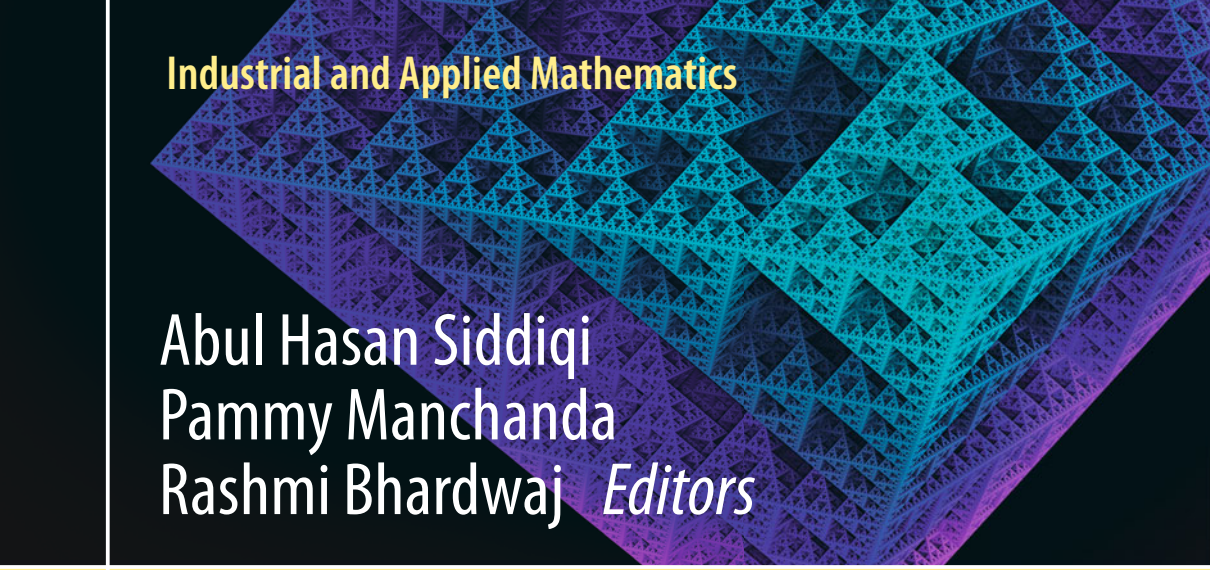

Industrial and Applied Mathematics

Abul Hasan Siddiqi
Pammy Manchanda
Rashmi Bhardwaj *Editors*

Mathematical Models, Methods and Applications

Springer

Industrial and Applied Mathematics

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

Abul Hasan Siddiqi · Pammy Manchanda
Rashmi Bhardwaj
Editors

Mathematical Models, Methods and Applications

Springer

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

Rashmi Bhardwaj
Guru Gobind Singh Indraprastha University
Dwarka
India

Pammy Manchanda
Guru Nanak Dev University
Amritsar, Punjab
India

ISSN 2364-6837 ISSN 2364-6845 (electronic)
Industrial and Applied Mathematics
ISBN 978-981-287-971-4 ISBN 978-981-287-973-8 (eBook)
DOI 10.1007/978-981-287-973-8

Library of Congress Control Number: 2015951755

Springer Singapore Heidelberg New York Dordrecht London

Printed on acid-free paper

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

Preface

The Indian Society of Industrial and Applied Mathematics (ISIAM) was established during a national symposium on differential equations in September 1990 at Aligarh Muslim University. Since then it has been organizing national and international conferences, seminars, workshops and symposiums in different parts of India. Proceedings of these academic activities have been published by reputed publishers including Longman (Pitman Research Notes in Mathematics), Kluwer Academic Publications (Now part of Springer Group), Taylor and Francis Publications, etc.

The present volume contains invited talks and some contributory talks of 11th International Biennial Conference on "Emerging Mathematical Methods, Models and Algorithms for Science and Technology" organized under the auspices of the society. This international conference was organized at Gautam Buddha University, National Capital Region, India, from December 15–16, 2012. This conference commemorates 125th birth year of the Mathematics Wizard Srinivasa Ramanujan. The conference was attended by more than 200 persons belonging to different specializations of mathematics, engineering, physics, computer science, information technology, and management studies coming from various states of India and countries like USA, Germany, France, Italy, Turkey, Saudi Arabia, and Oman. The conference was really interdisciplinary in nature, where applications of mathematical concepts to emerging technologies were focused.

The conference was inaugurated by Prof. Krishan Lal, President of the Indian National Science Academy (INSA) and eminent academicians such as Prof. H.P. Dikshit (Chairman EPCO. Institute of Environmental Studies, Govt. of Madhya Pradesh and former Vice-Chancellor of IGNOU), Prof. U.B. Desai, Director IIT Hyderabad (a renowned expert of information Technology and Tele Communication), Prof. N.K. Gupta, IIT Delhi (a renowned expert of Impact Problems and former Vice-President of INSA and Current President ISIAM), Prof. Moinuddin, Pro. Vice-Chancellor Delhi Technical University (Former Director NIT, Jalandhar), Prof. Aparajita Ojha, Director IIIT Jabalpur, Prof. Rajat Gupta, Director NIT, Srinagar, Prof. M. Brokate, former Dean School of Mathematical Sciences, Technical University Munich, Germany, Prof. R. Lozi,

CNRS & Nice University France et al. participated and delivered lectures. A special session on 125th birthday celebration of Ramanujan was also organized during the conference, and Prof. Dinesh Singh, Vice-Chancellor, Delhi University was the chief guest of this function.

On this occasion, Prof. U.B. Desai was conferred Dr. Zakir Husain Award 2011/2012 for his valuable contribution in the emerging areas like cyber physical systems, cognitive radio, wireless communication, wireless sensor networks, additive signal, and image processing. He has extensively used mathematical concepts such as wavelets and multiresolution analysis, artificial neural network, and fractals in his research works.

In the inaugural address, Prof. Krishan Lal highlighted the importance of mathematics for industrial and technological development of any nation. He expressed the serious concern of the scientists, engineers, and all well-wishers of our nation on dwindling standard of mathematics and especially applications of mathematics. He emphasized that the need of the hour is to attract talented young researchers towards applications of mathematics in emerging areas of science and technology. All invited speakers on this occasion echoed the same sentiment.

During the inaugural function, Prof. Pammy Manchanda, Convener Scientific Committee read the messages of the Hon'ble President of India, Hon'ble Union Minister of Communication and Information Technology, Minister of External Affairs, Minister of Water resources, Governors of Bengal, Jammu and Kashmir, Uttrakhand, Minister of State for Human Resource Development, and 10 other dignitaries including Prof. Barbara Lee Keyfitz, President International Council of Industrial and Applied Mathematics (www.iciam.org).

The invited and contributory talks published in the proceedings provide valuable information on certain current trends in mathematical models, methods, and algorithms. Rene Lozi discusses the cryptography-based chaos which provides a new mechanism for undersampling chaotic numbers obtained by the ring coupling of one-dimensional maps in Chap. 1. In Chap. 2, D.K. Chaturvedi provides the vital information about applications of soft computing techniques. Image decomposition–reconstruction is very important in image analysis and it has a wide range of applications in radar imaging which is discussed by Gaik Ambartsoumian and Venkateswaran P. Krishnan in Chaps. 3 and 4 respectively. Two-dimensional nonlinear elliptic boundary value problems by cubic spline approximation method is explored by R.K. Mohanty in Chap. 5. Application of Monte Carlo simulation to pricing of path-dependent European-type options is discussed by Siddhartha P. Chakrabarty in Chap. 6. Messaoud Boulbrachene's paper deals with the finite element approximation of the impluse control quasivariational inequality in Chap. 7. In Chap. 8, Chefi Triki and Nasr Al-Hinai give an overview of the Periodic Petrol Station Replenishment Problem. Mushahid Husain and Ayub Khan present their recent work in nanotechnology in Chap. 9. Chapter 10 contains results on generalized monotone mappings by R. Rais et al.

Rashmi Bhardwaj highlights the application of wavelet and fractal methods to environmental problems, especially problem of air and water pollution in Chap. 11. Mohd Ahmad Ansari provides an algorithm by context modeling of medical image

compression using discrete wavelet transform in Chap. 12. In Chap. 13, K. Srinivasa Rao (first DST-Ramanujan Professor) gives an elegant account of the life and work of Ramanujan, a creative genius. Sushil Kumar et al. study the dispersion in steady and oscillatory flows through curved channels with absorbing boundaries in Chap. 14. Noor-e-Zahra explains the basic ingredients of a new technology, compression—sensing in Chap. 15. Ruchira Aneja's paper is devoted to the emergence of shearlets and its applications in Chap. 16. Nagma Irfan et al. discuss the application of CAS wavelets in numerical evaluation of Hankel transforms arising in seismology in Chap. 17.

The main message conveyed through the conference is that mathematics has great potential to analyze and understand the challenging problems of nanotechnology, biotechnology, medical science, oil industry, environmental sciences, engineering, and financial technology. It has been emphasized throughout the conference that young researchers of the country should embark on those areas of mathematics which have significant applications in these fields.

I take this opportunity to thank Profs. Pammy Manchanda and Rashmi Bhardwaj coeditors of the proceedings for their valuable help.

Prof. Abul Hasan Siddiqi

Acknowledgments

Editors gratefully acknowledge the support of the Department of Science and Technology (DST), New Delhi; National Board of Higher Mathematics (NBHM), Mumbai; Council of Scientific and Industrial Research (CSIR), New Delhi; Defence Research Development Organization (DRDO), New Delhi; Duty Society, Aligarh; Aligarh Muslim University (AMU), Aligarh, and Gautam Buddha University (GBU), Greater Noida, India for the financial support. The administration of Gautam Buddha University deserves our full appreciation, particularly Dr. Sushil Kumar, Head of Mathematics Department for providing excellent facilities. We also take this opportunity to express our gratitude to the Springer, especially to Mr. Shamim Ahmed for his constant advice for improvement of the manuscript.

Abul Hasan Siddiqi
Pammy Manchanda
Rashmi Bhardwaj

Contents

About the Editors

Abul Hasan Siddiqi Ph.D. is a distinguished scientist and professor emeritus at School of Basic Sciences and Research at Sharda University, Greater Noida, India. He is also a visiting consultant at the International Centre for Theoretical Physics (ICTP) (Trieste, Italy), Sultan Qaboos University (Muscat, Oman), King Fahd University of Petroleum and Minerals (KFUPM) (Dhahran, Saudi Arabia) and several other well-known universities of the world. He has a long association with ICTP (a UNESCO institution) in capacities of short-time visitor, long-duration visitor, senior associate, guests of the director, senior associate, and ICTP visiting consultant to Turkey. He was awarded the German Academic Exchange Fellowship thrice to carry out mathematical research in Germany. He has published more than 100 research papers jointly with his research collaborators, 5 books and edited proceedings of 9 international conferences, as well as supervised 29 PhD scholars. He is the founder secretary of the Indian Society of Industrial and Applied Mathematics (ISIAM), which is celebrating its Silver Jubilee in January 2016. He is editor-in-chief of *Indian Journal of Industrial and Applied Mathematics*, published by ISIAM and Industrial and *Applied Mathematics*, a book series with Springer.

Pammy Manchanda Ph.D. is senior professor of mathematics at Guru Nanak Dev University, Amritsar, India. She has published 44 research papers in several international journals of repute, edited 2 proceedings of international conferences of ISIAM and co-authored 3 books. She has attended and delivered talks and chaired sessions at reputed academic conferences and workshops across the world, including ICIAM (1999–2015) and ICM since 2002. She was invited twice to the Industrial Mathematics Group of Prof Helmut Neunzert, Kaiserslautern University, Germany, and visited International Centre for Theoretical Physics (a UNESCO institution) (Trieste, Italy) many times to carry out her research activities. She is the joint secretary of the Indian Society of Industrial and Applied Mathematics (ISIAM) since 1999 and has been actively engaged in organizing international conferences by the society. She is managing editor of the *Indian Journal of Industrial and Applied Mathematics* and member of the editorial board of *Industrial and Applied Mathematics*, Springer book series.

Rashmi Bhardwaj Ph.D. is professor of mathematics, Guru Gobind Singh Indraprastha University (GGSIPU), New Delhi, India. She is also visiting associate of the Inter University Centre for Astronomy and Astrophysics (IUCAA). With 24 years of research and 19 years' teaching experience, she was awarded the Young Scientist Award from the All India Council for Technical Education (AICTE) and has received the Best Researcher Award by GGSIPU thrice. She established the Nonlinear Dynamics Research Lab and Mathematics Lab in GGSIPU. She is the Delhi University topper in M.Sc., acquiring 100% marks in "integral equation and boundary value problem". She has published over 80 research papers, 2 books and edited 1 proceedings of an international conference. She has supervised 10 PhD scholars and 5 research scholars are presently pursuing the PhD under her supervision. In addition to being a referee for many international journals of repute and a life member of several international organizations in the sphere of mathematical sciences, she is managing editor of the *Indian Journal of Industrial and Applied Mathematics.*

Summary

Dr. Zakir Husain Memorial Lecture; Smarter Societies: Cyber Physical System

We live in a highly connected world and connectivity is exploding. By 2015 we will have 15 billion devices connected to the Internet and by 2020 the number of connected devices is expected to reach 50 billion; in 2011 amount of data transmitted around the world exceeded 2 zettabytes, i.e., 2 × 10^21 bytes; by 2020 the world will generate 100 zettabytes. By 2017 a trillion wireless devices will be there serving 7 billion people. These numbers are mind-boggling, and they are creating not just technological challenges but also profound mathematical challenges. Many believe, to tackle the mathematical challenges of 50 billion Internet-connected devices or a trillion wireless devices it may require some new mathematics. Coupled with these mathematical challenges, the networked world is throwing new challenges in innovations and enhanced business opportunities. The challenges get compounded due to the deluge of information. Our psyche is governed by the networked world and with time we are moving to a *smarter society*.

The attributes of smarter society are (but not limited to): highly connected society, ubiquitous communication, strong connectivity between physical world and cyber world, everything will be connected to the Internet; data analytics will be backbone, and this will involve complex multivariable predictive algorithms and data interpretation involving high-level machine learning algorithms. In short, we will move towards a society where there will be seamless intelligent interaction between computers and humans.

This talk focused on a major subset of the smarter society, namely, cyber physical systems (CPS) or Internet of things (IoT). CPS and IoT are going to change the world in the coming days.

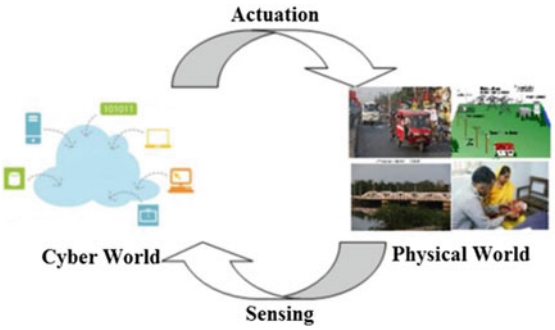

Fig. 1 Cyber physical system

Cyber physical system is a system which integrates the cyber world with the physical world using sensors and actuators; CPS closes the loop in the Internet. Applications of CPS will be there in all walks of life: agriculture, power systems, medical systems and health care, transportation, finance, smart structures, and many more. Many believe that the impact of CPS would be as big as or even bigger than the Internet (Fig. 1).

The key building blocks of CPS are: Communication, computing, control, sensing, and cognition. Communication, computing, and control are fairly mature; a lot of work needs to be done in sensors and cognition. Sensor technology has always worked in a niche domain and thus sensors are quite expensive—one needs major research to make sensors affordable and pervasive.

Internet of things (IoT) is very closely related to CPS. In IoT everything and anything is Internet enabled. IoT can be viewed as convergence of Internet, signal processing, VLSI, communication, and sensing. IoT has the same application domain as CPS.

CPS and IoT offer many technological and mathematical challenges. Below, very briefly, a description is provided for a few of the mathematical challenges.

One of key challenge is to have a *mathematical model for CPS*—this is challenging as CPS involves discrete components, continuous components, concurrent interactions between discrete and continuous components, and infinite execution. At present most of the work revolves around the use of hybrid automata for modeling CPS. A hybrid automata will model a CPS system with initial states H and a set of safe states M; this entails two key mathematical problems:

1. *Stability*: Does every execution of the CPS starting in any of the initial states from H always stay in the safe states M
2. *Reachability*: Does starting in any initial state from set H, the CPS system will always reach in finite steps (or asymptotically) the set of desirable states

It is likely that instead of hybrid automata, one may need a new kind of math to faithfully model a CPS system.

Smart green buildings are the thing of the future. A lot work involves technological challenges in building and making the system work. Nevertheless, to get a better understanding of smart green buildings and to take the idea forward, mathematical analysis is essential. We consider a smart building as:

1. A graph G = (V, E) with |V| = n and |E| = m.
2. Let some special vertices be designated as "entry" vertices and some as "exit" vertices (representing movement in the building). These vertices are not necessarily disjoint; the same vertex, at times can serve an entry vertex and at some other time as an exit vertex.
3. Each vertex represents a room and each edge represents a connection between two rooms.
4. People enter and leave the graph at the entry and exit vertices, say in a Poisson fashion.
5. The edges are weighted by the distance between the rooms and the probabilities of a person moving in either direction along the edge.
6. Each person that enters the building executes a random walk on the vertices and exits the system.
7. Define occupancy of a vertex as the number of people in the corresponding room at any given point in time.

Problem

Find the cumulative occupancy of any given room over a period of time or at any given time based on which the energy consumption is optimized.

In the above formulation, one can bring in other constraints like available energy (solar, battery, wind, etc.), available information on ambient conditions (temperature, humidity, wind velocity, solar lighting, etc.) and set up a realistic optimization problem. Given a smart room system description, one has the following mathematical challenge:

1. Model the system with sensor inputs (temperature, state of windows, PIR, etc.) and actuation strategies (for ACs, lights, fans, etc.) as possibly a hybrid system
2. Formally prove that the system maintains the desired ambient conditions
3. Prove the kind of energy saving that would be achieved with the system
4. Mathematically design optimal strategies to achieve the specified ambient conditions
5. Extend it to a system of interconnected rooms

Another problem in today's time is *green communication*. We have all tasted the benefits of cellular phones, but behind it there is tremendous energy consumption to keep the base stations alive. Tower diesel genset runs 3–5 h in urban areas and 8–10 h in rural areas. The estimated diesel consumption for all such cell towers is approximately, 3 billion liters per year. Cloud-based radio access network (CRAN) which can support mobility at very high speeds is likely to be a much more energy-efficient technology for mobile communication than the present technology. CRAN involves:

1. Opportunistic placement of cells (where fiber permits)
2. Large number of low-power antenna ports (ATP)
3. As low as 50-m interpoint spacing
4. Connected via fiber to a central controller known as remote base station (RBS)
5. Centralized (Cloud) baseband processing at RBSs
6. Fast hand-off at baseband level (implicit hand-off)

Some of the mathematical challenges that emerge are:

1. Self-optimization and self-organization of the network and mathematically modeling self-organizing networks
2. Modeling and analysis of large distributed multi-input-multi-output dynamically changing network
3. Centralized coordinated scheduling for antenna ports
4. Joint resource allocation among ATPs and users
5. Cross-layer optimization: Typically non-convex optimization

Cognitive radio (CR) is a technology which can very highly optimize the use of a scarce resource—*spectrum*. CR is defined as a transceiver that can combine its awareness of the environment with knowledge of its user's needs, and adapt its parameters intelligently to achieve reliable and spectrally efficient communication. CR involves sensing the spectrum continuously, and whenever a spectrum whole (white space) is available you transmit some bits of data. The CR user is typically the secondary user who either underlays its communication or interweaves its communication with respect to the primary user—thereby increasing spectral utilization and efficiency (Fig. 2).

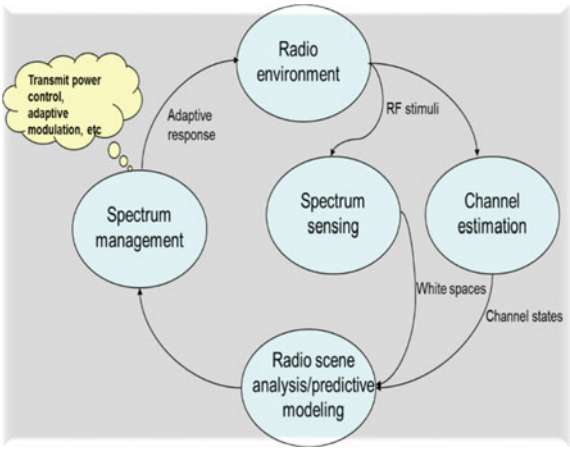

Fig. 2 Typical cognitive radio environment

At IIT Hyderabad, we have made measurements and showed that there are significant spectral holes even in GSM communication. Moreover, a complete CR networking technology and prototype has been developed at IIT Hyderabad for a

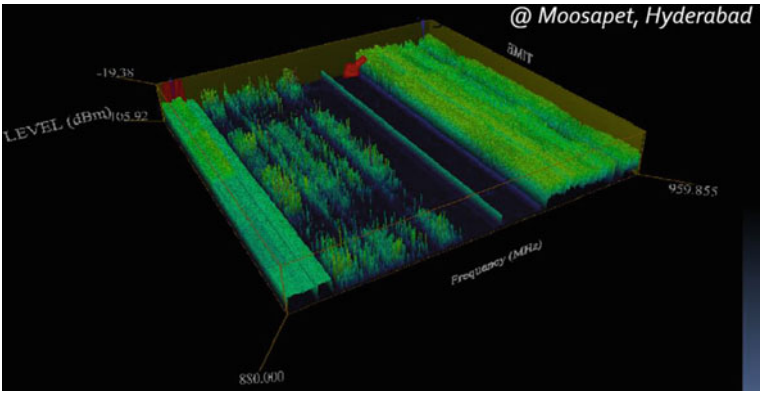

Fig. 3 Black spaces between greens show spectral holes in GSM band

CR base station that can exploit these spectral holes in GSM band. Any GSM handset will work as the CR receiver (Fig. 3).

There are many mathematical challenges in CR from the information theoretic perspective—the key challenge being to investigate information theoretic limits of CR networks with practical constrains. In particular one could focus on: User capacity (primary as well as secondary), data rates for primary as well secondary, and of course a measure of spectral efficiency.

In conclusion it should be mentioned that cyber physical systems and Internet of things are creating new technological challenges, and mathematical challenges. It is not enough that we build things, in order to go beyond we will need mathematical models and possibly new analysis methods. Perhaps, a new math to tackle the exploding connectivity and sensing

I would like to thank the Indian Society of Industrial and Applied Mathematics for conferring the Dr. Zakir Hussain award. I would like to acknowledge my team of faculty at IITH Hyderabad, who are working on cutting-edge CPS and IoT challenges and who helped with my presentation. Some of the team members are Dr. Zafar Khan, Dr. Kiran Kuchi, Dr. G.V. Sharma, Dr. Panduranga Rao, Dr. Rajalakshmi, Dr. Bheema Arjun, and many others.

Uday B. Desai
IIT Hyderabad

Chapter 1
Cryptography-Based Chaos via Geometric Undersampling of Ring-Coupled Attractors

René Lozi

Abstract We propose a new mechanism for undersampling chaotic numbers obtained by the ring coupling of one-dimensional maps. In the case of two coupled maps, this mechanism allows the building of a PRNG which passes all NIST tests. This new geometric undersampling is very effective for generating two parallel streams of pseudo-random numbers, as we show, computing carefully their properties, up to sequences of 10^{12} consecutives iterates of the ring-coupled mapping which provides more than 3.35×10^{10} random numbers in very short time. Both three- and four-dimensional cases can be managed in the same way. In addition, we recall a novel method of noise-resisting ciphering. The originality lies in the use of a chaotic pseudo-random number generator: several cogenerated sequences can be used at different steps of the ciphering process, as they present the strong property of being uncorrelated. Each letter of the initial alphabet of the plain text is encoded as a subinterval of $[-1, 1]$. The bounds of each interval are defined in function of the known bound of the additive noise. A pseudo-random sequence is used to enhance the complexity of the ciphering. The transmission consists of a substitution technique inside a chaotic carrier, depending on another cogenerated sequence. This novel noise-resisting ciphering method can be used with geometric undersampling when four mappings are coupled.

Keywords Cryptography · Chaos · Randomness · Undersampling · Pseudo-random number generator

R. Lozi (✉)
Université de NICE Sophia-Antipolis, Laboratoire J. A. Dieudonné,
UMR CNRS 7351, Parc Valrose, 06108 Nice, Cedex 02, France
e-mail: rlozi@unice.fr

A.H. Siddiqi et al. (eds.), *Mathematical Models, Methods and Applications*,
Industrial and Applied Mathematics, DOI 10.1007/978-981-287-973-8_1

1.1 Introduction

During the last decade, it has been emphasized that the undersampling of sequence of chaotic numbers is an efficient tool to build pseudo-random number generators (PRNG) [15]. Randomness appears to be an emergent property of complex systems of coupled chaotic maps [16]. Several kinds of coupling can be considered as ultra-weak coupling, ring coupling, etc. [17]. An ultra-weak coupling recovers chaotic properties of one-dimensional maps [12, 13] when computed with floating numbers or double-precision numbers. Chaotic undersampling with thresholds based on one component of the coupled system adds random properties to the chaotic sequences. Double threshold sampled sequence (i.e., using both thresholds of different nature) improves such random properties [14]. Ring coupling deals when p one-dimensional maps are constrained on a torus [5, 26], this coupling can directly provide random numbers without sampling or mixing, provided the number p of maps is large enough, although it is possible to combine these processes with it. However, in lower dimension two and three, the chaotic numbers are not equidistributed on the torus. Therefore we introduce a particular "geometric" undersampling based on the property of piecewise linearity of the invariant measure of the system of p one-dimensional ring-coupled maps. This new geometric undersampling is very effective for generating parallel streams of pseudo-random numbers with a very compact mapping.

Several applications in various fields (chaotic optimization, evolutionary algorithms, secure information transmission, chaotic cryptography, etc.) of such undersampling process can be found. In this article we focus on the latter ones.

- As the first example, we propose a novel noise-resisting ciphering based on a large number of uncorrelated chaotic sequences. These cogenerated sequences are actually used in several steps of the ciphering process. Noisy transmission conditions are considered with realistic assumptions. The efficiency of the proposed method for ciphering and deciphering is illustrated through numerical simulations based on ten coupled chaotic sequences [4].
- Another example is the use of such sequences in a chaotic encryption algorithm [27].

In Sect. 1.2, we briefly recall properties of chaotic mappings, when used alone or ultra-weakly coupled. Section 1.3 describes the route from chaos to randomness via chaotic undersampling, discovered during the last decade. In Sect. 1.4, we introduce geometric undersampling in the scope of ring-coupled mapping. In Sect. 1.5, we propose in addition, a novel method of noise-resisting ciphering. The originality lies in the use of a chaotic pseudo-random number generator: several cogenerated sequences can be used at different steps of the ciphering process, as they present the strong property of being uncorrelated. This novel noise-resisting ciphering method can be used with geometric undersampling when four mappings are coupled.

1.2 Recovering Chaotic Properties of Numerically Computed Chaotic Numbers

1.2.1 Numerical Approximation of Chaotic Numbers

Chaos theory studies the behavior of dynamical systems that are highly sensitive to initial conditions, an effect which is popularly referred to as the butterfly effect. Small differences in initial conditions (such as those due to rounding errors in numerical computation) yield widely diverging outcomes for chaotic systems, rendering long-term prediction impossible in general. This happens even though these systems are deterministic, meaning that their future behavior is fully determined by their initial conditions with no random elements involved. In other words, the deterministic nature of these systems does not make them predictable. The first example of such chaotic continuous system in the dissipative case was pointed out by the meteorologist E. Lorenz in 1963 [11].

In order to study numerically the properties of the Lorenz attractor, M. Hénon an astronomer of the Observatory of Nice, France, introduced in 1976 a simplified model of the Poincaré map of this attractor [9]. The Lorenz attractor being imbedded in dimension three, the corresponding Poincaré map is a mapping from the plane $\mathbb{R}^2$ into $\mathbb{R}^2$. Hence the Hénon mapping is also defined in dimension two and is associated to the dynamical system

$$\begin{cases} x_{n+1} = y_n + 1 - ax_n^2, \\ y_{n+1} = bx_n \end{cases} \tag{1.1}$$

which has been extensively studied for 36 years.

More simple dynamical systems in dimension one, on the interval $J = [-1, 1] \subset \mathbb{R}$ into itself

$$x_{n+1} = f_a(x_n), \tag{1.2}$$

corresponding to the logistic map

$$f_a \equiv L_a(x) = 1 - ax^2, \tag{1.3}$$

or the symmetric tent map

$$f_a \equiv T_a(x) = 1 - a|x|, \tag{1.4}$$

have also been fully explored in the hope of generating random numbers easily [24]. However, when a dynamical system is realized on a computer using floating point or double-precision numbers, the computation is of a discretization, where finite machine arithmetic replaces continuum state space. For chaotic dynamical systems in small dimension, the discretization often has collapsing effects to a fixed point or to short cycles [6].

It seems that the computation of numerical approximations of the periodic orbits leads to unpredictable and somewhat enigmatic results. As O.E. Lanford III [10] says "The reason is that because of the expansivity of the mapping the growth of roundoff error normally means that the computed orbit will remain near the true orbit with the chosen initial condition only for a relatively small number of steps typically of the order of the number of bits of precision with which the calculation is done. It is true that the above mapping like many 'chaotic' mappings satisfies a shadowing theorem [20, 21] which ensures that the computed orbit stays near to some true orbit over arbitrarily large numbers of steps. The flaw in this idea as an explanation of the behavior of computed orbits is that the shadowing theorem says that the computed orbit approximates some true orbit but not necessarily that it approximates a typical one."

The collapsing of iterates of dynamical systems or at least the existence of very short periodic orbits, their nonconstant invariant measure, and the easily recognized shape of the function in the phase space avoid the use of one-dimensional map (logistic, baker, or tent, etc.) as a pseudo-random number generator (see [18] for a survey).

Remark 1.1 However, the very simple implementation in computer program of chaotic dynamical systems led some authors to use it as a base of cryptosystem [2, 3]. In addition it seems that for some applications, chaotic numbers are more efficient than random numbers. That is the case for evolutionary algorithms [22, 25] or chaotic optimization [1].

In this paper, we show how to overcome the poor quality of chaotic generators using geometric undersampling. This special undersampling we introduce in this article is one of the other undersampling processes we have studied before. In order to explain the difference between these processes we give in Sect. 1.3 a brief survey of them. Before doing this survey, we have to show how to stabilize the chaotic properties of chaotic number when realized on a computer.

1.2.2 Very Long Periodic Orbits for Ultra-weakly Coupled Tent Map

The first step in order to preserve the genuine chaotic properties of the continuous models in numerical experiments is reached considering ultra-weak multidimensional coupling of p one-dimensional dynamical systems instead of solely a one-dimensional map.

1.2.2.1 Two-Coupled Symmetric Tent Map

In order to simplify the presentation below, we use as an example the symmetric tent map (1.4) with the parameter value $a = 2$, later denoted simply as f, even

though others as chaotic map of the interval, the logistic map, the baker transform, etc., can be used for the same purpose (as a matter of course, the invariant measure of the chaotic map considered is preserved).

When $p = 2$, the system is simply described by Eq. (1.5)

$$\begin{cases} x_{n+1} = (1-\varepsilon_1)f(x_n) + \varepsilon_1 f(y_n) \\ y_{n+1} = \varepsilon_2 f(x_n) + (1-\varepsilon_2)f(y_n) \end{cases}, \tag{1.5}$$

We use generally $\varepsilon_1 = 10^{-7}$, $\varepsilon_2 = 2\varepsilon_1$ when computations are done using floating points or $\varepsilon_1 = 10^{-14}$ for double-precision numbers. In both cases, with these numerical values, the collapsing effect disappears and the invariant measure of any component is the Lebesgue measure [12] as we show below. In the case of computation using floating points, starting from most initial condition, it is possible to find a Megaperiodic orbit (i.e., with period equal to 1,320,752). When computations are done with double-precision number, it is not possible to find any periodic orbit up to $n = 5 \times 10^{11}$ iterations. In [12], the computations have been performed on a Dell computer with a Pentium IV microprocessor using a Borland C compiler computing with ordinary (IEEE-754) double-precision numbers.

When ε_1 converges towards 0, the iterates of each component x_n and y_n of Eq. (1.5) converge to the Lebesgue measure (Fig. 1.1).

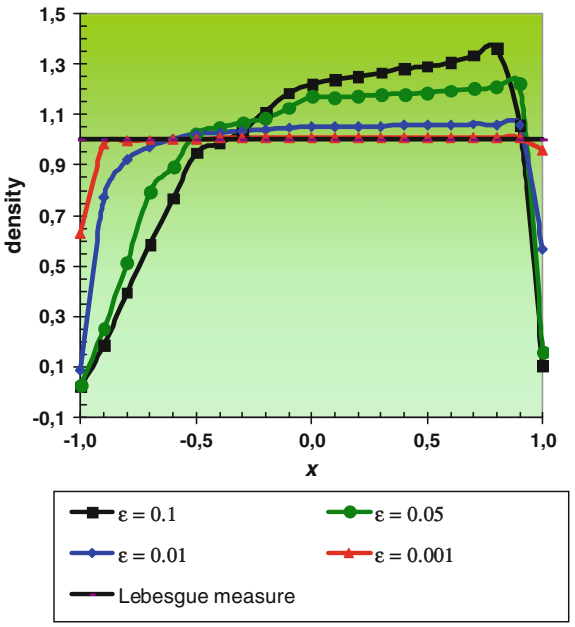

Fig. 1.1 Density of iterates of two-coupled symmetric tent maps, double precision, $N_{\text{disc}} = 10^5$, $\varepsilon_2 = 2\varepsilon_1$, $\varepsilon_1 = 10^{-1}$–10^{-3}, $N_{\text{iter}} = 10^8$, initial values $x_0 = 0.330$, $y_0 = 0.3387564$

1.2.2.2 *P*-Coupled Symmetric Tent Map

More generally, the coupling of p maps takes the form

$$X_{n+1} = F(X_n) = A \cdot \left(\underline{f}(X_n)\right), \tag{1.6}$$

where

$$\underline{f}(X_n) = \begin{pmatrix} f(x_n^1) \\ \vdots \\ \vdots \\ \vdots \\ f(x_n^p) \end{pmatrix}, X_n = \begin{pmatrix} x_n^1 \\ \vdots \\ \vdots \\ \vdots \\ x_n^p \end{pmatrix}, \tag{1.7}$$

and

$$A = \begin{pmatrix} \varepsilon_{1,1} = 1 - \sum_{j=2}^{j=p} \varepsilon_{1,j} & \varepsilon_{1,2} & \cdots & \varepsilon_{1,p-1} & \varepsilon_{1,p} \\ \varepsilon_{2,1} & \varepsilon_{2,2} = 1 - \sum_{j=1,j\neq 2}^{j=p} \varepsilon_{2,j} & \cdots & \varepsilon_{2,p-1} & \varepsilon_{2,p} \\ \vdots & \ddots & & \vdots & \vdots \\ \vdots & & \ddots & \vdots & \vdots \\ \varepsilon_{p,1} & \cdots & \cdots & \varepsilon_{p,p-1} & \varepsilon_{p,p} = 1 - \sum_{j=1}^{j=p-1} \varepsilon_{p,j} \end{pmatrix}, \tag{1.8}$$

with $\varepsilon_{i,i} = 1 - \sum_{j=1,j\neq i}^{j=p} \varepsilon_{i,j}$ on the diagonal (the matrix A is always a stochastic matrix iff the coupling constants verify $\varepsilon_{i,j} > 0$ for every i and j).

It is noteworthy that these families of very weakly coupled maps are more powerful than the usual formulas used to generate chaotic sequences, mainly because only additions and multiplications are used in the computation process and no division is required. Moreover, the computations are done using floating point or double-precision numbers, allowing the use of the powerful floating point unit (FPU) of the modern microprocessors. In addition, a large part of the computations can be parallelized taking advantage of the multicore microprocessors which appear on the market of laptop computers.

Moreover, a determining property of such coupled map is the high number of parameters used ($p \times (p-1)$ for p-coupled equations) which allows to choose it as cipher keys, when used in chaos-based cryptographic algorithms, due to the high sensitivity to the parameters values [16]. It can also be shown that using control theory techniques, synchronization of two systems (1.6), with p = 2 or 3, can be reached via exact (dead-beat) or asymptotic observers [8].

1.2.2.3 Computation of Approximated Invariant Measure

In order to assess numerical computations more accurately, we define an approximation $P_{M,N}(x)$ of the invariant measure also called the probability distribution function linked to the one-dimensional map f, when computed with floating numbers (or numbers in double precision). For this aim we consider a regular partition of M small intervals (boxes) r_i of J defined by

$$s_i = -1 + \frac{2i}{M}, \quad i = 0, M \tag{1.9}$$

$$r_i = [s_i\,,\, s_{i+1}[\,,\, i = 0,\, M-2 \text{ and } \quad r_{M-1} = [s_{M-1}, 1] \tag{1.10}$$

The length of each box is equal to $\frac{2}{M}$ and the r_i intervals form a partition of the interval J

$$J = \bigcup_{0}^{M-1} r_i \tag{1.11}$$

All iterates $f^{(n)}(x)$ belonging to these boxes are collected, after a transient regime of Q iterations decided a priori, (i.e., the first Q iterates are neglected). Once the computation of $N + Q$ iterates is completed, the relative number of iterates with respect to N/M in each box r_i represents the value $P_N(s_i)$. The approximated $P_N(x)$ defined in this article is then a step function with M steps. As M may vary, we define

$$P_{M,N}(s_i) = \frac{M}{N}(\#r_i) \tag{1.12}$$

where $\#r_i$ is the number of iterates belonging to the interval r_i. The approximate function $P_{M,N}(x)$ is normalized to 2 on the interval J.

$$P_{M,N}(x) = P_{M,N}(s_i), \quad \forall x \in r_i \tag{1.13}$$

In the case of p-coupled maps, we are interested by the distribution of each component $(x^1, x^2, x^3, \ldots, x^p)$ of X rather than the distribution of the variable X itself in J^p. We then consider the approximated probability distribution function $P_{M,N}(x^j)$ associated with one of the several components of $F(X)$ defined by (1.6), which are one-dimensional maps. In this paper, we equally use N_{disc} for M and N_{iter} for N, when they are more explicit.

The discrepancies E_1 (in norm L_1), E_2 (in norm L_2), and E_∞ (in norm L_∞) between $P_{N_{\text{disc}},\, N_{\text{iter}}}(x^j)$ and the Lebesgue measure, which is the invariant measure associated with the symmetric tent map, are defined by

$$E_{1,N_{\text{disc}},N_{\text{iter}}}(x^j) = \left\| P_{N_{\text{disc}},\,N_{\text{iter}}}(x^j) - 1 \right\|_{L_1} \tag{1.14}$$

$$E_{2,N_{\text{disc}},N_{\text{iter}}}(x^j) = \left\| P_{N_{\text{disc}},\,N_{\text{iter}}}(x^j) - 1 \right\|_{L_2} \tag{1.15}$$

$$E_{\infty,N_{\text{disc}},N_{\text{iter}}}(x^j) = \left\| P_{N_{\text{disc}},\,N_{\text{iter}}}(x^j) - 1 \right\|_{L_\infty} \tag{1.16}$$

As mentioned in earlier section, Fig. 1.1 shows the convergence of the density of iterates of the components of two-coupled symmetric tent maps to the Lebesgue measure when ε_1 converges towards 0. Moreover, for a fixed value of N_{disc} when the number N_{iter} increases, the discrepancy between $P_{N_{\text{disc}},\,N_{\text{iter}}}(x^j)$ and the Lebesgue measure is expected to converge towards 0, except if there exist periodic orbits of finite length lower than N_{iter} which captures the iterates. In this case whatsoever the value of N_{iter} is, the approximated distribution function converges to the distribution function of the periodic orbit, if it is unique, or to the average of the distribution functions of the periodic orbits observed, if not.

Figure 1.2 shows the errors $E_{1,N_{\text{disc}},N_{\text{iter}}}(x^1)$ versus the number of iterates of the approximated distribution functions, with respect to the first variable x^1, for two- and three-coupled symmetric tent maps. Same results are obtained for the other variables x^2 or x^3.

The three-coupled symmetric tent maps model considered here with very small value of ε_1, seems a sterling model of generator of chaotic numbers with a uniform distribution of these numbers over the interval J. It produces very long periodic orbits: Gigaperiodic orbits (i.e., with length of period between 10^9 and 10^{12}) when

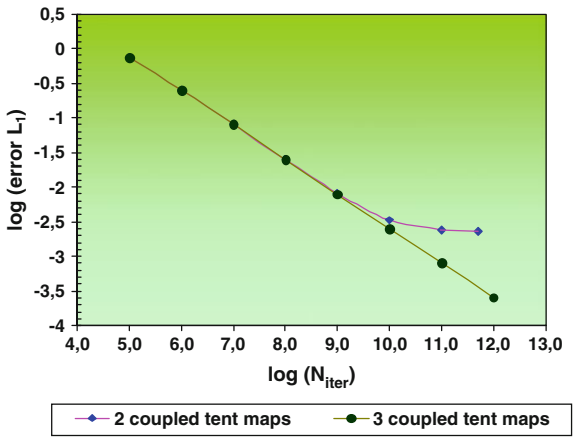

Fig. 1.2 Error $E_{1,N_{\text{disc}},N_{\text{iter}}}(x^1)$ for two- and three-coupled symmetric tent map, double precision, $N_{\text{disc}} = 10^5$, $\varepsilon_1 = 10^{-14}$, $\varepsilon_2 = 2\varepsilon_1$, $N_{\text{iter}} = 10^5$–10^{12}. Initial values $x_0^1 = 0.330$, $x_0^2 = 0.3387564$, $x_0^3 = 0.331353429$

computed with simple precision numbers, and orbits of unknown length when computed with double-precision numbers. However, these chaotic sequences are not at all random sequences.

1.3 The Route from Chaos to Pseudo-Randomness via Chaotic Undersampling

Chaotic numbers are not pseudo-random numbers, because the plot of the couples of any component $\left(x_n^l, x_{n+1}^l\right)$ of iterated points (X_n, X_{n+1}) in the corresponding phase plane reveals the map f used as one-dimensional dynamical systems to generate them via Eq. (1.6). Nevertheless, we have recently introduced a family of enhanced chaotic pseudo-random number generators (CPRNG) in order to compute faster a long series of pseudo-random numbers with a desktop computer [14, 15]. This family is based on the previous ultra-weak coupling which is improved in order to conceal the chaotic genuine function.

In this section, we describe briefly how this first process of undersampling, the chaotic one, works.

1.3.1 Chaotic Undersampling

In order to hide f in the phase space $\left(x_n^l,\ x_{n+1}^l\right)$, two mechanisms are used. The pivotal idea of the first mechanism is to sample chaotically the sequence $\left(x_0^l, x_1^l, x_2^l, \ldots, x_n^l, x_{n+1}^l, \ldots\right)$ generated by the lth component x^l, selecting x_n^l every time the value x_n^m of the mth component x^m, is strictly greater (or smaller) than a threshold $T \in J$, with $l \neq m$, for $1 \leq l, m \leq p$.

That is to say to extract the subsequence $\left(x_{n_{(0)}}^l, x_{n_{(1)}}^l, x_{n_{(2)}}^l, \ldots, x_{n_{(q)}}^l, x_{n_{(q+1)}}^l, \ldots\right)$ denoted here $a_l a_b\left(\overline{x_0}, \overline{x_1}, \overline{x_2}, \ldots, \overline{x_q}, \overline{x_{q+1}}, \ldots\right)$ of the original one, in the following way:Given that $1 \leq l,\ m \leq p,\ l \neq m$

$$\begin{cases} n_{(-1)} = -1 \\ \overline{x_q} = x_{n_{(q)}}^l, \text{with } n_{(q)} = \min\limits_{r \in \mathbb{N}} \left\{ r > n_{(q-1)} \middle| x_r^m > T \right\} \end{cases} \tag{1.17}$$

The sequence $\left(\overline{x_0}, \overline{x_1}, \overline{x_2}, \ldots, \overline{x_q}, \overline{x_{q+1}}, \ldots\right)$ is then the sequence of chaotic pseudo-random numbers.

The above mathematical formula can be best understood in algorithmic way. The pseudo-code, for computing iterates of (1.17) corresponding to N iterates of (1.6) is:

$$X_0 = \left(x_0^1, x_0^2, \ldots, x_0^{p-1}, x_0^p\right) = seed$$
$$n = 0;\ q = 0;$$
do { while $n < N$
do { while $\left(x_n^m \le T\right)$ compute $\left(x_n^1, x_n^2, \ldots, x_n^{p-1}, x_n^p\right)$; $n{+}{+}$}
compute $\left(x_n^1, x_n^2, \ldots, x_n^{p-1}, x_n^p\right)$; then $n(q) = n$; $\overline{x_q} = x^1_{n_{(q)}}$; $n{+}{+}$; $q{+}{+}$}

This chaotic sampling is possible due to the independence of each component of the iterated points X_n versus the others [13].

Remark 2.1 Albeit the number $N\text{Sampl}_{\text{iter}}$ of pseudo-random numbers $\overline{x_q}$ corresponding to the computation of N iterates is not known a priori, considering that the selecting process is again linked to the uniform distribution of the iterates of the tent map on J, this number is equivalent to $\frac{2N}{1-T}$.

1.3.2 Chaotic Mixing

A second mechanism can improve the unpredictability of the pseudo-random sequence generated as above, using synergistically all the components of the vector X_n, instead of two. Given $p - 1$ thresholds

$$T_1 < T_2 < \cdots < T_{p-1} \in J \tag{1.18}$$

and the corresponding partition $J_0 = [-1, T_1]$, $J_1 =]T_1, T_2[$, $J_k = [T_k, T_{k+1}[$ for $1 < k < p-1$, and $J_{p-1} = \left[T_{p-1}, 1\right[$, with

$$J = \bigcup_{k=0}^{p-1} J_k \tag{1.19}$$

(note that this partition of J is different from the regular previous one (1.11) used for the approximated distribution function).

The simple second mechanism is based on the chaotic undersampling combined with a chaotic mixing of the $p - 1$ sequences $\left(x_0^1, x_1^1, x_2^1, \ldots, x_n^1, x_{n+1}^1, \ldots\right)$, $\left(x_0^2, x_1^2, x_2^2, \ldots, x_n^2, x_{n+1}^2, \ldots\right)$,..., $\left(x_0^{p-1}, x_1^{p-1}, x_2^{p-1}, \ldots, x_n^{p-1}, x_{n+1}^{p-1}, \ldots\right)$,... using the last one $\left(x_0^p, x_1^p, x_2^p, \ldots, x_n^p, x_{n+1}^p, \ldots\right)$ in order to distribute the iterated points with respect to this given partition, defining the subsequence $\left(\overline{x_0}, \overline{x_1}, \overline{x_2}, \ldots, \overline{x_q}, \overline{x_{q+1}}, \ldots\right)$ (in pseudo-code) by

$$
\begin{aligned}
&X_0 = \left(x_0^1, x_0^2, \ldots, x_0^{p-1}, x_0^p\right) = seed\\
&n = 0;\ q = 0;\\
&\text{do } \{ \text{ while } n < N\\
&\quad \text{do } \{\text{while } \left(x_n^p \in J_0\right) \text{compute} \left(x_n^1, x_n^2, \ldots, x_n^{p-1}, x_n^p\right); n++\}\\
&\quad \text{compute } \left(x_n^1, x_n^2, \ldots, x_n^{p-1}, x_n^p\right)\\
&\quad \text{let } k \text{ be such that } x_n^p \in J_k \text{; then } n(q) = n;\quad \overline{x_q} = x_{n_{(q)}}^k;\ n++;\ q++\}
\end{aligned}
$$

Remark 2.2 In this case also, $N\text{Sampl}_{\text{iter}}$ is not known a priori, however, considering that the selecting process is linked to the uniform distribution of the iterates of the tent map on J, one has $N\text{Sampl}_{\text{iter}} \approx \frac{2N}{1-T_1}$.

Remark 2.3 This second mechanism is more or less linked to the whitening process [28, 29].

Remark 2.4 Actually, one can choose any of the components in order to sample and mix the sequence, not only the last one.

1.3.3 Enhanced Chaotic Undersampling

On can eventually improve the CPRG previously introduced with respect to the infinity norm instead of the L_1 or L_2 norms because the L_∞ norm is more sensitive than the others to reveal the concealed f [14]. For this purpose we introduce a second kind of threshold $T' \in \mathbb{N}$, together with $T_1, \ldots, T_{p-1} \in J$ such that the subsequence $\left(\overline{x_0}, \overline{x_1}, \overline{x_2}, \ldots, \overline{x_q}, \overline{x_{q+1}}, \ldots\right)$ is defined (in pseudo-code) by

$$
\begin{aligned}
&X_0 = \left(x_0^1, x_0^2, \ldots, x_0^{p-1}, x_0^p\right) = seed\\
&n = 0, q = 0;\\
&\text{do } \{ \text{ while } n < N\\
&\quad \text{do } \{\text{while } \left(n \le n_{(q-1)} + T' \quad and \quad x_n^p \in J_0\right)\\
&\qquad \text{compute } \left(x_n^1, x_n^2, \ldots, x_n^{p-1}, x_n^p\right); n++\}\\
&\quad \text{compute } \left(x_n^1, x_n^2, \ldots, x_n^{p-1}, x_n^p\right)\\
&\quad \text{let } k \text{ be such that } x_n^p \in J_k\\
&\quad \text{then } n(q) = n;\quad \overline{x_q} = x_{n_{(q)}}^k;\ n++;\ q++\}
\end{aligned}
$$

Remark 2.5 In this case also, $N\text{Sampl}_{\text{iter}}$ is not known a priori, it is very complicated to give an equivalent to it. However, considering that the selecting process is

linked to the uniform distribution of the iterates of the tent map on J, and to the second threshold T', it comes to $N\text{Sampl}_{\text{iter}} \leq \min\left\{\frac{2N}{1-T_1}, \frac{N}{T'}\right\}$.

Remark 2.6 The second kind of threshold T' can also be used with only the chaotic sampling, without the chaotic mixing.

1.3.4 A Window of Emergence of Randomness

In [15, 16], we show that if one consider the errors $E_{1,N_{\text{disc}},N_{\text{iter}}}(x) = \left\|P_{N_{\text{disc}},\,N_{\text{iter}}}(x) - 1\right\|_{L_1}$, $E_{2,N_{\text{disc}},N_{\text{iter}}}(x) = \left\|P_{N_{\text{disc}},\,N_{\text{iter}}}(x) - 1\right\|_{L_2}$, and $E_{\infty,N_{\text{disc}},N_{\text{iter}}}(x) = \left\|P_{N_{\text{disc}},\,N_{\text{iter}}}(x) - 1\right\|_{L_\infty}$ together with the correlated distribution functions which assess the independence of each component of the iterated points X_n versus the others, a window of emergence comes clearly into sight for the values $\varepsilon_1 \in \left[10^{-15}, 10^{-7}\right]$, in the case $p = 4$ and $\varepsilon_{i,j} = \varepsilon_i = i\,\varepsilon_1$. We have also performed NIST test developed by the National Institute of Standards and Technology [23], in order to check carefully the random nature of such numbers [7].

Then there is a route from chaos to randomness using the process of chaotic undersampling.

1.4 Geometric Undersampling

The previous route from chaos to randomness uses chaotic undersampling. It is possible to couple in another way p tent maps on the torus $J^p = [-1, 1]^p \subset \mathbb{R}^p$, which can directly provide random numbers without sampling or mixing, provided p is large enough, although it is possible to combine these processes with it. After reviewing this ring coupling in high dimension, we introduce the new geometric undersampling in order to obtain randomness with small values of p (for example $p = 2$).

1.4.1 Pseudo-Random Numbers Generated by Ring-Coupled Mapping

Consider the mapping defined on the p-dimensional torus $M_p : J^p \to J^p$

$$M_p \begin{pmatrix} x_n^1 \\ x_n^2 \\ \vdots \\ x_n^p \end{pmatrix} = \begin{pmatrix} x_{n+1}^1 \\ x_{n+1}^2 \\ \vdots \\ x_{n+1}^p \end{pmatrix} = \begin{pmatrix} 1 - 2\left|x_n^1\right| + k_1 \times x_n^2 \\ 1 - 2\left|x_n^2\right| + k_2 \times x_n^3 \\ \vdots \\ 1 - 2\left|x_n^p\right| + k_p \times x_n^1 \end{pmatrix} \tag{1.20}$$

with the parameters $k_i \in \{-1, 1\}$. In order to confine every variable x_n^j on J^p, we do, for every iteration, the transform

$$\begin{cases} \text{if } (x_{n+1}^j < -1) & \text{add} \quad 2 \\ \text{if } (x_{n+1}^j > 1) & \text{substract} \quad 2 \end{cases} \tag{1.21}$$

The particularity of this coupling is that each variable x^j is coupled only with itself and x^{j+1}, as displayed on Fig. 1.3a. At first glance, in order to enrich the random properties of the map, it could seem interesting to add supplementary cross couplings between these variables, as shown on Fig. 1.3b. However, in this case a cross coupling is inappropriate because it would increase the determinism against randomness, and therefore deteriorate the statistical properties which we are looking for.

To evaluate the random properties of these generators, the set of NIST tests have been used again.

The random properties validations of both a four-dimensional system and a ten-dimensional one have been carried out [5]. For this purpose, the chaotic carrier output needs to be quantized and binarized (0 and 1) in order to be validated as being random using NIST tests. Therefore, different methods of binarization (converting real signals into binary ones) have been implemented and compared.

A first 1-bit binarization has been applied to the system (1.21) output, defined as $y_n = x_n^j$ with $j \in [\![1, p]\!]$

$$\begin{cases} \text{if } (y_n \geq 0) & b = 1 \\ \text{else} & b = 0 \end{cases} \tag{1.22}$$

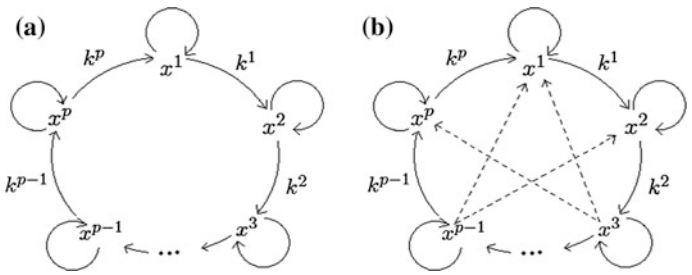

Fig. 1.3 **a** *Left* Ring coupling between the variables x^j. **b** *Right* Cross coupling between the variables x^j

RESULTS FOR THE UNIFORMITY OF P-VALUES AND THE PROPORTION OF PASSING SEQUENCES

generator is <data/lozi_10_positif.txt>

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	P-VALUE	PROPORTION	STATISTICAL TEST
8	5	13	9	9	12	6	19	8	11	0.102526	96/100	Frequency
11	16	9	10	10	10	14	6	8	6	0.437274	99/100	BlockFrequency
11	5	8	11	10	5	11	11	13	15	0.419021	97/100	CumulativeSums
8	6	17	10	10	6	7	11	15	10	0.213309	97/100	CumulativeSums
5	8	17	15	6	8	6	14	10	11	0.075719	99/100	Runs
11	11	10	13	9	5	8	8	15	10	0.637119	99/100	LongestRun
6	8	17	14	10	8	9	15	7	6	0.122325	99/100	Rank
9	10	9	13	10	10	9	8	12	10	0.991468	99/100	FFT
14	15	8	10	14	10	11	9	4	5	0.191687	98/100	NonOverlappingTemplate
10	8	11	9	9	13	7	12	10	11	0.964295	99/100	OverlappingTemplate
13	16	6	8	7	10	13	10	8	9	0.455937	100/100	Universal
9	10	12	8	10	11	5	14	11	10	0.816537	97/100	ApproximateEntropy
6	5	6	5	9	11	5	6	8	5	0.637119	65/66	RandomExcursions
3	5	6	7	10	10	9	6	4	6	0.407091	65/66	RandomExcursionsVariant
3	8	8	12	12	9	13	8	13	14	0.319084	100/100	Serial|
4	3	12	18	12	8	8	14	9	12	0.028817	100/100	LinearComplexity

Fig. 1.4 Example of NIST test for $k^i = (-1)^{i+1}$, $i = 1, 4$, each sequence of components satisfies the NIST test for randomness

The results showed to be highly sensitive to the type of binarization. Eventually, after testing several different methods, a 32-bit binarization has been chosen as being the most suitable solution. Because the system is confined to the p-dimensional torus J^p, 31 bits are assigned to represent the decimal part and 1 bit to the sign. To illustrate the results, the NIST tests for the four-dimensional system with parameters $k_i \in (-1)^{i+1}$ are shown in Fig. 1.4. The chosen conditions are: length of the original sequence = 10^8 bits, length of bit string = 10^6 bits, quantity of bit strings = 100. The output of the system has been arbitrary chosen as $y = x_n^4$.

Furthermore, as the results show their independence from the initial conditions, every bit string in this test is the resulting sequence of a different randomly chosen initial condition. The criterion for a successful test is that the p-value has to be superior to the significance level (0.01 for this case). For the present model, all tests were successful; thus the sequences can be accepted as being random.

1.4.2 Ring Coupling of Two-Dimensional Symmetric Tent Map

Although the system (1.20) is a good PRNG when $p \geq 4$, in lower dimension two and three, the chaotic numbers are not equidistributed on the torus (see Fig. 1.5).

In order to improve the ring-coupling mechanism in low dimension, we introduce now a new type of undersampling based on geometric nature of the invariant measure. We present this new mechanism which allows the emergence of randomness from chaos, in the simplest case, the two-dimensional ring mapping M_2 on the square J^2, with $k^1 = k^2 = 1$.

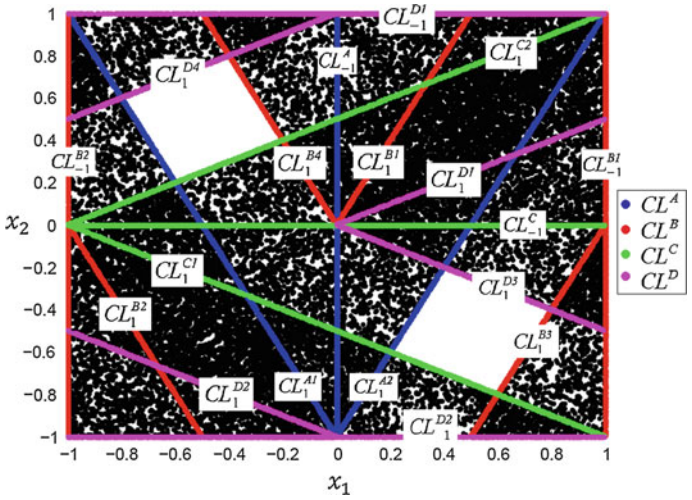

Fig. 1.5 Critical lines of the map M_2 on the torus J^2 (*a square*) [26]

Let M_2 be defined by

$$\begin{cases} x_{n+1}^1 = 1 - 2|x_n^1| + x_n^2 \\ x_{n+1}^2 = 1 - 2|x_n^2| + x_n^1 \end{cases} \tag{1.23}$$

$$\text{with} \quad \begin{cases} \text{if } (x_{n+1}^j < -1) & \text{add 2} \\ \text{if } (x_{n+1}^j > 1) & \text{substract 2} \end{cases} \tag{1.24}$$

1.4.2.1 Critical Lines

Figure 1.5 shows the distribution of the iterates of system (1.23) (the transient of the first 10^6 iterations has been cut off). It can be observed that the attractor contains regions where the point density is lower, and two lozenge-like holes. It is possible to define critical lines which form a partition of the square J^2. The critical lines CL [19] are singularities of dimension one and represent an important tool for the analysis of noninvertible maps. The holes on Fig. 1.5 are completely delimited by segments of the critical lines CL_1^{A1}, CL_1^{B4}, CL_1^{C2}, CL_1^{D4}, and CL_1^{A2}, CL_1^{B3}, CL_1^{C1}, CL_1^{D3}, defined below.

The critical lines separate regions of the phase space with different number of preimages (backward iterates). In the case of piecewise linear maps, they are the first iterates of the lines of discontinuity CL_{-1} of the system.

For the two-dimensional system (1.23), there are four groups of critical lines CL with preimages CL_{-1} given by

Critical lines A: CL_{-1}^{A}: $x^1 = 0$

$$\text{and} \begin{cases} CL_1^{A1} : x^2 = -2x^1 - 1 & \text{if } x^2 > 0 \\ CL_1^{A2} : x^2 = 2x^1 - 1 & \text{if } x^2 < 0 \end{cases} \tag{1.25}$$

Critical lines B: CL_{-1}^{B}: $x^1 = -1$

$$\text{and} \begin{cases} CL_1^{B1} : x^2 = 2x^1 & \text{if } x^2 < 0, x^1 \in [0, 0.5] \\ CL_1^{B2} : x^2 = -2x^1 - 2 & \text{if } x^2 > 0, x^1 \in [-1, -0.5] \\ CL_1^{B3} : x^2 = 2x^1 - 2 & \text{if } x^2 < 0, x^1 \in [0.5, 1] \\ CL_1^{B4} : x^2 = -2x^1 & \text{if } x^2 > 0, x^1 \in [-0.5, 0] \end{cases} \tag{1.26}$$

Critical lines C: CL_{-1}^{C}: $x^2 = 0$

$$\text{and} \begin{cases} CL_1^{C1} : x^2 = -\frac{1}{2}\left(x^1 + 1\right) & \text{if } x^1 > 0 \\ CL_1^{C2} : x^2 = \frac{1}{2}\left(x^1 + 1\right) & \text{if } x^1 < 0 \end{cases} \tag{1.27}$$

Critical lines D: CL_{-1}^{B}: $x^2 = -1$

$$\text{and} \begin{cases} CL_1^{D1} : x^2 = \frac{x^1}{2} & \text{if } x^1 < 0,\ x^2 \in [0, 0.5] \\ CL_1^{D2} : x^2 = -\frac{x^1}{2} - 1 & \text{if } x^1 > 0,\ x^2 \in [-1, -0.5] \\ CL_1^{D3} : x^2 = -\frac{x^1}{2} & \text{if } x^1 > 0,\ x^2 \in [-0.5, 0] \\ CL_1^{D4} : x^2 = \frac{x^1}{2} + 1 & \text{if } x^1 < 0,\ x^2 \in [0.5, 1] \end{cases} \tag{1.28}$$

1.4.2.2 Markov Partition of the Square

Our aim is first to use the partition defined by these critical lines in order to do a cell-to-cell analysis and, by the means of a Markov process, to compute explicitly the invariant measure of iterates associated to system (1.23). Figure 1.6 displays the 32 subregions of the square J^2, labeled from a to p and a′ to p′. For clarity of the presentation, we have labeled from (I) to (IV), the four quadrants of J^2.

Straightforward computation shows that the images of each region, by the mapping M_2, is one, two, or three regions of the same partition of the square J^2. Figure 1.7a, b display the images of the regions imbedded in the first quadrant (I). Figures 1.8a, b display the images of the regions imbedded in the second quadrant (II). The color is the same for every region and its corresponding image, except when two regions are mapped on the same region, in this case there is a mix of colors on the common part of the image.

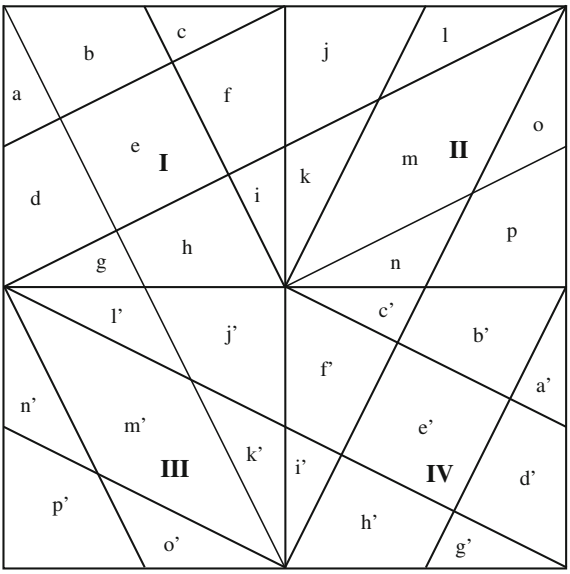

Fig. 1.6 The 32 subregions for a partition of the square J^2

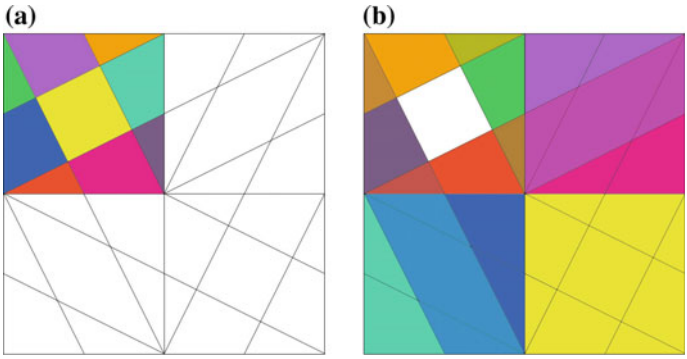

Fig. 1.7 **a** *Left* The nine regions *a*–*i* of quadrant (I). **b** *Right* The images M_2(I) of the nine regions of quadrant (I)

The overall correspondence between regions of the partition and their image is given by the Markov matrix M_a which is shown in Table 1.1. The computation of the coefficients of this matrix, which are rational numbers, is based on the ratios of surfaces of bounded regions.

In order to display the 32 × 32 matrix M_a on one page, we have labeled the coefficients using letters which are not related to the names of the regions.

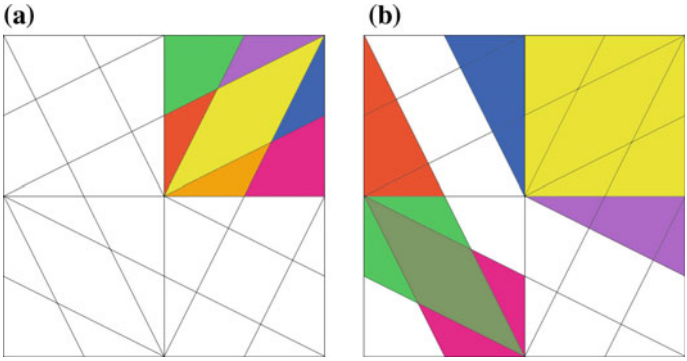

Fig. 1.8 **a** *Left* The seven regions *j*–*p* of quadrant (II). **b** *Right* The images M_2(II) of the seven regions of quadrant (II)

$$\begin{cases} o = \frac{1}{12}; p = \frac{1}{20}; q = \frac{3}{20}; r = \frac{4}{20}; \\ s = \frac{1}{9}; t = \frac{2}{9}; u = \frac{4}{9}; v = \frac{1}{6}; \\ w = \frac{1}{3}; x = \frac{1}{5}; y = \frac{3}{5}; z = \frac{2}{3}. \end{cases} \tag{1.29}$$

1.4.2.3 Exact Computation of Invariant Measure Associated to M_2

With the help of Markov matrix M_a, it is straightforward to compute explicitly the invariant measure associated to M_2. For every region on Fig. 1.6, we define a quantity of initial points called Q^i, i = 1, 32 uniformly scattered on it, and we compute its surface S_i. We normalize both quantities to $\sum_i Q^i = |Q| = 4$, and $\sum_i S_i = |S| = 4$. Hence it is possible to define the density of iterates on each region.

$$d^i = \frac{Q^i}{S_i} \tag{1.30}$$

Let $Q = \begin{pmatrix} Q^1 \\ \vdots \\ Q^{32} \end{pmatrix}$ and $D = \begin{pmatrix} d^1 \\ \vdots \\ d^{32} \end{pmatrix}$ be the vectors of quantities and densities obtained applying (1.30) to every region. Then starting from an arbitrary initial repartition of points on J^2, say $Q_0 = \begin{pmatrix} Q_0^1 \\ \vdots \\ Q_0^{32} \end{pmatrix}$, and applying repeatedly the equation

$$Q_{m+1} = M_a^t Q_m \tag{1.31}$$

Table 1.1 Markov matrix M_a

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	a'	b'	c'	d'	e'	f'	g'	h'	i'	j'	k'	l'	m'	n'	o'	p'
a				x		y			x																							
b										t	s	s	u		s																	
c	x	y	x																													
d																										t	s	s	u		s	
e																	p	q	p	q	r	q	p	q	p							
f																												s	u	s	s	t
g							x	y	x																							
h											s		u	s	s	t																
i	x			y			x																									
j																												v	z	v		
k	x			y			x																									
l																	x	y	x													
m										v	o	o	w	o	o	v																
n																							x	y	x							
o			x			y			x																							
p																											v		z		v	
a'																	x			y			x									
b'																											s		u	s	s	t
c'																							x	y	x							
d'										t	s	s	u	s																		
e'	p	q	p	q	r	q	p	q	p																							
f'												s	u	s	s	t																
g'																	x	y	x													
h'																										t	s	s	u	s		
i'																			x			y			x							
j'											v		z	v																		
k'																			x			y			x							
l'							x	y	x																							
m'																										v	o	o	w	o	o	v
n'																	x			y			x									
o'	x	y	x																													
p'												v	z		v																	

The sequence of vectors $\{Q_m\}_{m\in\mathbb{N}}$ converges to a limit vector $\overline{Q}$ which satisfies

$$\overline{Q} = M_a^t \overline{Q} \tag{1.32}$$

and gives the invariant measure, the density of which is the vector $\overline{D}$, using (1.30).

Numerical results

Starting from $Q_0 = \begin{pmatrix} Q_0^1 \\ \vdots \\ Q_0^{32} \end{pmatrix} = \begin{pmatrix} 1/8 \\ \vdots \\ 1/8 \end{pmatrix}, \overline{Q}$, it is obtained rapidly as

$$Q_{500} = \begin{pmatrix} Q_{500}^1 \\ Q_{500}^2 \\ Q_{500}^3 \\ Q_{500}^4 \\ \vdots \\ Q_{500}^{29} \\ Q_{500}^{30} \\ Q_{500}^{31} \\ Q_{500}^{32} \end{pmatrix} = \begin{pmatrix} 1/14 \\ 3/28 \\ 1/14 \\ 3/28 \\ \vdots \\ 4/7 \\ 3/28 \\ 3/28 \\ 1/7 \end{pmatrix} = \overline{Q}, \quad \text{which gives using (1.30),}$$

$$D_{500} = \begin{pmatrix} d_{500}^{1} \\ d_{500}^{2} \\ d_{500}^{3} \\ d_{500}^{4} \\ \vdots \\ d_{500}^{29} \\ d_{500}^{30} \\ d_{500}^{31} \\ d_{500}^{32} \end{pmatrix} = \begin{pmatrix} 10/7 \\ 5/7 \\ 10/7 \\ 5/7 \\ \vdots \\ 12/7 \\ 9/7 \\ 9/7 \\ 6/7 \end{pmatrix} = \overline{D}.$$

Remark 3.1 Computing directly this density and iterating (1.23) up to 10^{11} iterates, leads to the same result.

1.4.3 Geometric Undersampling

The exact computation of the density $\overline{D}$ of the invariant measure shows that this density is constant on each region. The geometric undersampling process consists of magnifying a square G included in one region (as for example the square $G = [0.36,\ 0.64] \times [0.36,\ 0.64]$ included in region m on Fig. 1.9), up to the size of the square J^2.

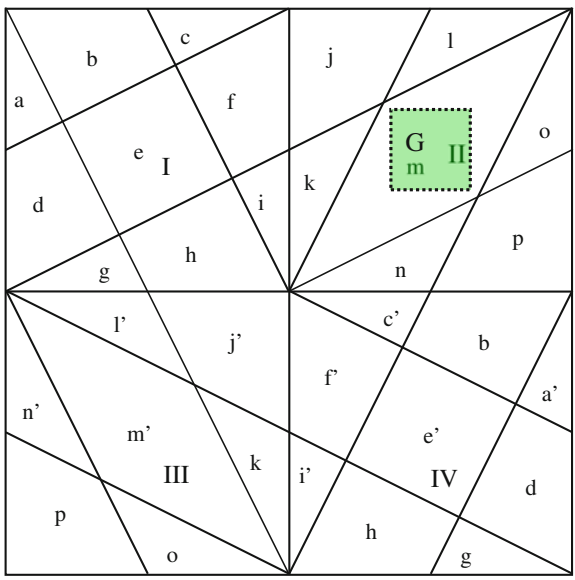

Fig. 1.9 The square $G = [0.36, 0.64] \times [0.36, 0.64]$ in which iterates of (1.23) are geometrically undersampled

1.4.3.1 Algorithm of Geometric Undersampling

Let $G = \left[x_l^1, x_r^1\right] \times \left[x_l^2, x_r^2\right]$ be the square in which we will undersample the iterates of (1.23) and, $x^1_{\text{mean}} = \frac{x_l^1 + x_r^1}{2}$, $x^2_{\text{mean}} = \frac{x_l^2 + x_r^2}{2}$. In algorithmic form, the pseudo-code to geometric undersample N iterates of (1.23) is:

$X_0 = \left(x_0^l, x_0^2\right) = seed$
$n = 0;$
do { while $n < N$ compute $\left(x_n^l, x_n^2\right)$;if $\left(x_n^l, x_n^2\right) \in G$ then
$\overline{x_q^l} = 2\left[\frac{x_n^l - x_{mean}^l}{x_r^l - x_l^l}\right]$, $\overline{x_q^2} = 2\left[\frac{x_n^2 - x_{mean}^2}{x_r^2 - x_l^2}\right]$; q=q+1; $n++$}

Remark 3.2 In this case, the undersampling process provides two streams of pseudo-random numbers.

Remark 3.3 In this case, $N\text{Sampl}_{\text{iter}}$ the number of geometrically undersampled iterates is not known a priori, however, considering that the selecting process is linked to the uniform distribution of the iterates of the tent map on J^2, one has $N\text{Sampl}_{\text{iter}} \approx \frac{\left(x_r^1 - x_l^1\right)^2}{4} \times d^m$, where d^m is the density of the measure in region m.

1.4.3.2 Numerical Tests

We have applied this process in the case of the square G of Fig. 1.9 with $N = 10^{12}$, which gives $N\text{Sampl}_{\text{iter}} \approx 3.35 \times 10^{10}$. Figure 1.10a displays the densities of the seven regions j, k, l, m, n, o, p of quadrant (II) which are equal to

$$\begin{cases} \overline{d}^j = \frac{6}{7}; \overline{d}^k = \frac{9}{7}; \overline{d}^l = \frac{9}{7}; \overline{d}^m = \frac{12}{7}; \\ \overline{d}^n = \frac{9}{7}; \overline{d}^o = \frac{9}{7}; \overline{d}^p = \frac{6}{7}; \end{cases}$$

Figure 1.10b shows the uniform density of iterates in the square $G = [0.36,\ 0.64] \times [0.36,\ 0.64]$ of quadrant (II). In Fig. 1.11, the square is magnified up to the size of the square J^2. The vertical scale is fitted near the invariant Lebesgue measure.

We have also used NIST test to confirm the random property of the geometrical undersampling process. They are all successful (Fig. 1.12).

