+C_2} \\ p_3^* &= \frac{C_2(V_{1,0}+V_{2,0}+V_{3,0})}{C_1+C_2} \end{cases} \quad \text{if } V_{1,0} + V_{2,0} + V_{3,0} \leq W_3 \ , \tag{5.30}$$

$$\begin{cases} r_1^* &= 0 \\ c_1^* &= 0 \\ p_1^* &= 0 \\ r_2^* &= 0 \\ c_2^* &= \frac{C_1(V_{1,0}+V_{2,0}+V_{3,0}-W_3)}{C_1+C_2} \\ p_2^* &= \frac{C_2(V_{1,0}+V_{2,0}+V_{3,0}-W_3)}{C_1+C_2} \\ r_3^* &= 0 \\ c_3^* &= \frac{C_1 W_3}{C_1+C_2} \\ p_3^* &= \frac{C_2 W_3}{C_1+C_2} \end{cases} \quad \text{if } W_3 \leq V_{1,0} + V_{2,0} + V_{3,0} \leq W_2 + W_3 \ , \tag{5.31}$$

$$\begin{cases} r_1^* &= 0 \\ c_1^* &= \frac{C_1(V_{1,0}+V_{2,0}+V_{3,0}-W_2-W_3)}{C_1+C_2} \\ p_1^* &= \frac{C_2(V_{1,0}+V_{2,0}+V_{3,0}-W_2-W_3)}{C_1+C_2} \\ r_2^* &= 0 \\ c_2^* &= \frac{C_1 W_2}{C_1+C_2} \\ p_2^* &= \frac{C_2 W_2}{C_1+C_2} \\ r_3^* &= 0 \\ c_3^* &= \frac{C_1 W_3}{C_1+C_2} \\ p_3^* &= \frac{C_2 W_3}{C_1+C_2} \end{cases} \quad \text{if } W_2 + W_3 \leq V_{1,0} + V_{2,0} + V_{3,0} \ . \tag{5.32}$$

Moreover $V_1^* = 0$, $V_2^* = 0$, $V_3^* = 20{,}000$. This value means that, initially, all the people are in node 1 and both nodes 2 and 3 are empty, but after a certain period of time, both nodes 1 and 2 are empty and everyone has reached node 3. One can see the flow of people through the two edges $(N_1 \to N_2)$; $(N_2 \to N_3)$, in Fig. 5.16.

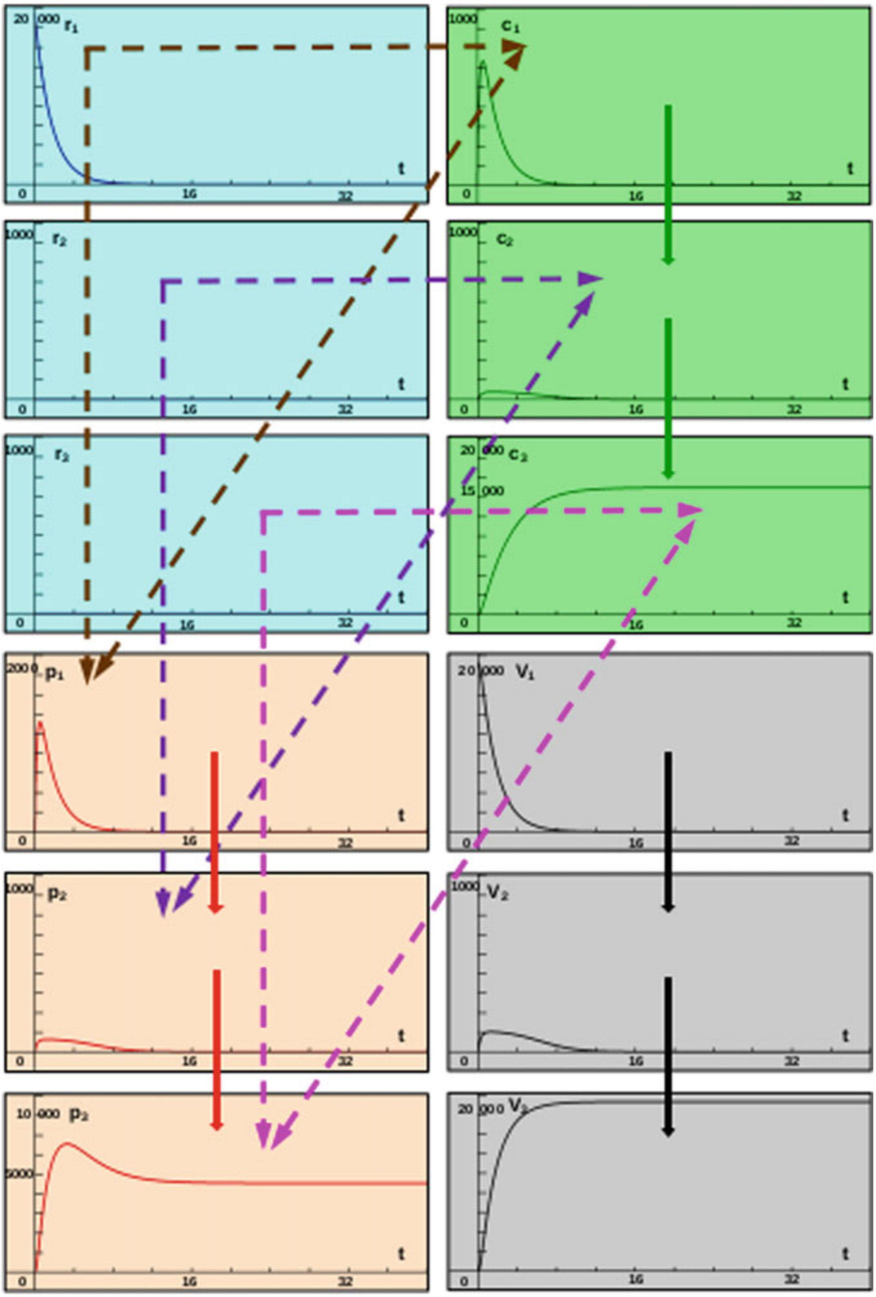


Fig. 5.16 Convergence towards the fixed point (0, 0, 0, 0, 0, 0, 0, 15,000, 5000)

5.4.4 Influence of the Intermediate Place Capacity on the Evacuation Dynamics

Intermediate places play a central role in the fluidity or, to the contrary, the congestion of movement between a dangerous place and a shelter place. This can be shown in the following numerical experiments: with the same parameter values, except for the size of node 2, we analyse the speed at which the people are emptying the place of the terrorist (node 1). We consider the following values of W_2 : 50, 100, 200, 1000 (Fig. 5.17).

In the case $W_2 = 1000$ (black curves), the flight of the entire population from place 1 to place 2 (which can be a small square) and then to place 3 is very fast; it lasts less than 10 min (Fig. 5.17c, f, i), because place 3 can foster the entire population. There is a massive influx of panicked people (Fig. 5.17d), which is greater than the controlled one (Fig. 5.17e), into place 2, which empties very quickly as the majority of controlled people reach the safe shelter (Fig. 5.17h). However, it can be noted that a significant number of the panicked population remains in the refuge place, and this number is only slowly decreased (see bump (Fig. 5.17g)).

This is explained by the fact that the flight dynamics are not hindered by obstacles or bottlenecks, and the fleeing populations have not enough time to change their behavioural state.

On the other hand, if $W_2 = 50$ (red curves), the evacuation of the total population from place 1 to place 2 and then place 3 is much slower (about 25 min instead of less

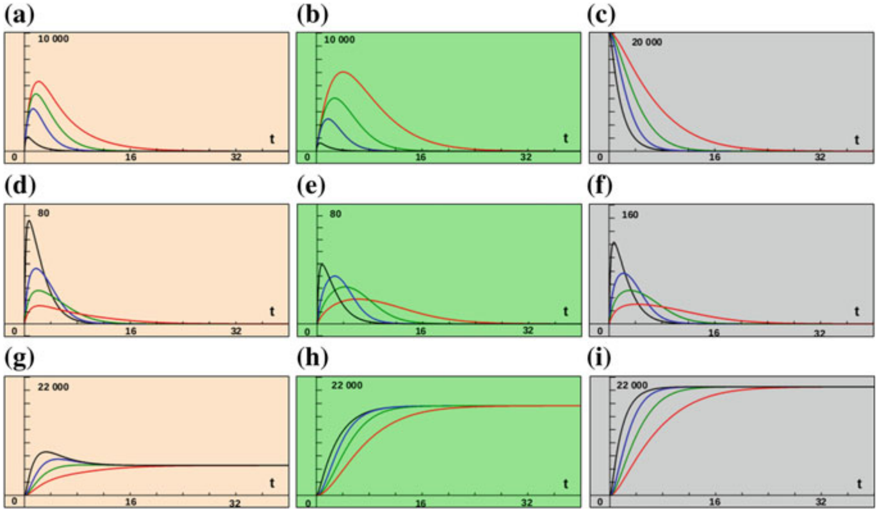


Fig. 5.17 Convergence towards the fixed point (0, 0, 0, 0, 0, 0, 0, 15,000, 5000) for the values of parameters $B_1 = 0.2$, $B_2 = 0.4$, $C_1 = 0.3$, $C_2 = 0.1$, $\eta_{c,1,2} = \eta_{c,2,3} = \eta_{p,1,2} = \eta_{p,2,3} = 0.005$, $V_{1,0} = 20{,}000$, $V_{2,0} = 0$, $V_{3,0} = 0$, $W_1 > V_1$, $W_3 = 20{,}500$ and the values of $W_2 = 50$ (red curves), $W_2 = 100$ (green curves), $W_2 = 200$ (blue curves), $W_2 = 1000$ (black curves)

than 10 min, (Fig. 5.17c, f, i)). The panicked population has time to calm down in place 1, because there is no new attack (it has been assumed that there is no domino effect, $s_1 = s_2 = 0$). It is thus a population that is mainly in a state of reasoned behaviour that arrives in place 3 (Fig. 5.17g, h).

It can be highlighted from this first analysis that the faster the speed of change of location (hence decreasing the vulnerability of populations), the faster this speed leads to significant flows of panic in both places 2 and 3. There is a paradox here: the fast self-safety movement of populations leads to situations of collective panic that are more difficult to manage. This fact must be taken into account by emergency services and emergency physicians.

Although the deaths that occur are not included in this CWPCR model version, one can imagine that the escape of panicked populations would induce more victims.

As said before, the simulation results shed light on the importance of the size of the intermediate places and their role in the fluidity or, to the contrary, on the congestion of movements between a dangerous place and a shelter place.

Depending on their respective size, intermediate places will modulate the dynamics and the speed of flow of the crowds. In this sense, they become strategic places both for the planners, who must think about the organization of the area to host public events, and also for the terrorists who can use these strategic places to multiply the effect of their harmful actions by trapping the flight movements between two areas of action.

This can be summarized in both Figs. 5.18 and 5.19, where the duration of evacuation time for 80% of people from place 1 is displayed according to two variables: the capacity of place 2 versus that of place 3. The arrow goes from long durations (in warm colours) to short durations (in cold colours). This representation is similar to a heat map. It is an easy way to identify a 'hot spot' (i.e. a configuration with an excessively long evacuation time) or, to the contrary, a more comfortable configuration with an acceptable evacuation duration time, which would save more lives.

In Fig. 5.18, the parameters that model the topological characteristics of the street linking node 1 to node 2, and node 2 to node 3, are set to $\eta_{c,1,2} = \eta_{c,2,3} = \eta_{p,1,2} = \eta_{p,2,3} = 0.0025$.

In contrast, in Fig. 5.19 their values are doubled $\eta_{c,1,2} = \eta_{c,2,3} = \eta_{p,1,2} = \eta_{p,2,3} = 0.005$, to model larger streets. In this case, shorter evacuation times are obtained.

One can see the necessary configurations that are required for a given evacuation time in Figs. 5.18 and 5.19. As an example, if we consider a duration between 20 and 25 min, in Fig. 5.18; this duration can be obtained with a narrow intermediate place that can accommodate 40–50 people, only if the capacity of place 3 is greater than 30,000 people. Instead the same duration is possible with a smaller place 3 (with a capacity between 16,500 and 17,000 people) if the capacity of place 2 is higher (between 90 and 100 people). That means that if planners who organize sites to host public events or festivals cannot enlarge the intermediate place (for example, due to the shape of a historic city center), they must establish a larger final evacuation shelter.

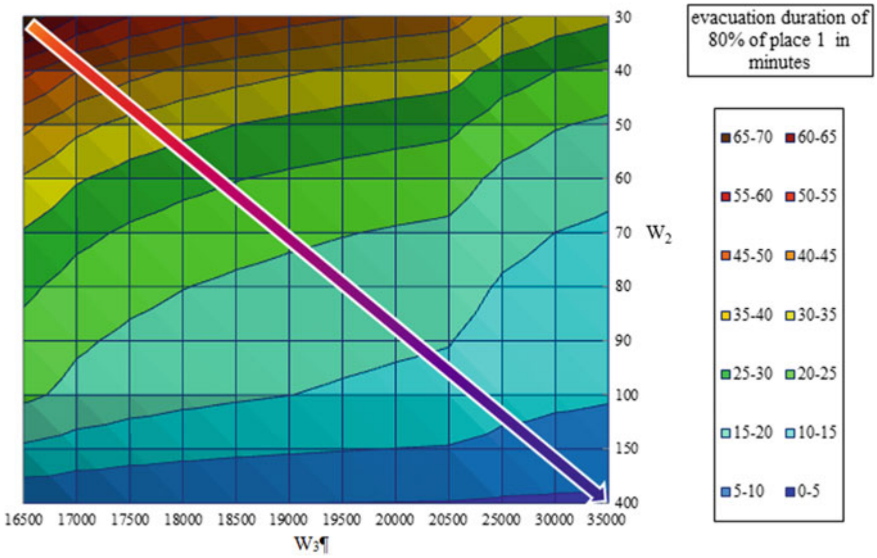


Fig. 5.18 Time required to evacuate 80% of population from place 1 according to W_2 and W_3 with $\eta = 0.0025$ and $V_{1,0} = 20{,}000$

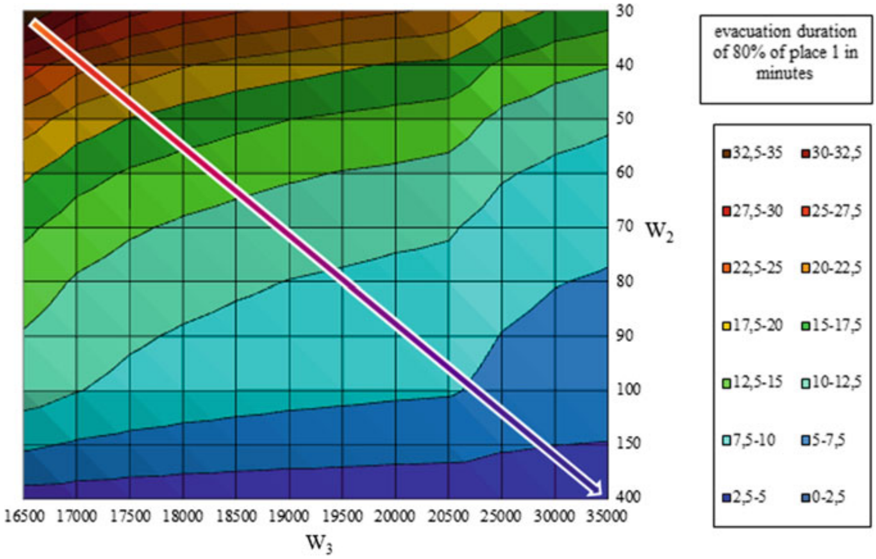


Fig. 5.19 Time required to evacuate 80% of population from place 1 according to W_2 and W_3 with $\eta = 0.005$ and $V_{1,0} = 20{,}000$

5.5 Conclusion

In this article, we have developed a new model of weighted human behaviour coupled on street and place networks, in the context of an urban terrorist attack, thus improving the PCR model [3, 4] with bottleneck coupling and by taking into account the capacity of every place and the number of people stranded in these places. The simulation results in a simple network with three nodes (places or public squares) and two edges (streets) that demonstrate the key role of the capacity of an intermediate place in the dynamics of evacuation from dangerous to safe places. This model is presented with a user-friendly graphical representation, which allows planners to accurately consider where to host public events in a specific territorial context.

Acknowledgements This work has been supported by the French government, through (i) the National Research Agency (ANR) under the Societal Challenge 9 'Freedom and security of Europe, its citizens and residents' with the reference number ANR-17-CE39-0008, co-financed by French Defence Procurement Agency (DGA) and The General Secretariat for Defence and National Security (SGDSN), and (ii) the IDEX UCA$^{\text{JEDI}}$ Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-15-IDEX-01. The authors thank gratefully Abby Cuttris in polishing the English.

References

1. Ph. Marchesin, *Introduction aux relations internationales* (Karthala, coll., Hommes et Sociétés, Paris, 2008)
2. J.-F. Colonna, Expériences virtuelles et virtualités expérimentales. Réseaux **11**(61) (1993)
3. D. Provitolo, E. Dubos-Paillard, N. Verdiere, V. Lanza, R. Charrier, C. Bertelle, M.A. Aziz-Alaoui, Les comportments humains en situation de catastrophe: de I'observation à la modélisation conceptuelle et mathematique. Cybergeo: Eur. J. Geogr. **735**, (2015), 23 p
4. N. Verdière, V. Lanza, R. Charrier, D. Provitolo, E. Dubos-Paillard, C. Bertelle, M.A. Aziz-Alaoui, Mathematical modeling of human behaviours during catastrophic events, in *ICCSA14, 23–26 June, Le Havre* (2014), 8 p
5. N. Verdière, G. Cantin, D. Provitolo, V. Lanza, E. Dubos-Paillard, R. Charrier, M.A. Aziz-Alaoui, C. Bertelle, Understanding and simulation of human behaviours in areas affected by disasters: from the observation to the conception of a mathematical model. Glob. J. Hum. Soc. Sci.: H Interdiscip. **15**(10), Version 1.0 (2015), 10 p
6. G. Cantin, N. Verdière, V. Lanza, M.A. Aziz-Alaoui, R. Charrier, C. Bertelle, D. Provitolo, E. Dubos-Paillard, Mathematical modeling of human behaviours during catastrophic events: stability and bifurcations. Int. J. Bifurc. Chaos **26**(10), 1630025 (2016), (20 pp.). https://doi.org/10.1142/S0218127416300251
7. J.D. Murray, *Mathematical Biology I: An Introduction* (Springer, New York, 2002)
8. H. Laborit, *La légende des comportements* (Flammarion, Paris, 1994)
9. N. George, L. Gamond, Premières impressions, L'essentiel Cerveau et Psycho: Les émotions au Pouvoir, No. 7 (2011)
10. R. Noto, P. Huguenard, A. Larcan, *Médecine de catastrophe* (Masson, Paris, 1994)
11. R. Soussignan, Un monde d'émotions, L'essentiel Cerveau et Psycho: Les émotions au pouvoir, No. 7 (2011)
12. T. Brosch, D. Sander, Les effects cognitifs des émotions, L'essentiel Cerveau et Psycho: Les émotions au pouvoir, No. 7 (2011)

13. L. Crocq, *Les paniques collectives* (Odile Jacob, Paris, 2013), 380 p
14. E. Hatfield, J.T. Cacioppo, R.L. Rapson, *Emotional Contagion* (Cambridge University Press, Cambridge, 1994)
15. D. Provitolo, Un exemple d'effets de dominos: la panique dans les catastrophes urbaines. Cybergéo: Revue européenne de géographie **328** (2005), 19 p, http://www.cybergeo.eu/index2998.html
16. L.D. Devi Vanumu, K.R. Rao, G. Tiwari, Fundamental diagrams of pedestrian flow characteristics: a review. Eur. Transp. Res. Rev. **9**(49), 49 (2017). https://doi.org/10.1007/s12544-017-0264-6
17. W. Daamen, S.P. Hoogendoorn, P.H.L. Bovy, First-order pedestrian traffic flow. Theory. Transp. Res. Rec. J. Transp. Res. Board **1934**, 43–52 (2005)
18. C. Dias, M. Sarvi, N. Shiwakoti, O. Ejtemai, M. Burd, Investigating collective escape behaviours in complex situations. Saf. Sci. **60**, 87–94 (2013)
19. T. Kretz, A. Grünebohm, M. Schreckenberg, Experimental study of pedestrian flow through a bottleneck. J. Stat. Mech.: Theory Exp. **2006**, P10014 (2006). https://doi.org/10.1088/1742-5468/2006/10/P10014
20. W. Liao, A. Seyfried, J. Zhang, M. Boltes, X. Zheng, Y. Zhao, Experimental study on pedestrian flow through wide bottleneck (The Conference on Pedestrian and Evacuation Dynamics 2014 (PED2014)). Transp. Res. Procedia **2**, 26–33 (2014)
21. A. Seyfried, B. Steffen, W. Klingsch, M. Boltes, The fundamental diagram of pedestrian movement revisited. J. Stat. Mech.: Theory Exp. **10**(10) (2005). https://doi.org/10.1088/1742-5468/2005/10/P10002
22. X.L. Zhang, W.G. Weng, H.Y. Yuan, J.G. Chen, Empirical study of a unidirectional dense crowd during a real mass event. Phys. A **392**, 2781–2791 (2013)
23. J.-B. Zhou, H. Chen, J. Yang, J. Yan, Pedestrian evacuation time model for urban metro hubs based on multiple video sequences data. Math. Probl. Eng. **2014**, Article ID 843096 (2014), 11 pp. https://doi.org/10.1155/2014/843096

Chapter 6
A New Model for Transient Flow in Gas Transportation Networks

Martin Gugat and Michael Herty

Abstract We consider the flow of gas through networks of pipelines. A hierarchy of models for the gas flow is available. The most accurate model is the pde system given by the 1-d Euler equations. For large-scale optimization problems, simplifications of this model are necessary. Here we propose a new model that is derived for high-pressure flows that are close to stationary flows. For such flows, we can make the assumption of constant gas velocity. Under this assumption, we obtain a model that allows transient gas flow rates and pressures. The model is given by a pde system, but in contrast to the Euler equations, it consists of linear equations. Based upon this model, the fast computation of transient large-scale gas network states is possible.

Keywords Optimal control · Gas network · Compressor · Flow rate · Pressure · Pressure drop · Pde constrained optimization · Feasibility

AMS Subject Classification 49J20 · 49M37

6.1 Introduction

The most accurate model for the gas flow in pipeline networks is the system of partial differential equations given by the 1-d Euler equations, see [1]. In practice, often networks with a large number of pipes occur. The problem of optimal control of the flow through such networks is a large-scale optimization problem. For the computational solution of this problem, simplifications of the model are necessary. Here we propose a new model that allows the computation of optimal controls on

M. Gugat (✉)
Lehrstuhl 2 für Angewandte Mathematik, University of Erlangen-Nuremberg,
Cauerstr. 11, 91058 Erlangen, Germany
e-mail: gugat@math.fau.de

M. Herty
RWTH Aachen University, Templergraben 55, 52056 Aachen, Germany
e-mail: herty@mathc.rwth-aachen.de

P. Manchanda et al. (eds.), *Mathematical Modelling, Optimization, Analytic and Numerical Solutions*, Industrial and Applied Mathematics,
https://doi.org/10.1007/978-981-15-0928-5_6

large graphs. The equations are derived under the assumption of constant gas velocity. The model allows transient gas flow rates and pressures, in particular at the points where the flow is controlled, namely, the boundary nodes and the compressors. This model allows the fast computation of the transient flows for large-scale gas network optimization problems. The method is based upon the ideas given in [3]. Similar as in [3], we derive an explicit representation of the system state that allows the computation of the flow rates in the edges and the time derivatives of the pressure at the nodes of the network graphs at a given point using only a finite number of algebraic operations.

A review of optimization problems in natural gas transportation systems is given in [10]. The optimal control of transient flow in gas networks based upon discretization with the method of lines has been discussed in [11]. An overview of challenges in optimal control problems for gas and fluid flow in networks of pipes and canals is given in [7].

This paper has the following structure. In Sect. 6.2, the pde model is introduced. Our model consists of a transport equation and an ordinary differential equation that describes the decay of the pressure along the pipes. Then conditions that describe the flow through the nodes of the network are introduced. The conditions require the conservation of mass and the continuity of the pressure in the nodes. In Sect. 6.3, the well-posedness of the system is analyzed. We represent the states using a vector of functions describing the flow rates along the pipes and a second vector of functions for the time derivatives of the pressures at the nodes of the network. In Sect. 6.4, a recursion for the system state is given that allows to derive an explicit representation.

6.2 The System

6.2.1 The Graph of the Network

Let a directed graph $G = (V, E)$ be given. The edges $e = (u, v) \in E$ of the graph correspond to pipes of length L^e that are modeled by intervals $[0, L^e]$. The edge $e = (u, v) \in E$ denotes the pipe that has its end zero at the vertex $u \in V$ and the end L^e at the vertex $v \in V$. For the flow rates of the gas through the pipe corresponding to the edge $e \in E$, we use the notation q^e $(e \in E)$. For $v \in V$, define $\sigma(v, e) = -1$ if the end zero of the pipe $[0, L^e]$ is located at the node $v \in V$ and $\sigma(v, e) = 1$ if the end L^e of the pipe corresponds to v.

For a node $v \in V$, let $E_0(v) \subset E$ denote the set of edges (pipes) that meet at the node v. Let $V_0 \subset V$ denote the set of nodes of the graph with at least two adjacent edges, that is, the interior nodes of the graph. Then for all $v \in V_0$, the stationary state satisfies the node conditions

$$\sum_{e \in E_0(v)} q^e \sigma(v, e) = 0, \tag{6.1}$$

which guarantees the conservation of the gas mass at the interior nodes.

The inflow into the network occurs at the boundary nodes $v \in V_\Gamma = V \setminus V_0$ with $\sigma(v, e) = -1$ where e is the unique adjacent edge. The boundary nodes where outflow occurs (consumer demand is satisfied) are in the set $\{v \in V_\Gamma : \sigma(v, e) = 1, e \in E_0(v)\}$.

Let D_e denote the diameter of pipe e. The gas transport velocity v^e for pipe e is given. We assume that the variations in time and space are small, so that we can consider v^e as a constant with respect to time.

Remark 6.1 It is well known that in the physical reality, the gas velocity is often not constant. In fact, the gas velocity that appears in the model is merely an average value of the velocity of the gas molecules in the sense of the kinetic theory of gases. Our assumption that the velocity of the gas is constant should be interpreted in a similar (average) sense. Gas pipelines are often operated in a neighborhood of a stationary state (see [6]). The corresponding velocity function can always be approximated with arbitrary precision by a piecewise constant function. We define the edges e of the graph in such a way that they correspond to the intervals that appear in the definition of such a piecewise constant function. Then in a sufficiently small neighborhood of the stationary state, the constant velocity values v^e that correspond to the approximation of the stationary state are a reasonable choice for the simplified model of the transient states that we derive below.

If v^e only depends on the space variable x, to travel through pipe e the gas needs the time

$$T^e = \frac{L^e}{\frac{1}{L^e} \int_0^{L^e} v^e(x)\, dx}.$$

In order to obtain an explicit solution of our state equation, we assume that there exists a real number $\delta > 0$ such that for all $e \in E$ there is a natural number k^e such that $T^e = k^e \delta$. In other words, we assume that for each pipe, the corresponding travel time is an integer multiple of the time δ. Note that each set of travel times $\{T_e : e \in E\}$ can be approximated with arbitrary high precision by times that satisfy this condition by making $\delta > 0$ sufficiently small. By inserting sufficiently many additional auxiliary nodes of degree two in the network, we can assume without restriction that for all $e \in E$ we have $T^e = \delta$. Moreover, by inserting sufficiently many additional auxiliary nodes of degree two in the network, we can also make all the pipes sufficiently short such that we can assume that v^e is a constant along each of these short pipes, that is, it does not depend on the time or the space variable.

6.2.2 The Partial Differential Equations

Let $p^e(t, x)$ denote the gas pressure in pipe $e \in E$ at the time t at the point $x \in [0, L^e]$. Let $f_g^e \geq 0$ denote the friction parameter and $D^e > 0$ the diameter for pipe e. Let α^e denote the angle of inclination of the pipe e and g the gravitational constant. To model the gas flow in a pipe $e \in E$ of the network, we use the partial differential equations

$$\partial_t q^e(t,x) + v^e \partial_x q^e(t,x) = 0, \tag{6.2}$$

$$\partial_x p^e(t,x) = -\frac{f_g^e}{2\,D^e}\,|v^e|\,q^e(t,x) - g\,\sin(\alpha^e)\,\frac{q^e(t,x)}{v^e}. \tag{6.3}$$

The first equation is a transport equation, and the second equation is used to compute the pressure drop along the pipe.

The first equation follows from the continuity equation

$$\partial_t\left(\frac{q^e}{v^e}\right) + \partial_x q = 0$$

under the assumption that the velocity $v^e > 0$ is constant. Under this assumption, the second equation follows from

$$\partial_t q^e + \partial_x\left(v^e q^e + p^e\right) = -\frac{f_g^e}{2D^e} q^e |v^e| - g\,\sin(\alpha^e)\,\frac{q^e(t,x)}{v^e}.$$

If the time derivative $\partial_t p$ exists, Eq. (6.3) implies

$$\partial_t p^e(t,L^e) - \partial_t p^e(t,0) = \left[\frac{f_g^e\,v^e\,|v^e|}{2\,D^e} + g\,\sin(\alpha^e)\right]\left[q^e(t,L^e) - q^e(t,0)\right]. \tag{6.4}$$

To compute the system state, we will work with (6.4) to compute the values of the time derivative of the pressure at the nodes, that is, $\partial_t p^e(t,0)$ and $\partial_t p^e(t,L^e)$ for each pipe $e \in E$.

The system (6.2), (6.3) is similar to the parabolic system considered in [8], but in contrast to [8] only first-order derivatives appear.

6.2.3 Node Conditions for the Pressure

In this section, we state the conditions that govern the behavior of the pressure in the nodes of the network. These conditions have already been considered, for example, in [1, 4]. The model is similar to the model for the flow in open channel networks presented in [9]. The stationary states in gas networks that satisfy the node conditions and the isothermal Euler equations for ideal gas are studied in [5].

First we consider an interior node $v \in V_0$.

Let $E_0^{in}(v) = \{e \in E_0(v) :\ \sigma(v,e) = 1\}$ denote the pipes where the end L^e is at the node v and $E_0^{out}(v) = \{e \in E_0(v) :\ \sigma(v,e) = -1\}$ denote the pipes where the end zero is at the node v. We assume that the pressure in the node is governed by the relation

$$p^d(t,0) = p^e(t,0) = p^f(t,L^f) = p^g(t,L^g) \tag{6.5}$$

for all $d,\ e \in E_0^{out}(v),\ f,\ g \in E_0^{in}(v)$. Due to (6.5), the value of the pressure at each node $v \in V$ is well defined, so we can introduce the notation

$$p^v(t) = p^e(t,0) = p^f(t,\ L^f) \tag{6.6}$$

for all $e \in E_0^{out}(v),\ f \in E_0^{in}(v)$. In order to obtain a complete system, we also need values for the boundary nodes where inflow into the network occurs. These are the nodes $v \in V_\Gamma = V \backslash V_0$ with $\sigma(v,e) = -1$. Here for $e \in E_0(v)$, we have the equation

$$q^e(t,0) = q^v(t). \tag{6.7}$$

At the other boundary nodes $v \in V_\Gamma = V \backslash V_0$ with $\sigma(v,e) = 1$ for $e \in E_0(v)$, we have the boundary conditions

$$p^e(t,L^e) = p^v(t). \tag{6.8}$$

We assume that for the boundary nodes $v \in V_\Gamma$, the input functions p^v, q^v, respectively, are given.

6.2.4 Initial Conditions for the Flow Rate and the Pressure

We assume that the initial flow rate is given by the conditions

$$q^e(0,x) = q_0^e(x), \quad x \in [0,L^e], \quad e \in E \tag{6.9}$$

with $q_0^e \in L^2(0,\ L^e)$ and that it is compatible with (6.1). Moreover, we assume that the initial pressure

$$p^e(0,L^e) = p_0^e$$

is also given in a way that is compatible with the ordinary differential Eqs. (6.3), (6.5) and q_0^e. Note that this is also an assumption on the values q_0^e, which must allow the existence of compatible pressure values.

6.3 Well-Posedness of the System

In this section, we present a representation of the system state for given initial data q_0^e compatible with the node conditions and input functions p^v, q^v. Our representation is based upon the functions α^e $(e \in E)$, β^v $(v \in V)$ that are defined in the following lemma.

Lemma 6.1 *Let a time $T > 0$ be given. Then there exists a unique set of functions $(\alpha^e)_{e\in E}$, $(\beta^v)_{v\in V}$ with $\alpha^e \in L^2(-\delta,\ T)$, $\beta^v \in L^2(0,\ T)$ that satisfy*

1. $\alpha^e(s) = q_0^e(-\nu^e s)$ *for all* $s \in [-\delta, 0)$.
2. *For all* $v \in V_\Gamma$ *with* $\sigma(v, e) = -1$, $e \in E_0(v)$ *and for all* $s \in [0,\ T]$,

$$\alpha^e(s) = q^v(s).$$

3. *For all* $v \in V_\Gamma$ *with* $\sigma(v, e) = 1$ *for* $e \in E_0(v)$ *and for all* $s \in [0,\ T]$,

$$\beta^v(s) = \partial_t p^v(s).$$

4. *For all* $t \in [0, T]$, $e = (v, w) \in E$, *that is* $e \in E_0^{out}(v)$ *and* $e \in E_0^{in}(w)$ *we have*

$$\beta^w(t) = \beta^v(t) + \left[\frac{f_g^e\ \nu^e\ |\nu^e|}{2\,D^e} + g\ \sin(\alpha^e)\right]\left[\alpha^e(t-\delta) - \alpha^e(t)\right].$$

5. *For all* $v \in V_0$ *and for all* $t \in [0, T]$

$$\sum_{e\in E_0^{in}(v)} \alpha^e(t-\delta) = \sum_{f\in E_0^{out}(v)} \alpha^f(t).$$

Proof For $t \in [-\delta, 0)$ and $e \in E$ the functions α^e are defined from the initial state q_0^e by *1*. Define the constants

$$\eta^e = \frac{f_g^e\ \nu^e\ |\nu^e|}{2\,D^e} + g\ \sin(\alpha^e).$$

At the boundary nodes $v \in V_\Gamma = V\backslash V_0$ with $\sigma(v, e) = 1$, the values of $\beta^v(t)$ are given by *3* for $t \in [0, T]$.

In order to define $\beta^v(t)$ for the interior nodes $v \in V_0$ and $t \in [0, \delta)$, we use the equations

$$\beta^v(t) = \frac{1}{\sum_{f\in E_0^{out}(v)} \frac{1}{\eta^f}}\cdot \tag{6.10}$$

$$\left[\sum_{f=(v,w)\in V} \frac{1}{\eta^f}\beta^w(t) + \left(\sum_{e\in E_0^{in}(v)} \alpha^e(t-\delta)\right) - \left(\sum_{f\in E_0^{out}(v)} \alpha^f(t-\delta)\right)\right].$$

We start from the nodes that have maximal distance one from the outflow nodes $v \in V_\Gamma = V\backslash V_0$ with $\sigma(v, e) = 1$ and then compute the values of β^v recursively by increasing the maximal distance from the outflow boundary nodes one by one.

Finally, we consider the boundary nodes $v \in V_\Gamma = V\backslash V_0$ with $\sigma(v, e) = -1$. Let $e = (u, v)$ be the adjacent edge. Then by *4* for $t \in [0, \delta)$ we have

$$\beta^v(t) = \beta^u(t) + \eta^e \left[\alpha^e(t-\delta) - \alpha^e(t)\right],$$

and the values of α^e are given by the boundary condition *2*.

In this way, we obtain $(\alpha^e(t))_{e\in E} \in (L^2(-\delta, 0))^E$ and $(\beta^v(t))_{v\in V} \in (L^2(0, \delta))^V$.

The values of α^e and $\beta^v(t)$ on the following intervals $[0, \delta)$, $[\delta, 2\delta)$, $[2\delta, 3\delta)$, respectively, are now obtained recursively in a similar way.

First, for $e = (v, w) \in E$, the values of α^e for $t \in [0, \delta)$ are computed by the equation

$$\alpha^e(t) = \alpha^e(t-\delta) + \frac{1}{\eta^e}\left[\beta^v(t) - \beta^w(t)\right] \tag{6.11}$$

that follows from *4*. If $e = (u, v)$ with $v \in V_\Gamma = V\backslash V_0$ with $\sigma(v, e) = -1$, this does not lead to a contradiction with *2* on account of the definition of $\beta^v(t)$ in the previous step of the recursion.

Now the values of $\beta^v(t)$ for $t \in [\delta, 2\delta)$ can be computed as in the previous step of the recursion, using the values of α^e on $[0, \delta)$.

With the functions α^e and β^v, we can give the explicit representation of the system state that is given in the following theorem.

Theorem 6.1 *Let a time $T > 0$ be given. For $v \in V$, let the boundary data $q_v(t)$, $\partial_t p_v(t)$, respectively, in $L^2(0, T)$, be given. Then the system governed by the initial condition (6.9), the node conditions (6.1), (6.5) for all $v \in V_0$, the boundary conditions (6.7), (6.8), respectively, for $v \in V_\Gamma$ and the partial differential Eqs. (6.2) and (6.4) for all $e \in E$ has a solution that is given by functions*

$$q^e(t, x) = \alpha^e\left(t - \frac{x}{\nu^e}\right) \tag{6.12}$$

and

$$\partial_t\, p^v(t) = \beta^v(t) \tag{6.13}$$

with α^e and β^v as in Lemma 6.1. Thus, we have for $e \in E_0^{in}(v)$

$$p^v(t) = p_0^e + \int_0^t \beta^v(s)\, ds. \tag{6.14}$$

The pressure values $p^e(t, x)$ outside the nodes can be obtained by integrating (6.3) along the edges $e \in E$ starting from a node. The solution satisfies the partial differential Eq. (6.2) in the sense of distributions.

Proof First, we check that the initial condition (6.9) holds. For $t = 0$, (6.12) yields $q^e(0, x) = \alpha^e(-x/\nu^e)$. By point *1* from Lemma 6.1, this implies $q^e(0, x) = q_0^e(x)$, that is, (6.9) holds.

The partial derivatives of q^e in the sense of distributions are given by $\partial_t q^e(t, x) = \alpha^{e\prime}(t - \frac{x}{\nu^e})$, $\partial_x q^e(t, x) = -\frac{1}{\nu^e}\alpha^{e\prime}(t - \frac{x}{\nu^e})$ where $\alpha^{e\prime}$ denotes the derivative of α^e.

Thus, the transport equation from (6.2) holds for all $e \in E$. Now we check that (6.4) also holds for all $e \in E$. By point *4* from Lemma 6.1, Definition (6.13) of the time derivative of the pressure p^v at the node $v \in V$ implies that (6.4) holds for $t > 0$.

Now we check that the node conditions (6.1) are satisfied. Point *5* in Lemma 6.1 yields

$$\sum_{e \in E_0^{in}(v)} q^e(t, L^e) = \sum_{f \in E_0^{out}(v)} q^f(t, 0)$$

that is (6.1) holds.

Point *2* in Lemma 6.1 implies that the boundary condition (6.7) holds.

Point *3* in Lemma 6.1 implies that the boundary condition (6.8) holds.

The node condition (6.5) holds since the pressure is only defined at the node $v \in V$ by (6.14) first, so (6.6) holds. Moreover since (6.4) holds, integrating (6.3) along the edges to obtain the values of the pressure in the edges $e \in E$ is compatible with (6.6).

6.4 A Recursion for the System State

Theorem 6.1 implies that the functions α^e and β^v that can be used to represent our system state by (6.12) and (6.13) satisfy an affine linear recursion for all $\tau \in [0, \delta)$. For $\tau \in [0, \delta)$, $e \in E$, $v \in V$ and $j \in \{0, 1, 2, 3, \ldots\}$ with $\tau + (j-1)\,\delta \le T$ define

$$\alpha_e^{(j)}(\tau) = \alpha^e\left(\tau + (j-1)\,\delta\right), \;\; \beta_v^{(j)}(\tau) = \beta^v\left(\tau + j\,\delta\right). \tag{6.15}$$

Then the functions $\alpha^{(j)} = \left(\alpha_e^{(j)}\right)_{e \in E}$, $\beta^{(j)} = \left(\beta_v^{(j)}\right)_{v \in V}$ are componentwise in $L^2(0, \delta)$ that is each component is in $L^2(0, \delta)$. Now point *1* from Lemma 6.1 implies

$$\alpha^{(0)}(\tau) = \left(\alpha^e(\tau - \delta)\right)_{e \in E} = (q_0^e(\nu^e\,(\delta - \tau)))_{e \in E}. \tag{6.16}$$

Now the values of $\beta^{(0)}(\tau)$ can be computed. We start with *3* that yields the values at the boundary nodes $v \in V_\Gamma$ with $\sigma(v,\ e) = 1$ from the boundary data $\partial_t p^v$. Going through paths to the interior nodes $v \in V_0$, the linear Eq. (6.10) yields the values at the adjacent nodes. Finally at the boundary nodes $v \in V_\gamma$ with $\sigma(v,\ e) = -1$, we can use *4* from Lemma 6.1, where the values of $\alpha^e(t)$ in for $t \in (0,\ \delta)$ at are obtained from the boundary condition *2*. Now

$$\alpha^{(1)}(\tau) = \left(\alpha^e(\tau)\right)_{e \in E}, \;\; \tau \in (0,\ \delta)$$

can be computed using Eq. (6.11) that is linear in the right-hand side. Now the values of $\beta^{(1)}(\tau)$ can be computed.

By repeating the construction we obtain for $j \in \{0, 1, 2, 3, \ldots\}$ a linear recursion of the form

$$\begin{pmatrix} \alpha^{(j+1)}(\tau) \\ \beta^{(j+1)}(\tau) \end{pmatrix} = A \begin{pmatrix} \alpha^{(j)}(\tau) \\ \beta^{(j)}(\tau) \end{pmatrix} + B u^{(j)}(\tau), \tag{6.17}$$

where $u^{(j)}(\tau) = \begin{pmatrix} (q^v\,(\tau + j\,\delta))_{v \in V_\Gamma, \sigma(v,\,e)=-1} \\ (\partial_t p^v\,(\tau + j\,\delta))_{v \in V_\Gamma, \sigma(v,\,e)=1} \end{pmatrix}$ contains the given boundary data. Here A denotes a linear map (that can be represented by a suitable time-independent matrix). Also B denotes a linear map, also represented by a matrix that is time-independent due to the definition of $u^{(j+1)}$. The first term in the sum provides the influence of the values of α from the past and the second term contains the influence of the inflow $u^{(j+1)}$. The matrices A and B contain the information on the structure of the graph G and the velocities on the edges.

By an induction argument (6.17) implies the following lemma.

Lemma 6.2 *For $\tau \in [0, \delta)$ and $j \in \{1, 2, 3, \ldots\}$ with $\tau + (j-1)\delta \leq T$ the function*

$$z^{(j+1)}(\tau) = \begin{pmatrix} \alpha^{(j+1)}(\tau) \\ \beta^{(j+1)}(\tau) \end{pmatrix}$$

that defines the system state on the network as in (6.12) and (6.13) can be computed using the following representation:

$$z^{(j)}(\tau) = \sum_{k=1}^{j} A^{j-k} B u^{(k-1)}(\tau) + A^j\, z^{(0)}(\tau). \tag{6.18}$$

6.5 Conclusion

We have presented a model for the flow in gas distribution networks. This model allows a fast and reliable computation of the state also for transient flow locally around a given stationary state. The model allows an explicit representation of the flow rates in the edges and the time derivatives of the pressure at the nodes of the graph as a linear function of the initial data and the boundary data that can be evaluated almost everywhere using a finite number of algebraic operations. For the solution of optimal control problems, the evaluation of gradients of the objective function is useful. Our model allows the derivation of an adjoint calculus similar as in [2].

Acknowledgements This work was supported by DFG grant SFB TRR 154, project C03. This paper took benefit from discussions at the Basque Center for Applied Mathematics.

References

1. M.K. Banda, M. Herty, A. Klar, Coupling conditions for gas networks governed by the isothermal Euler equations. Netw. Heterog. Media **1**, 295–314 (2006)
2. M. Gugat, Optimal nodal control of networked hyperbolic systems: evaluation of derivatives. Adv. Model. Optim. **7**, 9–37 (2005)
3. M. Gugat, Contamination source determination in water distribution networks. SIAM J. Appl. Math. **72**, 1772–1791 (2012)
4. M. Gugat, D. Wintergerst, Transient flow in gas networks: traveling waves. Int. J. Appl. Math. Comput. Sci. **28**, 341–348 (2018)
5. M. Gugat, F. Hante, M. Hirsch-Dick, G. Leugering, Stationary states in gas networks. NHM **10**, 295–320 (2015)
6. M. Gugat, D. Wintergerst, R. Schultz, Networks of pipelines for gas with nonconstant compressibility factor: stationary states. Comput. Appl. Math. **37**, 1066–1097 (2018)
7. F.M. Hante et al., Challenges in optimal control problems for gas and fluid flow in networks of pipes and canals: from modeling to industrial applications, in *Industrial Mathematics and Complex Systems*. Springer INdAM Series, ed. by P. Manchanda et al., to appear 2017
8. G. Leugering, G. Mophou, Instantaneous optimal control of friction dominated flow in a gas-network, DFG-AIMS-Workshop, in Mbour, Senegal, 13–16 March 2017, Birkhäuser, Basel (2017)
9. G. Leugering, E.J.P.G. Schmidt, On the modelling and stabilization of flows in networks of open canals. SIAM J. Control Optim. **41**, 164–180 (2002)
10. R.Z. Rios-Mercadoa, C. Borraz-Sanchez, Optimization problems in natural gas transportation systems: a state-of-the-art review. Appl. Energy **147**, 536–555 (2015)
11. A. Zlotnik, M. Chertkov, S. Backhaus, Optimal control of transient flow in natural gas networks, in *IEEE 54th Annual Conference on Decision and Control (CDC), Osaka, Japan*, 15–18 December 2015

Chapter 7
Mixed-Integer Optimal Control for PDEs: Relaxation via Differential Inclusions and Applications to Gas Network Optimization

Falk M. Hante

Abstract We show that mixed-integer control problems for evolution type partial differential equations can be regarded as operator differential inclusions. This yields a relaxation result including a characterization of the optimal value for mixed-integer optimal control problems with control constraints. The theory is related to partial outer convexification and sum-up rounding methods. The results are applied to optimal valve switching control for gas pipeline operations. A numerical example illustrates the approach.

Keywords Mixed-integer control · Partial differential equations · Switching control · Optimization · Differential inclusions · Relaxation · Gas networks

7.1 Introduction

The limitation to a finite number of possible control actions can be an important aspect in the context of optimal control. Such integer restrictions occur, for example, in autonomous driving in case of vehicles with gear shift power units [16], in contact problems such as robotic multi-arm transport [3] or in the operation of gas pipeline and water canal networks with valve switching [11]. Motivated by the latter, we address such mixed-integer optimal control problems involving evolution type partial differential equations (PDEs). This problem class includes in particular, not only optimal control of switched systems [31, 32] but also optimization of systems with coordinated activation of multiple actuators, for example, at different locations in space for certain distributed parameter systems [13, 15].

We show that these problems can be regarded as operator differential inclusions with set-valued but non-convex right-hand sides. A relaxation result based on extensions of the Filippov–Ważewski Theorem to operator differential inclusions relates

F. M. Hante (✉)
Lehrstuhl für Angewandte Mathematik 2, Department Mathematik,
Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
e-mail: falk.hante@fau.de

P. Manchanda et al. (eds.), *Mathematical Modelling, Optimization, Analytic and Numerical Solutions*, Industrial and Applied Mathematics,
https://doi.org/10.1007/978-981-15-0928-5_7

the non-convex problem to a convexified one. This yields a useful characterization of the value function for such mixed-integer optimal control problems. Multiplier representations of the convexified problem are closely related to a reformulation known as partial outer convexification [26, 27]. In this paper, we show that this approach extends to a rather general form of constraints on the controls of a form previously considered in [9]. It provides a convenient way to obtain near-optimal solutions, e.g., using sum-up rounding techniques as in [12, 13, 28] if the constraints are not imposed on the integer part.

Further, we show how this approach can be applied to optimize the operation of gas pipeline networks in nonstationary situations using valve switching as a control. While many stationary situations can be handled with algebraic models using tailored mixed-integer programming techniques [17], dynamic optimization remains a challenge [11] both in control theory and in industrial practice. Based on a recent result in [24], we show that the dynamic valve switching problem can be cast in the above framework. A numerical example is given to illustrate the performance of the relaxation technique combined with sum-up rounding.

The article is organized as follows. In Sect. 7.2, we introduce the problem class under consideration. In Sect. 7.3, we show that the problem is equivalent to a differential inclusion. Further, we provide the relaxation result and, closely related to that, a characterization of the optimal value function for the original problem. In Sect. 7.4, we apply these results to gas network operation. This includes various modeling aspects, an abstract problem formulation based on semigroup theory as well as a numerical example. A conclusion for this article and future research directions are given in Sect. 7.5.

7.2 Setting

In this section, we introduce the class of optimal control problems considered in this article.

Let Y be a separable Banach space, U and V be two complete and separable metric spaces, and $f : [t_0, t_f] \times Y \times U \times V \to Y$. We consider the control system

$$\dot{y}(t) = Ay(t) + f(t, y(t), u(t), v(t)),\ \ t \in (t_0, t_f) \text{ a.e.}, \tag{7.1}$$

where $[t_0, t_f]$ is a finite time horizon with $t_0 < t_f$, $A : D(A) \to Y$ is a generator of a strongly continuous semigroup $\{T(t)\}_{t \geq 0}$ of bounded linear operators on Y, where $u : [t_0, t_f] \to U$ and $v[t_0, t_f] \to V$ are two independent measurable control functions. Throughout the paper we consider the Lebesgue measure. Our main concern will be the confinement that v shall only take values from a finite non-empty subset $\mathscr{V} \subset V$. Without loss of generality, we may identify $\mathscr{V}$ with a set of integers $\{0, 1, \ldots, N-1\}$ and, in analogy to mixed-integer programming, we refer to (7.1) as a *mixed-integer control system*, where u represents ordinary controls and v integer controls. We will denote the set of measurable ordinary control functions $u : [t_0, t_f] \to U$ by $U_{[t_0,t_f]}$ and the set of measurable integer control functions $v : [t_0, t_f] \to \mathscr{V}$ by $V_{[t_0,t_f]}$. By the assumed finiteness of $\mathscr{V}$, we actually have $V_{[t_0,t_f]} = L^\infty(t_0, t_f; \mathscr{V})$.

In addition to the inherent integer confinement, we consider optional control restrictions of the form

$$u(t) \in \mathscr{U}^v(t),\ t \in (t_0, t_f) \text{ a.e.,} \tag{7.2}$$

where, for all $v \in V_{[t_0,t_f]}$, $\mathscr{U}^v$ is a set-valued map $\mathscr{U}^v\ [t_0, t_f] \rightrightarrows U$ and we consider an initial condition

$$y(t_0) = y_0, \tag{7.3}$$

where y_0 is a given initial state in Y.

Definition 7.1 We say that $y : [t_0, t_f] \to Y$ is a *solution of the mixed-integer control system* if there exists an ordinary control $u \in U_{[t_0,t_f]}$ and an integer control $v \in V_{[t_0,t_f]}$ such that $y \in C([t_0, t_f]; Y)$, satisfies the integral equation

$$y(t) = T(t - t_0)y(t_0) + \int_{t_0}^{t} T(t - s) f(s, y(s), u(s), v(s))\,\mathrm{d}s,\ t \in [t_0, t_f] \tag{7.4}$$

and (7.2) and (7.3) hold. We denote by $\mathscr{S}_{[t_0,t_f]}$ the set of all such solutions $y = y(u, v)$ defined on $[t_0, t_f]$.

In conjunction with the mixed-integer control system, we consider a cost function $\Phi : C([t_0, t_f]; Y) \times U_{[t_0,t_f]} \times V_{[t_0,t_f]} \to \mathbb{R} \cup \{\infty\}$ and define the *mixed-integer optimal control problem* as

$$\text{minimize } \Phi(y, u, v) \text{ subject to } y = y(u, v) \in \mathscr{S}_{[t_0,t_f]}. \tag{7.5}$$

We will study the corresponding *optimal value* given by

$$\nu = \inf\{\Phi(y, u, v) : y = y(u, v) \in \mathscr{S}_{[t_0,t_f]}\} \in \mathbb{R} \cup \{\pm\infty\} \tag{7.6}$$

in its dependency on a parameter λ varying in an interval $I \subset \mathbb{R}$ and acting on the initial value y_0, on the control constraint $U^v(t)$ and the cost function Φ of the mixed-integer control problem.

7.3 Relaxation via Differential Inclusions

In this section, we show that, for the problem class of Sect. 7.2 and unlike in general mixed-integer programming, the optimal value function coincides with the optimal value function for a suitably defined relaxed problem.

For the control system, we assume that

(i) the map $(y, u) \mapsto f(t, y, u, v)$ is continuous for a. e. $t \in (t_0, t_f)$ and all $v \in \mathscr{V}$.

(ii) there exists a function $k \in L^1(t_0, t_f)$ such that

(a) for a. e. $t \in (t_0, t_f)$, for all $v \in V_{[t_0,t_f]}$, for all $u \in \mathscr{U}^v(t)$ the map $y \mapsto f(t, y, u, v)$ is $k(t)$-Lipschitz on Y

(b) for a. e. $t \in (t_0, t_f)$,

$$\sup_{v \in V_{[t_0,t_f]}} \sup_{u \in \mathscr{U}^v(t)} \|f(t, 0, u, v)\|_Y \leq k(t). \tag{7.7}$$

(iii) the map $t \mapsto f(t, y, u, v)$ is strongly measurable on $[t_0, t_f]$ for all $y \in Y$, $u \in U$, and $v \in \mathscr{V}$.

For the cost function, we assume that

(iv) the function $\Phi(y, u, v)$ consists of terminal and integral cost

$$\Phi(y, u, v) = \varphi(y(t_f)) + \int_{t_0}^{t_f} L(t, y(t), u(t), v(t))\, \mathrm{d}t \tag{7.8}$$

with a locally Lipschitz continuous function $\varphi : Y \to \mathbb{R}$ and $L : [t_0, t_f] \times X \times U \times V \to \mathbb{R}$ satisfying (i)–(iii) for L in place of f.
For the control constraints, we assume that

(v) the set-valued map $\mathscr{U}^v$ is measurable with closed, non-empty images for all $v \in V_{[t_0,t_f]}$ and, for a. e. $t \in [t_0, t_f]$, the set

$$\bigcup_{v \in V_{[t_0,t_f]}} \mathscr{U}^v(t) \tag{7.9}$$

is closed.

In particular, under these assumptions, the integral in (7.4) is well-defined in the Lebesgue–Bochner sense and from the theory of abstract Cauchy problems [22], we obtain a solution y in $C([0, t_f]; Y)$ for all $y_0 \in Y$, $u \in U_{[t_0,t_f]}$ and $v \in V_{[t_0,t_f]}$.

The main result below is based on rewriting the mixed-integer optimal control problem as an operator differential inclusion. We will recall the essential aspects from the theory of operator differential inclusions from [6, 7] and apply these to obtain a characterization of the optimal value for the original problem by means of a convexified problem.

Consider a set valued map $G(t, y) : [t_0, t_f] \times Y \rightrightarrows Y$ and the operator differential inclusion

$$\dot{y}(t) \in Ay(t) + G(t, y(t)) \text{ a.e. in } [t_0, t_f], \quad y(t_0) = y_0 \in Y. \tag{7.10}$$

We define a solution of (7.10) as in [7].

Definition 7.2 A function $y \in C([t_0, t_f]; Y)$ is called a *mild trajectory* of (7.10) if there exists a Bochner integrable selection $g \in L^1(t_0, t_f; Y)$ of the map $t \mapsto G(t, y(t))$ and

$$y(t) = T(t - t_0)y(t_0) + \int_{t_0}^{t} T(t - s)g(s)\,\mathrm{d}s,\ t \in [t_0, t_f], \quad y(t_0) = y_0. \tag{7.11}$$

We recall that a set-valued map $H : Y \rightrightarrows Y$ is called L-Lipschitz on $K \subset Y$ if for all $y \in K$, $H(y) \neq \emptyset$, and

$$H(y) \subset H(\tilde{y}) + L\|y - \tilde{y}\|B,\ y, \tilde{y} \in Y, \tag{7.12}$$

where B denotes the closed unit ball in Y.

We will also need a measurable selection theorem and the direct image theorem as stated in [6] and proved in [1]. For the convenience of the reader, we restate the results below.

Lemma 7.1 *Consider complete separable metric spaces X and Y, a Carathéodory map $H : [t_0, t_f] \times X \to Y$ and a measurable set-valued map $W : [t_0, t_f] \rightrightarrows X$ with closed non-empty images. Then for every measurable map $h : [t_0, t_f] \to Y$ satisfying*

$$h(t) \in H(t, W(t)) \text{ a.e. in } [t_0, t_f] \tag{7.13}$$

there exists a measurable selection $w(t) \in W(t)$ s.t.

$$h(t) = H(t, w(t)) \text{ for a.e. } t \in [t_0, t_f]. \tag{7.14}$$

Lemma 7.2 *Let X be a complete separable metric space and $U : [t_0, t_f] \rightrightarrows X$ be a measurable set-valued map with closed images. Consider a Carathéodory set-valued map $G : [t_0, t_f] \times X$ to a complete separable metric space Y. Then, the map*

$$t \rightrightarrows \overline{G(t, U(t))} \tag{7.15}$$

is measurable on $[t_0, t_f]$.

For $y \in Y$, we now consider the set-valued map

$$t \mapsto F(t, y) = \{f(t, y, \bar{u}, v(t)) : \bar{u} \in \mathscr{U}^{v}(t),\ v \in V_{[t_0, t_f]}\}, \quad t \in [t_0, t_f] \text{ a.e.} \tag{7.16}$$

Note that the images of F are unions of infinitely many sets and are thus, in general, not closed. Still, using the above theory, one can obtain an equivalent representation of $\mathscr{S}_{[t_0, t_f]}$.

Theorem 7.1 *We have the representation*

$$\mathscr{S}_{[t_0,t_f]} = \Big\{ y \in C(0,t_f;Y) : \\ \dot{y}(t) \in Ay(t) + F(t,y(t)),\ t \in [t_0,t_f]\ a.e., \quad y(0) = y_0 \Big\}. \tag{7.17}$$

Proof Suppose that $y \in \mathscr{S}_{[t_0,t_f]}$ according to Definition 7.1. Then, the function $g(t) = f(t, y(t), u(t), v(t))$ is integrable and a selection of $F(t, y(t))$. Further, (7.4) and (7.11) coincide for this choice. Hence, y is a solution of the differential inclusion

$$\dot{y}(t) \in Ay(t) + F(t,y(t)),\ t \in [t_0,t_f] \text{ a.e.}, \quad y(0) = y_0 \tag{7.18}$$

according to Definition 7.2.

Now we show the converse. Let $y \in C(0,t_f;Y)$ be a solution of (7.18). Let $X = U \times V$ and define $H : [t_0,t_f] \times X \to Y$ by $H(t,(u,v)) = f(t,y(t),u,v)$. Then, by assumption (i)–(iii), H is measurable in t (by Lemma 7.2) and continuous in (u,v). Moreover, with defining $W : [t_0,t_f] \rightrightarrows U \times V$ by

$$W(t) = \bigcup_{v \in V_{[t_0,t_f]}} (\mathscr{U}^v(t), \{v(t)\}), \quad t \in [t_0,t_f] \text{ a.e.} \tag{7.19}$$

we have $g(t) \in F(t,y(t)) = H(t, W(t))$ with W being measurable with closed nonempty images due to assumption (v) and finiteness of $\mathscr{V}$. Lemma 7.1 then yields the existence of a measurable selection $[u,v](t) \in W(t)$, i.e., controls $u \in U_{[t_0,t_f]}$ and $v \in V_{[t_0,t_f]}$ such that $u(t) \in \mathscr{U}^v(t)$ and $g(t) = f(t,y(t),u(t),v(t))$ for a.e. $t \in [t_0,t_f]$. Again, (7.4) and (7.11) coincide for this g. Hence, $y \in \mathscr{S}_{[t_0,t_f]}$ according to Definition 7.1.

Based on a generalization of the Filippov–Ważewski Theorem, one proves the following relaxation theorem [6].

Lemma 7.3 *The mild solutions in $\mathscr{S}_{[t_0,t_f]}$ are dense in the solution set $\mathscr{S}^{\overline{\text{co}}}_{[t_0,t_f]}$ defined as the set of all mild trajectories of the relaxed inclusion*

$$\begin{aligned} &\dot{y} \in Ay(t) + \overline{\text{co}}\, F(t,y(t)), \\ &y(0) = y_0(\lambda), \end{aligned} \tag{7.20}$$

where $\overline{\text{co}}$ denotes the closed convex hull and density is to be understood in the metric of uniform convergence.

The density in Lemma 7.3 means that for every trajectory y of (7.20) and every $\delta > 0$, there exists a solution y' of (7.1) such that $\|y - y'\|_{C([0,t_f];Y)} \le \delta$.

Moreover, we recall the following argument.

Lemma 7.4 *Let Y be a metric space, $f : Y \to \mathbb{R}$ be a continuous function, $\Psi \subseteq Y$ an arbitrary set, Ξ be a dense subset of Ψ (in the metric of Y). Then*

$$\inf\{f(y) : y \in \Psi\} = \inf\{f(y) : y \in \Xi\}. \tag{7.21}$$

Proof Let $a = \inf\{f(x) : x \in \Psi\}$. Then $a \leq \inf\{f(y) : y \in \Xi\}$, because $\Xi \subset \Psi$. Suppose that $\inf\{f(x) : x \in \Xi\} > a$. Then, the sets $f^{-1}(B_{\frac{1}{n}}(a))$ are open in Ψ (because f is continuous on X and thus on Ψ) and non-empty in Ψ (because a is infimum of f on Ψ). By density of Ξ in Ψ, we find $y_n \in f^{-1}(B_{\frac{1}{n}}(a)) \cap \Xi$ such that $a \leq f(y_n) \leq a + \frac{1}{n}$. Thus $\lim_{n\to\infty} f(y_n) = a$, which contradicts $\inf\{f(x) : x \in \Xi\} > a$.

Lemmas 7.3 and 7.4 imply the following main result concerning a very useful characterization of the optimal value in mixed-integer optimal control.

Corollary 7.1 *The optimal value defined in* (7.6) *satisfies*

$$\nu = \inf\{\varphi(y(t_f)) : y \in \mathscr{S}^{\overline{\text{co}}}_{[t_0,t_f]}\}, \tag{7.22}$$

where $\mathscr{S}^{\overline{\text{co}}}_{[t_0,t_f]}$ *denotes the set of all mild trajectories of the relaxed problem* (7.20).

Remark 7.1 The representation (7.22) can be used to simplify computations of optimal solutions. For example, in the case that control restrictions are imposed only on the continuous control, i. e., $\mathscr{U}^v$ being independent of $v \in \mathscr{V}$, integer optimal controls can be obtained numerically by sum-up rounding strategies [27, 28] applied to multiplier representations of the convexified right-hand side [12, 13]. Alternatively, gradient descent methods can be applied to parameterizations of integer controls with switching times and mode sequences [25] or variable time transformations of the dynamical system [8].

7.4 Application to Gas Network Optimization

In this section, we apply the results from the previous section to gas network operation. To this end, we introduce the relevant modeling aspects, provide an abstract problem formulation based on semigroup theory and consider a numerical example.

7.4.1 Networks with Pipes, Valves, and Compressors

We consider a network of pipes modeled by a metric graph $G = (V, E)$ with nodes $V = (v_1, \ldots, v_m)$ and edges $E = (e_1, \ldots, e_n) \subseteq V \times V$ for some $m, n \in \mathbb{N}$. For each edge $e = (v_1, v_2) \in E$, we call v_1 the *left node* and v_2 the *right node* of e. We exclude self-loops, i.e., we require that $v_1 \neq v_2$ for any $e = (v_1, v_2) \in E$. Further, for any $v \in V$, we define

the *set of ingoing edges* by	$\delta^+ v = \{(v_1, v_2) \in E \mid v_2 = v\}$,
the *set of outgoing edges* by	$\delta^- v = \{(v_1, v_2) \in E \mid v_1 = v\}$,
the *set of incident edges* by	$\delta v = \delta^- v \cup \delta^+ v$.

The number $|\delta v|$ then is called the *degree* of node $v \in V$.

With each edge $e_j \in E$ of such a network, we associate a pipe with length $L^j > 0$ parameterized by $x \in [0, L^j]$. We consider the motion of a compressible nonviscous gas in the pipe associated with $e_j \in E$ being governed by the following system of partial differential equations

$$\begin{aligned} \partial_t \rho^j + \partial_x q^j &= 0, \\ \partial_t q^j + (c^j)^2 \partial_x \rho^j &= -\theta^j \frac{q^j |q^j|}{\rho^j} - g(h^j)' \rho^j, \end{aligned} \tag{7.23}$$

where ρ^j denotes the density in kgm^{-3} and q^j the flux $q^j = \rho^j v^j$ with v^j the velocity in ms^{-1}. This model assumes a constant speed of sound $c^j = \sqrt{R_s T_0^j Z(P^j, T_0^j)}$ for a constant gas compressibility factor $Z(P^j, T_0^j)$ and a constant temperature T_0^j with R_s being the specific gas constant. Moreover, in this model $g^j \approx 9.81$ is the gravitational constant and $(h^j)'$ the slope of the pipe, θ^j is a friction factor with $\theta^j = \frac{\lambda^j}{2D^j}$, where λ^j is coefficient for the roughness of the pipe, and D^j is the diameter of the pipe. Finally, in this model the gas pressure P^j in kgm^{-1} is given by $P^j = (c^j)^2 \rho^j$.

These equations are simplifications of the one-dimensional isothermal Euler equations [2, 20, 21, 30] used for description of the dynamics of natural gas in subsonic regimes with typical values such as $c^j \approx 340\,\text{ms}^{-1}$ and rather small velocities $|v| \leq 10\,\text{ms}^{-1}$ [14].

In order to simplify technical considerations, we assume that the density (and hence the pressure) and the flow remain within bounds $\rho \in [\underline{\rho}, \bar{\rho}]$ and $q \in [\underline{q}, \bar{q}]$ with $\underline{\rho} > 0$. Such bounds are typically required in pipeline operations and may even also be considered explicitly in the optimization [5, 29]. Moreover, we note that this semilinear pipe model exhibits two simple characteristics speeds $\lambda_1 = -c$ and $\lambda_2 = c$ for each edge $e_j \in E$. We set

$$z^j = \begin{pmatrix} \rho^j \\ q^j \end{pmatrix}, \quad A^j = -\begin{bmatrix} 0 & 1 \\ c_j^2 & 0 \end{bmatrix}, \quad f^j(z) = \left(0, -\theta^j \frac{\min\{q^j |q^j|, \bar{q}\}}{\max\{\rho^j, \underline{\rho}\}}\right)^\top$$

for each $j \in \{1, \ldots, n\}$, where $\bar{\rho} > 0$ and $\bar{q} > 0$ are suitable truncation parameters to simplify theoretical considerations.

With this, we can summarize the pipe model as

$$\partial_t z^j = A \partial_x z^j + f^j(z^j), \quad j = 1, \ldots, n.$$

Furthermore, we impose coupling conditions for the gas density and flow at the boundary of pipes corresponding to edges being incident to that node. To this end, we define for $v \in V$ and $e_j \in \delta v$

$$x(v, e_j) = \begin{cases} 0, & \text{if } e_j \in \delta^- v, \\ 1, & \text{if } e_j \in \delta^+ v. \end{cases}$$

For each node $v \in V$, we then impose a transmission condition for the density and a balance equation for the fluxes at the node. The transmission condition states that the density variables ρ^j weighted by given factors $\alpha \in (0, \infty)^{m \times 2}$ coincide for all incident edges $e \in \delta v$ and can be expressed as

$$\alpha^k_{x(v,e_k)} \rho^k(t, L_k x(v, e_k)) = \alpha^l_{x(v,e_l)} \rho^l(t, L_l x(v, e_l)), \quad \forall\, e_k, e_l \in \delta v,\ t \in [0, T].$$

The nodal balance equation for a given outflow function $q^v \colon [0, T] \to \mathscr{R}$ is similar to a classical Kirchhoff condition for the fluxes q^j and can be written as

$$\sum_{e_j \in \delta^+ v} q^j(t, L_j) - \sum_{e_j \in \delta^- v} q^j(t, 0) = q^v(t), \qquad t \in [0, T].$$

The above setting is general enough to model typical components of gas networks such as junctions, entires, exists, compressors, and valves. Junctions can be modeled as nodes v such that $q^v \equiv 0$ and $\alpha^k_{x(v,e_k)} = 1$ for all $e_k \in \delta v$. Entires and exits can be modeled as nodes v such that $\alpha^k_{x(v,e_k)} = 1$ for all $e_k \in \delta v$, but $q^v \not\equiv 0$. We refer to v as an entry node, if $q^v < 0$, or an exit node, if $q^v > 0$. Compressors can be modeled as nodes v with $q^v \equiv 0$ and $|\delta^+ v| = |\delta^- v| = 1$. A description established via the characteristic diagram based on measured specific changes in adiabatic enthalpy H_{ad} of the compression process yields the model

$$H_{\text{ad}} = \bar{Z} T_0 R_s \frac{\kappa}{\kappa - 1} \left(\left(\frac{\rho^l(0, t)}{\rho^k(L_k, t)} \right)^{\frac{\kappa - 1}{\kappa}} - 1 \right), \quad e_k \in \delta^+ v, e_l \in \delta^- v,\ t \in [0, T],$$

where κ is a compressor-specific constant, $\bar{Z}$ is the gas compressibility factor that is assumed to be constant and H_{ad} is within flow dependent and compressor-specific bounds obtained from the characteristic diagram [19]. In consistency with the pipe models, we assume that H_{ad} is given by a known reference $\bar{H}_{\text{ad}}$. Then we get

$$\rho^l(0, t) = \bar{\alpha} \rho^k(L_k, t), \quad e_k \in \delta^+ v, e_l \in \delta^- v,\ t \in [0, T]$$

with a compressor-specific factor

$$\bar{\alpha} = \left(1 + \frac{(\kappa - 1)\bar{H}_{\text{ad}}}{\kappa \bar{Z} T_0 R_s}\right)^{\frac{\kappa}{\kappa-1}}. \tag{7.24}$$

This yields $\alpha_1^k = 1$ and $\alpha_0^l = \bar{\alpha}$. Finally, valves can be modeled as short pipes. By relabeling, we may assume that the edges $e_1, \ldots, e_{n_v}$ model valves for some $n_v \in \mathbb{N}$ with $0 < L^j \ll 1$, $j = 1, \ldots, n_v$. For simplicity, we may also assume that $(h^j)' = 0$, for all $j = 1, \ldots, n_v$, i.e., valves are horizontal network elements. For some $\varepsilon > 0$, we consider the valve action

$$f^j(\rho, q, w^j) = \begin{cases} f^j(\rho, q), & \text{if } w^j = 1 \text{ (valve open)} \\ \frac{1}{\varepsilon} f^j(\rho, q), & \text{if } w^j = 0 \text{ (valve closed).} \end{cases} \tag{7.25}$$

The valve action (7.25) can also be expressed as

$$f^j(\rho, q, w^j) = w^j f^j(\rho, q) + (1 - w^j)\frac{1}{\varepsilon} f^j(\rho, q), \quad w^j \in \{0, 1\}. \tag{7.26}$$

We then consider for each $j \in \{1, \ldots, n_v\}$, the dynamics for $z^j = (\rho^j, q^j)^\top$ on e_j given by

$$\begin{aligned} z_t^j(t, x) + A^j z_x^j(t, x) &= f^j(z^j(t, x), w^j), \quad && t \in [0, T], x \in [0, L_j], \\ z^j(0, x) &= z_0^j(x), && x \in [0, L_j], \end{aligned} \tag{7.27}$$

and for all $j \in \{n_v + 1, \ldots, n\}$ the dynamics for $z^j = (\rho^j, q^j)^\top$ on e_j given by

$$\begin{aligned} z_t^j(t, x) + A^j z_x^j(t, x) &= f^j(z^j(t, x)), \quad && t \in [0, T], x \in [0, L_j], \\ z^j(0, x) &= z_0^j(x), && x \in [0, L_j]. \end{aligned} \tag{7.28}$$

As the objective for optimization, we consider a sum of costs for all pipes in the network

$$\begin{aligned} J = \sum_{j=1}^{n} \int_0^T \int_0^{L_j} &\gamma_1^j (\min\{\rho^j(t, x), \bar{\rho}\} - \rho_d^j(t, x))^2 \\ &+ \gamma_2^j (\min\{q^j(t, x), \bar{q}\} - q_d^j(t, x))^2 \, \mathrm{d}x \, \mathrm{d}t, \end{aligned} \tag{7.29}$$

where ρ_d^j and q_d^j are some desired states, and $\gamma_1^j, \gamma_2^j \geq 0$ are given constants, $j = 1, \ldots, n$. Of course, other cost functions are possible.

Remark 7.2 The above valve model is meaningful only for ε being sufficiently small, e.g., choosing $\varepsilon = \left(\min_j \theta^j\right)^2 \frac{\bar{q}}{\underline{\rho}}$. However, this modeling is mostly of theoretical interest since too small ε lead to very stiff problems in the numerics. Alternatively, valves may be modeled as controlled junctions using $\alpha \in \{0, 1\}$. This then yields a mixed-integer boundary control problem for which relaxation result such as the one obtained in Sect. 7.3 are still open, but for which our numerical experiments in Sect. 7.4.3 show that they may still hold. For a more detailed discussion of this and further challenges with respect to model switching, see [10–12, 24].

7.4.2 Abstract Problem Formulation

We will set up an abstract formulation of the gas network system and present a result about the existence and uniqueness of mild solutions. This will allow us to apply the theory from the previous section.

As the state and control space, we introduce

$$Y = \left[\bigotimes_{j=1}^{n} L^2([0, L_j], \mathscr{R}^2) \right] \otimes L^2([0, \infty), \mathscr{R}^m), \quad \mathscr{V} = \{0, 1\}^{n_v} \simeq \{0, \dots, 2^{n_v} - 1\},$$

with a corresponding state vector

$$y = ((z^1)^\top, \dots, (z^n)^\top, q^{v_1}, \dots, q^{v_m})^\top \in Y$$

and a corresponding control vector

$$v = (w^1, \dots, w^{n_v}) \in \{0, 1\}^{n_v} \in \mathscr{V}.$$

With y_0, we denote the initial state

$$y_0 = ((z_0^1)^\top, \dots, (z_0^n)^\top, q_0^{v_1}, \dots, q_0^{v_m})^\top.$$

Further, we introduce the operators

$$A = \begin{bmatrix} A^1 & & \\ & \ddots & \\ & & A^n \end{bmatrix} \frac{\partial}{\partial x}, \qquad B = \mathrm{i}_m \frac{\partial}{\partial x} \tag{7.30}$$

and the nonlinear mapping $f \colon Y \times \mathscr{V} \to Y$ defined by

$$f(y, v) = (f^1(y^1, w^1), \dots, f^{n_v}(y^{n_v}, w^{n_v}), f^{n_v+1}(y^{n_v+1}), \dots, f^n(y^n), 0, \dots, 0)^\top. \tag{7.31}$$

Moreover, we define the block diagonal operator $\mathrm{diag}(A, B)$ on the domain

$$\begin{aligned}
&D\left(\begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}\right) \\
&= \Big\{ y = (z^1, \dots, z^n, q^{v_1}, \dots, q^{v_m})^\top \in Y : \ y \text{ is absolutely continuous}, \\
&\quad \alpha^k_{x(v,e_k)} z^k_1(L_k x(v, e_k)) = \alpha^l_{x(v,e_l)} z^l_1(L_l x(v, e_l)) \quad \forall v \in V,\ e_k, e_l \in \delta v, \\
&\quad \sum_{e_j \in \delta^+ v} z^j_2(L_j) - \sum_{e_j \in \delta^- v} z^j_2(0) = q^v(0) \quad \forall v \in V \Big\}.
\end{aligned} \tag{7.32}$$

With that, we can write the gas network dynamics with valve switching control as an abstract mixed-integer control problem

$$\dot{y}(t) = \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix} y(t) + f(y(t), v(t)), \quad t \in [t_0, t_f], \tag{7.33}$$

and the initial condition $y(t_0) = y_0$ with $t_0 = 0$.

For the homogeneous part of (7.33), we have the following well-posedness result [24].

Theorem 7.2 *The operator*

$$\left(D\left(\begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}\right), \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}\right)$$

is the infinitesimal generator of a strongly continuous semigroup on Y.

We note that the function $f(y, v)$ as defined in (7.31) is bounded and globally Lipschitz continuous in y for each fixed $v \in \{0, 1\}^{n_v}$. In particular, the assumptions (i)–(iii) of Sect. 7.3 are satisfied. Moreover, we can define the following running costs

$$\begin{aligned}
L(t, y) = \sum_{j=1}^{n} \int_0^{L_j} &\gamma^j_1 (\min\{\rho^j(t, x), \bar{\rho}\} - \rho^j_d(t, x))^2 + \\
&\gamma^j_2 (\min\{q^j(t, x), \bar{q}\} - q^j_d(t, x))^2 \, \mathrm{d}x \, \mathrm{d}t.
\end{aligned} \tag{7.34}$$

The function L is globally Lipschitz continuous. Hence, the assumption (iv) of Sect. 7.3 is satisfied with $\varphi = 0$.

Finally, we do not consider constraints on v. Hence, the assumption (v) of Sect. 7.3 is obsolete in this case. Theorem 7.1 yields that the valve switching problem can be regraded as a differential inclusion $\dot{y} \in Ay + F(y)$ with the set valued map

$$F(y) = \{f(y, v) : v \in \{0, 1\}^{n_v}\}. \tag{7.35}$$

An application of Corollary 7.1 yields that optimal value is the same for the relaxed problem with a convexified function $F(y)$. From (7.26) we can see that the relaxed mixed-integer optimal control problem is of the form

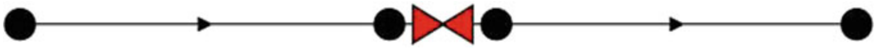


Fig. 7.1 A single valve scenario

Table 7.1 Numerical results for a single-valve scenario

Δt	Relaxed costs	Integer costs	Relative error
2.000	16.273	19.489	0.197
1.000	16.273	19.287	0.185
0.500	16.273	16.955	0.041
0.250	16.273	16.302	0.001

$$\min J(y) \quad \text{s.t. } \dot{y} = Ay + f(y, \tilde{v}), \; y(0) = y_0, \; \tilde{v} \in [0, 1]^{n_v}, \tag{7.36}$$

that is, a problem that can be assessed with standard methods.

7.4.3 Numerical Example

We consider a scenario with pipes of 10 km each which are coupled by a single valve, cf. Fig. 7.1. We choose $c = 340$, $\lambda = 0.01$ and impose a zero flow condition at the two boundary nodes and positive initial data. The objective is to close the valve as to "capture" most of the gas in the pipe on the right, modeled by tracking with $\rho_d = 0$ and $q_d = 0$ for the left pipe.

The minimization of (7.29) subject to the convexified problem (7.36) has been solved with a sequential quadratic programming method using finite differences for gradients and BFGS Hessian approximations applied to an explicit finite-volume-scheme for networked problems from [23, 24]. Integer feasibility is then obtained via sum up rounding [13, 27] using a step-size Δt. The numerical results are reported in Table 7.1. They confirm the theoretical result that the optimality gap between the relaxed and the rounded solution vanishes with Δt tending to zero.

7.5 Conclusion

We have shown that PDE mixed-integer optimal control problems can be regarded as non-convex operator differential inclusions. This includes a rather general form of constraints on the controls. Convexification yields a relaxation result and a useful characterization of the value function for such problems. Multiplier representations of the convexified problem are closely related to reformulations known as partial outer convexification. Combined, for example, with using sum-up rounding techniques, the approach yields a convenient way to obtain near-optimal solutions. We demonstrated this for the application of optimized operation of gas pipeline networks in nonstationary situations using valve switching as a control. We illustrated the approach using a numerical example.

Future research directions concern extensions of sum-up rounding strategies and similar decomposition techniques in order to treat more general control constraints. Also, the proper treatment of state constraints remains an important aspect in this context.

Acknowledgements This work was supported by the DFG grant CRC/Transregio 154, project A03.

References

1. J.-P. Aubin, H. Frankowska, *Set-Valued Analysis*. Systems and Control: Foundations and Applications, vol. 2 (Birkhauser Boston Inc, Boston, 1990)
2. J. Brouwer, I. Gasser, M. Herty, Gas pipeline models revisited: model hierarchies, nonisothermal models, and simulations of networks. Multiscale Model. Simul. **9**(2), 601–623 (2011)
3. M. Buss, M. Glocker, M. Hardt, O. von Stryk, R. Bulirsch, G. Schmidt, Nonlinear hybrid dynamical systems: modeling, optimal control, and applications, in *Modelling, Analysis, and Design of Hybrid Systems*, ed. by S. Engell, G. Frehse, E. Schnieder (Springer, Berlin 2002), pp. 311–335
4. S. Court, K. Kunisch, L. Pfeiffer, Hybrid optimal control problems for a class of semilinear parabolic equations. Discret. Contin. Dyn. Syst. - S **11**, 1031 (2018)
5. H. Egger, T. Kugler, W. Wollner, Numerical optimal control of instationary gas transport with control and state constraints. TRR 154 Preprint 214 (2017)
6. H. Frankowska, Value function in optimal control, in *Mathematical control theory, Part 1, 2 (Trieste, 2001)*. ICTP Lecture Notes, vol. VIII (Abdus Salam Int. Cent. Theoret. Phys., Trieste, 2002), pp. 516–653 (electronic)
7. H. Frankowska, A priori estimates for operational differential inclusions. J. Differ. Equ. **84**(1), 100–128 (1990)
8. M. Gerdts, A variable time transformation method for mixed-integer optimal control problems. Optim. Control Appl. Methods **27**(3), 169–182 (2006)
9. M. Gugat, F.M. Hante, Lipschitz continuity of the value function in mixed-integer optimal control problems. Math. Control Signals Syst. **29**(1), Art. 3, 15 (2017)
10. F.M. Hante, Stability and optimal control of switching PDE-dynamical systems, arXiv:1802.08143
11. F.M. Hante, G. Leugering, A. Martin, L. Schewe, M. Schmidt, Challenges in optimal control problems for gas and fluid flow in networks of pipes and canals: from modeling to industrial applications, in *Industrial Mathematics and Complex Systems: Emerging Mathematical Models, Methods and Algorithms*, ed. by P. Manchanda, R. Lozi, A. Hasan Siddiqi (Springer, Singapore, 2017), pp. 77–122
12. F.M. Hante, Relaxation methods for hyperbolic PDE mixed-integer optimal control problems. Optim. Control Appl. Methods **6**, 1103–1110 (2017). https://doi.org/10.1002/oca.2315
13. F.M. Hante, S. Sager, Relaxation methods for mixed-integer optimal control of partial differential equations. Comput. Optim. Appl. **55**(1), 197–225 (2013)
14. A. Herrán-González, J.M. De La Cruz, B. De Andrés-Toro, J.L. Risco-Martín, Modeling and simulation of a gas distribution pipeline network. Appl. Math. Model. **33**(3), 1584–1600 (2009)
15. O.V. Iftime, M.A. Demetriou, Optimal control of switched distributed parameter systems with spatially scheduled actuators. Autom. J. IFAC **45**(2), 312–323 (2009)

16. C. Kirches, S. Sager, H.G. Bock, J.P. Schlöder, Time optimal control of automobile test drives with gear shifts. Optim. Control Appl. Methods **31**(2), 137–153 (2010)
17. T. Koch, B. Hiller, M.E. Pfetsch, L. Schewe (eds.), *Evaluating Gas Network Capacities*. SIAM-MOS Series on Optimization (SIAM, 2015)
18. H.W.J. Lee, K.L. Teo, V. Rehbock, L.S. Jennings, Control parametrization enhancing technique for optimal discrete-valued control problems. Autom. J. IFAC **35**(8), 1401–1407 (1999)
19. B. Lendt, G. Cerbe, Grundlagen der Gastechnik. Hanser, 8th edn. (2016)
20. R.J. LeVeque, *Finite Volume Methods for Hyperbolic Problems* (Cambridge University Press, Cambridge, 2002)
21. J.R. LeVeque, *Numerical Methods for Conservation Laws* (Birkhäuser, Boston, 1992)
22. A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*. Applied Mathematical Sciences Series (Springer, New York, 1983)
23. F. Rüffer, F.M. Hante, Optimality conditions for switching operator differential equations. Proc. Appl. Math. Mech. 777–778 (2018)
24. F. Rüffer, V. Mehrmann, F.M. Hante, Optimal model switching for gas flow in pipe networks, in *Networks Heterogeneous Media*. **13**(4), 641–661 (2018)
25. F. Rüffer, F.M. Hante, Optimal switching for hybrid semilinear evolutions. Non-linear Anal. Hybrid Syst. **22**, 215–227 (2016)
26. S. Sager, *Numerical Methods for Mixed-Integer Optimal Control Problems* (Der andere Verlag, Tönning, Lübeck, Marburg, 2005)
27. S. Sager, Reformulations and algorithms for the optimization of switching decisions in non-linear optimal control. J. Process. Control. **19**(8), 1238–1247 (2009)
28. S. Sager, H.G. Bock, M. Diehl, The integer approximation error in mixed-integer optimal control. Math. Program. (2010). https://doi.org/10.1007/s10107-010-0405-3
29. J.M. Schmitt, S. Ulbrich, Optimal boundary control of hyperbolic balance laws with state constraints (2017)
30. J. Smoller, *Shock Waves and Reaction-Diffusion Equations*. Grundlehren der mathematischen Wissenschaften, vol. 258 (Springer, Berlin, 1983)
31. F. Zhu, P.J. Antsaklis, Optimal control of hybrid switched systems: a brief survey. Discret. Event Dyn. Syst. 1–20 (2014)
32. E. Zuazua, Switching control. J. Eur. Math. Soc. **13**, 85–117 (2011)

Chapter 8
Application of Solution of the Quantum Kinetic Equations for Information Technology and Renewable Energy Problem

Mukhayo Rasulova

Abstract In this paper, there is proved possibility application of the quantum kinetic equations toward the solution of the problems of information technology and renewable energy.

Keywords BBGKY hierarchy · General kinetic equation · Delta potential · Information technology · Renewable energy

8.1 Introduction

In 1872, for the first time, to describe the evolution of a classical particle, Ludwig Boltzmann introduced a kinetic equation [1] for the distribution function depending on the coordinate and momentum of the particles, called the following name. The other most well-known kinetic equation describing plasma evolution is the Vlasov equation [2]. In 1902, Gibbs introduced [3] the equation into statistical mechanics to describe the evolution of many interacting particles, which was derived as early as 1838 by Liouville. In 1946, starting from the Liouville kinetic equation, a chain of kinetic equations was formulated that relates the Liouville equation [4] and the Boltzmann equation and the Vlasov equation. This chain was included in the physical literature as the Bogolyubov–Born–Green–Kirkwood–Yvon chain [5], since it summarizes the attempts of different authors to generalize the kinetic equations for a single particle for the case of systems of many interacting particles. In quantum mechanics, particle dynamics is described by the Schrödinger equation [7]. To describe the physical phenomena in semiconductor physics, in solid-state physics, the Hartley [8] equation and other equations are used. To describe the evolution of systems of many interacting particles, the Liouville quantum kinetic equation is also used. Similarly, the classical case for connecting the equations for the one-particle

M. Rasulova (✉)
Institute of Nuclear Physics, Academy of Sciences Uzbekistan, Ulughbek, Tashkent 100214, Uzbekistan
e-mail: rasulova@live.com

P. Manchanda et al. (eds.), *Mathematical Modelling, Optimization, Analytic and Numerical Solutions*, Industrial and Applied Mathematics,
https://doi.org/10.1007/978-981-15-0928-5_8

case and the equation for the case of many particles, proceeding from the Liouville quantum equation, have been derived the BBGKY chain for the quantum kinetic equations [9] for the density matrices. As is known, the solution of the classical and quantum chain of BBGKY allows to determine the distribution function and the density matrix, respectively, in the classical and quantum cases. The definition of these solutions allows using these results to calculate the average values of physical quantities, characterizing the considered system of particles. In all areas of physics, the above equations are used to describe the dynamics of the corresponding particle systems. In this paper, on the basis of the Liouville quantum kinetic equation and the chain of quantum kinetic equations of BBGKY, we show the possibility of using them for information technology and for studying renewable energy.

8.2 Application of the Quantum Kinetic Equations for the Solution of Problems of Information Technology

To this end, we consider the chain of BBGKY quantum kinetic equations in a one-dimensional space bounded in Λ [9]:

$$
\begin{aligned}
i\frac{\partial \rho_s^{\Lambda}(t, x_1, \dots, x_s; x_1', \dots, x_s')}{\partial t} &= [H_s^{\Lambda}, \rho_s^{\Lambda}](t, x_1, \dots, x_s; x_1', \dots, x_s') \\
+\frac{N}{V}\left(1-\frac{s}{N}\right) Tr_{x_{s+1}} &\sum_{1\le i\le s} \left(\phi_{i,s+1}(|x_i - x_{s+1}|) - \phi_{i,s+1}(|x_i' - x_{s+1}|)\right) \\
&\times \rho_{s+1}^{\Lambda}(t, x_1, \dots, x_s, x_{s+1}; x_1', \dots, x_s', x_{s+1}),
\end{aligned}
\tag{8.1}
$$

with initial condition

$$
\rho_s^{\Lambda}(t, x_1, \dots, x_s; x_1', \dots, x_s')|_{t=0} = \rho_s^{\Lambda}(0, x_1, \dots, x_s; x_1', \dots, x_s').
$$

In (8.1) ρ is the density matrix, x is the one-dimensional coordinate of a particle, t-time, m-mass mass, $\hbar = 1$-Planck constant, H-Hamiltonian of the system, the s-number of particles, $\phi_{i,j}(|x_i - x_j|)$-the potential in this chain will be chosen as the delta function in the form:

$$
\delta(|x_i - x_j|) = \begin{cases} \infty \quad if \quad x_i{=}x_j, \\ 0 \quad if \quad x_i{\neq}x_j \end{cases}.
$$

Hamiltonian system has the form:

$$
H_s^{\Lambda}(x_1, \dots, x_s) = \sum_{1\le i\le s}\left(-\frac{1}{2m}\triangle_{x_i} + u^{\Lambda}(x_i)\right) + \sum_{1\le i<j\le s} \phi_{i,j}(|x_i - x_j|),
$$

where $\triangle_{x_i}$-laplacian and $\phi_{i,j}(|x_i - x_j|) = \delta(|x_i - x_j|)$.

To determine the solution of the chain (8.1), we introduce [10, 11] the space of nuclear operators B^{Λ} which is the Banach space of sequences of positively defined self-adjoint nuclear operators $\rho_s^{\Lambda}(x_1,\dots,x_s;x_1',\dots,x_s')$

$$\rho^{\Lambda}=\{\rho_0^{\Lambda},\rho_1^{\Lambda}(x_1;x_1'),\dots,\rho_s^{\Lambda}(x_1,\dots,x_s;x_1',\dots,x_s'),\dots\},$$

where ρ_0^{Λ} are complex numbers, $\rho_s^{\Lambda}\subset B_s^{\Lambda}$,

$$\rho_s^{\Lambda}(x_1,\dots,x_s;x_1',\dots,x_s')=0,\qquad when \qquad s>s_0,$$

s_0 bounded number and the norm is

$$|\rho^{\Lambda}|_1=\sum_{s=0}^{\infty}|\rho_s^{\Lambda}|_1.$$

and

$$|\rho_s^{\Lambda}|_1=sup\sum_{1\le i\le\infty}|(\rho_s^{\Lambda}\psi_i^s,\varphi_i^s)|,$$

where the upper bound is taken over all orthonormal systems of finite, twice differentiable functions with compact support $\{\psi_i^s\}$ and $\{\varphi_i^s\}$ in $L_2^s(\Lambda)$, $s\ge 1$ and $\left|\rho_0^{\Lambda}\right|_1=\left|\rho_0^{\Lambda}\right|$. Introducing operator

$$\left(\Omega(\Lambda)\rho^{\Lambda}\right)_s(x_1,\dots,x_s;x_1',\dots,x_s')=\frac{N}{V}\left(1-\frac{s}{N}\right)\times$$

$$\times\int_{\Lambda}\sum_i\rho_{s+1}^{\Lambda}(x_1,\dots,x_s,x_{s+1};x_1',\dots,x_s',x_{s+1})g_i^1(x_{s+1})\tilde{g}_i^1(x_{s+1})dx_{s+1},$$

and using the method of semigroups, on the basis of the Stone theorem on the specified space, we define a unique solution of a chain of quantum kinetic equations in the form

$$U^{\Lambda}(t)\rho_s^{\Lambda}(x_1,\dots,x_s;x_1',\dots,x_s')=$$

$$=(e^{\Omega(\Lambda)}e^{-iH^{\Lambda}t}e^{-\Omega(\Lambda)}\rho^{\Lambda}e^{iH^{\Lambda}t})_s(x_1,\dots,x_s;x_1',\dots,x_s'),\tag{8.2}$$

where

$$\rho_s^{\Lambda}(x_1,\dots,x_s;x_1',\dots,x_s')=\sum_i\psi_i(x_1,\dots,x_s)\psi_i^*(x_1',\dots,x_s').$$

and

$$\psi(x_1,\dots,x_s)=\frac{1}{s!}\sum_{\sigma}(-1)^{|\sigma|}exp\left(i\sum_{j=1}^{s}x_jk_{\sigma_j}\right)\times$$

$$\times exp\left[\frac{i}{2}\sum_{j>i}\theta(k_{\sigma_j}-k_{\sigma_i})\right],$$

in fundamental domain $F: x_1<x_2<\dots<x_s$ with eigenvalues $E_s=\sum_{i=1}^{s}k_i^2$ solving the equation

$$\psi|_{x_j=x_{k+0}}=\psi|_{x_j=x_{k-0}},$$

$$\left(\frac{\partial\psi}{\partial x_j}-\frac{\partial\psi}{\partial x_k}\right)|_{x_j=x_{k+0}}-\left(\frac{\partial\psi}{\partial x_j}-\frac{\partial\psi}{\partial x_k}\right)|_{x_j=x_{k-0}}=2c\psi|_{x_j=x_k},$$

for all $x_j=x_k$, j,k = 1,2,…,N and $j\neq k$ [12]. In [11] have been proved possibility to use formule (8.2) for information technology.

8.3 Application of the Quantum Kinetic Equations for the Solution of Problems of Renewable Energy

In the section of paper, Jaynes–Cummings Model (JCM) [13] is investigated in terms of the methods [14–16]. The system under consideration includes N two-level atoms, interacting with electromagnetic field mode. Hamiltonian of such systems in notations [14, 15] is given by

$$\hat{H}_t=\hat{H}_0+\sum_{j=1}^{N}\hbar\omega_0\hat{S}_j^z+\sum_{k=1}^{N}\hbar\omega_k\hat{b}_k^{\dagger}\hat{b}_k+$$

$$+e^{\varepsilon t}\sum_{k,j=1}^{N}\frac{\hbar g_k}{\sqrt{N}}(e^{ikx_j}\hat{b}_k\hat{S}_j^-(\mu)+e^{-ikx_j}\hat{b}_k^{\dagger}\hat{S}_j^+(\mu)),\tag{8.3}$$

where the first term describes the energy of free atom. This is given by

$$\hat{H}(S)=\hat{H}_0+\sum_{j=1}^{N}\hbar\omega_0\hat{S}_j^z.$$

Here $\hat{H}_0$ is the kinetic energy of atoms.

The second term corresponds to free electromagnetic field and is given by

$$\hat{H}(\Sigma) = \sum_{k=1}^{N} \hbar\omega_k \hat{b}_k^{\dagger}\hat{b}_k,$$

where $\hat{b}_k$ and $\hat{b}_k^{\dagger}$ are the operators of the annihilations and creations of photon with wave vector k. Here (S) and (Σ) denote (atom) and (field), correspondingly.

The last term corresponds to the interaction of atoms with the field and is given by

$$\hat{H}_t(S,\Sigma) = e^{\varepsilon t} \sum_{k,j=1}^{N} \frac{\hbar g_k}{\sqrt{N}} (e^{ikx_j}\hat{b}_k \hat{S}_j^{-}(\mu) + e^{-ikx_j}\hat{b}_k^{\dagger}\hat{S}_j^{+}(\mu)),$$

where g_k is the dipole coupling strength, N is the number of atoms, $\hat{S}_j^{-}(\mu) = \hat{S}_j^{+} + \mu\hat{S}_j^{-}$; $\varepsilon, \mu \in \mathscr{R}$; x_j is the radius vector of the jth atom and $0 \leq t$-time. In (8.3) $\hat{b}$ and $\hat{b}^{\dagger}$ are photon annihilation and creation operators, respectively, ω_0 is the splitting frequency between the atomic levels, $\hbar$ is Plank's constant, ω is the frequency of the field mode, g is the dipole coupling strength, and $\hat{S}^{+}$, $\hat{S}^{-}$, $\hat{S}^{z}$ are atomic spin operators satisfying the following commutation relation

$$[\hat{S}^{\pm}, \hat{S}^{z}] = \mp\hat{S}^{\pm}, \quad [\hat{S}^{+}, \hat{S}^{-}] = 2\hat{S}^{z},$$

where $\hat{S}^{+} = \frac{1}{2}(\hat{\sigma}_x + i\hat{\sigma}_y)$, $\hat{S}^{-} = \frac{1}{2}(\hat{\sigma}_x - i\hat{\sigma}_y)$, $\hat{S}^{z} = \frac{1}{2}\hat{\sigma}_z$; $\{\hat{\sigma}_x, \hat{\sigma}_y, \hat{\sigma}_z\}$ are Pauli's matrices.

By using the method of elimination of boson variables developed in [13], Liouville equation for operator $f(S)$ in Heisenberg representation can be converted to the following form:

$$\mathrm{Tr}_{(S)}\left(f(S)\frac{\partial\rho_t(S)}{\partial t} + \frac{\hat{H}(S)f(S) - f(S)\hat{H}(S)}{i\hbar}\rho_t(S)\right) =$$

$$= \sum_{k,j=1}^{N} \frac{g_k^2}{N} \int_{t_0}^{t} d\tau \mathrm{Tr}_{(S,\Sigma)} e^{-i\omega_k(t-\tau)} e^{\varepsilon(t+\tau)} \{N_k e^{ikx_j}\hat{S}_j^{+}(\tau,\mu)\times$$

$$\times[f(S_t), e^{-ikx_j}\hat{S}_j^{-}(t,\mu)] + (1+N_k)[e^{-ikx_j}\hat{S}_j^{-}(t,\mu), f(S_t)]\times$$

$$\times e^{ikx_j}\hat{S}_j^{+}(\tau,\mu)\}D_{t_0}(S,\Sigma) + \sum_{k,j=1}^{N} \frac{g_k^2}{N} \int_{t_0}^{t} d\tau \mathrm{Tr}_{(S,\Sigma)} e^{i\omega_k(t-\tau)} e^{\varepsilon(t+\tau)} \times$$

$$\times\{(1+N_k)e^{-ikx_j}\hat{S}_j^-(\tau,\mu)[f(S_t),e^{ikx_j}\hat{S}_j^+(t,\mu)]+$$

$$+N_k[e^{ikx_j}\hat{S}_j^+(t,\mu),f(S_t)]e^{-ikx_j}\hat{S}_j^-(\tau,\mu)\}D_{t_0}(S,\Sigma),$$

where $N_k=\frac{e^{-\beta\hbar\omega(k)}}{1-e^{-\beta\hbar\omega(k)}}$, and $f(S_t)$ is the dynamic value, $\rho(S)$ is the reduced density matrix of S system, $D(S,\Sigma)$ is the statistical operator of (S,Σ) system. In this paper, we used statistical approach and consider JCM from the point of view of many particle systems, thereby, we used collective operators and took into account the inhomogeneous Lorentz broadening and received formule intensity of emittting $(I(t)>0)$:

$$I_{\text{emit}}(t)=\frac{\hbar\omega_0\alpha}{4N}\left(N+\frac{\gamma}{\alpha}\right)^2 sech^2\frac{\eta(t-t_0)}{2},$$

and under the condition of $\langle\hat{S}^z(t)\rangle<-\frac{N}{2}$ and for tt_0, intensity of absorption $(I(t)<0)$ rate has the form:

$$I_{\text{abs}}(t)=-\frac{\hbar g^2\omega_0\alpha}{4N}\left(N+\frac{\gamma}{\alpha}\right)^2 cosech^2\frac{\eta(t-t_0)}{2}.$$

These latter formulas can be used to select efficient materials for converters of solar energy into electricity.

References

1. L. Boltzmann, *Wissenschaftliche Abhandlungen*, ed. by F. Hasenorl, J.A. Barth, vol. 2 (Leipzig, 1909)
2. A.A. Vlasov, JTEP **8** (1938)
3. J.W. Gibbs, Elementary Principles in Statistical Mechanics, Developed With Especial Reference to the Rational Foundation of Thermodynamics. Yale Bicentennial Publications, vol. XVIII (Scribner's Sons, 1902), p. 207
4. J. Liouville, J. de Math. **3**, 349 (1838)
5. N.N Bogolubov, Problems of a dynamical theory in statistical physics, Moscow (1946) in *Studies in Statistical Mechanics*, vol. 1, ed. by J. de Boer, G.E. Uhlenbeck (North-Holland, 1962)
6. E. Schrödinger, An undulatory theory of the mechanics of atoms and molecules. Phys. Rev. **28**(6), 1049–1070 (1926)
7. D.R. Hartree, Wave mechanics of an atom with a non-coulomb Central Field. Part II. Some results and discussion, in *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 24 (1928), pp. 111–132
8. N.N. Bogolyubov, Lectures on Quantum Statistics, London (1970); Selected papers, vol. 2 (Naukovo Dumka, Kyev, 1970)
9. N.N. Bogolyubov, N.N. Bogolubov, in *Introduction to Quantum Statistical Mechanics* (Nauka, Moscow, 1984)
10. D.Y. Petrina, *Mathematical Foundation of Quantum Statistical Mechanics, Continuous Systems* (Kluwer Academic Publishers, Dordrecht, 1995)
11. M.Yu. Rasulova, Appl. Math. Inf. Sci. **12**(4), 685–688 (2018)

12. E.H. Lieb, W. Liniger, Exact analysis of an interacting Bose gas. I: the general solution and the ground state. Phys. Rev. **130**, 1605–1616 (1963)
13. E.T. Jaynes, F.W. Cummings, Proc. IEEE **51**, 89 (1963)
14. N.N. Bogolyubov, N.N. Bogolyubov Jr., *Aspects of Polaron Theory* (Fizmatlit, Moscow, 2004)
15. N.N. Bogolyubov Jr., V.N. Plechko, A.S. Shumovsky, Phys. Elem. Part. At. Nuclei **14**, 1483 (1983)
16. N.N. Bogolyubov Jr., M.Y. Rasulova, I.A. Tishabaev, Theor. Math. Phys. **171**(1), 523–530 (2012)

Chapter 9
Inverse Problems Involving PDEs with Applications to Imaging

Taufiquar Khan

Abstract In this chapter, we introduce the general idea of inverse problems particularly with applications to imaging. We use two well-known imaging modalities namely electrical impedance and diffuse optical tomography to introduce and describe inverse problems involving PDEs. We also discuss the mathematical difficulties and challenges for image reconstruction in practice. We describe both deterministic and statistical regularization techniques including Gauss–Newton method, Bayesian inversion, and sparsity approaches to provide a broad exposure to the field.