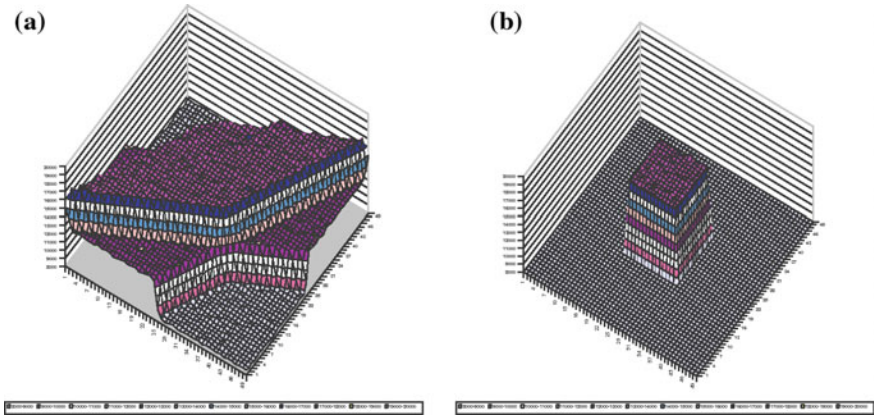

Fig. 1.10 **a** *Left* densities of the seven regions j, k, l, m, n, o, and p of quadrant (II). **b** *Right* Uniform density of iterates in the square $G = [0.36, 0.64] \times [0.36, 0.64]$ of quadrant (II)

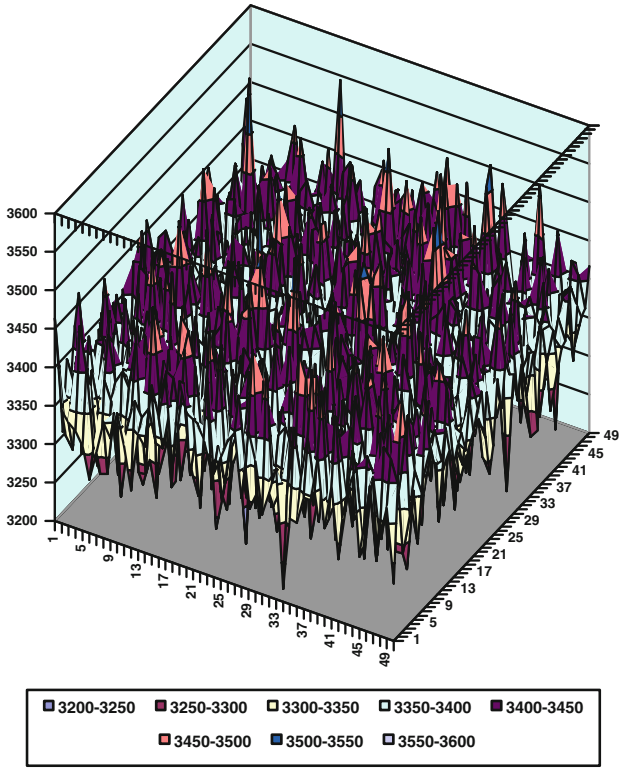

Fig. 1.11 Uniform density of iterates of the square $G = [0.36, 0.64] \times [0.36, 0.64]$ magnified to the square J^2

RESULTS FOR THE UNIFORMITY OF P-VALUES AND THE PROPORTION OF PASSING SEQUENCES

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	P-VALUE	PROPORTION	STATISTICAL TEST
12	9	7	9	12	11	8	11	13	8	0.924076	99/100	Frequency
1	4	3	4	7	5	9	16	16	35	0.000000 *	100/100	BlockFrequency
9	9	10	12	11	8	9	10	10	12	0.996335	99/100	CumulativeSums
10	9	12	12	9	7	10	10	9	12	0.983453	99/100	CumulativeSums
11	12	11	8	12	7	12	6	10	11	0.883171	99/100	Runs
9	9	13	8	9	8	17	8	10	9	0.595549	100/100	LongestRun
6	11	11	11	9	8	8	14	9	13	0.798139	100/100	Rank
15	10	7	8	8	8	15	16	7	6	0.153763	97/100	FFT
12	9	10	13	9	11	7	15	4	10	0.474986	98/100	NonOverlappingTemplate
12	6	10	6	13	6	8	8	17	14	0.145326	99/100	OverlappingTemplate
18	12	13	11	9	10	5	8	9	5	0.145326	99/100	Universal
11	8	12	11	11	14	8	10	7	8	0.883171	99/100	ApproximateEntropy
3	5	6	9	4	3	7	5	6	11	0.145326	59/59	RandomExcursions
7	6	6	2	6	7	6	7	4	8	0.637119	59/59	RandomExcursions
2	6	4	5	5	6	10	6	7	8	0.334538	59/59	RandomExcursionsVariant
8	15	13	12	9	12	13	5	9	4	0.224821	98/100	Serial
9	9	6	13	13	7	12	9	10	12	0.798139	99/100	LinearComplexity

The minimum pass rate for each statistical test with the exception of the random excursion (variant) test is approximately = 96 for a sample size = 100 binary sequences.

Fig. 1.12 Geometrical undersampling: each sequence of components satisfies the NIST test for randomness

1.5 Noise-Resisting Ciphering

As a first example, we propose a novel noise-resisting ciphering based on a large number of uncorrelated chaotic sequences. These cogenerated sequences are actually used in several steps of the ciphering process. Noisy transmission conditions are considered with realistic assumptions. The efficiency of the proposed method for ciphering and deciphering is illustrated through numerical simulations based on ten coupled chaotic sequences [5]. It can be also adapted to geometric undersampling, provided this undersampling is done in dimension four.

In this section, we detail the noise-resisting ciphered transmission principle, consisting of two steps: the ciphering process and the transmission process (see Figs. 1.13, 1.14). Both resort to the coupled chaotic pseudo-random generated sequences.

1.5.1 Ciphering Principle

We begin with some notations that will be used in the sequel. The *plain text* is denoted by $(t_k)_{k=1,\ldots,N}$: the letters t_k, for $k = 1, \ldots, N$ belong to the alphabet $\{l_1, \ldots, l_\pi\}$ composed of π letters.

The *ciphered text* is a sequence of real numbers denoted by y_k, $k = 1, \ldots, N$ and each y_k belongs to the interval $J = [-1, 1] \subset \mathbb{R}$. The transmitted signal (at the transmitter side) is denoted by s_n, while the received signal is $\hat{s}_n$ (at the receiver side).

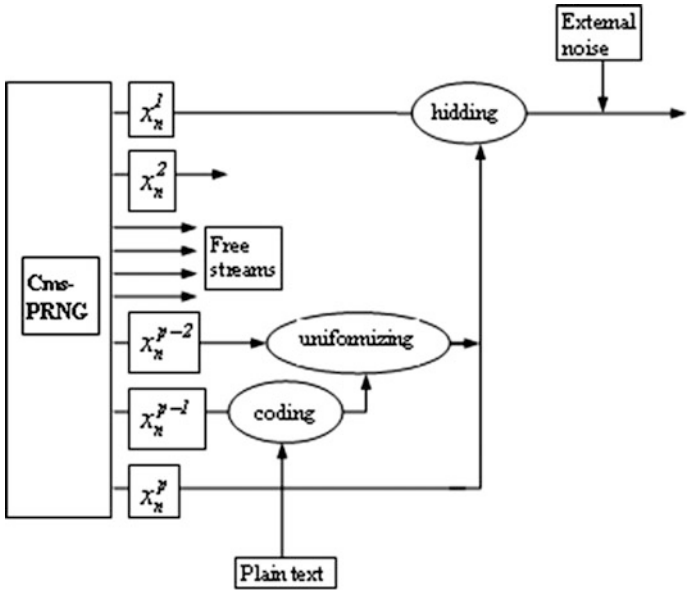

Fig. 1.13 General scheme of the ciphering and the ciphered transmission principle (coding and transmitting)

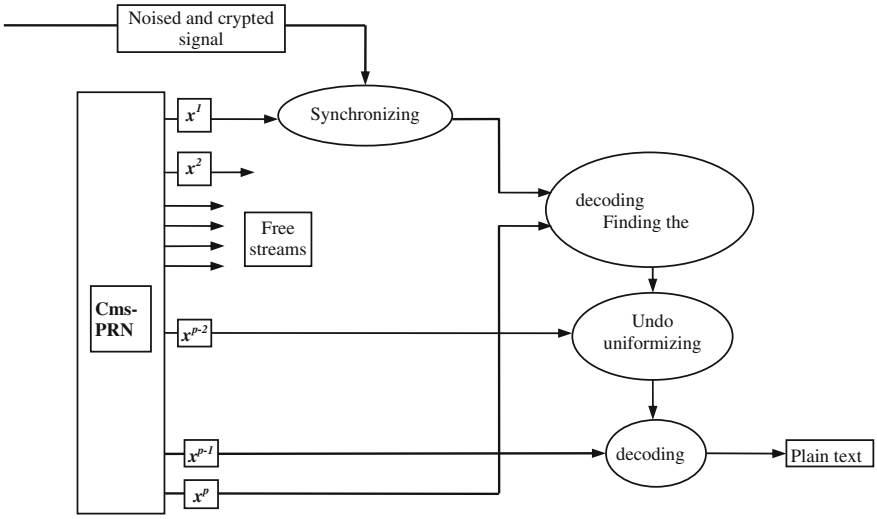

Fig. 1.14 General scheme of the ciphering and the ciphered transmission principle (receiving and decoding)

In this paper, we consider noisy transmission conditions, which means that $\hat{s}_n = s_n + \alpha_n$, where $\alpha_n > 0$ denotes an unknown additive noise at time n. We make the following classical assumption: the additive noise is bounded by a known bound K, which means that

$$\|s_n - \hat{s}_n\| = \alpha_n \leq K, \quad \forall n \geq 0 \tag{1.33}$$

We first detail how to transform each letter of the plain text t_k into a real number $y_k \in [-1, 1]$ with an original noise-resisting method. In the second step, the sequence $\{y_k\}$ will be transformed to obtain a uniform distribution on the interval $[-1, 1]$.

- Define a partition as follows:

$$[-1, 1] = \bigcup_{m=1,\pi} I_m \tag{1.34}$$

with a_m, b_m the bounds of each interval I_m, i.e., $I_m = [a_m, b_m]$.

In fact, owing to the presence of additive noise, not all real numbers inside I_m can be selected, one must add an interval of length K at each side of the interval I_m. Therefore some smaller intervals need to be defined.

- Define a subinterval I'_m to be included in the corresponding interval I_m such that

$$I'_m = \left[a'_m,\ b'_m\right] \subset I_m \tag{1.35}$$

and

$$\left[a'_m - K,\ b'_m + K\right] \subset I_m \tag{1.36}$$

where we recall that K is the upper bound of the noise, see (1.33).

Then the coding consists of random (i.e., with another pseudo-random sequence generated by (1.20): x_n^{p-1}, or the geometric undersampling in dimension four) choosing for each letter t_k of the plain text a real number y_k inside the interval I'_m (and not I_m) if $t_k = l_m$. Each interval I'_m corresponds to a letter l_m, for $m = 1, \ldots, \pi$. Remark that each letter has a frequency of apparition in the plain text, depending on the initial language. Therefore, one must carefully choose the length of each interval I'_m in proportion to the corresponding frequency of the letter l_m. An illustration is given by Fig. 1.15 for an alphabet with three letters: the letter A has a frequency of 10 %, the letter B of 30 %, and the letter C of 60 %.

- Once this first step of the coding is achieved, one has to ensure that the ciphered text has a random-like distribution inside $[-1, 1]$. With the aforementioned coding alone, this property cannot be ensured, as it can be seen in Fig. 1.16.

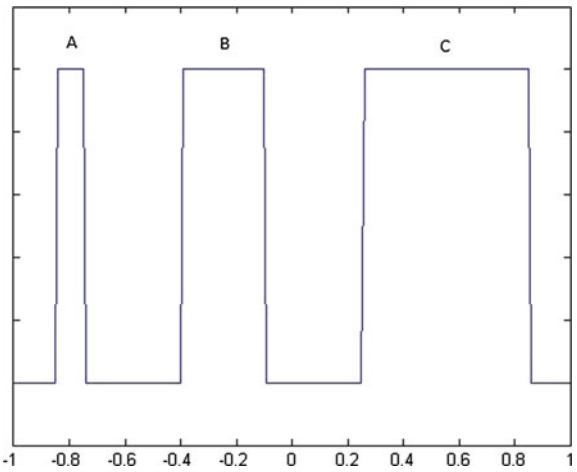

Fig. 1.15 Repartition of an alphabet of three letters

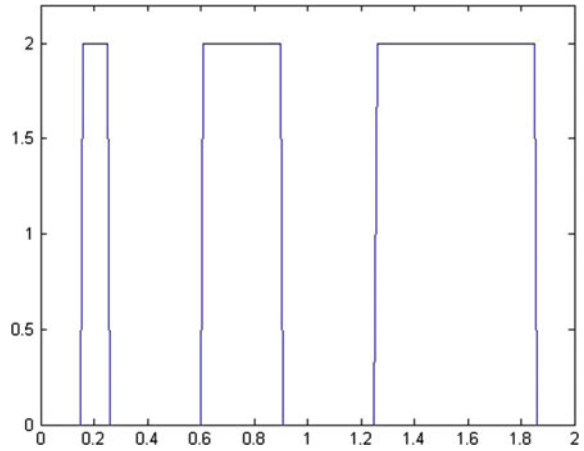

Fig. 1.16 Signal to be transmitted without transformation

Since one needs to leave some holes at the edges of the intervals I_m to resist the additive noise, the transmitted signal cannot have a random-like repartition. So we propose to transform the ciphered data y_k before transmitting it.

For all steps $n \in \mathbb{N}$ such that an encrypted letter is transmitted, we propose to transmit not directly y_n but:

$$\tilde{y}_n = \begin{cases} y_n + x_n^{p-2} & \text{if } y_n + x_n^{p-2} \in [-1, 1] \\ y_n + x_n^{p-2} + 2 & \text{if } y_n + x_n^{p-2} < -1 \\ y_n + x_n^{p-2} - 2 & \text{if } y_n + x_n^{p-2} > 1 \end{cases} \tag{1.37}$$

For simplicity of presentation, in the sequel, y_n will denote $\tilde{y}_n$, the ciphered message to transmit.

Then the obtained signal to transmit has the desired uniform repartition, as illustrated by Fig. 1.17.

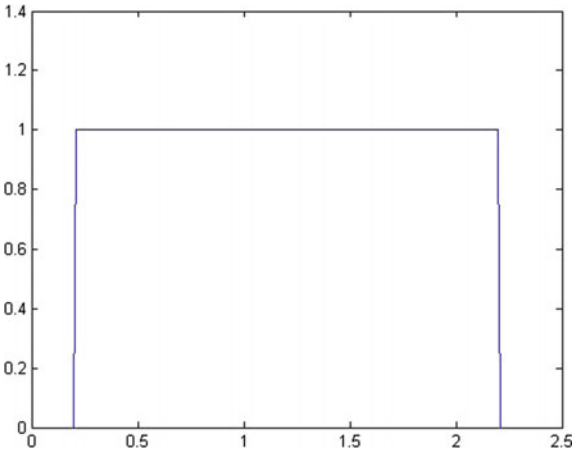

Fig. 1.17 Signal to be transmitted after transformation

1.5.2 Transmission Principle

We now present how to transmit the ciphered text using substitution method in a new pseudo-random sequence. The transmitted signal is denoted by s_n.

The ciphered text y_k, defined by (1.37), is not directly transmitted, it is chaotically hidden in a chaotic carrier signal as explained below.

The ciphering makes use of two coupled chaotic sequences: x_n^1 is used as chaotic carrier, while x_n^p is used to select the substitution times.

$$s_n = \begin{cases} x_n^1 & \text{if } x_n^p < T \\ y_{n(k)} & \text{if } x_n^p \geq T \end{cases} \tag{1.38}$$

where T is a predefined threshold. For example, as the x_n^p is equally distributed on the interval $[-1, 1]$, if one chooses $T = 0.8$, one ciphered letter will be transmitted in average of each ten elements of the sequence x_n^1. If one chooses $T = 0.98$, one element over 100 is replaced by a letter.

We do not detail here the sequence $k(n)$, as it is easily understandable that $k(n)$ increase by +1 each time $s_n = y_{k(n)}$ in order to transmit each element of the ciphered sequence y_k.

1.5.3 Decoding Principle

At the receiver end, suppose that the same PRNG defined by (1.20) is available. The transmitter and the authorized receiver have fixed the same parameters and same initial values; therefore the ciphering is a symmetrical one.

According to the substitution principle defined by (1.38) and the hypothesis (1.33) on the additive noise, the received signal can be expressed as

$$\hat{s}_n = x_n^1 + \alpha_n \text{ or } y_{k(n)} + \alpha_n \tag{1.39}$$

Since the initial conditions of the chaotic pseudo-random number generator (1.20) are assumed to be public, the receiver exactly knows when x_n^p is smaller or larger than the threshold T, so the receiver is able to reconstruct the sequence $\left(y_{k(n)} + \alpha_n\right)$ i.e., the sequence $y_q + \beta_q$, where $\beta_q = \alpha_n$ for $q = k(n)$.

As $\beta_q < K$, there exists $m \in \{1, 2, \ldots, \pi\}$, such that $\hat{s}_n \in I_m$.

The receiver, also, exactly knows the value of x_n^{p-2} and deduces from the rules (1.37) the value of y_q. Then the knowledge of the correspondence between the interval I_m and the letter l_m enables the receiver to retrieve the initial message.

1.5.4 Numerical Illustration

Now we summarize the main steps of the proposed algorithm:

(1) Choose the secret parameters $k_i = 1$ or $k_i = -1$, for $i \in \{1, 2, \ldots, p\}$.
(2) Define the initial conditions shared by the transmitter and the receiver.
(3) Iterate the PRNG (1.20) with the previous initial conditions, at both the transmitter and the receiver side.
(4) Apply the ciphering and transmission principle as detailed before.

The Fig. 1.18 shows the noisy signal at the receiver side (recall that the transmitted signal is given by Fig. 1.17). Notice that the Figs. 1.16, 1.17 and 1.18 represent our simulations with 10^9 iterations.

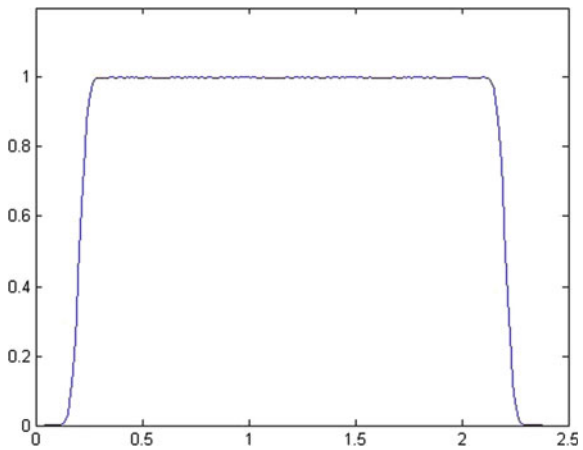

Fig. 1.18 Received noisy signal

1.6 Conclusion

We have proposed a new mechanism of undersampling of chaotic number obtained by the ring-coupling mechanism of one-dimensional maps. In the case of two coupled maps, this mechanism allows the building of a PRNG which passes all NIST tests.

This new geometric undersampling is very effective for generating two parallel streams of pseudo-random numbers, as we have shown, computing carefully their properties up to sequences of 10^{12} consecutives iterates of (1.23) which provides more than 3.35×10^{10} random numbers in very short time. In a forthcoming paper we will test both three- and four-dimensional cases.

In addition we have proposed a novel method of noise-resisting ciphering. The originality lies in the use of a chaotic pseudo-random number generator: several cogenerated sequences can be used at different steps of the ciphering process, as they present the strong property of being uncorrelated. Each letter of the initial alphabet of the plain text is encoded as a subinterval of $[-1, 1]$. The bounds of each interval are defined in function of the known bound of the additive noise. A pseudo-random sequence is used to enhance the complexity of the ciphering. The transmission consists of a substitution technique inside a chaotic carrier, depending on another cogenerated sequence. The efficiency of the proposed scheme is illustrated on some numerical simulations.

Cryptography is a wide field of research, in which the brilliant formulas of Srinivasan Ramanujan have been largely used. May be, it will be possible, in a near future, to link such formulas with chaos in the domain of emergent randomness.

References

1. Araujo E, Coelho LdS (2008) Particle swarm approaches using Lozi map chaotic sequences to fuzzy modelling of an experimental thermal-vacuum system. Appl Soft Comput 8:1354–1364
2. Ariffin MRK, Noorani MSM (2008) Modified Baptista type chaotic cryptosystem via matrix secret key. Phys Lett A 372:5427–5430
3. Baptista MS (1998) Cryptography with chaos. Phys Lett A 240:50–54
4. Cherrier E, Lozi R (2011) Noise-resisting ciphering based on a chaotic multi-stream pseudo-random number generator. In: Proceeding of ICITST, 2011, Abu-Dhabi, AUE, 11–14 Dec 2011
5. Espinel A, Taralova I, Lozi R (2013) New alternate ring-coupled map for multi-random number generation. J Nonlinear Syst Appl 4(3–4):64–69
6. Gora P, Boyarsky A, Islam MdS, Bahsoun W (2006) Absolutely continuous invariant measures that cannot be observed experimentally. SIAM J Appl Dyn Syst 5(1):84–90 (electronic)
7. Hénaff S, Taralova I, Lozi R (2009) Dynamical analysis of a new statistically highly performant deterministic function for chaotic signals generation. In: Proceeding of Physcon 2009, Catania, Italy, 1–4 Sept 2009 (IPACS open Access Electronic Library)
8. Hénaff S, Taralova I, Lozi R (2010) Exact and asymptotic synchronization of a new weakly coupled maps system. J Nonlinear Syst Appl 1(3–4):87–95

9. Hénon M (1976) A two-dimensional mapping with a strange attractor. Commun Math Phys 50:69–77
10. Lanford OE III (1998) Some informal remarks on the orbit structure of discrete approximations to chaotic maps. Exp Math 7(4):317–324
11. Lorenz EN (1963) Deterministic nonperiodic flow. J Atmos Sci 20:130–141
12. Lozi R (2006) Giga-periodic orbits for weakly coupled tent and logistic discretized maps. In: Siddiqi AH, Duff IS, Christensen O (eds.) Modern mathematical models, methods and algorithms for real world systems. Anamaya Publishers, New Delhi, India pp 80–124
13. Lozi R (2008) New enhanced chaotic number generators. Indian J Ind Appl Math 1(1):1–23
14. Lozi R (2009) Chaotic pseudo random number generators via ultra weak coupling of chaotic maps and double threshold sampling sequences. In: Proceedings of the 3rd international conference on complex systems and applications (ICCSA 2009), University of Le Havre, France, June 29–July 02, 2009, pp 20–24
15. Lozi R (2011) Complexity leads to randomness in chaotic systems. Mathematics in Science and technology: mathematical methods, models and algorithms in science and technology. In: Siddiqi AH, Singh RC, Manchanda P (eds) Proceedings of the satellite conference of ICM, 15–17 August 2010 Delhi, India, World Scientific Publisher, Singapore, pp 93–125
16. Lozi R (2012) Emergence of randomness from chaos. Int J Bifurc Chaos 22:2 1250021-1/1250021-15
17. Lozi R (2013) Chaotic mathematical circuitry. In: Adamatzky A, Chen G (eds) Chaos, CNN, memristors and beyond. World Scientific Publishing, Singapore, pp 307–323
18. Lozi R (2013) Can we trust in numerical computations of chaotic solutions of dynamical systems. In: Letellier Ch, Gilmore R (eds) Topology and dynamics of chaos, World Scientific Series in Nonlinear Science Series A, vol 84, pp 63–98
19. Mira C, Gardini L, Barugola A, Cathala J-C (1996) Chaotic dynamics in two-dimensional noninvertible maps, World Scientific Series on Nonlinear Science, Series A, vol 20
20. Palmer K (2000) Shadowing in dynamical systems: theory and applications. Kluwer Academic Publications, Dordrecht
21. Pilyugin SY 1999 Shadowing in dynamical systems. Lecture Notes in Mathematics, Springer, 1706, (1999)
22. Pluhacek M, Budikova V, Senkerik R, Oplatkova Z, Zelinka I (2012) Extended initial study on the performance of enhanced PSO algorithm with Lozi chaotic Map. In: Advances in intelligent systems and computing, vol 192, Nostradamus: modern methods of prediction, modeling and analysis of nonlinear systems, pp 167–177
23. Rukhin et al (2001) A statistical test suite for random and pseudorandom number generators for cryptographic applications, NIST http://csrc.nist.gov/rng/
24. Sprott JC (2003) Chaos and time-series analysis. Oxford University Press, Oxford
25. Tang TW, Allison A, Abbott D (2004) Parrondo's games with chaotic switching. In: Proceedings of the SPIE Noise in Complex Systems and Stochastic Dynamics II, vol 5471, pp 520–530, Maspalomas, Gran Canaria, Spain, 26–28 May 2004
26. Taralova I, Espinel A, Lozi R (2011) Dynamical and statistical analysis of a new Lozi function for random numbers generation. In: Proceeding of Physcon 2011, León, Spain, 5–8 Sept 2011 (IPACS open Access Electronic Library)
27. Taralova I, Hassad SEl, Lozi R (2012) Chaotic generator synthesis: dynamical and statistical analysis. In: Proceeding of ISTP 2012, London, UK, pp 56–59, 10–11 Dec 2012
28. Viega J (2003) Practical random number generation in software. In: Proceedings of 19th annual computer security applications conference, pp 129–140
29. Viega J, Messier M (2003) Secure programming cook book for C and C++. O'Reilly, Sebastopol

Chapter 2
Soft Computing Techniques and Their Applications

D.K. Chaturvedi

Abstract The modern science is still striving to develop consciousness-based machine. The forecasting is an intuition-based or consciousness-based problem. It is an important problem for planning, decision-making and designing of an appropriate controller for the systems. The paper deals with the synergism of soft computing techniques mainly artificial neural network, fuzzy logic systems, and genetic algorithms and their applications in forecasting.

Keywords Artificial neural network · Fuzzy systems · Genetic algorithms · Synergism of soft computing techniques · Forecasting

2.1 Introduction

In the last century, enormous industrial and technological developments had taken place. Technology had developed laterally well up to the biggest giant-sized complexes and also to the smallest molecular nano-mechanisms. Thus, having explored to the maxima of the two extreme fields, technology is exploring now vertically to reach the dizzy heights of soft computing, subtle soft computing, and the millennium wonder of reaching the almost unchartered height of evolving consciousness in computers (machines). This presentation makes its small and humble contribution to this new astounding scenario and possibly the greatest of all mechanical wonders, to transfer consciousness of man to machine [1]. Prior to World War II, numerical calculations were done with mechanical calculators. Simulated by military requirements during World War II, the first version modern digital computers began to make their appearance in late 1940s and early 1950s. During that pioneering period, a number of different approaches to digital computer organization and digital computing techniques were investigated. Primarily, as a result of the constraints imposed

D.K. Chaturvedi (✉)
Department of Electrical Engineering, D.E.I. (Deemed University), Dayalbagh, Agra
e-mail: dkc.foe@gmail.com
URL: http://www.works.bepress.com/dk_chaturvedi

A.H. Siddiqi et al. (eds.), *Mathematical Models, Methods and Applications*,
Industrial and Applied Mathematics, DOI 10.1007/978-981-287-973-8_2

by the available electronics technology, the designers of digital computers soon focused their attention on the concept of computer system architecture, which was championed by Dr. John Von Neumann, who first implemented it in the computer constructed for the Institute of Advanced Studies at Princeton. Because of the pervasiveness of the Von Neumann architecture in digital computers, during the 1950s and 1960s, most numerical analysts and other computer users concentrated their efforts on developing algorithms and software packages suitable to these types of computers. In 1960s and 1970s, there were numerous modifications and improvements to computers of the earlier generation. The "bottle neck" of Neumann computers was the memory buffer sizes and speeds on it. In the 1990s, there was a quantum leap in the size of computer memory and speeds. As a result of this, supercomputers have been developed, which could do lakhs of calculations within a fraction of a second. Supercomputers can also do all routine tasks, and it could handle it better with multi-coordination than a human being, and thus reducing a series of simple logical operations. It could store vast information and process the same in a flash. It does not also suffer from the human moods and many vagaries of mind.

But, the supercomputers cannot infer or acquire any knowledge from its information contents. It cannot think sensibly and talk intelligently. It could not recognize a person or could not relate his family background. On the other hand, as human beings, we continuously evolve our value judgment about the information we receive and instinctively process them. Our judgment is based on our feelings, tastes, knowledge, and experience. But computers are incapable of such judgments. A computer can be programmed (instructed), i.e., to generate poetry or music, but it cannot appraise or judge its quality.

Hence, there is a genuine and compulsory need for some other logic, which can handle such real-life scenario. In 1965, Prof. Lofti A. Zadeh at the University of California introduced an identification tool by which this degree of truth can be handled by fuzzy set theoretic approach. With the invention of fuzzy chips in 1980s fuzzy logic received a great boost in the industry.

Now in this twenty-first century fuzzy logic, artificial neural network (ANN), and evolutionary algorithms (EA) are receiving intensive attention in both academics and industry [1–15]. All these techniques are kept under one umbrella called "soft computing." Enormous research had already been done on soft computing techniques to identify a model and control of its different systems.

This paper deals with the synergism of soft computing techniques which are fuzzy logic, ANN, and EA for electrical load forecasting problem. The wavelet transform is used to decompose the past load pattern and used for training and testing of proposed method.

2.2 Wavelet Analysis

The underlying mathematical structure for wavelet bases of a function space is a multiscale decomposition of a signal, known as multi resolution or multiscale analysis. It is called the heart of wavelet analysis.

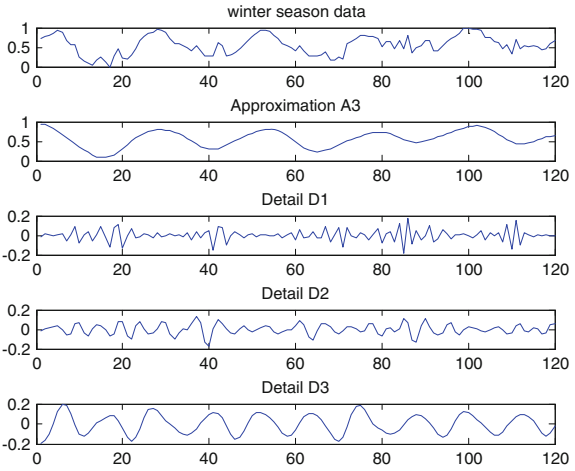

Fig. 2.1 Wavelet decomposition of hour load data into wavelet components

The first step of discrete wavelet transform corresponds to the mapping f to its wavelet coefficients and from these coefficients two components are received, namely a smooth version, named approximation and a second component that corresponds to the deviations or the so-called details of the signal. A decomposition of f into a low-frequency part a, and a high-frequency part d, is represented by $f = a_1 + d_1$. The same procedure is performed on a_1 in order to obtain decomposition in finer scales: $a_1 = a_2 + d_2$. A recursive decomposition for the low-frequency parts follows the directions that are illustrated in the following diagram.

$$\begin{array}{ccccccccc} f & ..... & a_1 & ..... & a_2 & ..... & a_3 & ..... & a_n \\ & \searrow & & \searrow & & \searrow & & \searrow & \\ & & d_1 & & d_2 & & d_3 & & d_4 ..... d_n \end{array}$$

The resulting low-frequency parts $a_1, a_2, \ldots a_N$ are approximations of f, and the high-frequency parts $d_1, d_2, \ldots d_N$ contain the details of f. Figure 2.1 illustrates a wavelet decomposition into four levels and corresponds to a_3, d_1, d_2, and d_3.

$$f = d_1 + d_2 + d_3 + \cdots + d_{N-1} + d_N + a_N.$$

2.3 Generalized Neural Network

In a simple neuron model the aggregation function is summation, which has been modified to obtain a generalized neuron network (GNN) model using fuzzy compensatory operators as aggregation operators to overcome the problems such as large number of neurons and layers required for complex function approximation, which affect not only the training time but also the fault tolerant capabilities of the artificial neural network (ANN) [2].

The common ANN is consisting of summation as aggregation function. As mentioned by Minsky and Parpet [16] in their book that linear perceptron could not be trained for non-separable problems. The multilayer ANN introduced to overcome the problems of perceptron and it was found that three-layer ANN could map any function. The three-layer ANN with simple back-propagation learning algorithm requires large training time. Then large number of back-propagation variants came up with time to improve its training performance. Basically, the training time of ANN depends on the number of unknown weights to be determined. This large number of unknown weights in huge ANN is required to map with complex functions. To obtain large number of weights, large number of training data is required. It is very difficult or sometimes impossible to collect accurate and sufficient training data for real-life problems. The noisy training data affect the testing performance of ANN.

The general structure of the common neuron is an aggregation function and its transformation through a filter. It is shown in the literature [4] that the ANNs can be universal function approximators for given input–output data. The common neuron structure has summation or product as the aggregation function with linear or nonlinear (sigmoid, radial basis, tangent hyperbolic, etc.) as the threshold function.

Different structures at neuron level have been tried to overcome above-mentioned drawbacks of ANN [1]. In this regard ANN consisting of Σ neurons (Σ-ANN), ANN consisting of Π neurons (Π-ANN), and combination of the above two have been tried and the results obtained are quite encouraging [1].

The proposed generalized neuron model shown in Fig. 2.2 has summation and product as aggregation and sigmoid and Gaussian as activation functions. The final

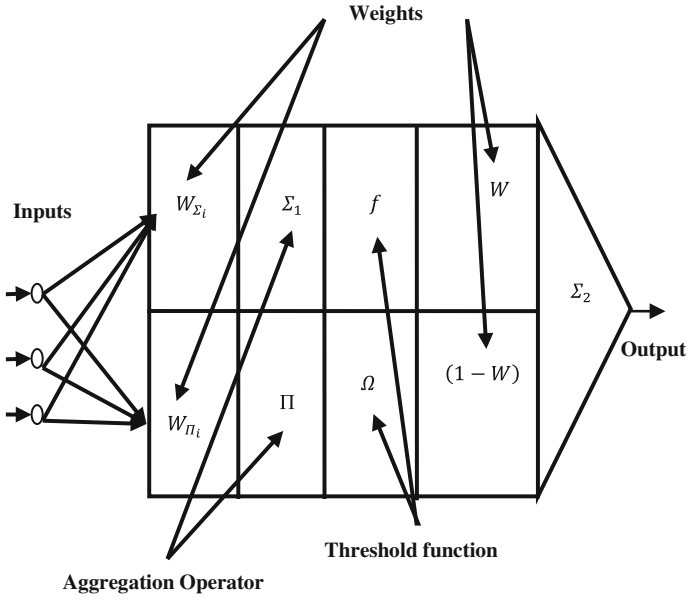

Fig. 2.2 Generalized neural network (GNN)

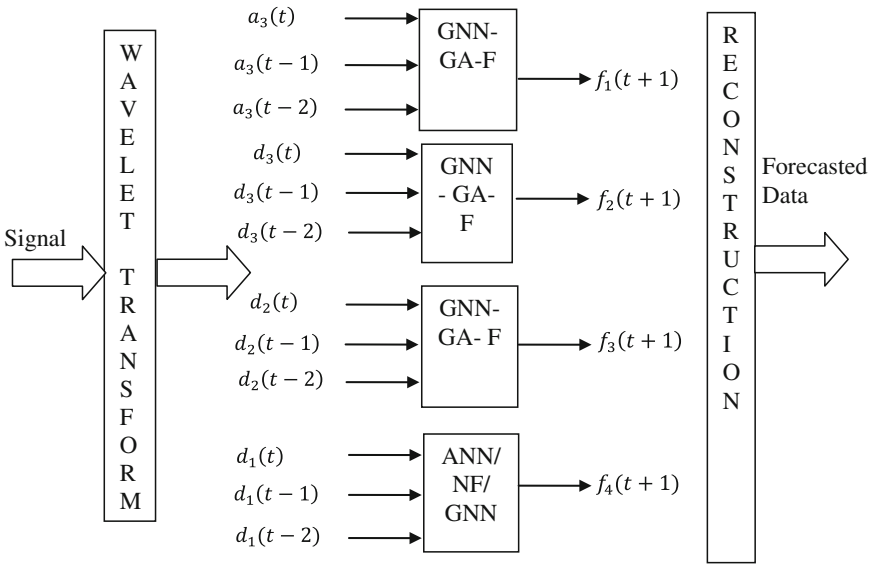

Fig. 2.3 Mechanism for short-term load forecasting

output of the neuron is a function of output of all activation functions. The learning of GNN is explained in [1].

The are many advantages of GNN such as less number of unknown weights, less training time, less number of training patterns, less complexity, and more flexibility.

The basic idea is to use the wavelet transforms and predict the data by synergism of soft computing techniques GNN-W-GA-F for individual coefficients of wavelet transform represented by a_3, d_1, d_2, and d_3. The input to the architecture to predict the wavelet coefficients is explained in Fig. 2.3.

2.4 Adaptive GA with Fuzzy System (GA-F)

Genetic algorithm (GA) simulates the strategy of evolution and survival of fittest. It is a powerful domain-free approach integrated with GNN as a learning tool. The GNN–GA integrated approach is applied to different problems to test this approach. It is well known that the GA optimization is slow and depends on the number of variables. To improve the convergence of GA, adaptive GA is developed, in which, the GA parameters are modified using fuzzy rules [5]. The initial parameters of GAF are given below:

GAF Parameters

- Population size: 50
- Initial crossover probability: 0.9

- Initial mutation probability: 0.1
- Selection operator: tournament selection
- Number of generations: 100

2.5 Short-Term Load Forecasting Using Generalized Neural Network-Wavelet-Genetic Algorithm-Fuzzy System (GNN-W-GA-F)

The neural network (NN) is widely used for short-term load forecasting applications in the past few decades. To improve the performance of ANN, GNN is developed. The GNN is then used to predict each wavelet component separately and combine the (predicted components) to get forecasted load.

The following steps are used in forecasting using GNN-W-GA-F.

Step-I Data collection

The electrical load data was collected from 33/11 kV substation of Dayalbagh Educational Institute (D.E.I.) Dayalbagh Agra. India has been recorded at every 1 h interval for each day for 1 year. The week containing no national holidays, Saturday, and Sunday, or religious holidays are not considered as desired data in the forecasting model. Furthermore, special holidays cannot be used as inputs since they have lower loads than a regular Monday to Friday and mislead the training.

Step-II Preprocessing of Data

The data collected in earlier step is preprocessed.

(a) Filtering of data

In this preprocessing of data, the data is de-noised, i.e., remove bad data. The data of Saturday and Sunday is removed, because the load patterns of these days are quite different and also they are not used in forecasting. Also the error data due to sensor problem or any other fault is removed.

(b) Normalization of data

The filtered and de-noised data is then used for electrical load forecasting after normalizing them.

The normalization range used in normalization process is from 0.1 to 0.9 and not in the range 0–1. This is because in extrapolation there is a tolerance of 0.1 on both sides.

Step-III Wavelet decomposition of Electrical load pattern

The wavelet transform is used to decompose the normalized electrical load pattern into a number of wavelet **components** as shown in Fig. 2.1. The original normalized signal of load demand is decomposed to high-

and low-frequency component by using **db8,** mother wavelet (**db8**) for calculating the coefficient of the details (***d***) and approximate (***a***) components.

Step-IV Selection of training pattern

The first step for training is obtaining an accurate and sufficient historical data after preprocessing. The data should be chosen that is relevant to the model. How well the data is chosen is the defining factor in how well the model output will match the event being modeled. There should be some correlation between the training data and the testing data. In the load data, for example, all the Monday's data look alike and this holds good for all the days of the week with some variations.

The wavelet-decomposed components are used for training.

The training patterns are consisting of decomposed wavelet components of given load pattern at time t, $t-1$, $t-2$ (past three points) as input and the forecasted wavelet component at $t+1$ as output. Hence, training patterns expressed as pair of set of input and output.

Training Pattern = [Input vector] $\rightarrow$
[Output Vector]

Roughly 85 % of total load data is used for training and rest 15 % load data is used for testing of models. The pseudo code of GNN-W-GA-F is given below.

```
Begin GNN-W-GAF
  Collect a set of data.
  Decompose the data into wavelet components
  Initialize parameters of GAF and GNN-W
     For I =1:Gmax
       while pop < popmax
         Evaluate the fitness using GNN-W
       end
       generation= generation + 1
       Modify crossover, mutation rate using FS
       select parental chromosomes
       Perform GA operators
       Get new population
     End
```

Step-V Forecasting using GNN-W

The forecasting models using GNN-W-GA-F for wavelet components have been used after proper training.

Step-VI Reconstruction of forecasted load

The forecasted load pattern is reconstructed after combining the wavelet components. In the comparisons of model performance, the load forecast accuracy is determined by RMSE.

2.6 Results and Discussions

The training of $\boldsymbol{a}_3$ component using GNN-W-GAF is shown in Fig. 2.4 as maximum fitness of GA-F. Actual load and forecasted load using **GNN-W-DA-F during testing** is given graphically in Fig. 2.5.

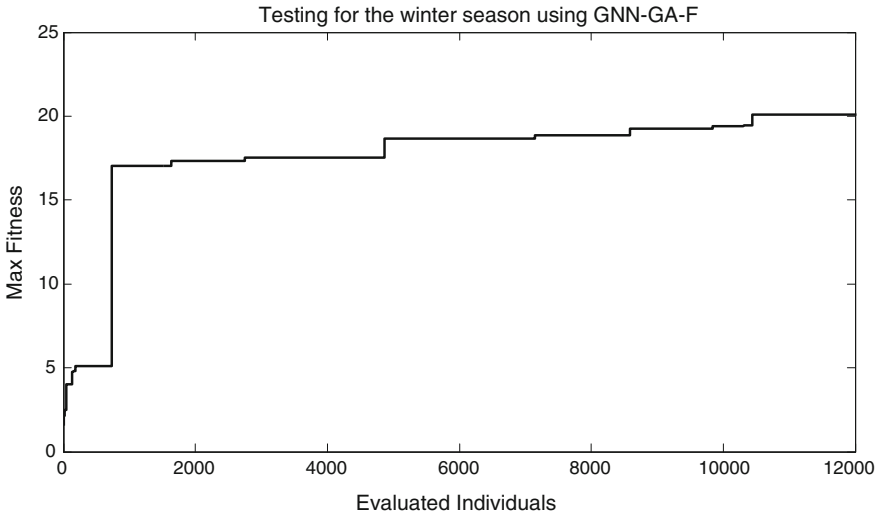

Fig. 2.4 Maximum Fitness of GA fuzzy during training of a_3 wavelet component of using GNN-W-GA-F

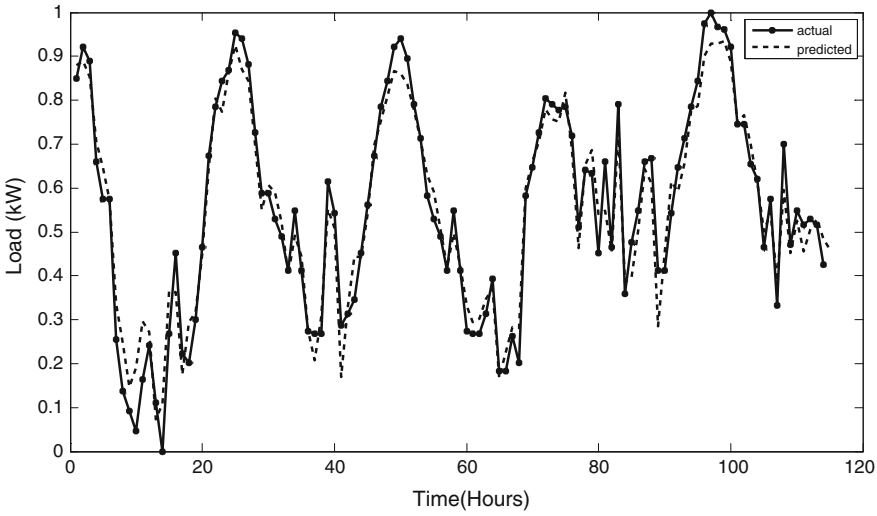

Fig. 2.5 Testing performance of GNN-W-GAF model

2.7 Conclusion

The paper deals with short-term electrical load forecasting problem using integrated approach of soft computing techniques and wavelet transform. The techniques and forecasting models were applied to datasets available from 33/11 kV substation of Dayalbagh Educational Institute, Dayalbagh, Agra, U.P. India. The soft computing technique, GNN-W-GA-F, has been applied to develop models for STLF. The integrated model, i.e., GNN-W-GAF gives the least RMSE in comparison to all the other ANN-based models.

References

1. Chaturvedi DK (2008) Soft computing techniques and applications to electrical engineering. Springer
2. Chaturvedi DK, Satsangi PS, Kalra PK (1999) New neuron model for simulating rotating electrical machines and load forecasting problems. Int J Elect Power Syst Res Elsevier Science, Ireland **52**:123–131
3. Chaturvedi DK, Satsangi PS, Kalra PK (2001) Fuzzified neural network approach for load forecasting problems. Int J Eng Intell Syst CRL Publishing, UK 9(1):3–9
4. Chaturvedi DK, Kumar R, Mohan M, Kalra PK (2001) Artificial neural network learning using improved genetic algorithm. J IE(I), EL, 82
5. Chaturvedi DK, Mohan M, Singh RK, Kalra PK (2004) Improved generalized neuron model for short term load forecasting. Int J Soft Computing—A Fusion of Foundations, Methodologies and Applications, Springer-Verlag Heidelberg 8(1):10–18
6. Chaturvedi DK, Satsangi PS, Kalra PK (1997) Short term load forecasting using generalized neural network (GNN) approach. J Inst Eng (India) 78:83–87
7. Chaturvedi DK (2009) Modeling and simulation of systems using MATLAB® and Simulink®. CRC Press (2009)
8. Chaturvedi DK, Premdayal SA, Chandiok A (2010) Load forecasting using soft computing. Int J Commun Netw Syst Sci 3:273–279
9. Chaturvedi DK, Singh AP (2011) Evapotranspiration model for irrigation scheduling. Int J Math Sci Appl 2(1):393–399
10. Chaturvedi DK, Premdayal SA (2012) Neural-wavelet based hybrid model for shortterm load forecasting. In: Proceedings of national conference on emerging trends in electrical, instrumentation and communication engineering (ETEIC—2012), Agra, 6th and 7th April 2012, pp 359–362
11. Chaturvedi DK, Chandiok A, Kamaruddin M (2009) Fuzzy system model of internet system exploitation and usage in society. IEEE Int Conf Innov Technol (IEEE- IMS/EMBS2009), New Delhi, 18–19th June 2009
12. Chaturvedi DK, Chandiok A, Siddiqui AH (2009) Prediction for the time series of annual rainfall using wavelets and neuro-fuzzy approach. Int workshop on Wavelets, Istanbul, Turkey, 4–7th June 2009
13. Chaturvedi DK, Siddiqi AH, Chandiok A, Agarwal S (2009) Annual rainfall forecasting using soft computing techniques. In: Proceedings of joint international conference on applied systems research and national systems conference, (NSC) Nov. 27–29, 2009 organized by D. E.I. (Deemed Univ.) Dayalbagh, Agra, India, TMH-2009:103–106

14. Banakar MFAA (2008) Artificial wavelet neural network and its application in Neurofuzzy models. Elsevier Appl Soft Comput
15. Sinha N, Lai LL, Ghosh PK, Yingnan M (2007) Wavelet-GA-ANN based hybrid model for accurate prediction of short-term load forecast. In: Proceedings international conference on intelligent systems applications to power systems (ISAP), 5–8 Nov, pp 1–8
16. Minsky M, Papert S (1968) Perceptions. The MIT Press, Cambridge, MA

Chapter 3
Integral Geometry and Mathematical Problems of Image Reconstruction

Gaik Ambartsoumian

Abstract Integral geometry is a branch of mathematics studying the representation of functions by their integrals along various curves and surfaces. Such tasks arise naturally in many problems of image reconstruction in medicine, remote sensing, non-destructive testing, and some other areas. In this paper, we give a short survey of mathematical models of several imaging modalities, which are based on generalized Radon transforms. We discuss the major mathematical problems arising in the study of these transforms, describe the known results, and state some open problems. The paper includes an extensive list of references providing further sources for interested readers.

Keywords Integral geometry · Image reconstruction · Generalized Radon transforms · Computer-assisted tomography (CAT) · Thermoacoustic tomography (TCT)

3.1 Introduction

In many imaging applications, the data collected by imaging devices correspond to the integrals of an unknown function along certain curves or surfaces. The mathematical task of image reconstruction in these cases becomes the stable recovery of that unknown function from its integrals [1–4]. This requires the inversion of so-called generalized Radon transforms (GRT) [3–6] (Fig. 3.1).

In this section, we describe several imaging modalities used in medicine and other fields, consider their mathematical models, and define the corresponding Radon transforms. We state the major problems arising in the study of these transforms the detailed discussion of which is presented in further sections.

G. Ambartsoumian (✉)
Department of Mathematics, University of Texas at Arlington,
P.O. Box 19408, Arlington, TX 76019-0408, USA
e-mail: gambarts@uta.edu

A.H. Siddiqi et al. (eds.), *Mathematical Models, Methods and Applications*,
Industrial and Applied Mathematics, DOI 10.1007/978-981-287-973-8_3

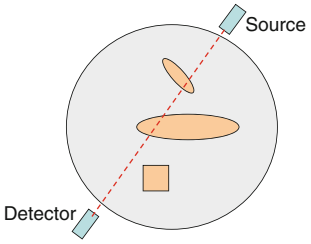

Fig. 3.1 A sketch of an X-ray passing through the body in CAT

One of the most famous medical imaging techniques is the computer-assisted tomography (CAT) which uses X-rays to produce images of the interior organs of the human body. In simple terms, the CAT process can be described as follows. An X-ray beam is sent through the body, and its intensity is measured at both the source and the detector (Fig. 3.1). It is known from physics that the change of intensity ΔI of the beam on a small interval Δx satisfies the following simple law:

$$\frac{\Delta I}{I} = f(x)\,\Delta x, \tag{3.1}$$

where $f(x)$ is the X-ray attenuation coefficient of the human body at point x and $I(x)$ is the intensity of the beam at point x. If the measured intensity at the source is I_0 and at the detector is I_1 then considering infinitesimal increments of Δx in (3.1) and solving the resulting differential equation one gets

$$\frac{I_1}{I_0} = \exp\left\{-\int_L f(x)\mathrm{d}x\right\}, \tag{3.2}$$

where the integral is taken along the line L passing through the source and the detector.

Hence, using the measurements of I_0 and I_1 at various locations of the source and the detector one can obtain the integrals of the unknown X-ray attenuation function $f(x)$ along various straight lines passing through the body. The grayscale graph of $f(x)$ is exactly what the CAT scanners use to create images of the interior organs. Thus, the mathematical task of image reconstruction in CAT is the recovery of $f(x)$ from its integrals along straight lines.

Definition 1 The Radon transform $\mathcal{R}$ maps a function f on $\mathbb{R}^2$ into the set of its integrals along lines in $\mathbb{R}^2$. In particular, if $\theta \in [0, 2\pi]$, $\omega = (\cos\theta, \sin\theta)$ and $s \in \mathbb{R}$, then

$$\mathcal{R}f(\theta, s) = \int\limits_{x\cdot\omega=s} f(x)\,\mathrm{d}x = \int\limits_{\omega^{\perp}} f(s\omega + y)\,\mathrm{d}y \tag{3.3}$$

is the integral of f along lines orthogonal to ω at (signed) distance $\mathcal{S}$ away from the origin.

Here f is assumed to be such that the integrals are well defined. For example, one can take $f \in \mathcal{S}(\mathbb{R}^2)$, the Schwartz space. For simplicity, in this article, we will consider mainly compactly supported functions f of certain smoothness. For details about most general classes of f for which the theory of GRT has been developed see [3–6].

In the discussion above, we assume that the source and the detector move within a fixed plane so that we get the integrals of f along all possible lines in that plane. Thus, the reconstructed 2D image will correspond to a cross-sectional view of the human body along that plane. Then one can vertically stack such cross-sectional images to create a 3D image of the body interior.

Remark 2 One can consider a setup where the source and the detector are not limited to a plane, and are placed at various locations in the space. In this case the CAT scanner measures the integrals of the X-ray attenuation coefficient f on $\mathbb{R}^3$ along lines in $\mathbb{R}^3$. The fully 3D recovery of f from such data is also an interesting problem, which will not be addressed in this paper. For more details on this we refer the interested reader to [1–4, 6]. Here we just mention that in dimensions $n = 3$ and higher, the transform integrating a function along lines in $\mathbb{R}^n$ is called X-ray transform, whereas the term Radon transform refers to the transform integrating f over hyperplanes in $\mathbb{R}^n$. Of course, for $n = 2$ these transforms coincide.

The problem of 2D-image reconstruction in CAT requires the inversion of the Radon transform defined above. In further sections, we will discuss the existence and uniqueness of such an inversion, the inversion formulas and algorithms, as well as their numerical implementations and stability issues. We continue this section with discussing a few other imaging modalities and corresponding GRT.

Thermoacoustic tomography (TCT) is one of the most promising novel medical imaging techniques [7–12]. The TCT scanner sends a short pulse of radio-frequency (RF) electromagnetic waves through the body heating up the tissue. The resulting thermo-elastic expansion of the tissue generates ultrasound waves which are measured by an array of transducers outside of the body. The measured data is then used to recover the RF-absorption coefficient $f(x)$, and the grayscale graph of which is used to generate the image of the body interior.

In a simple mathematical model of TCT, it is assumed that the RF energy is deposited instantaneously and uniformly through the body, and the speed of sound c is constant inside the body (see Fig. 3.2). Under these assumptions, the signal registered at a given transducer location p at some moment of time t corresponds to the integral of $f(x)$ along a sphere $S(p, r)$ centered at p and of radius $r = ct/2$ (e.g., see [7–12]).

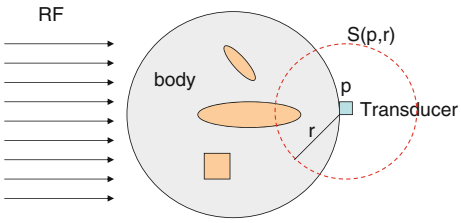

Fig. 3.2 A sketch of a simple TCT model

Hence, by placing the transducer at various locations and registering the generated ultrasound signals as functions of time, the TCT scanner effectively measures the integrals of the unknown image function f along a family of spheres. So the image reconstruction problem in TCT is mathematically equivalent to the recovery of f from its integrals along spheres.

One can place the transducers along a fixed planar curve and focus them so that they register only the signals coming from that plane. In that case, we get a problem of recovering a function f in $\mathbb{R}^2$ from its integrals along a two-dimensional family of circles in $\mathbb{R}^2$ (one parameter specifying the center of the circle along the given curve and the other one specifying its radius). As in the case of CAT, one can then vertically stack these 2D images to get a 3D image of the body interior. Of course, the fully 3D reconstruction of f from integrals over spheres in $\mathbb{R}^3$ is also possible.

Definition 3 The spherical Radon transform (SRT) $\mathcal{R}_S$ maps a function f on $\mathbb{R}^n$ into the set of its integrals along spheres in $\mathbb{R}^n$

$$\mathcal{R}_s f(p,r) = \int\limits_{|x-p|=r} f(x)\mathrm{d}\sigma(x), \tag{3.4}$$

where $\mathrm{d}\sigma(x)$ is area measure on the sphere $|x-p| = r$. (The transform is often called circular Radon transform when $n= 2$, with $\mathrm{d}\sigma(x)$ denoting the arc length measure on the circle.)

For SRT too, we assume in this article that f is a compactly supported function of certain smoothness.

The image reconstruction in TCT requires the inversion of the SRT, which will be discussed in the further sections. It is important to mention that SRT inversion is also needed in some other imaging modalities, e.g., in photo-acoustic tomography [11, 12], mono-static ultrasound reflection tomography [13, 14], radar and sonar imaging [15, 16], etc.

In bi-static ultrasound tomography (URT), an emitter sends spherical sound waves through the body and the backscattered echoes are registered by a receiver. Assuming that the speed of sound is constant in the body, the signal registered at any given moment by the receiver is generated by reflections from all those points for which the sum of their distances r_1 and r_1 to the emitter and the receiver is

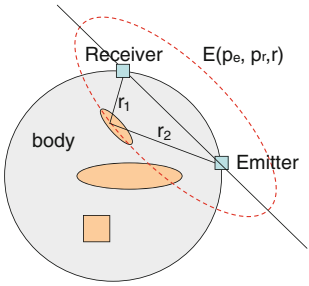

Fig. 3.3 A sketch of a simple bi-static URT model

constant (depending on time and sound speed) [17, 18]. In other words, those points are located on confocal ellipsoids of rotation in 3D (or ellipses in 2D) with foci at the emitter and receiver locations, and the problem of image reconstruction boils down to the inversion of a transform integrating functions along such ellipsoids in 3D (or ellipses in 2D) (see Fig. 3.3).

Definition 4 The elliptical Radon transform of $f(x)$, $x \in \mathbb{R}^n$ is defined as

$$\mathcal{R}_E f(p_e, p_r, r) = \int\limits_{|x-p_e|+|x-p_r|=r} f(x) \mathrm{d}\sigma(x), \tag{3.5}$$

where $\mathrm{d}\sigma(x)$ is the area measure on the ellipsoid $|x - p_e| + |x - p_r| = r$ (or arclength in 2D).

The rest of the paper is organized as follows. In Sect. 3.2, we discuss the known results and open problems related to the uniqueness of inversion of the generalized Radon transforms presented here. Section 3.3 is dedicated to the inversion formulas and algorithms of such transforms. Section 3.4 includes some additional remarks, and we finish the paper with acknowledgements.

3.2 Uniqueness

The Radon transform $\mathcal{R}f$ defined in Eq. (3.3) is a classical concept that has been extensively studied in twentieth century. It is well known that if $f \in \mathcal{S}(\mathbb{R}^2)$ then it is uniquely determined by $\mathcal{R}f(\theta, s)$, with $\theta \in [0, 2\pi]$, $s \in \mathbb{R}$ (e.g., see [3, 5, 6]). Various inversion formulas are known for this case, some of which will be presented in Sect. 3.3.

Notice, that the set of all lines in the plane is two-dimensional, so here we recover a function $f(x)$, $x \in \mathbb{R}^2$ of two variables from a two-dimensional dataset $\mathcal{R}f(\theta, s)$. The situation is drastically different for the spherical transform $\mathcal{R}_S f$, since the set of the circles in the plane is three-dimensional, and for the elliptic transform $\mathcal{R}_E f$, since the set of ellipses in the plane is five-dimensional. (Similar mismatch happens also if we consider $f(x)$ with $x \in \mathbb{R}^n$, $n \geq 3$.) Hence the inversion problems

for $\mathcal{R}_S f$ and $\mathcal{R}_E f$ are overdetermined. To match the dimensions of the domain of f and the GRT data, one needs to restrict both $\mathcal{R}_S f$ and $\mathcal{R}_E f$ to two-parameter families of circles and ellipses correspondingly. Consequently, the uniqueness results corresponding to these transforms depend on the choice of these restrictions, thus are more complicated, and less well understood.

While there are many different ways of restricting the degrees of freedom of $\mathcal{R}_S f$ to two, the imaging applications described above suggest a natural choice of restricting the centers of integration circles to a one-dimensional set M (e.g., a curve). From the imaging point, this curve can represent an arc of transducers surrounding the patient's body in TCT, or a flight trajectory of a plane in synthetic aperture radar imaging. Similarly, one can reduce the dimension of $\mathcal{R}_E f$ by two restricting the foci of the integration ellipses to a fixed simple curve (e.g., a circle). There will still be one extra dimension left in this case, and we choose to reduce it by forcing the distance between the foci (i.e., between the emitter and the receiver in URT) to be constant.

With these restrictions matching the dimensions of the domain of f and GRT data, we can now formulate several results about the unique inversion of these restricted transforms. Let us start with a definition

Definition 5 The spherical Radon transform $\mathcal{R}_S$ is said to be injective on a set M (M is called a set of injectivity) if for any $f \in C_c(\mathbb{R}^n)$ the condition $\mathcal{R}_S f(p, r) \equiv 0$ for all $r \in \mathbb{R}^+$ and all $p \in M$ implies $f \equiv 0$.

In other words, if M is an injectivity set, then the restricted transform $\mathcal{R}_S f(p, r)$, $p \in M$ can be uniquely inverted. Similarly,

Definition 6 The elliptical Radon transform $\mathcal{R}_E$ is said to be injective on a set M (M is called a set of injectivity) if for any $f \in C_c(\mathbb{R}^n)$ the condition $\mathcal{R}_E f(p_e, p_r, r) \equiv 0$ for all $r \in \mathbb{R}^+$ and all $p_e, p_r \in M$ implies $f \equiv 0$.

In the case of $n = 2$, a complete description of injectivity sets for $\mathcal{R}_S$ was given by Agranovsky and Quinto in [19]. To state their result, we need to introduce one more concept.

Definition 7 For any $n \in \mathbb{N}$ denote by Σ_n the Coxeter system of n lines $L_0, \ldots, L_{n-1}$ in the plane: $L_k = \{te^{i\pi k/n} | -\infty < t < \infty\}$ (here we identify the plane with the complex plane C).

In other words, a Coxeter system of n lines is a "cross" of n lines intersecting at the origin under equal angles. Then all injectivity sets of $\mathcal{R}_S$ in $\mathbb{R}^2$ can be described as follows.

Theorem 8 ([19]) *A set $M \in \mathbb{R}^2$ is a set of injectivity for the circular Radon transform $\mathcal{R}_S$ on $C_c(\mathbb{R}^2)$ if and only if M is not contained in any set of the form $\omega(\Sigma_N) \bigcup F$, where ω is a rigid motion in the plane, Σ_N is a Coxeter system of N lines, and F is a finite set.*

The authors of [19] also formulated the (still unproven) conjecture for higher dimensions.

Conjecture 9 ([19]) *The following condition is necessary and sufficient for M to be a set of injectivity for the spherical Radon transform $\mathcal{R}_S$ on $C_c(\mathbb{R}^n)$: M is not contained in any set of the form $\omega(\Sigma) \bigcup F$, where ω is a rigid motion of $\mathbb{R}^n$, Σ is the zero set of a homogeneous harmonic polynomial, and F is an algebraic subset in $\mathbb{R}^n$ of co-dimension at least 2.*

While the complete result in higher dimensions is still not proven, various partial results have been established in the last decade.

In [20], we used some PDE techniques developed in [21] to prove some very general results concerning geometry of non-injectivity sets of SRT, as well as reproved certain known results with much simpler means (namely, finite speed of propagation and domain of dependence for the wave equation). We formulate one of these results below.

Let S be an algebraic hypersurface that splits $\mathbb{R}^n$ into connected parts H^j, $j = 1, \ldots, m$. One can define the interior metric in H^j as follows:

$$d^j(p, q) = \inf\{\text{length of } \gamma\},$$

where the infimum is taken over all C^1-curves γ in H^j joining points $p, q \in H^j$.

Theorem 10 ([20]) *Let S and H^j be as above and $f \in C(\mathbb{R}^n)$ be such that $\mathcal{R}_S f(p, r) = 0$ for all $p \in S$, and all $r > 0$. Let also $x \in \bar{H}^j$, where $\bar{H}^j$ is the closure of H^j. Then*

$$\operatorname{dist}(x, \operatorname{supp} f \cap H^j) = \operatorname{dist}^j(x, \operatorname{supp} f \cap H^j) \leq \operatorname{dist}(x, \operatorname{supp} f \cap H^k),\ k \neq j,$$

where distances dist^j are computed with respect to the metrics d^j, while dist is computed with respect to the Euclidean metric in $\mathbb{R}^n$. In particular, for $x \in S$ and any j

$$\operatorname{dist}(x, \operatorname{supp} f \cap H^j) = \operatorname{dist}^j(x, \operatorname{supp} f \cap H^j) = \operatorname{dist}(x, \operatorname{supp} f), \quad j = 1, \ldots, m.$$

Notice, that the obtained necessary conditions for S to be a non-injectivity set not only hold in arbitrary dimensions, but also they do not require f to have compact support, and in fact do not impose any restriction on the behavior of f at infinity. One of the corollaries of the previous theorem is the following.

Theorem 11 ([20]) *Let $S \subset \mathbb{R}^n$ and $f(\neq 0) \in C_c(\mathbb{R}^n)$ be such that $\mathcal{R}_S f(p, r) = 0$ for any $p \in S$, and any $r > 0$. If the external boundary of the support of f is connected and its curvature is bounded from below, then S is a ruled hypersurface (union of a family of lines).*

If we could also show that all these lines pass through the same point (which can be easily done in 2D [20], but not so in higher dimensions), then this would immediately imply the validity of Conjecture 9 for this particular case (see [20]).

For further details about the injectivity sets of the spherical transform, we refer the reader to [20] and the references there.

Another interesting question related to the uniqueness of inversion of $\mathcal{R}_s$ is the possibility of unique reconstruction of f from $\mathcal{R}_s(p,r)$, where $p \in M$ for some specific set M of dimension 1, and r is restricted to an interval $0<r<r_0$ (as opposed to $r>0$ used before).

In [22], we proved that in the case when M is a circle of radius R, one can get such uniqueness results.

Theorem 12 ([22]) *Let $f(r,\theta)$ be an unknown continuous function supported inside the annulus $A(\varepsilon,R)=\{(r,\theta): r\in(\varepsilon,R),\ \theta\in[0,2\pi]\}$, where $0<\varepsilon<R$. If $\mathcal{R}_S f(\rho,\phi)$ is known for $\phi\in[0,2\pi]$ and $\rho\in[0,R-\varepsilon]$, then $f(r,\theta)$ can be uniquely recovered in $A(\varepsilon,R)$.*

Theorem 13 ([22]) *Let $f(r,\theta)$ be an unknown continuous function supported inside the annulus $A(R,3R)=\{(r,\theta): r\in(R,3R),\ \theta\in[0,2\pi]\}$. If $\mathcal{R}_S f(\rho,\phi)$ is known for $\phi\in[0,2\pi]$ and $\rho\in[0,R_1]$, where $0<R_1<2R$ then $f(r,\theta)$ can be uniquely recovered in $A(R,R_1)$.*

The paper also presented an exact inversion formula for SRT from this type of radially partial data for both interior and exterior problems. Some potential fields of application of the results of this work are intravascular ultrasound (IVUS) and Transrectal Ultrasound (TRUS) imaging, where the exterior problem appears naturally [23, 24].

For the case of the elliptic Radon transform, theorems similar to the Theorems 12 and 13 have been established in [25]. No comprehensive result similar to Theorem 8 is known for ERT at this time.

3.3 Inversion

Exact inversion formulas for Radon-type transforms can be roughly divided into two categories: closed backprojection type inversion formulas and expansions into series.

For the regular Radon transform $\mathcal{R}$, there are various explicit inversion formulae in the case, when $\mathcal{R}f(\psi,t)$ is known for all $\psi\in[0,2\pi]$ and all t (see [3]). For example, if $f\in\mathcal{S}(\mathbb{R}^2)$ (the Schwartz space), one of the most commonly used inversion formulae is the filtered backprojection:

$$f(x,y)=\frac{1}{4\pi}\int_0^{2\pi}\mathcal{H}\left(\mathcal{R}f_t^{'}\right)(\psi,x\cos\psi+y\sin\psi)\mathrm{d}\psi, \tag{3.6}$$

where $\mathcal{H}$ is the Hilbert transform defined by

$$\mathcal{H}h(t)=-\frac{i}{\sqrt{2\pi}}\int_{\mathbb{R}}\operatorname{sgn}(r)\,\widehat{h}(r)\,\mathrm{e}^{irt}\,\mathrm{d}r, \tag{3.7}$$

and $\hat{h}(r)$ is the Fourier transform of $h(t)$, i.e.,

$$\hat{h}(r) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} h(t) \mathrm{e}^{-irt} \mathrm{d}t. \tag{3.8}$$

A typical example of an inversion formula using series expansion is the one described by A. Cormack in his pioneering paper [26]. Let $f(\phi, r)$ be the image function in polar coordinates, $g(\theta, s) = \mathcal{R}(\theta, s)$ as in formula (3.3), with $\phi, \theta \in [0, 2\pi]$. Then one can expand both functions into Fourier series with respect to the corresponding angular variables:

$$f(\phi, r) = \sum_l f_l(r) \mathrm{e}^{il\phi}, \quad g(\theta, s) = \sum_l g_l(s) \mathrm{e}^{il\theta}.$$

Cormack showed in [26] that the lth Fourier coefficient of g depends only on lth Fourier coefficient of f, and that relation can be inverted, namely

$$g_l(s) = 2 \int_s^\infty T_{|l|}\left(\frac{s}{r}\right) \left(1 - \frac{s^2}{r^2}\right)^{-1/2} f_l(r) \mathrm{d}r$$

and

$$f_l(r) = -\frac{1}{\pi} \int_r^\infty (s^2 - r^2)^{-1/2} T_{|l|}\left(\frac{s}{r}\right) g_l'(s) \, \mathrm{d}s,$$

where $T_{|l|}$ are the Chebyshev polynomials of the first kind.

For SRT, the first exact inversion formulas used Fourier expansion techniques in 2D [13] and 3D [14] spherical acquisition geometry. For example, the result of Norton [13] can be stated as follows. Let $\mathcal{R}_S$ be the 2D spherical Radon transform on the plane that integrates functions compactly supported inside the unit disk D over all circles $|x - p| = \rho$ with centers $p = (\cos\theta, \sin\theta)$ located on the unit circle S. Consider the Fourier decomposition of $f(r, \phi)$ and $g(\rho, \theta)$ in angular variables

$$f(r, \phi) = \sum_{-\infty}^{\infty} f_k(r) \mathrm{e}^{ik\phi}, \quad g(\rho, \theta) = \mathcal{R}_S f(\rho, \theta) = \sum_{-\infty}^{\infty} g_n(\rho) \mathrm{e}^{in\theta}. \tag{3.9}$$

Since SRT commutes with rotations about the origin, the Fourier series expansion with respect to the polar angle partially diagonalizes the operator, and thus the nth Fourier coefficient $g_n(\rho)$ of $g = \mathcal{R}_S f$ will depend only on the nth coefficient f_n of the original f. It was shown in [13] that:

$$g_n(\rho) = 2\pi\rho \mathcal{H}_0\{J_n \mathcal{H}_n\{f_n\}\}, \tag{3.10}$$

where $(\mathcal{H}_n h)(\sigma)$ is the Hankel transform of an integer order n for a function $h(r)$ on $\mathbb{R}^+$

$$(\mathcal{H}_n h)(\sigma) = \int_0^\infty J_n(\sigma r) h(r) r \, \mathrm{d}r.$$

Here the standard notation J_n is used for Bessel functions. Using the fact that the Hankel transform is self-invertible one can now easily get an exact inversion of SRT by

$$f_k(r) = \mathcal{H}_k \left\{ \frac{1}{J_k(\sigma)} \mathcal{H}_0 \left[\frac{g_k(\rho)}{2\pi\rho} \right] \right\}.$$

In [14], the authors obtained a similar result in terms of spherical harmonics expansion of g in 3D.

In [22], we used similar techniques to derive exact inversion formulas for SRT of functions supported inside annuli, which are interior or exterior with respect to the data acquisition circle. The approach was based again on the rotation invariance of SRT, which allowed to diagonalize the operator and reduce the problem to the solution of an Abel-type integral equation with a special function kernel (in this case including Chebyshev polynomials). The authors are currently working on the extension of that result to 3D functions supported in spherical shells.

For the case of the elliptic Radon transform, similar results have been established in [25].

Another approach for obtaining exact inversion formulas for SRT in the form of series expansion was demonstrated by Kunyansky in [27]. Here the author used the expansion of the unknown function f in the basis of eigenvalues of Dirichlet Laplacian $-\Delta$, and the known relation between SRT and the solution of wave equation. The result in [27] holds for arbitrary closed acquisition surfaces, for which the eigenfunctions of the Dirichlet Laplacian are explicitly known (e.g., cube, finite cylinder, half-sphere).

Another important class of exact inversions includes closed form integral transform type formulas. For the case of SRT various backprojection type, formulas have been established by different authors in the spherical acquisition geometry [8, 21, 28–30].

For example, when $n = 3$ the authors of [21] proved the following Filtered Back Projection (FBP) formula

$$f(x) = \frac{-1}{8\pi^2} \left(\int_{|p|=1} \frac{1}{|x-p|} \mathcal{R}_S f''(p, |x-p|) \mathrm{d}p \right), \tag{3.11}$$

and in the case of $n = 2$ it was shown in [28] that if $|p| = 1$ then

$$f(x) = \frac{1}{(2\pi R_0)^2} \Delta_x \int_0^{2\pi} \int_0^{2R_0} r \mathcal{R}_S f(p, r) \log\left|r^2 - |x - p|^2\right| \mathrm{d}r \mathrm{d}\phi. \tag{3.12}$$

In planar geometry backprojection type, inversion formulas for functions that are even with respect to the plane were established in [9, 31, 32]. In case of cylindrical geometry, similar results were found in [10]. It should be noted, that these are the only acquisition geometries, for which closed form inversion formulas for SRT have been discovered. For more complicated geometries, one has to use either approximate inversion techniques described above, or the series expansion type approaches.

3.4 Additional Remarks

- Another interesting topic in integral geometry is the description of ranges of GRT's. Typically, the range of a Radon-type transform satisfies infinitely many conditions, in other words it has an infinite co-dimension in the space of smooth functions with corresponding variables. The knowledge of these conditions may be useful in applications to suppress the noise in data measurements, fill in some missing data with better approximation than mere zero-filling, determine malfunctioning hardware in the scanners, etc. To learn more about range descriptions of the Radon transform $\mathcal{R}$, we refer the reader to [3–6]. For the spherical Radon transform $\mathcal{R}_S$ check out [33–35] and the references there. The range description of the elliptical Radon transform $\mathcal{R}_E$ is still an open problem.
- The existence and uniqueness of an inversion for a GRT does not guarantee the possibility of an application of such an inversion on practise. In many setups, any existing inversion formula or algorithm is unstable, and the numerical implementations of these formulas lead to blurry images, and/or severe artifacts. To learn more about the stability issues of the inversion of GRT, we refer the reader to [3, 4, 11, 12, 16] and the references there.

Acknowledgments The author thanks Professor Abul Hasan Siddiqi and the other organizers of 11th Annual Conference of the Indian Society of Industrial and Applied Mathematics for the invitation to speak at the conference, the great hospitality, and the offer to publish this review paper in conference proceedings. This work was supported in part by US NSF Grant DMS 1109417.

References

1. Epstein CL (2008) Introduction to the mathematics of medical imaging. SIAM, Philadelphia
2. Kaka C, Slaney M (1988) Principles of computerized tomographic imaging. IEEE Press, New York
3. Natterer F (1986) The mathematics of computerized tomography. Wiley, New York
4. Natterer F, Wübbeling F (2001) Mathematical methods in image reconstruction. In: Monographs on mathematical modeling and computation, vol 5. SIAM, Philadelphia, PA
5. Ehrenpreis L (2003) The universality of the Radon transform. Oxford University Press, Oxford
6. Helgason S (1980) The Radon transform. Birkhäuser, Basel
7. Kuchment P, Kunyansky L (2008) A survey in mathematics for industry: mathematics of thermoacoustic tomography. Eur J Appl Math 19(2):191–224
8. Xu M, Wang L-H (2001) Time-domain reconstruction for thermoacoustic tomography in a spherical geometry. IEEE Trans Med Imag 21:814–822
9. Xu Y, Feng D, Wang L-H (2002) Exact frequency-domain reconstruction for thermoacoustic tomography: I Planar geometry. IEEE Trans Med Imag 21:823–828
10. Xu Y, Xu M, Wang L-H (2002) Exact frequency-domain reconstruction for thermoacoustic tomography: II Cylindrical geometry. IEEE Trans Med Imag 21:829–833
11. Xu Y, Wang L-H, Ambartsoumian G, Kuchment P (2004) Reconstructions in limited view thermoacoustic tomography. Med Phys 31(4):724–733
12. Xu Y, Wang L-H, Ambartsoumian G, Kuchment P (2009) Limited view thermoacoustic tomography. In: Wang L-H (ed) Photoacoustic imaging and spectroscopy. CRC Press
13. Norton SJ (1980) Reconstruction of a two-dimensional reflecting medium over a circular domain: exact solution. J Acoust Soc Am 67(4):1266–1273
14. Norton SJ, Linzer M (1981) Ultrasonic reflectivity imaging in three dimensions: exact inverse scattering solutions for plane, cylindrical, and spherical apertures. IEEE Trans Biomed Eng 28:202–220
15. Cheney M (2001) Tomography problems arising in synthetic aperture radar. In: Quinto ET et al Radon transforms and tomography. American Math. Soc., Providence, RI
16. Louis AK, Quinto ET (2000) Local tomographic methods. In: Sonar surveys on solution methods for inverse problems. Springer, Vienna pp 147–154
17. Mensah S, Franceschini E (2007) Near-field ultrasound tomography. J Acoust Soc Am 121(3):1423–1433
18. Mensah S, Franceschini E, Pauzin MC (2007) Ultrasound mammography. Nucl Instrum Meth Phys Res, Section A 571(3):52–55
19. Agranovsky ML, Quinto ET (1996) Injectivity sets for the Radon transform over circles and complete systems of radial functions. J Funct Anal 139:383–413
20. Ambartsoumian G, Kuchment P (2005) On the injectivity of the circular Radon transform. Inverse Prob 21:473–485
21. Finch D, Patch SK, Rakesh (2004) Determining a function from its mean values over a family of spheres. SIAM J Math Anal 35(5):1213–1240
22. Ambartsoumian G, Gouia-Zarrad R, Lewis M (2010) Inversion of the circular Radon transform on an annulus. Inverse Prob 26:105015
23. Cobbold R (2007) Foundations of biomedical ultrasound. Oxford University Press, New York
24. Patel U, Rickards D (2002) Handbook of transrectal ultrasound and biopsy of the prostate. Taylor & Francis
25. Ambartsoumian G, Krishnan V (2015) Inversion of a class of circular and elliptical radon transforms. Complex Analysis and Dynamical Systems VI (Part 1: PDE, Differential Geometry, Radon Transforms). Contemporary Mathematics, AMS
26. Cormack A (1963) Representation of a function by its line integrals, with some radiological applications. J Appl Phys 34(9):2722–2727
27. Kunyansky L (2007) A series solution and a fast algorithm for the inversion of the spherical mean Radon transform. Inverse Prob 23:s11–s20

28. Finch D, Haltmeier M, Rakesh (2007) Inversion of spherical means and the wave equation in even dimension. SIAM J Appl Math 68(3):392–412
29. Kunyansky L (2007) Explicit inversion formulas for the spherical mean Radon transform. Inverse Prob 23:373–383
30. Nguyen L (2009) A family of inversion formulas in thermoacoustic tomography. Inverse Prob Imag 3(4):649–675
31. Andersson L-E (1988) On the determination of a function from spherical averages. SIAM J Math Anal 19(1):214–232
32. Fawcett JA (1985) Inversion of *n*-dimensional spherical averages. SIAM J Appl Math 45 (2):336–341
33. Agranovsky ML, Kuchment P, Quinto ET (2007) Range descriptions for the spherical mean Radon transform. J Funct Anal 248:344–386
34. Ambartsoumian G, Kuchment P (2006) A range description for the planar circular Radon transform *SIAM*. J Math Anal 38(2):681–692
35. Finch D, Rakesh (2006) The range of the spherical mean value operator for functions supported in a ball. Inverse Prob 22:923–38

Chapter 4
Microlocal Analysis of Some Synthetic Aperture Radar Imaging Problems

Venkateswaran P. Krishnan

Abstract In this article, we analyze the microlocal properties of the linearized forward scattering operator $\mathcal{F}$, which arises in synthetic aperture radar imaging. A frequently applied imaging technique is to study the normal operator $\mathcal{F}^*\mathcal{F}$ ($\mathcal{F}^*$ is the L^2 adjoint of $\mathcal{F}$). However, such an imaging technique introduces artifacts in the image. We study the structure of these artifacts.

Keywords Synthetic aperture radar imaging · Singular Fourier integral operators · Elliptical radon transforms · Fold and blowdown singularities

MSC 2010 Classification 35S30 · 35S05 · 58J40 · 35A27

4.1 Introduction

In synthetic aperture radar (SAR) imaging, a region of interest on the surface of the earth is illuminated by electromagnetic waves from a moving airborne platform. One then tries to reconstruct an image of the region based on the measurement of backscattered waves. For an in-depth treatment of SAR imaging, we refer the reader to [3, 4]. SAR imaging is similar to other imaging problems such as sonar or seismic imaging where acoustic or pressure waves, respectively, are used to reconstruct objects on the ocean floor or underneath the surface of the earth [2, 5, 6, 23].

In monostatic SAR, the source and the receiver are located on the same moving airborne platform. In bistatic SAR, the source and the receiver are on independently moving airborne platforms. There are several advantages in considering such data acquisition geometries (ways of acquiring data). The receivers, compared to the transmitters, are passive and hence are more difficult to detect. Therefore, by

V.P. Krishnan (✉)
Centre for Applicable Mathematics, Tata Institute of Fundamental Research, Bangalore, India
e-mail: vkrishnan@math.tifrbng.res.in

A.H. Siddiqi et al. (eds.), *Mathematical Models, Methods and Applications*, Industrial and Applied Mathematics, DOI 10.1007/978-981-287-973-8_4

separating their locations, the receivers alone can be in an unsafe environment, while the transmitters are in a safe environment. Furthermore, bistatic SAR systems are more resistant to electronic countermeasures such as target shaping to reduce scattering in the direction of incident waves [21].

In this paper, we consider a bistatic SAR system where the antennas have poor directivity and hence the beams do not focus on targets on the ground. We assume that the transmitter and receiver traverse a one-dimensional curve and the backscattered data is measured at each point on this curve for a certain period of time. For the acquisition geometries we consider as in the monostatic SAR case, we show that with a weak scattering assumption, the linear scattering operator that relates the unknown function that models the object on the ground to the data at the receiver is a Fourier integral operator (FIO) [8, 17, 19, 28, 29]. Now, when $\mathcal{F}$ is an FIO, the canonical relation $\mathcal{C}_{\mathcal{F}}$ associated to $\mathcal{F}$ tells us how the singularities of the object are propagated to the data. The canonical relation $\mathcal{C}_{\mathcal{F}^*}$ of the L^2 adjoint $\mathcal{F}^*$ of $\mathcal{F}$ gives us information as to how the singularities in the data are propagated back to the reconstructed object. The microlocal analysis of singularities of the object is then done by analyzing the composition $\mathcal{C}_{\mathcal{F}^*} \circ \mathcal{C}_{\mathcal{F}}$. Such an analysis for monostatic SAR has been done by several authors [9, 10, 25, 27] and is fairly well understood. In their work [25], Nolan and Cheney showed that the composition of the linearized scattering operator with its L^2 adjoint is a singular pseudodifferential operator (Ψ DO) belonging to the class of Fourier integral operators associated with two cleanly intersecting Lagrangians [12–16, 18, 24]. Felea in her works [9, 10] further analyzed the properties of the composition of these operators. We would also like to mention the works of Yazici, Cheney, and their collaborators who have analyzed SAR imaging in a statistical framework [30–32].

In this article, we study the microlocal analysis for the bistatic SAR imaging problem for two different acquisition geometries; common offset SAR; and common midpoint SAR. We show that in each of these cases, artifacts are introduced in image reconstruction and describe the nature of these artifacts. The results presented in this article are based on the works [22] of the author done in collaboration with E.T. Quinto and [1] done in collaboration with G. Amabartsoumian, R. Felea, C. Nolan, and E.T. Quinto.

4.2 The Linearized Scattering Model

The linearized scattering model [4] has been the basis for several works on monostatic SAR imaging. Here, we derive the linearized scattering model for bistatic SAR imaging, based on slight modifications to the model derived for the monostatic case [4].

We assume that a bistatic SAR system is involved in imaging a scene. Let $\gamma_T(s)$ and $\gamma_R(s)$ for $s \in (s_0, s_1)$ be the trajectories of the transmitter and receiver,

respectively. The transmitter transmits electromagnetic waves that scatter off the target, which are then measured at the receiver. We are interested in obtaining a linearized model for this scattered signal.

The propagation of electromagnetic waves can be described by the scalar wave equation

$$\left(\Delta - \frac{1}{c^2}\partial_t^2\right)E(x,t) = -P(t)\delta(x - \gamma_T(s)), \tag{4.1}$$

where c is the speed of electromagnetic waves in the medium, $E(x, t)$ is each component of the electric field, and $P(t)$ is the transmit waveform sent to the transmitter antenna located at position $\gamma_T(s)$. The wave speed c is spatially varying due to inhomogeneities present in the medium. We assume that the background in which the electromagnetic waves propagate is free space. Therefore, c can be expressed as

$$\frac{1}{c^2(x)} = \frac{1}{c_0^2} + \tilde{V}(x),$$

where the constant c_0 is the speed of light in free space and $\tilde{V}(x)$ is the perturbation due to deviation from the background, which we would like to recover from backscattered waves.

Since the incident electromagnetic waves in typical radar frequencies attenuate rapidly as they penetrate the ground, we assume that $\tilde{V}(x)$ varies only on a two-dimensional surface. Therefore, we represent $\tilde{V}$ as a function of the form

$$\tilde{V}(x) = V(x)\delta_0(x_3),$$

where we assume for simplicity that the earth's surface is flat, represented by the $x = (x_1, x_2)$ plane.

The background Green's function g is then given by the solution to the following equation:

$$\left(\Delta - \frac{1}{c_0^2}\partial_t^2\right)g(x,t) = -\delta_0(x)\delta_0(t).$$

We can explicitly write g as

$$g(x,t) = \frac{\delta(t - |x|/c_0)}{4\pi|x|}.$$

Now, the incident field E^{in} due to the source $s(x,t) = -P(t)\delta(x - \gamma_T(s))$ is

$$\begin{aligned} E^{\text{in}}(x,t) &= -\int g(x-y, t-\tau)s(y,\tau)\mathrm{d}y\mathrm{d}\tau \\ &= \frac{P(t - |x - \gamma_T(s)|/c_0)}{4\pi|x - \gamma_T(s)|}. \end{aligned}$$

Let E denote the total field of the medium, $E = E^{\text{in}} + E^{\text{sc}}$. Then the scattered field can be written using the Lippman–Schwinger equation

$$E^{\text{sc}}(z,t) = -\int g(z-x, t-\tau)\partial_t^2 E(x,\tau)V(x)\mathrm{d}x\mathrm{d}\tau. \tag{4.2}$$

We linearize (4.2) by the first born approximation and write the linearized scattered wave field at receiver location $\gamma_R(s)$

$$\begin{aligned} E^{\text{sc}}_{\text{lin}}(\gamma_R(s),t) &= -\int g(\gamma_R(s) - x, t-\tau)\partial_t^2 E^{\text{in}}(x,\tau)V(x)\mathrm{d}x\mathrm{d}\tau \\ &= \int \frac{\delta(t - \tau - |x - \gamma_R(s)|/c_0)}{4\pi|x - \gamma_R(s)|}\left(\mathrm{e}^{-i\omega(\tau - |x-\gamma_T(s)|/c_0)}\frac{\omega^2 p(\omega)}{4\pi|x-\gamma_T(s)|}\right) \\ &\quad V(x)\mathrm{d}\omega\mathrm{d}x\mathrm{d}\tau, \end{aligned} \tag{4.3}$$

where p is the Fourier transform of P.

Now, integrating (4.3) with respect to τ, a linearized model for the scattered signal is as follows:

$$d(s,t) := E^{\text{sc}}_{\text{lin}}(\gamma_R(s),t) = \int \mathrm{e}^{-i\omega\left(t - \frac{1}{c_0}R(s,x)\right)}A(s,x,\omega)V(x)\mathrm{d}x\mathrm{d}\omega, \tag{4.4}$$

where

$$R(s,x) = |\gamma_T(s) - x| + |x - \gamma_R(s)| \tag{4.5}$$

and

$$A(s,x,\omega) = \omega^2 p(\omega)((4\pi)^2|\gamma_T(s) - x||\gamma_R(s) - x|)^{-1}.$$

This function includes terms that take into account the transmitted waveform and geometric spreading factors.

4.3 Common Offset SAR

In this section, we study the microlocal analysis of a SAR system in which the transmitter and receiver traverse a straight line above the ground offset by a constant distance at all times. All the results presented in this section are taken from the author's joint work with Quinto [22].

4.3.1 Transmitter and Receiver in a Linear Trajectory

In this section, let us assume that the trajectory of the transmitter is

$$\gamma_T : (s_0, s_1) \to \mathbb{R}^3, \quad \gamma_T(s) = (s + \alpha, 0, h)$$

and that of the receiver is

$$\gamma_R(s) : (s_0, s_1) \to \mathbb{R}^3, \quad \gamma_R(s) = (s - \alpha, 0, h).$$

Here, $\alpha > 0$ and $h > 0$ are fixed. From Eq. (4.4), the linearized model for the data at the receiver, for $s \in (s_0, s_1)$ and $t \in (t_0, t_1)$ is

$$d(s,t) = \int \mathrm{e}^{-i\omega\left(t - \frac{1}{c_0}(|x-\gamma_T(s)| + |x-\gamma_R(s)|)\right)} A(s,x,\omega) V(x) \mathrm{d}x \mathrm{d}\omega. \tag{4.6}$$

We multiply $d(s, t)$ by a smooth (infinitely differentiable) function $f(s, t)$ supported in a compact subset of $(s_0, s_1) \times (t_0, t_1)$. This compensates for the discontinuities in the measurements at the end points of the rectangle $(s_0, s_1) \times (t_0, t_1)$. For simplicity, let us denote the function $f \cdot d$ as d again. We then have

$$d(s,t) = \int \mathrm{e}^{-i\omega\left(t - \frac{1}{c_0}R(s,x)\right)} A(s,t,x,\omega) V(x) \mathrm{d}x \mathrm{d}\omega, \tag{4.7}$$

where now $A(s, t, x, \omega) = f(s, t)A(s, x, \omega)$.

Our method cannot image the point on the object that is "directly underneath" the transmitter and the receiver. That is, if the transmitter and receiver are at locations $(s + \alpha, 0, h)$ and $(s - \alpha, 0, h)$, then we cannot image the point $(s, 0, 0)$; see Remark 3.2. Therefore, we modify d in Eq. (4.7) by multiplying by another smooth function $g(s, t)$ such that

$$g \equiv 0 \quad \text{in a small neighborhood of} \left\{ \left(s, 2\frac{\sqrt{\alpha^2 + h^2}}{c_0} \right) : s_0 < s < s_1 \right\}. \tag{4.8}$$

For simplicity, again denote $g \cdot d$ as d and $g \cdot A$ as A. Consider,

$$\mathcal{F}_{co}V(s,t) := d(s,t) = \int \mathrm{e}^{-i\omega(t-\frac{1}{c_0}(|x-\gamma_T(s)|+|x-\gamma_R(s)|))}A(s,t,x,\omega)V(x)\mathrm{d}x\mathrm{d}\omega. \quad (4.9)$$

The subscript in $\mathcal{F}_{co}$ stands for common offset. For simplicity, let us denote the (s, t) space as Y.

We assume that the amplitude function A satisfies the following estimate: For every compact $K \subset Y \times X$ and for every nonnegative integer α and for every two-indexes $\beta = (\beta_1, \beta_2)$ and γ, there is a constant C such that

$$|\partial_\omega^\alpha \partial_s^{\beta_1} \partial_t^{\beta_2} \partial_x^\gamma A(s,t,x,\omega)| \leq C(1+|\omega|)^{2-\alpha}.$$

This assumption is satisfied if the transmitted waveform P in (4.1) which is approximately a Dirac delta distribution.

The phase function of the operator $\mathcal{F}_{co}$,

$$\psi(s,t,x,\omega) = -\omega\left(t - \frac{1}{c_0}(|x-\gamma_T(s)|+|x-\gamma_R(s)|)\right) \quad (4.10)$$

is positively homogeneous of degree 1 in ω.

We now analyze some properties of the canonical relation of the operator $\mathcal{F}_{co}$.

Proposition 3.1 *$\mathcal{F}_{co}$ is a Fourier integral operator of order 3/2 with canonical relation*

$$\begin{aligned}\mathcal{C}_{co} = \Bigg\{ &\left(s,t,-\frac{\omega}{c_0}\left(\frac{x_1-s-\alpha}{|x-\gamma_T(s)|}+\frac{x_1-s+\alpha}{|x-\gamma_R(s)|}\right),-\omega\right); \\ &\left(x_1,x_2,-\frac{\omega}{c_0}\left(\frac{x_1-s-\alpha}{|x-\gamma_T(s)|}+\frac{x_1-s+\alpha}{|x-\gamma_R(s)|}\right),-\frac{\omega}{c_0}\left(\frac{x_2}{|x-\gamma_T(s)|}+\frac{x_2}{|x-\gamma_R(s)|}\right)\right) \\ &: c_0 t = \sqrt{(x_1-s-\alpha)^2+x_2^2+h^2}+\sqrt{(x_1-s+\alpha)^2+x_2^2+h^2}, \quad \omega \neq 0\Bigg\}.\end{aligned} \quad (4.11)$$

Furthermore, (x_1, x_2, s, ω) is a global parameterization for $\mathcal{C}_{co}$.

Remark 3.2 Recall that we modified the amplitude function A to be 0 in a neighborhood of points "directly underneath the transmitter and receiver"; see (4.8). The exclusion of such points is required, as can be seen in the definition of the canonical relation (4.11) above. If the transmitter and receiver positions are $(s+\alpha, 0, h)$ and $(s-\alpha, 0, h)$, respectively, then for $(x_1, x_2) = (s, 0)$, the cotangent vector in the canonical relation corresponding to the point $(s, 0)$ is 0. Therefore, by making A to be 0 in a neighborhood of such points, we exclude a neighborhood of such points from the canonical relation in our analysis.

Proof This is a straightforward application of the theory of FIO. Since ψ in (4.10) is a nondegenerate phase function with $\partial_x\psi$ and $\partial_{s,t}\psi$ nowhere zero and the

amplitude A in (4.9) is of order 2, $\mathcal{F}$ is an FIO [19]. Since the amplitude is of order 2, the order of the FIO is 3/2 by [19, Definition 3.2.2]. By definition [19, Eq. (3.1.2)]

$$\mathcal{C}_{co} = \{(s,t,\partial_{s,t}\psi(x,s,t,\omega)),(x,-\partial_x\psi(x,s,t)) : \partial_\omega\psi(x,s,t,\omega) = 0\}.$$

A calculation using this definition establishes (4.11). Finally, it is easy to see that (x_1, x_2, s, ω) is a global parameterization of $\mathcal{C}_{co}$. □

In order to understand the microlocal mapping properties of $\mathcal{F}_{co}$ and $\mathcal{F}_{co}^*\mathcal{F}_{co}$, we consider the projections $\pi_L : T^*Y \times T^*X \to T^*Y$ and $\pi_R : T^*Y \times T^*X \to T^*X$. A good reference for mappings with singularities is the book [11].

Proposition 3.3 *The projection π_L restricted to $\mathcal{C}_{co}$ has a fold singularity on the set $\Sigma := \{(x_1, 0, s, \omega) : \omega \neq 0\}$.*

Proof The projection π_L is given by

$$\begin{aligned}&\pi_L(x_1,x_2,s,\omega)\\ &= \left(s, \frac{1}{c_0}(|x-\gamma_T(s)| + |x-\gamma_R(s)|), -\frac{\omega}{c_0}\left(\frac{x_1-s-\alpha}{|x-\gamma_T(s)|} + \frac{x_1-s+\alpha}{|x-\gamma_R(s)|}\right), -\omega\right)\end{aligned} \tag{4.12}$$

We have

$$\mathrm{d}\pi_L = \begin{pmatrix} 0 & 0 & 1 & 0 \\ \frac{1}{c_0}\left(\frac{x_1-s-\alpha}{|x-\gamma_T(s)|} + \frac{x_1-s+\alpha}{|x-\gamma_R(s)|}\right) & \frac{1}{c_0}\left(\frac{x_2}{|x-\gamma_T(s)|} + \frac{x_2}{|x-\gamma_R(s)|}\right) & * & 0 \\ -\frac{\omega}{c_0}\left(\frac{x_2^2+h^2}{|x-\gamma_T(s)|^3} + \frac{x_2^2+h^2}{|x-\gamma_R(s)|^3}\right) & \frac{\omega}{c_0}\left(\frac{(x_1-s-\alpha)x_2}{|x-\gamma_T(s)|^3} + \frac{(x_1-s+\alpha)x_2}{|x-\gamma_R(s)|^3}\right) & * & * \\ 0 & 0 & 0 & -1 \end{pmatrix}. \tag{4.13}$$

Then,

$$\det(\mathrm{d}\pi_L) = \frac{\omega}{c_0^2}x_2\left(\frac{1}{|x-\gamma_T(s)|^2} + \frac{1}{|x-\gamma_R(s)|^2}\right)\left(1 + \frac{(x_1-s)^2 + x_2^2 + h^2 - \alpha^2}{|x-\gamma_T(s)||x-\gamma_R(s)|}\right).$$

Now, Proposition 3.3 follows as a consequence of the following lemma. □

Lemma 3.4 *The term*

$$1 + \frac{(x_1-s)^2 + x_2^2 + h^2 - \alpha^2}{|x-\gamma_T(s)||x-\gamma_R(s)|}.$$

is positive for all $x \in \mathbb{R}^2$, $s \in \mathbb{R}$ *and* h *and* α *positive.*

Proof The second term on the right above is the angle between the vectors $(x_1 - s - \alpha, x_2, -h)$ and $(x_1 - s + \alpha, x_2, -h)$. Since these vectors are never parallel (due to $\alpha > 0$ and $h > 0$), we have that $1 + \frac{(x_1-s)^2 + x_2^2 + h^2 - \alpha^2}{|x-\gamma_T(s)||x-\gamma_R(s)|} > 0$. □

Now returning to the proof of the Proposition 3.3, we have that $\det(\mathrm{d}\pi_L) = 0$ if and only if $x_2 = 0$. Hence, $\det(\mathrm{d}\pi_L)$ vanishes on the set Σ and Lemma 3.4 again shows that $\mathrm{d}(\det(\mathrm{d}\pi_L))$ on Σ is nonvanishing. This implies that π_L drops rank by one simply on Σ. Alternately, one can also see that $(\mathrm{d}\pi_L)|_\Sigma$ has rank 3 by letting $x_2 = 0$ in (4.13). Furthermore, $\mathrm{d}\pi_L$ has full rank except on Σ, because $\det(\mathrm{d}\pi_L)$ is nonvanishing except on Σ.

Now, it remains to show that $T\Sigma \cap \mathrm{Kernel}\,(\mathrm{d}\pi_L) = \{0\}$. This follows from the fact that $\mathrm{kernel}\,(\mathrm{d}\pi_L) = \mathrm{span}\left(\frac{\partial}{\partial x_2}\right)$, but $T\Sigma = \mathrm{span}\left(\frac{\partial}{\partial x_1}, \frac{\partial}{\partial s}, \frac{\partial}{\partial \omega}\right)$. This concludes the proof of Proposition 3.3. □

Proposition 3.5 *Consider the projection* $\pi_R : T^*Y \times T^*X \to T^*X$. *The restriction of the projection to* $\mathcal{C}_{co}$ *has a blowdown singularity on* Σ.

Proof We have

$$\begin{aligned}&\pi_R(x_1, x_2, s, \omega)\\ &= \left(x_1, x_2, -\frac{\omega}{c_0}\left(\frac{x_1 - s - \alpha}{|x - \gamma_T(s)|} + \frac{x_1 - s - \alpha}{|x - \gamma_R(s)|}\right), -\frac{\omega}{c_0}\left(\frac{x_2}{|x - \gamma_T(s)|} + \frac{x_2}{|x - \gamma_R(s)|}\right)\right).\end{aligned} \tag{4.14}$$

Now,

$$\mathrm{d}\pi_R = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ * & * & -\frac{\omega}{c_0}\left(\frac{x_2^2 + h^2}{|x-\gamma_T(s)|^3} + \frac{x_2^2 + h^2}{|x-\gamma_R(s)|^3}\right) & -\frac{1}{c_0}\left(\frac{x_1 - s - \alpha}{|x-\gamma_T(s)|} + \frac{x_1 - s + \alpha}{|x-\gamma_R(s)|}\right) \\ * & * & \frac{\omega}{c_0}\left(\frac{(x_1 - s - \alpha)x_2}{|x-\gamma_T(s)|^3} + \frac{(x_1 - s + \alpha)x_2}{|x-\gamma_R(s)|^3}\right) & -\frac{1}{c_0}\left(\frac{x_2}{|x-\gamma_T(s)|} + \frac{x_2}{|x-\gamma_R(s)|}\right) \end{pmatrix}.$$

From this we see that $\mathrm{kernel}(\mathrm{d}\pi_R) \subset T\Sigma$. Since $\det(\mathrm{d}\pi_R) = \det(\mathrm{d}\pi_L)$, π_R drops rank by one simply along Σ. Therefore, the projection π_R has a blowdown singularity along Σ. □

We summarize what we have proved in this section by the following theorem:

Theorem 3.6 *The operator* $\mathcal{F}_{co}$ *defined in* (4.9) *is a Fourier integral operator of order 3/2. The canonical relation* $\mathcal{C}_{co}$ *associated to* $\mathcal{F}_{co}$ *defined in* (4.11) *satisfies the following: The projections* π_L *and* π_R *defined in* (4.12) *and* (4.14) *are a fold and blowdown, respectively.*

4.3.2 Image Reconstruction

Next, we study the composition of $\mathcal{F}_{co}$ with $\mathcal{F}_{co}^*$. This composition is given as follows:

$$\mathcal{F}_{co}^*\mathcal{F}_{co}V(x) = \int e^{i\left(\omega(t-\frac{1}{c_0}(|x-\gamma_R(s)|+|x-\gamma_R(s)|)) - \tilde{\omega}(t-\frac{1}{c_0}(|y-\gamma_T(s)|+|y-\gamma_R(s)|))\right)} \times \overline{A(x,s,t,\omega)}A(y,s,t,\tilde{\omega})V(y)\mathrm{d}s\mathrm{d}t\mathrm{d}\omega\mathrm{d}\tilde{\omega}\mathrm{d}y.$$

After an application of the method of stationary phase [17], we can write the kernel of the operator $\mathcal{F}_{co}^*\mathcal{F}_{co}$ as

$$K_{co}(x,y) = \int e^{i\frac{\omega}{c_0}(|y-\gamma_T(s)|+|y-\gamma_R(s)|-(|x-\gamma_T(s)|+|x-\gamma_R(s)|))}\tilde{A}(x,y,s,\omega)\mathrm{d}s\mathrm{d}\omega.$$

The phase function of the kernel $K_{co}(x,y)$ is

$$\phi(x,y,s,\omega) = \frac{\omega}{c_0}(|y-\gamma_T(s)|+|y-\gamma_R(s)|-(|x-\gamma_T(s)|+|x-\gamma_R(s)|)). \quad (4.15)$$

Let us denote the wavefront set [20] of a distribution f by $WF(f)$ and the twisted wavefront set by $WF'(f)$ [20].

Theorem 3.7

$$WF(K_{co})' \subset \Delta \cup \Lambda,$$

where $\Delta := \{(x_1,x_2,\xi_1,\xi_2;x_1,x_2,\xi_1,\xi_2)\}$ *and* $\Lambda := \{(x_1,x_2,\xi_1,\xi_2;x_1,-x_2,\xi_1,-\xi_2)\}$. *Here, for a point* $x=(x_1,x_2)$, (ξ_1,ξ_2) *are nonzero multiples of the vector* $(-\partial_{x_1}R(s,x),-\partial_{x_2}R(s,x))$, *where R is defined in* (4.5).

Proof Using the Hörmander–Sato Lemma [19], we have

$$WF(K_{co})' \subset$$

$$\left\{\left(x_1,x_2,-\frac{\omega}{c_0}\left(\frac{x_1-s-\alpha}{|x-\gamma_T(s)|}+\frac{x_1-s+\alpha}{|x-\gamma_R(s)|}\right),-\frac{\omega}{c_0}\left(\frac{x_2}{|x-\gamma_T(s)|}+\frac{x_2}{|x-\gamma_R(s)|}\right)\right);\right.$$
$$\left(y_1,y_2,-\frac{\omega}{c_0}\left(\frac{y_1-s-\alpha}{|y-\gamma_T(s)|}+\frac{y_1-s+\alpha}{|y-\gamma_R(s)|}\right),-\frac{\omega}{c_0}\left(\frac{y_2}{|y-\gamma_T(s)|}+\frac{y_2}{|y-\gamma_R(s)|}\right)\right):$$
$$|x-\gamma_T(s)|+|x-\gamma_R(s)| = |y-\gamma_T(s)|+|y-\gamma_R(s)|,$$
$$\left.\frac{x_1-s-\alpha}{|x-\gamma_T(s)|}+\frac{x_1-s+\alpha}{|x-\gamma_R(s)|} = \frac{y_1-s-\alpha}{|y-\gamma_T(s)|}+\frac{y_1-s+\alpha}{|y-\gamma_R(s)|}, \quad \omega \neq 0\right\}.$$

We now obtain a relation between (x_1, x_2) and (y_1, y_2). This is given by the following lemma. □

Lemma 3.8 *For all s, the set of all* (x_1, x_2), (y_1, y_2) *that satisfy*

$$|x - \gamma_T(s)| + |x - \gamma_R(s)| = |y - \gamma_T(s)| + |y - \gamma_R(s)|, \tag{4.16}$$

$$\frac{x_1 - s - \alpha}{|x - \gamma_T(s)|} + \frac{x_1 - s + \alpha}{|x - \gamma_R(s)|} = \frac{y_1 - s - \alpha}{|y - \gamma_T(s)|} + \frac{y_1 - s + \alpha}{|y - \gamma_R(s)|}. \tag{4.17}$$

necessarily satisfy the following relations: $x_1 = y_1$ *and* $x_2 = \pm y_2$.

Proof In order to show this, we will consider (4.16) and (4.17) as functions of $\mathbb{R}^3$ by replacing h in these expressions with $x_3 - h$. We then transform these expressions using the coordinates (4.18) and then set $x_3 = y_3 = 0$ to prove the lemma.

Consider the following change of coordinates, the so called prolate spheroidal coordinates:

$$\begin{aligned} x_1 &= s + \alpha \cosh\rho \cos\theta & y_1 &= s + \alpha \cosh\rho' \cos\theta' \\ x_2 &= \alpha \sinh\rho \sin\theta \cos\varphi & y_2 &= \alpha \sinh\rho' \sin\theta' \cos\varphi' \\ x_3 &= h + \alpha \sinh\rho \sin\theta \sin\varphi & y_3 &= h + \alpha \sinh\rho' \sin\theta' \sin\varphi' \end{aligned} \tag{4.18}$$

where s, $\alpha > 0$, and $h > 0$ are fixed and $\rho \in [0, \infty)$, $\theta \in [0, \pi]$, and $\varphi \in [0, 2\pi)$. This a well-defined coordinate system except for $\rho = 0$ and $\theta = 0, \pi$.

This coordinate system has also been used in the context of radar imaging by T. Dowling in his thesis [7].

In the coordinate system (4.18), we have

$$\begin{aligned} &|x - \gamma_T(s)| = \alpha(\cosh\rho - \cos\theta), \quad \tfrac{x_1 - s - \alpha}{|x - \gamma_T(s)|} = \tfrac{\cosh\rho\cos\theta - 1}{\cosh\rho - \cos\theta}, \\ &|x - \gamma_R(s)| = \alpha(\cosh\rho + \cos\theta), \quad \tfrac{x_1 - s + \alpha}{|x - \gamma_R(s)|} = \tfrac{\cosh\rho\cos\theta + 1}{\cosh\rho + \cos\theta}. \end{aligned} \tag{4.19}$$

The terms involving y are obtained similarly. Now (4.16) and (4.17) transform as follows:

$$2\cosh\rho = 2\cosh\rho\prime$$

$$\frac{\cosh\rho\cos\theta - 1}{\cosh\rho - \cos\theta} + \frac{\cosh\rho\cos\theta + 1}{\cosh\rho + \cos\theta} = \frac{\cosh\rho'\cos\theta' - 1}{\cosh\rho' - \cos\theta'} + \frac{\cosh\rho'\cos\theta' + 1}{\cosh\rho' + \cos\theta'}.$$

Using the first equality in the second equation, we have

$$\frac{\cos\theta}{\cosh^2\rho - \cos^2\theta} = \frac{\cos\theta'}{\cosh^2\rho - \cos^2\theta'}.$$

This gives $\cos\theta = \cos\theta'$. Therefore, $\theta = 2n\pi \pm \theta'$, which then gives $\sin\theta = \pm\sin\theta'$. Therefore, in terms of (x_1, x_2) and (y_1, y_2), we have $(x_1 = y_1)$ and $(x_2 = \pm y_2)$. □

Now, to finish the proof of Theorem 3.7, when $x_1 = y_1$ and $x_2 = y_2$, there is contribution to $WF(K_{co})'$ contained in the diagonal set given by $\Delta := \{(x_1, x_2, \xi_1, \xi_2; x_1, x_2, \xi_1, \xi_2)\}$ and when $x_I = y_1$ and $x_1 = -y_2$, we have a contribution to $WF(K_{co})'$ contained in Λ, where $\Lambda := \{(x_1, x_2, \xi_1, \xi_2; x_1, -x_2, \xi_1, -\xi_2)\}$. □

From an imaging point of view, the result of Theorem 3.7 shows that artifacts are introduced in image reconstruction. The true singularities are given by the Lagrangian Δ and the artifact singularities are given by the Lagrangian Λ. A more detailed analysis, as done in [22], shows that the strengths of the true singularities are the same as that of the artifacts. We do not provide the details of this analysis here, but instead refer to our work [22].

4.4 Common Midpoint SAR

In this section, we assume that both the transmitter and receiver are at the same height $h > 0$ above the ground, $x_3 = 0$, at all times and move in opposite directions at equal speeds along the line parallel to the x_1 axis and containing the common midpoint (0, 0, h). Such a model arises when considering signals which have been scattered from a wall within the vicinity of a scatterer and can be understood in the context of the method of images; see [26] for more details. The material in this section is taken from the author's joint work with Ambartsoumian et al. [1].

Let $\gamma_T(s) = (s, 0, h)$ and $\gamma_R(s) = (-s, 0, h)$ for $s \in (0, \infty)$ be the trajectories of the transmitter and receiver, respectively.

The linearized model for the scattered signal we will use in this article is from [26]

$$d(s,t) := \mathcal{F}_{cm} V(s,t) = \int \mathrm{e}^{-i\omega(t-\frac{1}{c_0}R(s,x))} a(s,x,\omega) V(x) \mathrm{d}x \mathrm{d}\omega$$

for $(s, t) \in Y = (0, \infty) \times (0, \infty)$. As in the common offset case, the subscript in $\mathcal{F}_{cm}$ refers to common midpoint. Here, $V(x) = V(x_1, x_2)$ is the function modeling the object on the ground, $R(s, x)$ is the bistatic distance

$$R(s,x) = |\gamma_T(s) - x| + |x - \gamma_R(s)|,$$

c_0 is the speed of electromagnetic wave in free space and the amplitude term a is given by

$$a(s,x,\omega) = \frac{\omega^2 p(\omega)}{16\pi^2 |\gamma_T(s) - x||\gamma_R(s) - x|},$$

where p is the Fourier transform of the transmitted waveform.

4.4.1 Preliminary Modifications on the Scattered Data

For simplicity, from now on we will assume that $c_0 = 1$. To make the composition of $\mathcal{F}_{cm}$ with its L^2 adjoint $\mathcal{F}_{cm}^*$ to be well-defined, we multiply $d(s, t)$ by an infinitely differentiable function $f(s, t)$ identically equal to 1 in a compact subset of $(0, \infty) \times (0, \infty)$ and supported in a slightly bigger compact subset of $(0, \infty) \times (0, \infty)$. We rename $f \cdot d$ as d again.

As we will see below, our method cannot image a neighborhood of the common midpoint. That is, if the transmitter and receiver are at $(s, 0, h)$ and $(-s, 0, h)$, respectively, we cannot image a neighborhood of the origin on the horizontal plane of the earth, $x_3 = 0$. Therefore, we modify d further by considering a smooth function $g(s, t)$ such that

$$g(s,t) = 0 \text{ for } (s,t) : |t - 2\sqrt{s^2 + h^2}| < 20\varepsilon^2/h, \tag{4.20}$$

where $\varepsilon > 0$ is given. Again we let $g \cdot d$ to be d and $g \cdot a$ to be a. Our forward operator is

$$\mathcal{F}_{cm}V(s,t) = \int \mathrm{e}^{-i\varphi(s,t,x,\omega)} a(s,t,x,\omega) V(x)\, \mathrm{d}x\mathrm{d}\omega \tag{4.21}$$

where

$$\varphi(s,t,x,\omega) = \omega\left(t - \sqrt{(x_1 - s)^2 + x_2^2 + h^2} - \sqrt{(x_1 + s)^2 + x_2^2 + h^2}\right). \tag{4.22}$$

From now on, we will denote the ground (the plane $x_3 = 0$) by X, thus the points on X will be denoted as $x = (x_1, x_2)$.

We assume that the amplitude function $a \in S^{m+\frac{1}{2}}$, i.e., it satisfies the following estimate: For every compact set $K \subset Y \times X$, nonnegative integer α, and two-indexes $\beta = (\beta_1, \beta_2)$ and γ, there is a constant C such that

$$|\partial_\omega^\alpha \partial_s^{\beta_1} \partial_t^{\beta_2} \partial_x^\gamma a(s,t,x,\omega)| \le C(1 + |\omega|)^{m + (1/2) - \alpha}. \tag{4.23}$$

This assumption is satisfied if the transmitted waveform from the antenna is approximately a Dirac delta distribution.

With these modifications, we show that $\mathcal{F}_{cm}$ is a Fourier integral operator of order m and study the properties of the natural projection maps from the canonical relation of $\mathcal{F}_{cm}$.

4.4.2 Analysis of the Operator $\mathcal{F}_{cm}$

In this section, we prove the following Theorem 4.1, the proof of which is in Lemma 4.4 and Proposition 4.6.

Theorem 4.1 *Let $\mathcal{F}_{cm}$ be as in (4.21). Then,*

(1) *$\mathcal{F}_{cm}$ is an FIO of order m.*
(2) *The canonical relation $\mathcal{C}_{cm}$ associated to $\mathcal{F}_{cm}$ is given by*

$$\begin{aligned}\mathcal{C}_{cm} = \Bigg\{ \Bigg(& s, t, -\omega\left(\frac{x_1 - s}{\sqrt{(x_1 - s)^2 + x_2^2 + h^2}} - \frac{x_1 + s}{\sqrt{(x_1 + s)^2 + x_2^2 + h^2}}\right), -\omega; \\ & x_1, x_2, -\omega\left(\frac{x_1 - s}{\sqrt{(x_1 - s)^2 + x_2^2 + h^2}} + \frac{x_1 + s}{\sqrt{(x_1 + s)^2 + x_2^2 + h^2}}\right), \\ & -\omega\left(\frac{x_2}{\sqrt{(x_1 - s)^2 + x_2^2 + h^2}} + \frac{x_2}{\sqrt{(x_1 + s)^2 + x_2^2 + h^2}}\right)\Bigg) : \\ & s > 0, t = \sqrt{(x_1 - s)^2 + x_2^2 + h^2} + \sqrt{(x_1 + s)^2 + x_2^2 + h^2}, \\ & x \neq 0 \text{ and } \omega \neq 0\Bigg\},\end{aligned} \tag{4.24}$$

and $\mathcal{C}_{cm}$ has global parameterization

$$(0, \infty) \times \left(\mathbb{R}^2 \backslash 0\right) \times (\mathbb{R} \backslash 0) \ni (s, x_1, x_2, \omega) \mapsto C.$$

(3) *Let $\pi_L : \mathcal{C}_{cm} \to T^*Y$ and $\pi_R : \mathcal{C}_{cm} \to T^*X$ be the left and right projections, respectively. Then, π_L and π_R drop rank simply by one on a set $\Sigma = \Sigma_1 \cup \Sigma_2$ where in the coordinates (s, x, ω), $\Sigma_1 = \{(s, x_1, 0, \omega) | s > 0, |x_1| > \varepsilon', \omega \neq 0\}$, and $\Sigma_2 = \{(s, 0, x_2, \omega) | s > 0, |x_2| > \varepsilon', \omega \neq 0\}$ for $0 < \varepsilon'$ small enough.*
(4) *π_L has a fold singularity along Σ.*
(5) *π_R has a blowdown singularity along Σ.*

Remark 4.2 Note that due to the function $g(s, t)$ of (4.20) in the amplitude, it is enough to consider only points in $\mathcal{C}_{cm}$ that are strictly away from $\{(s, 0, \omega) : s > 0, \omega \neq 0\}$. This is reflected in the definitions of Σ_1 and Σ_2, where $|x_1|$ and $|x_2|$, respectively, are strictly positive.

Remark 4.3 Note that $\mathcal{C}_{cm}$ is even with respect to both x_1 and x_2. In other words, $\mathcal{C}_{cm}$ is a four-to-one relation. This observation suggests that π_L (respectively, π_R) has two fold (respectively, blowdown) sets. See Proposition 4.6.

Lemma 4.4 *$\mathcal{F}_{cm}$ is an FIO of order m with the canonical relation $\mathcal{C}_{cm}$ given by*

$$\begin{aligned}\mathcal{C}_{cm} = \Bigg\{ \Bigg(& s, t, -\omega \left(\frac{x_1 - s}{\sqrt{(x_1 - s)^2 + x_2^2 + h^2}} - \frac{x_1 + s}{\sqrt{(x_1 + s)^2 + x_2^2 + h^2}} \right), -\omega; \\ & x_1, x_2, -\omega \left(\frac{x_1 - s}{\sqrt{(x_1 - s)^2 + x_2^2 + h^2}} + \frac{x_1 + s}{\sqrt{(x_1 + s)^2 + x_2^2 + h^2}} \right), \\ & -\omega \left(\frac{x_2}{\sqrt{(x_1 - s)^2 + x_2^2 + h^2}} + \frac{x_2}{\sqrt{(x_1 + s)^2 + x_2^2 + h^2}} \right) \Bigg) : \\ & s > 0, t = \sqrt{(x_1 - s)^2 + x_2^2 + h^2} + \sqrt{(x_1 + s)^2 + x_2^2 + h^2}, \\ & x \in \mathbb{R}^2 \backslash \{0\}, \omega \neq 0 \Big\}. \end{aligned} \tag{4.25}$$

We note that $(0, \infty) \times (\mathbb{R}^2 \backslash 0) \times (\mathbb{R} \backslash 0) \ni (s, x_1, x_2, \omega) \mapsto \mathcal{C}_{cm}$ is a global parametrization of $\mathcal{C}_{cm}$.

We will use the coordinates (s, x, ω) in this lemma from now on to describe $\mathcal{C}_{cm}$ and subsets of $\mathcal{C}_{cm}$.

Proof The phase function φ (Eq. 4.22) is nondegenerate with $\partial_x \varphi$, $\partial_{s,t} \varphi$ nowhere 0 whenever $\partial_\omega \varphi = 0$. We should mention that $\nabla \partial_\omega \varphi \neq 0$. (Note that in order for $\partial_x \varphi$ to be nowhere 0, we require exclusion of the common midpoint from our analysis). This observation is needed to show that $\mathcal{F}$ is a FIO rather than just a Fourier integral distribution. Recalling that a satisfies amplitude estimates (4.23), we conclude that $\mathcal{C}_{cm}$ is an FIO [29]. Also, since a is of order $m + \frac{1}{2}$, the order of the FIO is m [8, Definition 3.2.2]. By definition, [19, Eq. (3.1.2)]

$$\mathcal{C}_{cm} = \{((s, t, \partial_s \varphi, \partial_t \varphi); (x, -\partial_x \varphi)) : \partial_\omega \varphi = 0\}.$$

A calculation using this definition establishes (4.25). Furthermore, it is easy to see that (s, x_1, x_2, ω) is a global parametrization of $\mathcal{C}_{cm}$. □

Remark 4.5 In the SAR application, a has order 2 which makes operator $\mathcal{F}_{cm}$ of order $\frac{3}{2}$. But from now on, we will consider that $\mathcal{F}_{cm}$ has order m.

Proposition 4.6 *Denoting the restriction of the left and right projections to $\mathcal{C}_{cm}$ by π_L and π_R, respectively, we have*

(1) *π_L and π_R drop rank by one on a set $\Sigma = \Sigma_1 \cup \Sigma_2$. Here, we use the global coordinates from Lemma 4.4.*
(2) *π_L has a fold singularity along Σ.*
(3) *π_R has a blowdown singularity along Σ.*

Proof Let $A = \sqrt{(x_1 - s)^2 + x_2^2 + h^2}$ and $B = \sqrt{(x_1 + s)^2 + x_2^2 + h^2}$. We have

$$\pi_L(x_1, x_2, s, \omega) = \left(s, A + B, -\left(\frac{x_1 - s}{A} - \frac{x_1 + s}{B}\right)\omega, -\omega\right)$$

and

$$\mathrm{d}\pi_L = \begin{pmatrix} 0 & 0 & 1 & 0 \\ \frac{x_1 - s}{A} + \frac{x_1 + s}{B} & \frac{x_2}{A} + \frac{x_2}{B} & * & 0 \\ -\omega\left(\frac{x_2^2 + h^2}{A^3} - \frac{x_2^2 + h^2}{B^3}\right) & \omega\left(\frac{(x_1 - s)x_2}{A^3} - \frac{(x_1 + s)x_2}{B^3}\right) & * & * \\ 0 & 0 & 0 & -1 \end{pmatrix}$$

where * denotes derivatives that are not needed for the calculation. The determinant is

$$\det \mathrm{d}\pi_L = \frac{4x_1 x_2 s\omega}{A^2 B^2}\left(1 + \frac{(x_1^2 - s^2 + x_2^2 + h^2)}{AB}\right) \tag{4.26}$$

We have that $s > 0$ and the number in the parenthesis is a positive number by Lemma 4.7.

Therefore, π_L drops rank by one on $\Sigma = \Sigma_1 \cup \Sigma_2$. To show that $\mathrm{d}(\det(\mathrm{d}\pi_L))$ is nowhere 0 on Σ, one uses the product rule in (4.26) and the fact that the differential of $\frac{4x_1 x_2 s\omega}{A^2 B^2}$ is never 0 on Σ and the inequality in Lemma 4.7.

On Σ_1, the kernel of $\mathrm{d}\pi_L$ is $\frac{\partial}{\partial x_2}$ which is transversal to Σ_1 and on Σ_2 the kernel of $\mathrm{d}\pi_L$ is $\frac{\partial}{\partial x_1}$, which is transversal to Σ_2. This means that π_L has a fold singularity along Σ.

Similarly,

$$\pi_R(x_1, x_2, s, \omega) = \left(x_1, x_2, -\left(\frac{x_1 - s}{A} + \frac{x_1 + s}{B}\right)\omega, -\left(\frac{x_2}{A} + \frac{x_2}{B}\right)\omega\right).$$

Then,

$$\mathrm{d}\pi_R = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ * & * & \omega\left(\frac{x_2^2 + h^2}{A^3} - \frac{x_2^2 + h^2}{B^3}\right) & -\left(\frac{x_1 - s}{A} + \frac{x_1 + s}{B}\right) \\ * & * & -\omega\left(\frac{(x_1 - s)x_2}{A^3} - \frac{(x_1 + s)x_2}{B^3}\right) & -\left(\frac{x_2}{A} + \frac{x_2}{B}\right) \end{pmatrix}$$

has the same determinant so that π_R drops rank by one on Σ and the kernel of $\mathrm{d}\pi_R$ is a linear combination of $\frac{\partial}{\partial \omega}$ and $\frac{\partial}{\partial s}$ which are tangent to both Σ_1 and Σ_2. This means that π_R has a blowdown singularity along Σ. □

Lemma 4.7 *For all* $s \neq 0$,

$$1 + \frac{x_1^2 - s^2 + x_2^2 + h^2}{|x - \gamma_T(s)||x - \gamma_R(s)|} > 0.$$

Proof The proof is similar to that of Lemma 3.4 and is left to the reader. □

4.4.3 Analysis of the Normal Operator $\mathcal{F}_{cm}^* \mathcal{F}_{cm}$

We have

$$\mathcal{F}_{cm}^* \mathcal{F}_{cm} V(x) = \int \mathrm{e}^{i\omega(t-(|x-\gamma_T(s)|+|x-\gamma_R(s)|))-\widetilde{\omega}(t-(|y-\gamma_T(s)|+|y-\gamma_R(s)|))} \overline{a(s,t,x,\omega)} a(s,t,y,\tilde{\omega}) V(y) \mathrm{d}s\mathrm{d}t\mathrm{d}\omega\mathrm{d}\tilde{\omega}\mathrm{d}y.$$

After an application of the method of stationary phase [17] in t and $\tilde{\omega}$, the Schwartz kernel of this operator is

$$K_{cm}(x,y) = \int \mathrm{e}^{i\omega(|y-\gamma_T(s)|+|y-\gamma_R(s)|-(|x-\gamma_T(s)|+|x-\gamma_R(s)|))} \tilde{a}(x,y,s,\omega)\, \mathrm{d}s\mathrm{d}\omega. \tag{4.27}$$

Note that $\tilde{a} \in S^{2m+1}$ since we assume $a \in S^{m+1/2}$.

Let the phase function of the kernel K_{cm} be denoted as

$$\Phi = \omega(|y - \gamma_T(s)| + |y - \gamma_R(s)| - (|x - \gamma_T(s)| + |x - \gamma_R(s)|)). \tag{4.28}$$

Theorem 4.8 *The wavefront set of the kernel* K_{cm} *of* $\mathcal{F}_{cm}^* \mathcal{F}_{cm}$ *satisfies,*

$$WF(K_{cm})' \subset \Delta \cup C_1 \cup C_2 \cup C_3,$$

where Δ *is the diagonal in* $T^*X \times T^*X$ *and the Lagrangians* C_i *for* $i = 1, 2, 3$ *are the graphs of the following functions* χ_i *for* $i = 1, 2, 3$ *on* T^*X

$$\begin{aligned}\chi_1(x,\xi) =& (x_1, -x_2, \xi_1, -\xi_2), \chi_2(x,\xi) \\ =& (-x_1, x_2, -\xi_1, \xi_2) \text{ and } \chi_3 = \chi_1 \circ \chi_2.\end{aligned}$$

Proof In order to find the wavefront set of the kernel K, we consider the canonical relation $\mathcal{C}_{cm}^t \circ \mathcal{C}_{cm}$ of $\mathcal{F}_{cm}^* \mathcal{F}_{cm}$: $\mathcal{C}_{cm}^t \circ \mathcal{C}_{cm} = \{(x,\xi;y,\eta) | (x,\xi;s,t,\sigma,\tau) \in \mathcal{C}_{cm}^t; (s,t,\sigma,\tau;y,\eta) \in \mathcal{C}_{cm}\}$. We have that $(s,t,\sigma,\tau;y,\eta) \in \mathcal{C}_{cm}$ implies

$$\begin{aligned} t &= \sqrt{(y_1-s)^2+y_2^2+h^2}+\sqrt{(y_1+s)^2+y_2^2+h^2} \\ \sigma &= \tau\left(\frac{y_1-s}{\sqrt{(y_1-s)^2+y_2^2+h^2}}-\frac{y_1+s}{\sqrt{(y_1+s)^2+y_2^2+h^2}}\right) \\ \eta_1 &= \tau\left(\frac{y_1-s}{\sqrt{(y_1-s)^2+y_2^2+h^2}}+\frac{y_1+s}{\sqrt{(y_1+s)^2+y_2^2+h^2}}\right) \\ \eta_2 &= \tau\left(\frac{y_2}{\sqrt{(y_1-s)^2+y_2^2+h^2}}+\frac{y_2}{\sqrt{(y_1+s)^2+y_2^2+h^2}}\right) \end{aligned} \tag{4.29}$$

and $(x,\xi;s,t,\sigma,\tau)\in\mathcal{C}^t_{cm}$ implies

$$\begin{aligned} t &= \sqrt{(x_1-s)^2+x_2^2+h^2}+\sqrt{(x_1+s)^2+x_2^2+h^2} \\ \sigma &= \tau\left(\frac{x_1-s}{\sqrt{(x_1-s)^2+x_2^2+h^2}}-\frac{x_1+s}{\sqrt{(x_1+s)^2+x_2^2+h^2}}\right) \\ \xi_1 &= \tau\left(\frac{x_1-s}{\sqrt{(x_1-s)^2+x_2^2+h^2}}+\frac{x_1+s}{\sqrt{(x_1+s)^2+x_2^2+h^2}}\right) \\ \xi_2 &= \tau\left(\frac{x_2}{\sqrt{(x_1-s)^2+x_2^2+h^2}}+\frac{x_2}{\sqrt{(x_1+s)^2+x_2^2+h^2}}\right). \end{aligned} \tag{4.30}$$

From the first two relations in (4.29) and (4.30), we have

$$\begin{aligned} &\sqrt{(y_1-s)^2+y_2^2+h^2}+\sqrt{(y_1+s)^2+y_2^2+h^2} \\ &\quad=\sqrt{(x_1-s)^2+x_2^2+h^2}+\sqrt{(x_1+s)^2+x_2^2+h^2} \end{aligned} \tag{4.31}$$

and

$$\begin{aligned} &\frac{y_1-s}{\sqrt{(y_1-s)^2+y_2^2+h^2}}-\frac{y_1+s}{\sqrt{(y_1+s)^2+y_2^2+h^2}} \\ &\quad=\frac{x_1-s}{\sqrt{(x_1-s)^2+x_2^2+h^2}}-\frac{x_1+s}{\sqrt{(x_1+s)^2+x_2^2+h^2}}. \end{aligned} \tag{4.32}$$

We will use the prolate spheroidal coordinates to solve for x and y. We let

$$\begin{array}{ll} x_1 = s\cosh\rho\cos\phi & y_1 = s\cosh\rho'\cos\phi' \\ x_2 = s\sinh\rho\sin\phi\cos\theta & y_2 = s\sinh\rho'\sin\phi'\cos\theta' \\ x_3 = h + s\sinh\rho\sin\phi\sin\theta & y_3 = h + s\sinh\rho'\sin\phi'\sin\theta' \end{array} \tag{4.33}$$

with $\rho > 0$, $0 \leq \phi \leq \pi$, and $0 \leq \theta < 2\pi$.

In this case, $x_3 = 0$ and we use it to solve for h. Hence,

$$(x_1 - s)^2 + x_2^2 + h^2 = s^2(\cosh\rho - \cos\phi)^2$$

and

$$(x_1 + s)^2 + x_2^2 + h^2 = s^2(\cosh\rho + \cos\phi)^2.$$

Noting that $s > 0$ and $\cosh\rho \pm \cos\phi > 0$, the first relation given by (4.31) in these coordinates become

$$s(\cosh\rho - \cos\phi) + s(\cosh\rho + \cos\phi) = s(\cosh\rho' - \cos\phi') + s(\cosh\rho' + \cos\phi')$$

from which we get

$$\cosh\rho = \cosh\rho' \Rightarrow \rho = \rho'.$$

The second relation given by (4.32) becomes

$$\frac{\cosh\rho\cos\phi - 1}{\cosh\rho - \cos\phi} - \frac{\cosh\rho\cos\phi + 1}{\cosh\rho + \cos\phi} = \frac{\cosh\rho\cos\phi' - 1}{\cosh\rho - \cos\phi'} - \frac{\cosh\rho\cos\phi' + 1}{\cosh\rho + \cos\phi'}.$$

After simplification, we get

$$\frac{\sin^2\phi}{\cosh^2\rho - \cos^2\phi} = \frac{\sin^2\phi'}{\cosh^2\rho - \cos^2\phi'}$$

which implies

$$(\cosh^2\rho - 1)(\sin^2\phi - \sin^2\phi') = 0.$$

Thus, $\sin\phi = \pm\sin\phi' \Rightarrow \phi = \pm\phi', \pi \pm \phi'$.

We remark that $\cos\theta = \pm\sqrt{1 - \frac{h^2}{s^2\sinh^2\rho\sin^2\phi}} = \pm\cos\theta'$ and note that $x_3 = 0$ implies that $\sin(\phi) \neq 0$, so that division by $\sin(\phi)$ is allowed here. We also remark that it is enough to consider $\cos\theta = \cos\theta'$ as no additional relations are introduced by considering $\cos\theta = -\cos\theta'$.

Now, we go back to x and y coordinates.

If $\phi' = \phi$ then $x_1 = y_1,\ x_2 = y_2, \xi_i = \eta_i$ for $i = 1, 2$. For these points, the composition, $\mathcal{C}^t_{cm} \circ \mathcal{C}_{cm} \subset \Delta = \{(x, \xi; x, \xi)\}$.

If $\phi' = -\phi$ then, $x_1 = y_1,\ -x_2 = y_2, \xi_1 = \eta_1, -\xi_2 = \eta_2$. For these points, the composition, $\mathcal{C}^t_{cm} \circ \mathcal{C}_{cm}$ is a subset of $C_1 = \{(x_1, x_2, \xi_1, \xi_2; x_1, -x_2, \xi_1, -\xi_2)\}$ which is the graph of $\chi_1(x, \xi) = (x_1, -x_2, \xi_1, -\xi_2)$. This in the base space represents the reflection about the x_1 axis.

If $\phi' = \pi - \phi$ then, $-x_1 = y_1,\ x_2 = y_2, -\xi_1 = \eta_1, \xi_2 = \eta_2$. For these points, the composition $\mathcal{C}^t_{cm} \circ \mathcal{C}_{cm}$ is a subset of $C_2 = \{(x_1, x_2, \xi_1, \xi_2; -x_1, x_2, -\xi_1, \xi_2)\}$ which is the graph of $\chi_2(x, \xi) = (-x_1, x_2, -\xi_1, \xi_2)$. This in the base space represents the reflection about the x_2 axis.

If $\phi' = \pi + \phi$ then, $-x_1 = y_1,\ -x_2 = y_2, -\xi_1 = \eta_1, -\xi_2 = \eta_2$. For these points, $\mathcal{C}^t_{cm} \circ \mathcal{C}_{cm}$ is a subset of $C_3 = \{(x_1, x_2, \xi_1, \xi_2; -x_1, -x_2, -\xi_1, -\xi_2)\}$ which is the graph of $\chi_3(x, \xi) = (-x_1, -x_2, -\xi_1, -\xi_2)$. This in the base space represents the reflection about the origin.

Notice that $\chi_1 \circ \chi_1 = \mathrm{Id},\ \chi_2 \circ \chi_2 = \mathrm{Id},\ \chi_1 \circ \chi_2 = \chi_3$.

Hence, we have shown that $\mathcal{C}^t_{cm} \circ \mathcal{C}_{cm} \subset \Delta \cup C_1 \cup C_2 \cup C_3$.

As in the common offset case, Theorem 4.8 shows that artifacts are introduced in image reconstruction. However, in this case, each true singularity introduces three additional artifact singularities (these added singularities are described by the Lagrangians C_1, C_2 and C_3). One can show that the strengths of the added singularities are the same as that of the true ones. We do not include the details of this analysis in the current article. We refer the interested reader to our paper [1]. □

Acknowledgments The author thanks Prof. Abul Hasan Siddiqi, Prof. Pammy Manchanda and the organizers for the invitation to give a talk in the 11th Annual Conference of the Indian Society of Industrial and Applied Mathematics titled, "Emerging Mathematical Methods, Models and Algorithms for Science and Technology" at Gautam Buddha University (GBU) on December 15 and 16, 2012, commemorating the 125th birth year of Srinivasa Ramanujan, and for the warm hospitality during his stay on GBU campus. He also thanks Prof. Siddiqi for inviting to contribute an article in the proceedings. The material here is a slightly expanded version of the talk given at this conference. The author was partially supported by NSF Grant DMS 1109417.

References

1. Ambartsoumian G, Felea R, Krishnan VP, Nolan C, Quinto ET (2013) A class of singular Fourier integral operators in synthetic aperture radar imaging. J. Funct. Anal. 264(1):246–269
2. Andersson L-E (1988) On the determination of a function from spherical averages. SIAM J Math Anal 19:214–232
3. Cheney M (2001) A mathematical tutorial on synthetic aperture radar. SIAM Rev 43(2):301–312 (electronic)
4. Cheney M, Borden B (2009) Fundamentals of Radar Imaging. CBMS-NSF regional conference series in applied mathematics, vol 79. Society for Industrial and Applied Mathematics

5. Cohen J, Bleistein H (1979) Velocity inversion procedure for acoustic waves. Geophysics 44:1077–1085
6. de Hoop MV (2003) Microlocal analysis of seismic inverse scattering. In: Inside out: inverse problems and applications. Math Sci Res Inst Publ 47:219–296. Cambridge Univeristy Press, Cambridge
7. Dowling T (2009) Radar imaging using multiply scattered waves. In: Ph.D. thesis, University of Limerick, Ireland
8. Duistermaat JJ (2011) Fourier integral operators. Modern Birkhäuser Classics. Birkhäuser/Springer, New York. Reprint of the 1996 original
9. Felea R (2005) Composition of Fourier integral operators with fold and blowdown singularities. Comm Partial Differ Equ 30(10–12):1717–1740
10. Felea R (2007) Displacement of artefacts in inverse scattering. Inverse Prob 23(4):1519–1531
11. Golubitsky M, Guillemin V (1973) Stable mappings and their singularities. Springer, New York. Graduate Texts Math 14
12. Greenleaf A, Uhlmann G (1989) Nonlocal inversion formulas for the X-ray transform. Duke Math J 58(1):205–240
13. Greenleaf A, Uhlmann G (1990) Composition of some singular Fourier integral operators and estimates for restricted X-ray transforms. Ann Inst Fourier (Grenoble) 40(2):443–466
14. Greenleaf A, Uhlmann G (1990) Estimates for singular Radon transforms and pseudodifferential operators with singular symbols. J Funct Anal 89(1):202–232
15. Greenleaf A, Uhlmann G (1990) Microlocal techniques in integral geometry. In: Integral geometry and tomography (Arcata, CA, 1989). Contemp Math 113:121–135. Amer Math Soc. Providence, RI
16. Greenleaf A, Uhlmann G (1991) Composition of some singular Fourier integral operators and estimates for restricted X-ray transforms. II Duke Math J 64(3):415–444
17. Grigis A, Sjöstrand J (1994) Microlocal analysis for differential operators. London mathematical society lecture note series, vol 196. Cambridge University Press, Cambridge. An introduction
18. Guillemin V, Uhlmann G (1981) Oscillatory integrals with singular symbols. Duke Math J 48 (1):251–267
19. Hörmander L (1971) Fourier integral operators. I Acta Math 127(1–2):79–183
20. Hörmander L (2003) The analysis of linear partial differential operators. I. Classics in mathematics. Springer, Berlin, 2003. Distribution theory and Fourier analysis, Reprint of the second (1990) edition (Springer, Berlin; MR1065993 (91 m:35001a))
21. Horne AM, Yates G (2002) Bistatic synthetic aperture radar. pp 6–10
22. Krishnan VP, Quinto ET (2011) Microlocal aspects of common offset synthetic aperture radar imaging. Inverse Probl Imaging 5(3):659–674
23. Louis AK, Quinto ET (2000) Local tomographic methods in SONAR. In: Colton D, Engl H, Louis A, McLaughlin J, Rundell W (eds) Surveys on solution methods for inverse problems. Springer, Vienna/New York, pp 147–154
24. Melrose RB, Uhlmann GA (1979) Lagrangian intersection and the Cauchy problem. Commun Pure Appl Math 32(4):483–519
25. Nolan CJ, Cheney M (2004) Microlocal analysis of synthetic aperture radar imaging. J Fourier Anal Appl 10(2):133–148
26. Nolan CJ, Cheney M, Dowling T, Gaburro R (2006) Enhanced angular resolution from multiply scattered waves. Inverse Prob 22(5):1817–1834
27. Stefanov P, Uhlmann G (2013) Is a curved flight path in SAR better than a straight one? SIAM J Appl Math 73(4):1596–1612
28. Trèves F (1980) Introduction to pseudodifferential and Fourier integral operators, vol 1. Plenum Press, New York. Pseudodifferential operators, The University Series in Mathematics
29. Trèves F (1980) Introduction to pseudodifferential and Fourier integral operators, vol 2. Plenum Press, New York. Fourier integral operators, The University Series in Mathematics

30. Yarman CE, Yazici B (2008) Synthetic aperture hitchhiker imaging. IEEE Trans Image Process 17(11):2156–2173
31. Yarman CE, Yazici B, Cheney M (2008) Bistatic synthetic aperture radar imaging for arbitrary flight trajectories. IEEE Trans Image Process 17(1):84–93
32. Yazici B, Cheney M, Yarman CE (2006) Synthetic-aperture inversion in the presence of noise and clutter. Inverse Prob 22(5):1705–1729

Chapter 5
Cubic Spline Approximation for Two-Dimensional Nonlinear Elliptic Boundary Value Problems

R.K. Mohanty

Abstract We report a new 9-point compact discretization of order two in y- and order four in x-directions, based on cubic spline approximation, for the solution of two-dimensional nonlinear elliptic partial differential equations of the form

$$A(x,y)\frac{\partial^2 u}{\partial x^2} + B(x,y)\frac{\partial^2 u}{\partial y^2} = f\left(x, y, u, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}\right), (x,y) \in \Omega$$

defined in the domain $\Omega = \{(x,y) : 0 < x, y < 1\}$ with boundary $\partial\Omega$, where $A(x,y) > 0$ and $B(x,y) > 0$ in Ω. The corresponding Dirichlet boundary conditions are prescribed by

$$u(x,y) = \psi(x,y), \quad (x,y) \in \partial\Omega$$

The main spline relations are presented and incorporated into solution procedures for elliptic partial differential equations. Available numerical methods based on cubic spline approximations for the numerical solution of nonlinear elliptic equations are of second-order accurate. Although 9-point finite difference approximations of order four accurate for the solution of nonlinear elliptic differential equations are discussed in the past, but these methods require five evaluations of the function f. In this piece of work, using the same number of grid points and three evaluations of the function f, we have derived a new stable cubic spline method of order 2 in y- and order 4 in x-directions for the solution of nonlinear elliptic equation. However, for a fixed parameter $(\Delta y/\Delta x^2)$, the proposed method behaves like a fourth order method. The accuracy of the proposed method is exhibited from the computed results. The proposed method is applicable to Poisson's equation and two-dimensional Navier-Stokes' equations of motion in polar coordinates, which is

R.K. Mohanty (✉)
Department of Applied Mathematics, South Asian University, Akbar Bhawan, Chanakyapuri 110021, New Delhi, India
e-mail: rmohanty@sau.ac.in

A.H. Siddiqi et al. (eds.), *Mathematical Models, Methods and Applications*, Industrial and Applied Mathematics, DOI 10.1007/978-981-287-973-8_5

main highlight of the work. The convergence analysis of the proposed cubic spline approximation for the nonlinear elliptic equation is discussed and we have shown under appropriate conditions the proposed method converges. Some physical examples and their numerical results are provided to justify the advantages of the proposed method.

Keywords Nonlinear elliptic equation · Cubic spline approximation · Poisson's equation in polar coordinates · Diffusion-convection equation · Burgers' equation · Reynolds number

5.1 Introduction

We consider the 2D nonlinear elliptic partial differential equation

$$A(x,y)\frac{\partial^2 u}{\partial x^2}+B(x,y)\frac{\partial^2 u}{\partial y^2}=f\left(x,y,u,\frac{\partial u}{\partial x},\frac{\partial u}{\partial y}\right),(x,y)\in\Omega \tag{5.1}$$

defined in the bounded domain $\Omega=\{(x,y):0<x,y<1\}$ with boundary $\partial\Omega$, where $A(x,y)>0$ and $B(x,y)>0$ in Ω.

The corresponding Dirichlet boundary conditions are prescribed by

$$u(x,y)\ =\ \psi(x,y),\quad (x,y)\in\partial\Omega \tag{5.2}$$

We assume that for $0<x,y<1$,

$$(i)\qquad f\left(x,y,u,\frac{\partial u}{\partial x},\frac{\partial u}{\partial y}\right)\text{ is continuous,} \tag{5.3a}$$

$$(ii)\qquad \frac{\partial f}{\partial u},\frac{\partial f}{\partial u_x},\frac{\partial f}{\partial u_y}\text{ exist and are continuous,} \tag{5.3b}$$

$$(iii)\qquad \frac{\partial f}{\partial u}\geq 0,\left|\frac{\partial f}{\partial u_x}\right|\leq G\text{ and }\left|\frac{\partial f}{\partial u_y}\right|\leq H \tag{5.3c}$$

where G and H are positive constants (see [1]). Further, we may also assume that the coefficients $u(x,y),A(x,y)$ and $B(x,y)$ are sufficiently smooth and their required higher order partial derivatives exist in the solution domain Ω.

The main aim of this work is to use cubic spline function and its certain properties, which are then used to approximate the differential equation (5.1) to obtain the numerical solution. We use only 9-point compact cell three evaluations of the function f (Fig. 5.1).

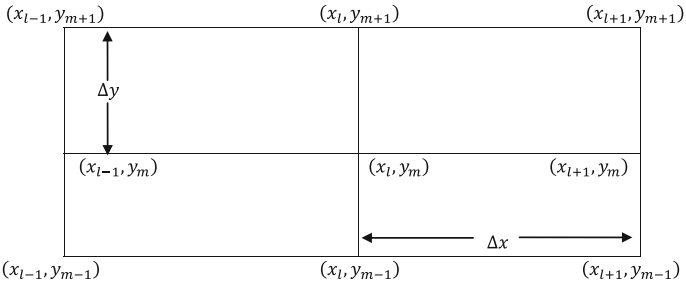

Fig. 5.1 9-Point computational network

We consider our region of interest, a rectangular domain $\Omega = [0, 1] \times [0, 1]$. A grid with spacing $\Delta x > 0$ and $\Delta y > 0$ in the directions x- and y-respectively are first chosen, so that the mesh points (x_l, y_m) are defined as $x_l = l\Delta x$ and $y_m = m\Delta y$, $l = 0, 1, \ldots, N+1$, $m = 0, 1, \ldots, M+1$, where N and M are positive integers such that $(N+1)\Delta x = 1$ and $(M+1)\Delta y = 1$.

Let us denote the mesh ratio parameter by $p = (\Delta y/\Delta x) > 0$. For convergence of the numerical scheme it is essential that our parameter remains in the range $0 < \sqrt{6}p < 1$. Let $U_{l,m}$ and $u_{l,m}$ be the exact and approximation solution values of $u(x, y)$ at the grid point (x_l, y_m), respectively. Similarly, let $A_{l,m} = A(x_l, y_m)$ and $B_{l,m} = B(x_l, y_m)$ be the exact values of $A(x, y)$ and $B(x, y)$ at the grid point (x_l, y_m), respectively.

Let $S_m(x)$ is a piecewise cubic polynomial defined in $x_{l-1} \leq x \leq x_l$, which satisfies

$$S''_m(x) = \frac{(x_l - x)}{\Delta x} M_{l-1,m} + \frac{(x - x_{l-1})}{\Delta x} M_{l,m}, x \in [x_{l-1}, x_l] \tag{5.4}$$

where $M_{l,m} = S''_m(x_l)$ and $m_{l,m} = S'_m(x_l)$. Integrating (5.4) twice and using the interpolating conditions $S_m(x_{l-1}) = u_{l-1,m}$ and $S_m(x_l) = u_{l,m}$, we obtain the cubic spline interpolating polynomial

$$\begin{aligned} S_m(x) = {} & \frac{(x_l - x)^3}{6\Delta x} M_{l-1,m} + \frac{(x - x_{l-1})^3}{6\Delta x} M_{l,m} + \left(u_{l-1,m} - \frac{\Delta x^2}{6} M_{l-1,m}\right)\left(\frac{x_l - x}{\Delta x}\right) \\ & + \left(u_{l,m} - \frac{\Delta x^2}{6} M_{l,m}\right)\left(\frac{x - x_{l-1}}{\Delta x}\right), x_{l-1} \leq x \leq x_l; \\ & l = 1, 2, \ldots, N+1, m = 0, 1, \ldots, M+1 \end{aligned} \tag{5.5}$$

which satisfies at mth-line parallel to x-axis the following properties.

(i) $S_m(x)$ coincides with a polynomial of degree three on each $[x_{l-1}, x_l]$, $l = 1, 2, \ldots, N+1$, $m = 0, 1, \ldots, M+1$;
(ii) $S_m(x) \in C^2[0, 1]$, and
(iii) $S_m(x_l) = u_{l,m}$, $l = 0, 1, \ldots, N+1$, $m = 0, 1, \ldots, M+1$.

Denote:
$m_{l,m} = S_m^{'}(x_l) = U_{xl,m}$. First derivative of (5.5) is given by

$$S_m'(x_l) = U_{xl,m} = \frac{U_{l,m} - U_{l-1,m}}{\Delta x} + \frac{\Delta x}{6}\left[M_{l-1,m} + 2M_{l,m}\right], x \in [x_{l-1}, x_l]$$
$$S_m'(x_l) = U_{xl,m} = \frac{U_{l+1,m} - U_{l,m}}{\Delta x} - \frac{\Delta x}{6}\left[M_{l+1,m} + 2M_{l,m}\right], x \in [x_l, x_{l+1}]$$

The continuity of first derivative implies

$$M_{l-1,m} + 4M_{l,m} + M_{l+1,m} = \frac{6}{\Delta x^2}\left(U_{l+1,m} - 2U_{l,m} + U_{l-1,m}\right)$$
$$M_{l,m} = S_m''(x_l) = U_{xxl,m} = \frac{1}{A_{l,m}}\left[-B_{l,m}U_{yyl,m} + f\left(x_l, y_m, U_{l,m}, m_{l,m}, U_{yl,m}\right)\right],$$
$$S_m'(x_{l+1}) = U_{xl+1,m} = \frac{U_{l+1,m} - U_{l,m}}{\Delta x} + \frac{\Delta x}{6}\left[M_{l,m} + 2M_{l+1,m}\right]$$

and $S_m^{'}(x_{l-1}) = U_{xl-1,m} = \frac{U_{l,m} - U_{l-1,m}}{\Delta x} - \frac{\Delta x}{6}\left[M_{l,m} + 2M_{l-1,m}\right]$.

Note that these are important properties of the cubic spline function $S_m(x)$ which are used in building up the numerical scheme.