Keywords Inverse problem involving PDEs · Ill-posed inverse problems in imaging · Electrical impedance tomography · Diffuse optical tomography

9.1 Introduction

The field of inverse problem is a fairly mature area of research and was initially motivated by industrial problems [1]. The growth of this research area has been tremendous in the last two decades. It is predicted that the use of inverse problems in applications in developing countries will grow due to significant demand from the industrial sector. Therefore, inverse problems relevant to developing countries should be getting a lot more attention. For example, the application of inverse problems in biomedical imaging is relevant in developing countries. Breast cancer is the most common cancer after skin-related cancers in the US. In 2012, 232,714 women in the US were diagnosed with breast cancer, of which 43,909 women died from breast cancer whereas 144,937 women in India were diagnosed with breast cancer during 2012, of which 70,218 women died from breast cancer with a significantly higher rate of death in India than the US. This is partly due to the failure to detect breast cancer early. X-ray mammography is still the predominant method for detection. However,

T. Khan (✉)
School of Mathematical and Statistical Sciences, Clemson University, Clemson, SC 20631, USA
e-mail: khan@clemson.edu

P. Manchanda et al. (eds.), *Mathematical Modelling, Optimization, Analytic and Numerical Solutions*, Industrial and Applied Mathematics,
https://doi.org/10.1007/978-981-15-0928-5_9

X-rays has drawbacks in terms of harmful radiation exposure and high false-positive rate in younger women. On the other hand, Diffuse Optical Tomography (DOT) has potential to be a cheaper alternative particularly useful in developing countries. However, the mathematics and computational challenges of this highly nonlinear ill-posed inverse problem is still making it not possible to use in clinical applications.

Let us describe the difference between a forward and an inverse problem. A forward problem is given by a process where given/known model parameters q we need to predict the solution or the output/data $u(q)$. For example, q could be parameters involving a partial differential equation (PDE) and u would be the solution of the PDE. Forward problems involving PDEs have been around hundreds of years where mathematicians and scientists are looking for appropriate solutions given for known parameters q. Many practical applications in fields such as biology, medicine, ecology, geophysics, flexible structures, as well as industry including biomedical imaging involve distributed parameter systems, i.e., partial differential equations (PDEs) [1] where the solution u is measured but the model parameters q are unknown. In order to validate models and/or control theoretical issues for these systems, it is often necessary to determine these model parameters q, such as coefficients, boundary terms, initial conditions, and control inputs, given a source or forcing function f and observations of the system z for which a quantitative model is sought. In general the inverse parameter estimation problem requires minimizing an error or a cost functional $J(q; z, f)$. In most cases a regularization procedure is also required to generate a well-posed minimization problem (existence, uniqueness, and stability of the inverse problem with respect to data). If an inverse problem suffers from existence, uniqueness, or stability, it is referred to as an ill-posed problem. Therefore, a typical reformulation of an ill-posed inverse parameter estimation problem requires optimizing $J(q; z, f) + \psi(q)$ where ψ is a regularization functional. Therefore, for ill-posed inverse problems, one must investigating several mathematical issues: (i) the choice of appropriate error or cost functional J; (ii) the algorithm to find the parameter q in the appropriate abstract function space given an initial parameter guess q_0; (iii) the selection of the best regularization operator ψ for inverse stability.

Here is an outline of the chapter. In Sect. 9.2, we describe Electrical Impedance Tomography (EIT) as an example of an ill-posed inverse problem involving PDEs, and in Sect. 9.3, we describe Diffuse Optical Tomography (DOT) as another example of an ill-posed inverse problem. In Sect. 9.4, we provide details of computational challenges including regularization. In Sect. 9.5, we conclude with a summary of future trends in solving ill-posed inverse problems.

9.2 Electrical Impedance Tomography

The Electrical Impedance Tomography (EIT) problem involves measuring electrical voltages on the smooth boundary $\partial\Omega$ to determine the spatially varying electrical conductivity q within the bounded region $\Omega \subseteq R^d (d = 2, 3)$. We assume q is strictly positive, isotropic and bounded conductivity with no current sources inside Ω. The

EIT forward problem is typically given by the following elliptic partial differential equation,

$$-\operatorname{div}(q\nabla u) = 0 \ \text{ in } \Omega, \tag{9.1}$$

where $u \in H^1(\Omega)$ is the electric potential and q is known.

In an EIT experiment, an electrical current (Neumann data) f on $\partial\Omega$ is applied and then the resulting electrical potential (Dirichlet data) g on $\partial\Omega$ is measured. The data collected then provides information and is used to approximate q from a set of EIT experiments using different input currents [2–4]. EIT can be applied to the monitoring of oil and gas mixtures in oil pipelines [6], noninvasive medical imaging [7, 8].

9.2.1 Analytical Setting for the EIT Model

We define the following Neumann and Dirichlet boundary value problems

$$\begin{aligned} -\operatorname{div}(q\nabla u) &= 0 \quad \text{in } \Omega, \\ q\frac{\partial u}{\partial n} &= f \quad \text{on } \partial\Omega, \end{aligned} \tag{9.2}$$

and

$$\begin{aligned} -\operatorname{div}(q\nabla u) &= 0 \quad \text{in } \Omega, \\ u &= g \quad \text{on } \partial\Omega. \end{aligned} \tag{9.3}$$

Let the conductivity $q \in Q$, an appropriate metric space. q is assumed to be bounded below and above, i.e. $0 < c_1 \le q \le c_2 < \infty$. In addition, denote by $\Gamma_D u$ the Dirichlet trace operator, i.e., the restriction of u to the boundary

$$\begin{aligned} \Gamma_D \ : \ X &\to Z \\ u &\mapsto \Gamma_D u. \end{aligned}$$

As usual in EIT, we restore uniqueness of the solution u of the Neumann problem (9.2) by requiring that the Dirichlet trace $\Gamma_D u$ satisfies

$$\int_{\partial\Omega} \Gamma_D u(s)ds = 0. \tag{9.4}$$

Note that to ensure the solvability of the Neumann problem (9.2) the current f must satisfy the integrability condition, which, in the absence of a source term, reads $\int_{\partial\Omega} f(s)ds = 0$. The associated linear forward operator of the Neumann problem, which maps an input current f to the solution u, is denoted by

$$F_N^q \ : W \to X \tag{9.5}$$

$$f \mapsto u \ \text{ solves (9.2).} \tag{9.6}$$

The linear operator F_D^q for the Dirichlet problem (9.3) can be defined analogously. The NtD map can be written as $\Gamma_D F_N^q$. The weak formulation of the Neumann problem (9.2) becomes

$$\int_\Omega q\nabla F_N^q(f) \cdot \nabla v dx = \int_{\partial\Omega} f \Gamma_D v ds \tag{9.7}$$

for a suitable set of test functions v. The integral on the boundary should be understood in the sense of duality pairing, i.e., $f \in H^{-1/2}(\partial\Omega)$ and $\Gamma_D v \in H^{1/2}(\partial\Omega)$ yield $\int_{\partial\Omega} f \Gamma_D v ds = \langle f, \Gamma_D v\rangle_{H^{-1/2} \times H^{1/2}}$.

There are several natural choices for the spaces X, Y and W. To this end, we introduce

$$X = \tilde{H}^1(\Omega) = \left\{ u \in L_2(\Omega) \mid \int_\Omega q(x)|\nabla u(x)|^2 dx < \infty, \int_{\partial\Omega} \Gamma_D u(s) ds = 0 \right\}. \tag{9.8}$$

Because of Eq. (9.4), the following bilinear form defines a scalar product on this space

$$\langle u, v\rangle_{\tilde{H}^1} = \int_\Omega q\nabla u \cdot \nabla v dx. \tag{9.9}$$

We use the Dirichlet forward operator F_D^q as an extension operator and define the following space of functions on the boundary $\partial\Omega$

$$Z = \tilde{H}^{1/2}(\partial\Omega) = \left\{ g \in L_2(\partial\Omega) \mid \int_\Omega q(x)|\nabla F_D^q(g)(x)|^2 dx < \infty, \int_{\partial\Omega} g(s) ds = 0 \right\} \tag{9.10}$$

together with its scalar product

$$\langle g, \varphi\rangle_{\tilde{H}^{1/2}} = \int_\Omega q(x)\nabla F_D^q(g)(x) \cdot \nabla F_D^q(\varphi)(x) dx. \tag{9.11}$$

The Dirichlet-to-Neumann (DtN) operator $q\frac{\partial}{\partial n} F_D^q$ is well defined on $\tilde{H}^{1/2}(\partial\Omega)$, and we introduce

$$W = \tilde{H}^{-1/2}(\partial\Omega) = \left\{ f \mid f = q\frac{\partial}{\partial n} F_D^q(g), \ g \in \tilde{H}^{1/2}(\partial\Omega) \right\} \tag{9.12}$$

together with its scalar product

$$\langle f, \psi\rangle_{\tilde{H}^{-1/2}} = \int_\Omega q\nabla F_N^q(g) \cdot \nabla F_N^q(\psi) dx. \tag{9.13}$$

We observe that $f \in \tilde{H}^{-1/2}(\partial\Omega)$ is the Neumann trace for $u = F_D^q(g)$. This implies $-\operatorname{div}(q\nabla u) = 0$ and the integrability condition for Neumann problems yields

$$\int_{\partial\Omega} f ds = 0 \text{ and } \tilde{H}^{-1/2}(\partial\Omega) = \left\{ f \in H^{-1/2}(\partial\Omega) | \int_{\partial\Omega} f ds = 0 \right\}. \tag{9.14}$$

Furthermore, the natural choice for a metric space Q is as follows:

$$Q = \left\{ q \in L^\infty(\Omega) | 0 < c_1 \le q \le c_2 < \infty \right\}. \tag{9.15}$$

9.2.2 Inverse Problem in EIT

The forward problem uses knowledge of conductivity parameter q to find the boundary data associated with a given source. The inverse problem instead uses knowledge of the source and boundary data and find the conductivity interior to the object. The goal is to estimate the conductivity distribution q from all pairs of current and voltage measurements. The identification of the parameter q can be formulated as the following minimization problem for the cost functional

$$J(q) = ||\mathscr{F}(q) - g_\delta||^2_{L_2(\partial\Omega)} \tag{9.16}$$

where g_δ approximate the exact data $g = \mathscr{F}(q)$ with the accuracy δ, i.e.,

$$||g - g_\delta|| < \delta \tag{9.17}$$

However, because of the ill-posedness of the problem, regularization is needed and one of the well-known regularization is Tikhonov's regularization mainly,

$$J_\alpha(q) = ||\mathscr{F}(q) - g_\delta||^2_{L_2(\partial\Omega)} + \alpha||q - q^*||^2_{L_2(\Omega)} \tag{9.18}$$

where λ is the regularization parameter and q^* is the prior or background parameter.

There are several other regularization approaches (see Sect. 9.4), for example for a particular EIT application total variation (TV) regularization functional may be the most appropriate,

$$J_\alpha(q) = \|\mathscr{F}(q) - g^\delta\|^2_{L_2(\partial\Omega)} + \alpha\|\nabla q\|_{L_1(\Omega)} \tag{9.19}$$

where α is a TV regularization parameter. The regularization parameter α is typically determined on a trial and error basis.

Why is the Inverse Problem Ill-Posed?

The inverse problem here is ill-posed because the EIT model is an elliptic PDE and elliptic forward operator is a highly smooth operator which means the information about the parameter q is diffused as it travels to the boundary i.e., the boundary measurements don't have enough information about an inhomogeneity located far away from the boundary. In fact, the information from the data at the boundary about the inhomogeneity is exponentially decayed away from the boundary [9]. This means that if an inhomogeneity is closer to the boundary then the data on the boundary provides better information for a reconstruction of the inhomogeneity. If an inhomogeneity is far away from the boundary, then the data on the boundary does not provide sufficient information due to the exponential decay [9]. So the very fact that elliptic equations can handle rough parameters q and still solve the forward problem for u resulting in a very smooth solution, in turn, making the inverse problem ill-posed. Therefore, EIT is an extremely challenging inverse problem. We will discuss more about the nonlinearity and ill-posedness of inverse problems while discussing DOT below.

Connection to DOT

EIT is a close cousin of DOT because the forward problem for both is elliptic with DOT being worse in the sense that DOT requires estimating two functions mainly $q = (D, \mu_a)$ rather one parameter for EIT. However, DOT is an important modality in practice because DOT can be used as an alternative to X-ray in detecting cancer in the breast and the brain.

9.3 Diffuse Optical Tomography

In optical imaging, low-energy visible light is used to illuminate biological tissue. The illumination of the tissue is modeled as a photon transport phenomenon. The process is described by the most widely applied equation in optical imaging, the radiative transfer, or transport equation (RTE) [10, 11]. RTE is an integro-differential equation for photon density and has spatially dependent diffusion and absorption parameters as coefficients. These coefficients are a priori unknown for a particular tissue sample of an individual who is being examined for cancer. Therefore, the problem is to infer from the measurements of the photon density on the boundary the absorption and diffusion coefficients inside the tissue. This estimate helps determine the location and size of the abnormality in the tissue.

9.3.1 Radiative Transport Equation

Let $\Omega \subset R^n$, with $n = 2, 3$ and with boundary $\partial\Omega$, $\nu(x)$ denote the outward unit normal to $\partial\Omega$ at the point $x \in \partial\Omega$, and $\Gamma_\pm$ is defined as,

$$\Gamma_\pm := \left\{(x, s, t) \in \partial\Omega \times S^{n-1} \times [0, T], \pm\nu(x) \cdot s > 0\right\}.$$

For example, one may assume a geometry as shown in Fig. 9.1. Then the RTE is given by,

$$\begin{aligned}\frac{1}{c}\frac{\partial u}{\partial t}(x, s, t) + s \cdot \nabla u(x, s, t) + a(x)u(x, s, t)& \\ - b(x)\int_{S^{n-1}} \Theta(s \cdot s')u(x, s', t)ds' = f(x, s, t)&\end{aligned} \tag{9.20}$$

together with the initial and boundary conditions,

$$u(x, s, 0) = 0 \text{ in } \Omega \times S^{n-1} \tag{9.21}$$

$$u(x, s, t) = 0 \text{ on } \Gamma_- \tag{9.22}$$

where $u(x, s, t)$ describes the density function of particles (photons) which travel in Ω at time t though the point x in the direction $s \in S^{n-1}$, unit sphere in R^n. The parameter a is the total cross section or attenuation, and b is the scattering cross section. The difference $\mu := a - b$ has the physical meaning of an absorption cross section. The parameters a, b and μ are assumed to be real, nonnegative and bounded functions of the position. The parameters a and b are the sought for tissue parameters and c is the velocity of light. The function Θ is the scattering phase function characterizing the intensity of a wave incident in direction s' scattered in the direction s. It is assumed to be a real, nonnegative function and is normalized to one. The inverse problem is to recover a and b from measurements of some given functionals of the outgoing density $u_j|_{\Gamma_+}$ at the boundary $\partial\Gamma$ for m_s different set of source distribution f_j, $j = 1, \ldots, m_s$.

9.3.2 DOT Model

Simpler deterministic models can be derived from RTE by expanding the density u and source f in spherical harmonics and retaining a limited number of terms [12–15]. Due to the prevalence of scattering, the flux is essentially isotropic a small distance away from the sources; i.e., it depends only linearly on s. Thus, we may describe the process adequately by the first two moments of u. We get diffusion approximation P_0. Frequency-domain diffusion approximation can easily be obtained by Fourier transforming the time-domain equation. Furthermore, the diffusion approximation

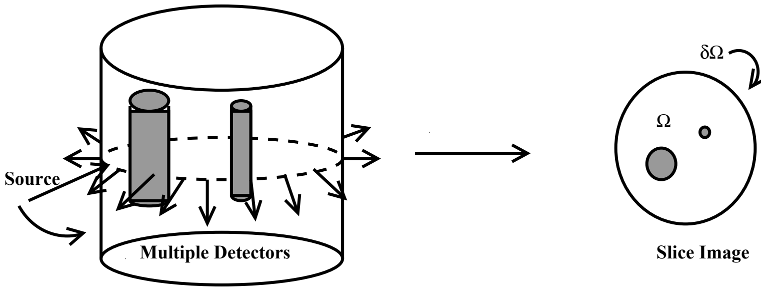


Fig. 9.1 Tomographic imaging

to the radiative transfer model can be written in the time independent (dc) case as in [15],

$$-\nabla \cdot D\nabla u + \mu_a u = 0. \tag{9.23}$$

The associated boundary condition is

$$u + 2D\frac{\partial u}{\partial \nu} = f, x \in \partial\Omega. \tag{9.24}$$

If we let Ω be the domain under consideration with surface $\partial\Omega$, the weak forward problem corresponding to Eq. (9.23) is to find $u \in H^1(\Omega)$ such that for all $v \in H^1(\Omega)$, the following variational equation is satisfied,

$$\int_\Omega D\nabla v \cdot \nabla u dx + \int_\Omega v\mu_a u dx + \int_{\partial\Omega} \frac{1}{2} vuds = \int_{\partial\Omega} vfds. \tag{9.25}$$

Now, we can define the forward problem as: given sources f_j in $\partial\Omega$ and q in Q, a vector of model parameters, for example the coefficient of diffusion D and the coefficient of absorption μ_a (i.e. $q = (D, \mu_a)^T$) that belongs to a parameter set Q, find the data u on $\partial\Omega$ and the inverse problem as: given data z on $\partial\Omega$ find q. We can recast the forward problem in an abstract setting with u in an appropriate abstract space H, and f represents a source or a forcing distribution. In general, measurement of u may not be possible, only some observable part $z = \mathscr{C}u(q)$ of the actual state $u(q)$ may be measured. In this abstract setting, the objective of the inverse or parameter estimation problem is to choose a parameter q^* in Q, that minimizes an error criterion or cost functional $J(u(q), \mathscr{C}u(q), q)$ over all possible q in Q subject to $u(q)$ satisfying the diffusion approximation. A typical observation operator is,

$$\mathscr{C}^f u(q) = \left\{ -D\frac{\partial u}{\partial \nu}(x_i; q, f) \right\}_{i=1}^m \tag{9.26}$$

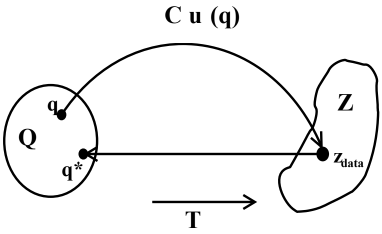


Fig. 9.2 Nonlinear mapping

where x_i is in $\partial\Omega$, m is the number of measurements, and the second equality comes from the boundary condition (9.24). A typical Tikhonov cost functional J_λ is given as,

$$J_\lambda(q) = \frac{1}{2}\sum_{j=1}^{m_s}\sum_{i=1}^{m} w_{ij}\left|\mathscr{C}_i^{f_j}u(q) - z_i^{f_j}\right|^2 + \lambda\|q - q_0\|^2 \tag{9.27}$$

where $z_i^{f_j}$ is the measured data at the boundary for a given source f_j, w_{ij} is the weight for ij-th data and λ is the Tikhonov regularization parameter. As shown in Fig. 9.2, composing $u(q)$ and $\mathscr{C}u(q)$ we obtain the parameter-to-output mapping:

$$T : Q \to Z$$

such that $Tq = \mathscr{C}u(q)$, where Z is the space of measurements. This is the nonlinear mapping of DOT in abstract setting. The map from Q to Z is nonlinear because the solution of a partial differential equation is a nonlinear function of its coefficients $q = (D, \mu_a)$.

9.4 Computational Aspect and the Regularization of the Ill-Posed Problem

There are various approaches for solving this nonlinear ill-posed problems, we outline a few approaches in this section. In general, for complex geometries, the analytic solution is intractable. Therefore, one requires numerical solutions. The finite element method (FEM) is somewhat more versatile because of its ease in complex geometries and modeling boundary effects. The FEM is a variational method used to approximate the solution by a family of finite-dimensional basis functions. Then the forward problem is reduced to one of linear algebra and one computes the approximate solution using FEM codes. The FEM is derived by projecting the weak form of (9.25) onto a finite-dimensional function space. For example, the finite-dimensional function space could be the set of continuous and twice differentiable, piecewise cubic polynomials and obtain a system of equations.

More precisely, for example in EIT, we can project the infinite dimensional solution space X into a finite dimensional subspace $X_K \subset X$ which means to restrict u and v above to lie in X_K rather than X. Let us assume that there exists a family of basis functions $\phi_m(x)$ for $m = 1, \ldots, K$ for X_K. Then let $u_K = \sum_{m=1}^{K} c_m \phi_m(x)$ to be the approximation to u and we want the weak formulation to be satisfied for the K test functions $\phi_k = v$ for $k = 1, \ldots, K$. Now plugging this into the weak formulation we obtain the system of equations for c_m:

$$\sum_{m=1}^{K} \left(\int_{\Omega} q \nabla \phi_m \cdot \nabla \phi_k dx \right) c_m = \int_{\partial \Omega} f \Gamma_{D} \phi_k ds \tag{9.28}$$

for $k = 1, \ldots, K$. If we denote

$$\begin{aligned} A_{mk} &= \int_{\Omega} q \nabla \phi_m \cdot \nabla \phi_k dx \\ F_k &= \int_{\partial \Omega} f \Gamma_{D} \phi_k ds \end{aligned}$$

then we can solve for the solution $c = (c_m)_{m=1}^{K}$, which depends nonlinearly on q using the linear system $Ac = F$ where A is a $K \times K$ matrix and F is a vector of length K. We can also approximate the infinite dimensional parameter space Q by a finite dimensional subspace $Q_M \subset Q$. The solution to the finite dimensional problem is the solution q_M in Q_M that is closest to infinite dimensional optimization problem for example $J_\alpha(q)$. Therefore, we assume that there exists a family of basis functions $\psi_k(x)$ for $k = 1, \ldots, M$ for Q_M. Then set $q_M = \sum_{k=1}^{M} q_k \psi_k(x)$ to be the approximation to q and we arrive at the finite dimensional nonlinear optimization problem for $q_M = (q_k)_{k=1}^{M}$:

$$\hat{J}_{\alpha_1, \alpha_2}(q_M) = \sum_{k=1}^{\hat{K}} |\hat{g}_k(j) - \hat{g}_k^\delta|^2 + \alpha_1 \sum_{i=1}^{M} \delta_i |q_i - q_i^*|^p + \alpha_2 \sum_{j=1}^{Z} d_j | \triangle_j q_M|. \tag{9.29}$$

where $1 \leq p \leq 2$ and δ_i's are weights, $\hat{g}_k(f) = \Gamma_D u_K^{q_M}(f)[\hat{x}_k]$ is the trace of the solution evaluated at $\hat{K}$ boundary points $\hat{x}_k$, $\hat{g}_k^\delta$ is the noisy voltage data collected at the boundary point $\hat{x}_k$, the second term is the approximation to the TV term using a nearest neighbor approximation where Z is the number of nearest neighbors and d_j is the distance between two neighbors [16–18].

EIT is well-known to suffer from a high degree of nonlinearity and severe ill-posedness [19, 20]. Therefore, regularization is required to produce reasonable electrical impedance images. Most reconstruction methods are deterministic, such as the factorization method [22], d-bar method [23], and variational type methods for least squares fitting. The variational type methods use an iterative type method of a linearized model or fully nonlinear model such as sparsity constraints [19, 20], iteratively regularized Gauss Newton method [21]. These analytical methods can be effective in determining specific conductivity, but statistical inversion methods [24] can offer an alternative approach. In [25], Kaipio et al. optimizes the current patterns based on criteria regarding functionals of the posterior covariance matrix. The Bayesian approach has also been used to study the errors from model reduction and partially unknown geometry [26, 27]. The sparsity regularization for statistical inversion enforces the ℓ_p prior to the expansion coefficients for a certain basis like the deterministic approaches for EIT [19, 20, 29].

Since we have a nonlinear forward and inverse operator, any iterative algorithm requires computing the jacobian of the forward operator. Under the regularity assumptions on the domain and the coefficients, the forward operator can be shown to be differentiable.

Theorem The operator $\mathscr{F}$ which maps q to the solution $u(q) \in H$ of the forward problem with current f is Frèchet differentiable. If $\eta \in L^\infty(\Omega)$ is such that $q + \eta \in \mathscr{Q}$, then the derivative $\mathscr{F}(q)\eta = w$ satisfies the following variational problem

$$b(w, v) = -\int_\Omega \eta \nabla u \nabla v dx \tag{9.30}$$

for all $v \in H$, where $u = \mathscr{F}(q)$. Using the theorem above, computing the jacobian is explained in [20].

9.4.1 *Iteratively Regularized Gauss–Newton Method*

Suppose λ_k is a sequence of regularizing parameters ([30]). A general algorithm is given by [30] using a line search procedure with a variable step size α_k such that

$$0 < \alpha_k \leq 1 \tag{9.31}$$

yielding the following Iteratively Regularized Gauss–Newton (IRGN) algorithm

$$q_{k+1} = q_k - \alpha_k(\mathscr{F}'(q_k)^T\mathscr{F}'(q_k) + \lambda_k W_2)^{-1}\{\mathscr{F}'(q_k)^T(\mathscr{F}(q_k) - g_\delta) + \lambda_k W_2(q_k - q^*)\} \tag{9.32}$$

where W_2 is a preconditioning matrix. Due to the inexact nature of g_δ, we adopt a stopping rule presented in [31] to terminate the iterations at the first index $\mathscr{K} = \mathscr{K}(\delta)$, such that

$$||\mathscr{F}(q_{\mathscr{K}}) - g_\delta||^2 \leq \rho\delta < ||\mathscr{F}(q_k) - g_\delta||^2, \quad 0 \leq k \leq \mathscr{K}, \quad \rho > 1 \tag{9.33}$$

The line search parameter α_k is chosen appropriately with a search direction using a backtracking strategy until either the strong or weak Wolfe conditions are satisfied [32], or a maximum number of backtracking steps have been taken.

9.4.2 The Statistical Inverse Problem

Inverse problems are typically written in terms of a minimization problem. After discretization, we can also write the finite-dimensional inverse problem in terms of the posterior density of the conductivity q_M given the measurements g_K on $\partial\Omega$. In other words, if we know the density of the conductivity q_M given the measurements g_K we can obtain the expected value of the conductivity given the measurements. This estimate is a reasonable point estimate of the solution of the ill-posed inverse problem. In the statistical setting, one derives the posterior density of the finite-dimensional version of the conductivity mainly q_M given the finite-dimensional measurement $g_{\hat{K}}$ on $\partial\Omega$ [16–18].

9.4.2.1 The Posterior Density

The density of q_M^* is usually called prior density. This is because it contains all the information about q_M^* that we believe to be true. Here we assume that

$$\pi_{q_M^*}(q_M) \propto \chi_A(q_M)\exp[-\alpha R(q_M)], \tag{9.34}$$

where $R(\cdot)$ is a regularization function, $\alpha > 0$ a constant and $\chi_A(q_M)$ a indicator function with $A = [0,\infty)^n$. In the following section we discuss several common choices for the regularizing function $R(\cdot)$.

9.4.3 Regularization Functions

In this section, we discuss several choices for the regularization function $R(\cdot)$. We have several choices for the regularization function $R(\cdot)$ which are used in the analytical and the statistical setting.

9.4.3.1 The ℓ_p Regularizations

The idea of the ℓ_p regularization is to force the difference of the parameters q_M and the background q_M^b mainly $q_M - q_M^b$ to be sparse with respect to some basis. For example, when using EIT to reconstruct an object mainly made of concrete we would choose q_M^b to be the typical conductivity of concrete. The ℓ_p regularization $R_{\ell_p}(q_M)$ is defined as

$$R_{\ell_p}(q_M) := \sum_{i=1}^{n} c_i |q_M(i) - q_M^b(i)|^p, \tag{9.35}$$

where c_i represent weights and $0 < p \leq 2$ a constant [24]. In theory a good choice for the weights would be large values at the boundary and exponentially decreasing values towards the center of Ω. This is because the variance is smaller on the boundary than in the center [9]. The ℓ_p regularization enforces sparsity when $0 < p \leq 1$ and enforces smoothness when $p \geq 2$.

9.4.3.2 The Total Variation Regularization

The idea behind the total variation regularization is to obtain smooth images. This is meaningful in most practical applications. The total variation regularization is defined as

$$R_{TV_c}(q) := \int_{\Omega} |\nabla q| dx, \tag{9.36}$$

where q the continuous version of the parameter of interest q_M. The discrete analog for a two-dimensional body of the total variation regularization R_{TV_c} [24] is

$$R_{TV}(q_M) := \sum_{i=1}^{h} l_i |\triangle_i q_M|, \tag{9.37}$$

where l_i is defined as the length of the edge corresponding to the ith adjacent pixel and

$$\triangle_i = (0, 0, ..., 0, 1_{a_{(1,i)}}, 0, ..., 0, -1_{a_{(2,i)}}, 0, ..., 0), \tag{9.38}$$

with $a = (a_{(j,i)})_{i=1, j\in\{1,2\}}^{h}$ is the set containing the numbers of all adjacent pixel tuples $(a_{(1,i)}, a_{(2,i)})$.

9.4.4 *The Markov Chain Monte Carlo Method*

In the previous sections, we discussed the posterior density with several meaningful prior densities (regularizations). Hence to obtain a good estimate for q_M^* based on the measurements $g_{\hat{K}}$. That is, the algorithm seeks to find the Bayesian estimate $E(q_M^*|g_{\hat{K}}) = \int_{R^n} q_M \pi_{q_M^*}(q_M|g_{\hat{K}}) dq_M$.

Given that the posterior density $\pi_{q_M^*}(q_M|g_{\hat{K}})$ does not have a closed form, there is no direct method of finding the Bayesian estimate $E(q_M^*|g_{\hat{K}})$. Therefore, one uses the Markov Chain Monte Carlo Method (MCMC) to generate a large random sample $\{q_M^{(i)}\}_{i=1}^{N}$ from the posterior density $\pi_{q_M^*}(q_M|g_{\hat{K}})$ and then approximate the Bayesian estimate by its sample mean,

$$E(q_M^*|g_{\hat{K}}) = \int_{R^n} q_M \pi_{q_M^*}(q_M|g_{\hat{K}}) dq_M \approx \frac{1}{N} \sum_{i=1}^{N} q_M^{(i)}. \tag{9.39}$$

Typical algorithms to generate such a random samples from a posterior density are the Gibbs sampler or the Metropolis-Hastings algorithm [18, 33].

9.5 Conclusion

In this chapter, we have introduced inverse problems involving PDEs for imaging applications. We described the challenges and difficulties of solving an ill-posed inverse problem. We have provided details of two very well-known examples namely EIT and DOT. A range of both deterministic and statistical regularization approaches have been exposed and summarized in this chapter. The computational approaches have been discussed without any rigorous analysis of the convergence rates. The future challenges in this area can be overcome by reformulating the inverse question either using a regularization approach such as a combination of smoothness and sparsity or statistical approaches or weakening the inverse question itself. The progress in the area of inverse problems in imaging is expected to be in the intersection of computation, statistical, and analytical approaches to understand the input out put behavior of the inverse operator. One of the latest approaches proposed involve machine learning and training methods for inverse problems. There are already significant interests in that direction in the latest literature however many questions about machine learning and adaptive training methods are still open and subject to future research.

Acknowledgements The author would like to thanks the organizers of ISIAM particularly Professor Pammy Manchanda and Professor Abul Hasan Siddiqi for editing the book chapters and with the opportunity to attend the conference at Amritsar that brought together both pure and applied mathematicians to help provide the platform to discuss various ideas for applied math research.

References

1. H.T. Banks, K. Kunisch, *Estimation Techniques for Distributed Parameter Systems* (Birkhauser, Basel, 2001)
2. M. Cheney, D. Isaacson, J.C. Newell, Electrical impedance tomography. SIAM Rev. **41**(1), 85–101 (1999)
3. L. Borcea, Electrical impedance tomography. Inverse Probl. **18**(6), R99–R136 (2002)
4. M. Hanke, M. Brühl, Recent progress in electrical impedance tomography. Inverse Probl. **19**(6), S65–S90 (2003)
5. W. Daily, A. Ramirez, D. LaBrecque, J. Nitao, Electrical resistivity tomography of vadose water movement. Water Resour. Res. **28**(5), 1429–1442 (1992)
6. O. Isaksen, A.S. Dico, E.A. Hammer, A capacitance-based tomography system for interface measurement in separation vessels. Meas. Sci. Technol. **5**(10), 1262 (1994)
7. D.S. Holder (ed.), *Electrical Impedance Tomography: Methods, History and Applications* (Institute of Physics Publishing, Bristol, 2005)
8. R.H. Bayford, Bioimpedance tomography (electrical impedance tomography). Annu. Rev. Biomed. Eng. **8**, 63–91 (2006)

9. V.P. Palamodov, Gabor analysis of the continuum model for impedance tomography. Arkiv för Matematik **40**(1), 169–187 (2002)
10. R. Chandrasekhar, *Radiation Transfer* (Clarendon, Oxford, 1950)
11. A. Ishimaru, *Single Scattering and Transport Theory (Wave Propogation and Scattering in Random Media I)* (Academic, New York, 1978)
12. H.W. Lewis, Multiple scattering in an infinite medium. Phys. Rev. **78**, 526–529 (1950)
13. H. Bremmer, Random volume scattering. Radiat. Sci. J. Res. **680**, 967–981 (1964)
14. S.R. Arridge, J.C. Hebden, Optical imaging in medicine: 2. Modelling and reconstruction. Phys. Med. Biol. **42**, 841–853 (1997)
15. S.R. Arridge, Optical tomography in medical imaging: topical review. Inverse Probl. **15**, R41–R93 (1999)
16. T. Strauss, Statistical inverse problems in electrical impedance and diffuse optical tomography. Doctoral dissertation, Clemson University (2015)
17. T. Strauss, T. Khan, Statistical inversion in electrical impedance tomography using mixed total variation and non-convex ℓ_p regularization prior. J. Inverse Ill-Posed Probl. **23**(5), 529–542 (2015)
18. T. Strauss, X. Fan, S. Sun, T. Khan, Statistical inversion of absolute permeability in single-phase darcy flow. Proc. Comput. Sci. **51**, 1188–1197 (2015)
19. B. Jin, P. Maass, An analysis of electrical impedance tomography with applications to Tikhonov regularization. ESAIM: Control, Optim. Calc. Variat. **18**(04), 1027–1048 (2012)
20. B. Jin, T. Khan, P. Maass, A reconstruction algorithm for electrical impedance tomography based on sparsity regularization. Int. J. Numer. Methods Eng. **89**(3), 337–353 (2012)
21. T. Khan, A. Smirnova, 1D inverse problem in diffusion based optical tomography using iteratively regularized Gauss-Newton algorithm. Appl. Math. Comput. **161**(1), 149–170 (2005)
22. A. Kirsh, N. Grinberg, *The Factorization Method for Inverse Problems* (Oxford University Press, Oxford, 2008)
23. D. Isaacson, J.L. Mueller, J.C. Newell, S. Siltanen, Reconstructions of chest phantoms by the D-bar method for electrical impedance tomography. IEEE Trans. Med. Imaging **23**(7), 821–828 (2004)
24. J.P. Kaipio, V. Kolehmainen, E. Somersalo, M. Vauhkonen, Statistical inversion and Monte Carlo sampling methods in electrical impedance tomography. Inverse Probl. **16**(5), 1487 (2000)
25. J.P. Kaipio, A. Seppänen, E. Somersalo, H. Haario, Posterior covariance related optimal current patterns in electrical impedance tomography. Inverse Probl. **20**(3), 919 (2004)
26. A. Nissinen, L.M. Heikkinen, J.P. Kaipio, The Bayesian approximation error approach for electrical impedance tomography experimental results. Meas. Sci. Technol. **19**(1), 015501 (2008)
27. A. Nissinen, L.M. Heikkinen, V. Kolehmainen, J.P. Kaipio, Compensation of errors due to discretization, domain truncation and unknown contact impedances in electrical impedance tomography. Meas. Sci. Technol. **20**(10), 105504 (2009)
28. J.M. Bardsley, MCMC-based image reconstruction with uncertainty quantification. SIAM J. Sci. Comput. **34**(3), A1316–A1332 (2012)
29. B. Jin, P. Maass, Sparsity regularization for parameter identification problems. Inverse Probl. **28**(12), 123001 (2012)
30. A. Smirnova, R.A. Renaut, T. Khan, Convergence and application of a modified iteratively regularized Gauss-Newton algorithm. Inverse Probl. **23**, 1547–1563 (2007)
31. A.B. Bakushinsky, A. Smirnova, On application of generalized discrepancy principle to iterative methods for nonlinear ill-posed problems. Numer. Funct. Anal. Optim. **26**, 35–48 (2005)
32. J. Nocedal, S.J. Wright, *Numerical Optimization* (Springer, New York, 1999)
33. S. Chib, E. Greenberg, Understanding the metropolis-hastings algorithm. Am. Stat. **49**(4), 327–335 (1995)

Chapter 10
Critical Growth Elliptic Problems with Choquard Type Nonlinearity: A Survey

K. Sreenadh and T. Mukherjee

Abstract This article deals with a survey of recent developments and results on Choquard equations where we focus on the existence and multiplicity of solutions of the partial differential equations which involves the nonlinearity of the convolution type. Because of its nature, these equations are categorized under the nonlocal problems. We give a brief survey on the work already done in this regard following which we illustrate the problems we have addressed. Seeking the help of variational methods and asymptotic estimates, we prove our main results.

Keywords Hardy–Littlewood–Sobolev inequality · Critical exponent problem · Nonlocal elliptic equations

10.1 A Brief Survey

We devote our first section on briefly glimpsing the results that have already been proved in the context of existence and multiplicity of solutions of the Choquard equations. Consider the problem

$$-\Delta u + u = (I_\alpha * |u|^p)|u|^{p-2}u \text{ in } \mathbb{R}^n \tag{10.1}$$

where $u : \mathbb{R}^n \to \mathbb{R}$ and $I_\alpha : \mathbb{R}^n \to \mathbb{R}$ is the Riesz potential defined by

$$I_\alpha(x) = \frac{\Gamma\left(\frac{n-\alpha}{2}\right)}{\Gamma\left(\frac{\alpha}{2}\right)\pi^{\frac{n}{2}}2^\alpha|x|^{n-\alpha}}$$

K. Sreenadh (✉)
Indian Institute of Technology Delhi, Hauz Khaz, New Delhi 110016, India
e-mail: sreenadh@maths.iitd.ac.in

T. Mukherjee
Tata Institute of Fundamental Research (TIFR) Centre of Applicable Mathematics, Bangalore, India
e-mail: tulimukh@gmail.com

P. Manchanda et al. (eds.), *Mathematical Modelling, Optimization, Analytic and Numerical Solutions*, Industrial and Applied Mathematics,
https://doi.org/10.1007/978-981-15-0928-5_10

for $\alpha \in (0, n)$ and Γ denotes the Gamma function. Equation (10.1) is generally termed as Choquard equations or the Hartree type equations. It has various physical significance. In the case $n = 3$, $p = 2$ and $\alpha = 2$, (10.1) finds its origin in a work by S.I. Pekar describing the quantum mechanics of a polaron at rest [63]. Under the same assumptions, in 1976 P. Choquard used (10.1) to describe an electron trapped in its own hole, in a certain approximation to Hartree–Fock theory of one component plasma [46]. Following standard critical point theory, we expect that solutions of (10.1) can be viewed as critical points of the energy functional

$$J(u) = \frac{1}{2}\int_{\mathbb{R}^n} (|\nabla u|^2 + u^2) - \frac{1}{2p}\int_{\mathbb{R}^n} (I_\alpha * |u|^p)|u|^p.$$

It is clear from the first term that naturally we have to take $u \in H^1(\mathbb{R}^n)$ which makes the first and second term well defined. Now the question is whether the third term is well defined and sufficiently smooth over $H^1(\mathbb{R}^n)$? For this, we recall the following Hardy–Littlewood–Sobolev inequality.

Theorem 10.1.1 *Let $t, r > 1$ and $0 < \mu < n$ with $1/t + \mu/n + 1/r = 2$, $f \in L^t(\mathbb{R}^n)$ and $h \in L^r(\mathbb{R}^n)$. Then there exists a constant $C(t, n, \mu, r)$, independent of f, h such that*

$$\int_{\mathbb{R}^n}\int_{\mathbb{R}^n} \frac{f(x)h(y)}{|x-y|^\mu}\mathrm{d}x\mathrm{d}y \le C(t, n, \mu, r)\|f\|_{L^t}\|h\|_{L^r}. \tag{10.2}$$

If $t = r = \frac{2n}{2n-\mu}$ then

$$C(t, n, \mu, r) = C(n, \mu) = \pi^{\frac{\mu}{2}} \frac{\Gamma\left(\frac{n}{2} - \frac{\mu}{2}\right)}{\Gamma\left(n - \frac{\mu}{2}\right)} \left\{\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma(n)}\right\}^{-1+\frac{\mu}{n}}.$$

The inequality in (10.2) is achieved if and only if $f \equiv (constant)\ h$ and

$$h(x) = A(\gamma^2 + |x-a|^2)^{\frac{-(2n-\mu)}{2}}$$

for some $A \in \mathbb{C}$, $0 \ne \gamma \in \mathbb{R}$ and $a \in \mathbb{R}^n$.

For $u \in H^1(\mathbb{R}^n)$, let $f = h = |u|^p$, then by Theorem 10.1.1,

$$\int_{\mathbb{R}^n}\int_{\mathbb{R}^n} \frac{|u(x)|^p|u(y)|^p}{|x-y|^{n-\alpha}}\mathrm{d}x\mathrm{d}y \le C(t, n, \mu, p)\left(\int_{\mathbb{R}^n} |u|^{\frac{2np}{n+\alpha}}\right)^{1+\frac{\alpha}{n}}.$$

This is well defined if $u \in L^{\frac{2np}{n+\alpha}}(\mathbb{R}^n)$. By the classical Sobolev embedding theorem, the embedding $H^1(\mathbb{R}^n) \hookrightarrow L^r(\mathbb{R}^n)$ is continuous when $r \in [2, 2^*]$, where $2^* = \frac{2n}{n-2}$. This implies $u \in L^{\frac{2np}{n+\alpha}}(\mathbb{R}^n)$ if and only if

$$2_\alpha := \frac{n+\alpha}{n} \le p \le \frac{n+\alpha}{n-2} := 2_\alpha^*. \tag{10.3}$$

The constant 2_α is termed as the lower critical exponent and 2^*_α is termed as the upper critical exponent in the sense of Hardy–Littlewood–Sobolev inequality. Then we have the following result.

Theorem 10.1.2 *If $p \in (1, \infty)$ satisfies (10.3), then the functional J is well defined and continuously Fréchet differentiable on the Sobolev space $H^1(\mathbb{R}^n)$. Moreover, if $p \geq 2$, then the functional J is twice continuously Fréchet differentiable.*

This suggests that it makes sense to define the solutions of (10.1) as critical points of J. A remarkable feature in Choquard nonlinearity is the appearance of a lower nonlinear restriction: the lower critical exponent $2_\alpha > 1$. That is the nonlinearity is superlinear.

10.1.1 Existence and Multiplicity Results

Definition 10.1.3 A function $u \in H^1(\mathbb{R}^n)$ is said to be a weak solution of (10.1) if it satisfies

$$\int_{\mathbb{R}^n} (\nabla u \nabla v + uv)\, dx + \int_{\mathbb{R}^n} \left(\int_{\mathbb{R}^n} \frac{|u(y)|^p}{|x-y|^\alpha}\, dy \right) |u|^{p-2}uv\, dx = 0$$

for each $v \in H^1(\mathbb{R}^n)$.

Definition 10.1.4 We define a solution $u \in H^1(\mathbb{R}^n)$ to be a groundstate of the Choquard equation (10.1) whenever it is a solution that minimizes the functional J among all nontrivial solutions.

In [54] V. Moroz and J. Van Schaftingen studied the existence of groundstate solutions and their asymptotic behaviour using concentration-compactness lemma. The groundstate solution has been identified as infimum of J on the Nehari manifold

$$\mathscr{N} = \{u \in H^1(\mathbb{R}^n) : \langle J'(u), u \rangle = 0\}$$

which is equivalent to prove that the mountain pass minimax level $\inf\limits_{\gamma \in \Gamma} \sup\limits_{[0,1]} J \circ \gamma$ is a critical value. Here the class of paths Γ is defined by $\Gamma = \{\gamma \in C([0, 1]; H^1(\mathbb{R}^n)) : \gamma(0) = 0 \text{ and } J(\gamma(1)) < 0\}$. Precisely, they proved the following existence result.

Theorem 10.1.5 *If $2_\alpha < p < 2^*_\alpha$ then there exists a nonzero weak solution $u \in W^{1,2}(\mathbb{R}^n)$ of (10.1) which is a groundstate solution of (10.1).*

They have also proved the following Pohozaev identity:

Proposition 10.1.6 *Let $u \in H^2_{loc}(\mathbb{R}^n) \cap W^{1,\frac{2np}{n+\alpha}}(\mathbb{R}^n)$ is a weak solution of the equation*

$$-\Delta u + u = (I_\alpha * |u|^p)|u|^{p-2}u \text{ in } \mathbb{R}^n$$

then
$$\frac{n-2}{2}\int_{\mathbb{R}^n} |\nabla u|^2 + \frac{n}{2}\int_{\mathbb{R}^n} |u|^2 = \frac{n+\alpha}{2p}\int_{\mathbb{R}^n} (I_\alpha * |u|^p)|u|^p.$$

Pohozaev identity for some Choquard type nonlinear equations has also been studied in [21, 51]. Using Proposition 10.1.6, they proved the following nonexistence result.

Theorem 10.1.7 *If $p \le 2_\alpha$ or $p \ge 2_\alpha^*$ and $u \in H^1(\mathbb{R}^n) \cap L^{\frac{2np}{n+\alpha}}(\mathbb{R}^n)$ such that $\nabla u \in H^1_{loc}(\mathbb{R}^n) \cap L^{\frac{2np}{n+\alpha}}_{loc}(\mathbb{R}^n)$ satisfies (10.1) weakly, then $u \equiv 0$.*

Next important thing to note is the following counterpart of Brezis–Leib lemma:

If the sequence $\{u_k\}$ converges weakly to u in $H^1(\mathbb{R}^n)$, then

$$\lim_{k\to\infty}\int_{\mathbb{R}^n} (I_\alpha * |u_k|^p)|u_k|^p - (I_\alpha * |u-u_k|^p)|u-u_k|^p = \int_{\mathbb{R}^n} (I_\alpha * |u|^p)|u|^p. \tag{10.4}$$

One can find its proof in [52, 54]. Equation (10.4) plays a crucial role in obtaining the solution where there is a lack of compactness.

Next coming to the positive solutions, in [8] authors studied the existence of solutions for the following equation

$$-\Delta u + V(x)u = (|x|^{-\mu} * F(u))f(u),\ u > 0 \text{ in } \mathbb{R}^n,\ u \in D^{1,2}(\mathbb{R}^n) \tag{10.5}$$

where F denotes primitive of f, $n \ge 3$ and $\mu \in (0, n)$. Assumptions on the potential function V and the function f are as follows:

(i) $\lim\limits_{s\to 0^+} \dfrac{sf(s)}{s^q} < +\infty$ for $q \ge 2^* = \frac{2n}{n-2}$
(ii) $\lim\limits_{s\to\infty} \dfrac{sf(s)}{s^p} = 0$ for some $p \in \left(1, \frac{2(n-\mu)}{n-2}\right)$ when $\mu \in (1, \frac{n+2}{2})$,
(iii) There exists $\theta > 2$ such that $1 < \theta F(s) < 2f(s)s$ for all $s > 0$,
(iv) V is a nonnegative continuous function.

Define the function $\mathscr{V} : [1, +\infty) \to [0, \infty)$ as

$$\mathscr{V}(R) = \frac{1}{R^{(q-2)(n-2)}} \inf_{|x|\ge R} |x|^{(q-2)(n-2)} V(x)$$

Motivated by the articles [11–13], authors proved the following result in [8].

Theorem 10.1.8 *Assume that $0 < \mu < \frac{n+2}{2}$ and $(i)-(iv)$ hold. If there exists a constant $\mathscr{V}_0 > 0$ such that if $\mathscr{V}(R) > \mathscr{V}_0$ for some $R > 1$, then (10.5) admits a positive solution.*

Taking $p = 2$ in (10.1), Ghimenti, Moroz and Schaftingen [56] established existence of a least action nodal solution which appeared as the limit of minimal action nodal solutions for (10.1) when $p \searrow 2$. They proved the following theorem by constructing a Nehari nodal set and minimizing the corresponding energy functional over this set.

Theorem 10.1.9 *If* $\alpha \in ((n-4)^+, n)$ *and* $p = 2$ *then* (10.1) *admits a least action nodal solution.*

In [74], Zhang et al. proved the existence of infinitely many distinct solutions for the following generalized Choquard equation using the index theory

$$-\Delta u + V(x)u = \left(\int_{\mathbb{R}^n} \frac{Q(y)F(u(y))}{|x-y|^\mu}\, dy\right) Q(x)f(u(x)) \text{ in } \mathbb{R}^n \tag{10.6}$$

where $\mu \in (0, n)$, V is periodic, f is either odd or even and some additional assumptions. Although Theorem 10.1.7 holds, when $p = 2_\alpha$ in (10.1), Moroz and Schaftingen in [53] proved some existence and nonexistence of solutions for the problem

$$-\Delta u + V(x)u = (I_\alpha * |u|^{2_\alpha})|u|^{2_\alpha - 2}u \text{ in } \mathbb{R}^n \tag{10.7}$$

where the potential $V \in L^\infty(\mathbb{R}^n)$ and must not be a constant. They proved existence of a nontrivial solution if

$$\liminf_{|x|\to\infty} (1 - V(x))|x|^2 > \frac{n^2(n-2)}{4(n+1)}$$

and gave necessary conditions for existence of solutions of (10.7). Because 2_α is the lower critical exponent in the sense of Theorem 10.1.1, a lack of compactness occurs in minimization technique. So concentration-compactness lemma and Brezis Lieb type lemma plays an important role. Equation (10.7) was reconsidered by Cassani, Schaftingen and Zhang in [17] where they gave necessary and sufficient condition for the existence of positive ground state solution depending on the potential V. In [1], authors addressed the topic of existence of ground state solutions, existence and multiplicity of the semiclassical solutions and their concentration behaviour related to the following singularly perturbed Choquard equation

$$-\varepsilon^2 \Delta u + V(x)u = \varepsilon^{\mu-3} \left(\int_{\mathbb{R}^3} \frac{Q(y)G(u(y))}{|x-y|^\mu}\, dy\right) Q(x)g(u) \text{ in } \mathbb{R}^3$$

where $\varepsilon > 0$, $\mu \in (0, 3)$, V, Q are continuous functions on $\mathbb{R}^3$, G denotes primitive of g which has critical growth. We alo refer [25–27, 29, 30, 67] to readers as some relevant contribution on this topic.

Very recently, in [36], authors studied some existence and multiplicity results for the following critical growth Kirchhoff-Choquard equations

$$-M(\|u\|^2)\Delta u = \lambda u + (I_\alpha * |u|^{2^*_\mu})|u|^{2^*_\mu - 2}u \text{ in } \Omega,\ u = 0 \text{ on } \partial\Omega$$

where $M(t) \sim at + bt^\theta$, $\theta \geq 1$ for some constants a and b.

Now let us consider the critical dimension case that is $n = 2$ commonly known as the Trudinger–Moser case. When $n = 2$, the critical Sobolev exponent becomes infinity and the embedding goes as $W^{1,2}(\mathbb{R}^2) \hookrightarrow L^q(\mathbb{R}^2)$ for $q \in [2, \infty)$ whereas $W^{1,2}(\mathbb{R}^2) \not\hookrightarrow L^\infty(\mathbb{R}^2)$. The following *Trudinger–Moser inequality* plays a crucial role when $n = 2$.

Theorem 10.1.10 *For $u \in W_0^{1,2}(\mathbb{R}^2)$,*

$$\int_{\mathbb{R}^2} [\exp(\alpha |u|^2) - 1]\, dx < \infty.$$

Moreover if $\|\nabla u\|_2 \leq 1$, $\|u\|_2 \leq M$ and $\alpha < 4\pi$ then there exists a $C(\alpha, M) > 0$ such that

$$\int_{\mathbb{R}^2} [\exp(\alpha |u|^2) - 1]\, dx < C(M, \alpha).$$

Motivated by this, the nonlinearity in this case is an appropriate exponential function. The following singularly perturbed Choquard equation

$$-\varepsilon^2 \Delta u + V(x)u = \varepsilon^{\mu-2} \left(|x|^{-\mu} * F(u)\right) f(u) \text{ in } \mathbb{R}^2 \tag{10.8}$$

was studied by Alves et al. in [7]. Here $\mu \in (0, 2)$, V is a continuous potential, ε is a positive parameter, f has critical exponential growth in the sense of Trudinger–Moser and F denotes its primitive. Under appropriate growth assumptions on f, authors in [7] proved existence of a ground state solution to (10.8) when $\varepsilon = 1$ and V is periodic and also established the existence and concentration of semiclassical ground state solutions of (10.8) with respect to ε. An existence result for Choquard equation with exponential nonlinearity in $\mathbb{R}^2$ has also been proved in [6]. The Kirchhoff-Choquard problems in this case are studied in the work [9]. Very recently, Yang in [73] established an existence and concentration behaviour of solutions for Choquard equations in $\mathbb{R}^2$ with critical growth.

Now let us consider the Choquard equations in the bounded domains. In particular, consider the Brezis–Nirenberg type problem for Choquard equation

$$-\Delta u = \lambda u + \left(\int_\Omega \frac{|u|^{2^*_\mu}(y)}{|x-y|^\mu}\, dy \right) |u|^{2^*_\mu - 2} u \text{ in } \Omega,\ u = 0 \text{ on } \partial\Omega \tag{10.9}$$

where Ω is bounded domain in $\mathbb{R}^n$ with Lipschitz boundary, $\lambda \in \mathbb{R}$ and $2^*_\mu = \dfrac{2n-\mu}{n-2}$ which is the critical exponent in the sense of Hardy–Littlewood–Sobolev inequality. These kind of problems are motivated by the celebrated paper of Brezis and Nirenberg [15]. Gao and Yang in [32] proved existence of nontrivial solution to (10.9) for $n \geq 4$ in case λ is not an eigenvalue of $-\Delta$ with Dirichlet boundary condition and for a suitable range of λ when $n = 3$. They also proved the nonexistence result when Ω is a star shaped region with respect to origin. Here, the best constant for the embedding is defined as

$$S_{H,L} := \inf\left\{\int_{\mathbb{R}^n} |\nabla u|^2 : \ u \in H^1(\mathbb{R}^n), \ \int_{\mathbb{R}^n} (|x|^{-\mu} * |u|^{2^*_\mu})|u|^{2^*_\mu} dx = 1\right\}.$$

They showed that the minimizers of $S_{H,L}$ are of the form $U(x) = \left(\frac{b}{b^2+|x-a|^2}\right)^{\frac{n-2}{2}}$ where a, b are appropriate constants. We remark that $U(x)$ is the Talenti function which also forms minimizers of S, the best constant in the embedding $H^1_0(\Omega)$ into $L^{2^*}(\Omega)$. Let us consider the family $U_\varepsilon(x) = \varepsilon^{\frac{2-n}{2}} U(\frac{x}{\varepsilon})$. Using Brezis–Lieb lemma, in [32] it was shown that every Palais Smale sequence is bounded and the first critical level is

$$c < \frac{n+2-\mu}{4n-2\mu} S_{H,L}^{\frac{2n-\mu}{n+2-\mu}}.$$

If $Q_\lambda := \inf\limits_{u \in H^1_0(\Omega)\setminus\{0\}} \frac{\int_\Omega |\nabla u|^2 - \lambda u}{\int_\Omega (|x|^{-\mu} * |u|^{2^*_\mu})|u|^{2^*_\mu} dx}$, then $Q_\lambda < S_{H,L}$ which can be shown using U_ε's. Then using Mountain pass lemma and Linking theorem depending on the dimension n, existence of first solution to (10.9) is shown. The nonexistence result was proved after establishing a Pohozaev type identity. Gao and Yang also studied Choquard equations with concave-convex power nonlinearities in [31] with Dirichlet boundary condition.

Very recently, the effect of topology of domain on the solution of Choquard equations has been studied by some researchers. Ghimenti and Pagliardini [34] proved that the number of positive solution of the following Choquard equation

$$-\Delta u - \lambda u = \left(\int_\Omega \frac{|u|^{p_\varepsilon}(y)}{|x-y|^\mu}\, dy\right) |u|^{p_\varepsilon - 2}u, \ u > 0 \text{ in } \Omega, \quad u = 0 \text{ in } \mathbb{R}^n \setminus \Omega \tag{10.10}$$

depends on the topology of the domain when the exponent p_ε is very close to the critical one. Precisely, they proved.

Theorem 10.1.11 *There exists $\bar{\varepsilon} > 0$ such that for every $\varepsilon \in (0, \bar{\varepsilon}]$, Problem (10.10) has at least $cat_\Omega(\Omega)$ low energy solutions. Moreover, if it is not contractible, then there exists another solution with higher energy.*

Here $cat_\Omega(\Omega)$ denotes the Lusternik–Schnirelmann category of Ω. They used variational methods to look for critical points of a suitable functional and proved a multiplicity result through category methods. This type of result was historically introduced by Coron for local problems in [10]. Another significant result in this regard has been recently obtained by authors in [37]. Here they showed existence of a high energy solution for

$$-\Delta u = \left(\int_\Omega \frac{|u|^{2^*_\mu}(y)}{|x-y|^\mu}\, dy\right) |u|^{2^*_\mu - 2}u \text{ in } \Omega, \ u = 0 \text{ on } \partial\Omega,$$

where Ω is an annular type domain with sufficiently small inner hole.

10.1.2 Radial Symmetry and Regularity of Solutions

Here, we try to give some literature on radially symmetric solutions and regularity of weak solutions constructed variationally for Choquard equations.

First we come to the question of radially symmetric solutions. Is all the positive solutions for the equation

$$\Delta u - \omega u + (|x|^{-\mu} * |u|^{2_\alpha})p|u|^{2_\alpha - 2}u = 0,\ \omega > 0,\ u \in H^1(\mathbb{R}^n) \tag{10.11}$$

are radially symmetric and monotone decreasing about some fixed point? This was an open problem which was settled by Ma and Zhao [50] in case $2 \leq p < \frac{2n-\mu}{n-2}$ and some additional assumptions. The radial symmetry and uniqueness of minimizers corresponding to some Hartree equation has also been investigated in [33]. Recently, Wang and Yi [70] proved that if $u \in C^2(\mathbb{R}^n) \cap H^1(\mathbb{R}^n)$ is a positive radial solution of (10.1) with $p = 2$ and $\alpha = 2$ then u must be unique. Using Ma and Zhao's result, they also concluded that the positive solutions of (10.1) in this case is uniquely determined, up to translations in the dimension $n = 3, 4, 5$. Huang et al. in [42] proved that (10.1) with $n = 3$ has at least one radially symmetric solutions changing sign exactly k-times for each k when $p \in (2.5, 5)$. Taking $V \equiv 1$ in (10.6) and f satisfies almost necessary the upper critical growth conditions in the spirit of Berestycki and Lions, Li and Tang [45] very recently proved that (10.6) has a ground state solution, which has the constant sign and is radially symmetric with respect to some point in $\mathbb{R}^n$. They used the Pohozaev manifold and a compactness lemma by Strauss to conclude their main result. For further results regarding Choquard equations, we suggest readers to refer [57] which extensively covers the existing literature on the topic. Very recently, in [37] authors studied the classification problem and proved that all positive solutions of the following equation are radially symmetric: for $p = 2^*_\mu$

$$-\Delta u = (I_\mu * |u|^p)|u|^{p-2}u \text{ in } \mathbb{R}^n. \tag{10.12}$$

They observed that the solutions of this problem satisfy the integral system of equations

$$\begin{aligned} u(x) &= \int_{\mathbb{R}^n} \frac{u^{p-1}(y)v(y)}{|x-y|^{N-2}}\, dy, u \geq 0 \text{ in } \mathbb{R}^n \\ v(x) &= \int_{\mathbb{R}^n} \frac{u^p(y)}{|x-y|^{N-\mu}}\, dy, v \geq 0 \text{ in } \mathbb{R}^n. \end{aligned} \tag{10.13}$$

By obtaining the regularity estimates and using moving method they proved the following result:

Theorem 10.1.12 *Every nonnegative solution $u \in D^{1,2}(\mathbb{R}^N)$ of equation (10.12) is radially symmetric, monotone decreasing and of the form*

$$u(x) = \left(\frac{c_1}{c_2 + |x - x_0|^2}\right)^{\frac{N-2}{2}}$$

for some constants $c_1, c_2 > 0$ and some $x_0 \in \mathbb{R}^N$.

Next we recall some regularity results for the problem (10.1). Fix $\alpha \in (0, n)$ and consider the problem (10.1), then in [54] authors showed the following-

Theorem 10.1.13 *If $u \in H^1(\mathbb{R}^n)$ solves (10.1) weakly for $p \in (2_\alpha, 2^*_\alpha)$ then $u \in L^1(\mathbb{R}^n) \cap C^2(\mathbb{R}^n)$, $u \in W^{2,s}(\mathbb{R}^n)$ for every $s > 1$ and $u \in C^\infty(\mathbb{R}^n \setminus u^{-1}\{0\})$.*

The classical bootstrap method for subcritical semilinear elliptic problems combined with estimates for Riesz potentials allows them to prove this result. Precisely, they first proved that $I_\alpha * |u|^p \in L^\infty(\mathbb{R}^n)$ and using the Calderón–Zygmund theory they obtain $u \in W^{2,r}(\mathbb{R}^n)$ for every $r > 1$. Then the proof of Theorem 10.1.13 followed from application of Morrey–Sobolev embedding and classical Schauder regularity estimates. In [55], author extended a special case of the regularity result by Brezis and Kato [14] for the Choquard equations. They proved the following-

Theorem 10.1.14 *If $H, K \in L^{\frac{2n}{\alpha}}(\mathbb{R}^n) \cap L^{\frac{2n}{\alpha+2}}(\mathbb{R}^n)$ and $u \in H^1(\mathbb{R}^n)$ solves*

$$-\Delta u + u = (I_\alpha * Hu)K \text{ in } \mathbb{R}^n$$

then $u \in L^p(\mathbb{R}^n)$ for every $p \in \left[2, \frac{2n^2}{\alpha(n-2)}\right)$.

They proved it by establishing a nonlocal counterpart of Lemma 2.1 of [14] in terms of the Riesz potentials. After this, they showed that the convolution term is a bounded function that is $I_\alpha * |u|^p \in L^\infty(\mathbb{R}^n)$. Therefore,

$$| - \Delta u + u| \leq C(|u|^{\frac{\alpha}{n}} + |u|^{\frac{\alpha+2}{n-2}})$$

that is the right hand side now has subcritical growth with respect to the Sobolev embedding. So by the classical bootstrap method for subcritical local problems in bounded domains, it is deduced that $u \in W^{2,p}_{loc}(\mathbb{R}^n)$ for every $p \geq 1$. Moreover it holds that if (10.1) admits a positive solution and p is an even integer then $u \in C^\infty$, refer [23, 43, 44]. Using appropriate test function and results from [55], Gao and Yang in [31] established the following regularity and L^∞ estimate for problems on bounded domains-

Lemma 10.1.15 *Let u be the solution of the problem*

$$- \Delta u = g(x, u) \text{ in } \Omega, \quad u \in H^1_0(\Omega), \tag{10.14}$$

where g is satisfies $|g(x,u)| \leq C(1+|u|^p) + \left(\int_\Omega \frac{|u|^{2^*_\mu}}{|x-y|^\mu} dy\right) |u|^{2^*_\mu - 2}u$, $\mu \in (0,n)$, $1 < p < 2^* - 1$ *and* $C > 0$ *then* $u \in L^\infty(\Omega)$.

As a consequence of this lemma we can obtain $u \in C^2(\bar{\Omega})$ by adopting the classical L^p regularity theory of elliptic equations. Lopes and Maris in [48] proved radially symmetrical of minimizers of a generalized Choquard functional

$$E(u) = \frac{1}{2}\int_{\mathbb{R}^n} |\nabla u|^2 dx - \int_{\mathbb{R}^n}\int_{\mathbb{R}^n} \frac{F(u(y))(F(u(x))}{|x-y|^{n-2}} dxdy + \int_{\mathbb{R}^n} H(u(x))dx$$

under the constraint $Q(u) = \int_{\mathbb{R}^n} G(u(x))dx = constant \neq 0$, where $n \geq 3$. They have also proved some regularity results in Lemma [48].

10.1.3 Choquard Equations Involving the $p(.)$-Laplacian

Firstly, let us consider the quasilinear generalization of the Laplace operator that is the p-Laplace operator defined as

$$-\Delta_p u := -\nabla \cdot (|\nabla u|^{p-2}\nabla u), \ 1 < p < \infty.$$

The Choquard equation involving $-\Delta_p$ has been studied in [3–5]. In [4], Alves and Yang studied concentration behaviour of solutions for the following quasilinear Choquard equation

$$-\varepsilon^p \Delta_p u + V(x)|u|^{p-2}u = \varepsilon^{\mu-n}\left(\int_{\mathbb{R}^n} \frac{Q(y)F(u(y))}{|x-y|^\mu}\right) Q(x)f(u) \text{ in } \mathbb{R}^n \tag{10.15}$$

where $1 < p < n, n \geq 3, 0 < \mu < n$, V and Q are two continuous real valued functions on $\mathbb{R}^n$, $F(s)$ is the primitive function of $f(s)$ and ε is a positive parameter. Taking $Q \equiv 1$, Alves and Yang also studied (10.15) in [3]. Recently, Alves and Tavares proved a version of Hardy–Littlewood–Sobolev inequality with variable exponent in [2] in the spirit of variable exponent Lebesgue and Sobolev spaces. Precisely, for $p(x), q(x) \in C^+(\mathbb{R}^n)$ with $p^- := \min\{p(x), 0\} > 1$, and $q^- > 1$, the following holds:

Theorem 10.1.16 *Let* $h \in L^{p^+}(\mathbb{R}^n) \cap L^{p^-}(\mathbb{R}^n)$, $g \in L^{q^+}(\mathbb{R}^n) \cap L^{q^-}(\mathbb{R}^n)$ *and* $\lambda : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ *be a continuous function such that* $0 \leq \lambda^+ \leq \lambda^- < n$ *and*

$$\frac{1}{p(x)} + \frac{\lambda(x,y)}{n} + \frac{1}{q(y)} = 2, \ \forall x, y \in \mathbb{R}^n.$$

Then there exists a constant C independent of h and g such that

$$\left|\int_{\mathbb{R}^n}\int_{\mathbb{R}^n}\frac{h(x)g(y)}{|x-y|^{\mu}}\,dxdy\right| \leq C(\|h\|_{p^+}\|g\|_{q^+}+\|h\|_{p^-}\|g\|_{q^-}).$$

In the spirit of Theorem 10.1.16, authors in [2] proved existence of a solution $u \in W^{1,p(x)}(\mathbb{R}^n)$ to the following quasilinear Choquard equation using variational methods under the subcritical growth conditions on $f(x,u)$:

$$-\varDelta_{p(x)}u+V(x)|u|^{p(x)-2}u=\left(\int_{\varOmega}\frac{F(x,u(x))}{|x-y|^{\lambda(x,y)}}\,dx\right)f(y,u(y)) \text{ in } \mathbb{R}^n, \tag{10.16}$$

where $\varDelta_{p(x)}$ denotes the $p(x)$-Laplacian defined as $-\varDelta_{p(x)}u := -div(|\nabla u|^{p(x)-2}\nabla u)$, V, p, f are real valued continuous functions and F denotes primitive of f with respect to the second variable.

10.2 Choquard Equations Involving the Fractional Laplacian

In this section, we summarize our contributions related to the existence and multiplicity results concerning different Choquard equations, in separate subsections. We employ the variational methods and used some asymptotic estimates to achieve our goal. While dealing with critical exponent in the sense of Hardy–Littlewood–Sobolev inequality, we always consider the upper critical exponent. We denote $\|\cdot\|_r$ as the $L^r(\varOmega)$ norm.

10.2.1 Brezis–Nirenberg Type Existence Results

The fractional Laplacian operator $(-\varDelta)^s$ on the set of the Schwartz class functions is defined as

$$(-\varDelta)^s u(x) = -\text{P.V.}\int_{\mathbb{R}^n}\frac{u(x)-u(y)}{|x-y|^{n+2s}}\,dy$$

(up to a normalizing constant), where P.V. denotes the Cauchy principal value, $s \in (0,1)$ and $n > 2s$. The operator $(-\varDelta)^s$ is the infinitesimal generator of Lévy stable diffusion process. The equations involving this operator arise in the modelling of anomalous diffusion in plasma, population dynamics, geophysical fluid dynamics, flames propagation, chemical reactions in liquids and American options in finance. Motivated by (10.9), in [59] we considered the following doubly nonlocal equation involving the fractional Laplacian with noncompact nonlinearity

$$(-\Delta)^s u = \left(\int_\Omega \frac{|u|^{2^*_{\mu,s}}}{|x-y|^\mu} \mathrm{d}y\right) |u|^{2^*_{\mu,s}-2}u + \lambda u \;\text{ in } \Omega, \quad u = 0 \;\text{ in } \mathbb{R}^n \setminus \Omega, \quad (10.17)$$

where Ω is a bounded domain in $\mathbb{R}^n$ with Lipschitz boundary, λ is a real parameter, $s \in (0, 1)$, $2^*_{\mu,s} = \dfrac{2n-\mu}{n-2s}$, $0 < \mu < n$ and $n > 2s$. Here, $2^*_{\mu,s}$ appears as the upper critical exponent in the sense of Hardy–Littlewood–Sobolev inequality when the function is taken in the fractional Sobolev space $H^s(\mathbb{R}^n) := \{u \in L^2(\mathbb{R}^n) : \|(-\Delta)^{\frac{s}{2}} u\|_2 < \infty\}$ which is continuously embedded in $L^{2^*_s}(\mathbb{R}^n)$ where $2^*_s = \dfrac{2n}{n-2s}$. For more details regarding the fractional Sobolev spaces and its embeddings, we refer [62]. Following are the main results that we have proved.

Theorem 10.2.1 *Let $n \geq 4s$ for $s \in (0, 1)$, then (10.17) has a nontrivial weak solution for every $\lambda > 0$ such that λ is not an eigenvalue of $(-\Delta)^s$ with homogenous Dirichlet boundary condition in $\mathbb{R}^n \setminus \Omega$.*

Theorem 10.2.2 *Let $s \in (0, 1)$ and $2s < n < 4s$, then there exist $\bar{\lambda} > 0$ such that for any $\lambda > \bar{\lambda}$ different from the eigenvalues of $(-\Delta)^s$ with homogenous Dirichlet boundary condition in $\mathbb{R}^n \setminus \Omega$, (10.17) has a nontrivial weak solution.*

Theorem 10.2.3 *Let $\lambda < 0$ and $\Omega \not\equiv \mathbb{R}^n$ be a strictly star shaped bounded domain (with respect to origin) with $C^{1,1}$ boundary, then (10.17) cannot have a nonnegative nontrivial solution.*

Consider the space X defined as

$$X = \left\{u |\; u : \mathbb{R}^n \to \mathbb{R} \text{ is measurable, } u|_\Omega \in L^2(\Omega) \text{ and } \frac{(u(x)-u(y))}{|x-y|^{\frac{n}{2}+s}} \in L^2(Q)\right\},$$

where $Q = \mathbb{R}^{2n} \setminus (\mathscr{C}\Omega \times \mathscr{C}\Omega)$ and $\mathscr{C}\Omega := \mathbb{R}^n \setminus \Omega$ endowed with the norm

$$\|u\|_X = \|u\|_{L^2(\Omega)} + [u]_X\,,$$

where

$$[u]_X = \left(\int_Q \frac{|u(x)-u(y)|^2}{|x-y|^{n+2s}}\, \mathrm{d}x\mathrm{d}y\right)^{\frac{1}{2}}.$$

Then we define $X_0 = \{u \in X : u = 0 \text{ a.e. in } \mathbb{R}^n \setminus \Omega\}$ and we have the Poincare type inequality: there exists a constant $C > 0$ such that $\|u\|_{L^2(\Omega)} \leq C[u]_X$, for all $u \in X_0$. Hence, $\|u\| = [u]_X$ is a norm on X_0. Moreover, X_0 is a Hilbert space and $C^\infty_c(\Omega)$ is dense in X_0. For details on these spaces and variational setup we refer to [65].

Definition 10.2.4 We say that $u \in X_0$ is a weak solution of (10.17) if

$$\int_Q \frac{(u(x)-u(y))(\varphi(x)-\varphi(y))}{|x-y|^{n+2s}}\,dxdy = \int_\Omega\int_\Omega \frac{|u(x)|^{2^*_{\mu,s}}|u(y)|^{2^*_{\mu,s}-2}u(y)\varphi(y)}{|x-y|^{\mu}}\,dxdy + \lambda\int_\Omega u\varphi\,dx, \text{ for every } \varphi\in X_0.$$

The corresponding energy functional associated to the problem (10.17) is given by

$$I_\lambda(u)=I(u):=\frac{\|u\|^2}{2}-\frac{1}{22^*_{\mu,s}}\int_\Omega\int_\Omega \frac{|u(x)|^{2^*_{\mu,s}}|u(y)|^{2^*_{\mu,s}}}{|x-y|^{\mu}}\,dxdy-\frac{\lambda}{2}\int_\Omega |u|^2dx.$$

Using Hardy–Littlewood–Sobolev inequality, we can show that $I\in C^1(X_0,\mathbb{R})$ and the critical points of I corresponds to weak solution of (10.17). We define

$$S_s^H := \inf_{H^s(\mathbb{R}^n)\setminus\{0\}} \frac{\displaystyle\int_{\mathbb{R}^n}\int_{\mathbb{R}^n}\frac{|u(x)-u(y)|^2}{|x-y|^{n+2s}}\,dxdy}{\left(\displaystyle\int_{\mathbb{R}^n}\int_{\mathbb{R}^n}\frac{|u(x)|^{2^*_{\mu,s}}|u(y)|^{2^*_{\mu,s}}}{|x-y|^{\mu}}dxdy\right)^{\frac{1}{2^*_{\mu,s}}}}$$

as the best constant which is achieved if and only if u is of the form

$$C\left(\frac{t}{t^2+|x-x_0|^2}\right)^{\frac{n-2s}{2}}, \quad \text{for all } x\in\mathbb{R}^n,$$

for some $x_0\in\mathbb{R}^n$, $C>0$ and $t>0$. Moreover, $S_s^H\ C(n,\mu)^{\frac{1}{2^*_{\mu,s}}}=S_s$, where S_s is the best constant of the Sobolev imbedding $H^s(\mathbb{R}^n)$ into $L^2(\mathbb{R}^n)$. Using suitable translation and dilation of the minimizing sequence, we proved.

Lemma 10.2.5 *Let*

$$S_s^H(\Omega) := \inf_{X_0\setminus\{0\}} \frac{\displaystyle\int_Q \frac{|u(x)-u(y)|^2}{|x-y|^{n+2s}}\,\mathrm{d}x\mathrm{d}y}{\left(\displaystyle\int_\Omega\int_\Omega\frac{|u(x)|^{2^*_{\mu,s}}|u(y)|^{2^*_{\mu,s}}}{|x-y|^{\mu}}\mathrm{d}x\mathrm{d}y\right)^{\frac{1}{2^*_{\mu,s}}}}.$$

Then $S_s^H(\Omega)=S_s^H$ and $S_s^H(\Omega)$ is never achieved except when $\Omega=\mathbb{R}^n$.

Since (10.17) has a lack of compactness due to the presence of the critical exponent, we needed a Brezis–Lieb type lemma which can be proved in the spirit of (10.4) as follows-

$$\int_{\mathbb{R}^n}\int_{\mathbb{R}^n} \frac{|u_k(x)|^{2^*_{\mu,s}}|u_k(y)|^{2^*_{\mu,s}}}{|x-y|^{\mu}}\,\mathrm{d}x\mathrm{d}y - \int_{\mathbb{R}^n}\int_{\mathbb{R}^n} \frac{|(u_k-u)(x)|^{2^*_{\mu,s}}|(u_k-u)(y)|^{2^*_{\mu,s}}}{|x-y|^{\mu}}\,\mathrm{d}x\mathrm{d}y$$
$$\to \int_{\mathbb{R}^n}\int_{\mathbb{R}^n} \frac{|u(x)|^{2^*_{\mu,s}}|u(y)|^{2^*_{\mu,s}}}{|x-y|^{\mu}}\,\mathrm{d}x\mathrm{d}y \ \text{ as } k \to \infty$$

where $\{u_k\}$ is a bounded sequence in $L^{2^*_s}(\mathbb{R}^n)$ such that $u_k \to u$ almost everywhere in $\mathbb{R}^n$ as $n \to \infty$. Next, we prove the following properties concerning the compactness of Palais–Smale sequences. If $\{u_k\}$ is a Palais–Smale sequence of I at c. Then

(i) $\{u_k\}$ must be bounded in X_0 and its weak limit is a weak solution of (10.17),
(ii) $\{u_k\}$ has a convergent subsequence if

$$c < \frac{n+2s-\mu}{2(2n-\mu)}(S_s^H)^{\frac{2n-\mu}{n+2s-\mu}}.$$

Let us consider the sequence of eigenvalues of the operator $(-\Delta)^s$ with homogenous Dirichlet boundary condition in $\mathbb{R}^n \setminus \Omega$, denoted by

$$0 < \lambda_1 < \lambda_2 \le \lambda_3 \le \cdots \le \lambda_j \le \lambda_{j+1} \le \dots$$

and $\{e_j\}_{j\in\mathbb{N}} \subset L^\infty(\Omega)$ be the corresponding sequence of eigen functions. We also consider this sequence of e_j's to form an orthonormal basis of $L^2(\Omega)$ and orthogonal basis of X_0. We then dealt with the cases $\lambda \in (0, \lambda_1)$ and $\lambda \in (\lambda_r, \lambda_{r+1})$ separately. We assume $0 \in \Omega$ and fix $\delta > 0$ such that $B_\delta \subset \Omega \subset B_{\hat{k}\delta}$, for some $\hat{k} > 1$. Let $\eta \in C^\infty(\mathbb{R}^n)$ be such that $0 \le \eta \le 1$ in $\mathbb{R}^n$, $\eta \equiv 1$ in B_δ and $\eta \equiv 0$ in $\mathbb{R}^n \setminus \Omega$. For $\varepsilon > 0$, we define the function u_ε as follows

$$u_\varepsilon(x) := \eta(x)U_\varepsilon(x),$$

for $x \in \mathbb{R}^n$, where $U_\varepsilon(x) = \varepsilon^{-\frac{(n-2s)}{2}}\left(\frac{u^*\left(\frac{x}{\varepsilon}\right)}{\|u^*\|_{2^*_s}}\right)$ and $u^*(x) = \alpha\left(\beta^2 + \left|\frac{x}{S_s^{\frac{1}{2s}}}\right|^2\right)^{-\frac{n-2s}{2}}$ with $\alpha \in \mathbb{R}\setminus\{0\}$, $\beta > 0$. We obtained the following important asymptotic estimates

Proposition 10.2.6 *The following estimates holds true:*

$$\int_{\mathbb{R}^n} \frac{|u_\varepsilon(x)-u_\varepsilon(y)|^2}{|x-y|^{n+2s}}\,\mathrm{d}x\mathrm{d}y \le \left((C(n,\mu))^{\frac{n-2s}{2n-\mu}} S_s^H\right)^{\frac{n}{2s}} + O(\varepsilon^{n-2s}),$$

$$\left(\int_\Omega\int_\Omega \frac{|u_\varepsilon(x)|^{2^*_{\mu,s}}|u_\varepsilon(y)|^{2^*_{\mu,s}}}{|x-y|^{\mu}}\,\mathrm{d}x\mathrm{d}y\right)^{\frac{n-2s}{2n-\mu}} \le (C(n,\mu))^{\frac{n(n-2s)}{2s(2n-\mu)}}(S_s^H)^{\frac{n-2s}{2s}} + O(\varepsilon^n),$$

and

$$\left(\int_\Omega\int_\Omega \frac{|u_\varepsilon(x)|^{2^*_{\mu,s}}|u_\varepsilon(y)|^{2^*_{\mu,s}}}{|x-y|^\mu}\,\mathrm{d}x\mathrm{d}y\right)^{\frac{n-2s}{2n-\mu}} \geq \left((C(n,\mu))^{\frac{n}{2s}}(S^H_s)^{\frac{2n-\mu}{2s}} - O\left(\varepsilon^n\right)\right)^{\frac{n-2s}{2n-\mu}}.$$

When $n \geq 4s$, we proved that the energy functional I_λ satisfies the Mountain pass geometry if $\lambda \in (0, \lambda_1)$ and Linking Theorem geometry if $\lambda \in (\lambda_r, \lambda_{r+1})$. Also in both the cases, Proposition 10.2.6 helped us to show that for small enough $\varepsilon > 0$

$$\frac{\|u_\varepsilon\|^2 - \lambda\int_\Omega |u_\varepsilon|^2\mathrm{d}x}{\left(\int_\Omega\int_\Omega \frac{|u_\varepsilon(x)|^{2^*_{\mu,s}}|u_\varepsilon(y)|^{2^*_{\mu,s}}}{|x-y|^\mu}\,\mathrm{d}x\mathrm{d}y\right)^{\frac{n-2s}{2n-\mu}}} < S^H_s. \tag{10.18}$$

Then the proof of Theorem 10.2.1 follows by applying Mountain Pass Lemma and Linking Theorem. On the other hand when $2s < n < 4s$, (10.18) could be proved only when $\lambda > \bar{\lambda}$ for some suitable $\bar{\lambda} > 0$, when $\varepsilon > 0$ is sufficiently small. Hence again applying Mountain Pass Lemma and Linking Theorem in this case too, we prove Theorem 10.2.2. To prove Theorem 10.2.3, we first prove that if $\lambda < 0$ then any solution $u \in X_0$ of (10.17) belongs to $L^\infty(\Omega)$ which implied that when Ω is a $C^{1,1}$ domain then $u/\delta^s \in C^\alpha(\bar{\Omega})$ for some $\alpha > 0$ (depending on Ω and s) satisfying $\alpha < \min\{s, 1-s\}$, where $\delta(x) = \mathrm{dist}(x, \partial\Omega)$ for $x \in \Omega$. Then using $(x.\nabla u)$ as a test function in (10.17), we proved the following Pohozaev type identity-

Proposition 10.2.7 *If $\lambda < 0$, Ω be bounded $C^{1,1}$ domain and $u \in L^\infty(\Omega)$ solves* (10.17)*, then*

$$\frac{2s-n}{2}\int_\Omega u(-\Delta)^s u\,\mathrm{d}x - \frac{\Gamma(1+s)^2}{2}\int_{\partial\Omega}\left(\frac{u}{\delta^s}\right)^2(x.\nu)\mathrm{d}\sigma$$
$$= -\left(\frac{2n-\mu}{22^*_{\mu,s}}\int_\Omega\int_\Omega \frac{|u(x)|^{2^*_{\mu,s}}|u(y)|^{2^*_{\mu,s}}}{|x-y|^\mu}\,\mathrm{d}x\mathrm{d}y + \frac{\lambda n}{2}\int_\Omega |u|^2\mathrm{d}x\right),$$

where ν denotes the unit outward normal to $\partial\Omega$ at x and Γ is the Gamma function.

Using Proposition 10.2.7, Theorem 10.2.3 easily followed.

10.2.2 Magnetic Choquard Equations

Very recently Lü [49] studied the problem

$$(-i\nabla + A(x))^2 u + (g_0 + \mu g)(x)u = (|x|^{-\alpha} * |u|^p)|u|^{p-2}u,\ u \in H^1(\mathbb{R}^n, \mathbb{C}), \tag{10.19}$$

where $n \geq 3$, $\alpha \in (0, n)$, $p \in \left(\frac{2n-\alpha}{n}, \frac{2n-\alpha}{n-2}\right)$, $A = (A_1, A_2, \ldots, A_n) : \mathbb{R}^n \to \mathbb{R}^n$ is a vector (or magnetic) potential such that $A \in L^n_{\mathrm{loc}}(\mathbb{R}^n, \mathbb{R}^n)$ and A is continuous at 0,

g_0 and g are real valued functions on $\mathbb{R}^n$ satisfying some necessary conditions and $\mu > 0$. He proved the existence of ground state solution when $\mu \geq \mu^*$, for some $\mu^* > 0$ and concentration behaviour of solutions as $\mu \to \infty$. Salazar in [64] showed existence of vortex type solutions for the stationary nonlinear magnetic Choquard equation

$$(-i\nabla + A(x))^2u + W(x)u = (|x|^{-\alpha} * |u|^p)|u|^{p-2}u \text{ in } \mathbb{R}^n,$$

where $p \in \left[2, 2^*_\alpha\right)$ and $W : \mathbb{R}^n \to \mathbb{R}$ is bounded electric potential. Under some assumptions on decay of A and W at infinity, Cingloni, Sechi and Squassina in [21] showed existence of family of solutions. Schrödinger equations with magnetic field and Choquard type nonlinearity has also been studied in [22, 23]. But the critical case in (10.19) was still open which motivated us to study the problem $(P_{\lambda,\mu})$ in [60]:

$$(P_{\lambda,\mu}) \begin{cases} (-i\nabla + A(x))^2u + \mu g(x)u = \lambda u + (|x|^{-\alpha} * |u|^{2^*_\alpha})|u|^{2^*_\alpha - 2}u \text{ in } \mathbb{R}^n \\ u \in H^1(\mathbb{R}^n, \mathbb{C}) \end{cases}$$

where $n \geq 4$, $2^*_\alpha = \frac{2n-\alpha}{n-2}$, $\alpha \in (0, n)$, $\mu > 0$, $\lambda > 0$, $A = (A_1, A_2, \ldots, A_n) : \mathbb{R}^n \to \mathbb{R}^n$ is a vector(or magnetic) potential such that $A \in L^n_{loc}(\mathbb{R}^n, \mathbb{R}^n)$ and A is continuous at 0 and $g(x)$ satisfies the following assumptions:

(g1) $g \in C(\mathbb{R}^n, \mathbb{R})$, $g \geq 0$ and $\Omega :=$ interior of $g^{-1}(0)$ is a nonempty bounded set with smooth boundary and $\overline{\Omega} = g^{-1}(0)$.

(g2) There exists $M > 0$ such that $\mathscr{L}\{x \in \mathbb{R}^n : \ g(x) \leq M\} < +\infty$, where $\mathscr{L}$ denotes the Lebesgue measure in $\mathbb{R}^n$.

Let us define $-\nabla_A := (-i\nabla + A)$ and

$$H^1_A(\mathbb{R}^n, \mathbb{C}) = \left\{u \in L^2(\mathbb{R}^n, \mathbb{C}) \ : \ \nabla_A u \in L^2(\mathbb{R}^n, \mathbb{C}^n)\right\}.$$

Then $H^1_A(\mathbb{R}^n, \mathbb{C})$ is a Hilbert space with the inner product

$$\langle u, v\rangle_A = \text{Re}\left(\int_{\mathbb{R}^n} (\nabla_A u \overline{\nabla_A v} + u\overline{v}) \, \mathrm{d}x\right),$$

where $\text{Re}(w)$ denotes the real part of $w \in \mathbb{C}$ and $\bar{w}$ denotes its complex conjugate. The associated norm $\| \cdot \|_A$ on the space $H^1_A(\mathbb{R}^n, \mathbb{C})$ is given by

$$\|u\|_A = \left(\int_{\mathbb{R}^n} (|\nabla_A u|^2 + |u|^2) \, \mathrm{d}x\right)^{\frac{1}{2}}.$$

We call $H^1_A(\mathbb{R}^n, \mathbb{C})$ simply $H^1_A(\mathbb{R}^n)$. Let $H^{0,1}_A(\Omega, \mathbb{C})$ (denoted by $H^{0,1}_A(\Omega)$ for simplicity) be the Hilbert space defined by the closure of $C^\infty_c(\Omega, \mathbb{C})$ under the scalar product $\langle u, v\rangle_A = \text{Re}\left(\int_\Omega (\nabla_A u \overline{\nabla_A v} + u\overline{v}) \, \mathrm{d}x\right)$, where $\Omega =$ interior of $g^{-1}(0)$. Thus norm on $H_A^{0,1}(\Omega)$ is given by

$$\|u\|_{H_A^{0,1}(\Omega)} = \left(\int_\Omega (|\nabla_A u|^2 + |u|^2)\, \mathrm{d}x\right)^{\frac{1}{2}}.$$

Let $E = \left\{u \in H_A^1(\mathbb{R}^n) : \int_{\mathbb{R}^n} g(x)|u|^2\, \mathrm{d}x < +\infty\right\}$ be the Hilbert space equipped with the inner product

$$\langle u, v\rangle = \mathrm{Re}\left(\int_{\mathbb{R}^n} \left(\nabla_A u \overline{\nabla_A v}\, \mathrm{d}x + g(x)u\bar{v}\right)\, \mathrm{d}x\right)$$

and the associated norm $\|u\|_E^2 = \int_{\mathbb{R}^n} \left(|\nabla_A u|^2 + g(x)|u|^2\right)\, \mathrm{d}x$. Then $\|\cdot\|_E$ is clearly equivalent to each of the norm $\|u\|_\mu^2 = \int_{\mathbb{R}^n} \left(|\nabla_A u|^2 + \mu g(x)|u|^2\right)\, \mathrm{d}x$ for $\mu > 0$. We have the following well known *diamagnetic inequality* (for detailed proof, see [47], Theorem 7.21).

Theorem 10.2.8 *If* $u \in H_A^1(\mathbb{R}^n)$*, then* $|u| \in H^1(\mathbb{R}^n, \mathbb{R})$ *and*

$$|\nabla|u|(x)| \leq |\nabla u(x) + iA(x)u(x)| \text{ for a.e. } x \in \mathbb{R}^n.$$

So for each $q \in [2, 2^*]$, there exists constant $b_q > 0$ (independent of μ) such that

$$|u|_q \leq b_q\|u\|_\mu, \text{ for any } u \in E, \tag{10.20}$$

where $|\cdot|_q$ denotes the norm in $L^q(\mathbb{R}^n, \mathbb{C})$ and $2^* = \frac{2n}{n-2}$ is the Sobolev critical exponent. Also $H_A^1(\Omega) \hookrightarrow L^q(\Omega, \mathbb{C})$ is continuous for each $1 \leq q \leq 2^*$ and compact when $1 \leq q < 2^*$. We denote $\lambda_1(\Omega) > 0$ as the best constant of the embedding $H_A^{0,1}(\Omega)$ into $L^2(\Omega, \mathbb{C})$ given by

$$\lambda_1(\Omega) = \inf_{u \in H_A^{0,1}(\Omega)} \left\{\int_\Omega |\nabla_A u|^2\, \mathrm{d}x : \int_\Omega |u|^2\, \mathrm{d}x = 1\right\}$$

which is also the first eigenvalue of $-\Delta_A := (-i\nabla + A)^2$ on Ω with boundary condition $u = 0$. In [60], we consider the problem

$$(P_\lambda) \quad (-i\nabla + A(x))^2 u = \lambda u + (|x|^{-\alpha} * |u|^{2_\alpha^*})|u|^{2_\alpha^*-2}u \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega$$

and proved the following main results:

Theorem 10.2.9 *For every* $\lambda \in (0, \lambda_1(\Omega))$ *there exists a* $\mu(\lambda) > 0$ *such that* $(P_{\lambda,\mu})$ *has a least energy solution* u_μ *for each* $\mu \geq \mu(\lambda)$.

Theorem 10.2.10 *Let* $\{u_m\}$ *be a sequence of nontrivial solutions of* (P_{λ,μ_m}) *with* $\mu_m \to \infty$ *and* $I_{\lambda,\mu_m}(u_m) \to c < \frac{n+2-\alpha}{2(2n-\alpha)} S_A^{\frac{2n-\alpha}{n+2-\alpha}}$ *as* $m \to \infty$*. Then* u_m *concentrates at a solution of* (P_λ).

We give some definitions below-

Definition 10.2.11 We say that a function $u \in E$ is a weak solution of $(P_{\lambda,\mu})$ if

$$\mathrm{Re}\left(\int_{\mathbb{R}^n} \nabla_A u \overline{\nabla_A v}\, \mathrm{d}x + \int_{\mathbb{R}^n} (\mu g(x) - \lambda) u\bar{v}\, \mathrm{d}x - \int_{\mathbb{R}^n} (|x|^{-\alpha} * |u|^{2^*_\alpha})|u|^{2^*_\alpha - 2} u\bar{v}\, \mathrm{d}x\right) = 0$$

for all $v \in E$.

Definition 10.2.12 A solution u of $(P_{\lambda,\mu})$ is said to be a least energy solution if the energy functional

$$I_{\lambda,\mu}(u) = \int_{\mathbb{R}^n} \left(\frac{1}{2}\left(|\nabla_A u|^2 + (\mu g(x) - \lambda)|u|^2\right) - \frac{1}{22^*_\alpha}(|x|^{-\alpha} * |u|^{2^*_\alpha})|u|^{2^*_\alpha}\right) \mathrm{d}x$$

achieves its minimum at u over all the nontrivial solutions of $(P_{\lambda,\mu})$.

Definition 10.2.13 A sequence of solutions $\{u_k\}$ of (P_{λ,μ_k}) is said to concentrate at a solution u of (P_λ) if a subsequence converges strongly to u in $H^1_A(\mathbb{R}^n)$ as $\mu_k \to \infty$.

We first proved the following Lemma.

Lemma 10.2.14 *Suppose $\mu_m \geq 1$ and $u_m \in E$ be such that $\mu_m \to \infty$ as $m \to \infty$ and there exists a $K > 0$ such that $\|u_m\|_{\mu_m} < K$, for all $m \in \mathbb{N}$. Then there exists a $u \in H^{0,1}_A(\Omega)$ such that (upto a subsequence), $u_m \rightharpoonup u$ weakly in E and $u_m \to u$ strongly in $L^2(\mathbb{R}^n)$ as $m \to \infty$.*

Then we define an operator $T_\mu := -\Delta_A + \mu g(x)$ on E given by

$$\big(T_\mu(u), v\big) = \mathrm{Re}\left(\int_{\mathbb{R}^n} (\nabla_A u \overline{\nabla_A v} + \mu g(x) u\bar{v})\, \mathrm{d}x\right).$$

Clearly T_μ is a self adjoint operator and if $a_\mu := \inf \sigma(T_\mu)$, i.e. the infimum of the spectrum of T_μ, then a_μ can be characterized as

$$0 \leq a_\mu = \inf\{\big(T_\mu(u), u\big) : \ u \in E, \ \|u\|_{L^2} = 1\} = \inf\{\|u\|^2_\mu : \ u \in E, \ \|u\|_{L^2} = 1\}.$$

Then considering a minimizing sequence of a_μ, we were able to prove that for each $\lambda \in (0, \lambda_1(\Omega))$, there exists a $\mu(\lambda) > 0$ such that $a_\mu \geq (\lambda + \lambda_1(\Omega))/2$ whenever $\mu \geq \mu(\lambda)$. As a consequence

$$\big((T_\mu - \lambda)u, u\big) \geq \beta_\lambda \|u\|^2_\mu$$

for all $u \in E$, $\mu \geq \mu(\lambda)$, where $\beta_\lambda := (\lambda_1(\Omega) - \lambda)/(\lambda_1(\Omega) + \lambda)$. We fix $\lambda \in (0, \lambda_1(\Omega))$ and $\mu \geq \mu(\lambda)$. Using standard techniques, we established the following concerning any Palais Smale sequence $\{u_k\}$ of $I_{\lambda,\mu}$-

(i) $\{u_m\}$ must be bounded in E and its weak limit is a solution of $(P_{\lambda,\mu})$,
(ii) $\{u_m\}$ has a convergent subsequence when c satisfies

$$c \in \left(-\infty, \frac{n+2-\alpha}{2(2n-\alpha)} S_A^{\frac{2n-\alpha}{n+2-\alpha}}\right)$$

where S_A is defined as follows

$$S_A = \inf_{u \in H_A^1(\mathbb{R}^n)\setminus\{0\}} \frac{\displaystyle\int_{\mathbb{R}^n} |\nabla_A u|^2 \, dx}{\displaystyle\int_{\mathbb{R}^n} (|x|^{-\alpha} * |u|^{2_\alpha^*})|u|^{2_\alpha^*} \, dx}.$$

Using asymptotic estimates and using the family $U_\varepsilon(x) = (n(n-2))^{\frac{n-2}{4}} \left(\frac{\varepsilon}{\varepsilon^2+|x|^2}\right)^{\frac{n-2}{4}}$, we showed that-

Theorem 10.2.15 *If $g \geq 0$ and $A \in L^n_{loc}(\mathbb{R}^n, \mathbb{R}^n)$, then the infimum S_A is attained if and only if curl $A \equiv 0$.*

Our next step was to introduce the Nehari manifold

$$\mathcal{N}_{\lambda,\mu} = \left\{u \in E \setminus \{0\} : \; \langle I'_{\lambda,\mu}(u), u\rangle = 0\right\}$$

and consider the minimization problem $k_{\lambda,\mu} := \inf_{u \in \mathcal{N}_{\lambda,\mu}} I_{\lambda,\mu}(u)$. Using the family $\{U_\varepsilon\}$, we showed that

$$k_{\lambda,\mu} < \frac{n+2-\alpha}{2(2n-\alpha)} S_A^{\frac{2n-\alpha}{n+2-\alpha}}.$$

Then the proof of Theorem 10.2.9 followed by using the Ekeland Variational Principle over $\mathcal{N}_{\lambda,\mu}$. The proof of Theorem 10.2.10 followed from Lemma 10.2.14 and the Brezis–Lieb type lemma for the Riesz potentials.

Remark 10.1 These results can be generalized to the problems involving fractional magnetic operators:

$$(P^s_{\lambda,\mu}) \begin{cases} (-\Delta)^s_A u + \mu g(x) u = \lambda u + (|x|^{-\alpha} * |u|^{2^*_{\alpha,s}})|u|^{2^*_{\alpha,s}-2}u \text{ in } \mathbb{R}^n, \\ u \in H^s_A(\mathbb{R}^n, \mathbb{C}) \end{cases}$$

where $n \geq 4s$, $s \in (0,1)$ and $\alpha \in (0,n)$. Here $2^*_{\alpha,s} = \frac{2n-\alpha}{n-2s}$ is the critical exponent in the sense of Hardy–Littlewood–Sobolev inequality. We assume the same conditions on A and g as before. For $u \in C_c^\infty(\Omega)$, the fractional magnetic operator $(-\Delta)^s_A$, up to a normalization constant, is defined by

$$(-\Delta)^s_A u(x) = 2 \lim_{\varepsilon \to 0^+} \int_{\mathbb{R}^n \setminus B_\varepsilon(x)} \frac{u(x) - e^{i(x-y)\cdot A\left(\frac{x+y}{2}\right)}u(y)}{|x-y|^{n+2s}} dy$$

for all $x \in \mathbb{R}^n$. With proper functional setting, we can prove the existence and concentration results for the problem $(P^s_{\lambda,\mu})$ employing the same arguments as in the local magnetic operator case.

10.2.3 Singular Problems Involving Choquard Nonlinearity

The paper by Crandal, Rabinowitz and Tartar [24] is the starting point on the semilinear problem with singular nonlinearity. A lot of work has been done related to the existence and multiplicity results for singular problems, see [39–41]. Using splitting Nehari manifold technique, authors in [40] studied the existence of multiple solutions of the problem:

$$-\Delta u = \lambda u^{-q} + u^p,\ u > 0 \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega, \tag{10.21}$$

where Ω is smooth bounded domain in $\mathbb{R}^n, n \geq 1, p = 2^* - 1, \lambda > 0$ and $0 < q < 1$. In [39], Haitao studied the equation (10.21) for $n \geq 3$, $1 < p \leq 2^* - 1$ and showed the existence of two positive solutions for maximal interval of the parameter λ using monotone iterations and mountain pass lemma. But the singular problem involving Choquard nonlinearity was completely open until we studied the following problem in [58]

$$(P_\lambda): \qquad -\Delta u = \lambda u^{-q} + \left(\int_\Omega \frac{|u(y)|^{2^*_\mu}}{|x-y|^\mu}\mathrm{d}y\right)|u|^{2^*_\mu - 2}u,\ u > 0 \text{ in } \Omega,\ u = 0 \text{ on } \partial\Omega,$$

where $\Omega \subset \mathbb{R}^n, n > 2$ be a bounded domain with smooth boundary $\partial\Omega, \lambda > 0,\ 0 < q < 1,\ 0 < \mu < n$ and $2^*_\mu = \frac{2n-\mu}{n-2}$. The main difficulty in treating (P_λ) is the presence of singular nonlinearity along with critical exponent in the sense of Hardy–Littlewood–Sobolev inequality which is nonlocal in nature. The energy functional no longer remains differentiable due to presence of singular nonlinearity, so usual minimax theorems are not applicable. Also the critical exponent term being nonlocal adds on the difficulty to study the Palais–Smale level around a nontrivial critical point.

Definition 10.2.16 We say that $u \in H^1_0(\Omega)$ is a positive weak solution of (P_λ) if $u > 0$ in Ω and

$$\int_\Omega (\nabla u \nabla \psi - \lambda u^{-q}\psi)\,\mathrm{d}x - \int_\Omega\int_\Omega \frac{|u(x)|^{2^*_\mu}|u(y)|^{2^*_\mu - 2}u(y)\psi(y)}{|x-y|^\mu}\,\mathrm{d}x\mathrm{d}y = 0 \tag{10.22}$$

for all $\psi \in C^\infty_c(\Omega)$.

We define the functional associated to (P_λ) as $I : H^1_0(\Omega) \to (-\infty, \infty]$ by

$$I(u)=\frac{1}{2}\int_\Omega |\nabla u|^2\,\mathrm{d}x-\frac{\lambda}{1-q}\int_\Omega |u|^{1-q}\mathrm{d}x-\frac{1}{22^*_\mu}\int_\Omega\int_\Omega \frac{|u(x)|^{2^*_\mu}|u(y)|^{2^*_\mu}}{|x-y|^\mu}\,\mathrm{d}x\mathrm{d}y,$$

for $u\in H^1_0(\Omega)$. For each $0<q<1$, we set $H_+=\{u\in H^1_0(\Omega): u\geq 0\}$ and

$$H_{+,q}=\{u\in H_+ : u\not\equiv 0,\ |u|^{1-q}\in L^1(\Omega)\}=H_+\setminus\{0\}.$$

For each $u\in H_{+,q}$ we define the fiber map $\phi_u:\mathbb{R}^+\to\mathbb{R}$ by $\phi_u(t)=I_\lambda(tu)$. Then we proved the following:

Theorem 10.2.17 *Assume* $0<q<1$ *and let* Λ *be a constant defined by*

$$\Lambda=\sup\big\{\lambda>0 : \textit{for each } u\in H_{+,q}\backslash\{0\},\ \phi_u(t) \textit{ has two critical points in } (0,\infty)$$
$$\textit{and } \sup\left\{\int_\Omega |\nabla u|^2\,\mathrm{d}x\ :\ u\in H_{+,q}, \phi_u'(1)=0,\ \phi_u''(1)>0\right\}\leq (2^*_\mu S_{H,L}^{2^*_\mu})^{\frac{1}{2^*_\mu-1}}\right\}.$$

Then $\Lambda>0$.

Using the variational methods on the Nehari manifold, we proved the following multiplicity result.

Theorem 10.2.18 *For all* $\lambda\in(0,\Lambda)$, (P_λ) *has two positive weak solutions* u_λ *and* v_λ *in* $C^\infty(\Omega)\cap L^\infty(\Omega)$.

We also have that if u is a positive weak solution of (P_λ), then u is a classical solution in the sense that $u\in C^\infty(\Omega)\cap C(\bar{\Omega})$. We define $\delta:\Omega\to[0,\infty)$ by $\delta(x)=\inf\{|x-y| : y\in\partial\Omega\}$, for each $x\in\Omega$.

Theorem 10.2.19 *Let* u *be a positive weak solution of* (P_λ), *then there exist* $K,\ L>0$ *such that* $L\delta\leq u\leq K\delta$ *in* Ω.

We define the Nehari manifold

$$\mathcal{N}_\lambda=\{u\in H_{+,q} | \big\langle I'(u),u\big\rangle=0\}$$

and show that I is coercive and bounded below on $\mathcal{N}_\lambda$. It is easy to see that the points in $\mathcal{N}_\lambda$ are corresponding to critical points of ϕ_u at $t=1$. So, we divided $\mathcal{N}_\lambda$ in three sets corresponding to local minima, local maxima and points of inflexion

$$\mathcal{N}_\lambda^+=\{t_0u\in\mathcal{N}_\lambda |\ t_0>0,\ \phi_u'(t_0)=0,\ \phi_u''(t_0)>0\},$$
$$\mathcal{N}_\lambda^-=\{t_0u\in\mathcal{N}_\lambda |\ t_0>0,\ \phi_u'(t_0)=0,\ \phi_u''(t_0)<0\}$$

and $\mathcal{N}_\lambda^0=\{u\in\mathcal{N}_\lambda|\phi_u'(1)=0,\ \phi_u''(1)=0\}$. We aimed at showing that the minimizers of I over $\mathcal{N}^+$ and $\mathcal{N}^-$ forms a weak solution of (P_λ). We briefly describe the key steps to show this. Using the fibering map analysis, we proved that there exist $\lambda_*>0$ such that for each $u\in H_{+,q}\backslash\{0\}$, there is unique t_1 and t_2 with the property that $t_1<t_2$,

$t_1u \in \mathcal{N}_\lambda^+$ and $t_2u \in \mathcal{N}_\lambda^-$, for all $\lambda \in (0, \lambda_*)$. This implied Theorem 10.2.17. Also $\mathcal{N}_\lambda^0 = \{0\}$ for all $\lambda \in (0, \lambda_*)$. Then it is easy to see that $\sup\{\|u\| : u \in \mathcal{N}_\lambda^+\} < \infty$ and $\inf\{\|v\| : v \in \mathcal{N}_\lambda^-\} > 0$. Suppose u and v are minimizers of I on $\mathcal{N}_\lambda^+$ and $\mathcal{N}_\lambda^-$ respectively. Then for each $w \in H_+$, we showed $u^{-q}w, v^{-q}w \in L^1(\Omega)$ and

$$\int_\Omega (\nabla u \nabla w - \lambda u^{-q} w)\, \mathrm{d}x - \int_\Omega \int_\Omega \frac{|u(y)|^{2^*_\mu}|u(x)|^{2^*_\mu - 2}u(x)w(x)}{|x-y|^\mu}\, \mathrm{d}y\mathrm{d}x \geq 0, \tag{10.23}$$

$$\int_\Omega (\nabla v \nabla w - \lambda v^{-q} w)\, \mathrm{d}x - \int_\Omega \int_\Omega \frac{|v(y)|^{2^*_\mu}|u(x)|^{2^*_\mu - 2}v(x)w(x)}{|x-y|^\mu}\, \mathrm{d}y\mathrm{d}x \geq 0. \tag{10.24}$$

Particularly, $u,\ v > 0$ almost everywhere in Ω. Then the claim followed using the Gateaux differentiability of I. Lastly, the proof of Theorem 10.2.18 followed by proving that I achieves its minimum over the sets $\mathcal{N}_\lambda^+$ and $\mathcal{N}_\lambda^-$.

In the regularity section, firstly, we showed that (10.22) holds for all $\psi \in H_0^1(\Omega)$ and each positive weak solution u of (P_λ) belongs to $L^\infty(\Omega)$. Under the assumption that there exist $a \geq 0$, $R \geq 0$ and $q \leq s < 1$ such that $\Delta\delta \leq R\delta^{-s}$ in $\Omega_a := \{x \in \Omega, \delta(x) \leq a\}$, using appropriate test functions, we proved that there exist $K > 0$ such that $u \leq K\delta$ in Ω. To get the lower bound on u with respect to δ, following result from [16] plays a crucial role.

Lemma 10.2.20 *Let Ω be a bounded domain in $\mathbb{R}^n$ with smooth boundary $\partial\Omega$. Let $u \in L^1_{loc}(\Omega)$ and assume that for some $k \geq 0$, u satisfies, in the sense of distributions*

$$-\Delta u + ku \geq 0 \ in\ \Omega, \quad u \geq 0 \ in\ \Omega.$$

Then either $u \equiv 0$, or there exists $\varepsilon > 0$ such that $u(x) \geq \varepsilon\delta(x),\ x \in \Omega$.

Additionally, we also prove that the solution can be more regular in a restricted range of q.

Lemma 10.2.21 *Let $q \in (0, \frac{1}{n})$ and let $u \in H_0^1(\Omega)$ be a positive weak solution of (P_λ), then $u \in C^{1+\alpha}(\bar{\Omega})$ for some $0 < \alpha < 1$.*

10.3 System of Equations with Choquard Type Nonlinearity

In this section, we briefly illustrate some existence and multiplicity results proved concerning the system of Choquard equations with nonhomogeneous terms. We consider the nonlocal operator that is the fractional Laplacian and since the Choquard nonlinearity is also a nonlocal one, such problems are often called 'doubly nonlocal problems'. We employ the method of Nehari manifold to achieve our goal.

10.3.1 Doubly Nonlocal p-Fractional Coupled Elliptic System

The p-fractional Laplace operator is defined as

$$(-\Delta)_p^s u(x) = 2 \lim_{\varepsilon \searrow 0} \int_{|x|>\varepsilon} \frac{|u(x)-u(y)|^{p-2}(u(x)-u(y))}{|x-y|^{n+sp}}\, dy,\ \forall x \in \mathbb{R}^n,$$

which is nonlinear and nonlocal in nature. This definition matches to linear fractional Laplacian operator $(-\Delta)^s$, up to a normalizing constant depending on n and s, when $p=2$. The operator $(-\Delta)_p^s$ is degenerate when $p>2$ and singular when $1<p<2$. For details, refer [62]. Our concern lies in the nonhomogenous Choquard equations and system of equations. Recently, authors in [66, 72] showed multiplicity of positive solutions for a nonhomogeneous Choquard equation using Nehari manifold. The motivation behind such problems lies in the famous article by Tarantello [68] where author used the structure of associated Nehari manifold to obtain the multiplicity of solutions for the following nonhomogeneous Dirichlet problem on bounded domain Ω

$$-\Delta u = |u|^{2^*-2}u + f \text{ in } \Omega,\ u = 0 \text{ on } \partial\Omega.$$

System of elliptic equations involving p-fractional Laplacian has been studied in [18, 19] using Nehari manifold techniques. Very recently, Guo et al. [38] studied a nonlocal system involving fractional Sobolev critical exponent and fractional Laplacian. There are not many results on elliptic systems with nonhomogeneous nonlinearities in the literature but we cite [20, 28, 69] as some very recent works on the study of fractional elliptic systems.

Motivated by these articles, we consider the following nonhomogenous quasilinear system of equations with perturbations involving p-fractional Laplacian in [61]:

Let $p \geq 2, s \in (0,1), n > sp,$ $\mu \in (0,n),$ $\frac{p}{2}\left(2-\frac{\mu}{n}\right) < q < \frac{p_s^*}{2}\left(2-\frac{\mu}{n}\right)$, $\alpha, \beta, \gamma > 0$,

$$(P)\begin{cases} (-\Delta)_p^s u + a_1(x)u|u|^{p-2} = \alpha(|x|^{-\mu} * |u|^q)|u|^{q-2}u + \beta(|x|^{-\mu} * |v|^q)|u|^{q-2}u \\ \qquad\qquad\qquad\qquad +f_1(x) \text{ in } \mathbb{R}^n, \\ (-\Delta)_p^s v + a_2(x)v|v|^{p-2} = \gamma(|x|^{-\mu} * |v|^q)|v|^{q-2}v + \beta(|x|^{-\mu} * |u|^q)|v|^{q-2}v \\ \qquad\qquad\qquad\qquad +f_2(x) \text{ in } \mathbb{R}^n, \end{cases}$$

where $0 < a_i \in C^1(\mathbb{R}^n, \mathbb{R})$, $i = 1, 2$ and $f_1, f_2 : \mathbb{R}^n \to \mathbb{R}$ are perturbations. Here $p_s^* = \frac{np}{n-sp}$ is the critical exponent associated with the embedding of the fractional Sobolev space $W^{s,p}(\mathbb{R}^n)$ into $L^{p_s^*}(\mathbb{R}^n)$. Wang et. al in [71] studied the problem (P) in the local case $s = 1$ and obtained a partial multiplicity results. We improved their results and showed the multiplicity results with a weaker assumption (10.25) of f_1 and f_2 below. For $i = 1, 2$ we introduce the spaces

$$Y_i := \left\{ u \in W^{s,p}(\mathbb{R}^n) : \int_{\mathbb{R}^n} a_i(x)|u|^p \, dx < +\infty \right\}$$

then Y_i are Banach spaces equipped with the norm

$$\|u\|_{Y_i}^p = \int_{\mathbb{R}^n}\int_{\mathbb{R}^n} \frac{|u(x)-u(y)|^p}{|x-y|^{n+sp}} dxdy + \int_{\mathbb{R}^n} a_i(x)|u|^p dx.$$

We define the product space $Y = Y_1 \times Y_2$ which is a reflexive Banach space with the norm

$$\|(u,v)\|^p := \|u\|_{Y_1}^p + \|v\|_{Y_2}^p,$$

for all $(u, v) \in Y$. We assume the following condition on a_i, for $i = 1, 2$

(A) $a_i \in C(\mathbb{R}^n)$, $a_i > 0$ and there exists $M_i > 0$ such that $\mu(\{x \in \mathbb{R}^n : a_i \leq M_i\}) < \infty$.

Then under the condition (A) on a_i, for $i = 1, 2$, we get Y_i is continuously imbedded into $L^r(\mathbb{R}^n)$ for $r \in [p, p_s^*]$. To obtain our results, we assumed the following condition on perturbation terms:

$$\int_{\mathbb{R}^n} (f_1 u + f_2 v) < C_{p,q} \left(\frac{2q+p-1}{4pq} \right) \|(u,v)\|^{\frac{p(2q-1)}{2q-p}} \tag{10.25}$$

for all $(u, v) \in Y$ such that

$$\int_{\mathbb{R}^n} \left(\alpha(|x|^{-\mu} * |u|^q)|u|^q + 2\beta(|x|^{-\mu} * |u|^q)|v|^q + \gamma(|x|^{-\mu} * |v|^q)|v|^q \right) dx = 1$$

and

$$C_{p,q} = \left(\frac{p-1}{2q-1} \right)^{\frac{2q-1}{2q-p}} \left(\frac{2q-p}{p-1} \right).$$

It is easy to see that $2q > p\left(\frac{2n-\mu}{n}\right) > p-1 > \frac{p-1}{2p-1}$ which implies $\frac{2q+p-1}{4pq} < 1$. So (10.25) implies that

$$\int_{\mathbb{R}^n} (f_1 u + f_2 v) < C_{p,q} \|(u,v)\|^{\frac{p(2q-1)}{2q-p}} \tag{10.26}$$

which we used more frequently rather than our actual assumption (10.25). Now, the main results goes as follows.

Theorem 10.3.1 *Suppose* $\frac{p}{2}\left(\frac{2n-\mu}{n}\right) < q < \frac{p}{2}\left(\frac{2n-\mu}{n-sp}\right)$, $\mu \in (0, n)$ *and* (A) *holds true. Let* $0 \not\equiv f_1, f_2 \in L^{\frac{p}{p-1}}(\mathbb{R}^n)$ *satisfies* (10.25) *then* (P) *has at least two weak* *solutions, in which one forms a local minimum of J on Y. Moreover if $f_1, f_2 \geq 0$ then this solution is a nonnegative weak solution.*

If $u, \phi \in W^{s,p}(\mathbb{R}^n)$, we use the notation $\langle u, \phi\rangle$ to denote

$$\langle u, \phi\rangle := \int_{\mathbb{R}^n}\int_{\mathbb{R}^n} \frac{(u(x)-u(y))|u(x)-u(y)|^{p-2}(\phi(x)-\phi(y))}{|x-y|^{n+sp}}dxdy.$$

Definition 10.3.2 A pair of functions $(u, v) \in Y$ is said to be a weak solution to (P) if

$$\begin{aligned}
&\langle u, \phi_1\rangle + \int_{\mathbb{R}^n} a_1(x)u|u|^{p-2}\phi_1\, dx + \langle v, \phi_2\rangle + \int_{\mathbb{R}^n} a_2(x)v|v|^{p-2}\phi_2\, dx\\
&\quad - \alpha\int_{\mathbb{R}^n}(|x|^{-\mu} * |u|^q)u|u|^{q-2}\phi_1\, dx - \gamma\int_{\mathbb{R}^n}(|x|^{-\mu} * |v|^q)v|v|^{q-2}\phi_2\, dx\\
&\quad - \beta\int_{\mathbb{R}^n}(|x|^{-\mu} * |v|^q)u|u|^{q-2}\phi_1\, dx - \beta\int_{\mathbb{R}^n}(|x|^{-\mu} * |u|^q)v|v|^{q-2}\phi_2\, dx\\
&\quad - \int_{\mathbb{R}^n}(f_1\phi_1 + f_2\phi_2)\, dx = 0,\ \forall\ (\phi_1, \phi_2) \in Y.
\end{aligned}$$

Thus, we define the energy functional corresponding to (P) as

$$\begin{aligned}
J(u, v) &= \frac{1}{p}\|(u, v)\|^p - \frac{1}{2q}\int_{\mathbb{R}^n}\left(\alpha(|x|^{-\mu} * |u|^q)|u|^q + \beta(|x|^{-\mu} * |u|^q)|v|^q\right) dx\\
&\quad - \frac{1}{2q}\int_{\mathbb{R}^n}\left(\beta(|x|^{-\mu} * |v|^q)|u|^q + \gamma(|x|^{-\mu} * |v|^q)|v|^q\right) dx - \int_{\mathbb{R}^n}(f_1u + f_2v)dx\\
&= \frac{1}{p}\|(u, v)\|^p - \frac{1}{2q}\int_{\mathbb{R}^n}\left(\alpha(|x|^{-\mu} * |u|^q)|u|^q + 2\beta(|x|^{-\mu} * |u|^q)|v|^q\right.\\
&\quad \left.+\gamma(|x|^{-\mu} * |v|^q)|v|^q\right) dx - \int_{\mathbb{R}^n}(f_1u + f_2v)dx.
\end{aligned}$$

Clearly, weak solutions to (P) corresponds to the critical points of J. To find the critical points of J, we constraint our functional J on the Nehari manifold

$$\mathcal{N} = \{(u, v) \in Y :\ (J'(u, v), (u, v)) = 0\},$$

where

$$\begin{aligned}
(J'(u, v), (u, v)) =& \|(u, v)\|^p - \int_{\mathbb{R}^n}\left(\alpha(|x|^{-\mu} * |u|^q)|u|^q + 2\beta(|x|^{-\mu} * |u|^q)|v|^q\right.\\
&\left.+\gamma(|x|^{-\mu} * |v|^q)|v|^q\right) dx - \int_{\mathbb{R}^n}(f_1u + f_2v)dx.
\end{aligned}$$

Clearly, every nontrivial weak solution to (P) belongs to $\mathcal{N}$. Denote $I(u, v) = (J'(u, v), (u, v))$ and subdivide the set $\mathcal{N}$ into three sets as follows:

$$\mathcal{N}^{\pm} = \{(u, v) \in \mathcal{N} : \pm(I'(u, v), (u, v)) > 0\},$$

$$\mathcal{N}^{0} = \{(u, v) \in \mathcal{N} : (I'(u, v), (u, v)) = 0\}.$$

Then $\mathcal{N}^0$ contains the element $(0, 0)$ and $\mathcal{N}^+ \cup \mathcal{N}^0$ and $\mathcal{N}^- \cup \mathcal{N}^0$ are closed subsets of Y. For $(u, v) \in Y$, we define the fibering map $\varphi : (0, \infty) \to \mathbb{R}$ as $\varphi(t) = J(tu, tv)$. One can easily check that $(tu, tv) \in \mathcal{N}$ if and only if $\varphi'(t) = 0$, for $t > 0$ and $\mathcal{N}^+$, $\mathcal{N}^-$ and $\mathcal{N}^0$ can also be written as

$$\mathcal{N}^{\pm} = \{(tu, tv) \in \mathcal{N} : \varphi''(t) \gtrless 0\}, \text{ and } \mathcal{N}^{0} = \{(tu, tv) \in \mathcal{N} : \varphi''(t) = 0\}.$$

We showed that J becomes coercive and bounded from below on $\mathcal{N}$. By analyzing the fiber maps $\varphi_{u,v}(t)$ we proved that if (10.25) holds, then $\mathcal{N}_0 = \{(0, 0)\}$ and $\mathcal{N}^-$ is a closed set. By Lagrange multiplier method, we showed that minimizers of J over $\mathcal{N}^+$ and $\mathcal{N}^-$ are the weak solutions of (P). So our problem reduced to minimization problem is given below.

$$\Upsilon^+ := \inf_{(u,v)\in\mathcal{N}^+} J(u, v), \text{ and } \Upsilon^- := \inf_{(u,v)\in\mathcal{N}^-} J(u, v).$$

Using again the map $\varphi_{u,v}$, we could show that $\Upsilon^+ < 0$ whereas $\Upsilon^- > 0$. Our next task was to consider

$$\Upsilon := \inf_{(u,v)\in\mathcal{N}} J(u, v)$$

and show that there exist a constant $C_1 > 0$ such that $\Upsilon \le -\frac{(2q-p)(2qp-2q-p)}{4pq^2} C_1$. Our next result was crucial one, which concerns another minimization problem.

Lemma 10.3.3 *For* $0 \neq f_1, f_2 \in L^{\frac{p}{p-1}}(\mathbb{R}^n)$,

$$\inf_{Q} \left(C_{p,q} \|(u, v)\|^{\frac{p(2q-1)}{2q-p}} - \int_{\mathbb{R}^n} (f_1 u + f_2 v)\, dx \right) := \delta$$

is achieved, where $Q = \{(u, v) \in Y : L(u, v) = 1\}$. *Also if* f_1, f_2 *satisfies* (10.25), *then* $\delta > 0$.

After this, using the Ekeland variational principle we proved the existence of a Palais Smale sequence for J at the levels Υ and Υ^-. Keeping this altogether, we could prove that Υ and Υ^- are achieved by some functions (u_0, v_0) and (u_1, v_1) where (u_0, v_0) lies in $\mathcal{N}^+$ and forms a local minimum of J. The non negativity of (u_i, v_i) for $i = 0, 1$ was showed using the modulus function $(|u_i|, |v_i|)$ and their corresponding fiber maps. Hence we conclude our main result, Theorem 10.3.1.

10.3.2 Doubly Nonlocal System with Critical Nonlinearity

In this section, we illustrate our results concerning a system of Choquard equation with Hardy–Littlewood–Sobolev critical nonlinearity which involves the fractional Laplacian. Precisely, we consider the following problem in [35]

$$(P_{\lambda,\delta})\begin{cases}(-\varDelta)^s u = \lambda|u|^{q-2}u + \left(\int_\varOmega \frac{|v(y)|^{2^*_\mu}}{|x-y|^\mu}\,\mathrm{d}y\right)|u|^{2^*_\mu-2}u \text{ in } \varOmega\\ (-\varDelta)^s v = \delta|v|^{q-2}v + \left(\int_\varOmega \frac{|u(y)|^{2^*_\mu}}{|x-y|^\mu}\,\mathrm{d}y\right)|v|^{2^*_\mu-2}v \text{ in } \varOmega\\ \qquad u = v = 0 \text{ in } \mathbb{R}^n\setminus\varOmega,\end{cases}$$

where $\varOmega$ is a smooth bounded domain in $\mathbb{R}^n$, $n > 2s$, $s \in (0,1)$, $\mu \in (0,n)$, $2^*_\mu = \dfrac{2n-\mu}{n-2s}$ is the upper critical exponent in the Hardy–Littlewood–Sobolev inequality, $1 < q < 2, \lambda, \delta > 0$ are real parameters. As we illustrated some literature on system of elliptic equation involving fractional Laplacian in the last subsection, it was an open question regarding the existence and multiplicity result for system of Choquard equation with Hardy–Littlewood–Sobolev critical nonlinearity, even in the local case $s = 1$. Using the Nehari manifold technique, we prove the following main result.

Theorem 10.3.4 *Assume* $1 < q < 2$ *and* $0 < \mu < n$ *then there exists positive constants* $\varTheta$ *and* $\varTheta_0$ *such that*

1. *if* $\mu \leq 4s$ *and* $0 < \lambda^{\frac{2}{2-q}} + \delta^{\frac{2}{2-q}} < \varTheta$, *the system* $(P_{\lambda,\delta})$ *admits at least two nontrivial solutions,*
2. *if* $\mu > 4s$ *and* $0 < \lambda^{\frac{2}{2-q}} + \delta^{\frac{2}{2-q}} < \varTheta_0$, *the system* $(P_{\lambda,\delta})$ *admits at least two nontrivial solutions.*

Moreover, there exists a positive solution for $(P_{\lambda,\delta})$.

Consider the product space $Y := X_0 \times X_0$ endowed with the norm $\|(u,v)\|^2 := \|u\|^2 + \|v\|^2$. For notational convenience, if $u, v \in X_0$ we set

$$B(u,v) := \int_\varOmega (|x|^{-\mu} * |u|^{2^*_\mu})|v|^{2^*_\mu}.$$

Definition 10.3.5 We say that $(u,v) \in Y$ is a weak solution to $(P_{\lambda,\delta})$ if for every $(\phi,\psi) \in Y$, it satisfies

$$(\langle u,\phi\rangle + \langle v,\psi\rangle) = \int_\varOmega (\lambda|u|^{q-2}u\phi + \delta|v|^{q-2}v\psi)\mathrm{d}x \\ + \int_\varOmega (|x|^{-\mu} * |v|^{2^*_\mu})|u|^{2^*_\mu-2}u\phi\,\mathrm{d}x + \int_\varOmega (|x|^{-\mu} * |u|^{2^*_\mu})|v|^{2^*_\mu-2}v\psi\,\mathrm{d}x.$$

Equivalently, if we define the functional $I_{\lambda,\delta} : Y \to \mathbb{R}$ as

$$I_{\lambda,\delta}(u) := \frac{1}{2}\|(u,v)\|^2 - \frac{1}{q}\int_\Omega (\lambda|u|^q + \delta|v|^q) - \frac{2}{22^*_\mu}B(u,v)$$

then the critical points of $I_{\lambda,\delta}$ correspond to the weak solutions of $(P_{\lambda,\delta})$. We set

$$\tilde{S}^H_s = \inf_{(u,v)\in Y\setminus\{(0,0)\}} \frac{\|(u,v)\|^2}{\left(\int_\Omega (|x|^{-\mu} * |u|^{2^*_\mu})|v|^{2^*_\mu}\,\mathrm{d}x\right)^{\frac{1}{2^*_\mu}}} = \inf_{(u,v)\in Y\setminus\{(0,0)\}} \frac{\|(u,v)\|^2}{B(u,v)^{\frac{1}{2^*_\mu}}}$$

and show that $\tilde{S}^H_s = 2S^H_s$. We define the set

$$\mathcal{N}_{\lambda,\delta} := \{(u,v) \in Y \setminus \{0\} : \; (I'_{\lambda,\delta}(u,v),(u,v)) = 0\}$$

and find that the functional $I_{\lambda,\delta}$ is coercive and bounded below on $\mathcal{N}_{\lambda,\delta}$. Consider the fibering map $\varphi_{u,v} : \mathbb{R}^+ \to \mathbb{R}$ as $\varphi_{u,v}(t) = I_{\lambda,\delta}(tu,tv)$ which gives another characterization of $\mathcal{N}_{\lambda,\delta}$ as follows

$$\mathcal{N}_{\lambda,\delta} = \{(tu,tv) \in Y \setminus \{(0,0)\} : \; \varphi'_{u,v}(t) = 0\}$$

because $\varphi'_{u,v}(t) = (I'_{\lambda,\delta}(tu,tv),(u,v))$. Naturally, our next step is to divide $\mathcal{N}_{\lambda,\delta}$ into three subsets corresponding to local minima, local maxima and point of inflexion of $\varphi_{u,v}$ namely

$$\mathcal{N}^{\pm}_{\lambda,\delta} := \{(u,v) \in \mathcal{N}_{\lambda,\delta} : \; \varphi''_{u,v}(1) \gtrless 0\} \text{ and } \mathcal{N}^0_{\lambda,\delta} := \{(u,v) \in \mathcal{N}_{\lambda.\delta} : \; \varphi''_{u,v}(1) = 0\}.$$

As earlier, the minimizers of $I_{\lambda,\delta}$ on $\mathcal{N}^+_{\lambda,\delta}$ and $\mathcal{N}^-_{\lambda,\delta}$ forms nontrivial weak solutions of $(P_{\lambda,\delta})$. Then we found a threshold on the range of λ and δ so that $\mathcal{N}_{\lambda,\delta}$ forms a manifold. Precisely we proved.

Lemma 10.3.6 *For every $(u,v) \in Y \setminus \{(0,0)\}$ and λ, δ satisfying $0<\lambda^{\frac{2}{2-q}}+\delta^{\frac{2}{2-q}} < \Theta$, where Θ is equal to*

$$\left[\frac{2^{2^*_\mu-1}(C^n_s)^{\frac{22^*_\mu-q}{2-q}}}{C(n,\mu)}\left(\frac{2-q}{22^*_\mu-q}\right)\left(\frac{22^*_\mu-2}{22^*_\mu-q}\right)^{\frac{22^*_\mu-2}{2-q}} S_s^{\frac{q(2^*_\mu-1)}{2-q}+2^*_\mu}|\Omega|^{-\frac{(2^*_s-q)(22^*_\mu-2)}{2^*_s(2-q)}}\right]^{\frac{1}{2^*_\mu-1}} \tag{10.27}$$

then there exist unique $t_1, t_2 > 0$ such that $t_1 < t_{max}(u,v) < t_2$, $(t_1u, t_1v) \in \mathcal{N}^+_{\lambda,\delta}$ and $(t_2u, t_2v) \in \mathcal{N}^-_{\lambda,\delta}$. Moreover, $\mathcal{N}^0_{\lambda,\delta} = \emptyset$. As a consequence, we infer that for any λ, δ satisfying $0 < \lambda^{\frac{2}{2-q}} + \delta^{\frac{2}{2-q}} < \Theta$,

$$\mathcal{N}_{\lambda,\delta} = \mathcal{N}^+_{\lambda,\delta} \cup \mathcal{N}^-_{\lambda,\delta}.$$

After this, we prove that any Palais Smale sequence $\{(u_k, v_k)\}$ for $I_{\lambda,\delta}$ must be bounded in Y and its weak limit forms a weak solution of $(P_{\lambda,\delta})$. We define the following

$$l_{\lambda,\delta} = \inf_{\mathcal{N}_{\lambda,\delta}} I_{\lambda,\delta} \text{ and } l^{\pm}_{\lambda,\delta} = \inf_{\mathcal{N}^{\pm}_{\lambda,\delta}} I_{\lambda,\delta}.$$

We fix $0 < \lambda^{\frac{2}{2-q}} + \delta^{\frac{2}{2-q}} < \Theta$ and showed that $l_{\lambda,\delta} \leq l^{+}_{\lambda,\delta} < 0$ and $\inf\{\|(u,v)\| : (u,v) \in \mathcal{N}^{-}_{\lambda,\delta}\} > 0$.

To prove the existence of first solution, we first show that there exists a $(PS)_{l_{\lambda,\delta}}$ sequence $\{(u_k, v_k)\} \subset \mathcal{N}_{\lambda,\delta}$ for $I_{\lambda,\delta}$ using the Ekeland variational principle and then prove that $l^{+}_{\lambda,\delta}$ is achieved by some function $(u_1, v_1) \in \mathcal{N}^{+}_{\lambda,\delta}$. Moreover $u_1, v_1 > 0$ in Ω and for each compact subset K of Ω, there exists a $m_K > 0$ such that $u_1, v_1 \geq m_K$ in K. Thus, we obtain a positive weak solution (u_1, v_1) of $(P_{\lambda,\delta})$.

On the other hand, proof of existence of second solution has been divided into two parts- $\mu \leq 4s$ and $\mu > 4s$. In the case $\mu \leq 4s$, using the estimates in Proposition 10.2.6, we could reach the first critical level as follows:

$$\sup_{t\geq 0} I_{\lambda,\delta}((u_1, v_1) + t(w_0, z_0)) < c_1 := I_{\lambda,\delta}(u_1, v_1) + \frac{n-\mu+2s}{2n-\mu}\left(\frac{C_s^n \tilde{S}_s^H}{2}\right)^{\frac{2n-\mu}{n-\mu+2s}}$$

for some nonnegative $(w_0, z_0) \in Y \setminus \{(0,0)\}$. This implied $l^{-}_{\lambda,\delta} < c_1$. Whereas to show the same thing in the case $\mu > 4s$, we had to take another constant $\Theta_0 \leq \Theta$ and the same estimates as in Proposition 10.2.6. Consequently, we prove that there exists a $(u_2, v_2) \in \mathcal{N}^{-}_{\lambda,\delta}$ such that $l^{-}_{\lambda,\delta}$ is achieved, hence gave us the second solution. From this, we concluded the proof of Theorem 10.3.4.

10.4 Some Open Questions

Here we state some open problems in this direction.

1. H^1 versus C^1 local minimizers and global multiplicity result: Consider energy functional defined on $H^1_0(\Omega)$ given by $\Phi(u) = \frac{\|u\|^2}{2} - \lambda \int_\Omega F(x,u)$ where F is the primitive of f. When $|f(u)| \leq C(1+|u|^p)$ for $p \in [1, 2^*]$, Brezis and Nirenberg in [16] showed that a local minimum of Φ in $C^1(\Omega)$-topology is also a local minimum in the $H^1_0(\Omega)$-topology. Such property of the functional corresponding to Choquard type nonlinearity and singular terms is still not addressed.
2. Variable exponent problems: As pointed out in Sect. 10.1.3, existence of a solution for problem (10.16) has been studied in [2] but the question of multiplicity of solutions for variable exponent Choquard equations is still open.
3. p-Laplacian critical problems: The Critical exponent problem involving the p-Laplacian and Choquard terms is an important question. This requires the study of minimizers of $S_{H,L}$. Also, the regularity of solutions and the global multiplicity results for convex-concave nonlinearities is worth exploring.

4. Hardy-Sobolev operators and nonlocal problems: The doubly critical problems arise due to the presence of two noncompact terms. Hardy-Sobolev operator is defined as $-\Delta_p u - \frac{\mu|u|^{p-2}u}{|x|^2}$. Here the critical growth Choquard terms in the equations require the minimizers and asymptotic estimates to study the compactness of minimizing sequences. The existence and multiplicity results are good questions to explore in this case.

References

1. C.O. Alves, F. Gao, M. Squassina, M. Yang, Singularly perturbed critical Choquard equations. J. Differ. Equ. **263**(7), 3943–3988 (2017)
2. C.O. Alves, L.S. Tavares, A Hardy-Littlewood-Sobolev type inequality for variable exponents and applications to quasilinear Choquard equations involving variable exponent. Mediterr. J. Math. **16**, 55 (2019), https://arxiv.org/pdf/1609.09558.pdf
3. C.O. Alves, M. Yang, Multiplicity and concentration of solutions for a quasilinear Choquard equation, J. Math. Phys. **55**, 061502, 21 (2014)
4. C.O. Alves, M. Yang, Existence of semiclassical ground state solutions for a generalized Choquard equation. J. Differ. Equ. **257**, 4133–4164 (2014)
5. C.O. Alves, M. Yang, Investigating the multiplicity and concentration behaviour of solutions for a quasi-linear Choquard equation via the penalization method. Proc. R. Soc. Edinburgh Sect. A. **146**, 23–58 (2016)
6. C.O. Alves, M. Yang, Existence of solutions for a nonlocal variational problem in $\mathbb{R}^2$ with exponential critical growth. J. Convex Anal. **24**, 1197–1215 (2017)
7. C.O. Alves, D. Cassani, C. Tarsi, M. Yang, Existence and concentration of ground state solutions for a critical nonlocal Schrödinger equation in $\mathbb{R}^n$. J. Differ. Equ. **261**, 1933–1972 (2016)
8. C.O. Alves, M.G. Figueiredo, M. Yang, Existence of solutions for a nonlinear Choquard equation with potential vanishing at infinity. Adv. Nonlinear Anal. **5**, 331–345 (2016)
9. R. Arora, J. Giacomoni, T. Mukherjee, K. Sreenadh, n-Kirchhoff Choquard equation with exponential nonlinearity. Nonlinear Anal. **186**, 113–144 (2019), https://arxiv.org/pdf/1810.00583.pdf
10. A. Bahri, J.-M. Coron, On a nonlinear elliptic equation involving the critical Sobolev exponent: the effect of the topology of the domain. Commun. Pure Appl. Math. **41**, 253–294 (1988)
11. V. Benci, C.R. Grisanti, A.M. Micheletti, Existence and non existence of the ground state solution for the nonlinear Schrödinger equations with $V(\infty) = 0$. Topol. Methods Nonlinear Anal. **26**, 203–219 (2005)
12. V. Benci, C.R. Grisanti, A.M. Micheletti, Existence of solutions for the nonlinear Schrödinger equation with $V(\infty) = 0$. Progr. Nonlinear Differ. Equ. Appl. **66**, 53–65 (2005)
13. H. Berestycki, P.L. Lions, Nonlinear scalar field equations. I Existence of a ground state. Arch. Ration. Mech. Anal. **82**, 313–346 (1983)
14. H. Brézis, T. Kato, Remarks on the Schrödinger operator with singular complex potentials. J. Math. Pures Appl. **9**, 137–151 (1979)
15. H. Brézis, L. Nirenberg, Positive solutions of nonlinear elliptic equations involving critical Sobolev exponents. Commun. Pure Appl. Math. **36**, 437–477 (1983)
16. H. Brézis, L. Nirenberg, H^1 versus C^1 local minimizers. C. R. Acad. Sci. Paris. **317**, 465–472 (1993)
17. D. Cassani, J.V. Scahftingen, J. Zhang, Groundstates for Choquard type equations with Hardy-Littlewood-Sobolev lower critical exponent, *Proceedings of the Royal Society of Edinburgh: Section A Mathematics*, pp. 1–24, https://doi.org/10.1017/prm.2018.135, https://arxiv.org/pdf/1709.09448.pdf

18. W. Chen, S. Deng, The Nehari manifold for a fractional p-Laplacian system involving concave-convex nonlinearities. Nonlinear Anal. Real World Appl. **27**, 80–92 (2016)
19. W. Chen, M. Squassina, Critical Nonlocal Systems with Concave-Convex Powers. Adv. Nonlinear Stud. **16**, 821–842 (2016)
20. W. Choi, On strongly indefinite systems involving the fractional Laplacian. Nonlinear Anal. **120**, 127–153 (2015)
21. S. Cingolani, S. Secchi, M. Squassina, Semi-classical limit for Schrödinger equations with magnetic field and Hartree-type nonlinearities. Proc. R. Soc. Edinb. Sect. A **140**, 973–1009 (2010)
22. S. Cingolani, M. Clapp, S. Secchi, Multiple solutions to a magnetic nonlinear Choquard equation. Z. Angrew. Math. Phys. **63**, 233–248 (2012)
23. S. Cingolani, M. Clapp, S. Secchi, Intertwining semiclassical solutions to a Schrödinger- Newton system. Discret. Contin. Dyn. Syst. Ser. S. **6**, 891–908 (2013)
24. M.G. Crandall, P.H. Rabinowitz, L. Tartar, On a Dirichlet problem with a singular nonlinearity. Commun. Part. Differ. Equ. **2**, 193–222 (1977)
25. Y. Ding, F. Gao, M. Yang, Semiclassical states for Choquard type equations with critical growth: critical frequency case. https://arxiv.org/pdf/1710.05255.pdf
26. L. Du, F. Gao, M. Yang, Existence and qualitative analysis for nonlinear weighted Choquard equations, https://arxiv.org/pdf/1810.11759.pdf
27. L. Du, M. Yang, Uniqueness and nondegeneracy of solutions for a critical nonlocal equation. Discrete Contin. Dyn. Syst. **39**(10), 5847–5866 (2019), https://arxiv.org/pdf/1810.11186.pdf
28. L.F.O. Faria, O.H. Miyagaki, F.R. Pereira, M. Squassina, C. Zhang, The Brezis-Nirenberg problem for nonlocal systems. Adv. Nonlinear Anal. **5**, 85–103 (2016)
29. F. Gao, E.D. da Silva, M. Yang, J. Zhou, Existence of solutions for critical Choquard equations via the concentration compactness method, *Proceedings of the Royal Society of Edinburgh: Section A Mathematics*, (2018), pp. 1–34, https://arxiv.org/abs/1712.08264, to appear in Proc. R. Soc. Edinb., A Math
30. F. Gao, M. Yang, A strongly indefinite Choquard equation with critical exponent due to the Hardy-Littlewood-Sobolev inequality. Commun. Contemp. Math. **20**(4), 1750037, (22 pages) (2018)
31. F. Gao, M. Yang, On nonlocal Choquard equations with Hardy Littlewood Sobolev critical exponents. J. Math. Anal. Appl. **448**, 1006–1041 (2017)
32. F. Gao, M. Yang, On the Brezis Nirenberg type critical problem for nonlinear Choquard equation. Sci. Chi. Math. **61**, 1219–1242 (2018), https://doi.org/10.1007/s11425-016-9067-5
33. V. Georgiev, G. Venkov, Symmetry and uniqueness of minimizers of Hartree type equations with external Coulomb potential. J. Differ. Equ. **251**, 420–438 (2011)
34. M. Ghimenti, D. Pagliardini, Multiple positive solutions for a slightly subcritical Choquard problem on bounded domains. Calc. Var. Partial Dif. **58**, 167 (2019), https://arxiv.org/pdf/1804.03448.pdf
35. J. Giacomoni, T. Mukherjee, K. Sreenadh, Doubly nonlocal system with Hardy-Littlewood-Sobolev critical nonlinearity. J. Math. Anal. Appl. **467**, 638–672 (2018)
36. Goel, D., Sreenadh, K.: Kirchhoff equations with Hardy-Littlewood-Sobolev critical nonlinearity. Nonlinear Anal. **186**, 162–186 (2019)
37. D. Goel, V. Radulescu, K. Sreenadh, Coron problem for nonlocal equations involving Choquard nonlinearity. Adv. Nonlinear Stud. (2019), https://doi.org/10.1515/ans-2019-2064, https://arxiv.org/pdf/1804.08084.pdf
38. Z. Guo, S. Luo, W. Zou, On critical systems involving frcational Laplacian. J. Math. Anal. Appl. **446**, 681–706 (2017)
39. Y. Haitao, Multiplicity and asymptotic behavior of positive solutions for a singular semilinear elliptic problem. J. Differ. Equ. **189**, 487–512 (2003)
40. N. Hirano, C. Saccon, N. Shioji, Existence of multiple positive solutions for singular elliptic problems with concave and convex nonlinearities. Adv. Differ. Equ. **9**, 197–220 (2004)
41. N. Hirano, C. Saccon, N. Shioji, Brezis-Nirenberg type theorems and multiplicity of positive solutions for a singular elliptic problem. J. Differ. Equ. **245**, 1997–2037 (2008)

42. Z. Huang, J. Yang, W. Yu, Multiple nodal solutions of nonlinear choquard equations. Electron. J. Differ. Equ. **268**, 1–18 (2017)
43. Y. Lei, On the regularity of positive solutions of a class of Choquard type equations. Math. Z. **273**, 883–905 (2013)
44. Y. Lei, Qualitative analysis for the static Hartree-type equations. SIAM J. Math. Anal. **45**, 388–406 (2013)
45. G.-D. Li, C.-L. Tnag, Existence of ground state solutions for Choquard equation involving the general upper critical Hardy-Littlewood-Sobolev nonlinear term. Commun. Pure Appl. Anal. **18**, 285–300 (2019)
46. E.H. Lieb, Existence and uniqueness of the minimizing solution of Choquards nonlinear equation. Stud. Appl. Math. **57** 93-105 (1976/77)
47. E.H. Lieb, M. Loss, *Analysis*, 2nd edn. (AMS, 2001)
48. O. Lopes, M. Maris, Symmetry of minimizers for some nonlocal variational problems. J. Funct. Anal. **254**, 535–592 (2008)
49. D. Lü, Existence and concentration behavior of ground state solutions for magnetic nonlinear Choquard equations. Commun. Pure Appl. Anal. **15**, 1781–1795 (2016)
50. L. Ma, L. Zhao, Classification of positive solitary solutions of the nonlinear Choquard equation. Arch. Rational Mech. Anal. **195**, 455–467 (2010)
51. G.P. Menzala, On the nonexistence of solutions for an elliptic problem in unbounded domains. Funkcial. Ekvac. **26**, 231–235 (1983)
52. C. Mercuri, V. Moroz, J.V. Schaftingen, Groundstates and radial solutions to nonlinear SchrödingerPoissonSlater equations at the critical frequency. J. Calc. Var. **55**, 146 (2016)
53. V. Moroz, J.V. Schaftingen, Groundstates of nonlinear Choquard equations: Hardy Littlewood Sobolev critical exponent. Commun. Contemp. Math. **17**, 1550005 (12 pages) (2015)
54. V. Moroz, J.V. Schaftingen, Groundstates of nonlinear Choquard equations: existence, qualitative properties and decay asymptotics. J. Funct. Anal. **265**, 153–184 (2013)
55. V. Moroz, J.V. Schaftingen, Existence of groundstates for a class of nonlinear Choquard equations. Trans. Am. Math. Soc. **367**, 6557–6579 (2015)
56. V. Moroz, J.V. Schaftingen, Least action nodal solutions for the quadratic Choquard equation. Proc. Am. Math. Soc. **145**, 737–747 (2017)
57. V. Moroz, J.V. Schaftingen, A guide to the Choquard equation. J. Fixed Point Theory Appl. **19**, 773–813 (2017)
58. T. Mukherjee, K. Sreenadh, Positive solutions for nonlinear Choquard equation with singular nonlinearity. Complex Var. Elliptic Equ. **62**, 1044–1071 (2017)
59. T. Mukherjee, K. Sreenadh, Fractional Choquard Equation with critical nonlinearities. Nonlinear Differ. Equ. Appl. **24**, 63 (2017)
60. T. Mukherjee, K. Sreenadh, On Concentration of least energy solutions for magnetic critical Choquard equations. J. Math. Anal. Appl. **464**, 402–420 (2018)
61. T. Mukherjee, K. Sreenadh, On doubly nonlocal p-fractional coupled elliptic system. Topol. Methods Nonlinear Anal. **51**, 609–636 (2018)
62. E.D. Nezza, G. Palatucci, E. Valdinoci, Hitchhikers guide to the fractional sobolev spaces. Bull. Sci. Math. **136**, 521–573 (2012)
63. S. Pekar, *Untersuchung über die Elektronentheorie der Kristalle* (Akademie Verlag, Berlin, 1954)
64. D. Salazar, Vortex-type solutions to a magnetic nonlinear Choquard equation. Z. Angew. Math. Phys. **66**, 663–675 (2015)
65. R. Servadei, E. Valdinoci, The Brezis-Nirenberg result for the fractional laplacian. Trans. Am. Math. Soc. **367**, 67–102 (2015)
66. Z. Shen, F. Gao, M. Yang, Multiple solutions for nonhomogeneous Choquard equation involving Hardy Littlewood Sobolev critical exponent. Z. Angew. Math. Phys. **68**, 61 (2017)
67. Z. Shen, F. Gao, M. Yang, On critical Choquard equation with potential well. Discret. Contin. Dyn. Syst. A **38**(7), 3669–3695 (2018)
68. G. Tarantell, On nonhomogeneous elliptic equations involving critical Sobolev exponent. Ann. Inst. H. Poincaré Anal. Non Linéaire. **9**, 281–304 (1992)

69. K. Wang, J. Wei, On the uniqueness of solutions of a nonlocal elliptic system. Math. Ann. **365**, 105–153 (2016)
70. T. Wang, T. Yi, Uniqueness of positive solutions of the Choquard type equations. Appl. Anal. **96**, 409–417 (2017)
71. J. Wang, Y. Dong, Q. He, L. Xiao, Multiple positive solutions for a coupled nonlinear Hartree type equations with perturbations. J. Math. Anal. Appl. **450**, 780–794 (2017)
72. T. Xie, L. Xiao, J. Wang, Exixtence of multiple positive solutions for Choquard equation with perturbation. Adv. Math. Phys. **2015**, 760157 (2015)
73. M. Yang, Semiclassical ground state solutions for a Choquard type equation in R^2 with critical exponential growth. ESAIM: Control., Optim. Calc. Var. **24**, 177–209 (2018)
74. H. Zhang, J. Xu, F. Zhang, Existence and multiplicity of solutions for a generalized Choquard equation. Comput. Math. Appl. **73**, 1803–1814 (2017)

Chapter 11
Wavelet Galerkin Methods for Higher Order Partial Differential Equations

B. V. Rathish Kumar and Gopal Priyadarshi

Abstract In this paper, we develop efficient and accurate wavelet Galerkin methods for higher order partial differential equations. Compactly supported Daubechies wavelets are used for spatial discretization, whereas stable finite difference methods are used for temporal discretization. The exact values of two-term connection coefficients are effectively used for the evaluation of integrals consisting of higher order derivatives. For the nonlinear elliptic partial differential equations, we have employed quasilinearization technique to obtain the nonlinear wavelet coefficients. Sparse GMRES solver is used to solve linear system of equations obtained after spatial and temporal discretization. Error analysis has been carried out to ensure the convergence of the proposed method. Finally, the method is successfully tested on few linear and nonlinear 1D and 2D PDEs.

Keywords Wavelet Galerkin method · Higher order PDEs · Daubechies wavelets · Nonlinear PDEs · GMRES method

11.1 Introduction

This paper is concerned with fully discrete wavelet Galerkin methods for higher order partial differential equations of the form

$$F(x, u(x), \nabla u(x), \Delta u(x), \nabla \Delta u(x), \Delta^2 u(x)) = f(x), \tag{11.1}$$

$$F(x, t, u(x,t), u_t(x,t), \nabla u(x,t), \Delta u(x,t), \nabla \Delta u(x,t), \Delta^2 u(x,t)) = f(x,t), \tag{11.2}$$

B. V. Rathish Kumar (✉) · G. Priyadarshi
Indian Institute of Technology Kanpur, Kanpur, Uttar Pradesh, India
e-mail: bvrk@iitk.ac.in

G. Priyadarshi
e-mail: gopalpriyadarshi8@gmail.com

P. Manchanda et al. (eds.), *Mathematical Modelling, Optimization, Analytic and Numerical Solutions*, Industrial and Applied Mathematics,
https://doi.org/10.1007/978-981-15-0928-5_11

and

$$F(x, t, u(x, t), u_{tt}(x, t), \nabla u(x, t), \Delta u(x, t), \nabla \Delta u(x, t), \Delta^2 u(x, t)) = f(x, t), \tag{11.3}$$

where $x \in \mathbb{R}^n,\ t \in [0, T],\ u(x, t)$ is the unknown solution and F may be a linear or nonlinear function.

In particular, we consider linear and nonlinear biharmonic equation, fourth-order diffusion equation and fourth-order wave equation with one-periodic boundary conditions given by

$$\Delta^2 u(x) = f(x) \tag{11.4}$$

$$\Delta^2 u(x) + u^2(x) = f(x) \tag{11.5}$$

$$\left.\begin{array}{r} u_t(x, t) + \eta \Delta^2 u(x, t) = f(x, t) \\ u(x, 0) = g(x) \end{array}\right\} \tag{11.6}$$

$$\left.\begin{array}{r} u_t(x, t) + \eta \Delta^2 u(x, t) + u^2(x, t) = f(x, t) \\ u(x, 0) = g(x) \end{array}\right\} \tag{11.7}$$

$$\left.\begin{array}{r} u_{tt}(x, t) + \eta \Delta^2 u(x, t) = f(x, t) \\ u(x, 0) = g(x) \\ u_t(x, 0) = h(x) \end{array}\right\} \tag{11.8}$$

$$\left.\begin{array}{r} u_{tt}(x, t) + \eta \Delta^2 u(x, t) + u^2(x, t) = f(x, t) \\ u(x, 0) = g(x) \\ u_t(x, 0) = h(x) \end{array}\right\}. \tag{11.9}$$

These PDEs have many applications in various areas of science and engineering. For example, biharmonic equation appears in continuum mechanics, elasticity, dynamical system, and fluid dynamics, whereas fourth-order diffusion equation arise in material science, computer graphics, and image processing. In the study of vibration of beams and thin plates, fourth-order wave equation plays an important role.

In the last few decades, wavelets have emerged as a powerful tool to solve partial differential equations numerically. For the first time, Beylkin et al. [1, 2] realized that certain operators may have sparse multiscale representation in terms of wavelets. Thereafter, Glowinski et al. [3] used wavelets in the Galerkin framework to solve linear and nonlinear elliptic, parabolic, and hyperbolic problems. Qian et al. [4] used wavelet Galerkin solver with an adaptation of capacitance matrix method to solve Helmholtz equation in nonseparable domain. Amaratunga et al. [5] proposed wavelet Galerkin method based on Daubechies wavelets for solving one-dimensional partial differential equations with periodic boundary conditions. Adaptive wavelet methods have been investigated by several researchers, e.g., Masson et al. [6, 7], Urban et al. [8], Stevenson et al. [9, 10], and DeVore et al. [11].

There is not much literature on the wavelet methods for higher order partial differential equations. In 2012, Shi et al. [12] proposed collocation method based on Haar wavelets to solve multidimensional biharmonic and Poisson equations. Bertoluzza et al. [13] constructed a mixed Lagrange–Hermite interpolating wavelet family for solving fourth-order elliptic equation, in particular, one-dimensional Euler–Bernoulli beam equation. Qian et al. [14] proposed wavelet capacitance matrix method to solve biharmonic equation with nonseparable boundary conditions. They used Daubechies wavelets to achieve a spectral convergence rate. Dahlke et al. [15] developed a numerical scheme based on Deslauriers–Dubuc fundamental functions and Newton's method to solve nonlinear elliptic partial differential equations. Recently, Priyadarshi et al. [16, 17] investigated wavelet-based numerical methods to solve higher order elliptic PDEs.

Wavelet-based methods have been extensively studied for the second-order parabolic and hyperbolic partial differential equations. Rathish et al. [18–22] developed various numerical methods, e.g., wavelet Taylor–Galerkin method, three-step wavelet Galerkin method, and time accurate pseudo-wavelet method to solve second-order parabolic and hyperbolic PDEs. They derived a priori error estimates [23, 24] using spectral decomposition theorem and wavelet approximation properties. However, these schemes are limited to second-order PDEs and largely to one- and two-dimensional problems only. Alam et al. [25] proposed space–time adaptive wavelet method to solve second-order nonlinear parabolic PDEs. They have found that the proposed method used roughly 18 times less grid points and roughly 4 times faster than a dynamically adaptive time marching scheme. Henn et al. [26] proposed a numerical method based on finite difference and multigrid approach to solve fourth-order diffusion equation with an application to image processing. Recently, Rathish et al. [27] proposed a wavelet-based numerical method for higher order parabolic PDEs. Fourth-order wave equations with dissipative and nonlinear strain terms have been investigated by Yacheng et al. [28]. The investigation is based on potential well methods. Decay estimate for fourth-order wave equation has been studied by Levandosky et al. [29]. They have obtained both $L^p - L^q$ estimates and space–time integrability estimates on the solutions of linear wave equation.

Based on compactly supported Daubechies wavelets, we develop wavelet Galerkin methods to solve linear and nonlinear PDEs. Various attractive properties, such as compact support, orthogonality, high-order vanishing moments, and arbitrary regularity make Daubechies wavelets a natural choice for the numerical solution of PDEs. To compute the integrals consisting of higher order derivatives, we exploit the exact values of two-term connection coefficients [30] which make the computation easier and accurate. For the nonlinear elliptic case, we exploit the property of wavelet coefficients obtained from the linear problem to get wavelet coefficients for the nonlinear problem. We derive error estimate using wavelet approximation results and Sobolev space theory. Finally, numerical results are provided to demonstrate the accuracy of the proposed methods.

The content of this paper is organized as follows. In Sect. 11.2, we provide a brief background of wavelets, in particular, Daubechies wavelets and some standard results which have been used throughout the paper. In Sect. 11.3, we develop wavelet

Galerkin methods for higher order linear and nonlinear partial differential equations. Error estimates are also derived in this section. In Sect. 11.4, we present some numerical results which ensure the accuracy of the proposed method. A brief conclusion is presented in Sect. 11.5.

11.2 Basic Background

In this section, we provide a basic background of wavelet, in particular, Daubechies wavelet which has been used extensively throughout this paper. For more details on Daubechies wavelet, one may refer to [31, 32].

The continuous wavelet family is defined as

$$\psi_{m,n}(x) = |m|^{-1/2}\psi\left(\frac{x-n}{m}\right), \ m, n \in \mathbb{R}, m \neq 0,$$

where the translation parameter, n, and the dilation parameter, m, vary continuously. If we take the translation and dilation parameters knm^{-J} and m^{-J}, respectively, where $m > 1, n > 0$, J and k are positive integers, then we obtain a family of discrete wavelets

$$\psi_{J,k}(x) = |m|^{J/2}\psi(m^J x - nk), \quad x \in \mathbb{R}.$$

In the special case, $m = 2$ and $n = 1$, the family of functions $\psi_{J,k}(x)$ forms an orthonormal basis for $L^2(\mathbb{R})$.

Multiresolution Analysis: A multiresolution analysis is a sequence of nested subspaces V_J $(J \in \mathbb{Z})$ of $L^2(\mathbb{R})$ which satisfies the following conditions:

(i) $V_J \subset V_{J+1}, \qquad \forall J \in \mathbb{Z},$
(ii) $f \in V_J \Leftrightarrow f(2(\cdot)) \in V_{J+1}, \qquad \forall J \in \mathbb{Z},$
(iii) $\bigcap_{J\in\mathbb{Z}} V_J = \{0\},$
(iv) $\overline{\bigcup_{J\in\mathbb{Z}} V_J} = L^2(\mathbb{R}),$
(v) There exists a function $\phi \in V_0$, known as scaling function, such that $\{\phi(\cdot - k): k \in \mathbb{Z}\}$ forms an orthonormal basis for V_0.

Define

$$\phi_{J,k}(x) = 2^{J/2}\phi(2^J x - k), \qquad J, k \in \mathbb{Z}.$$

Using the multiresolution analysis, it can be proved that the set $\{\phi_{J,k}(x) \in L^2(\mathbb{R}) \mid k \in \mathbb{Z}\}$ forms an orthonormal basis for V_J.

Since, $\phi(x) \in V_0 \subset V_1$, we have a two-scale relation

$$\phi(x) = \sqrt{2}\sum_{k=-\infty}^{\infty} a_k\phi(2x - k), \tag{11.10}$$

where

$$a_k = \sqrt{2}\int_{-\infty}^{\infty} \phi(x)\phi(2x-k)dx.$$

Similarly, $\psi(x) \in W_0 \subset V_1$, we get the following:

$$\psi(x) = \sqrt{2}\sum_{k=-\infty}^{\infty} b_k\phi(2x-k), \tag{11.11}$$

where

$$b_k = \sqrt{2}\int_{-\infty}^{\infty} \psi(x)\phi(2x-k)dx.$$

For Daubechies scaling function, only finitely many filter coefficients (a_k) are nonzero (see [31]). Hence, Eq. (11.10) becomes

$$\phi(x) = \sqrt{2}\sum_{k=0}^{D-1} a_k\phi(2x-k),$$

where $D(=2r)$ is called the wavelet genus and $a_0, a_1, \ldots, a_{D-1}$ are called the filter coefficients.

Similarly, for Daubechies wavelet, only finitely many filter coefficients (b_k) are nonzero (see [31]). Hence, Eq. (11.11) becomes

$$\psi(x) = \sqrt{2}\sum_{k=0}^{D-1} b_k\phi(2x-k).$$

The filter coefficients a_k and b_k are related in the following way:

$$b_k = (-1)^k a_{D-1-k}, \qquad k = 0, 1, \ldots, D-1.$$

Properties of Daubechies Wavelet

- For $r = 1$, we recover the Haar wavelet.
- The length of the support of Dbr is $(2r-1)$.
- The number of vanishing moments of Dbr is r.
- $Dbr \in C^{\mu r}$ where $\mu \approx 0.2$ for large r.

The orthogonal projection $P_J : L^2(\mathbb{R}) \to V_J$ is given by

$$P_J(f) = \sum_{k\in\mathbb{Z}} \langle f, \phi_{J,k}\rangle \phi_{J,k},$$

where $\langle\ .\ \rangle$ denotes the standard inner product in $L^2(\mathbb{R})$.

11.2.1 Periodized Wavelet

Based on the technique developed by Meyer [33], we define periodic scaling function and periodic wavelet. Let $\phi \in L^2(\mathbb{R})$ and $\psi \in L^2(\mathbb{R})$ be scaling function and wavelet from a multiresolution analysis. For any $J, k \in \mathbb{Z}$, the one-periodic scaling function is defined as

$$\tilde{\phi}_{J,k}(x) = \sum_{n=-\infty}^{\infty} \phi_{J,k}(x+n) = 2^{J/2} \sum_{n=-\infty}^{\infty} \phi(2^J(x+n)-k), \quad x \in [0,1],$$

and the one-periodic wavelet

$$\tilde{\psi}_{J,k}(x) = \sum_{n=-\infty}^{\infty} \psi_{J,k}(x+n) = 2^{J/2} \sum_{n=-\infty}^{\infty} \psi(2^J(x+n)-k), \quad x \in [0,1].$$

Approximation space $\tilde{V}_J$ is defined as

$$\tilde{V}_J = \overline{\text{span}\{\tilde{\phi}_{J,k}(x) \mid k = 0, 1, \ldots, 2^J - 1\}}, \;\; x \in [0,1].$$

It can be easily observed that the family of $\tilde{V}_J$'s forms an MRA for $L^2[0,1]$.

For the higher dimensional problem, we define the approximation space as tensor product of $\tilde{V}_J$. For example, in 2D case, the approximation space $\tilde{X}_J$ is defined as

$$\tilde{X}_J = \overline{\text{span}\{\tilde{\phi}_{J,k}(x)\tilde{\phi}_{J,l}(y) \mid k, l = 0, 1, \ldots, 2^J - 1\}}, \;\; (x,y) \in [0,1] \times [0,1].$$

Two-Term Connection Coefficients

In order to solve higher order partial differential equations, we have to deal with the integrals consisting higher order derivatives. So, we define two-term connection coefficients as follows:

$$\Gamma^{d_1,d_2}_{J,k,l} = \int_{-\infty}^{\infty} \phi^{d_1}_{J,k}(x)\phi^{d_2}_{J,l}(x)dx, \qquad J, k, l \in \mathbb{Z},$$

where d_1, d_2 are order of differentiation and $\phi_{J,k}, \phi_{J,l}$ are Daubechies scaling functions.

Change of variables and repeated integration by parts yields

$$\Gamma^{d_1,d_2}_{J,k,l} = (-1)^{d_1} 2^{Jd} \Gamma^{0,d}_{0,0,l-k},$$

where $d = d_1 + d_2$. Therefore, it is sufficient to consider only one order of differentiation and one shift of parameter.

Hence, we define

$$\Gamma^d_n = \int_{-\infty}^{\infty} \phi(x)\phi^d_n(x)dx,$$

where $\phi_n^d(x) = \phi^d(x-n)$.

Since Daubechies scaling functions are highly oscillatory in nature, using standard numerical quadrature is impractical for computing two-term connection coefficients. An exact method to evaluate two-term connection coefficients has been developed by Latto et al. [30].

Wavelet Approximation Results

Let the Daubechies scaling function be k-regular, that is, for each $n \in \mathbb{N}$ there exists c_n such that for all multi-index α, $|\alpha| \leq k$, the following condition holds:

$$|D^\alpha \phi(x)| \leq c_n (1+|x|)^{-n}. \tag{11.12}$$

Lemma 11.1 (see [34]) *Let k is fixed then, for any $0 < m \leq k+1$, there exists a constant $c > 0$ such that for all $u \in H^m(\Omega)$ and $J \in \mathbb{N}$*

$$\|u - P_J u\|_{H^0(\Omega)} \leq c 2^{-Jm} \|u\|_{H^m(\Omega)}, \tag{11.13}$$

where $P_J u$ is the orthogonal projection of u.

It is clear from Lemma 11.1 that the approximation error tends to zero as we go to higher and higher resolution level provided the function is sufficiently smooth.

11.3 Wavelet Galerkin Methods for Higher Order PDEs

In this section, we will describe wavelet Galerkin methods for higher order partial differential equations. For the spatial discretization, we have used Daubechies wavelets, whereas stable finite difference schemes are used for temporal discretization.

Let us consider linear elliptic partial differential equation

$$\frac{d^4 u}{dx^4}(x) = f(x), \qquad x \in \mathbb{R}, \tag{11.14}$$

with one-periodic boundary conditions. It is assumed that f is one periodic.

Since we are looking for one-periodic solution u, it is sufficient to consider u on the unit interval.

The Daubechies approximation for u is given by

$$u_J(x) = \sum_{k=0}^{2^J-1} a_{J,k} \tilde{\phi}_{J,k}(x), \tag{11.15}$$

where $a_{J,k}$ are wavelet coefficients.

Discretizing Eq. (11.14) by replacing $\dfrac{d^4u}{dx^4}$ with the following Daubechies approximation in the subspace $\tilde{V_J}$:

$$\frac{d^4u_J}{dx^4}(x) = \sum_{k=0}^{2^J-1}(a_{J,k})^*\tilde{\phi}_{J,k}(x), \tag{11.16}$$

where

$$(a_{J,k})^* = \sum_{n=2-D}^{D-2}(a_{J,\langle n+k\rangle_{2^J}})2^{4J}\Gamma_n^4, \tag{11.17}$$

and obtain the following equation:

$$\sum_{k=0}^{2^J-1}(a_{J,k})^*\tilde{\phi}_{J,k}(x) = f(x). \tag{11.18}$$

Multiplying Eq. (11.18) by $\tilde{\phi}_{J,p}(x)$ and integrating over unit interval and using orthonormality property, we obtain

$$(a_{J,p})^* = f_{J,p}, \quad p = 0, 1, \ldots, 2^J - 1, \tag{11.19}$$

where

$$f_{J,p} = \int_0^1 f(x)\tilde{\phi}_{J,p}(x)dx. \tag{11.20}$$

Using Eq. (11.17) in Eq. (11.19), we get

$$\sum_{n=2-D}^{D-2}(a_{J,\langle n+p\rangle_{2^J}})2^{4J}\Gamma_n^4 = f_{J,p}, \quad p = 0, 1, \ldots, 2^J - 1. \tag{11.21}$$

The set of equations obtained from (11.21) leads to a matrix equation

$$\mathbf{Ma = f}, \tag{11.22}$$

where $\mathbf{M}$ is a $2^J \times 2^J$ symmetric matrix consisting of two-term connection coefficients, $\mathbf{a}$ is a column vector consisting of wavelet coefficients $(a_{J,p})$, and $\mathbf{f}$ is a column vector consisting of $f_{J,p}$. $f_{J,p}$'s are calculated using appropriate quadrature formula.

Let us consider nonlinear elliptic partial differential equation

$$\frac{d^4u}{dx^4}(x) + u^2(x) = f(x), \qquad x \in \mathbb{R}, \tag{11.23}$$

with one-periodic boundary conditions. It is assumed that f is one periodic.

For the nonlinear problem, we apply quasilinearization technique to handle nonlinear terms arising after spatial discretization.

As done in the linear case, we obtain

$$(a_{J,p})^* + (a_{J,p})^2 = f_{J,p}, \quad p = 0, 1, \ldots, 2^J - 1. \tag{11.24}$$

Using Eq. (11.17) in Eq. (11.24), we obtain

$$\sum_{n=2-D}^{D-2} (a_{J,\langle n+p\rangle_{2^J}}) 2^{4J} \Gamma_n^4 + (a_{J,p})^2 = f_{J,p}, \quad p = 0, 1, \ldots, 2^J - 1. \tag{11.25}$$

The set of Eq. (11.25) leads to a nonlinear matrix equation which is difficult to solve. To overcome this difficulty, we first solve linear system of equations given by

$$\sum_{n=2-D}^{D-2} (a_{J,\langle n+p\rangle_{2^J}}) 2^{4J} \Gamma_n^4 + a_{J,p} = f_{J,p}, \quad p = 0, 1, \ldots, 2^J - 1. \tag{11.26}$$

Obtain the wavelet coefficients $a_{J,p}$ and name it $a_{J,p}^{old}$. Rewrite Eq. (11.24) as

$$(a_{J,p}^{new})^* + (a_{J,p}^{new})(a_{J,p}^{old}) = f_{J,p}, \;\; p = 0, 1, \ldots, 2^J - 1, \tag{11.27}$$

where $(a_{J,p}^{new})$ is wavelet coefficient of nonlinear problem.

In vector notation, we can write

$$(\mathbf{a}^{\mathbf{Jnew}})^* + (\mathbf{a}^{\mathbf{Jnew}})(\mathbf{a}^{\mathbf{Jold}}) = \mathbf{f}, \tag{11.28}$$

where

$$\mathbf{a}^{\mathbf{Jnew}} = [a_{J,0}^{new}, a_{J,1}^{new} \ldots, a_{J,2^J-1}^{new}]^T,$$

$$\mathbf{a}^{\mathbf{Jold}} = [a_{J,0}^{old}, a_{J,1}^{old} \ldots, a_{J,2^J-1}^{old}]^T,$$

$$\mathbf{f} = [f_{J,0}, f_{J,1} \ldots, f_{J,2^J-1}]^T.$$

The matrix equation is given by

$$(M_1 + M_2)\mathbf{a}^{\mathbf{Jnew}} = \mathbf{f}, \tag{11.29}$$

where M_1 is a $2^J \times 2^J$ matrix corresponding to $(\mathbf{a}^{\mathbf{Jnew}})^*$ and M_2 is a diagonal matrix whose diagonal entries are the elements of $\mathbf{a}^{\mathbf{Jold}}$. The matrix equation (11.29) is solved using GMRES iterative solver. Once we obtain $\mathbf{a}^{\mathbf{Jnew}}$ then again rename it $\mathbf{a}^{\mathbf{Jold}}$ and solve (11.28). We repeat this process until the maximum norm error of $\mathbf{a}^{\mathbf{Jnew}}$ and $\mathbf{a}^{\mathbf{Jold}}$ becomes less than some prescribed tolerance ε.

Theorem 11.1 *Let u and u_J be the exact and approximate solution, then*

$$\|u - u_J\|_{L^2} \le C2^{-4J}|u|_{H^4}, \qquad u \in H^4.$$

The above result can be obtained using standard Cea's lemma and wavelet projection estimate.

Let us consider linear parabolic partial differential equations

$$\left.\begin{aligned} \frac{\partial u}{\partial t}(x,t) + \eta\frac{\partial^4 u}{\partial x^4}(x,t) = f(x,t),\\ u(x,0) = g(x), \end{aligned}\right\} \qquad (x,t) \in \mathbb{R} \times [0,T] \tag{11.30}$$

with one-periodic boundary conditions. It is assumed that η is a positive constant, whereas f and g are one-periodic square-integrable functions.

Following the same process as done in the linear elliptic case, we get

$$\frac{d}{dt}a_{J,p}(t) + \eta(a_{J,p}(t))^* = f_{J,p}(t), \;\; p = 0, 1, \ldots, 2^J - 1,$$

where

$$f_{J,p}(t) = \int_0^1 f(x,t)\tilde{\phi}_{J,p}(x)dx.$$

In vector notation

$$\left.\begin{aligned} \frac{d}{dt}\mathbf{a}(t) + \eta\mathbf{a}^*(t) = \mathbf{f}(t)\\ \mathbf{a}(0) = \mathbf{a}_g \end{aligned}\right\},$$

where

$$\mathbf{a}(t) = [a_{J,0}(t), a_{J,1}(t), \ldots, a_{J,2^J-1}(t)]^T,$$

$$\mathbf{a}^*(t) = [(a_{J,0}(t))^*, (a_{J,1}(t))^*, \ldots, (a_{J,2^J-1}(t))^*]^T,$$

$$\mathbf{f}(t) = [f_{J,p}(t), p = 0, 1, \ldots, 2^J - 1]^T,$$

$$\mathbf{a}_g = [g_{J,p}, p = 0, 1, \ldots, 2^J - 1]^T,$$

$$f_{J,p}(t) = \int_0^1 f(x,t)\tilde{\phi}_{J,p}(x)dx,$$

$$g_{J,p} = \int_0^1 g(x)\tilde{\phi}_{J,p}(x)dx.$$

Applying backward Euler scheme for time discretization and get

$$\frac{\mathbf{a}_{n+1}-\mathbf{a}_n}{\triangle t}+\eta\mathbf{X}a_{n+1}=\mathbf{f_{n+1}},$$

where $\mathbf{a}_n=\mathbf{a}(n\triangle t)$. This leads to the recursive relation

$$\mathbf{a}_{n+1}=(I+\eta\triangle t\mathbf{X})^{-1}(\mathbf{a}_n+\mathbf{f_{n+1}}\triangle t), \tag{11.31}$$

where $\mathbf{X}$ is a $2^J\times 2^J$ symmetric matrix consisting connection coefficients and $\mathbf{a}_n$ is a column vector consisting unknown wavelet coefficients. $\mathbf{a}(0)$ is calculated using appropriate quadrature rule.

Using Eq. (11.31), we obtain wavelet coefficients and subsequently the solution at desired time.

Let us consider nonlinear parabolic partial differential equation

$$\left.\begin{array}{r}\dfrac{\partial u}{\partial t}(x,t)+\eta\dfrac{\partial^4 u}{\partial x^4}(x,t)+u^2(x,t)=f(x,t)\\ u(x,0)=g(x)\end{array}\right\}. \tag{11.32}$$

Applying finite difference scheme and following the same process as done in the case of linear parabolic PDEs, we get

$$\frac{\mathbf{a}_{n+1}-\mathbf{a}_n}{\triangle t}+\eta\mathbf{X}\mathbf{a}_{n+1}=\mathbf{f_{n+1}}-\mathbf{U_n},$$

where $\mathbf{U_n}=\left[\int_0^1 u_n^2\tilde{\phi}_{J,p}(x)dx\,:\,p=0,1,\ldots,2^J-1\right]$. This leads to the recursive relation

$$\mathbf{a}_{n+1}=(I+\eta\triangle t\mathbf{X})^{-1}(\mathbf{a}_n+(\mathbf{f_{n+1}}-\mathbf{U_n})\triangle t),$$

where $\mathbf{X}$ is a $2^J\times 2^J$ symmetric matrix consisting of connection coefficients and $\mathbf{a}_n$ is a column vector consisting of unknown wavelet coefficients. $\mathbf{U_0}$ is calculated using appropriate quadrature rule.

Error Analysis of a Fully Discrete Wavelet Galerkin Scheme

Theorem 11.2 *Let u^n and u_J^n be the exact and Daubechies solution at time t_n, then*

$$\max_{0\le n\le p}\|u^n-u_J^n\|=O(k+2^{-Js}),$$

where $u\in H^s$ and k is the time step.

We provide the key steps for the proof of the above theorem.

- Compare u_J not directly to u, but with elliptic projection y_J.
- $\|u-u_J\|\le\|u-y_J\|+\|y_J-u_J\|$.

- Choose u_J^0 such that $\|u^0 - u_J^0\| = O(2^{-Js})$.
- Obtain consistency error:

$$\left\langle \frac{y_J^{n+1} - y_J^n}{k}, v \right\rangle + a(y_J^{n+1}, v) - \langle f^{n+1}, v\rangle = \langle x^n, v\rangle, \text{ where}$$

$$x^n = \left(\frac{u^{n+1} - u^n}{k} - \frac{\partial u^{n+1}}{\partial t} + \frac{y_J^{n+1} - y_J^n}{k} - \frac{u^{n+1} - u^n}{k} \right).$$

- Estimate x^n using Taylor's theorem

$$\|x^n\| \leq c\left(k\|\frac{\partial^2 u}{\partial t^2}\|_{L^\infty([0,T],L^2)} + 2^{-Js}\|\frac{\partial u}{\partial t}\|_{L^\infty([0,T],H^s)} \right) = M.$$

- Choose $z_J = y_J - u_J$ and apply Cauchy–Schwarz inequality to get

$$\|z_J^{n+1}\| \leq k\|x^n\| + \|z_J^n\|.$$

- By repeated iteration

$$\max_{0\leq n\leq p} \|z_J^n\| \leq \|z_J^0\| + TM.$$

- Use elliptic projection estimate to get

$$\max_{0\leq n\leq p} \|u^n - u_J^n\| = O(k + 2^{-Js}).$$

Let us consider linear fourth-order wave equation

$$\left.\begin{aligned} \frac{\partial^2 u}{\partial t^2}(x,t) + \eta\frac{\partial^4 u}{\partial x^4}(x,t) &= f(x,t),\\ u(x,0) &= g(x),\\ \frac{\partial u}{\partial t}(x,0) &= h(x), \end{aligned}\right\} (x,t) \in \mathbb{R} \times [0,T] \tag{11.33}$$

with one-periodic boundary conditions. It is assumed that η is a positive constant, whereas $g(x)$, $h(x)$, and $f(x,t)$ are one-periodic functions.

Following the same process as done in the linear parabolic case, we get

$$\frac{d^2}{dt^2}a_{J,p}(t) + \eta(a_{J,p}(t))^* = f_{J,p}(t).$$

In vector notation

$$\left.\begin{aligned} \frac{d^2}{dt^2}\mathbf{a}(t) + \eta\mathbf{a}^*(t) &= \mathbf{f}(t)\\ \mathbf{a}(0) &= \mathbf{a}_g\\ \mathbf{a_t}(0) &= \mathbf{a}_h, \end{aligned}\right\}$$

where

$$\mathbf{a}_h = \{h_{J,p}, \quad p = 0, 1, 2, ..., 2^J - 1\},$$

$$h_{J,p} = \int_0^1 h(x)\tilde{\phi}_{J,p}(x)dx.$$

Applying finite difference formula for temporal discretization, we get

$$\frac{\mathbf{a}_{n+1} - 2\mathbf{a}_n + \mathbf{a}_{n-1}}{\triangle t^2} + \eta \mathbf{Y}\mathbf{a}_{n+1} = \mathbf{f_{n+1}},$$

where $\mathbf{a}_n = \mathbf{a}(n\triangle t)$. This leads to the recursive relation

$$\mathbf{a}_{n+1} = (I + \eta\triangle t^2\mathbf{Y})^{-1}(2\mathbf{a}_n - \mathbf{a}_{n-1} + \triangle t^2\mathbf{f_{n+1}}). \tag{11.34}$$

From the initial condition, we have

$$\mathbf{a_{-1}} = \mathbf{a_1} - 2\triangle t\mathbf{a_h}.$$

So

$$\mathbf{a}_1 = (2I + \eta\triangle t^2\mathbf{Y})^{-1}(2\mathbf{a}_0 + 2\triangle t\mathbf{a_h} + \triangle t^2\mathbf{f_1}).$$

Using Eq. (11.34), we obtain the wavelet coefficients and subsequently solution at desired time.

Let us consider nonlinear fourth-order wave equation

$$\left.\begin{array}{r} \frac{\partial^2 u}{\partial t^2}(x,t) + \eta\frac{\partial^4 u}{\partial x^4}(x,t) + u^2(x,t) = f(x,t) \\ \mathbf{a}(0) = \mathbf{a}_g \\ \mathbf{a_t}(0) = \mathbf{a}_h. \end{array}\right\}$$

Following the same process as done in the linear case, we get

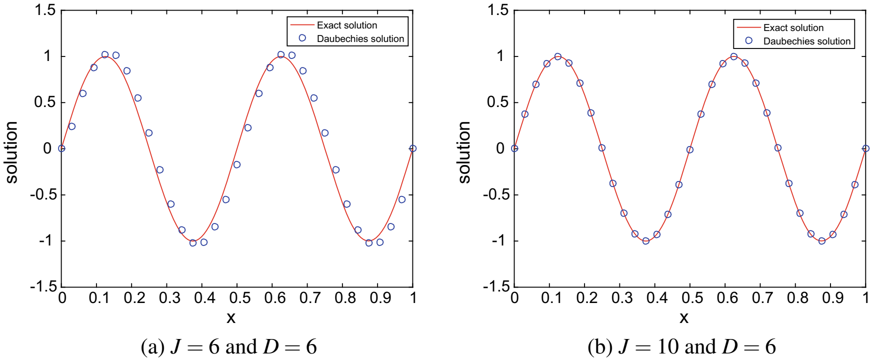
$$\frac{\mathbf{a}_{n+1} - 2\mathbf{a}_n + \mathbf{a}_{n-1}}{\triangle t^2} + \eta \mathbf{Y}\mathbf{a}_{n+1} = \mathbf{f_{n+1}} - \mathbf{U_n} \tag{11.35}$$

$\mathbf{U_n} = \left[\int_0^1 u_n^2\tilde{\phi}_{J,p}(x)dx : p = 0, 1, \ldots, 2^J - 1\right]$. Equation (11.35) leads to the recursive relation

$$\mathbf{a}_{n+1} = (I + \eta\triangle t^2\mathbf{Y})^{-1}(2\mathbf{a}_n - \mathbf{a}_{n-1} + \triangle t^2(\mathbf{f_{n+1}} - \mathbf{U_n})). \tag{11.36}$$

From the initial condition, we have

$$\mathbf{a_{-1}} = \mathbf{a_1} - 2\triangle t\mathbf{a_h}.$$



(a) $J = 6$ and $D = 6$

(b) $J = 10$ and $D = 6$

Fig. 11.1 Comparison between the exact and Daubechies solutions

Table 11.1 Numerical error at various resolution levels

J	$\|u - u_J\|_{J,\infty}$
7	8.5×10^{-2}
8	4.2×10^{-2}
9	2.1×10^{-2}
10	1.0×10^{-2}
11	5.3×10^{-3}
12	2.6×10^{-3}

So

$$\mathbf{a}_1 = (2I + \eta \triangle t^2 \mathbf{Y})^{-1}(2\mathbf{a}_0 + 2\triangle t \mathbf{a_h} + \triangle t^2(\mathbf{f_1} - \mathbf{U_0})).$$

Using Eq. (11.36), we obtain the wavelet coefficients and subsequently solution at desired time.

11.4 Numerical Results

Example 1 Consider the linear elliptic partial differential equation

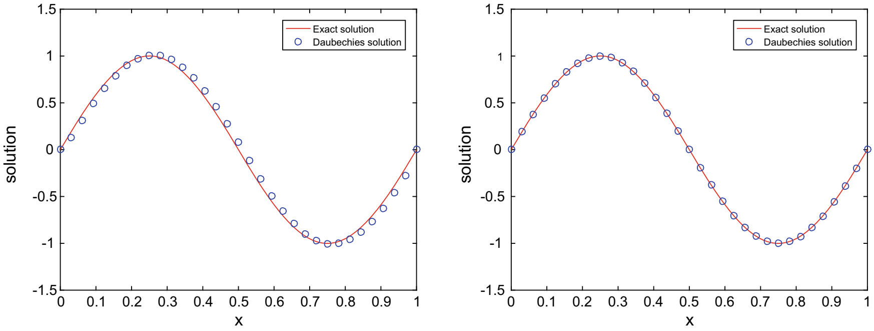
$$\frac{d^4u}{dx^4}(x) = 256\pi^4 \sin(4\pi x), \quad x \in \mathbb{R} \tag{11.37}$$

with one-periodic boundary conditions.

The exact solution of (11.37) is

$$u(x) = \sin(4\pi x).$$

The exact and Daubechies solutions at different resolution levels are reported in Fig. 11.1, whereas Table 11.1 presents max norm error at different resolution levels.



(a) $J = 6$ and $D = 6$

(b) $J = 10$ and $D = 6$

Fig. 11.2 Comparison between the exact and Daubechies solutions

Table 11.2 Numerical error at various resolution levels

J	$\|u - u_J\|_{J,\infty}$
7	4.0×10^{-2}
8	1.9×10^{-2}
9	1.0×10^{-2}
10	5.0×10^{-3}
11	2.5×10^{-3}
12	1.2×10^{-3}

Due to large oscillation in the solution u, we are achieving a good accuracy at high resolution level (see, Table 11.1).

Example 2 Consider the nonlinear elliptic partial differential equation

$$\frac{d^4u}{dx^4}(x) + u^2(x) = 16\pi^4 \sin(2\pi x) + \sin^2(2\pi x), \; x \in \mathbb{R} \tag{11.38}$$

with one-periodic boundary conditions.

The exact solution of (11.38) is

$$u(x) = \sin(2\pi x).$$

The exact and Daubechies solutions at different resolution levels are reported in Fig. 11.2, whereas Table 11.2 presents max norm error at different resolution levels.

It is observed that maximum error reduces with increasing level of resolution.

Example 3 Consider the 2D biharmonic equation

$$\frac{\partial^4 u}{\partial x^4}(x, y) + 2\frac{\partial^4 u}{\partial x^2 \partial y^2}(x, y) + \frac{\partial^4 u}{\partial y^4}(x, y) = 400\pi^4 \sin(4\pi x) \sin(2\pi y), \tag{11.39}$$

with one-periodic boundary conditions.

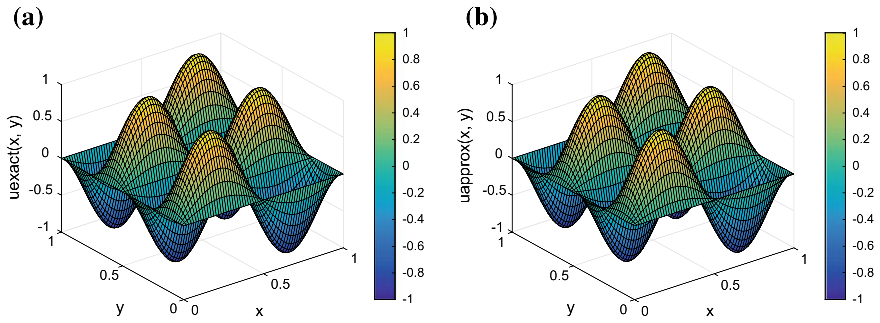


Fig. 11.3 **a** Exact solution. **b** Daubechies solution at $J = 7$ and $D = 6$

Table 11.3 Numerical error at various resolution levels

J	$\|u - u_J\|_{J,\infty}$
7	8.0×10^{-2}
8	4.1×10^{-2}
9	2.1×10^{-2}
10	1.0×10^{-2}
11	5.1×10^{-3}
12	2.5×10^{-3}

The exact solution of (11.39) is

$$u(x, y) = \sin(4\pi x)\sin(2\pi y).$$

The exact and Daubechies solutions at resolution level 8 are reported in Fig. 11.3. Table 11.3 presents max norm error at different resolution levels.

Example 4 Consider the linear parabolic partial differential equation

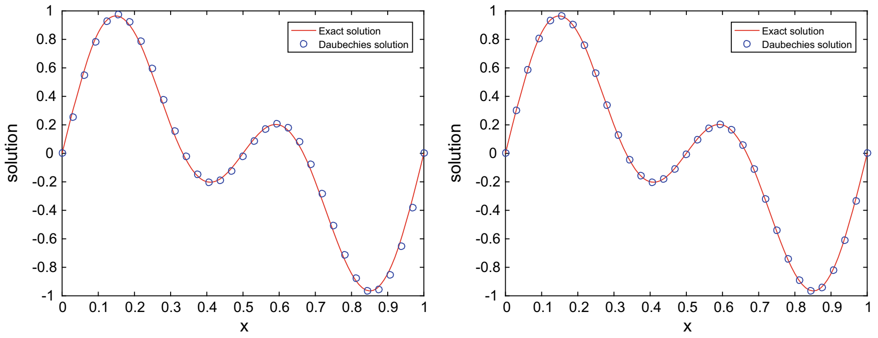
$$\left.\begin{aligned} \frac{\partial u}{\partial t}(x,t) + \frac{1}{16\pi^4}\frac{\partial^4 u}{\partial x^4}(x,t) &= 15e^{-t}\sin(4\pi x), \\ u(x,0) &= \sin(2\pi x) + \sin(4\pi x), \end{aligned}\right\} \quad (x,t) \in \mathbb{R} \times [0,T] \tag{11.40}$$

with one-periodic boundary conditions.

The exact solution of (11.40) is

$$u(x, t) = e^{-t}(\sin(2\pi x) + \sin(4\pi x)).$$

Figure 11.4 presents the exact and Daubechies solutions at time $t = 0.6$. Max norm errors at various time points and resolution levels are reported in Table 11.4.



(a) $J=7, D=6$ and $k=10^{-2}$ (b) $J=9, D=6$ and $k=10^{-2}$

Fig. 11.4 Comparison between the exact and Daubechies solutions

Table 11.4 Numerical error at various time points and resolution levels

J	k	At time	Max norm error
9	10^{-2}	0.2	3.0×10^{-2}
		0.4	2.4×10^{-2}
		0.6	2.0×10^{-2}
		0.8	1.6×10^{-2}
10	10^{-2}	0.2	1.5×10^{-2}
		0.4	1.2×10^{-2}
		0.6	1.0×10^{-2}
		0.8	8.3×10^{-3}
11	10^{-2}	0.2	7.5×10^{-3}
		0.4	6.2×10^{-3}
		0.6	5.1×10^{-3}
		0.8	4.1×10^{-3}

Note: Due to large oscillation in the solution, we have achieved a good accuracy at very high resolution level (see Table 11.4).

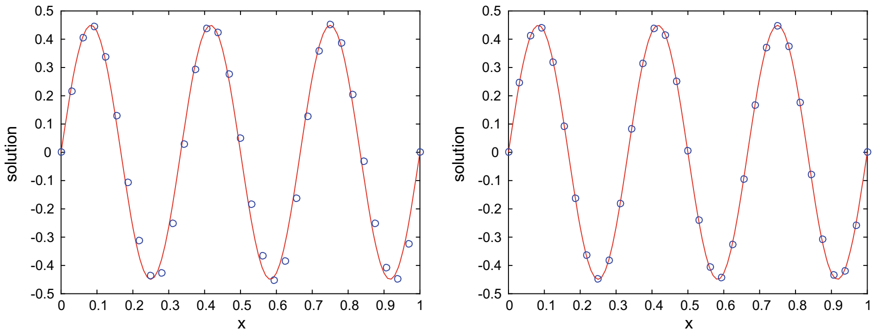
Example 5 Consider the nonlinear parabolic partial differential equation

$$\left.\begin{array}{r}\dfrac{\partial u}{\partial t}(x,t)+\dfrac{1}{1296\pi^4}\dfrac{\partial^4 u}{\partial x^4}(x,t)+u^2(x,t)=e^{-2t}\sin^2(6\pi x),\\ u(x,0)=\sin(6\pi x),\end{array}\right\}\ (x,t)\in\mathbb{R}\times[0,T] \tag{11.41}$$

with one-periodic boundary conditions.

The exact solution of (11.41) is

$$u(x,t)=e^{-t}\sin(6\pi x).$$



(a) $J = 7, D = 6$ and $k = 10^{-3}$

(b) $J = 10, D = 6$ and $k = 10^{-3}$

Fig. 11.5 Comparison between the exact and Daubechies solutions

Table 11.5 Numerical error at various time points and resolution levels

J	k	At time	Max norm error
8	10^{-2}	0.2	6.0×10^{-2}
		0.4	5.0×10^{-2}
		0.6	4.7×10^{-2}
		0.8	4.3×10^{-2}
9	10^{-2}	0.2	3.0×10^{-2}
		0.4	2.5×10^{-2}
		0.6	2.3×10^{-2}
		0.8	2.1×10^{-2}
10	10^{-2}	0.2	1.5×10^{-2}
		0.4	1.2×10^{-2}
		0.6	1.1×10^{-2}
		0.8	1.0×10^{-2}

Figure 11.5 presents the exact and Daubechies solutions at time $t = 0.8$. Max norm errors at various time points and resolution levels are reported in Table 11.5.

Due to large oscillation in the solution u, we have to calculate the solution at very high resolution level to achieve a good accuracy (see Table 11.5).

Example 6 Consider the fourth-order diffusion equation in 2D

$$\left.\begin{aligned} \frac{\partial u}{\partial t}(x, y, t) + \frac{1}{1024\pi^4}\Delta^2 u(x, y, t) = 0, \\ u(x, y, 0) = \sin(4\pi x)\sin(4\pi y), \end{aligned}\right\} \quad (x, y, t) \in \mathbb{R} \times \mathbb{R} \times [0, T] \quad (11.42)$$

with one-periodic boundary conditions.

The exact solution of (11.42) is

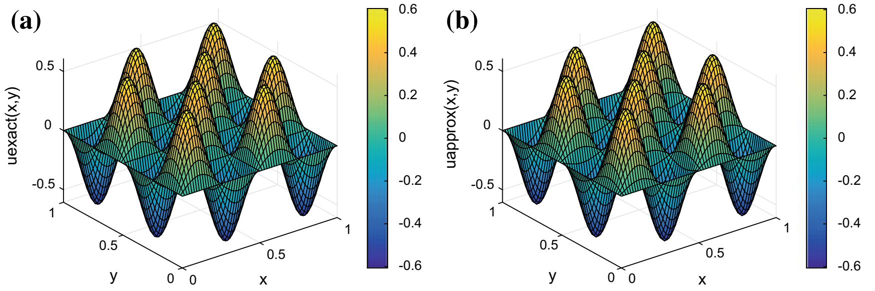


Fig. 11.6 **a** Exact solution. **b** Daubechies solution at $J = 7$ and $D = 6, k = 10^{-2}$

Table 11.6 Numerical error at various time points and resolution levels

J	k	At time	Max norm error
8	10^{-2}	0.2	4.1×10^{-2}
		0.4	3.2×10^{-2}
		0.6	2.6×10^{-2}
		0.8	2.2×10^{-2}
9	10^{-2}	0.2	2.0×10^{-2}
		0.4	1.6×10^{-2}
		0.6	1.3×10^{-2}
		0.8	1.1×10^{-2}
10	10^{-2}	0.2	1.0×10^{-2}
		0.4	8.0×10^{-3}
		0.6	6.5×10^{-3}
		0.8	5.5×10^{-3}

$$u(x, y, t) = e^{-t} \sin(4\pi x) \sin(4\pi y).$$

Figure 11.6a and b presents the exact and Daubechies solutions at time $t = 0.5$. Max norm errors at various time points and resolution levels are reported in Table 11.6.

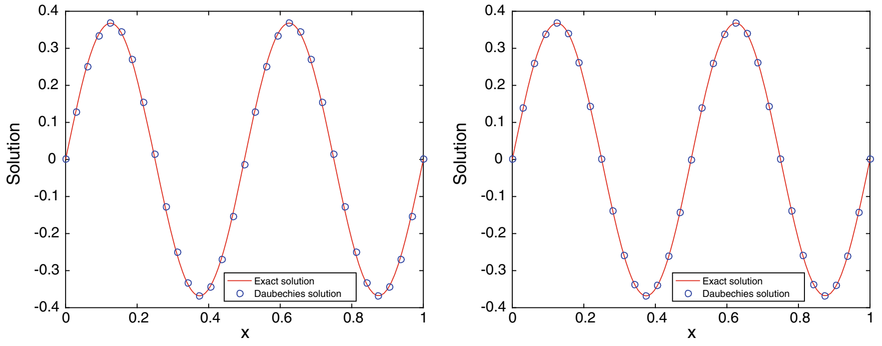
Example 7 Consider the fourth-order linear wave equation

$$\left.\begin{aligned} \frac{\partial^2 u}{\partial t^2}(x,t) + \frac{1}{64\pi^4}\frac{\partial^4 u}{\partial x^4}(x,t) &= 8e^{-2t}\sin(4\pi x), \\ u(x,0) &= \sin(4\pi x), \\ u_t(x,0) &= -2\sin(4\pi x), \end{aligned}\right\} \quad (x,t) \in \mathbb{R} \times [0,T] \tag{11.43}$$

with one-periodic boundary conditions.

The exact solution of (11.43) is

$$u(x, t) = e^{-2t} \sin(4\pi x).$$



(a) $J=8, D=6$ and $\Delta t=10^{-3}$ (b) $J=11, D=6$ and $\Delta t=10^{-3}$

Fig. 11.7 Comparison between the exact and Daubechies solutions at $t=0.5$

Table 11.7 Numerical error at various time points and resolution levels

J	Δt	At time	Max norm error
8	10^{-3}	0.25	2.9×10^{-2}
		0.50	1.8×10^{-2}
		0.75	1.1×10^{-2}
9	10^{-3}	0.25	1.5×10^{-2}
		0.50	9.0×10^{-3}
		0.75	5.5×10^{-3}
10	10^{-3}	0.25	7.4×10^{-3}
		0.50	4.5×10^{-3}
		0.75	2.7×10^{-3}

Figure 11.7 presents the exact and Daubechies solutions at time $t=0.5$. Max norm errors at various time points and resolution levels are reported in Table 11.7.

Example 8 Consider the fourth-order nonlinear wave equation

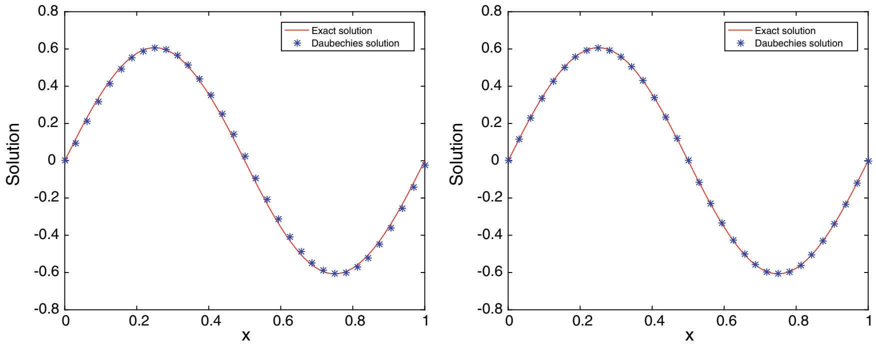
$$\left.\begin{array}{r}\frac{\partial^2 u}{\partial t^2}(x,t)+\frac{1}{8\pi^4}\frac{\partial^4 u}{\partial x^4}(x,t)+u^2(x,t)=3e^{-t}\sin(2\pi x)+e^{-2t}\sin^2(2\pi x),\\ u(x,0)=\sin(2\pi x),\\ u_t(x,0)=-\sin(2\pi x),\end{array}\right\}\quad (x,t)\in\mathbb{R}\times[0,T] \tag{11.44}$$

with one-periodic boundary conditions.

The exact solution of (11.44) is

$$u(x,t)=e^{-t}\sin(2\pi x).$$

Figure 11.8 presents the exact and Daubechies solutions at time $t=0.5$. Max norm errors at various time points and resolution levels are reported in Table 11.8.



(a) $J = 7, D = 6$ and $\Delta t = 10^{-3}$

(b) $J = 10, D = 6$ and $\Delta t = 10^{-3}$

Fig. 11.8 Comparison between the exact and Daubechies solutions at $t = 0.5$

Table 11.8 Numerical error at various time points and resolution levels

J	Δt	At time	Max norm error
8	10^{-3}	0.25	1.9×10^{-2}
		0.50	1.5×10^{-2}
		0.75	1.1×10^{-2}
9	10^{-3}	0.25	9.6×10^{-3}
		0.50	7.5×10^{-3}
		0.75	5.8×10^{-3}
10	10^{-3}	0.25	4.8×10^{-3}
		0.50	3.7×10^{-3}
		0.75	3.0×10^{-3}

11.5 Conclusion

In this paper, we have developed wavelet Galerkin methods for linear and nonlinear partial differential equations. For the efficient and accurate evaluation of integrals consisting of derivatives or product of derivatives, we have used the two-term connection coefficients table. Sparse GMRES solver has been used to solve linear system of algebraic equations. Error estimates are derived in order to show the convergence of proposed method. Finally, numerical results are presented and it is shown that the numerical results are in good agreement with exact results.

References

1. G. Beylkin, On the representation of operators in bases of compactly supported wavelets. SIAM J. Numer. Anal. **6**, 1716–1740 (1992)
2. G. Beylkin, R. Coifman, V. Rokhlin, Fast wavelet transforms and numerical algorithms. Commun. Pure Appl. Math. **44**, 141–183 (1991)
3. R. Glowinski, W. Lawton, M. Ravachol, E. Tenenbaum, Wavelet solutions of linear and nonlinear elliptic, parabolic and hyperbolic problems in one dimension, in *Proceedings of 9th International Conference on Numerical Methods in Applied Sciences and Engineering* (SIAM, Philadelphia, 1990)
4. S. Qian, J. Weiss, Wavelets and the numerical solution of boundary value problems. Appl. Math. Lett. **6**, 47–52 (1993)
5. A. Amaratunga, J. Williams, S. Qian, J. Weiss, Wavelet Galerkin solutions for one-dimensional partial differential equations. Int. J. Numer. Meth. Eng. **37**, 2703–2716 (1994)
6. A. Cohen, R. Masson, Wavelet methods for second-order elliptic problems, preconditioning and adaptivity. SIAM. J. Sci. Comput. **21**, 1006–1026 (1999)
7. A. Cohen, R. Masson, Wavelet adaptive method for second order elliptic problems: boundary conditions and domain decomposition. Numer. Math. **86**, 193–238 (2000)
8. S. Kestler, K. Urban, Adaptive wavelet methods on unbounded domains. J. Sci. Comput. **53**, 342–376 (2012)
9. T. Gantumur, H. Harbrecht, R. Stevenson, An optimal adaptive wavelet method without coarsening of the iterands. Math. Comput. **76**, 615–629 (2007)
10. R. Stevenson, An optimal adaptive finite element method. SIAM J. Numer. Anal. **42**, 2188–2217 (2005)
11. A. Cohen, W. Dahmen, R. DeVore, Adaptive wavelet methods for elliptic operator equations: convergence rates. Math. Comput. **70**, 27–75 (2001)
12. Z. Shi, Y.Y. Cao, Q.J. Chen, Solving 2D and 3D Poisson equations and biharmonic equations by the Haar wavelet method. Appl. Math. Model. **36**, 5143–5161 (2012)
13. S. Bertoluzza, V. Perrier, A new construction of boundary interpolating wavelets for fourth order problems. Acta Applicandae Mathematicae **152**(33), 56 (2017)
14. S. Qian, J. Weiss, Wavelets and the numerical solution of partial differential equations. J. Comput. Phys. **106**, 155–175 (1993)
15. S. Dahlke, M. Lindemann, G. Teschke, M. Zhariy, M. J. Soares, P. Cerejeiars, U. Kähler, A convergent numerical scheme for nonlinear elliptic partial differential equations, Technical report
16. G. Priyadarshi, B.V. Rathish Kumar, Wavelet Galerkin method for fourth order linear and nonlinear differential equations. Appl. Math. Comput. **327**, 8–21 (2018)
17. B.V. Rathish Kumar, G. Priyadarshi, Wavelet Galerkin method for fourth order multidimensional elliptic partial differential equations. Int. J. Wavelets Multiresolut. Inf. Process **16** (2018)
18. B.V. Rathish Kumar, M. Mehra, A wavelet-Taylor Galerkin method for parabolic and hyperbolic partial differential equations. Int. J. Comput. Meth. **2**, 75–97 (2005)
19. B.V. Rathish Kumar, M. Mehra, A time-accurate pseudo-wavelet scheme for parabolic and hyperbolic PDE's. Nonlinear Anal. **63**, 345–356 (2005)
20. B. V. Rathish Kumar, M. Mehra, A three-step wavelet Galerkin method for parabolic and hyperbolic partial differential equations. Int. J. Comput. Math. **83**, 143–157 (2006)
21. M. Mehra, B.V. Rathish Kumar, Time accurate solution of advection diffusion problems by wavelet Taylor Galerkin method. Commun. Numer. Meth. Eng. **21**, 313–326 (2005)
22. M. Mehra, B.V. Rathish Kumar, Time accurate fast three step wavelet Galerkin method for partial differential equations. Int. J. Wavelets Multiresolut. Inf. Process. **4**, 65–79 (2006)
23. M. Mehra, B.V. Rathish Kumar, Error estimates for time accurate wavelet based schemes for hyperbolic partial differential equations. Int. J. Wavelets Multiresolut. Inf. Process. **5**, 667–678 (2007)
24. M. Mehra, B.V. Rathish Kumar, Error estimates for linear parabolic PDEs solved by wavelet based Taylor Galerkin schemes. Int. J. Wavelets Multiresolut. Inf. Process. **7**, 143–162 (2009)

25. J.M. Alam, N.K.R. Kevlahan, O.V. Vasilyev, Simultaneous space-time adaptive wavelet solution of nonlinear parabolic differential equations. J. Comput. Phys. **214**, 829–857 (2006)
26. S. Henn, A Multigrid method for a fourth order diffusion equation with application to image processing. SIAM J. Sci. Comput. **27**, 831–849 (2005)
27. G. Priyadarshi, B.V. Rathish Kumar, Wavelet Galerkin schemes for higher order time dependent partial differential equations. Numer. Methods Partial Differ. Equ. **34**, 982–1008 (2018)
28. L. Yacheng, X. Runzhang, A class of fourth order wave equations with dissipative and nonlinear strain terms. J. Differ. Equ. **244**, 200–228 (2008)
29. S.P. Levandosky, Decay estimates for fourth order wave equations. J. Differ. Equ. **143**, 360–413 (1998)
30. A. Latto, H.L. Resnikoff, E. Tenenbaum, The evaluation of connection coefficients of compactly supported wavelets, in *Proceedings of the French-USA Workshop on Wavelets and Turbulence* (Princeton University, Princeton, 1991)
31. I. Daubechies, Orthonormal basis of compactly supported wavelets. Commun. Pure Appl. Math. **41**, 909–996 (1988)
32. I. Daubechies, *Ten Lectures on Wavelets* (SIAM, Philadelphia, 1992)
33. Y. Meyer, *Ondelettes et Operateurs* (Hermann, Paris, 1990)
34. W. Dahmen, S. Prossdorf, R. Schneider, Wavelet approximation methods for pseudodifferential equation I: stability and convergence. Math. Zeit. **215**, 583–620 (1994)

Chapter 12
Resilience and Dynamics of Coral Reefs Impacted by Chemically Rich Seaweeds and Unsustainable Fishing

Samares Pal and Joydeb Bhattacharyya

Abstract Coral reefs are globally threatened by numerous natural and anthropogenic impacts. The proliferation of seaweeds in coral reefs is one of the most common and significant reasons for the decline of healthy corals. Some seaweeds release chemicals that are harmful to corals. The chemicals released by toxic seaweeds damage corals in areas of direct contact. While herbivorous reef fish play an important role in preventing the overgrowth of seaweeds on corals, unsustainable fishing of herbivores disrupts the ecological balance in coral reefs. This induces changes in the community structure from the dominant reef-building corals to one by seaweeds. We have considered a mathematical model of interactions between coral, toxic seaweeds, and herbivores to investigate the phase shifts from coral- to seaweed-dominated states. We investigate how seaweed toxicity and overfishing trigger negative effects on the ecological resilience of coral reefs through trophic cascades. It is observed that in the presence of seaweed toxicity and unsustainable fishing, the system can exhibit an irreversible dynamics through hysteresis cycles. Further, we employ Mawhin's coincidence degree theory to investigate the existence of a unique positive almost periodic solution of the nonautonomous version of our model by incorporating synchronous or asynchronous seasonal variations in different parameters. The results from computer simulations have potential applications to control the overgrowth of seaweeds in coral reefs as well as to prevent coral bleaching.