At the grid point (x_l, y_m), we use the notation

$$A_{pq} = \frac{\partial^{p+q} A(x_l, y_m)}{\partial x^p \partial y^q}, \text{etc.}$$

We consider the following approximations:

$$\bar{U}_{yl,m} = \left(U_{l,m+1} - U_{l,m-1}\right)/(2\Delta y) \tag{5.6a}$$

$$\bar{U}_{yl+1,m} = \left(U_{l+1,m+1} - U_{l+1,m-1}\right)/(2\Delta y) \tag{5.6b}$$

$$\bar{U}_{yl-1,m} = \left(U_{l-1,m+1} - U_{l-1,m-1}\right)/(2\Delta y) \tag{5.6c}$$

$$\bar{U}_{yyl,m} = \left(U_{l,m+1} - 2U_{l,m} + U_{l,m-1}\right)/\Delta y^2 \tag{5.6d}$$

$$\bar{U}_{yyl+1,m} = \left(U_{l+1,m+1} - 2U_{l+1,m} + U_{l+1,m-1}\right)/\Delta y^2 \tag{5.6e}$$

$$\bar{U}_{yyl-1,m} = \left(U_{l-1,m+1} - 2U_{l-1,m} + U_{l-1,m-1}\right)/\Delta y^2 \tag{5.6f}$$

$$\bar{m}_{l,m} = \bar{U}_{xl,m} = \left(U_{l+1,m} - U_{l-1,m}\right)/(2\Delta x) \tag{5.7a}$$

$$\bar{m}_{l+1,m} = \bar{U}_{xl+1,m} = \left(3U_{l+1,m} - 4U_{l,m} + U_{l-1,m}\right)/(2\Delta x) \tag{5.7b}$$

$$\bar{m}_{l-1,m} = \bar{U}_{xl-1,m} = \left(-3U_{l-1,m} + 4U_{l,m} - U_{l+1,m}\right)/(2\Delta x) \tag{5.7c}$$

$$\bar{U}_{xxl,m} = \left(U_{l+1,m} - 2U_{l,m} + U_{l-1,m}\right)/\Delta x^2 \tag{5.7d}$$

$$\bar{F}_{l,m} = f\left(x_l, y_m, U_{l,m}, \bar{m}_{l,m}, \bar{U}_{yl,m}\right) \tag{5.8a}$$

$$\bar{F}_{l+1,m} = f\left(x_{l+1}, y_m, U_{l+1,m}, \bar{m}_{l+1,m}, \bar{U}_{yl+1,m}\right) \tag{5.8b}$$

$$\bar{F}_{l-1,m} = f\left(x_{l-1}, y_m, U_{l-1,m}, \bar{m}_{l-1,m}, \bar{U}_{yl-1,m}\right) \tag{5.8c}$$

$$\bar{M}_{l,m} = \frac{1}{A_{00}}\left[-B_{00}\bar{U}_{yyl,m} + \bar{F}_{l,m}\right] \tag{5.9a}$$

$$\bar{M}_{l+1,m} = \frac{1}{A_{00}}\left(1 - \frac{\Delta x A_{10}}{A_{00}}\right)\left[-B_{l+1,m}\bar{U}_{yyl+1,m} + \bar{F}_{l+1,m}\right] \tag{5.9b}$$

$$\bar{M}_{l-1,m} = \frac{1}{A_{00}}\left(1 + \frac{\Delta x A_{10}}{A_{00}}\right)\left[-B_{l-1,m}\bar{U}_{yyl-1,m} + \bar{F}_{l-1,m}\right] \tag{5.9c}$$

$$\overline{\overline{m}}_{l+1,m} = \overline{\overline{U}}_{xl+1,m} = \frac{U_{l+1,m} - U_{l,m}}{\Delta x} + \frac{\Delta x}{6}\left[\bar{M}_{l,m} + 2\bar{M}_{l+1,m}\right] \tag{5.10a}$$

$$\overline{\overline{m}}_{l-1,m} = \overline{\overline{U}}_{xl-1,m} = \frac{U_{l,m} - U_{l-1,m}}{\Delta x} - \frac{\Delta x}{6}\left[\bar{M}_{l,m} + 2\bar{M}_{l-1,m}\right] \tag{5.10b}$$

$$\begin{aligned}\widehat{U}_{xl,m} = \bar{U}_{xl,m} &- \frac{\Delta x}{12A_{00}}\left[\bar{F}_{l+1,m} - \bar{F}_{l-1,m}\right] + \frac{\Delta x}{12A_{00}}B_{00}\left[\bar{U}_{yyl+1,m} - \bar{U}_{yyl-1,m}\right] \\ &+ \frac{\Delta x^2}{6}\frac{A_{10}}{A_{00}}\bar{U}_{xxl,m} + \frac{\Delta x^2}{6}\frac{B_{10}}{A_{00}}\bar{U}_{yyl,m}\end{aligned} \tag{5.10c}$$

$$\overline{\overline{F}}_{l+1,m} = f\left(x_{l+1}, y_m, U_{l+1,m}, \overline{\overline{m}}_{l+1,m}, \bar{U}_{yl+1,m}\right) \tag{5.11a}$$

$$\overline{\overline{F}}_{l-1,m} = f\left(x_{l-1}, y_m, U_{l-1,m}, \overline{\overline{m}}_{l-1,m}, \bar{U}_{yl-1,m}\right) \tag{5.11b}$$

$$\widehat{\overline{F}}_{l,m} = f\left(x_l, y_m, U_{l,m}, \widehat{U}_{xl,m}, \bar{U}_{yl,m}\right) \tag{5.11c}$$

The cubic spline approximations (5.9a)–(5.10c) are discussed in details in [2]. Then at each internal grid point (x_l, y_m), the cubic spline method with accuracy of $O(\Delta y^2+\Delta y^2\Delta x^2+\Delta x^4)$ for the solution of nonlinear elliptic partial differential Eq. (5.1) may be written as

$$\begin{aligned} L_u \equiv\, & p^2\left[A_{00}-\frac{\Delta x^2}{6}\frac{A_{10}}{A_{00}}A_{10}+\frac{\Delta x^2}{12}A_{20}\right]\delta_x^2 U_{l,m} \\ & +\frac{\Delta y^2}{12}\left[\left(1-\frac{\Delta x A_{10}}{A_{00}}\right)B_{l+1,m}\bar{U}_{yyl+1,m}+\left(1+\frac{\Delta x A_{10}}{A_{00}}\right)B_{l-1,m}\bar{U}_{yyl-1,m}+10B_{l,m}\bar{U}_{yyl,m}\right] \\ & =\frac{\Delta y^2}{12}\left[\left(1-\frac{\Delta x A_{10}}{A_{00}}\right)\overline{\overline{F}}_{l+1,m}+\left(1+\frac{\Delta x A_{10}}{A_{00}}\right)\overline{\overline{F}}_{l-1,m}+10\widehat{\widehat{F}}_{l,m}\right]+\widehat{T}_{l,m}, \\ & l=1,2,\ldots,N, m=1,2,\ldots,M \end{aligned} \tag{5.12}$$

where, $\delta_x U_l = \left(U_{l+\frac{1}{2}}-U_{l-\frac{1}{2}}\right)$ and $\mu_x U_l = \frac{1}{2}\left(U_{l+\frac{1}{2}}+U_{l-\frac{1}{2}}\right)$ are the central and average difference operators with respect to x-direction and the local truncation error $\widehat{T}_{l,m} = O(\Delta y^4+\Delta y^4\Delta x^2+\Delta y^2\Delta x^4)$.

We may re-write (5.12) as

$$\begin{aligned} & \lambda_1\left(U_{l+1,m}+U_{l-1,m}\right)+\lambda_2\left(U_{l,m+1}+U_{l,m-1}\right) \\ & +\lambda_3\left(U_{l+1,m+1}+U_{l+1,m-1}+U_{l-1,m+1}+U_{l-1,m-1}-\left(24p^2+20\right)U_{l,m}\right) \\ & =\frac{\Delta y^2}{12}\left[\overline{\overline{F}}_{l+1,m}+\overline{\overline{F}}_{l-1,m}+10\widehat{\widehat{F}}_{l,m}\right]+\widehat{T}_{l,m} \end{aligned} \tag{5.13}$$

where $\lambda_1 = p^2-\frac{2}{12}, \lambda_2 = \frac{10}{12}, \lambda_3 = \frac{1}{12}$.

The condition which is usually imposed on Eq. (5.13) is that $\lambda_1 > 0$, $\lambda_2 > 0$ and $\lambda_3 > 0$, i.e., $0 < \sqrt{6}p < 1$.

5.2 Derivation of the Method

For the derivation of the numerical method (5.12) for the solution of partial differential Eq. (5.1), we follow the ideas given by Jain and Aziz [2].

At the grid point (x_l, y_m), we may write the differential Eq. (5.1) as

$$A_{l,m}U_{xxl,m}+B_{l,m}U_{yyl,m} = f\left(x_l, y_m, U_{l,m}, U_{xl,m}, U_{yl,m}\right) \equiv F_{l,m} \tag{5.14}$$

Using Taylor's expansion, we may write the approximations as

$$\bar{U}_{yl,m} = \frac{\left(U_{l,m+1} - U_{l,m-1}\right)}{(2\Delta y)} = U_{yl,m} + \frac{\Delta y^2}{6} U_{03} + O\left(\Delta y^4\right)$$

$$\bar{U}_{yl+1,m} = \frac{\left(U_{l+1,m+1} - U_{l+1,m-1}\right)}{(2\Delta y)} = U_{yl+1,m} + \frac{\Delta y^2}{6} U_{03} + \frac{\Delta x \cdot \Delta y^2}{6} U_{13} + O\left(\Delta y^2 \Delta x^2\right)$$

$$\bar{U}_{yl-1,m} = \frac{\left(U_{l-1,m+1} - U_{l-1,m-1}\right)}{(2\Delta y)} = U_{yl-1,m} + \frac{\Delta y^2}{6} U_{03} - \frac{\Delta x \cdot \Delta y^2}{6} U_{13} + O\left(\Delta y^2 \Delta x^2\right)$$

$$\bar{U}_{yyl,m} = \frac{\left(U_{l,m+1} - 2U_{l,m} + U_{l,m-1}\right)}{\Delta y^2} = U_{yyl,m} + \frac{\Delta y^2}{12} U_{04} + O\left(\Delta y^4\right)$$

$$\begin{aligned}\bar{U}_{yyl+1,m} &= \frac{\left(U_{l+1,m+1} - 2U_{l+1,m} + U_{l+1,m-1}\right)}{\Delta y^2} \\ &= U_{yyl+1,m} + \frac{\Delta y^2}{12} U_{04} + \frac{\Delta x \cdot \Delta y^2}{12} U_{14} + O\left(\Delta y^2 \Delta x^2\right)\end{aligned}$$

$$\begin{aligned}\bar{U}_{yyl-1,m} &= \frac{\left(U_{l-1,m+1} - 2U_{l-1,m} + U_{l-1,m-1}\right)}{\Delta y^2} \\ &= U_{yyl-1,m} + \frac{\Delta y^2}{12} U_{04} - \frac{\Delta x \cdot \Delta y^2}{12} U_{14} + O\left(\Delta y^2 \Delta x^2\right)\end{aligned}$$

$$\bar{U}_{xxl,m} = \frac{\left(U_{l+1,m} - 2U_{l,m} + U_{l-1,m}\right)}{\Delta x^2} = U_{xxl,m} + \frac{\Delta x^2}{12} U_{40} + O\left(\Delta x^4\right)$$

$$\bar{m}_{l,m} = \bar{U}_{xl,m} = \frac{\left(U_{l+1,m} - U_{l-1,m}\right)}{(2\Delta x)} = U_{xl,m} + \frac{\Delta x^2}{6} U_{30} + O\left(\Delta x^4\right)$$

$$\bar{m}_{l+1,m} = \bar{U}_{xl+1,m} = \frac{\left(3U_{l+1,m} - 4U_{l,m} + U_{l-1,m}\right)}{(2\Delta x)} = U_{xl+1,m} - \frac{\Delta x^2}{3} U_{30} + O\left(\Delta x^3\right)$$

$$\bar{m}_{l-1,m} = \bar{U}_{xl-1,m} = \frac{\left(-3U_{l-1,m} + 4U_{l,m} - U_{l+1,m}\right)}{(2\Delta x)} = U_{xl-1,m} - \frac{\Delta x^2}{3} U_{30} - O\left(\Delta x^3\right)$$

Using Taylor series expansion about the grid point (x_l, y_m), from Eq. (5.1) we obtain

$$\begin{aligned}L_u = \frac{\Delta y^2}{12}\left[\left(1 - \frac{\Delta x A_{10}}{A_{00}}\right) F_{l+1,m} + \left(1 + \frac{\Delta x A_{10}}{A_{00}}\right) F_{l-1,m} + 10 F_{l,m}\right] \\ + O\left(\Delta y^4 + \Delta y^4 \Delta x^2 + \Delta y^2 \Delta x^4\right); l = 1, 2, \ldots, N, m = 1, 2, \ldots M\end{aligned} \tag{5.15}$$

Let us denote $\alpha_{l,m} = \left(\frac{\partial f}{\partial U_x}\right)_{l,m}$, then

$$\begin{aligned}\bar{F}_{l,m} &= f\left(x_l, y_m, U_{l,m}, \bar{m}_{l,m}, \bar{U}_{yl,m}\right) \\ &= f\left(x_l, y_m, U_{l,m}, U_{xl,m} + \frac{\Delta x^2}{6} U_{30} + O\left(\Delta x^4\right), U_{yl,m} + O\left(\Delta y^2\right)\right) \\ &= F_{l,m} + \frac{\Delta x^2}{6} U_{30} \alpha_{l,m} + O\left(\Delta y^2 + \Delta x^4\right)\end{aligned}$$

Similarly,

$$\begin{aligned}\bar{F}_{l+1,m} &= f\left(x_{l+1}, y_m, U_{l+1,m}, \bar{m}_{l+1,m}, \bar{U}_{yl+1,m}\right)\\ &= F_{l+1,m} - \frac{\Delta x^2}{3} U_{30}\alpha_{l,m} + O\left(\Delta x^3 + \Delta y^2\right)\\ \bar{F}_{l-1,m} &= f\left(x_{l-1}, y_m, U_{l-1,m}, \bar{m}_{l-1,m}, \bar{U}_{yl-1,m}\right)\\ &= F_{l-1,m} - \frac{\Delta x^2}{3} U_{30}\alpha_{l,m} + O\left(-\Delta x^3 + \Delta y^2\right)\end{aligned}$$

We have also $M_{l,m} = S''_m(x_l) = U_{xxl,m}$

$$\begin{aligned}\Rightarrow \bar{M}_{l,m} &= \bar{U}_{xxl,m} = \frac{1}{A_{00}}\left[-B_{00}\bar{U}_{yyl,m} + \bar{F}_{l,m}\right]\\ \bar{M}_{l+1,m} &= \bar{U}_{xxl+1,m} = \frac{1}{A_{l+1,m}}\left[-B_{l+1,m}\bar{U}_{yyl+1,m} + \bar{F}_{l+1,m}\right]\\ &= \frac{1}{\left(A_{l,m} + \Delta x A_{10} + \frac{\Delta x^2}{2}A_{20} + \cdots\right)}\left[-B_{l+1,m}\bar{U}_{yyl+1,m} + \bar{F}_{l+1,m}\right]\\ &= \frac{1}{A_{l,m}\left(1 + \Delta x\frac{A_{10}}{A_{l,m}} + \cdots\right)}\left[-B_{l+1,m}\bar{U}_{yyl+1,m} + \bar{F}_{l+1,m}\right]\\ &= \frac{1}{A_{l,m}}\left(1 + \Delta x\frac{A_{10}}{A_{l,m}} + \cdots\right)^{-1}\left[-B_{l+1,m}\bar{U}_{yyl+1,m} + \bar{F}_{l+1,m}\right]\\ &= \frac{1}{A_{l,m}}\left(1 - \Delta x\frac{A_{10}}{A_{l,m}} + \cdots\right)\left[-B_{l+1,m}\bar{U}_{yyl+1,m} + \bar{F}_{l+1,m}\right]\\ &= \frac{1}{A_{00}}\left(1 - \frac{\Delta x A_{10}}{A_{00}}\right)\left[-B_{l+1,m}\bar{U}_{yyl+1,m} + \bar{F}_{l+1,m}\right]\end{aligned}$$

Similarly,

$$\bar{M}_{l-1,m} = \frac{1}{A_{00}}\left(1 + \frac{\Delta x A_{10}}{A_{00}}\right)\left[-B_{l-1,m}\bar{U}_{yyl-1,m} + \bar{F}_{l-1,m}\right]$$

Further, using spline relation we have

$$\begin{aligned}\overline{\overline{m}}_{l+1,m} &= \overline{\overline{U}}_{xl+1,m} = \frac{U_{l+1,m} - U_{l,m}}{\Delta x} + \frac{\Delta x}{6}\left[\bar{M}_{l,m} + 2\bar{M}_{l+1,m}\right]\\ &= m_{l+1,m} + O\left(\Delta x^3 + \Delta y^2\right)\end{aligned}$$

and

$$\begin{aligned}\overline{\overline{m}}_{l-1,m} = \overline{\overline{U}}_{xl-1,m} &= \frac{U_{l,m} - U_{l-1,m}}{\Delta x} - \frac{\Delta x}{6}\left[\bar{M}_{l,m} + 2\bar{M}_{l-1,m}\right] \\ &= m_{l-1,m} + O\left(-\Delta x^3 + \Delta y^2\right).\end{aligned}$$

Now we need $O(\Delta y^2 + \Delta y^2 \Delta x^2 + \Delta x^4)$-approximation for $\widehat{U}_{xl,m}$.
Let us consider

$$\begin{aligned}\widehat{U}_{xl,m} = \bar{U}_{xl,m} &+ a\Delta x\left[\bar{F}_{l+1,m} - \bar{F}_{l-1,m}\right] + b\Delta x\left[\bar{U}_{yyl+1,m} - \bar{U}_{yyl-1,m}\right] \\ &+ c\Delta x^2 \bar{U}_{xxl,m} + d\Delta x^2 \bar{U}_{yyl,m}\end{aligned}$$

where 'a', 'b', 'c' and 'd' are free parameters to be determined. By the help of the approximations defined earlier, we obtain

$$\begin{aligned}\widehat{U}_{xl,m} &= U_{xl,m} + \frac{\Delta x^2}{6} U_{30} + a\Delta x\left[F_{l+1,m} - F_{l-1,m}\right] + b\Delta x\left[\bar{U}_{yyl+1,m} - \bar{U}_{yyl-1,m}\right] \\ &\quad + c\Delta x^2 U_{xxl,m} + d\Delta x^2 U_{yyl,m} + O\left(\Delta y^2 + \Delta y^2 \Delta x^2 + \Delta x^4\right) \\ &= m_{l,m} + \frac{\Delta x^2}{6}\left[(1 + 12aA_{00})U_{30} + 12(aB_{00} + b)U_{12} + (6c + 12aA_{10})U_{20} + (12aB_{10} + 6d)U_{02}\right] \\ &\quad + O\left(\Delta y^2 + \Delta y^2 \Delta x^2 + \Delta x^4\right)\end{aligned}$$

$\widehat{U}_{xl,m} = m_{l,m} + O(\Delta y^2 + \Delta y^2 \Delta x^2 + \Delta x^4)$, the coefficient of Δx^2 must be zero which means

$$\begin{aligned}1 + 12aA_{00} &= 0 \\ aB_{00} + b &= 0 \\ 6c + 12aA_{10} &= 0\end{aligned}$$

and

$$12aB_{10} + 6d = 0.$$

From above, it is easily seen that, $a = -\frac{1}{12A_{00}}, b = \frac{B_{00}}{12A_{00}}, c = \frac{A_{10}}{6A_{00}}$ and $d = \frac{B_{10}}{6A_{00}}$,
Hence

$$\begin{aligned}\widehat{U}_{xl,m} &= \bar{U}_{xl,m} - \frac{\Delta x}{12A_{00}}\left[\bar{F}_{l+1,m} - \bar{F}_{l-1,m}\right] + \frac{\Delta x}{12A_{00}} B_{00}\left[\bar{U}_{yyl+1,m} - \bar{U}_{yyl-1,m}\right] \\ &\quad + \frac{\Delta x^2}{6}\frac{A_{10}}{A_{00}}\bar{U}_{xxl,m} + \frac{\Delta x^2}{6}\frac{B_{10}}{A_{00}}\bar{U}_{yyl,m} \\ &= m_{l,m} + O\left(\Delta y^2 + \Delta y^2 \Delta x^2 + \Delta x^4\right)\end{aligned}$$

Now,

$$\begin{aligned}
\overline{\overline{F}}_{l+1,m} &= f\left(x_{l+1}, y_m, U_{l+1,m}, \overline{\overline{m}}_{l+1,m}, \bar{U}_{yl+1,m}\right) \\
&= f(x_{l+1}, y_m, U_{l+1,m}, m_{l+1,m} + O\left(\Delta x^3 + \Delta y^2\right), U_{yl+1,m} + O\left(\Delta y^2\right)) \\
&= F_{l+1,m} + O\left(\Delta x^3 + \Delta y^2\right) \\
\overline{\overline{F}}_{l-1,m} &= f\left(x_{l-1}, y_m, U_{l-1,m}, \overline{\overline{m}}_{l-1,m}, \bar{U}_{yl-1,m}\right) \\
&= F_{l-1,m} + O\left(-\Delta x^3 + \Delta y^2\right) \\
\widehat{\widehat{F}}_{l,m} &= f\left(x_l, y_m, U_{l,m}, \widehat{U}_{xl,m}, \bar{U}_{yl,m}\right) \\
&= f\left(x_l, y_m, U_{l,m}, m_{l,m}, U_{yl,m}\right) + O\left(\Delta y^2 + \Delta y^2 \Delta x^2 + \Delta x^4\right) \\
&= F_{l,m} + O\left(\Delta x^4 + \Delta y^2 + \Delta y^2 \Delta x^2\right)
\end{aligned}$$

Finally, using the preceding approximations, from (5.12) and (5.15), we obtain the local truncation error as $\widehat{T}_{l,m} = O(\Delta y^4 + \Delta y^4 \Delta x^2 + \Delta y^2 \Delta x^4)$.

Note that, the Dirichlet boundary conditions are given by (5.2). Incorporating the boundary conditions, we can write the cubic spline method (5.12) in a tri-block diagonal matrix form. If the differential equation (5.1) is linear, we can solve the linear system using block Gauss-Seidel iterative method; in the nonlinear case, we can use block Newton-Raphson iterative method to solve the nonlinear system. The details of the convergence analysis has been discussed in [3].

5.3 Application to Singular Problems

Consider the two spatial dimensions elliptic partial differential equation

$$\frac{\partial^2 u}{\partial x^2} + B(x)\frac{\partial^2 u}{\partial y^2} = D(x)\frac{\partial u}{\partial x} + g(x, y), 0 < x, y < 1 \tag{5.16}$$

subject to appropriate Dirichlet boundary conditions prescribed.

The coefficients $B(x), D(x)$ and function $g(x, y) \in C^2(\Omega)$, where $C^m(\Omega)$ denotes the set of all functions of x and y with continuous partial derivatives up to order m, in Ω.

On applying formula (5.12) to the elliptic equation (5.15), we obtain the following difference scheme

$$\begin{aligned}
p^2 \delta_x^2 U_{l,m} + \frac{\Delta y^2}{12}\left[B_{l+1}\bar{U}_{yyl+1,m} + B_{l-1}\bar{U}_{yyl-1,m} + 10B_l\bar{U}_{yyl,m}\right] \\
= \frac{\Delta y^2}{12}\left[D_{l+1}\overline{\overline{U}}_{xl+1,m} + D_{l-1}\overline{\overline{U}}_{xl-1,m} + 10D_l\widehat{U}_{xl,m}\right] \\
+ \frac{\Delta y^2}{12}\left[g_{l+1,m} + g_{l-1,m} + 10g_{l,m}\right] + \widehat{T}_{l,m}
\end{aligned} \tag{5.17}$$

Note that scheme (5.17) is of $O(\Delta y^2+\Delta y^2\Delta x^2+\Delta x^4)$. However, this scheme fails to compute at $l = 1$, when the coefficients $B(x)$, $D(x)$ and/or $g(x, y)$ involve terms like $1/x, 1/x^2, 1/xy^3$ and so forth. For an example, if $D(x) = 1/x$, then $D_{l-1} = 1/x_{l-1}$ which blows to infinity at $l = 1$ (since $x_0 = 0$).

So, in order to handle the singularity at $x = 0$, we modify scheme (5.17) such that the order and accuracy of the solution is retained throughout the solution region.

For this purpose, we would need the following approximations:

$$D_{l\pm1} = D_{00} \pm \Delta x D_{10} + \frac{\Delta x^2}{2} D_{20} \pm O(\Delta x^3)$$

$$B_{l\pm1} = B_{00} \pm \Delta x B_{10} + \frac{\Delta x^2}{2} B_{20} \pm O(\Delta x^3)$$

$$g_{l\pm1,m} = g_{00} \pm \Delta x g_{10} + \frac{\Delta x^2}{2} g_{20} \pm O(\Delta x^3)$$

where $g_{l,m} = g_{00} = g(x_l, y_m)$ etc.

Now, substituting above approximations in the difference scheme (5.17) and merging the higher order terms in local truncation error, we obtain the modified scheme as

$$\begin{aligned}
&\left[-12p^2 + \frac{4\Delta y^2}{3} D_{10} - \Delta y^2 D_{00}^2\right] \delta_x^2 U_{l,m} \\
&\quad + \left[p\Delta y\left(6D_{00} + \frac{\Delta x^2}{2} D_{20}\right) - \frac{\Delta y^2 \Delta x}{6} D_{00} D_{10}\right] (2\mu_x \delta_x) U_{l,m} \\
&\quad + \left[-12B_{00} - \Delta x^2 B_{20} - \frac{2\Delta x^2}{3} B_{00} D_{10} + \Delta x^2 B_{10} D_{00}\right] \delta_y^2 U_{l,m} - [B_{00}] \delta_x^2 \delta_y^2 U_{l,m} \\
&\quad + \left[-\Delta x B_{10} + \frac{\Delta x}{2} B_{00} D_{00}\right] \delta_y^2 (2\mu_x \delta_x) U_{l,m} \\
&\quad = -\Delta y^2 \left[12g_{00} + \Delta x^2 (g_{20} - D_{00} g_{10} + \frac{2}{3} D_{10} g_{00})\right] \\
&l = 1(1)N, m = 1(1)M
\end{aligned} \tag{5.18}$$

Note that, the modified scheme (5.18) is of $O(\Delta y^2+\Delta y^2\Delta x^2+\Delta x^4)$ accurate and applicable to both singular and non-singular elliptic differential equations.

Now, consider the Poisson's equation

$$\frac{\partial^2 u}{\partial r^2} + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2} + \frac{1}{r}\frac{\partial u}{\partial r} = \left[4 - \pi^2\right] \cos(\pi\theta), 0 < r, \theta < 1 \tag{5.19}$$

The above equation represents two-dimensional Poisson's equation in cylindrical polar coordinates in $r-\theta$ plane. This problem arises in the simulation of certain semi-bounded plasmas where the electric potential u are to be computed. Replacing the variables (x,y) by (r, θ) and substituting $B_{00} = 1/r_l^2, B_{10} = -2/r_l^3, B_{20} = 6/r_l^4, D_{00} = -1/r_l, D_{10} = B_{00}, D_{20} = B_{10}$ in (5.18), we obtain $O(\Delta y^2+\Delta y^2\Delta x^2+\Delta x^4)$ scheme for the solution of the elliptic equation (5.19).

Similarly, for the 2D Poisson's equation in cylindrical polar coordinates in r–z plane

$$\frac{\partial^2 u}{\partial r^2}+\frac{\partial^2 u}{\partial z^2}+\frac{1}{r}\frac{\partial u}{\partial r}=\cosh z\left[2\cosh r+\frac{1}{r}\sinh r\right], 0<r,z<1 \tag{5.20}$$

We replace the variables (x, y) by (r, z) and setting $B_{00}=1, B_{10}=0=B_{20}, D_{00}=-1/r_l, D_{10}=1/r_l^2, D_{20}=-2/r_l^3$ in (5.18), we can get $O(\Delta y^2+\Delta y^2\Delta x^2+\Delta x^4)$ scheme for the solution of elliptic equation (5.20).

Next consider the Convection-Diffusion equation

$$\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}=\beta\frac{\partial u}{\partial x} \tag{5.21}$$

where $\beta>0$ is a constant and magnitude of β determines the ratio of convection to diffusion. Substituting $B(x)=1.0, D(x)=\beta$ and $g(x,y)=0$ in the difference scheme (5.18) and simplifying, we obtain a difference scheme of $O(\Delta y^2+\Delta y^2\Delta x^2+\Delta x^4)$ accuracy for the solution of the convection-diffusion equation (5.21).

$$\begin{aligned}&\gamma_0 u_{l,m}+\gamma_1 u_{l+1,m}+\gamma_2 u_{l-1,m}+\gamma_3 u_{l,m+1}\\&\quad+\gamma_4 u_{l,m-1}+\gamma_5 u_{l+1,m+1}+\gamma_6 u_{l+1,m-1}\\&\quad+\gamma_7 u_{l-1,m+1}+\gamma_8 u_{l-1,m-1}=0,\\&\quad[l=1(1)N, m=1(1)M]\end{aligned} \tag{5.22}$$

where the coefficients $\gamma_w, w=0,1,2,\ldots 8$ are defined by

$$\begin{aligned}\gamma_0&=24p^2+20+8p^2R^2,\\\gamma_1&=-\left[12p^2-2-12p^2R+4p^2R^2+2R\right],\\\gamma_2&=-\left[12p^2-2+12p^2R+4p^2R^2-2R\right],\\\gamma_3&=\gamma_4=-10,\\\gamma_5&=\gamma_6=-(1-R),\\\gamma_7&=\gamma_8=-(1+R)\end{aligned}$$

where $R=\beta h/2$ is called the *Cell Reynolds number*.

The discretizations result in NM number of linear equations in NM unknowns. Incorporating the boundary conditions, the above system may be expressed in the matrix form as

$$\boldsymbol{Au}=\boldsymbol{b} \tag{5.23}$$

where $\boldsymbol{A}$ is a tri-block diagonal matrix of order $(NM\times NM)$, $\boldsymbol{u}$ is the solution vector and $\boldsymbol{b}$ is the right-hand side column vector arising from the boundary values of problem (5.21).

The coefficient matrix $\boldsymbol{A}$ has a block tri-diagonal structure,

$$\boldsymbol{A} = \text{tri}[-\boldsymbol{L} \quad \boldsymbol{D} - \boldsymbol{U}]_{NM\times NM}$$

with the sub-matrices $-\boldsymbol{L}$, $\boldsymbol{D}$ and $-\boldsymbol{U}$ each of order $(N \times M)$ given by

$$-\boldsymbol{L} = \text{tri}[\,\gamma_8 \quad \gamma_4 \quad \gamma_6\,] = -\boldsymbol{U}, \boldsymbol{D} = \text{tri}[\,\gamma_2 \quad \gamma_0 \quad \gamma_1\,]$$

The iteration matrices of the block Jacobi and block Gauss-Seidel methods are described by

$$\boldsymbol{G}_J = \boldsymbol{D}^{-1}(\boldsymbol{L}+\boldsymbol{U}) \text{ and } \boldsymbol{G}_{GS} = (\boldsymbol{D}-\boldsymbol{L})^{-1}\boldsymbol{U}$$

It can be verified that $\gamma_0 > 0$ and $\gamma_w < 0$ for $w = 1, 2, .., 8$ assuming the diffusion dominated case i.e. $R \leq 1$ and taking $p \geq 1/\sqrt{6}$. One can also easily verify that

$$\gamma_0 = \sum_{w=1}^{8} |\gamma_w|$$

which implies that the coefficient matrix $\boldsymbol{A}$ generated from is weakly diagonally dominant. Since $\boldsymbol{A}$ is irreducible (as its directed graph is strongly connected), we conclude that it is an M-matrix and hence monotone [4].

Now, applying the Jacobi iteration method to the system of Eq. (5.22), we get the iterative scheme for $s = 0, 1, 2\ldots$.

$$\begin{aligned}\left[24p^2\left(1+\frac{R^2}{3}\right)+20\right]u_{l,m}^{(s+1)} &= (1-R)u_{l+1,m+1}^{(s)} + \left[12p^2\left(1-R+\frac{R^2}{3}\right)-2(1-R)\right]u_{l+1,m}^{(s)} \\ &+ (1-R)u_{l+1,m-1}^{(s)} + (1+R)u_{l-1,m+1}^{(s)} + 10u_{l,m-1}^{(s)} \\ &+ \left[12p^2\left(1+R+\frac{R^2}{3}\right)-2(1+R)\right]u_{l-1,m}^{(s)} + (1+R)u_{l-1,m-1}^{(s)} + 10u_{l,m+1}^{(s)}\end{aligned} \tag{5.24}$$

where $u_{l,m}^{(s+1)}$ and $u_{l,m}^{(s)}$ are the successive approximations for $u_{l,m}$ at $(s+1)$th and sth iterations, respectively.

We examine the stability of the Jacobi iteration method by studying the behaviour of the error equation. Let us assume that an error $\varepsilon_{l,m}^{(s)}$ exists at each mesh point (x_l, y_m) at the sth iteration and is of the form

$$\varepsilon_{l,m}^{(s)} = \xi^s A^l B^m \sin\left(\frac{\pi a l}{N+1}\right)\sin\left(\frac{\pi b m}{M+1}\right), 1 \leq a \leq N,\ 1 \leq b \leq M \tag{5.25}$$

where A and B are arbitrary constants and ξ is the propagating factor which determines the rate of growth or decay of the errors. The necessary and sufficient condition for the iterative method to be stable is

$$|\xi| < 1, 1 \le a \le N,\ 1 \le b \le M \tag{5.26}$$

The corresponding error equation is

$$\begin{aligned}\left[24p^2\left(1+\frac{R^2}{3}\right)+20\right]\varepsilon_{l,m}^{(s+1)} &= (1-R)\left(\varepsilon_{l+1,m+1}^{(s)}+\varepsilon_{l+1,m-1}^{(s)}\right)\\ &+\left[12p^2\left(1-R+\frac{R^2}{3}\right)-2(1-R)\right]\varepsilon_{l+1,m}^{(s)}\\ &+(1+R)\left(\varepsilon_{l-1,m+1}^{(s)}+\varepsilon_{l-1,m-1}^{(s)}\right)+10\left(\varepsilon_{l,m-1}^{(s)}+\varepsilon_{l,m+1}^{(s)}\right)\\ &+\left[12p^2\left(1+R+\frac{R^2}{3}\right)-2(1+R)\right]\varepsilon_{l-1,m}^{(s)}\end{aligned} \tag{5.27}$$

Substituting (5.25) in error equation (5.27), we obtain the characteristic equation

$$\begin{aligned}&\left(24p^2+20+8p^2R^2\right)\xi\sin\left(\frac{\pi al}{N+1}\right)\sin\left(\frac{\pi bm}{M+1}\right)\\ &=\sin\left(\frac{\pi al}{N+1}\right)\sin\left(\frac{\pi bm}{M+1}\right)\left[10\left(B+B^{-1}\right)\cos\left(\frac{\pi b}{M+1}\right)\right.\\ &\quad+\left(B+B^{-1}\right)\cos\left(\frac{\pi a}{N+1}\right)\cos\left(\frac{\pi b}{M+1}\right)\left[A(1-R)+A^{-1}(1+R)\right]\\ &\quad\left.+\cos\left(\frac{\pi a}{N+1}\right)\left[\left(12p^2-2+4p^2R^2\right)\left(A+A^{-1}\right)+\left(12p^2R-2R\right)\left(A^{-1}-A\right)\right]\right]\\ &\quad+\sin\left(\frac{\pi al}{N+1}\right)\cos\left(\frac{\pi bm}{M+1}\right)\left[10\left(B-B^{-1}\right)\sin\left(\frac{\pi b}{M+1}\right)\right.\\ &\quad\left.+\left(B-B^{-1}\right)\cos\left(\frac{\pi a}{N+1}\right)\sin\left(\frac{\pi b}{M+1}\right)\left[A(1-R)+A^{-1}(1+R)\right]\right]\\ &+\cos\left(\frac{\pi al}{N+1}\right)\sin\left(\frac{\pi bm}{M+1}\right)\left[\left(B+B^{-1}\right)\cos\left(\frac{\pi b}{M+1}\right)\sin\left(\frac{\pi al}{N+1}\right)\left[A(1-R)-A^{-1}(1+R)\right]\right.\\ &\quad+\sin\left(\frac{\pi a}{N+1}\right)\left[\left(12p^2-2+4p^2R^2\right)\left(A-A^{-1}\right)-\left(12p^2R-2R\right)\left(A+A^{-1}\right)\right]\\ &\quad\left.+\cos\left(\frac{\pi al}{N+1}\right)\cos\left(\frac{\pi bm}{M+1}\right)\left[\left(B-B^{-1}\right)\sin\left(\frac{\pi al}{N+1}\right)\sin\left(\frac{\pi b}{M+1}\right)\left[A(1-R)-A^{-1}(1+R)\right]\right]\end{aligned} \tag{5.28}$$

Comparing both sides, we get

$$10\left(B-B^{-1}\right)+\left(B-B^{-1}\right)\cos\left(\frac{\pi a}{N+1}\right)\left[A(1-R)+A^{-1}(1+R)\right]=0$$

$$\left(B+B^{-1}\right)\cos\left(\frac{\pi b}{M+1}\right)\left[A(1-R)-A^{-1}(1+R)\right]+\left(12p^2-2+4p^2R^2\right)\left(A-A^{-1}\right)-\left(12p^2R-2R\right)\left(A+A^{-1}\right)=0$$

and

$$\left(B-B^{-1}\right)\left[A(1-R)-A^{-1}(1+R)\right]=0$$

On solving, we get $B = 1$ and

$$A=\left\{\frac{(1+R)\cos\left(\frac{\pi b}{M+1}\right)+6p^2-1+2p^2R^2+6p^2R-R}{(1-R)\cos\left(\frac{\pi b}{M+1}\right)+6p^2-1+2p^2R^2-6p^2R+R}\right\}^{1/2}$$

The propagating factor becomes

$$\begin{aligned}&\left(24p^2+20+8p^2R^2\right)\xi\\&\quad=2\cos\left(\frac{\pi b}{M+1}\right)\left[10+\cos\left(\frac{\pi a}{N+1}\right)\left(A(1-R)+A^{-1}(1+R)\right)\right]\\&\quad+\cos\left(\frac{\pi a}{N+1}\right)[(12p^2-2+4p^2R^2-12p^2R+2R)A\\&\quad+\left(12p^2-2+4p^2R^2+12p^2R-2R\right)A^{-1}]\end{aligned}$$

Now the largest value of $\cos\left(\frac{\pi al}{N+1}\right)$ and $\cos\left(\frac{\pi bm}{M+1}\right)$ occur when $a = b = 1$.

$$\begin{aligned}\left(24p^2+20+8p^2R^2\right)\xi&=20\cos\left(\frac{\pi}{M+1}\right)\\&+2\cos\left(\frac{\pi}{N+1}\right)\left[(1-R)\cos\left(\frac{\pi}{M+1}\right)+6p^2-1+2p^2R^2-6p^2R+R\right]A\\&+2\cos\left(\frac{\pi}{N+1}\right)A^{-1}\left[(1+R)\cos\left(\frac{\pi}{M+1}\right)+6p^2-1+2p^2R^2+6p^2R-R\right]\end{aligned}$$

Substituting the value of A and simplifying, the propagating factor ξ_J for the Jacobi iteration method is obtained as

$$\xi_J = \frac{1}{5+6p^2\left(1+\frac{R^2}{3}\right)}\left\{5\cos\left(\frac{\pi}{M+1}\right)\right.$$
$$+\cos\left(\frac{\pi}{N+1}\right)\sqrt{6p^2\left(1-R+\frac{R^2}{3}\right)-(1-R)\left(1-\cos\left(\frac{\pi}{M+1}\right)\right)} \quad (5.29)$$
$$\left.\times\sqrt{6p^2\left(1+R+\frac{R^2}{3}\right)-(1+R)\left(1-\cos\left(\frac{\pi}{M+1}\right)\right)}\right\}.$$

Thus, the Jacobi Iteration method is stable for those values of R such that $|\xi_J| < 1$ and the rate of convergence of the Jacobi iteration method is given by

$$v_J = -\log\xi_J$$

Similarly, applying the Gauss-Siedel iteration method into the system of Eq. (5.22) we can obtain propagating factor of for the Gauss-Seidel iteration method.

Consequently, the spectral radii ρ of the block Jacobi and block Gauss-Seidel matrices are related by

$$\rho(\boldsymbol{G}_{GS}) = \rho(\boldsymbol{G}_J)^2 \quad (5.30)$$

Hence, the associated iteration

$$\boldsymbol{u}^{(k+1)} = \boldsymbol{G}\boldsymbol{u}^{(k)} + \boldsymbol{c} \quad (5.31)$$

converges for any initial guess where, $\boldsymbol{G}$ is Jacobi or Gauss-Seidel iteration matrix.

5.4 Numerical Illustrations

Substituting the central difference approximations in the differential equation (5.1), we obtain a central difference scheme of $O(\Delta y^2+\Delta x^2)$ of the form

$$A_{l,m}\bar{U}_{xxl,m} + B_{l,m}\bar{U}_{yyl,m} = f\left(x_l, y_m, U_{l,m}, \bar{U}_{xl,m}, \bar{U}_{yl,m}\right) + O\left(\Delta y^2 + \Delta x^2\right)$$

Numerical experiments are carried out to illustrate our method and to demonstrate computationally its convergence. We solve the following two-dimensional elliptic boundary value problems on unequal mesh both on rectangular and cylindrical polar coordinates whose exact solutions are known to us. The Dirichlet boundary conditions can be obtained using the exact solutions as a test procedure.

We also compare our method with the central difference scheme and the methods discussed in [5] in terms of solution accuracy. In all cases, we have taken the initial guess $\boldsymbol{u}(x_l, y_m) = 0$. The iterations were stopped when the absolute error tolerance became $\leq 10^{-10}$.

Example 1 $\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \beta \frac{\partial u}{\partial x}$, $0 < x, y < 1$ (Convection-diffusion equation)

The exact solution is given by

$$u(x, y) = \mathrm{e}^{\frac{\beta x}{2}} \frac{\sin \pi y}{\sinh \sigma} \left[2\mathrm{e}^{\frac{\beta}{2}} \sinh \sigma x + \sinh \sigma (1 - x) \right],$$

where $\sigma^2 = \pi^2 + \frac{\beta^2}{4}$. The maximum absolute errors for u are tabulated in Table 5.1. Figure 5.2a, b demonstrate a comparison of the plots of the numerical and exact solution of $u(x, y)$ for the values $\beta = 30$ and $\gamma = (\Delta y/\Delta x^2) = 20$.

Example 2 (Poisson's equation in polar coordinates)

$$(a) \qquad \frac{\partial^2 u}{\partial r^2} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \theta^2} + \frac{1}{r} \frac{\partial u}{\partial r} = \left[4 - \pi^2 \right] \cos(\pi \theta), \quad 0 < r, \theta < 1$$

The exact solutions are given by $u(r, \theta) = r^2 \cos \pi \theta$.

$$(b) \qquad \frac{\partial^2 u}{\partial r^2} + \frac{\partial^2 u}{\partial z^2} + \frac{1}{r} \frac{\partial u}{\partial r} = \cosh z \left[2 \cosh r + \frac{1}{r} \sinh r \right], \quad 0 < r, z < 1$$

The exact solutions are given by $u(r, z) = \cosh r \cosh z$.

The maximum absolute errors for u are tabulated in Table 5.2. A comparison of the plots of the numerical and exact solution of u for the value $\gamma = (\Delta y/\Delta x^2) = 20$ is shown in the Fig. 5.3a, b.

Example 3 (Steady-state Burgers' Model Equation)

$$\varepsilon \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right) = u \left(\frac{\partial u}{\partial x} + \frac{\partial u}{\partial y} \right) + \mathrm{e}^x \sin \left(\frac{\pi y}{2} \right) \left[\varepsilon \left(1 - \frac{\pi^2}{4} \right) - \mathrm{e}^x \left(\sin \left(\frac{\pi y}{2} \right) + \frac{\pi}{2} \cos \left(\frac{\pi y}{2} \right) \right) \right], 0 < x, y < 1$$

Table 5.1 Example 1: the maximum absolute errors ($\gamma = \frac{\Delta y}{\Delta x^2} = 20$)

Δx	Proposed $O(\Delta y^2 + \Delta y^2 \Delta x^2 + \Delta x^4)$-method			$O(\Delta y^4 + \Delta y^2 \Delta x^2 + \Delta x^4)$-method		
	$\beta = 10$	$\beta = 20$	$\beta = 30$	$\beta = 10$	$\beta = 20$	$\beta = 30$
$\frac{1}{10}$	0.1062E-01	0.1823E-01	0.4618E-01	0.7918E-01	0.8220E-01	0.8998E-01
$\frac{1}{20}$	0.6971E-03	0.1213E-02	0.3922E-02	0.4919E-02	0.5155E-02	0.5676E-02
$\frac{1}{40}$	0.4352E-04	0.7360E-04	0.2312E-03	0.3102E-03	0.3214E-03	0.3511E-03

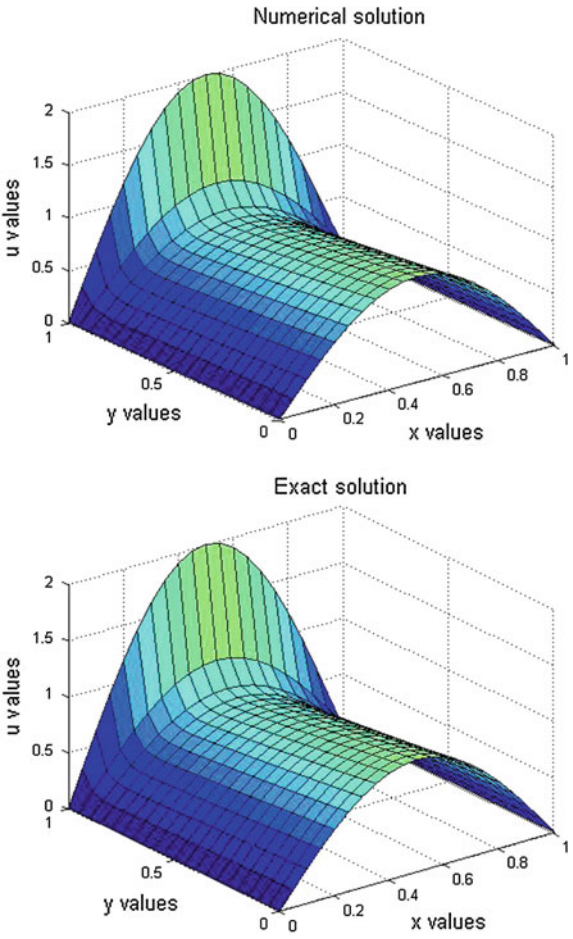

Fig. 5.2 Comparison of plots of solution of Example 1. **a** Convection-diffusion equation $\gamma = 20$, $\beta = 30$ (numerical solution). **b** Convection-diffusion equation $\gamma = 20$, $\beta = 30$ (exact solution)

Table 5.2 Example 2: the maximum absolute errors ($\gamma = \frac{\Delta y}{\Delta x^2} = 20$)

Δx	Proposed $O(\Delta y^2 + \Delta y^2 \Delta x^2 + \Delta x^4)$-method		$O(\Delta y^2 + \Delta y^2 \Delta x^2 + \Delta x^4)$-method discussed	
	Ex. 2(a)	Ex. 2(b)	Ex. 2(a)	Ex. 2(b)
$\frac{1}{10}$	0.2976E-02	0.3574E-03	0.5018E-02	0.6662E-03
$\frac{1}{20}$	0.1917E-03	0.2343E-04	0.3072E-03	0.4155E-04
$\frac{1}{40}$	0.1202E-04	0.1448E-05	0.1911E-04	0.2614E-05

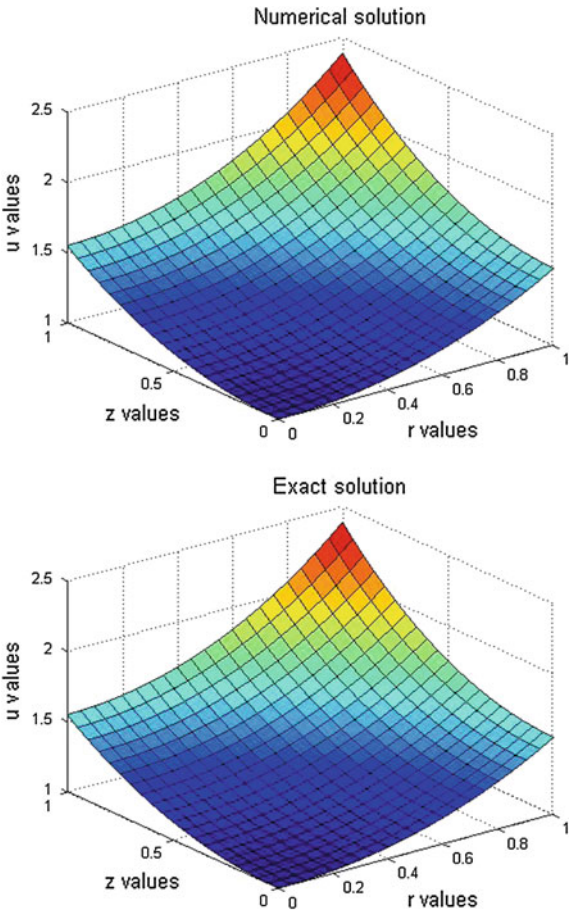

Fig. 5.3 Comparison of plots of solution of Example 2. **a** Poisson's equation (r–z plane) $\gamma = 20$ (numerical solution). **b** Poisson's equation (r–z plane) $\gamma = 20$ (exact solution)

where $R_e = \varepsilon^{-1} > 0$ is called Reynolds number. The exact solution is given by $u(x, y) = e^x \sin\left(\frac{\pi y}{2}\right)$. The maximum absolute errors for u are tabulated in Table 5.3 for various values of R_e. Figure 5.4a, b demonstrate a comparison of the plots of the numerical and exact solution of $u(x,y)$ for the values $R_e = 100$ and $\gamma = (\Delta y/\Delta x^2) = 20$.

Finally, Table 5.4 shows that our method works as a fourth order method with fixed mesh parameter $\gamma = \Delta y/\Delta x^2$. The order of convergence may be obtained by using the formula

$$\log\left(\frac{e_{\Delta x_1}}{e_{\Delta x_2}}\right) / \log\left(\frac{\Delta x_1}{\Delta x_2}\right)$$

Table 5.3 Example 3: the maximum absolute errors ($\gamma = \frac{\Delta y}{\Delta x^2} = 20$)

Δx	Proposed $O(\Delta y^2 + \Delta y^2 \Delta x^2 + \Delta x^4)$-method		$O(\Delta y^4 + \Delta y^2 \Delta x^2 + \Delta x^4)$-method discussed in	
	$R_e = 10$	$R_e = 100$	$R_e = 10$	$R_e = 100$
$\frac{1}{10}$	0.1022E-01	0.8190E-02	0.4242E-01	0.1244E-01
$\frac{1}{20}$	0.5887E-03	0.7330E-03	0.2510E-02	0.7711E-03
$\frac{1}{40}$	0.3683E-04	0.4347E-04	0.1516E-03	0.4748E-04

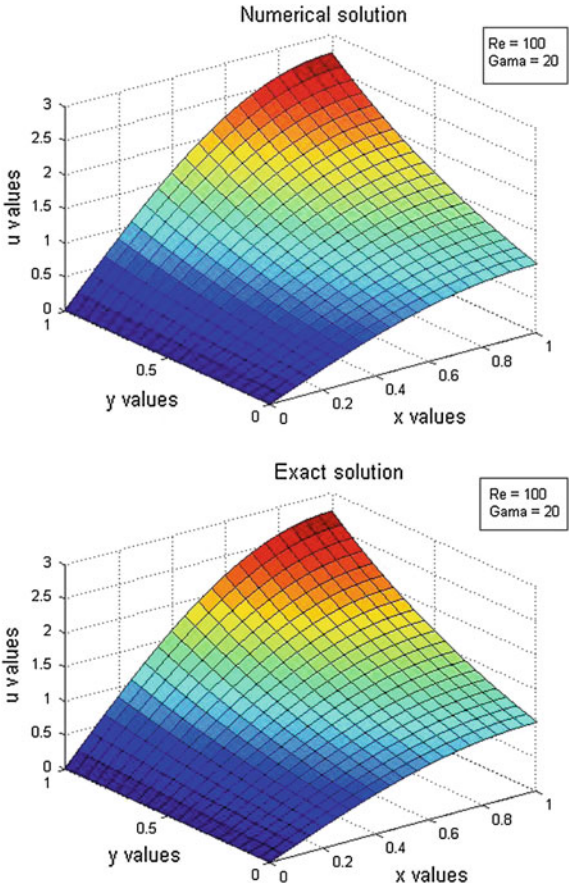

Fig. 5.4 comparison of plots of solution of Example 3. **a** Steady-state Burger's equation $\gamma = 20$, $R_e = 100$ (numerical solution). **b** Steady-state Burger's equation $\gamma = 20$, $R_e = 100$ (exact solution)

where $e_{\Delta x_1}$ and $e_{\Delta x_2}$ are maximum absolute errors for two uniform mesh widths Δx_1 and Δx_2, respectively. For computation of order of convergence of the proposed method, we have considered errors for last two values of Δx, i.e., $\Delta x_1 = \frac{1}{20}, \Delta x_2 = \frac{1}{40}$ for the above discussed elliptic partial differential equations.

Table 5.4 Fourth order convergence: $\Delta x_1 = \frac{1}{20}$, $\Delta x_2 = \frac{1}{40}$, $\gamma = \frac{\Delta y}{\Delta x^2} = 20$

Example	Parameters	Order of the method
1	$\beta = 10$	4.00
	$\beta = 20$	4.04
	$\beta = 30$	4.08
2	Ex. 2(a)	4.00
	Ex. 2(b)	4.01
3	$R_e = 10$	4.00
	$R_e = 100$	4.07
	$\alpha = 20$	3.99

5.5 Conclusion and Observations

Available numerical methods based on cubic spline approximations for the numerical solution of non-elliptic equations are of $O(\Delta y^2+\Delta x^2)$ accurate. Although 9-point finite difference approximations of $O(\Delta y^4+\Delta y^2\Delta x^2+\Delta x^4)$ accurate for the solution of nonlinear and quasi-linear elliptic differential equations are available in the literature, but these methods require five evaluations of the function f. In this article, using the same number of grid points and three evaluations of the function f, we have derived a new stable cubic spline method of $O(\Delta y^2+\Delta y^2\Delta x^2+\Delta x^4)$ accuracy for the solution of nonlinear elliptic equation (5.1). However, for a fixed parameter $\gamma = \frac{\Delta y}{\Delta x^2}$, the proposed method behaves like a fourth order method. The accuracy of the proposed method is exhibited from the computed results. The proposed method is applicable to Poisson's equation in polar coordinates, and two-dimensional Burgers' equation, which is main highlight of the work.

References

1. Jain MK, Jain RK, Mohanty RK (1989) A fourth order difference method for elliptic equations with non linear first derivative terms. Numer Meth Partial Diff Eqs 5:87–95
2. Jain MK, Aziz T (1983) Cubic spline solution of two point boundary value problems with significant first derivatives. Comput Meth Appl Mech Eng 39:83–91
3. Appl. Math. Modell. 37:155–171
4. Varga RS (2000) Matrix iterative analysis. Springer, New York
5. Mohanty RK, Karaa S, Arora U (2006) Fourth order nine point unequal mesh discretization for the solution of 2-D non linear elliptic partial differential equations. Neural Parallel Sci Comput 14:453–470

Chapter 6
Pricing of Path-Dependent European-Type Options Using Monte Carlo Simulation

Siddhartha P. Chakrabarty

Abstract This expository article highlights the significance of Monte Carlo simulation in pricing of options. We discuss the various types of financial derivatives, particularly options and their classifications. The discrete and continuous time models for the underlying assets are dwelled upon. We consider a geometric Brownian motion (GBM) based model for stock price process and discuss the payoffs of plain vanilla as well as path-dependent European-type options, namely, barrier, lookback, and Asian. We mention the option pricing formula for plain vanilla European option and describe the Monte Carlo approach to option pricing with illustrative algorithms and results for some of these options.

Keywords Option pricing · Monte Carlo simulation

Mathematics Subject Classification 91G60

6.1 Introduction

Financial derivatives have come to occupy a position of great importance in the global financial markets, especially in terms of growth, diversity and volume. They are used for a variety of purposes, the most common ones being hedging, speculation, and arbitrage [1] and are usually traded over-the-counter (OTC) or are exchange traded. Some of the typical derivative contracts are options, forwards and futures, and swaps [1–3]. The term "financial derivative" is motivated by the fact that the values of these derivatives are derived from more basic underlying assets or securities. The ever increasing complexity of payoff structures of derivatives, especially the OTC traded ones, have resulted in the use of sophisticated mathematical techniques in the

S.P. Chakrabarty (✉)
Department of Mathematics, Indian Institute of Technology Guwahati, Guwahati 781039, Assam, India
e-mail: pratim@iitg.ernet.in

A.H. Siddiqi et al. (eds.), *Mathematical Models, Methods and Applications*, Industrial and Applied Mathematics, DOI 10.1007/978-981-287-973-8_6

pricing and valuation of such derivatives. In this expository article, we discuss one such technique, namely, Monte Carlo simulation and its application to pricing of path-dependent European-type options.

Options are one of the most common and popular financial derivatives [1–5]. These are derivative contacts which give the owner the right to buy or sell the underlying asset, for a certain price, on or before a certain date. Options can broadly be classified as a call option (where the holder or the buyer of the option has the right to buy the underlying asset from the writer or the seller of the option) and a put option (where the holder or the buyer of the option has the right to sell the underlying asset to the writer or the seller of the option). Another classification of options is European and American. In the case of the former, the holder of the option can exercise it only at a fixed future time called the expiration date. In case of the latter, however, this exercise by the holder can take place at any time up to and including the expiration date. Both these types of options can either be plain vanilla or the more sophisticated path-dependent ones.

Since options confer a right to the holder of the option and imposes an obligation on the writer of the option, the former must pay a "premium" to the latter, in lieu of this right. This "premium" is referred to as the "price" of the option. The determination of the correct and fair option price is an important problem in today's financial markets. While options have a wide variety of securities as underlying, we shall solely focus on options which have stocks as the underlying. Modeling of stocks has been done in both discrete and continuous time settings. While discrete time models like the binomial model have been used extensively, the continuous time geometric Brownian motion (GBM) model for asset pricing is more commonly used in literature.

The seminal paper of Black and Scholes [6, 7], that appeared in 1973, was a significant and important breakthrough in this area. In this paper, the authors for the first time gave a closed form option pricing formula for European options, in continuous time. Cox, Ross, and Rubinstein [8, 9], in 1979, presented a simple discrete time model for pricing of options. They obtained a formula for pricing of options and derived the Black–Scholes formula as a limiting case of their formula, based on symmetric random walk.

6.2 Model for Asset Price and Options

For the purpose of this article, we will consider the GBM model [2–5, 7] for the underlying stock of the option under consideration. The model is given by the following stochastic differential equation for the stock price process s_t,

$$\mathrm{d}s_t = \mu s_t \mathrm{d}t + \sigma s_t \mathrm{d}w_t \tag{6.1}$$

where μ is the drift and σ is the volatility of the stock prices. Here, w_t is the Wiener process, under the risk-neutral measure $\mathbb{P}$. While μ and σ have been taken to be

constant here, they can be dependent both on time as well the stock price. Note that, the Wiener process, w_t is a random variable with: (i) $w_0 = 0$, (ii) for $0 \leq s < t$, $w_t - w_s$ follows the normal distribution $\mathcal{N}(0, t-s)$ and (iii) increments of w_t over nonoverlapping time intervals are independent of each other.

The pricing of European-type options on a stock (following the GBM), primarily depends on the expected payoff from the option at expiration date. This is due to the fact that exercising of a European-type of option can take place only at expiration date, which will be denoted by T. In case of a plain vanilla European call option, the holder of the option has the right to buy the underlying stock from the writer, for a fixed price called the strike price, which will be represented by k. It is obvious that the holder will exercise only if the asset price s_T at expiration is at least the strike price k, in which case the profit for the holder will be $s_T - k$. Else the option will expire. Thus, the payoff will be $\max\{s_T - k, 0\}$. Similarly, the payoff for a plain vanilla European put option is given by $\max\{k - s_T, 0\}$. We will now discuss the payoff functions for three different path-dependent or exotic European options, whose pricing using the Monte Carlo simulation will be done later [4, 10].

A barrier option is an option whose payoff is switching in nature and depends on whether the underlying asset prices cross a predefined threshold level during the lifetime of the option. There are basically four such types of barrier call options. A down-and-out barrier call option has the payoff for a European call option provided the asset price does not go below a prespecified barrier $b < s_0$ and zero if it does. A down-and-in barrier call option has the payoff for a European call option provided the asset price goes below a prespecified barrier $b < s_0$ and zero if it does not. An up-and-out barrier call option has the payoff for a European call option provided the asset price does not go above a prespecified barrier $b > s_0$ and zero if it does. An up-and-in barrier call option has the payoff for a European call option provided the asset price goes above a prespecified barrier $b > s_0$ and zero if it does not. The payoffs for barrier put options are similar.

A lookback option is one whose payoff depends on either the maximum or the minimum price of the underlying asset during the lifetime of the option. The fixed strike lookback call and put option have payoffs $\max\{s_{\max} - k, 0\}$ and $\max\{k - s_{\min}, 0\}$, respectively. On the other hand, floating strike lookback call and put option have payoffs $\max\{s_T - s_{\min}, 0\}$ and $\max\{s_{\max} - s_T, 0\}$, respectively. Here, $s_{\max}$ and $s_{\min}$ denote the maximum and the minimum stock prices, respectively, during the lifetime of the option.

While barrier and lookback options focus on some fixed barrier and extreme values for the stock, Asian options take into account the average pattern of the price of the stock. The payoffs for Asian options are functions of the average price of the stock during the lifetime of the option. We mention four different types of Asian options in continuous time. The average price Asian call and put options have payoffs $\max\left(\frac{1}{T}\int_0^T s_\tau \mathrm{d}\tau - k, 0\right)$ and $\max\left(k - \frac{1}{T}\int_0^T s_\tau \mathrm{d}\tau, 0\right)$, respectively. Similarly, the average strike Asian call option and put options have payoffs $\max\left(s_T - \frac{1}{T}\int_0^T s_\tau \mathrm{d}\tau, 0\right)$ and $\max\left(\frac{1}{T}\int_0^T s_\tau \mathrm{d}\tau - s_T, 0\right)$, respectively.

6.3 Pricing of Options Using Monte Carlo Simulation

Black and Scholes [6], provided the following closed form formula [2, 4, 7] for the price of a European call and put option,

$$c(s,t) = N(d_+)s - N(d_-)K\mathrm{e}^{-r(T-t)}$$
$$p(s,t) = N(-d_-)K\mathrm{e}^{-r(T-t)} - N(-d_+)s.$$

where

$$d_+ = \frac{\ln\left(\frac{s}{k}\right) + \left(\mu + \frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}} \text{ and } d_- = \frac{\ln\left(\frac{s}{k}\right) + \left(\mu - \frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}}$$

This solution was obtained by analytically solving the famous Black–Scholes equation, with appropriate final and boundary conditions for European call and put options. In case of most options, however, such closed form pricing formula cannot be obtained. One can then resort to numerical techniques for PDEs [11] and solve variations of the Black–Scholes equation that arise in cases of such options. This approach, however, has limitations, in terms of an appropriate numerical scheme, especially when there are several underlying assets instead of only one (as is the case with plain vanilla options) and also for path-dependent options. The most practical approach to the determination of option price, then, is to resort to Monte Carlo simulation.

Monte Carlo simulation is used to determine the expected value of a random variable, by generating a large number of independent sample random variables [10–14]. In the case of option pricing with a stock as the underlying asset, a large number of sample stock paths are generated and their corresponding payoffs at the expiration is determined. The price of the option is then given by the risk-neutral discounting of the average or expectation of these variables. For the purpose of generation of sample paths, we use the following solution of Eq. (6.1) that can be obtained using Ito's lemma,

$$s_t = s_0\mathrm{e}^{\left(\mu - \frac{\sigma^2}{2}\right)t + \sigma w_t} = s_0\mathrm{e}^{\left(\mu - \frac{\sigma^2}{2}\right)t + \sigma\sqrt{t}z} \tag{6.2}$$

where $z \sim \mathcal{N}(0,1)$. Note that $w_t = \sqrt{t}z \sim \mathcal{N}(0,t)$. Once the payoff for a large number (say M) of sample paths is generated, the option price based on these paths is given by the risk-neutral valuation formula

$$m_v = \mathrm{e}^{-rT}E[\text{Payoff}_i] = \mathrm{e}^{-rT}\left[\frac{1}{M}\sum_{i=1}^{M}\text{Payoff}_i\right],$$

where Payoff_i is the payoff from the ith path, r is the risk-free rate, and $\mu = r$ under risk-neutral valuation. In case of path-dependent options, we need to keep track of

the stock price at all time points (say $N + 1$ with the length of each of the N time intervals taken to be $\Delta t = T/N$, though one could take nonuniform time intervals as a more general case) for each sample path $i = 1 : M$.

Also, the sample variance is given as

$$\sigma_v^2 = \frac{1}{M-1}\sum_{i=1}^{M}(\text{Payoff}_i - m_v)^2.$$

Note that the confidence interval (say 95 %) for the option price is

$$\left[m_v - 1.96\frac{\sigma_v}{\sqrt{M}}, m_v + 1.96\frac{\sigma_v}{\sqrt{M}}\right].$$

We outline the simulation algorithm for one case each of plain vanilla European, barrier, lookback, and Asian options. The other cases can be dealt with on similar lines.

1. European put option
   for $i = 1 : M$
   generate sample $z_i \sim \mathcal{N}(0,1)$
   set $s_i = s_0 \mathrm{e}^{\left(\mu - \frac{1}{2}\sigma^2\right)T + \sigma\sqrt{T}z_i}$
   set Payoff$_i = \max(k - s_i, 0)$
   end
   Price of option, $m_v = \mathrm{e}^{-rT}\left[\frac{1}{M}\sum_{i=1}^{M}\textbf{Payoff}_i\right]$
   Sample variance, $\sigma_v^2 = \frac{1}{M-1}\sum_{i=1}^{M}(\textbf{Payoff}_i - m_v)^2$
   95 % **confidence interval for the price of the option**

$$\left[m_v - 1.96\frac{\sigma_v}{\sqrt{M}}, m_v + 1.96\frac{\sigma_v}{\sqrt{M}}\right].$$

2. Up-and-out barrier call option
   for $i = 1 : M$
   for $j = 0 : N - 1$
   generate sample $z_j \sim \mathcal{N}(0,1)$
   set $s_{i,j+1} = s_{i,j}\mathrm{e}^{\left(\mu - \frac{1}{2}\sigma^2\right)\Delta t + \sigma\sqrt{\Delta t}z_j}$
   end
   set $s_{i,\max} = \max_{0 \le j \le N} s_{i,j}$
   if $s_{i,\max} < b$ **set Payoff**$_i = \max(s_{i,T} - k, 0)$
   else set Payoff$_i = 0$
   end

Price of option, $m_v = \mathrm{e}^{-rT}\left[\frac{1}{M}\sum_{i=1}^{M} \textbf{Payoff}_i\right]$
Sample variance, $\sigma_v^2 = \frac{1}{M-1}\sum_{i=1}^{M}(\textbf{Payoff}_i - m_v)^2$
95 % **confidence interval for the price of the option**

$$\left[m_v - 1.96\frac{\sigma_v}{\sqrt{M}}, m_v + 1.96\frac{\sigma_v}{\sqrt{M}}\right].$$

3. Floating strike lookback call option
for $i = 1 : M$
for $j = 0 : N - 1$
generate sample $z_j \sim \mathcal{N}(0,1)$
set $s_{i,j+1} = s_{i,j}\mathrm{e}^{\left(\mu - \frac{1}{2}\sigma^2\right)\Delta t + \sigma\sqrt{\Delta t}z_j}$
end
set $s_{i,\min} = \min_{0 \le j \le N} s_{i,j}$
set $\textbf{Payoff}_i = \max(s_{i,T} - s_{i,\min}, 0)$
end
Price of option, $m_v = \mathrm{e}^{-rT}\left[\frac{1}{M}\sum_{i=1}^{M} \textbf{Payoff}_i\right]$
Sample variance, $\sigma_v^2 = \frac{1}{M-1}\sum_{i=1}^{M}(\textbf{Payoff}_i - m_v)^2$
95 % **confidence interval for the price of the option**

$$\left[m_v - 1.96\frac{\sigma_v}{\sqrt{M}}, m_v + 1.96\frac{\sigma_v}{\sqrt{M}}\right].$$

4. Average price Asian put option
for $i = 1 : M$
for $j = 0 : N - 1$
generate sample $z_j \sim \mathcal{N}(0,1)$
set $s_{i,j+1} = s_{i,j}\mathrm{e}^{\left(\mu - \frac{1}{2}\sigma^2\right)\Delta t + \sigma\sqrt{\Delta t}z_j}$
end
set $s_{i,\text{avg}} = \sum_{j=0}^{N} s_{i,j}\frac{\Delta t}{T}$
set $\textbf{Payoff}_i = \max(k - s_{i,\text{avg}}, 0)$
end
Price of option, $m_v = \mathrm{e}^{-rT}\left[\frac{1}{M}\sum_{i=1}^{M} \textbf{Payoff}_i\right]$
Sample variance, $\sigma_v^2 = \frac{1}{M-1}\sum_{i=1}^{M}(\textbf{Payoff}_i - m_v)^2$
95 % **confidence interval for the price of the option**

$$\left[m_v - 1.96\frac{\sigma_v}{\sqrt{M}}, m_v + 1.96\frac{\sigma_v}{\sqrt{M}}\right].$$

Table 6.1 Results for Monte Carlo simulation

Option type	Option price	Sample variance	95 % confidence interval
European put	20.3997	364.4944	[19.8705, 20.9289]
Up-and-out barrier call	6.8101	239.8059	[6.3808, 7.2393]
Floating strike lookback call	23.7967	532.7144	[23.1570, 24.4365]
Average price Asian put	17.8306	187.0266	[17.4515, 18.2097]

6.4 Results and Discussion

We ran the simulations for the four different path-dependent options outlined in Sect. 6.3 using MatLab™. For this purpose, we took the number of sample paths to be $M = 5000$ and the number of time intervals to be $N = 1000$. The parameters used in the simulation were $s(0) = 100$, $r = 6$ %, $\sigma = 30$ %, $k = 120$, and $b = 250$. The expiration time was $T = 1$, thereby resulting in the length of time intervals (taken to be uniform) $\Delta t = T/N = 10^{-3}$. The results from the simulation are given in Table 6.1

The simulation results are for option prices in a certain confidence interval. This is a disadvantage of this method, since the price is not unique but lies in a certain range. The numerical PDE approach to the same pricing problems, which potentially leads to a unique option value, poses challenges in terms of implementation, thereby limiting the usage of this approach in practical situation. Consequently, from the point of view of implementation, the Monte Carlo simulation technique is preferred among the commonly used techniques.

References

1. Hull JC (2006) Options, futures and other derivatives. Prentice Hall, New Delhi
2. Capinski M, Zastawniak T (2003) Mathematics for finance: an introduction to financial engineering. Springer, New York
3. Cvitanic J, Zapatero F (2004) introduction to the economics and mathematics of financial markets. Prentice Hall, New Delhi
4. Wilmott P, Howison S, Dewynne J (1995) The mathematics of financial derivatives. Cambridge University Press, Cambridge
5. Roman S (2004) Introduction to the mathematics of finance: from risk management to options pricing. Springer, London
6. Black F, Scholes M (1973) The pricing of options and corporate liabilities. J Polit Econ 81:637–659
7. Shreve S (2004) Stochastic calculus for finance: continuous-time models. Springer, New York
8. Cox JC, Ross SA, Rubinstein M (1979) Option pricing: a simplified approach. J Financ Econ 7:229–263
9. Shreve S (2004) Stochastic calculus for finance: the binomial asset pricing model. Springer, Hamburg
10. Higham DJ (2004) An introduction to financial option valuation: mathematics, stochastics and computation. Cambridge University Press, Cambridge
11. Seydel R (2006) Tools for computational finance. Springer, Berlin

12. Glasserman P (2003) Monte Carlo methods in financial engineering. Springer, New York
13. Boyle PP (1977) Options: A Monte Carlo approach. J Financ Econ 4(3):323–338
14. Broadie M, Glasserman P (1996) Estimating security price derivatives using simulation. Manag Sci 42(2):269–285

Chapter 7
On the Finite Element Approximation of the Impulse Control Quasivariational Inequality

Messaoud Boulbrachene

Abstract In this paper, we develop a new approach for the standard finite element approximation in the maximum norm for the impulse control quasivariational inequality. We establish the optimal convergence order combining the Bensoussan–Lions algorithm and the concepts of subsolution and discrete.

Keywords Quasivariational inequalities · Bensoussan–Lions algorithm · Subsolution · Finite element · Discrete regularity · L^{∞} error estimate

2000 Mathematics Subject Classification Primary 65N30 · 65N15

7.1 Introduction

In this paper, we are interested in the standard finite element approximation in the maximum norm of the following quasivariational inequality (QVI)

$$\begin{cases} a(u, v-u) \geqq (f, v-u) \, \forall v \in H_0^1(\Omega) \\ u \leq Mu, \; v \leq Mu \end{cases} \tag{7.1.1}$$

Here, Ω is a bounded convex domain of $\mathbb{R}^N$, $N \geq 1$, with boundary $\partial\Omega$, $(.,.)$ denotes the scalar product in $L^2(\Omega)$, f is a nonnegative right-hand side in $L^{\infty}(\Omega)$, $a(.,.)$ denotes the bilinear form associated with an elliptic second order differential operator $\mathcal{A}$, and M is a nonlinear operator from $L^{\infty}(\Omega)$ into itself defined as

$$Mu = k + \inf u(x+\xi), \xi \geq 0, x+\xi \in \bar{\Omega}\}; k > 0 \tag{7.1.2}$$

M. Boulbrachene (✉)
Department of Mathematics and Statistics, Sultan Qaboos University,
P.O. Box 36, Muscat 123, Oman
e-mail: boulbrac@squ.edu.om

A.H. Siddiqi et al. (eds.), *Mathematical Models, Methods and Applications*,
Industrial and Applied Mathematics, DOI 10.1007/978-981-287-973-8_7

Problem (7.1.1) is analogous to the obstacle problem where the obstacle function is replaced with an implicit one, depending upon the solution sought. The terminology quasivariational inequality being chosen is a result of this remark.

This QVI arises in impulse control problems: an introduction to impulse control with numerous examples and applications can be found in Bensoussan and Lions [1].

Its numerical approximation in the L^∞ norm has recently gained a high interest in computational finance (see [2, 3]).

Let τ_h denote a regular and quasiuniform triangulation of $\Omega; h > 0$ is the mesh size. Let $\mathbb{V}_h$ denote the finite element space consisting of continuous piecewise linear functions vanishing on $\partial\Omega$, $\{\varphi_i\}, i = 1, 2, \ldots m(h)$, the basis functions of $\mathbb{V}_h$, and π_h, the usual restriction operator.

The discrete counterpart of (7.1.1) consists of seeking $u_h \in \mathbb{V}_h$ such that

$$\begin{cases} a(u_h, v - u_h) \geqq (f, v - u_h) \forall v \in \mathbb{V}_h \\ u_h \leq \pi_h M u_h, v \leq \pi_h M u_h \end{cases} \tag{7.1.3}$$

Under $W^{2,p} - regularity$ of the continuous solution, the following error estimate

$$\|u - u_h\|_\infty \leq Ch|\log h|$$

was obtained by Loinger [4] in the one-dimensional case $(N = 1)$, and by Cortey Dumont [5] for $N \geq 1$.

In this paper, we improve on the above results and obtain a sharp error estimate (for $N \geq 1$), i.e.,

$$\|u - u_h\|_\infty \leq Ch^2|\log h|^2 \tag{7.1.4}$$

For this, we develop a new approach, which combines the concept of subsolution in variational inequalities (VI): w is a continuous subsolution if

$$\begin{cases} a(w, v) \leq (f, v) \forall v \in H_0^1(\Omega), v \geq 0 \\ w \leq \psi \end{cases} \tag{7.1.5}$$

(respect. w_h is a discrete subsolution), if

$$\begin{cases} a(w_h, \varphi_i) \leq (f, \varphi_i) \forall \varphi_i, i = 1, \ldots, m(h) \\ w_h \leq \pi_h \psi \end{cases} \tag{7.1.6}$$

and the concept of "discrete regularity": a discrete solution ζ_h of a variational inequality is regular in the discrete sense if it satisfies

$$|a(\zeta_h, \varphi_i)| \leq C\|\varphi_i\|_{L^1(\Omega)}$$

This new concept of "discrete regularity," introduced in [6], can be regarded as the discrete counterpart of the Lewy–Stampacchia regularity estimate $\|Au\|_\infty \leq C$,

extended to the variational form through the $L^\infty - L^1$ duality. It plays a major role in deriving the optimal error estimate as this will be shown in the sequel of the paper.