Keywords Invasion · Phase shift · Hysteresis · Resilience · Seasonality · Harvesting

S. Pal (✉)
Department of Mathematics, University of Kalyani, Nadia 741235, WB, India
e-mail: samaresp@gmail.com

J. Bhattacharyya
Department of Mathematics, Karimpur Pannadevi College, Nadia 741152, WB, India
e-mail: b.joydeb@gmail.com

P. Manchanda et al. (eds.), *Mathematical Modelling, Optimization, Analytic and Numerical Solutions*, Industrial and Applied Mathematics,
https://doi.org/10.1007/978-981-15-0928-5_12

12.1 Introduction

Coral reefs are among the most species-rich and productive, yet vulnerable marine habitats around the world. Exposed to numerous natural and anthropogenic stresses, coral reef ecosystems do not necessarily respond smoothly to gradual changes in slow variables, which are often caused by these stresses. Instead, they can switch rapidly into a new regime when a threshold level of a controlling parameter in the system is passed, a process termed as regime or phase shift [1, 2]. The resilience of coral reefs can be thought of as the ability of reefs to resist and recover from recurrent stresses without switching to an alternative stable state. Seaweeds play many important ecosystem functioning roles in coral reefs. Despite their importance in coral reefs, the proliferation of seaweeds on coral reefs is increasingly related to the reduction in the resilience of coral reefs [3, 4]. Done [5] and Bellwood et al. [6] observed that the loss of resilience of coral reefs corresponding to the reduction in the adaptive capacity of coral reefs can lead to a shift of regime from coral-dominated state to an alternate stable state, typically dominated by seaweeds or other benthic organisms [7, 8]. Phase shifts in coral reefs are largely the result of seaweeds displacing corals by means of shading and allelopathic chemical defences [9, 10]. Also, faster growing seaweeds dominate coral reefs by reducing the available space for the successful settlement of corals [11, 12]. Although the growth rate of corals is less compared to that of seaweeds, once bleached, corals can return to dominate seaweeds within a decade [13].

The resilience of an ecosystem is a dynamic property of the system that changes through time. Natural and anthropogenic stresses often lead to a slow erosion of resilience of the ecosystem, which goes unnoticed until a perturbation that could have been absorbed previously leads to a catastrophic shift into a different regime. There are two views of resilience recognized in the ecological literature—ecological resilience and engineering resilience. Ecological resilience, described by Holling [14], deals with systems having multiple attractors. It can be defined as the ability of the system to absorb disturbances without being shifted to an alternative basin of attraction. According to Walker et al. [15], ecological resilience is characterized by the four key features—latitude, resistance, precariousness, and panarchy. The first three can be applicable to a system that makes it up, whereas panarchy describes cross-scale interactions and how perturbations at one scale may create regime shifts at some other scale of observation. Declines in reef-building corals have been reported across different regions due to rapid loss of herbivores and high macroalgal toxicity [16]. In coral reef ecosystems, there is a competition between seaweeds and corals to encroach the available space in seabed and when there is a decline in herbivore biomass, seaweeds overgrow corals by allelopathic interactions and by depriving the corals of essential sunlight [12, 17–19]. There are at least two alternate attractors on some coral reef ecosystems, one dominated by seaweeds and the other is dominated by corals. The ecological resilience of these systems is the amount of disturbance that the systems can absorb without switching to an alternative steady state. Since each alternative regime is stabilized by a distinct set of feedbacks, reverting back the system from seaweeds-dominated regime to the regime dominated by corals becomes difficult. It is, therefore, evident that the threshold for a change in regime and its subsequent recovery can become different, a phenomenon called hysteresis.

On the other hand, engineering resilience is applicable for ecosystems having a single attractor and is quantified as the return rate to the steady state after a small perturbation. In nearly pristine coral reefs [20], the macroalgal cover is low compared to coral cover and the herbivore grazing helps in increasing the resilience of the coral-dominated reef. As observed by the researchers [21–23], overfishing of herbivores in coral reefs is one of the reasons for the proliferation of seaweeds on coral reefs that cause a permanent change in the regime dominated by seaweeds. If the growth of macroalgae in coral reefs is not kept in check by the grazers, it becomes difficult to shift the system back toward the coral-dominated state, resulting in a seaweeds-dominated single attractor state. In this situation, the resilience of the system can be measured by the time of return to the seaweeds-dominated steady state after an arbitrary perturbation.

Several seaweeds species contain varying levels of harmful hydrophobic compounds that damage coral tissues [24, 25]. As observed by Andras et al. [26] and Rasher et al. [27], the presence of toxic seaweeds in coral reefs leads to a significant reduction in fecundity and an increase in the mortality rate of corals. Researchers [24] found that toxic seaweeds *Chlorodesmis fastigiata*, when in contact with the coral species *A. millepora*, produce a concoction of chemicals which is lethal to corals. Our proposed model is an extension of the models studied in [34, 35] under the assumption that seaweeds recruits externally from the surrounding seascape and produce toxins which are lethal to corals. The complexity in the model formulation is due to the complexity in the endosymbiotic relationship between corals and microalgae together with the competitive but nonconsumptive direct interactions between corals and macroalgae. As observed by Underwood et al. [28], many reefs are demographically independent and the hydrodynamics associated with these reefs restrict the movement of coral larvae. Thus, we exclude the immigration of coral larvae in our model. We assume that herbivorous Parrotfish follow a logistic growth with macroalgal-biomass-dependent carrying capacity in the absence of harvesting. We analyze the stability and bifurcations by linearizing the system about the equilibrium points, using the techniques previously adopted in [35]. The conditions for stability of the system is determined based on macroalgal toxicity and the harvesting rates of the herbivores. Further, we include the effect of seasonality in the model by means of periodic fluctuation of the model parameters.

In this paper, the main emphasis will be put in studying the effect of overfishing of herbivorous Parrotfish in phase shift from coral to seaweed-dominated systems as well as to examine the feedback in the seaweed-coral-herbivore interaction triangle on the dynamics of coral reef ecosystem. The effects of seasonality on parameters of the interacting species of our proposed model have been reported in this study.

12.2 The Basic Model

In this paper, we consider a mathematical model to investigate the dynamics of corals (C), algal turf (T), and toxic seaweeds (M) competing for a particular area on the seabed in presence of herbivorous Parrotfish (P). We assume that seaweeds can survive in the system irrespective of the abundance of corals. Ignoring the existence of an empty seabed, we have $M(t) + C(t) + T(t) = M(0) + C(0) + T(0) = c_0$ (constant) at any instant t.

The following assumptions are made in formulating the model:

(H_1) Corals are overgrown by seaweeds, at a rate α.

(H_2) Seaweeds spread vegetatively over algal turfs at a rate a.

(H_3) The rate of colonization of newly immigrated seaweeds on turf algae is b.

(H_4) The recruitment rate of corals on algal turfs is r.

(H_5) Seaweeds and corals have natural mortality rates d_1 and d_2, respectively.

(H_6) Mortality rate of corals from seaweed toxicity [24] is γ.

(H_7) The maximum grazing intensity of Parrotfish, in absence of harvesting, is g.

(H_8) The growth rate of Parrotfish is s.

(H_9) The grazing intensity $\frac{gP}{k}$ of Parrotfish is proportional to the abundance of Parrotfish relative to its maximum carrying capacity k with maximal grazing rate, g. The loss of seaweed cover and subsequent recolonization of algal turfs due to grazing is at a rate $\frac{gMP}{k(M+T)}$.

(H_{10}) The harvesting rate of Parrotfish is h.

A schematic diagram of the system is given in Fig. 12.1. The equations representing reef dynamics in presence of grazing are given by:

$$\begin{aligned}
\frac{dM}{dt} &= M\left\{\alpha C - \frac{gP}{k(M+T)} - d_1\right\} + (aM+b)T \\
\frac{dC}{dt} &= C\left\{rT - (\alpha+\gamma)M - d_2\right\} \\
\frac{dT}{dt} &= M\left\{\frac{gP}{k(M+T)} + d_1\right\} + d_2C + \gamma MC - T(aM + rC + b) \\
\frac{dP}{dt} &= P\left[s\left\{1 - \frac{P}{k(M+T)}\right\} - h\right]
\end{aligned} \tag{12.1}$$

where $M(0) > 0$, $C(0) \geq 0$, $T(0) > 0$, and $P(0) \geq 0$.

The parameters considered in the system (12.1) are nonnegative and are given in Table 12.1. The parameters related to macroalgae and corals were derived from empirical studies in the Leeward Islands [29], southern Caribbean [30], and Central America [31]. The experimental observations by Box et al. [3] shows that the of corals grow in cage control and in no-cage control at a rate (r) 1 cm yr^{-1} and 0.55 cm yr^{-1}, respectively. The yearly mortality rate (d_2) of coral exposed to seaweeds is 0.42 ± 1.62. In Table 12.1, we have considered the mortality rate of coral as 0.24 yr^{-1} which lies well within the estimated 95% confidence interval for the yearly mortality rate. The toxic seaweed-induced mortality (γ) of corals is taken as 0.1 yr^{-1}. Thus, the

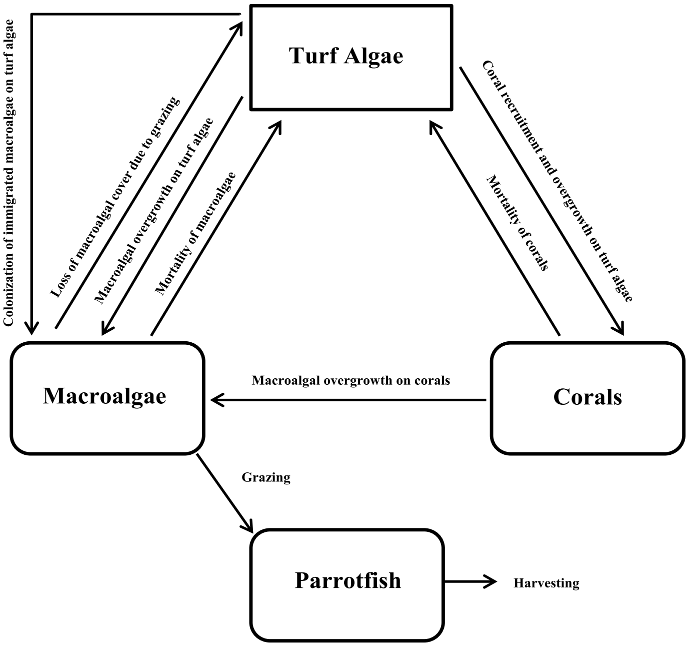

Turf Algae
Colonization of immigrated macroalgae on turf algae
Loss of macroalgal cover due to grazing
Macroalgal overgrowth on turf algae
Mortality of macroalgae
Coral recruitment and overgrowth on turf algae
Mortality of corals
Macroalgae
Macroalgal overgrowth on corals
Corals
Grazing
Parrotfish
Harvesting

Fig. 12.1 Schematic representation of coral-macroalgal competition for occupying turf algae in presence of Parrotfish

estimated mortality rate of corals is $0.34\,\text{yr}^{-1}$. The overgrowth rate of seaweeds on corals (α) is taken from the experimental observations by Lirman [10]. Mumby et al. [31] observed that in absence of environmental perturbations, the vegetative growth rate (a) of seaweeds on algal turf is $1.2\,\text{yr}^{-1}$. The researchers [31] also observed that the growth rate of seaweeds during hurricane period becomes $0.35\,\text{yr}^{-1}$. This gives an estimated average growth rate of seaweeds as $0.77\,\text{yr}^{-1}$. Further, Mumby et al. [31] observed that during hurricane impact, there is a severe loss in the seaweed cover and estimated the average annual loss as 0.083. We have considered the seaweed mortality rate (d_1) to be a combination of natural mortality rate and mortality rate due to external perturbations as $0.1\,\text{yr}^{-1}$. The rate of colonization (b) of seaweeds on algal turf, the intrinsic growth rate (s) of Parrotfish, and the maximum grazing rate (g) of Parrotfish are taken from [32]. We have chosen the harvesting parameter as nonnegative.

Without any loss of generality, we may assume that $c_0 = 1$. Then from (12.1) we obtain

$$\begin{aligned}\frac{dM}{dt} &= M\left\{\alpha C - \frac{gP}{k(1-C)} - d_1\right\} + (aM+b)(1-M-C) \equiv f^1\\ \frac{dC}{dt} &= C\left\{r(1-M-C) - (\alpha+\gamma)M - d_2\right\} \equiv f^2 \qquad (12.2)\end{aligned}$$

$$\frac{dP}{dt} = P\left[s\left\{1 - \frac{P}{k(1-C)}\right\} - h\right] \equiv f^3$$

where $0 < M(0) < 1, 0 \leq C(0) < 1$ and $P(0) \geq 0$.

The right-hand sides of the system (12.2) are smooth functions of M, C, P and the parameters. The existence and uniqueness of solutions of the system (12.2) hold in the positive octant as long as the variables and the parameters are nonnegative.

Lemma 12.2.1 *For all $\epsilon > 0$, there exists $t_\epsilon > 0$ such that all the solutions of (12.2) enter into the set $\left\{(M, C, P) \in \boldsymbol{R}^3 : M(t) + C(t) + P(t) < 1 + k + \epsilon\right\}$ whenever $t \geq t_\epsilon$.*

Proof We have $\frac{d}{dt}(P(t)) \leq sP(t)\left(1 - \frac{P(t)}{k}\right)$.

Let $u(t)$ be the solution of $\frac{d}{dt}u(t) = su(t)\left(\frac{k-u(t)}{k}\right)$, satisfying $u(0) = P(0)$.

Then $u(t) = k\left(1 + \frac{k-P(0)}{P(0)}e^{-st}\right)^{-1} \to k$ as $t \to \infty$.

Applying the standard theorem of differential inequality it follows that $\lim_{t\to\infty} \sup P(t) \leq k$.

Also, $M(t) + C(t) + T(t) = 1$ for all $t \geq 0$ implies $M(t) + C(t) + P(t) \leq 1 + k$ as $t \to \infty$.

12.3 Equilibria and Their Stability

In this section, we determine the equilibrium solutions of the model and investigate the effect of parameters on the stability of the system at the biologically feasible equilibria.

The system (12.2) possesses the following equilibria:
(i) Coral and Parrotfish-free equilibrium $E_0 = (M_0, 0, 0)$, where $M_0 = \frac{a-b-d_1+\sqrt{(a+b-d_1)^2+4bd_1}}{2a}$.

E_0 always exists;

(ii) Parrotfish-free equilibrium $E_1 = (M_1, C_1, 0)$, where $C_1 = p + qM_1$, $p = 1 - \frac{d_2}{r}$, $q = -\frac{r+\alpha+\gamma}{r}$ and
$M_1 = \frac{(r-d_2)\alpha-rd_1+ad_2+b(\alpha+\gamma)+\sqrt{\{(r-d_2)\alpha-rd_1+ad_2+b(\alpha+\gamma)\}^2+4bd_2\{(\alpha-a)(\alpha+\gamma)+r\alpha\}}}{2\{(\alpha-a)(\alpha+\gamma)+r\alpha\}}$.
E_1 exists if $(\alpha - a)(\alpha + \gamma) + r\alpha > 0$;

(iii) coral-free equilibrium in presence of seaweeds and Parrotfish $E_2 = (M_2, 0, P_2)$, where
$M_0 = \frac{a-b-d_1-g\left(1-\frac{h}{s}\right)+\sqrt{\left\{a-b-d_1-g\left(1-\frac{h}{s}\right)\right\}^2+4ab}}{2a}$ and $P_2 = k\left(1 - \frac{h}{s}\right)$. E_2 exists if $h < s$;

(iv) interior equilibrium $E^* = (M^*, C^*, P^*)$, where M^* is a positive root of the equation $\sum_{i=1}^{3} a_i M^{4-i} = 0$, $C^* = p_1 + q_1 M^*$ and $P^* = p_2 + q_2 M^*$, where $a_1 = \frac{k}{r^2}(r + \alpha + \gamma)[(a - \alpha)(\alpha + \gamma) - r\alpha]$, $a_2 = k\alpha q_1(1 - p_1) - kq_1(\alpha p_1 - d_1) - gq_2 + bk(1 + q_1) - ak(1 - p_1)(1 + 2q_1)$, $a_3 = k(\alpha p_1 - d_1)(1 - p_1) - gp_2 -$

$bk(1-p_1)(2q_1+1)+ak(1-p_1)^2, a_4 = bk(1-p_1)^2, p_1 = 1-\frac{d_2}{r}, p_2 = \frac{k(s-h)(1-p_1)}{s}, q_1 = -\frac{r+\alpha+\gamma}{r}$ and $q_2 = -\frac{q_1 k(s-h)}{s}$. E^* exists uniquely if $r > (\alpha+\gamma)\left(\frac{a-\alpha}{\alpha}\right)$.

At E_0 the eigenvalues of the Jacobian matrix of the system (12.2) are $-\sqrt{(a-b-d_1)^2+4ab}, r-d_2-M_0(r+\alpha+\gamma)$ and $s-h$. Therefore, all the eigenvalues of the Jacobian matrix are negative if $r+\alpha+\gamma > \frac{r-d_2}{M_0}$ and $h > s$. This gives the following lemma.

Lemma 12.3.1 *The system (12.2) is locally asymptotically stable at E_0 if $\gamma > \gamma_1$ and $h > s$, where $\gamma_1 = \frac{r-d_2}{M_0} - (r+\alpha)$.*

Therefore, with high macroalgal toxicity and high rate of harvesting of Parrotfish, the system stabilizes at seaweeds-dominated steady state with complete elimination of coral and Parrotfish.

Lemma 12.3.2 *If $\alpha > \frac{1}{2}\left\{a+b-d_1+\sqrt{(a+b-d_1)^2+4ad_1}\right\}$ and $h > s$, the system (12.2) undergoes a transcritical bifurcation at E_0 when γ crosses γ_1.*

Proof At $\gamma = \gamma_1$, we have

$$J_0 = \begin{pmatrix} -\sqrt{(a-b-d_1)^2+4ab} & (\alpha-a)M_0-b & -\frac{gM_0}{k} \\ 0 & 0 & 0 \\ 0 & 0 & s-h \end{pmatrix}$$

Therefore, the zero eigenvalue of the Jacobian matrix is simple.

Let V_1 and W_1 be the eigenvectors corresponding to the zero eigenvalue for J_0 and J_0^T, respectively.

Then we obtain $V_1 = \begin{pmatrix} \frac{(\alpha-a)M_0-b}{\sqrt{(a-b-d_1)^2+4ab}} & 1 & 0 \end{pmatrix}^T$ and $W_1 = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}^T$. Let us express the system (12.2) in the form $\dot{X} = f(X;\gamma)$, where $X = \begin{pmatrix} M & C & P \end{pmatrix}^T$ and $f(X;\gamma) = \begin{pmatrix} f^1 & f^2 & f^3 \end{pmatrix}^T$

Then $W^T f_\gamma(M_0, 0, 0; \gamma_1) = 0$ and so no saddle-node bifurcation occurs at E_0 when γ crosses γ_1.

Also, $Df_\gamma(M_0, 0, 0; \gamma_1)V_1 = \begin{pmatrix} 0 & -M_0 & 0 \end{pmatrix}^T$ and so $W^T[Df_\gamma(M_0, 0, 0; \gamma_1)V_1] = -M_0 < 0$.

Now, we have $D^2 f(M_0, 0, 0; \gamma_1)(V_1, V_1) = \begin{pmatrix} \frac{2(\alpha-a)\{(\alpha-a)M_0-b\}}{\sqrt{(a-b-d_1)^2+4ab}} - \frac{2a\{(\alpha-a)M_0-b\}^2}{(a-b-d_1)^2+4ab} \\ -2r - \frac{2(r-d_2)\{(\alpha-a)M_0-b\}}{M_0\sqrt{(a-b-d_1)^2+4ab}} \\ 0 \end{pmatrix}$

This gives $W^T[D^2 f(M_0, 0, 0; \gamma_1)(V_1, V_1)] = -2r - \frac{2(r-d_2)\{(\alpha-a)M_0-b\}}{M_0\sqrt{(a-b-d_1)^2+4ab}}$.

If $\alpha > \frac{1}{2}\left\{a+b-d_1+\sqrt{(a+b-d_1)^2+4ad_1}\right\}$ holds then $W^T[D^2 f(M_0, 0, 0; \gamma_1)(V_1, V_1)] < 0$.

Therefore, if $\alpha > \frac{1}{2}\left\{a+b-d_1+\sqrt{(a+b-d_1)^2+4ad_1}\right\}$ and $h > s$ are satisfied, by Sotomayor theorem [33] it follows that the system (12.2) undergoes a transcritical bifurcation at E_0 when γ crosses γ_1.

The Jacobian $J_1 \equiv J(E_1)$ of the system (12.2) evaluated at an interior equilibrium E_1 is

$$J_1 = \begin{pmatrix} (\alpha C_1 - d_1)(1+M_1) & (\alpha - a)M_1 - b & -\frac{gM_1}{k(1-C_1)^2} \\ -(r+\alpha+\gamma)C_1 & -rC_1 & 0 \\ 0 & 0 & s-h \end{pmatrix}$$

At E_1 one eigenvalue of the Jacobian matrix of the system (12.2) is $s-h$ and the other two eigenvalues are given by the equation $\lambda^2 + \lambda\{rC_1 + (d_1 - \alpha C_1)(1 + M_1)\} + (r+\alpha+\gamma)C_1\{(\alpha-a)M_1 - b\} - rC_1(\alpha C_1 - d_1)(1+M_1) = 0$.

Now $rC_1 + (d_1 - \alpha C_1)(1+M_1) > (r-2\alpha)C_1 > 0$ if $r > 2\alpha$ and $(r+\alpha+\gamma)\{(\alpha-a)M_1 - b\} - r(\alpha C_1 - d_1)(1+M_1) > \{(r-a)d_1 + (\alpha+\gamma)(\alpha-a)\}$ $M_1 - b(r+\alpha+\gamma) - r(\alpha - d_1) > 0$ if $a < \frac{rd_1+\alpha(\alpha+\gamma)}{\alpha+\gamma+d_1}$ and $\frac{b(r+\alpha+\gamma)+r(\alpha-d_1)}{(r-a)d_1+(\alpha+\gamma)(\alpha-a)} < M_1 < 1$. This gives the following lemma.

Lemma 12.3.3 *If $r > 2\alpha$, $a < \frac{rd_1+\alpha(\alpha+\gamma)}{\alpha+\gamma+d_1}$, $\frac{b(r+\alpha+\gamma)+r(\alpha-d_1)}{(r-a)d_1+(\alpha+\gamma)(\alpha-a)} < M_1 < 1$ and $h > s$, the system (12.2) is locally asymptotically stable at E_1.*

Therefore, with high rate of harvesting of Parrotfish, high recruitment rate of corals, and low seaweed growth rate on turf algae, corals, and seaweeds can coexist even in absence of Parrotfish.

At E_2 the eigenvalues of the Jacobian matrix of the system (12.2) are $-\sqrt{\left\{a-b-d_1-g\left(1-\frac{h}{s}\right)\right\}^2+4ab}$, $r - d_2 - M_2(r+\alpha+\gamma)$ and $h-s$.

Therefore, all the eigenvalues of the Jacobian matrix are negative if $r+\alpha+\gamma > \frac{r-d_2}{M_2}$ and $h < s$. This gives the following lemma.

Lemma 12.3.4 *The system (12.2) is locally asymptotically stable at E_2 if $\gamma > \gamma_2$ and $h < s$, where $\gamma_2 = \frac{r-d_2}{M_2} - (r+\alpha)$.*

Therefore, with high seaweed toxicity and low rate of harvesting of Parrotfish, the system stabilizes at seaweeds-dominated steady state in presence of Parrotfish with complete elimination of corals.

Lemma 12.3.5 *If $h < s$ and $\gamma_2 \neq \frac{2r\sqrt{\left\{a-b-d_1+g\left(1-\frac{h}{s}\right)\right\}^2+4ab}}{(\alpha-a)M_2-b} - (r+\alpha)$, the system (12.2) undergoes a transcritical bifurcation at E_2 when γ crosses γ_2.*

Proof At $\gamma = \gamma_2$, we have

$$J_2 = \begin{pmatrix} -\sqrt{\left\{a-b-d_1-g\left(1-\frac{h}{s}\right)\right\}^2+4ab} & \left\{\alpha-a-g\left(1-\frac{h}{s}\right)\right\}M_2 - b & -\frac{gM_2}{k} \\ 0 & 0 & 0 \\ 0 & -sk\left(1-\frac{h}{s}\right)^2 & h-s \end{pmatrix}$$

Therefore, the zero eigenvalue of the Jacobian matrix is simple.

Let V_2 and W_2 be the eigenvectors corresponding to the zero eigenvalue for J_2 and J_2^T, respectively.

Then we obtain $V_2 = \left(\frac{(\alpha-a)M_2-b}{\sqrt{\left\{a-b-d_1+g\left(1-\frac{h}{s}\right)\right\}^2+4ab}} \; 1 \; \frac{k(h-s)}{s} \right)^T$ and $W_2 = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}^T$.

Since $W_2^T f_\gamma(E_2; \gamma_2) = 0$, it follows that no saddle-node bifurcation occurs at E_2 when γ crosses γ_2.

Also, $Df_\gamma(E_2; \gamma_2)V_2 = \begin{pmatrix} 0 & -M_2 & 0 \end{pmatrix}^T$ and so $W_2^T[Df_\gamma(E_2; \gamma_2)V_2] = -M_2 < 0$. This gives $W_2^T[D^2 f(E_2; \gamma_2)(V_2, V_2)] = -2r - \frac{2(r-d_2)\{(\alpha-a)M_2-b\}}{M_2\sqrt{\left\{a-b-d_1-g\left(1-\frac{h}{s}\right)\right\}^2+4ab}}$.

If $\gamma_2 \neq \frac{2r\sqrt{\left\{a-b-d_1+g\left(1-\frac{h}{s}\right)\right\}^2+4ab}}{(\alpha-a)M_2-b} - (r+\alpha)$ holds, then $W_2^T[D^2 f(E_2; \gamma_2)(V_2, V_2)] \neq 0$.

Therefore, if $h < s$ and $\gamma_2 \neq \frac{2r\sqrt{\left\{a-b-d_1+g\left(1-\frac{h}{s}\right)\right\}^2+4ab}}{(\alpha-a)M_2-b} - (r+\alpha)$ hold, the system (12.2) undergoes a transcritical bifurcation at E_2 when γ crosses γ_2.

The interior equilibrium E^* is persistent if the boundary equilibria E_0, E_1, and E_2 repel interior trajectories. We see that the boundary equilibria of the system (12.2) are unstable if $\gamma < \min\{\gamma_1, \gamma_2\}$ and $h < s$. Also, the system is bounded. The following lemma gives the condition of persistence of the system (12.2) at E^*:

Lemma 12.3.6 *The system (12.2) is persistent at E^* if $\gamma < \min\{\gamma_1, \gamma_2\}$ and $h < s$.*

Therefore, with low seaweed toxicity level and low rate of harvesting of Parrotfish, all the organisms in the system coexists.

The Jacobian $J^* \equiv J(E^*)$ of the system (12.2) evaluated at an interior equilibrium E^* is

$$J^* = \begin{pmatrix} -aM^* - \frac{b(1-C^*)}{M^*} & (\alpha-a)M^* - b - \frac{gM^*P^*}{k(1-C^*)^2} & -\frac{gM^*}{k(1-C^*)} \\ -(r+\alpha+\gamma)C^* & -rC^* & 0 \\ 0 & -\frac{sP^{*2}}{k(1-C^*)^2} & -\frac{sP^*}{k(1-C^*)} \end{pmatrix}$$

The characteristic equation of the Jacobian J^* of the system (12.2) is $\lambda^3 + A_1\lambda^2 + A_2\lambda + A_3 = 0$, where

$A_1 = rC^* + aM^* + \frac{b(1-C^*)}{M^*} + \frac{gP^*}{k(1-C^*)^2}$,

$A_2 = \frac{rsC^*P^*}{k(1-C^*)} + \left\{aM^* + \frac{b(1-C^*)}{M^*}\right\}\left\{rC^* + \frac{sP^*}{k(1-C^*)}\right\}$
$\quad + (r+\alpha+\gamma)\left\{(\alpha-a)M^*C^* - bC^* - \frac{gM^*C^*P^*}{k(1-C^*)^2}\right\}$,

$A_3 = \frac{s(r+\alpha+\gamma)C^*P^*}{k(1-C^*)}\{(\alpha-a)M^* - b\}$.

The system is locally asymptotically stable at E^* if $A_1A_2 > A_3$.

At $\alpha = a + \frac{b}{M^*} = \alpha^*$(say), we have $A_2(\alpha^*) = \frac{rsC^*P^*}{k(1-C^*)} + \left\{aM^* + \frac{b(1-C^*)}{M^*}\right\}\left\{rC^* + \frac{sP^*}{k(1-C^*)}\right\} - \frac{gM^*C^*P^*(r+\alpha^*+\gamma)}{k(1-C^*)^2}$. Therefore, at $\alpha = \alpha^*$ if $A_2(\alpha^*) > 0$, then the Jacobian J^* of the system (12.2) has a simple zero eigenvalue.

Table 12.1 Parameter values used in the numerical analysis

Parameters	Description of parameters	Value	Ref.
α	Rate of overgrowth of seaweeds on coral	0.1	[10, 32, 34]
r	Recruitment rate of corals on turf algae	0.55	[3, 32]
a	Rate of vegetative spread of seaweed over algal turfs	0.77	[31, 32]
b	Colonization rate of newly immigrated seaweeds on algal turf	0.005	[32]
d_1	Natural mortality rate of seaweeds	0.1	[31, 35]
d_2	Natural mortality rate of corals	0.24	[3, 35]
γ	Toxin-induced death rate of corals	0.1	[35]
s	Intrinsic growth rate of Parrotfish	0.49	[32]
k	Maximal carrying capacity of Parrotfish	1	[32]
g	Maximal seaweeds-grazing rate of Parrotfish	0.5	[32]
h	Harvesting rate of Parrotfish	0.05	–

Let V^* and W^* are the eigenvectors corresponding to the zero eigenvalue for J^* and J^{*T}, respectively. Then we obtain $V^* = \left(\frac{-r}{r+\alpha^*+\gamma}\ \ 1\ -P^*\right)^T$ and $W^* = \left(1\ -\frac{aM^*+\frac{b(1-C^*)}{M^*}}{C^*(r+\alpha^*+\gamma)}\ -\frac{gM^*}{sP^*}\right)^T$.

Due to the complexity in the algebraic expressions involved, we will use numerical simulations to verify that $W^{*T} f_\alpha(E^*;\alpha^*) \neq 0$ and $W^{*T}[D^2 f(E^*;\alpha^*)(V^*, V^*)] \neq 0$. In this case, the system undergoes a saddle-node bifurcation at E^* when α crosses α^*. By analyzing the system (12.2) we are able to show that a sharp transition with hysteresis can be achieved by varying some of the parameter values.

To identify the impact of seaweed toxicity on corals, we plot the solutions of the nullcline equations projected onto the $C-\gamma$ plane (Fig. 12.2), yielding a bifurcation diagram. The curves of stable interior equilibria are shown in black, stable boundary equilibrium E_2 are shown in blue and unstable equilibria are shown in red. The region I represents monostability at E^* for $0 \leq \gamma < \gamma_2 = \frac{r-d_2}{M_2} - (r+\alpha) = 0.1469$ for all nonnegative initial conditions. In this region, the system will ultimately arrive at a coral-dominated state corresponding to low levels of seaweeds in presence of Parrotfish. The bistable region is represented by the region II for $\gamma_2 < \gamma < \gamma^* = 0.1506$. Once the seaweed toxicity level surpasses the threshold γ^*, the system arrives at a seaweeds-dominated and coral-depleted stable state in presence of Parrotfish, represented by the region III of monostability at E_2. Hysteresis will result, with low seaweed cover followed by an increase in the seaweed cover above the critical threshold γ^*. A backward shift occurs only if the seaweed toxicity level is reduced far enough to reach the other bifurcation point γ_2.

To study the ecological resilience of the system at a particular point on the equilibrium curve in the bistable region II, we consider an arbitrary point A on the curve of stable interior equilibrium in the bistable region. If there is a drop of coral cover from the stable coexistence steady state A upto or just beyond the unstable coexistence state (in red), then there will be a shift of regime to the coral-free steady state in presence of Parrotfish. In this case, the latitude at A is defined as the distance (L) between the stable coexistence state A and its basin boundary. Also, due to the increase of seaweed toxicity beyond γ^*, there is a shift of regime by overcoming the "resistance" of the coexistence state and eroding the size of its basin of attraction. The resistance of the system at A can be defined as the minimum additional seaweed toxicity level required for the complete elimination of corals and is denoted by R. The latitude (L) component of resilience is measured in terms of coral cover and the resistance (R) component of resilience is measured in terms of seaweed toxicity level. The precariousness (Pr) of the system at A is defined as the current position and trajectory of the system in the basin of attraction relative to the edge and can be measured as the linear distance from A to the point of saddle-node bifurcation. The ecological resilience of the system at A can be represented as a combination of the latitude component vector, the resistance component vector and the precariousness component vector at A. The bistable region II in Fig. 12.2 is a representation of the three aspects of ecological resilience of our system at A in terms of seaweed toxicity and coral cover. With low seaweed toxicity, the system has a coral-dominated single attractor. In this case the rate of recovery from small perturbations is an indicator of engineering resilience. From Fig. 12.3a it follows that the recovery time in the monostable region I after an arbitrary perturbation is least in absence of seaweed toxicity and increases due to the increase of seaweed toxicity. Consequently, the engineering resilience of the system in the monostable region I decreases due to the increase of seaweed toxicity. From Fig. 12.3b it follows that the ecological resilience of the system at the interior equilibrium is maximum when seaweed toxicity level is less than γ_2 and decreases in the bistable region II due to the increase of seaweed toxicity. The ecological resilience of the coexistence steady state becomes minimum when seaweed toxicity level approaches the threshold value γ^*. In this case, slight increase in seaweed toxicity leads to a catastrophic shift of regime to a seaweeds-dominated ecosystem in presence of Parrotfish.

From Fig. 12.3c we see that for $0 \leq \gamma < \gamma_2$ two eigenvalues of the Jacobian of the system at E_2 are negative and one eigenvalue is positive, i.e., the fixed point E_2 is unstable. The stability of the system changes when γ crosses γ_2. All the eigenvalues becomesnegative for $\gamma > \gamma_2$, representing the stability of the system at E_2. From Fig. 12.3d it follows that the system (12.2) has a stable node at E^* for $\gamma < \gamma^*$. Also, E^* ceased to exist for $\gamma > \gamma^*$. Thus, there are changes in the stability of the system when γ crosses γ_2 and γ^*. We use numerical simulations to determine the nature of bifurcations at $\gamma = \gamma_2$ and $\gamma = \gamma^*$.

At $\gamma = \gamma_2$, we have $E_2 = (0.389, 0, 0.898)$ and

$$J_2 = \begin{pmatrix} -0.4281 & -0.4853 & -0.1945 \\ 0 & 0 & 0 \\ 0 & -0.3951 & -0.44 \end{pmatrix}$$

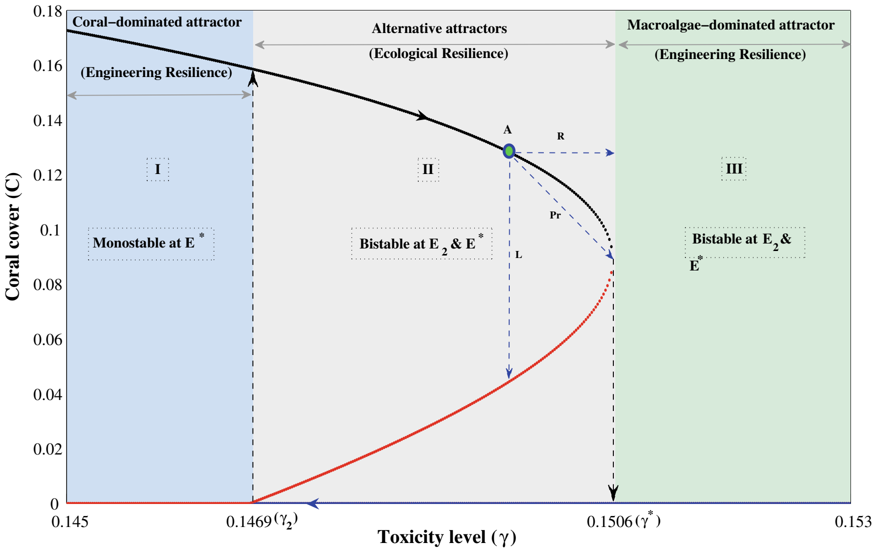


Fig. 12.2 Bifurcation diagram of γ versus the equilibrium value of coral cover, where $h < s$. Coral-dominated stable interior equilibria are indicated by black curves, seaweeds-dominated stable equilibria E_2 are indicated by blue curves, and unstable equilibria by red curves

has a simple zero eigenvalue. Also, we obtain $V_2 = \begin{pmatrix} -0.7257 & 1 & -0.8980 \end{pmatrix}^T$, $W_2 = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}^T$, $W_2^T f_\gamma(E_2; \gamma_2) = 0$, $W_2^T [Df_\gamma(E_2; \gamma_2)V_2] = -0.389 < 0$ and $W_2^T [D^2 f(E_2; \gamma_2)(V_2, V_2)] = 0.0565 > 0$, satisfying the conditions of transcritical bifurcation at E_2 when γ crosses γ_2.

At $\gamma = \gamma^*$, we have $E^* = (0.326, 0.0891, 0.818)$ and

$$J^* = \begin{pmatrix} -0.3907 & -0.4291 & -0.179 \\ -0.0713 & -0.049 & 0 \\ 0 & -0.3951 & -0.44 \end{pmatrix}$$

with eigenvalues 0, -0.5469 and -0.3328. Also, we obtain $V^* = \begin{pmatrix} 0.8769 & 0.1256 & 0.4641 \end{pmatrix}^T$, $W^* = \begin{pmatrix} 0.4742 & -0.3852 & -0.7917 \end{pmatrix}^T$, $W^{*T} f_\gamma(E^*; \gamma^*) = 0.0112 > 0$ and $W^{*T}[D^2 f(E^*; \gamma^*)(V^*, V^*)] = -0.5502 < 0$, satisfying the conditions of saddle-node bifurcation at E^* when γ crosses γ^*.

To identify the impact of seaweed toxicity on coral cover with high rate of harvesting of Parrotfish ($h > s$) and low seaweed recruitment rate on turf algae, we represent a bifurcation diagram in Fig. 12.4a with γ as an active parameter. Coordinates of stable boundary equilibria E_1 are shown in green, stable boundary equilibrium E_0 are shown in cyan and unstable equilibria are shown in red. The region IV represents monostability at E_1 for $0 \le \gamma < \gamma_1 = \frac{r-d_2}{M_0} - (r + \alpha) = 0.0226$ for all nonnegative initial conditions. In this region, the system will ultimately arrive at a coral-dominated state corresponding to low levels of seaweeds in absence of Parrot-

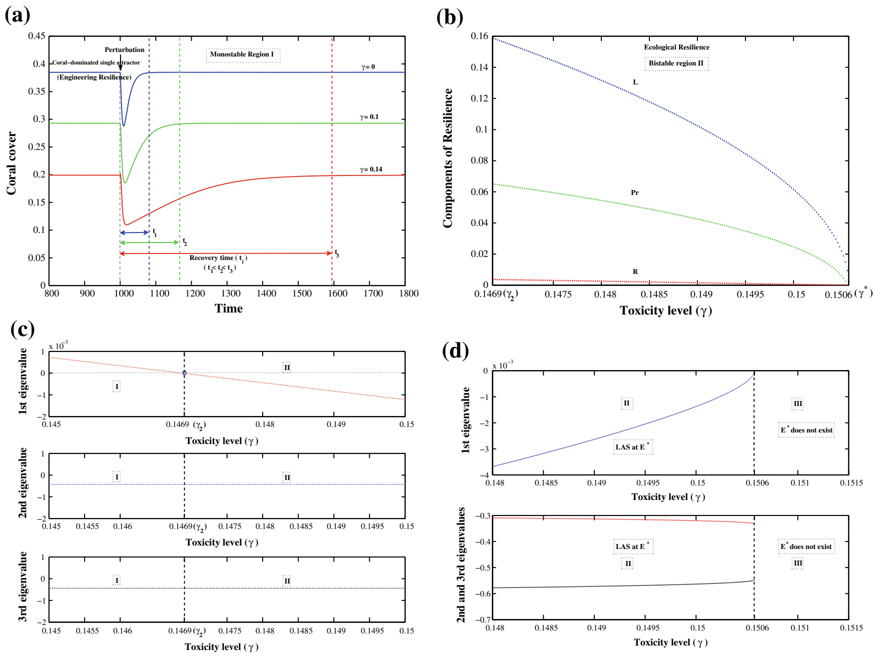


Fig. 12.3 **a** Change in the engineering resilience of the system in the monostable region I with γ as an active parameter. **b** Change in the ecological resilience of the system in the bistable region II with γ as an active parameter. **c** Eigenvalues for the coral-free equilibrium E_2 as functions of γ. **d** Eigenvalues for the interior equilibrium E^* as functions of γ

fish. Once the seaweed toxicity level surpasses the threshold γ_1, the system arrives at a seaweeds-dominated and coral-depleted stable state in absence of Parrotfish, represented by the region V of monostability at E_0. We use numerical simulations to determine the nature of bifurcation at $\gamma = \gamma_1$.

At $\gamma = \gamma_1$, we have $E_0 = (0.4609, 0, 0)$ and

$$J_0 = \begin{pmatrix} -0.144 & -0.0394 & -0.2304 \\ 0 & 0 & 0 \\ 0 & 0 & -0.01 \end{pmatrix}$$

has a simple zero eigenvalue. Also, we obtain $V_1 = \begin{pmatrix}-0.2737 & 1 & 0\end{pmatrix}^T$, $W_1 = \begin{pmatrix}0 & 1 & 0\end{pmatrix}^T$, $W_1^T f_\gamma(E_0; \gamma_1) = 0$, $W_1^T [Df_\gamma(E_0; \gamma_1)V_1] = -0.4609 < 0$ and $W_1^T [D^2 f(E_0; \gamma_1)(V_1, V_1)] = -0.7319 < 0$, satisfying the conditions of transcritical bifurcation at E_0 when γ crosses γ_1.

From Fig. 12.4b it is observed that with high rate of harvesting of Parrotfish (viz. $h = 0.1$), the system is seaweeds-dominated and stable even with low toxicity level of seaweeds. The decrease in the rate of harvesting of Parrotfish increases the latitude component of resilience due to the increase of coral cover. Also, the

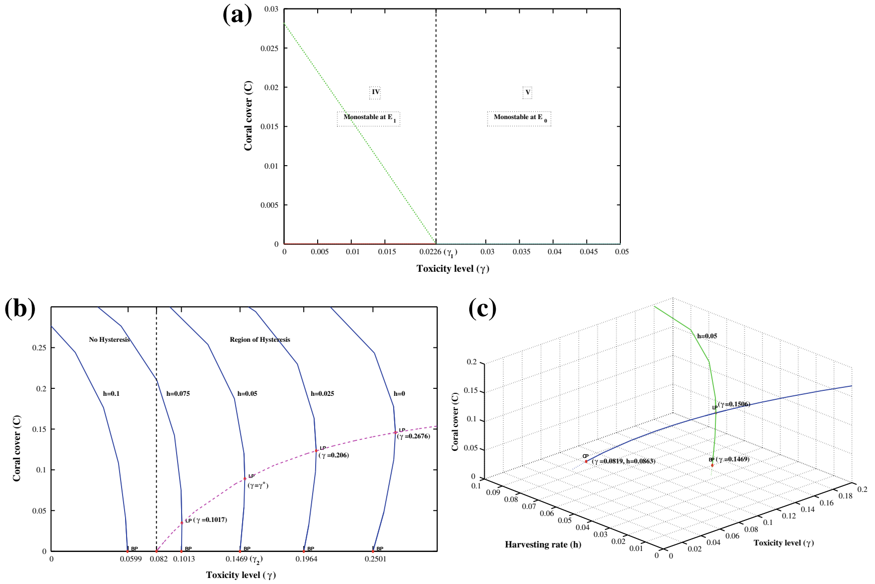


Fig. 12.4 **a** Bifurcation diagram of γ versus the equilibrium value of coral cover with $h = 0.5$ ($h > s$) and $a = 0.077$. Coral-dominated stable equilibria E_1 are indicated by green curves, seaweeds-dominated stable equilibria E_0 are indicated by cyan curves and unstable equilibria by red curves. **b** Bifurcation diagram of γ versus the equilibrium value of coral cover for different values of h. **c** Two-parameter bifurcation diagram with γ and h as active parameters

resistance component of resilience of coral-dominated regime is increased even with the increase of seaweed toxicity, measured by taking the difference of the values of γ at the saddle-node bifurcating point (LP) and at transcritical bifurcating point (BP) for a particular value of h. With γ and h as active parameters, the ecological resilience of the system becomes minimum when rate of harvesting of Parrotfish is greater than $h = 0.0863$ where the saddle-node curve meets the parameter axis at $\gamma = 0.082$, generating a cusp point (CP) at their point of intersection. Figure 12.4c gives a two-parameter bifurcation diagram with γ and h as active parameters, representing a cusp point at $(\gamma, h) = (0.082, 0.0863)$ on the saddle-node curve.

The impact of seaweed overgrowth rate on coral cover is given by the solutions of the nullcline equations projected onto the $C - \alpha$ plane (Fig. 12.5). The region VI represents monostability at E^* for $0 \leq \alpha < \alpha_* = \frac{r-d_2}{M_2} - (r + \gamma) = 0.0412$, representing the coexistence steady state for all nonnegative initial conditions. In this region, the system will ultimately arrive at a coral-dominated state corresponding to low levels of seaweeds in presence of Parrotfish. The bistable region is represented by the region VII for $\alpha_* < \alpha < \alpha^* = 0.043$. In this region, all the trajectories of the system will arrive at E^* or E_2 depending upon the initial conditions. Once the rate of seaweed overgrowth rate on corals surpasses the threshold α^*, the system arrives at a seaweeds-dominated and coral-depleted stable state in presence of Parrotfish,

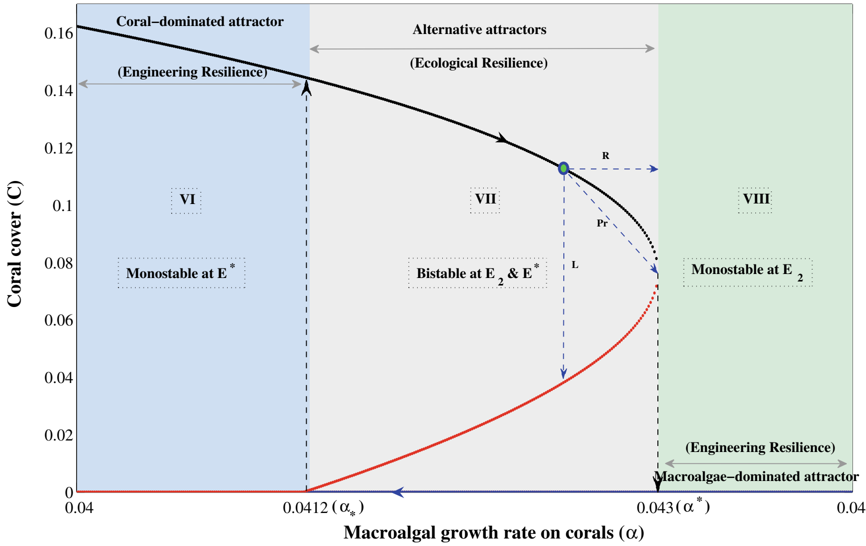


Fig. 12.5 Bifurcation diagram of α versus the equilibrium value of coral cover with $g = 0.43$ and $h < s$. Coral-dominated stable interior equilibria are indicated by black curves, seaweeds-dominated stable equilibria E_2 are indicated by blue curves and unstable equilibria by red curves

represented by region $VIII$ of monostability at E_2. Hysteresis will result, with low seaweed cover followed by an increase in the seaweed cover above a critical threshold α^*. A backward shift occurs only if the seaweed immigration rate is reduced far enough to reach the other bifurcation point α_*. With low seaweed growth rate on coral cover (viz. $\alpha = 0.01$), the system has a coral-dominated single attractor and the corresponding rate of recovery from small perturbations quantifies the engineering resilience of the system. From Fig. 12.6a it follows that the recovery time in the monostable region VI after an arbitrary perturbation increases due to the increase of seaweed growth rate on coral cover, and so, the engineering resilience of the system in the monostable region VI decreases due to the increase of seaweed growth rate. From Fig. 12.6b it follows that the ecological resilience of the system at the interior equilibrium in the bistable region VII is maximum when seaweed overgrowth rate on corals is less than α_* and decreases due to the increase of seaweed overgrowth on corals. The ecological resilience becomes minimum when seaweed overgrowth rate on corals approaches the threshold value α^*. In this case, slight increase in α leads to a catastrophic shift of regime to a seaweeds-dominated ecosystem in presence of Parrotfish.

From Fig. 12.6c it is observed that with low grazing rate of Parrotfish (viz. $g = 0.43$), the system is seaweeds-dominated and stable even with low seaweed overgrowth rate on corals. In this case, the system undergoes a sudden change in transition from coral-seaweeds coexistence steady state to coral-depleted steady state when α crosses α^*. The increase of grazing rate of Parrotfish increases the

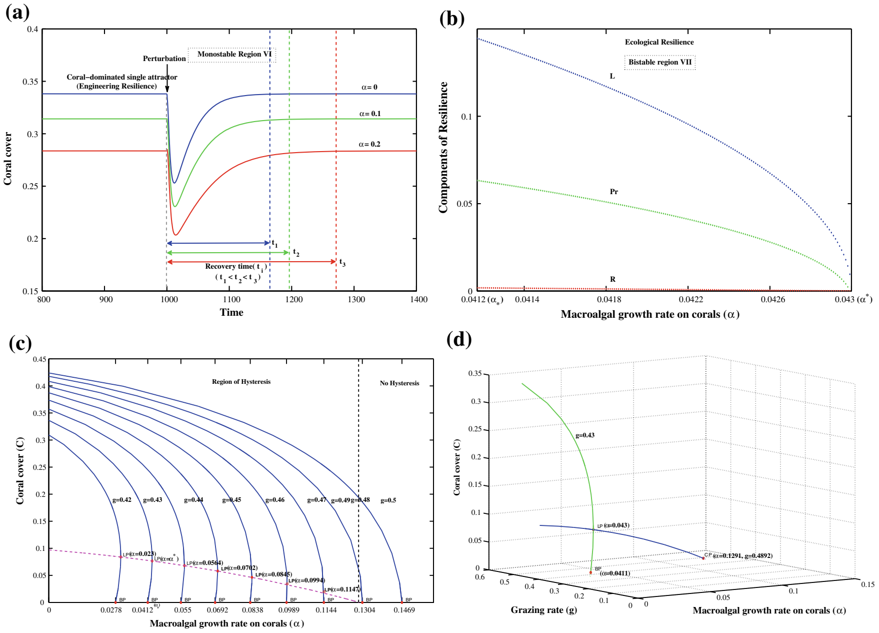


Fig. 12.6 **a** Change in the engineering resilience of the system in the monostable region VI with α as an active parameter. **b** Change in the ecological resilience of the system in the bistable region VII with α as an active parameter. **c** Bifurcation diagram of α versus the equilibrium value of coral cover for different values of g. **d** Two-parameter bifurcation diagram with α and g as active parameters

latitude component of ecological resilience of coral-dominated bistable regime VII due to the increase of coral cover. Also, the resistance component of resilience of coral-dominated regime is increased even with high seaweed overgrowth on corals, measured by taking the difference of the values of α at the saddle-node bifurcating point (LP) and at transcritical bifurcating point (BP) for a particular value of g (cf. Fig. 12.6b). With α and g as active parameters, the resilience of the system with high seaweed overgrowth on corals becomes maximum when the grazing intensity crosses the threshold $g = 0.4892$ where the saddle-node curve meets the parameter axis at $\alpha = 0.1291$, generating a cusp point (CP) at the point of intersection. Figure 12.6d gives a two-parameter bifurcation diagram with α and g as active parameters, representing the cusp point at $(\alpha, g) = (0.1291, 0.4892)$ on the saddle-node curve.

The grazing rate g depends on the abundance of Parrotfish and is thus subjected to variation with changes in available refuge and food abundance. To identify the impact of changes in grazing intensity on coral cover, in Fig. 12.7, we plot the solutions of the nullcline equations in the $C - g$ plane with $\gamma = 0.25$, yielding a bifurcation diagram. The region IX represents monostability at E_2 for $0 \leq g < g^* = 0.549$, representing seaweeds-dominated and coral-depleted state in presence of Parrotfish for all nonnegative initial conditions. The bistable region is represented by the region

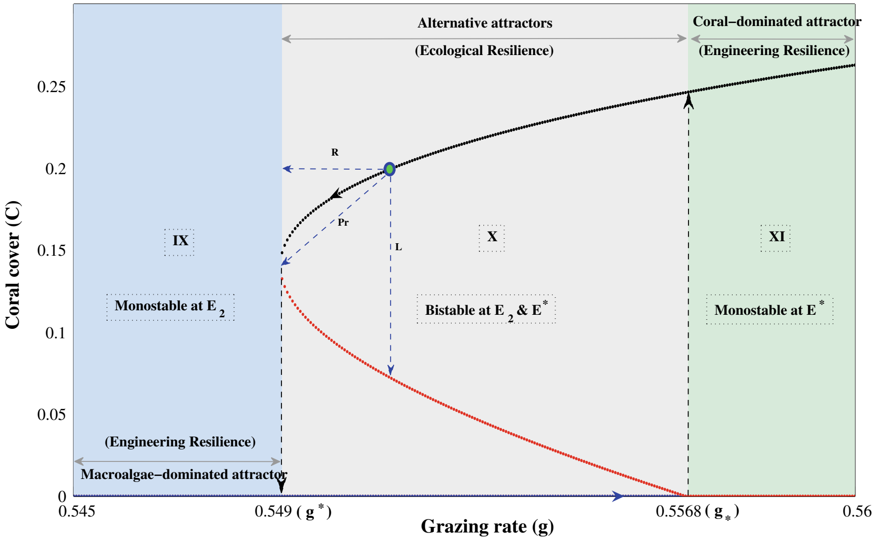


Fig. 12.7 Bifurcation diagram of g versus the equilibrium value of coral cover with $\gamma = 0.25$ and $h < s$. Coral-dominated stable interior equilibria are indicated by black curves, seaweeds-dominated stable equilibria E_2 are indicated by blue curves and unstable equilibria by red curves

X for $g^* < g < g_* = 0.5568$. In this region, all the trajectories of the system will arrive at E^* or E_2 depending upon the initial conditions. Once the grazing intensity surpasses the threshold $g = g_*$, the system arrives at the coexistence stable state, represented by region XI of monostability at E^*. From Fig. 12.8a it follows that the ecological resilience of the system at the interior equilibrium is maximum when grazing intensity exceeds $g = g_*$ and decreases in the bistable region X due to the decrease of grazing intensity. The ecological resilience becomes minimum when grazing intensity of Parrotfish approaches the threshold value g^*. In this case, slight decrease in g leads to a catastrophic shift of regime to a seaweeds-dominated ecosystem in presence of Parrotfish. With high rate of grazing by herbivores (viz. $g = 0.6$), the system stabilizes at coral-dominated single attractor and the corresponding rate of recovery from small perturbations gives the measure of engineering resilience of the system in the monostable region XI. From Fig. 12.8b it follows that the recovery time in the monostable region XI after an arbitrary perturbation decreases due to the increase of grazing intensity. Consequently, the engineering resilience of the system in the monostable region XI increase due to the increase of grazing intensity of Parrotfish.

From Fig. 12.8c it is observed that with low grazing rate of Parrotfish the system is seaweeds-dominated and stable even with low seaweed toxicity. The increase of seaweed toxicity decreases the latitude component of resilience of coral-dominated regime due to the decrease of coral cover. Also, the resistance component of resilience of coral-dominated regime is decreased even with high grazing rate of Parrotfish,

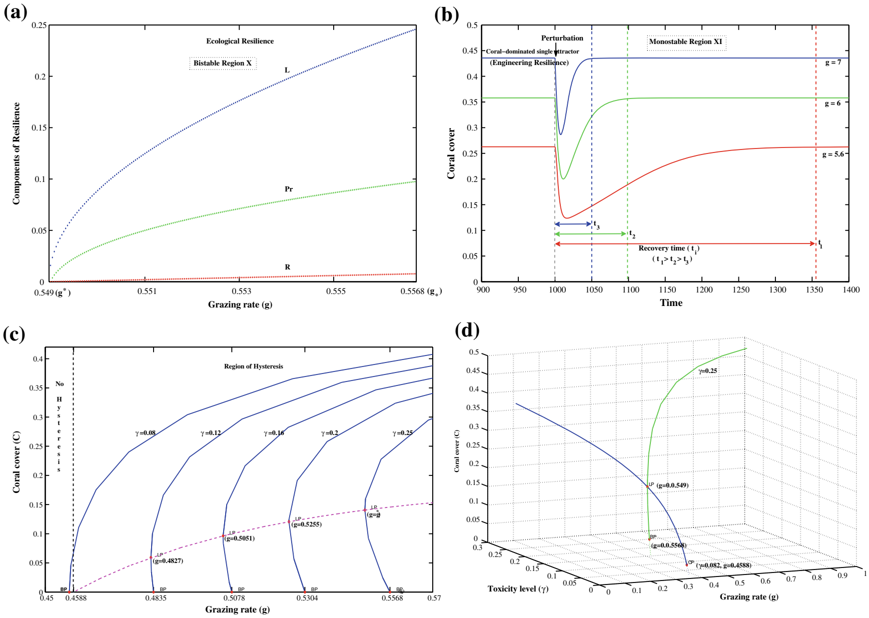


Fig. 12.8 **a** Change in the ecological resilience of the system in the bistable region X with g as an active parameter. **b** Change in the engineering resilience of the system in the monostable region XI with g as an active parameter. **c** Bifurcation diagram of g versus the equilibrium value of coral cover for different values of γ. **d** Two-parameter bifurcation diagram with g and γ as active parameters

measured by taking the difference of the values of g at the saddle-node bifurcating point (LP) and at transcritical bifurcating point (BP) for a particular value of γ. With g and γ as active parameters, the ecological resilience of the system becomes maximum when the seaweed toxicity is less than the threshold value $\gamma = 0.082$ where the saddle-node curve meets the parameter axis at $g = 0.4588$, generating a cusp point (CP) at the point of intersection. Figure 12.8d gives a two-parameter bifurcation diagram with g and γ as active parameters, representing the cusp point at $(g, \gamma) = (0.4588, 0.082)$ on the saddle-node curve.

To identify the effect of harvesting of Parrotfish on coral cover in presence of high seaweed toxicity, in Fig. 12.9, we plot the equilibrium values projected onto the $C - h$ plane with $\gamma = 0.22$. The system is monostable at E^* in region XII for $0 < h < h_* = 0.0138$. The bistable region is represented by the region $XIII$ for $h_* < h < h^* = 0.0191$. For $h > h^*$, the system becomes monostable at the seaweeds-dominated and coral-depleted steady state as depicted by the region XIV. Hysteresis occurs in the bistable region $XIII$ with low seaweed cover followed by an increase in the seaweed cover above the critical threshold $h = h^*$. A backward shift occurs only if the rate of harvesting of Parrotfish reduced far enough to reach the other bifurcation point $h = h_*$. With low rate of harvesting by herbivores (viz. $h = 0.01$), the system stabilizes at coral-dominated single attractor. In this case, the rate of

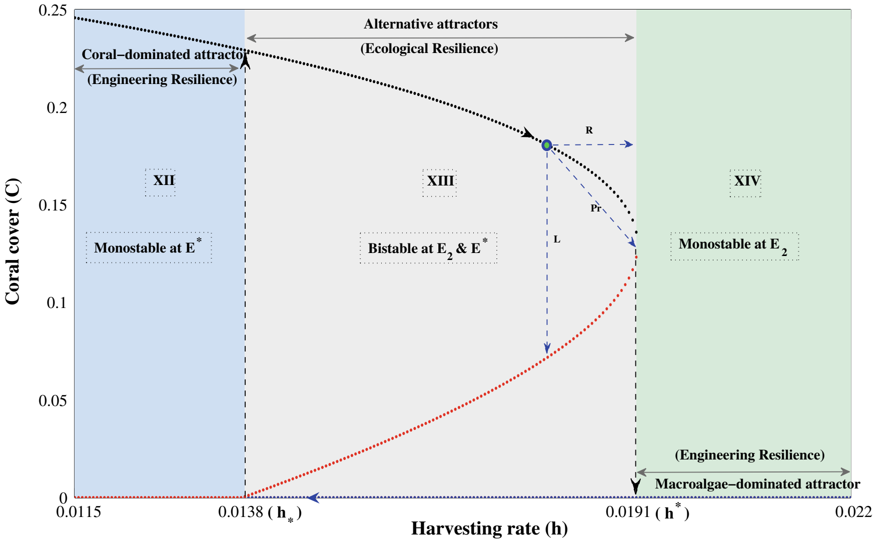


Fig. 12.9 **a** Bifurcation diagram of h versus the equilibrium value of coral cover with $\gamma = 0.22$. Coral-dominated stable interior equilibria are indicated by black curves, seaweeds-dominated stable equilibria E_2 are indicated by blue curves and unstable equilibria by red curves

recovery from small perturbations quantifies engineering resilience of the system in the monostable region XII. From Fig. 12.10a it follows that the recovery time in the monostable region XII after an arbitrary perturbation is least in absence of harvesting of Parrotfish and increases due to the increase of harvesting. Consequently, the engineering resilience of the system in the monostable region XII decreases due to the increase of harvesting of Parrotfish. From Fig. 12.10b it follows that the ecological resilience of the system at the interior equilibrium is minimum when the harvesting rate of Parrotfish exceeds $h = h^*$ and increases in the bistable region $XIII$ due to the decrease of harvesting rate. The resilience becomes maximum when rate of harvesting is lowered below $h = h_*$.

From Fig. 12.10c it is observed that with high seaweed toxicity (viz. $\gamma = 0.22$), the system is seaweeds-dominated and stable even with low harvesting rate of Parrotfish followed by a sudden change in transition from coral-seaweeds coexistence steady state to coral-depleted steady state. The decrease of seaweed toxicity increases the latitude component of resilience of coral-dominated regime due to the increase in coral cover. Also, the resistance component of resilience of coral-dominated regime is increased even with the increase in harvesting of Parrotfish, measured by taking the difference of the values of h at the saddle-node bifurcating point (LP) and at transcritical bifurcating point (BP) for a particular value of γ. With h and γ as active parameters, the resilience of the system becomes maximum when the seaweed toxicity level is less than $\gamma = 0.0854$ where the saddle-node curve meets the parameter axis at $\gamma = 0.0736$, generating a cusp point (CP) at their point of inter-

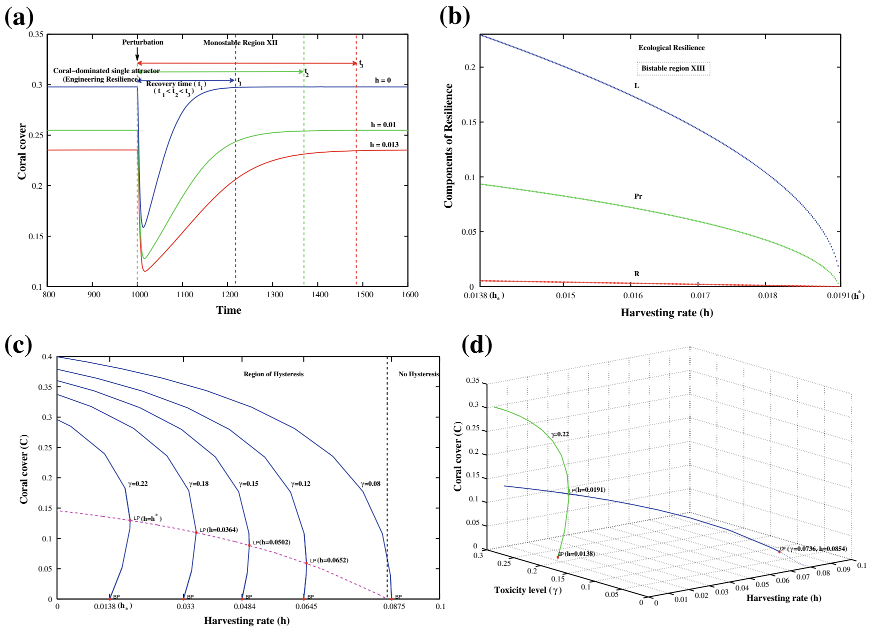


Fig. 12.10 **a** Change in the engineering resilience of the system in the monostable region XII with h as an active parameter. **b** Change in the ecological resilience of the system in the bistable region $XIII$ with h as an active parameter. **c** Bifurcation diagram of h versus the equilibrium value of coral cover for different values of γ. **d** Two-parameter bifurcation diagram with h and γ as active parameters

section. Figure 12.10d gives a two-parameter bifurcation diagram with h and γ as active parameters, representing the cusp point at $(h, \gamma) = (0.0854, 0.0736)$ on the saddle-node curve.

The effect of high seaweed toxicity and colonization of seaweeds on algal turf is shown in Fig. 12.11a, where we plot the solutions of the nullcline equations projected onto the $C - b$ plane with $\gamma = 0.2$. The region XV represents monostability at E^* for $0 \leq b < b_* = 0.0343$, representing coral-seaweeds coexistence steady state for all nonnegative initial conditions. In this region, the system will ultimately arrive at a coral-dominated state corresponding to low levels of seaweeds. The bistable region is represented by the region XVI for $b_* < b < b^* = 0.0394$. Once the rate of seaweed immigration surpasses the threshold $b = b^*$, the system arrives at a seaweeds-dominated and coral-depleted stable state in presence of Parrotfish, represented by region $XVII$ of monostability at E_2. Hysteresis will result, with low seaweed cover followed by an increase in the seaweed cover above a critical threshold $b = b^*$. A backward shift occurs only if the seaweed immigration rate is reduced far enough to reach the other bifurcation point $b = b_*$. From Fig. 12.11b it follows that the ecological resilience of the system at the interior equilibrium is minimum when the seaweed immigration rate exceeds $b = b^*$ and increases in the bistable region XVI due to

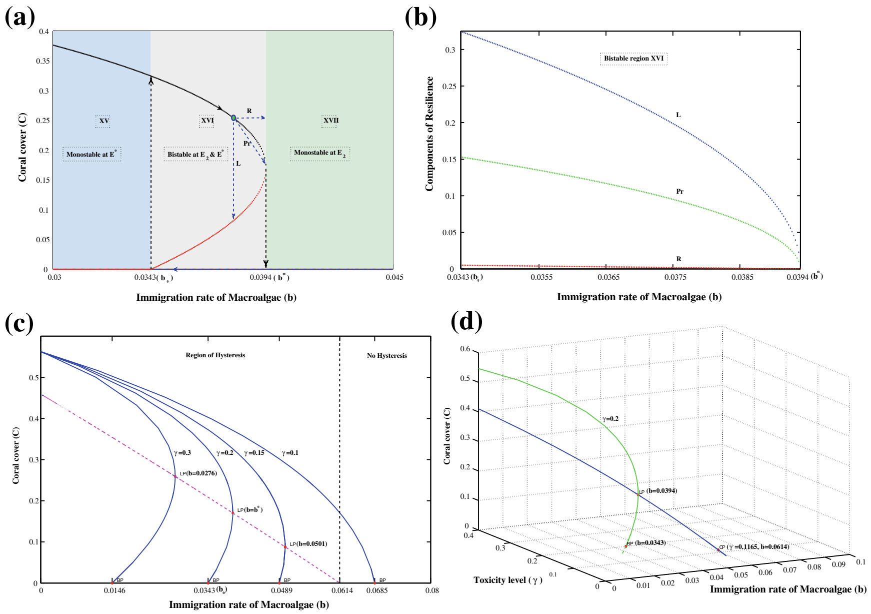


Fig. 12.11 **a** Bifurcation diagram of b versus the equilibrium value of coral cover with $\gamma = 0.2$. Coral-dominated stable interior equilibria are indicated by black curves, seaweeds-dominated stable equilibria E_2 are indicated by blue curves, and unstable equilibria by red curves. **b** Change in the ecological resilience of the system with b as an active parameter. **c** Bifurcation diagram of b versus the equilibrium value of coral cover for different values of γ. **d** Two-parameter bifurcation diagram with b, and γ as active parameters

the decrease of immigration rate. The resilience becomes maximum when rate of immigration is lowered below $b = b_*$.

From Fig. 12.11c it is observed that with high seaweed toxicity (viz. $\gamma = 0.3$), the system is seaweeds-dominated and stable even with low colonization rate of seaweeds on algal turf followed by a sudden change in transition from coral-seaweeds coexistence steady state to coral-depleted steady state. The decrease of seaweed toxicity increases the resilience of coral-dominated regime even with the increase in colonization rate of seaweeds, measured by taking the difference of the values of b at the saddle-node bifurcating point (LP) and at transcritical bifurcating point (BP) for a particular value of γ. With b and γ as active parameters, the ecological resilience of the system becomes maximum when the seaweed toxicity level is less than $b = 0.0614$ where the saddle-node curve meets the parameter axis at $\gamma = 0.1165$, generating a cusp point (CP) at their point of intersection. Figure 12.11d gives a two-parameter bifurcation diagram with b and γ as active parameters, representing the cusp point at $(b, \gamma) = (0.0614, 0.1165)$ on the saddle-node curve.

12.4 Seasonally Forced System

Coral reef ecosystems are subject to increasing environmental fluctuations. Seasonality, which is a kind of periodic fluctuation varying with changing seasons, is proposed to be considered in our model to describe more realistic relationships between seaweeds, corals, and reef-herbivores. As observed by the researchers [36, 37], seaweeds are highly seasonal in their occurrence, growth, and reproduction. To include the seasonal influence on bioactivity of seaweeds in coral reefs, we consider the seasonal forcing as a sinusoidal function of relevant parameters of our model. Considering the growth and immigration rate of seaweeds as periodically varying function of time due to seasonal variations, we adopt $\alpha(t) = \alpha(1 + \epsilon_1 \sin(\omega t)), a(t) = a(1 + \epsilon_2 \sin(\omega t))$, and $b(t) = b(1 + \epsilon_3 \sin(\omega t))$, where $\alpha\epsilon_1$, $a\epsilon_2$, and $b\epsilon_3$ are the amplitudes and ω is the frequency of sinusoidal perturbations in α, a, and b, respectively. Also, considering the seasonal variations in seaweed toxicity, we adopt $\gamma(t) = \gamma(1 + \epsilon_4 \sin(\omega t))$, where $\gamma\epsilon_4$ is the amplitude of sinusoidal perturbations in γ. Variability in grazing pressure has also been related to seasonal changes in the abundance and productivity of seaweeds. We, therefore, consider the maximal grazing rate g and maximal carrying capacity k of Parrotfish as periodically varying function of time due to seasonal variations by adopting $g(t) = g(1 + \epsilon_5 \sin(\omega t))$ and $k(t) = k(1 + \epsilon_6 \sin(\omega t))$. Seasonally varying growth rate of corals has been reported by many researchers [38]. Since coral distribution is negatively associated with seaweed abundance, we choose $r(t) = r(1 + \epsilon_7 \sin(\omega t + \phi))$, where the parameter ϕ $(0 \leq \phi \leq 2\pi)$, can be interpreted as a difference in phase angle between the seasonality in the growth rates of corals and seaweeds. Since the parameters are necessarily positive, we have $0 \leq \epsilon_i \leq 1$ $(i = 1, \ldots, 7)$.

Considering the seasonally varying parameters of the system (12.2), we propose a nonautonomous system as follows:

$$
\begin{aligned}
\frac{dM}{dt} &= M\left\{\alpha(t)C - \frac{g(t)P}{k(t)(1-C)} - d_1\right\} + (a(t)M + b(t))(1 - M - C)\\
\frac{dC}{dt} &= C\left\{r(t)(1 - M - C) - (\alpha(t) + \gamma(t))M - d_2\right\}\\
\frac{dP}{dt} &= P\left[s\left\{1 - \frac{P}{k(t)(1-C)}\right\} - h\right]
\end{aligned}
\tag{12.3}
$$

where $\alpha(t), \gamma(t), a(t), b(t), g(t), k(t), r(t)$ are all positive ω-periodic functions in $[0, \infty)$; d_1, d_2, s, h are time-independent positive parameters.

In order to study the existence of a unique positive almost periodic solution for the system (12.3), we will establish sufficient conditions based on Gaines and Mawhin's [39] coincidence degree theory. We will summarize some basic results form [39] that will be important for this section. Let X and Z be real Banach spaces, $L : DomL \subset X \to Z$ be a linear mapping, and $N : X \to X$ be a continuous mapping. The mapping L is a Fredholm mapping of index zero if $dimKerL = codimImL < \infty$ and ImL is closed in Z. If L is a Fredholm mapping of index zero and there exist continuous projections $P : X \rightarrow X$ and $Q : Z \rightarrow Z$ such that $ImP = KerL$ and $KerQ = ImL = Im(I - Q)$, it follows that $L_{|DomL \cap KerP} : (I - P)X \rightarrow ImL$ is invertible. Let K_P be its inverse mapping. If Ω is an open bounded subset of X, the mapping N is L-compact on $\bar{\Omega}$ if $QN(\bar{\Omega})$ is bounded and $K_P(I - Q)N : P \rightarrow X$ is compact. Since ImQ is isomorphic to $KerL$, there is an isomorphism $J : ImQ \rightarrow KerL$.

Lemma 12.4.1 (Mawhin's continuation theorem [39]) *Let L be a Fredholm mapping of index zero and let N be L-compact on $\bar{\Omega}$. Suppose that*
(i) $Lx \neq \lambda Nx$ for any $x \in \partial\Omega$ and $\lambda \in (0, 1)$;
(ii) $QNx \neq 0$ for any $x \in \partial\Omega \cap KerL$;
(iii) $deg\{JQN, \Omega \cap KerL, 0\} \neq 0$.
Then the operator equation $Lx = Nx$ has at least one solution in $DomL \cap \bar{\Omega}$.

Supposing that $f(t), t \in [0, \infty)$ is a continuous function with period ω, we denote

$$f^L = \min_{t\in[0,\omega]}\{f(t)\},\ f^M = \max_{t\in[0,\omega]}\{f(t)\},\ \bar{f} = \frac{1}{\omega}\int_0^{\omega} f(t)dt$$

Following the boundedness of the system (12.2) we have $M(t) + C(t) + P(t) < 1 + k^M$ and so the system (12.3) is also bounded.

Lemma 12.4.2 *The system (12.3) has at least one ω-periodic solution if $\bar{a} > \bar{b} + \bar{\alpha} - d_1$, $\bar{r} > d_2$ and $s > h$ hold.*

Proof Let us consider the following system:

$$\begin{aligned}
\frac{du_1(t)}{dt} &= \alpha(t)e^{u_2(t)} - \frac{g(t)e^{u_3(t)}}{k(t)\left(1 - e^{u_2(t)}\right)} + \left(a(t) + b(t)e^{-u_1(t)}\right)\left(1 - e^{u_1(t)} - e^{u_2(t)}\right) - d_1 \\
\frac{du_2(t)}{dt} &= r(t)\left(1 - e^{u_1(t)} - e^{u_2(t)}\right) - (\alpha(t) + \gamma(t))\,e^{u_1(t)} - d_2 \\
\frac{du_3(t)}{dt} &= s\left\{1 - \frac{e^{u_3(t)}}{k(t)\left(1 - e^{u_2(t)}\right)}\right\} - h
\end{aligned} \tag{12.4}$$

where all functions are defined as ones in system (12.3). It is easy to see that if system (12.4) has one ω-periodic solution $\left(u_1^*(t), u_2^*(t), u_3^*(t)\right)^T$, then $(M^*(t), C^*(t), P^*(t))^T = \left(e^{u_1^*(t)}, e^{u_2^*(t)}, e^{u_3^*(t)}\right)^T$ is a positive ω-periodic solution of system (12.3). Therefore, to complete the proof it suffices to show that system (12.4) has a ω-periodic solution.

Since $0 < M^*, C^* < 1$, we must have $u_1^*, u_2^* < 0$.

Taking $X{=}Y = \left((u_1(t), u_2(t), u_3(t))^T \in C(R, R^3) : u_i(t + \omega) = u_i(t), t \in R,\right.$ $i = 1, 2, 3)$, we define

$$\parallel (u_1(t), u_2(t), u_3(t))^T \parallel = \Sigma_{i=1}^{3} \max_{t\in[0,\omega]} |u_i(t)|,$$

where $|.|$ denotes the Euclidian norm. Then X and Y both are Banach spaces when they are endowed with the norm $\|.\|$.

Let $L : DomL \cap X$, $L\,(u_1(t), u_2(t), u_3(t))^T = \left(\frac{du_1(t)}{dt}, \frac{du_2(t)}{dt}, \frac{du_3(t)}{dt}\right)^T$, where $DomL = \left\{(u_1(t), u_2(t), u_3(t))^T \in C^1(R, R^3)\right\}$, $N : X \to X$ and

$$N\begin{pmatrix} u_1(t) \\ u_2(t) \\ u_3(t) \end{pmatrix} = \begin{pmatrix} \alpha(t)e^{u_2(t)} + \frac{g(t)e^{u_3(t)}}{k(t)\left(e^{u_2(t)}-1\right)} - d_1 + \left(a(t) + b(t)e^{-u_1(t)}\right)\left(1 - e^{u_1(t)} - e^{u_2(t)}\right) \\ r(t)\left(1 - e^{u_1(t)} - e^{u_2(t)}\right) - \left(\alpha(t) + \gamma(t)\right)e^{u_1(t)} - d_2 \\ s\left\{1 + \frac{e^{u_3(t)}}{k(t)\left(e^{u_2(t)}-1\right)}\right\} - h \end{pmatrix}$$

With these notations system (12.4) can be written in the form $Lu = Nu$, $u \in X$.

Now define two projectors $P : X \to X$ and $Q : Y \to Y$ as

$$P\begin{pmatrix} u_1(t) \\ u_2(t) \\ u_3(t) \end{pmatrix} = Q\begin{pmatrix} u_1(t) \\ u_2(t) \\ u_3(t) \end{pmatrix} = \begin{pmatrix} \frac{1}{\omega}\int_0^{\omega} u_1(t)dt \\ \frac{1}{\omega}\int_0^{\omega} u_2(t)dt \\ \frac{1}{\omega}\int_0^{\omega} u_3(t)dt \end{pmatrix}, \begin{pmatrix} u_1(t) \\ u_2(t) \\ u_3(t) \end{pmatrix} \in X = Y.$$

Then P and Q are continuous projectors such that $ImP = KerL$, $KerQ = ImL = Im(I - Q)$.

Obviously, we have $KerL = R^3$, $ImL = \left((u_1, u_2, u_3)^T \in Y : \int_0^{\omega} u_i(t)dt = 0, i = 1, 2, 3\right)$ is closed in Y, and $dimKerL = codimImL = 3$. Therefore L is a Fredholm mapping of index zero.

Furthermore, the generalized inverse (to L) $K_P : ImL \to DomL \cap KerP$ exists and is given by

$$K_P\begin{pmatrix} u_1(t) \\ u_2(t) \\ u_3(t) \end{pmatrix} = \begin{pmatrix} \int_0^t u_1(s)ds - \frac{1}{\omega}\int_0^{\omega}\int_0^t u_1(s)dsdt \\ \int_0^t u_2(s)ds - \frac{1}{\omega}\int_0^{\omega}\int_0^t u_2(s)dsdt \\ \int_0^t u_3(s)ds - \frac{1}{\omega}\int_0^{\omega}\int_0^t u_3(s)dsdt \end{pmatrix}$$

Accordingly, $QN : X \to Y$ and $K_P(I - Q)N : X \to X$ lead

$$(QN)u = \begin{pmatrix} \frac{1}{\omega}\int_0^{\omega}\left[\alpha(t)e^{u_2(t)} + \frac{g(t)e^{u_3(t)}}{k(t)\left(e^{u_2(t)}-1\right)} - d_1 + \left(a(t) + \frac{b(t)}{e^{u_1(t)}}\right)\left(1 - e^{u_1(t)} - e^{u_2(t)}\right)\right]dt \\ \frac{1}{\omega}\int_0^{\omega}\left[r(t)\left(1 - e^{u_1(t)} - e^{u_2(t)}\right) - \left(\alpha(t) + \gamma(t)\right)e^{u_1(t)} - d_2\right]dt \\ \frac{1}{\omega}\int_0^{\omega}\left[s\left\{1 + \frac{e^{u_3(t)}}{k(t)\left(e^{u_2(t)}-1\right)}\right\} - h\right]dt \end{pmatrix}$$

and

$$K_P(I - Q)Nu = \begin{pmatrix} \int_0^t\left[\alpha(s)e^{u_2(s)} + \frac{g(s)e^{u_3(s)}}{k(s)\left(e^{u_2(s)}-1\right)} - d_1 + \left(a(s) + \frac{b(s)}{e^{u_1(s)}}\right)\left(1 - e^{u_1(s)} - e^{u_2(s)}\right)\right]ds \\ \int_0^{\omega}\left[r(s)\left(1 - e^{u_1(s)} - e^{u_2(s)}\right) - \left(\alpha(s) + \gamma(s)\right)e^{u_1(s)} - d_2\right]ds \\ \int_0^{\omega}\left[s\left\{1 + \frac{e^{u_3(s)}}{k(s)\left(e^{u_2(s)}-1\right)}\right\} - h\right]ds \end{pmatrix}$$

$$-\begin{pmatrix} \frac{1}{\omega}\int_0^\omega\int_0^t \left[\alpha(s)e^{u_2(s)} + \frac{g(s)e^{u_3(s)}}{k(s)\left(e^{u_2(s)}-1\right)} - d_1 + \left(a(s)+\frac{b(s)}{e^{u_1(s)}}\right)\left(1-e^{u_1(s)}-e^{u_2(s)}\right)\right] dsdt \\ \frac{1}{\omega}\int_0^\omega\int_0^t \left[r(s)\left(1-e^{u_1(s)}-e^{u_2(s)}\right) - \left(\alpha(s)+\gamma(s)\right)e^{u_1(s)} - d_2\right] dsdt \\ \frac{1}{\omega}\int_0^\omega\int_0^t \left[s\left\{1+\frac{e^{u_3(s)}}{k(s)\left(e^{u_2(s)}-1\right)}\right\} - h\right] dsdt \end{pmatrix}$$

$$-\begin{pmatrix} \left(\frac{t}{\omega}-\frac{1}{2}\right)\int_0^\omega \left[\alpha(s)e^{u_2(s)} + \frac{g(s)e^{u_3(s)}}{k(s)\left(e^{u_2(s)}-1\right)} - d_1 + \left(a(s)+\frac{b(s)}{e^{u_1(s)}}\right)\left(1-e^{u_1(s)}-e^{u_2(s)}\right)\right] ds \\ \left(\frac{t}{\omega}-\frac{1}{2}\right)\int_0^\omega \left[r(s)\left(1-e^{u_1(s)}-e^{u_2(s)}\right) - \left(\alpha(s)+\gamma(s)\right)e^{u_1(s)} - d_2\right] ds \\ \left(\frac{t}{\omega}-\frac{1}{2}\right)\int_0^\omega \left[s\left\{1+\frac{e^{u_3(s)}}{k(s)\left(e^{u_2(s)}-1\right)}\right\} - h\right] ds \end{pmatrix}$$

Clearly, QN and $K_P(I-Q)N$ are continuous by the Lebesgue theorem and, furthermore, by the Arzela–Ascoli theorem, it follows that $QN(\bar{\Omega})$ and $\overline{K_P(I-Q)N(\bar{\Omega})}$ are relatively compact for any open bounded set $\Omega \subset X$. Hence N is L-compact on $\bar{\Omega}$ for any open bounded set $\Omega \subset X$. Corresponding to operator equation $Lu = \lambda Nu$, $\lambda \in (0, 1)$, we have

$$\begin{aligned} \frac{du_1(t)}{dt} &= \lambda\left[\alpha(t)e^{u_2(t)} + \frac{g(t)e^{u_3(t)}}{k(t)\left(e^{u_2(t)}-1\right)} - d_1 + \left(a(t)+b(t)e^{-u_1(t)}\right)\left(1-e^{u_1(t)}-e^{u_2(t)}\right)\right] \\ \frac{du_2(t)}{dt} &= \lambda\left[r(t)\left(1-e^{u_1(t)}-e^{u_2(t)}\right) - \left(\alpha(t)+\gamma(t)\right)e^{u_1(t)} - d_2\right] \\ \frac{du_3(t)}{dt} &= \lambda\left[s\left\{1+\frac{e^{u_3(t)}}{k(t)\left(e^{u_2(t)}-1\right)}\right\} - h\right] \end{aligned} \tag{12.5}$$

Suppose that $(u_1(t), u_2(t), u_3(t))^T \in X$ is a solution of (12.5) for a certain $\lambda \in (0, 1)$. By integrating (12.5) over the interval $[0, \omega]$ we obtain

$$\frac{1}{\omega}\int_0^\omega \left[\left(\alpha(t)-a(t)\right)e^{u_2} - a(t)e^{u_1} - \frac{g(t)e^{u_3}}{k(t)\left(1-e^{u_2}\right)} + \frac{b(t)\left(1-e^{u_2}\right)}{e^{u_1}}\right]dt = \bar{a} - \bar{b} + d_1 \tag{12.6}$$

$$\frac{1}{\omega}\int_0^\omega \left[r(t)\left(e^{u_1(t)}+e^{u_2(t)}\right) + \left(\alpha(t)+\gamma(t)\right)e^{u_1(t)}\right]dt = \bar{r} - d_2 \tag{12.7}$$

$$\frac{1}{\omega}\int_0^\omega \frac{e^{u_3(t)}}{k(t)\left(1-e^{u_2(t)}\right)}dt = s - h \tag{12.8}$$

From (12.5) to (12.8) we obtain

$$\int_0^\omega \left|\frac{u_1}{dt}\right| dt < 2\omega(\bar{a}+\bar{b}+d_1), \int_0^\omega \left|\frac{u_2}{dt}\right| dt < 2\omega(\bar{r}+d_2), \int_0^\omega \left|\frac{u_3}{dt}\right| dt < 2\omega(s+h) \tag{12.9}$$

Since $(u_1(t), u_2(t), u_3(t))^T \in X$, there exists $\xi_i, \eta_i \in [0, \omega]$, $(i = 1, 2, 3)$ such that

$$u_i\left(\xi_i\right) = \min_{t\in[0,\omega]} u_i(t), u_i\left(\eta_i\right) = \max_{t\in[0,\omega]} u_i(t), (i = 1, 2, 3)$$

From (12.7) we get

$$\omega(\bar{r}-d_2) \geq \int_0^{\omega} (\alpha(t)+\gamma(t))\, e^{u_1(t)}dt \geq \omega(\bar{\alpha}+\bar{\beta})e^{u_1(\xi_1)} \Rightarrow u_1\,(\xi_1) \leq \ln\left(\frac{\bar{r}-d_2}{\bar{\alpha}+\bar{\gamma}}\right) = H_{11}$$

and

$$\omega(\bar{r}-d_2) \geq \int_0^{\omega} r(t)e^{u_2(t)}dt \geq \omega\bar{r}e^{u_2(\xi_2)} \Rightarrow u_2\,(\xi_2) \leq \ln\left(\frac{\bar{r}-d_2}{\bar{r}}\right) = H_{21},$$

where $\bar{r} > d_2$. Then, we have,
$u_1(t) \leq u_1\,(\xi_1) + \int_0^{\omega} \left|\frac{du_1(t)}{dt}\right| dt \leq H_{11} + 2\omega(\bar{a}+\bar{b}+d_1)$ and
$u_2(t) \leq u_2\,(\xi_2) + \int_0^{\omega} \left|\frac{du_2(t)}{dt}\right| dt \leq H_{21} + 2\omega(\bar{r}+d_2)$.

From (12.6) we get

$$\omega(\bar{a}-\bar{b}+d_1) \leq \bar{\alpha}\omega e^{u_2(\eta_2)} + \bar{b}\omega e^{-u_1(\xi_1)}$$

$$\Rightarrow u_2\,(\eta_2) \geq \ln\left\{\frac{\bar{a}+d_1-\bar{b}\left(1+e^{-u_1(\xi_1)}\right)}{\bar{\alpha}}\right\} = H_{22},$$

where $u_1\,(\xi_1) \geq \ln\left(\frac{\bar{b}}{\bar{a}-\bar{b}+d_1}\right)$ and $\bar{b} < \bar{a}+d_1$.

Therefore, $u_2(t) \geq u_2\,(\eta_2) - \int_0^{\omega} \left|\frac{du_2(t)}{dt}\right| dt \geq H_{22} - 2\omega(\bar{r}+d_2)$ and so

$$\max_{t\in[0,\omega]} |u_2(t)| \leq \max\{|H_{21}+2\omega(\bar{r}+d_2)|, |H_{22}-2\omega(\bar{r}+d_2)|\} = B_2$$

Again, from (12.6) we get

$$\omega(\bar{a}-\bar{b}+d_1) \leq \bar{\alpha}\omega e^{u_2(\eta_2)} + \bar{a}\omega e^{u_1(\eta_1)} + \bar{b}\omega e^{-u_1(\xi_1)}$$

$$\Rightarrow u_1\,(\eta_1) \geq \ln\left\{\frac{\bar{a}+d_1-\bar{\alpha}-\bar{b}\left(1+e^{-u_1(\xi_1)}\right)}{\bar{a}}\right\} = H_{12},$$

where $u_1\,(\xi_1) \geq \ln\left(\frac{\bar{b}}{\bar{a}-\bar{\alpha}-\bar{b}+d_1}\right)$ and $\bar{b}+\bar{\alpha} < \bar{a}+d_1$.

Therefore, $u_1(t) \geq u_1\,(\eta_1) - \int_0^{\omega} \left|\frac{du_1(t)}{dt}\right| dt \geq H_{12} - 2\omega(\bar{a}+\bar{b}+d_1)$ and so

$$\max_{t\in[0,\omega]} |u_1(t)| \leq \max\left\{|H_{11}+2\omega(\bar{a}+\bar{b}+d_1)|, |H_{12}-2\omega(\bar{a}+\bar{b}+d_1)|\right\} = B_1$$

From (12.8) we get

$$\omega(s-h) = \int_0^{\omega} \frac{e^{u_3(t)}}{k(t)\left(1-e^{u_2(t)}\right)}dt \geq \frac{\omega e^{u_3(\xi_3)}}{\bar{k}} \Rightarrow u_3\,(\xi_3) \leq \ln\{\bar{k}(s-h)\} = H_{31},$$

where $s > h$ so that $u_3(\xi_3) \leq \ln\{\bar{k}(s-h)\} + \int_0^\omega \left|\frac{du_3(t)}{dt}\right| dt \leq H_{31} + 2\omega(s+h)$.

Also, (12.8) gives

$$\omega(s-h) \leq \int_0^\omega \frac{e^{u_3(t)}}{k(t)\left(1-e^{u_2(\eta_2)}\right)} dt \leq \frac{\omega e^{u_3(\eta_3)}}{\bar{k}\left(1-e^{u_2(\eta_2)}\right)}$$

$$\Rightarrow u_3(\eta_3) \geq \ln\left\{\bar{k}(s-h)\left(1-e^{u_2(\eta_2)}\right)\right\} = H_{32},$$

where $u_2(\eta_2) < 0$ and so

$$\max_{t\in[0,\omega]} |u_3(t)| \leq \max\{|H_{31} + 2\omega(s+h)|, |H_{32} - 2\omega(s+h)|\} = B_3$$

Clearly, H_{ij} and B_i are independent of λ for $i = 1, 2, 3$ and $j = 1, 2$.

Denote $\tilde{B} = \sum_{i=1}^3 B_i + B_0$, where B_0 is chosen sufficiently large so that each solution $\left(v_1^*, v_2^*, v_3^*\right)^T$ with $v_i^* > 0$ $(i = 1, 2, 3)$ of the system of algebraic equations

$$\begin{aligned} \frac{\bar{g}v_3}{\bar{k}(v_2-1)} + \bar{\alpha}v_2 &= \bar{a}(v_1+v_2-1) + \frac{\bar{b}}{v_1}(v_2-1) + \bar{b} + d_1 \\ (\bar{r}+\bar{\alpha}+\bar{\gamma})v_1 + \bar{r}v_2 &= \bar{r} - d_2 \\ sv_3 &= \bar{k}(s-h)(1-v_2) \end{aligned} \tag{12.10}$$

satisfies $\| \left(\ln(v_1^*), \ln(v_2^*), \ln(v_3^*)\right)^T \| = \Sigma_{i=1}^3 |\ln(v_i^*)| < \tilde{B}$, provided that the system (12.10) has solutions.

Now, we take $\Omega = \left\{(u_1, u_2, u_3)^T \in X : \|(u_1, u_2, u_3)^T\| < \tilde{B}\right\}$. Thus, condition (i) of Lemma 4.1 is satisfied. When $(u_1, u_2, u_3)^T \in \partial\Omega \cap KerL = \partial\Omega \cap R^3$, $(u_1, u_2, u_3)^T$ is a constant vector in R^3 with $|u_1| + |u_2| + |u_3| = \tilde{B}$. If system (12.10) has at least one solution, then we have

$$QN\begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} \frac{\bar{g}e^{u_3}}{\bar{k}(e^{u_2}-1)} - \bar{a}\left(e^{u_1}+e^{u_2}-1\right) - \bar{b}e^{-u_1}\left(e^{u_2}-1\right) - \bar{b} - d_1 + \bar{\alpha}e^{u_2} \\ \bar{r}\left(e^{u_1}+e^{u_2}\right) + (\bar{\alpha}+\bar{\gamma})e^{u_1} - \bar{r} + d_2 \\ \frac{e^{u_3}}{\bar{k}(e^{u_2}-1)} - \frac{h}{s} + 1 \end{pmatrix}$$

If system (12.9) does not have a solution, then we can directly derive

$$QN\begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

Thus, condition (ii) in Lemma 4.1 is satisfied.

In order to compute the Brouwer degree, let us consider the homotopy $H_\mu(u) = \mu QN(u) + (1-\mu)G(u)$ for $\mu \in [0, 1]$, where

$$G\begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} \bar{\alpha} e^{u_2} - \bar{b} - d_1 \\ (\bar{\alpha} + \bar{\gamma})\, e^{u_1} - \bar{r} \\ \frac{e^{u_3}}{\bar{k}(e^{u_2}-1)} - 1 \end{pmatrix}$$

Then we have $0 \notin H_\mu(\partial\Omega \cap KerL)$ for $\mu \in [0, 1]$. Moreover, one can easily show that the algebraic equation $G(u) = 0$ has a unique solution $(u_1^*, u_2^*, u_3^*) = \left(\ln\left(\frac{\bar{r}}{\bar{\alpha}+\bar{\gamma}}\right), \ln\left(\frac{\bar{b}+d_1}{\bar{\alpha}}\right), \ln\left(\frac{\bar{b}+d_1-\bar{\alpha}}{\bar{\alpha}/\bar{k}}\right)\right) \in R^3$.

By the invariance property of homotopy, direct calculation produces

$$deg(G, \Omega \cap KerL, 0) = sng_{(u_1^*,u_2^*,u_3^*) \in QN^{-1}\{0\}}[detDG(u)]$$

$$= sng \begin{vmatrix} 0 & \bar{\alpha} e^{u_2} & 0 \\ (\bar{\alpha} + \bar{\gamma})\, e^{u_1} & 0 & 0 \\ 0 & \frac{e^{u_2+u_3}}{\bar{k}(e^{u_2}-1)^2} & \frac{e^{u_3}}{\bar{k}(e^{u_2}-1)} \end{vmatrix} = -1 \neq 0,$$

where $DG(u)$ is the Jacobian matrix of G in u.

Thus system (12.5) has at least one ω-periodic solution. Then the condition (iii) of Lemma (4.1) holds, as a consequence, the system (12.4) has at least one positive ω-periodic solution. This completes the proof.

To study the seasonal variation of the growth of seaweeds, corals, and Parrotfish we consider the rate parameters as a sinusoidal function with a period of 1 year so that $\omega = \frac{2\pi}{365} = 0.01721$. Simulating the nonautonomous system (12.3) we observe that there exists a positive periodic solution with different phase differences and all the positive periodic solutions initiating from different initial values converge to a single periodic solution (cf. Fig. 12.12).

12.5 Discussion

In this paper, we have investigated the dynamics of coral reef benthic system in which seaweeds and corals are competing to occupy turf algae in presence of herbivorous Parrotfish. We assume that seaweeds immigrate from other areas of the sea bed while corals do not engage in immigration. We analyze the stability and bifurcations by linearizing the system about the equilibrium points, using the techniques previously adopted in [35]. The conditions for stability of the system is determined based on macroalgal toxicity and the harvesting rates of the herbivores. On analyzing our proposed model we observe that the system is capable of exhibiting the existence of two stable configurations of the community under identical environmental conditions, allowing saddle-node bifurcations along with hysteresis cycles. The transition between the branches of stable coexistence steady states is not reversible but exhibits hysteresis when the grazing rate of herbivores and seaweed growth rate cross some certain thresholds. It is observed that with low seaweed toxicity, the system exhibits two alternative stable states. With high toxicity level, the system becomes

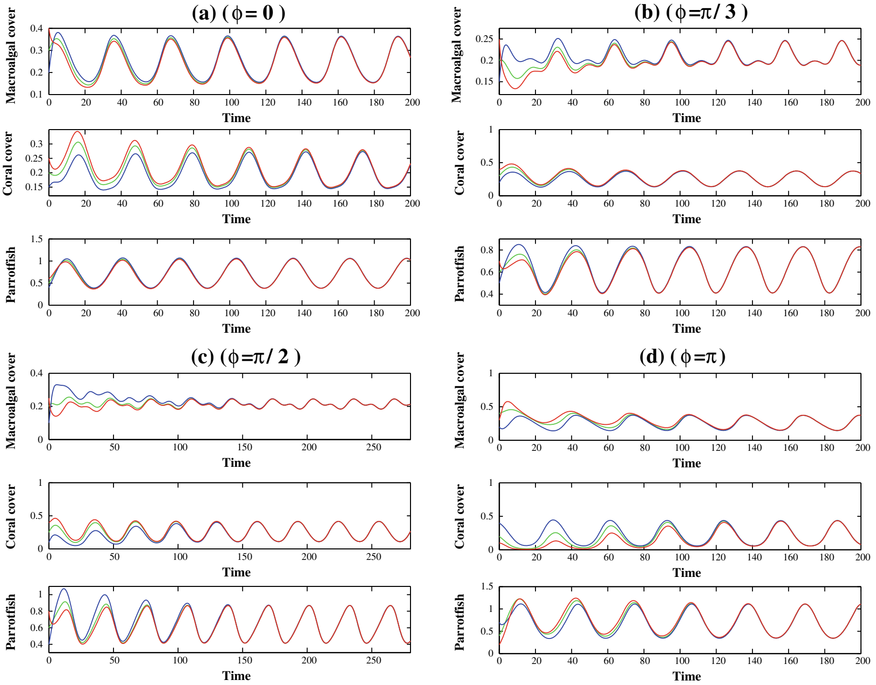


Fig. 12.12 Dynamic behavior of the nonautonomous system (12.3) with different initial conditions (represented by red, green and blue) and with different phase differences for $\epsilon_i = 0.5$ ($i = 1, \ldots, 7$), $\omega = 0.01721$ and other parameter values as in Table 12.1

locally asymptotically stable at coral-free equilibrium followed by a sudden change of transition and associated hysteresis effect, justifying the observations of [24] that allelopathy can suppress coral resilience by preventing coral recovery. It is observed that the system exhibits a sudden change of transition associated with saddle-node bifurcation and hysteresis effects when the immigration rate of seaweeds crosses some certain threshold. Also, with low grazing intensity of herbivores can lead to a sudden change of regime from a coral-dominated regime to one that is dominated by seaweeds. Moreover, overfishing of Parrotfish marks the transition from a coral-seaweeds bistable regime to a seaweeds-dominated regime. Based on the existence of two or multiple alternative stable configurations, we have defined different components of ecological resilience and illustrated the components inscribed in each of the bifurcation plots. We simulate the measures of ecological resilience by varying the key parameters of the model. For the monostable scenario where the ecological resilience is no longer applicable, we have defined the engineering resilience and evaluated the resilience of the system by giving some perturbation in the system at a given point of time. The perturbations and the corresponding measure of engineering resilience are illustrated as a time series plot of the system. Further, we include the effect of seasonal variations as sinusoidal functions of the biological parameters of our model to study the dynamics of the nonautonomous system. We obtain sufficient conditions for the existence of positive periodic solutions and observed that a positive periodic solution with different phase differences and all the positive periodic solutions initiating from different initial values converge to a single periodic solution.

Throughout the article, an attempt is made to search for a suitable way to restore corals from possible bleaching effect and maintain a healthy coral reef ecosystem. From analytical and numerical observations, it is seen that a sudden shift of transition from the coral-dominated regime to the seaweed-dominated regime can happen due to the reduction in herbivory. Further, analytical and numerical simulations demonstrate the following conclusions:

(i) With the increase of seaweed growth rate on corals, the resilience of coral-dominated coexistence steady state gradually decreases until the growth rate of seaweeds reaches a critical threshold. It is observed that as the seaweed growth rate is increased, two interior equilibria approach each other, collide and undergo mutual annihilation, leading to a catastrophic shift of regime to a seaweeds-dominated and coral-depleted steady state in presence of Parrotfish.

(ii) With low toxicity of seaweeds, the system becomes stable at the coral-dominated monostable regime. An increase of seaweed toxicity reduces the resilience of the system and determines two possible stable regimes depending upon the initial conditions. With high toxicity of seaweeds, the system becomes monostable at the seaweed-dominated equilibrium followed by the complete elimination of live corals, justifying the experimental observations of Bonaldo and Hay [24] that toxic seaweeds can exhibit significant negative impact on coral species.

(iii) With high seaweed toxicity, the increase of grazing rate of herbivores increases the resilience of the coral-dominated regime, signifying the importance of grazers in coral reefs affected by the allelopathy of seaweeds.

(iv) The system will be seaweed-dominated for low seaweed grazing intensity of herbivores. It is observed that a sudden shift of transition from the coral-dominated regime to the seaweed-dominated regime can happen due to the reduction in herbivory. An increase in the grazing intensity of herbivores increases the resilience of the coral-dominated regime.

(v) The resilience of the coexistence state decreases owing to the increase of seaweed toxicity, the increase in the rate of harvesting of herbivores, the increase of seaweed external immigration rate and the decrease of grazing intensity of herbivores.

Moreover, we have observed that there is a gradual decrease in the toxicity-tolerance level of the stable coexistence state with a steady increase in the harvesting rate of herbivores. Also, a sharp decrease in the toxicity-tolerance level of the stable coexistence state can occur even with a slight decrease of herbivore grazing intensity and a slight increase in the immigration rate of toxic seaweeds. The gradual decrease of ecological resilience followed by the emergence of a monostable regime can be taken into consideration as an early warning signal for a catastrophic shift of regime in coral reefs.

Acknowledgements SP is supported by the grants from Science & Technology and Biotechnology Department, Govt. of West Bengal, India (No. ST/P/S&T/16G-22/2018). JB acknowledges financial support in the form of research grants from Science & Technology and Biotechnology Department, Govt. of West Bengal, India (No. ST/P/S&T/16G-06/2018).

References

1. M. Scheffer, S.M. Carpenter, Catastrophic regime shifts in ecosystems: linking theory to observation. Trends Ecol. Evol. **18**, 648–656 (2003)
2. S.R. Dudgeon, R.B. Aronson, J.F. Bruno, W.F. Precht, Phase shifts and stable states on coral reefs. Mar. Ecol. Prog. Ser. **413**, 201–216 (2010)
3. S.J. Box, P.J. Mumby, Effect of macroalgal competition on growth and survival of juvenile Caribbean corals. Mar. Ecol. Prog. Ser. **342**, 139–149 (2007)
4. T.P. Hughes, N.A.J. Graham, J.B.C. Jackson, P.J. Mumby, R.S. Steneck, Rising to the challenge of sustaining coral-reef resilience. Trends Ecol. Evol. **25**(11), 633–642 (2010)
5. T.J. Done, Phase shifts in coral reef communities and their ecological significance. Hydrobiologia **247**, 121–132 (1992)
6. D.R. Bellwood, T.P. Hughes, C. Folke, M. Nystrom, Confronting the coral reef crisis. Nature **429**, 827–833 (2004)
7. J.F. Bruno, H. Swetman, W.F. Precht, E.R. Selig, Assessing evidence of phase shifts from coral to macroalgal dominance on coral reefs. Ecology **90**(6), 1478–1484 (2009)
8. M.A. Albins, M.A. Hixon, Invasive Indo-Pacific lionfish (Pterois Volitans) reduce recruitment of Atlantic coral-reef fishes. Mar. Ecol. Prog. Ser. **367**, 233–238 (2008)
9. A.J. Cheal, M.A. MacNeil, E. Cripps, M.J. Emslie, M. Jonker, B. Schaffelke, H. Sweatman, Coral-macroalgal phase shifts or reef resilience: links with diversity and functional roles of herbivorous fishes on the Great Barrier Reef. Coral Reefs **29**, 1005–1015 (2010)
10. D. Lirman, Competition between macroalgae and corals: effects of herbivore exclusion and increased algal biomass on coral survivorship and growth. Coral Reefs **19**, 392–399 (2001)
11. C.L. Birrell, L.J. McCook, B.L. Willis, G.A. Diaz-Pulido, Effects of benthic algae on the replenishment of corals and the implications for the resilience of coral reefs. Ocean. Mar. Biol.: Annu. Rev. **46**, 25–63 (2008)
12. J. Jompa, L.J. McCook, Effects of competition and herbivory on interactions between a hard coral and a brown alga. J. Exp. Mar. Biol. Ecol. **271**, 25–39 (2002)
13. V.J. Harriott, S.A. Banks, Latitudinal variation in coral communities in eastern Australia: a qualitative biophysical model of factors regulating coral reefs. Coral Reefs **21**, 83–90 (2002)
14. C.S. Holling, Resilience and stability of ecological systems. Annu. Rev. Ecol. Syst. **4**, 1–23 (1973)
15. B. Walker, C.S. Holling, S.R. Carpenter, A. Kinzing, Resilience, adaptability and transformability in social-ecological systems. Ecol. Soc. **9**(2), 5 (2004)
16. P.J. Mumby, R.S. Steneck, Coral reef management and conservation in light of rapidly evolving ecological paradigms. Trends Ecol. Evol. **23**(10), 555–563 (2008)
17. P.L. Antonelli, Nonlinear allometric growth I. Perfectly cooperative systems. Math. Model. **4**(4), 367–372 (1983)
18. L.J. McCook, J. Jompa, G. Diaz-Pulido, Competition between corals and algae on coral reefs: a review of evidence and mechanisms. Coral Reefs **19**, 400–417 (2001)
19. M.M. Nugues, R.P.M. Bak, Differential competitive abilities between Caribbean coral species and a brown alga: a year of experiments and a long-term perspective. Mar. Ecol. Prog. Ser. **315**, 75–86 (2006)
20. K.L. Barott, J.E. Caselle, E.A. Dinsdale, A.M. Friedlander, J.E. Maragos, D. Obura, F.L. Rohwer, S.A. Sandin, J.E. Smith, B. Zgliczynski, The lagoon at Caroline/Millennium Atoll, Republic of Kiribati: natural history of a nearly pristine ecosystem. PLoS One **5**(6), e10950 (2010)

21. J.A. Morris, J.L. Akins, A. Barse, D. Cerino, D.W. Freshwater, S.J. Green, R.C. Munoz, C. Paris, P.E. Whitefield, Biology and Ecology of Invasive Lionfishes, Pterois miles and Pterois volitans. Gulf Caribb. Fish. Inst. **61**, 1–6 (2009)
22. J.E. Smith, C.L. Hunter, C.M. Smith, The effects of top-down versus bottom-up control on benthic coral reef community structure. Oecologia **163**(2), 497–507 (2010)
23. J.W. McManus, J.F. Polsenberg, Coral-algal phase shifts on coral reefs: ecological and environmental aspects. Prog. Ocean. **60**, 263–279 (2004)
24. R.M. Bonaldo, M.E. Hay, Seaweed-coral interactions: variance in seaweed allelopathy, coral susceptibility, and potential effects on coral resilience. PLoS One **9**(1), e85786 (2014)
25. C.L. Birrell, L.J. McCook, B.L. Willis, L. Harrington, Chemical effects of macroalgae on larval settlement of the broadcast spawning coral Acropora millepora. Mar. Ecol. Prog. Ser. **362**, 129–137 (2008)
26. T.D. Andras, T.S. Alexander, A. Gahlena, R.M. Parry, F.M. Fernandez, J. Kubanek, M.D. Wang, M.E. Hay, Seaweed allelopathy against coral: surface distribution of seaweed secondary metabolites by imaging mass spectrometry. J. Chem. Ecol. **38**, 1203–1214 (2012)
27. D.B. Rasher, E.P. Stout, S. Engel, J. Kubanek, M.E. Hay, Macroalgal terpenes function as allelopathic agents against reef corals. Proc. Natl. Acadademy Sci. **108**(43), 17726–17731 (2011)
28. J.N. Underwood, L.D. Smith, M.J.H. Oppen, J.P. Gilmour, Ecologically relevant dispersal of corals on isolated reefs: implications for managing resilience. Ecol. Appl. **19**(1), 18–29 (2009)
29. J.C. Bythell, E.H. Gladfelter, M. Bythell, Chronic and catastrophic natural mortality of three common Caribbean reef corals. Coral Reefs **12**, 143–152 (1993)
30. E.H. Meesters, I. Wesseling, R.P.M. Bak, Coral colony tissue damage in six species of reef-building corals: partial mortality in relation with depth and surface area. J. Sea Res. **37**, 131–144 (1997)
31. P.J. Mumby, N.L. Foster, E.A.G. Fahy, Patch dynamics of coral reef macroalgae under chronic and acute disturbance. Coral Reefs **24**, 681–692 (2005)
32. T. Elmhirst, S.R. Connolly, T.P. Hughes, Connectivity, regime shifts and the resilience of coral reefs. Coral Reefs **28**, 949–957 (2009)
33. L. Perko, *Differential Equations and Dynamical Systems*, 3rd edn. (Springer, New York, 2001)
34. J.C. Blackwood, A. Hastings, P.J. Mumby, The effect of fishing on hysteresis in Caribbean coral reefs. Theor. Ecol. **5**, 105–114 (2012)
35. J. Bhattacharyya, S. Pal, Hysteresis in coral reefs under macroalgal toxicity and overfishing. J. Biol. Phys. **41**(2), 151–172 (2015)
36. G. Diaz-Pulido, J. Garzón-Ferreira, Seasonality in algal assemblages on upwelling-influenced coral reefs in the colombian caribbean. Bot. Mar. **45**, 284–292 (2002)
37. C.D. Lefèvre, D.R. Bellwood, Seasonality and dynamics in coral reef macroalgae: variation in condition and susceptibility to herbivory. Mar. Biol. **157**(5), 955–965 (2010)
38. C.J. Crossland, Seasonal variations in the rates of calcification and productivity in the coral Acropora formosa on a high-latitude reef. Mar. Ecol. Prog. Ser. **15**, 135–140 (1984)
39. R.E. Gaines, J.L. Mawhin, *Coincidence Degree and Non-Linear Differential Equations* (Springer, Berlin, 1977)

Chapter 13
Multigrid Methods for the Simulations of Surfactant Spreading on a Thin Liquid Film

Satyananda Panda and Aleksander Grm

Abstract A multigrid approach is proposed in this work for the simulations of surfactant spreading on a thin liquid film. The model equations for the descriptions of the surfactant dynamics are the coupled nonlinear partial differential equations in radial coordinate. The finite volume method on a uniform grid is used for the discretization of the governing equations in which the fluxes are discretized implicitly. The discretized system is solved using the nonlinear multigrid method such as the full approximation scheme. The obtained simulation results are discussed and validated with existing results.

Keywords Thin liquid film · Surfactant transport · Multigrid methods

13.1 Introduction

The simulation of surfactant dynamics on a thin liquid film has important applications in many areas of engineering and sciences, for example, surfactant replacement therapy [1], pulmonary drug delivery [2], crude oil recovery [3], ocular surfactant and blinking dynamics [4], etc. Such flows can be described by the system of partial differential equations (PDEs), which consists of equations for conservation of mass, momentum, and surfactant transport. The free surface boundary conditions support these equations, which are difficult in general for the solution. But the slenderness of the fluid domain enables the simplification of the full two/three-dimensional mathematical model using lubrication analysis. Subsequently, the surfactant flow can be

S. Panda (✉)
National Institute of Technology Calicut, NIT(P.O), Kozhikode
673601, Kerala, India
e-mail: satyanand@nitc.ac.in

A. Grm
Faculty of Maritime Studies and Transportation, University of Ljubljana,
Pot pomorscakov 4, 6320 Portoroz, Slovenia
e-mail: aleksander.grm@fpp.uni-lj.si

P. Manchanda et al. (eds.), *Mathematical Modelling, Optimization, Analytic and Numerical Solutions*, Industrial and Applied Mathematics,
https://doi.org/10.1007/978-981-15-0928-5_13

predicted by a one-dimensional coupled system of partial differential equations for the free surface height and the surfactant concentration. A closed-form solution of such combined system of equations is not available, except for certain simplified conditions, and they must be solved using appropriate numerical techniques. The process is transient and needs long simulation times for the fine grid. In this paper, a nonlinear multigrid method based on the finite volume method is presented. The method is applied for the solutions of the coupled PDEs derived by Gaver et al. [5] to predict the dynamics of free fluid surface and surfactant concentration distribution. We demonstrate that the proposed method is more accurate, efficient, and robust.

The paper is structured as follows. Section 13.2 deals with the description of the model proposed by Gaver and Grotberg [5] for the flow of surfactant concentration and the film thickness. In the next section, we describe the discretization procedure based on the finite volume method of the governing PDEs on radial coordinate. The nonlinear multigrid method is then discussed for the solution of the discrete nonlinear equations. In the penultimate section, we show the simulation results, and the validation and the mesh refinement analysis are performed. The last section presents the concluding remarks.

13.2 Model Description

The governing nondimensional equations which describe the axisymmetric spreading of an insoluble surfactant of a thin liquid film of Newtonian incompressible fluid are the system of nonlinear time-dependent partial differential equations for the film thickness $h(r, t)$ and surfactant concentration $\gamma(r, t)$. Here r is the radial coordinate and t is the time from the release of surfactant. The formulation of the problem, as well as the notations used, is from [1, 5, 6]

$$\frac{\partial h}{\partial t} = \frac{1}{r}\frac{\partial}{\partial r}\left(G\,r\frac{h^3}{3}\frac{\partial h}{\partial r} - r\frac{\partial \sigma}{\partial \gamma}\frac{\partial \gamma}{\partial r}\right) \tag{13.1}$$

$$\frac{\partial \gamma}{\partial t} = \frac{1}{r}\frac{\partial}{\partial r}\left(\frac{1}{\mathrm{Pe}}r\frac{\partial \gamma}{\partial r} - r\gamma h\frac{\partial \sigma}{\partial \gamma}\frac{\partial \gamma}{\partial r} + \frac{G}{2}r\gamma h^2\frac{\partial h}{\partial r}\right), \tag{13.2}$$

where G is the gravitational parameter, and Pe is the surface Peclet number.

The equation of state which describes the relationship between the surfactant concentration γ and the surface tension σ is defined by

$$\sigma(\gamma) = (\beta + 1)\left[1 + \Theta(\beta)\gamma\right]^{-3} - \beta\ , \tag{13.3}$$

with $\Theta(\beta) = ((\beta + 1)/\beta)^{1/3} - 1$. It should be noted that the constitutive Eq. (13.3) used by [7, 8] has the following properties: the surface tension is a monotonically decreasing function of surfactant concentration, and the value of the dimensionless surface tension lies between 0 and 1 and $\sigma(0) = 1$.

As reported in [5], the film is initially flat and locally contaminated by a spot of insoluble surfactant which prompts $h(r, t = 0) = 1$, and the starting condition for the surfactant distribution is given by

$$\gamma(r, t = 0) = \begin{cases} \gamma_{max} \ , \ (r \le RI) \\ \gamma_{max}\left(0.5\cos\left(\frac{\pi(r-RI)}{(1-RI)}\right) + 0.5\right) \ , \ (RI < r \le 1) \\ 0 \ , \ (r > RI) \ , \end{cases} \tag{13.4}$$

where we used $\gamma_{max} = 1$ and $RI = 0.7$ in the following.

Since the insoluble surfactant spreads radially on the surface, it is symmetric in any plane perpendicular to the surface. Due to this symmetry, the insoluble surfactant is treated as an axisymmetric body, and only half of the domain is considered in the analysis. Thus, because of the symmetry of the problem, we suppose that the film thickness and surfactant concentration are smooth at the origin. Accordingly, the gradient of all field variables must vanish about the axis $r = 0$, i.e.,

$$\frac{\partial h}{\partial r}(0, t) = 0, \quad \frac{\partial \gamma}{\partial r}(0, t) = 0. \tag{13.5}$$

For the far away condition, we also assume that

$$h(+\infty, t) = 1, \quad \gamma(+\infty, t) = 0, \tag{13.6}$$

where $+\infty$ stands for the limit of the computational domain. In the following, we consider the computational domain spans over the length 4 (nondimensional) which is sufficiently large enough such that the endpoint does not affect the spreading dynamics.

13.3 Discretization

The discretization of the coupled nonlinear partial differential equations (PDEs) (13.1) and (13.2) subject to initial and boundary conditions is performed using finite volume method [9] on radial coordinates. The details of the finite volume discretization are given in [6]. We describe in brief here for the completeness. In order to solve this set of PDEs, we first adapt Eqs. (13.1), (13.2) in conservative form as follows:

$$r\frac{\partial h}{\partial t} = \frac{\partial}{\partial r}\left(G\, r\frac{h^3}{3}\frac{\partial h}{\partial r} - r\frac{\partial \sigma}{\partial \gamma}\frac{\partial \gamma}{\partial r}\right) \tag{13.7}$$

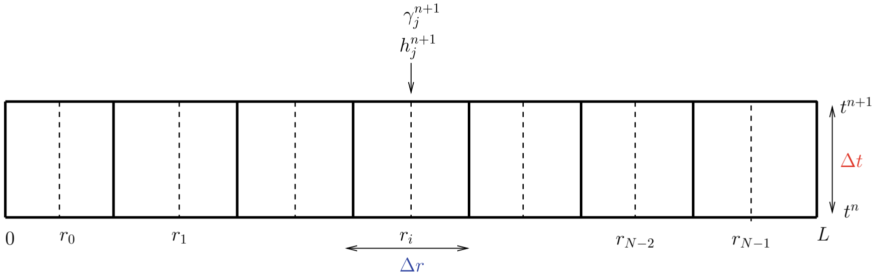


Fig. 13.1 Finite volume mesh and notations

$$r\frac{\partial \gamma}{\partial t} = \frac{\partial}{\partial r}\left(\frac{1}{\mathrm{Pe}} r\frac{\partial \gamma}{\partial r} - r\gamma h\frac{\partial \sigma}{\partial \gamma}\frac{\partial \gamma}{\partial r} + \frac{G}{2}r\gamma h^2\frac{\partial h}{\partial r}\right). \tag{13.8}$$

We discretize the flow domain $[0, L]$ into N equal size grid cells of size $\Delta r = L/N$ (see Fig. 13.1). We define the center of the cell r_j as $r_j = \Delta r/2 + j\Delta r$, $j = 0, 1, \ldots, N-1$. The cell edges of the cell j are located at $r_{j-1/2} = r_j - \Delta r/2$ and $r_{j+1/2} = r_j + \Delta r/2$. In order to set up discrete equations, the functions h and γ are approximated over the cell $[r_{j-1/2}, r_{j+1/2}]$, i.e.,

$$h_j(t) \sim h(r_j, t) = \frac{1}{\Delta r}\int_{r_{j-1/2}}^{r_j+1/2} h(r, t)dr \tag{13.9}$$

and

$$\gamma_j(t) \sim \gamma(r_j, t) = \frac{1}{\Delta r}\int_{r_{j-1/2}}^{r_j+1/2} \gamma(r, t)dr. \tag{13.10}$$

The discrete relations for $h_j(t)$ and $\gamma_j(t)$ for $j = 1, 2, \ldots, N-2$ are obtained by integrating the governing Eqs. (13.7) and (13.8) over the interval $[r_{j-1/2}, r_{j+1/2}]$, i.e.,

$$\int_{r_j-\frac{1}{2}\Delta r}^{r_j+\frac{1}{2}\Delta r} r\frac{\partial h}{\partial t}dr = \int_{r_j-\frac{1}{2}\Delta r}^{r_j+\frac{1}{2}\Delta r} \frac{\partial}{\partial r}\left(G\,r\frac{h^3}{3}\frac{\partial h}{\partial r} - r\frac{\partial \sigma}{\partial \gamma}\frac{\partial \gamma}{\partial r}\right)dr, \tag{13.11}$$

and

$$\int_{r_j-\frac{1}{2}\Delta r}^{r_j+\frac{1}{2}\Delta r} r\frac{\partial \gamma}{\partial t}dr = \int_{r_j-\frac{1}{2}\Delta r}^{r_j+\frac{1}{2}\Delta r} \frac{\partial}{\partial r}\left(\frac{1}{\mathrm{Pe}} r\frac{\partial \gamma}{\partial r} - r\gamma h\frac{\partial \sigma}{\partial \gamma}\frac{\partial \gamma}{\partial r} + \frac{G}{2}r\gamma h^2\frac{\partial h}{\partial r}\right)dr. \tag{13.12}$$

Following the integration procedure given in [6], we obtain

$$\int_{r_j-\frac{1}{2}\Delta r}^{r_j+\frac{1}{2}\Delta r} r\frac{\partial h}{\partial t}dr = \Delta r\frac{\partial}{\partial t}\left(\frac{1}{8}r_{j+1}h_{j+1} + \frac{6}{8}r_j h_j + \frac{1}{8}r_{j-1}h_{j-1}\right), \tag{13.13}$$

and similarly for the left-hand side of Eq. (13.12).

Finally, we obtain the following discrete equations after integrating the right-hand side of Eqs. (13.11) and (13.12),

$$\Delta r\frac{\partial}{\partial t}\left(\frac{1}{8}r_{j+1}h_{j+1} + \frac{6}{8}r_j h_j + \frac{1}{8}r_{j-1}h_{j-1}\right) = F^{n+1}_{r_j+1/2} - F^{n+1}_{r_j-1/2}\ , \tag{13.14}$$

and

$$\Delta r\frac{\partial}{\partial t}\left(\frac{1}{8}r_{j+1}\gamma_{j+1} + \frac{6}{8}r_j \gamma_j + \frac{1}{8}r_{j-1}\gamma_{j-1}\right) = G^{n+1}_{r_j+1/2} - G^{n+1}_{r_j-1/2}. \tag{13.15}$$

The discrete flux functions $F^{n+1}_{r_j+1/2}$ and $G^{n+1}_{r_j+1/2}$ are given by

$$F^{n+1}_{r_j+1/2} = F\left(r_j + \frac{1}{2}\Delta r, t^{n+1}\right) \quad \text{and} \quad G^{n+1}_{r_j+1/2} = G\left(r_j + \frac{1}{2}\Delta r, t^{n+1}\right)\ . \tag{13.16}$$

The face values are evaluated as the mid-values of the two neighboring nodal values, i.e.,

$$h(r_{j+1/2}, t^{n+1}) = \frac{1}{2}\left(h^{n+1}_j + h^{n+1}_j\right) \tag{13.17}$$

and the forward differences are used for the evaluation of gradient, i.e.,

$$\frac{\partial h}{\partial r}(r_{j+1/2}, t^{n+1}) = \frac{1}{\Delta r}\left(h^{n+1}_{j+1} - h^{n+1}_j\right). \tag{13.18}$$

We approximate all time derivative terms using forward differences, e.g.,

$$\frac{\partial h_j}{\partial t} = \frac{h^{n+1}_j - h^n_j}{\Delta t}. \tag{13.19}$$

At the boundary nodes r_0 and r_{N-1}, the discretized equations are derived applying the boundary conditions (13.5) and (13.6). We additionally assume that at the boundary nodes, the value of the time derivative term which is outside the cell is zero.

13.4 Numerical Approach—Nonlinear Multigrid Method (FAS)

The large system of equations obtained by time and space discretization of the partial differential equations has to be solved for every time step. There are many linear and nonlinear solvers to solve such system. The multigrid strategies are a great class of iterative solvers for the solution of the discretized PDEs. The main idea of the multigrid techniques was first presented by Southwell [10], where he depicted an application which solves on a coarse framework and afterward interpolated the solution on a fine grid to enhance the initial guess. Brandt [11] introduced systematically the multigrid methods and their applications. There are two fundamental kinds of multigrid strategies: geometric and algebraic. This work is concerned about geometric multigrid techniques, in which geometric data with respect to the problem are utilized to form a solution algorithm. The geometric multigrid method operates on the hierarchy of the grids. The algebraic multigrid works on the principle of the multigrid method but does not require the grid information [12, 13]. In this work, we have developed a nonlinear multigrid [14] algorithm based on the finite volume method for the solution of the nonlinear system of equations described in Sect. 13.3. We described the procedure in detail.

Let $\mathscr{L}(\mathbf{y}) = \mathbf{f}$ be the given nonlinear system of equations. After discretization with grid size Δr, we get a system of nonlinear equations $\mathscr{L}^{\Delta r}(\mathbf{y}^{\Delta r}) = \mathbf{f}^{\Delta r}$, where $\mathscr{L}$ is a nonlinear operator. As per discretization given in Sect. 13.3, we have $\mathbf{y}^{\Delta r} = (y_1, y_2, \dots, y_{2N-1}, y_{2N})^*$ and $\mathbf{f}^{\Delta r}(\mathbf{y}) = (f_1(\mathbf{y}), f_2(\mathbf{y}), \dots, f_{2N-1}(\mathbf{y}), f_{2N}(\mathbf{y}))^*$. Here the superscript star $(*)$ denotes the transpose operator. The solution of this nonlinear system can be obtained using any nonlinear solver. In this work, we propose a nonlinear multigrid method, known as the full approximation scheme (FAS) to obtain the numerical solution of the discrete equations.

The method begins with an initial guess and then three Newton–Raphson iterations are applied for all the internal nodes on such a grid as the pre-smoother in order to smooth the high-frequency error. The residual on the finest grid is calculated. The next step is restricting both residual and the value of $\mathbf{y}^{\Delta r}$ onto a coarser grid. Additionally, the modified right-hand side is also obtained and stored on the coarse grid. This process is recursively called on every grid (except the coarsest grid), until the coarsest grid is reached. Then the coarse grid problem is solved exactly by using the Newton–Raphson method. Since the main reason of using the full approximate storage (FAS) is to store the actual value of $\mathbf{y}$ on every grid (including the coarsest grid), we will obtain an exact solution at the coarsest grid. By subtracting this solution from the restricted value of $\mathbf{y}$ from the finer grid, we can obtain the error term. Such error is interpolated recursively back to a finer grid, and simple correction is applied by adding the old value $\mathbf{y}$ on that grid with this interpolated error on every grid. This interpolation process runs until the finest grid is reached. Finally, three Newton–Raphson iterations are taken as the post-smoother on each grid immediately after the coarse grid correction. This procedure is called the nonlinear multigrid V-cycle [14]. The algorithm is given below.

FAS Algorithm—A Nonlinear Multigrid (NLMG) Method (v-Cycle) for the System of PDEs

Let γ_1 and γ_2 be the number of iterations performed by Newton–Raphson method as pre- and post-smothers, respectively. Here Δr is the grid length on a given grid $\Omega_{\Delta r}$ and Δt is the known time step.

Function: $\mathbf{y}^{\Delta r}$ = FASNLMG $\left(\mathbf{y}^{\Delta r}, \mathbf{f}^{\Delta r}, \mathscr{L}^{\Delta r}(\mathbf{y}^{\Delta r}), \Delta t\right)$:

- Given an initial guess $\mathbf{y}_o^{\Delta r}$, relax γ_1 times on $\mathscr{L}^{\Delta r}(\mathbf{y}^{\Delta r}) = \mathbf{f}^{\Delta r}$.
- Compute the fine grid residual $\mathbf{r}_{res}^{\Delta r} = \mathbf{f}^{\Delta r} - \mathscr{L}^{\Delta r}(\mathbf{y}^{\Delta r})$.
- Restrict fine grid residual to the coarse grid as $\mathbf{r}_{res}^{2\Delta r} = R_{2\Delta r}^{\Delta r}\mathbf{r}_{res}^{\Delta r}$, where $R_{2\Delta r}^{\Delta r}$ is a full weighted average operator.
- Initialize coarse guess: $\mathbf{y}_0^{2\Delta r} = \tilde{R}_h^{2\Delta r}\tilde{\mathbf{y}}^{\Delta r}$, where $\tilde{R}_{\Delta r}^{2\Delta r}$ is the restriction operator.
- Compute the coarse right-hand side vector: $\mathbf{f}^{2\Delta r} = \mathscr{L}^{2\Delta r}\left(\mathbf{y}_0^{2\Delta r}\right) + \mathbf{r}_{res}^{2\Delta r}$.
- if $\Omega_{2\Delta r}$ is the coarsest grid, then solve: $\mathscr{L}^{2\Delta r}(\mathbf{y}^{2\Delta r}) = \mathbf{f}^{2\Delta r}$ for $\mathbf{y}^{2\Delta r}$.
- else $\mathbf{y}^{2\Delta r}$=FASNLMG $\left(\mathbf{y}^{2\Delta r}, \mathbf{f}^{2\Delta r}, \mathscr{L}^{2\Delta r}(\mathbf{y}^{2\Delta r}), \Delta t\right)$, endif.
- Compute the error: $\mathbf{e}^{2\Delta r} = \mathbf{y}^{2\Delta r} - \mathbf{y}_0^{2\Delta r}$.
- Interpolate the error approximation to the fine grid: $\mathbf{e}^{\Delta r} = I_{2\Delta r}^{\Delta r}\mathbf{e}^{2\Delta r}$, where $I_{2\Delta r}^{\Delta r}$ is the linear interpolation operator.
- Correct the solution on the fine grid: $\mathbf{y}^{\Delta r} = \mathbf{y}^{\Delta r} + \mathbf{e}^{\Delta r}$.
- Relax on the new solution γ_2 times.

The nonlinear multigrid (NLMG) algorithm includes the step of solving system of nonlinear equations for a specific time step with Newton–Raphson (NR) method. In NR method, we have to provide the information of Jacobian. In our case, the Jacobian is approximated with the finite difference method.

13.5 Result and Discussions

The proposed NLMG algorithm was implemented in Matlab, and the results are obtained with an accuracy of order 10^{-6}. Figure 13.2a, b shows the numerical results obtained using NLMG for $G = 1$, $Pe = 10$, $\beta = 5$, $RI = 0.5$, and $\gamma_{max} = 1$. The results show the evolution of film thickness over time. The total simulation time is 1 (nondimensional). The figure shows that the film thickness at the center of the domain decreases due to the liquid having high surfactant concentration are draining away. The surfactant concentration distribution is given at the right panel (Fig. 13.2b) for different times. The figure shows that the spreading of surfactant distribution increases with the advancement of time.

13.5.1 Validation

To exhibit the effective implementation of the NLMG technique, the numerical results are first compared to those obtained by Gaver and Grotberg in [5]. The NLMG

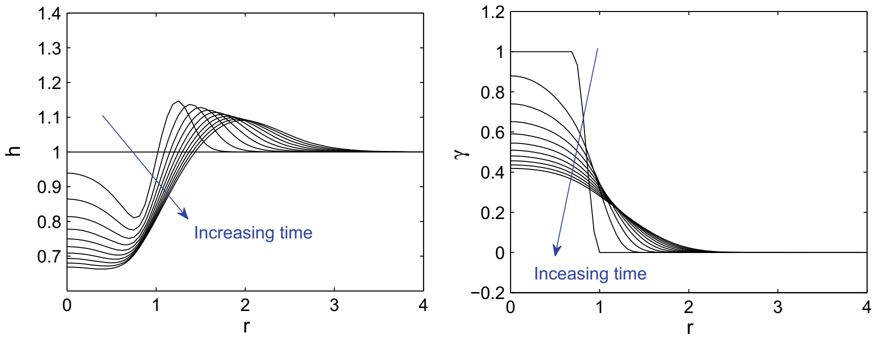


Fig. 13.2 Numerical results computed for film thickness $h(r, t)$ and surfactant concentration $\gamma(r, t)$ at several times using the proposed NLMG method for $G = 1$, $Pe = 10$, $\beta = 5$, $RI = 0.7$, and $\gamma_{max} = 1$. The total simulation time is 1 (dimensionless time unit) with increment of 0.1

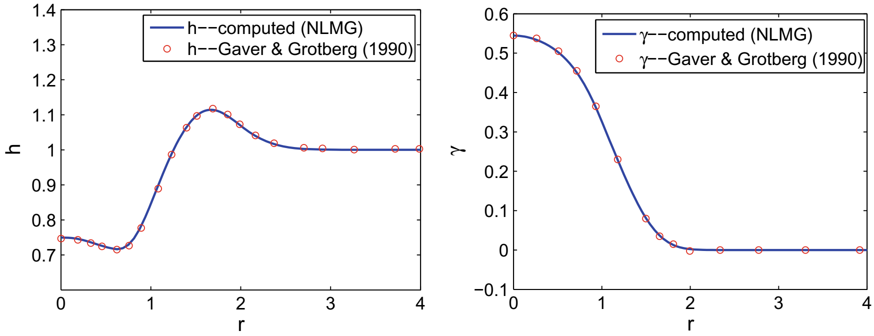


Fig. 13.3 Comparison of the results obtained by Gaver et al. [5] and the numerical simulation using NLMG method for $G = 1$, $Pe = 10$, $\beta = 5$, $RI = 0.7$, $\gamma_{max} = 1$ and $t = 0.5$: left: film thickness distribution; **b** surfactant concentration distribution

simulation was performed with the grid points $2^6 + 1$ and time step $\Delta t = 0.001$. Figure 13.3 demonstrates the numerical results obtained using the NLMG method (solid line) for $G = 1$, $Pe = 10$, $\beta = 5$, $RI = 0.7$, $\gamma_{max} = 1$, and t $= 0.5$ (nondimensional). The figure shows the NLMG results that are in good agreement, which provides the necessary confidence that the NLMG method has been correctly implemented.

Although the numerical scheme validates implementation and compares well with the solution in the available literature, we need to make sure that the solution is also independent of mesh and time resolution. The mesh and time resolution analyses are shown in the following subsection.

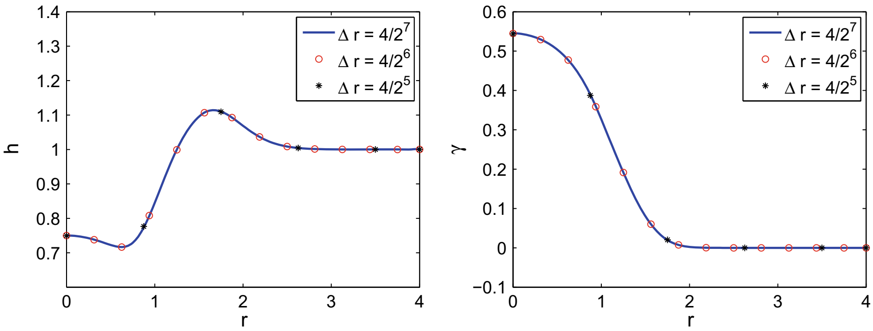


Fig. 13.4 Grid convergence test: **a** film thickness and **b** surfactant concentration at time $t = 0.5$ with time step $\Delta t = 0.001$

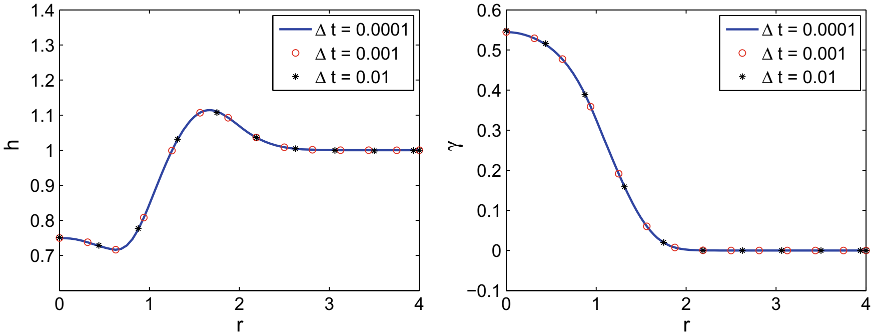


Fig. 13.5 Time convergence test with grid size $\Delta r = 4/2^6$: **a** film thickness and **b** surfactant concentration

13.5.2 Mesh and Time Convergence Study

The mesh and time convergence studies were conducted on NLMG method to find out optimum mesh size to balance between the accuracy and computational easiness. In the present analysis, three different mesh sizes were used for the solutions, and it can be observed that the results (Fig. 13.4) are independent of the mesh grid. Thus, it validates that the convergence of the NLMG solution is independent of the grid mesh size. Similarly, for the time-independent study, the results for the film thickness height and the surfactant concentration distribution are plotted (Fig. 13.5) for the three different time step sizes keeping the grid size $\Delta r = 4/2^6$ constant. The results at different time steps are indistinguishable, which is another advantage to the proposed method to compute solution at the faster time.

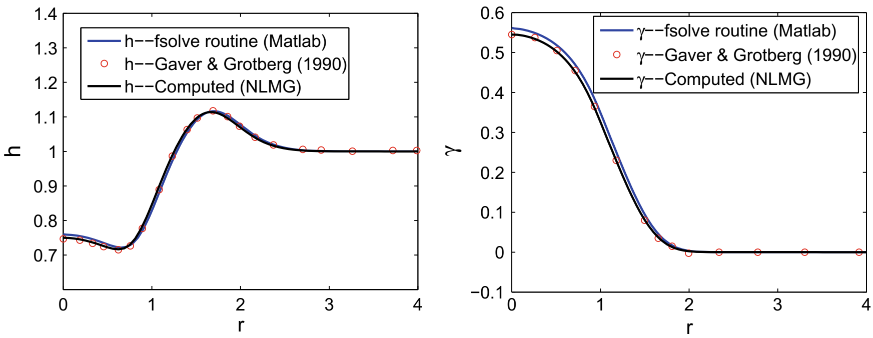


Fig. 13.6 Comparison among results of Gaver et al. [5] and numerical results by fsolve (Matlab) and NLMG

13.5.3 Comparison with Matlab Solver

In the following, we compare the performance of the NLMG method with Matlab [15] nonlinear solver *fsolve*, as illustrated in Fig. 13.6. The figure plots the film thickness and surfactant concentration distributions at time $t = 0.5$ (nondimensional). For the comparison, we solve the discretized system using Matlab *fsolve* routine. This routine solves the nonlinear equations using trust-region-dogleg (Levenberg–Marquardt) method. The number of grid points chosen is equal to $2^6 + 1$, and the time step is $\Delta = 0.001$. It can be observed that the solution obtained with the *fsolve* does not agree well with the solution of Gaver and Grotberg [5] whereas the NLMG solver produces the exact result that matches well. It is further observed that the NLMG solver performs well even with larger time step and confirms the robustness of the multigrid numerical method.

13.5.4 Mesh Refinement Analysis

We display in Table 13.1 a mesh refinement analysis for the film thickness h and the surfactant concentration γ. We ran the NLMG scheme for the grid points $2^{32} + 1$, $2^{64} + 1$ and $2^{128} + 1$. For the analysis, we denote $\theta^{(M)}$ as the variable θ computed with $2^M + 1$ grid points, and its relative error is estimated by comparison to the most refined computation, i.e.,

$$e\left(\theta^{(M)}\right) = \frac{\left\|\theta^{(M)} - \theta^{(128)}\right\|_2}{\left\|\theta^{(128)}\right\|_2}, \tag{13.20}$$

where $\|.\|_2$ is the L^2 norm. The relative error for the film thickness and the surfactant distribution is computed using Eq. (13.20) and given in Table 13.1. The table shows

Table 13.1 Mesh refinement analysis for the film thickness (h) and the surfactant concentration (γ) at various times for $G = 1$, $Pe = 10$, $\beta = 5$, $RI = 0.7$, and $\gamma_{max} = 1$

$Time\,(t)$	$e\left(h^{(32)}\right)$	$e\left(h^{(64)}\right)$	Rate	$e\left(\gamma^{(32)}\right)$	$e\left(\gamma^{(64)}\right)$	Rate
0.1	0.0013	3.7781e-04	1.7828	0.0027	8.0066e-04	1.7537
0.5	0.0011	3.1301e-04	1.8132	0.0036	9.9039e-04	1.8619
1.0	9.2714e-04	2.7631e-04	1.7465	0.0034	9.5498e-04	1.8320
2.0	8.9214e-04	2.8451e-04	1.6488	0.0031	8.9824e-04	1.7871

that both the variables attain the superlinear rate of convergence. It can be further noted that the superlinear accuracy is achieved for all the times, but the convergence rate decreases after time $t = 0.5$ (nondimensional) for both the film thickness and surfactant variables. This can be attributed to the fact that at the origin the surfactant is initially higher results in lower the surface tension that drives the fluid away from the center at the faster speed. With the advancement of time, the spreading rate of surfactant concentration distribution is slower, which decreases the movement of fluid from the origin as it approaches the steady state.

13.6 Validation Case with Sharp Changes in Gradient in the Free Surface Profile and Surfactant Concentration

As one intuitively expects, the free surface of the thin film and concentration can also experience a shock-type structure in the absence of surface tension, and the question arises as to whether the proposed NLMG method can predict the sharp changes in gradient in the free surface profile and surfactant concentration. To address this question, we consider a model given in [16, 17] in the absence of surface tension for the description of the spreading of an insoluble surfactant on the free surface of a thin liquid film. The governing nonlinear partial differential equations in Cartesian coordinates system for the free surface height $h = h(x, t)$ of the thin liquid film and the surfactant concentration $\Gamma = \Gamma(x, t)$ are

$$\frac{\partial h}{\partial t} + \frac{\partial Q}{\partial x} = 0, \tag{13.21}$$

and

$$\frac{\partial \Gamma}{\partial t} + \frac{\partial P}{\partial x} = 0, \tag{13.22}$$

where

$$Q = -\frac{h^2}{2}\frac{\partial \Gamma}{\partial x} + \frac{h^3}{3}, \quad \text{and} \quad P = -h\Gamma\frac{\partial \Gamma}{\partial x} + \frac{h^2}{2}\Gamma.$$

The boundary conditions are

$$h(0,t) = 1, \quad \Gamma(0,t) = 1 \tag{13.23}$$

and

$$\frac{\partial h}{\partial x} \mid_{x=0,x=\infty} = 0, \qquad \frac{\partial \Gamma}{\partial x} \mid_{x=0,x=\infty} = 0. \tag{13.24}$$

The initial film profile is considered as exponential:

$$h(x,0) = e^{-x^2}, \tag{13.25}$$

and for the exogenous surfactant, we have

$$\Gamma(x,0) = e^{-x^2}. \tag{13.26}$$

Authors Momoniat et al. [17] have developed a higher order numerical scheme based on finite volume approximation for the solution of Eqs. (13.21) and (13.22). A BDF approximation of order four is used for the time derivative, and the fluxes are approximated explicitly by the three-point central difference scheme with Roe–Sweby flux limiter.

For the validation purpose, we follow the same finite volume discretization as described in Sect. 13.3 to discretize the spatial derivatives. The discrete equations are in the form:

$$h_j^{n+1} = h_j^n + \frac{\Delta t}{\Delta x}\left(Q_{j+1/2}^{n+1} - Q_{j-1/2}^{n+1}\right), \tag{13.27}$$

and

$$\Gamma_j^{n+1} = \Gamma_j^n + \frac{\Delta t}{\Delta x}\left(P_{j+1/2}^{n+1} - P_{j-1/2}^{n+1}\right). \tag{13.28}$$

The fluxes are approximated implicitly, i.e.,

$$Q_{j+1/2}^{n+1} = Q\left(x_j + \frac{1}{2}\Delta x, t^{n+1}\right), \quad \text{and} \quad P_{j+1/2}^{n+1} = P\left(x_j + \frac{1}{2}\Delta x, t^{n+1}\right)$$

with

$$\left(\frac{\partial \Gamma}{\partial x}\right)_{j+1/2}^{n+1} = \frac{\left(\Gamma_{j+1}^{n+1} - \Gamma_j^{n+1}\right)}{\Delta x},$$

and

$$h_{j+1/2}^{n+1} = \frac{1}{2}\left(h_j^{n+1} + h_{j+1}^{n+1}\right).$$

The discretized Eqs. (13.27) and (13.28) with prescribed initial and boundary conditions are solved using the developed NLMG solver. For the fine resolution

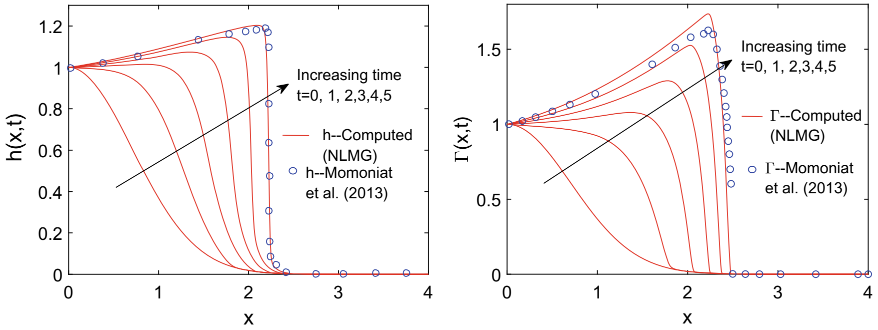


Fig. 13.7 Comparison of results of Momoniat et al. [17] and numerical results by NLMG at time t = 5

of the solution, the number of grid points chosen is equal to $2^{12} + 1$, and the time step is $\Delta t = 0.001$. The numerical results obtained with the proposed algorithm are compared to those obtained by Momoniat et al. [17] at time t = 5. Specifically, the data of results of Fig. 2 in [17] are read here for validation purpose. Figure 13.7 shows the free surface profile of the thin film (left panel) and the surfactant distribution (right panel) at different times.

It is apparent from these profiles that the steep changes in gradient in film thickness and concentration variations occur on a much shorter length scale which will lead to a more difficult test case for the NLMG algorithm, as one would expect. The reason is that the derivative terms in the finite volume discretization are of lower order approximation, and thus, a higher order finite volume discretization will be explored in future for the correct resolution of the solution. In spite of these limitations, the NLMG solver is capturing the shock front for the small time step without introducing any limiters.

13.7 Concluding Remarks

In this work, a nonlinear multigrid method based on finite volume method in radial coordinate is presented to solve the surfactant-driven thin liquid film equations derived by Gaver et al. [5]. We have demonstrated that the developed algorithm is quite robust and stable in the sense that it solves the discretized nonlinear equations for the large time step compared to ordinary methods. The simulation results of the nonlinear multigrid code were validated with the existing results of Gaver and Grotberg [5]. The time effectiveness of NLMG solver was explored compared to the Matlab solver *fsolve*. The results were very impressive showing speedup of few orders compared to the Matlab solver. Also, the accuracy of the Matlab solver was not satisfactory. Expectedly, not every free surface profile and surfactant concentration can be captured with a first-order numerical scheme with multigrid solver when the changes in the gradient of these profiles are very sharp. There is obvious limitation to the proposed finite volume discretization scheme when there is a sharp change in the gradient of the solution variables. Therefore, in order to have a shock-capturing NLMG solver, the planned work will develop a higher order finite volume discretization scheme.

In spite of these limitations, the proposed NLMG solver is very fast, accurate, and robust for the particular problems and can be implemented in many different areas of numerical solutions for highly nonlinear time-dependent PDE systems.

References

1. D.P. Gaver, J.B. Grotberg, Droplet spreading on a thin viscous film. J. Fluid Mech. **235**, 399–414 (1992)
2. J.B. Grotberg, Pulmonary flow and transport phenomena. Annu. Rev. Fluid Mech. **26**, 529–571 (1994)
3. D.O. Shan, R.S. Schechter, *Improved Oil Recovery by Surfactant and Polymer Flooding* (Academic, New York, 1977)
4. R.J. Braun, Dynamics of the tear film. Annu. Rev. Fluid Mech. **44**, 267–297 (2012)
5. D.P. Gaver, J.B. Grotberg, The dynamics of a localized surfactant on a thin film. J. Fluid Mech. **213**, 127–148 (1990)
6. M. Sellier, S. Panda, Unraveling surfactant transport on a thin liquid film. Wavemotion **70**, 183–194 (2017)
7. M.S. Borgas, J.B. Grotberg, Monolayer flow on a thin film. J. Fluid Mech. **193**, 151–170 (1988)
8. E.R. Swanson, S.L. Strickland, M. Shearer, K.E. Daniels, Surfactant spreading on a thin liquid film: reconciling models and experiments. J. Eng. Math. 1–17 (2014)
9. H.K. Versteeg, W. Malalasekera, *An Introduction to Computational Fluid Dynamics: The Finite Volume Method* (Pearson Education, Harlow, 1995)
10. R.V. Southwell, *Relaxation Methods in Theoretical Physics* (Clarendon Press, Oxford, 1946)
11. A. Brandt, Multi-level adaptive solutions to boundary-value problems. Math. Comput. **31**, 333–390 (1977)
12. G. Haase, U. Langer, Multigrid methods: from geometrical to algebraic versions, in *Modern Methods in Scientific Computing and Applications*. NATO Science Series (Series II: Mathematics, Physics and Chemistry), vol. 75, ed. by A. Bourlioux, M.J. Gander, G. Sabidussi (Springer, Dordrecht, 2002)
13. U. Trottenberg, C. Oosterlee, A. Schuller, *Multigrid* (Academic, New York, 2001)
14. W.L. Briggs, V.E. Henson, S.F. Mccormick, *A Multigrid Tutorial*, 2nd edn. (SIAM, Philadelphia, 2000)
15. The MathWorks Inc., MATLAB R2011b documentation, The MathWorks Inc. (2011)
16. E.R. Peterson, M. Shearer, Simulation of spreading surfactant on a thin liquid film. Appl. Math. Comput. **218**(9), 5157–5167 (2012)
17. E. Momoniat, M.M. Rashidi, R.S. Herbst, Numerical investigation of thin film spreading driven by surfactant using upwind schemes. Math. Probl. Eng. **2013**, Article ID 325132 (2013), 8 pp

Chapter 14
Hopf Bifurcation in a Mathematical Model of Tuberculosis with Delay

Eenezer Bonyah, Fahad Al Basir and Santanu Ray

Abstract Tuberculosis is an air-borne infectious disease which is transmitted to one another through the respiratory system and mostly occurs due to close contact with an infected person. Here, a mathematical of SIR type is proposed for the dynamics of tuberculosis with the effect of treatment and time delay. The level of treatment is assumed proportional to the number of infected people reported to the health organization. The equilibria and stability analysis has been carried out using qualitative theory. This paper provides some vital information such as the basic reproduction number R_0 and the stability of equilibrium points. Hopf bifurcation at the endemic steady states has been analyzed taking delay as the main parameter. Numerical simulations fulfill analytical outcomes. We found that large time delay in treatment can cause problems and it should be avoided.

Keywords Mathematical model · Delay differential equation · Basic reproduction number R_0 · Stability · Hopf bifurcation

14.1 Introduction

Tuberculosis is known to have killed humans than any other disease of mankind and the infection rate is higher than any other disease in the world [1]. In Africa TB is considered as a dangerous disease and many nongovernmental organizations are

E. Bonyah
Department of Information Technology Education, University of Education Winneba, Kumasi-campus, Winneba, Ghana

F. Al Basir (✉)
Department of Mathematics, Asansol Girls' College, Asansol-4, Asansol, West Bengal 713304, India
e-mail: fahadbasir@gmail.com

S. Ray
Systems Ecology & Ecological Modeling Laboratory, Department of Zoology, Visva-Bharati, Santiniketan 731235, India

P. Manchanda et al. (eds.), *Mathematical Modelling, Optimization, Analytic and Numerical Solutions*, Industrial and Applied Mathematics,
https://doi.org/10.1007/978-981-15-0928-5_14

helping to manage the death associated with this disease. There are instances where drugs for curing TB are given free to patients however, in some communities in Africa due to cultural practices that has not been successful. The developed countries are also facing the challenges of TB. For instance, in 2015 alone about 10.4 million people were contracted with Mycobacterium tuberculosis (Mtb), of which 1.8 million people perished due to the disease [2, 3]. Nearly 80% of the current report cases of TB in the world crop up in 22 high burden countries noted for a high incidence rate from 59 to 1003 per 100,000 people. China and India alone account for 38% of the total TB cases in the world. The one-third of the world's population is infected with TB and this shows the seriousness of TB [4].

Time delays on the treatment of TB has a serious consequence on the spread and the control of the disease. This is because it reduces the chance of survival, the cost of treatment and also the productivity of the individuals. It can be connected with both patients' altitude and medical health facilities [4]. TB has become a major health problem in both developing and developed countries due to drug resistance [5]. Of all the advancements in medicine and technology TB still remains one of the major causes of death in many high incidence countries. Thus, no country is safe irrespective of the health care system [6].

Mathematical modeling has become a powerful tool for examining dynamics of diseases in order to provide clear information on the spread and control of many infectious diseases [7, 8]. Several mathematical models in different forms have been constructed to study the dynamics of TB [9–13]. Houben et al. [11] constructed a mathematical model to examine the feasibility of achieving the 2025 global TB target among three countries South Africa, China, and India. Li et al. [12] proposed a mathematical model incorporating mixed cross infection in public farms. In [14], Blower developed a mathematical model to study the intrinsic dynamics of TB. Jia et al. [15] have examined the impact of immigration on the transmission of TB while Bhunu [16] constructed a TB model with chemoprophylaxis. Cohen and Murray [17] has developed a mathematical model to explore the multi drug resistant M. tuberculosis of heterogeneous fitness.

Delay in treatment plays a crucial role in the survival of TB patient however, there have been few mathematical models on delay, [6, 18] and the references therein. The effect of delay in the treatment of TB in many of the Sub-Saharan countries cannot be quantified because of the geographical and cultural arrangement [19]. Even in some communities it is wrong to send a sick person to the hospital utill the gods have been consulted. In some cases, poverty is so high such that many of them cannot meet the cost for treatment where TB treatment is not free. In Ghana, for example, the government has made the treatment of this disease for free and also a constant public education is being undertaken.

In this article, we assume the level of control in the form of treatment for the control of the disease. Moreover, it is assumed that the level of control is proportional to the number of infected people reported to the health organization. The aim here is to examine the effect of time delay in the treatment of TB which is vital for it controls such as in determining drug dosage, efficacy of drugs and others. Consequently, a delay model is formulated and analyzed. Numerically, we have shown the main results.

14.2 Mathematical Model Formulation

Following assumptions are made for the formulation of desired mathematical model.

- The model sub-partitions the entire human population at time t into the following sub-populations of susceptible individuals who are not yet suffering tuberculosis, $S(t)$, infected individuals who are infected with tuberculosis, $I(t)$ and recovered individuals are those infected but have recovered through treatment $R(t)$.
- The recruitment rate into the susceptible population is denoted by Λ. The effective contact infection between the susceptible individuals and infected individuals is denoted by β_1.
- The natural mortality rate of a human is μ and disease induced death rate is denoted by d.
- The level of treatment is proportional to the number of infected individual modeled via the term $\alpha f(I)$ where α is the maximum level of treatment and $f(x)$ is an increasing function of $I(t)$ and $0 \leq f(I) \leq 1$.
- The recovery rate of the infected individuals is $\alpha f(I)$ and the rate individuals recovered loss immunity and become infected through contact with infected ones is denoted by β_2.

The following equations depict the various interactions between the compartments:

$$\begin{aligned}
\frac{dS}{dt} &= \Lambda - \beta_1 IS - \mu S,\\
\frac{dI}{dt} &= \beta_1 IS - (\mu + d)I - \alpha f(I(t-\tau))I + \beta_2 IR,\\
\frac{dR}{dt} &= \alpha f(I(t-\tau))I - \beta_2 IR - \mu R.
\end{aligned} \tag{14.1}$$

with initial conditions:

$$S(\theta) > 0,\ I(\theta) > 0,\ R(\theta) > 0,\ \theta \in [-\tau, 0]. \tag{14.2}$$

By the fundamental theory of functional differential equations [20], we know that there is a unique solution $(S(t), I(t), R(t))$ to system (14.1) with the initial conditions given in (14.2).

Some basic properties such as positive invariance and boundedness of the solutions are discussed through the following theorems.

Theorem 14.2.1 *All the solution of (14.1) with initial conditions (14.2) are positive.*

Proof The system (14.1) can be written as:

$$\frac{dX}{dt} = g(X, t, \tau) = \begin{pmatrix} g_1 \\ g_2 \\ g_3 \end{pmatrix} = \begin{pmatrix} \Lambda - \beta_1 IS - \mu S \\ \beta_1 IS - (\mu + d)I - \alpha f(I(t-\tau))I + \beta_2 IR \\ \alpha f(I(t-\tau))I - \beta_2 IR - \mu R \end{pmatrix} \tag{14.3}$$

where, $X(t) = (X_1(t), X_1(t), X_1(t))^T = (S(t), I(t), R(t))^T$.

It is easy to check in system (14.1) that whenever choosing $X(\theta) \in \mathbb{R}_+$ such that $S = 0, I = 0, R = 0$, then

$$g_i(X)|_{x_i=0, X\in R^3_+} \geq 0, \text{i} = 1, 2, 3.$$

with $x_1(t) = S(t), \quad x_2(t) = I(t), \quad x_3(t) = R(t)$.

Using the *Lemma 1* in [21], *Theorem* 1.1 in [20], we can conclude that any solution of (14.1) with $X(\theta) \in \mathbb{C}$, say $X(t) = X(t, X(\theta))$, is such that $X(\theta) \in \mathbb{R}^3_+$ for all $t \geq 0$. Hence the solution of the system of system (14.1) exist in the region $\mathbb{R}^3_+$ and all solutions remain nonnegative for all $t > 0$.

The following theorem characterize the boundedness of solutions of the model system (14.1).

Theorem 14.2.2 *All the solutions of (14.1), that are initiated from $\mathbb{R}^3_+$, will be confined in the region*

$$\Gamma = \left\{(S, I, R) \in \mathbb{R}^3_+ : 0 \leq S + I + R \leq \frac{\Lambda}{\mu}\right\}. \tag{14.4}$$

14.2.1 Existence of Steady States

The system (14.1) has two equilibria, namely:

(i) the disease-free equilibrium point $E_0(\frac{\Lambda}{\mu}, 0, 0)$
(ii) the endemic equilibrium point, $E_*(S^*, I^*, R^*)$ where

$$S^* = \frac{\Lambda}{\beta I^* + \mu}, \quad R^* = \frac{\alpha f(I^*)I^*}{\beta I^* + \mu} \tag{14.5}$$

and I^* is the positive root of

$$\Lambda\beta_1 I(\beta_2 I + \mu) - I[(\mu + d) + \alpha f(I)](\beta_1 I + \mu) \cdot (\beta_2 I + \mu) + \beta_2 \alpha f(I) I(\beta_1 I + \mu) = 0.$$

14.3 Stability of Equilibria

In this subsection, we determine the local stability of the endemic equilibrium by finding the eigenvalues of the Jacobian matrix. For this we need the Characteristic equation of Jacobian matrix.

14.3.1 Characteristic Equation

Linearizing the system (14.1) about any point $E(S, I, R)$, we get:

$$\frac{dX}{dt} = FX(t) + QX(t - \tau). \tag{14.6}$$

Here F, Q are 3×3 matrices, given as below:

$$F = [F_{ij}] = \begin{bmatrix} -\beta_1 I - \mu & -\beta_1 S & 0 \\ \beta_1 I & \beta_1 S - \mu - d - \alpha f(I) + \beta_2 R & \beta_2 I \\ 0 & \alpha f(I) - \beta_2 R & -\beta_2 I - \mu \end{bmatrix}.$$

$$Q = [Q_{ij}] = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -\alpha f'(I) I & 0 \\ 0 & \alpha f'(I) I & 0 \end{bmatrix}, \tag{14.7}$$

The characteristic equation of the delay system (14.1) is given by,

$$H(\rho) =\mid \rho I - P - e^{-\rho\tau} Q \mid= 0.$$

This gives,

$$H(\rho, \tau) = \rho^3 + m_1\rho^2 + m_2\rho + m_3 + e^{-\rho\tau}[m_4\rho^2 + m_5\rho + m_6] = 0. \tag{14.8}$$

Here,

$$\begin{aligned} &m_1 = -[F_{11} + F_{22} + F_{33}], \quad m_2 = F_{22}F_{33} + F_{11}F_{33} + F_{11}F_{22} - F_{21}F_{12}, \\ &m_3 = F_{11}F_{22}F_{33} + F_{12}F_{21}F_{33}, \quad m_4 = -Q_{22}, \quad m_5 = Q_{22}F_{33} + F_{11}Q_{22} \\ &m_6 = F_{11}Q_{22}F_{33}. \end{aligned}$$

For $\tau = 0$, (14.8) becomes

$$H(\rho, 0) = \rho^3 + (m_1 + m_4)\rho^2 + (m_2 + m_5)\rho + (m_3 + a_6) = 0. \tag{14.9}$$

Let

$$\sigma_1 = (m_1 + m_4), \quad \sigma_2 = (m_2 + m_5), \quad \sigma_3 = (m_3 + a_6), \tag{14.10}$$

then the following results can be obtained.

Theorem 14.3.1 *Let*

$$R_0 = \frac{\beta_1 \Lambda}{\mu(d+\mu+\alpha)}, \tag{14.11}$$

then disease-free equilibrium E_0 is locally asymptotically stable when $R_0 < 1$ and unstable otherwise. Transcritical bifurcation occurs at $R_0 = 1$.

Remark Infection cannot spread across the population if $R_0 < 1$ and the disease-free state is locally asymptotically stable. Each infected human produces more than one secondary infected individuals if $R_0 > 1$ and as a result of this, the invasion is always possible which makes the disease-free state unstable.

Theorem 14.3.2 *The coexistence equilibrium point $E^*(S^*, I^*, R^*)$ is stable if the following conditions are satisfied:*

$$\sigma_1 > 0,\ \sigma_3 > 0,\ \sigma_1\sigma_2 - \sigma_3 > 0$$

where, $\sigma_i, i = 1, 2, 3$ are defined in (14.10).

Proof The characteristic for $\tau = 0$ becomes

$$\rho^3 + \sigma_1\rho^2 + \sigma_2\rho + \sigma_3 = 0, \tag{14.12}$$

Thus, if the conditions stated in the theorem hold then using the Routh–Hurwitz criteria, the coexistence equilibrium E^* of the system (14.1) is locally asymptotically stable for $\tau = 0$.

14.3.2 Length of Delay and Hopf Bifurcation

For $\tau > 0$, the characteristic equation is a transcendental equation in ρ. It is known that E^* is locally asymptotically stable if all the roots of the corresponding characteristic equation have negative real parts and unstable if purely imaginary roots exist.

Assuming $\rho = i\omega$ as a root of the Eq. (14.8) and separating the real and imaginary parts, we obtain

$$\begin{aligned} m_3 - \omega^2 &= m_4\omega^2 \cos\omega\tau - m_5\omega\sin\omega\tau - m_6\cos\omega\tau, \\ -\omega^3 + m_3\omega &= -m_4\omega^2\sin\omega\tau - m_5\omega\cos\omega\tau + m_6\sin\omega\tau. \end{aligned} \tag{14.13}$$

It follows from (14.13) that

$$\omega^6 + \delta_1\omega^4 + \delta_2\omega^2 + \delta_3 = 0, \tag{14.14}$$

where

$$\delta_1 = m_1^2 - 2m_2 - m_4^2,\ \delta_2 = m_2^2 + 2m_4m_6 - 2m_1m_3 - m_5^2,\ \delta_3 = m_3^2 - m_6^2.$$

Let $\omega^2 = a$, then the Eq. (14.14) becomes

$$F(a) = a^3 + \delta_1 a^2 + \delta_2 a + \delta_3 = 0, \tag{14.15}$$

If $\delta_1 > 0,\ \delta_2 > 0$ and $\delta_3 > 0$, we can claim there exists no w such that iw is the eigenvalue of the characteristic Eq. (14.8). Therefore, the real parts of all the eigenvalues of (14.8) are negative for all $\tau \geq 0$ and thud system is stable for all $\tau \geq 0$. Therefore, we have the following theorem.

Theorem 14.3.3 *If the following conditions:* $\sigma_1 > 0,\ \sigma_3 > 0,\quad \sigma_1\sigma_2 - \sigma_3 > 0$ *and* $\delta_1 \geq 0,\ \delta_3 \geq 0,\ \delta_2 > 0$ *are satisfied then the infected steady state* E^* *is asymptotically stable for all* $\tau \geq 0$*.*

Now, if $\delta_3 < 0$, then there exists a positive root a_0 of (14.15) for which the characteristic equation has a pair of purely imaginary roots $\pm i\omega_0$. Then Eq. (14.14) possesses a pair of purely imaginary roots $\pm i\omega_0$.

Now, suppose that (14.15) has positive roots and are denoted by a_i, $i = 1,\ 2,\ 3$. Then (14.14) has three positive roots, $\omega_i = \sqrt{a_i}$, $i = 1,\ 2,\ 3$. From Eq. (14.13), we obtain the value of τ as

$$\tau_k^n = \frac{1}{\omega_0}\left(\cos^{-1}\frac{(m_5 - m_4m_1)\omega_0^4 + (m_3m_4 + m_1m_6 - m_2m_5)\omega_0^2 - m_6m_3}{m_4^2\omega_0^4 + (m_5^2 - 2m_6m_4)\omega_0^2 + m_6^2} + 2n\pi\right), \tag{14.16}$$

for $k = 0, 1, 2, 3$ and $n = 1, 2, \ldots$. Thus $\pm i\omega_k$ is a pair of purely imaginary roots of (14.8).

Let

$$\tau_0 = \tau_{k_0}^{n_0} = \min_{n\geq 0, 1\leq k\leq 3}\{\tau_k^n\},\ \omega_0 = \omega_{k_0}.$$

The above results can be given in the following theorem (Fig. 14.1).

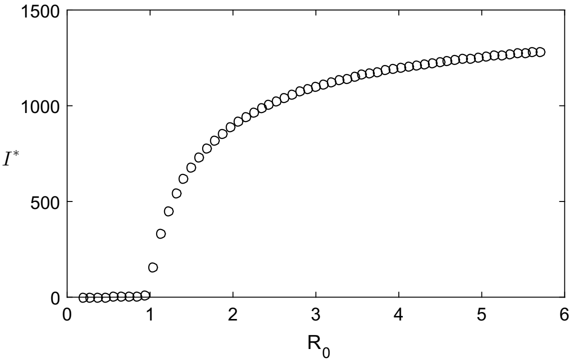


Fig. 14.1 Forward transcritical bifurcation: Steady state values of infected human is plotted with parameters values as $\Lambda = 40$, $d = 0.012,\ \mu = 0.01$, $\beta_2 = 0.00002, \alpha = 0.2$, $\beta_1 \in [0.000002,\ 0.00006]$

Theorem 14.3.4 *Suppose that the interior equilibrium point E^* exists and is locally asymptotically stable for $\tau = 0$ and if either, $\delta_3 < 0$ or $\delta_3 \geq 0$ and $\delta_2 < 0$, then E^* is asymptotically stable when $\tau < \tau_0$ and unstable when $\tau > \tau_0$. When $\tau = \tau_0$, Hopf bifurcation occurs provided the following transversality condition is satisfied,*

$$3\omega_0^4 + 2\delta_1\omega_0^2 + \delta_2 > 0.$$

Proof We only need to prove the transversality condition only. Denoting $\rho = \rho(\tau)$, differentiating (14.8), we have

$$\left(\frac{\mathrm{d}\rho(\tau)}{\mathrm{d}\tau}\right)^{-1} = -\frac{3\rho^2 + 2m_1\rho + m_2}{\rho(\rho^3 + m_1\rho^2 + m_2\rho + m_3)} + \frac{2m_4\rho + m_5}{\rho(m_4\rho^2 + m_5\rho + m_6)} - \frac{\tau}{\rho},$$

which leads to

$$\begin{aligned}
&\operatorname{sign}\left\{\operatorname{Re}\left(\frac{\mathrm{d}\rho}{\mathrm{d}\tau}\right)_{\tau=\tau_j}\right\} \\
&= \operatorname{sign}\left\{\operatorname{Re}\left(\frac{\mathrm{d}\rho}{\mathrm{d}\tau}\right)^{-1}_{\tau=\tau_j}\right\} \\
&= \operatorname{sign}\left\{3\omega_0^4 + (2m_1^2 - 4m_2 - 2m_4^2)\omega_0^2 + m_2^2 + 2m_4m_5 - 2m_1m_3 - m_5^2\right\} \\
&= \operatorname{sign}\left\{3\omega_0^4 + 2\delta_1\omega_0^2 + \delta_2\right\}.
\end{aligned} \tag{14.17}$$

Thus, the transversality condition holds and Hopf bifurcation occurs at $\tau = \tau_0$. Hence the theorem.

14.4 Numerical Simulations

The numerical simulations of the system (14.1) are carried out in this section to explore the dynamics of the model. The effect of increasing delay of treatment is critically investigated in this section to confirm some of the theoretical findings already established in previous sections. We have taken $f(I) = I/(1 + I)$ for numerical simulations.

In Figure 14.1, we have seen that the disease will persist if $R_0 > 1$, system will be disease free for $R_0 < 1$, and Transcritcal bifurcation will occur at $R_0 = 1$.

Figure 14.2 shows that the infected individual decreases for increasing values of α (the level of treatment). This may suggest that individuals become aware of TB and adopt strategies to prevent infection in the population.

The endemic equilibrium of the system (14.1) without delay ($\tau = 0$) is asymptotically stable if the given parameter values as the conditions of Theorem 14.3.2 are satisfied. The system populations initially exhibit a small amount of oscillation become stable when the delay is smaller than its critical value τ^* (Fig. 14.3). This

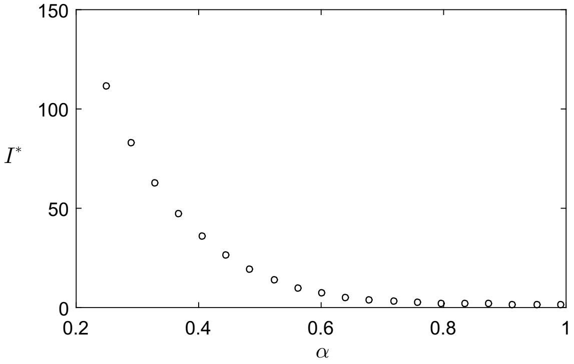


Fig. 14.2 Effect of 'level of treatment' on the system populations. Steady state value of infected human (I^*) is plotted for a range of values of α. The parameters values are: $\Lambda = 40$, $\beta_1 = 0.00025$, $d = 0.012$, $\mu = 0.01$, $\beta_2 = 0.00002$ and $\alpha \in (0.2, 1)$

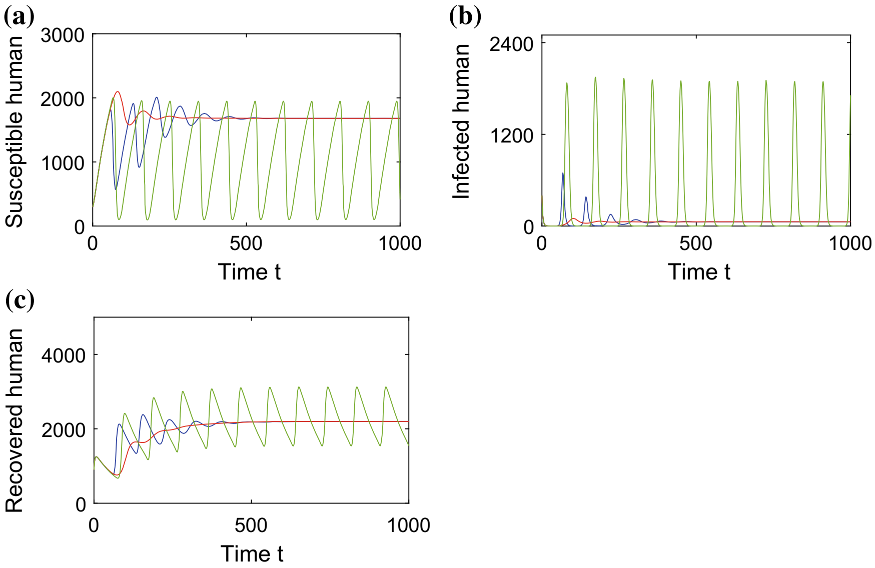


Fig. 14.3 Numerical solution of the system for different values of τ. Parameter values used for the simulation are the same as in Fig. 14.2

suggests that making a prediction with regard to the epidemic size may not be too difficult.

Figure 14.3 also depicts periodic oscillation for ($\tau > \tau^* \approx 32.3$) i.e. endemic equilibrium is unstable for ($\tau > \tau^*$) (Theorem 14.3.4). This implies that in some instances the number of infective will be rising and other time may be in decreasing which will result in the difficulty of estimating the actual size of the epidemic. This suggests that all state variables bifurcate into periodic solution at for ($\tau = \tau^*$) (Fig. 14.4).

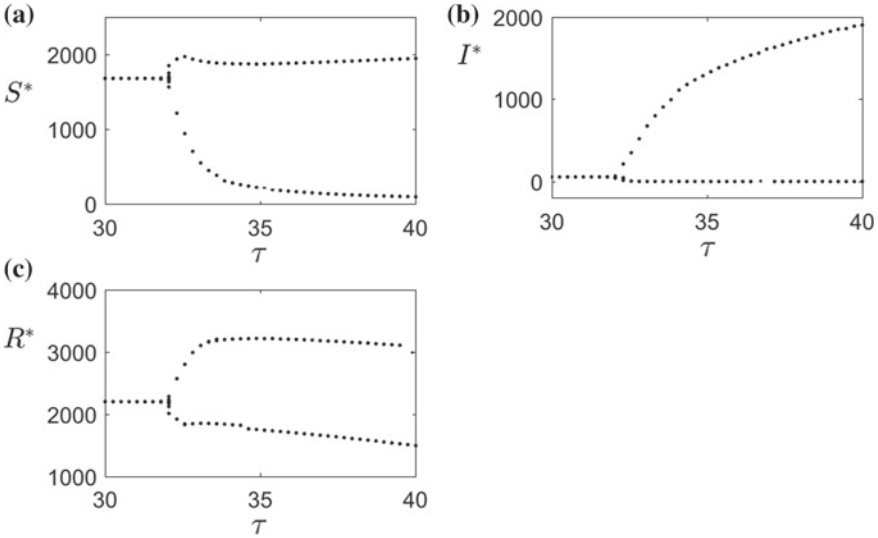


Fig. 14.4 Hopf bifurcation solution of the system for $\tau = 13$. The parameters are same as Fig. 14.3

14.5 Discussion and Conclusion

In this chapter, a mathematical is developed for the dynamics of tuberculosis epidemic using a set of delay differential equations with an aim to examine the effect of time delay in controlling the disease. Recovery is assumed to be proportional to the level of treatment/control of the disease also assumed as a delayed process.

The model system has two equilibria: the disease-free equilibrium and the endemic equilibrium. Disease-free equilibrium is stability for $R_0 < 1$ and becomes unstable if $R_0 > 1$. Endemic equilibrium become feasible for $R_0 > 1$ and undergoes Hopf bifurcation if the delay parameter crosses a threshold value, τ^*.

In conclusion, the persistence of long period of oscillation if the lag period is increased has a serious negative effect on controlling TB epidemic in the communities.

References

1. L. Ramakrishnan, Revisiting the role of the granuloma in tuberculosis. Nat. Rev. Immunol. **12**, 352–366 (2012)
2. World Health Organization, Global tuberculosis report 2016. Technical Report (World Health Organization, Geneva, Switzerland, 2016)
3. C. Herrera, S. Lima, R. Munoz, G. Ramos, A. Rodriguez, C. Salzberg, A model describing the response of immune system to Mycobacterium tuberculosis. Department of Biometrics Technical Report Series: BU-1364-M (Cornell University, Biometrics Department, 1996)

4. Y. Zhao, M. Li, S. Yuan, Analysis of transmission and control of tuberculosis in mainland China, 2005–2016, based on the age-structure mathematical model. Int. J. Environ. Res. Public Health **14**, 1192 (2017)
5. James M. Trauer, Justin T. Denholm, Emma S. McBryde, Construction of a mathematical model for tuberculosis transmission in highly endemic regions of the Asia-Pacific. J. Theor. Biol. **358**, 74–84 (2014)
6. P.W. Uys, R.M. Warren, P.D. van Helden, A threshold value for the time delay to TB diagnosis. PLoS ONE **2**(8), ID e757 (2007)
7. F.A. Basir, Dynamics of infectious diseases with media coverage and two time delay. Math. Model. Comput. Simul. **10**(6), 770–783 (2018)
8. F.A. Basir, S. Ray, E. Venturino, Role of media coverage and delay in controlling infectious diseases: a mathematical model. Appl. Math. Comput. **337**, 372–385 (2018)
9. C.T. Sreeramareddy, K.V. Panduru, J. Menten, J. Van den Ende, Time delays in diagnosis of pulmonary tuberculosis: a asystematic review of literature. BMC Infect. Dis. **9**, article 91 (2009)
10. S.M. Blower, C.L. Daley, Problems and solutions for the stop tb partnership. Lancet Infect. Dis. **2**, 374–376 (2002)
11. M.G. Houben, C.Y. Wu, A.S. Rhines, J.T. Denholm, G.B. Gomez, P. Hippner, Feasibility of achieving the 2025 WHO global tuberculosis target in South Africa, China, and India: a combined analysis of mathematical models. Lancet Glob. Health **4**, 806–815 (2016)
12. M.T. Li, G.Q. Sun, Y.F. Wu, J. Zhang, Z. Jin, Transmission dynamics of a multi-group brucellosis model with mixed cross infection in public farm. Appl. Math. Comput. **237**, 582–594 (2014)
13. H.T. Waaler, A. Gese, S. Anderson, The use of mathematical models in the study of the epidemiology of tuberculosis. Am. J. Public Health **52**, 1002–1013 (1962)
14. S.M. Blower, A.R. McLean, T.C. Porco, The intrinsic transmission dynamics of tuberculosis epidemics. Nat. Med. **8**, 815–821 (1995)
15. Z.W. Jia, G.Y. Tang, Z. Jin, Modeling the impact of immigration on the epidemiology of tuberculosis. Theor. Popul. Biol. **73**, 437–448 (2008)
16. C.P. Bhunu, W. Garira, Z. Mukandavire, M. Zimba, Tuberculosis transmission model with chemoprophylaxis and treatment. Bull. Math. Biol. **70**, 1163–1191 (2008)
17. T. Cohen, M. Murray, Modeling epidemics of multi drug resistant M. tuberculosis of heterogeneous fitness. Nat. Med. **10**(10), 1117–1121 (2004)
18. R.M.G.J. Houben, D.W. Dowdy, A. Vassall, T. Cohen, M.P. Nicol, R.M. Granich, J.E. Shea, How can mathematical models advance tuberculosis control in high HIV prevalence settings? Int. J. Tuberc. Lung Dis. **18**(5), 509–514 (2014)
19. J.S. Cristiana, H. Maurer, D.F.M. Torres, Optimal control of a Tuberculosis model with state and control delays **14**(1), 321–337 (2017)
20. J. Hale, *Theory of Functional Differential Equations* (Springer, Berlin, 1977)
21. X. Yang, L. Chen, J. Chen, Permanence and positive periodic solution for the single species nonautonomus delay diffusive model. Comput. Math. Appl. **32**, 109–116 (1996)

Chapter 15
Treatment of Psoriasis by Interleukin-10 Through Impulsive Control Strategy: A Mathematical Study

Amit Kumar Roy and Priti Kumar Roy

Abstract Psoriasis is characterized by anomalous growth of keratinocytes (skin cells), which occurs due to abrupt signaling within immune cells and cytokines. The most significant immune cells, T cells go through differentiation with interaction of dendritic cells (DCs) to produce Type 1 T helper cell (Th_1) and Type 2 T helper cell (Th_2) subtypes. In psoriatic progression dynamics, the inflammation effect of Th_1 mediated cytokines (pro-inflammatory) are responsible for the abnormal growth of keratinocytes. In this measure, the effect of anti-inflammatory cytokines secreted by Th_2 subtype partially downregulate the growth of epidermal cell. In this research article, we have constructed a five-dimensional mathematical model involving T cells, dendritic cells, Th_1, Th_2, and keratinocyte cell populations for better understanding the development of psoriatic lesions. Moreover, we have evaluated the role of Th_1, Th_2, and interplay of various cytokine networks in Psoriasis through a set of nonlinear differential equations. Our analytical study reveals the preconditions for disease persistence and also validates the stability criteria of endemic equilibrium for the disease. Furthermore, we have used one-dimensional impulsive differential equation to examine the effects of different levels of biologic (Interleukin-10) for different dosing intervals in keratinocytes cell population. We have also examined the qualitative behavior of keratinocyte by considering two different values of the parameter corresponding to the reduction of keratinocyte due to the impact of drug (IL-10). We have also found the perfect dosing intervals of biologic (Interleukin-10) that could tolerate the keratinocytes at the desired level. Finally, our analytical and numerical computations reveal that the use of IL-10 through impulsive way is proven better treatment compared with other trivial therapeutic policies for psoriatic patients.

Keywords Th_1 · Th_2 · Cytokines · Keratinocytes · Biologic · Impulsive approach

A. K. Roy · P. K. Roy (✉)
Department of Mathematics, Centre for Mathematical Biology and Ecology,
Jadavpur University, Kolkata 700032, India
e-mail: pritiju@gmail.com

A. K. Roy
e-mail: amit.jumath@gmail.com

P. Manchanda et al. (eds.), *Mathematical Modelling, Optimization, Analytic and Numerical Solutions*, Industrial and Applied Mathematics,
https://doi.org/10.1007/978-981-15-0928-5_15

15.1 Introduction

Psoriasis is an autoimmune disorder which persists with hyper-proliferation of the dermal cells with multifarious dermatological symptoms. Recently, it is a serious dermatological disorder which has a deep disagreeable effect on patient's social, mental, and physical happiness globally. Scientists now believe that at least 10 percent of the general population inherits the genes that create a predisposition to psoriasis development [1]. Certain environmental factors may also trigger the psoriasis onset, causing the disease to become active. These environmental triggers vary from person to person and sometimes it becomes a privilege for a patient. However, etiology of psoriasis remains unclear till date, but substantial evidence for recurring immune imbalance indicates psoriasis development.

Human immune cells, T cell and dendritic cell (DCs) take vital accountability for creating the hyper-proliferation of keratinocyte which is the causal fact of the disease psoriasis. T cell is a kind of lymphocyte (a subtype of white blood cell) that plays a central role in cell-mediated immunity, arise in the bone marrow and migrate to thymus gland to mature [2]. Dendritic cells are particular antigen-presenting cells and important intermediaries of immunity originated from monocyte and dendritic cell progenitor in bone marrow [3]. Naive T cells (T cells that have not yet encountered antigen) undergo a differentiation with the interaction of DCs to produce Th_1 and Th_2 subtypes under certain cytokine environments [2, 4]. If the naive T cells interact with DCs in Interleukin 12 (IL-12) dominated region, it results in T cells to differentiate into a large amount of Th_1 cells that secrete pro-inflammatory cytokines, viz., Interferon-gamma (IFN-γ), Transforming growth factor-beta (TGF-β), and Tumor necrosis factor-alpha (TNF-α). At the time of naive T cells differentiation, if the periphery is Interleukin 4 (IL-4) conquered, it results in the enrichment of the density of Th_2 cells, which secrets anti-inflammatory cytokines family, viz., Interleukin 4 (IL-4) and Interleukin 10 (IL-10). In the presence of pro-inflammatory cytokines, the proliferation of Th_2 cells is downregulated and on the other hand, the proliferation of Th_1 cells is upregulated [4, 5]. Nowadays, psoriasis is treated as Th_1 cells mediated skin disorder, characterized by the overproduction of IFN-γ, TNF-α, and TGF-β [6]. Keratinocyte is a principal epidermal cell, expected as major target tissue of TGF-β and it differentiates by the influence of TGF-β signaling [7]. IFN-γ is a multifunctional and immunomodulatory cytokine, which activate keratinocyte by the possess of biochemical requirements [8]. TNF-α alone is not capable to provoke immunologic reaction but in combination with IL-17A, IL-17C, and other cytokines, it forms strong synergies [9, 10]. Under this strong synergism, the expression of IL-17R is increased by keratinocyte, which gives the significant response in hyper-proliferation of keratinocyte [10, 11]. However, there is increasing evidence that IL-4 gives pleiotropic effects on the immune system and directly suppress Th_1 mediated inflammation on keratinocyte. It is conveyed that IL-10 cytokine stimulates the enlargement of anti-inflammatory cytokines by inhibiting the IFN-γ production. Although anti-inflammatory cytokines secreted from Th_2 cells negatively regulate the keratinocytes population, yet overexpression of various pro-inflammatory cytokines play a central role in the disease progression [12].

Clinically, it is accepted that the turnover time for the epidermis in psoriatic case is 7 days (normal turnover time for the keratin layer is 2 days) and also a doubling of the proliferative cell population in psoriasis is from 27,000 to 52,000 cells/mm [13, 14]. Microarray analysis performed by Johnson-Huang et al. reported that IFN-γ and TNF-α are the key regulatory cytokines in psoriasis development [15, 16]. Mussi et al. already developed an experimental study on the level of serum TNF-α which is significantly high for psoriasis using enzyme-linked immunosorbent assay (ELISA) kits [17]. Using ELISA method, Baran et al. specified that the concentrations of TGF-β were dramatically increased for patients with psoriasis [18]. Promising new therapies are mainly pro-inflammatory cytokines inhibitor implicated in psoriasis [19]. Through clinical trial program, many biological agents (Alefacept, Efalizumab, Etanercept, Infliximab, and Adalimumab) are globally accepted as a safe and effective drug for patients with psoriasis [19]. Recently, many biological and clinical experimenters suggested that injection of anti-inflammatory cytokines (IL-4 and IL-10) may be a successful treatment for psoriasis [20, 21]. They have also suggested that receiving of IL-10 (20 ug/kg of body weight, 3 times per week) may reduce about 90% of initial psoriasis area within 50 days [20, 22–24].

During the last decade, some mathematical models are being developed using Ordinary Differential Equation (ODE) as well as Partial Differential Equation (PDE) on the disease dynamics of psoriasis introducing different cell population of T cells, dendritic cells and keratinocytes along with cytokine influence [25–28]. Roy et al. also studied the mathematical model on psoriasis based on Fractional Order Differential Equation (FDE) and they discussed about the control of the disease using the negative feedback loop [4, 29–31]. In the previous study of psoriasis in the mathematical aspect, it was considered that T Cells and dendritic Cells play a vital role in the disease dynamics and all disease control approaches were based on some hypothetical assumptions. In this research article, we have emphasized the effect of Th$_1$ and Th$_2$ on the hyper-proliferation of keratinocytes through pro-inflammatory and anti-inflammatory cytokines network. We have also studied our proposed mathematical system introducing the therapeutic agent (IL-10). In the disease control strategy, we have considered the one-dimensional growth equation of keratinocyte which represents the maximum density of epidermal cell, present during the disease progression. Furthermore, we have studied the keratinocyte proliferation under IL-10 therapy using modified impulsive method. Our analytical and numerical analysis reveals that using IL-10 in perfect dose with some fixed time interval may reduce more than 90% of psoriatic plaque more quickly.

The article begins with an overall introductory section; then, we have formulated the mathematical model based on suitable assumptions and the basic property of formulated model has been also discussed in Sect. 15.2. In Sect. 15.3, we have studied the model system analytically which explores the existence and stability criteria of endemic equilibria. In Sect. 15.4, we have investigated the Keratinocyte density using impulse therapeutic approach under fixed IL-10 injecting process. Section 15.5 presents the numerical simulation of system dynamics for without and with therapy.

In Sect. 15.6, we have discussed about the consequences of outcomes which we have found out in different sections and we have drawn the conclusion of this research work, in Sect. 15.7.

15.2 The Model

15.2.1 Model Formulation with Suitable Assumptions

We develop a mathematical model of psoriasis by introducing different cells to reflect the cell-biological relationships in expressing the disease. In order to develop the mathematical model we have considered the schematic diagram (Fig. 15.1). Here, $T(t)$, $D(t)$, $T_1(t)$, $T_2(t)$, and $K(t)$ represent the densities of naive T cells, dendritic cells, Th$_1$ cells, Th$_2$ cells, and epidermal keratinocytes, respectively, at any time t. The following assumptions are considered to develop our mathematical model.

(A) Naive T cells and DCs strictly originated from bone marrow and got mature at thymus. The accumulation rate of naive T cells and DCs in the area proximity at the suitable management are assumed a_L and a_D, respectively. We assume that the proliferation of naive T cells is logistic, where ρ indicates the maximum proliferation

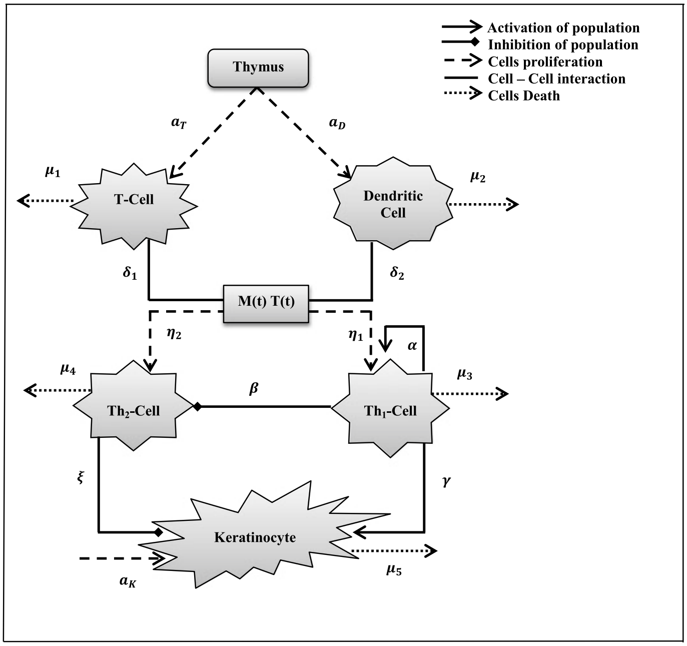


Fig. 15.1 Schematic diagram of the interactions between the components of the model

rate and T$_{max}$ stands for the maximum stage of T cell proliferation [32]. T cells and DCs are two different types of immune cells with dissimilar features in the human immune system. Therefore, we have considered different activation rates of T cells and DCs for justification of our mathematical model. The rate at which T cells bind with dendritic cells is denoted as δ_1. On the other hand, δ_2 is the rate of stimulation of dendritic cells with T cells. In mathematical perceptive, the interaction obey the *Law of Mass Action*. Under mixing homogeneity, the combined interaction of naive T cells and DCs contributes to the subtype of T cells (Th$_1$ and Th$_2$). The per capita removal rates of T cells and DCs are assumed as μ_1 and μ_2, respectively, throughout normal sequence.

(B) Th$_1$ ($T_1(t)$) and Th$_2$ ($T_2(t)$) cells are furnished due to cytokine conformational changes of naive T cells after the interaction with dendritic cells (DCs). We assume that η_1 and η_2 are the rates of accumulation of Th$_1$ and Th$_2$, respectively. Note that the summation of activation rates of T cells and DCs is greater than the total rates of accumulation of Th$_1$ and Th$_2$, which is expressed as ($\delta_1 + \delta_2 \geq \eta_1 + \eta_2$). Th$_1$ cells proliferation is upregulated in the presence of pro-inflammatory cytokine released by itself. Here, we consider that α is the regulation rate of Th$_1$ cells in the presence of pro-inflammatory cytokine. Again, we consider that at a rate β, Th$_2$ cell is downregulated due to the effect of pro-inflammatory cytokine released by Th$_1$. Natural death rates of Th$_1$ and Th$_2$ cells are noted by μ_3 and μ_4, respectively, due to normal cell death.

(C) Psoriasis is characterized by hyper-proliferation of keratinocytes due to over expression of pro-inflammatory cytokines released by Th$_1$ cells. Keratinocytes' proliferation may be reduced to a certain level by the effect of anti-inflammatory cytokines secreted by Th$_2$ cells. In the keratinocyte density, it is to be noted that release factors of Th$_1$ cells increase the keratinocytes' proliferation at a rate γ. We also assume that ξ be the anti-inflammatory cytokines' effect on keratinocytes by Th$_2$ cells. Here, we consider a_K is the constant growth of keratinocytes due to cell migration from the dermal layer to epidermal layer and the removal rate of keratinocytes is considered as μ_5.

Assembling the above three assumptions (A, B, C), we can formulate the following mathematical model:

$$
\begin{aligned}
\frac{dT}{dt} &= a_T + \rho T\left(1 - \frac{T}{T_{max}}\right) - \delta_1 TD - \mu_1 T,\\
\frac{dD}{dt} &= a_D - \delta_2 TD - \mu_2 D,\\
\frac{dT_1}{dt} &= \eta_1 TD + \alpha T_1 - \mu_3 T_1,\\
\frac{dT_2}{dt} &= \eta_2 TD - \beta T_1 T_2 - \mu_4 T_2,\\
\frac{dK}{dt} &= a_K + \gamma T_1 K - \xi T_2 K - \mu_5 K,
\end{aligned}
\qquad (15.1)
$$

where $T(0) > 0$, $D(0) > 0$, $T_1(0) > 0$, $T_2(0) > 0$, and $K(0) > 0$ are initial conditions.

15.2.2 Model Properties

The right-hand sides of system (15.1) are smooth and nonlinear functions of the variable T, D, T_1, T_2, and K and also the parameters are always nonnegative. Henceforth, the system dynamics is assuredly bounded in the positive octant and the considered cells' concentration is less than a pre-assumed quantity. In the following theorem, we wish to clarify that the solution of the dynamical system is bounded.

Theorem 15.1 *The solutions of system (15.1), which satisfy the initial conditions, i.e., $T(t) > 0, D(t) > 0, T_1(t) > 0, T_2(t) > 0, and K(t) > 0$ for all $t > 0$. The region $\Omega \subset \mathscr{R}^5_+$ is positively invariant and attracting with respect to system (15.1). Where*

$$\Omega = \Bigg\{(T, D, T_1, T_2, K) \in \mathscr{R}^5_+ : 0 \le T \le \frac{a_T}{\mu_1 - \rho}, 0 \le D \le \frac{a_D}{\mu_2},$$
$$0 \le T_1 \le \frac{\eta_1 a_T a_D}{(\mu_1 - \rho)\mu_2(\mu_3 - \alpha)}, 0 \le T_2 \le \frac{\eta_2 a_T a_D(\mu_3 - \alpha)}{\beta\eta_1 a_T a_D + \mu_2\mu_4(\mu_1 - \rho)(\mu_3 - \alpha)},$$
$$0 \le K \le \frac{a_K(\mu_1 - \rho)\mu_2(\mu_3 - \alpha)}{(\mu_1 - \rho)\mu_2\mu_5(\mu_3 - \alpha) - \gamma\eta_1 a_T a_D}\Bigg\}.$$

Proof From the first equation of system (15.1), we can predict the upper threshold of T cells exist in psoriatic patients.

$$\frac{dT}{dt} = a_T + \rho T\left(1 - \frac{T}{T_{max}}\right) - \delta_1 T D - \mu_1 T,$$
$$\le a_T - (\mu_1 - \rho)T. \tag{15.2}$$

Now, by solving the above inequality (15.2) for a long time interval and for positive $(\mu_1 - \rho)$, we get $T(t) \le \frac{a_T}{\mu_1 - \rho}$, the maximum density of T cells present in the case of psoriasis. In a similar manner, we also determine from the second equation of the system (15.1), density of dendritic cell at any time cannot exceed the ratio of it's constant accumulation and natural death, i.e., $D(t) \le \frac{a_D}{\mu_2}$.

Now considering the third equation of model (15.1), we get the following:

$$\frac{dT_1}{dt} = \eta_1 T D + \alpha T_1 - \mu_3 T_1.$$

We get the following inequation by putting the maximum density of T cell and DCs.

$$\frac{dT_1}{dt} \leq \eta_1 \frac{a_T a_D}{(\mu_1 - \rho)\mu_2} - (\mu_3 - \alpha)T_1, \tag{15.3}$$

solving the above inequality (15.3), we get

$$T_1(t) \leq \frac{\eta_1 a_T a_D}{(\mu_1 - \rho)\mu_2(\mu_3 - \alpha)} + \left(T_1(0) - \frac{\eta_1 a_T a_D}{(\mu_1 - \rho)\mu_2(\mu_3 - \alpha)}\right)e^{-(\mu_3 - \alpha)t}.$$

For the positive value of $(\mu_3 - \alpha)$ and for long time period, we get

$$T_1(t) \leq \frac{\eta_1 a_T a_D}{(\mu_1 - \rho)\mu_2(\mu_3 - \alpha)}. \tag{15.4}$$

Similarly, using the maximum value of density of T cell, Dcs, and Th$_1$, we also get the upper threshold of Th$_2$ density from the fourth equation of our formulated model (15.1)

$$T_2(t) \leq \frac{\eta_2 a_T a_D(\mu_3 - \alpha)}{\beta\eta_1 a_T a_D + \mu_2\mu_4(\mu_1 - \rho)(\mu_3 - \alpha)}. \tag{15.5}$$

From the last equation of system (15.1)

$$\frac{dK}{dt} = a_K + \gamma T_1 K - \xi T_2 K - \mu_5 K. \tag{15.6}$$

It is to be mentioned here that the reducing effect of anti-inflammatory cytokine over the keratinocytes by Th$_2$ is low in amount for the case of psoriatic patients. So neglecting the negative effect of Th$_2$ on keratinocyte and also considering the maximum density level of Th$_1$ cell, we get the following inequation from the above Eq. (15.6)

$$\frac{dK}{dt} \leq a_K - \left(\mu_5 - \frac{\gamma\eta_1 a_T a_D}{(\mu_1 - \rho)\mu_2(\mu_3 - \alpha)}\right)K, \tag{15.7}$$

in order to find the maximum value of Keratinocyte's density in psoriatic patient we solve the above inequation (15.7) for long time period and by considering the positive value of $\left(\mu_5 - \frac{\gamma\eta_1 a_T a_D}{(\mu_1-\rho)\mu_2(\mu_3-\alpha)}\right)$, we get

$$K(t) \leq \frac{a_K(\mu_1 - \rho)\mu_2(\mu_3 - \alpha)}{(\mu_1 - \rho)\mu_2\mu_5(\mu_3 - \alpha) - \gamma\eta_1 a_T a_D}. \tag{15.8}$$

From the above discussion and the inequations (15.4), (15.5), (15.8), we can conclude that all cell populations are bounded in positive octant and $\Omega \subset \mathscr{R}^5_+$ is positively invariant and attracting with respect to the system (15.1).

15.3 Equilibrium Analysis

15.3.1 Existence Condition

In this system, we consider an interior equilibrium point $E(T^*, D^*, T_1^*, T_2^*, K^*)$, where the disease persists in the population. The interior equilibrium point is obtained by setting equations of the system to zero. We then solve for state variables in terms of T^* and obtain the following:

$$\begin{aligned}
D^* &= \frac{a_D}{\delta_2 T^* + \mu_2};\\
T_1^* &= \frac{\eta_1 a_D T^*}{\mathscr{A}};\\
T_2^* &= \frac{\eta_2 a_D T^*(\mu_3 - \alpha)}{\beta\eta_1 a_D T^* + \mu_4 \mathscr{A}};\\
K^* &= \frac{a_K \mathscr{A}[\beta\eta_1 a_D T^* + \mu_4 \mathscr{A}]}{[\mu_5 \beta\eta_1 a_D T^* + \mu_4\mu_5 \mathscr{A} + \xi\eta_2 T^* a_D(\mu_3 - \alpha)]\mathscr{A} - \gamma\eta_1 T^* a_D[\beta\eta_1 T^* a_D + \mu_4 \mathscr{A}]};
\end{aligned} \tag{15.9}$$

where $\mathscr{A} = (\delta_2 T^* + \mu_2)(\mu_3 - \alpha)$ and T^* is the positive root of the following cubic equation:

$$a_3(T^*)^3 - a_2(T^*)^2 - a_1 T^* - a_0 = 0, \tag{15.10}$$

where

$$a_3 = \frac{\rho\delta_2}{T_{max}},\ a_2 = \left(\rho\delta_2 - \frac{\rho\mu_2}{T_{max}}\right),\ a_1 = (a_T\delta_1 + \rho\mu_2 - \delta_1 a_D - \mu_1),\ a_0 = a_T\mu_2.$$

Since a_0 is always positive, there exists at least one positive root of Eq. (15.10). From the above mathematical expression, we can conclude the existence condition of the endemic equilibrium (E) by the following proposition

Proposition 15.1 *At least an endemic equilibrium (E) of our formulated mathematical model (15.1) exists, if the positive root (T^*) of the Eq. (15.10) satisfies the inequality $\mathscr{A} > 0$, i.e., $\mu_3 > \alpha$.*

15.3.2 Stability Criteria

The Jacobian matrix for the endemic equilibrium of model system (15.1) is given by

$$J(T^*, D^*, T_1^*, T_2^*, K^*) = \begin{bmatrix} -\frac{a_T}{T^*} - \frac{\rho T^*}{T_{max}} & -\delta_1 T^* & 0 & 0 & 0 \\ -\delta_2 D^* & -\frac{a_D}{D^*} & 0 & 0 & 0 \\ \eta_1 D^* & \eta_1 T^* & \alpha - \mu_3 & 0 & 0 \\ \eta_2 D^* & \eta_2 T^* & -\beta T_2^* & -\frac{\eta_2 T^* D^*}{T_2^*} & 0 \\ 0 & 0 & \gamma K^* & -\xi K^* & -\frac{a_K}{K^*} \end{bmatrix}.$$

$J(T^*, D^*, T_1^*, T_2^*, K^*)$ can be expressed as a block diagonal matrix by: $J(T^*, D^*, T_1^*, T_2^*, K^*) = \begin{bmatrix} J_{11} & O \\ J_{21} & J_{22} \end{bmatrix}$.

Where $J_{11} = \begin{bmatrix} -\frac{a_T}{T^*} - \frac{\rho T^*}{T_{max}} & -\delta_1 T^* & 0 \\ -\delta_2 D^* & -\frac{a_D}{D^*} & 0 \\ \eta_1 D^* & \eta_1 T^* & \alpha - \mu_3 \end{bmatrix}$, $J_{21} = \begin{bmatrix} \eta_2 D^* & \eta_2 T^* & -\beta T_2^* \\ 0 & 0 & \gamma K^* \end{bmatrix}$ and $J_{22} = \begin{bmatrix} -\frac{\eta_2 T^* D^*}{T_2^*} & 0 \\ -\xi K^* & -\frac{a_K}{K^*} \end{bmatrix}$.