The finite element error analysis stands on the construction of a continuous sequence of subsolutions $(\beta_n^{(h)})$ such that

$$\beta_n^{(h)} \leq u_n \text{ and } \left\|\beta_n^{(h)} - u_{nh}\right\|_\infty \leq Ch^2|\log h|^2$$

and a discrete sequence of subsolutions (α_{nh}) such that

$$\alpha_{nh} \leq u_{nh} \text{ and } \|u_n - \alpha_{nh}\|_\infty \leq Ch^2|\log h|^2,$$

where (u_n) is the Bensoussan–Lions algorithm and (u_{nh}) is its finite element counterpart.

7.2 Background

7.2.1 Assumptions and Definitions

We begin with introducing some notations and assumptions. We are given sufficiently smooth functions

$$a_{jk}(x), b_k(x), a_0(x) x \in \bar{\Omega} \tag{7.2.1}$$

such that

$$\sum_{1 \leq j,k \leq N} a_{jk}(x)\xi_j\xi_k \geqq \alpha|\xi|^2; (x \in \bar{\Omega}, \xi \in R^N, \alpha > 0) \tag{7.2.2}$$

$$a_0(x) \geqq c_0 \geq 0\,(x \in \bar{\Omega}; c_0 > 0) \tag{7.2.3}$$

We define the second order elliptic operator

$$\mathcal{A} = \sum_{1 \leq j,k \leq N} -\frac{\partial}{\partial x_j}\left(a_{jk}(x)\frac{\partial}{\partial x_k}\right) + \sum_{k=1}^{N} b_k(x)\frac{\partial}{\partial x_k} + a_0(x) \tag{7.2.4}$$

and the associated bilinear form $\forall u, v \in H^1(\Omega)$

$$a(u,v) = \int_\Omega \left(\sum_{1 \leq j,k \leq N} a_{jk}(x)\frac{\partial u}{\partial x_j}\frac{\partial v}{\partial x_k} + \sum_{k=1}^{N} b_k(x)\frac{\partial u}{\partial x_k}v + a_0(x)uv\right)\mathrm{d}x \tag{7.2.5}$$

which we assume to be coercive

$$a(v,v) \geq \delta \|v\|^2_{H^1(\Omega)}, \delta > 0, \forall v \in H^1(\Omega) \tag{7.2.6}$$

We assume that

$$\text{If } \psi \in C(\bar{\Omega}) \text{ then } \partial(\psi) \in C(\bar{\Omega}) \tag{7.2.7}$$

$$\text{If } u \text{ and } v \in C(\bar{\Omega}) \text{ then } \|Mu - Mv\|_{C(\bar{\Omega})} \leq \|u - v\|_{C(\bar{\Omega})} \tag{7.2.8}$$

$$\text{If } u \in W^{1,\infty}(\Omega) \text{ then } \|Mu\|_{W^{1,\infty}(\Omega)} \leq C\|u\|_{W^{1,\infty}(\Omega)} \tag{7.2.9}$$

Throughout the paper, we will introduce several variational inequalities of obstacle type. The following are some useful, related definitions and properties.

7.2.2 *Continuous Variational Inequality*

Let g in $L^\infty(\Omega)$ and ψ in $W^{1,\infty}(\Omega)$ such that $\psi \geq 0$ on $\partial\Omega$. The following problem is called variational inequality:

$$\begin{cases} a(\omega, v-\omega) \geqq (g, v-\omega) \forall v \in H_0^1(\Omega) \\ v \leq \psi, w \leq \psi \end{cases} \tag{7.2.10}$$

Thanks to [1], problem (7.2.10) has a unique solution.

Definition 1 w is said to be a subsolution for the VI (7.2.10) if

$$\begin{cases} a(w,v) \leq (g,v) \forall v \in H_0^1(\Omega), v \geq 0 \\ w \leq \psi \end{cases} \tag{7.2.11}$$

Theorem 1 [1] *The solution ω of the VI (7.2.10) is the least upper bound of the set of subsolutions.*

Theorem 2 [1] *Let ψ and $\tilde{\psi}$ in $W^{1,\infty}(\Omega)$, and ω and $\tilde{\omega}$ be the corresponding solutions to (7.2.10). Then,*

$$\|\omega - \tilde{\omega}\|_\infty \leq C\left\|\psi - \tilde{\psi}\right\|_\infty$$

Lemma 1 (The continuous Levy–Stampacchia Inequality) *Let ψ in $H^1(\Omega)$ such that $\psi \geq 0$ on $\partial\Omega$. Let also ω be the solution of (7.2.10) such that $\mathcal{A}\omega \geq h$ (in the sense of $H^{-1}(\Omega)$), where $h \in L^2(\Omega)$. Then,*

$$g \geq \mathcal{A}\psi \geq g \wedge h$$

Theorem 3 [1] *Under conditions of lemma 1, the solution* ω *of* (7.2.10) *is in* $W^{2,p}(\Omega)$ *for all* $p \geq 2, p < \infty, \mathcal{A}\omega \in L^{\infty}(\Omega)$.

7.2.3 Discrete Variational Inequality

The following is the corresponding discrete variational inequality

$$\begin{cases} a(\omega_h, v - \omega_h) \geqq (g, v - \omega_h) \forall v \in \mathbb{V}_h \\ v \leq \pi_h \psi, w \leq \pi_h \psi \end{cases} \tag{7.2.12}$$

w_h is said to be a discrete subsolution if

$$\begin{cases} a(w_h, \varphi_i) \leq (g, \varphi_i) \forall \varphi_i, i = 1, \ldots, m(h) \\ w_h \leq \pi_h \psi \end{cases} \tag{7.2.13}$$

Under the **discrete maximum assumption** (**d.m.p**), the stiffness matrix $a(\varphi_i, \varphi_j)$ is an M-Matrix (this will be thoroughly explained in Sect. 7.3), we have

Theorem 4 *The solution* ω_h *of the VI* (7.2.13) *is the least upper bound of the set of discrete subsolutions.*

Theorem 5 *Let* ψ *and* $\tilde{\psi}$ *in* $W^{1,\infty}(\Omega)$, *and* ω_h *and* $\tilde{\omega}_h$ *the corresponding solutions to* (7.2.12). *Then,*

$$\|\omega_h - \tilde{\omega}_h\|_{\infty} \leq C \left\|\psi - \tilde{\psi}\right\|_{\infty}$$

Theorem 6 (The discrete Levy–Stampacchia inequality) *Let* ω_h *be the solution of the discrete VI* (7.2.12). *Then,*

$$(g, \varphi_i) \geq a(\omega_h, \varphi_i) \geq a(\psi, \varphi_i) \wedge (g, \varphi_i) \forall \varphi_i, i = 1, \ldots, m(h)$$

7.3 The Impulse Control QVI

7.3.1 The Continuous QVI

The existence of a unique solution for QVI (7.1.1) can be achieved, making use of the method of upper and lower solutions (see [7]).

Indeed, one can define the fixed point mapping

$$T : L^{\infty}(\Omega) \to L^{\infty}(\Omega)$$
$$z \to Tz = \zeta,$$

where ζ solves the following VI

$$\begin{cases} a(\zeta, v - \zeta) \geqq (f, v - \zeta) \forall v \in H_0^1(\Omega) \\ \zeta \leq Mz, v \leq Mz \end{cases} \tag{7.3.1}$$

Let u_0 be the solution of the equation

$$a(u_0, v) = (f, v) \forall v \in H_0^1(\Omega) \tag{7.3.2}$$

Thanks to [7], (7.3.2) has a unique solution which belongs to $W^{2,p}(\Omega)$. The Bensoussan–Lions algorithm is constructed as follows: starting from u_0, solution of the above equation, we define the sequence

$$u_n = Tu_{n-1}, n = 1, 2, \ldots \tag{7.3.3}$$

Theorem 7 [7] *Assume that $f \geq f_0 > 0$. Then, the sequence $\{u_n\}_{n \geq 0}$ converges decreasingly to the unique solution of the QVI (7.1.3). Moreover, there exists $0 < \mu < 1$ such that*

$$\|u_n - u\|_{\infty} \leq \mu^n \|u_0\|_{\infty} \tag{7.3.4}$$

Remark 1 One can also start from $\breve{u}_0 = 0$ and generate an increasing sequence $\breve{u}_n = T\breve{u}_{n-1}$, $n = 1, 2, \ldots$ which converges geometrically to the solution of the QVI (7.1.1).

7.3.2 The Discrete QVI

For the sake of finite element discretization, we will assume that Ω is polyhedral. Let then τ_h be a regular and quasiuniform triangulation of Ω into triangles; $h > 0$ be the mesh size. For each $K \in \tau_h$, denote by $P_1(K)$ the set of polynomials on K with degree not more than 1. The P_1, conforming finite element space, is given as

$$V_h = \left\{ v : v \in H_0^1(\Omega) \cap C(\bar{\Omega}), v_{/K} \in P_1(K), \forall K \in \tau_h \right\}$$

Let $M_i, 1 \leq i \leq m(h)$ denote the vertices of the triangulation τ_h, and let $\varphi_i, 1 \leq i \leq m(h)$, denote the functions of V_h which satisfies

$$\varphi_i(M_j) = \delta_{ij}, 1 \leq i, j \leq m(h)$$

So that the function φ_i form a basis of V_h. $\forall v \in H^1(\Omega) \cap C(\bar{\Omega})$, the function

$$r_h v(x) = \sum_{i=1}^{m(h)} v(M_i)\varphi_i(x)$$

represents the interpolate of v over τ_h.

The existence of a solution for QVI (7.1.3) can be obtained similarly to that of the continuous case. Indeed, we construct a discrete fixed point mapping

$$\begin{aligned} T_h &: L^\infty(\Omega) \to \mathbb{V}_h \\ z &\to T_h z = \zeta_h, \end{aligned}$$

where ζ_h solves the following discrete VI

$$\begin{cases} a(\zeta_h, v - \zeta_h) \geqq (f, v - \zeta_h) \forall v \in \mathbb{V}_h \\ \zeta_h \leq \pi_h M z, v \leq \pi_h M z \end{cases} \tag{7.3.5}$$

Now, starting from u_{0h}, solution of the equation

$$a(u_{0h}, v) = (f, v) \forall v \in \mathbb{V}_h \tag{7.3.6}$$

we construct the discrete version of the Bensoussan–Lions algorithm

$$u_{nh} = T u_{n-1h}, n = 1, 2, \ldots \tag{7.3.7}$$

The convergence analysis of the discrete algorithm will require that the stiffness matrix is an M-Matrix.

Definition 2 A real matrix $d \times d$ $C = (c_{ij})$ with $c_{ij} \leq 0$, $\forall i \neq j$, $1 \leq i, j \leq d$, is called an M-Matrix, if C is nonsingular and $C^{-1} \geq 0$ (i.e., all entries of its inverse are nonnegative).

Denoted by A is the matrix with generic coefficient

$$A_{ij} = a(\varphi_i, \varphi_j), 1 \leq i, j \leq m(h) \tag{7.3.8}$$

Because the bilinear form $a(.,.)$ is coercive, we have

$$A \text{ is positive definite} \tag{7.3.9}$$

and

$$A_{ii} > 0 \, \forall i = 1, \ldots, m(h) \tag{7.3.10}$$

Furthermore, if the matrix (a_{jk}) involved in the bilinear form (7.2.5) is symmetric $(a_{jk} = a_{kj})$, then mesh conditions for which the off-diagonal entries of B satisfy

$$A_{ij} \leq 0, \forall i \neq j, 1 \leq i, j \leq m(h) \tag{7.3.11}$$

can be found in [8]. By combining (7.3.9), (7.3.10), and (7.3.11), we have the following lemma.

Lemma 2 *The matrix A is an M-Matrix.*

Proof See [8]. □

Theorem 8 *Let conditions of Lemma 2 hold. Then, the sequence* $\{u_{nh}\}_{n \geq 0}$ *converges to the unique solution of the* (7.1.3). *Moreover, there exists* $0 < \mu < 1$ *such that*

$$\|u_{nh} - u_h\|_\infty \leq \mu^n \|u_{0h}\|_\infty \tag{7.3.12}$$

7.4 The Finite Element Error Analysis

The establishment of the optimal error estimate (7.1.4), in which the concept of "discrete regularity" will play a crucial role, rests on several lemmas and theorems.

7.4.1 The Discrete Regularity

Consider the VI

$$\begin{cases} a(\omega_h, v - \omega_h) \geqq (g, v - \omega_h) \forall v \in \mathbb{V}_h \\ v \leq \pi_h \psi \, , \omega_h \leq \pi_h \psi \end{cases}$$

Assumption We assume that there exists a constant C independent of h such that

$$|a(\omega_h, \varphi_i)| \leq C \|\varphi_i\|_{L^1(\Omega)} \, \forall i = 1, 2, \ldots, m(h) \tag{7.4.1}$$

Lemma 3 [6] *Under assumption* (7.4.1)*, there exists a family of right-hands side* $\left\{g^{(h)}\right\}_{h>0} \in L^{\infty}(\Omega)$ *such that*

$$\left\|g^{(h)}\right\|_{\infty} \leq C, \forall h$$

and

$$a(\omega_h, v) = (g^{(h)}, v) \forall v \in \mathbb{V}_h \tag{7.4.2}$$

Theorem 9 *Let conditions of Lemma 3 hold. Then, there exist two continuous sequences* $\left(g_n^{(h)}\right)_{n\geq 1}$ *and* $\left(\omega_n^{(h)}\right)_{n\geq 1}$*, and a constant independent of h and n such that*

$$\left\|g_n^{(h)}\right\|_{\infty} \leq C$$

and

$$a(\omega_n^{(h)}, v) = (g_n^{(h)}, v) \forall v \in H_0^1(\Omega)$$

Proof The proof will be carried out by induction. For $n = 1$, let ω_{1h} be the solution of the VI

$$\begin{cases} a(\omega_{1h}, v - \omega_{1h}) \geqq (f, v - \omega_{1h}) \forall v \in \mathbb{V}_h \\ v \leq \pi_h M \omega_0^{(h)},\ \omega_{1h} \leq \pi_h M \omega_0^{(h)} \end{cases},$$

where $\omega_0^{(h)} = u_0$ is the solution of

$$a(u_0, v) = (f, v) \forall v \in H_0^1(\Omega)$$

So, by the discrete Levy–Stampachia inequality, we have

$$-(f, \varphi_i) \wedge a(Mu_0, \varphi_i) \leq a(\omega_{1h}, \varphi_i) \leq (f, \varphi_i)$$

or

$$-(f, \varphi_i) \wedge \langle \mathcal{A}(Mu_0), \varphi_i \rangle \leq a(\omega_{1h}, \varphi_i) \leq (f, \varphi_i)$$

and using ([7], pp. 366–376), there exists a constant c such that $\mathcal{A}(Mu_0) \geq -c$. Hence,

$$-(f, \varphi_i) \wedge (-c, \varphi_i) \leq a(\omega_{1h}, \varphi_i) \leq (f, \varphi_i),$$

which implies

$$|a(\omega_{1h}, \varphi_i)| \leq C\|\varphi_i\|_{L^1(\Omega)}$$

So, making use of Lemma 3, there exists a family of right-hands side $\left\{g_1^{(h)}\right\} \in L^\infty(\Omega)$ such that

$$\begin{cases} (i)\ \left\|g_1^{(h)}\right\|_\infty \leq C \\ (ii)\ a(\omega_{1h}, v) = (g_1^{(h)}, v) \forall v \in \mathbb{V}_h \end{cases},$$

which enables us to define $\omega_1^{(h)}$ as the solution of the equation

$$a(\omega_1^{(h)}, v) = (g_1^{(h)}, v) \forall v \in H_0^1(\Omega)$$

and

$$\left\|\omega_1^{(h)}\right\|_{W^{2,p}(\Omega)} \leq C$$

Now, let us consider ω_{2h} to be the solution of the VI

$$\begin{cases} a(\omega_{2h}, v - \omega_{2h}) \geqq (f, v - \omega_{2h})\ \forall v \in \mathbb{V}_h \\ v \leq \pi_h M\omega_1^{(h)},\ \omega_{1h} \leq \pi_h M\omega_1^{(h)} \end{cases} \tag{7.4.3}$$

So, using the discrete Levy–Stampachia inequality, we have

$$-(f, \varphi_i) \wedge a(M\omega_1^{(h)}, \varphi_i) \leq a(\omega_{2h}, \varphi_i) \leq (f, \varphi_i)$$

or

$$-(f, \varphi_i) \wedge \left\langle \mathcal{A}(M\omega_1^{(h)}), \varphi_i \right\rangle \leq a(\omega_{2h}, \varphi_i) \leq (f, \varphi_i)$$

and using ([7], pp. 366–376), as above, there exists a constant c such that $\mathcal{A}(M\omega_1^{(h)}) \geq -c$, and therefore

$$-(f, \varphi_i) \wedge (-c, \varphi_i) \leq a(\omega_{2h}, \varphi_i) \leq (f, \varphi_i),$$

which implies

$$|a(\omega_{2h}, \varphi_i)| \leq C\|\varphi_i\|_{L^1(\Omega)}$$

So, making use of Lemma 3, there exists a family of right-hands side $\left\{g_2^{(h)}\right\} \in L^\infty(\Omega)$ such that

$$\begin{cases} (i)\left\|g_2^{(h)}\right\|_\infty \leq C \\ (ii)\, a(\omega_{2h}, v) = (g_2^{(h)}, v) \forall v \in \mathbb{V}_h \end{cases}$$

and we can therefore define $\omega_2^{(h)}$ such that

$$a(\omega_2^{(h)}, v) = (g_2^{(h)}, v) \forall v \in H_0^1(\Omega)$$

and

$$\left\|\omega_2^{(h)}\right\|_{W^{2,p}(\Omega)} \leq C$$

Hence, by induction, there exists $\left\{g_n^{(h)}\right\} \in L^\infty(\Omega)$ such that the solution ω_{nh} of the VI

$$\begin{cases} a(\omega_{nh}, v - \omega_{nh}) \geqq (f, v - \omega_{nh}) \forall v \in \mathbb{V}_h \\ v \leq \pi_h M \omega_{n-1}^{(h)},\ \omega_{1h} \leq \pi_h M \omega_{n-1}^{(h)} \end{cases} \tag{7.4.4}$$

satisfies

$$a(\omega_{nh}, v) = (g_n^{(h)}, v) \forall v \in \mathbb{V}_h$$

and therefore $\omega_n^{(h)}$ such that

$$a(\omega_n^{(h)}, v) = (g_n^{(h)}, v) \forall v \in H_0^1(\Omega)$$

and

$$\left\|\omega_n^{(h)}\right\|_{W^{2,p}(\Omega)} \leq C \qquad \square$$

In the light of the above, let

$$\omega_{nh} = \partial_h(M\omega_{n-1}^{(h)}), n = 1, 2, \ldots \tag{7.4.5}$$

Lemma 4 *We have*

$$\left\|\omega_n^{(h)} - \omega_{nh}\right\|_\infty \leq Ch^2|\log h|^2 \tag{7.4.6}$$

Proof We know that

$$a(\omega_n^{(h)}, v) = (g_n^{(h)}, v) \forall v \in H_0^1(\Omega)$$

and

$$a(\omega_{nh}, v) = (g_n^{(h)}, v) \forall v \in \mathbb{V}_h$$

So, since

$$\left\|\omega_n^{(h)}\right\|_{W^{2,p}(\Omega)} \leq C,$$

making use of standard maximum norm error estimate [9], we get (7.4.6) □

In the light of the above, one can define the following sequences of variational inequalities

$$\omega_{nh} = \partial_h(M\omega_{n-1}^{(h)}), n = 1, 2, \ldots \tag{7.4.7}$$

Lemma 5 *We have*

$$\left\|\omega_n^{(h)} - \omega_{nh}\right\|_\infty \leq Ch^2|\log h| \tag{7.4.8}$$

Proof We know that

$$a(\omega_n^{(h)}, v) = (g_n^{(h)}, v) \forall v \in H_0^1(\Omega)$$

and

$$a(\omega_{nh}, v) = (g_n^{(h)}, v) \forall v \in \mathbb{V}_h$$

So, since

$$\left\|\omega_n^{(h)}\right\|_{W^{2,p}(\Omega)} \leq C$$

making use of standard maximum norm error estimate [9], we get (7.4.8) □

Remark 2 Estimate (7.4.6) holds for at least the operator $-\Delta + cI$, with c as a positive constant.

Lemma 6 *Let conditions of Lemma 5 hold. Then, we have*

$$\|\omega_{nh} - u_{nh}\|_\infty \leq Ch^2|\log h|^2 \tag{7.4.9}$$

Proof We proceed by induction $n = 1$:

$$\|\omega_{1h} - u_{1h}\|_\infty \leq \left\|Mu_0^{(h)} - Mu_{0h}\right\|_\infty \leq \left\|u_0^{(h)} - u_{0h}\right\|_\infty \leq Ch^2|\log h|^2$$

Now, assume that

$$\|\omega_{n-1h} - u_{n-1h}\|_\infty \leq Ch^2|\log h|^2$$

Then, applying Lemma 5 we get

$$\begin{aligned}\|\omega_{nh} - u_{nh}\|_\infty &\leq \left\|M\omega_{n-1}^{(h)} - Mu_{n-1h}\right\|_\infty \leq \left\|\omega_{n-1}^{(h)} - u_{n-1h}\right\|_\infty \\ &\leq \left\|\omega_{n-1}^{(h)} - \omega_{n-1h}\right\|_\infty + \|\omega_{n-1} - u_{n-1h}\|_\infty \\ &\leq Ch^2|\log h|^2 + Ch^2|\log h|^2 \\ &\leq Ch^2|\log h|^2\end{aligned}$$

□

Theorem 10 *There exists a constant independent of h and n such that*

$$\|u_n - u_{nh}\|_\infty \leq Ch^2|\log h|^2 \tag{7.4.10}$$

The proof rests on the construction of a sequence of continuous subsolutions and a sequence of discrete subsolutions.

7.4.2 Construction of Subsolutions

Consider the sequence of continuous VIs $\bar{u}_n = \partial(M\omega_{n-1}^{(h)})$

$$\begin{cases} a(\bar{u}_n, v - \bar{u}_n) \geqq (g, v - \bar{u}_n)\, \forall v \in H_0^1(\Omega) \\ v \leq M\omega_{n-1}^{(h)},\ \omega_{1h} \leq M\omega_{n-1}^{(h)}, n \geq 1 \end{cases} \tag{7.4.11}$$

and the sequence of discrete VIs $\bar{u}_{nh} = \partial_h(Mu_{n-1})$

$$\begin{cases} a(\bar{u}_{nh}, v - \bar{u}_{nh}) \geqq (g, v - \bar{u}_{nh})\, \forall\, v \in \mathbb{V}_h \\ v \leq \pi_h Mu_{n-1},\ \omega_{1h} \leq \pi_h Mu_{n-1}, n \geq 1 \end{cases} \tag{7.4.12}$$

Lemma 7 *We have*

$$\|\bar{u}_n - \omega_{nh}\|_\infty \leq Ch^2|\log h|^2 \tag{7.4.13}$$

and

$$\|u_n - \bar{u}_{nh}\|_\infty \leq Ch^2|\log h|^2 \tag{7.4.14}$$

Proof Since $\bar{u}_{nh} = \partial_h(M\omega_{n-1}^{(h)})$ and $\omega_{nh} = \partial_h(M\omega_{n-1}^{(h)})$ are the discrete counterparts of $\bar{u}_n$ and u_n, respectively, applying standard L^∞error estimate for elliptic VIs (see [10])—we get (7.4.13) and (7.4.14). □

Theorem 11 *There exist a sequence* $(\beta^n)_{n\geq 1}$ *of continuous subsolutions such that*

$$\begin{cases} (i)\ \beta^n \leq u_n \\ (ii)\ \|\beta^n - u_{nh}\|_\infty \leq Ch^2|\log h|^2 \end{cases}$$

and a discrete sequence of subsolutions $(\alpha_{nh})_{n\geq 1}$ *such that*

$$\begin{cases} (i)\ \alpha_{nh} \leq u_{nh} \\ (ii)\ \|u_n - \alpha_{nh}\|_\infty \leq Ch^2|\log h|^2 \end{cases}$$

Proof We proceed by induction.

Construction of β^1. Consider the VI

$$\begin{cases} a(\bar{u}_1, v - \bar{u}_1) \geqq (f, v - \bar{u}_1) \forall v \in H_0^1(\Omega) \\ v \leq M\omega_0^{(h)}, \omega_{1h} \leq M\omega_0^{(h)} \end{cases}$$

So,

$$\begin{cases} a(\bar{u}_1, v) \leq (f, v) \forall v \in H_0^1(\Omega), v \geq 0 \\ v \leq M\omega_0^{(h)},\ \omega_{1h} \leq M\omega_0^{(h)} \end{cases}$$

But,

$$\begin{aligned} \bar{u}_1 &\leq M\omega_0^{(h)} \\ &\leq M\omega_0^{(h)} - M\omega_{0h} + M\omega_{0h} \\ &\leq \left\|M\omega_0^{(h)} - M\omega_{0h}\right\|_\infty + M\omega_{0h} \\ &\leq \left\|\omega_0^{(h)} - \omega_{0h}\right\|_\infty + M\omega_{0h} \\ &\leq Ch^2|\log h| + M\omega_{0h} \end{aligned}$$

then, $\bar{u}_1$ is a subsolution for the VI with obstacle

$$\psi = Ch^2|\log h| + M\omega_{0h}$$

Let $\bar{U}_1 = \partial(Ch^2|\log h| + M\omega_{0h})$ be the solution of such VI. Then, as $u_1 = \partial(Mu_0)$, applying Theorem 2, we get

$$\begin{aligned}\|u_1 - \bar{U}_1\|_\infty &\leq Ch^2|\log h| + \|Mu_0 - M\omega_{0h}\|_\infty \\ &\leq Ch^2|\log h| + \|u_0 - \omega_{0h}\|_\infty \\ &\leq Ch^2|\log h| + Ch^2|\log h| \\ &\leq Ch^2|\log h|\end{aligned}$$

Hence, Theorem 1 implies that

$$\bar{u}_1 \leq \bar{U}_1 \leq u_1 + Ch^2|\log h|$$

So, putting

$$\beta^1 = \bar{u}_1 - Ch^2|\log h|^2$$

we get

$$\beta^1 \leq u_1$$

and, using (7.4.9) and (7.4.13), we obtain

$$\begin{aligned}\left\|\beta^1 - u_{1h}\right\|_\infty &\leq \left\|\bar{u}_1 - Ch^2|\log h| - u_{1h}\right\|_\infty \\ &\leq \|\bar{u}_1 - u_{1h}\|_\infty + Ch^2|\log h| \\ &\leq \|\bar{u}_1 - \omega_{1h}\|_\infty + \|\omega_{1h} - u_{1h}\|_\infty \\ &\leq Ch^2|\log h|^2 + Ch^2|\log h| \\ &\leq Ch^2|\log h|^2\end{aligned}$$

Construction of α_{1h}. Consider the VI

$$\begin{cases} a(\bar{u}_{1h}, v - \bar{u}_{1h}) \geqq (f, v - \bar{u}_{1h}) \forall v \in \mathbb{V}_h \\ v \leq \pi_h Mu_0, \omega_{1h} \leq \pi_h Mu_0 \end{cases}$$

So,

$$\begin{cases} a(\bar{u}_{1h}, \varphi_i) \geqq (f, \varphi_i) \forall v \in \varphi_i \\ v \leq \pi_h Mu_0, \omega_{1h} \leq \pi_h Mu_0 \end{cases}$$

But,

$$\begin{aligned}\bar{u}_{1h} &\leq \pi_h M u_0 - \pi_h M u_{0h} + \pi_h M u_{0h} \\ &\leq \|\pi_h M u_0 - \pi_h M u_{0h}\|_\infty + \pi_h M u_{0h} \\ &\leq \|u_0 - u_{0h}\|_\infty + \pi_h M u_{0h} \\ &\leq Ch^2|\log h| + \pi_h M u_{0h}\end{aligned}$$

Hence, $\bar{u}_{1h}$ is a subsolution for the VI with obstacle

$$\psi = Ch^2|\log h| + \pi_h M u_{0h}$$

Let $\bar{U}_{1h} = \partial_h(Ch^2|\log h| + \pi_h M u_{0h})$ be the solution of such VI. Then, as $u_{1h} = \partial_h(\pi_h M u_{0h})$, making use of Theorem 5, we have

$$\begin{aligned}\|\bar{U}_{1h} - u_{1h}\|_\infty &\leq \left\|Ch^2|\log h| + \pi_h M u_{0h} - \pi_h M u_{0h}\right\|_\infty \\ &\leq Ch^2|\log h|\end{aligned}$$

So, using Theorem 4, we have

$$\bar{u}_{1h} \leq \bar{U}_{1h} \leq u_{1h} + Ch^2|\log h|$$

Now, putting

$$\alpha_{1h} = \bar{u}_{1h} - Ch^2|\log h|$$

we have

$$\alpha_{1h} \leq u_{1h}$$

and, using (7.4.14), we get

$$\begin{aligned}\|\alpha_{1h} - u_1\|_\infty &\leq \|\alpha_{1h} - u_1\|_\infty \\ &\leq \left\|\bar{u}_{1h} - Ch^2|\log h| - u_1\right\|_\infty \\ &\leq \|\bar{u}_{1h} - u_1\|_\infty + Ch^2|\log h| \\ &\leq Ch^2|\log h|^2 + Ch^2|\log h| \\ &\leq Ch^2|\log h|^2\end{aligned}$$

Now, combining the above, we get

$$\begin{aligned}u_1 &\leq \alpha_{1h} + Ch^2|\log h|^2 \\ &\leq u_{1h} + Ch^2|\log h|^2 \\ &\leq \beta^1 + Ch^2|\log h|^2 \\ &\leq u_1 + Ch^2|\log h|^2\end{aligned}$$

Thus,

$$\|u_1 - u_{1h}\|_\infty \leq Ch^2|\log h|^2$$

Now, assume that

$$\|u_{n-1} - u_{n-1h}\|_\infty \leq Ch^2|\log h|^2 \tag{7.4.15}$$

Construction of β^n. Consider the VI

$$\begin{cases} a(\bar{u}_n, v - \bar{u}_n) \geqq (f, v - \bar{u}_n) \forall v \in H_0^1(\Omega) \\ v \leq M\omega_{n-1}^{(h)}, \omega_{1h} \leq M\omega_{n-1}^{(h)} \end{cases}$$

Then,

$$\begin{cases} a(\bar{u}_n, v) \leq (f, v) \forall v \in H_0^1(\Omega), v \geq 0 \\ v \leq M\omega_{n-1}^{(h)}, \omega_{1h} \leq M\omega_{n-1}^{(h)} \end{cases}$$

So, using (7.4.6), we get

$$\begin{aligned} \bar{u}_n &\leq M\omega_{n-1}^{(h)} \\ &\leq M\omega_{n-1}^{(h)} - M\omega_{n-1h} + M\omega_{n-1h} \\ &\leq \left\|M\omega_{n-1}^{(h)} - M\omega_{n-1h}\right\|_\infty + M\omega_{n-1h} \\ &\leq \left\|\omega_{n-1}^{(h)} - \omega_{n-1h}\right\|_\infty + M\omega_{n-1h} \\ &\leq Ch^2|\log h| + M\omega_{n-1h} \end{aligned}$$

Hence, $\bar{u}_n$ is a subsolution for the VI with obstacle

$$\psi = Ch^2|\log h|^2 + M\omega_{n-1}$$

Let $\bar{U}_n = \partial(Ch^2|\log h|^2 + M\omega_{n-1})$ be the solution of such a VI. Then, as $u_n = \partial(Mu_{n-1})$, making use of Theorem 2, we have

$$\begin{aligned} \|\bar{U}_n - u_n\|_\infty &\leq Ch^2|\log h|^2 + \|M\omega_{n-1h} - Mu_{n-1}\|_\infty \\ &\leq Ch^2|\log h|^2 + \|\omega_{n-1h} - u_{n-1}\|_\infty \\ &\leq Ch^2|\log h|^2 + \|\omega_{n-1h} - u_{n-1h}\|_\infty + \|u_{n-1h} - u_{n-1}\|_\infty \\ &\leq Ch^2|\log h|^2 + Ch^2|\log h| + Ch^2|\log h| \\ &\leq Ch^2|\log h|^2 \end{aligned}$$

So, using Theorem 1 and (7.4.15) we have

$$\bar{u}_n \leq \bar{U}_n \leq u_n + Ch^2|\log h|^2$$

and, putting

$$\beta^n = \bar{u}_n - Ch^2|\log h|^2$$

we get

$$\beta^n \leq u_n$$

Finally, using (7.4.9) and (7.4.13), we obtain

$$\begin{aligned}
\|\beta^n - u_{nh}\|_\infty &\leq \left\|\bar{u}_n - Ch^2|\log h|^2 - u_{nh}\right\|_\infty \\
&\leq \|\bar{u}_n - u_{nh}\|_\infty + Ch^2|\log h|^2 \\
&\leq \|\bar{u}_n - \omega_{nh}\|_\infty + \|\omega_{nh} - u_{nh}\|_\infty + Ch^2|\log h|^2 \\
&\leq Ch^2|\log h|^2 + Ch^2|\log h|^2 + Ch^2|\log h|^2 \\
&\leq Ch^2|\log h|^2
\end{aligned}$$

Construction of α_{nh}. Consider the VI

$$\begin{cases} a(\bar{u}_{nh}, v - \bar{u}_{nh}) \geqq (f, v - \bar{u}_{nh}) \forall v \in \mathbb{V}_h \\ v \leq \pi_h M u_{n-1}, \omega_{1h} \leq \pi_h M u_{n-1} \end{cases}$$

So,

$$\begin{cases} a(\bar{u}_{nh}, \varphi_i) \leq (f, \varphi_i) \forall v \in \varphi_i \\ v \leq \pi_h M u_{n-1},\ \omega_{1h} \leq \pi_h M u_{n-1} \end{cases}$$

But,

$$\begin{aligned}
\bar{u}_{nh} &\leq \pi_h M u_{n-1} - \pi_h M u_{n-1h} + \pi_h M u_{n-1h} \\
&\leq \|\pi_h M u_{n-1} - \pi_h M u_{n-1h}\|_\infty + \pi_h M u_{n-1h} \\
&\leq \|u_{n-1} - u_{n-1h}\|_\infty + \pi_h M u_{n-1h} \\
&\leq Ch^2|\log h|^2 + \pi_h M u_{n-1h}
\end{aligned}$$

Hence, $\bar{u}_{nh}$ is a subsolution for the VI with obstacle

$$\psi = Ch^2|\log h|^2 + \pi_h M u_{n-1h}$$

Let $\bar{U}_{nh}$ be the solution of such VI. Then as $u_{nh} = \partial_h(\pi_h M u_{n-1h})$, making use of Theorem 5, we have

$$\|\bar{U}_{nh} - u_{nh}\|_\infty \leq Ch^2|\log h|^2$$

So, using Theorem 4, we have

$$\bar{u}_{nh} \leq \bar{U}_{nh} \leq u_{nh} + Ch^2|\log h|^2$$

Now, putting

$$\alpha_{nh} = \bar{u}_{nh} - Ch^2|\log h|^2$$

we have

$$\alpha_{nh} \leq u_{nh}$$

and, using (7.4.14), we get

$$\begin{aligned}\|\alpha_{nh} - u_n\|_\infty &\leq \left\|\bar{u}_{nh} - Ch^2|\log h|^2 - u_n\right\|_\infty \\ &\leq \|\bar{u}_{nh} - u_n\|_\infty + Ch^2|\log h|^2 \\ &\leq Ch^2|\log h| + Ch^2|\log h|^2 \\ &\leq Ch^2|\log h|^2\end{aligned}$$

Now, combining the above, we obtain

$$\begin{aligned}u_n &\leq \alpha_{nh} + Ch^2|\log h|^2 \\ &\leq u_{nh} + Ch^2|\log h|^2 \\ &\leq \beta^n + Ch^2|\log h|^2 \\ &\leq u_n + Ch^2|\log h|^2\end{aligned}$$

Thus,

$$\|u_n - u_{nh}\|_\infty \leq Ch^2|\log h|^2$$ □

Now, making use of estimates (7.3.4), (7.3.12), and (7.4.10), we are in the position to derive the main result of this paper.

Theorem 12

$$\|u - u_h\|_\infty \leq Ch^2 |\log h|^2 \tag{7.4.16}$$

Proof Making use of (7.3.4), (7.3.12), and (7.4.10), we have

$$\begin{aligned} \|u - u_h\|_\infty &\leq \|u - u_n\|_\infty + \|u_n - u_{nh}\|_\infty + \|u_{nh} - u_h\|_\infty \\ &\leq \mu^n \|u_0\|_\infty + Ch^2 |\log h|^2 + \mu^n \|u_{0h}\|_\infty \end{aligned}$$

So, passing to the limit on both sides $(n \to \infty)$, we get

$$\|u - u_h\|_\infty \leq Ch^2 |\log h|^2$$

□

References

1. Bensoussan A, Lions JL (1982) Application of variational inequalities in stochastic control. Dunod, Paris
2. Baccarin S, Sanfelici S (2006) Optimal impulse control on an unbounded domain with nonlinear cost functions. CMS 3(1):81–100
3. Baccarin S (2009) Optimal impulse control for a multidimensional cash management system with generalized functions. Eur J Oper Res 196:198–206
4. Loinger E (1980) A finite element approach to quasi-variational inequality. Calcolo 17:197–209
5. Cortey Dumont P (1980) Approximation numerique d' une IQV liee a des problemes de gestion de stock. RAIRO, Num. Anal. 4, 335–346
6. Cortey Dumont P (1983) Contribution a l' approximation des inequations variationnelles en norme L^∞. CR Acad Sci Paris Ser I Math 296(17):753–756
7. Bensoussan A, Lions JL (1984) Impulse control and quasi-variational inequalities. Gauthier Villars, Paris
8. Lu C, Huang W, Qiu J (2014) Maximum principle in linear finite element approximations of anisotropic diffusion–convection–reaction problems. Numer Math 127(3):515–537
9. Nitsche J (1977) L^∞-convergence of finite element approximations. Mathematical Aspects of finite element methods. Lect Notes Math 606:261–274
10. Cortey Dumont P (1985) On the finite element approximation in the L^∞-norm of variational inequalities with nonlinear operators. Numer Math 47:45–57

Chapter 8
The Periodic Petrol Station Replenishment Problem: An Overview

Chefi Triki and Nasr Al-Hinai

Abstract This paper focuses on the periodic aspects within the *Petrol Station Replenishment Problem* when defined on an extended planning horizon of t working days. It has the aim of surveying the scientific literature on this topic and giving an overview of the modeling issues, mathematical formulations, and solution approaches related to the *Periodic Petrol Station Replenishment Problem* (*PPSRP*).

Keywords Petrol delivery · Integer optimization models · Periodicity constraints · Vehicle routing problems

8.1 Introduction

Companies operating in the field of delivery of petrol to the stations often face challenging problems. Usually, these problems cannot efficiently be solved by simply using common sense or just relying on the experience of the logistics operator. For the company to remain competitive in the market, more sophisticated mathematical tools and software packages are necessary. One of the main problems faced by these companies is related to the planning of the petrol distribution to the interested stations. This theme, known in the scientific literature as the Petrol Station Replenishment Problem (PSRP), has attracted the interest of several researchers who proposed different methods for its solution. Interested readers are referred for example to [1, 3–6, 9–12, 18, 20, 21]. Many decision aspects of the PSRP have been modeled in these works such as sizing the transportation fleet, defining the routing of each tank-truck, assigning the stations to be served to the appropriate tank-trucks, etc.

C. Triki · N. Al-Hinai (✉)
Department of Mechanical and Industrial Engineering, Sultan Qaboos University, Muscat, Oman
e-mail: nhinai@squ.edu.om

A.H. Siddiqi et al. (eds.), *Mathematical Models, Methods and Applications*, Industrial and Applied Mathematics, DOI 10.1007/978-981-287-973-8_8

Nevertheless, a major drawback in all of the above research works is the fact that they were only considering one single day as planning horizon. However, recognizing that the problem is multi-period by nature and including this aspect into the solution approach may achieve a superior efficiency in minimizing the delivery costs.

Only few scholars have considered, indeed, the multi-period nature of the PSRP. These researchers have taken into account that most of the stations should not be served at each day of the t-day planning horizon, but rather at a specified number of times, which depends on their storage capacity as well as petrol demand.

Specifically, Taqa Allah et al. have considered a single depot and an unlimited homogeneous tank-truck fleet [22]. Accordingly, they have proposed some construction and improvement heuristics to solve the multi-period variant of the PSRP. Motivated by a real-life application, Malepart et al. have solved a vendor inventory management variant of the multi-period PSRP [17]. Their model incorporated the option of giving the distribution company the possibility of choosing the quantity to be delivered to some stations at each replenishment. The third work is due to Cornillier et al. who proposed a heuristic that contains a route construction and truck loading procedures as well as a route packing procedure [9]. They limited the number of stations to be visited at each tank-truck route to only two stations but they suggested two procedures enabling the anticipation or the postponement of deliveries. Finally, recently Triki has considered, while solving the PSRP, not just the multi-period aspect but specifically the periodic nature of the problem [23]. Periodicity there means that each station i must be served f_i times within the time horizon by choosing the replenishment days among the feasible schedules for station i with the objective of minimizing the total distance traveled by the tank-trucks. Triki has defined the Periodic PSRP (PPSRP) and has proposed several integer programming-based heuristics for assigning first the service schedules, then the routing for each tank-truck and each day and finally an improvement technique to further reduce the delivery cost.

This paper mainly focuses on the periodic aspects within the PSRP and has the aim of giving an overview of the modeling features, mathematical formulations, and solution methods related to the periodic PSRP. It is organized as follows. Section 8.2 is devoted to surveying the modeling issues related to the PPSRP and on how to define the feasible schedules describing the periodicity constraints. Section 8.3 formally defines the PPSRP, presents an integer optimization model for its formulation, and discusses the different available solution approaches. Finally, some directions for future research developments will be drawn in Sect. 8.4.

8.2 Modeling Aspects Related to the PPSRP

While modeling a real-life PPSRP, different aspects of the problem could be considered. These aspects lead to different optimization models and consequently diverse solution approaches. Hence, this section defines these aspects and constraints and highlights how they are incorporated into the problem formulation.

8.2.1 Classification of the PPSRP

PPSRP models come in a variety of forms, which can be classified with respect to the following criteria:

- *Number of commodities*: the delivery activity can involve one single commodity (such as just petrol) or simultaneously multiple commodities (such as normal petrol, unleaded petrol, diesel, etc.).
- *Number of tank-trucks*: can be fixed a priori, but in some cases can also constitute a decision variable of the problem.
- *Type of tank-trucks*: trucks can be identical (homogeneous fleet), for example, all having the same capacity, or can have different characteristics (heterogeneous fleet) allowing to decide on the truck route assignment. Moreover, tank-trucks can have single or multiple compartments.
- *Number of depots*: petrol replenishment problems become more challenging when more than one depot has to be considered and more specifically, if truck could be assigned dynamically to any of the depots.
- *Level of dynamism*: service requests from the stations can be all known at the beginning of the planning horizon (static problems) or can be revealed only during the service time (dynamic problems) necessitating, thus, a continuous adjustments of the routes.
- *Structure of the routes*: there may be precedence/priority constraints of the stations, constraints on the maximum number of stations to be served per day/route, or constraints related to the maximum driving distance per truck.
- *Time windows*: stations can impose time limits within which the service should take place. These additional constraints may lead to further decisions related to repositioning strategies or waiting policies of the trucks.

8.2.2 Periodicity Schedules Definition

Different alternative representations can be used in order to describe the periodicity and, thus, to incorporate the resulting schedules into the optimization models. All the representations lead to the definition, for each station, say i, of a set of possible schedules of f_i service days that are feasible for that station. Three different representations have been proposed in the literature and will be detailed below. While the set of possible schedules is enumerated explicitly by each station in the first technique, a preprocessing phase is needed in the successive two techniques in order to form an explicit set of the feasible schedules (called also combinations, sequences, etc.).

8.2.2.1 Predetermined Schedules

This representation, adopted by [8], is the simplest and most used technique for representing periodicity. It consists of explicitly specifying all the allowable alternatives that define the set of schedules and hence the decision maker should select only one of them.

8.2.2.2 Periodic Schedules

Each station specifies, in this technique, the number of service visits f_i during the t-day planning horizon, and then that station must be visited every t/f_i days. If, for example, $f_i = 3$ and $t = 6$, then station i must be periodically visited every $t/f_i = 2$ days up to the end of the planning horizon. This means that, during the 6-day planning horizon, the feasible service schedules for station i are $\{(1, 3, 5), (2, 4, 6)\}$. The ratio t/f_i defines the cardinality of the set of feasible schedules for station i. This representation technique has been used by authors like [7, 13, 19].

8.2.2.3 Multi-stage Network-Based Schedules

In this representation, station i must be visited once during each time interval of r_i days, so $f_i = t/r_i$. Furthermore, additional constraints impose that at least l_i days and at most u_i days must elapse between two successive visits. This technique, proposed by [16], can be represented as an acyclic multi-stage network corresponding to each station i. The nodes of stage k represent the allowable alternative days to execute the kth visit to station i, whereas each edge represents two possible successive visit days. The set of feasible schedules is thus defined by all the paths between the nodes of the first and the last stages of the directed network. An example of a network structure with $t = 9$, $r_i = 3$, $l_i = 2$, and $u_i = 4$ is represented in Fig. 8.1, which defines for station i the following set of feasible schedules: $\{(1, 4, 7), (1, 4, 8), (1, 5, 7), (1, 5, 8), (1, 5, 9), (2, 4, 7), (2, 4, 8), (2, 5, 7), (2, 5, 8), (2, 5, 9), (2, 6, 8), (2, 6, 9), (3, 5, 7), (3, 5, 8), (3, 5, 9), (3, 6, 8), (3, 6, 9)\}$.

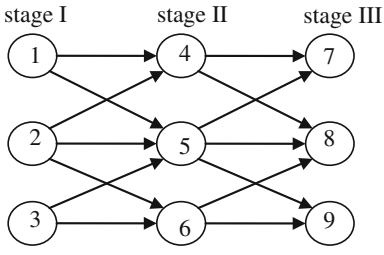

Fig. 8.1 Multi-stage network-based schedules

8.3 Mathematical Formulations

As discussed in Sect. 8.1, despite its importance in industrial applications, the PPSRP is still in its infancy and only a limited number of works have tried to take advantage from extending the planning horizon while solving the petrol replenishment activities. However, the PPSRP is tightly related to the well-known periodic vehicle routing problem (PVRP) that has been intensively studied over the past three decades (see for example the excellent survey by [15] and the references therein). The enormous advances achieved in investigating the PVRP may be very useful for tackling the PPSRP since the two problems have many similarities. The PPSRP is, indeed, basically a PVRP that includes side constraints that take into account the specifications of the products to be delivered and additional operational restrictions related to the drivers' shifts, trucks assignment to the routes, etc.

8.3.1 PVRP Versus PPSRP

The problems related to the distribution of material goods from a set of deposits to a set of customers are generally known as the *Vehicle Routing Problem* (VRP). It consists of serving a set of customers by using the available fleet of vehicles that are located in one or more depots. The solution of a VRP is then represented as a set of routes, that begin and end in the depots, and to which is assigned a single vehicle. The resulting routes are optimal only if all customers' demands are met, all operational constraints are satisfied and the objective of minimizing the total cost of transportation is achieved. There are many variants of VRPs but the simplest and most studied one is the *Capacitated Vehicle Routing Problem* (CVRP). In this problem, all customers are characterized by demand quantities that are deterministic and known a priori. Solving the CVRP means finding a set of minimum cost routes (the cost of a route is its length or its service time) that ensures that the sum of the demands of the customers visited in a route must not exceed the capacity of the vehicle performing the service. Figure 8.2 reports an example of a feasible CVRP solution showing four routes ($C1$, ..., $C4$), the capacity of the assigned vehicles ($Q1$, ..., $Q4$) and the demand at each node of the network.

The VRP becomes even more challenging when the planning horizon is extended to several days, instead of limiting it to a single day which defines the *Periodic Vehicle Routing Problem* (PVRP). During the extended planning horizon, each customer must be served exactly a certain number of times, identified by its frequency of service.

The PPSRP has usually the same objective function as the PVRP and they also share the following operational constraints related to the periodicity aspects:

- Each vehicle must start its route, on each day of the planning horizon, from the depots and returns to them at the end of the working day; multiple visits to the same depot in the same day may be also acceptable.

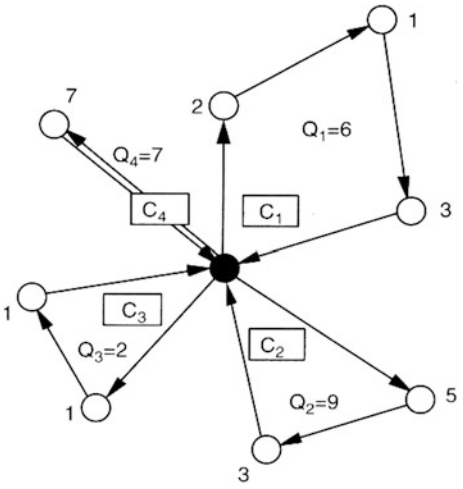

Fig. 8.2 Example of four feasible routes of a CVRP (the numbers next to the nodes are the demands and $Q1$, …, $Q4$ are the capacities of the vehicles)

- Each customer has to be visited by at most one vehicle every day;
- A vehicle can visit more than one customer in the same route, but the total demand cannot exceed its maximum capacity;
- Each customer must be visited exactly a prespecified number of times during the planning horizon, as defined by its frequency of service.

However, the PPSRP involves additional operational constraints that are usually strongly dependent on the application under consideration, such as

- Some tank-trucks cannot be partially filled because of the dynamic stability requirements, especially when performing difficult routes.
- Some trucks may have different compartments in the same tank in order to allow delivering different petroleum products at the same trip.
- Safety considerations that are usually particularly restrictive while transporting flammable products such as petrol.
- Access limitations that prohibit the movement of the tank-trucks on certain roads and forbid some drivers to enter certain zones (such as inexperienced drivers to the airports or to military zones).

As a consequence, it is difficult to characterize all the PPSRPs by a typical optimization formulation since every instance has its ad-hoc features that should be specifically studied in details in order to define the more appropriate mathematical model. In the sequel, we will focus on the basic variant of the PPSRP in order to present an optimization model that could be considered as general as possible. This model is nothing but a PVRP formulation since it does not take into account the particular requirements deriving from the product to be delivered (petrol) that are, as mentioned above, application dependent.

8.3.2 Optimization Model

In order to mathematically represent a general variant of the PPSRP, consider a connected undirected graph $G = (V, A)$, where $V = \{1, 2, \ldots, n\}$ is the set of nodes (stations) and the set $A = \{(i,j) : i,j = 1, .., n, i \neq j\}$ represents the edges connecting the nodes. Considering a connected graph means that we assume that each node is spatially connected to all other nodes through at least one route. Vertex 1 represents the depot, i.e., the node where is located the heterogeneous fleet of tank-trucks and from where all the routes are started and finished. To each tank-truck $NV = 1, \ldots, m$ is associated a capacity denoted by Q_{NV}. To each edge (i,j) of A, is associated a traveling cost c_{ij} (we assume here, for simplicity, that it is independent from the truck to be used) to move from node i to node j. On the other hand, to each node $i = 1, \ldots, n$ of V correspond a nonnegative demand d_i and a frequency of service f_i that indicates the exact number of service days along the planning horizon. Consequently, to each node i is associated a set of schedules C_i of feasible service days determined by using one of the techniques analyzed above. For this purpose, let us define a binary constant a_{rl} that takes value 1 if day l belongs to a certain service schedule r, and 0 otherwise.

Concerning the decision variables of the model, we define two sets of variables. The first set describes the tank-truck routes as follows:

$$\begin{cases} x_{ijvl} = 1 & \text{if edge } i-j \text{ is part of the route performed by truck } v \text{ in day } l \\ x_{ijvl} = 0 & \text{otherwise,} \end{cases}$$

whereas the second set of variables identifies the schedule assigned to each station

$$\begin{cases} y_{ir} = 1 & \text{if service schedule } r \text{ is assigned to station } i \\ y_{ir} = 0 & \text{otherwise.} \end{cases}$$

From the above variables definition it is clear that the model should perform two tasks. First, it identifies the best schedules to be assigned to each station and then, it builds the routes of each truck and for each day of the planning horizon. The mathematical model of the basic variant of the PPSRP could be represented thus, as follows:

$$\text{MIN} \quad z = \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{v=1}^{NV} \sum_{l=1}^{t} c_{ij} x_{ijvl} \tag{8.1}$$

Subject to

$$\sum_{r \in C_i}^{n} y_{ir} = 1 \quad i = 2, \ldots, n \tag{8.2}$$

$$\sum_{j=1}^{n}\sum_{v=1}^{NV} x_{ijvl} - \sum_{r\in C_i} a_{rl} y_{ir} = 0 \quad i = 2,\ldots,n; l = 1,\ldots,t \tag{8.3}$$

$$\sum_{i=1}^{n} x_{ipvl} - \sum_{j=1}^{n} x_{pjvl} = 0 \quad v = 1,\ldots,NV; p = 1,\ldots,n; l = 1,\ldots,t \tag{8.4}$$

$$\sum_{i=1}^{n} d_i \left(\sum_{j=1}^{n} x_{ijvl} \right) \leq Q_v \quad v = 1,\ldots,NV; l = 1,\ldots,t \tag{8.5}$$

$$\sum_{v_i\in S}\sum_{v_j\in S} x_{ijvl} \leq |S| - 1 \quad v = 1,\ldots,NV; l = 1,\ldots,t; S\subseteq N - \{1\}; |S| \geq 2 \tag{8.6}$$

$$x_{ijvl} \in \{0,1\} \quad i,j = 1,\ldots,n;\ v = 1,\ldots,NV; l = 1,\ldots,t \tag{8.7}$$

$$y_{ir} \in \{0,1\} \quad i = 1,\ldots,n; r \in \mathrm{C_i} \tag{8.8}$$

The objective function (8.1) minimizes the sum of the traveling costs for all the routes performed by all the trucks during all the days of the time horizon. The set of constraints (8.2) ensures that only one schedule among the feasible ones is assigned to each station (excluding, of course, the depot, i.e., $i = 1$). Constraints (8.3) ensure that each station is visited only on the days corresponding to the assigned schedule. Constraints (8.4) force each truck arriving to a node to leave it on the same day. Constraints (8.5) are the capacity limitations on each tank-truck. Constraints (8.6) prevent against the creation of undesirable subtours in the routing solution and finally, constraints (8.7) and (8.8) force all the decision variables to be binary.

8.3.3 Solution Approaches

Often, the one-day PSRP instances are already characterized by large-scale optimization models that make their solution with exact methods quite difficult [9]. When we add to this complexity, even the periodicity aspect related to the extended planning horizon, the problem becomes extremely hard to be faced by exact solution methods. Indeed, Triki has reported in [23] how it was not possible to get an exact solution of his test problem consisting of 14 heterogeneous tank-trucks, 38 petrol stations, and a planning horizon of six days by using general purpose state-of-the-art optimization software. After more than 30 h CPU time and a huge number of iterations, the solver has given a fault memory error and has succeeded to generate only a feasible upper bound solution. Consequently, to the best of our knowledge, no exact algorithms are available in the literatures that are specifically designed to solve the PPSRP. However, again some insights can derive from the PVRP literature where only recently Baldacci et al. have proposed an exact algorithm and several lower bounds for the problem [2]. They succeeded to solve

randomly generated test problems having up to 199 customers and covering a time horizon of five days.

Since it is very difficult to solve the PPSRP exactly, the attention of all the works available in the literature has been devoted to developing heuristic approaches. The solution strategy that seems to be more appropriate in this context belongs to the class "Group-Before, Route-After" (or Cluster-First, Route-Second). This class of algorithms splits the PPSRP into two distinct phases: in the first, each station is assigned a feasible schedule that takes into account its required frequency of service; In the second phase, the algorithm builds the optimal route for each truck by considering, for each day of the planning horizon, only the subgraph involving the stations assigned to that day [10, 17, 22, 23].

Metaheuristics can be also considered as a promising approach to solve routing problems with periodicity restrictions. However, we are not aware of any metaheuristic method that has been specifically developed to solve the PPSRP. Again, we should rely on the advances achieved in the context of the PVRP for which metaheuristics, such as, Tabu search [13] and genetic algorithms [14, 24] have been used for its solution.

8.4 Conclusions

Nowadays, most of the petrol distribution companies schedule the delivery of petrol to the stations by considering a single-day planning horizon. This paper has the aim of showing how extending the horizon to t working days may ensure important savings but to be paid usually by further complexity in the underlying optimization problems to be solved. This defines the PPSRP that has been shown to present many affinities with the well-known PVRP. The two problems share many common characteristics, but the nature of the products to be transported (petroleum) and the related restrictions usually generate additional challenges while solving the PPSRP. This explains the reason why the PPSRP did not reach yet the maturity level achieved by its general purpose counterpart and opens new research directions from different points of view.

References

1. Avella P, Boccia M, Sforza A (2004) Solving a fuel delivery problem by heuristic and exact approaches. Eur J Oper Res 152:170–179
2. Baldacci R, Bartolini E, Mingozzi A, Valletta A (2011) An exact algorithm for the period routing problem. Oper Res 59(1):228–241
3. Ben Abdelaziz F, Roucairol C, Bacha C (2002) Deliveries of liquid fuels to SNDP gas stations using vehicles with multiple compartments. In: Systems man and cybernetics IEEE international conference. Hammamet, Tunisia

4. Boctor F, Renaud J, Cornillier F (2011) Trip packing in petrol stations replenishment. Omega 39:86–98
5. Brown GG, GravesGW (1981) Realtime dispatch of petroleum tank trucks. Manag Sci 27:19–32
6. Brown GG, Ellis CJ, Graves GW, Ronen D (1987) Realtime wide area dispatch of Mobil tank trucks. Interfaces 17(1):107–120
7. Chao I-M, Golden BL, Wasil E (1995) A new heuristic for the period traveling salesman problem. Comput Oper Res 22:553–565
8. Christofides N, Beasley JE (1984) The period routing problem. Networks 14:237–256
9. Cornillier F, Boctor F, Laporte G, Renaud J (2008) An exact algorithm for the petrol station replenishment problem. J Optim Res Soc 59:607–615
10. Cornillier F, Boctor F, Laporte G, Renaud J (2008) A heuristic for the multi-period petrol station replenishment problem. Eur J Optim Res 191:295–305
11. Cornillier F, Laporte G, Boctor F, Renaud J (2009) The petrol station replenishment problem with time windows. Comput Oper Res 36:919–935
12. Cornillier F, Boctor F, Renaud J (2012) Heuristics for the multi-depot petrol station replenishment problem with time windows. Eur J Optim Res 220:361–369
13. Cordeau JF, Gendreau M, Laporte G (1997) A tabu search heuristic for periodic and multi-depot vehicle routing problems. Networks 30:105–119
14. Drummond LMA, Ochi LS, Vianna DS (2001) Asynchronous parallel metaheuristic for the period vehicle routing problem. Future Gener Comput Syst 17(4):379–386
15. Francis P, Smilowitz K, Tzur M (2008) The period vehicle routing problem and its extensions. In: Golden BL, Raghavan S, Wasil E (eds) The vehicle routing problem: latest advances and new challenges, vol 43. Springer, Berlin
16. Gaudioso M, Paletta G (1992) A heuristic for the periodic vehicle routing problem. Transp Sci 26:86–92
17. Malépart V, Boctor F, Renaud J, Labilois S (2003) Nouvelles approches pour l'approvisionnement des stations d'essence. Revue Française de Gestion Industrielle 22:15–31
18. Ng WL, Leung SH, Lam JP, Pan SW (2008) Petrol delivery tanker assignment and routing: a case study in Hong Kong. J Optim Res Soc 59:1191–1200
19. Paletta G, Triki C (2004) Solving the asymmetric traveling salesman problem with periodic constraints. Networks 44:31–37
20. Rizzoli A, Casagrande N, Donati A, Gambardella L, Lepori D, Montemanni R, Pina P, Zaffalon M (2003) Planning and optimisation of vehicle routes for fuel oil distribution. In: MODSIM international congress on modelling and simulation. Townsville, Australia
21. Surjandari I, Rachman A, Dianawati F, Wibowo RP (2011) Petrol delivery assignment with multi-product, multi-depot, split deliveries and time windows. Int J Modeling Optim 1(5):375–379
22. Taqaallah D, Renaud J, Boctor FF (2000) Le problème d'approvisionnement des stations d'essence. APII-JESA J. Européen des Systèmes Automatisés 34:11–33
23. Triki C (2013) Solution methods for the periodic petrol replenishment problem, The journal of engineering research. Accepted for publication (subject to modifications)
24. Vidal T, Crainic TG, Gendreau M, Lahrichi N, Rei W (2012) A hybrid genetic algorithm for multidepot and periodic vehicle routing problems. Oper Res 60(3):611–624

Chapter 9
Nanotechnology and Mathematics "Study of Non-linear Dynamic Vibration in Single Walled Carbon Nanotubes (SWNTs)"

Mushahid Husain and Ayub Khan

Abstract This paper discusses some aspects of the applied nonlinear mathematics that are used to solve the problems in nanotechnology. The equation of motion of nanoscopic systems, which is a nonlinear dynamical process, is discussed in the current study. It is observed that in the nonresonant response, the amplitude remains constant up to the second order of approximation.

Keywords Nanotechnology · Nonlinear dynamics · Carbon nanotube · SWNT · MWNT

9.1 Introduction

Nanotechnology is a rich source of intensity problems in applied mathematics. There are a number of problems in nanotechnology that may be solved using different mathematical homogenization methods. Microscopic boundary conditions for flow over this surface can be investigated mathematically. Expressions can be obtained in several limiting cases relating roughness and local slip to macroscopic slip boundary condition and show that this can significantly affect micro and nanoscale flows in some circumstances. The equation of motion of nanoscopic systems is generically nonlinear and frequently operates in a regime, where a linear approximation is not justified. The comprehension of the nonlinear dynamical process in nanosystems is a new field of research that is certainly of considerable technological importance. Nonlinear dynamics can be applied to solve the concept of aging effect in carbon nanotubes [1, 2] based devices.

M. Husain (✉)
Centre for Nanoscience and Nanotechnology, Jamia Millia Islamia, New Delhi, India
e-mail: mush_phys@rediffmail.com

A. Khan
Department of Mathematics, Jamia Millia Islamia, New Delhi, India

A.H. Siddiqi et al. (eds.), *Mathematical Models, Methods and Applications*, Industrial and Applied Mathematics, DOI 10.1007/978-981-287-973-8_9

Single-walled carbon nanotubes (SWNTs) are nanometer-diameter cylinders consisting of a single graphene sheet wrapped up to form a tube. The history of single-walled carbon nanotubes (SWNTs) began with the discovery of multiwalled carbon nanotubes (MWNTs). First SWNT made of just one layer of carbon atoms were created independently in 1993 by Iijima and Donald Bethune of IBM [3]. These are basically tubes of graphite and are normally capped at the ends. The caps are formed due to mixing in some pentagons with the hexagons. The theoretical minimum diameter of a SWNT is around 0.4 nm, which is about as long as two silicon atoms side by side. The average diameter tends to be around 1.2 nm on the basis of available literature. SWNT are more pliable than their multiwalled counterparts. They can be twisted, flattened, and bent into small circles or a round shape bends without breaking. The unique electronic properties of carbon nanotubes offer great intellectual challenges and potentials for new applications. Experiments and theoretical calculations have shown that depending only on diameter and helicity, single-walled carbon nanotubes (SWNTs) can be either metallic or semiconducting [4, 5].

Radial breathing modes (RBMs) is unique to CNTs without any counterpart in graphene sheets. The phonon modes in general and the RBM modes in particular have already received some attention in theoretical work [6]. To investigate the physical and mechanical performances of nanostructures, different approaches have been adopted [7–9]. When radius-to-thickness ratio of SWNTs is large, it may be treated as an elastic model. The nonlinear dynamic response of zig-zag (Fig. 9.1) (where $m = 0$) SWNTs under the effect of radial impulse has been studied [10]. In the course of studies, the response has been seen into the two cases namely, nonresonant and resonant. In the nonresonant case, it has been observed that the amplitude of the vibration remains constant up to the second order of approximation, while in the resonant case, there have been obtained, the resonant solution for the parametric and main frequencies. In the numerical part, plots exhibit the chaotic behavior of the nonlinear vibrations.

As in the close neighborhood of resonant solutions, there is a possibility of chaotic behavior. Therefore the analytic estimation of resonance solutions prevents us to employ the hit and trial technique of selecting the parameter to probe the chaotic behavior computationally.

The equation of motion of nonlinear planar oscillation of zig-zag $(n, 0)$ (Fig. 9.2) CNTs under the influence of radial impulse is written as [10]

$$\ddot{C}_n + \omega_n^2 C_n = -\frac{\eta^2 n^2}{(n^2+1)} \lambda_o \left(2 - n^2\right) \sin(\mu_o \tau) C_n, \tag{9.1}$$

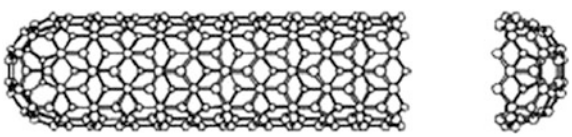

Fig. 9.1 Structure of a zig-zag SWCNTs

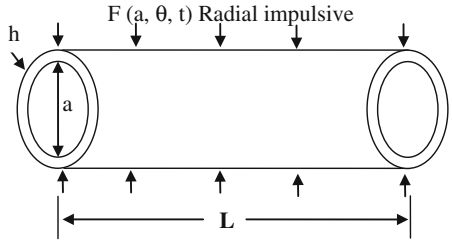

Fig. 9.2 The continuum shell model of CNT where h is the effective thickness and a is the radius of midsurface of CNTs, L/a is much larger than unit

$$\ddot{C}_n + \omega_n^2 C_n = -\eta\kappa_n \sin(\mu_o \tau) C_n, \tag{9.2}$$

$$\text{where } \kappa_n = \frac{n^2\lambda_o(2-n^2)}{(n^2+1)}. \tag{9.3}$$

C_n is generalized in extensional amplitude

$$\omega_n^2 = \frac{\eta^2(n^2-1)\alpha^2}{n^2+1} \tag{9.4}$$

$$\alpha^2 = \frac{R^2}{12a^2},\ \mu_o = 1, \eta = \frac{v_o}{c},\ \tau = \frac{ct}{\alpha},\ \lambda_o = 1 \tag{9.5}$$

For elastic motion, $\eta \approx 0$ then Eq. (9.2) will become

$$\ddot{C}_n + \omega_n^2 C_n \approx 0. \tag{9.6}$$

To study the non-resonant vibrations, the generating solution of (9.6) is given as

$$C_n = a\cos\phi \text{ and } \phi = \omega_n \tau + \phi^*, \tag{9.7}$$

where amplitude a and phase ϕ^* are to be determined by the initial conditions. The solution of (9.1) is obtained in the form of

$$C_n = a\cos\phi + \eta u_1(a,\phi,\tau) + \eta^2 u_2(a,\phi,\tau) + \cdots, \tag{9.8}$$

where a and ϕ are determined by the differential equations:

$$\frac{da}{d\tau} = \eta A_1(a) + \eta^2 A_2(a) + \cdots. \tag{9.9}$$

$$\frac{d\phi}{d\tau} = \omega_n + \eta B_1(a) + \eta^2 B_2(a) + \cdots. \tag{9.10}$$

After the first order of approximation, the solutions are

$$C_n = a\cos\phi,\ \frac{\mathrm{d}a}{\mathrm{d}\tau} = 0,\ \frac{\mathrm{d}\phi}{\mathrm{d}\tau} = \omega_n, \tag{9.11}$$

and in the second order of approximation, the solutions are

$$C_n = a\cos\phi + \eta\frac{a\kappa_n\cos\phi\sin(\mu_0\tau)}{\mu_0(2\omega_n+\mu_0)},\ \frac{\mathrm{d}a}{\mathrm{d}\tau} = 0 \text{ and } \frac{\mathrm{d}\phi}{\mathrm{d}\tau} = \omega_n. \tag{9.12}$$

The solution in the second approximation indicates the phenomenon of non-resonance. Also, it is observed that the amplitude is constant up to second order of approximation.

In case of resonant vibrations, the behavior of the dynamical system is studied in the neighborhood of the resonance. For $\eta = 0$, the generating solution of Eq. (9.6) is given as,

$$C_n = a\cos\phi,\ \ \phi = \frac{\tau}{k} + \theta, \tag{9.13}$$

where amplitude a and phase angle θ are determined by the following equations,

$$\frac{\mathrm{d}a}{\mathrm{d}\tau} = \eta A_1(a,\theta), \tag{9.14}$$

$$\frac{\mathrm{d}\theta}{\mathrm{d}\tau} = \omega_n - \frac{1}{k} + \eta B_1(a,\theta), \tag{9.15}$$

$$\frac{\mathrm{d}\phi}{\mathrm{d}\tau} = \omega_n + \eta B_1(a,\theta), \tag{9.16}$$

where $A_1(a,\theta)$, $B_1(a,\theta)$ are particular solutions with respect to θ.

During such investigations, it is observed that in the nonresonant response, the amplitude remains constant up to the second order of approximation. The computational studies based on the phase plots, time series, Poincare surface of sections, Poincare maps, and the graphs of resonant solutions reveal that the nonlinear response of SWNTs is chaotic when parameters η and ω_n are increased. On the other hand, the increasing values of the parameter k_n changes the behavior of the system from chaotic to regular. Thus, this conjecture enables to conclude that the above cited parameters η and ω_n are significantly responsible for chaotic (or aging) phenomena in SWNTs [10].

Acknowledgments Thanks are due to Department of Electronics and Information Technology (DeitY), Ministry of Communication & Information Technology, Government of India for providing financial assistance in the form of Major Research Project

References

1. Iijima S (1991) Nature 56:354
2. Iijima S, Ichihashi T (1993) Nature 363:603
3. Bethune DS, Kiang CH, de Vries MS, Gorman G (1993) Nature 363:605
4. Wildoer JWG, Venema JWG, Rinzler AG, Smalley RE (1998) Dekker. Nature 391:59
5. Odom TW, Huang J, Kim P, Lieber CM (1998) Nature 391:62
6. Yakobson BI, Brsbec CJ, Bernholc J (1996) Phys Rev Lett 76:2511
7. Li X, Diao B, Bhushan B (2002) Mater Charact 48:11
8. Ru CQ (2000) Phys Rev B 62:9973
9. Ru CQ (2001) J Appl Phys 89:3426
10. Khan A, Husain S, Shehzad M, Qadri SB, Husain M (2012) J Comput Theor Nanosci 9:360

Chapter 10
Generalized Monotone Mappings with Applications

R. Ahmad, A.H. Siddiqi, M. Dilshad and M. Rahaman

Abstract In this work, we introduce a generalized monotone mapping and we call it $H(\bullet,\bullet)$-cocoercive mapping. Then, we have extended this concept of $H(\bullet,\bullet)$-cocoercive mapping to $H(\bullet,\bullet)$-η-cocoercive mapping. Further, we have proved some of the properties of $H(\bullet,\bullet)$-cocoercive and $H(\bullet,\bullet)$-η-cocoercive mappings and finally apply these concepts to solve some generalized variational inclusions and system of variational inclusions.

Keywords Generalized monotonicity · Lipschitz continuity · Variational inclusions · System · Algorithm · Resolvent operator

10.1 Introduction

Fang and Huang [1] introduced H-monotone mappings for solving a system of variational inclusions involving a combination of H-monotone and strongly monotone mappings based on the resolvent mapping technique. The notion of H-monotonicity has revitalized the theory of maximal monotone mappings in many directions. Verma [2] introduced A-monotone mappings with applications to solve a system of nonlinear variational inclusions. Zou and Huang [3] introduced and studied $H(\bullet,\bullet)$-accretive mappings and applied them to solve variational inclusions and system of variational inclusions. For more details, we refer to [4–12].

Various concepts of generalized monotone mappings have been introduced in the literature. Cocoercive mappings which are generalized form of monotone

R. Ahmad (✉) · M. Dilshad · M. Rahaman
Department of Mathematics, Aligarh Muslim University, Aligarh 202002, India
e-mail: raisain123@rediffmail.com

A.H. Siddiqi
S-10, Administrative Block, Gautam Buddha University, Gautam Budh Nagar, Greater Noida, NCR, India

A.H. Siddiqi et al. (eds.), *Mathematical Models, Methods and Applications*, Industrial and Applied Mathematics, DOI 10.1007/978-981-287-973-8_10

mappings are defined by Tseng [13], Magnanti and Perakis [14] and Zhu and Marcotte [15].

We introduce a generalized monotone mapping and call it $H(\cdot,\cdot)$-cocoercive mapping. We also define its resolvent operator with some of its properties. We apply these new concepts to find the solutions of a generalized variational inclusions and system of variational inclusions.

By taking into account the fact that η-cocoercivity is an intermediate concept that lies between η-monotonicity and strong η-monotonicity, we extend the notion of $H(\cdot,\cdot)$-cocoercive mapping, we call it as $H(\cdot,\cdot)$-η-cocoercive mapping.

We define the resolvent operator of $H(\cdot,\cdot)$-η-cocoercive mapping with its properties and one numerical example through Matlab programming is also constructed. We apply the concept of $H(\cdot,\cdot)$-η-cocoercive mapping to solve a variational-like inclusion problem in real Banach spaces and a generalized variational-like inclusion problem in q-uniformly smooth Banach spaces.

10.2 $H(\cdot,\cdot)$-Cocoercive Mapping

In this section, we define a new generalized monotone mapping and we call it $H(\cdot,\cdot)$-cocoercive mapping. We discuss some of its properties.

Definition 2.1 Let $A, B: X \to X$, $H: X \times X \to X$ be three single-valued mappings. Then $M: X \to 2^X$ is said to be $H(\cdot,\cdot)$-cocoercive mapping with respect to mappings A and B (or simply $H(\cdot,\cdot)$-cocoercive in the sequel), if M is cocoercive and $(H(A, B) + \lambda M)(X) = X$, for every $\lambda > 0$.

Remark 2.1 Since cocoercive mappings includes monotone operators our definition is more general than definition of $H(\cdot,\cdot)$-accretive mapping [3]. It is easy to check that $H(\cdot,\cdot)$-cocoercive mappings provide a unified framework for the existing $H(\cdot,\cdot)$-monotone, H-monotone operators in Hilbert spaces and $H(\cdot,\cdot)$-accretive, H-accretive operators in Banach spaces.

Example 2.1 Let $X = \mathbb{R}^2$ with usual inner product. Let $A, B : \mathbb{R}^2 \to \mathbb{R}^2$ be defined by

$$Ax = (2x_1 - 2x_2, -2x_1 + 4x_2),$$
$$By = (-y_1 + y_2, -y_2), \text{ for all } x = (x_1, x_2), y = (y_1, y_2) \in \mathbb{R}^2.$$

Suppose that $H(A, B) : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}^2$ is defined by

$$H(Ax, By) = Ax + By, \text{ for all } x, y \in \mathbb{R}^2.$$

Then $H(A, B)$ is cocoercive with respect to A with constant $\frac{1}{6}$ and relaxed cocoercive with respect to B with constant $\frac{1}{2}$, since

$$\begin{aligned}
&\langle H(Ax,u)-H(Ay,u),x-y\rangle\\
&\quad=\langle Ax-Ay,x-y\rangle\\
&\quad=\langle(2x_1-2x_2,-2x_1+4x_2)-(2y_1-2y_2,-2y_1+4y_2),(x_1-y_1,x_2-y_2)\rangle\\
&\quad=\langle(2(x_1-y_1)-2(x_2-y_2),-2(x_1-y_1)+4(x_2-y_2)),(x_1-y_1,x_2-y_2)\rangle\\
&\quad=2(x_1-y_1)^2+4(x_2-y_2)^2-4(x_1-y_1)(x_2-y_2)
\end{aligned}$$

and

$$\begin{aligned}
\|\,Ax-Ay\,\|^2&=\langle((2x_1-2x_2,-2x_1+4x_2)-(2y_1-2y_2,-2y_1+4y_2)),\\
&\qquad((2x_1-2x_2,-2x_1+4x_2)-(2y_1-2y_2,-2y_1+4y_2))\rangle\\
&=8(x_1-y_1)^2+20(x_2-y_2)^2-24(x_1-y_1)(x_2-y_2)\\
&\le 12(x_1-y_1)^2+24(x_2-y_2)^2-24(x_1-y_1)(x_2-y_2)\\
&=6\{2(x_1-y_1)^2+4(x_2-y_2)^2-4(x_1-y_1)(x_2-y_2)\}\\
&=6\{\langle H(u,Ax)-H(u,Ay),x-y\rangle\},
\end{aligned}$$

which implies that

$$\langle H(Ax,u)-H(Ay,u),x-y\rangle\ge\frac{1}{6}\|\,Ax-Ay\,\|^2,$$

i.e., $H(A, B)$ is cocoercive with respect to A with constant $\frac{1}{6}$.

Further,

$$\begin{aligned}
\langle H(u,Bx)-H(u,By),x-y\rangle&=\langle Bx-By,x-y\rangle\\
&=\langle(-x_1+x_2,-x_2)-(-y_1+y_2,-y_2),(x_1-y_1,x_2-y_2)\rangle\\
&=\langle(-(x_1-y_1)+(x_2-y_2),-(x_2-y_2)),(x_1-y_1,x_2-y_2)\rangle\\
&=-(x_1-y_1)^2-(x_2-y_2)^2+(x_1-y_1)(x_2-y_2)\\
&=-\{(x_1-y_1)^2+(x_2-y_2)^2-(x_1-y_1)(x_2-y_2)\},
\end{aligned}$$

and

$$\begin{aligned}
\|\,Bx-By\,\|^2&=\langle(-(x_1-y_1)+(x_2-y_2),-(x_2-y_2)),\\
&\qquad(-(x_1-y_1)+(x_2-y_2),-(x_2-y_2)\rangle\\
&=(x_1-y_1)^2+2(x_2-y_2)^2-2(x_1-y_1)(x_2-y_2)\\
&\le 2\{(x_1-y_1)^2+(x_2-y_2)^2-(x_1-y_1)(x_2-y_2)\}\\
&=2(-1)\langle H(Bx,u)-H(By,u),x-y\rangle
\end{aligned}$$

which implies that

$$\langle H(u,Bx) - H(u,By), x-y\rangle \geq -\frac{1}{2}\parallel Bx - By \parallel^2,$$

i.e., $H(A, B)$ is relaxed cocoercive with respect to B with constant $\frac{1}{2}$.

Example 2.2 Let X, A, B and H are same as in Example 2.1 and let $M : \mathbb{R}^2 \to \mathbb{R}^2$ be define by $M(x_1, x_2) = (0, x_2)$, for all $(x_1, x_2) \in \mathbb{R}^2$. Then it is easy to check that M is cocoercive and $(H(A,B) + \lambda M)(\mathbb{R}^2) = \mathbb{R}^2$, for all $\lambda > 0$, that is, M is $H(\bullet,\bullet)$-cocoercive mapping with respect to A and B.

Example 2.3 Let $X = \mathbb{S}^2$, where $\mathbb{S}^2$ denotes the space of all 2×2 real symmetric matrices. Let $H(Ax, By) = x^2 - y$, for all $x, y \in \mathbb{S}^2$ and $M = I$. Then for $\lambda = 1$, we have

$$(H(A,B) + M)(x) = x^2 - x + x = x^2,$$

but

$$\begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix} \notin (H(A,B) + M)(\mathbb{S}^2),$$

because $\begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix}$ is not the square of any 2×2 real symmetric matrix. Hence M is not $H(\bullet,\bullet)$-cocoercive with respect to A and B.

Since $H(\bullet,\bullet)$-cocoercive mappings are more general than maximal monotone operators, we give the following characterization of $H(\bullet,\bullet)$-cocoercive mappings.

Proposition 2.1 Let $H(A, B)$ be cocoercive with respect to A with constant $\mu > 0$, relaxed cocoercive with respect to B with constant $\gamma > 0$, A is α-expansive, B is β-Lipschitz continuous and $\mu > \gamma$, $\alpha > \beta$. Let $M{:}X \to 2^X$ be $H(\bullet,\bullet)$-cocoercive mapping. If the following inequality

$$\langle x - y, u - v\rangle \geq 0$$

holds for all $(v, y) \in \text{Graph}(M)$, then $x \in Mu$, where

$$\text{Graph}(M) = \{(u,x) \in X \times X : x \in M(u)\}.$$

Proof Suppose that there exists some (u_0, x_0) such that

$$\langle x_0 - y, u_0 - v\rangle \geq 0, \quad \text{for all} \quad (v,y) \in \text{Graph}(M). \tag{2.1}$$

Since M is $H(\bullet,\bullet)$-cocoercive, we know that $(H(A, B) + \lambda M)(X) = X$ holds for every $\lambda > 0$ and so there exists $(u_1, x_1) \in \text{Graph}(M)$ such that

$$H(Au_1, Bu_1) + \lambda x_1 = H(Au_0, Bu_0) + \lambda x_0 \in X. \tag{2.2}$$

It follows from (2.1) and (2.2) that

$$\begin{aligned}
0 \le \langle \lambda x_0 + H(Au_0, Bu_0) - \lambda x_1 - H(Au_1, Bu_1), u_0 - u_1 \rangle \\
0 \le \lambda \langle x_0 - x_1, u_0 - u_1 \rangle &= -\langle H(Au_0, Bu_0) - H(Au_1, Bu_1), u_0 - u_1 \rangle \\
&= -\langle H(Au_0, Bu_0) - H(Au_1, Bu_0), u_0 - u_1 \rangle \\
&\quad - \langle H(Au_1, Bu_0) - H(Au_1, Bu_1), u_0 - u_1 \rangle \\
&\le -\mu \parallel Au_0 - Au_1 \parallel^2 + \gamma \parallel Bu_0 - Bu_1 \parallel^2 \\
&\le -\mu\alpha^2 \parallel u_0 - u_1 \parallel^2 + \gamma\beta^2 \parallel u_0 - u_1 \parallel^2 \\
&= -(\mu\alpha^2 - \gamma\beta^2) \parallel u_0 - u_1 \parallel^2 \le 0,
\end{aligned}$$

which gives $u_1 = u_0$, since $\mu > \gamma$, $\alpha > \beta$. By (2.2), we have $x_1 = x_0$. Hence, $(u_0, x_0) = (u_1, x_1) \in \text{Graph}(M)$ and so $x_0 \in Mu_0$. □

Theorem 2.1 *Let X be a real Hilbert space and M:X → 2^X be a maximal monotone operator. Suppose that H:X × X → X be a bounded cocoercive and semi-continuous with respect to A and B. Let H:X × X → X is also cocoercive with respect to A with constant μ > 0 and relaxed cocoercive with respect to B with constant γ > 0. The mapping A is α-expansive and B is β-Lipschitz continuous. If μ > γ and α > β, then M is H(•,•)-cocoercive.*

Proof For the proof we refer to [3]. □

Theorem 2.2 *Let H(A, B) be a cocoercive with respect to A with constant μ > 0 and relaxed cocoercive with respect to B with constant γ > 0, A is α-expansive and B is β-Lipschitz continuous, μ > γ and α > β. Let M be an H(•,•)-cocoercive mapping. Then the operator $(H(A, B) + \lambda M)^{-1}$ is single-valued.*

Proof For any given $u \in X$, let $x, y \in (H(A, B) + \lambda M)^{-1}(u)$. It follows that

$$-H(Ax, Bx) + u \in \lambda Mx$$

and

$$-H(Ay, By) + u \in \lambda My.$$

As M is cocoercive (thus monotone), we have

$$\begin{aligned}
0 \le \langle -H(Ax, Bx) + u - (-H(Ay, By) + u), x - y \rangle \\
&= -\langle H(Ax, Bx) - H(Ay, By), x - y \rangle \\
&= -\langle H(Ax, Bx) - H(Ay, Bx) + H(Ay, Bx) - H(Ay, By), x - y \rangle \\
&= -\langle H(Ax, Bx) - H(Ay, Bx), x - y \rangle \\
&\quad - \langle H(Ay, Bx) - H(Ay, By), x - y \rangle.
\end{aligned} \tag{2.3}$$

Since H is cocoercive with respect to A with constant $\mu > 0$ and relaxed cocoercive with respect to B with constant $\gamma > 0$, A is α-expansive and B is β-Lipschitz continuous, thus (2.3) becomes

$$0 \leq -\mu\alpha^2 \parallel x - y \parallel^2 + \gamma\beta^2 \parallel x - y \parallel^2 = -(\mu\alpha^2 - \gamma\beta^2) \parallel x - y \parallel^2 \leq 0, \quad (2.4)$$

since $\mu > \gamma$, $\alpha > \beta$. Thus, we have $x = y$ and so $(H(A, B) + \lambda M)^{-1}$ is single-valued. □

Definition 2.2 Let $H(A, B)$ be cocoercive with respect to A with constant $\mu > 0$ and relaxed cocoercive with respect to B with constant $\gamma > 0$, A is α-expansive and B is β-Lipschitz continuous and $\mu > \gamma$, $\alpha > \beta$. Let M be $H(\bullet,\bullet)$-cocoercive mapping. The resolvent operator $R_{\lambda,M}^{H(\cdot,\cdot)}: X \to X$ is defined by

$$R_{\lambda,M}^{H(\cdot,\cdot)}(u) = (H(A, B) + \lambda M)^{-1}(u), \text{ for all } u \in X. \quad (2.5)$$

Now, we prove the Lipschitz continuity of resolvent operator defined by (2.5) and estimate its Lipschitz constant.

Theorem 2.3 *Let $H(A, B)$ be cocoercive with respect to A with constant $\mu > 0$, relaxed cocoercive with respect to B with constant $\gamma > 0$, A is α-expansive and B is β-Lipschitz continuous and $\mu > \gamma$, $\alpha > \beta$. Let M be $H(\bullet,\bullet)$-cocoercive mapping. Then, the resolvent operator $R_{\lambda,M}^{H(\cdot,\cdot)}: X \to X$ is $\frac{1}{\mu\alpha^2 - \gamma\beta^2}$-Lipschitz continuous, that is*

$$\parallel R_{\lambda,M}^{H(\cdot,\cdot)}(u) - R_{\lambda,M}^{H(\cdot,\cdot)}(v) \parallel \leq \frac{1}{\mu\alpha^2 - \gamma\beta^2} \parallel u - v \parallel, \text{for all } u, v \in X.$$

Proof Let u and v be any given points in X. It follows from (2.5) that

$$R_{\lambda,M}^{H(\cdot,\cdot)}(u) = (H(A, B) + \lambda M)^{-1}(u),$$

and

$$R_{\lambda,M}^{H(\cdot,\cdot)}(v) = (H(A, B) + \lambda M)^{-1}(v).$$

This implies that

$$\frac{1}{\lambda}(u - H(A(R_{\lambda,M}^{H(\cdot,\cdot)}(u)), B(R_{\lambda,M}^{H(\cdot,\cdot)}(u)))) \in M(R_{\lambda,M}^{H(\cdot,\cdot)}(u)),$$

and

$$\frac{1}{\lambda}(v - H(A(R_{\lambda,M}^{H(\cdot,\cdot)}(v)), B(R_{\lambda,M}^{H(\cdot,\cdot)}(v)))) \in M(R_{\lambda,M}^{H(\cdot,\cdot)}(v)).$$

For the sake of clarity, we take

$$Pu = R^{H(\cdot,\cdot)}_{\lambda,M}(u), Pv = R^{H(\cdot,\cdot)}_{\lambda,M}(v).$$

Since M is cocoercive (hence monotone), we have

$$\frac{1}{\lambda}\langle u - H(A(Pu), B(Pu)) - (v - H(A(Pv), B(Pv))), Pu - Pv\rangle \geq 0,$$
$$\frac{1}{\lambda}\langle u - v - H(A(Pu), B(Pu)) + H(A(Pv), B(Pv)), Pu - Pv\rangle \geq 0,$$

which implies that

$$\langle u - v, Pu - Pv\rangle \geq \langle H(A(Pu), B(Pu)) - H(A(Pv), B(Pv)), Pu - Pv\rangle.$$

Further, we have

$$\begin{aligned}
\| u - v \| \| Pu - Pv \| &\geq \langle u - v, Pu - Pv\rangle \\
&\geq \langle H(A(Pu), B(Pu)) - H(A(Pv), B(Pv)), Pu - Pv\rangle \\
&= \langle H(A(Pu), B(Pu)) - H(A(Pv), B(Pu)) \\
&\quad + H(A(Pv), B(Pu)) - H(A(Pv), B(Pv)), Pu - Pv\rangle \\
&= \langle H(A(Pu), B(Pu)) - H(A(Pv), B(Pu)), Pu - Pv\rangle \\
&\quad + \langle H(A(Pv), B(Pu)) - H(A(Pv), B(Pv)), Pu - Pv\rangle \\
&\geq \mu \| A(Pu) - A(Pv) \|^2 - \gamma \| B(Pu) - B(Pv) \|^2 \\
&\geq \mu\alpha^2 \| Pu - Pv \|^2 - \gamma\beta^2 \| Pu - Pv \|^2,
\end{aligned}$$

and so

$$\| u - v \| \| Pu - Pv \| \geq (\mu\alpha^2 - \gamma\beta^2) \| Pu - Pv \|^2,$$

thus,

$$\| Pu - Pv \| \leq \frac{1}{\mu\alpha^2 - \gamma\beta^2} \| u - v \|,$$

that is,

$$\| R^{H(\cdot,\cdot)}_{\lambda,M}(u) - R^{H(\cdot,\cdot)}_{\lambda,M}(v) \| \leq \frac{1}{\mu\alpha^2 - \gamma\beta^2} \| u - v \|, \text{ for all } u, v \in X.$$

This completes the proof. □

10.3 Generalized Variational Inclusions

In this section, we apply $H(\bullet,\bullet)$-cocoercive mapping for solving a generalized variational inclusion problem.

We consider the problem of finding $u \in X$ and $w \in T(u)$ such that

$$0 \in w + M(g(u)), \tag{3.1}$$

where $g: X \to X$, $M:X \to 2^X$ and $T: X \to CB(X)$ be the mappings. Problem (3.1) is introduced and studied by Huang [16] in the setting of Banach spaces.

Lemma 3.1 *(u, w), where $u \in X$, $w \in T(u)$, is a solution of the problem (3.1), if and only if (u, w) is a solution of the following equation*:

$$g(u) = R_{\lambda,M}^{H(\cdot,\cdot)}[H(A(gu), B(gu)) - \lambda w], \tag{3.2}$$

where $\lambda > 0$ is a constant.

Proof By using the definition of resolvent operator $R_{\lambda,M}^{H(\cdot,\cdot)}$, the conclusion follows directly. □

Based on (3.2), we construct the following algorithm.

Algorithm 3.1 *For any $u_0 \in X$, $w_0 \in T(u_0)$, compute the sequences* $\{u_n\}$ and $\{w_n\}$ by iterative schemes:

$$g(u_{n+1}) = R_{\lambda,M}^{H(\cdot,\cdot)}[H(A(gu_n), B(gu_n)) - \lambda w_n], \tag{3.3}$$

$$w_n \in T(u_n), \| w_n - w_{n+1} \| \leq (1 + \frac{1}{n+1})\mathcal{H}(T(u_n), T(u_{n+1})), \tag{3.4}$$

for all $n = 0, 1, 2, \ldots,$ and $\lambda > 0$ is a constant.

Theorem 3.1 *Let X be a real Hilbert space and A, B, $g:X \to X$, $H:X \times X \to X$ be the single-valued mappings. Let $T:X \to CB(X)$ be a set-valued mapping. Suppose that $M:X \to 2^X$ be the set-valued, $H(\bullet,\bullet)$-cocoercive mapping. Assume that*

1. T is δ-Lipschitz continuous in the Häusdorff metric $\mathcal{H}(\bullet,\bullet)$;
2. $H(A, B)$ is cocoercive with respect to A with constant $\mu > 0$ and relaxed cocoercive with respect to B with constant $\gamma > 0$;
3. A is α-expansive;
4. B is β-Lipschitz continuous;
5. g is λ_g-Lipschitz continuous and ξ-strongly monotone;
6. $H(A, B)$ is r_1-Lipschitz continuous with respect to A and r_2-Lipschitz continuous with respect to B;
7. $(r_1 + r_2)\lambda_g < [(\mu\alpha^2 - \gamma\beta^2)\xi - \lambda\delta]; \mu > \gamma, \alpha > \beta.$

Then, the generalized variational inclusion problem (3.1) has a solution (u, w) with $u \in X$, $w \in T(u)$ and the iterative sequences $\{u_n\}$ and $\{w_n\}$ generated by Algorithm 3.1 converge strongly to u and w, respectively.

Proof Since T is δ-Lipschitz continuous, it follows from Algorithm 3.1 that

$$\begin{aligned} \| w_n - w_{n+1} \| &\leq \left(1 + \frac{1}{n+1}\right) \mathcal{H}(T(u_n), T(u_{n+1})) \\ &\leq \left(1 + \frac{1}{n+1}\right) \delta \| u_n - u_{n+1} \|, \end{aligned} \tag{3.5}$$

for $n = 0, 1, 2, \ldots$

Using the ξ-strong monotonicity of g, we have

$$\begin{aligned} \| g(u_{n+1}) - g(u_n) \| \| u_{n+1} - u_n \| &\geq \langle g(u_{n+1}) - g(u_n), u_{n+1} - u_n \rangle \\ &\geq \xi \| u_{n+1} - u_n \|^2, \end{aligned}$$

which implies that

$$\| u_{n+1} - u_n \| \leq \frac{1}{\xi} \| g(u_{n+1}) - g(u_n) \|. \tag{3.6}$$

Now, we estimate $\| g(u_{n+1}) - g(u_n) \|$ by using the Lipschitz continuity of $R_{\lambda,M}^{H(\cdot,\cdot)}$.

$$\begin{aligned} \| g(u_{n+1}) - g(u_n) \| &= \| R_{\lambda,M}^{H(\cdot,\cdot)}[H(A(gu_n), B(gu_n)) - \lambda w_n] \\ &\quad - R_{\lambda,M}^{H(\cdot,\cdot)}[H(A(gu_{n-1}), B(gu_{n-1})) - \lambda w_{n-1}] \| \\ &\leq \frac{1}{\mu\alpha^2 - \gamma\beta^2} \| H(A(gu_n), B(gu_n)) - H(A(gu_{n-1}), B(gu_{n-1}))) \| \\ &\quad + \frac{\lambda}{\mu\alpha^2 - \gamma\beta^2} \| w_n - w_{n-1} \| \\ &\leq \frac{1}{\mu\alpha^2 - \gamma\beta^2} \| H(A(gu_n), B(g(u_n)) - H(A(gu_{n-1})), B(gu_n)) \| \\ &\quad + \frac{1}{\mu\alpha^2 - \gamma\beta^2} \| H(A(gu_{n-1}), B(gu_n)) - H(A(gu_{n-1}), B(gu_{n-1})) \| \\ &\quad + \frac{\lambda}{\mu\alpha^2 - \gamma\beta^2} \| w_n - w_{n-1} \|. \end{aligned} \tag{3.7}$$

Since $H(A, B)$ is r_1-Lipschitz continuous with respect to A and r_2-Lipschitz continuous with respect to B, g is λ_g-Lipschitz continuous and using (3.5), (3.7) becomes

$$\| g(u_{n+1}) - g(u_n) \| \leq \frac{r_1\lambda_g}{\mu\alpha^2 - \gamma\beta^2} \| u_n - u_{n-1} \| + \frac{r_2\lambda_g}{\mu\alpha^2 - \gamma\beta^2} \| u_n - u_{n-1} \| + \frac{\lambda}{\mu\alpha^2 - \gamma\beta^2}(1 + \frac{1}{n})\delta \| u_n - u_{n-1} \|,$$

or

$$\| g(u_{n+1} - g(u_n) \| \leq \left[\frac{r_1\lambda_g}{\mu\alpha^2 - \gamma\beta^2} + \frac{r_2\lambda_g}{\mu\alpha^2 - \gamma\beta^2} + \frac{\lambda}{\mu\alpha^2 - \gamma\beta^2}(1 + \frac{1}{n})\delta\right] \| u_n - u_{n-1} \|. \tag{3.8}$$

Using (3.8), (3.6) becomes

$$\| u_{n+1} - u_n \| \leq \theta_n \| u_n - u_{n-1} \|, \tag{3.9}$$

where

$$\theta_n = \frac{(r_1 + r_2)\lambda_g + \lambda\delta(1 + 1/n)}{(\mu\alpha^2 - \gamma\beta^2)\xi}.$$

Letting

$$\theta = \frac{(r_1 + r_2)\lambda_g + \lambda\delta}{(\mu\alpha^2 - \gamma\beta^2)\xi}.$$

We know that $\theta_n \to \theta$ and $n \to \infty$. From assumption (vii), it is easy to see that $\theta < 1$. Therefore, it follows from (3.9) that $\{u_n\}$ is a Cauchy sequence in X. Since X is a Hilbert space, there exists $u \in X$ such that $u_n \to u$ as $n \to \infty$. From (3.5), we know that $\{w_n\}$ is also a Cauchy sequence in X, thus there exists $w \in X$ such that $w_n \to w$ as $n \to \infty$. By the continuity of g, $R_{\lambda,M}^{H(\cdot,\cdot)}$, H, A, B, and T and Algorithm 3.1, we have

$$g(u) = R_{\lambda,M}^{H(\cdot,\cdot)}[H(A(gu), B(gu)) - \lambda w].$$

Now, we prove that $w \in T(u)$. In fact, since $w_n \in T(u_n)$, we have

$$\begin{aligned} d(w, T(u)) &\leq \| w - w_n \| + d(w_n, T(u)) \\ &\leq \| w - w_n \| + \mathcal{H}(T(u_n), T(u)) \\ &\leq \| w - w_n \| + \delta \| u_n - u \| \to 0, \quad \text{as } n \to \infty, \end{aligned}$$

which implies that $d(w, T(u)) = 0$. Since $T(u) \in CB(X)$, it follows that $w \in T(u)$. By Lemma 3.1, it follows that (u, w) is a solution of problem (3.1). This completes the proof. □

10.4 System of Variational Inclusions

In this section, we study a system of variational inclusions involving $H(\cdot,\cdot)$-cocoercive mapping.

Let X_1 and X_2 be two real Hilbert spaces and let F: $X_1 \times X_2 \to X_1$, G: $X_1 \times X_2 \to X_2$, H_1: $X_1 \times X_1 \to X_1$, H_2: $X_2 \times X_2 \to X_2$, A_1, B_1: $X_1 \to X_1$, A_2, B_2: $X_2 \to X_2$ be the single-valued mappings. Let M: $X_1 \to 2^{X_1}$ be a set-valued, $H_1(A_1, B_1)$-cocoercive mapping and N: $X_2 \to 2^{X_2}$ be a set-valued, $H_2(A_2, B_2)$-cocoercive mapping. Find $(a, b) \in X_1 \times X_2$ such that

$$\begin{cases} 0 \in F(a,b) + M(a), \\ 0 \in G(a,b) + N(b). \end{cases} \tag{4.1}$$

Problem (4.1) is called system of variational inclusions.

If $M : X_1 \to 2^{X_1}$ is (H_1, η)-monotone and $N : X_2 \to 2^{X_2}$ is (H_2, η)-monotone, then problem (4.1) includes the problem considered and studied by Fang et al. [17].

It is clear that for suitable choices of operators involved in the formulation of problem (4.1), one can obtain many systems of variational inequalities and variational inclusions exist in the literature.

Lemma 4.1 *Let X_1 and X_2 be two real Hilbert spaces. Let F:$X_1 \times X_2 \to X_1$, G: $X_1 \times X_2 \to X_2$, A_1, B_1:$X_1 \to X_1$, A_2, B_2:$X_2 \to X_2$ be the single-valued mappings. Let H_1:$X_1 \times X_1 \to X_1$ be a single-valued mapping such that $H_1(A_1, B_1)$ is cocoercive with respect to A_1 with constant $\mu_1 > 0$ and relaxed cocoercive with respect to B_1 with constant $\gamma_1 > 0$, A_1 is α_1-expansive and B_1 is β_1-Lipschitz continuous, $\alpha_1 > \beta_1$ and $\mu_1 > \gamma_1$. Let H_2:$X_2 \times X_2 \to X_2$ be also a single-valued mapping such that $H_2(A_2, B_2)$ is cocoercive with respect to A_2 with constant $\mu_2 > 0$ and relaxed cocoercive with respect to B_2 with constant $\gamma_2 > 0$, A_2 is α_2-expansive and B_2 is β_2-Lipschitz continuous, $\alpha_2 > \beta_2$ and $\mu_2 > \gamma_2$. Let M:$X_1 \to 2^{X_1}$ is set-valued, $H_1(\cdot,\cdot)$-cocoercive mapping and N:$X_2 \to 2^{X_2}$ is set-valued, $H_2(\cdot,\cdot)$-cocoercive mapping. Then $(a, b) \in X_1 \times X_2$ is a solution of problem (4.1) if and only if (a, b)* satisfies the following:

$$\begin{cases} a = R_{\lambda,M}^{H_1(\cdot,\cdot)}[H_1(A_1(a), B_1(a)) - \lambda F(a,b)], \\ b = R_{\rho,N}^{H_2(\cdot,\cdot)}[H_2(A_2(b), B_2(b)) - \rho G(a,b)], \end{cases}$$

where $\lambda > 0$ and $\rho > 0$ are two constants.

Proof The conclusion can be obtained directly from the definitions of $R_{\lambda,M}^{H_1(\cdot,\cdot)}$ and $R_{\rho,N}^{H_2(\cdot,\cdot)}$. □

Based on Lemma 4.1, we now define an iterative algorithm for approximating a solution of problem (4.1).

Algorithm 4.1 Let X_1, X_2, A_1, A_2, B_1, B_2, H_1, H_2, M, N, F, and G are same as Lemma 4.1. For any given initial $(a_0, b_0) \in X_1 \times X_2$, we define the following iterative scheme:

$$\begin{cases} a_{n+1} = R_{\lambda,M}^{H_1(\cdot,\cdot)}[H_1(A_1(a_n), B_1(a_n)) - \lambda F(a_n, b_n)], \\ b_{n+1} = R_{\rho,N}^{H_2(\cdot,\cdot)}[H_2(A_2(b_n), B_2(b_n)) - \rho G(a_n, b_n)], \end{cases}$$

for $n = 0, 1, 2, \ldots\ldots$, where $\lambda > 0$ and $\rho > 0$ are two constants.

Now, we show the existence of solution of problem (4.1) and analyze the convergence of iterative Algorithm 4.1.

Theorem 4.1 *Let X_1 and X_2 be two real Hilbert spaces. Let A_1, B_1:$X_1 \to X_1$, A_2, B_2:$X_2 \to X_2$ be the single-valued mappings. Let H_1:$X_1 \times X_1 \to X_1$ be a single-valued mapping such that $H_1(A_1, B_1)$ is cocoercive with respect to A_1 with constant $\mu_1 > 0$ and relaxed cocoercive with respect to B_1 with constant $\gamma_1 > 0$, A_1 is α_1-expansive and B_1 is β_1-Lipschitz continuous, $\alpha_1 > \beta_1$ and $\mu_1 > \gamma_1$. Let H_2: $X_2 \times X_2 \to X_2$ be also a single-valued mapping such that $H_2(A_2, B_2)$ is cocoercive with respect to A_2 with constant $\mu_2 > 0$ and relaxed cocoercive with respect to B_2 with constant $\gamma_2 > 0$, A_2 is α_2-expansive and B_2 is β_2-Lipschitz continuous, $\alpha_2 > \beta_2$ and $\mu_2 > \gamma_2$. Let $M : X_1 \to 2^{X_1}$ is set-valued, $H_1(\bullet,\bullet)$-cocoercive mapping and $N : X_2 \to 2^{X_2}$ is set-valued, $H_2(\bullet,\bullet)$-cocoercive mapping. Assume that $H_1(A_1, B_1)$ is r_1-Lipschitz continuous with respect to A_1 and r_2-Lipschitz continuous with respect to B_1, F:$X_1 \times X_2 \to X_1$ is τ_1-Lipschitz continuous with respect to the first argument and τ_2-Lipschitz continuous with respect to the second argument, $H_2(A_2, B_2)$ is r_3-Lipschitz continuous with respect to A_2 and r_4-Lipschitz continuous with respect to B_2, G:$X_1 \times X_2 \to X_2$ is τ_1'-Lipschitz continuous with respect to first argument and τ_2'-Lipschitz continuous with respect to second argument. $F(x, \cdot)$ is m_1-strongly monotone with respect to $H_1(A_1, B_1)$ and $G(\cdot, y)$ is m_2-strongly monotone with respect to $H_2(A_2, B_2)$. If the following conditions are satisfied:*

$$\begin{cases} 0 < \dfrac{\sqrt{(r_1 + r_2)^2 - 2\lambda m_1 + \lambda^2 \tau_1^2}}{\mu_1 \alpha_1^2 - \gamma_1 \beta_1^2} + \dfrac{\rho \tau_1'}{\mu_2 \alpha_2^2 - \gamma_2 \beta_2^2} < 1, \\ 0 < \dfrac{\sqrt{(r_3 + r_4)^2 - 2\rho m_2 + \rho^2 \tau_2'^2}}{\mu_2 \alpha_2^2 - \gamma_2 \beta_2^2} + \dfrac{\lambda \tau_2}{\mu_1 \alpha_1^2 - \gamma_1 \beta_1^2} < 1. \end{cases} \tag{4.2}$$

Then the problem (4.1) admits a solution $(a, b) \in X_1 \times X_2$ and the sequence $\{(a_n, b_n)\}$ generated by Algorithm 4.1 converges strongly to a solution (a, b) of problem (4.1).

Proof From Algorithm 4.1 and Theorem 2.3, we have

$$\begin{aligned}
\| a_{n+1} - a_n \| &= \| R_{\lambda,M}^{H_1(\cdot,\cdot)}[H_1(A_1(a_n), B_1(a_n)) - \lambda F(a_n, b_n)] \\
&\quad - R_{\lambda,M}^{H_1(\cdot,\cdot)}[H_1(A_1(a_{n-1}), B_1(a_{n-1})) - \lambda F(a_{n-1}, b_{n-1})] \| \\
&\leq \frac{1}{\mu_1\alpha_1^2 - \gamma_1\beta_1^2} \| H_1(A_1(a_n), B_1(a_n)) - \lambda F(a_n, b_n) \\
&\quad - [H_1(A_1(a_{n-1}), B_1(a_{n-1})) - \lambda F(a_{n-1}, b_{n-1})] \| \\
&= \frac{1}{\mu_1\alpha_1^2 - \gamma_1\beta_1^2} \| [H_1(A_1(a_n), B_1(a_n)) - H_1(A_1(a_{n-1}), B_1(a_{n-1}))] \\
&\quad - \lambda[F(a_n, b_n) - F(a_{n-1}, b_n) + F(a_{n-1}, b_n) - F(a_{n-1}, b_{n-1}] \| \\
&\leq \frac{1}{\mu_1\alpha_1^2 - \gamma_1\beta_1^2} \| [H_1(A_1(a_n), B_1(a_n)) - H_1(A_1(a_{n-1}), B_1(a_{n-1}))] \\
&\quad - \lambda[F(a_n, b_n) - F(a_{n-1}, b_n)] \| \\
&\quad + \frac{\lambda}{\mu_1\alpha_1^2 - \gamma_1\beta_1^2} \| F(a_{n-1}, b_n) - F(a_{n-1}, b_{n-1}) \| .
\end{aligned} \tag{4.3}$$

Further,

$$\begin{aligned}
&\| [H_1(A_1(a_n), B_1(a_n)) - H_1(A_1(a_{n-1}), B_1(a_{n-1}))] - \lambda[F(a_n, b_n) - F(a_{n-1}, b_n)] \|^2 \\
&\quad \leq \| H_1(A_1(a_n), B_1(a_n)) - H_1(A_1(a_{n-1}), B_1(a_{n-1})) \|^2 \\
&\qquad - 2\lambda\langle H_1(A_1(a_n), B_1(a_n)) - H_1(A_1(a_{n-1}), B_1(a_{n-1})), F(a_n, b_n) - F(a_{n-1}, b_n)\rangle \\
&\qquad + \lambda^2 \| F(a_n, b_n) - F(a_{n-1}, b_n) \|^2 .
\end{aligned}$$

Since $H_1(A_1, B_1)$ is r_1-Lipschitz continuous with respect to A_1 and r_2-Lipschitz continuous with respect to B_1, we have

$$\begin{aligned}
\| H_1(A_1(a_n), B_1(a_n)) - H_1(A_1(a_{n-1}), B_1(a_{n-1})) \| &= \| H_1(A_1(a_n), B_1(a_n)) - H_1(A_1(a_{n-1}), B_1(a_n)) \\
&\quad + H_1(A_1(a_{n-1}), B_1(a_n)) - H_1(A_1(a_{n-1}), B_1(a_{n-1})) \| \\
&\leq \| H_1(A_1(a_n), B_1(a_n)) - H_1(A_1(a_{n-1}), B_1(a_n)) \| \\
&\quad + \| H_1(A_1(a_{n-1}), B_1(a_n)) - H_1(A_1(a_{n-1}), B_1(a_{n-1})) \| \\
&\leq r_1 n - a_{n-1} \| + r_2 \| a_n - a_{n-1} \| \\
&= (r_1 + r_2) \| a_n - a_{n-1} \| .
\end{aligned} \tag{4.5}$$

As $F(x, \cdot)$ is strongly monotone with respect to $H_1(A_1, B_1)$, we have

$$\begin{aligned}
&-\langle H_1(A_1(a_n), B_1(a_n)) - H_1(A_1(a_{n-1}), B_1(a_{n-1})), F(a_n, b_n) - F(a_{n-1}, b_n)\rangle \\
&\qquad \leq -m_1 \| a_n - a_{n-1} \|^2 .
\end{aligned} \tag{4.6}$$

Using the τ_1-Lipschitz continuity of $F(\bullet,\bullet)$ with respect to first argument, we obtain

$$\| F(a_n, b_n) - F(a_{n-1}, b_n) \| \leq \tau_1 \| a_n - a_{n-1} \| . \tag{4.7}$$

Combining (4.5)–(4.7) with (4.5), we obtain

$$\begin{aligned} \| [H_1(A_1(a_n), B_1(a_n)) - H_1(A_1(a_{n-1}), B_1(a_{n-1}))] - \lambda[F(a_n, b_n) - F(a_{n-1}, b_n)] \|^2 \\ \leq [(r_1 + r_2)^2 - 2\lambda m_1 + \lambda^2 \tau_1^2] \| a_n - a_{n-1} \|^2, \end{aligned}$$

which implies that

$$\begin{aligned} \| [H_1(A_1(a_n), B_1(a_n)) - H_1(A_1(a_{n-1}), B_1(a_{n-1}))] - \lambda[F(a_n, b_n) - F(a_{n-1}, b_n)] \| \\ \leq \sqrt{(r_1 + r_2)^2 - 2\lambda m_1 + \lambda^2 \tau_1^2} \| a_n - a_{n-1} \| . \end{aligned} \tag{4.8}$$

Also as $F(\bullet,\bullet)$ is τ_2-Lipschitz continuous with respect to second argument, we have

$$\| F(a_{n-1}, b_n) - F(a_{n-1}, b_{n-1}) \| \leq \tau_2 \| b_n - b_{n-1} \| . \tag{4.9}$$

Using (4.8) and (4.9), (4.3) becomes

$$\begin{aligned} \| a_{n+1} - a_n \| \leq \frac{\sqrt{(r_1 + r_2)^2 - 2\lambda m_1 + \lambda^2 \tau_1^2}}{\mu_1 \alpha_1^2 - \gamma_1 \beta_1^2} \| a_n - a_{n-1} \| \\ + \frac{\lambda \tau_2}{\mu_1 \alpha_1^2 - \gamma_1 \beta_1^2} \| b_n - b_{n-1} \| . \end{aligned} \tag{4.10}$$

In a similar way, we estimate

$$\begin{aligned} \| b_{n+1} - b_n \| &= \| R_{\rho,N}^{H_2(\cdot,\cdot)}[H_2(A_2(b_n), B_2(b_n)) - \rho G(a_n, b_n)] \\ &\quad - R_{\rho,N}^{H_2(\cdot,\cdot)}[H_2(A_2(b_{n-1}), B_2(b_{n-1})) - \rho G(a_{n-1}, b_{n-1})] \| \\ &\leq \frac{1}{\mu_2 \alpha_2^2 - \gamma_2 \beta_2^2} \| H_2(A_2(b_n), B_2(b_n)) - H_2(A_2(b_{n-1}), B_2(b_{n-1}))] \\ &\quad - \rho[G(a_n, b_n) - G(a_n, b_{n-1}) + G(a_n, b_{n-1}) - G(a_{n-1}, b_{n-1})] \| \\ &\leq \frac{1}{\mu_2 \alpha_2^2 - \gamma_2 \beta_2^2} \| [H_2(A_2(b_n), B_2(b_n)) - H_2(A_2(b_{n-1}), B_2(b_{n-1}))] \\ &\quad - \rho[G(a_n, b_n) - G(a_n, b_{n-1})] \| \\ &\quad + \frac{\rho}{\mu_2 \alpha_2^2 - \gamma_2 \beta_2^2} \| [G(a_n, b_{n-1}) - G(a_{n-1}, b_{n-1})] \| . \end{aligned} \tag{4.11}$$

Using the same arguments as for (4.8), we have

$$\| H_2(A_2(b_n), B_2(b_n)) - H_2(A_2(b_{n-1}), B_2(b_{n-1}))] - \rho[G(a_n, b_n) - G(a_n, b_{n-1})] \| \leq \sqrt{(r_3 + r_4)^2 - 2\rho m_2 + \rho^2 \tau_2'^2} \| b_n - b_{n-1} \| . \tag{4.12}$$

As $G(\bullet,\bullet)$ is τ_1'-Lipschitz continuous with respect to first argument, we have

$$\| G(a_n, b_{n-1}) - G(a_{n-1}, b_{n-1}) \| \leq \tau_1' \| a_n - a_{n-1} \| . \tag{4.13}$$

Combining (4.12), (4.13) with (4.11), we have

$$\| b_{n+1} - b_n \| \leq \frac{\sqrt{(r_3 + r_4)^2 - 2\rho m_2 + \rho^2 \tau_2'^2}}{\mu_2 \alpha_2^2 - \gamma_2 \beta_2^2} \| b_n - b_{n-1} \| + \frac{\rho \tau_1'}{\mu_2 \alpha_2^2 - \gamma_2 \beta_2^2} \| a_n - a_{n-1} \| . \tag{4.14}$$

Combining (4.10) and (4.14), we have

$$\begin{aligned} &\| a_{n+1} - a_n \| + \| b_{n+1} - b_n \| \\ &\leq \left[\frac{\sqrt{(r_1 + r_2)^2 - 2\lambda m_1 + \lambda^2 \tau_1^2}}{\mu_1 \alpha_1^2 - \gamma_1 \beta_1^2} + \frac{\rho \tau_1'}{\mu_2 \alpha_2^2 - \gamma_2 \beta_2^2} \right] \| a_n - a_{n-1} \| \\ &\quad + \left[\frac{\sqrt{(r_3 + r_4)^2 - 2\rho m_2 + \rho^2 \tau_2'^2}}{\mu_2 \alpha_2^2 - \gamma_2 \beta_2^2} + \frac{\lambda \tau_2}{\mu_1 \alpha_1^2 - \gamma_1 \beta_1^2} \right] \| b_n - b_{n-1} \| \\ &\leq \theta [\| a_n - a_{n-1} \| + \| b_n - b_{n-1} \|], \end{aligned} \tag{4.15}$$

where

$$\theta = \max\{ \frac{\sqrt{(r_1 + r_2)^2 - 2\lambda m_1 + \lambda^2 \tau_1^2}}{\mu_1 \alpha_1^2 - \gamma_1 \beta_1^2} + \frac{\rho \tau_1'}{\mu_2 \alpha_2^2 - \gamma_2 \beta_2^2}, \frac{\sqrt{(r_3 + r_4)^2 - 2\rho m_2 + \rho^2 \tau_2'^2}}{\mu_2 \alpha_2^2 - \gamma_2 \beta_2^2} + \frac{\lambda \tau_2}{\mu_1 \alpha_1^2 - \gamma_1 \beta_1^2} \}.$$

By (4.2), $\theta < 1$ and (4.15) implies that $\{a_n\}$ and $\{b_n\}$ both are Cauchy sequences. Therefore, $\{(a_n, b_n)\}$ converges to a solution (a, b) of problem (4.1). This completes the proof. □

10.5 $H(\cdot,\cdot)$-η-Cocoercive Mapping

In this section, we define $H(\cdot,\cdot)$-η-cocoercive mapping and discuss some of its properties.

Definition 5.1 Let $A, B{:}E \to E$, H, $\eta{:}E \times E \to E$ be the single-valued mappings. Then $M{:}E \to 2^E$ is said to be $H(\cdot,\cdot)$-η-cocoercive mapping with respect to the mappings A and B (or simply $H(\cdot,\cdot)$-η-cocoercive in the sequel), if M is η-cocoercive and $(H(A, B) + \lambda M)(E) = E$, for every $\lambda > 0$.

Following example shows that $H(\cdot,\cdot)$ is η-cocoercive with respect to A with constant $\frac{1}{3}$ and relaxed η-cocoercive with respect to B with constant $\frac{1}{2}$.

Example 5.1 Let us consider $E = \mathbb{R}^2$. Let $A, B : \mathbb{R}^2 \to \mathbb{R}^2$ are defined by

$$A(x_1, x_2) = (x_1, 3x_2), B(y_1, y_2) = (-y_1, -y_1 - y_2), \text{ for all } (x_1, x_2), (y_1, y_2) \in \mathbb{R}^2.$$

Suppose $H(A, B), \eta : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}^2$ are defined as

$$H(Ax, By) = Ax + By,\ \eta(x, y) = x - y, \quad \text{for all } x, y \in \mathbb{R}^2.$$

Then

$$\begin{aligned}\langle H(Ax, u) - H(Ay, u), \eta(x, y)\rangle &= \langle Ax + u - Ay - u, x - y\rangle \\ &= \langle Ax - Ay, x - y\rangle \\ &= \langle((x_1, 3x_2) - (y_1, 3y_2)), (x_1 - y_1, x_2 - y_2)\rangle \\ &= \langle(x_1 - y_1, 3(x_2 - y_2)), (x_1 - y_1, x_2 - y_2)\rangle \\ &= (x_1 - y_1)^2 + 3(x_2 - y_2)^2\end{aligned}$$

and

$$\begin{aligned}\| Ax - Ay \|^2 = \| (x_1 - y_1, 3(x_2 - y_2)) \|^2 &= (x_1 - y_1)^2 + 9(x_2 - y_2)^2 \\ &\leq 3(x_1 - y_1)^2 + 9(x_2 - y_2)^2 \\ &= 3\{(x_1 - y_1)^2 + 3(x_2 - y_2)^2\} \\ &= 3\{\langle H(Ax, u) - H(Ay, u), \eta(x, y)\rangle\}\end{aligned}$$

that is, $\langle H(Ax, u) - H(Ay, u), \eta(x, y)\rangle \geq \frac{1}{3} \| Ax - Ay \|^2$, which implies that H is η-cocoercive with respect to A with constant $\frac{1}{3}$. Also

$$\begin{aligned}\langle H(u, Bx) - H(u, By), \eta(x, y)\rangle &= \langle Bx - By, x - y\rangle \\ &= \langle((-x_1, -x_1 - x_2) - (-y_1, -y_1 - y_2)), (x_1 - y_1, x_2 - y_2)\rangle \\ &= \langle(-(x_1 - y_1), -(x_1 - y_1) - (x_2 - y_2)), (x_1 - y_1, x_2 - y_2)\rangle \\ &= -(x_1 - y_1)^2 - (x_1 - y_1)(x_2 - y_2) - (x_2 - y_2)^2 \\ &= -\{(x_1 - y_1)^2 + (x_1 - y_1)(x_2 - y_2) + (x_2 - y_2)^2\}\end{aligned}$$

and

$$\begin{aligned}\| Bx - By \|^2 &= \| (-(x_1 - y_1), -(x_1 - y_1) - (x_2 - y_2)) \|^2 \\ &= (x_1 - y_1)^2 + ((x_1 - y_1) + (x_2 - y_2))^2 \\ &= (x_1 - y_1)^2 + (x_1 - y_1)^2 + (x_2 - y_2)^2 + 2(x_1 - y_1)(x_2 - y_2) \\ &\leq 2(x_1 - y_1)^2 + 2(x_2 - y_2)^2 + 2(x_1 - y_1)(x_2 - y_2) \\ &= 2\{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_1 - y_1)(x_2 - y_2)\} \\ &= 2\{-\langle H(u, Bx) - H(u, By), \eta(x, y)\rangle\},\end{aligned}$$

that is, $\langle H(u, Bx) - H(u, By), \eta(x, y)\rangle \geq -\frac{1}{2} \| Bx - By \|^2$, which implies that H is relaxed η-cocoercive with respect to B with constant $\frac{1}{2}$.

Example 5.2 Let $E = \mathbb{R}^2$ and A, B, $H(A, B)$, and η are same as in Example 5.1. Suppose that $M{:}E \to 2^E$ is defined by

$$M(x_1, x_2) = (x_1, 0), \text{ forall } (x_1, x_2) \in \mathbb{R}^2.$$

Then it is easy to check that M is η-cocoercive and

$$(H(A, B) + \lambda M)(\mathbb{R}^2) = \mathbb{R}^2, \text{ for all } \lambda > 0,$$

which shows that M is $H(\bullet,\bullet)$-η-cocoercive with respect to A and B.

We have the following Matlab programming for Example 5.1.

Numerical Example 5.3 Let $E = \mathbb{R}^2$ and $A, B : \mathbb{R}^2 \to \mathbb{R}^2$ be defined by

$$\begin{aligned}A(x_1, x_2) = (x_1, 3x_2), B(y_1, y_2) = (-y_1, -y_1 - y_2), \text{ for all } x = (x_1, x_2), y \\ = (y_1, y_2) \in \mathbb{R}^2.\end{aligned}$$

Let $H(A, B), \eta : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}^2$ be defined by

$$H(Ax, By) = Ax + By, \eta(x, y) = x - y, \text{ for all } x, y \in \mathbb{R}^2.$$

$$\begin{aligned}x_1 &= \text{input}('\text{enter the vector} x_1 :');\\ x_2 &= \text{input}('\text{enter the vector} x_2 :');\\ y_1 &= \text{input}('\text{enter the vector} y_1 :');\\ y_2 &= \text{input}('\text{enter the vector} y_2 :');\end{aligned}$$

Assign values:

$$\begin{aligned}T_1 &= x_1; T_2 = 3. * x_2;\\ P_1 &= y_1; P_2 = 3. * y_2;\\ u_1 &= x_1 - y_1; u_2 = x_2 - y_2;\end{aligned}$$

compute $\langle H(Ax, u) - H(Ay, u), \eta(x, y)\rangle = W$, where

$$W_1 = u_1.^2; W_2 = 3. * u_2.^2; W = W_1 + W_2$$

compute square of norm $\| Ax - Ay \|^2 = M_1$, where

$$L_1 = u_1.^2; L_2 = 9. * u_2.^2;$$
$$M_1 = L_1 + L_2; M = (1/3). * M_1$$

then it is easy to check that H is η-cocoercive with respect to A with constant $\frac{1}{3}$, that is, $W \geq M$. Next, compute $\langle H(u, Bx) - H(u, By), \eta(x, y)\rangle = V$.

$$V = -(u_1.^2 + u_1. * 3. * u_2 + 3. * u_2.^2)$$

compute square of norm $\| Bx - By \|^2 = B_1$, where

$$B_1 = (2. * (u_1.^2) + (3. * u_2).^2 + (2. * u_1). * (3. * u_2));$$
$$Z = -(1/2). * B_1$$

then it is easy to check that H is relaxed η-cocoercive with respect to B with constant $\frac{1}{2}$, that is, $V \geq Z$.

Example 5.4 Let $E = C[0, 1]$, space of all real-valued continuous function define over closed interval [0, 1] with the norm

$$\| f \| = \max_{t\in[0,1]} |f(t)|.$$

Let $A, B{:}E \to E$ are defined by

$$A(f) = e^f \text{ and } B(g) = e^{-g}, \quad \text{for all } f, g \in E.$$

Let $H(A, B){:}E \times E \to E$ is defined as

$$H(A(f), B(g)) = A(f) + B(g), \text{ for all } f, g \in E.$$

Suppose that $M = I$, where I is the identity mapping. Then for $\lambda = 1$, we have

$$\| (H(A, B) + M)(f) \| = \| A(f) + B(f) + f \| = \max_{t\in[0,1]} |e^{f(t)} + e^{-f(t)} + f(t)| > 0,$$

which means that $0 \notin (H(A, B) + M)(E)$ and thus M is not $H(\bullet,\bullet)$-η-cocoercive mapping with respect to A and B.

Theorem 5.1 *Let $H(A, B)$ be η-cocoercive with respect to A with constant $\mu > 0$ and relaxed η-cocoercive with respect to B with constant $\gamma > 0$, A is α-expansive*

and B is β-Lipschitz continuous and $\mu > \gamma$, $\alpha > \beta$. Let $M{:}E \to 2^E$ be $H(\bullet,\bullet)$-η-cocoercive mapping. If the following inequality

$$\langle x - y, \mathcal{J}_q(\eta(u,v))\rangle \geq 0,$$

holds for all $(v,y) \in \mathrm{Graph}(M)$, then $x \in Mu$, where $\mathrm{Graph}(M) = \{(u,x) \in E \times E : x \in Mu\}$.

Proof Suppose that there exists some (u_0, x_0) such that

$$\langle x_0 - y, \mathcal{J}_q(\eta(u_0,v))\rangle \geq 0, \quad \text{for all } (v,y) \in \mathrm{Graph}(M). \tag{5.1}$$

Since M is $H(\bullet,\bullet)$-η-cocoercive mapping, we know that $(H(A, B) + \lambda M)(E) = E$, holds for every $\lambda > 0$ and so there exists $(u_1, x_1) \in \mathrm{Graph}(M)$ such that

$$H(Au_1, Bu_1) + \lambda x_1 = H(Au_0, Bu_0) + \lambda x_0 \in E. \tag{5.2}$$

It follows from (5.1) and (5.2) that

$$\begin{aligned}
0 &\leq \lambda\langle x_0 - x_1, \mathcal{J}_q(\eta(u_0,u_1))\rangle \\
&= -\langle H(Au_0,Bu_0) - H(Au_1,Bu_1), \mathcal{J}_q(\eta(u_0,u_1))\rangle \\
&= -\langle H(Au_0,Bu_0) - H(Au_1,Bu_1), \mathcal{J}_q(\eta(u_0,u_1))\rangle \\
&\quad - \langle H(Au_1,Bu_0) - H(Au_1,Bu_1), \mathcal{J}_q(\eta(u_0,u_1))\rangle \\
&\leq -\mu \parallel Au_0 - Au_1 \parallel^q + \gamma \parallel Bu_0 - Bu_1 \parallel^q \\
&\leq -\mu\alpha^q \parallel u_0 - u_1 \parallel^q + \gamma\beta^q \parallel u_0 - u_1 \parallel^q \\
&= -(\mu\alpha^q - \gamma\beta^q) \parallel u_0 - u_1 \parallel^q \leq 0,
\end{aligned}$$

which gives $u_1 = u_0$, since $\mu > \gamma$ and $\alpha > \beta$. By (5.2), we have $x_1 = x_0$. Hence $(u_0, x_0) = (u_1, x_1) \in \mathrm{Graph}(M)$ and so $x_0 \in Mu_0$. □

Theorem 5.2 *Let $H(A, B)$ be η-cocoercive with respect to A with constant $\mu > 0$ and relaxed η-cocoercive with respect to B with constant $\gamma > 0$, A is α-expansive and B is β-Lipschitz continuous, $\mu > \gamma$ and $\alpha > \beta$. Let M be $H(\bullet,\bullet)$-η-cocoercive mapping. Then the operator $(H(A, B) + \lambda M)^{-1}$ is single-valued.*

Definition 5.2 Let $H(A, B)$ be η-cocoercive with respect to A with constant $\mu > 0$ and relaxed η-cocoercive with respect to B with constant $\gamma > 0$, A is α-expansive and B is β-Lipschitz continuous, $\mu > \gamma$ and $\alpha > \beta$. Let M be $H(\bullet,\bullet)$-η-cocoercive mapping. The resolvent operator $R_{\lambda,M}^{H(\cdot,\cdot)-\eta}{:}E \to E$ is defined by

$$R_{\lambda,M}^{H(\cdot,\cdot)-\eta}(u) = (H(A,B) + \lambda M)^{-1}(u), \quad \text{for all } u \in E. \tag{5.3}$$

The following theorem shows that the resolvent operator defined by (5.3) is Lipschitz continuous.

Theorem 5.3 *Let H(A, B) be η-cocoercive with respect to A with constant $\mu > 0$ and relaxed η-cocoercive with respect to B with constant $\gamma > 0$, A is α-expansive, B is β-Lipschitz continuous and η is τ-Lipschitz continuous and $\mu > \gamma$, $\alpha > \beta$. Let M be an H(•,•)-η-cocoercive mapping. Then the resolvent operator $R_{\lambda,M}^{H(\cdot,\cdot)-\eta}$:E → E is $\frac{\tau^{q-1}}{\mu\alpha^q - \gamma\beta^q}$-Lipschitz continuous, that is*

$$\| R_{\lambda,M}^{H(\cdot,\cdot)-\eta}(u) - R_{\lambda,M}^{H(\cdot,\cdot)-\eta}(v) \| \leq \frac{\tau^{q-1}}{\mu\alpha^q - \gamma\beta^q} \| u - v \|, \quad \text{for all } u, v \in E.$$

10.6 Variational-Like Inclusions

In this section, we study a variational-like inclusion problem.

Let $H, N, \eta: E \times E \to E$, $A, B: E \to E$ be the single-valued mappings and $T, Q: E \to CB(E)$ be the set-valued mappings. Let $M: E \to 2^E$ be set-valued, H(•,•)-η-cocoercive mapping. Then, we consider the following problem:

Find $u \in E$, $w \in T(u)$, $v \in Q(u)$ such that

$$0 \in N(w, v) + M(u). \tag{6.1}$$

We call problem (6.1), a variational-like inclusion problem.

If $H(\bullet,\bullet) = H(\cdot)$ and M is H-accretive mapping, then problem (6.1) is introduced and studied by Chang et al. [4]. It is clear that for suitable choices of operators involved in the formulation of (6.1), one can obtain many variational inclusions studied in recent past.

Lemma 6.1 *The (u, w, v), where $u \in E$, $w \in T(u)$, $v \in Q(u)$ is a solution of problem (6.1) if and only if (u, w, v) is a solution of the following equation*:

$$u = R_{\lambda,M}^{H(\cdot,\cdot)-\eta}[H(A(u), B(u)) - \lambda N(w, v)], \tag{6.2}$$

where $\lambda > 0$ is a constant.

Proof Proof is straightforward by the use of definition of resolvent operator.

Based on Lemma 6.1, we define the following algorithm for approximating a solution of variational-like inclusion problem (6.1).

Algorithm 6.1 For any $u_0 \in E$, $w_0 \in T(u_0)$, $v_0 \in Q(u_0)$, compute the sequences $\{u_n\}$, $\{w_n\}$, and $\{v_n\}$ by the following iterative scheme:

$$u_{n+1} = R_{\lambda,M}^{H(\cdot,\cdot)-\eta}[H(A(u_n), B(u_n)) - \lambda N(w_n, v_n)]; \tag{6.3}$$

$$w_n \in T(u_n), \| w_n - w_{n+1} \| \leq \mathcal{H}(T(u_n), T(u_{n+1})); \tag{6.4}$$

$$v_n \in Q(u_n), \| v_n - v_{n+1} \| \leq \mathcal{H}(Q(u_n), Q(u_{n+1})); \tag{6.5}$$

for all $n = 0, 1, 2, \ldots\ldots$ and $\lambda > 0$ is a constant.

Lemma 6.2 [18] *Let E be a real Banach space and $\mathcal{J} : E \to 2^{E^*}$ be a normalized duality mapping. Then for any $x, y \in E$, $j(x + y) \in \mathcal{J}(x + y)$*

1. $\| x+y \|^2 \leq \| x \|^2 + 2\langle y, j(x+y)\rangle$;
2. $\langle x - y, j(x) - j(y)\rangle \leq 2D^2_{\rho E}(4 \| x - y \| /D)$, where $D = \sqrt{\| x \|^2 + \| y \|^2 /2}$

Theorem 6.1 *Let E be a real Banach space. Let $H, N, \eta : E \times E \to E$, $A, B : E \to E$ be the single-valued mappings and $T, Q : E \to CB(E)$ be the set-valued mappings. Let $M : E \to 2^E$ be a set-valued, $H(\bullet,\bullet)$-η-cocoercive mapping. Let*

1. *T is $\mathcal{H}$-Lipschitz continuous with constant λ_T and Q is $\mathcal{H}$-Lipschitz continuous with constant λ_Q;*
2. *$H(A, B)$ is η-cocoercive with respect to A with constant $\mu > 0$ and relaxed η-cocoercive with respect to B with constant $\gamma > 0$;*
3. *A is α-expansive and B is β-Lipschitz continuous;*
4. *$H(A, B)$ is r_1-Lipschitz continuous with respect to A and r_2-Lipschitz continuous with respect to B;*
5. *N is t_1-Lipschitz continuous with respect to T in the first argument and t_2-Lipschitz continuous with respect to Q in the second argument;*
6. *η is τ-Lipschitz continuous;*
7. *N is strongly η-accretive with respect to T in the first argument and strongly η-accretive with respect to Q in the second argument with constants τ_1 and τ_2, respectively.*

Suppose that the following condition is satisfied:

$$\sqrt{r_1^2 + 2\lambda(t_1\lambda_T + t_2\lambda_Q)[r_1 + \lambda(t_1\lambda_T + t_2\lambda_Q) + \tau] - 2\lambda(\tau_1 + \tau_2)} < \frac{\mu\alpha^2 - \gamma\beta^2}{\tau} - r^2,$$
$$\mu\alpha^2 - \gamma\beta^2 > \tau r^2, \mu > \gamma, \alpha > \beta. \tag{6.6}$$

Then there exist $u \in E$, $w \in T(u)$, and $v \in Q(u)$ satisfying the variational-like inclusion problem (6.1) and the iterative sequences $\{u_n\}$, $\{w_n\}$ and $\{v_n\}$ generated by Algorithm 6.1 converge strongly to u, w and v, respectively.

Proof Since T is $\mathcal{H}$-Lipschitz continuous with constant λ_T and Q is $\mathcal{H}$-Lipschitz continuous with constant λ_Q, it follows from Algorithm 6.1 that

$$\| w_n - w_{n+1} \| \leq \mathcal{H}(T(u_n), T(u_{n+1})) \leq \lambda_T \| u_n - u_{n+1} \|, \tag{6.7}$$

and

$$\| v_n - v_{n+1} \| \leq H(Q(u_n), Q(u_{n+1})) \leq \lambda_Q \| u_n - u_{n+1} \| . \tag{6.8}$$

By using Algorithm 6.1 and Lipschitz continuity of resolvent operator $R_{\lambda,M}^{H(\cdot,\cdot)-\eta}$, we have

$$\begin{aligned}
\| u_{n+1} - u_n \| =& \| R_{\lambda,M}^{H(\cdot,\cdot)-\eta}[H(Au_n, Bu_n) - \lambda N(w_n, v_n)] \\
& - R_{\lambda,M}^{H(\cdot,\cdot)-\eta}[H(Au_{n-1}, Bu_{n-1}) - \lambda N(w_{n-1}, v_{n-1})] \| \\
\leq & \frac{\tau}{\mu\alpha^2 - \gamma\beta^2} \| H(Au_n, Bu_n) - H(Au_{n-1}, Bu_{n-1}) \\
& - \lambda[N(w_n, v_n) - N(w_{n-1}, v_{n-1})] \| \\
\leq & \frac{\tau}{\mu\alpha^2 - \gamma\beta^2} \| H(Au_n, Bu_n) - H(Au_{n-1}, Bu_n) \\
& - \lambda[N(w_n, v_n) - N(w_{n-1}, v_{n-1})] \| \\
& + \frac{\tau}{\mu\alpha^2 - \gamma\beta^2} \| H(Au_{n-1}, Bu_n) - H(Au_{n-1}, Bu_{n-1}) \| .
\end{aligned} \tag{6.9}$$

Using Lemma 6.2, we have

$$\begin{aligned}
& \| H(Au_n, Bu_n) - H(Au_{n-1}, Bu_n) - \lambda[N(w_n, v_n) - N(w_{n-1}, v_{n-1})] \|^2 \\
\leq & \| H(Au_n, Bu_n) - H(Au_{n-1}, Bu_n) \|^2 - 2\lambda\langle N(w_n, v_n) - N(w_{n-1}, v_{n-1}), \\
& j[H(Au_n, Bu_n) - H(Au_{n-1}, Bu_n) - \lambda(N(w_n, v_n) - N(w_{n-1}, v_{n-1}))]\rangle \\
= & \| H(Au_n, Bu_n) - H(Au_{n-1}, Bu_n) \|^2 - 2\lambda\langle N(w_n, v_n) - N(w_{n-1}, v_{n-1}), \\
& j[H(Au_n, Bu_n) - H(Au_{n-1}, Bu_n) - \lambda(N(w_n, v_n) - N(w_{n-1}, v_{n-1}))] \\
& - j(\eta(u_n, u_{n-1}))\rangle - 2\lambda\langle N(w_n, v_n) - N(w_{n-1}, v_{n-1}), j(\eta(u_n, u_{n-1}))\rangle \\
\leq & \| H(Au_n, Bu_n) - H(Au_{n-1}, Bu_n) \|^2 + 2\lambda \| N(w_n, v_n) - N(w_{n-1}, v_{n-1}) \| \\
& \times \| H(Au_n, Bu_n) - H(Au_{n-1}, Bu_n) \| + \lambda \| N(w_n, v_n) - N(w_{n-1}, v_{n-1}) \| \\
& + \| \eta(u_n, u_{n-1}) \|] - 2\lambda\langle N(w_n, v_n) - N(w_{n-1}, v_{n-1}), j(\eta(u_n, u_{n-1}))\rangle .
\end{aligned} \tag{6.10}$$

As $H(\bullet,\bullet)$ is r_1-Lipschitz continuous with respect to A, we have

$$\| H(Au_n, Bu_n) - H(Au_{n-1}, Bu_n) \| \leq r_1 \| u_n - u_{n-1} \| . \tag{6.11}$$

Since N is t_1-Lipschitz continuous with respect to T with respect to first argument and t_2-Lipschitz continuous with respect to Q with respect to the second argument and T is λ_T-Lipschitz continuous and Q is λ_Q-Lipschitz continuous, we have

$$\begin{aligned}
&\| N(w_n, v_n) - N(w_{n-1}, v_{n-1}) \| \\
&\quad = \| N(w_n, v_n) - N(w_{n-1}, v_n) + N(w_{n-1}, v_n) - N(w_{n-1}, v_{n-1}) \| \\
&\quad \leq \| N(w_n, v_n) - N(w_{n-1}, v_n) \| + \| N(w_{n-1}, v_n) - N(w_{n-1}, v_{n-1}) \| \\
&\quad \leq t_1 \| w_n - w_{n-1} \| + t_2 \| v_n - v_{n-1} \| \\
&\quad \leq t_1 \mathcal{H}(T(u_n), T(u_{n-1})) + t_2 \mathcal{H}(Q(u_n), Q(u_{n-1})) \\
&\quad \leq t_1 \lambda_T \| u_n - u_{n-1} \| + t_2 \lambda_Q \| u_n - u_{n-1} \| \\
&\quad = (t_1 \lambda_T + t_2 \lambda_Q) \| u_n - u_{n-1} \| .
\end{aligned} \tag{6.12}$$

As η is τ-Lipschitz continuous, we have

$$\| \eta(u_n, u_{n-1}) \| \leq \tau \| u_n - u_{n-1} \| . \tag{6.13}$$

Since N is strongly η-accretive with respect to T and strongly η-accretive with respect to Q in the first and second arguments with constants τ_1 and τ_2, respectively, we have

$$\begin{aligned}
&\langle N(w_n, v_n) - N(w_{n-1}, v_{n-1}), j(\eta(u_n, u_{n-1})) \rangle \\
&\quad = \langle N(w_n, v_n) - N(w_{n-1}, v_n), j(\eta(u_n, u_{n-1})) \rangle \\
&\qquad + \langle N(w_{n-1}, v_n) - N(w_{n-1}, v_{n-1}), j(\eta(u_n, u_{n-1})) \rangle \\
&\quad \geq \tau_1 \| u_n - u_{n-1} \|^2 + \tau_2 \| u_n - u_{n-1} \|^2 \\
&\quad \geq (\tau_1 + \tau_2) \| u_n - u_{n-1} \|^2 .
\end{aligned} \tag{6.14}$$

Using (6.11)–(6.14), (6.10) becomes

$$\begin{aligned}
&\| H(Au_n, Bu_n) - H(Au_{n-1}, Bu_{n-1}) - \lambda[N(w_n, v_n) - N(w_{n-1}, v_{n-1})] \|^2 \\
&\quad \leq r_1^2 \| u_n - u_{n-1} \|^2 + 2\lambda(t_1 \lambda_T + t_2 \lambda_Q) \| u_n - u_{n-1} \| [r_1 \| u_n - u_{n-1} \| \\
&\qquad + \lambda(t_1 \lambda_T + t_2 \lambda_Q) \| u_n - u_{n-1} \| + \tau \| u_n - u_{n-1} \|] - 2\lambda((\tau_1 + \tau_2)) \| u_n - u_{n-1} \|^2 \\
&\quad = r_1^2 \| u_n - u_{n-1} \|^2 + 2\lambda(t_1 \lambda_T + t_2 \lambda_Q) \| u_n - u_{n-1} \| \times \\
&\quad [\{r_1 + \lambda(t_1 \lambda_T + t_2 \lambda_Q) + \tau\} \| u_n - u_{n-1} \|] - 2\lambda(\tau_1 + \tau_2) \| u_n - u_{n-1} \|^2 \\
&\quad = r_1^2 \| u_n - u_{n-1} \|^2 + 2\lambda(t_1 \lambda_T + t_2 \lambda_Q)[r_1 + \lambda(t_1 \lambda_T + t_2 \lambda_Q) \\
&\qquad + \tau] \| u_n - u_{n-1} \|^2 - 2\lambda(\tau_1 + \tau_2) \| u_n - u_{n-1} \|^2 \\
&\quad = [r_1^2 + 2\lambda(t_1 \lambda_T + t_2 \lambda_Q)[r_1 + \lambda(t_1 \lambda_T + t_2 \lambda_Q) + \tau] \\
&\qquad - 2\lambda(\tau_1 + \tau_2)] \| u_n - u_{n-1} \|^2 .
\end{aligned} \tag{6.15}$$

Using r_2-Lipschitz continuity of $H(\bullet, \bullet)$ with respect to B and (6.15), (6.9) becomes

$$\| u_{n+1} - u_n \| \leq \theta \| u_n - u_{n-1} \|, \tag{6.16}$$

where

$$\theta = \frac{\tau}{\mu\alpha^2 - \gamma\beta^2}\sqrt{\theta_1} + \frac{\tau r_2}{\mu\alpha^2 - \gamma\beta^2},$$

and

$$\theta_1 = [r_1^2 + 2\lambda(t_1\lambda_T + t_2\lambda_Q)[r_1 + \lambda(t_1\lambda_T + t_2\lambda_Q) + \tau] - 2\lambda(\tau_1 + \tau_2)]. \tag{6.17}$$

From (6.6), it follows that $\theta < 1$, so $\{u_n\}$ is a Cauchy sequence in E, thus, there exists a $u \in E$ such that $u_n \to u$ as $n \to \infty$. Also from (6.7) and (6.8), it follows that $\{w_n\}$ and $\{v_n\}$ are also Cauchy sequences in E, thus there exist w and v in E such that $w_n \to w$, $v_n \to v$ as $n \to \infty$. By the continuity of $R_{\lambda,M}^{H(\cdot,\cdot)-\eta}$, H, A, B, η, N, T, and Q, it follows from Algorithm 6.1 that

$$u = R_{\lambda,M}^{H(\cdot,\cdot)-\eta}[H(A(gu), B(gu)) - \lambda N(w, v)].$$

Now, we prove that $w \in T(u)$. In fact, since $w_n \in T(u_n)$, we have

$$\begin{aligned} d(w, T(u)) &\leq \| w - w_n \| + d(w_n, T(u)) \\ &\leq \| w - w_n \| + \mathcal{H}(T(u_n), T(u)) \\ &\leq \| w - w_n \| + \lambda_T \| u_n - u \| \to 0, \text{ as } n \to \infty, \end{aligned}$$

which means that $d(w, T(u)) = 0$. Since $T(u) \in CB(E)$, it follows that $w \in T(u)$. Similarly, we can show that $v \in Q(u)$. By Lemma 6.1, we conclude that (u, w, v) is a solution of variational-like inclusion problem (6.1). This completes the proof. □

10.7 Generalized Variational-like Inclusions

In this section, we solve a generalized variational-like inclusion problem. We take E to be q-uniformly smooth Banach space throughout this section.

Let $H, N, W, \eta: E \times E \to E$, $A, B, g: E \to E$ be the single-valued mappings and T, Q, R, $S: E \to CB(E)$ be the set-valued mappings. Let $M: E \to 2^E$ be set-valued, $H(\bullet, \bullet)$-η-cocoercive mapping. Then, we consider the following problem:

Find $u \in E$, $x \in T(u)$, $y \in Q(u)$, $z \in R(u)$, $v \in S(u)$ such that

$$0 \in N(x, y) - W(z, v) + M(g(u)). \tag{7.1}$$

Problem (7.1) is called generalized variational-like inclusion problem.

Lemma 7.1 *(u, x, y, z, v), where $u \in E$, $x \in T(u)$, $y \in Q(u)$, $z \in R(u)$, $v \in S(u)$ is a solution of problem (7.1) if and only if (u, x, y, z, v) is the solution of the following equation*:

$$g(u) = R_{\lambda,M}^{H(\cdot,\cdot)-\eta}[H(A(gu), B(gu)) - \lambda\{N(x,y) - W(z,v)\}], \tag{7.2}$$

where $\lambda > 0$ is a constant.

Proof Proof is a direct consequence of definition of resolvent operator.□

Based on (7.2), we have the following iterative algorithm.

Algorithm 7.1 For any given $u_0 \in E$, $x_0 \in T(u_0)$, $y_0 \in Q(u_0)$, $z_0 \in R(u_0)$, $v_0 \in S(u_0)$, compute the sequences $\{u_n\}$, $\{x_n\}$, $\{y_n\}$, $\{z_n\}$, and $\{v_n\}$ by the following iterative procedure:

$$g(u_{n+1}) = R_{\lambda,M}^{H(\cdot,\cdot)-\eta}[H(A(gu_n), B(g(u_n)) - \lambda\{N(x_n, y_n) - W(z_n, u_n)\}]; \tag{7.3}$$

$$\| x_{n+1} - x_n \| \leq \mathcal{H}(T(u_{n+1}), T(u_n)); \tag{7.4}$$

$$\| y_{n+1} - y_n \| \leq \mathcal{H}(Q(u_{n+1}), Q(u_n)); \tag{7.5}$$

$$\| z_{n+1} - z_n \| \leq \mathcal{H}(R(u_{n+1}), R(u_n)); \tag{7.6}$$

$$\| v_{n+1} - v_n \| \leq \mathcal{H}(S(u_{n+1}), S(u_n)); \tag{7.7}$$

where $n = 0, 1, 2, \ldots\ldots,$ and $\lambda > 0$ is a constant.

Lemma 7.2 [19] *Let E be a real uniformly smooth Banach space. Then E is q-uniformly smooth if and only if there exists a constant $C_q > 0$ such that, for all x, $y \in E$,*

$$\| x+y \|^q \leq \| x \|^q + q\langle y, \mathcal{J}_q(x)\rangle + C_q \| y \|^q .$$

Theorem 7.1 *Let E be a q-uniformly smooth Banach space. Let A, B, $g{:}E \to E$, H, N, W, $\eta{:}E \times E \to E$ be the single-valued mappings. Let T, Q, R, $S{:}E \to CB(E)$ be the set-valued mappings and $M{:}E \to 2^E$ be the set-valued, $H(\bullet,\bullet)$-η-cocoercive mapping. Suppose that*

1. g is δ-strongly accretive and λ_g-Lipschitz continuous;
2. N is Lipschitz continuous with respect to first argument with constant λ_{N_1} and Lipschitz continuous with respect to second argument with constant λ_{N_2}, strongly η-accretive with respect to T and Q with constant t;
3. W is Lipschitz continuous with respect to the first argument with constant λ_{W_1} and Lipschitz continuous with respect to second argument with constant λ_{W_2};
4. η is τ-Lipschitz continuous, A is α-expansive and B is β-Lipschitz continuous;
5. $H(A, B)$ is η-cocoercive with respect to A with constant $\mu > 0$ and relaxed η-cocoercive with respect to B with constant $\gamma > 0$, r_1-Lipschitz continuous with respect to A and r_2-Lipschitz continuous with respect to B;
6. T, Q, R, S are $\mathcal{H}$-Lipschitz continuous mappings with constants λ_T, λ_Q, λ_R and λ_S, respectively.

Suppose that the following conditions are satisfied:

$$\left[\sqrt[q]{(r_1+r_2)^q\lambda_g^q - q\lambda t + q\lambda(\lambda_{N_1}\lambda_T + \lambda_{N_2}\lambda_Q)[(r_1+r_2)^{q-1}\lambda_g^{q-1} + \tau^{q-1}]}\right] + \lambda^q C_q(\lambda_{N_1}\lambda_T + \lambda_{N_2}\lambda_Q)^q < \left[\frac{\delta}{\tau^q}(\mu\alpha^q - \gamma\beta^q) - \lambda(\lambda_{W_1}\lambda_R + \lambda_{W_2}\lambda_S)\right], \tag{7.8}$$

$$\frac{\delta}{\tau^q}(\mu\alpha^q - \gamma\beta^q) > \lambda(\lambda_{W_1}\lambda_R + \lambda_{W_2}\lambda_S), \mu > \gamma, \alpha > \beta.$$

Then there exist $u \in E$, $x \in T(u)$, $y \in Q(u)$, $z \in R(u)$, and $v \in S(u)$ satisfying the generalized variational-like inclusion problem (7.1) and the iterative sequences $\{u_n\}$, $\{x_n\}$, $\{y_n\}$, $\{z_n\}$ and $\{v_n\}$ generated by Algorithm 7.1 converge strongly to u, x, y, z, and v, respectively.