Submatrix J_{22} has two negative eigenvalues, viz., $-\frac{a_K}{K^*}$ and $-\beta T_1^* - \mu_4$. So the stability criteria of the dynamical system around the interior equilibrium (E) depends on the eigenvalues of J_{11}. After expanding the matrix J_{11} in order to develop the characteristic equation in form,

$$(\alpha - \mu_3 - \lambda)\left(\lambda^2 + \lambda\left(\frac{a_T}{T^*} + \frac{\rho T^*}{T_{max}} + \frac{a_D}{D^*}\right) + \left(\frac{a_T a_D}{T^* D^*} + \frac{\rho T^* a_D}{T_{max} D^*} - \delta_1 \delta_2 T^* D^*\right)\right) = 0 \quad (15.11)$$

From the existence criteria of the interior equilibrium, the death rate of Th$_1$ is higher compared with the pro-inflammatory cytokine effect over Th$_1$, so it is obvious that $(\alpha - \mu_3) < 0$. Now, by considering Eq. (15.11) and using *Routh–Hurwitz criteria* [33, 34], we can state that the interior equilibrium will be locally asymptotically stable if $\left(\frac{a_T a_D}{T^* D^*} + \frac{\rho T^* a_D}{T_{max} D^*} - \delta_1 \delta_2 T^* D^*\right) > 0$.

Proposition 15.2 *Along with the existence condition (i.e., $\mathscr{A} > 0$), if $\frac{a_T a_D}{T^* D^*} + \frac{\rho T^* a_D}{T_{max} D^*} > \delta_1 \delta_2 T^* D^*$, then the interior equilibrium $E = (T^*, D^*, T_1^*, T_2^*, K^*)$ is locally asymptotically stable.*

15.4 Impulse Therapeutic Approaches

In this section, we analyze our drug-induced system using modified impulsive method for a better understanding of drug dynamics [35, 36]. Here, we study the effects of IL-10 through impulsive way with a fixed time interval to control the keratinocytes' cell population. During the therapy period of biologic (IL-10), taken through injection, the cell density of keratinocytes are made less by some proportion r. Here, we assume that the injections are taken at a fixed time interval and the effects of IL-10 on Th$_1$ and Th$_2$ are not considered. Now by taking the maximum density of keratinocyte, the one-dimensional impulsive differential equation takes the form:

$$\begin{aligned} \frac{dK}{dt} &\leq a_K - \left(\mu_5 - \frac{\gamma \eta_1 a_T a_D}{(\mu_1 - \rho)\mu_2(\mu_3 - \alpha)}\right)K, \text{ for } t \neq t_k \\ \Delta K &= -rK, \text{ for } t = t_k \text{ where } k = 1, 2, 3, \ldots, n. \end{aligned} \quad (15.12)$$

For mathematical simplicity, we use the notation $\mathscr{P}$ instead of the large expression $\left(\mu_5 - \frac{\gamma\eta_1 a_T a_D}{(\mu_1-\rho)\mu_2(\mu_3-\alpha)}\right)$. Here for single impulsive cycle $t_k \le t \le t_{k+1}$, the solution of the Eq. (15.12) is

$$K(t_{k+1}^-) = \frac{a_K}{\mathscr{P}}\left[1 - e^{-\mathscr{P}(t_{n+1}-t_n)}\right] + K(t_n^+)e^{-\mathscr{P}(t_{n+1}-t_n)}. \tag{15.13}$$

Where $K(t_k^-)$ is the value immediately before and $K(t_k^+)$ is the value immediately after the impulse therapy. Now, for different successive time interval, solutions become

$$\begin{aligned}
K(t_1^-) &= \frac{a_K}{\mathscr{P}},\\
K(t_1^+) &= (1-r)\frac{a_K}{\mathscr{P}},\\
K(t_2^-) &= (1-r)\frac{a_K}{\mathscr{P}}e^{-\mathscr{P}(t_2-t_1)} + \frac{a_k}{\mathscr{P}}\left[1 - e^{-\mathscr{P}(t_2-t_1)}\right],\\
K(t_2^+) &= (1-r)^2\frac{a_K}{\mathscr{P}}e^{-\mathscr{P}(t_2-t_1)} + (1-r)\frac{a_K}{\mathscr{P}}\left[1 - e^{-\mathscr{P}(t_2-t_1)}\right],\\
K(t_3^-) &= \frac{a_K}{\mathscr{P}}\Big[(1-r)^2e^{-\mathscr{P}(t_3-t_1)} + (1-r)e^{-\mathscr{P}(t_3-t_2)} - (1-r)e^{-\mathscr{P}(t_3-t_1)}\\
&\quad +1 - e^{-\mathscr{P}(t_3-t_2)}\Big],\\
K(t_3^+) &= \frac{a_K}{\mathscr{P}}\Big[(1-r)^3e^{-\mathscr{P}(t_3-t_1)} + (1-r)^2e^{-\mathscr{P}(t_3-t_2)} - (1-r)^2e^{-\mathscr{P}(t_3-t_1)}\\
&\quad +(1-r) - (1-r)e^{-\mathscr{P}(t_3-t_2)}\Big],\\
K(t_4^-) &= \frac{a_K}{\mathscr{P}}\Big[(1-r)^3e^{-\mathscr{P}(t_4-t_1)} + (1-r)^2e^{-\mathscr{P}(t_4-t_2)} + (1-r)e^{-\mathscr{P}(t_4-t_3)} +\\
&\quad 1 - (1-r)^2e^{-\mathscr{P}(t_4-t_1)} - (1-r)e^{-\mathscr{P}(t_4-t_2)} - e^{-\mathscr{P}(t_4-t_3)}\Big],\\
K(t_4^+) &= \frac{a_K}{\mathscr{P}}\Big[(1-r)^4e^{-\mathscr{P}(t_4-t_1)} + (1-r)^3e^{-\mathscr{P}(t_4-t_2)} + (1-r)^2e^{-\mathscr{P}(t_4-t_3)} +\\
&\quad (1-r)^3e^{-\mathscr{P}(t_4-t_1)} - (1-r)^2e^{-\mathscr{P}(t_4-t_2)} - (1-r)e^{-\mathscr{P}(t_4-t_3)} + (1-r)\Big].\\
&\quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots
\end{aligned}$$

The general solution becomes

$$\begin{aligned}
K(t_n^-) = \frac{a_K}{\mathscr{P}}\Big[&(1-r)^{(n-1)}e^{-\mathscr{P}\,(t_n-t_1)} + (1-r)^{(n-2)}e^{-\mathscr{P}\,(t_n-t_2)} + \cdots\\
&+(1-r)e^{-\mathscr{P}\,(t_n-t_{n-1})} + 1 - (1-r)^{(n-2)}e^{-\mathscr{P}\,(t_n-t_1)} - (1-r)^{(n-3)}e^{-\mathscr{P}\,(t_n-t_2)} - \cdots\\
&-e^{-\mathscr{P}\,(t_n-t_{n-1})}\Big]
\end{aligned} \tag{15.14}$$

$$\begin{aligned} K(t_n^+) = \frac{a_K}{\mathscr{P}}\Big[&(1-r)^n e^{-\mathscr{P}\,(t_n-t_1)} + (1-r)^{(n-1)} e^{-\mathscr{P}\,(t_n-t_2)} + \cdots \\ &+(1-r)^2 e^{-\mathscr{P}\,(t_n-t_{n-1})} - (1-r)^{(n-1)} e^{-\mathscr{P}\,(t_n-t_1)} - (1-r)^{(n-2)} e^{-\mathscr{P}\,(t_n-t_2)} - \cdots \\ &-(1-r)e^{-\mathscr{P}\,(t_n-t_{n-1})} + (1-r)\Big] \end{aligned} \tag{15.15}$$

The above general solutions (15.14), (15.15) help to predict the maximal Keratinocytes present in the formation of psoriasis, just before and after was injection taken. Note that the solutions do not depend on the time between two consecutive drug doses.

15.4.1 System Under Fixed IL-10 Injecting Process

For a fixed time period, i.e., $\tau = t_{n+1} - t_n$ is constant, then we have

$$\begin{aligned} K(t_n^-) &= \frac{a_K}{\mathscr{P}}\Big[1 + (1-r)e^{-\mathscr{P}\tau} + (1-r)^2 e^{-2\mathscr{P}\tau} + \cdots + (1-r)^{n-1} e^{-(n-1)\mathscr{P}\tau} \\ &\qquad -e^{-\mathscr{P}\tau}\Big(1 + (1-r)e^{-\mathscr{P}\tau} + \cdots + (1-r)^{n-2} e^{-(n-2)\mathscr{P}\tau}\Big)\Big] \\ &= \frac{a_K}{\mathscr{P}}\Big[\frac{1-(1-r)^n e^{-n\mathscr{P}\tau}}{1-(1-r)e^{-\mathscr{P}\tau}} - e^{-\mathscr{P}\tau}\frac{1-(1-r)^{n-1}e^{-(n-1)\mathscr{P}\tau}}{1-(1-r)e^{-\mathscr{P}\tau}}\Big] \\ \lim_{n\to\infty} K(t_n^-) &= \frac{a_K}{\mathscr{P}}\Big[\frac{1}{1-(1-r)e^{-\mathscr{P}\tau}} - e^{-\mathscr{P}\tau}\frac{1}{1-(1-r)e^{-\mathscr{P}\tau}}\Big] \\ &= \frac{a_K}{\mathscr{P}}\Big[\frac{1-e^{-\mathscr{P}\tau}}{1-(1-r)e^{-\mathscr{P}\tau}}\Big]. \end{aligned}$$

This is the density of keratinocytes before taking the drug dose, in the long term. Now, after applying drug, the density of the keratinocyte will be expressed by the following expression:

$$\begin{aligned} \lim_{n\to\infty} K(t_n^+) &= (1-r)\lim_{n\to\infty} K(t_n^-), \\ &= (1-r)\frac{a_K}{\mathscr{P}}\Big[\frac{1-e^{-\mathscr{P}\tau}}{1-(1-r)e^{-\mathscr{P}\tau}}\Big]. \end{aligned} \tag{15.16}$$

After the long-term biologic therapy, to keep the keratinocyte density (from the Eq. (15.16)) below a certain threshold $\tilde{K}$ (the normal keratinocyte density), we need the maximum time interval of applying two consecutive doses, i.e., τ_{max}, which must satisfy

$$\tau < (1-r)\frac{1}{\mathscr{P}}\ln\Big[\frac{a_K - (1-r)\tilde{K}\mathscr{P}}{a_K - \tilde{K}\mathscr{P}}\Big]$$

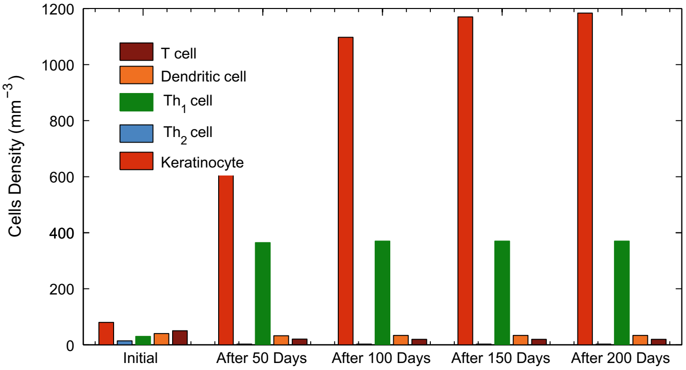


Fig. 15.2 Qualitative behavior of system variables (T cell, DCs, Th$_1$, Th$_2$, and keratinocyte) are demonstrated by the bar diagram

$$\tau \; < \; (1-r)\frac{1}{\left(\mu_5 - \frac{\gamma\eta_1 a_T a_D}{(\mu_1-\rho)\mu_2(\mu_3-\alpha)}\right)} \ln\left[\frac{a_K - (1-r)\tilde{K}\left(\mu_5 - \frac{\gamma\eta_1 a_T a_D}{(\mu_1-\rho)\mu_2(\mu_3-\alpha)}\right)}{a_K - \tilde{K}\left(\mu_5 - \frac{\gamma\eta_1 a_T a_D}{(\mu_1-\rho)\mu_2(\mu_3-\alpha)}\right)}\right] \equiv \tau_{max} \text{ (say)}. \tag{15.17}$$

The maximum period mentioned by the Eq. (15.18) between two consecutive IL-10 injections must be required to maintain the keratinocyte density below $\tilde{K}$. The threshold value $\tilde{K}$ must satisfy

$$\tilde{K} < \frac{a_K}{\left(\mu_5 - \frac{\gamma\eta_1 a_T a_D}{(\mu_1-\rho)\mu_2(\mu_3-\alpha)}\right)}. \tag{15.18}$$

It follows that, in the case of fixed IL-10 injecting process, we can derive a maximal gap of injection (15.18), which is fixed and that may keep the concentration of keratinocyte strictly below the threshold described by the Eq. (15.16).

15.5 Numerical Simulations

In the previous sections, we have used several analytic tools for theoretical analysis of the formulated mathematical model and also, we studied the system behavior introducing biologic. In this section, we execute the numerical simulation of our model system on the basis of analytical outcomes. The values we assigned to each parameter are collected from different papers, listed in Table 15.1. Initial values of the cells density are assigned as $T(0) = 25$, $D(0) = 20$, $T_1(0) = 15$, $T_2(0) = 7$, and $K(0) = 20$. Here, we have tried to emphasize the cells' interaction toward the psoriatic expression and also, dynamical behavior of the cell components were numerically evaluated under impulsive approach with IL-10 therapy. Numerical simulations are done using Mathworks MATLAB (version 7.6.0).

Table 15.1 Parameters value using for numerical simulation

Parameter	Assigned value	Range	References
a_T	$9\,\text{mm}^{-3}\text{Day}^{-1}$	$9\text{–}15\,\text{mm}^{-3}\text{Day}^{-1}$	[25, 30]
a_D	$12\,\text{mm}^{-3}\text{Day}^{-1}$	$12\text{–}14\,\text{mm}^{-3}\text{Day}^{-1}$	[30, 31]
ρ	$0.03\ \text{Day}^{-1}$	$0.03\ \text{Day}^{-1}$	[32]
T_{max}	$1500\,\text{mm}^{-3}$	$1500\,\text{mm}^{-3}$	[32]
δ_1	$0.07\ \text{Day}^{-1}$	$0.005\text{–}0.15\ \text{Day}^{-1}$	[27, 29]
δ_2	$0.08\ \text{Day}^{-1}$	$0.00004\text{–}0.4\ \text{Day}^{-1}$	[30, 31]
μ_1	$0.1\ \text{Day}^{-1}$	$0.007\text{–}0.1\ \text{Day}^{-1}$	[27, 29]
η_1	$0.05\ \text{Day}^{-1}$	Estimated	[37]
η_2	$0.03\ \text{Day}^{-1}$	Estimated	[37]
α	$0.04\ \text{Day}^{-1}$	Estimated	[38, 39]
μ_2	$0.05\ \text{Day}^{-1}$	$0.002\text{–}0.05\ \text{Day}^{-1}$	[27, 31]
β	$0.02\ \text{Day}^{-1}$	–	Assumed
μ_3	$0.24\ \text{Day}^{-1}$	$0.24\ \text{Day}^{-1}$	[40]
μ_4	$0.24\ \text{Day}^{-1}$	$0.24\ \text{Day}^{-1}$	[40]
γ	$0.0001\ \text{Day}^{-1}$	–	Assumed
ξ	$0.04\ \text{Day}^{-1}$	–	Assumed
μ_5	$0.88\ \text{Day}^{-1}$	$0.04\text{–}0.9\ \text{Day}^{-1}$	[25, 27]
a_K	$30\,\text{mm}^{-3}\text{Day}^{-1}$	Estimated	[14, 41]

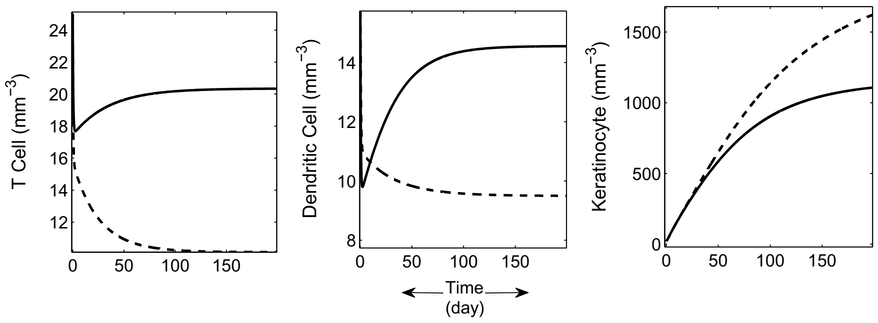


Fig. 15.3 Dynamical behavior of T cell, DCs, and keratinocyte are plotted with respect to time for different activation rates of T cell and dendritic cell (δ_1 and δ_2)

In Fig. 15.2, we investigate the qualitative behavior of considered cells between 200 days. From Fig. 15.2, it is manifested that due to interaction between T cell and dendritic cell both population will be decreased chronologically. Both T cell and DCs density reached a stable condition after 50 days approximately. For the effect of pro-inflammatory cytokine, Th$_1$ is upregulated and Th$_2$ is subjected to suppressed condition. After initial 50 days, both the population (Th$_1$ and Th$_2$) will be stable. For the Th$_1$ mediated cytokines, the keratinocyte population is increased but the

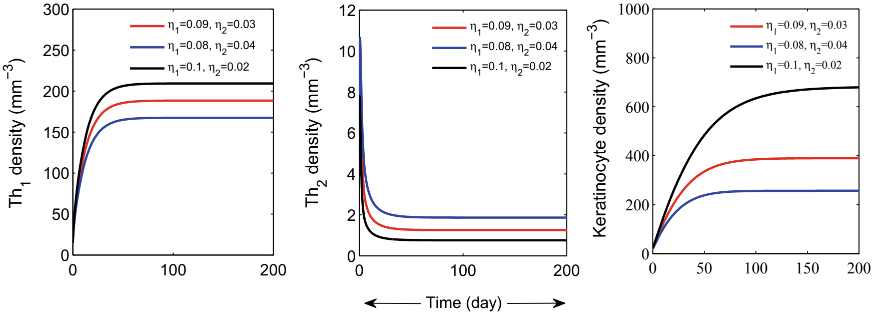


Fig. 15.4 Qualitative behavior of Th$_1$, Th$_2$, and keratinocyte are plotted with respect to time for different accumulation rates of Th$_1$ and Th$_2$ (η_1 and η_2) from naive T cell

rate becomes slow due to the effect of anti-inflammatory cytokines released by Th$_2$. Eventually, keratinocytes population will be stable between 180 days.

In Fig. 15.3, the dynamical nature of T cells, dendritic cells, and keratinocytes are demonstrated with respect to time for different activation rates of T cell (δ_1) and dendritic cell (δ_2). This figure manifests that for low activation rate of T cell and dendritic cell ($\delta_1 = 0.06$ and $\delta_2 = 0.07$), rate of synapse formation among DCs and T cells slows down. Due to slow synapse formation, the density of both immune cells initially reduce but after a short time interval, they chronologically increase. For low activation rate, T cells and dendritic cells reach a stable condition after 100 days at density level 21 mm^{-3} and 14.5 mm^{-3}. On the other hand, if the activation rates are high ($\delta_1 = 0.08$ and $\delta_2 = 0.09$), the population of T cell and dendritic cell will be reduced to reach a stable situation at density level 2 and 5 mm^{-3}, respectively. Since the growth of keratinocyte depends upon Th$_1$ and Th$_2$ regulation, so it is also indirectly dependent on the activation rate of T cell and dendritic cell. Hence, the keratinocyte density reaches 1150 mm^{-3} for low activation rate. Conversely, for higher activation rate, the growth of keratinocyte is dramatically increased to density level 1600 mm^{-3}. It is to be noted that due to the presence of Th$_2$, the density deflection of keratinocyte is low in comparison with the other two immune cells.

Th$_1$, Th$_2$ and keratinocyte are plotted with respect to time for different accumulation rates of Th$_1$ and Th$_2$ (η_1 and η_2) in Fig. 15.4, other parameters are taken as same in Table 15.1. For the higher accumulation rate of Th$_1$ ($\eta_1 = 0.1$) and lower accumulation rate of Th$_2$ ($\eta_2 = 0.02$), Th$_1$ density reaches at the level 210 mm^{-3} and Th$_2$ density reduces to 1 mm^{-3}. In this case, we notice a startling change in keratinocyte concentration which reaches a level of 1300 mm^{-3} due to high inflammation effect of Th$_1$. From this figure, it is also clear that for low accumulation rate of Th$_1$ ($\eta_1 = 0.08$) and higher accumulation rate of Th$_2$ ($\eta_2 = 0.04$), Th$_1$ density reduced to 160 mm^{-3} and Th$_2$ density increased to level 2 mm^{-3}. In that case, we observe that the concentration level of Th$_2$ is more efficient to maintain the balancing between pro-inflammatory and anti-inflammatory cytokine which ultimately reduces the keratinocyte density to level 510 mm^{-3}. This figure also manifest that due to high

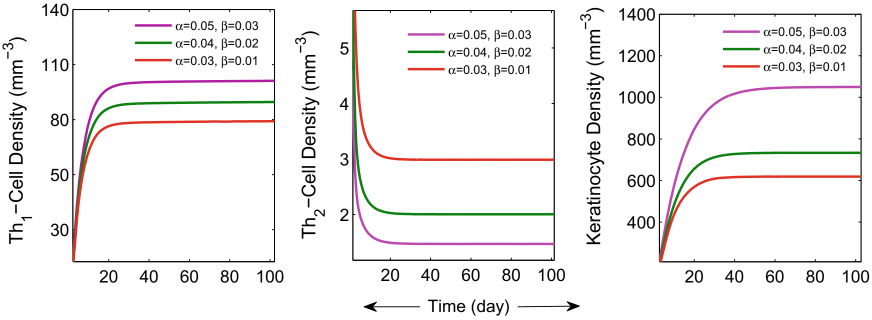


Fig. 15.5 Dynamical behavior of Th_1, Th_2, and keratinocyte are plotted with respect to time under different effects of pro-inflammatory cytokine on Th_1 and Th_2 (α and β)

inflammation effect of pro-inflammatory cytokine for psoriatic patient, a little bit of increment of Th_1 accumulation has a great impact on keratinocyte concentration.

In Fig. 15.5, we demonstrate how the density of Th_1, Th_2, and keratinocyte depend on both inflammation effect (negative and positive) of Th_1 cell-mediated cytokines, represented by α and β. It is to be noticed that when the inflammation effects are considered as $\alpha = 0.04$ and $\beta = 0.02$, then the cells density of Th_1 and Th_2 reach a stable concentration level of 185 mm^{-3} and 2 mm^{-3} , respectively, and under this circumstance, keratinocyte reaches a density level of 725 mm^{-3}. From this figure, it is clear that due to higher inflammation efficacy ($\alpha = 0.05$ and $\beta = 0.03$) on Th_1 and Th_2, keratinocyte shows a hyper-proliferative nature and it becomes stable at 1050 mm^{-3}. Furthermore, for the low inflammation effect ($\alpha = 0.03$ and $\beta = 0.01$), Th_1 density gradually decreases and Th_2 density chronologically increases to a certain level. Under this well managed condition, keratinocyte goes into a suppressed situation (density of keratinocyte <625 mm^{-3}). This figure mainly emphasize an abnormal deflection of keratinocyte due to higher inflammation effect of pro-inflammatory cytokine. The anti-inflammation effect of Th_2 mediated cytokine is lower to compare with the pro-inflammatory effect on keratinocyte cell population through Th_1 mediated cytokine, it results in the changes of keratinocyte density that is not smooth for different effects of pro-inflammatory cytokine.

In Fig. 15.6, we manifest that how the nature of keratinocytes population changes under impulsive therapeutic approach with respect to time t and we also plot a comparison simulation between with control and without control situation of keratinocyte. In Fig. 15.6a, the solid line indicates the keratinocyte density during therapy period and the dotted lines indicate the upper threshold of normal growth of keratinocyte. We evaluate the upper boundary of keratinocyte density as 272 mm^{-3}, using the Eq. (15.18) and considering the parameter from Table 15.1. This mathematically evaluated upper threshold is represented by a red dotted line. The natural growth of keratinocyte is also estimated to be 195200 mm^{-3} in comparison with the ratio of the other cells taken in our model, represented by the green dotted line [14, 42]. It is evident from this figure, for impulse dosing ($r = 0.005$ and dosing

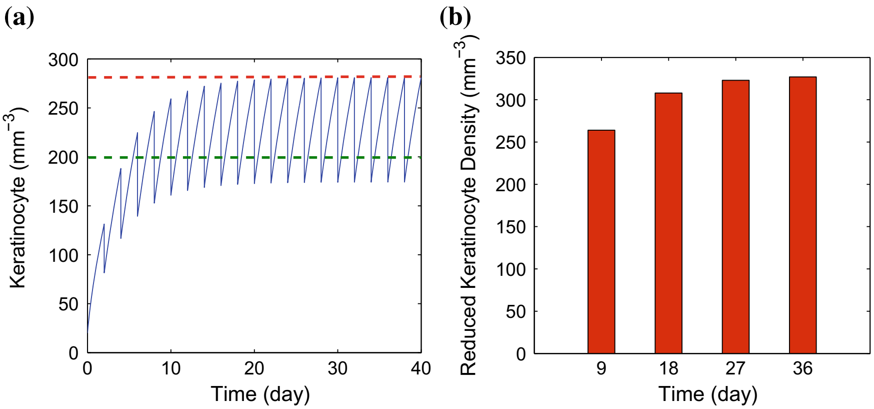


Fig. 15.6 Simulations of keratinocyte density under fixed IL-10 injecting process by considering the value of the parameter corresponding to reduction of keratinocyte due to the impact of drug (IL-10), i.e., $r = 0.005$ and dosing interval 4.5 days. **a** Qualitative behavior of keratinocyte is plotted with respect to time. **b** Reduction of keratinocyte density for different days interval due to the effect of IL-10

interval(τ) $= 4.5$), the keratinocyte density is just below the threshold (272 mm^{-3}). For this proposed drug dose, keratinocyte satisfies the mathematical threshold but biologically estimated threshold is not gratified accurately. After initial 30 days of injecting, keratinocytes show an oscillatory behavior to a fixed magnitude of 170–272 mm^{-3}. In Fig. 15.6b, we emphasize the reduction of keratinocyte during IL-10 therapy for initial 36 days. From this figure, it is clear that initially parameter corresponding to the reduction of keratinocyte due to the impact of drug (IL-10) is low due to the presence of a high amount of pro-inflammatory cytokine. Within 20 days of treatment, IL-10 reduces the Th$_1$ cell density in a noticeable amount; after that, it directly downregulates the keratinocyte density. From this figure, it is observed that after 27 days, the reduced amount of keratinocyte is more than 300 mm^{-3}.

Figure 15.7 depicts the dynamical behavior of keratinocyte under perfect biologic dose (parameter corresponding to the reduction of keratinocyte due to the impact of drug (IL-10), i.e., $r = 0.005$ and considering dosing interval 2.8 days) for a psoriatic patient and simulation results of keratinocyte's retrenchment from initial to under treatment condition. In Fig. 15.7a, it is clear that when we fix the value of r as 0.005 and dosing interval 2.8 days, keratinocytes maintain its density level under 200 mm^{-3}. For this particular IL-10 dosing, keratinocytes density satisfied the biological estimation as well as mathematical evaluation. The keratinocyte density reaches a stable condition within 36 days and oscillates with a fixed amplitude between 120200 mm^{-3}. Figure 15.7b, demonstrates the reduction amount of keratinocyte during IL-10 therapy for first 36 days. Here, we simulate the reduced amount of keratinocyte density after every 9 days of treatment, evaluated by subtracting the density of keratinocyte of with treatment policy from that of in the

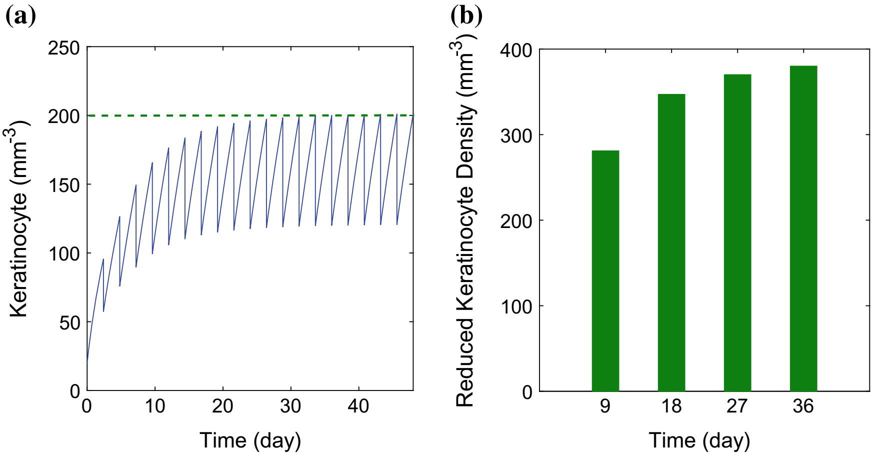


Fig. 15.7 Simulations of keratinocyte density under fixed IL-10 injecting process by considering the value of the parameter corresponding to the reduction of keratinocyte due to the impact of drug (IL-10), i.e., $r = 0.005$ and dosing interval 2.8 days. **a** Qualitative behavior of keratinocyte is plotted with respect to time. **b** Reduction of keratinocyte density for different days due to the effect of IL-10

without treatment policy. After initial 9 days, the consolidated amount of keratinocyte (280 mm^{-3}) is near about 70% of hyper-proliferative keratinocyte density. From this figure, it is also clear that after 30 days of biologic treatment ($r = 0.005$ and considering dosing interval 2.8 days) keratinocyte density suppressed in a noticeable amount (380 mm^{-3}), i.e., above 90% of its prime density.

15.6 Discussion

In this manuscript, we have studied the role of immune cells (T cell, DCs, Th$_1$, and Th$_2$) along with inflammation effect of cytokine network for psoriatic skin by considering a mathematical model. In our analytical study, we obtain the existence condition of interior equilibrium which describes that the natural death rate of Th$_1$ cell (μ_3) is greater than the inflammatory effect on itself by Th$_1$, indicated by α. We also established the stability criterion of interior equilibrium depending on *Routh–Hurwitz criteria*, which validate the existence condition along with biological restrictions. Here, we have discussed about cytokine treatment which gives a better impact on disease pathogenesis using some mathematical tools and numerical simulations. To control the hyper-proliferative nature of keratinocyte, we use biologic (IL-10) and consider it's direct effect on keratinocytes using the modified impulsive method. Our mathematical outcomes depict that if we fix the time gap of two consecutive doses less than τ_{max}, then the disease will be under control. We also analytically determined the threshold for keratinocyte ($\tilde{K}$) and we also get the estimated value of normal keratinocyte proliferation ($200\,\mathrm{mm}^{-3}$) rate from various clinical studies. Our numerical results reveal that if the dose interval of IL-10 is fixed as 2.8 days ($<\tau_{max}$) and the parameter corresponding to the reduction of Keratinocyte due to the impact of drug (IL-10), i.e., $r = 0.005$, the keratinocyte density can be reduced to a certain level which validated both the mathematical and clinical outcomes. It is also observed that considering the effect of IL-10 on keratinocyte through modified impulsive method, it is possible to control the keratinocyte density within about 36 days.

15.7 Conclusion

Here, we introduce IL-10 as a biologic (drug) from the cell-biological point of view as hyper-proliferation of keratinocytes is the main cause of psoriasis. We demonstrate the impact of this drug on keratinocyte population only by considering a suitable value of the parameter (r) which corresponds to the reduction of keratinocyte due to the impact of drug (IL-10). In the present study, we assume a very small value of this parameter (r) to avoid the side effects of this highly sensitive cytokine treatment and also to reduce the cost of drug administration. We use impulsive control strategy among all the control strategies to make drug administration more realistic and biologically relevant. Our analytical and numerical results reveal that keratinocyte population reduces significantly by applying biologic (IL-10), which suggests this drug as a potential drug for better treatment of psoriasis. In fact, our study suggests that using IL-10, it is possible to remove more than 90% of psoriatic plaque within 5 weeks and this will be a burning challenge for clinical and experimental researchers in the future.

Acknowledgements The research work is financially supported by DST-PURSE PROJECT (Phase-II), Government of India, Department of Mathematics, Jadavpur University, Kolkata-700032, India.

References

1. World Health Organization, Global report on psoriasis 2016, *WHO Library Cataloguing-in-Publication Data* (2016)
2. T.J. Kindt, R.A. Goldsby, B.A. Osborne, J. Kuby, *Kuby Immunology* (Macmillan, London, 2007)
3. K. Liu, M.C. Nussenzweig, Origin and development of dendritic cells. Immunol. Rev. **234**(1), 45–54 (2010)
4. A.K. Roy, P.K. Roy, E. Grigorieva, Mathematical insights on psoriasis regulation: role of Th 1 and Th 2 cells. Math. Biosci. Eng. **15**(3), 717–738 (2018)
5. A. Coondoo, The role of cytokines in the pathomechanism of cutaneous disorders. Indian J. Derm. **57**(2), 90 (2012)

6. A. Balato, F. Ayala, M. Schiattarella, M. Megna, N. Balato, S. Lembo, *Pathogenesis of Psoriasis: The Role of Pro-inflammatory Cytokines Produced by Keratinocytes* (INTECH Open Access Publisher, London, 2012)
7. J.H. Mao, E.F. Saunier, J.P. de Koning, M.M. McKinnon, M.N. Higgins, H.T. Yang, A. Balmain, R.J. Akhurst, Genetic variants of Tgfb1 act as context-dependent modifiers of mouse skin tumor susceptibility. Proc. Natl. Acad. Sci. **103**(21), 8125–8130 (2006)
8. A. Cavani, G. Girolomoni, Interferon-?-stimulated human keratinocytes express the genes necessary for the production of peptide-loaded MHC class II molecules. J. Investig. Derm. **110**(2), 138–142 (1998)
9. A. Chiricozzi, E. Guttman-Yassky, M. Surez-Farinas, K.E. Nograles, S. Tian, I. Cardinale, S. Chimenti, J.G. Krueger, Integrative responses to IL-17 and TNF-a in human keratinocytes account for key inflammatory pathogenic circuits in psoriasis. J. Investig. Derm. **131**(3), 677–687 (2011)
10. A. Johnston, Y. Fritz, S.M. Dawes, D. Diaconu, P.M. Al-Attar, A.M. Guzman, C.S. Chen, W. Fu, J.E. Gudjonsson, T.S. McCormick, N.L. Ward, Keratinocyte overexpression of IL-17C promotes psoriasiform skin inflammation. J. Immunol. **190**(5), 2252–2262 (2013)
11. J. Baliwag, D.H. Barnes, A. Johnston, Cytokines in psoriasis. Cytokine **73**(2), 342–350 (2015)
12. A.K. Roy, F. Al Basir, P.K. Roy, A vivid cytokines interaction model on psoriasis with the effect of impulse biologic (TNF-α inhibitor) therapy. J. Theor. Biol. **474**, 63–77 (2019)
13. P. Goodwin, S. Hamilton, L. Fry, The cell cycle in psoriasis. Br. J. Derm. **90**(5), 517–524 (1974)
14. G.D. Weinstein, J.L. McCullough, P.A. Ross, Cell kinetic basis for pathophysiology of psoriasis. J. Investig. Derm. **82**(6), 623–628 (1984)
15. L.M. Johnson-Huang, M. Surez-Farias, K.C. Pierson, J. Fuentes-Duculan, I. Cueto, T. Lentini, M. Sullivan-Whalen, P. Gilleaudeau, J.G. Krueger, A.S. Haider, M.A. Lowes, A single intradermal injection of IFN-? induces an inflammatory state in both non-lesional psoriatic and healthy skin. J. Investig. Derm. **132**(4), 1177–1187 (2012)
16. A. Mussi, C. Bonifati, M. Carducci, G. D'Agosto, F. Pimpinelli, D. D'Urso, L. D'Auria, M. Fazio, F. Ameglio, Serum TNF-alpha levels correlate with disease severity and are reduced by effective therapy in plaque-type psoriasis. J. Biol. Regul. Homeost. Agents **11**(3), 115–118 (1996)
17. A. Mussi, C. Bonifati, M. Carducci, G. D'Agosto, F. Pimpinelli, D. D'Urso, L. D'Auria, M. Fazio, F. Ameglio, Serum TNF-alpha levels correlate with disease severity and are reduced by effective therapy in plaque-type psoriasis. J. Biol. Regul. Homeost. Agents **11**(3), 115–118 (1997)
18. P. Nockowski, J.C. Szepietowski, M. Ziarkiewicz, E. Baran, Serum concentrations of transforming growth factor beta 1 in patients with psoriasis vulgaris. Acta Derm.Venerologica Croat.: ADC **12**(1), 2–6 (2003)
19. K.A. Papp, The long-term efficacy and safety of new biological therapies for psoriasis. Arch. Derm. Res. **298**(1), 7–15 (2006)
20. K. Asadullah, W. Sterry, H.D. Volk, Interleukin-10 therapyreview of a new approach. Pharmacol. Rev. **55**(2), 241–269 (2003)
21. K. Ghoreschi, P. Thomas, S. Breit, M. Dugas, R. Mailhammer, W. van Eden, R. van der Zee, T. Biedermann, J. Prinz, M. Mack, U. Mrowietz, Interleukin-4 therapy of psoriasis induces Th2 responses and improves human autoimmune disease. Nat. Med. **9**(1), 40–46 (2003)
22. K. Reich, M. Bruck, A. Grafe, C. Vente, C. Neumann, C. Grabe, Treatment of psoriasis with interleukin-10. J. Investig. Derm. **111**(6), 1235–1236 (1998)
23. M. Friedrich, W.D. Dcke, A. Klein, S. Philipp, H.D. Volk, W. Sterry, K. Asadullah, Immunomodulation by interleukin-10 therapy decreases the incidence of relapse and prolongs the relapse-free interval in psoriasis. J. Investig. Derm. **118**(4), 672–677 (2002)
24. I.B. McInnes, G.G. Illei, C.L. Danning, C.H. Yarboro, M. Crane, T. Kuroiwa, R. Schlimgen, E. Lee, B. Foster, D. Flemming, C. Prussin, IL-10 improves skin disease and modulates endothelial activation and leukocyte effector function in patients with psoriatic arthritis. J. Immunol. **167**(7), 4075–4082 (2001)

25. A. Datta, D.K. Kesh, P.K. Roy, Effect of CD4+ T-cells and CD8+ T-cells on psoriasis: a mathematical study. IMHOTEP: Afr. J. Pure Appl. Math. **3**(1), 1–11 (2016)
26. N.J. Savill, R. Weller, J.A. Sherratt, Mathematical modelling of nitric oxide regulation of rete peg formation in psoriasis. J. Theor. Biol. **214**(1), 1–16 (2002)
27. P.K. Roy, A. Datta, Impact of perfect drug adherence on immunopathogenic mechanism for dynamical system of psoriasis. BIOMATH **2**(1), 1212101 (2013)
28. E. Grigorieva, E. Khailov, P. Deignan, Optimal treatment strategies for control model of psoriasis, in *2017 Proceedings of the Conference on Control and its Applications*, Society for Industrial and Applied Mathematics (2017), pp. 86–93
29. C.A.O. Xianbing, A. Datta, F.A. Basir, P.K. Roy, Fractional-order model of the disease psoriasis: a control based mathematical approach. J. Syst. Sci. Complex. **29**(6), 1565–1584 (2016)
30. P.K. Roy, A. Datta, S. Rana, The Fractional-order differential equation model of psoriatic pathogenesis: a mathematical study. Afr. Diaspora J. Math., New Series **15**(2), 35–46 (2013)
31. P.K. Roy, A. Datta, Negative feedback control may regulate cytokines effect during growth of keratinocytes in the chronic plaque of psoriasis: a mathematical study. Int. J. Appl. Math. **25**(2), 233–254 (2012)
32. A. Datta, P.K. Roy, T-cell proliferation on immunopathogenic mechanism of psoriasis: a control based theoretical approach. Control Cybern. **42** (2013)
33. E.J. Routh, *A Treatise on the Stability of a Given State of Motion: Particularly Steady Motion* (Macmillan and Company, London, 1877)
34. A. Hurwitz, On the conditions under which an equation has only roots with negative real parts. Sel. Pap. Math. Trends Control Theory **65**, 273–284 (1964)
35. V. Lakshmikantham, D.D. Bainov, P.S. Simeonov, *Theory of Impulsive Differential Equations* (World Scientific, Singapore, 1989)
36. R.J. Smith, P. Cloutier, J. Harrison, A. Desforges, A mathematical model for the eradication of Guinea worm disease. In Understanding the Dynamics of Emerging and Re-Emerging Infectious Diseases Using Mathematical Models, 37/661(2), 133–156 (2012)
37. G. Magombedze, S. Eda, V.V. Ganusov, Competition for antigen between Th1 and Th2 responses determines the timing of the immune response switch during Mycobaterium avium subspecies paratuberulosis infection in ruminants. PLoS Comput. Biol. **10**(1), e1003414 (2014)
38. Y. Kogan, Z. Agur, M. Elishmereni, A mathematical model for the immunotherapeutic control of the Th1/Th2 imbalance in melanoma. Discr. Cont. Dyn. Syst. Ser. B **18**(4), 1017–1030 (2013)
39. Y. Kim, S. Lee, Y.S. Kim, S. Lawler, Y.S. Gho, Y.K. Kim, H.J. Hwang, Regulation of Th1/Th2 cells in asthma development: a mathematical model. Math. Biosci. Eng. **10**(4), 1095–1133 (2013)
40. R. Fernandez-Botran, V.M. Sanders, T.R. Mosmann, E.S. Vitetta, Lymphokine-mediated regulation of the proliferative response of clones of T helper 1 and T helper 2 cells. J. Exp. Med. **168**(2), 543–558 (1988)
41. P.K. Denman, D.S. McElwain, D.G. Harkin, Z. Upton, Mathematical modelling of aerosolised skin grafts incorporating keratinocyte clonal subtypes. Bull. Math. Biol. **69**(1), 157–179 (2007)
42. G.D. Weinstein, J.L. McCullough, P.A. Ross, Cell kinetic basis for pathophysiology of psoriasis. J. Investig. Derm. **85**(6), 579–583 (1985)

Chapter 16
On Fractional Partial Differential Equations of Diffusion Type with Integral Kernel

A. Akilandeeswari, K. Balachandran and N. Annapoorani

Abstract The main purpose of this work is to investigate the existence of solution of the fractional partial differential equations of diffusion type with integral kernel. The existence of solutions of the problem with Dirichlet boundary condition is established by using the Leray–Schauder fixed point theorem and Arzela–Ascoli theorem under suitable assumptions. Then, the result is generalized for Neumann boundary condition with the help of Green's identity.

Keywords Existence and uniqueness · Fractional derivatives and integrals · Partial Integrodifferential equations · Fixed point theorem

2010 Mathematics Subject Classification 34A12 · 26A33 · 45J05 · 47H10

16.1 Introduction

To model a process with delay, it is not sufficient to employ an ordinary or partial differential equations. An approach to resolve this problem is to use integrodifferential equations. In some fields such as nuclear reactor dynamics and thermoelasticity, we need to reflect the effect of the memory of the systems in the model. If such systems are modeled using partial differential equation, the effect of past history is ignored. Therefore in order to incorporate the memory effect in such systems, an integral term in the partial differential equation is introduced and this leads to a partial integrodifferential equation [7]. In recent years, due to the novel surprising insights and framework of fractional calculus, the fractional partial integrodifferential equations have been scrutinized by several authors. Historically, the origins of fractional calculus can be traced back to the end of the seventeenth century, the time when Newton and Leibniz developed the foundations of differential and integral calculus. It extends the differentiation and integration of integer order to an arbitrary order

A. Akilandeeswari (✉) · K. Balachandran · N. Annapoorani
Department of Mathematics, Bharathiar University, Coimbatore 641046, India
e-mail: akilamathematics@gmail.com

P. Manchanda et al. (eds.), *Mathematical Modelling, Optimization, Analytic and Numerical Solutions*, Industrial and Applied Mathematics,
https://doi.org/10.1007/978-981-15-0928-5_16

and concatenates these two operators. To be precise, it consists of integrodifferential operators with the convolution type integrals and power-law type weakly singular kernels. An imperative feature is that fractional derivatives and integrals are non local, since it depends on all of its historical states. This is very effective when the system has a longterm memory and any evaluation point depends on the past values of the function. For example, the use of half derivatives and integrals lead to the formulation of certain electrochemical problems which are more economical and useful than the classical approach in terms of Fick's law of diffusion. Some of the applications of fractional calculus in interdisciplinary sciences can be found in [30, 34]. During the last few decades, fractional differentiation is drawing huge consideration toward physical and biological behaviors. The reason behind using fractional order differential equation is that it is naturally related to systems with memory which exists in most biological systems and fractional order system response ultimately converges to the integer-order equations. The elementary theory and some applications of fractional differential equations are widely covered in [14, 26, 29] and for the books associated with fractional differential equations, see [18, 21, 24, 28]. The applications of fractional derivatives in reservoir engineering problems are given in [23]. Jesus et al. [19] investigated the fractional model of the electrical impedance for botanical elements according to Bode and polar diagrams. A review of some applications of fractional derivatives in continuum and statistical mechanics is given by Carpinteri and Mainardi [11]. Next, we propose some of the works concerning the solvability of fractional differential equations. For instance, Balachandran et al. [8, 9] studied the existence results for several kinds of fractional integrodifferential equations in a Banach space using a fixed point technique. In [36], Zhang et al. investigated the existence of nonnegative solutions for nonlinear fractional differential equations with nonlocal fractional integrodifferential boundary conditions on an unbounded domain by using the Leray–Schauder nonlinear alternative theorem. The differential transform method was applied to fractional integrodifferential equations in [6] to solve those equations analytically. To know more details about the existence of solutions of integrodifferential equation, see the papers [1, 2, 5, 12, 20] and for fractional partial integrodifferential equations refer [3, 4]. In this paper, we extend the results of [25] to fractional order partial integrodifferential equation of diffusion type with integral kernel.

16.2 Basic Concepts

Now, we present the definitions of some well-known fractional operators that play an important role in fractional calculus. For any $n-1<\alpha<n, n\in\mathbb{N}$, the Rieman–Liouville fractional integral operator is defined as follows:

Definition 16.2.1 ([21]) The partial Riemann–Liouville fractional integral operator of order α with respect to t of a function $f(x,t)$ is defined by

$$I^{\alpha} f(x,t) = \frac{1}{\Gamma(\alpha)} \int\limits_{0}^{t} \frac{f(x,s)}{(t-s)^{n-\alpha}} \, \mathrm{d}s.$$

where $f(\cdot, t)$ is an integrable function.

The most popular definition of fractional calculus is Riemann–Liouville fractional derivative definition, which is basic for the Caputo fractional derivative. It is written as follows:

Definition 16.2.2 ([21]) The partial Riemann–Liouville fractional derivative of order α of a function $f(x,t)$ with respect to t is of the form

$$\frac{\partial^{\alpha}}{\partial t^{\alpha}} f(x,t) = \frac{1}{\Gamma(n-\alpha)} \frac{\partial^{n}}{\partial t^{n}} \int\limits_{0}^{t} \frac{f(x,s)}{(t-s)^{\alpha-n+1}} \, \mathrm{d}s.$$

where the function $f(\cdot, t)$ has absolutely continuous derivatives up to order $(n-1)$.

Since the Riemann–Liouville fractional derivative of a constant is a function, to overcome this difficulty, Caputo [10] reformulated the Riemann–Liouville fractional derivative to handle integer order initial conditions, in the following way.

Definition 16.2.3 ([21]) The Caputo partial fractional derivative of order α with respect to t of a function $f(x,t)$ is defined as

$$\frac{{}^{C}\partial^{\alpha}}{\partial t^{\alpha}} f(x,t) = \frac{1}{\Gamma(n-\alpha)} \int\limits_{0}^{t} \frac{1}{(t-s)^{\alpha-n+1}} \frac{\partial^{n} f(x,s)}{\partial s^{n}} \, \mathrm{d}s.$$

where the function $f(\cdot, t)$ has absolutely continuous derivatives up to order $(n-1)$.

To know the properties of these operators, see the books [21, 28] and for more facts on the geometric and physical interpretation of fractional derivatives with Riemann–Liouville and Caputo types, see [17, 22]. There has been a significant development in ordinary and partial differential equations involving both Riemann–Liouville and Caputo fractional derivatives in the past few years, for instance, see the papers of Gejji and Jafari [16], Furati and Tatar [15]. The Riemann Liouville and Caputo fractional derivatives are linked by the following relationship:

$$\frac{{}^{C}\partial^{\alpha}}{\partial t^{\alpha}} f(x,t) = \frac{\partial^{\alpha}}{\partial t^{\alpha}} f(x,t) - \sum_{k=0}^{n-1} \frac{t^{k-\alpha}}{\Gamma(k+1-\alpha)} \frac{\partial^{k}}{\partial t^{k}} f(x,0).$$

Before looking at the existence result of fractional partial integrodifferential equations, we introduce some basic results that are inherently tied to existence theory.

Lemma 16.2.1 ((Leray–Schauder fixed point theorem) [25]) *If U is a closed bounded convex subset of a Banach space X and $T : U \to U$ is completely continuous, then T has at least a fixed point in U.*

Lemma 16.2.2 ((Arzela–Ascoli Theorem) [25]) *Assume that K is a compact set in $\mathbb{R}^n$, $n \geq 1$, then a set $S \subset C(K)$ is relatively compact in $C(K)$ if and only if the functions in S are uniformly bounded and equicontinuous on K.*

Lemma 16.2.3 ((Green's Identity) [13]) *Let Ω be a bounded domain in $\mathbb{R}^m$ with smooth boundary $\partial\Omega$. Then, for any $u, v \in C^2(\Omega)$,*

$$\int_\Omega v\Delta u \,\mathrm{d}x = \int_{\partial\Omega} v\frac{\partial u}{\partial n}\,\mathrm{d}s - \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x,$$

where n is the outward unit normal to the boundary $\partial\Omega$ and $\mathrm{d}s$ is the element of arc length. For the special case $v = 1$,

$$\int_\Omega \Delta u \,\mathrm{d}x = \int_{\partial\Omega} \frac{\partial u}{\partial n}\,\mathrm{d}s. \tag{16.1}$$

This is called Green's first identity.

16.3 Fractional Partial Differential Equations with Kernel

Let Ω be a bounded subset of an m-dimensional space with smooth boundary and let $J = [0, T]$. Consider the fractional partial integrodifferential equation of the form

$$\begin{aligned}\frac{{}^C\partial^\alpha u}{\partial t^\alpha} &= a(t)\Delta u(x,t) + \int_0^t h(t-s)\Delta u(x,s)\,\mathrm{d}s + f\big(t, u(x,t)\big)\\ &\quad + \int_0^t g(t,s,u(x,s))\,\mathrm{d}s,\ \ t \in J,\end{aligned} \tag{16.2}$$

with the initial condition

$$u(x,0) = u_0(x), \qquad x \in \ \Omega,$$

where $0 < \alpha < 1$, $h : J \to \mathbb{R}$ is a positive kernel and $f : J \times \mathbb{R} \to \mathbb{R}$ is a nonlinear function. This (16.2) is a special case of integrodifferential equation of motion of fractional Maxwell fluid with zero pressure. This type of equation appears in the investigation of viscoelastic property. This equation gets attention from the fact that the fractional derivatives are used to depict the viscoelasticity phenomena with little amount of constraints. A viscoelastic fractional order mathematical model of a human root dentin is proposed by Petrovic et al. in [27]. Fractional partial integrodifferential equation has also been applied in the study of signal processing, turbulence, plasma physics, and in many other fields, for instance, see [31–33]. The integral equation corresponding to (16.2) can be written as

$$\begin{aligned} u(x,t) = u_0(x) &+ \frac{1}{\Gamma(\alpha)}\int_0^t (t-s)^{\alpha-1} a(s)\Delta u(x,s)\,\mathrm{d}s \\ &+ \frac{1}{\Gamma(\alpha)}\int_0^t (t-s)^{\alpha-1}\Big(\int_0^s h(s-\tau)\Delta u(x,\tau)\,\mathrm{d}\tau\Big)\,\mathrm{d}s \\ &+ \frac{1}{\Gamma(\alpha)}\int_0^t (t-s)^{\alpha-1} f\big(s,u(x,s)\big)\,\mathrm{d}s \\ &+ \frac{1}{\Gamma(\alpha)}\int_0^t (t-s)^{\alpha-1}\left(\int_0^s g(s,\tau,u(x,\tau))\,\mathrm{d}\tau\right)\mathrm{d}s. \end{aligned} \tag{16.3}$$

Next, we present some hypotheses which will be used to prove our main result.

(H1) $a(t)$ is continuous on J and $a(t) \in L^{1/\beta}(0,t)$, for all $t \in J$ and some $\beta \in (0,\alpha)$. That is, $\left(\int_0^t (a(s))^{\frac{1}{\beta}}\,\mathrm{d}s\right)^{\beta} \leq C_1$.

(H2) $f(t,u)$ is continuous with respect to u, Lebesgue measurable with respect to t and satisfies

$$\frac{\int\limits_{\Omega} \phi(x) f(t,u)\,\mathrm{d}x}{\int\limits_{\Omega} \phi(x)\,\mathrm{d}x} \leq f\left(t, \frac{\int\limits_{\Omega} \phi(x) u(x,t)\,\mathrm{d}x}{\int\limits_{\Omega} \phi(x)\,\mathrm{d}x}\right),$$

for some function $\phi(x)$.

(H3) There exists an integrable function $m_1(t) : J \to [0,\infty)$ such that

$$\| f(t,u) \| \leq m_1(t)\|u\|,$$

where $m_1(t) \in L^{1/\beta}(0,t)$, for all $t \in J$ and $\left(\int_0^t (m_1(s))^{\frac{1}{\beta}}\,\mathrm{d}s\right)^{\beta} \leq C_2$, for β as in (H1) and $C_2 \geq 0$.

(H4) $g(t,s,u)$ is continuous with respect to u, Lebesgue measurable with respect to t and also satisfies the inequality

$$\frac{\int\limits_{\Omega} \phi(x) g(t,s,u)\,\mathrm{d}x}{\int\limits_{\Omega} \phi(x)\,\mathrm{d}x} \leq g\left(t, s, \frac{\int\limits_{\Omega} \phi(x) u(x,t)\,\mathrm{d}x}{\int\limits_{\Omega} \phi(x)\,\mathrm{d}x}\right).$$

(H5) There exists an integrable function $m_2(t,s) : J \times J \to [0,\infty)$, such that

$$\| g(t,s,u) \| \leq m_2(t,s)\|u\|,$$

and for a nonnegative integer C_3,

$$\Big(\int_0^t \Big(\int_0^s m_2(s,\tau)\,\mathrm{d}\tau\Big)^{\frac{1}{\beta}}\,\mathrm{d}s\Big)^{\beta} \leq C_3.$$

(H6) The integral kernel satisfies

$$\Big(\int_0^t \Big(\int_0^s h(s-\tau)\,\mathrm{d}\tau\Big)^{\frac{1}{\beta}}\Big)^{\beta} \leq C_4,$$

where $C_4 \geq 0$.

16.3.1 Dirichlet Boundary Condition

This section is consecrated to the existence of solution of (16.2) with Dirichlet boundary condition

$$u(x,t) = 0, \qquad (x,t) \in \partial\Omega \times J, \tag{16.4}$$

where $\partial\Omega$ is the boundary of Ω. In order to achieve the required result, consider the following eigenvalue problem:

$$\left.\begin{aligned} \Delta u + \lambda u &= 0,\ (x,t) \in \Omega \times J, \\ u &= 0,\ (x,t) \in \partial\Omega \times J, \end{aligned}\right\} \tag{16.5}$$

where λ is a constant not depending on the variables x and t. The theory of eigenvalue problems is well known [35]. Thus for $x \in \Omega$, the smallest eigenvalue λ_1 of the problem (16.5) is positive and the corresponding eigenfunction is $\phi(x) \geq 0$. Now, we define the function $U(t)$ as

$$U(t) = \frac{\int\limits_{\Omega} u(x,t)\phi(x)\,\mathrm{d}x}{\int\limits_{\Omega} \phi(x)\,\mathrm{d}x}. \tag{16.6}$$

The main theorem is as follows:

Theorem 16.3.1 *Assume that there exists a $\beta \in (0,\alpha)$ for some $0 < \alpha < 1$ such that (H1)–(H3) and (H6) hold. For any constant $b > 0$, suppose that*

$$r_1 = \min\left\{T, \left[\frac{\Gamma(\alpha)b}{(\|U(0)\| + b)(\lambda_1(C_1 + C_4) + C_2 + C_3)}\left(\frac{\alpha-\beta}{1-\beta}\right)^{1-\beta}\right]^{\frac{1}{\alpha-\beta}}\right\}.$$

Then there exists at least one solution for the initial value problem (16.2) on $\Omega \times [0, r_1]$.

Proof Our first aim is to prove that the initial value problem (16.2) has a solution if and only if the equation

$$\begin{aligned} U(t) = U(0) &- \frac{\lambda_1}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} a(s) U(s)\, \mathrm{d}s \\ &- \frac{\lambda_1}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} \Big(\int_0^s h(s-\tau) U(\tau)\, \mathrm{d}\tau \Big)\, \mathrm{d}s \\ &+ \frac{1}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} f(s, U(s))\, \mathrm{d}s \\ &+ \frac{1}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} \bigg(\int_0^s g\big(s, \tau, U(\tau)\big)\, \mathrm{d}\tau \bigg)\, \mathrm{d}s \end{aligned} \tag{16.7}$$

has a solution.

Step 1. We start the proof by assuming $u(x, t)$ to be a solution of (16.3). On integrating both sides of (16.3) with respect to $x \in \Omega$, we get

$$\begin{aligned} \int_\Omega u(x, t)\, \mathrm{d}x = \int_\Omega u_0(x)\, \mathrm{d}x &+ \frac{1}{\Gamma(\alpha)} \int_\Omega \int_0^t (t-s)^{\alpha-1} a(s) \Delta u(x, s)\, \mathrm{d}s\, \mathrm{d}x \\ &+ \frac{1}{\Gamma(\alpha)} \int_\Omega \int_0^t (t-s)^{\alpha-1} \Big(\int_0^s h(s-\tau) \Delta u(x, \tau)\, \mathrm{d}\tau \Big)\, \mathrm{d}s\, \mathrm{d}x \\ &+ \frac{1}{\Gamma(\alpha)} \int_\Omega \int_0^t (t-s)^{\alpha-1} f(s, u(x, s))\, \mathrm{d}s\, \mathrm{d}x \\ &+ \frac{1}{\Gamma(\alpha)} \int_\Omega \int_0^t (t-s)^{\alpha-1} \bigg(\int_0^s g\big(s, \tau, u(x, \tau)\big)\, \mathrm{d}\tau \bigg)\, \mathrm{d}s\, \mathrm{d}x. \end{aligned} \tag{16.8}$$

Combining (16.6) and assumptions (H2) and (H6), (16.8) we get

$$\begin{aligned} U(t) \le U(0) &+ \frac{1}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} a(s) \Delta U(s)\, \mathrm{d}s \\ &+ \frac{1}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} \Big(\int_0^s h(s-\tau) \Delta U(\tau)\, \mathrm{d}\tau \Big)\, \mathrm{d}s \\ &+ \frac{1}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} f(s, U(s))\, \mathrm{d}s \\ &+ \frac{1}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} \bigg(\int_0^s g\big(s, \tau, U(\tau)\big)\, \mathrm{d}\tau \bigg)\, \mathrm{d}s. \end{aligned} \tag{16.9}$$

Let $K = \{U : U \in C(J, \ \mathbb{R}), \ \parallel U(t) - U(0) \parallel \le b\}$ and define an operator $T : C(J, \ \mathbb{R}) \to C(J, \ \mathbb{R})$ by

$$
\begin{aligned}
TU(t) = U(0) &+ \frac{1}{\Gamma(\alpha)}\int_0^t (t-s)^{\alpha-1} a(s)\Delta U(s)\,\mathrm{d}s \\
&+ \frac{1}{\Gamma(\alpha)}\int_0^t (t-s)^{\alpha-1}\Big(\int_0^s h(s-\tau)\Delta U(\tau)\,\mathrm{d}\tau\Big)\,\mathrm{d}s \\
&+ \frac{1}{\Gamma(\alpha)}\int_0^t (t-s)^{\alpha-1} f(s,U(s))\,\mathrm{d}s \\
&+ \frac{1}{\Gamma(\alpha)}\int_0^t (t-s)^{\alpha-1}\bigg(\int_0^s g\big(s,\tau,U(\tau)\big)\,\mathrm{d}\tau\bigg)\,\mathrm{d}s. \qquad (16.10)
\end{aligned}
$$

Clearly, $U(0) \in K$. This means that K is nonempty. From our construction of K, we can say that K is closed and bounded. Now, for any $U_1, U_2 \in K$ and for any $a_1, a_2 \geq 0$ such that $a_1 + a_2 = 1$,

$$
\begin{aligned}
\| \; a_1 U_1(t) + a_2 U_2(t) - U(0) \; \| &\leq a_1 \; \| \; U_1(t) - U(0) \; \| + a_2 \; \| \; U_2(t) - U(0) \; \| \\
&\leq a_1 b + a_2 b = b.
\end{aligned}
$$

Thus $a_1U_1 + a_2U_2 \in K$. Therefore K is a nonempty closed convex set. Next, we move on to verify that T maps K into itself.

$$
\begin{aligned}
\| \; TU(t) - TU(0) \; \| \leq\; & \frac{\lambda_1}{\Gamma(\alpha)}(\|U(0)\| + b)\int_0^t (t-s)^{\alpha-1}\|a(s)\|\,\mathrm{d}s \\
&+ \frac{\lambda_1}{\Gamma(\alpha)}(\|U(0)\| + b)\int_0^t (t-s)^{\alpha-1}\Big(\int_0^s h(s-\tau)\,\mathrm{d}\tau\Big)\,\mathrm{d}s \\
&+ \frac{1}{\Gamma(\alpha)}\int_0^t (t-s)^{\alpha-1}\|f(s,U(s))\|\,\mathrm{d}s \\
&+ \frac{1}{\Gamma(\alpha)}\int_0^t (t-s)^{\alpha-1}\bigg(\int_0^s \|g\big(s,\tau,U(\tau)\big)\|\mathrm{d}\tau\bigg)\,\mathrm{d}s.
\end{aligned}
$$

Making use of Holder’s inequality and the assumptions, for any $U \in K$, we can establish

$$
\begin{aligned}
\| \; TU(t) - TU(0) \; \| \leq\; & \frac{\lambda_1 C_1}{\Gamma(\alpha)}(\|U(0)\| + b)\bigg(\int_0^t \big((t-s)^{\alpha-1}\big)^{\frac{1}{1-\beta}}\,\mathrm{d}s\bigg)^{1-\beta} \\
&+ \frac{\lambda_1 C_4}{\Gamma(\alpha)}(\|U(0)\| + b)\bigg(\int_0^t \big((t-s)^{\alpha-1}\big)^{\frac{1}{1-\beta}}\,\mathrm{d}s\bigg)^{1-\beta} \\
&+ \frac{1}{\Gamma(\alpha)}\int_0^t m_1(s)(t-s)^{\alpha-1}\|U(s)\|\,\mathrm{d}s \\
&+ \frac{1}{\Gamma(\alpha)}\int_0^t (t-s)^{\alpha-1}\bigg(\int_0^s m_2(s,\tau)\|U(s)\|\,\mathrm{d}\tau\bigg)\,\mathrm{d}s
\end{aligned}
$$

$$
\begin{aligned}
&\le \frac{(\|U(0)\|+b)\,\lambda_1 C_1}{\Gamma(\alpha)}\left(\frac{1-\beta}{\alpha-\beta}\right)^{1-\beta} r_1^{\alpha-\beta} + \frac{(\|U(0)\|+b)\,\lambda_1 C_4}{\Gamma(\alpha)}\left(\frac{1-\beta}{\alpha-\beta}\right)^{1-\beta} r_1^{\alpha-\beta} \\
&+ \frac{(\|U(0)\|+b)\,C_2}{\Gamma(\alpha)}\left(\frac{1-\beta}{\alpha-\beta}\right)^{1-\beta} r_1^{\alpha-\beta} + \frac{(\|U(0)\|+b)\,C_3}{\Gamma(\alpha)}\left(\frac{1-\beta}{\alpha-\beta}\right)^{1-\beta} r_1^{\alpha-\beta} \\
&= \frac{(\|U(0)\|+b)\,(\lambda_1(C_1+C_4)+C_2+C_3)}{\Gamma(\alpha)}\left(\frac{1-\beta}{\alpha-\beta}\right)^{1-\beta} r_1^{\alpha-\beta} \\
&\le b, \quad t \in [0, r_1].
\end{aligned}
$$

Now, define a sequence $\{U_k(t)\}$ in K such that

$$U_0(t) = U(0) \text{ and } U_{k+1}(t) = U_k(t), \quad k = 0, 1, 2, \ldots$$

Since K is closed, there exists a subsequence $\{U_{k_i}(t)\}$ of $U_k(t)$ and $\widetilde{U}(t) \in K$ such that

$$\lim_{k_i \to \infty} U_{k_i}(t) = \widetilde{U}(t). \tag{16.11}$$

Then, Lebesgue's dominated convergence theorem yields

$$
\begin{aligned}
\widetilde{U}(t) = \widetilde{U}(0) &- \frac{\lambda_1}{\Gamma(\alpha)}\int_0^t (t-s)^{\alpha-1} a(s)\widetilde{U}(s)\,\mathrm{d}s \\
&- \frac{\lambda_1}{\Gamma(\alpha)}\int_0^t (t-s)^{\alpha-1}\Big(\int_0^s h(s-\tau)\widetilde{U}(\tau)\,\mathrm{d}\tau\Big)\,\mathrm{d}s \\
&+ \frac{1}{\Gamma(\alpha)}\int_0^t (t-s)^{\alpha-1} f(s, \widetilde{U}(s))\,\mathrm{d}s \\
&+ \frac{1}{\Gamma(\alpha)}\int_0^t (t-s)^{\alpha-1}\left(\int_0^s g\big(s, \tau, \widetilde{U}(s)\big)\,\mathrm{d}\tau\right)\mathrm{d}s.
\end{aligned}
$$

Next, we claim that T is continuous.
Step 2. Let $\{U_m(t)\}$ be a converging sequence in K to $U(t)$. Then, for any $\varepsilon > 0$, let

$$\|U_m(t) - U(t)\| \le \frac{\Gamma(\alpha)\varepsilon}{4\lambda_1 C r_1^{\alpha-\beta}}\left(\frac{\alpha-\beta}{1-\beta}\right)^{1-\beta}, \tag{16.12}$$

where $C = \max\{C_1, C_4\}$. By assumption (H2) and (H4),

$$f(t, U_m(t)) \longrightarrow f(t, U(t)) \quad \text{and} \quad g(t, s, U_m(t)) \longrightarrow g(t, s, U(t))$$

for each $t \in [0, r_1]$. Therefore, for any $\varepsilon > 0$, we can take

$$\big\|f(t, U_m(t)) - f(t, U(t))\big\| \le \frac{\alpha\Gamma(\alpha)\varepsilon}{4r_1^{\alpha}}\left(\frac{\alpha-\beta}{1-\beta}\right)^{1-\beta}, \tag{16.13}$$

$$\left\| g(t, s, U_m(t)) - g(t, s, U(t)) \right\| \leq \frac{\Gamma(\alpha)\varepsilon}{4Tr_1^{\alpha}} \left(\frac{\alpha - \beta}{1 - \beta}\right)^{1-\beta}. \tag{16.14}$$

Employing (16.12) and (16.13) and simplifying, we have

$$\begin{aligned}
\|TU_m(t) - TU(t)\| \leq{}& \frac{\lambda_1 C_1}{\Gamma(\alpha)} \left(\frac{1-\beta}{\alpha-\beta}\right)^{1-\beta} r_1^{\alpha-\beta} \|U_m(t) - U(t)\| \\
&+ \frac{\lambda_1 C_4}{\Gamma(\alpha)} \left(\frac{1-\beta}{\alpha-\beta}\right)^{1-\beta} r_1^{\alpha-\beta} \|U_m(t) - U(t)\| \\
&+ \frac{r_1^{\alpha}}{\alpha\Gamma(\alpha)} \left\| f\left(s, U_m(s)\right) - f\left(s, U(s)\right) \right\| \\
&+ \frac{r_1^{\alpha}}{\Gamma(\alpha)} \left(\frac{1-\beta}{\alpha-\beta}\right)^{1-\beta} \int_0^s \left\| g(t, s, U_m(t)) - g(t, s, U(t)) \right\| \mathrm{d}s \\
\leq{}& \varepsilon.
\end{aligned}$$

Since ε can be arbitrarily small, taking limit $m \to \infty$ implies T is continuous.
Step 3. Moreover, for $U \in K$,

$$\begin{aligned}
\| TU(t) \| &\leq \|U(0)\| + \frac{\lambda_1(C_1 + C_4) + C_2 + C_3}{\Gamma(\alpha)} (\|U(0)\| + b) \left(\frac{1-\beta}{\alpha-\beta}\right)^{1-\beta} r_1^{\alpha-\beta} \\
&\leq \|U(0)\| + b.
\end{aligned}$$

Hence, TK is uniformly bounded and so T is completely continuous. At this point, it remains to show that T maps K into an equicontinuous family.
Step 4. Now, let $U \in K$ and $t_1, t_2 \in J$. Then, if $0 < t_1 < t_2 \leq r_1$, by the assumptions (H1)–(H6), we obtain

$$\begin{aligned}
&\| TU(t_1) - TU(t_2) \| \\
&\quad\leq \frac{\lambda_1}{\Gamma(\alpha)} (\|U(0)\| + b) \int_0^{t_1} \left((t_2 - s)^{\alpha-1} - (t_1 - s)^{\alpha-1}\right) \|a(s)\| \,\mathrm{d}s \\
&\quad+ \frac{\lambda_1}{\Gamma(\alpha)} (\|U(0)\| + b) \int_{t_1}^{t_2} (t_2 - s)^{\alpha-1} \|a(s)\| \,\mathrm{d}s \\
&\quad+ \frac{\lambda_1}{\Gamma(\alpha)} (\|U(0)\| + b) \int_0^{t_1} \left((t_2 - s)^{\alpha-1} - (t_1 - s)^{\alpha-1}\right) \left(\int_0^s h(s - \tau) \,\mathrm{d}\tau\right) \mathrm{d}s \\
&\quad+ \frac{\lambda_1}{\Gamma(\alpha)} (\|U(0)\| + b) \int_{t_1}^{t_2} (t_2 - s)^{\alpha-1} \left(\int_0^s h(s - \tau) \,\mathrm{d}\tau\right) \mathrm{d}s \\
&\quad+ \frac{1}{\Gamma(\alpha)} \int_0^{t_1} \left((t_2 - s)^{\alpha-1} - (t_1 - s)^{\alpha-1}\right) \left\| f(s, U(s)) \right\| \mathrm{d}s \\
&\quad+ \frac{1}{\Gamma(\alpha)} \int_{t_1}^{t_2} (t_2 - s)^{\alpha-1} \left\| f(s, U(s)) \right\| \mathrm{d}s \\
&\quad+ \frac{1}{\Gamma(\alpha)} \left\| \int_0^{t_1} \left((t_2 - s)^{\alpha-1} - (t_1 - s)^{\alpha-1}\right) \left(\int_0^s g\left(s, \tau, U(\tau) \,\mathrm{d}\tau\right)\right) \mathrm{d}s \right\|
\end{aligned}$$

$$
\begin{aligned}
&\quad + \frac{1}{\Gamma(\alpha)} \left\| \int_{t_1}^{t_2} (t_2-s)^{\alpha-1}\Big(\int_0^s g\big(s,\tau,U(\tau)\,\mathrm{d}\tau\big)\Big)\,\mathrm{d}s \right\| \\
&\quad \le \frac{\lambda_1 C_1}{\Gamma(\alpha)}(\|U(0)\|+b)\left(\int_0^{t_1}\Big((t_2-s)^{\alpha-1}-(t_1-s)^{\alpha-1}\Big)^{\frac{1}{1-\beta}}\,\mathrm{d}s\right)^{1-\beta} \\
&\quad + \frac{\lambda_1 C_1}{\Gamma(\alpha)}(\|U(0)\|+b)\left(\int_{t_1}^{t_2}((t_2-s)^{\alpha-1})^{\frac{1}{1-\beta}}\,\mathrm{d}s\right)^{1-\beta} \\
&+ \frac{\lambda_1 C_4}{\Gamma(\alpha)}(\|U(0)\|+b)\left(\int_0^{t_1}\Big((t_2-s)^{\alpha-1}-(t_1-s)^{\alpha-1}\Big)^{\frac{1}{1-\beta}}\,\mathrm{d}s\right)^{1-\beta} \\
&+ \frac{\lambda_1 C_4}{\Gamma(\alpha)}(\|U(0)\|+b)\left(\int_{t_1}^{t_2}((t_2-s)^{\alpha-1})^{\frac{1}{1-\beta}}\,\mathrm{d}s\right)^{1-\beta} \\
&+ \frac{(\|U(0)\|+b)}{\Gamma(\alpha)}\left(\int_0^{t_1}\Big((t_2-s)^{\alpha-1}-(t_1-s)^{\alpha-1}\Big)^{\frac{1}{1-\beta}}\,\mathrm{d}s\right)^{1-\beta}\left(\int_0^{t}(m_1(s))^{\frac{1}{\beta}}\,\mathrm{d}s\right)^{\beta} \\
&+ \frac{1}{\Gamma(\alpha)}(\|U(0)\|+b)\left(\int_{t_1}^{t_2}\Big((t_2-s)^{\alpha-1}\Big)^{\frac{1}{1-\beta}}\,\mathrm{d}s\right)^{1-\beta}\left(\int_0^{t}(m_1(s))^{\frac{1}{\beta}}\,\mathrm{d}s\right)^{\beta} \\
&+ \frac{C_3}{\Gamma(\alpha)}(\|U(0)\|+b)\left(\int_0^{t_1}\Big((t_2-s)^{\alpha-1}-(t_1-s)^{\alpha-1}\Big)^{\frac{1}{1-\beta}}\,\mathrm{d}s\right)^{1-\beta} \\
&+ \frac{C_3}{\Gamma(\alpha)}(\|U(0)\|+b)\left(\int_{t_1}^{t_2}((t_2-s)^{\alpha-1})^{\frac{1}{1-\beta}}\,\mathrm{d}s\right)^{1-\beta} \\
&\le \frac{\lambda_1(C_1+C_4)+C_2+C_3}{\Gamma(\alpha)}(\|U(0)\|+b)\left(\int_0^{t_1}\Big((t_2-s)^{\alpha-1}-(t_1-s)^{\alpha-1}\Big)^{\frac{1}{1-\beta}}\,\mathrm{d}s\right)^{1-\beta} \\
&+ \frac{\lambda_1(C_1+C_4)+C_2+C_3}{\Gamma(\alpha)}(\|U(0)\|+b)\left(\int_{t_1}^{t_2}\Big((t_2-s)^{\alpha-1}\Big)^{\frac{1}{1-\beta}}\,\mathrm{d}s\right)^{1-\beta}.
\end{aligned}
$$

Clearly, T maps K into an equicontinuous family of functions and it is noted that T is completely continuous by Ascoli–Arzela theorem. Then, applying Leray–Schauder fixed point theorem, we achieve that T has a fixed point in K which is a solution of (16.2).

16.3.2 Neumann Boundary Condition

Next, our aim is to show the existence of solutions of (16.2) with Neumann boundary condition instead of Dirichlet boundary condition. That is,

$$
\frac{\partial u(x,t)}{\partial n} = 0, \qquad (x,t) \in \partial\Omega \times J, \tag{16.15}
$$

where n is an outward unit normal. Now, we define the function $V(t)$ by

$$V(t) = \frac{\int\limits_{\Omega} u(x,t)\,\mathrm{d}x}{\int\limits_{\Omega} \mathrm{d}x}. \tag{16.16}$$

The following theorem asserts the existence of solution of (16.2) with Neumann boundary conditions (16.15).

Theorem 16.3.2 *Assume that there exists a* $\beta \in (0, \alpha)$ *for some* $0 < \alpha < 1$ *such that (H2) and (H3) hold. For any constant* $b > 0$*, suppose that*

$$r_2 = \min\left\{ T, \left[\frac{\Gamma(\alpha) b}{(\|V(0)\| + b)C_2} \left(\frac{\alpha - \beta}{1-\beta} \right)^{1-\beta} \right]^{\frac{1}{\alpha-\beta}} \right\}.$$

Then, there exists at least one solution for the initial value problem (16.2) on $\Omega \times [0, r_2]$.

Proof In order to prove the existence of solutions of (16.2), it is enough to show that the equation

$$\begin{aligned} V(t) = V(0) &+ \frac{1}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} f(s, V(s))\,\mathrm{ds} \\ &+ \frac{1}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} \left(\int_0^s g\big(s, \tau, V(\tau)\big)\,\mathrm{d}\tau \right) \mathrm{d}s \end{aligned} \tag{16.17}$$

has a solution.

Step 1. Assume $u(x,t)$ to be a solution of (16.2). Then, it follows that $u(x,t)$ is a solution of (16.3). Now, integrating both sides of Eq. (16.3) with respect to $x \in \Omega$, we are led to

$$\begin{aligned} \int_\Omega u(x,t)\,\mathrm{d}x = \int_\Omega u_0(x)\,\mathrm{d}x &+ \frac{1}{\Gamma(\alpha)} \int_\Omega \int_0^t (t-s)^{\alpha-1} a(s) \Delta u(x,s)\,\mathrm{d}s\,\mathrm{d}x \\ &+ \frac{1}{\Gamma(\alpha)} \int_\Omega \int_0^t (t-s)^{\alpha-1} \Big(\int_0^s h(s-\tau) \Delta u(x,\tau)\,\mathrm{d}\tau \Big)\,\mathrm{d}s\,\mathrm{d}x \\ &+ \frac{1}{\Gamma(\alpha)} \int_\Omega \int_0^t (t-s)^{\alpha-1} f(s, u(x,s))\,\mathrm{d}s\,\mathrm{d}x \\ &+ \frac{1}{\Gamma(\alpha)} \int_\Omega \int_0^t (t-s)^{\alpha-1} \left(\int_0^s g\big(s, \tau, u(x,\tau)\big)\,\mathrm{d}\tau \right) \mathrm{d}s\,\mathrm{d}x. \end{aligned} \tag{16.18}$$

Combining Green's identity and the Neumann boundary condition, the assumption (H2), (16.18) can be written as

$$V(t) \leq V(0) + \frac{1}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} f(s, V(s))\, \mathrm{ds} + \frac{1}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} \left(\int_0^s g\big(s, \tau, V(\tau)\big)\, \mathrm{d}\tau \right) \mathrm{d}s. \tag{16.19}$$

Now, an operator $P : C(J, \mathbb{R}) \to C(J, \mathbb{R})$ is defined by

$$PV(t) = V(0) + \frac{1}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} f(s, V(s))\, \mathrm{ds} + \frac{1}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} \left(\int_0^s g\big(s, \tau, V(\tau)\big)\, \mathrm{d}\tau \right) \mathrm{d}s. \tag{16.20}$$

Next, we have to prove that the operator P maps K into itself. From the above equation, we observe that

$$\| \; PV(t) - PV(0) \; \| \leq \frac{1}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} \|f(s, V(s))\|\, \mathrm{ds} + \frac{1}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} \left(\int_0^s \|g\big(s, \tau, V(\tau)\big)\| \mathrm{d}\tau \right) \mathrm{d}s.$$

Then, by using the Holder inequality and the assumptions (H2) and (H3), we obtain

$$\begin{aligned}
\| \; PV(t) - PV(0) \; \| &\leq \frac{1}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} \big\| f(s, V(s)) \big\|\, \mathrm{d}s \\
&\quad + \frac{1}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} \left(\int_0^s \|g(s, \tau, V(\tau))\|\, \mathrm{d}\tau \right) \mathrm{d}s.
\end{aligned}$$

$$\begin{aligned}
&\leq \frac{1}{\Gamma(\alpha)} \int_0^t m_1(s)(t-s)^{\alpha-1} \left(\|V(s)\|\right) \mathrm{d}s \\
&+ \frac{1}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} \left(\int_0^s m_2(s, \tau)\|V(s)\|\, \mathrm{d}\tau \right) \mathrm{d}s \\
&\leq \frac{1}{\Gamma(\alpha)} \left(\|V(0)\| + b\right) \left(\int_0^t \left((t-s)^{\alpha-1}\right)^{\frac{1}{1-\beta}}\, \mathrm{d}s \right)^{1-\beta} \left(\int_0^t (m_1(s))^{\frac{1}{\beta}}\, \mathrm{d}s \right)^{\beta} \\
&+ \frac{1}{\Gamma(\alpha)} \left(\|V(0)\| + b\right) \left(\int_0^t \left((t-s)^{\alpha-1}\right)^{\frac{1}{1-\beta}} \mathrm{d}s \right)^{1-\beta} \left(\int_0^t \left(\int_0^s m_2(s, \tau)\, \mathrm{d}\tau \right)^{\frac{1}{\beta}} \mathrm{d}s \right)^{\beta} \\
&\leq \frac{\left(\|V(0)\| + b\right) C_2}{\Gamma(\alpha)} \left(\frac{1-\beta}{\alpha-\beta} \right)^{1-\beta} r_2^{\alpha-\beta} + \frac{\left(\|V(0)\| + b\right) C_3}{\Gamma(\alpha)} \left(\frac{1-\beta}{\alpha-\beta} \right)^{1-\beta} r_2^{\alpha-\beta} \\
&= \frac{\left(\|V(0)\| + b\right) (C_2 + C_3)}{\Gamma(\alpha)} \left(\frac{1-\beta}{\alpha-\beta} \right)^{1-\beta} r_2^{\alpha-\beta} \\
&\leq b, \quad t \in [\,0, r_2\,].
\end{aligned}$$

Since K is closed, we next define a sequence $\{V_k(t)\}$ in K which has a subsequence $\{V_{k_i}(t)\}$ such that

$$\lim_{k_i \to \infty} V_{k_i}(t) = \widetilde{V}(t). \tag{16.21}$$

Thus, by Lebesgue's dominated convergence, we obtain

$$\begin{aligned}\widetilde{V}(t) = \widetilde{V}(0) &+ \frac{1}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} f\big(s, \widetilde{V}(s)\big)\, \mathrm{d}s \\ &+ \frac{1}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} \left(\int_0^s g\big(s, \tau, \widetilde{V}(s)\big)\, \mathrm{d}\tau \right) \mathrm{d}s.\end{aligned}$$

Now, we intend to show that P is continuous.

Step 2. Let $\{V_m(t)\}$ be a converging sequence in K to $V(t)$. Therefore, for any $\varepsilon > 0$ and for each $t \in [\,0, r_2\,]$, let

$$\big\| f\big(t, V_m(t)\big) - f\big(t, V(t)\big) \big\| \le \frac{\alpha \Gamma(\alpha) \varepsilon}{2 r_2^{\alpha}} \left(\frac{\alpha - \beta}{1-\beta} \right)^{1-\beta}, \tag{16.22}$$

$$\big\| g(t, s, V_m(t)) - g(t, s, V(t)) \big\| \le \frac{\Gamma(\alpha) \varepsilon}{2 T r_2^{\alpha}} \left(\frac{\alpha - \beta}{1-\beta} \right)^{1-\beta}. \tag{16.23}$$

Making use of (16.13) and then simplifying, we have

$$\begin{aligned}\| P V_m(t) - P V(t) \| \le{}& \frac{r_2^{\alpha}}{\alpha \Gamma(\alpha)} \left(\frac{1-\beta}{\alpha - \beta} \right)^{1-\beta} \| f\,(s, V_m(s)) - f\,(s, V(s))\, \| \\ &+ \frac{r_2^{\alpha}}{\Gamma(\alpha)} \left(\frac{1-\beta}{\alpha - \beta} \right)^{1-\beta} \int_0^s \big\| g(t, s, V_m(t)) - g(t, s, V(t)) \big\|\, \mathrm{d}s \\ \le{}& \varepsilon.\end{aligned}$$

Taking limit $m \to \infty$, for sufficiently small ε, P is continuous.

Step 3. Moreover, for $V \in K$,

$$\begin{aligned}\|\, P V(t)\, \| &\le \| V(0) \| + \frac{(C_2 + C_3)}{\Gamma(\alpha)} (\| V(0) \| + b) \left(\frac{1-\beta}{\alpha-\beta} \right)^{1-\beta} r_2^{\alpha-\beta} \\ &\le \| V(0) \| + b.\end{aligned}$$

Hence, PK is uniformly bounded. Now, it remains to show that P maps K into an equicontinuous family.

Step 4. Now, let $V \in K$ and $t_1, t_2 \in J$. Then, if $0 < t_1 < t_2 \le r_2$, by the assumptions (H2) and (H3), we obtain

$$
\begin{aligned}
&\| PV(t_1) - PV(t_2) \| \\
&\quad \leq \frac{1}{\Gamma(\alpha)} \int_0^{t_1} \Big((t_2 - s)^{\alpha-1} - (t_1 - s)^{\alpha-1} \Big) \big\| f(s, V(s)) \big\| \, \mathrm{d}s \\
&\quad + \frac{1}{\Gamma(\alpha)} \int_{t_1}^{t_2} (t_2 - s)^{\alpha-1} \big\| f(s, V(s)) \big\| \, \mathrm{d}s \\
&\quad + \frac{1}{\Gamma(\alpha)} \left\| \int_0^{t_1} \Big((t_2 - s)^{\alpha-1} - (t_1 - s)^{\alpha-1} \Big) \Big(\int_0^s g\big(s, \tau, V(\tau) \, \mathrm{d}\tau\big) \Big) \, \mathrm{d}s \right\| \\
&\quad + \frac{1}{\Gamma(\alpha)} \left\| \int_{t_1}^{t_2} (t_2 - s)^{\alpha-1} \Big(\int_0^s g\big(s, \tau, V(\tau) \, \mathrm{d}\tau\big) \Big) \, \mathrm{d}s \right\| \\
&\quad \leq \frac{(\|V(0)\| + b)}{\Gamma(\alpha)} \left(\int_0^{t_1} \Big((t_2 - s)^{\alpha-1} - (t_1 - s)^{\alpha-1} \Big)^{\frac{1}{1-\beta}} \mathrm{d}s \right)^{1-\beta} \left(\int_0^t (m_1(s))^{\frac{1}{\beta}} \, \mathrm{d}s \right)^{\beta} \\
&\quad + \frac{(\|V(0)\| + b)}{\Gamma(\alpha)} \left(\int_{t_1}^{t_2} ((t_2 - s)^{\alpha-1})^{\frac{1}{1-\beta}} \, \mathrm{d}s \right)^{1-\beta} \left(\int_0^t (m_1(s))^{\frac{1}{\beta}} \, \mathrm{d}s \right)^{\beta} \\
&\quad + \frac{1}{\Gamma(\alpha)} (\|V(0)\| + b) \left(\int_0^{t_1} \Big((t - s)^{\alpha-1} \Big)^{\frac{1}{1-\beta}} \mathrm{d}s \right)^{1-\beta} \left(\int_0^{t_1} \Big(\int_0^s m_2(s, \tau) \, \mathrm{d}\tau \Big)^{\frac{1}{\beta}} \mathrm{d}s \right)^{\beta} \\
&\quad + \frac{1}{\Gamma(\alpha)} (\|V(0)\| + b) \left(\int_{t_1}^{t_2} \Big((t - s)^{\alpha-1} \Big)^{\frac{1}{1-\beta}} \mathrm{d}s \right)^{1-\beta} \left(\int_{t_1}^{t_2} \Big(\int_0^s m_2(s, \tau) \, \mathrm{d}\tau \Big)^{\frac{1}{\beta}} \mathrm{d}s \right)^{\beta} \\
&\quad \leq \frac{C_2}{\Gamma(\alpha)} (\|V(0)\| + b) \left(\int_0^{t_1} ((t_2 - s)^{\alpha-1} - (t_1 - s)^{\alpha-1})^{\frac{1}{1-\beta}} \, \mathrm{d}s \right)^{1-\beta} \\
&\quad + \frac{C_2}{\Gamma(\alpha)} (\|V(0)\| + b) \left(\int_{t_1}^{t_2} ((t_2 - s)^{\alpha-1})^{\frac{1}{1-\beta}} \, \mathrm{d}s \right)^{1-\beta} \\
&\quad + \frac{C_3}{\Gamma(\alpha)} (\|V(0)\| + b) \left(\int_0^{t_1} \Big((t_2 - s)^{\alpha-1} - (t_1 - s)^{\alpha-1} \Big)^{\frac{1}{1-\beta}} \mathrm{d}s \right)^{1-\beta} \\
&\quad + \frac{C_3}{\Gamma(\alpha)} (\|V(0)\| + b) \left(\int_{t_1}^{t_2} ((t_2 - s)^{\alpha-1})^{\frac{1}{1-\beta}} \, \mathrm{d}s \right)^{1-\beta} .
\end{aligned}
$$

Thus, P maps K into an equicontinuous family of functions. Then, as in the previous case, from Leray–Schauder fixed point theorem, we conclude that P has a fixed point in K which is a solution of (16.2).

Conclusion

In this chapter, we consider fractional integrodifferential equation describing the motion of fractional Maxwell fluid with zero pressure. This equation gets attention from the fact that the fractional derivatives are used to depict the viscoelasticity phenomena with little amount of constraints. Since this equation has a positive kernel with diffusion term, this is different from the integrodifferential equation considered in [3]. Further, our equation is not a particular case of the equation discussed in [3].

Acknowledgements The authors are thankful to the referees for their valuable suggestions for the improvisation of the paper. The first author is thankful for the financial support under UGC-SAP-DRS-III project (Grant No. G2/10810/2017) and the second author is thankful to UGC-BSR Faculty fellowship for carrying out this work.

References

1. K.S. Akiladevi, K. Balachandran, On fractional delay integrodifferential equations with four-point multiterm fractional boundary conditions. Acta Math. Univ. Comen. **86**, 187-204 (2017)
2. K.S. Akiladevi, K. Balachandran, J.K. Kim, Existence results for neutral fractional integrodifferential equations with fractional integral boundary conditions. Nonlinear Funct. Anal. Appl. **19**, 251-270 (2014)
3. A. Akilandeeswari, K. Balachandran, J.J. Trujillo, M. Rivero, On the solutions of partial integrodifferential equations of fractional order. Tiblisi Math. J. **10**, 19–29 (2017)
4. A. Akilandeeswari, K. Balachandran, N. Annapoorani, Existence of solutions of fractional partial integrodifferential equations with Neumann boundary condition. Nonlinear Funct. Anal. Appl. **22**, 711–722 (2017)
5. N. Annapoorani, K. Balachandran, Existence of solutions of partial neutral integrodifferential equations. Carpathian J. Math. **26**, 134–145 (2010)
6. A. Arikoglu, I. Ozkol, Solution of fractional integrodifferential equations by Fourier transform method. Chaos Solitons Fractals **40**, 521–529 (2009)
7. I. Aziz, I. Khan, Numerical solution of partial integrodifferential equations of diffusion type. Math. Probl. Eng. **2017**, 1–11 (2017)
8. K. Balachandran, N. Annapoorani, Existence results for impulsive neutral evolution integrodifferential equations with infinite delay. Nonlinear Anal.: Hybrid Syst. **3**, 674–684 (2009)
9. K. Balachandran, J.J. Trujillo, The nonlocal Cauchy problem for nonlinear fractional integrodifferential equations in Banach spaces. Nonlinear Anal. **72**, 4587–4593 (2010)
10. M. Caputo, Linear models of dissipation whose Q is almost frequency independent-II. Geophys. J. R. Astron. Soc. **13**, 529–539 (1967)
11. A. Carpinteri, F. Mainardi, *Fractals and Fractional Calculus in Continuum Mechanics* (Springer, New York, 1997)
12. M. De La Sena, V. Hedayati, Y.G. Atani, S. Rezapour, The existence and numerical solution for a k-dimensional system of multi-term fractional integro-differential equations. Nonlinear Anal.: Model. Control **22**, 188–209 (2017)
13. L.C. Evans, *Partial Differential Equations* (American Mathematical Society, Providence, 1998)
14. M.A. Ezzat, Theory of fractional order in generalized thermoelectric MHD. Appl. Math. Model. **35**, 4965–4978 (2011)
15. K.M. Furati, N.E. Tatar, Behavior of solutions for a weighted Cauchy-type fractional differential problem. J. Fract. Calc. **28**, 23–42 (2005)
16. V.D. Gejji, H. Jafari, Boundary value problems for fractional diffusion-wave equation. Australian J. Math. Anal. Appl. **3**, 1–8 (2006)
17. M.A.E. Herzallah, A.M.A. El-Sayed, D. Baleanu, On the fractional order diffusion-wave process. Romanian J. Phys. **55**, 274–284 (2010)
18. R. Hilfer, *Applications of Fractional Calculus in Physics* (World Scientific Publishing, Singapore, 2000)
19. I.S. Jesus, J.A.T. Machado, J.B. Cunha, Fractional electrical impedances in botanical elements. J. Vib. Control **14**, 1389–1402 (2008)
20. B. Kamalapriya, K. Balachandran, N. Annapoorani, Existence results for fractional integrodifferential equations. Nonlinear Funct. Anal. Appl. **22**, 641–653 (2017)
21. A.A. Kilbas, H.M. Srivasta, J.J. Trujillo, *Theory and Applications of Fractional Differential Equations* (Elsevier, Amstrdam, 2006)

22. X. Li, C. Xu, A space-time spectral method for the time-fractional diffusion equation. SIAM J. Numer. Anal. **47**, 2108–2131 (2009)
23. A.D. Obembe, H.Y. Al-Yousef, M.E. Hosssin, S.A. Abu-Khamsin, Fractional derivatives and their applications in reservoir engineering problems: a review. J. Petrol. Sci. Eng. **157**, 312–327 (2017)
24. M.D. Ortigueira, *Fractional Calculus for Scientists and Engineers* (Springer, New York, 2011)
25. Z. Ouyang, Existence and uniqueness of the solutions for a class of nonlinear fractional order partial differential equations with delay. Comput. Math. Appl. **61**, 860–870 (2011)
26. M. Ozalp, I. Koca, A fractional order nonlinear dynamical model of interpersonal relationships. Adv. Differ. Equ. **2012**, 1–7 (2012)
27. L.M. Petrovic, D.T. Spasic, T.M. Atanackovic, On a mathematical model of a human root dentin. Dent. Mater. **21**, 125–128 (2005)
28. I. Podlubny, *Fractional Differential Equations* (Academic Press, New York, 1999)
29. S.Z. Rida, A.M.A. El-Sayed, A.A.M. Arfa, Effect of bacterial memory dependent growth by using fractional derivatives reaction diffusion chemotactic model. J. Stat. Phys. **140**, 797–811 (2010)
30. B. Ross, A brief history and exposition of the fundamental theory of fractional calculus. *Fractional Calculus and Its Applications*, vol. 57 (1975), pp. 1–36
31. S.G. Samko, A.A. Kilbas, O.I. Marichev, *Fractional Integrals and Derivatives: Theory and Applications* (Gordon and Breach Science Publishers, Yverdon, 1993)
32. H. Schiessel, R. Metzler, A. Blumen, T.F. Nonnenmacher, Generalized viscoelastic models: their fractional equations with solutions. J. Phys. A: Math. Gen. **28**, 6567–6584 (1995)
33. W.R. Schneider, W. Wyss, Fractional diffusion and wave equations. J. Math. Phys. **30**, 134–144 (1989)
34. X. Su, Boundary value problem for a coupled system of nonlinear fractional differential equations. Appl. Math. Lett. **22**, 64–69 (2009)
35. V.S. Vladimirov, *Equations of Mathematical Physics* (Marcel Dekker, New York, 1981)
36. L. Zhang, B. Ahmad, G. Wang, R.P. Agarwal, M. Al-Yami, W. Shammakh, Nonlocal integrodifferential boundary value problem for nonlinear fractional differential equations on unbounded domain. Abstr. Appl. Anal. **2013**, 1–5 (2013)

Chapter 17
Mathematical Study on Human Cells Interaction Dynamics for HIV-TB Co-infection

Suman Dolai, Amit Kumar Roy and Priti Kumar Roy

Abstract Co-infection of Tuberculosis (caused by Mycobacterium tuberculosis bacteria) and HIV (caused by Human Immunodeficiency virus) remains a global burden on public health system and poses particular diagnostic and therapeutic challenges. Due to co-infection, HIV speeds up the progression from latent to active TB and TB bacteria also accelerates the progress of HIV infection which ultimately leads to serious condition in individuals. In this research work, we formulate a six compartment mathematical model on the HIV-TB co-infection dynamics incorporating Macrophage (active and infected), T-cell (active and infected), Virus, and Bacteria population. Moreover, we explore the accelerating effect of both pathogens on each other in mathematical perceptive. Our analytical study reveals the conditions for the persistence of co-infection and also validates the stability criteria of equilibrium points for the disease. We also evaluate the disease-free condition using next generation method, expressed by the basic reproductive ratio (R_0). Moreover, our analytical and numerical simulations manifest the influence of certain key parameters on the threats posed by the impact of HIV-TB co-infection.

Keywords HIV · TB · Co-infection · Basic reproductive ratio · Sensitivity analysis

17.1 Introduction

Tuberculosis (TB) is one of the main reasons for human death among all infectious diseases; close to 2 million people passed away at the end of 2016 [1]. World Health Organization (WHO) reported that one-third of the world's total population live with TB latency but only 5–10% individuals will advance with active TB disease [2–4]. On the other hand, according to UNAIDS; 1.8 million people became infected with HIV

S. Dolai · A. K. Roy · P. K. Roy (✉)
Department of Mathematics, Centre for Mathematical Biology and Ecology,
Jadavpur University, Kolkata 700032, India
e-mail: pritiju@gmail.com

P. Manchanda et al. (eds.), *Mathematical Modelling, Optimization, Analytic and Numerical Solutions*, Industrial and Applied Mathematics,
https://doi.org/10.1007/978-981-15-0928-5_17

and approximately 1.0 million died globally due to AIDS in the last year [5]. AIDS is a syndrome caused by the virus HIV which alters the immune system and makes people much more vulnerable to infections [6]. Due to the damaged immune system, the HIV infected individuals can become afflict with active TB disease within weeks to months which is 20–30 times greater than among those without HIV infection [7]. Estimated by the WHO in 2016, one million new TB cases ascended among people who were HIV positive and about 374,000 people died worldwide [8].

Human immune system is the prime shield that is accountable to defense against foreign body particle (virus, bacteria, and parasites); made by different organs, cells, and cytokines [9]. The two most important immune cells are Macrophage (Bone marrow derived) and T-cell (derived from human cord blood hematopoietic stem cells) which are different types of white blood cell [10–12]. Like all retroviruses, HIV-1 attacks T-cells as well as macrophages with two types of CD4+ receptors (CXCR4 and CCR5) and infects those two key immune cells [2, 13, 14]. This infection process is categorized into three steps: First, HIV attacks to the body immune system through transmission process; next, HIV exists in latent stage; and finally, there is AIDS when individual has extreme viral load [15]. Moreover, Cytotoxic T lymphocytes (CD8+ T-cells) are believed to play a major role in killing virus levels in asymptomatic period (primary stage) of HIV infection [16]. Furthermore, Mycobacterium Tuberculosis (TB bacteria) progresses so slowly that it could be misdiagnosed initially and Macrophages act as a primary reservoir cell for this bacterial growth. After activation by T-cell, Macrophages engulf TB bacteria by detecting them with toll-like receptors [2, 17] and kill them by producing reactive oxygen species (nitric-oxide) [2, 3]. These TB bacteria can survive inside the macrophage; they replicate themselves more and more until the macrophages burst and attain the active stage [2, 18]. During co-infection, the activation of TB from latent stage is more prominent due to huge loss of macrophage cell in presence of HIV. HIV decreases the ability of macrophage to produce the nitric oxide and also distract to engulf TB bacteria due to loss of CD4+ T-cell [2, 19, 20]. Tumor necrosis factor alpha ($\text{TNF} - \alpha$) is the pro-inflammatory cytokine released for controlling TB bacterial growth which ultimately helps to enhance HIV replication in co-infected individuals [2, 21]. Upregulation of co-receptors (CXCR4 and CCR5) expressions on CD4+ T-cell and downregulated CCL5 ligand by TB bacteria permit to increase the virus replication [2, 22]. In this way, both virus and bacteria give a great positive impact on each other to develop active stage in a short time.

During last few decades, many clinical and experimental studies have been performed on HIV-TB co-infection disease. Pawlowski et al. suggested that the risk of developing TB from latent to active is approx. 20 fold during co-infection [2]. Selwyn et al. investigated that seven of the eight cases of tuberculosis occurred in HIV infected individuals through a prior positive PPD test [23]. Diedrich et al. reported that there was a 5–160 fold increase in plasma viral titers during acute infection with TB, which are 2.5 times higher in HIV+ individuals upon TB diagnosis [24]. Though many clinical and experimental studies have been done on HIV-TB co-infection but its mathematical outlook is highly anticipated.

In this direction, some mathematical works have been done related to co-infection of HIV and TB from epidemiological and cell dynamical aspect. Naresh et al. investigated the effect of TB on HIV infected people by formulating a four-dimensional mathematical model focusing on population dynamics and discussed about some key parameters on spread of the disease [25]. Boralin G. et al. discussed the effect of treatment on the TB in HIV/TB co-infection by constructing a six-dimensional epidemiological model [26] and Cristiana J. Silva et al. extended the work in which they investigated the treatment of TB, HIV, and TB/HIV co-infection separately [27]. In recent years, some mathematical models focusing on various treatment process of HIV/TB co-infection have been established [28, 29]. Roger et al. developed a eight-dimensional mathematical model mentioning the joint dynamics of HIV and TB in a pseudo-competitive environment, at the population level [30]. However, all the above articles are based on population dynamics, but our aim is to investigate the dynamical behavior of human immune cells when two diseases coexist in human body. In this direction, Kirschner et al. first instigated a four-dimensional mathematical model based on cell–cell interaction dynamics (macrophage, T-cell, virus, and bacteria) [18] and Magombedze et al. extended the work by considering two types of T-cells and macrophage cells (uninfected and infected) [31]. However, these works fail to demonstrate about the acceleration hypothesis between virus and bacteria during co-infection in mathematical conjecture. In this research work, we have proposed a six-dimensional mathematical model based on cell dynamical system introducing two acceleration parameters which indicates the cross talk between virus and bacteria. Our analytical and numerical studies show that how the key parameters (acceleration parameters) play a crucial role in disease pathogenesis.

This article is started with a general introductory section and in Sect. 17.2, we have introduced our models and analyzed the model properties. In Sect. 17.3, the equilibrium points, stability conditions, and sensitivity analysis have been discussed and numerical simulations with varying parameters are given in Sect. 17.4. In Sect. 17.5, we summarize the results of our analysis with some concluding remarks.