Proof Since g is δ-strongly accretive, we have

$$\begin{aligned} \| g(u_{n+1}) - g(u_n) \| \| u_{n+1} - u_n \|^{q-1} &\geq \langle g(u_{n+1}) - g(u_n), \mathcal{J}_q(u_{n+1} - u_n)\rangle \\ &\geq \delta \| u_{n+1} - u_n \|^q . \end{aligned} \tag{7.9}$$

From (7.9), we get

$$\| u_{n+1} - u_n \| \leq \frac{1}{\delta} \| g(u_{n+1}) - g(u_n) \| . \tag{7.10}$$

By Algorithm 7.1 and Theorem 5.3, we have

$$\begin{aligned} \| g(u_{n+1}) - g(u_n) \| &= \| R_{\lambda,M}^{H(\cdot,\cdot)-\eta}[H(A(gu_n), B(gu_n)) - \lambda\{N(x_n, y_n) - W(z_n, v_n)\}] \\ &\quad - R_{\lambda,M}^{H(\cdot,\cdot)-\eta}[H(A(gu_{n-1}), B(gu_{n-1})) - \lambda\{N(x_{n-1}, y_{n-1}) - W(z_{n-1}, v_{n-1})\}] \| \\ &\leq \frac{\tau^{q-1}}{\mu\alpha^q - \gamma\beta^q} \| H(A(gu_n), B(gu_n)) - H(A(gu_{n-1}), B(gu_{n-1})) \\ &\quad - \lambda\{N(x_n, y_n) - N(x_{n-1}, y_{n-1})\} \\ &\quad - \lambda\{W(z_n, v_n) - W(z_{n-1}, v_{n-1})\} \| \\ &\leq \frac{\tau^{q-1}}{\mu\alpha^q - \gamma\beta^q} \| H(A(gu_n), B(gu_n)) - H(A(gu_{n-1}), B(gu_{n-1})) \\ &\quad - \lambda\{N(x_n, y_n) - N(x_{n-1}, y_{n-1})\} \| \\ &\quad + \frac{\tau^{q-1}\lambda}{\mu\alpha^q - \gamma\beta^q} \| W(z_n, v_n) - W(z_{n-1}, v_{n-1}) \| . \end{aligned} \tag{7.11}$$

Using Lipschitz continuity of N with constant λ_{N_1} with respect to first argument and λ_{N_2} with respect to second argument and $\mathcal{H}$-Lipschitz continuity of T and Q with constants λ_T and λ_Q, respectively, we have

$$\begin{aligned}\| N(x_n, y_n) - N(x_{n-1}, y_{n-1}) \| &= \| N(x_n, y_n) - N(x_{n-1}, y_n) \\ &\quad + N(x_{n-1}, y_n) - N(x_{n-1}, y_{n-1}) \| \\ &\leq \| N(x_n, y_n) - N(x_{n-1}, y_n) \| \\ &\quad + \| N(x_{n-1}, y_n) - N(x_{n-1}, y_{n-1}) \| \\ &\leq \lambda_{N_1} \| x_n - x_{n-1} \| + \lambda_{N_2} \| y_n - y_{n-1} \| \\ &\leq \lambda_{N_1} \mathcal{H}(T(u_n), T(u_{n-1})) + \lambda_{N_2} \mathcal{H}(Q(u_n), Q(u_{n-1})) \\ &\leq \lambda_{N_1} \lambda_T \| u_n - u_{n-1} \| + \lambda_{N_2} \lambda_Q \| u_n - u_{n-1} \| \\ &= (\lambda_{N_1} \lambda_T + \lambda_{N_2} \lambda_Q) \| u_n - u_{n-1} \| . \end{aligned} \tag{7.12}$$

Also, as $H(A, B)$ is r_1-Lipschitz continuous with respect to A and r_2-Lipschitz continuous with respect to B and g is λ_g-Lipschitz continuous, we have

$$\| H(A(gu_n), B(gu_n)) - H(A(gu_{n-1}), B(gu_{n-1})) \| \leq (r_1 + r_2)\lambda_g \| u_n - u_{n-1} \| . \tag{7.13}$$

By using Lemma 7.2, (7.12), (7.13) and strong η-accretivity of N with respect to T and Q with constant t and τ-Lipschitz continuity of η, we have

$$\begin{aligned}&\| H(A(gu_n), B(gu_n)) - H(A(gu_{n-1}), B(gu_{n-1})) - \lambda\{N(x_n, y_n) - N(x_{n-1}, y_{n-1})\} \|^q \\ &\quad \leq (A(gu_n), B(gu_n)) - H(A(gu_{n-1}), B(g(u_{n-1})) \|^q \\ &\qquad - q\lambda\langle N(x_n, y_n) - N(x_{n-1}, y_{n-1}), \mathcal{J}_q(\eta(u_n, u_{n-1}))\rangle \\ &\qquad - q\lambda\langle N(x_n, y_n) - N(x_{n-1}, y_{n-1}), \mathcal{J}_q[H(A(gu_n), B(gu_n)) \\ &\qquad - H(A(gu_{n-1}), B(gu_{n-1}))] - \mathcal{J}_q(\eta(u_n, u_{n-1}))\rangle \\ &\qquad + \lambda^q C_q(x_n, y_n) - N(x_{n-1}, y_{n-1}) \|^q \\ &\quad \leq (r_1 + r_2)^q \lambda_g^q \| u_n - u_{n-1} \|^q - q\lambda t \| u_n - u_{n-1} \|^q + q\lambda \| N(x_n, y_n) - N(x_{n-1}, y_{n-1}) \| \\ &\qquad \times [\| H(A(gu_n) -, B(gu_n)) - H(A(gu_{n-1}), B(gu_{n-1})) \|^{q-1} + \| \eta(u_n, u_{n-1}) \|^{q-1}] \\ &\qquad + \lambda^q C_q(\lambda_{N_1} \lambda_T + \lambda_{N_2})^q \| u_n - u_{n-1} \|^q \\ &\quad \leq (r_1 + r_2)^q \lambda_g^q \| u_n - u_{n-1} \|^q - q\lambda t \| u_n - u_{n-1} \|^q + q\lambda(\lambda_{N_1} \lambda_T + \lambda_{N_2} \lambda_Q) \| u_n - u_{n-1} \| \\ &\qquad \times [(r_1 + r_2)^{q-1} \lambda_g^{q-1} \| u_n - u_{n-1} \|^{q-1} + \tau^{q-1} \| u_n - u_{n-1} \|^{q-1}] \\ &\qquad + \lambda^q C_q(\lambda_{N_1} \lambda_T + \lambda_{N_2} \lambda_Q)_n^q - u_{n-1} \|^q \\ &\quad = [(r_1 + r_2)^q \lambda_g^q - q\lambda t + q\lambda(\lambda_{N_1} \lambda_T + \lambda_{N_2} \lambda_Q)[(r_1 + r_2)^{q-1} \lambda_g^{q-1} + \tau^{q-1}] \\ &\qquad + \lambda^q C_q(\lambda_{N_1} \lambda_T + \lambda_{N_2} \lambda_Q)^q] \| u_n - u_{n-1} \|^q . \end{aligned} \tag{7.14}$$

Using Lipschitz continuity of W with respect to first argument with constant λ_{W_1} and with respect to second argument with constant λ_{W_2} and $\mathcal{H}$-Lipschitz continuity of R and S with constants λ_R and λ_S, respectively, we obtain

$$\| W(z_n, v_n) - W(z_{n-1}, v_{n-1}) \| \leq (\lambda_{W_1} \lambda_R + \lambda_{W_2} \lambda_S) \| u_n - u_{n-1} \| . \tag{7.15}$$

In view of (7.14) and (7.15), (7.11) becomes

$$\| g(u_n) - g(u_{n-1}) \| \leq [\frac{\tau^{q-1}}{\mu\alpha^q - \gamma\beta^q} (\sqrt[q]{(r_1+r_2)^q\lambda_g^q - q\lambda t + q\lambda(\lambda_{N_1}\lambda_T + \lambda_{N_2}\lambda_Q)} \overline{[(r_1+r_2)^{q-1}\lambda_g^{q-1} + \tau^{q-1}] + \lambda^q C_q(\lambda_{N_1}\lambda_T + \lambda_{N_2}\lambda_Q)^q}) + \frac{\tau^{q-1}\lambda}{\mu\alpha^q - \gamma\beta^q}(\lambda_{W_1}\lambda_R + \lambda_{W_2}\lambda_S)] \| u_n - u_{n-1} \| . \tag{7.16}$$

Using (7.16), (7.10) becomes

$$\| u_{n+1} - u_n \| \leq \theta \| u_n - u_{n-1} \|, \tag{7.17}$$

where

$$\theta = \frac{1}{\delta}[\frac{\tau^{q-1}}{\mu\alpha^q - \gamma\beta^q} (\sqrt[q]{(r_1+r_2)^q\lambda_g^q - q\lambda t + q\lambda(\lambda_{N_1}\lambda_T + \lambda_{N_2}\lambda_Q)} \overline{[(r_1+r_2)^{q-1}\lambda_g^{q-1} + \tau^{q-1}] + \lambda^q C_q(\lambda_{N_1}\lambda_T + \lambda_{N_2}\lambda_Q)^q}) + \frac{\tau^{q-1}\lambda}{\mu\alpha^q - \gamma\beta^q}(\lambda_{W_1}\lambda_R + \lambda_{W_2}\lambda_S)].$$

By condition (7.8), $\theta < 1$ and hence $\{u_n\}$ is Cauchy sequence in E, so there exists $u \in E$ such that $u_n \to u$ as $n \to \infty$. Using (7.4)–(7.7) of Algorithm 7.1 and $\mathcal{H}$-Lipschitz continuity of T, Q, R and S with constants λ_T, λ_Q, λ_R and λ_S, respectively, we have

$$\begin{aligned}
&\| x_{n+1} - x_n \| \leq \mathcal{H}(T(u_{n+1}), T(u_n)) \leq \lambda_T \| u_{n+1} - u_n \|;\\
&\| y_{n+1} - y_n \| \leq \mathcal{H}(Q(u_{n+1}), Q(u_n)) \leq \lambda_Q \| u_{n+1} - u_n \|;\\
&\| z_{n+1} - z_n \| \leq \mathcal{H}(R(u_{n+1}), R(u_n)) \leq \lambda_R \| u_{n+1} - u_n \|;\\
&\| v_{n+1} - n_n \| \leq \mathcal{H}(S(u_{n+1}), S(u_n)) \leq \lambda_S \| u_{n+1} - u_n \|;
\end{aligned}$$

which shows that the sequences $\{x_n\}$, $\{y_n\}$, $\{z_n\}$ and $\{v_n\}$ are all Cauchy sequences in E, so there exist x, y, z, and $v \in E$ such that $x_n \to x$, $y_n \to y$, $z_n \to z$ and $v_n \to v$, as $n \to \infty$. By continuity of mappings H, A, B, N, W, $R_{\lambda,M}^{H(\cdot,\cdot)-\eta}$ and Algorithm 7.1, it follows that

$$g(u) = R_{\lambda,M}^{H(\cdot,\cdot)-\eta}[H(A(gu), B(gu)) - \lambda\{N(x,y) - W(z,v)\}].$$

It remain to show that $x \in T(u)$. In fact, since $x_n \in T(u_n)$. we have

$$\begin{aligned} d(x, T(u)) &\leq \| x - x_n \| + d(x_n, T(u)) \\ &\leq \| x - x_n \| + \mathcal{H}(T(u_n), T(u)) \\ &\leq \| x - x_n \| + \lambda_T \| u_n - u \| \to 0, \text{ as } n \to \infty, \end{aligned}$$

which implies that $d(x, T(u)) = 0$, since $T(u) \in CB(E)$, it follows that $x \in T(u)$. Similarly, we can show that $y \in Q(u)$, $z \in R(u)$ and $v \in S(u)$. This completes the proof. □

References

1. Fang YP, Huang NJ (2004) H-monotone operators and systems of variational inclusions. Comm Appl Nonlinear Anal 11(1):93–101
2. Verma RU (2005) A-monotonicity and applications to nonlinear variational inclusion problems. J Appl Math Stochastic Anal 17(2):193–195
3. Zhu D, Marcotte P (1996) Cocoercivity and its role in the convergence of iterative schemes for solving variational inequalities. SIAM J Optim 6(3):714–726
4. Chang SS, Cho YJ, Lee BS, Jung IH (2000) Generalized set-valued variational inclusions in Banach spaces. J Math Anal Appl 246:409–422
5. Fang YP, Cho YJ, Kim JK (2004) (H, η)-accretive operator and approximating solutions for systems of variational inclusions in Banach spaces. Reprint
6. Fang YP, Huang NJ (2004) Accretive operators and resolvent operator technique for solving variational inclusions in Banach spaces. Appl Math Lett 17:647–653
7. Fang YP, Huang NJ (2005) Iterative algorithm for a system of variational inclusions involving H-accretive operators in Banach spaces. Acta Math Hungar 108:183–195
8. Fang YP, Huang NJ (2003) Approximate solutions for nonlinear operator inclusions with (H, η)-monotone operators. Research report, Sichuan University
9. Hassouni A, Moudafi A (1994) A perturbed algorithm for variational inclusions. J Math Anal Appl 185:706–712
10. Huang NJ, Fang YP (2001) Generalized m-acceretive mapping in Banach spaces. J Sichuan Univ 38(4):591–592
11. Nadler SB Jr (1969) Multivalued contraction mappings. Pacific J Math 30:475–488
12. Verma RU (2004) Generalized system of relaxed cocoercive variational inequalities and projection method. J Optim Theory Appl 12(1):203–210
13. Tseng P (1990) Further applications of splitting algorithm to decomposition in variational inequalities and convex programming. Math Prog 48:249–264
14. Magananti TL, Perakis G (1993) Convergence conditions for variational inequality algorithms. Massachusetts Institute of Technology, OR, pp. 282–293
15. Zou YZ, Huang NJ (2008) H(•,•)-accretive operator with an application for solving variational inclusions in Banach spaces. Appl Math Comput 204:809–816
16. Huang NJ (2001) A new class of generalized set-valued implicit variational inclusions in Banach spaces with an application. Comput Math Appl 41:937–943
17. Fang YP, Huang NJ, Thompson HB (2005) A new system of variational inclusions with (H, η)-monotone operators in Hilbert spaces. Comput Math Appl 49:365–374
18. Petryshyn WV (1997) A characterization of strictly convexity of Banach spaces and other uses of duality mappings. J Funct Anal 6:282–291
19. Xu HK (1991) Inequalities in Banach spaces with application. Nonlinear Anal 16(12):1127–1138

Chapter 11
Wavelet and Fractal Methods with Environmental Applications

Bhardwaj Rashmi

Abstract In this paper, the Wavelet and Fractal Methods with environmental applications were discussed. Fractal dimension is a ratio providing a statistical index of complexity comparing how detail in a pattern (strictly speaking, a fractal pattern) changes with the scale at which it is measured. It has also been characterized as a measure of the space-filling capacity of a pattern that tells how a fractal scales different from the space it is embedded in; a fractal dimension does not have to be an integer. Hurst exponent is a numerical estimate for predictability of time series. It is defined as relative tendency of time series to either regress for longer term mean value or 'cluster' in direction. It is related to fractal dimension, which gives measure for roughness of surface. Predictability increases when fractal dimension becomes less than 1.5 or more than 1.5. In the former case, persistence behavior is observed, while in the latter, an anti-persistence. If one of these indices comes close to 0, then corresponding process approximates usual Brownian motion and is therefore unpredictable. If it becomes close to 1, process is said to be predictable. In environmental sciences, these methods are applied for studying the behavior of air pollutants and water pollutants. It is observed that each of the air pollution parameters CO, NO, NO_2, O_3, and SO_2 at each monitoring station follows an anti-persistent behavior. Also, each of the water-quality parameters COD, BOD, DO, WT, AMM, TKN, TC, FC, and pH follows Brownian motion and thus behavior is unpredictable. It is concluded that Brownian time series behavior exists for air and water pollutants.

Keywords Fractal dimension · Predictability index · Hurst exponent · Brownian motion · Air and water pollutants

B. Rashmi (✉)
University School of Basic and Applied Sciences, Non-Linear Dynamics Research Lab, Guru Gobind Singh Indraprastha University, Dwarka 110078, Delhi, India
e-mail: rashmib22@gmail.com

A.H. Siddiqi et al. (eds.), *Mathematical Models, Methods and Applications*, Industrial and Applied Mathematics, DOI 10.1007/978-981-287-973-8_11

11.1 Introduction

The problem of air and water pollution is increasing tremendously day by day in all the metropolitan cities by exponential increase in vehicles, emission from industries, and unplanned urbanization. Therefore, evaluation of a suitable method for predicting and monitoring the pollution is very important. Air and water pollutions in Delhi city draw an attention of NGO's Environmentalists, Researchers and Government.

Delhi, the Capital of India, is largest metropolis by area and the second-largest metropolis by population in India. It is located at 28° 22′ 48″ N and 77° 7′ 12″ E. Its area is 1,484 km^2 (573 sq mi) of which 783 km^2 (302 sq mi) is rural and 700 km^2 (270 sq mi) is urban. Its maximum length is 51.9 km (32 mi) and maximum width is 48.48 km (30 mi). The population of Delhi is 1, 27, 09, 458 (as per Census 2001). It is drained by river Yamuna. Yamuna river accounts for more than 70 % of Delhi's water supplies and about 57 million people depend on river water for their daily usage [6]. Nizamuddin is approximately 14 km downstream from Wazirabad barrage at Delhi and 410 km from Yamunotri. Pollution in river water is continuously increasing due to urbanization, industrialization, population growth, etc. Many rivers are dying due to pollution which is an alarming signal.

The rapid population growth along with the high rate of urbanization as also industrialization and an increase in motorized transport has resulted in an increase in the levels of various air pollutants, namely (1) oxides of sulfur, (2) oxides of nitrogen, (3) suspended particulate matter, (4) respirable suspended particulate matter, (5) carbon monoxide, (6) lead, (7) ozone, (8) benzene, and (9) hydrocarbons. Vehicles, thermal power plants, and large- and small-scale industrial units in Delhi were the major sources of these pollutants.

Rangarajan et al. [12] discussed the wavelet-based analysis of meteorological parameters using daily mean value of pressure, temperature, relative humidity, wind speed data, and applications of wavelet and fractal methods to meteorological problems. Siddiqi et al. [13] discussed the wavelet-based Hurst exponent and fractal dimension analysis of Saudi climatic dynamics. Siddiqi [14] discussed in detail the wavelet and fractal methods in Science and Engineering. Nunnari [10] modeled air pollution time series using wavelet functions and genetic algorithms. Vela'squez et al. [16] discussed the spatial variability of the Hurst exponent for the daily rainfall series in the state of Zacatecas Mexio. Cannistraro and Ponterio [4] discussed the analysis of air quality in the outdoor environment of the city of Messina by an application of the pollution index method. Continuous wavelet and wavelet transform in time and frequency domain have been considered to analyze air pollution parameters such as SO_2 and smoke concentration as discussed by Can et al. [3]. Carbone et al. [5] calculated Hurst exponent of several time series by dynamical implementation of a recently proposed detrending moving average scaling technique.

The water quality at Nizamuddin (Delhi) has the impact of industrial, sewerage, and domestic discharge from Haryana and Delhi [7, 8]. Regression equations can be

used to estimate constituent concentrations. Constituent concentrations can be used by water-quality managers for comparison of current water-quality conditions to water-quality standards. Examination of stream flow and physical properties of water that act as surrogates for constituents of interest also helps for collection of water-quality samples [9, 11, 15]. Fractal dimension and predictability analysis are used to predict the behavior of water-quality parameters [1, 2]. It has been observed that regional climatic models would not be able to predict local climate as it deals with averaged quantities and that precipitation during the south-west monsoon is affected by temperature and pressure variability during the preceding winter. Time series can be modeled by a stochastic process possessing long-range correlation.

The analysis of major air pollutants such as carbon monoxide (CO), nitrogen oxide (NO), nitrogen dioxide (NO_2), ozone (O_3), and sulfur dioxide (SO_2) recorded at each of the different locations at Delhi College of Engineering (Industrial area), ITO-Crossing(Commercial area), Siri fort (Residential area) and by mobile van in Delhi. Data, monitored by mobile van throughout the city, can be treated as average pollution of a mix location. Today, we are living in a water-starved world. Water is an essential element for life. Most of countries fulfill the requirement of water from river water and ground water. Pollution in river water draws attention of government, public, NGOs, and environmentalists in India and world over. Water-quality parameters such as COD (Chemical Oxygen Demand), BOD (Biochemical Oxygen Demand), DO (Dissolved Oxygen), WT (Water Temperature), AMM (Free Ammonia), TKN (Total Kjeldahl Nitrogen), TC (Total Coliform), FC (Fecal Coliform), and PH (Potential of Hydrogen) monitored at Nizamuddin bridge-mid stream (Delhi) of Yamuna River in India for last 10 years have been used.

This paper deals with the estimation of Hurst exponent, fractal dimension, and predictability index using wavelet method for air and water pollutants measured by Central Pollution Control Board (CPCB). The daily averaged values of each air pollutant at different locations of Delhi for a period of 4 years (August 2006–July 2010) have been considered for the study. Also, the monthly average values of each water-quality parameter monitored at Nizamudin mid-bridge stream of Yamuna river for a period of 10 years have been considered for the study.

11.2 Methodology

This paper deals with the analysis of air pollutants and water pollutants through Hurst exponent, predictability index, and fractal dimension using wavelet method.

11.2.1 Hurst Exponent (H)

It refers to the index of dependence. It quantifies the relative tendency of a time series either to regress strongly to the mean or to cluster in a direction. The values

of the Hurst exponent range between 0 and 1. A value of 0.5 indicates a true random walk (a Brownian time series). In a random walk, there is no correlation between any element and a future element. A Hurst exponent value H, $0.5 < H < 1$, indicates "persistent behavior" (a positive autocorrelation). If there is an increase from time step t_{i-1} to t_i, there will probably be an increase from t_i to t_{i+1}. The same is true for decrease, where a decrease will tend to follow a decrease. A Hurst exponent value, H $0 < H < 0.5$, will exist for a time series with "anti-persistent behavior" (or negative autocorrelation). Here, an increase will tend to be followed by a decrease or decrease will be followed by an increase. This behavior is sometimes called "mean reversion."

$$H = \left| \frac{b_{yx} - 1}{2} \right|$$

Also, Hurst exponent can be calculated using power-law decay:

$$p(k) = Ck^{-\alpha}$$

where C is a constant and $p(k)$ is the autocorrelation function with lag k. The Hurst exponent is related to the exponent alpha in the equation by the relation

$$H = 1 - \frac{\alpha}{2}$$

11.2.2 Fractal Dimension (D)

It is a statistical quantity that gives an indication of how completely a fractal appears to fill space, as one zooms down to finer and finer scales.

$$D = 2 - H$$

Also fractal dimension is calculated from the Haussdorf dimension. The Haussdorf dimension D_H, in a metric space, is defined as

$$D_H = -\lim_{\varepsilon \to 0} \frac{\ln[N(\varepsilon)]}{\ln \varepsilon}$$

where $N(\varepsilon)$ is the number of open balls of a radius ε needed to cover the entire set. An open ball with center P and radius ε, in a metric space with metric d, is defined as set of all points x such that $\mathrm{d}(P, x) < \varepsilon$.

11.2.3 Predictability Index (PI)

It describes the behavior of time series:

$$\mathrm{PI} = 2|D - 1.5|$$

PI value increases when D value becomes less than or greater than 1.5. In the former case, persistence behavior is observed, while in the latter, an anti-persistence. If one of these indices comes close to 0, then the corresponding process approximates the Brownian motion and is therefore unpredictable.

11.3 Results and Discussion

Hurst exponent, fractal dimension, and predictability indices have been used to estimate the air and water pollution levels in Delhi.

11.3.1 Air Pollutants

This paper deals with the analysis of major air pollutants such as CO, NO, NO_2, O_3, and SO_2, recorded at Delhi College of Engineering (DCE), ITO-Crossing, Siri fort, and by mobile van in Delhi. DCE is the Industrial area of the city, ITO-Crossing is the Commercial area, and Siri fort is the residential area. Data are also monitored by mobile van throughout the city which can be treated as average pollution of a mix location. Daily averaged values for each air pollutant mentioned above are used for a period of last 4 years from August 2006 to July 2010 which have been considered for the study of Hurst exponent, predictability index, and fractal dimension.

11.3.2 Time Series of Air Pollution Parameters

The daily averaged values of CO, NO, NO_2, O_3, and SO_2 are recorded at each of DCE, ITO, Siri fort, and by mobile van in Delhi India. DCE is the Industrial area in the city, ITO-Crossing is the Commercial area, and Siri fort is the residential area. Data are also monitored by mobile van throughout the city which can be treated as average pollution of a mix location (Figs. 11.1, 11.2, 11.3 and 11.4).

Table 11.1 gives the estimation of Hurst exponent of each air pollutant using wavelet method calculated at different locations in Delhi. Table 11.2 predicts the fractal dimension (D) for each of the air pollutants such as CO, NO, NO_2, O_3, and SO_2 monitored at DCE, ITO, Siri fort, and by mobile van in Delhi from the Hurst

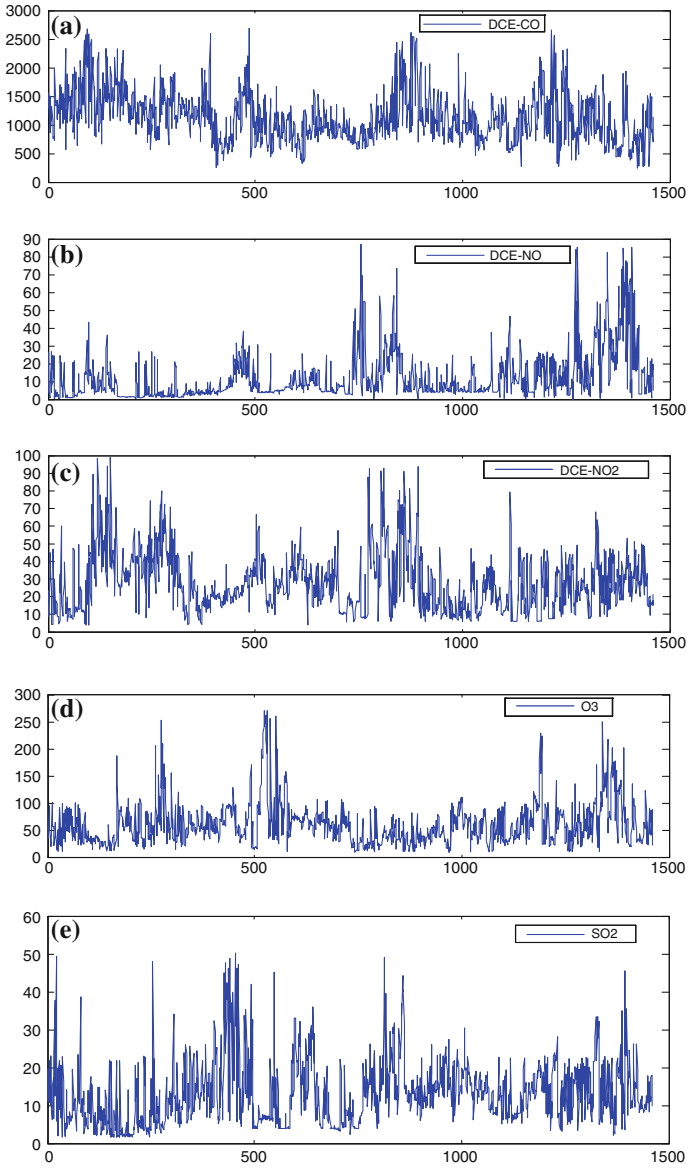

Fig. 11.1 **a–e** Represent the time series of CO, NO, NO_2, O_3, and SO_2 monitored at DCE, Delhi

exponent using formula $D = 2 - H$. Table 11.3 shows the predictability index for each of the air pollutants such as CO, NO, NO_2, O_3, and SO_2 monitored at DCE, ITO, Siri fort, and by mobile van in Delhi.

From Tables 11.1, 11.2, and 11.3, it is observed that for

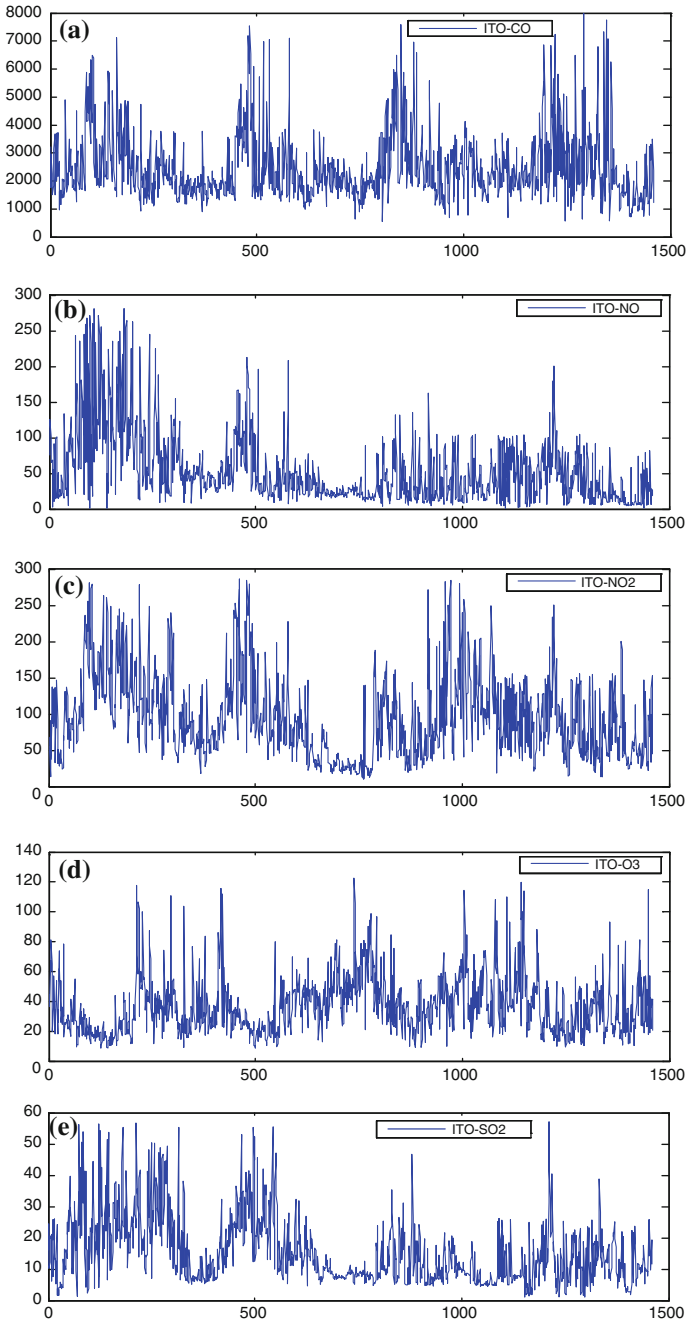

Fig. 11.2 **a–e** Represent the time series of CO, NO, NO_2, O_3, and SO_2 monitored at ITO, Delhi

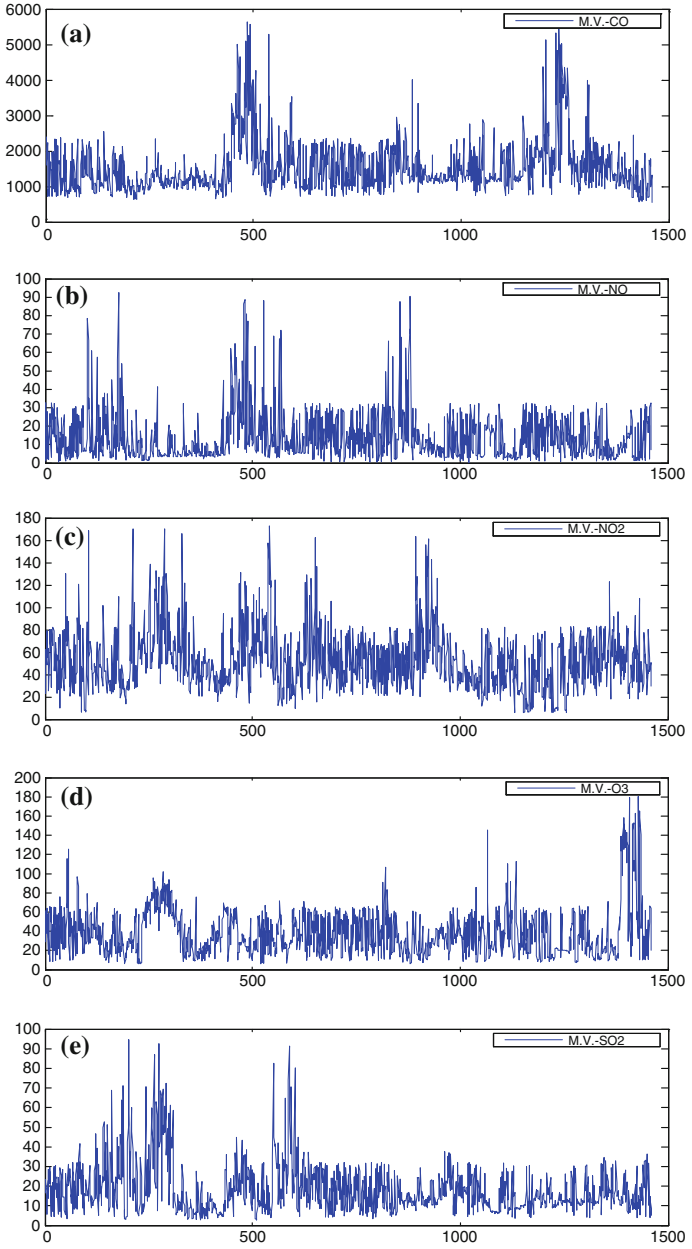

Fig. 11.3 **a–e** Represent the time series of CO, NO, NO_2, O_3, and SO_2 monitored by mobile van, Delhi

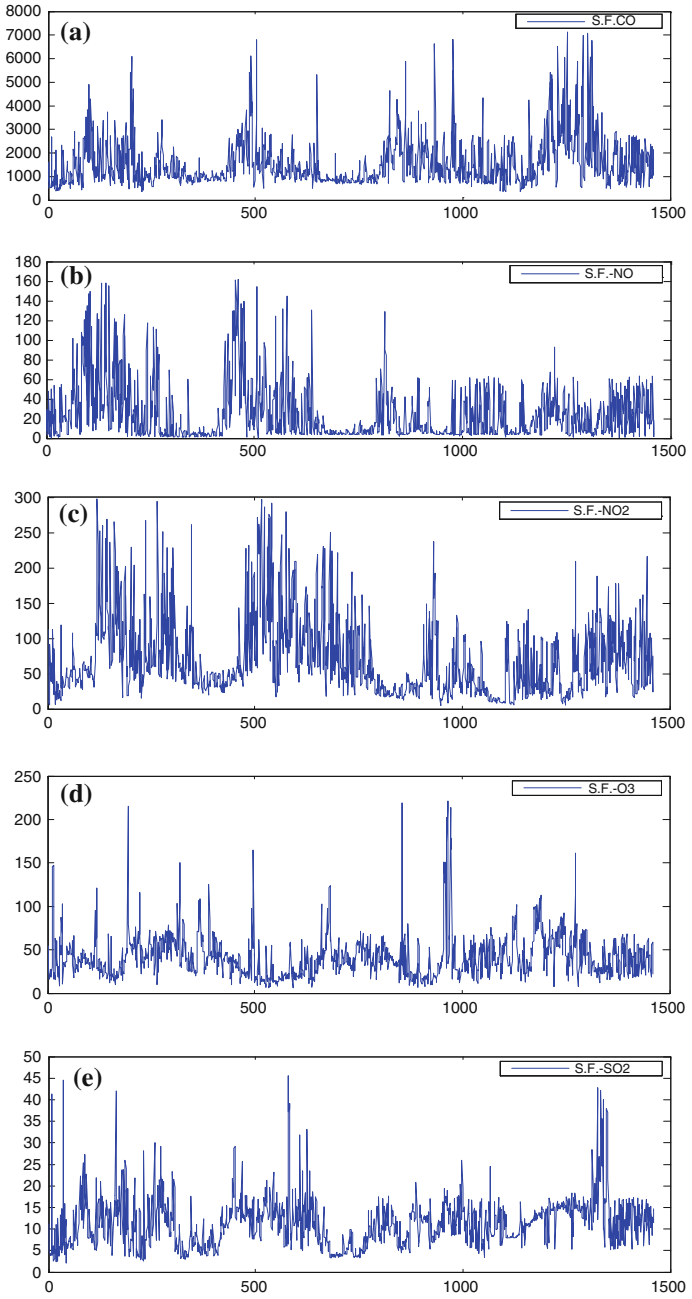

Fig. 11.4 **a–e** Represent the time series of CO, NO, NO_2, O_3, and SO_2 monitored at Siri fort, Delhi

Table 11.1 Hurst exponents for each air pollutant monitored at Delhi College of engineering, ITO-Crossing, Siri fort, and by mobile van in Delhi, India

Values of Hurst exponent (H)				
	DCE	ITO	SF	MV
CO	0.170	0.222	0.190	0.065
NO	0.216	0.272	0.284	0.230
NO_2	0.269	0.314	0.033	0.046
O_3	0.139	0.223	0.206	0.217
SO_2	0.112	0.187	0.165	0.118

Table 11.2 Fractal dimension for each air pollutant monitored at Delhi College of engineering, ITO-Crossing, Siri fort, and by mobile van in Delhi, India

Fractal dimension $D = 2 - H$				
	DCE	ITO	SF	MV
CO	1.830	1.778	1.810	1.935
NO	1.784	1.728	1.716	1.770
NO_2	1.731	1.686	1.967	1.954
O_3	1.861	1.777	1.794	1.783
SO_2	1.888	1.813	1.835	1.882

Table 11.3 Predictability index for each air pollutant monitored at Delhi College of engineering, ITO-Crossing, Siri fort, and by mobile van in Delhi, India

Predictability indices				
	DCE	ITO	SF	MV
CO	0.33	0.278	0.31	0.435
NO	0.284	0.228	0.216	0.27
NO_2	0.231	0.186	0.467	0.454
O_3	0.361	0.277	0.294	0.283
SO_2	0.388	0.313	0.335	0.382

CO Hurst exponent value at all locations lies between 0 and 0.5, and thus has an anti-persistent behavior. Fractal dimension value at all locations lies between 1.5 and 2, which shows an anti-persistent behavior. Also the value of predictability index at all locations lies between 0 and 0.5, and confirms the anti-persistent behavior. It follows Brownian motion and thus the future is unpredictable.

NO Hurst exponent value at all locations is less than 0.5, and thus has an anti-persistent behavior. The values of fractal dimension are greater than 1.5, which show an anti-persistent behavior. Also the value of predictability index at all locations lies between 0 and 0.5, and confirms the anti-persistent behavior. It follows Brownian motion and thus the future is unpredictable.

NO_2 Hurst exponent value at all locations is less than 0.5, and thus has anti-persistent behavior. The values of fractal dimension are greater than 1.5, which show an anti-persistent behavior. Also the value of predictability index at all locations lies between 0 and 0.5, and confirms the anti-persistent behavior. It follows Brownian motion and thus the future is unpredictable.

O_3 Hurst exponent value at all locations lies between 0 and 0.5, and thus has an anti-persistent behavior. Fractal dimension value at all locations lies between 1.5

and 2, which shows an anti-persistent behavior. Also the value of predictability index at all locations lies between 0 and 0.5, and confirms the anti-persistent behavior. It follows Brownian motion and thus the future is unpredictable.
SO_2 Hurst exponent value at all locations lies between 0 and 0.5, and thus has an anti-persistent behavior. Fractal dimension value at all locations lies between 1.5 and 2, which shows an anti-persistent behavior. Also the value of predictability index at all locations lies between 0 and 0.5, and confirms the anti-persistent behavior. It follows Brownian motion and thus the future is unpredictable.

It is observed that Hurst exponent value for each air pollutant recorded at each monitoring station is less than 0.5, which shows an anti-persistent behavior, i.e., an increase in time tends to a decrease in parameter value and vice versa. Fractal dimension value for each air pollutant recorded at each monitoring station lies between 1.5 and 2 which indicate that time series of air pollutant are more jagged than random. Predictability indices for each air pollutant at each location are less than 0.5 which indicate for the existence of usual Brownian motion. Therefore, it can be concluded that each air pollutant has and anti-persistent behavior, and therefore future trend is unpredictable.

11.3.3 Water Pollutants

The monthly average value of last 10 years of water-quality parameters pH (Potential of Hydrogen), COD (Chemical Oxygen Demand), BOD (Biochemical Oxygen Demand), AMM (Free Ammonia), TKN (Total Kjeldahl Nitrogen), DO (Dissolved Oxygen), and WT (Water Temperature) monitored at Nizamuddin bridge-mid stream of Yamuna river in Delhi (India) has been considered for the study of Hurst exponent, predictability index, and fractal dimension. Daubechies wavelet at level 5 (Db_5) is used to get the finer approximation and decomposition. One-dimensional discrete wavelet analysis of water-quality parameters such as pH, BOD, COD, DO, AMM, TKN, WT, TC, and FC for Yamuna river at Nizamuddin bridge-mid Stream, Delhi (India) has been discussed. Db_5 wavelet decomposition of each data was presented in seven parts namely s, a_5, d_1, d_2, d_3, d_4, and d_5 where "s" represents signal or raw data; low-frequency part "a_5" gives an approximate of signal at level 5; and high-frequency parts d_1, d_2, d_3, d_4, and d_5 contains the detail of "s" at different levels, respectively.

Discrete Daubechies wavelets at level 5 (Db_5) for each water-quality parameter have been plotted in Figs. 11.5, 11.6, 11.7, 11.8, 11.9, 11.10, 11.11, 11.12 and 11.13.

Table 11.4 gives the details of s, a_5, d_1, d_2, d_3, d_4, and d_5 using one-dimensional discrete wavelet analysis of water-quality parameters such as pH, BOD, COD, DO, AMM, TKN, WT, TC, and FC for Yamuna river at Nizamuddin bridge-mid Stream, Delhi (India).

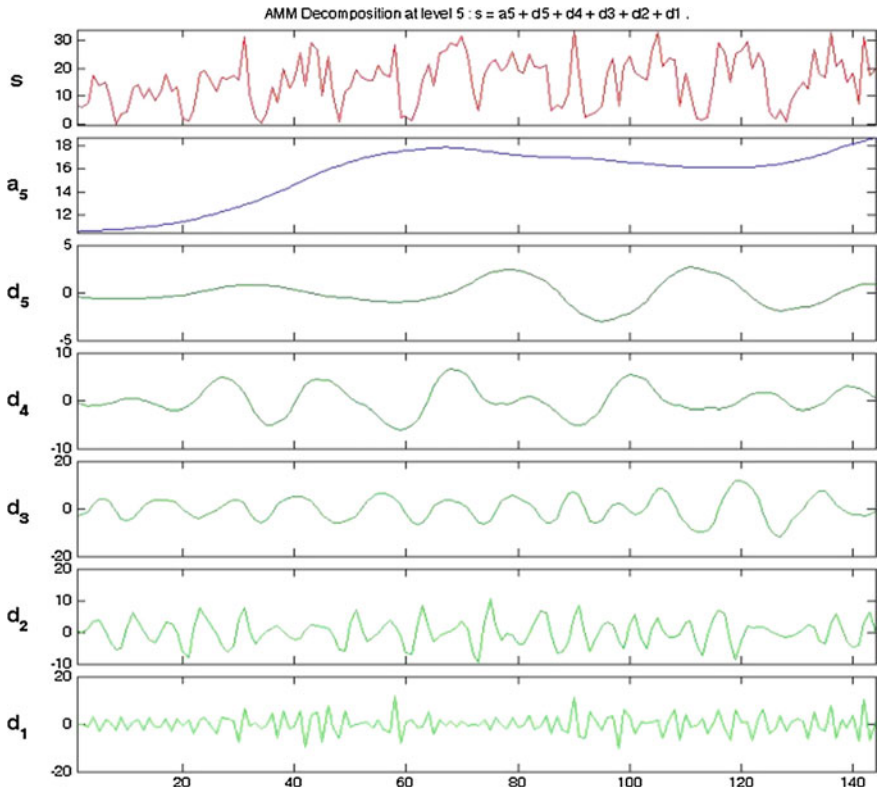

Fig. 11.5 1D discrete wavelet analysis of AMM

Table 11.5 gives the regression equation, Hurst exponent, fractal dimension, and predictability index for each water-quality parameter monitored at Nizamudin bridge-mid stream of Yamuna river.

Using Db_5 analysis and fractal dimension at Nizamuddin, Delhi (India), it is observed that:

AMM The first part of Fig. 11.5 shows that maximum value of AMM is 30 and lower frequency approximation at level 5 (a_5) varies from 10 to 18. Hurst exponent value and predictability index lies between 0 and 0.5, and thus has an anti-persistent behavior. Fractal dimension value lies between 1.5 and 2, and thus shows an anti-persistent behavior. Thus the future is unpredictable.

BOD Figure 11.6 explains BOD parameter's signal and gives maximum value as 50 and lower frequency approximation at level 5 (a_5) of this signal varies from 10 to 30. Hurst exponent and predictability index value lies between 0 and 0.5, and thus has an anti-persistent behavior. Fractal dimension value lies between 1.5 and 2, and thus shows an anti-persistent behavior. Thus the future is unpredictable.

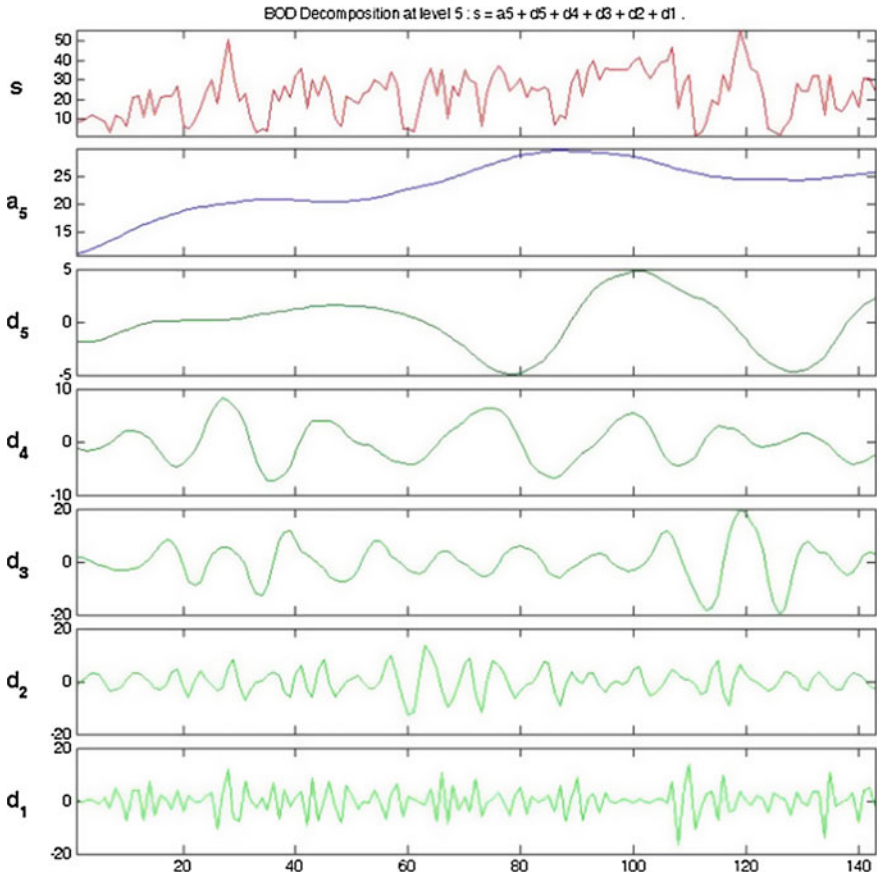

Fig. 11.6 1D discrete wavelet analysis of BOD

COD Figure 11.7 explains COD parameter's signal and gives maximum value as 120 and lower frequency approximation at level 5 (a_5) varies from 54 to 76. Hurst exponent value and predictability index lies between 0 and 0.5, and thus has an anti-persistent behavior. Fractal dimension value lies between 1.5 and 2, and thus shows an anti-persistent behavior. Thus the future is unpredictable.
FC Figure 11.9 explains FC parameter's signal and gives maximum value as 20 and lower frequency approximation at level 5 (a_5) varies from 0 to 14.
DO Figure 11.8 describes DO parameter's signal and gives maximum value as 7 and lower frequency approximation at level 5 (a_5) varies from 0.2 to 1.4. Hurst exponent and predictability index value lies close to 0.5, and thus has Brownian motion. Fractal dimension value is close to 1.5 which shows Brownian motion. Thus future is unpredictable and has Brownian motion.

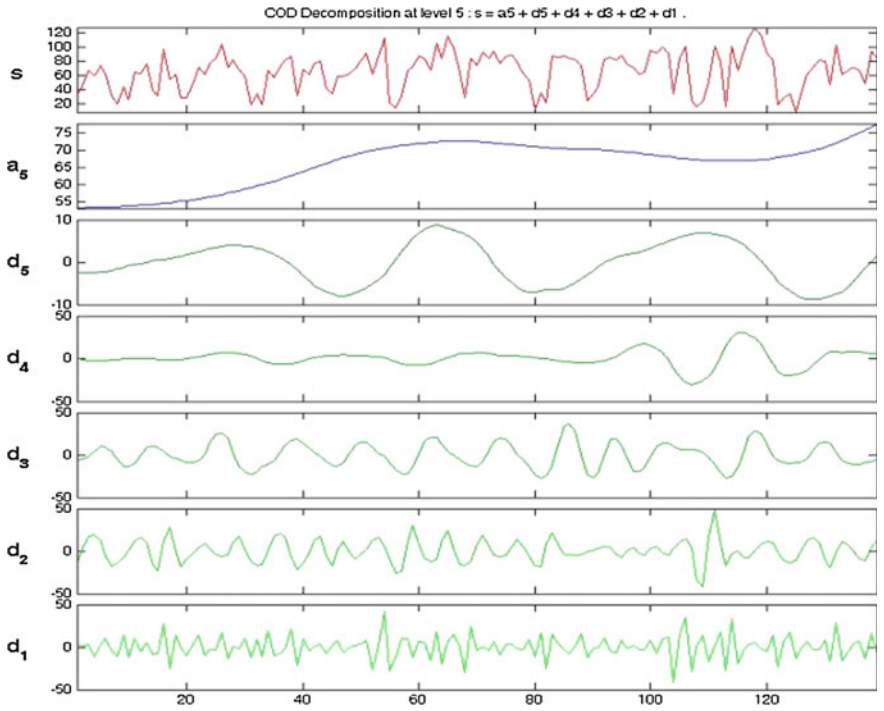

Fig. 11.7 1D discrete wavelet analysis of COD

pH Figure 11.10 depicts pH parameter's signal and gives maximum value as 8.5 and lower frequency approximation at level 5 (a_5) varies from 7.3 to 7.7. Hurst exponent and predictability index value lies close to 0.5, and thus has Brownian motion. Fractal dimension value is close to 1.5 which shows Brownian motion. It follows Brownian motion and future is unpredictable.
TC First part of Fig. 11.11 explains TC parameter's signal and gives maximum value as 8 and lower frequency approximation at level 5 (a_5) varies from 0 to 10.
TKN Figure 11.12 explains TKN parameter's signal and gives maximum value as 40 and lower frequency approximation at level 5 (a_5) varies from 18 to 26. Hurst exponent value and predictability index lies between 0 and 0.5, and thus has an anti-persistent behavior. Fractal dimension value lies between 1.5 and 2, and thus shows an anti-persistent behavior. Thus the future is unpredictable.
WT Figure 11.13 explains WT parameter's signal and gives maximum value as 35 and lower frequency approximation at level 5 (a_5) varies from 24 to 27. Hurst exponent and predictability index value lie close to 0.5, and thus has Brownian motion. Fractal dimension value is close to 1.5 which shows Brownian motion. It follows Brownian motion and future is unpredictable.

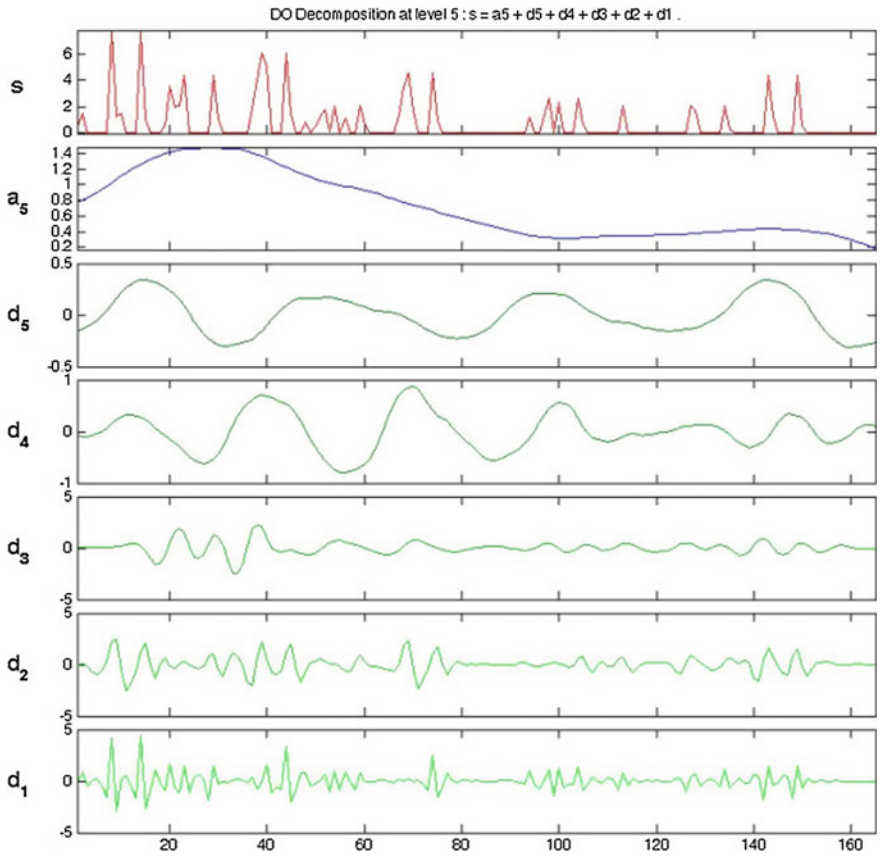

Fig. 11.8 1D discrete wavelet analysis of DO

It is observed that Hurst exponent value for water pollutants COD, BOD, AMM, and TKN is less than 0.5 which shows an anti-persistent behavior, i.e., an increase in time tends to a decrease in parameter value and vice versa. Also for pH, DO, and WT, value is close to 0.5 which shows Brownian behavior. Fractal dimension values for COD, BOD, AMM, and TKN lie between 1.5 and 2, which indicate that time series of water pollutant are more jagged than random. Also for pH, DO, and WT, value is close to 0.5 which shows Brownian behavior. Predictability index for each water parameter is close to 0.5 which indicate for the existence of usual Brownian motion. Thus it can be concluded that pH, DO, and WT follow the Brownian time series behavior (true random walk) and parameters COD, BOD, AMM, and TKN follow an anti-persistence behavior (or negative correlation) and the future trend is unpredictable.

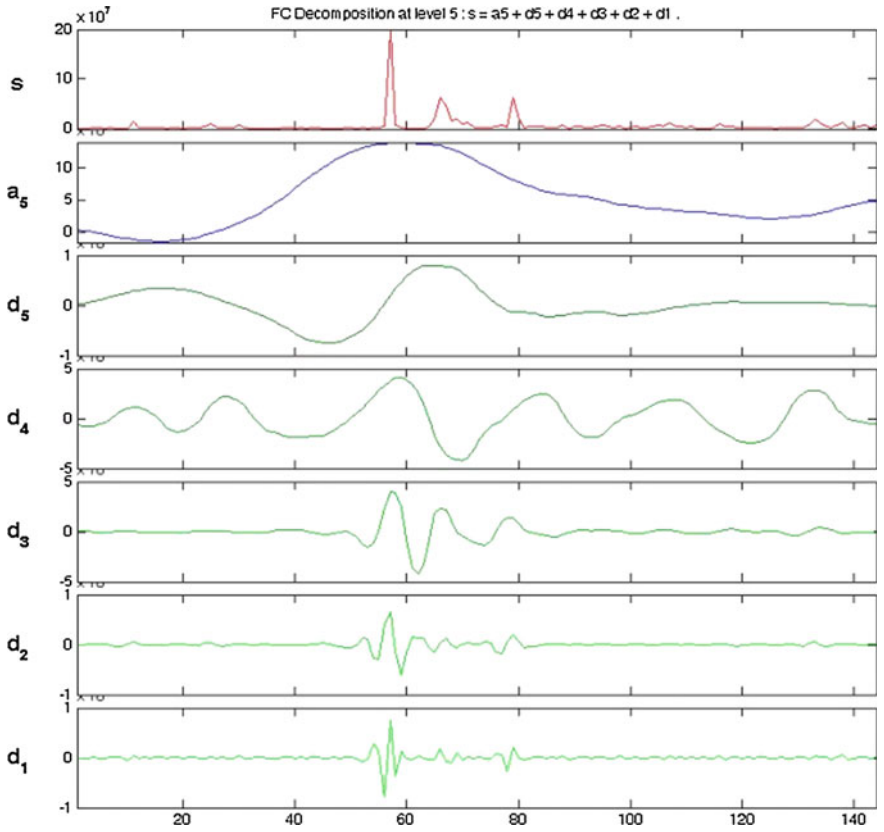

Fig. 11.9 1D discrete wavelet analysis of FC

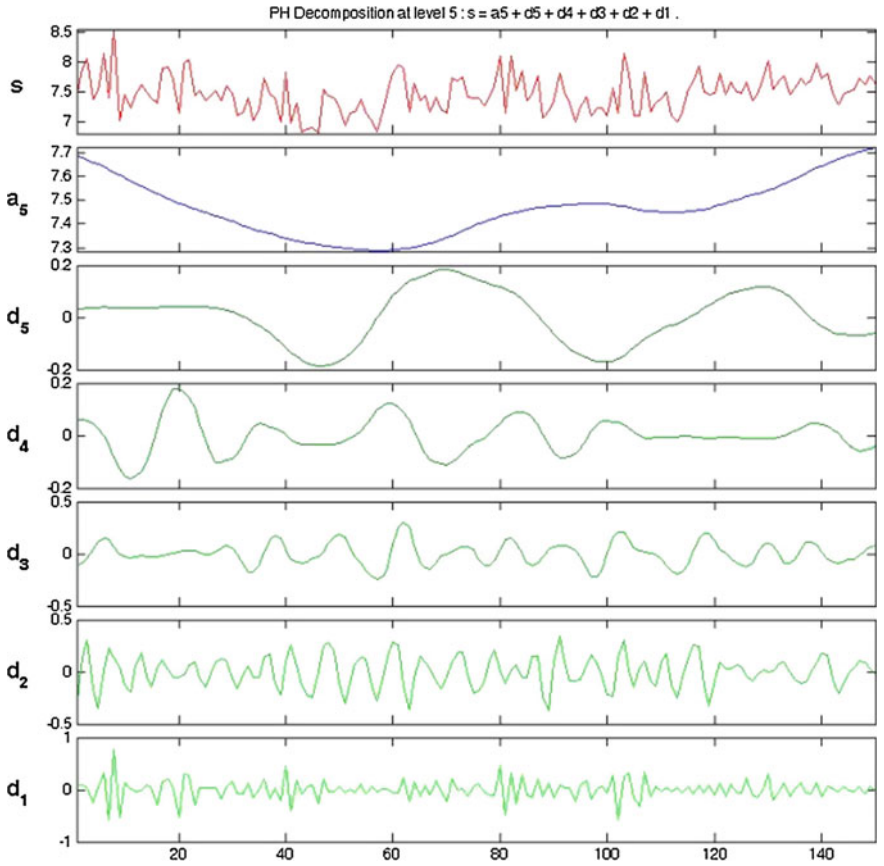

Fig. 11.10 1D discrete wavelet analysis of pH

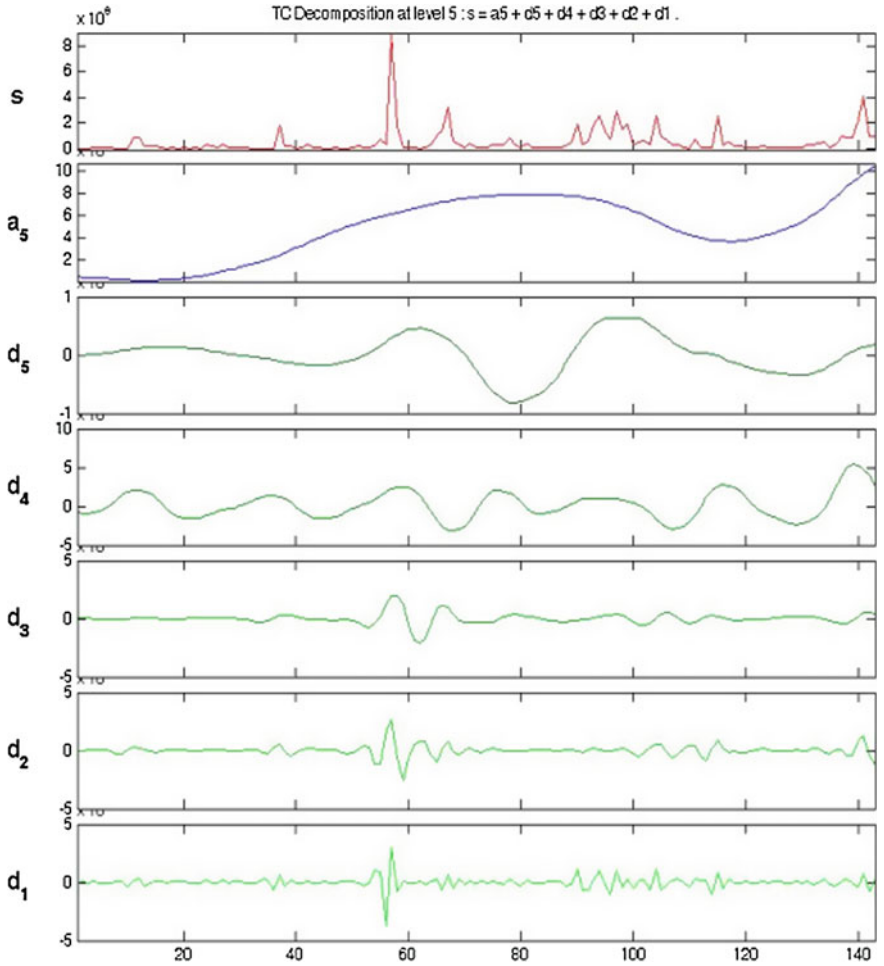

Fig. 11.11 1D discrete wavelet analysis of TC

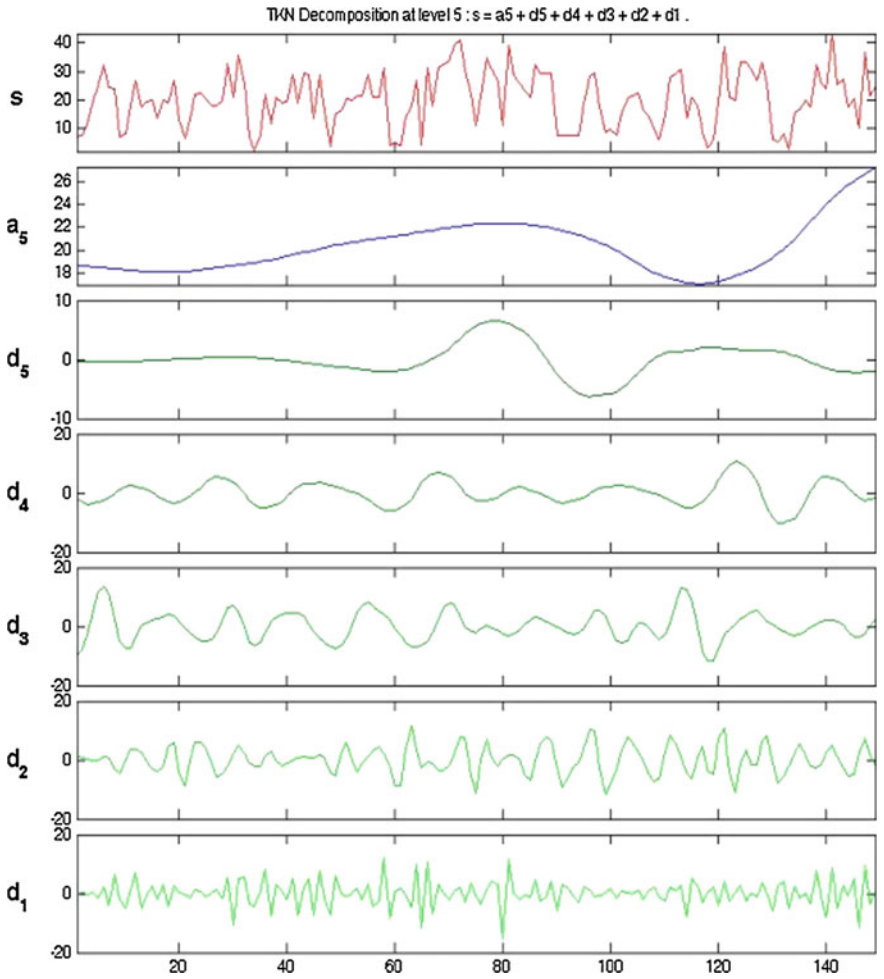

Fig. 11.12 1D discrete wavelet analysis of TKN

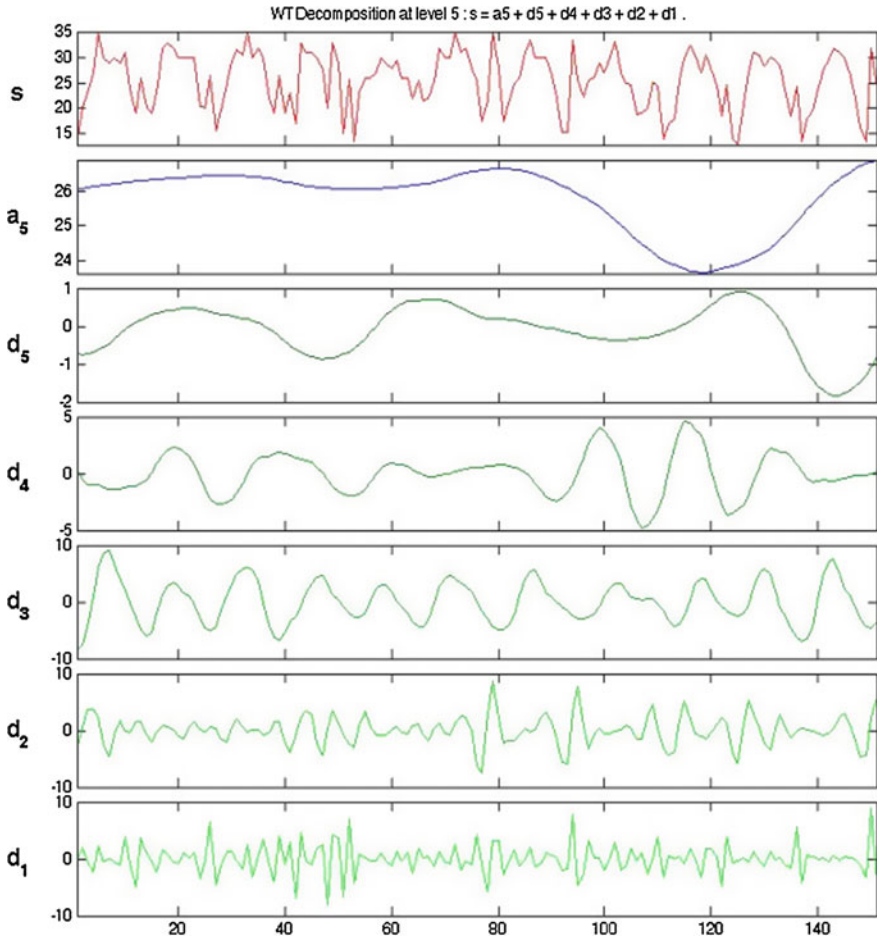

Fig. 11.13 1D discrete wavelet analysis of WT

Table 11.4 1D Daubechies wavelet level (5) parameters values

Water-quality parameters	Range of 's'	Range of 'a_5'	Range of 'd_5'	Range of 'd_4'	Range of 'd_3'	Range of 'd_2'	Range of 'd_1'
AMM	0 to 30	10 to 18	−5 to +5	−10 to +10	−20 to +20	−10 to +10	−20 to +20
BOD	0 to 60	10 to 30	−5 to +5	−10 to +10	−20 to +20	−20 to +20	−20 to +20
COD	0 to 120	54 to 76	−10 to +10	−50 to +50	−50 to +50	−50 to +50	−50 to +50
DO	0 to 7	0.2 to 1.4	−0.5 to 0.5	−1 to +1	−5 to +5	−5 to +5	−5 to +5
FC	0 to 20	0 to 14	−1 to +1	−5 to +5	−5 to +5	−1 to +1	−1 to +1
pH	0 to 8.5	7.3 to 7.7	−0.2 to 0.2	−0.2 to 0.2	−0.5 to 0.5	−0.5 to 0.5	−1 to +1
TC	0 to 8	0 to 10	−1 to +1	−5 to +10	−5 to +5	−5 to +5	−5 to +5
TKN	0 to 40	18 to 26	−10 to +10	−20 to +20	−20 to +20	−20 to +20	−20 to +20
WT	0 to 35	24 to 27	−2 to 1	−5 to +5	−10 to +10	−10 to +10	−10 to +10

Table 11.5 Regression equations, Hurst exponent, fractal dimension, and predictability index of water parameter

Parameters	Regression Eq. Monthly	R square	H (abs)	D (Fractal)	PI
pH	$y = 0.0011x + 7.4062$	0.0137	0.49945	1.50055	0.0011
COD	$y = 0.1509x + 55.932$	0.0368	0.42455	1.57545	0.1509
BOD	y = 0.0761x + 16.957	0.0502	0.46195	1.53805	0.0761
AMM	$y = 0.0638x + 11.567$	0.0651	0.4681	1.5319	0.0638
TKN	$y = 0.0388x + 18.15$	0.0189	0.4806	1.5194	0.0388
DO	$y = -0.0092x + 1.199$	0.052	0.5046	1.4954	0.0092
WT	$y = -0.0155x + 26.62$	0.0087	0.50775	1.49225	0.0155

11.4 Conclusion

In this paper, estimation of Hurst exponent, fractal dimension, and predictability index using wavelet method have been studied for air and water pollutants. For major air pollutants such as CO, NO, NO_2, O_3, and SO_2, recorded at Delhi College of Engineering (DCE), ITO-Crossing, Siri fort, and by mobile van in Delhi, the values of Hurst exponent, fractal dimension, and predictability index have been estimated. The monthly average values of water-quality parameters pH (Potential of Hydrogen), COD (Chemical Oxygen Demand), BOD (Biochemical Oxygen Demand), AMM (Free Ammonia), TKN (Total Kjeldahl Nitrogen), DO (Dissolved Oxygen), and WT (Water Temperature) monitored at Nizamuddin bridge-mid Stream of Yamuna river in Delhi (India) have been considered for the estimation of Hurst exponent, predictability index, and fractal dimension.

For each air pollutant recorded at each monitoring station, it is observed that Hurst exponent value is less than 0.5 which shows an anti-persistent behavior, i.e., an increase in time tends to a decrease in parameter value and vice versa. Fractal

dimension value lies between 1.5 and 2, which indicates that time series of air pollutant are more jagged than random. Predictability index is less than 0.5 which indicates for the existence of usual Brownian motion. Therefore, it can be concluded that each air pollutant has an anti-persistent behavior, and therefore future trend is unpredictable.

It is observed that Hurst exponent value for water pollutants COD, BOD, AMM, and TKN is less than 0.5 which shows an anti-persistent behavior, i.e., an increase in time tends to a decrease in parameter value and vice versa. Also for pH, DO, and WT, the value is close to 0.5 which shows Brownian behavior. Fractal dimension values for COD, BOD, AMM, and TKN lie between 1.5 and 2, which indicate that time series of water pollutant are more jagged than random. Also for pH, DO, and WT, the value is close to 0.5 which shows Brownian behavior. Predictability index for each water parameter is close to 0.5, which indicates for the existence of usual Brownian motion. Thus it can be concluded that pH, DO, and WT follow the Brownian time series behavior (true random walk) and parameters COD, BOD, AMM, and TKN follow and an anti-persistence behavior (or negative correlation) and the future trend is unpredictable.

From analysis of Hurst exponent, fractal dimension, and predictability index, it is concluded that the air quality parameters CO, NO, NO_2, O_3, and SO_2 follow the anti-persistent behavior; water-quality parameters pH, DO, and WT follow the Brownian time series behavior; and COD, BOD, AMM, and TKN follow an anti-persistence behavior. Thus the future trend for each water and air pollutant is unpredictable.

Acknowledgments Author is thankful to Central Pollution Control Board (CPCB), Government of India for providing the research data and Guru Gobind Singh Indraprastha University, Delhi (India) for providing research facilities.

References

1. Bhardwaj R, Parmar KS (2013) Water quality index and fractal dimension analysis of water Parameters. Int J Environ Sci Tech 10:151–164
2. Bhardwaj R, Singh K (2014) Water quality management using statistical analysis and time series prediction model. Appl Water Sci 4:425–434
3. Can Z, Aslan Z, Oguz O, Siddiqi AH (2005) Wavelet transform of meteorological parameters and gravity waves. Ann Geophys 23:659
4. Cannistraro G, Ponterio L (2009) Analysis of air quality in the outdoor environment of the city of messina by an application of the pollution index method. Int J Civ Environ Eng 1(4):214
5. Carbone A, Castelli G, Stanley HE (2004) Time-dependent Hurst-exponent in financial time series. Physica-A 344:267
6. CPCB, Water quality status of Yamuna river (1999–2005) 2006 Central Pollution Control Board, Ministry of Environment & Forests, Assessment and Development of River Basin Series: ADSORBS/41/2006-07
7. Hur J, Lee TH, Lee BM (2011) Estimating the removal efficiency of refractory dissolved organic matter in wastewater treatment plants using a fluorescence technique. Environ Technol 32:1843–1850

8. Koh YKK, Chiu TY, Boobis A, Cartmell E, Scrimshaw MD, Lester JN (2008) Treatment and removal strategies for estrogens from wastewater. Environ Technol 29:245–267
9. Korashey R (2009) Using regression analysis to estimate water quality constituents in Bahr El Baqar Drain. J Appl Sci Res 5:1067–1076
10. Nunnari G (2004) Modeling air pollution time-series by using wavelet functions and genetic algorithms. Soft Comput 8:173
11. Psargaonkar A, Gupta A, Devotta S (2008) Multivariate analysis of ground water resources in Ganga-Yamuna Basin (India). J Environ Sci Eng 50:215–222
12. Rangarajan G, Sant DA (2004) Fractal dimension analysis of indian climatic dynamics. Chaos, Solitons Fractals 19:285
13. Rehman S, Siddiqi AH (2009) Wavelet based Hurst exponent and fractal dimension analysis of Saudi climatic dynamics. Chaos, Solitons Fractals 40:1081–1090
14. Siddiqi AH (Lead Editors) (2003/2004) Arabian J Sci Eng, Part 1 28(1C), Part1129(11 C). Thematic issue on Wavelet and Fractal Methods in Science and Engineering
15. Vassilis Z, Antonopoulos M, Mitsiou AK (2001) Statistical and trend analysis of water quality and quantity data for the Strymon river in Greece. Hydrol Earth Syst Sci 5:679–691
16. Vela'squez Valle MA, Garcia GM, Cohen IS, Oleschko LK, Corral JAR, Korvin G (2013) Spatial Variability of the Hurst exponent for the daily rain fall series in the state of Zacatecas Mexico. Am Meteorol Soc 52(12):2771–2780

Chapter 12
A Novel Algorithm by Context Modeling of Medical Image Compression with Discrete Wavelet Transform

M.A. Ansari

Abstract To overcome the storage, transmission bandwidth, picture archiving and communication constraints and the limitations of the conventional compression methods, the medical imagery needs to be compressed selectively to reduce the transmission time and storage costs while maintaining the high diagnostic image quality. The selective medical image compression provides high spatial resolution and contrast sensitivity requirements for the diagnostic purpose. To fulfill these requirements, a novel approach of context modeling of medical image compression based on discrete wavelet transform has been proposed in this work. In medical images, contextual region is an area which contains the most useful and important information and must be coded carefully without appreciable distortion. The proposed method yields significantly better compression rates with better image quality than the general methods of compression defined in terms of image quality metrics performance. The experimental results have been tested on ultrasound medical images and the results have been compared with the results of standard general Scaling, Maxshift, Implicit, and EBCOT methods of selective image coding where it has been found that the proposed algorithm gives better and improved results based on subjective and objective image quality metrics analysis.