17.2 The Model

17.2.1 The Deterministic Model

We develop a six-dimensional mathematical model of HIV-TB co-infection by introducing different cells to reflect the cell-biological relationships in expressing the disease. Here, $M(t)$, $M_i(t)$, $T(t)$, $T_i(t)$, $V(t)$, and $B(t)$ represent the densities of Macrophage cell, Infected Macrophage cell, T-cell, Infected T-cell, Virus population, and Bacteria population at any time t, respectively. Now, the deterministic mathematical model is given as follows:

$$\begin{aligned}
\frac{dM}{dt} &= S_M - \lambda_1 BM - \lambda_2 MV - \mu_M M,\\
\frac{dM_i}{dt} &= \lambda_1 BM + \lambda_2 MV - a_1 M_i - a_2 M_i - \mu_{M_i} M_i,\\
\frac{dT}{dt} &= S_T - \lambda_3 TV - \mu_T T,\\
\frac{dT_i}{dt} &= \lambda_3 TV - a_3 T_i - \mu_{T_i} T_i,\\
\frac{dV}{dt} &= N_3 a_3 T_i + N_2 a_2 M_i + \gamma_1 VB - k_1 VT - \mu_V V,\\
\frac{dB}{dt} &= N_1 a_1 M_i + \gamma_2 VB - k_2 BM - \mu_B B,
\end{aligned} \tag{17.1}$$

where $M(0) > 0$, $M_i(0) \geq 0$, $T(0) > 0$, $T_i(0) \geq 0$, $V(0) \geq 0$, and $B(0) \geq 0$ are the initial conditions.

In system (17.1), the first equation illustrates the growth dynamics of macrophage. Here, S_M is the constant production of macrophages from bone marrow through thymus. λ_1 and λ_2 are symbolized the rate at which MTB and HIV infect the macrophages, respectively. The last term of the first equation expresses the death term of macrophage and μ_M is the natural decay rate. The second equation of system (17.1) represents the growth rate of infected macrophage. The diseases (HIV and TB both) replicate inside the infected cells more and more until those cells burst and after bursting of host cells, diseases come out from the cell. Here, a_1 and a_2 indicate the bursting rate of infected macrophage cell due to bacteria and virus, respectively. The natural death rate of infected macrophage is denoted by μ_{M_i}.

In the third equation, S_T and μ_T indicate the constant accumulation rate and mortality rate of T-cell. At a rate λ_3 HIV attacks T-cell and convert into infected one. The fourth equation stands for the growth equation of infected T-cell. a_3 represents the bursting rate of infected T-cell and μ_{T_i} denotes death rate.

The last two equations denote the dynamics of virus and bacteria populations. In the fifth equation, N_3 and N_2 specify the virus production rate due to destroying of infected T-cell and infected macrophage, respectively. k_1 is the rate at which virus killed by T-cell (specially killed by CD8+ T-cell). N_1 is the production rate of bacteria for bursting of MTB specific infected macrophage. k_2 represents the killing rate of bacteria by macrophages. γ_1 and γ_2 represent the accelerating growth rate of virus and bacteria by one another, respectively. μ_V and μ_B denote the death rate of virus and bacteria, respectively.

17.2.2 Boundedness

Let $\Omega = \{(M, M_i, T, T_i, V, B) \in R^6 \colon 0 < (M + M_i)(t) \leq \frac{S_M}{\mu_1}, 0 < (T + T_i)(t) \leq \frac{S_T}{\mu_2}$ and $0 < (V + B)(t) \leq \frac{\mu}{2\gamma} - a\}$ is a positive invariant subset of R^6.

The right-hand sides of system (17.1) are smooth and nonlinear functions of the variable M, M_i, T, T_i, V, and B and also the parameters are always nonnegative. Henceforth, the system dynamics is assuredly bounded in the positive octant and the considered cells concentration are less than a pre-assumed quantity. In the following theorem, we wish to clarify that the solution of the dynamical system is bounded.

Theorem 17.1 *The solutions of the system (17.1) with initial conditions satisfy* $M(t)>0$*,* $M_i(t)>0$*,* $T(t)>0$*,* $T_i(t)>0$*,* $V(t)>0$*, and* $B(t)>0$ *for all* $t > 0$*. The region* $\Omega \subset R^6_+$ *is positively invariant and attracting with respect to system (17.1).*

Proof Adding first two equation of our mathematical model, we get

$$\frac{d(M+M_i)}{dt} = S_M - (a_1 + a_2 + \mu_{M_i})M_i - \mu_M M.$$

From this equation, it follows that

$$\frac{d(M+M_i)}{dt} \leq S_M - \mu_1(M + M_i),$$

where $\mu_1 = min\{(a_1 + a_2 + \mu_{M_i}), \mu_M\}$.

Now, solving the above inequality, we get

$$(M + M_i)(t) \leq \frac{S_M}{\mu_1} + \left(\frac{S_M}{\mu_1} - M(0)\right) \exp^{-\mu_1 t}.$$

For long time interval, we also obtain $(M + M_i)(t) \leq \frac{S_M}{\mu_1}$, the maximum value of active and infected macrophage present in the case of co-infection.

Again we add third and fourth equation of our model and taking $\mu_2 = min\{\mu_T, (a_3 + \mu_{T_i})\}$, we get

$$\frac{d(T+T_i)}{dt} \leq S_T - \mu_2(T + T_i).$$

Now, solving the above inequality, we get

$$(T + T_i)(t) \leq \frac{S_T}{\mu_2} + \left(\frac{S_T}{\mu_2} - T(0)\right) \exp^{-\mu_2 t}.$$

Now,we get $(T + T_i)(t) \leq \frac{S_T}{\mu_2}$ the maximum value of active and infected T-cell. Similarly, using the maximum value of M, M_i, T, and T_i cells, we also get from the fifth and sixth equation of our model:

$$\begin{aligned}
\frac{d(V+B)}{dt} &= N_1 a_1 M_i + N_2 a_2 M_i + N_3 a_3 T_i + \gamma_1 VB + \gamma_2 VB - k_1 VT - k_2 BM - \mu_V V - \mu_B B,\\
&\leq X + \gamma_1 VB + \gamma_2 VB - \mu_v V - \mu_B B,\\
&\leq X + \gamma VB + \gamma VB + \gamma V^2 + \gamma B^2 - \mu(V + B),
\end{aligned}$$

where $X = \frac{N_3 a_3 S_T}{\mu_2} + \frac{N_2 a_2 S_M}{\mu_1} + \frac{N_1 a_1 S_M}{\mu_1}$, $\mu = min\{\mu_V,\ \mu_B\}$ and $\gamma = max\{\gamma_1, \gamma_2\}$.

Now, solving the above inequality, we get the threshold value of virus and bacteria population as follows:
$(V + B)(t) \leq \frac{\mu}{2\gamma} - a$, where $a^2 = \frac{\mu^2}{4\gamma^2} - \frac{X}{\gamma}$ and $\frac{\mu}{2\gamma} - a > 0$.

Hence, the system is bounded in the region $\Omega \subset R^6$ with the initial conditions $M(t){>}0,\ M_i(t){\geq}0,\ T(t){>}0,\ T_i(t){\geq}0,\ V(t){\geq}0$, and $B(t){\geq}0$.

17.3 Equilibrium Analysis

The endemic equilibrium $E^* = (M^*, M_i^*, T^*, T_i^*, V^*, B^*)$ is obtained by setting equation of the system to zero. Then, the values of M^*, M_i^*, T^*, T_i^*, V^*, B^* are given follows:
$M^* = \frac{(a_1+a_2+\mu_{M_i})M_i^*}{\lambda_1 B^* + \lambda_2 V^*}$,
$T^* = \frac{(a_3+\mu_{T_i})T_i^*}{V^*}$,
$V^* = \frac{N_3 a_3 T_i^* + N_2 a_2 M_i^*}{k_1 T^* + \mu_V - \gamma_1 B^*}$,
$B^* = \frac{N_1 a_1 M_i^*}{k_2 M^* + \mu_B - \gamma_2 V^*}$,
where $T_i^* = \frac{\lambda_3 T^* V^*}{a_3+\mu_{T_i}}$ and $M_i^* = \frac{\lambda_1 B^* M^* + \lambda_2 M^* V^*}{a_1+a_2+\mu_{M_i}}$.

Now, the disease-free equilibrium point is $E_0 = (M_1, 0, T_1, 0, 0, 0)$ where $M_1 = \frac{S_M}{\mu_M}$ and $T_1 = \frac{S_T}{\mu_T}$.

17.3.1 Stability of the Endemic Equilibrium

The Jacobian matrix for the endemic equilibrium of model system (17.1) is given by

$$J(E^*) = \begin{bmatrix} -\lambda_1 B^* - \lambda_2 V^* - \mu_M & 0 & 0 & 0 & -\lambda_2 M^* & -\lambda_1 M^* \\ \lambda_1 B^* + \lambda_2 V^* & -a_1 - a_2 - \mu_{M_i} & 0 & 0 & \lambda_2 M^* & \lambda_1 M^* \\ 0 & 0 & \lambda_3 V^* - \mu_T & 0 & \lambda_3 T^* & 0 \\ 0 & 0 & \lambda_3 V^* & -a_3 - \mu_{T_i} & \lambda_3 T^* & 0 \\ 0 & N_2 a_2 & -k_1 V^* & N_3 a_3 & \gamma_1 B^* - k_1 T^* - \mu_V & \gamma_1 V^* \\ -k_2 B^* & N_1 a_1 & 0 & 0 & \gamma_2 B^* & \gamma_2 V^* - k_2 M^* - \mu_B \end{bmatrix}$$

Let, element of the above matrix $J(E^*)$ are in the form of a_{ij} where $\{i$ and $j \in (1, 2, \ldots, 6)\}$.

The characteristic polynomial of the above matrix is

$$det(J - uI_6) = u^6 + Au^5 + Bu^4 + Cu^3 + Du^2 + Eu + F.$$

Here, A, B, C, D, E, and F are coefficients of the above polynomial. See the Appendix for the value of these coefficients.

Let $H_1 = \begin{pmatrix} A \end{pmatrix}$, $H_2 = \begin{pmatrix} A & 1 \\ 0 & B \end{pmatrix}$, $H_3 = \begin{pmatrix} A & 1 & 0 \\ C & B & A \\ 0 & 0 & C \end{pmatrix}$, $H_4 = \begin{pmatrix} A & 1 & 0 & 0 \\ C & B & A & 1 \\ 0 & D & C & B \\ 0 & 0 & 0 & D \end{pmatrix}$,

$H_5 = \begin{pmatrix} A & 1 & 0 & 0 & 0 \\ C & B & A & 1 & 0 \\ E & D & C & B & A \\ 0 & 0 & E & D & C \\ 0 & 0 & 0 & 0 & E \end{pmatrix}$, and $H_6 = \begin{pmatrix} A & 1 & 0 & 0 & 0 & 0 \\ C & B & A & 1 & 0 & 0 \\ F & E & D & C & B & A \\ 0 & F & E & D & C & B \\ 0 & 0 & 0 & F & E & D \\ 0 & 0 & 0 & 0 & 0 & F \end{pmatrix}$ are all Hurwitz matrix.

Lemma *All the roots of the characteristic equation are negative or negative real part if the determinants of all the Hurwitz matrices are positive, i.e., $det(H_j) > 0$, $j = 1, 2 \ldots 6$. Thus, from the Routh–Hurwitz criterion [32], the system is asymptotically stable if $det(H_j) > 0, \quad j = 1, 2 \ldots 6$.*

17.3.2 Reproduction Number

$a = a_1 + a_2 + \mu_{M_i}$,
$b = a_3 + \mu_{T_i}$,
$c = k_1 T_1 + \mu_V$ and
$d = k_2 M_1 + \mu_B$ are some preassigned parameters.

The linearisation of the second, fourth, fifth, and sixth equation of the model at the disease-free equilibrium E_0 can be written as $\frac{dY}{dt} = (F - V)Y$, where $Y = [M_i, T_i, V, B]^T$,

$$F = \begin{pmatrix} 0 & 0 & \lambda_2 M_1 & \lambda_1 M_1 \\ 0 & 0 & \lambda_3 T_1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \text{ and } V = \begin{pmatrix} a & 0 & 0 & 0 \\ 0 & b & 0 & 0 \\ -N_2 a_2 & -N_3 a_3 & c & 0 \\ -N_1 a_1 & 0 & 0 & d \end{pmatrix}.$$

The basic reproduction number, R_0 is determined by the method of next generation matric (van den Driessche and Watmough, 2002). Therefore, to find R_0, we must find the dominant eigenvalue of FV^{-1} where $FV^{-1} =$

$$\begin{bmatrix} \frac{\lambda_2 N_2 a_2 M_1}{ac} + \frac{\lambda_1 N_1 a_1 M_1}{ad} & \frac{\lambda_2 N_3 a_3 M_1}{bc} & \frac{\lambda_2 M_1}{c} & \frac{\lambda_1 M_1}{d} \\ \frac{\lambda_3 N_2 a_2 T_1}{ac} & \frac{\lambda_3 N_3 a_3 T_1}{bc} & \frac{\lambda_3 T_1}{c} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

The characteristic equation of the above matrix is given by
$\lambda^2(\lambda^2 - P\lambda + Q) = 0$,
where $P = \frac{\lambda_1 N_1 a_1 M_1}{ad} + \frac{\lambda_2 N_2 a_2 M_1}{ac} + \frac{\lambda_3 N_3 a_3 T_1}{bc}$
and $Q = \frac{\lambda_1 N_1 a_1 M_1 \lambda_3 N_3 a_3 T_1}{abcd}$.
The dominant eigenvalue is denoted by $\rho(FV^{-1})$, which is also R_0. Hence,
$R_0 = \frac{P + (p^2 - 4Q)^{1/2}}{2}$.

17.3.3 Stability of Disease-Free Equilibrium

Theorem 17.2

$$\begin{aligned}
A_3 &= a+b+c+d;\\
A_2 &= ab+ac+ad+bc+bd+cd-\lambda_1 a_1 N_1 M_1-\lambda_2 a_2 N_2 M_1-\lambda_3 a_3 N_3 T_1;\\
A_1 &= abc+abd+acd+bcd-(b+c)\lambda_1 a_1 N_1 M_1-(b+d)\lambda_2 a_2 N_2 M_1-(a+d)\lambda_3 a_3 N_3 T_1;\\
A_0 &= abcd+\lambda_1 a_1 N_1 M_1 \lambda_3 a_3 N_3 T_1-bc\lambda_1 a_1 N_1 M_1-bd\lambda_2 a_2 N_2 M_1-ad\lambda_3 a_3 N_3 T_1.
\end{aligned}$$

If $H = min\{(k_1 - \lambda_3 N_3), (\mu_B - \lambda_1 N_1), (\mu_V - \lambda_2 N_2)$, and $(A_3 A_2 A_1 - A_1^2 - A_3^2 A_0)\}{>}0$ then the disease-free equilibrium is asymptotically stable.

Proof The Jacobian matrix at the disease-free equilibrium of model system (17.1) is given by $J(E_0) = \begin{bmatrix} -\mu_M & 0 & 0 & 0 & -\lambda_2 M_1 & -\lambda_1 M_1 \\ 0 & -a_1 - a_2 - \mu_{M_i} & 0 & 0 & \lambda_2 M_1 & \lambda_1 M_1 \\ 0 & 0 & -\mu_T & 0 & -\lambda_3 T_1 & 0 \\ 0 & 0 & 0 & -a_3 - \mu_{T_i} & \lambda_3 T_1 & 0 \\ 0 & N_2 a_2 & 0 & N_3 a_3 & -k_1 T_1 - \mu_V & 0 \\ 0 & N_1 a_1 & 0 & 0 & 0 & -k_2 M_1 - \mu_B \end{bmatrix}$.

After expanding with respect to the term, we get the characteristic polynomial of the jacobian matrix as follows:

$$det(J - \lambda I) = (\lambda + \mu_M)(\lambda + \mu_T)(\lambda^4 + A_3\lambda^3 + A_2\lambda^2 + A_1\lambda + A_0).$$

By assumption μ_M and μ_T are all strictly positive, so it suffices to examine the fourth degree equation

$$\lambda^4 + A_3\lambda^3 + A_2\lambda^2 + A_1\lambda + A_0 = 0.$$

where $A_3 = a + b + c + d > 0$ trivially.

Again, we show that $A_0 > 0$ and $A_2 A_3 - A_1 > 0$ if $k_1 > \lambda_3 N_3, \mu_B > \lambda_1 N_1, \mu_v > \lambda_2 N_2$. From the *Routh–Hurwitz criteria* (R-H criteria), if $A_3{>}0$ and $A_2 A_3 - A_1 > 0$, $A_0{>}0$ and $A_3 A_2 A_1 - A_1^2 - A_3^2 A_0{>}0$, then the above equation has roots with negative real part. If these conditions exist, then disease-free equilibrium E_0 is asymptotically stable.

Remark 1 From this theorem, we can show the stability condition of disease-free equilibrium point and get three valid biological results $k_1 > \lambda_3 N_3, \mu_B > \lambda_1 N_1, \mu_v > \lambda_2 N_2$ by solving the stability criterion.

Remark 2 If $R_0 < 1$, then the disease-free equilibrium point is asymptotically stable and if it is greater than one, the endemic equilibrium point exists.

Remark 3 If $k_1 > \lambda_3 N_3, \mu_B > \lambda_1 N_1$, and $\mu_V > \lambda_2 N_2$, then the reproduction number $R_0 < 1$.

Table 17.1 Sensitivity analysis of parameters

Parameter	Sensitivity index of R_0 w.r.t parameters	Positive or negative
S_M	−0.4289	−
λ_1	0.1575	+
λ_2	0.2689	+
μ_M	−0.4406	−
a_1	−0.000337	−
a_2	0.1866	+
μ_{M_i}	−0.18629	−
S_T	−0.6538	−
λ_3	0.57124	Most positive
μ_T	−0.57145	−
a_3	0.57123	+
μ_{T_i}	−0.91399	Most negative
N_2	0.5425	+
N_3	0.2112	+
k_1	−0.8492	−
μ_V	−0.6988	−
N_1	0.01598	+
k_2	−0.1259	−
μ_B	−0.0058	−

17.3.4 Sensitivity Analysis

In this section, we use sensitivity analysis to investigate the impact of various intervention measure. By this method, we can identify the parameters that have high impact on the basic reproductive ratio R_0, as well as on the disease transmission. Here, we derive the sensitivity index by using partial rank correlation coefficients (PRCC) of the basic reproductive ratio with respect to parameters. According to [33], the normalized forward sensitivity index of R_0 with respect to a parameter c is defined as follows:

$$\prod_{c}^{R_0} = \frac{\delta R_0}{\delta c} \times \frac{c}{R_0}. \tag{17.2}$$

In Table 17.1, we have written sensitivity index of R_0 w.r.t parameters by using above formula. This demonstrates that R_0 is most negatively sensitive to the mortality of infected macrophage (μ_{T_i}), meaning that if we increase the value of μ_{T_i}, that can reduce new cases and disease prevalence. On the other hand, λ_3 is most positive effect on R_0, i.e., if we increase the value of λ_3, then the value of R_0 increases.

Table 17.2 Parameters value using for numerical simulation

Parameter	Assigned value	Range	References
S_M	$12\,\text{mm}^{-3}\text{Day}^{-1}$	$12–14\,\text{mm}^{-3}\text{Day}^{-1}$	[34]
λ_1	$0.0003\,\text{mm}^3\text{Day}^{-1}$	–	Assumed
λ_2	$0.0000022\,\text{mm}^3\text{Day}^{-1}$	0.000002– $0.000025\,\text{mm}^3\text{Day}^{-1}$	[18, 35]
μ_M	$0.011\ \text{Day}^{-1}$	$0.011–0.05\ \text{Day}^{-1}$	[31, 34]
a_1	$0.00002\ \text{Day}^{-1}$	–	[31]
a_2	$0.015\ \text{Day}^{-1}$	–	Assumed
μ_{M_i}	$0.011\ \text{Day}^{-1}$	–	[31]
S_T	$12\,\text{mm}^{-3}\text{Day}^{-1}$	$9\text{-}15\ \text{mm}^{-3}\text{Day}^{-1}$	[34]
λ_3	$0.000024\,\text{mm}^3\text{Day}^{-1}$	–	[18]
μ_T	$0.05\ \text{Day}^{-1}$	$0.007–0.1\ \text{Day}^{-1}$	[18, 34]
a_3	$0.1\ \text{Day}^{-1}$	–	Assumed
μ_{T_i}	$0.025\ \text{Day}^{-1}$	–	[35]
N_2	500	100–1000	[18]
N_3	500	100–1000	[18]
γ_1	$0.03\ \text{mm}^3\text{Day}^{-1}$	–	Assumed
k_1	$0.00074\,\text{mm}^3\text{Day}^{-1}$	–	[18]
μ_V	$2.4\ \text{Day}^{-1}$	–	[18]
N_1	50	–	[31, 36]
γ_2	$0.007\,\text{mm}^3\text{Day}^{-1}$	–	Assumed
k_2	$0.5\,\text{mm}^3\text{Day}^{-1}$	–	[18]
μ_B	$0.5\ \text{Day}^{-1}$	–	[18]

17.4 Numerical Simulation

In this section, we study the numerical simulations of our model system on the basis of analytical findings. Our numerical studies were done using the MathWorks MATLAB 2016a. For numerical simulations, we take a set of parameter values given in Table 17.2. Some parameter values are taken from different journals, some are estimated, and remaining values are assumed. We choose the initial values in ratio dependant according to cardinal rule of scientific hypothesis.

From Fig. 17.1, we investigate the qualitative behavior of considered cells (Macrophage, T-cell, virus, and bacteria) between time intervals 400 days and also observe the time period for steady state. In Fig. 17.1a, initially macrophage decreases dramatically due to the negative effect of both pathogens on it. After 200 days, macrophage concentration increases in a small amount and reaches a stable density level $200\,\text{mm}^{-3}$ due to positive impact of CTL cells. As a result, infected macrophage chronologically increases and reaches steady concentration level about $579\,\text{mm}^{-3}$. From Fig. 17.1b, we observe that the gradient of infected T-cell gains the highest

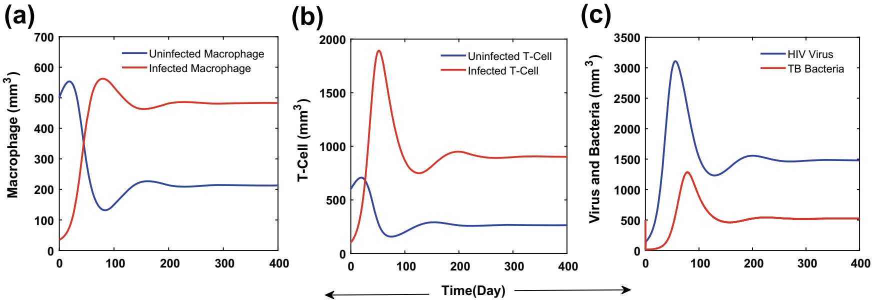


Fig. 17.1 Graphs of numerical solutions showing propagation of macrophages, T-cell, Bacteria, and virus in Co-infection with parameter values given in Table (17.1): **a** Macrophage and Infected Macrophages, **b** T-cell and Infected T-cell **c** Bacteria and Virus Population. Initial values are $M(0) = 500$, $M_i(0) = 35$, $T(0) = 600$, $T_i(0) = 100$, $V(0) = 50$, and $B(0) = 10$

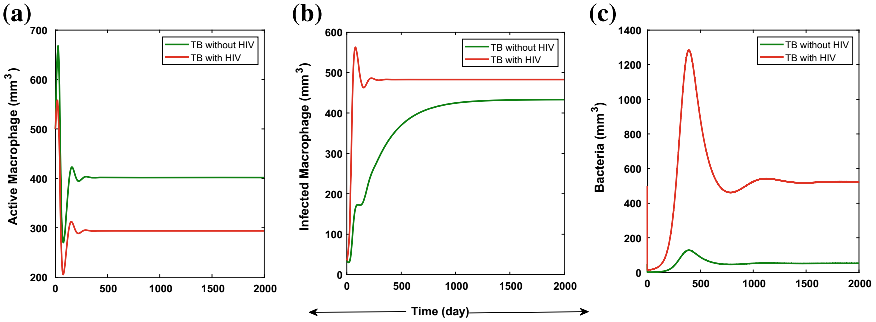


Fig. 17.2 Qualitative behavior of Active Macrophage cell, Infected Macrophage cell, and Bacteria population with initial conditions $M(0) = 500$, $M_i(0) = 35$, and $B(0) = 10$: **a** Macrophage cells, **b** Infected Macrophage cells, and **c** Bacteria population

level density 1800 mm^{-3} between 100 days. Then, it declines overly to reach stable condition due to the CTL response. Figure 17.1c demonstrates that virus and bacteria population reach the top most concentration level (3000 mm^{-3} and 1300 mm^{-3}, respectively) after initial 90 days due to high acceleration effect and then the trajectories become stable after 250 days.

In Fig. 17.2, we plot macrophage, infected macrophage, and bacteria population with respect to time and study the dynamical nature of the trajectories for two different conditions (TB without HIV and TB with HIV). From Fig. 17.2a, the green trajectory of macrophage (for the case of co-infection) decreases more than the red trajectory of that (for the case of TB without HIV) due to the enhancement effect of HIV. After 500 days, both trajectories of macrophage reach the stable density level at 300 and 430.7 mm^{-3}. For the same reason in Fig. 17.2b, it is noted that the infected macrophage reaches at the density level 560 mm^{-3} very fast during co-infection compare with when HIV is absent. Figure 17.2c illustrates the dynamical behavior of bacterial growth for co-infection (red line) and without co-infection (green line).

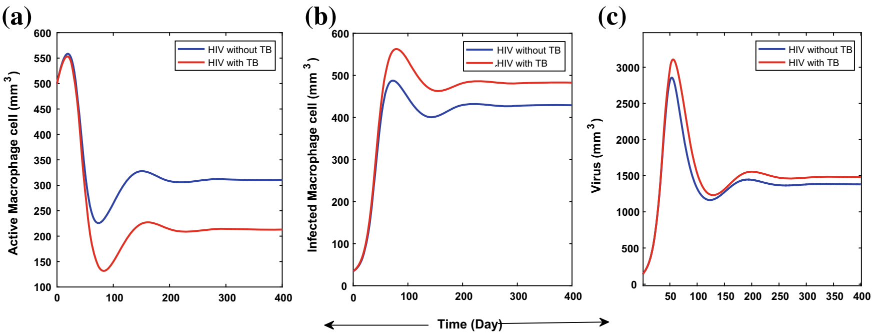


Fig. 17.3 Dynamical behavior of Macrophage cell, Infected Macrophage cell, and Virus population: **a** Macrophage cells, **b** Infected Macrophage cells, and **c** Virus population. Initial conditions are $M(0) = 500$, $M_i(0) = 35$, and $V(0) = 50$ and parameter values are taken from Table 17.1

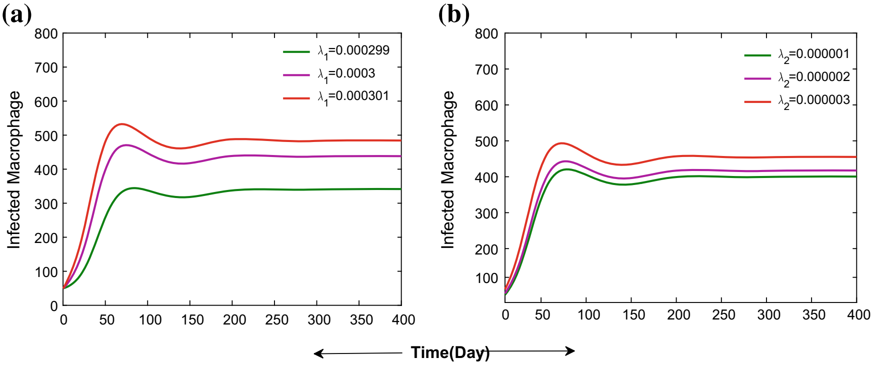


Fig. 17.4 Graph of numerical simulations for increment of Infected macrophage cells. The parameter values are $a_1 = 0.00002$, $a_2 = 0.015$ and $\mu_{M_i} = 0.011$ with initial conditions $M(0) = 500$, $M_i(0) = 35$, $V(0) = 50$, and $B(0) = 10$

During co-infection, the increment of bacterial growth is very fast and reaches the highest density level at 1300 mm^{-3} between 500 days but for the case of TB without HIV gain the maximum density level (100 mm^{-3}) in the same time interval. Then, bacterial density falls chronologically and gains a stable condition after 1500 days.

Figure 17.3 manifests the qualitative behavior of macrophage, infected macrophage along with virus population between time intervals 400 days and demonstrates two trajectories (Red and Blue color based on co-infection and without co-infection dynamics, respectively). Figure 17.3a shows the trajectories of macrophages decrease initially due to effect of HIV and co-infection, respectively and reach the steady state level at 200 and 310 mm^{-3} after 300 days. From Fig. 17.3b, it is investigated that for the same reason trajectories of infected macrophages increase to reach a steady state density level 350 and 300 mm^{-3} after 300 days. In Fig. 17.3c, it is prominent that the increment of viral growth is higher during the co-infection.

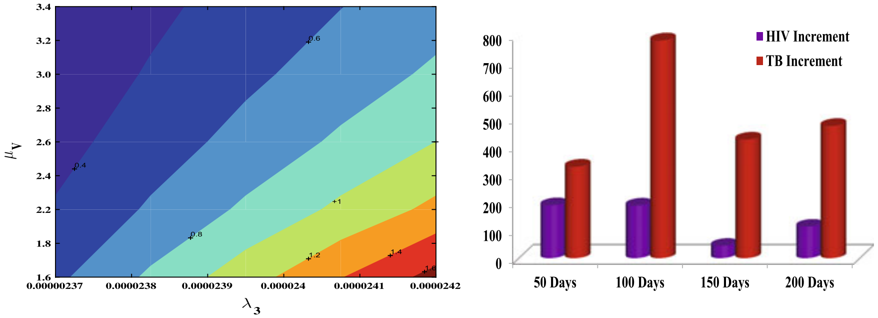


Fig. 17.5 Basic reproduction ratio R_0 (Left Panel). Increment of HIV and TB due to acceleration effect on each other (Right Panel). The parameter values are taken from the above Table 17.1

Figure 17.4 shows the change in qualitative behavior of infected macrophage with respect to time for different values of λ_1 (0.0003, 0.000301, and 0.000299) and λ_2 (0.000001, 0.0000002, and 0.000003) when $\lambda_2 = 0.0000022$ and $\lambda_1 = 0.0003$, respectively. From Fig. 17.4a (the value of λ_2 is fixed), it is clear that for the lower value of λ_1 (0.000299), the density of infected macrophage is in level 350 mm^{-3} and if we increase the value of λ_1 in small amount (0.000001), the density increase approx. 80 mm^{-3}. Furthermore, for the greater value of λ_1, the density of infected macrophage increases approx. 50 mm^{-3}. Similarly, we vary the λ_2 when the value of λ_1 is fixed, we get different trajectories from which the deflection form lower to upper trajectories is 45 mm^{-3}. From Fig. 17.4a and b, it is clear that λ_1 is more effective to infect macrophage than λ_2.

Figure 17.5 (Left panel) gives the contour plot of the basic reproduction ratio (R_0) as a function of μ_V (natural mortality rate of virus) and λ_3 (the rate at which T-cell infected by HIV). The figure illustrates that the change of the parameter R_0 as μ_V and λ_3 vary. We observe that when the ratio of rate at which T-cell infected by HIV and natural death rate of virus becomes less than unity, our system will be locally asymptotically stable indicated by dark blue region. It becomes unstable when the ratio will be greater than unity located in the figure in all parts other than the blue region. Analytically, we obtain three results connecting different parameters, but we cannot get relation between some parameters for checking the behavior of this threshold value. So, here we take different two parameters to check qualitative behavior of it. From this figure, it is also clear that if the value of λ_3 is below a certain level (approximate $\lambda_3 = 0.0000239$), the system is always stable even if higher level of μ_V. For other values of the parameters, however, R_0 is relatively stable with respect to variations.

In Fig. 17.5 (Right panel), we elaborate the growth increment (difference between growth for co-infection and without co-infection) of virus and bacteria between 200 days. From this figure, we can say that the increment of bacterial growth due to co-infection is much more than the growth of viral population. It is also manifested by this figure, after 100 days, the increment of TB is maximum and then the acceleration effect of HIV on TB gradually decreases due to CTL response. From this figure it is clear that reactivation of TB due to co-infection is greater than that of HIV.

17.5 Discussion

In this research work, we have analyzed the role of two immune cells (Macrophage cell and T-cell) along with various cytokines effect on the dynamics of HIV/TB co-infection. In our analytical study, we have verified the existence condition of disease-free and endemic equilibrium depending on the basic reproductive ratio (R_0), using next generation method. We have studied the stability criteria of disease-free as well as endemic equilibrium using Routh–Hurwitz criterion point depending on some key parameters. However, our stability analysis of disease-free equilibrium demonstrates that the mortality rate of bacteria (μ_B) is greater than the product of bacterial infection rate of macrophage (λ_1) and birth rate of new bacteria due to MTB specific macrophages (N_1). We also analytically reveal an important cell-biological phenomenon for the disease-free condition: the killing rate of virus due to CTL response (k_1) is always greater than the product of the viral infection of T-cell (λ_3) and production rate of virus due to infected T-cell (N_3). Our numerical outcomes are associated with analytical results which allow more precise prediction about reactivation of both pathogens by each other due to co-infection. Our numerical simulation shows that the accelerating effect on TB due to presence of HIV is greater than the effect on HIV by TB. Finally, our analysis speculates that if TB is effectively treated in the areas where HIV is widespread, then the rate of AIDS related deaths can be slowed down and the life span of dually infected patients will be larger.

Acknowledgements Suman Dolai was the recipient of JUNIOR RESEARCH FELLOWSHIP (CSIR), File No-09/096(0904)/2017-EMR-I. Amit Kumar Roy acknowledges DST-PURSE PROJECT (Phase-II), Government of India, Department of Mathematics, Jadavpur University, Kolkata-700032, India.

17.6 Appendix

$A = -\sum a_{ii}$
$B = \sum a_{ii}a_{jj} - \sum a_{ij}a_{ji}$
$C = \sum a_{ij}a_{ji}a_{kk} - \sum a_{ii}a_{jj}a_{kk} - \sum a_{ij}a_{jk}a_{ki}$
$D = \sum a_{ii}a_{jj}a_{kk}a_{ll} + \sum a_{ij}a_{ji}a_{ki}a_{ll} + \sum a_{ij}a_{ji}a_{kl}a_{lk} - \sum a_{ij}a_{ji}a_{kk}a_{ll} - \sum a_{ij}a_{jk}$
$a_{kl}a_{li}$
$E = \sum a_{ij}a_{ji}a_{kk}a_{ll}a_{mm} - \sum a_{ii}a_{jj}a_{kk}a_{ll}a_{mm} - \sum a_{ij}a_{jk}a_{kl}a_{li}a_{mm} - \sum a_{ij}a_{jk}a_{ki}$
$a_{ll}a_{mm} - \sum a_{ij}a_{ji}a_{kl}a_{lk}a_{mm} + \sum a_{ij}a_{jk}a_{kl}a_{lm}a_{mi}$
$F = \sum a_{ii}a_{jj}a_{kk}a_{ll}a_{mm}a_{nn} + \sum a_{ij}a_{jk}a_{kl}a_{lm}a_{mi}a_{nn} + \sum a_{ij}a_{ji}a_{kl}a_{lk}a_{mm}a_{nn} +$
$\sum a_{ij}a_{jk}a_{ki}a_{lm}a_{mn}a_{nl} - \sum a_{ij}a_{ji}a_{kk}a_{ll}a_{mm}a_{nn} - \sum a_{ij}a_{ji}a_{kl}a_{lm}a_{mk}a_{nn}$

Here, A, B, C, D, E, and F follows a rule: $i \neq j \neq k \neq l \neq m \neq n$. In a_{ij}, if $i = 1$, then j can go to 6 or 5. Following same rule for k, l, m, n. Similarly, 2 can go 5 or 6. 3 can go 5. 4 can go 5 or 3. 5 can go 2 or 3 or 4 or 6 and 6 can go 1 or 2 or 5. In $a_{ij}a_{ji}$, let $i = 1$ then j can go 5 or 6. But, $a_{15}a_{51}$ does not exist because 5 cannot go to 1.

Let $B = \sum a_{ii}a_{jj} - \sum a_{ij}a_{ji}$

By above rule $B = \sum_{i=1}^{5} a_{ii} \sum_{j=i+1}^{6} a_{jj} - \{a_{16}a_{61} + a_{26}a_{62} + a_{25}a_{52} + a_{35}a_{53} + a_{45}a_{54} + a_{56}a_{65}\}$.

References

1. World Health Organization, Global tuberculosis report 2017, Available: http://www.who.int/en/news-room/fact-sheets/detail/tuberculosis
2. A. Pawlowski, M. Jansson, M. Skld, M.E. Rottenberg, G. Killenius, Tuberculosis and HIV co-infection. PLoS Pathog. **8**(2), e1002464 (2012)
3. R.M. Houben, P.J. Dodd, The global burden of latent tuberculosis infection: a re-estimation using mathematical modelling. PLoS Med. **13**(10), e1002152 (2016)
4. P.L. Liu, J.L. Flynn, Understanding latent tuberculosis: a moving target. J. Immunol. **185**(1), 15–22 (2010)
5. UNAIDS (2017) report, Available: https://www.hiv.gov/hiv-basics/overview/data-and-trends/global-statistics
6. G.A.C. Blood, Human immunodeficiency virus (HIV). Transfus. Med. Hemotherapy **43**(3), 203 (2016)
7. C.C. Chang, M. Crane, J. Zhou, J.J. Post, B.A. Cameron, A.R. Lloyd, A. Jaworowski, M.A. French, S.R. Lewin, HIV and co-infections. Immunol. Rev. **254**(1), 114–142 (2013)
8. World Health Organization (WHO), Global tuberculosis report 2017, http://apps.who.int/iris/bitstream/10665/259366/1/9789241565516-eng.pdf. Accessed 10 Jan 2017
9. K.C. McCullough, A. Summerfield, Basic concepts of immune response and defense development. ILAR J **46**(3), 230–240 (2005)
10. V. Trouplin, N. Boucherit, L. Gorvel, F. Conti, G. Mottola, E. Ghigo, Bone marrow-derived macrophage production. J. Vis. Exp.: JoVE **81** (2013)
11. R.N. La Motte-Mohs, E. Herer, J.C. Ziga-Pflcker, Induction of T-cell development from human cord blood hematopoietic stem cells by Delta-like 1 in vitro. Blood **105**(4), 1431—1439 (2005)
12. M.S. Blumenreich, *The White Blood Cell and Differential Count, Clinical Methods: The History, Physical, and Laboratory Examinations [Internet]*, 3rd edn. (Butterworths, Boston, 1990)
13. L. Wu, N.P. Gerard, R. Wyatt, H. Choe, C. Parolin, N. Ruffing, A. Borsetti, A.A. Cardoso, E. Desjardin, W. Newman, C. Gerard, CD4-induced interaction of primary HIV-1 gp120 glycoproteins with the chemokine receptor CCR-5. Nature **384**(6605), 179 (1996)
14. B. Lee, M. Sharron, L.J. Montaner, D. Weissman, R.W. Doms, Quantification of CD4, CCR5, and CXCR4 levels on lymphocyte subsets, dendritic cells, and differentially conditioned monocyte-derived macrophages. Proc. Natl. Acad. Sci. **96**(9), 5215–5220 (1999)
15. A.N. Chatterjee, P.K. Roy, Anti-viral drug treatment along with immune activator IL-2: a control-based mathematical approach for HIV infection. Int. J. Control **85**(2), 220–237 (2012)
16. P.K. Roy, A.N. Chatterjee, T-cell proliferation in a mathematical model of CTL activity through HIV-1 infection. Proc. World Congr. Eng. **I**, 615–620 (2010)
17. S. Gordon, Pattern recognition receptors: doubling up for the innate immune response. Cell **111**(7), 927–930 (2002)
18. D. Kirschner, Dynamics of Co-infection withM. tuberculosisand HIV-1. Theor. Popul. Biol. **55**(1), 94–109 (1999)

19. C.K. Kwan, J.D. Ernst, HIV and tuberculosis: a deadly human syndemic. Clin. Microbiol. Rev. **24**(2), 351–376 (2011)
20. S. Pathak, T. Wentzel-Larsen, B. Åsjö, Effects of in vitro HIV-1 infection on mycobacterial growth in peripheral blood monocyte-derived macrophages. Infect. Immun. **78**(9), 4022–4032 (2010)
21. E.M. Shankar, R. Vignesh, R. EllegAard, M. Barathan, Y.K. Chong, M.K. Bador, D.V. Rukumani, N.S. Sabet, A. Kamarulzaman, V. Velu, M. Larsson, HIV Mycobacterium tuberculosis co-infection: a danger-couple model of disease pathogenesis. Pathog. Dis. **70**(2), 110–118 (2014)
22. D. Wolday, B. Tegbaru, A. Kassu, T. Messele, R. Coutinho, D. van Baarle, F. Miedema, Expression of chemokine receptors CCR5 and CXCR4 on CD4+ T cells and plasma chemokine levels during treatment of active tuberculosis in HIV-1-coinfected patients. JAIDS J. Acquir. Immune Defic. Syndr. **39**(3) 265–271 (2005)
23. P.A. Selwyn, D. Hartel, V.A. Lewis, E.E. Schoenbaum, S.H. Vermund, R.S. Klein, A.T. Walker, G.H. Friedland, A prospective study of the risk of tuberculosis among intravenous drug users with human immunodeficiency virus infection. N. Engl. J. Med. **320**(9), 545–550 (1989)
24. C.R. Diedrich, J.L. Flynn, HIV-1/mycobacterium tuberculosis coinfection immunology: how does HIV-1 exacerbate tuberculosis?. Infect. Immun. **79**(4), 1407–1417 (2011)
25. R. Naresh, D. Sharma, A. Tripathi, Modelling the effect of tuberculosis on the spread of HIV infection in a population with density-dependent birth and death rate. Math. Comput. Model. **50**(7–8), 1154–1166 (2009)
26. G. Bolarin, I.U. Omatola, A mathematical analysis of HIV/TB Co-infection model. Appl. Math. **6**(4), 65–72(2016)
27. C.J. Silva, D.F. Torres, A TB-HIV/AIDS coinfection model and optimal control treatment (2015). arXiv:1501.03322
28. A. Mallela, S. Lenhart, N.K. Vaidya, HIVTB co-infection treatment: modeling and optimal control theory perspectives. J. Comput. Appl. Math. **307** 143–161 (2016)
29. T.D. Awoke, M.K. Semu, Optimal control strategy for TB-HIV/AIDS Co-infection model in the presence of behaviour modification. Processes **6**(5), 48 (2018)
30. L.I.W. Roeger, Z. Feng, C. Castillo-Chavez, Modeling TB and HIV co-infections. Math. Biosci. Eng. **6**(4), 815–837 (2009)
31. G. Magombedze, W. Garira, E. Mwenje, In-vivo mathematical study of co-infection dynamics of HIV-1 and Mycobacterium tuberculosis. J. Biol. Syst. **16**(03), 357–394 (2008)
32. A. Hurwitz, On the conditions under which an equation has only roots with negative real parts. Sel. Pap. Math. Trends Control. Theory **65**, 273–284 (1964)
33. G.J. Abiodun, N. Marcus, K.O. Okosun, P.J.Witbooi, A model for control of HIV/AIDS with parental care. Int. J. Biomath. **6**(02), 1350006 (2013)
34. A.K. Roy, P.K. Roy, E. Grigorieva, Mathematical insights on psoriasis regulation: role of Th 1 and Th 2 cells. Math. Biosci. Eng. **15**(3), 717–738 (2018)
35. T. Shiri, W. Garira, S.D. Musekwa, A two-strain HIV-1 mathematical model to assess the effects of chemotherapy on disease parameters. MBE **2**, 811–832 (2005)
36. S. Marino, D.E. Kirschner, The human immune response to Mycobacterium tuberculosis in lung and lymph node. J. Theor. Biol. **227**(4), 463–486 (2004)

Chapter 18
Relative Controllability of Nonlinear Fractional Damped Delay Systems with Multiple Delays in Control

P. Suresh Kumar

Abstract This paper is concerned with the relative controllability of fractional damped dynamical systems with multiple delays in control for finite-dimensional spaces. Sufficient conditions for controllability are obtained using Schauder's fixed point theorem and the controllability Grammian matrix which is defined by the Mittag-Leffer matrix function. An example is provided to illustrate the theory.

Keywords Controllability · Fractional differential equations · Mittag-Leffler matrix function · Laplace transform

18.1 Introduction

Nowadays it is the realm of physicists and mathematicians who investigate the usefulness of non-integer order derivatives and integrals in different areas of physics and mathematics. It is a successful tool for describing complex quantum field dynamical systems, dissipation and long-range phenomena that cannot be well illustrated using ordinary differential operators. Many models are reformulated and expressed in terms of fractional differential equations so that their physical meaning will be incorporated in the mathematical models more realistically. In fact, fractional calculus attracts many physicists, biologists, engineers, and mathematicians for its interdisciplinary applications which are elegantly modeled with the help of fractional derivatives and it was conceptualized in connection with the infinitesimal calculus. Delay differential equations are often solved using numerical methods, asymptotic methods and graphical tools. Number of attempts have been made to find an analytical solution for delay differential equations by solving the characteristic equation under different conditions [16].

Controllability is one of the important qualitative aspects of a dynamical system. It is used to influence an object's behavior so as to accomplish the desired goal.

P. Suresh Kumar (✉)
Department of Mathematics, National Institute of Technology, Calicut 673601, Kerala, India
e-mail: sureshkumarp.maths@gmail.com

P. Manchanda et al. (eds.), *Mathematical Modelling, Optimization, Analytic and Numerical Solutions*, Industrial and Applied Mathematics,
https://doi.org/10.1007/978-981-15-0928-5_18

Analysis of the control problems of fractional delay dynamical system is much more advanced. The control problems involving the delay in state variables are not developed much. Controllability of delay dynamical systems was studied by Wiess [19]. Chung [5] investigated the controllability of linear time-varying systems with delay. Controllability of nonlinear delay dynamical systems is studied by Dauer [6]. Klamka [8] addressed the constrained controllability of semilinear delayed systems. A sliding mode control for linear fractional systems with input and state delays is studied by Si-Ammour [14]. Balachandran et al. [1–4] investigated the controllability of damped dynamical systems with multiple delays in control. Controllability criteria for linear fractional differential systems with state delay and impulse are studied by Zhang et al. [20]. Wang [18] proposed a numerical method for delayed fractional-order differential equations. Explicit representations of solutions of linear delay systems are studied by Shu [13]. Morgado [9] analyzed and proposed numerical methods for fractional differential equations with delay. Recently controllability of a fractional delay dynamical systems and fractional systems with time-varying delays in control is studied by Joice Nirmala et al. [10, 11]. He et al. [7] addressed the controllability of fractional damped dynamical systems with delay in control. Suresh Kumar et al. [17] studied the controllability of nonlinear fractional Langevin delay systems by assuming the conditions $0 < \alpha, \beta \leq 1$ and $\alpha + \beta > 1$. In Caputo differential operators do not satisfy the semigroup property. We can apply the only fractional integral definition. Hence in the present manuscript, we consider $0 < \beta \leq 1 < \alpha \leq 2$. So, both the problems are different by formation in the fractional sense even though they are similar in the integer case. Moreover constrained controllability of fractional linear systems with delays in control is discussed by Sikora and Klamka [15]. Motivated by this, the main aim of the present article is to present controllability of nonlinear fractional damped delay dynamical systems with multiple delays in control of order $0 < \beta \leq 1 < \alpha \leq 2$.

In this paper, we discuss the controllability of linear fractional damped delay dynamical system by utilizing the solution representation. Further, sufficient conditions for the controllability of nonlinear fractional damped delay systems are established by using Schauder's fixed point theorem. Numerical examples with simulations are provided to illustrate the theory.

18.2 Preliminaries

In this section, we introduce the definitions and preliminary results from fractional calculus which are used throughout this paper.

Definition 18.1 The Caputo fractional derivative of order $\alpha \in \mathbb{C}$ with $1 < \alpha \leq 2$, for a suitable function f is defined as

$$^{C}D^{\alpha}_{0+} f(t) = \frac{1}{\Gamma(2-\alpha)} \int_0^t (t-s)^{1-\alpha} f^{(2)}(s)\mathrm{d}s.$$

For brevity, the Caputo fractional derivative $^{C}D^{\alpha}_{0+}$ is taken as $^{C}D^{\alpha}$.

Definition 18.2 The Mittag-Leffler functions of various type are defined by

$$
\begin{aligned}
E_{\alpha}(z) &= \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\alpha k + 1)}, && (\alpha > 0, z \in \mathbb{C}).\\
E_{\alpha,\beta}(z) &= \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\alpha k + \beta)}, && (\alpha, \beta > 0, z \in \mathbb{C}),\\
E^{\gamma}_{\alpha,\beta}(z) &= \sum_{k=0}^{\infty} \frac{(\gamma)_k z^k}{k!\Gamma(\alpha k + \beta)},
\end{aligned}
$$

where $(\gamma)_n$ is a Pochhammer symbol which is defined as $\gamma(\gamma+1)\dots(\gamma+n-1)$ and $(\gamma)_n = \frac{\Gamma(\gamma+n)}{\Gamma(\gamma)}$. For an $n \times n$ matrix A

$$
\begin{aligned}
E_{\alpha,\beta}(A) &= \sum_{k=0}^{\infty} \frac{A^k}{\Gamma(\alpha k + \beta)}, \ \alpha, \beta > 0,\\
E_{\alpha,1}(A) &= E_{\alpha}(A) \text{ with } \beta = 1.
\end{aligned}
$$

Definition 18.3 ([12]) The formal definition of the Laplace transform of a function $f(t)$ of a real variable $t \in \mathbb{R}^+ = (0, \infty)$ is given by

$$\mathscr{L}\{f(t)\} = \int_0^{\infty} e^{-st} f(t)\mathrm{d}t, \quad s \in \mathbb{C}.$$

The convolution operator of two functions $f(t)$ and $g(t)$ given on $\mathbb{R}^+$ is defined for $x \in \mathbb{R}^+$ by the integral

$$(f * g)(t) = \int_0^t f(t-s)g(s)\mathrm{d}s.$$

The Laplace transform of a convolution is given by

$$\mathscr{L}\{f(t) * g(t)\} = \mathscr{L}\{f(t)\}\mathscr{L}\{g(t)\}.$$

Let $\mathscr{L}\{f(t)\} = F(s)$ and $\mathscr{L}\{g(t)\} = G(s)$. The inverse Laplace transform of product of two functions $F(s)$ and $G(s)$ is defined by

$$\mathscr{L}^{-1}\{F(s)G(s)\} = \mathscr{L}^{-1}\{F(s)\} * \mathscr{L}^{-1}\{G(s)\}.$$

The Laplace transforms of Mittag-Leffler functions are defined as

$$\mathscr{L}[E_{\alpha,1}(\pm\lambda t^{\alpha})](s) = \frac{s^{\alpha-1}}{(s^{\alpha} \mp \lambda)}, \quad Re(\alpha) > 0,$$

$$\mathscr{L}[t^{\beta-1}E_{\alpha,\beta}(\pm\lambda t^{\alpha})](s) = \frac{s^{\alpha-\beta}}{(s^{\alpha} \mp \lambda)}, \quad Re(\alpha) > 0, \quad Re(\beta) > 0,$$
$$\mathscr{L}[t^{\beta-1}E^{\gamma}_{\alpha,\beta}(\pm\lambda t^{\alpha})](s) = \frac{s^{\alpha\gamma-\beta}}{(s^{\alpha} \mp \lambda)^{\gamma}}, \quad Re(\alpha) > 0, \quad Re(\beta) > 0.$$

18.3 Linear System with Multiple Delays in Control

Consider the linear fractional damped delay dynamical system with multiple delays of the form

$$\begin{aligned} {}^{C}D^{\alpha}x(t) - A\,{}^{C}D^{\beta}x(t) &= Bx(t) + Cx(t-\tau) + \sum_{i=0}^{M} D_i u(h_i(t)), \quad t \in J : [0, T], \\ x(t) &= \phi(t), \quad -\tau < t \le 0, \\ x'(0) &= q_0, \end{aligned} \tag{18.1}$$

where $0 < \beta \le 1 < \alpha \le 2$, $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, A, B and C are $n \times n$ matrices and D_i for $n \times m$ matrices for $i = 0, 1, 2, \ldots, M$. Assume the following conditions:

(H1) The functions $h_i : J \to \mathbb{R}, i = 0, 1, 2, \ldots M$ are twice differentiable and strictly increasing in J. Moreover

$$h_i(t) \le t, \quad \text{for } i = 0, 1, 2, \ldots M, \quad \text{for all} \quad t \in J, \tag{18.2}$$

(H2) Introduce the time lead functions $r_i(t) : [h_i(0), h_i(T)] \to [0, T], i = 0, 1, 2, \ldots M$, such that $r_i(h_i(t)) = t$ for $t \in J$. Further $h_0(t) = t$ and for $t = T$. The following inequality holds

$$\begin{aligned} h_M(T) \le h_{M-1}(T) \le \ldots h_{m+1}(T) \le 0 = h_m(T) < h_{m-1}(T) = \ldots \\ = h_1(T) = h_0(T) = T. \end{aligned} \tag{18.3}$$

(H3) Let $h > 0$ be given. For functions $u : [-h, T] \to \mathbb{R}^n$ and $t \in J$, we use the symbol u_t denote the function on $[-h, 0]$ defined by $u_t(s) = u(t + s)$, for $s \in [-h, 0)$.

The following definitions of complete state of the system (18.1) at time t and relative controllability are assumed.

Definition 18.4 The set $y(t) = \{x(t), u_t\}$ is the complete state of the system (18.1) at time t.

Definition 18.5 System (18.1) is said to be relatively controllable on $[0, T]$ if, for every complete state $y(t)$ and every $x_1 \in \mathbb{R}^n$ there exists a control $u(t)$ defined on $[0, T]$ such that the solution of system (18.1) satisfies $x(T) = x_1$.

Here the complete state $y(0)$ and the vector $x_1 \in \mathbb{R}^n$ are chosen arbitrarily. The solution of the system (18.1) can be written [11] as

$$\begin{aligned} x(t) = {} & X_{\alpha-\beta}(t)\phi(0) - AX_{\alpha-\beta,\alpha-\beta+1}(t)\phi(0) + tX_{\alpha-\beta,2}(t)q_0 \\ & + C\int_{-\tau}^{0}(t-s-\tau)^{\alpha-1}X_{\alpha-\beta,\alpha}(t-s-\tau)\phi(s)\mathrm{d}s \\ & + \int_0^t (t-s)^{\alpha-1}X_{\alpha-\beta,\alpha}(t-s)\sum_{i=0}^{M} D_i u(h_i(s))\mathrm{d}s. \end{aligned} \tag{18.4}$$

Using the time lead functions $r_i(t)$, we have

$$x(t) = x_L(t;\phi) + \sum_{i=0}^{M}\int_{h_i(0)}^{h_i(t)}(t-r_i(s))^{\alpha-1}X_{\alpha-\beta,\alpha}(t-r_i(s))D_i\dot{r}_i(s)u(s)\mathrm{d}s,$$

where

$$\begin{aligned} x_L(t;\phi) = {} & X_{\alpha-\beta}(t)\phi(0) - AX_{\alpha-\beta,\alpha-\beta+1}(t)\phi(0) + tX_{\alpha-\beta,2}(t)q_0 \\ & + C\int_{-\tau}^{0}(t-s-\tau)^{\alpha-1}X_{\alpha-\beta,\alpha}(t-s-\tau)\phi(s)\mathrm{d}s. \end{aligned}$$

By using the inequality (18.3) we get

$$\begin{aligned} x(t) = {} & x_L(t;\phi) + \sum_{i=0}^{m}\int_{h_i(0)}^{0}(t-r_i(s))^{\alpha-1}X_{\alpha-\beta,\alpha}(t-r_i(s))D_i\dot{r}_i(s)u_0(s)\mathrm{d}s \\ & + \sum_{i=0}^{m}\int_0^t (t-r_i(s))^{\alpha-1}X_{\alpha-\beta,\alpha}(t-r_i(s))D_i\dot{r}_i(s)u(s)\mathrm{d}s \\ & + \sum_{i=m+1}^{M}\int_{h_i(0)}^{h_i(t)}(t-r_i(s))^{\alpha-1}X_{\alpha-\beta,\alpha}(t-r_i(s))D_i\dot{r}_i(s)u_0(s)\mathrm{d}s. \end{aligned}$$

For simplicity, let us write the solution as

$$x(t) = x_L(t;\phi) + G(t) + \sum_{i=0}^{M}\int_0^t (t-r_i(s))^{\alpha-1}X_{\alpha-\beta,\alpha}(t-r_i(s))D_i\dot{r}_i(s)u(s)\mathrm{d}s, \tag{18.5}$$

where

$$G(t) = \sum_{i=0}^{m}\int_{h_i(0)}^{0}(t-r_i(s))^{\alpha-1}X_{\alpha-\beta,\alpha}(t-r_i(s))D_i\dot{r}_i(s)u_0(s)\mathrm{d}s$$

$$+\sum_{i=m+1}^{M}\int_{h_i(0)}^{h_i(t)}(t-r_i(s))^{\alpha-1}X_{\alpha-\beta,\alpha}(t-r_i(s))D_i\dot{r}_i(s)u_0(s)\mathrm{d}s.$$

Now let us define the controllability Grammian matrix by

$$W=\sum_{i=0}^{m}\int_0^T(T-r_i(s))^{2(\alpha-1)}(X_{\alpha-\beta,\alpha}(T-r_i(s))D_i\dot{r}_i(s))(X_{\alpha-\beta,\alpha}(T-r_i(s))D_i\dot{r}_i(s))^*\mathrm{d}s.$$

Theorem 18.1 *The linear system (18.1) is relatively controllable on* $[0,T]$ *if and only if the controllability Grammian matrix is positive definite for some* $T>0$.

Proof Assume that W is positive definite. Define the control function by

$$u(t)=(T-r_i(t))^{\alpha-1}(X_{\alpha-\beta,\alpha}(T-r_i(t))D_i\dot{r}_i(t))^*W^{-1}[x_1-x_L(T;\phi)-G(T)], \tag{18.6}$$

where the complete state $y(0)$ and the vector $x_1\in\mathbb{R}^n$ are chosen arbitrary. Taking $t=T$ in (18.5) and by using (18.6), we have $x(T)=x_1$. Then

$$y^*Wy=0,$$

that is,

$$y^*\bigg[\sum_{i=0}^{m}\int_0^T(T-r_i(s))^{2(\alpha-1)}(X_{\alpha-\beta,\alpha}(T-r_i(s))D_i\dot{r}_i(s))(X_{\alpha-\beta,\alpha}(T-r_i(s))D_i\dot{r}_i(s))^*\mathrm{d}s\bigg]y=0,$$

which implies

$$y^*\sum_{i=0}^{m}(T-r_i(s))^{\alpha-1}(X_{\alpha-\beta,\alpha}(T-r_i(s))D_i\dot{r}_i(s))=0,\ \text{on } [0,T].$$

Consider the zero initial function $\phi=0$ and $u_0=0$ on $[-h,0]$ and the final point $x_1=y$. Since the system is controllable there exists a control $u(t)$ on J that steers the response to $x_1=y$. For $\phi=0$, $x_L(T,\phi)=0$, $G(t)=0$. On the other hand

$$y=x_L(T)=\sum_{i=0}^{m}\int_0^T(T-r_i(s))^{\alpha-1}X_{\alpha-\beta,\alpha}(T-r_i(s))D_i\dot{r}_i(s)u(s)\mathrm{d}s.$$

Then

$$y^*y=\sum_{i=0}^{m}\int_0^T y^*(T-r_i(s))^{\alpha-1}X_{\alpha-\beta,\alpha}(T-r_i(s))D_i\dot{r}_i(s)u(s)\mathrm{d}s=0.$$

This contradicts for $y\neq 0$. Hence W is nonsingular.

18.4 Nonlinear Systems with Multiple Delays in Control

Consider the nonlinear fractional damped delay dynamical system with multiple delays in control of the form

$$
\begin{aligned}
{}^{C}D^{\alpha}x(t) - A\,{}^{C}D^{\beta}x(t) &= Bx(t) + Cx(t-\tau) + \sum_{i=0}^{M} D_i u(h_i(t)) + f(t, x(t), x(t-\tau), u(t)),\\
x(t) &= \phi(t),\\
x'(0) &= q_0, \quad -\tau < t \leq 0, \qquad (18.7)
\end{aligned}
$$

where $0 < \beta \leq 1 < \alpha \leq 2$, $x \in \mathbb{R}^n$ is a state vector, $u \in \mathbb{R}^m$ is a control vector, A, B, C are $n \times n$ matrices, D_i for $i = 0, 1, 2, \dots M$, are $n \times m$ matrices and $f : J \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is a continuous function. Further we impose the following assumption:

Let Q be the Banach space of continuous $\mathbb{R}^n \times \mathbb{R}^m$ valued functions defined on the interval J with the norm

$$\|(x, u)\| = \|x\| + \|u\|,$$

where $\|x\| = \sup\{x(t) : t \in J\}$ and $\|u\| = \sup\{u(t) : t \in J\}$. That is $Q = C_n(J) \times C_m(J)$, where $C_n(J)$ is the Banach space of continuous $\mathbb{R}^n$ valued functions defined on the interval J with the sup norm.

Similar to the linear system, the solution of nonlinear system (18.7) using time lead function $r_i(t)$ is given as

$$
\begin{aligned}
x(t) = x_L(t;\phi) + G(t) + \sum_{i=0}^{m} \int_0^t (t - r_i(s))^{\alpha-1} X_{\alpha-\beta,\alpha}(t - r_i(s)) D_i \dot{r}_i(s) u(s)\mathrm{d}s\\
+ \int_0^t (t-s)^{\alpha-1} X_{\alpha-\beta,\alpha}(t-s) f(s, x(s), x(s-\tau), u(s))\mathrm{d}s. \qquad (18.8)
\end{aligned}
$$

Theorem 18.2 *Let the continuous function f satisfy the condition*

$$\lim_{|p|\to\infty} \frac{|f(t,p)|}{|p|} = 0 \qquad (18.9)$$

uniformly in $t \in J$ and suppose that the system (18.1) is relatively controllable on J. Then the system (18.7) is relatively controllable on J.

Proof Let $\phi(t)$ be continuous on $[-\tau, 0]$ and let $x_1 \in \mathbb{R}^n$. Let Q be the Banach space of all continuous functions

$$(x, u) : [-\tau, T] \times [0, T] \to \mathbb{R}^n \times \mathbb{R}^m,$$

with the norm

$$\|(x,u)\| = \|x\| + \|u\|,$$

where $\|x\| = \{\sup|x(t)| \text{ for } t \in [-\tau, T]\}$ and $\|u\| = \{\sup|u(t)| \text{ for } t \in [0, T]\}$.

The solution of (18.7) using time lead function $r_i(t)$ is given by

$$\begin{aligned} x(t) = x_L(t;\phi) + G(t) + \sum_{i=0}^{M} \int_0^t (t - r_i(s))^{\alpha-1} X_{\alpha-\beta,\alpha}(t - r_i(s)) D_i \dot{r}_i(s) u(s) \mathrm{d}s \\ + \int_0^t (t-s)^{\alpha-1} X_{\alpha-\beta,\alpha}(t-s) f(s, x(s), x(s-\tau), u(s)) \mathrm{d}s. \end{aligned} \tag{18.10}$$

Let us assume

$$\begin{aligned} a_i &= \sup\|X_{\alpha-\beta,\alpha}(T - r_i(s))\|, b_i = \|\dot{r}_i(s)\|, i = 0, 1, 2, \ldots, M, \upsilon = \sup\|u_0(s)\|, \\ \vartheta &= \sup\|X_{\alpha-\beta,\alpha}(T - s)\|, \mu = \sum_{i=0}^{m} a_i b_i \|D_i\| N_i + \sum_{i=m+1}^{M} a_i b_i \|D_i\| M_i, \\ c_1 &= 4[a_i b_i \|D_i^*\|]\|W^{-1}\|\upsilon(\alpha-\beta)^{-1}T^{\alpha-\beta}, d_1 = 4[a_i b_i \|D_i^*\|]\|W^{-1}\|[|x_1 + \gamma + \mu|], \\ a &= \max\{b(\alpha-\beta)^{-1}T^{\alpha-\beta}\|D_i\|, 1\}, b = \sum_{i=0}^{m} a_i b_i L_i, c_2 = 4\vartheta(\alpha-\beta)^{-1}T^{\alpha-\beta}, d_2 = 4[\gamma + \upsilon\mu], \\ N_i &= \int_{h_(0)}^{0} (T - r_i(s))^{\alpha-1} \mathrm{d}s, M_i = \int_{h_(0)}^{h_i(T)} (T - r_i(s))^{\alpha-1} \mathrm{d}s, \\ L_i &= \int_0^T (T - r_i(s))^{\alpha-1} \mathrm{d}s, c = \max\{c_1, c_2\}, d = \max\{d_1, d_2\}, \end{aligned}$$

and

$$\sup = \{\sup|f(t, x(t), x(t-\tau), u(t))|, t \in J\}.$$

Define $\Psi : Q \to Q$ by

$$\Psi(x,u) = (z, v),$$

where

$$\begin{aligned} v(t) = (T - r_i(t))^{\alpha-1}(x_{\alpha-\beta,\alpha}(T - r_i(t))(D_i)^* \dot{r}_i(t))^* W^{-1}\Bigg[x_1 - x_L(T;\phi) \\ - \sum_{i=0}^{m} \int_{h_i(0)}^{0} (T - r_i(s))^{\alpha-1} X_{\alpha-\beta,\alpha}(T - r_i(s)) D_i \dot{r}_i(s) u_0(s) \mathrm{d}s \\ + \sum_{i=m+1}^{M} \int_0^T (T - r_i(s))^{\alpha-1} X_{\alpha-\beta,\alpha}(T - r_i(s)) D_i \dot{r}_i(s) u_0(s) \mathrm{d}s \\ - \int_0^T (T-s)^{\alpha-1} X_{\alpha-\beta,\alpha}(T-s) f(s, x(s), x(s-\tau), u(s)) \mathrm{d}s\Bigg], \end{aligned}$$

and

$$\begin{aligned} z(t) = x_L(t;\phi) &- \sum_{i=0}^{m} \int_{h_i(0)}^{0} (t-r_i(s))^{\alpha-1} X_{\alpha-\beta,\alpha}(t-r_i(s)) D_i \dot{r}_i(s) u_0(s)\mathrm{d}s \\ &+ \sum_{i=0}^{m} \int_{0}^{t} (t-r_i(s))^{\alpha-1} X_{\alpha-\beta,\alpha}(t-r_i(s)) D_i \dot{r}_i(s) v(s)\mathrm{d}s \\ &+ \sum_{i=m+1}^{M} \int_{\alpha_0}^{t} (t-r_i(s))^{\alpha-1} X_{\alpha-\beta,\alpha}(t-r_i(s)) D_i \dot{r}_i(s) u_0(s)\mathrm{d}s \\ &+ \int_{0}^{t} (t-s)^{\alpha-1} X_{\alpha-\beta,\alpha}(t-s) f(s, x(s), x(s-\tau), u(s))\mathrm{d}s. \end{aligned}$$

Then

$$\begin{aligned} |v(t)| &\le \|D_i^*\| a_i b_i \|w^{-1}\| [\|x_1\| + \gamma + \mu] + a_i b_i \|D_i^*\| \|W^{-1}\| \vartheta(\alpha-\beta)^{-1} T^{\alpha-\beta}, \\ &\le \frac{1}{4a}(d + c \sup|f|) \end{aligned}$$

and

$$\begin{aligned} |z(t)| &\le \gamma + \upsilon\mu + \left(\sum_{i=0}^{m} a_i b_i \|D_i\| L_i \alpha^{-1} T^{\alpha-\beta} \right) v(s) + \vartheta(\alpha-\beta)^{-1} T^{\alpha-\beta} \sup|f|, \\ &\le \frac{d}{2} + \frac{c}{2} \sup|f|. \end{aligned}$$

Further P maps

$$Q(r) = \left\{ (z,v) \in Q : \|z\| \le \frac{r}{2} \text{ and } \|v\| \le \frac{r}{2} \right\}$$

into itself and has a fixed point by the Schauder's fixed point theorem such that $P(z,v) = (z,v) = (x,u)$. Hence we have

$$\begin{aligned} x(t) = x_L(t;\phi) + G(t) &+ \sum_{i=0}^{M} \int_{0}^{t} (t-r_i(s))^{\alpha-1} X_{\alpha-\beta,\alpha}(t-r_i(s)) D_i \dot{r}_i(s) u(s)\mathrm{d}s \\ &+ \int_{0}^{t} (t-s)^{\alpha-1} X_{\alpha-\beta,\alpha}(t-s) f(s, x(s), x(s-\tau), u(s))\mathrm{d}s. \end{aligned} \tag{18.11}$$

for $t \in J$ and $x(t) = \phi(t)$ for $t \in [-\tau, 0]$ and

$$x(T) = x_1.$$

Hence the system (18.7) is relatively controllable on J.

18.5 Example

Example 18.1 Consider the nonlinear fractional damped delay dynamical system

$$\begin{aligned} {}^{C}D^{\alpha}x(t) - A\,{}^{C}D^{\beta}x(t) &= Bx(t) + Cx(t-1) + D_0u(t) + D_1u(t-1) + f(t, x(t), x(t-1), u(t)),\\ x(t) &= \phi(t),\\ x'(0) &= q_0, \quad -1 < t \leq 0, \end{aligned} \tag{18.12}$$

The solution of the above problem (18.12) using Laplace transform we get

$$\begin{aligned} x(t) = &\sum_{n=0}^{[t]}\Big[B^n(t-n)^{\alpha n}E^{n+1}_{\alpha-\beta,\alpha n+1}(A(t-n)^{\alpha-\beta}) + C^n(t-n)^{\alpha n}E^{n+1}_{\alpha-\beta,\alpha n+1}(A(t-n)^{\alpha-\beta})\Big]\phi(0)\\ &-A\sum_{n=0}^{[t]}\Big[B^n(t-n)^{\alpha n+\alpha-\beta}E^{n+1}_{\alpha-\beta,\alpha n+\alpha-\beta+1}(A(t-n)^{\alpha-\beta})\\ &+C^n(t-n)^{\alpha n+\alpha-\beta}E^{n+1}_{\alpha-\beta,\alpha n+\alpha-\beta+1}(A(t-n)^{\alpha-\beta})\Big]\phi(0)\\ &+\sum_{n=0}^{[t]}\Big[B^n(t-n)^{\alpha n+1}E^{n+1}_{\alpha-\beta,\alpha n+2}(A(t-n)^{\alpha-\beta}\\ &+C^n(t-n)^{\alpha n+1}E^{n+1}_{\alpha-\beta,\alpha n+2}(A(t-n)^{\alpha-\beta})\Big]y_0\\ &+C\sum_{n=0}^{[t]}B^n\int_{-1}^{0}(t-s-n-1)^{\alpha n+\alpha-1}E^{n+1}_{\alpha-\beta,\alpha}(A(t-n)^{\alpha-\beta})\phi(s)\mathrm{ds} \end{aligned}$$

$$\begin{aligned} &+C^n\int_{-1}^{0}(t-s-n-1)^{\alpha n+\alpha-1}E^{n+1}_{\alpha-\beta,\alpha}(A(t-n)^{\alpha-\beta})\phi(s)\mathrm{ds}\\ &+\sum_{n=0}^{[t]}\Big[B^n\int_{0}^{t-n}(t-s-n)^{\alpha n+\alpha-1}E^{n+1}_{\alpha-\beta,\alpha}(A(t-n)^{\alpha-\beta})\\ &+C^n\int_{0}^{t-n}(t-s-n)^{\alpha n+\alpha-1}E^{n+1}_{\alpha-\beta,\alpha}(A(t-n)^{\alpha-\beta})\Big]D\dot{r}_iu(s)\mathrm{ds}\\ &+\sum_{n=0}^{[t]}\Big[B^n\int_{0}^{t-n}(t-s-n)^{\alpha n+\alpha-1}E^{n+1}_{\alpha-\beta,\alpha}(A(t-n)^{\alpha-\beta})\\ &+C^n\int_{0}^{t-n}(t-s-n)^{\alpha n+\alpha-1}E^{n+1}_{\alpha-\beta,\alpha}(A(t-n)^{\alpha-\beta})\Big]f(s, x(s), x(s-1), u(s)\mathrm{ds}, \end{aligned}$$

where $[\cdot]$ is the greatest integer function. Now consider the controllability on $[0,\ 1]$. Here $[t]$=0; and let $\alpha = \frac{3}{2}$, β=$\frac{1}{2}$, $h = 1$, $A = \begin{pmatrix} 0 & -1\\ 1 & 0 \end{pmatrix}$, $B = \begin{pmatrix} 1 & 1\\ -1 & 0 \end{pmatrix}$,

$C = \begin{pmatrix} 1 & -1 \\ 0 & 2 \end{pmatrix}$, $x(t) = \phi(t) \in \mathbb{R}^2$ and $x(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix}$ with initial conditions $\phi(0) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, $y_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and final condition $x(1) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, $f(t, x(t), x(t-1), u(t)) = \frac{x(t)+x(t-1)}{x^2(t)+x^2(t-1)+u(t)}$. By applying Laplace transform on both sides of the equation, we get the solution as therefore the solution of (18.12) on [0,1] is

$$\begin{aligned} x(t) = {} & 2E_{\alpha-\beta}(At^{\alpha-\beta})\phi(0) - 2t^{\alpha-\beta}AE_{\alpha-\beta,\alpha-\beta+1}(At^{\alpha-\beta})\phi(0) + 2tE_{\alpha-\beta,2}(At^{\alpha-\beta})y_0 \\ & +2C\int_{-1}^{0}(t-s-1)^{\alpha-\beta}E_{\alpha-\beta,\alpha}(A(t-s-1)^{\alpha-\beta})\phi(s)\mathrm{ds} \\ & +2\int_0^t (t-r_i(s))^{\alpha-1}E_{\alpha-\beta,\alpha}(A(t-r_i(s))^{\alpha-\beta})D\dot{r_i}u(s)\mathrm{ds} \\ & +2\int_0^t (t-r_i(s))^{\alpha-1}E_{\alpha-\beta,\alpha}(A(t-r_i(s))^{\alpha-\beta})f(s, x(s), x(s-1), u(s)\mathrm{ds}, \end{aligned}$$

and on further simplification

$$\begin{aligned} x(t) = {} & 2E_{\alpha-\beta}(At^{\alpha-\beta})\phi(0) - 2t^{\alpha-\beta}AE_{\alpha-\beta,\alpha-\beta+1}(At^{\alpha-\beta})\phi(0) + 2tE_{\alpha-\beta,2}(At^{\alpha-\beta})y_0 \\ & +2t^{\alpha-1}(t)^{\alpha-\beta}E_{\alpha-\beta,\alpha}(A(t)^{\alpha-\beta})\phi(0) \\ & +2\int_0^t (t-r_i(s))^{\alpha-1}E_{\alpha-\beta,\alpha}(A(t-r_i(s))^{\alpha-\beta})D\dot{r_i}u(s)\mathrm{ds} \\ & +2\int_0^t (t-r_i(s))^{\alpha-1}E_{\alpha-\beta,\alpha}(A(t-r_i(s))^{\alpha-\beta})f(s, x(s), x(s-1), u(s)\mathrm{ds}, \end{aligned}$$

By simple matrix calculation, we have the controllability Grammian matrix as

$$W = \begin{pmatrix} 26.6369 & -19.9353 \\ -19.9353 & 57.6070 \end{pmatrix} > 0,$$

which is positive definite. Hence the system (18.12) is controllable on [0, 1]. Therefore, the linear system of (18.12) is controllable on [0, 1]. And the nonlinear function $f(t, x(t), x(t-1), u(t))$ satisfies the hypothesis of Theorem (18.2) and hence the nonlinear system (18.12) steering from the initial point $\phi_0 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ to a desire state $x_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ during [0, 1]. Hence the nonlinear system (18.12) is relatively controllable.

18.6 Conclusion

This paper deals with the relative controllability of nonlinear fractional damped delay systems with multiple delays in control. In [10, 11] the authors have studied the problem of order $0 < \alpha \leq 1$. In this paper, we considered two different orders α

and β which satisfy $0 < \beta \leq 1 < \alpha \leq 2$. Sufficient conditions for the controllability results are established using Schauder's fixed point theorem. Also, the controllability of nonlinear fractional damped delay system with multiple delays in control are discussed. An example is provided to illustrate the theory.

References

1. K. Balachandran, Controllability of nonlinear fractional delay dynamical systems with multiple delays in control. *Lecture Notes in Electrical Engineering* (2016), pp. 321–332
2. K. Balachandran, J. Kokila, J.J. Trujillo, Relative controllability of fractional dynamical systems with multiple delays in control. Comput. Math. Appl. **64**, 3037–3045 (2012)
3. K. Balachandran, V. Govindaraj, L. Rodíguez-Germa, J.J. Trujillo, Controllability results for nonlinear fractional-order dynamical systems. J. Optim. Theory Appl. **156**, 33–44 (2013)
4. K. Balachandran, V. Govindaraj, M. Rivero, J.J. Trujillo, Controllability of fractional damped dynamical systems. Appl. Math. Comput. **257**, 66–73 (2015)
5. D.H. Chyung, Controllability of linear time-varying systems with delay. IEEE Trans. Automat. Control **16**, 493–495 (1971)
6. J.P. Dauer, R.D. Gahl, Controllability of nonlinear delay systems. J. Optim. Theory Appl. **21**, 59–68 (1977)
7. B.B. He, H.C. Zhou, C.H. Kou, The controllability of fractional damped dynamical systems with control delay. Commun. Nonlinear Sci. Numer. Simul. **32**, 190–198 (2016)
8. J. Klamka, Constrained controllability of semilinear delayed systems. Bull. Pol. Acad. Sci. Tech. Sci. Electron. Electrotech. **49**, 505–515 (2001)
9. M.L. Morgado, N.J. Ford, P.M. Lima, Analysis and numerical methods for fractional differential equation with delay. J. Comput. Appl. Math. **252**, 159–168 (2013)
10. R.J. Nirmala, Relative controllability of nonlinear fractional delay dynamical systems with time varying delay in control. *Lecture Notes in Electrical Engineering*, pp. 369–380
11. R.J. Nirmala, K. Balachandran, L. Rodriguez–Germa, J.J. Trujillo, Controllability of nonlinear fractional delay dynamical systems. Rep. Math. Phys. **77**, 87–104 (2016)
12. J.L. Schiff, *The Laplace Transform Theory and Applications* (Springer, New York, 1999)
13. F.C. Shu, On explicit representations of solutions of linear delay systems. Appl. Math. E-Notes **13**, 120–135 (2013)
14. A. Si-Ammour, S. Djennoune, M. Bettayeb, A sliding mode control for linear fractional systems with input and state delays. Commun. Nonlinear Sci. Numer. Simul. **14**, 2310–2318 (2009)
15. B. Sikora, J. Klamka, Constrained controllability of fractional linear systems with delays in control. Syst. Control Lett. **106**, 9–15 (2017)
16. H. Smith, *An Introduction to Delay Differential Equations with Application to Life Sciences* (Springer, New York, 2011)
17. P. Sureshkumar, K. Balachandran, N. Annapoorani, Controllability of nonlinear fractional Langevin delay systems. Nonlinear Anal. Model. Cont. **23**, 321–340 (2018)
18. Z. Wang, A numerical method for delayed fractional order differential equations. J. Appl. Math. Article ID 256071 (2013). https://doi.org/10.1155/2013/256071
19. L. Wiess, On the controllability of delayed differential systems. SIAM J. Control **5**, 575–587 (1967)
20. H. Zhang, J. Cao, W. Jiang, Controllability criteria for linear fractional differential systems with state delay and impulse. J. Appl. Math. Article ID 146010 (2013). https://doi.org/10.1155/2013/567089

Chapter 19
A Graphical User Interface-Based Fingerprint Recognition

Rohit Khokher and Ram Chandra Singh

Abstract Biometric authentication is a process of establishing an individual's identity through measurable characteristics of their behavior, anatomy, or physiology. Fingerprint recognition is a biometric technology that has been extensively used in a various range of contexts from immigration control on airports, transactions in banks, applying for a driving license, a passport to Aadhar card in India, and personal computing. In recent emerging technologies, the usability aspects of system design have received less attention rather than technical aspects. The researches on fingerprint have shown many challenges for users like placing fingers to capture fingerprints, system feedback, and instructions to use fingerprint systems. This paper proposes a Graphical User Interface (GUI) system for studying various operations in recognizing fingerprints for biometric identification of individuals using an iterative, participative design approach. During this process, several different layouts have been identified. The fingerprint GUI provides facility to users to use by clicking on the buttons on the front-end interface of the system. The coding for the back-end interface functions is written in MATLAB. This study has been tested over DB1 of FVC2006 database. The dataset consists of 1800 images captured by electric field sensor at 250 dpi. The volunteers were asked to put their fingers naturally on the acquisition device and no constraints were enforced to guarantee a minimum quality in the images. The minutiae and texture features of fingerprints have been studied and the results show 100% matching of an individual from the collected database. Fingerprint recognition using GUI is reliable and easy to understand the operations and results more efficiently.

Keywords Graphical User Interface (GUI) · Fingerprint recognition biometrics · Enhancement · Feature extraction · Noises · Filters · Similarity measures

R. Khokher (✉)
Vidya College of Engineering, Baghpat Road, Meerut 250002, Uttar Pradesh, India
e-mail: khokherrohit@gmail.com

R. C. Singh
School of Basic Sciences and Research, Sharda University, Greater Noida 201310, Uttar Pradesh, India
e-mail: rcsingh_physics@yahoo.com

P. Manchanda et al. (eds.), *Mathematical Modelling, Optimization, Analytic and Numerical Solutions*, Industrial and Applied Mathematics,
https://doi.org/10.1007/978-981-15-0928-5_19

19.1 Introduction

In today's life, information and communication technologies (ICT) are spreading widely throughout the globe in daily routine activities. Therefore, the security of these systems is the most important challenge. Identification of genuine users of these systems is the need for secure systems. A user can be identified in three different ways—Token-based, knowledge-based, and biometrics identifications. Token-based identification requires the presence of physical objects like ID card, pass, etc. to authenticate a user as where in knowledge-based identification it relies on nonobvious information like passwords, personal identification numbers (PINs) to confirm the authenticity of an individual. In contrast, biometric identification considers physical, behavioral, or anatomical characteristics of the user to authenticate the identity. The use of biometrics for identification is increasing day by day because of its features used for authentication as it cannot be stolen or lost [1–3].

Nowadays biometric authentication technology is being used both in commercial and public sectors. According to International Biometrics Group (IBG), the usage of biometrics will be doubled in size over the next five years and there are numerous trends that support IBG [4]. The secure user identification is an international trend that is being used worldwide that uses public facing implementations of biometric systems such as the immigration control in US, Dubai, Malaysia, etc.; identity card scheme in the United Kingdom; and Aadhar card in India. To secure the information in IT world the usage of biometric technology has emerged as a powerful tool to secure the information. There are many challenges associated with the use of biometrics such as enrolment or registration and authentication processes. During enrolment process, the biometric traits of an individual are stored in the database. In identification or authentication process, the data of enrolled traits in the database are matched with the input data to verify an individual's identity. Generally, in process of automated identity verification through biometrics, the users have no familiarity with the technology being used during the authentication process. This motivated us to develop a Graphical User Interface (GUI) for fingerprint recognition. This interface would help the users to understand the results visually of the operations performed by a click on the button of the panel.

The fingerprint biometric systems are the most commonly used commercially available biometric system as it has achieved the accuracy of 100% [5–10]. Further, there is a need for a system that explains the complete process of fingerprint recognition using GUI. During last few years, design of the interface for fingerprint systems has received an increased amount of attention from the industry. The motivation behind developing this system is to help students, young researchers, and users to understand the process of fingerprint recognition. The GUI proposed in this study for fingerprint recognition includes the enhancement operations, feature extractions, plotting of histograms, noising and denoising and computation of similarity measures and performance parameters. The organization of rest of the paper is as follows: Sect. 19.2 describes an overview of proposed GUI, the functionalities of the buttons of the interface are discussed in Sects. 19.3, 19.4 deals with the results obtained in the authentication process and Sect. 19.5 ends with the overall conclusion and the future scope of this GUI. The MATLAB functions have been used to develop this GUI.

19.2 System Development

A biometric system is essentially a pattern recognition system that operates by acquiring biometric data from an individual, extracting a feature set from the acquired data, and comparing this feature set against the template set in the database. Depending on the application context, a biometric system may operate either in verification mode or in identification mode. In the verification mode, the system validates a person's identity by comparing the captured biometric data with his/her own biometric template(s) stored in the centralized system database. In the identification mode, the system recognizes an individual by searching the templates of all the person in the database for a match. Therefore, the system conducts a one-to-many comparison to establish an individual's identity. A biometric system is designed using the following five main modules. Enrolment module is the first module where user's biometric data is captured using sensor. The second module is image enhancement module, which improves the visibility of any portion or feature of the image and suppresses the information in other parts. It is done after enrolment is completed. It includes brightening, sharpening, adjusting contrast, etc., so that the image is usable for further processing. Feature extraction module is the third module, where the acquired biometric data is processed to extract a set of salient and discriminatory features. In this study, the position and orientation of minutiae points (local ridges and valley singularities) in a fingerprint are extracted. The fourth module is matching module in which the features that have been extracted during recognition are compared against the stored template(s) to generate matching scores. The number of matching minutiae between the input and the template fingerprint image is determined and a matching score is computed in this work. This module also encapsulates a decision-making module, in which a user's claimed identity is confirmed (verification) or a user's identity is established (identification) based on the matching score. The last module is system database module, which is used to store the biometric templates of the enrolled users in the centralized database of the biometric system.

In this study, a GUI has been developed to support the understanding of aforesaid operations that are used in fingerprint recognition. The GUI for fingerprint recognition with 2-axis and 8 panels has been developed using MATLAB and is shown in Fig. 19.1. The first axis is for the input image and the second is to display the output image after performing the operations on the input image. Panel-1 is for basic operations that contains *Load Image* and *Reset* buttons. *Load Image* button is to import the image for processing and *Reset* button is used to restart the processing. Panel-2 is the image enhancement panel that contains *Normalization*, *Orient. & Ridge Freq.*, *Filtering*, *Bin. & Thin.*, and *Masking* buttons. Panel-3 is the histogram panel of GUI with two buttons, namely, *Histogram* and *Histogram EQV* which are used for the

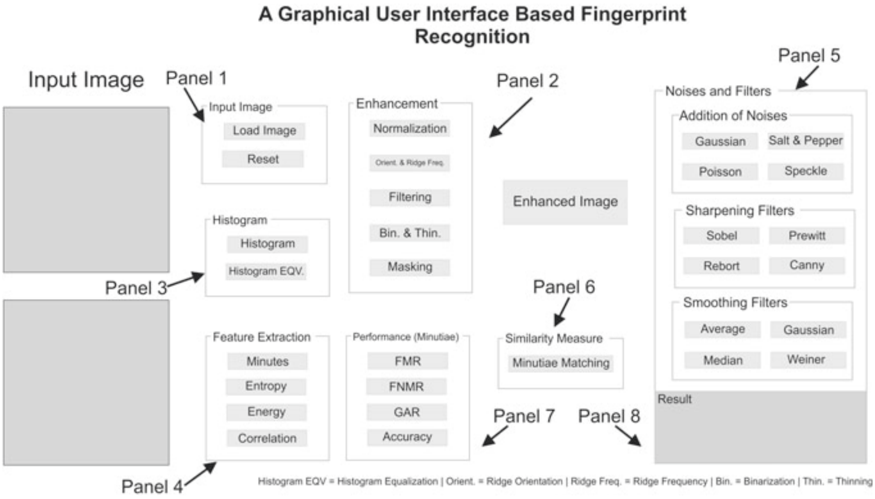


Fig. 19.1 GUI of fingerprint recognition

adjustment of intensity of the input image. Panel-4 is used for feature extraction such as *Minutiae*, *Entropy*, *Energy*, and *Correlations* of a fingerprint image to recognize an individual. Addition of artificial noises like *Gaussian*, *Salt and Pepper*, *Poisson*, and *Speckle* and their removal using various sharpening and smoothing filters have been shown in Panel-5.

Panel-6 is developed for similarity measure and Panel-7 is to measure the performance parameters of the input image. Both of these contain buttons for evaluating False Match Rate (*FMR*), Genuine Acceptance Rate (*GAR*), False Non-Match Rate (*FNMR*), and *Accuracy*. Panel-8 is used for display of numerical values of various operations.

19.3 Proposed System Functions

In this study, a prototype GUI of fingerprint recognition has been developed and tested on DB1 of FVC2006 database which consists of 1800 fingerprint images. The functionalities of fingerprint GUI system are as follows.

19.3.1 Panel-1: Basic Operation Panel

This panel contains *Load Image* and *Reset* buttons which are used to upload the input fingerprint image from the database using *imread*() function and to restart the process of the system, respectively.

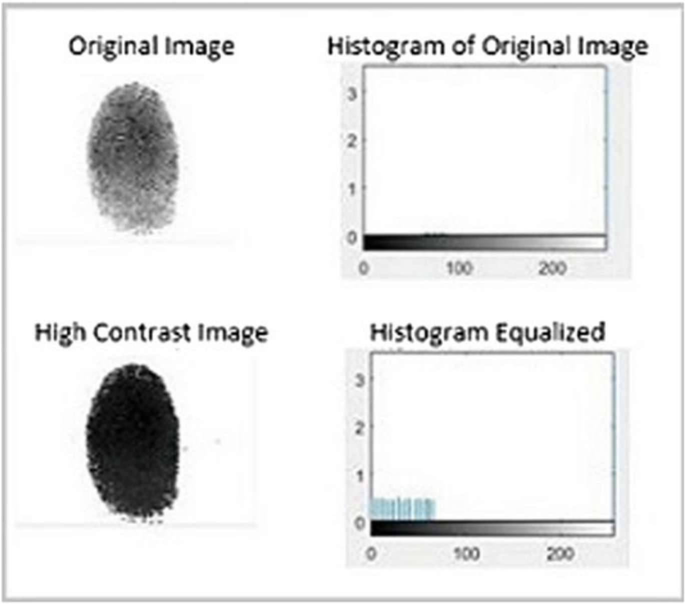


Fig. 19.2 Histogram and histogram equalization of an image

19.3.2 Panel-2: Histogram

Histogram and histogram equalization are the techniques that provide a sophisticated method for modifying the dynamic range and contrast of an image by altering that image such that its intensity of the desired shape. Histogram technique may employ nonlinear and non-monotonic transfer functions to map between pixel intensity values in the input and output images. Histogram of an image represents relative frequency of occurrence of various gray levels. In a 2-dimensional plot, x-axis represents gray levels and y-axis represents the number of pixels in each gray level. Histogram equalization employs a monotonic, nonlinear mapping which reassigns the intensity values of the pixel in the input image such that the output image contains a uniform distribution of intensities. Therefore, histogram equalization generally is used to enhance the contrast of an image [11]. Figure 19.2 shows histogram and histogram equalized image of a fingerprint. The function *imhist*() is used to generate the histogram for the fingerprint image and *histeq*() is used to equalize the histogram, respectively.

19.3.3 Panel-3: Image Enhancements

The fingerprint image enhancement algorithm, involves a set of intermediate operations which are applied to input fingerprint image, generates output enhanced fingerprint image [12, 13]. A gray-level fingerprint image, *im*, is defined as N×N matrix where *im* (i, j) represents the intensity of the pixel at ith row and jth column. As per FBI recommendation the fingerprint images should be scanned at a resolution of 500 dpi. The mean and variance of the gray-level fingerprint image, *im*, are defined, respectively, as

$$\bar{im} = \sum_{i=1}^{N} \frac{im_i}{N},$$

and

$$var(im) = \sum_{i=1}^{N} \frac{(im_i - \bar{im})^2}{N}$$

Normalization operation on fingerprint image is applied to obtain a pre-specified mean and variance. This operation is performed using *ridgesegment*(*im*, *blksze*, *thresh*) function where *im* is fingerprint image to be segmented, *blksze* is the block size over which the standard deviation is determined and *thresh* is threshold of standard deviation for ridge region. This operation can be performed using the *Normalization* button. An orientation image, $O(i, j)$, represents the local ridge frequency at pixel (i, j) and the size of this image would also be N×N. This operation is performed for a block rather than at every pixel. Therefore, the normalized image is divided into a set of M×M nonoverlapping blocks.

To calculate the ridge frequency, the function *ridgefreq*(*im*, *mask*, *orientim*, *blksze*, *windsze*, *minWaveLength*, *maxWaveLength*) is used where *im* is normalized fingerprint image, *mask* defines ridge regions obtained from *ridgesegment*(), *orientim* is ridge-oriented fingerprint image obtained from *ridgeorient*() function, *blksze* is size of image block to be used, *windsze* is window length used to identify peaks, *minWaveLength* and *maxWaveLength* are minimum and maximum ridge wavelengths. To obtain *orientim*, the function *ridgeorient*(*im*, *gradientsigma*, *blocksigma*, *orientsmoothsigma*) is used where *gradientsigma* is used to compute image gradients, *blocksigma* is sigma of the Gaussian weighting used to sum the gradient moments and *orientsmoothsigma* is used to smooth the final orientation vector field. *Orient. & Ridge Freq.* button can be used for this operation.

The configurations of parallel ridges and furrows with well-defined frequency and orientation in a fingerprint image provide useful features. A band-pass filter is used to remove the undesired noise and preserve the true ridge and furrow structure. The filtered image can be obtained using the function *ridgefilter*(*im*, *orientim*, *freqim*, *kx*, *ky*, *showfilter*) where *freqim* is ridge frequency image obtained from *ridgefreq*(), *kx*, and *ky* are scale factors specifying the filters sigma relative to the wavelength of the filter, *kx* controls the sigma in x-direction which is along the filter and hence

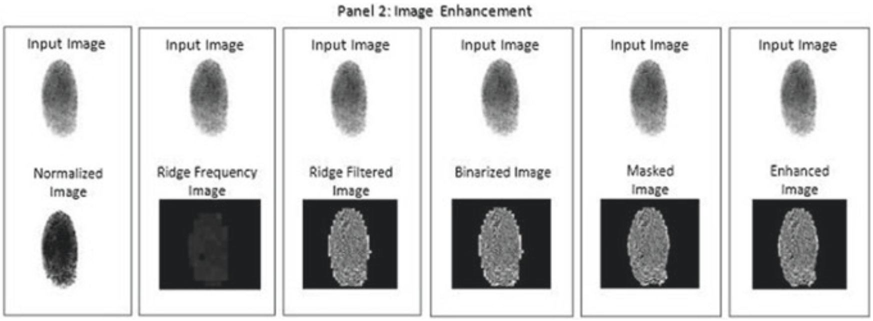


Fig. 19.3 Image enhancement processes

control the bandwidth of the filter and *ky* controls the sigma across the filter and hence controls the orientational selectivity of the filter, and *showfilter* is an optional flag having values either 0 or 1. This operation is performed using the *Filtering* button of GUI. The image obtained from the filtering operation is binarized and thinned to make it more suitable for feature extractions. The success of these operations depends on the difference between the mean of the considered blocks. The binary image then submitted to the thinning algorithm which reduces the ridge thickness to one pixel wide. The operation that converts a gray-scale image into a binary image is known as binarization. Pixels with the value 0 are displayed as black and with 1 are displayed as white and thinning operation is performed using the function *bwmorph*(). *Bin. & Thin.* button is used for this operation. The last operation to get an enhanced image is masking which is performed to obtain the mask region by classifying each block in the fingerprint image into recoverable and unrecoverable blocks. The output images of the above discussed operations are shown in Fig. 19.3.

All the enhancement operations can be performed together through *Enhanced Image* button on the GUI panel.

19.3.4 Panel-4: Feature Extraction

The feature of an image is defined as a function of one or more measurements each of which specifies some quantifiable properties of the image and is computed to quantify significant characteristics of the image. One can say that feature extraction is a process of extracting the information from the image such that distinctive properties of extracted features help in differentiating between the categories of input patterns [14, 15]. Feature extraction is also preferred to reduce the cost of feature measurements, to increase classifier efficiency, and allows higher matching accuracy to identify an individual. The features extracted in this study for fingerprints are minutiae, energy, entropy, and correlation.

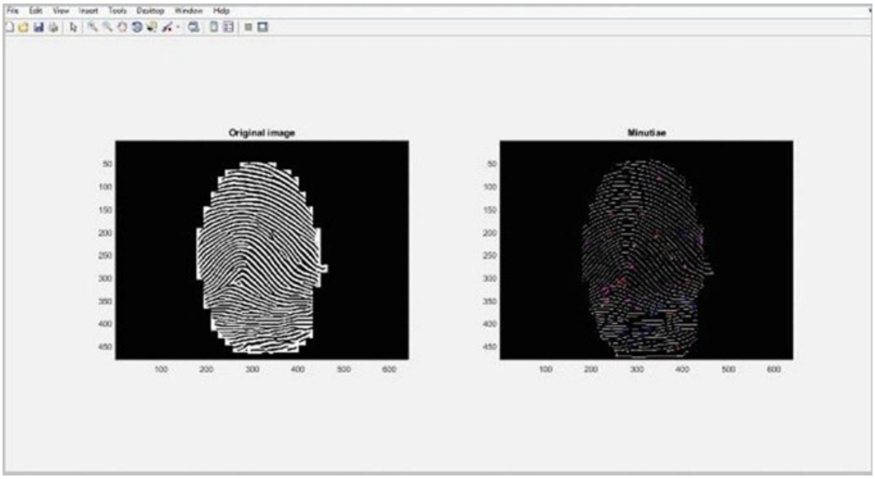


Fig. 19.4 Minutiae of fingerprint

19.3.4.1 Minutiae Features

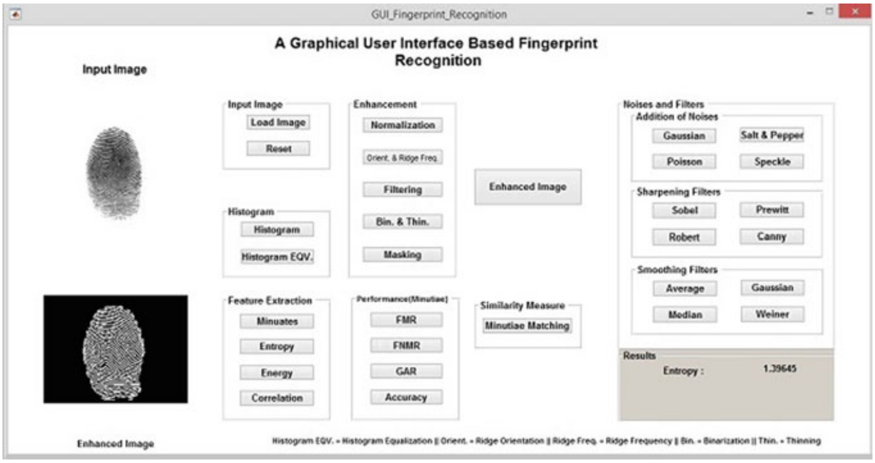
Most of the fingerprint scanned technologies are based on minutiae-based techniques that represent the fingerprint by its local features like ridge terminations and ridge bifurcations. Ridge termination is a point where a ridge ends abruptly and ridge bifurcation is the point where a ridge forks or diverges into branch ridges. Collectively these features are called minutiae [16]. A good quality fingerprint image typically contains 40–100 minutiae. Two fingerprints match if their minutiae points match and this approach is being intensively used in the available commercial fingerprint biometric system. In this study the function *ext_finger*(*im*, *display_flag*) is used to compute minutiae where *im* is fingerprint image and *display_flag* is a flag to display an image. The minutiae's of the input fingerprint can be seen using the *Minutiae* button on GUI as shown in Fig. 19.4.

19.3.4.2 Entropy

The entropy of an image is defined as a measure of the average information content. In other words, it is a statistical measure of randomness that can be used to characterize the texture of an image. Mathematically it is defined as

$$E_n = -\sum_{i=1}^{k} P_i \log_2 P_i \ ,$$

where P_i is the ith frequency value generated from k-bin normalized intensity histogram of the image. The normalized values are computed by dividing each frequency



Entropy : 1.39645

Histogram EQV. = Histogram Equalization || Orient. = Ridge Orientation || Ridge Freq. = Ridge Frequency || Bin. = Binarization || Thin. = Thinning

Fig. 19.5 Entropy computed of a fingerprint

count by the sum of the pixels in the image using the equation

$$P_i = \frac{f_i}{N}$$

Here, f_i is the ith frequency value of the histogram and N is the total number of pixels. The entropy of the fingerprint image can be seen in the result panel using the *Entropy* button. For a good quality fingerprint image, the entropy should have a lower value (Fig. 19.5).

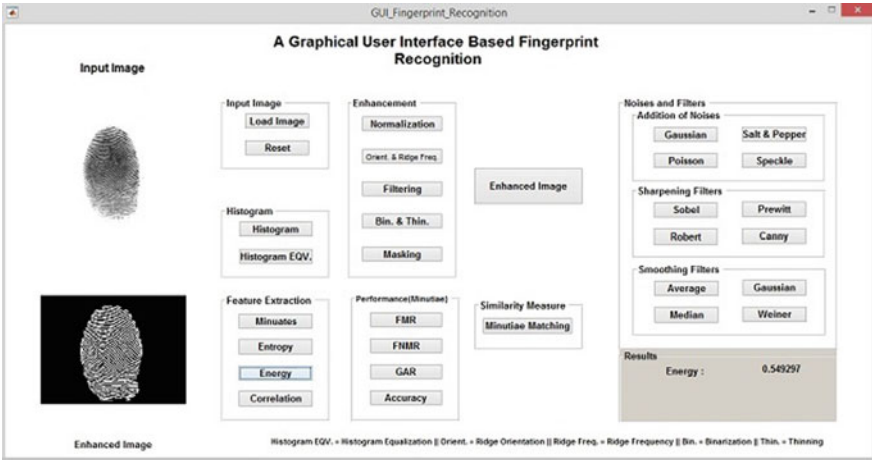
19.3.4.3 Energy

Energy in image processing has different meanings depending on the context. There are more than one definitions of energy in image processing as it depends on the context where it is being used. In fingerprint image, energy is used to describe a measure of information while formulating an operation under a probability framework. It is defined as

$$\sum_{i,j} p(i,\, j)^2$$

where $p(i, j)$ represents the probability of ith row and jth column pixel in the image.

The function *graycoprops*(*GLCM*, *properties*) is used to compute the energy of a given fingerprint where *GLCM* is gray-level co-occurrence matrix obtained by the pre-defined function *graycomatrix*() and properties is a constant value. For example, to compute energy, the value for properties will be "*energy*", for entropy and correlation, the values will be *entropy* and *correlation*, respectively. The values of these



Results

Energy : 0.549297

Histogram EQV. = Histogram Equalization || Orient. = Ridge Orientation || Ridge Freq. = Ridge Frequency || Bin. = Binarization || Thin. = Thinning

Fig. 19.6 Energy computed from fingerprint image

parameters for a given fingerprint image can be seen in the result panel using their respective button (Fig. 19.6).

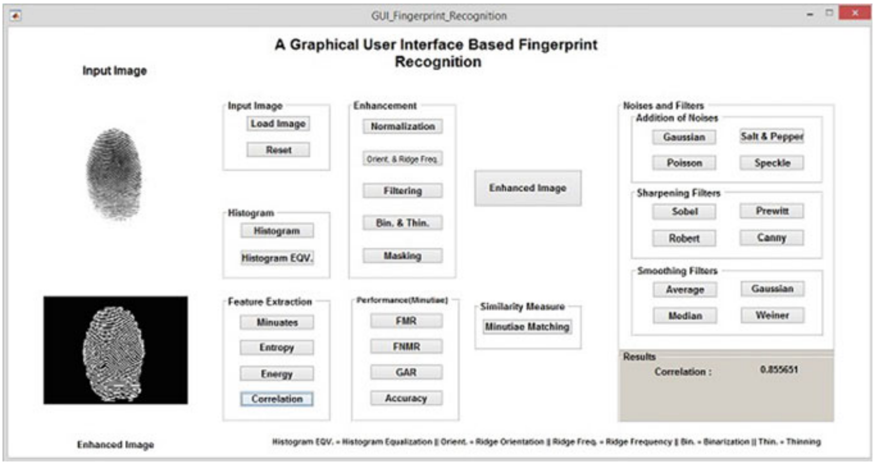
19.3.4.4 Correlation

Correlation is a method for establishing the degree of probability that a linear relationship exists between two measured quantities [17]. In 1895, Karl Pearson defined the Pearson product-moment correlation coefficient, r. Pearson's correlation coefficient was the first formal correlation measure and is widely used in statistical analysis, pattern recognition, and image processing. For monochrome digital images, the Pearson's correlation coefficient is defined as

$$r = \frac{\sum_i (x_i - x_m)(y_i - y_m)}{\sqrt{\sum_i (x_i - x_m)^2}\sqrt{\sum_i (y_i - y_m)^2}}$$

where x_i and y_i are intensity values of ith pixel in first and second image, respectively. Also, x_m and y_m are mean intensity values of first and second image, respectively. The correlation coefficient has the value $r = 1$ if the two images are absolute identical, $r = 0$ if they are completely uncorrelated and $r = -1$ if they are completely anti-correlated. The Pearson product-moment correlation coefficient is a dimensionless index which is invariant to linear transformations of either variable.

In this study, the correlation of fingerprint image is computed using *Correlation* button. The function *graycoprops*(*GLCM*, *properties*), where *GLCM* is gray-level co-occurrence matrix obtained by the pre-defined function *graycomatrix*() and properties will be given as "*correlation*", will be executed to compute the correlation. The



Correlation : 0.855651

Histogram EQV. = Histogram Equalization || Orient. = Ridge Orientation || Ridge Freq. = Ridge Frequency || Bin. = Binarization || Thin. = Thinning

Fig. 19.7 Correlation computed from fingerprint image

result of correlation of a fingerprint image is shown in result section of the interface which is shown in Fig. 19.7.

19.3.5 *Panel-5: Noises and Filters*

This panel describes the types of noises that are generally introduced during the acquisition and transmission of fingerprint images and their removal using the smoothing and sharpening filters.

19.3.5.1 Noises

Noise is an evitable problem of image processing which occurs in the image during image acquisition or image transmission. Noises can be differentiated on the basis of their characteristics like intensity, wavelength, etc. and can be introduced in the image to study their effects on the image [18]. Few prominent noises used in this study are discussed here briefly.

Gaussian Noise

The Gaussian noise is generally uniformly distributed over the image. The distribution function of Gaussian noise is given by

$$p(z) = \frac{e^{\frac{(z-\mu)^2}{2a^2}}}{\sqrt{2\pi\sigma}}$$

where z is gray-level, μ is average or mean of a function, σ is standard deviation of a noise and $p(z)$ is the probability density function. In an image default value of Gaussian noise has zero mean value and 0.05 variance. The Gaussian noise can be computed using *Gaussian* button.

Salt and Pepper Noise

Salt and pepper noise is sparsely occurring white and black pixels in an image. The pixels of an image is corrupted or not can be represented by a probability function q whose values lie in the range $0 \leq q \leq 1$. A system can introduce salt and pepper noise in an image by setting a fraction $q/2$ randomly for black and another $q/2$ fraction for white pixels. This noise in an input image can be seen using *Salt & Pepper* button.

Poisson Noise

Poisson noise is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space in an image. If these events occur with a nonconstant rate independent of time. This noise in the image can be seen using *Poisson* button of the panel.

Speckle Noise

Speckle is a granular noise that inherently exists in the image that degrades the quality of the image. The large area of surfaces, synthetic, or natural is extremely rough. Images obtained from these surfaces generally suffer from speckle noise. The function available in MATLAB to introduce these noises is *imnoise*(*im*, *type*, *parameters*) where *im* is fingerprint image, *type* defines the type of noise to be introduced and *parameters* is an optional parameter to the function that defines the intensity ranging from 0 to 1. The noises discussed above are shown in Fig. 19.8.

19.3.5.2 Sharpening Filters

Sharpening is a technique for increasing the sharpness of an image which is a combination of two factors: resolution and acutance [19]. Resolution is the number of the pixels in an image, i.e., higher the resolution, more pixels are required to sharpen the image. Acutance is the measure of the contrast at an edge of the image. Edges that have more contrast appear to have a more defined edge to the human visual system. Image sharpening refers to enhancement technique that highlights the edges, line structures, and fine details in an image. High-pass filters are used for sharpening of an image. The high-pass filters which have been used in this study are Sobel, Prewitt,

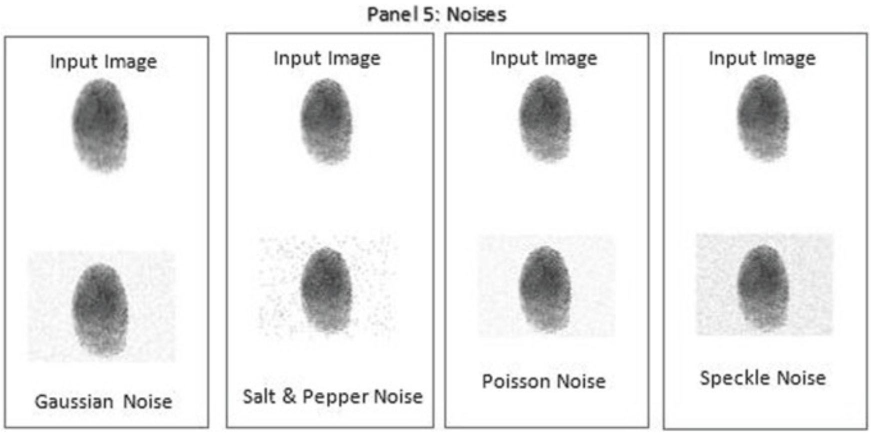


Fig. 19.8 Noises added to the fingerprint image

Robert, and Canny filters. Sobel filter is used to detect the edges in the image. The Sobel filter reduces the visibility of those regions in the image where the intensity changes slowly which allows to highlight the edges. A 2-D gradient is computed by Sobel operator to find out the edge strength at each point of the image. Normally a 3×3 matrix is used as a gradient along the x-axis and y-axis. Prewitt filter is another operator which is used for edge detection in the image by calculating the gradients of the image intensity at each point. The resulted image shows the smooth or abrupt changes in the image at that point that helps to detect the edge and its orientation. This filter detects edges horizontally and vertically and is, therefore, computationally inexpensive in comparison to Sobel. Robert and Canny filters are also used to detect the edges in the image.

Function which is used to compute the gradients for these filters is *edge*(*im*, *type*, *thresh*) where *im* is fingerprint image, *type* defines the type of filter to be used and *thresh* is an optional parameter to the function. The output of sharpening filter can be seen in Fig. 19.9.

19.3.5.3 Smoothing Filters

Smoothing filters are low-pass filters that are often used to reduce noise within an image or to produce a less pixelated image [20]. Average, Gaussian, Median, and Wiener are the four filters that have been used in this study. These filters can be used to replace each data point by local average of surrounding data points in an image. Function *filter2*(*fspecial*(*type*, *hsize*), *im*) has been used to reduce the noise from the image where *fspecial*() function has parameters *type* is average filter and *hsize* is the size of filter and *im* is the fingerprint image. The function *imgaussfilt*(*im*) is used to remove the noise using Gaussian filter. The *medfilt2*(*im*, [*m n*]) function is used to

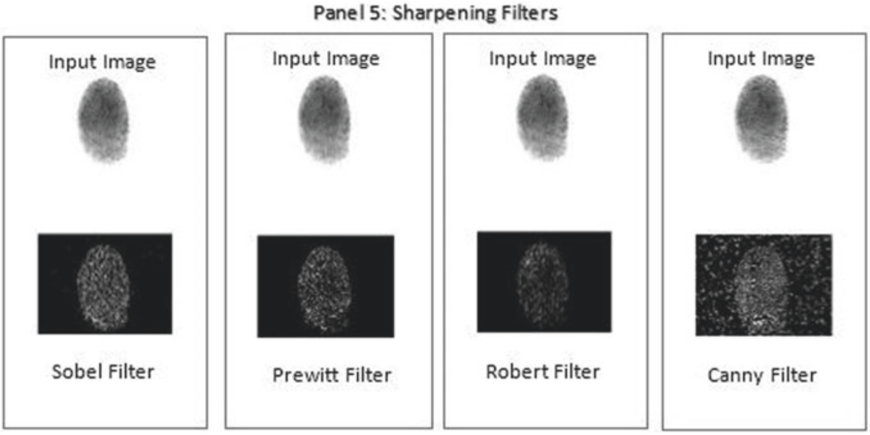


Fig. 19.9 Sharpening filters

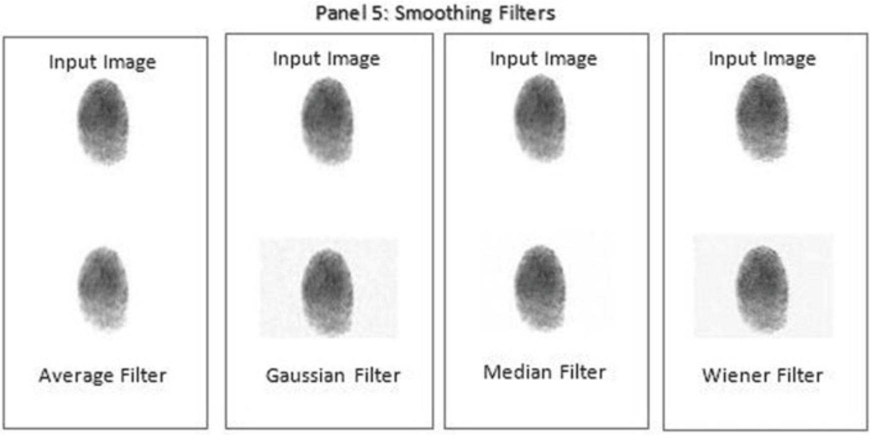


Fig. 19.10 Smoothing filters

smooth the fingerprint image and removing noise using median filter. The Wiener filter uses function *wiener2*(*im*, [*m n*]) to remove the noise from a fingerprint image. The output of all these filters can be obtained using their respective buttons on the interface that has been shown in Fig. 19.10.

19.3.6 Panel-6: Similarity Measure

Biometric system using fingerprints as a biometric trait stores a user's fingerprint data in the form of a template and compact form of an image. The template is considered to be an accurate representation of user's biometric feature. A fingerprint template

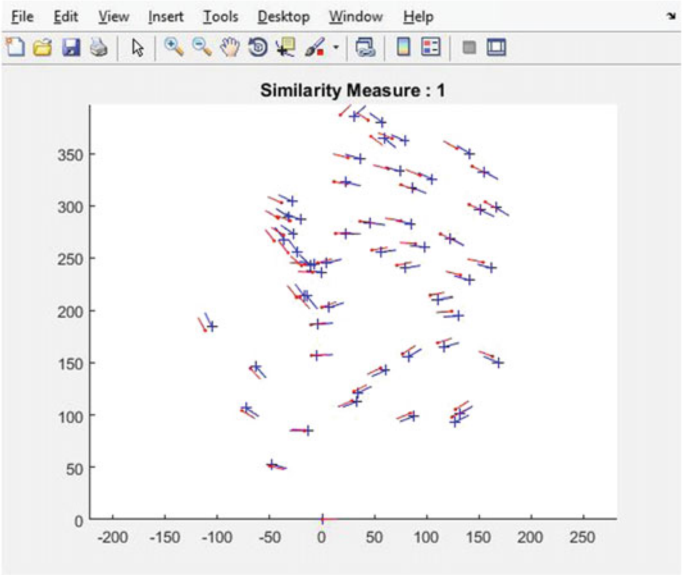


Fig. 19.11 Similarity measure

contains spatial information about the minutiae, which are generally between 30 and 100 points. A match of 6–8 minutiae points is usually considered sufficient for verification of an individual's fingerprint. The similarity measure does not provide the inference that two templates in the database belong to the same person rather it is an indicator of the level of difficulty of recognizer in comparing the two templates [21, 22]. A pair of template having an extreme similarity value, i.e., either too low or too high should be correctly classified by the biometric system with ease and pairs with intermediate values are going to take a greater computation from matcher to classify. The similarity measure gives good result in cases where the matching is either extremely too low or too high. For perfect match, the matching probability should be extremely high and for mismatch is should be extremely low. The function which is used in this study to compute the similarity metric is *match*(*M1*, *M2*, *display_flag*) where *M1* is minutiae of the input image, *M2* is minutiae in the database, and *display_flag* is the flag that defines the display of the image. Figure 19.11 shows the similarity measure of the input and the matched images which can be seen using *Minutiae Matching* button.

19.3.7 Panel-7: Performance Parameters

This panel is used to evaluate the performance parameters of fingerprint recognition system. In this study, False Match Rate (FMR), False Non-Match Rate (FNMR), Genuine Acceptance Rate (GAR), Equal Error Rate (ERR), and Accuracy have been computed. The FMR is the percentage of invalid inputs that are incorrectly accepted (match between input and a nonmatching template). The FNMR is the percentage of valid inputs that are incorrectly rejected (fails to detect a match between input and matching template). The accuracy of the system could also be expressed in terms of ERR which is the value of the FNMR at a particular threshold when it is equal to the FMR [23–25].

19.3.8 Panel-8: Results

This panel is used to display various numerical values computed for different operations performed for fingerprint study.

19.4 Results and Discussion

In this study, various performance parameters have been computed to test the performance and efficiency of the proposed GUI of fingerprint recognition system. A dataset of 1800 fingerprint images has been stored in the database to compute parameters like FAR, FRR ERR, and Accuracy. The False Acceptance Rate (FAR) is the measure of the likelihood that the biometric security system incorrectly provides an access to unauthorized user. The result shows that when the threshold is low the system accepts fingerprints of imposter users but when the threshold value starts increasing the probability of providing access to imposter users starts decreasing because more feature matchings are required for identification. This study shows that zero FAR is achieved at the threshold 0.45, i.e., no fingerprint of imposter users will be accepted after this threshold.

The False Rejection Rate (FRR) is defined when the biometric security system incorrectly rejects access to authorized users. The result of this parameter shows that when the threshold value is high the system rejects the genuine users but as the threshold value starts decreasing the probability of rejecting authorized users starts decreasing. Some of the possible reasons for rejecting an authorized user at some high threshold value could be variations in the skin conditions, impression conditions such as scars, humidity, dirt, and nonuniform contact with the biometric system during capturing of fingerprint images, variable pressure on the fingerprint capturing device by user and variable area of contact with the device. In this study, zero FRR is achieved at threshold 0.19, i.e., no authorized user will be rejected after this threshold value.

The Equal Error Rate (EER) is termed as the crossover point on a graph that has both FAR and FRR curves plotted. The crossover point shows that the probability of accepting and rejecting an imposter and an authorized user will be same at this point. Performance parameters EER, FAR, and FRR has been calculated and plotted

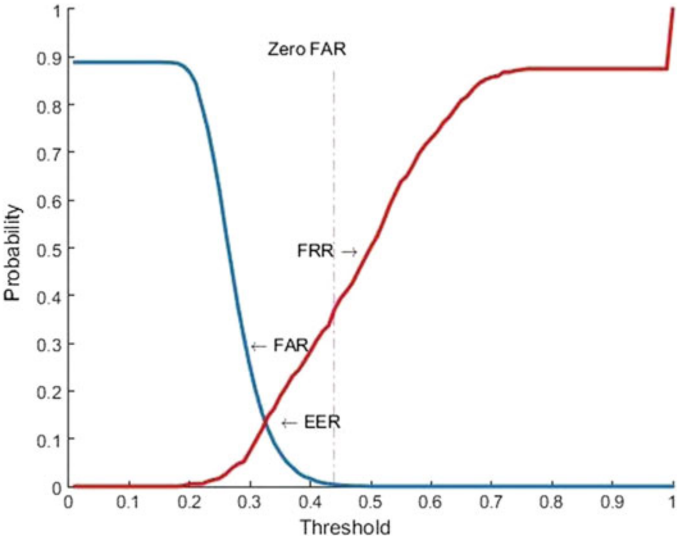


Fig. 19.12 Equal error rate (ERR)

for threshold value from 0 to 1 on the graph as shown in Fig. 19.12. Here, the EER lies near the threshold 0.32, i.e., the system will accept and reject the same number of imposter users and authorized users at this point, respectively. Figure 19.12 indicates the inverse relationship of FAR and FRR rates by plotting them against each other at different thresholds. The FAR is given by the percentage of comparisons between different fingerprints where the system has accepted the imposter users. The FRR is given by the percentage of comparisons between different samples of the same fingerprint where the system has rejected the authorized users. The point at which these two probabilities cross is called EER. The steps involved to compute the performance parameters are given below:

1. Enroll n different fingerprint records for the first time in the database. Every record must be unique and labeled as $F = \{f_1, f_2, f_3, \ldots, f_n\}$. Here the value of n is 1800.
2. Enroll same fingerprints of an individual again and label them as $P = \{p_1, p_2, p_3, \ldots, p_n\}$.
3. Perform verification of all fingerprints from F against all records from P. Hence, there will be a 1800×1800 matching probability, i.e., 3240000 matching results for every pair.
4. Analyze for genuine distribution having the same label.
5. Analyze imposter distribution having different label.
6. Calculate FAR, FRR, and Accuracy using steps 1–5.

The FRR is the expected probability that f_i is not matched with p_i, i.e.,

Table 19.1 False reject rate

Label	Should: accept	Should: reject
Reality: accepted	TA(**1800**)	FA(0)
Reality: rejected	FR(**0**)	TR(3238200)

Table 19.2 False accept rate

Label	Should: accept	Should: reject
Reality: accepted	TA(1800)	FA(**0**)
Reality: rejected	FR(0)	TR(**3238200**)

$$FRR = \frac{FR}{(FR + TA)} \times 100$$

where *FR* is total number of false reject and *TA* is total number of acceptance. The computed *FRR* is found to be 0% and shown in Table 19.1. The FAR is the expected probability that f_i will be falsely declared to match p_i.

$$FAR = \frac{FA}{(FA + TR)} \times 100$$

where, *FA* is total number of false acceptance and *TR* is the total number of rejections. The computed FAR is found to be 0% and shown in Table 19.2. The accuracy of the system is calculated using the formula

$$Accuracy = \left(1 - \frac{(FRR + FAR)}{2}\right) \times 100$$

The accuracy is reported to be 100%.

19.5 Conclusion

Fingerprint recognition is one of the reliable and well-known biometrics recognition techniques. In this study, a GUI has been developed to recognize fingerprints of an individual using MATLAB. The proposed system will help the young researchers to understand various basic operations on images like histogram, histogram equalization, image enhancement, feature extraction, noise addition and removal, filtration of an image, computation of performance parameters, and similarity measures etc. visually. One of the best advantage of a GUI is that one can use it without having prior knowledge in this area. For example, if someone wishes to extract features of a fingerprint, they need not know about algorithms rather they can extract them on select of options provided onto the interface. In addition, GUIs are user-friendly that make learning intuitive, attractive, and interactive. In this study, various features have been extracted, namely, minutiae, entropy, energy, and correlation. These features have been used to obtain a match between the trained fingerprint image and the input test fingerprint image. The similarity measures have been computed for input test fingerprint image and the template which helped to recognize the exact match with the template. The proposed system has shown satisfactory results with 100% accuracy in recognizing a fingerprint. In the future work, we intend to analyze and test the results on larger datasets and to extend this GUI by providing the functionalities to compute other features of a fingerprint image that can help in recognizing an individual.

References

1. A. Jain, L. Hong, S. Pankanti, Biometric identification. Mag. Commun. ACM **43**(2), 90–98 (2000)
2. A.K. Jain, R. Bolle, S. Pankanti, *Biometrics: Personal Identification in Networked Society* (Springer Science & Business Media, Berlin, 2006)
3. K. Renaud, Evaluating authentication mechanisms, in *Security and Usability*, ed. by L. F. Cranor, S. Garfinkel (O'Reilly, Sebastopol, 2005)
4. International Biometrics Group (IBG): Biometrics market and industry report 2009–2014 (2009)
5. S.C. Dass, Fingerprint-based recognition. Int. Stat. Rev. **81**(2), 175–187 (2013)
6. E. Marasco, A. Ross, A survey on antispoofing schemes for fingerprint recognition systems. J. ACM Comput. surv. (CSUR) **47**(2), 28 (2015)
7. C. Sousedik, C. Busch, Presentation attack detection methods for fingerprint recognition systems: a survey. IET Biom. **3**(4), 219–233 (2014)
8. H. Hasan, S. Abdul-Kareem, Fingerprint image enhancement and recognition algorithms: a survey. Neural Comput. Appl. **23**(6), 1605–1610 (2013)
9. F. Liu, D. Zhang, L. Shen, Study on novel curvature features for 3D fingerprint recognition. Neurocomputing **168**, 599–608 (2015)
10. R.D. Labati, A. Genovese, V. Piuri, F. Scotti, Contactless fingerprint recognition: a neural approach for perspective and rotation effects reduction, in *IEEE Workshop on Computational Intelligence in Biometrics and Identity Management (CIBIM)* (2013), pp. 22–30
11. T. Celik, Two-dimensional histogram equalization and contrast enhancement. Pattern Recognit. **45**(10), 3810–3824 (2012)
12. I.G. Babatunde, A.B. Kayode, A.O. Charles, O. Olatubosun, Fingerprint image enhacement segmentation to thinning. Int. J. Adv. Comput. Sci. Appl. **3**(1), 15–24 (2012)
13. J.S. Bartunek, M.G. Nilsson, B. Sallberg, I. Claesson, Adaptive fingerprint image enhancement with emphasis on preprocessing of data. IEEE Trans. Image Process. **22**, 644–656 (2013)
14. R. Gayathri, P. Ramamoorthy, A fingerprint and palmprint recognition approach based on multiple feature extraction. Eur. J. Sci. Res. **76**(4), 514–526 (2012)
15. A. Shrivastava, D.K. Srivastava, Fingerprint identification using feature extraction: a survey, in *Proceedings of 2014 International Conference on Contemporary Computing and Informatics* (2014), pp. 522–525
16. Z. Jin, A.B.J. Teoh, B.-M. Goi, Y.-H. Tay, Biometric cryptosystems: a new biometric key binding and its implementation for fingerprint minutiae-based representation. Pattern Recognit. **56**, 50–62 (2016)
17. J. Peng, Q. Li, Q. Han, X. Niu, Feature-level fusion of finger biometrics based on multi-set canonical correlation analysis, in *Biometric Recognition Lecture Notes in Computer Science (LNCS)*, vol. 8232 (2013), pp. 216–224

18. S.-H. Ju, H.-S. Seo, S.-H. Han, J.-C. Ryou, J. Kwak, A study on user authentication methodology using numeric password and fingerprint biometric information. BioMed. Res. Int. **2013**(1), 427542 (2013)
19. M.A.U. Khan, T.M. Khan, Fingerprint image enhancement using data driven directional filter bank. Optik- Int. J. Light Electron Optics **124**(23), 6063–6068 (2013)
20. P. Sutthiwichaiporn, V. Areekul, Adaptive boosted spectral filtering for progressive fingerprint enhancement. Pattern Recognit. **46**(9), 2465–2486 (2013)
21. J. Torres-Sospedra, R. Montoliu, S. Trilles, O. Belmonte, J. Huerta, Comprehensive analysis of distance and similarity measures for wi-fi fingerprint indoor positioning systems. Expert Syst. Appl. **42**(23), 9263–9278 (2015)
22. J. Galbally, S. Marcel, J. Fierrez, Image quality assessment for fake biometric detection: application to iris, fingerprint, and face recognition. IEEE Trans. Image Process. **23**(2), 710–724 (2014)
23. Z. Jin, A.B.J. Teoh, T.S. Ong, C. Tee, Fingerprint template protection with minutiae-based bit-string for security and privacy preserving. Expert. Syst. Appl. **39**(6), 6157–6167 (2012)
24. H. Benaliouche, M. Touahria, Comparative study of multimodal biometric recognition by fusion of iris and fingerprint. Sci. World J. **2014**, 829369 (2013)
25. M.M. Ali, V.H. Mahale, P.L. Yannawar, A.T. Gaikwad, Fingerprint recognition for personal identification and verification based on minutiae matching, in *IEEE 6th International Conference on Advanced Computing (IACC)* (2016)

Chapter 20
Existence and Stability Results for Stochastic Fractional Delay Differential Equations with Gaussian Noise

P. Umamaheswari, K. Balachandran and N. Annapoorani

Abstract In this paper, the existence and uniqueness of solutions of stochastic fractional delay differential equations is obtained by using Picard–Lindelöf successive approximation scheme. Further, the stability results are established using the Mittag-Leffler function. Examples are provided to illustrate the theory.

20.1 Introduction

Stochastic Differential Equations (SDEs) are natural extension of deterministic. These equations play an important role in characterizing many physical, biological, and engineering problems. They are important from the viewpoint of applications since they incorporate randomness into the mathematical description of the phenomena and provide a more accurate description of it. Therefore, the theory of SDEs has developed quickly and the investigation for SDEs has attracted considerable attention [7, 17, 19]. On the other hand, fractional differential equations [6, 13, 18] describe the dynamical behavior of real-life phenomena more accurately than integer-order equations because of its ability to describe systems with memory and hereditary properties. It generalizes the concepts of derivative and integral of a function to a noninteger order.

The motivation for considering fractional differential equations with random elements comes from the fact that many phenomena in science that have been modeled by fractional differential equations have some uncertainty. Therefore, it is important to analyze the solution of stochastic fractional differential equations. These equations have physical applications in many fields such as turbulence, heterogeneous flows

P. Umamaheswari (✉) · K. Balachandran · N. Annapoorani
Department of Mathematics, Bharathiar University, Coimbatore 641046, India
e-mail: umamaths.tamil@gmail.com

K. Balachandran
e-mail: kb.maths.bu@gmail.com

N. Annapoorani
e-mail: pooranimaths@gmail.com

P. Manchanda et al. (eds.), *Mathematical Modelling, Optimization, Analytic and Numerical Solutions*, Industrial and Applied Mathematics,
https://doi.org/10.1007/978-981-15-0928-5_20

and materials, viscoelasticity, and electromagnetic theory [25, 26]. Many authors [7, 14, 19, 21] discussed the existence and uniqueness of the solution of stochastic differential equations. Pedju and Ladde [20] studied the existence of solutions of stochastic fractional differential equations using an independent set of time scales.

The concept of stability is extremely important because almost every workable control system is designed to be stable. It means that the system remains in a constant state unless affected by an external action and returns to a constant state when the external action is removed. Balachandran et al. [1], Luo [16] and Khasminskii [12] discussed the stability of stochastic differential equations. Taniguchi [24] discussed the exponentially asymptotic stability of the stochastic evolution equations. Exponential stability for stochastic neutral partial functional differential equations was obtained by Govindan using semigroup theory [8, 10]. Stability of fractional dynamical systems is studied by many authors [5, 11, 22]. Delay differential equations are often used as tools in several areas of applied mathematics including the study of epidemics, population dynamics, automation, control theory, industrial robotics, and so on. The literature related to the existence of solutions of fractional order delay differential equations is extensive. See, for instance, [3, 4]. For stochastic equations with delay, one can refer [9, 28]. In this paper, we prove the existence of solutions of stochastic fractional delay differential equations and stability analysis of such equations.

20.2 Preliminaries

Now we present a few well-known concepts of fractional and stochastic differential equations.

Definition 20.1 *(Riemann–Liouville fractional integral)* The Riemann–Liouville fractional integral operator of order $\alpha > 0$ of a function $f \in L^1(\mathbb{R}^+)$ is defined as

$$I_{0+}^{\alpha} f(t) = \frac{1}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} f(s)\, ds, \; t > 0, \tag{20.1}$$

where $\Gamma(\cdot)$ is the Euler gamma function.

Definition 20.2 *(Riemann–Liouville fractional derivative)* The Riemann–Liouville fractional derivative of order $\alpha > 0$, $n-1 < \alpha < n$, $n \in \mathbb{N}$, is defined as

$$D_{0+}^{\alpha} f(t) = \left(\frac{d}{dt}\right)^n I_{0+}^{n-\alpha} f(t) = \frac{1}{\Gamma(n-\alpha)} \left(\frac{d}{dt}\right)^n \int_0^t (t-s)^{n-\alpha-1} f(s)\, ds, \tag{20.2}$$

where the function $f(t)$ has absolutely continuous derivatives up to order $(n-1)$.

Definition 20.3 *(Caputo fractional derivative)* The Caputo fractional derivative of order $\alpha > 0, n-1 < \alpha < n, n \in \mathbb{N}$, is defined as

$$^{C}D^{\alpha}_{0+}f(t) = \frac{1}{\Gamma(n-\alpha)} \int_0^t (t-s)^{n-\alpha-1} f^{(n)}(s)\, ds, \tag{20.3}$$

where the function $f(t)$ has absolutely continuous derivatives upto order n.

Definition 20.4 *(Mittag-Leffler Function)* The one-parameter Mittag-Leffler function is defined by

$$E_\alpha(z) = \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\alpha k + 1)}, \; (z \in \mathbb{C}, \; Re(\alpha) > 0). \tag{20.4}$$

A two-parameter function of the Mittag-Leffler type is defined by

$$E_{\alpha,\beta}(z) = \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\alpha k + \beta)}, \; (z, \beta \in \mathbb{C}, \; Re(\alpha) > 0). \tag{20.5}$$

In particular, when $\beta = 1$ then $E_{\alpha,1}(z) = E_\alpha(z)$. The Mittag Leffler function of a matrix A is defined by

$$E_{\alpha,\beta}(At) = \sum_{k=0}^{\infty} \frac{(At)^k}{\Gamma(\alpha k + \beta)}, \; (\alpha, \beta > 0, \; A \in \mathbb{R}^{n \times n}).$$

Definition 20.5 *(Stochastic Process)* A collection $\{X(t) | \, t \geq 0\}$ of random variables is called a stochastic process.

Definition 20.6 *(Chebyshev's Inequality)* If X is a random variable and $1 \leq p < \infty$, then

$$\mathbb{P}(|X| \geq \lambda) \leq \frac{1}{\lambda^p} \mathbb{E}(|X|^p) \; for \; all \; \lambda > 0.$$

Lemma 20.1 (Borel Cantelli Lemma) *If* $\{A_k\} \subset \mathscr{F}$ *and* $\sum_{k=1}^{\infty} \mathbb{P}(A_k) < \infty$, *then*

$$\mathbb{P}\left(\lim_{k \to \infty} \sup \; A_k \right) = 0.$$

Theorem 20.1 *(i) If* $\{X_n\}_{n=1}^{\infty}$ *is a submartingale, then*

$$\mathbb{P}\left(\max_{1 \leq k \leq n} X_k \geq \lambda \right) \leq \mathbb{E}(X_n^+)$$

for all $n = 1, 2, \ldots$ *and* $\lambda > 0$.

(ii) If $\{X_n\}_{n=1}^{\infty}$ *is a martingale and* $1 < p < \infty$*, then*

$$\mathbb{E}\left(\max_{1\le k\le n}|X_k|^p\right) \le \left(\frac{p}{p-1}\right)^p \mathbb{E}(|X_n|^p)$$

for all $n = 1, 2, \ldots$.

20.3 Existence and Uniqueness

In this section, we prove the existence and uniqueness of solution of nonlinear stochastic fractional delay differential equations with Gaussian noise. Here the successive approximation technique is used to obtain the existence of the solution [27]. For convenience $x(t,\omega), t \ge 0$ and $\omega \in \Omega$ can be written as $x(t)$ throughout this paper. Consider the stochastic fractional delay differential equation of the form

$$\left.\begin{aligned} {}^{C}D^{\alpha}x(t) &= b(t,x(t),x(t-\delta)) + \sigma(t,x(t),x(t-\delta))\frac{dW(t)}{dt},\ \ t \in J = [0,T] \\ x(t) &= \xi(t),\ \ t \in [-\delta, 0], \end{aligned}\right\} \tag{20.6}$$

where $\alpha \in (1/2, 1),\ \ \delta > 0,\ \ b \in C(J \times \mathbb{R}^n \times \mathbb{R}^n, \mathbb{R}^n),\ \sigma \in C(J \times \mathbb{R}^n, \mathbb{R}^{n\times m})$ and $W = \{W(t), t \ge 0\}$ is an m-dimensional Brownian motion on a complete probability space $(\Omega, \mathscr{F}, \mathscr{P})$. We can rewrite the Eq. (20.6) in its equivalent integral form as

$$\begin{aligned} x(t) = \xi(0) &+ \frac{1}{\Gamma(\alpha)}\int_0^t (t-s)^{\alpha-1} b(s, x(s), x(s-\delta))\,ds \\ &+ \frac{1}{\Gamma(\alpha)}\int_0^t (t-s)^{\alpha-1}\sigma(s, x(s), x(s-\delta))\,dW(s). \end{aligned} \tag{20.7}$$

Theorem 20.2 (Existence and Uniqueness) *Assume that* $(t,x) \in J \times \mathbb{R}^n$, $\alpha \in (1/2, 1)$, $b \in C(J \times \mathbb{R}^n \times \mathbb{R}^n, \mathbb{R}^n)$, $\sigma \in C(J \times \mathbb{R}^n \times \mathbb{R}^n, \mathbb{R}^{n\times m})$*, and* $W = \{W(t), t \ge 0\}$ *is an* m*-dimensional Brownian motion on a complete probability space* $(\Omega, \mathscr{F}, \mathscr{P})$*. Suppose the following inequalities hold:*
(i) Linear growth condition:

$$|b(t,x,y)|^2 + |\sigma(t,x,y)|^2 \le K_1^2(1 + |x|^2 + |y|^2) \tag{20.8}$$

$$|y|^2 \le K_2^2(1 + |x|^2) \tag{20.9}$$

for some constant $K_1,\ K_2 > 0$.
(ii) The Lipschitz condition:

$$|b(t,x_1,y_1)-b(t,x_2,y_2)|^2+|\sigma(t,x_1,y_1)-\sigma(t,x_2,y_2)|^2\le L_1^2(|x_1-x_2|^2+|y_1-y_2|^2) \tag{20.10}$$

$$|y_1-y_2|^2\le L_2^2(|x_1-x_2|^2) \tag{20.11}$$

for some constant $L_1,\ L_2>0$.

Let $\xi(0)$ *be a random variable defined on* $(\Omega,\mathscr{F},\mathscr{P})$ *and independent of the* σ*-algebra* $\mathscr{F}_s^t\subset\mathscr{F}$ *generated by* $\{W(s),t\ge s\ge 0\}$ *and such that* $\mathbb{E}|\xi(0)|^2<\infty$. *Then the initial value problem (20.6) has a unique solution which is* t*-continuous with the property that* $x(t,\omega)$ *is adapted to the filtration* $\mathscr{F}_t^{x_0}$ *generated by* x_0 *and* $\{W(s)(\cdot),s\le t\}$ *and*

$$\sup_{0\le t\le T}\mathbb{E}[|x(t)|^2]<\infty. \tag{20.12}$$

Proof ***Existence***: First, we establish the existence of solution of the initial value problem. Let us define $y(t)=x(t-\delta)$, $x^{(0)}(t)=\xi(0)$ and $x^{(k)}(t)=x^{(k)}(t,\omega)$ inductively as follows:

$$\begin{aligned}x^{(k+1)}(t)=\xi(0)&+\frac{1}{\Gamma(\alpha)}\int_0^t(t-s)^{\alpha-1}b(s,x^{(k)}(s),y^{(k)}(s))\,ds\\&+\frac{1}{\Gamma(\alpha)}\int_0^t(t-s)^{\alpha-1}\sigma(s,x^{(k)}(s),y^{(k)}(s))\,dW(s)\end{aligned} \tag{20.13}$$

for $k=0,1,2,\ldots$. If, for fixed $k\ge 0$, the approximation $x^{(k)}(t)$ is $\mathscr{F}_t$-measurable and continuous on J, then it follows from (20.8)–(20.11), that the integrals in (20.13) are meaningful and that the resulting process $x^{(k+1)}(t)$ is $\mathscr{F}_t$-measurable and continuous on J. As $x^{(0)}(t)$ is obviously $\mathscr{F}_t$-measurable and continuous on J, it follows by induction that so too is each $x^{(k)}(t)$ for $k=1,2,\ldots$.

Since $\xi(0)$ is $\mathscr{F}_t$-measurable with $\mathbb{E}(|\xi(0)|^2)<\infty$, it is clear that

$$\sup_{0\le t\le T}\mathbb{E}(|x^{(0)}(t)|^2)<\infty.$$

Applying the algebraic inequality $(a+b+c)^2\le 3(a^2+b^2+c^2)$, the Cauchy–Schwarz inequality, the Itô isometry, and the linear growth condition (20.8) we obtain from (20.13) that

$$\begin{aligned}\mathbb{E}(|x^{(k+1)}(t)|^2)\le 3\mathbb{E}[|\xi(0)|^2]&+\frac{3}{(\Gamma(\alpha))^2}\frac{T^{2\alpha-1}}{2\alpha-1}\mathbb{E}\left[\int_0^t\left|b(s,x^{(k)}(s),y^{(k)}(s)\right|^2ds\right]\\&+\frac{3}{(\Gamma(\alpha))^2}\frac{T^{2\alpha-1}}{2\alpha-1}\mathbb{E}\left[\int_0^t\left|\sigma(s,x^{(k)}(s),y^{(k)}(s))\right|^2ds\right].\end{aligned}$$

Therefore

$$\mathbb{E}(|x^{(k+1)}(t)|^2) \leq 3\mathbb{E}[|\xi(0)|^2] + 2K_1^2(1+K_2^2)\frac{3}{(\Gamma(\alpha))^2}\frac{T^{2\alpha-1}}{2\alpha-1}
\mathbb{E}\left(\int_0^t \left(1+|x^{(k)}(s)|^2\right)\, ds\right),$$

for $k = 0,\ 1,\ 2, \ldots$. By induction, we have

$$\sup_{0\leq t\leq T} \mathbb{E}(|x^{(k)}(t)|^2) \leq C_0 < \infty,$$

for $k = 1,\ 2,\ 3, \ldots$ and C_0 is a positive constant. Let

$$d^{(k)}(t) = \mathbb{E}(|x^{(k+1)}(t) - x^{(k)}(t)|).$$

We claim that

$$d^{(k)}(t) \leq \frac{(Mt)^{(k+1)}}{(k+1)!}, \quad \text{for all} \quad k = 0, 1, 2, \ldots, \tag{20.14}$$

for some constants M, depending in K_1, K_2, L_1, L_2, and $\xi(0)$. From Eq. (20.13) by applying the Schwarz inequality, Itô isometry, and the Lipschitz condition (20.10) and (20.11) we obtain

$$\begin{aligned}
d^{(k)}(t) &= \mathbb{E}[|x^{(k+1)}(t) - x^{(k)}(t)|^2] \\
&\leq \frac{2}{(\Gamma(\alpha))^2}\frac{T^{2\alpha-1}}{2\alpha-1}\int_0^t \mathbb{E}\left[\left|b(s,x^{(k)}(s),y^{(k)}(s)) - b(s,x^{(k-1)}(s),y^{(k-1)}(s))\right|^2\right] ds \\
&\quad + \frac{2}{(\Gamma(\alpha))^2}\frac{T^{2\alpha-1}}{2\alpha-1}\int_0^t \mathbb{E}\left[\left|\sigma(s,x^{(k)}(s),y^{(k)}(s)) - \sigma(s,x^{(k-1)}(s),y^{(k-1)}(s))\right|^2\right] ds \\
&\leq 2\frac{L_1^2}{(\Gamma(\alpha))^2}(1+L_2^2)\frac{T^{2\alpha-1}}{2\alpha-1}\int_0^t \mathbb{E}\left[\left|x^{(k)}(s) - x^{(k-1)}(s)\right|^2\right] ds \\
&\quad + 2\frac{L_1^2}{(\Gamma(\alpha))^2}(1+L_2^2)\frac{T^{2\alpha-1}}{2\alpha-1}\int_0^t \mathbb{E}\left[\left|x^{(k)}(s) - x^{(k-1)}(s)\right|^2\right] ds.
\end{aligned} \tag{20.15}$$

By applying again the Schwarz inequality, the Itô isometry together with the growth conditions (20.8) and (20.9) for $k = 0$,

$$\begin{aligned}
d^{(0)}(t) &= \mathbb{E}[|x^{(1)}(t) - x^{(0)}(t)|^2] \\
&\leq \frac{2}{(\Gamma(\alpha))^2}\mathbb{E}\left(\left|\int_0^t (t-s)^{\alpha-1} b(s,x^{(0)}(s),y^{(0)}(s))\, ds\right|^2\right) \\
&\qquad + \frac{2}{(\Gamma(\alpha))^2}\mathbb{E}\left(\left|\int_0^t (t-s)^{\alpha-1}\sigma(s,x^{(0)}(s),y^{(0)}(s))\, dW(s)\right|^2\right) \\
&\leq \frac{2}{(\Gamma(\alpha))^2}\frac{T^{2\alpha-1}}{2\alpha-1}\int_0^t \mathbb{E}\left[\left|b(s,x^{(0)},y^{(0)}(s)(s))\right|^2\right] ds
\end{aligned}$$

$$
\begin{aligned}
&+\frac{2}{(\Gamma(\alpha))^2}\frac{T^{2\alpha-1}}{2\alpha-1}\int_0^t \mathbb{E}\left[\left|\sigma(s,x^{(0)}(s),y^{(0)}(s))\right|^2\right]ds\\
&\leq K_1^2(1+K_2^2)\frac{4}{(\Gamma(\alpha))^2}\frac{T^{2\alpha-1}}{2\alpha-1}\mathbb{E}\left(\int_0^t(1+|x_0|^2)ds\right)\\
&\leq K_1^2(1+K_2^2)\frac{4}{(\Gamma(\alpha))^2}\frac{T^{2\alpha-1}}{2\alpha-1}(t)(1+\mathbb{E}(|x_0|^2)). \qquad (20.16)
\end{aligned}
$$

Now, for $k=1$, replacing $\mathbb{E}[|x^{(1)}(t)-x^{(0)}(t)|^2]$ in the inequality (20.15) with the value on the right-hand side of inequality (20.16) and integrating, we obtain

$$
\begin{aligned}
\mathbb{E}[|x^{(2)}(t)-x^{(1)}(t)|^2] &\leq L_1^2(1+L_2^2)\frac{4}{(\Gamma(\alpha))^2}\frac{T^{2\alpha-1}}{2\alpha-1}\int_0^t\mathbb{E}[|x^{(1)}(s)-x^{(0)}(s)|^2]\,ds\\
&\leq K_1^2(1+K_2^2)(1+\mathbb{E}(|x_0|^2))\left(L^2\frac{4}{(\Gamma(\alpha))^2}\frac{T^{2\alpha-1}}{2\alpha-1}\right)^2\int_0^t sds\\
&\leq K^2(1+\mathbb{E}(|x_0|^2))\left(L^2\frac{4}{(\Gamma(\alpha))^2}\frac{T^{2\alpha-1}}{2\alpha-1}\right)^2\times\frac{t^2}{2!}, \qquad (20.17)
\end{aligned}
$$

where $L^2=L_1^2(1+L_2^2)$ and $K^2=K_1^2(1+K_2^2)$. For $k=2$, proceeding as before, we have

$$
\mathbb{E}[|x^{(3)}(t)-x^{(2)}(t)|^2]\leq K^2(1+\mathbb{E}(|x_0|^2))\left(L^2\frac{4}{(\Gamma(\alpha))^2}\frac{T^{2\alpha-1}}{2\alpha-1}\right)^3\times\frac{t^3}{3!}. \qquad (20.18)
$$

Thus, by the principle of mathematical induction, we have

$$
d^{(k)}(t)=\mathbb{E}[|x^{(k+1)}(t)-x^{(k)}(t)|^2]\leq\frac{BM^{k+1}t^{(k+1)}}{(k+1)!},\ k=0,1,2,\ldots,\ 0\leq t\leq T, \qquad (20.19)
$$

where $B=K^2(1+\mathbb{E}|x_0|^2)$ and $M=\left(L^2\frac{4}{(\Gamma(\alpha))^2}\frac{T^{2\alpha-1}}{2\alpha-1}\right)$ is a constant depending only on α, T, L^2, and $\mathbb{E}|x_0|^2$. Note that

$$
\begin{aligned}
&\max_{0\leq t\leq T}|x^{(k+1)}(t)-x^{(k)}(t)|^2\\
&\leq 2\max_{0\leq t\leq T}\int_0^t(t-s)^{\alpha-1}|b(s,x^{(k)}(s),y^{(k)}(s))-b(s,x^{(k-1)}(s),y^{(k-1)}(s))|^2ds\\
&+2\max_{0\leq t\leq T}\int_0^t(t-s)^{\alpha-1}|\sigma(s,x^{(k)}(s),y^{(k)}(s))-\sigma(s,x^{(k-1)}(s),y^{(k-1)}(s))|^2dW(s).
\end{aligned}
$$

Taking expectation on both sides, we have

$$
\mathbb{E}\left(\max_{0\leq t\leq T}|x^{(k+1)}(t)-x^{(k)}(t)|^2\right)\leq 2L^2\frac{T^{2\alpha-1}}{2\alpha-1}\mathbb{E}\left(\max_{0\leq t\leq T}\int_0^t|x^{(k)}(s)-x^{(k-1)}(s)|^2ds\right)
$$

$$+\, 2\mathbb{E}\left(\max_{0\le t\le T}\int_0^t (t-s)^{\alpha-1}|\sigma(s,x^{(k)}(s),y^{(k)}(s))-\sigma(s,x^{(k-1)}(s),y^{(k-1)}(s))|^2 dW(s)\right)$$

Using second part of the Theorem 20.1 gives

$$\begin{aligned}\mathbb{E}\left(\max_{0\le t\le T}|x^{(k+1)}(t)-x^{(k)}(t)|^2\right) &\le 2L^2\frac{T^{2\alpha-1}}{2\alpha-1}\mathbb{E}\left(\int_0^T |x^{(k)}(s)-x^{(k-1)}(s)|^2 ds\right)\\ &+ 8L^2\frac{T^{2\alpha-1}}{2\alpha-1}\mathbb{E}\left(\int_0^T |x^{(k)}(s)-x^{(k-1)}(s)|^2 ds\right)\\ &\le B\frac{M^{k+1}}{(k+1)!}T^{(k+1)}, \qquad (20.20)\end{aligned}$$

where B is a constant depending on L and T. By using Chebyshev's inequality gives

$$\mathbb{P}\left(\max_{0\le t\le T}|x^{(k+1)}(t)-x^{(k)}(t)|^2 > \frac{1}{k^2}\right) \le \frac{1}{(1/k^2)^2}\mathbb{E}\left(\max_{0\le t\le T}|x^{(k+1)}(t)-x^{(k)}(t)|^2\right).$$

Using the Eq. (20.20) and summing up the resultant inequalities gives,

$$\sum_{k=0}^{\infty}\mathbb{P}\left(\max_{0\le t\le T}|x^{(k+1)}(t)-x^{(k)}(t)|^2 > \frac{1}{k^2}\right) \le \sum_{k=0}^{\infty}\frac{BM^{k+1}k^4T^{(k+1)}}{(k+1)!}.$$

where the series on the right side converges by ratio test. Hence the series on the left side also converges, so by the Borel–Cantelli lemma, we conclude that $\max_{0\le t\le T}\left(|x^{(k+1)}(t)-x^{(k)}(t)|^2\right)$ converges to 0, almost surely, that is, the successive approximations $x^{(k)}(t)$ converge, almost surely, uniformly on J to a limit $x(t)$ defined by

$$\lim_{n\to\infty}\left(x^{(0)}(t)+\sum_{k=1}^{n}(x^{(k)}(t)-x^{(k-1)}(t))\right) = \lim_{n\to\infty}x^{(n)}(t) = x(t). \qquad (20.21)$$

From (20.13), we have

$$\begin{aligned}x(t) = \xi(0) &+ \frac{1}{\Gamma(\alpha)}\int_0^t (t-s)^{\alpha-1}b(s,x(s),y(s))\,ds\\ &+ \frac{1}{\Gamma(\alpha)}\int_0^t (t-s)^{\alpha-1}\sigma(s,x(s),y(s))\,dW(s). \qquad (20.22)\end{aligned}$$

for all $t\in J$. This completes the proof of the existence of solution of (20.6).
Uniqueness: The uniqueness follows from the Itô isometry, the Lipschitz conditions (20.10).
Let $x(t,\omega)$ and $y(t,\omega)$ be solution processes through the initial data $(0,\xi(0))$ and $(0,\nu(0))$, respectively, that is, $x(0,\omega)=\xi(0)(\omega)$ and $y(0,\omega)=\nu(0)(\omega)$, $\omega\in\Omega$.

Let

$$a(s,\omega)=b\left(s,x_1(s),y_1(s)\right)-b\left(s,x_2(s),y_2(s)\right),$$
$$\gamma(s,\omega)=\sigma\left(s,x_1(s),y_1(s)\right)-\sigma\left(s,x_2(s),y_2(s)\right).$$

Then by virtue of the Schwarz inequality and the Itô isometry, we have

$$\begin{aligned}\mathbb{E}[|x(t)-y(t)|^2]\le{}& \frac{3}{(\Gamma(\alpha))^2}\mathbb{E}[|\xi(0)-\nu(0)|^2]+\frac{3}{(\Gamma(\alpha))^2}\frac{t^{2\alpha-1}}{2\alpha-1}\mathbb{E}\left[\int_0^t |a(s,\omega)|^2ds\right]\\ &+\frac{3}{(\Gamma(\alpha))^2}\frac{t^{2\alpha-1}}{2\alpha-1}\mathbb{E}\left[\int_0^t |\gamma(s,\omega)|^2ds\right]\\ \le{}& \frac{3}{(\Gamma(\alpha))^2}\mathbb{E}[|\xi(0)-\nu(0)|^2]+2L^2\frac{3}{(\Gamma(\alpha))^2}\frac{t^{2\alpha-1}}{2\alpha-1}\int_0^t \mathbb{E}[|x(s)-y(s)|^2]\,ds.\end{aligned}$$

We define $v(t)=\mathbb{E}[|x(t)-y(t)|^2]$. Then the function v satisfies $v(t)\le F+A\int_0^t v(s)ds$, where $F=\frac{3}{(\Gamma(\alpha))^2}\mathbb{E}[|\xi(0)-\nu(0)|^2]$ and $A=2L^2\frac{3}{(\Gamma(\alpha))^2}\frac{t^{2\alpha-1}}{2\alpha-1}$. By the application of the Gronwall inequality, we conclude that

$$v(t)\le F\exp(At).$$

Now assume that $\xi(0)=\nu(0)$. Then $F=0$ and so $v(t)=0$ for all $t\ge 0$. That is,

$$\mathbb{E}[|x(t)-y(t)|^2]=0.$$

Which gives

$$\int_0^t |x(t)-y(t)|^2d\mathbb{P}=0.$$

This implies that $x(t)=y(t)$ a.s for all $t\in J$. That is

$$P\Big\{|x(t,\omega)-y(t,\omega)|=0 \qquad \text{for all } t\in J\Big\}=1,$$

that is, the solution is unique. This completes the proof of existence and uniqueness of solution of the given stochastic fractional differential equation (20.6).

20.4 Stability Analysis

In this section, we study the exponentially asymptotic stability in the quadratic mean of a trivial solution. Consider the following stochastic fractional nonlinear system of the form

$$\left.\begin{aligned} {}^{C}D^{\alpha}x(t) &= Ax(t) + f(t, x(t), x(t-\delta)) + \sigma(t, x(t), x(t-\delta))\frac{dW(t)}{dt}, \\ x(t) &= \xi(t),\ \ t \in [-\delta, 0] \end{aligned}\right\} \tag{20.23}$$

where $\alpha \in (1/2, 1)$, $f \in C(J \times \mathbb{R}^n \times \mathbb{R}^n, \mathbb{R}^n)$, $\sigma \in C(J \times \mathbb{R}^n \times \mathbb{R}^n, \mathbb{R}^{n\times m})$ and $W = \{W(t), t \geq 0\}$ is an m-dimensional Brownian motion on a complete probability space $\Omega \equiv (\Omega, \mathscr{F}, P)$, $A \in \mathbb{R}^{n\times n}$ is a diagonal stability matrix. Assume from now on that $f(t, 0, 0) = \sigma(t, 0, 0) \equiv 0$ a.e t so that Eq. (20.23) admits a trivial solution.

Definition 20.7 The trivial solution of Eq. (20.23) is said to be exponentially stable in the quadratic mean if there exist positive constants C, ν such that

$$\mathbb{E}(|x(t)|^2) \leq C\mathbb{E}(|\xi(0)|^2)\exp(-\nu t), \quad t \geq 0.$$

The following lemmas are necessary to obtain the main results. For that, we assume the following hypothesis:

(H1) There exists a constant $M > 0$ such that for $t \geq 0$,

$$|E_{\alpha,\beta}(At^{\alpha})| \leq Me^{-at},$$

where $0 < \alpha < 1$ and $\beta = 1, 2$ and α.

Lemma 20.2 *Assume that the hypothesis (H1) holds. Then for any stochastic process* $F : [0, \infty) \to \mathbb{R}^n$ *which is strongly measurable with* $\int_0^T \mathbb{E}|F(t)|^2 ds < \infty$, $0 < T \leq \infty$, *the following inequality holds for* $0 < t \leq T$,

$$\mathbb{E}\left|\int_0^t E_{\alpha,\beta}(A(t-s)^{\alpha})F(s)\, ds\right|^2 \leq (M^2/a)\int_0^t \exp(-a(t-s))\mathbb{E}|F(s)|^2 ds,$$

where $\alpha \in (1/2, 1)$ *and* $\beta = 1, 2,$ *and* α.

Proof Assume that the hypothesis (H1) holds; that is there exists a constants $a > 0$ and $M > 0$ such that for $t \geq 0$

$$|E_{\alpha,\beta}(At^{\alpha})| \leq Me^{-at},$$

where $0 < \alpha < 1$ and $\beta = 1, 2,$ and α. By the Hölder inequality, we obtain, for $0 < t \leq T$,

$$\begin{aligned} &\mathbb{E}\left|\int_0^t E_{\alpha,\beta}(A(t-s)^{\alpha})F(s)\,\mathrm{d}s\right|^2 \\ &\qquad \leq \mathbb{E}\left(\int_0^t M\exp(-(a/2)(t-s))\exp(-(a/2)(t-s))|F(s)|\,\mathrm{d}s\right)^2 \end{aligned}$$

$$\leq \mathbb{E}\left(\int_0^t M\exp(-(a/2)(t-s))\mathrm{d}s\right)^2 \mathbb{E}\left(\int_0^t \exp(-(a/2)(t-s))|F(s)|\,\mathrm{d}s\right)^2$$
$$\leq (M^2/a)\int_0^t \exp(-a(t-s))\mathbb{E}(|F(s)|^2)\,\mathrm{d}s,$$

which complete the proof of the lemma.

Lemma 20.3 *Assume that the hypothesis (H3) holds. Then for any B_t-adapted predictable process $\Phi : [0,\infty) \to \mathbb{R}^n$ with $\int_0^T \mathbb{E}|\Phi(s)|^2 ds < \infty,\ t \geq 0$, the following inequality holds for $0 < t \leq T$,*

$$\mathbb{E}\left|\int_0^t E_{\alpha,\beta}(A(t-s)^\alpha)\Phi(s)\,dW(s)\right|^2 \leq M^2\int_0^t \exp(-a(t-s))\mathbb{E}|\Phi(s)|^2 ds,$$

where $\alpha \in (1/2, 1)$ and $\beta = 1, 2$ and α.

The proof is similar to the previous Lemma.

Theorem 20.3 *Let the assumptions of Theorem 20.2 holds. Then the solution of equation (20.23) is exponentially stable in the quadratic mean provided*

$$a > \beta = \beta(a, K, M) = \frac{3M^2(K_1^2/a + k_1^2)(1 + K_2^2)T^{2\alpha-1}}{2\alpha - 1}.$$

Proof The integral form of the Eq. (20.23) can be given by [2, 15]

$$x(t) = E_\alpha(At^\alpha)\xi(0) + \int_0^t (t-s)^{\alpha-1}E_{\alpha,\alpha}(A(t-s)^\alpha)f(s, x(s), x(s-\delta))\,ds$$
$$+ \int_0^t (t-s)^{\alpha-1}E_{\alpha,\alpha}(A(t-s)^\alpha)\sigma(s, x(s), x(s-\delta))\,dW(s). \tag{20.24}$$

By using Hölder inequality and Lemmas 20.2 and 20.3, we get

$$\mathbb{E}|x(t)|^2 \leq 3M^2\exp(-at)\mathbb{E}|\xi(0)|^2$$
$$+ 3(M^2/a)\frac{T^{2\alpha-1}}{2\alpha-1}\int_0^t \exp(-a(t-s))\mathbb{E}|f(s, x(s), x(s-\delta))|^2 ds$$
$$+ 3M^2\frac{T^{2\alpha-1}}{2\alpha-1}\int_0^t \exp(-a(t-s))\mathbb{E}|\sigma(s, x(s), x(s-\delta))|^2 ds.$$

Linear growth assumption (20.8) when $f(t,0,0) = \sigma(t,0,0) \equiv 0$ a.e t yields

$$\begin{aligned}\exp(at)\mathbb{E}|x(t)|^2 \leq 3M^2\mathbb{E}|\xi(0)|^2 &+ 3(M^2/a)K_1^2(1+K_2^2)\frac{T^{2\alpha-1}}{2\alpha-1}\int_0^t \exp(as)\mathbb{E}|x(s))|^2 ds \\ &+3M^2K_1^2(1+K_2^2)\frac{T^{2\alpha-1}}{2\alpha-1}\int_0^t \exp(as)\mathbb{E}|x(s))|^2 ds. \\ \leq 3M^2\mathbb{E}|\xi(0)|^2 + 3M^2(K_1^2/a+K_1^2)(1+K_2^2)&\frac{T^{2\alpha-1}}{2\alpha-1}\int_0^t \exp(as)\mathbb{E}|x(s))|^2 ds\end{aligned}$$

Applying Gronwall's inequality, we obtain

$$\exp(at)\mathbb{E}|x(t)|^2 \leq 3M^2\mathbb{E}|\xi(0)|^2 \exp\left(3M^2(K_1^2/a+K_1^2)(1+K_2^2)\frac{T^{2\alpha-1}}{2\alpha-1}t\right)$$

Consequently,

$$\mathbb{E}|x(t)|^2 \leq C\mathbb{E}|\xi(0)|^2 \exp(-\nu t),\ \ t \geq 0, \tag{20.25}$$

where $\nu = a - \beta\ \ and\ \ C = 3M^2$.

20.5 Examples

Example 20.1 Consider the following stochastic fractional delay differential equation of the form:

$$\left.\begin{aligned} {}^C D^{0.6}x(t) + 0.2x(t) &= -\frac{t^2 y(t)}{\Gamma(3-\alpha)} + t^2\frac{dW(t)}{dt},\ \ t \in J \\ x(t) &= 0.t \in [-1,0] \end{aligned}\right\} \tag{20.26}$$

Here $f(t, x(t), y(t)) = -0.2x(t) - \dfrac{t^2 y(t)}{\Gamma(3-\alpha)}$, $\sigma(t, x(t), y(t)) = t^2$. It can be easily seen that $f(t, x(t), y(t))$ and $\sigma(t, x(t), y(t))$ satisfies the condition of (20.8), (20.9), (20.10), and (20.11) of Theorem 20.2 for $\alpha = 0.6$. Hence by the Theorem 20.2 the stochastic fractional delay differential equation (20.26) has a unique solution. Also, Eq. (20.26) satisfy the condition of Theorem 20.3. So from Theorem 20.3 the stochastic fractional differential equation with $A = -0.2$ is exponentially stable.

Example 20.2 Consider the nonlinear stochastic fractional system, for $t \in [0, T]$,

$$\begin{aligned} {}^C D^{\frac{3}{4}}x(t) &= \begin{pmatrix} 1 & -1 \\ 0 & 2 \end{pmatrix} x(t) + t^3 y(t) + \begin{pmatrix} \frac{1}{1+t} \\ e^{\frac{sin(x_2)}{10(1+t)}} \end{pmatrix} \frac{\mathrm{d}W(t)}{\mathrm{d}t}, \\ x(0) &= x_0, \end{aligned} \tag{20.27}$$

where $x(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix}$ and $\alpha = \frac{3}{4}$. It can be easily seen that $f(t, x(t), y(t))$ and $\sigma(t, x(t), y(t))$ satisfies the condition of (20.8), (20.9), (20.10), and (20.11) of Theorem 20.2. Hence by the Theorem 20.2 the stochastic fractional delay differential equation (20.27) has a unique solution. Also Eq. (20.27) satisfies the condition of Theorem 20.3 which gives the exponential stable.

Conclusion

The existence and uniqueness of solution of the nonlinear stochastic fractional delay differential equation with Gaussian noise is obtained. To study the nonlinear system, an equivalent nonlinear integral equation to the nonlinear stochastic fractional delay differential equation is given. Using the integral equation, the sufficient conditions for ensuring the stability of the stochastic fractional differential equations with Gaussian noise is established. The Picard–Lindelöf method of successive approximation technique is used to obtain the results. Examples are provided to illustrate the theory developed.

Acknowledgements The work of the first author is supported by the DST PURSE program of Government of India.

References

1. K. Balachandran, K. Sumathy, J.K. Kim, Existence and stability of solutions of general stochastic integral equations. Nonlinear Funct. Anal. Appl. **12**, 219–235 (2007)
2. K. Balachandran, M. Matar, J.J. Trujillo, Note on controllability of linear fractional dynamical systems. J. Control Decis. **3**, 267–279 (2016)
3. S. Bhalekar, Stability analysis of a class of fractional delay differential equations. Pramana: J. Phys. **81**, 215–224 (2013)
4. S. Bhalekar, Stability and bifurcation analysis of a generalized scalar delay differential equation. Chaos **26**, 084306 (2016)
5. K. Diethelm, *The Analysis of Fractional Differential Equations* (Springer, New York, 2010)
6. K. Diethelm, K. Ford, Analysis of fractional differential equations. J. Math. Anal. Appl. **265**, 229–248 (2002)
7. L.C. Evans, *An Introduction to Stochastic Differential Equations* (American Mathematical Society, Providence, 2014)
8. T.E. Govindan, Existence and stability of solutions of stochastic semilinear functional differential equations. Stoch. Anal. Appl. **20**, 1257–1280 (2002)
9. T.E. Govindan, Stability of mild solutions of stochastic evolution equations with variable delay. Stoch. Anal. Appl. **21**, 1059–1077 (2003)
10. T.E. Govindan, Almost sure exponential stability for stochastic neutral partial functional differential equations. Stochastics **77**, 139–154 (2005)
11. R.W. Ibrahim, Stability of fractional differential equations. Int. J. Math. Comput. Sci. **7**, 11–16 (2013)
12. R. Khasminskii, *Stochastic Stability of Differential Equations* (Springer, London, 2012)

13. A.A. Kilbas, H.M. Srivastava, J.J. Trujillo, *Theory and Applications of Fractional Differential Equations* (Elsevier, Amsterdam, 2006)
14. P.E. Kloeden, E. Platen, *Numerical Solution of Stochastic Differential Equations* (Springer, New York, 1992)
15. R.M. Lizzy, K. Balachandran, J.J. Trujillo, Controllability of nonlinear stochastic fractional neutral systems with multiple time varying delay in control. Chaos Solitons Fractals **102**, 162–167 (2017)
16. J. Luo, Exponential stability for stochastic neutral partial functional differential equations. J. Math. Anal. Appl. **355**, 414–425 (2009)
17. X. Mao, Numerical solutions of stochastic functional differential equations. LMS J. Comput. Math. **6**, 141–161 (2003)
18. K.S. Miller, B. Ross, *An Introduction to the Fractional Calculus and Fractional Equations* (Wiley, New York, 1993)
19. B. Øksendal, *Stochastic Differential Equations*. An Introduction with Applications (Springer, Heidelberg, 2003)
20. J.C. Pedjeu, G.S. Ladde, Stochastic fractional differential equations: modeling, method and analysis. Chaos Solitons Fractals **45**, 279–293 (2012)
21. J.C. Pedjeu, S. Sathananthan, Fundamental properties of stochastic integrodifferential equations-I, existence and uniqueness results. Int. J. Pure. Appl. Math. **7**, 337–355 (2003)
22. D. Qian, C. Li, R.P. Agarwal, P.J.Y. Wong, Stability analysis of fractional differential system with Riemann-Liouville derivative. Math. Comput. Model. **52**, 862–874 (2010)
23. T. Taniguchi, Successive approximations to solutions of stochastic differential equations. J. Differ. Equ. **96**, 152–169 (1992)
24. T. Taniguchi, Asymptotic stability theorems of semilinear stochastic evolution equations in Hilbert spaces. Stoch. Stoch. Rep. **53**, 41–52 (1995)
25. P. Umamaheswari, K. Balachandran, N. Annapoorani, On the solution of stochastic fractional integrodifferential equations. Nonlinear Funct. Anal. Appl. **22**, 335–354 (2017)
26. P. Umamaheswari, K. Balachandran, N. Annapoorani, Existence of solution of stochastic fractional integrodifferential equations. Discontin. Nonlinearity Complex. **7**, 55–65 (2018)
27. T. Yamada, On the successive approximation of solutions of stochastic differential equations. Kyoto J. Math. **21**, 501–515 (1981)
28. W. Zhu, J. Huang, Z. Zhao, Exponential stability of stochastic systems with delay and Poisson jumps. Math. Probl. Eng. 903821 (2014), 10 pp

Chapter 21
Asymptotic Stability of Implicit Fractional Volterra Integrodifferential Equations

Kausika Chellamuthu

Abstract In this paper, we discuss the stability of fractional Volterra integrodifferential equations using a method based on eigenvalue criterion. Lyapunov's definition of stability is used and of the two methods, Lyapunov's first or indirect method is used to prove the stability results. Some sufficient conditions ensuring asymptotic stability of the system involving implicit fractional derivative are established. Examples are provided to demonstrate the effectiveness of the method.

Keywords Caputo derivative · Fractional differential equations · Mittag-Leffler function · Asymptotic stability

21.1 Introduction

Fractional calculus has attracted the attention of a large number of mathematicians, physicists and engineers in the recent years. The past three decades have seen a considerable number of interesting and novel applications of fractional differential equations in physics, biology, chemistry, engineering, finance, other recently developed sciences and even psychology, where fractional differential equations capture human behaviour more rationally.

Recently, the stability of fractional differential equations is gaining attention due to its importance in control theory, since every controllable system is designed to be stable. In 1996, Matignon [20] firstly gave the stability result on linear autonomous fractional differential systems from a control-theoretic point of view. This result formed the basis for further research concerning stability issues [12, 20, 22]. There are various methods to solve stability problems. Some of the early works on integrodifferential equations were done using Lyapunov theory [19], but the construction of the Liapunov functional is a demanding art. Many methods thereafter came to bypass this difficulty. The study by Burton [7] showed the effectiveness of the fixed point

K. Chellamuthu (✉)
Department of Mathematics, Bharathiar University, Coimbatore 641046, India
e-mail: kausika.rs@buc.edu.in

P. Manchanda et al. (eds.), *Mathematical Modelling, Optimization, Analytic and Numerical Solutions*, Industrial and Applied Mathematics,
https://doi.org/10.1007/978-981-15-0928-5_21

techniques in dealing with the stability problems shedding light on systems with nonunique solutions. Qian et al. [26] proposed an eigenvalue criterion to check the stability of fractional differential equations. This point of view helps to study the stability of fractional differential equation in parallel with the ordinary differential equation as one may observe the importance of Mittag Leffler functions as eigenfunctions of fractional differential equations analogous to how exponential functions are to ordinary differential equations as far as autonomous systems are concerned. Priyadharsni [25] studied the stability of fractional neutral and integrodifferential equations in a similar way.

Also, there have been various notions of stability given by different authors. The Ulam Hyers and Ulam Hyers Rassias stability of nonlinear Volterra integrodifferential equations has been discussed [13, 27, 28], in both integer- and fractional-order cases. A note on Mittag-Leffler stability can be seen in [18] and references therein. Balachandran et al. [1–5, 14] studied existence and other qualitative behaviours of nonlinear integrodifferential equations in both integer and fractional orders. Benchohra and Lazreg [6] studied Ulam Hyers stability of nonlinear fractional differential equations with an implicit derivative. Coming to the nonlinear fractional implicit differential equations, fewer works are reported [16, 21, 32] and to the best of our knowledge no work has been reported using Lyapunov's indirect method of stability investigation. In Lyapunov's first or indirect method, the nonlinear equation is linearized and then the stability results are transferred from the linear to nonlinear equation using appropriate growth conditions on the nonlinear terms. This method is adopted to investigate the stability results in our work. In this paper, we study the following fractional Volterra integrodifferential equation with an implicit derivative, where the fractional derivative is taken in the sense of Caputo.

$$\begin{aligned} {}^C D^\alpha x(t) &= Ax(t) + \int_0^t K(t,s)G(x(s), {}^C D^\beta x(s))\,\mathrm{d}s, \quad 0<\alpha,\beta<1, \;\; t\in J, \\ x(0) &= x_0, \end{aligned} \tag{21.1}$$

where $x \in \mathbb{R}^n$, $J := [0, T]$, $A \in \mathbb{R}^{n\times n}$, $K : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}$ and $G : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$.

21.2 Preliminaries

Definition 21.1 The operator I^α defined on $L_1[0, T]$ by

$$I^\alpha f(t) := \frac{1}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} f(s)\,\mathrm{d}s, \quad 0<\alpha<1,$$

is called the Riemann–Liouville fractional integral operator of order α.

Definition 21.2 The Caputo fractional differential operator ${}^C D^\alpha$ is defined by

$$ {}^{C}D^{\alpha}f(t) := \frac{1}{\Gamma(1-\alpha)}\int_0^t (t-s)^{-\alpha}f'(s)\,\mathrm{d}s \quad 0<\alpha<1, $$

whenever $f' \in L_1[0,T]$.

Definition 21.3 Let $\alpha \in \mathbb{C}$. The function E_α defined by

$$ E_\alpha(z) := \sum_{j=0}^{\infty} \frac{z^j}{\Gamma(\alpha j+1)} $$

whenever the series converges is called the Mittag-Leffler function of order α.

Definition 21.4 Let $\alpha, \beta \in \mathbb{C}$. The function $E_{\alpha,\beta}$ defined by

$$ E_{\alpha,\beta}(z) := \sum_{j=0}^{\infty} \frac{z^j}{\Gamma(\alpha j+\beta)} $$

whenever the series converges is called the two-parameter Mittag-Leffler function with parameters α, β.

Laplace transforms of Mittag-Leffler functions which are useful in next section are given below [23]:

- $\mathscr{L}\{t^{\beta-1}E_{\alpha,\beta}(\pm\lambda t^\alpha)\}(s) = \dfrac{s^{\alpha-\beta}}{s^\alpha \mp \lambda}$,
- $\mathscr{L}\{t^{\alpha+\beta-1}E_{\alpha,\alpha+\beta}(\pm\lambda t^\alpha)\}(s) = \dfrac{s^{-\beta}}{s^\alpha \mp \lambda}$.

Derivatives of some Mittag-Leffler-type functions are given below:

- $\dfrac{\mathrm{d}}{\mathrm{d}t}(E_\alpha(At^\alpha)) = At^{\alpha-1}E_{\alpha,\alpha}(At^\alpha)$,
- $\dfrac{\mathrm{d}}{\mathrm{d}t}\Big[(t-s)^\alpha E_{\alpha,\alpha+1}(A(t-s)^\alpha)\Big] = (t-s)^{\alpha-1}E_{\alpha,\alpha}[A(t-s)^\alpha]$,

where $\mathscr{Re}(\alpha) > 0$, $\mathscr{Re}(\beta) > 0$, $t \geq 0$ and $\lambda \in \mathbb{R}$.

Theorem 21.1 ([11]) *Assume that* $f : [0,\infty) \to \mathbb{R}$ *is such that* $\mathscr{L}\{f\}$ *exists on* $[s_0,\infty)$ *with some* $s_0 \in \mathbb{R}$. *Let* $\alpha > 0$ *and* $m := [\alpha]$. *Then, for* $s > max\{0, s_0\}$, *we have*

$$ \mathscr{L}\{I^\alpha f(s)\} = \frac{1}{s^\alpha}\mathscr{L}\{f(s)\} $$

and

$$ \mathscr{L}\{{}^{C}D^\alpha f(s)\} = s^\alpha \mathscr{L}\{f(s)\} - \sum_{k=1}^{m} s^{\alpha-k}f^{(k-1)}(0). $$

Now we state the stability concepts.

Definition 21.5 ([11])

(a) The zero solution of the differential equation (21.1) is called "stable" if, for any $\epsilon > 0$ there exists some $\delta > 0$ such that the solution of the initial value problem consisting of the differential equation (21.1) and the initial condition $x(0) = x_0$ satisfies $\|x(t)\| < \epsilon$ for all $t \geq 0$ whenever $\|x_0\| < \delta$.

(b) The zero solution of the differential equation (21.1) is called "asymptotically stable" if it is stable and there exists some $\gamma > 0$ such that $\lim_{t\to 0} \|x(t)\| = 0$ whenever $\|x_0\| < \gamma$.

Well-posedness of our initial value problem asserts that under the usual continuity and Lipschitz assumptions on f, the solution of a fractional differential equation does not change much over some finite interval if we perturb the initial values by a small magnitude. The notion of stability is the extension of this idea to unbounded intervals. The trivial solution is stable if a small change in initial value leads to a small change of the solution over the complete positive half-plane. Asymptotic stability is even stronger as it requires the solution of the perturbed problem not only to remain close to the original solution but also converge to the latter.

It is well known that the homogeneous linear fractional differential equation with constant coefficients

$$\begin{aligned} {}^{C}D^{\alpha}x(t) &= Ax(t), \quad 0 < \alpha < 1, \\ x(0) &= x_0, \end{aligned} \tag{21.2}$$

where A is an arbitrary $n \times n$ matrix has the following property.

Theorem 21.2 ([20])

(a) The solution $x(t) = 0$ of the system (21.2) is asymptotically stable if and only if all eigenvalues $\lambda_j (j = 1, 2, \ldots, n)$ of A satisfy $|arg\lambda_j| > \alpha\pi/2$.

(b) The solution $x(t) = 0$ of the system (21.2) is stable if and only if the eigenvalues satisfy $|arg\lambda_j| \geq \alpha\pi/2$ and all the eigenvalues with $|arg\lambda_j| = \alpha\pi/2$ have a geometric multiplicity that coincides with their algebraic multiplicity.

Note that in the limiting case $\alpha \to 1$, we recover the well known classical results. To discuss asymptotic stability, mainly we need the following two theorems. One is asymptotic expansions of the Mittag-Leffler functions and the other is a Grownwall-type inequality.

Theorem 21.3 ([11]) *If $0 < \alpha < 2$, β is an arbitrary complex number and μ is an arbitrary real number such that*

$$\frac{\pi\alpha}{2} < \mu < \min\{\pi, \pi\alpha\},$$

then, for an arbitrary $p \geq 1$, the following expansion holds:

$$E_{\alpha,\beta}(z) = -\sum_{k=1}^{p} \frac{z^{-k}}{\Gamma(\beta - \alpha k)} + O\big(|z|^{-1-p}\big), \quad |z| \to \infty, \quad \mu \leq |arg(z)| \leq \pi. \tag{21.3}$$

Theorem 21.4 ([9] [Grownwall-type inequality]) *If*

$$x(t) \leq h(t) + \int_{t_0}^{t} k(s)x(s)\mathrm{d}s, \;\; t \in [t_0, T),$$

where all the functions involved are continuous on $[t_0, T)$, $T \leq \infty$ *and* $k(t) \geq 0$, *then* $x(t)$ *satisfies*

$$x(t) \leq h(t) + \int_{t_0}^{t} h(s)k(s) \exp\Big[\int_{s}^{t} k(u)\mathrm{d}u\Big]\mathrm{d}s, \;\; t \in [t_0, T).$$

If, in addition, $h(t)$ *is non-decreasing, then*

$$x(t) \leq h(t) \exp\Big(\int_{t_0}^{t} k(s)\mathrm{d}s\Big), \;\; t \in [t_0, T).$$

We need the following estimates on Mittag Leffler type functions that is going to arise in the solution representation of our problem.

Lemma 21.1 ([26]) *If all the eigenvalues of A satisfy* $|arg(spec(A))| > \alpha\pi/2$, $0 < \alpha < 1$,

(i) $\|t^{\alpha-1}E_{\alpha,\alpha}(At^{\alpha})\| \to 0$ *as* $t \to \infty$.
(ii) $\exp\Big\{M \int_0^t \|s^{\alpha-1}E_{\alpha,\alpha}(As^{\alpha})\|\mathrm{d}s\Big\}$ *is bounded.*

In view of the above lemma, one obtains the following analogous conditions for $E_{\alpha,\alpha+1}(At^{\alpha})$ by substituting the corresponding asymptotic expansions and following a similar line of proof.

Lemma 21.2 *If all the eigenvalues of A satisfy* $|arg(spec(A))| > \alpha\pi/2$, $0 < \alpha < 1$, *for arbitrary* $\gamma \geq 0$, *then*

(i) $\|t^{\alpha}E_{\alpha,\alpha+1}(At^{\alpha})\| \to 0$ *as* $t \to \infty$.
(ii) $\exp\Big\{M \int_0^t \|s^{\alpha}E_{\alpha,\alpha+1}(As^{\alpha})\|\mathrm{d}s\Big\}$ *is bounded.*

Proof By definition, the series $t^{\alpha}E_{\alpha,\alpha}(\lambda t^{\alpha})$ is always greater than or equal to the series $t^{\alpha}E_{\alpha,\alpha+1}(\lambda t^{\alpha})$. Hence by asymptotic expansion of the first series for $0 < \alpha < 2$, $\beta = \alpha$,

$$\frac{\pi\alpha}{2} < \mu < \min\{\pi, \pi\alpha\},$$

for an arbitrary $p \geq 2$, the following expansion holds:

$$t^{\alpha} E_{\alpha,\alpha}(\lambda t^{\alpha}) = -\sum_{k=2}^{p} \frac{t^{\alpha}(\lambda t^{\alpha})^{-k}}{\Gamma(\alpha-\alpha k)} + O(|\lambda t^{\alpha}|^{-1-p}), \quad |t| \to \infty, \quad \mu \leq |arg(\lambda)| \leq \pi. \tag{21.4}$$

Hence for $p = 3$, we have

$$\begin{aligned} t^{\alpha} E_{\alpha,\alpha}(\lambda t^{\alpha}) &= -\frac{t^{\alpha}}{\lambda^2 t^{2\alpha} \Gamma(\alpha-2\alpha)} - \frac{t^{\alpha}}{\lambda^3 t^{3\alpha} \Gamma(\alpha-3\alpha)} + O\Big(\frac{t^{\alpha}}{|\lambda t^{\alpha}|^4}\Big) \\ &= \frac{-1}{\lambda^2 t^{\alpha} \Gamma(-\alpha)} - \frac{1}{\lambda^3 t^{2\alpha} \Gamma(-2\alpha)} + O\Big(\frac{1}{|\lambda|^4 t^{3\alpha}}\Big) \\ &\to 0 \text{ as } t \to \infty. \end{aligned}$$

Hence $\|t^{\alpha} E_{\alpha,\alpha+1}(\lambda t^{\alpha})| \leq \|t^{\alpha} E_{\alpha,\alpha}(\lambda t^{\alpha})\| \to 0$ as $t \to \infty$, whenever $|arg(\lambda)| > \frac{\alpha\pi}{2}$. Coming to the matrix portion, by some simple calculations, we see that $\|t^{\alpha} E_{\alpha,\alpha+1}(At^{\alpha})\| \to 0$ whenever $|arg(spec(A))| > \frac{\alpha\pi}{2}$. The result (ii) can be obtained through a series of steps similar to the proof as in [26].

21.3 Main Results

Consider the nonlinear fractional Volterra integrodifferential equation of the form

$$\begin{aligned} {}^{C}D^{\alpha}x(t) &= Ax(t) + \int_0^t K(t,s)G(x(s), {}^{C}D^{\beta}x(s))\,\mathrm{d}s, \quad 0 \leq \alpha, \beta \leq 1, \quad t \in J, \\ x(0) &= x_0, \end{aligned} \tag{21.5}$$

where $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times n}$, $K : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}$ and $G : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$.

We assume the following hypotheses.

(H1) There exists a continuous function $m_1 : [0, T] \to [0, \infty)$ such that

$$\|K(t,s)\| \leq m_1(t)$$

and a continuous non-decreasing function $\Omega : (0, \infty) \to (0, \infty)$ such that

$$\|G(y)\| \leq \Omega(\|x\| + \|y\|)$$

and

$$\int_r^T m(s)\mathrm{d}s < \int_r^{\infty} \frac{\mathrm{d}s}{\Omega(s)}.$$

(H2) There exists a constant $M > 0$ and a continuous function $m_2 : [0, T] \to [0, \infty)$ such that

$$\frac{k_1}{\Gamma(1-\beta)}t^{-\beta} + \frac{n_2}{\Gamma(1-\beta)}\int_0^t (t-\xi)^{-\beta} m_1(\xi)\, \Omega(r(\xi))\, \mathrm{d}\xi \leq Mm_2(t)\, \Omega(r(t)),$$

where

$$\begin{aligned} n_1 &= \sup \|E_\alpha(At^\alpha)\|,\ t \in J, \\ n_2 &= \sup \|(t-s)^\alpha E_{\alpha,\alpha+1}(A(t-s)^\alpha)\|,\ t, s \in J, \\ n_3 &= \sup \|At^{\alpha-1} E_{\alpha,\alpha}(At^\alpha)\|, t \in J, \\ n_4 &= \sup \|(t-s)^{\alpha-1} E_{\alpha,\alpha}[A(t-s)^\alpha]\|,\ t, s \in J, \\ m(t) &= n_2 m_1(t) + Mm_2(t), \\ r &= n_1 \|x_0\|. \end{aligned}$$

(H3) The kernel $K(t, s)$ is bounded by a constant say, $M_1 > 0$, and the nonlinear function $G(x, y)$ is bounded in the sense that $\|G(x, y)\| \leq M_2(\|x\| + \|y\|)$, where $M_2 > 0$ is a constant.

Lemma 21.3 *Consider the nonlinear system (21.5). Under the hypotheses (H1)–(H2), the norm functions $\|x(t)\|$ and $\|^C D^\beta x(t)\|$ are bounded.*

Proof The solution of (21.5) by using the Laplace transform technique is given by

$$x(t) = E_\alpha(At^\alpha)x_0 + \int_0^t (t-s)^\alpha E_{\alpha,\alpha+1}(A(t-s)^\alpha)K(s,\tau)G(^C D^\beta x(s))\, \mathrm{d}s.$$

By the hypotheses, we have

$$\|x(t)\| \leq n_1\|x_0\| + n_2 \int_0^t m_1(s)\Omega(\|x\| + \|^C D^\beta x\|)\, \mathrm{d}s.$$

Let us take the right-hand side of the inequality as $r_1(t)$; then

$$r_1'(t) = n_2 m_1(t)\Omega(\|x\| + \|^C D^\beta x\|). \tag{21.6}$$

Now

$$x'(t) = At^{\alpha-1}E_{\alpha,\alpha}(At^\alpha)x_0 + \int_0^t (t-s)^{\alpha-1}E_{\alpha,\alpha}(A(t-s)^\alpha)K(s,\tau)G(x(s), {}^C D^\beta x(s))\, \mathrm{d}s$$

and
$$\begin{aligned} \|x'(t)\| &\leq n_3\|x_0\| + n_4 \int_0^t m_1(s)\Omega(\|x\| + \|^C D^\beta x\|)\mathrm{d}s \\ &\equiv k_2 + n_4 \int_0^t m_1(s)\Omega(\|x\|) + \|^C D^\beta x\|)\, \mathrm{d}s. \end{aligned} \tag{21.7}$$

Hence it follows that

$$
\begin{aligned}
\|^C D^\beta x(t)\| &\leq \frac{1}{\Gamma(1-\beta)} \int_0^t (t-s)^{-\beta} \|x'(s)\| \,\mathrm{d}s \\
&\leq \frac{k_2}{\Gamma(1-\beta)} \Big[\frac{t^{1-\beta}}{1-\beta}\Big] \\
&\quad + \frac{n_4}{\Gamma(1-\beta)(1-\beta)} \int_0^t (t-\xi)^{1-\beta} m_1(\xi) \Omega(\|x\| + \|^C D^\beta x\|) \,\mathrm{d}\xi \\
&\leq \frac{k_2}{\Gamma(2-\beta)} t^{1-\beta} + \frac{n_4}{\Gamma(2-\beta)} \int_0^t (t-\xi)^{1-\beta} m_1(\xi) \Omega(\|x\| + \|^C D^\beta x\|) \,\mathrm{d}\xi.
\end{aligned}
$$

Let us take the right hand side of the inequality as $r_2(t)$; then

$$
r_2'(t) = \frac{k_2}{\Gamma(1-\beta)} t^{-\beta} + \frac{n_4}{\Gamma(1-\beta)} \int_0^t (t-\xi)^{-\beta} m_1(\xi) \Omega(\|x\| + \|^C D^\beta x\|) \,\mathrm{d}\xi. \tag{21.8}
$$

Now we see that $r_1(0) = n_1 \|x_0\| = r$, as given in the hypothesis and $r_2(0) = 0$. Let $w(t) = r_1(t) + r_2(t)$, $t \in J$. Then $w(0) = r_1(0) + r_2(0) = r$ and

$$
w'(t) = r_1'(t) + r_2'(t) \leq m(t) \Omega(w(t)).
$$

Then, for each $t \in J$,

$$
\int_r^t \frac{w'(s)}{\Omega[w(s)]} \,\mathrm{d}s \leq \int_r^T m(s) \,\mathrm{d}s
$$

which implies

$$
\int_{w(0)}^{w(t)} \frac{\mathrm{d}s}{\Omega(s)} \leq \int_r^T m(s) \,\mathrm{d}s < \int_r^\infty \frac{\mathrm{d}s}{\Omega(s)}.
$$

The above inequality implies that there exists a constant K such that $w(t) = r_1(t) + r_2(t) \leq K$, $t \in J$, and hence

$$
\|x(t)\| + \|^C D^\beta x(t)\| \leq K.
$$

We establish the main result now.

Theorem 21.5 *Assume that the hypotheses (H1)–(H3) hold. Then the nonlinear system* (21.5) *is asymptotically stable whenever the eigenvalues of A satisfy*

$$
\|arg(spec(A))\| > \frac{\alpha\pi}{2}. \tag{21.9}
$$

Proof The solution of the system is given by

$$x(t) = E_\alpha(At^\alpha)x_0 + \int_0^t (t-s)^\alpha E_{\alpha,\alpha+1}(A(t-s)^\alpha)K(s,\tau)G(x(s), {}^C D^\beta x(s))\,\mathrm{d}s,$$

hence

$$\|x(t)\| \le \|E_\alpha(At^\alpha)x_0\| + \int_0^t \|(t-s)^\alpha E_{\alpha,\alpha+1}(A(t-s)^\alpha)\| \|K(s,\tau)\| \|G(x(s), {}^C D^\beta x(s))\|\,\mathrm{d}s.$$

By using the hypothesis (H3) and Lemma 21.3, we have

$$\begin{aligned}\|x(t)\| &\le \|E_\alpha(At^\alpha)x_0\| + M_1M_2\int_0^t \|(t-s)^\alpha E_{\alpha,\alpha+1}(A(t-s)^\alpha)\|(\|x\| + \|{}^C D^\beta x\|)\,\mathrm{d}s,\\ &\le \|E_\alpha(At^\alpha)x_0\| + M_1M_2K\int_0^t \|(t-s)^\alpha E_{\alpha,\alpha+1}(A(t-s)^\alpha)\|\,\mathrm{d}s,\\ &\le \|E_\alpha(At^\alpha)x_0\| + M_1M_2KK'\int_0^t \|(t-s)^\alpha E_{\alpha,\alpha+1}(A(t-s)^\alpha)\| \|x\|\,\mathrm{d}s,\end{aligned}$$

for a suitable choice of a constant $K' > 0$. By using the Grownwall-type inequality, we have

$$\|x(t)\| \le \|E_\alpha(At^\alpha)x_0\| \Big[\exp\Big\{M_1M_2KK'\int_0^t \|(t-s)^\alpha E_{\alpha,\alpha+1}(A(t-s)^\alpha)\|\mathrm{d}s\Big\}\Big].$$

By inequality (21.9) and hence by Lemma 21.2, the term

$$\exp\Big\{M_1M_2KK'\int_0^t \|(t-s)^\alpha E_{\alpha,\alpha+1}(A(t-s)^\alpha)\|\mathrm{d}s\Big\}$$

is bounded. Also, by Theorem 21.2,

$$\|E_\alpha(At^\alpha)\| \to 0 \text{ as } t \to \infty.$$

Hence $\|x(t)\| \to 0$ as $t \to \infty$ for any non-zero initial value x_0. Hence the zero solution of the system (21.5) is asymptotically stable.

21.4 Examples

We give some examples to validate the theory.

Example 21.1 Consider the nonlinear Volterra integrodifferential equation

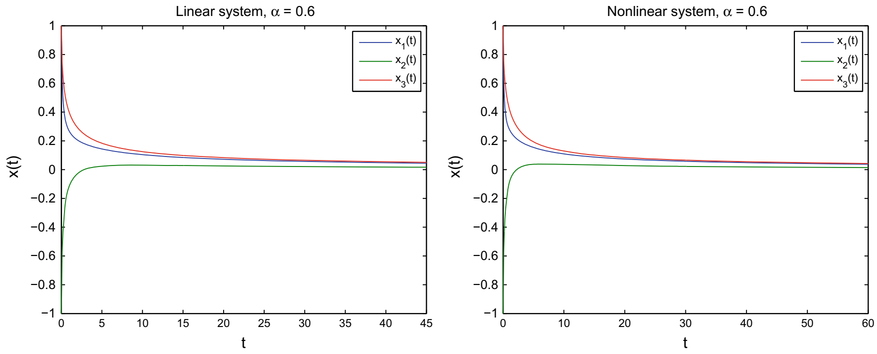


Fig. 21.1 The nonlinear system (21.10) is asymptotically stable

$$
\begin{aligned}
{}^{C}D^{\alpha}x(t) &= Ax(t) + \int_0^t K(t,s)G(x(s), {}^{C}D^{\beta}x(s))\,\mathrm{d}s, \quad t \in J, \\
x(0) &= x_0,
\end{aligned}
\qquad (21.10)
$$

where $\alpha = 0.6$, $\beta = 0.3$, $A = \begin{pmatrix} -2 & 1 & 1/2 \\ 1/2 & -1 & 4/5 \\ 0 & 1/5 & -1 \end{pmatrix}$, $K(t,s) = \dfrac{1}{e^{3(t-s)}}$, $G(x, {}^{C}D^{\beta}x) = {}^{C}D^{0.3}x^3(s)$ and $x(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{pmatrix}$ with $x_0 = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}$.

The eigenvalues of matrix A are -2.3699, 0.3820, and -1.2481. Then

$$|arg(-2.3699)| = |\tan^{-1}\left(\frac{0}{-2.3699}\right)| = \pi > \frac{3\pi}{10},$$

$$|arg(-0.3820)| = |\tan^{-1}\left(\frac{0}{-0.3820}\right)| = \pi > \frac{3\pi}{10}$$

and

$$|arg(-1.2481)| = |\tan^{-1}\left(\frac{0}{-1.2481}\right)| = \pi > \frac{3\pi}{10}.$$

The eigenvalue criterion is satisfied. Also the kernel $\dfrac{1}{e^{3(t-s)}}$ is continuous and bounded; the nonlinear function ${}^{C}D^{0.3}x^3(s)$ is continuous and both satisfy growth conditions given in the hypotheses of Theorem 21.5. Hence the system (21.10) is asymptotically stable as can be observed in Fig. 21.1. The convergence rate of the solution is different for different values of α, as can be seen in Fig. 21.2.

Example 21.2 Consider the nonlinear Volterra integrodifferential equation

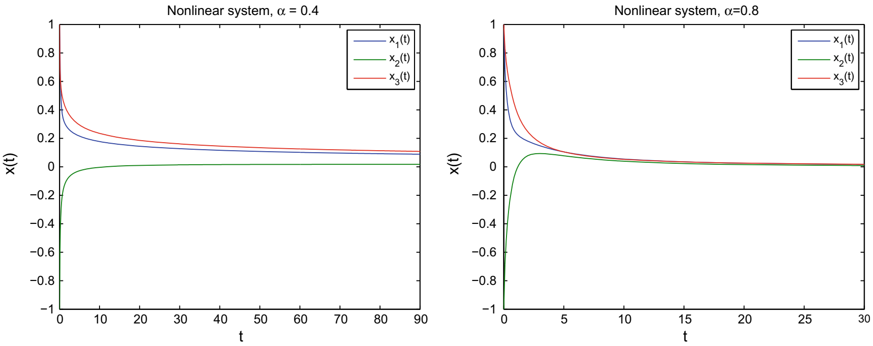


Fig. 21.2 Asymptotic stability of the nonlinear system (21.10), for different values of α

$$
\begin{aligned}
{}^{C}D^{\alpha}x(t) &= Ax(t) + \int_0^t K(t,s)G(x(s), {}^{C}D^{\beta}x(s))\,\mathrm{d}s, \quad t \in J, \qquad (21.11)\\
x(0) &= x_0,
\end{aligned}
$$

where $\alpha = 0.65$, $\beta = 0.8$, $A = \begin{pmatrix} 1+2i & 1/4 \\ 3/5 & 1+3i \end{pmatrix}$, $K(t,s) = \dfrac{\sin(t-s)}{e^{4t}}$, $G(x, {}^{C}D^{\beta}x) = {}^{C}D^{0.8}(\sin(x(s)))$ and $x(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix}$ with $x_0 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$.

The eigenvalues of the matrix A are $1 + 2.1838i$ and $1 + 2.8612i$. Then the required eigenvalue criterion is

$$|arg(1 + 1838i)| = 1.1414 > \frac{\alpha\pi}{2} = \alpha(1.5708)$$

and

$$|arg(1 + 8162i)| = 1.2296 > \frac{\alpha\pi}{2} = \alpha(1.5708).$$

The above two inequalities are satisfied only if $\alpha < \min(0.7266, 0.7828) = 0.7266$. Also the kernel $\dfrac{\sin(t-s)}{e^{t}}$ is continuous and bounded; the nonlinear function ${}^{C}D^{0.8}(\sin(x(s)))$ is continuous and both satisfy growth conditions given in the hypotheses of Theorem 21.5. Hence for all values of $\alpha < 0.7266$, the nonlinear system (21.11) is asymptotically stable as it can be observed in Fig. 21.3. Also, it is to be noted that for $\alpha > 0.73$, the system is unstable as shown in Fig. 21.4. Here we observe that the first order system is unstable but still it's fractional counterpart is stable for the prescribed values of α.

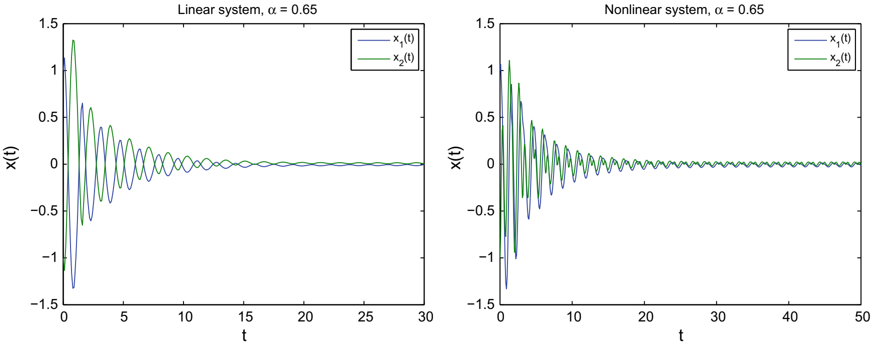
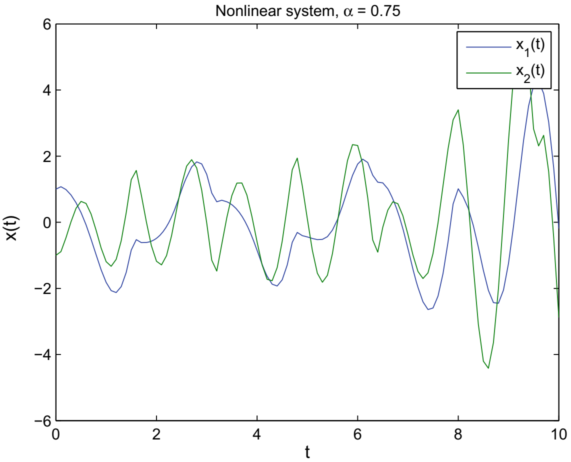


Fig. 21.3 The nonlinear system (21.11) is asymptotically stable for $\alpha < 0.73$

Fig. 21.4 The nonlinear system (21.11) is unstable for $\alpha > 0.73$



21.5 Conclusion

We have obtained a set of sufficient conditions to establish the asymptotic stability of the implicit fractional Volterra integrodifferential equation. However, the conditions can be improved further. Lyapunov's indirect method is followed to establish the asymptotic stability. The graphical examples provided illustrate the effectiveness of considering fractional systems over integer-order systems.

References

1. K. Balachandran, Controllability of nonlinear Volterra integrodifferential systems. Kybernetika **25**, 505–508 (1989)
2. K. Balachandran, S. Divya, Controllability of nonlinear implicit fractional integrodifferential systems. Int. J. Appl. Math. Comput. Sci. **24**, 713–722 (2014)

3. K. Balachandran, S. Divya, M. Rivero, J.J. Trujillo, Controllability of nonlinear implicit neutral fractional Volterra integrodifferential systems. J. Vib. Control **22**, 2165–2172 (2016)
4. K. Balachandran, V. Govindaraj, L. Rodrguez-Germa, J.J. Trujillo, Controllability results for nonlinear fractional-Order dynamical systems. J. Optim. Theory Appl. **156**, 33–44 (2013)
5. K. Balachandran, S. Kiruthika, J.J. Trujillo, Existence results for fractional impulsive integrodifferential equations in Banach spaces. Commun. Nonlinear Sci. Numer. Simul. **16**, 1970–1977 (2011)
6. M. Benchohra, J.E. Lazreg, On stability for nonlinear implicit fractional differential equations. Matematiche (Catania) **LXX** 49–61 (2015)
7. T.A. Burton, *Stability and Periodic Solutions of Ordinary and Functional Differential Equations* (Academic Press Inc, Orlando, 1985)
8. T.A. Burton, *Volterra Integral and Differential Equations* (Elsevier, Amsterdam, 2005)
9. C. Corduneanu, *Principles of Integral and Differential Equations* (Allyn and Bacon, Boston, 1971)
10. S.K. Choi, B. Kang, K. Namjip, Stability for Caputo fractional differential systems. Abstr. Appl. Anal. **2014**, 6 (2014)
11. K. Diethelm, *The Analysis of Fractional Differential Equations* (Springer, Berlin, 2004)
12. K. Diethelm, N.J. Ford, A.D. Freed, A predictor-corrector approach for the numerical solution of fractional differential equations. Nonlinear Dynam. **29**, 3–22 (2002)
13. M. Janfada, G. Sadeghi, Stability of the Volterra integrodifferential equation. Folia Math. **18**, 11–20 (2013)
14. B. Kamalapriya, K. Balachandran, N. Annapoorani, Existence results for fractional integrodifferential equations. Nonlinear Funct. Anal. Appl. **22**, 641–653 (2017)
15. A.A. Kilbas, H.M. Srivastava, J.J. Trujillo, *Theory and Applications of Fractional Differential Equations* (Elsevier, Amsterdam, 2006)
16. K.D. Kucche, S.D. Sutar, Stability via successive approximation for nonlinear implicit differential equations. Moroccan J. Pure Appl. Math. **3**, 36–54 (2017)
17. C. Li, F. Zhang, A survey on the stability of fractional differential equations. Eur. Phys. J. Spec. Top. **193**, 27–47 (2011)
18. Y. Li, Y.Q. Chen, I. Podlubny, Mittag-Leffler stability of fractional order nonlinear dynamic systems. Automatica **45**, 1965–1969 (2009)
19. Y. Li, Y. Chen, I. Podlubny, Stability of fractional-order nonlinear dynamic systems: Lyapunov direct method and generalized Mittag-Leffler stability. Comput. Math. Appl. **59**, 1810–1821 (2010)
20. D. Matignon, Stability results for fractional differential equations with applications to control processing. Comput. Engg. Sys. Appl. **2**, 963–968 (1996)
21. J.J. Nieto, A. Ouahab, V. Venktesh, Implicit fractional differential equation via the Liouville-Caputo derivative. Mathematics **3**, 318–411 (2015)
22. M. Odibat, S. Momani, An algorithm for the numerical solutions of differential equations of fractional order. J Appl. Math. Inf. **26**, 15–27 (2008)
23. I. Podlubny, *Fractional Differential Equations* (Academic Press, New York, 1999)
24. I. Podlubny, Geometric and physical interpretation of fractional integration and fractional differentiation. Fract. Calc. Appl. Anal. **5**, 367–386 (2002)
25. S. Priyadharsini, Stability of fractional neutral and integrodifferential systems. J. Fract. Calc. Appl. **7**, 87–102 (2016)
26. D. Qian, C. Li, R.P. Agarwal, P.J.Y. Wong, Stability analysis of fractional differential system with Riemann-Liouville derivative. Math. Comput. Modelling **52**, 862–874 (2010)
27. S. Sevgin, H. Sevli, Stability of a nonlinear Volterra Integrodifferential equation via a fixed point approach. J. Nonlinear Sci. Appl. **9**, 200–207 (2016)
28. J.V.C. Sousa, E.C. Oliveira, Ulam Hyers stability of a nonlinear fractional Volterra integro-differential equation. Appl. Math. Lett. **81**, 5–56 (2008)
29. C. Tunc, Asymptotic stability and boundedness criteria for nonlinear retarded Volterra Integrodifferential equations. J. King Saud Univ. Sci. (in press) (2017)

30. C. Tunc, S.R. Mohammed, On the stability and instability of functional Volterra integrodifferential equations of first order. Bull. Math. Anal. Appl. **9**, 151–160 (2017)
31. J. Vanualailai, S. Nakagiri, Stability of a system of Volterra integrodifferential equations. J. Math. Anal. Appl. **281**, 602–619 (2003)
32. R. Zhang, S. Yang, S. Feng, Stability analysis of a class of differential systems with Riemann-Liouville derivative. IEEE/CAA J. Autom. Sinica **4**, 1–7 (2007)