Keywords Transmission bandwidth · Maxshift method · EBCOT (Embedded block coding with optimized truncation) · Contextual modeling · Medical image compression · Image quality metrics

M.A. Ansari, Senior Member IEEE, affiliated from Islamic University (on EOL from Gautam Buddha University, Gr. Noida, India).

M.A. Ansari (✉)
Faculty of Engineering, Islamic University, Madinah, KSA
e-mail: mahmadiitr@gmail.com

A.H. Siddiqi et al. (eds.), *Mathematical Models, Methods and Applications*,
Industrial and Applied Mathematics, DOI 10.1007/978-981-287-973-8_12

12.1 Introduction

The conventional medical image compression techniques suffer from low compression rates, high distortions, and poor reconstruction image quality. The biggest challenge before the current medical image compression is to retain the high diagnostic image quality and to meet the high compression efficiency requirements. Therefore, the basic goal of medical image compression should be to reduce the bit rate as much as possible to enhance the compression efficiency while maintaining an acceptable diagnostic image quality. Perfect reconstruction, quality scalability, and the region of interest (ROI) based coding are the basic features needed for the teleradiology and telemedicine applications. However, medical images acquired from various modalities, such as Magnetic Resonance Imaging (MRI), Computed Tomography (CT), Ultrasonography (US), X-Ray Imaging (XR), Nuclear Medicine (NM), Computed Radiography (CR), and Digital Subtraction Angiography (DSA) comprise huge amount of data rendering them impractical for storage and transmission. For an example, a digital mammogram with pixel size of 50 μm is of size 5000 × 5000 pixels with 12 bits per pixel, and needs about 37.5 MB for storage without compression and similarly a single analog mammogram may be digitized at 4096 × 4096 pixels × 16 bpp of file size over 33 MB [1]. The storage cost differences for various compression states are more intuitive to understand. It will be paying 50 % less for storage media with 2:1 compression and a whopping 90 % less with 10:1 compression: an administrator's dream solution to reducing prices which picture archiving and communication systems (PACS) costs [2].

The JPEG (Joint Photographic Expert Group) [3] is one of the oldest and mostly used standard for still image compression which is based on discrete cosine transform (DCT) proposed by Ahmed in 1974 [4] but it suffers from the blocking artifacts and does not posses the multiresolution property like the wavelet transform, therefore, in the literature, many wavelet transform (WT) [5] based image coders have been proposed such as embedded zero-tree wavelets (EZW) [6], set partitioning in hierarchical trees (SPIHT) [7], EBCOT [8], morphological representation of wavelet data (MRWD) of Servetto et al. (99) and group testing for wavelets (GTW) of Hong and Ladner (02) developed for the JPEG2000 compression standard (ISO, 2000) [9, 10]. Whereas, WT analyzes images with recursive decomposition procedure applied to the low frequency component only [5], wavelet packet transform applies the decomposition to both the low- and high-frequency components resulting in the entire family of subband decompositions [11]. Most of the ROI coding methods are wavelet-based compression techniques. One range of popular ROI coding schemes are based on SPIHT [12, 13]. Several ROI coding methods based on bit-plane coding have also been proposed in [13–15].

Though the lossless compression can do the perfect reconstruction requirement, however, it will have very low compression rate, i.e., 4:1 maximum and will not be able to handle the bulk amount of medical data generated everyday which has to be stored and transmitted through PACS. So, it will not solve the purpose. On the other

hand, the lossy compression methods will give better compression rates (up to 50:1) but may lose some diagnostic information which will never be tolerated by the medical community. So few new trends like ROI coding, context modeling, and content-based coding are developing in this context to meet out the high compression rates and the perfect reconstruction of the medical imagery. The very first method of the ROI coding incorporated in JPEG2K is the scaling method, which is based on EBCOT where the bits representing the wavelet coefficients contributing to the contextual region are shifted upward by a user-defined value [14]. This allows for the coding of ROI with any desired quality compared to the background (BG). The multiple ROIs are allowed each with its own corresponding scaling value and the ROI shape is limited to the circles and rectangles. The ROI coordinates and shift values are signaled in the bit stream [16–18]. The basic requirement in scaling method is that the ROI mask needs to be generated both on the encoder and the decoder side. The scaling method has two major drawbacks. First, it needs to encode and transmit the shape information of the ROIs. Thereby increases the algorithm complexity. Second, if arbitrary ROI shapes are desired, the shape coding will consume a large number of bits, which significantly decreases the overall coding efficiency [17, 18]. In the part I, JPEG2K proposes Maxshift method based on EBCOT, which is a particular case of the general scaling based method when the scaling value is so large that there is no overlapping between BG and ROI biplanes, but it lacks the flexibility to allow an arbitrary scaling value to define the relative importance of the ROI and the BG coefficients [19]. The other serious problem of the Maxshift method is bit stream overflow which results in some loss in the BG portion [20] and the third problem is the own scale value for multiple ROIs. Therefore, it is very difficult to support the different degrees of interest during multiple ROI coding and transmission [18].

The proposed context-based compression is an excellent method which gives better performance with comparable computational efficiency and the high visual quality of the reconstructed image. The contextual coding allows selected parts of an image (contextual region) to be coded with higher quality as compared to the background (patient information and image parameters). It is done using a contextual region of interest (CROI) mask with the priority adjustment. The CROI is identified by the segmentation and interactive methods and the binary mask for the CROI region is generated in the wavelet domain and describes which quantized wavelet coefficients must be encoded with higher quality. It depends on CROI specification in the image domain and the DWT filter. In the progressive transmission, the CROI is transmitted and decoded before the background (BG). Therefore, CROI coding is capable of delivering high reconstruction image quality over user-specified spatial regions in a limited time, compared to compression of the entire image [15]. In the proposed contextual coding, the basic objective is to achieve better compression rates by applying different compression thresholds for the wavelet coefficients of each DWT band (BG and CROI), while conventional image compression methodologies utilizing the DWT apply it to the whole image. That is, in contextual coding, different compression rates are applied to the wavelet

coefficients in the different CROIs, respectively, resulting in the high diagnostic quality of the image with sufficient information retained in the BG. Further, CROI coding provides an excellent trade-off between image quality and the compression ratio. In order to find an optimal technique for medical image compression, an experimental study is conducted to qualitatively judge the efficacy of contextual approach in comparison with the EBCOT, Maxshift, Implicit, and SPIHT compression techniques and it is found that CBDWT reconstructed images outperform the above methods in terms of rate distortion and the visual quality. The performance parameters have been discussed in detail by many authors and may be referred in [15, 21–23]. These output performance parameters have been analyzed quantitatively and plotted which clearly show the improved performance of the proposed CBDWT method at low bit rates (high compression rates) in comparison to the methods discussed in [13, 15, 19].

12.2 The Wavelet Transform-Based Coding

The main difference between the WT-based and DCT-based transform coding system is the omission of transform coder's sub-image processing stages. Because wavelet transforms are both computationally efficient and inherently local (i.e., their basis functions are limited in duration), subdivision of the original image is not required. The removal of the subdivision step eliminates the blocking artifact. Wavelet coding techniques are based on the idea that the coefficient of a transform which decorrelates the pixels of an image can be coded more efficiently than the original pixels themselves [24]. The computed transform converts a large portion of the original image to horizontal, vertical, and diagonal decomposition coefficients with zero mean and Laplacian-like distribution. The 9/7 tap biorthogonal filters [16], which produce floating point wavelet coefficients, are widely used in image compression techniques to generate a wavelet transform [25–27]. The wavelet coefficients are uniformly quantized by dividing by a user-specified parameter and rounding off to the nearest integer. Typically, the majority of coefficients with small values are quantized to zero by this step. The zeroes in the resulting sequence are run-length encoded, and Huffman and arithmetic coding are performed on the resulting sequence. The various subbands blocks of coefficients are coded separately, which improves the overall compression [24]. If the quantization parameter is increased, more coefficients are quantized to zero, the remaining ones are quantized more coarsely, the representation accuracy decreases, and the compression ratio increases consequently. Since the input image needs to be divided into blocs in DCT-based compression, correlation across the block boundaries is not eliminated. This results in 'blocking artifacts' particularly at low bit rates. Whereas in wavelet coding, there is no need to block the input image and its basis functions have variable length hence wavelet coding schemes at higher compression avoid blocking artifacts. The basic structure of wavelet-based compression algorithm is shown in Fig. 12.1.

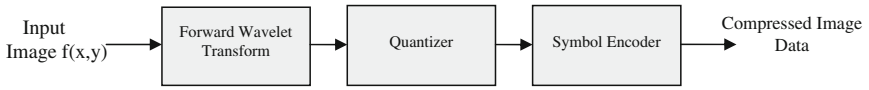

Fig. 12.1 Wavelet compression model

12.2.1 The Discrete Wavelet Transform

Wavelet transforms are based on 'basis functions.' Unlike the Fourier transform, whose basis functions are sinusoids, wavelet transforms are based on small waves, called 'wavelets' of varying frequency and limited duration. Wavelets are the foundation of a powerful signal processing approach, called Multiresolution Analysis (MRA). As its name implies, the multiresolution theory is concerned with the representation and analysis of signals (or images) at more than one resolution. Hence features that might go undetected at one resolution may be easy to spot at another. The Wavelet analysis is based on two important functions, viz., the scaling function and the wavelet function. Calculating wavelet coefficients at every possible scale is a fair amount of work, and it generates lot of data. If we choose only a subset of scales and positions at which to make our calculations, it turns out, rather remarkably, that if we choose scales and positions based on powers of two—so-called *dyadic* scales and positions—then our analysis will be much more efficient and just as accurate. If the function being expanded is a sequence of numbers, like samples of a continuous function $f(x)$, the resulting coefficients are called the discrete wavelet transform (DWT) of $f(x)$ [28].

In MRA, a scaling function, $\varphi(x)$, is used to create a series of approximations of a function or image, each differing by a factor of 2 from its nearest neighboring approximations. The wavelets are then used to encode the difference in information between adjacent approximations. Consider a set of expansion functions composed of integer translations and binary scalings of the real, square-integrable function $\varphi(x)$; that is, the set $\{\varphi_{j,k}(x)\}$ where [28];

$$\varphi_{j,k}(x) = 2^{j/2}\varphi(2^j x - k) \tag{12.1}$$

for all j, $k \in Z$ and $\varphi(x) \in L^2$ (**R**). Here, k determines the position of $\varphi_{j,k}(x)$ along the x-axis, j determines $\varphi_{j,k}(x)$'s width—how broad or narrow it is along the x-axis. Because the shape of $\varphi_{j,k}(x)$ changes with j, $\varphi(x)$ is called a 'scaling function.' Given a scaling function that meets the MRA requirements, we can define a wavelet function $\psi(x)$ that, together with its integer translates and binary scalings spans the difference between any two adjacent scaling subspaces V_j and V_{j+1}. The set $\{\psi_{j,k}(x)\}$ of wavelets are defined as [28]:

$$\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k) \tag{12.2}$$

There exist various extensions of the one-dimensional wavelet transform to higher dimensions. In two dimensions, a two-dimensional scaling function, $\varphi(x, y)$, and three-dimensional wavelets, $\psi^H(x, y)$, $\psi^V(x, y)$ and $\psi^D(x, y)$, are required. Each is a product of a one-dimensional scaling function φ and corresponding wavelet ψ. Excluding products that produce 1D results, like $\varphi(x)\ \psi(x)$, the four remaining products produce the separable scaling function $\varphi(x, y)$. In 2-D wavelet analysis, a scaling function $\varphi(x, y)$ is defined such that;

$$\varphi(x, y) = \varphi(x)\varphi(y) \tag{12.3}$$

where, $\varphi(x)$ is a one-dimensional scaling function. Let $\psi(x)$ be the one-dimensional wavelet associated with the scaling function. Then, the three-, two- D wavelets are defined as:

$$\psi^H(x, y) = \varphi(x)\psi(y) \tag{12.4}$$

$$\psi^V(x, y) = \psi(x)\varphi(y) \tag{12.5}$$

$$\psi^D(x, y) = \psi(x)\psi(y) \tag{12.6}$$

where, H, V, and D stand for "horizontal", "vertical" and "diagonal" respectively. These wavelets measure functional variations intensity or gray-level variations for images—along different directions : ψ^H measures variations along columns(e.g., horizontal edges), ψ^V responds to variations along rows(e.g., vertical edges), and ψ^D corresponds to variations along diagonals. The 2D multiresolution analysis (MRA) decomposition is completed in two steps. First, using $\varphi(x)$ and $\psi(x)$ in the x direction, $f(x, y)$ (an image) is decomposed into two parts, a smooth approximation and a detail. Next, the two parts are analyzed in the same way using $\varphi(y)$ and $\psi(y)$ in the y direction. As a result, four channel outputs are produced, one channel is A1 $f(x, y)$, the level one smooth approximation of $f(x, y)$, through $\varphi(x)\ \varphi(y)$ processing, the other three channels are $D_1^{(H)} f(x, y)$, $D_1^{(V)} f(x, y)$ and $D_1^{(D)} f(x, y)$, the details of the image. Level two results are obtained after decomposing A1 $f(x, y)$ progressively.

Given separable 2D scaling and wavelet functions, extension of the one-dimensional DWT to 2D is straight forward. We first define the scaled and translated basis functions [28]:

$$\varphi_j, m, n(x, y) = 2^{j/2}\varphi(2^j x - m, 2^j y - n), \tag{12.7}$$

$$\psi_j^i, m, n(x, y) = 2^{j/2}\psi(2^j x - m, 2^j y - n);\ i = \{H, V, D\} \tag{12.8}$$

where index '*i*' identifies the directional wavelets in above equations. The DWT of function $f(x, y)$ of size M × N is then given by:

$$W_{\varphi}(j_0, m, n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y) \varphi_{j0,m,n}(x, y); \quad i = \{H, V, D\} \tag{12.9}$$

$$W_{\psi}^{i}(j, m, n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y) \psi_{j,m,n}^{i}(x, y) \tag{12.10}$$

where, 'j_0' is an arbitrary starting scale and the $W_{\varphi}(j_0,m,n)$ coefficients define an approximation of $f(x, y)$ at scale 'j_0'. The $W_{\psi}^{i}(j_0,m,n)$ coefficients add horizontal, vertical and diagonal details for scales $j \leq j_0$. Normally, $j_0 = 0$ and select $N = M = 2^J$ so that $j = 0, 1, 2 \ldots J-1$ and $m, n = 1, 2, \ldots 2^{j-1}$. Given the W_{φ} and W_{ψ}^{i} of Eqs. (12.7–12.10), $f(x,y)$ is obtained via the inverse DWT.

$$\begin{aligned} f(x, y) = & \frac{1}{\sqrt{MN}} \sum_{m} \sum_{n} W\varphi(j_0, m, n) \varphi_{j_0,m,n}(x, y) \\ & + \frac{1}{\sqrt{MN}} \sum_{i=H,V,D} \sum_{j=j_0}^{\infty} \sum_{m} \sum_{n} W_{\psi}^{i}(j, m, n) \psi_{j,n,m}^{i}(x, y) \end{aligned} \tag{12.11}$$

The decomposition process using DWT is represented in Fig. 12.2 in the block diagram form to illustrate the decomposition of image into the high- and low-frequency components.

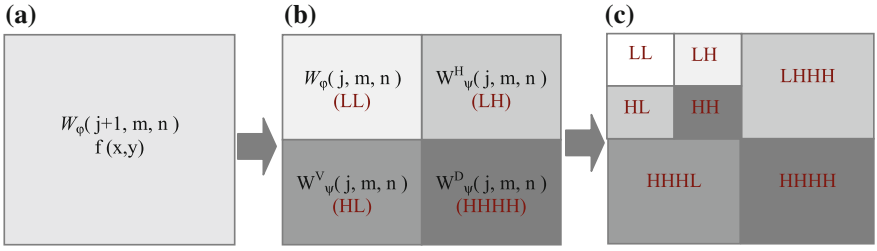

Fig. 12.2 Two-level decomposition in 2-D DWT. **a** Original image function level 0. **b** Level one decomposition. **c** Two-level decomposition

12.2.2 Algorithm for Wavelet-Based Coding

1. The first step of encoding process is to level shift the pixels of an image by subtracting $2^{(m-1)}$, where 2^m is number of gray levels in the image.
2. The one-dimensional DWT of the rows and the columns of the image can then be computed.
3. After the DWT has been computed, the total number of transform coefficients is equal to the number of samples in the original image but the important visual information is concentrated in a few coefficients.
4. The final step of encoding process is to code the quantized coefficients arithmetically on a bit-plane basis.
5. The decoding process simply inverts the operation of encoding.
6. After decoding the arithmetically coded coefficients, a user-selected number of the original image's subbands are reconstructed. The de-normalized coefficients are then inverse transformed and level shifted to yield an approximation of the original image.

12.3 Selection Criterion of the Contextual Region

12.3.1 Interactive Approach

ROI coding is one of the features of the JPEG2000 standard. This feature allows users to define regions within an image to be coded and transmitted in better quality and with less distortion than the rest of the image. ROIs are manually defined in JPEG2000, then wavelets are used to compress the ROI at a higher bitrate than the rest of the image and the ROI wavelet coefficients are upshifted before the actual transmission occurs. ROI is especially useful when using progressive transmission of the image; in such a scenario, the ROI is transmitted first and the background information is transmitted later. The receiver progressively reconstructs the image and can interrupt the transmission at any time; yet, the ROI will still have the highest quality of all the regions in the image [29]. A number of ROI coding methods have been proposed, such as the ROI coding in JPEG2000 and ROI coding based on SPIHT [7, 12, 13]. Although, ROI coding techniques are used extensively, most of the ROI identification itself is done manually [29], i.e., with human intervention.

12.3.2 The Mathematical Approach

The selection of the contextual region (CROI) is one of the most important parts of the coding for the implementation of any context-based algorithm. The ultrasound scanners produce conical images containing a lot of background information which

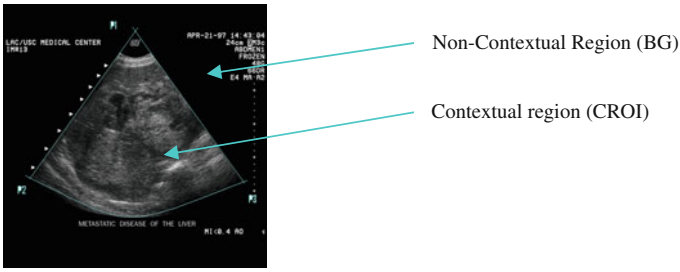

Fig. 12.3 Original US image (heart) and the modeled CROI with three points P_1, P_2, and P_3

comprises patient information and the image parameters. The actual image acquired by the probe sensor is of the form of a fan-shaped window, placed over a black rectangle, and centered and aligned with its top border. This is the diagnostically useful image area and it corresponds to only about 50–60 % [30] of the area of the full image as shown in Fig. 12.3. Thus before applying any contextual compression technique, it is necessary to model the useful contextual image area.

This can be achieved by automatically detecting three points P_1 (uppermost), P_2 (leftmost) and P_3 (rightmost), as shown in Fig. 12.3, knowing that P_1 can be found toward the top center point of the image. The other two are pointed to by bright markers (generated by the scanner) and are symmetrically located left and right of P_1. A raster scan of the original image will detect these points [30]. The fan-shaped CROI can now be modeled as the sector of a circle centered at P_1 and bounded by two straight lines intersecting at P_1 and crossing the circle at P_2 and P_3, respectively. This can be expressed by the parametric equation of a straight line passing through $P_1(x_1, y_1)$ and $P_2(x_2, y_2)$:

$$P(t) = P_1 + t(P_2 - P_1) \tag{12.12}$$

where, 't' is a variable in the form of 'x' and $P(t)$ is a variable in the form of 'y' in the straight line equation. The parametric equation of a circle with center at $P_1(x_1, y_1)$ and with radius 'r' is given by:

$$(x - x_1)^2 + (y - y_1)^2 = r^2 \tag{12.13}$$

The above-stated method of locating the CROI is a mathematical approach and can help in the implementation of the contextual algorithm. However, this technique of locating the CROI is limited to the US images only.

12.3.3 Segmentation Approach

Segmentation is another approach of separating out the contextual region, i.e., CROI and the non-contextual region (BG) from a medical image. For this purpose,

many segmentation techniques like edge detection or region growing segmentation methods are useful. However, they cannot be directly applied to medical images because they contain large amounts of noise induced during the acquisition process. To overcome this problem, the image is preprocessed prior to segmentation which involves adaptive smoothing [31]. The fundamental idea behind it is to iteratively convolve the image we want to smooth with a very small (3 × 3) averaging mask whose coefficients reflect, at each point, the degree of continuity of the image intensity function. Two effects can be observed during adaptive smoothing: one is the sharpening of the edges which will eventually become the boundaries of constant intensity regions; the other is the smoothing within each region [30].

In the segmentation-based CROI coding, based on region growing technique, the aim is to group spatially connected pixels lying within a small dynamic gray level range. The region growing procedure starts with a single pixel, called the seed pixel. Each of the seed's four-connected (neighbor) pixels are checked with a region growing (or inclusion) condition. If the condition is satisfied, the neighbor pixel is included in the region. The four neighbors of the newly added neighbor pixel are then checked for inclusion in the region. This recursive procedure is continued until no spatially connected pixel meets the growing condition. A new region growing procedure is then started with the next pixel of the image which is not already a member of a region; the procedure ends when every pixel in the image has been included in one of the regions grown [1]. After locating the CROI, a suitable compression algorithm is applied in such a way that after reconstructing the image, the quality of a CROI is superior as compared to the background area BG.

12.3.3.1 Region-Based Segmentation

The segmentation techniques are based on finding the regions directly [28]. Let R represents the entire image region. Segmentation can be viewed as a process that partitions R into n subregions, namely R_1, R_2…, R_n, such that;

1. $\bigcup_{i=1}^{n} R_i = R$
2. R_i is a connected region, $i = 1, 2\ldots, n$.
3. $R_i \bigcap R_j = \phi$ for all i and j, where $i \neq j$.
4. $P(R_i)$ = TRUE for $i = 1, 2\ldots,n$.
5. $P(R_i \bigcup R_j)$ = FALSE for $i \neq j$.

Here $P(R_i)$ is a logical predicate defined over the points in set R_i and $\emptyset$ is the null set. *Condition* 1 indicates that segmentation must be complete; that is, every pixel must be in a region. *Condition* 2 requires that points in a region must be connected in some predefined sense. *Condition* 3 indicates that the regions must be disjoint. *Condition* 4 deals with the properties that must be satisfied by the pixels in a segmented region. Finally, *Condition* 5 indicates that regions R_i and R_j are different in the sense of predicate P.

12.4 The Region of Interest (ROI) Coding Schemes

12.4.1 Generation of ROI Mask

When an image is coded with an emphasis of ROI, it is necessary to identify the wavelet coefficients needed for the reconstruction of the ROI. Thus, the ROI mask is introduced to indicate which wavelet coefficients have to be transmitted exactly in order for the receiver to reconstruct the ROI [32]. Once an arbitrarily shaped ROI is defined by user, generation of the ROI mask is performed for rows and columns at each decomposition level. The process is then repeated for the remaining levels until the entire wavelet tree is processed. The wavelet coefficients that are required to reconstruct a pixel are selected with dependency on the wavelet length [13].

For the generation of parent of ROI mask (PROI), the ROI coding algorithm discussed in [12] examines whether each node is necessary for the decoder to reconstruct the ROI, i.e., whether it is an ROI coefficient, before testing whether it is significant (node test). Then, if the node is not an ROI coefficient, its node test is skipped and performed later. After the node tests for all ROI coefficients, the encoder examines whether descendants of each node are significant (descendant test) without testing whether descendants of the node are the ROI coefficients, i.e., whether the node is a PROI coefficient [13].

12.4.2 EBCOT Coding

The ROI coding is one of the functionalities in the JPEG2K image compression standard. Its coding paradigm is based on the embedded block coding with optimized truncation (EBCOT) algorithm [8, 15]. The scalability for progressive transmission in JPEG2K is based on the discrete wavelet transform (DWT) and EBCOT. DWT is used to exploit spatial redundancy and to impart resolution scalability. The DWTs used in JPEG2K are the nonreversible wavelet transform for lossy compression and the reversible integer wavelet transform (IWT) for lossy and lossless compression [9, 10]. When a quality progressive bit stream is transmitted to a client, the image quality of ROI is expected to improve more rapidly than the BG. This is achieved with emphasis on ROI by identifying the wavelet coefficients needed for reconstruction of the spatial region of interest, and encoding them with higher priority. To identify ROI coefficients, an ROI mask is generated by tracing the inverse wavelet transform backwards. The Fig. 12.4 depicts an ROI mask for two-level wavelet decomposition [15]. Detailed calculation of the ROI mask can be found in [8]. It is observed in Fig. 12.4 that each ROI code block may contain ROI coefficients and BG coefficients. JPEG2K Part 1 provides two mechanisms for assigning higher priority to the ROI. One is the Maxshift method, in which wavelet coefficients involved with the reconstruction of ROI are scaled up prior to

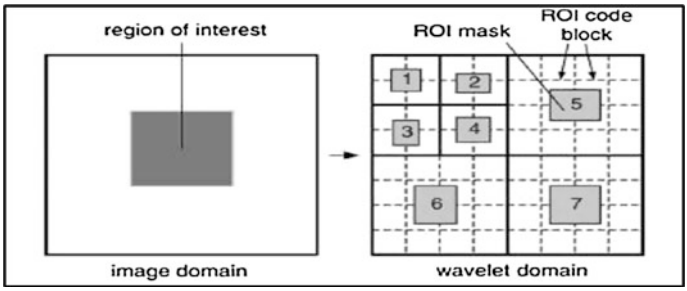

Fig. 12.4 ROI region and ROI mask

quantization. The second mechanism is to adjust the cost function of the rate distortion optimization algorithm so that code blocks whose wavelet coefficients contribute to the ROI are assigned a higher priority [9, 10, 15].

12.4.3 The General Scaling-Based Method

The very first method of the ROI coding in JPEG2K is the general scaling method of contextual coding based on EBCOT where the bits representing the wavelet coefficients contributing to the contextual region are shifted upward by a user-defined value [14] as shown in Fig. 12.5. The general scaling-based method places ROI associated bits in the higher bit planes by scaling the bit planes of ROI coefficients up, so that ROI coefficients can be coded first in the embedded bit plane coding. This method allows the use of arbitrary scaling value, so allows fine control on the relative importance between ROI and BG [18]. The scaling method has two major drawbacks. First, it needs to encode and transmit the shape information of the ROIs. This rapidly increases the algorithm complexity. Second, if arbitrary ROI shapes are desired, the shape coding will consume a large number of bits, which significantly decreases the overall coding efficiency [17, 33]. Further, the general scaling-based method requires the generation of an ROI mask and the distinction of ROI/BG coefficients at both encoder and decoder sides. This increases decoder complexity and processing overhead.

12.4.4 The Maxshift Method

The Maxshift method is efficient for ROI reconstruction supported by JPEG2K Part I which is a particular case of the general scaling-based method where the scaling value is so large that there is no overlapping between BG and ROI bitplanes.

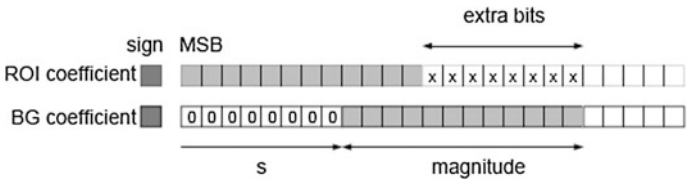

Fig. 12.5 ROI scaling operation where '*s*' is the scaling value

All the significant bits associated with the ROI after scaling will be in higher bitplanes than all the significant bits associated with the background. Therefore, ROI shape is implicit for the decoder in this method, and arbitrarily shaped ROI coding can be supported [18] but it has the drawback that the BG becomes recognizable only after ROI is completely reconstructed [15]. In the Maxshift method, prior to bitplane coding, the bitplanes of the ROI coefficients are scaled up by the desired amount so that coefficients associated with the ROI are placed in higher bit planes. If the scaling value is '*s*', the wavelet coefficient $a'(u, v)$ supplied to the bit-plane entropy coder is given by:

$$a'(u,v) = \begin{cases} a(u,v), & M(u,v) = 0 \\ 2^s a(u,v), & M(u,v) = 1 \end{cases} \tag{12.14}$$

The scaling factor is selected to ensure there is no overlap between BG and ROI bit planes, as depicted in Fig. 12.6. When the entropy coder encodes the code block containing ROI coefficients, the encoded ROI bits appear before the BG bits. Then rate control builds a layer progressive bit stream in which information pertaining to ROI precedes that of the BG.

The main strength of Maxshift is its fast ROI reconstruction. It also lifts the restriction on ROI shape. Drawbacks of the Maxshift method include the increase in coding time, and BG information is received only after full ROI reconstruction. The flow chart of Maxshift algorithm is given in Fig. 12.7 [34]. Some problems associated with this method are as follows [18].

(i) It lacks the flexibility to allow an arbitrary scaling value to define the relative importance of the ROI and the BG coefficients. This means in all the subbands, no information about the BG coefficients can be received until the ROI coefficients has been fully decoded, even if detail is imperceptible random noise or unnecessary information [19].
(ii) In Maxshift method, bit stream overflow problem is serious. For some medical image, the scaling value can exceed 16. This means if the implementation precision is only 32 bits, we will lose some of the least significant BG bitplanes. This results in some loss in the BG portion [20].
(iii) When there are multiple ROIs in an image, any ROI cannot have its own scaling value. Therefore, it is very difficult to support the different degrees of interest during Multiple ROI coding and transmission [18].

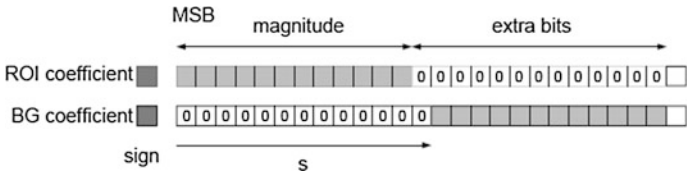

Fig. 12.6 Maxshift method of ROI coding where '*s*' is scaling of the ROI coefficients

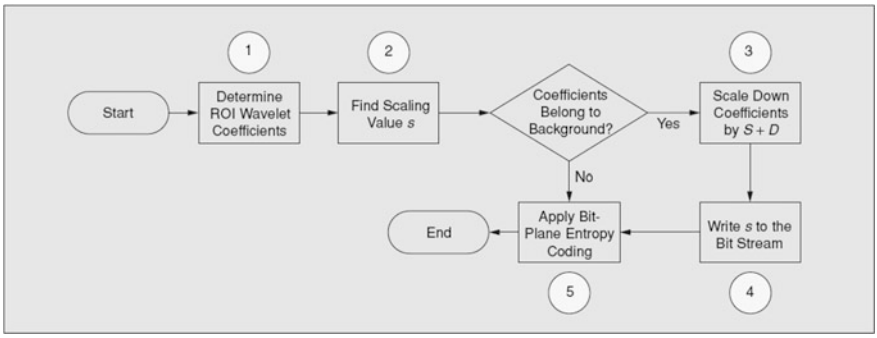

Fig. 12.7 Flowchart of the MAXSHIFT ROI coding procedure. (*1*) The set of wavelet coefficients that belong to ROI is determined. (*2*) The scaling value (*s*) and the magnitude of the largest wavelet coefficient not in the ROI, i.e., contained in the background, are calculated. (*3*) The background coefficients are scaled down by '*s*'. (*4*) The '*s*' value is added to the bit stream. (*5*) The bit-plane entropy coding is applied as usual [33]

(iv) The main problem of the Maxshift method is that the bitplanes of all ROI coefficients must be encoded before accessing bitplanes of the background. So the Maxshift does not have the flexibility for an arbitrary scaling value to define the relative importance of the ROI and the BG wavelet coefficients as in the general scaling-based method. Additionally, the EBCOT algorithm applied by JPEG2000 also increases complexity of the ROI coding [33].

12.4.5 Implicit ROI Coding

In contrast to the Maxshift method, the implicit ROI encoding method is designed to take full advantage of EBCOT and achieves ROI emphasis in the bit stream ordering process [35]. In EBCOT, each quality layer comprises an arbitrary contribution from the embedded bit stream of each code block of each subband. Thus ROI emphasis is possible by including relatively larger contributions from code blocks involved in the ROI reconstruction to quality layers. The main advantage of the implicit ROI encoding is its low complexity. The method itself is

straightforward and easy to implement. More than that, no bitplane scaling is involved at either the encoding or the decoding side. However, the implicit ROI encoding has the disadvantage of slow ROI reconstruction. This is because the priority arrangement is made on a block-by-block basis and some ROI code blocks may contain a large amount of background information. To calculate the distortion, these background coefficients in ROI code blocks are assigned the same priority as ROI coefficients, and are coded in tandem [15].

12.5 The Proposed Context Based DWT (CBDWT) Compression Method

The present work is based on the context modeling of DWT coefficients for medical image compression. In the proposed method, the segmentation and interactive methods of selecting the diagnostically important contextual region of interest (CROI) mask is used to separate out BG and CROI from the given test image and they are encoded separately on the priority basis by DWT-based algorithm as given in Fig. 12.8. The CROI transformed coefficients are first quantized followed by progressively encoding and then the BG coefficients are quantized and encoded. Thus, the CROI coefficients are first transmitted for decoding followed by the BG coefficients. The Image processing toolbox of Matlab7.2 has been used for the implementation of the algorithm after performing several iterations and steps of preprocessing. The wavelet-based context modeled structure, representing the intersubband dependency, provides excellent rate distortion and the encoding process can be terminated as soon as the desired bit rate and the compression rate is achieved. The other advantage of the DWT-based coding schemes is that these allow for modification of transmission order to place more emphasis on CROI.

Figure 12.9 shows the proposed contextual coding model where (a) Represents the entire image (A+B) in which both the contextual region (CROI) and the background (BG) are present (b) Separated contextual region (CROI) denoted by 'A', and (c) Extracted Background (BG), i.e., the rest of the image is present denoted by 'B'. Figure 12.10a represents the actual image and the separated CROI and BG from this image are shown through (b) and (c), respectively. The reconstructed image is shown in (d). For the lesser complexity purpose and the safe

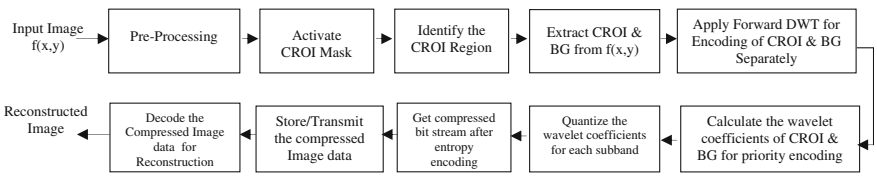

Fig. 12.8 Proposed CBDWTT compression algorithm

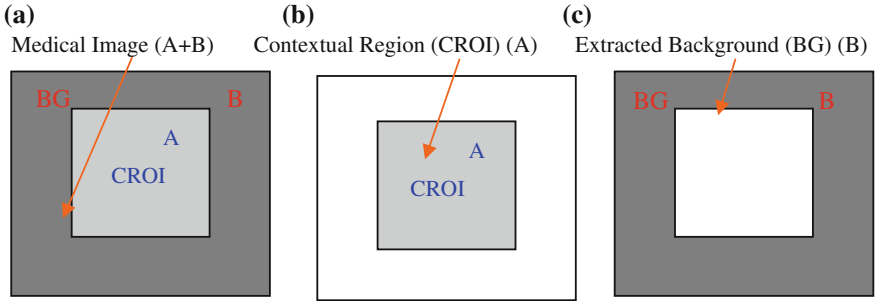

Fig. 12.9 **a** Represents the entire image (*A+B*) in which both the contextual region (CROI) and the background (BG) is present. **b** Separated contextual region (CROI) denoted by '*A*'. **c** Extracted background (BG), i.e., the rest of the image is present denoted by '*B*'

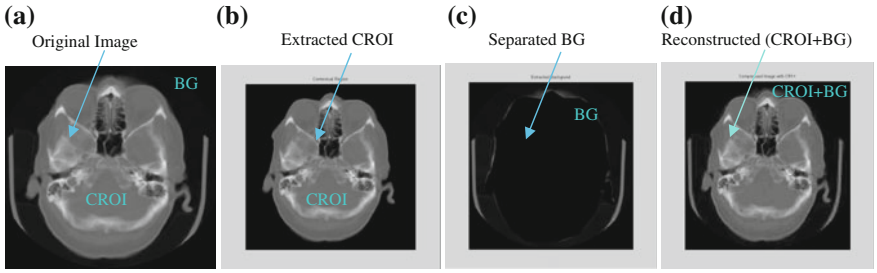

Fig. 12.10 Separation of CROI and BG from an CT image and its reconstructed image. **a** Represents the entire original image in which both the contextual region (CROI) and the background (BG), i.e., the rest of the image is present. **b** Separated contextual region (CROI) which contains the diagnostic information. **c** Extracted background (BG) contains rest of the image. **d** Reconstructed image at compression ratio CR and bit rate bpp

diagnosis, the entire diagnostic image area has been selected as CROI as shown in Fig. 12.10b. Depending upon the type of image and need of the diagnosis, the shape and the size of CROI may vary. As it has been suggested in [14] more than one CROI can also be selected and the most important region can be compressed with the best quality.

After separating out the image area CROI and the BG, the context-based DWT (CBDWT) compression algorithm is applied as shown in Fig. 12.8 in block diagram. The CROI is compressed with high bit rate, i.e., low compression ratio (CR) to have a high diagnostic quality whereas the BG is compressed with high compression ratio, i.e., low bit rate. Thereby, as a result, the over all compression ratios are good along with the improved quality of the reconstructed image. After encoding of the CROI and the BG separately, they are merged to get an overall compression ratio and the performance of the reconstructed image. The CBDWT

algorithm's results have been compared with other region of interest coding methods given in [15] and Table 12.5. The performance of these compression algorithms is checked with the known parameters like, CR, bpp, MSE, PSNR, CoC, and the visual quality of the reconstructed image by using the human visual system (HVS). The different stages of the proposed algorithm are shown in Fig. 12.8 and the steps of the proposed algorithm are given in the following Sect. 12.5.1 proposed CBDWT algorithm.

12.5.1 The Basic Steps of the Proposed CBDWT Algorithm

/......./

Step 1: Identify the contextual region (CROI) from the test image.

Step 2: Set CROI mask on source image f(x,y).

Step 3: Generate the CROI Mask.

Step 4: Obtain the corresponding mask in transform domain.

Step 5: Split CROI and the background (BG) from test image f(x,y).

Step 6: Select the compression methodology for CROI and BG, respectively.

Step 7: Calculate the wavelet coefficients of CROI & BG separately for priority encoding of the test image f(x,y).

Step 8: Obtain the bit allocation for each region i.e. CROI and BG.

Step 9: Quantize the wavelet coefficients for each subband of each region (CROI & BG) by the bit allocation resulting from Step.8, and send the quantized coefficients to entropy encoder progressively.

Step 10: Compress the CROI region with very low CR (high bit rate)lossy or near lossless by quantization in step 9.

Step 11: Compress the BG region with very high CR (low bpp) and lossy by quantization in step 9.

Step 12: Multiplex the entropy coded coefficients and CROI mask in order to make bit stream.

Step 13: Get the compressed bit stream of f(x,y) for CROI & BG, separately.

Step 14: Decode the image f(x,y)(Decoding process is the inverse order of the encoding process).

Step 15: Merge the CROI and the BG regions to get decoded image $\hat{f}(x,y)$.

Step 16: Calculate performance parameters of CROI, BG & Image $\hat{f}(x,y)$.

Step 17: Check the Image quality by the HVS and the Correlation.

Step 18: Compare and analyze output compression performance parameters.

Step 19: Plot histograms of $\hat{f}(x,y)$ and Image f(x,y) for correlation.

Step 20: Repeat the process by changing bpp till the desired quality and the required compression performance parameters (CPP) are achieved.

Step 21: End the process if satisfactory image quality metrics parameters and image quality are achieved.

/......./

12.5.2 Image Quality Metrics (IQM) Indices

The IQM performance parameters are of paramount importance for any image compression method on the basis of which the efficiency of the compression algorithm is measured. The main IQM parameters are—bit rate (bpp), compression ratio (CR), mean square error (MSE), peak signal to noise ratio (PSNR), and correlation coefficient (CoC). The detailed study of IQM parameters may be referred in [21, 22].

(i) The CR of a compressed image is defined as the ratio of the size of the original image data in bits to the size of compressed image data in bits and mathematically it is expressed as:

$$\text{CR} = \frac{\text{size of original image in bits}}{\text{size of compressed image in bits}} \tag{12.15}$$

(ii) The frequent quality measure used for evaluation of the distortion in a compressed image is the MSE. The MSE of a reconstructed image is given by the mean of the squares of difference between the original image and the reconstructed image pixels given as:

$$\text{MSE} = \frac{1}{MN}\sum_{x=1}^{M}\sum_{y=1}^{N}\left\{\left|f(x,y) - \hat{f}(x,y)\right|^2\right\} \tag{12.16}$$

where; $f(x, y)$ is the original image pixel value and $f(x,y)$ is the reconstructed image pixel value and the size of image is $M \times N$.

(iii) The PSNR has been accepted as a widely used measure of quality in the field of medical image compression. For each filtering operation, the measurement of ability to reduce the noise is defined by PSNR and it is the most appropriate parameter to judge the quality of compression. Higher the values of PSNR better the compression quality and vice versa. The PSNR is defined as:

$$\text{PSNR} = 10\log_{10}\left\{\frac{(255)^2}{\text{MSE}}\right\}\text{in dB} \tag{12.17}$$

(iv) The CoC suggests how closely the reconstructed image is correlated with an original image on a scale of 0–1. The closure the value of CoC to 1, higher the correlation of a compressed image to the original image is there, and vice versa. The CoC is defined as:

$$\text{CoC} = \frac{\sum_{x=1}^{m}\sum_{y=1}^{n} f(x,y)\hat{f}(x,y)}{\sqrt{\sum_{x=1}^{m}\sum_{y=1}^{n} f(x,y)^2}\sqrt{\sum_{x=1}^{m}\sum_{y=1}^{n} \hat{f}(x,y)^2}} \tag{12.18}$$

(v) The % improvement in any of the above performance parameter may be calculated as:

$$\%\ \text{Improvement} = \frac{\text{Standard method value} - \text{Proposed Method value}}{\text{Standard method value}} \times 100 \tag{12.19}$$

For example,

$$\%\ \text{Improvement in PSNR} = \frac{\text{Standard method PSNR} - \text{Proposed Method PSNR}}{\text{Standard method PSNR}} \times 100 \tag{12.20}$$

12.6 Results and Discussion

In this work, two aspects of wavelet-based compression, i.e., context-based CBDWT (contextual) and non-context-based DWT (general) have been considered. We have taken five different types of US images for the test purpose as shown in Fig. 12.11 and their dimensions are given in Table 12.1. We have varied compressions at various ratios from 8.0018:1 to 128.4705:1 and bpp from 1.0 to 0.0625

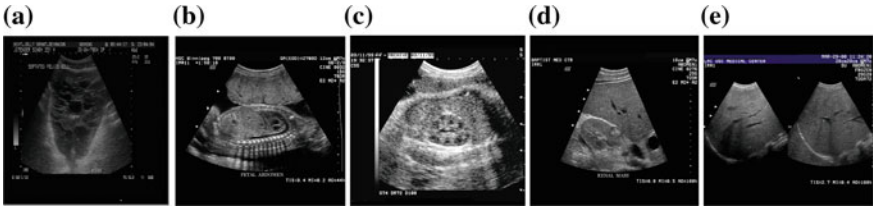

Fig. 12.11 Test US images considered. **a** US image—abdomen. **b** US image—fetal abdomen. **c** US image—Kidney. **d** US image—renal mass. **e** US image—liver (*source* (**a**) Jolly Grant Medical College, Dehradun—India—Radiology Department and (**b**)–(**e**) http://www.gehealthcare.com/usen/ultrasound/images)

Table 12.1 Radiology ultrasound (US) test images dimensions and their sizes in MBs

S. no.	Modality	Image dimension	Gray level (bits)	Size in bytes	Avg. size before compression (KB)
1	US 1	667 × 505	8	336,835	330
2	US 2	500 × 500	8	250,000	245
3	US 3	500 × 500	8	250,000	245
4	US 4	500 × 500	24	750,054	732
5	US 5	500 × 500	24	750,054	732

Table 12.2 Discrete wavelet transform (DWT) compression results (non-contextual)

S.no	bpp	Thr.	% of Transform Coeffs. Zeroed	% Energy retained in Comp. image	CR	RMSE	MSE	PSNR (dB)	COC
Image 12.1 (US)									
1	1.00	05	77.337428	99.953629	8.0018	4.4089	19.4382	35.2442	0.998693
2	0.50	10	86.609407	99.899008	16.0073	5.1673	26.7005	33.8656	0.998205
3	0.25	15	89.116633	99.852033	32.0293	5.9707	35.6491	32.6103	0.997602
4	0.125	20	90.381921	99.803962	64.1174	6.4790	41.9774	31.9007	0.997176
5	0.0625	25	91.105154	99.758311	128.4705	11.1248	123.7605	27.2050	0.991786
Image 12.2 (US)									
1	1.00	05	72.794800	99.965587	8.0316	4.7146	22.2279	34.6618	0.996799
2	0.50	10	80.438000	99.847422	16.0640	5.8724	34.4847	32.7545	0.995030
3	0.25	15	84.347600	99.675294	32.1318	7.2138	52.0395	30.9675	0.992490
4	0.125	20	86.710000	99.469579	64.2772	7.9565	63.3058	30.1164	0.990857
5	0.0625	25	88.189200	99.257358	128.6090	14.6258	213.9137	24.8284	0.968958
Image 12.3 (US)									
1	1.00	05	62.808000	99.983941	8.0316	4.8556	23.5772	34.4059	0.998784
2	0.50	10	73.424400	99.923911	16.0641	6.0814	36.9834	32.4507	0.998092
3	0.25	15	79.206000	99.832220	32.1313	7.5707	57.3149	30.5481	0.997042
4	0.125	20	83.035200	99.712728	64.2792	8.4030	70.6096	29.6422	0.996355
5	0.0625	25	85.587600	99.580198	128.6090	15.3593	235.9088	24.4034	0.987744
Image 12.4 (US)									
1	1.00	10	83.695200	99.856947	8.0005	4.3930	19.2988	35.2755	0.996570
2	0.50	20	88.569600	99.497280	16.0018	5.8153	33.8181	32.8393	0.993977
3	0.25	30	90.393600	99.107591	32.0072	7.1766	51.5041	31.0124	0.990814
4	0.125	40	91.278800	98.731780	64.0287	8.0043	64.0686	30.0644	0.988557
5	0.0625	50	91.768800	98.381878	128.1148	13.3977	179.4979	25.5902	0.967792
Image 12.5 (US)									
1	1.00	10	84.415600	99.798974	8.0004	4.6357	21.4893	34.8086	0.996052
2	0.50	20	89.371600	99.436238	16.0017	5.9700	35.6412	32.6113	0.993439
3	0.25	30	90.777600	99.131637	32.0068	7.1168	50.6487	31.0851	0.990662
4	0.125	40	91.398400	98.861179	64.0274	7.8675	61.8969	30.2141	0.988570
5	0.0625	50	91.762800	98.593629	128.1173	13.1881	173.9250	25.7272	0.967720

for the proposed CBDWT algorithm and the conventional DWT method. The compression performance parameters (IQM), namely bpp, CR, RMSE, MSE, PSNR, CoC, % transform coefficients zeroed and the % energy retained has been calculated for the conventional DWT algorithm and the proposed CBDWT algorithm and are listed in tabular form in Tables 12.2 and 12.3 respectively. A comparative analysis of the IQM parameters for the proposed CBDWT algorithm and the conventional DWT method has been depicted in Table 12.4. As a comparison of our results with the standard methods, the PSNR versus bpp variations have been compared with the results given in [15], EBCOT, Implicit and Maxshift methods as shown in Table 12.5 and are plotted in Fig. 12.20 and the proposed CBDWT's PSNR performance (both subjective and objective) in the bpp range of 1.0–0.0625 is improved as compared to other standard methods given in [13, 15, 23].

Table 12.3 Context based (CBDWT) compression results

Image Seq.	Entire image parameters					'CROI' performance parameters					
	bpp	Final CR	MSE	PSNR (dB)	CoC	CR CROI	% of CROI Transform Coeffs. Zeroed	% Energy CROI Retained	MSE	PSNR	CoC
US 1	1.00	08.0004	67.8459	29.8156	0.982827	6.8673	79.837804	99.984838	1.3545	46.8129	0.999378
	0.50	16.0018	76.5021	29.2941	0.980652	10.3715	87.439847	99.915275	2.9945	43.3675	0.998598
	0.25	32.0068	91.1384	28.5338	0.976844	13.2560	91.506280	99.812427	4.4569	41.6405	0.997902
	0.125	64.0293	119.2071	27.3678	0.969623	15.6268	93.656117	99.704210	5.8599	40.4519	0.997234
	0.0625	128.1173	156.0226	26.1989	0.960702	17.9454	94.996582	99.591914	7.2077	39.5528	0.996590
US 2	1.00	8.0005	87.7818	28.6968	0.987135	5.7225	78.992903	99.994333	0.3981	52.1306	0.999935
	0.50	16.0018	109.3778	27.7415	0.983946	7.7239	84.697996	99.973132	2.7252	43.7768	0.999486
	0.25	32.0073	134.5383	26.8423	0.980196	9.5249	87.869796	99.940010	4.5150	41.5842	0.999142
	0.125	64.0293	176.5479	25.6622	0.973955	10.9529	89.919382	99.897786	6.5743	39.9523	0.998744
	0.0625	128.1173	258.1818	24.0115	0.961867	12.3567	91.399103	99.847091	8.5986	38.7865	0.998355
US 3	1.00	8.0018	81.3243	29.0286	0.995776	3.9587	68.171174	99.997745	0.5084	51.0689	0.999974
	0.50	16.0073	101.2525	28.0767	0.994730	5.2032	75.487550	99.988928	1.2308	47.2289	0.999929
	0.25	32.0293	126.5600	27.1078	0.993410	6.1246	79.992135	99.973774	4.0530	42.0530	0.999755
	0.125	64.1096	170.8996	25.8034	0.991086	7.0476	83.173230	99.952627	8.3900	38.8932	0.999488
	0.0625	128.4705	261.0524	23.9635	0.986356	7.9340	85.475573	99.927308	11.9241	37.3665	0.999268
US 4	1.00	8.0007	70.1807	29.6686	0.987279	7.4182	83.725220	99.995347	0.3150	53.1482	0.999938
	0.50	16.0017	88.1514	28.6785	0.983989	9.8159	87.786854	99.976858	2.2234	44.6607	0.999490
	0.25	32.0068	104.4005	27.9438	0.981000	12.0083	90.198595	99.946109	3.8128	42.3184	0.999118
	0.125	64.0293	137.3348	26.7530	0.974936	13.7605	91.839551	99.905005	5.6151	40.6373	0.998696
	0.0625	128.1173	208.9214	24.9310	0.961908	15.4464	93.036484	99.854968	7.3432	39.4719	0.998291

(continued)

Table 12.3 (continued)

Image Seq.	Entire image parameters					'CROI' performance parameters					
	bpp	Final CR	MSE	PSNR (dB)	CoC	CR CROI	% of CROI Transform Coeffs. Zeroed	% Energy CROI Retained	MSE	PSNR	CoC
US 5	1.00	8.9395	74.7135	29.3968	0.985957	6.3722	79.195967	99.992384	0.4034	52.0740	0.999911
	0.50	17.8813	91.5689	28.5133	0.982755	8.7449	85.194931	99.963633	2.6233	43.9424	0.999332
	0.25	35.7709	110.4883	27.6976	0.979148	10.9972	88.775361	99.915256	4.4328	41.6640	0.998861
	0.125	71.5776	144.0596	26.5454	0.972735	12.8753	91.108092	99.853130	6.3849	40.0793	0.998354
	0.0625	133.2896	206.7742	24.9758	0.960865	14.7510	92.779079	99.779576	8.3267	38.9260	0.997849

Table 12.4 Comparative analysis of IQM parameters for DWT and CBDWT

CODEC →			DWT				CBDWT			
S.no.	Image seq.	bpp	MSE	PSNR (dB)	CoC	% Eng. Retained	MSE	PSNR (dB)	CoC	% Eng. Retained
1	US1	1.00	19.4382	35.2442	0.998693	99.953629	1.3545	46.8129	0.999378	99.984838
2		0.50	26.7005	33.8656	0.998205	99.899008	2.9945	43.3675	0.998598	99.915275
3		0.25	35.6491	32.6103	0.997602	99.852033	4.4569	41.6405	0.997902	99.812427
4		0.125	41.9774	31.9007	0.997176	99.803962	5.8599	40.4519	0.997234	99.704210
5		0.0625	123.7605	27.2050	0.991786	99.758311	7.2077	39.5528	0.996590	99.591914
6	US2	1.00	22.2279	34.6618	0.996799	99.965587	0.3981	52.1306	0.999935	99.994333
7		0.50	34.4847	32.7545	0.995030	99.847422	2.7252	43.7768	0.999486	99.973132
8		0.25	52.0395	30.9675	0.992490	99.675294	4.5150	41.5842	0.999142	99.940010
9		0.125	63.3058	30.1164	0.990857	99.469579	6.5743	39.9523	0.998744	99.897786
10		0.0625	213.9137	24.8284	0.968958	99.257358	8.5986	38.7865	0.998355	99.847091
11	US3	1.00	23.5772	34.4059	0.998784	99.983941	0.5084	51.0689	0.999974	99.997745
12		0.50	36.9834	32.4507	0.998092	99.923911	1.2308	47.2289	0.999929	99.988928
13		0.25	57.3149	30.5481	0.997042	99.83222	4.0530	42.0530	0.999755	99.973774
14		0.125	70.6096	29.6422	0.996355	99.712728	8.3900	38.8932	0.999488	99.952627
15		0.0625	235.9088	24.4034	0.987744	99.580198	11.9241	37.3665	0.999268	99.927308
16	US4	1.00	19.2988	35.2755	0.99657	99.856947	0.315	53.1482	0.999938	99.995347
17		0.50	33.8181	32.8393	0.993977	99.49728	2.2234	44.6607	0.99949	99.976858
18		0.25	51.5041	31.0124	0.990814	99.107591	3.8128	42.3184	0.999118	99.946109
19		0.125	64.0686	30.0644	0.988557	98.73178	5.6151	40.6373	0.998696	99.905005
20		0.0625	179.4979	25.5902	0.967792	98.381878	7.3432	39.4719	0.998291	99.854968
21	US5	1.00	21.4893	34.8086	0.996052	99.798974	0.4034	52.074	0.999911	99.992384
22		0.50	35.6412	32.6113	0.993439	99.436238	2.6233	43.9424	0.999332	99.963633
23		0.25	50.6487	31.0851	0.990662	99.131637	4.4328	41.6640	0.998861	99.915256
24		0.125	61.8969	30.2141	0.988570	98.861179	6.3849	40.0793	0.998354	99.853130
25		0.0625	173.925	25.7272	0.967720	98.593629	8.3267	38.9260	0.997849	99.779576

Table 12.5 Comparison of CROI's PSNR in (dB) for different methods

S. no.	Bit rate	EBCOT	Implicit	Maxshift	Yang et al. [15]	DWT	CBDWT
1	0.0625	23.399	24.191	27.684	24.486	24.8284	38.7865
2	0.125	24.959	27.259	30.089	28.178	30.1164	39.9523
3	0.25	27.698	30.869	33.486	32.583	30.9675	41.5842
4	0.5	31.524	35.882	38.911	38.097	32.7545	43.7768
5	1.0	37.775	43.168	47.711	37.775	34.6618	52.1306

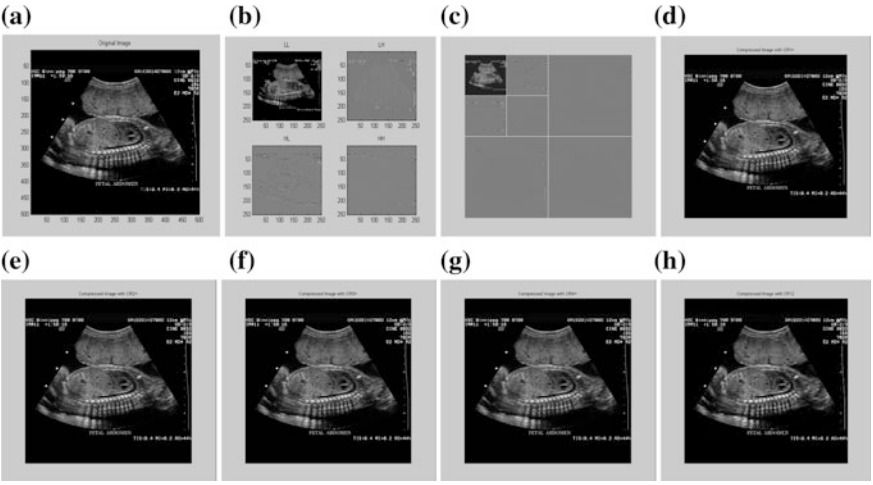

Fig. 12.12 US images compressed with DWT at different bit rates and CRs. **a** Original filtered image. **b** Level one decomposition. **c** Level two decomposition. **d** Compressed Image with bpp = 1.00. **e** Compressed image with bpp = 0.5. **f** Compressed image with bpp = 0.25. **g** Compressed image with bpp = 0.125. **h** Compressed image with bpp = 0.0625

The compression output results of the conventional DWT algorithm and the proposed CBDWT algorithm are shown through Figs. 12.12a–h and 12.17 a–l respectively for different bpp and CRs for US2. In Fig. 12.13a–d, the histograms and the envelope plots of the original image (US2) and the reconstructed image by DWT method at compression ratio 8.0018:1 are plotted which correlate the pixel densities and the gray levels (maximum gray level value is 255 and the minimum is 0 for an 8 bit image) of the original and the reconstructed images. The Fig. 12.18a–d and 12.19a–d represents the histogram and envelope plots of the original image (US2), contextual region (CROI), background (BG), and the reconstructed image at compression ratio (CR3) 32.1318:1. These histogram plots correlate the pixel densities and the gray levels of the original and the reconstructed images and give the information of the truncation of the low gray level coefficients. The graphical DWT and CBDWT results of different IQM parameters (PSNR, CoC, MSE, CR) versus bit

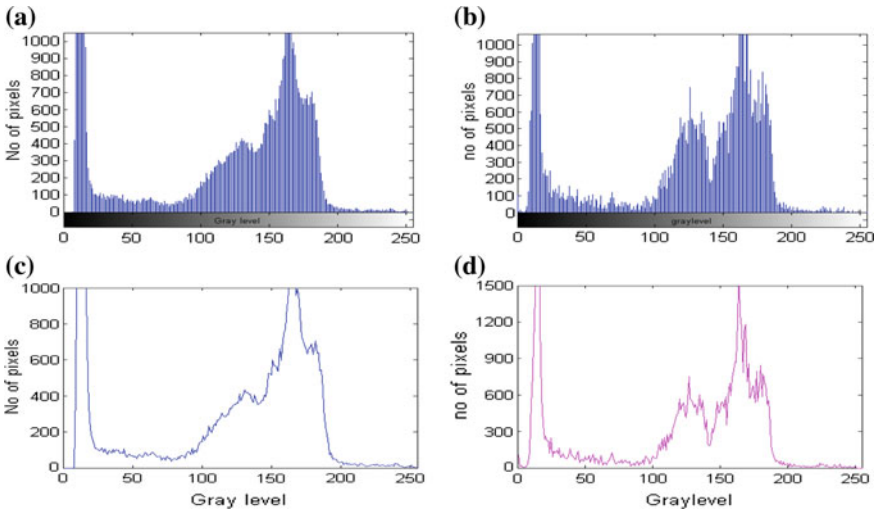

Fig. 12.13 Histograms of US image. **a** Histogram of original image. **b** Histogram of reconstructed image (CR = 8.0005). **c** Envelope original image. **d** Envelope of reconstructed image @ CR = 8.0005

rate variations for different US images considered are shown through Figs. 12.14a–f and 12.15a–d respectively. The comparative IQM performance parameters variations of DWT and CBDWT algorithms for different US image sequences are shown through Fig. 12.16a–d where the proposed CBDWT algorithm's results outperform the DWT compression performance.

In proposed contextual region of interest (CROI) coding method by CBDWT algorithm, both the desired aspects of medical image compression have been taken care of, i.e., high quality of compressed image and the high compression rates. As shown in Fig. 12.17k, the original US2 image has been reconstructed after decompression with a very high CoC (0.998744) with a PSNR value of 39.9523 at bpp = 0.125 and CR = 64.0293 which is the best available in all the methods like DWT, EBCOT, Implicit and Maxshift methods [13, 15, 23]. This compression will reduce almost 64 times the storage and the transmission cost! [2]. Now Imagine, at the CR = 128.1173 and bpp = 0.0625, the storage and transmission cost will reduce to almost 128 times with an acceptable image quality. So, in future the contextual coding will be a feasible solution of huge medical image data to store and transmit without losing the image quality. The proposed CBDWT method performs excellently at low bit rates as compared to other similar methods which are shown in Table 12.5 and Fig. 12.20. As shown in Fig. 12.15a–d the IQM performance (CR, MSE, PSNR CoC) with bpp in the CROI region is better as compared to the entire image area. The BG area is compressed heavily as compared to the CROI from CR = 5.7225 to 12.3567 to maintain the over all moderate CR = 8.0018:1 to 128.4705:1, so as a consequence the over all good CR is achieved which is a very high CR as compared to existing standard methods [13, 15, 23] along with the

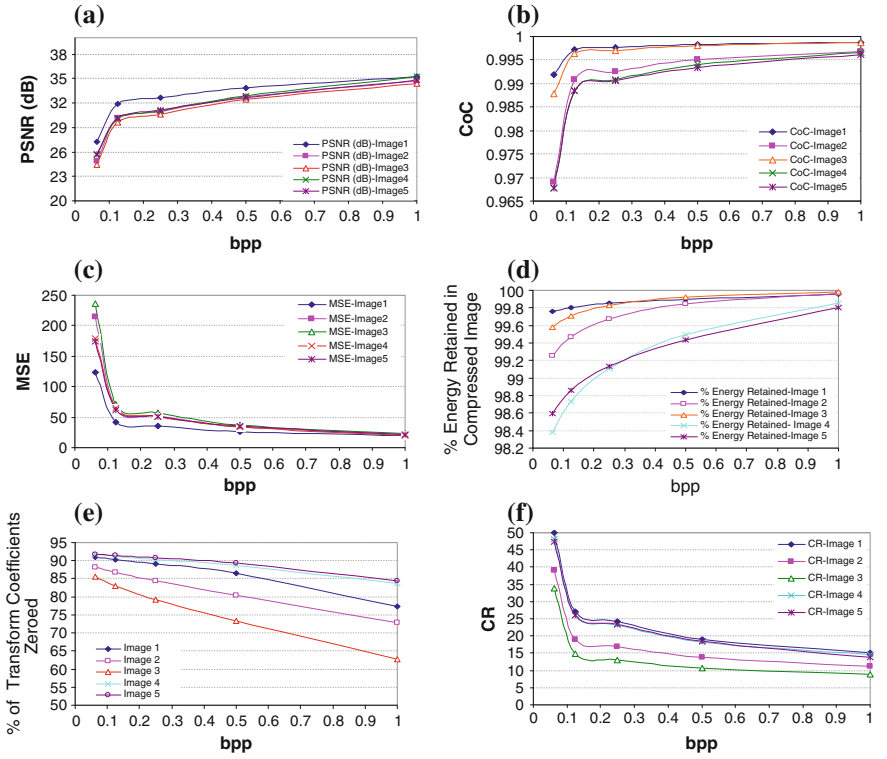

Fig. 12.14 IQM performance variations versus bpp for **a** PSNR. **b** CoC. **c** MSE. **d** % Energy retained. **e** % Transformed coefficients. **f** Compression ratio (CR)

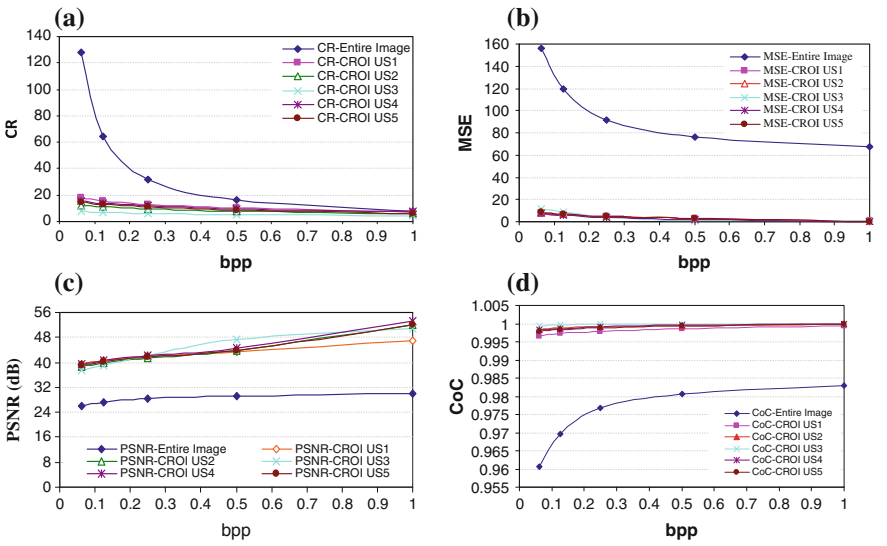

Fig. 12.15 IQM performance variations versus bpp for **a** CR. **b** MSE. **c** PSNR. **d** CoC

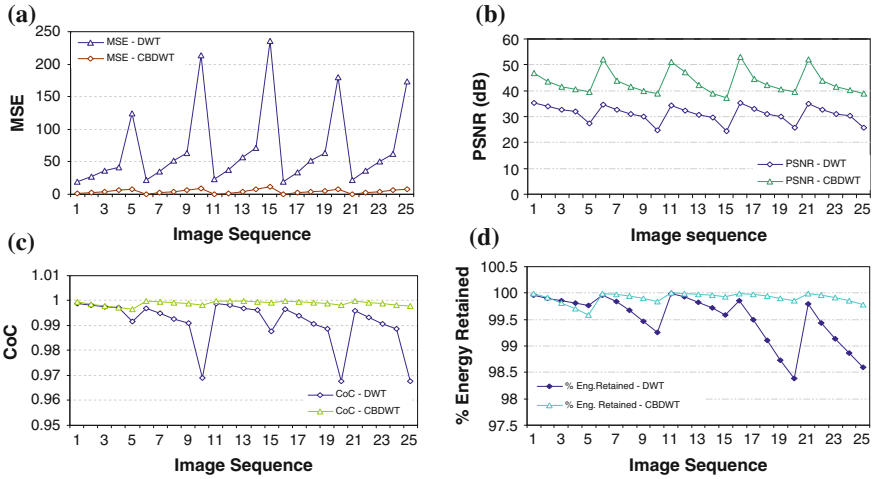

Fig. 12.16 IQM performance parameters variations for DWT and CBDWT taken as all test US images. **a** MSE. **b** PSNR. **c** CoC. **d** % Energy retained

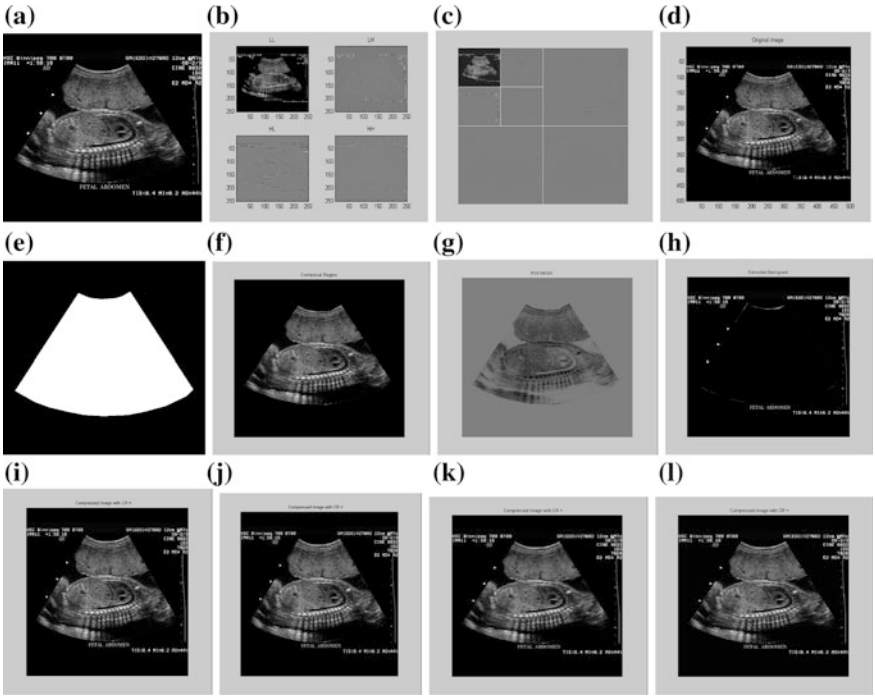

Fig. 12.17 US images compressed with CBDWT at different bpp and CRs. **a** Original US image. **b** Level one filtering. **c** Level two. **d** Filtered image. **e** ROI mask generated. **f** CROI separated. **g** Difference CROI. **h** Separated background. **i** Compressed (ROI+BG): CR2. **j** Compressed: CR3. **k** Compressed: CR4. **l** Compressed: CR5

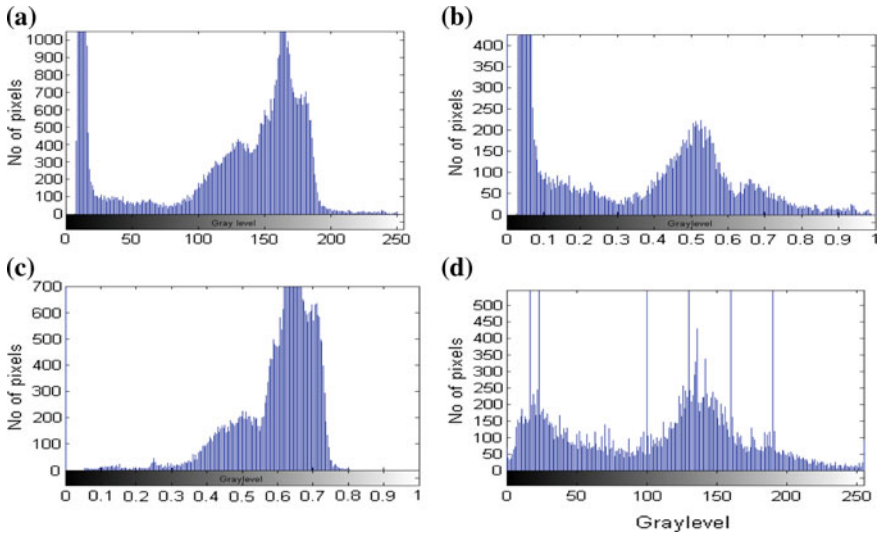

Fig. 12.18 Histogram analysis of US image. **a** Original image histogram. **b** Contextual region. **c** Background region. **d** Reconstructed image @ CR = 32.1318

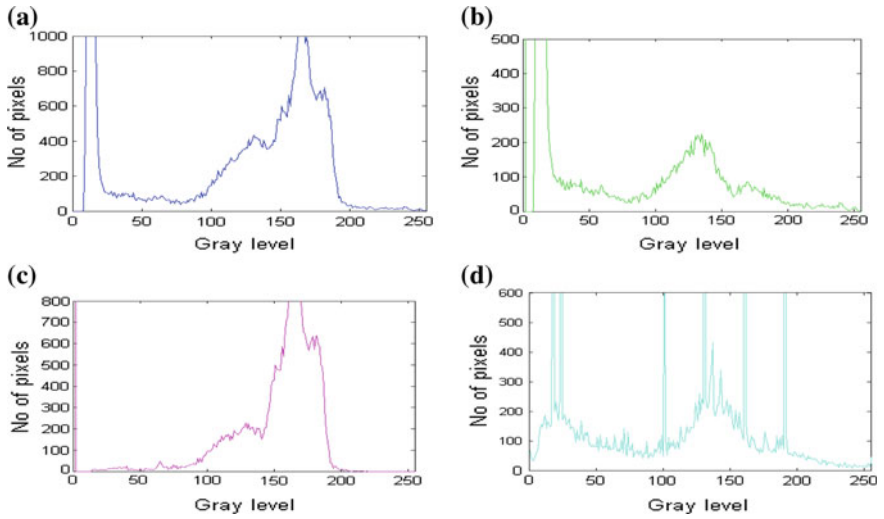

Fig. 12.19 Graphical analysis. **a** Envelope of original image. **b** Envelope of contextual region. **c** Envelope of background region. **d** Envelope of reconstructed image @ 32.1318:1

appreciable image quality. A comparison of DWT and CBDWT compressed image is shown in Fig. 12.21 which shows that the CBDWT compressed image is far better in visual quality as compared to the DWT compressed image for the same CR value.

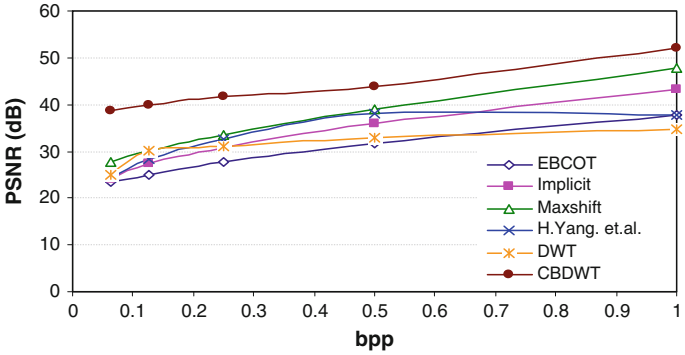

Fig. 12.20 Comparison of ROI's PSNR (dB) for different methods for ultrasound Image

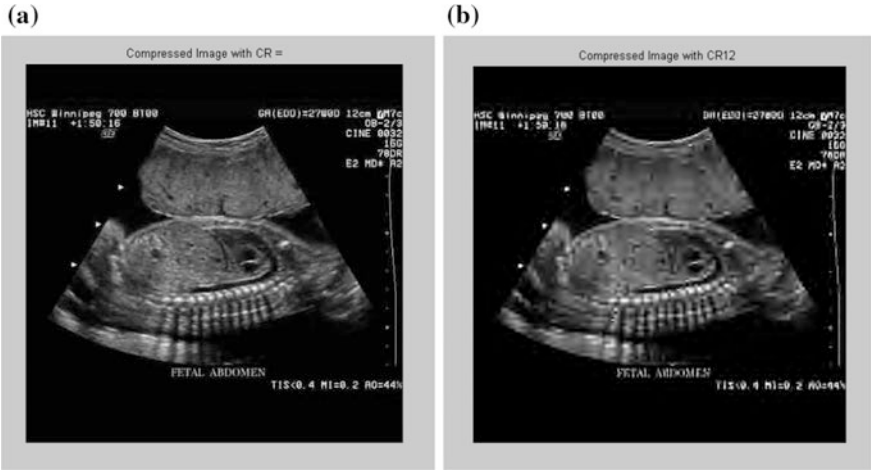

Fig. 12.21 (i) The comparison of the visual quality of the DWT and CBDWT compressed images at a common bpp = 0.0625. **a** The CBDWT compressed image. **b** The DWT compressed image. (ii) The CBDWT compressed image has a CR value of 128.1173 and PSNR 38.7865 dB while the DWT compressed image has a CR value of 128.6090 and PSNR 24.8284 dB for the same bpp = 0.0625

12.7 Conclusions

The practical limitations of the conventional medical image compression methods and new trends in the technology have given a new way of context-based medical image compression. In this paper, the proposed wavelet transform-based CBDWT (context-based discrete wavelet transform) coding method has given excellent results by maintaining the desired image quality at low bit rates as well as the high compression ratios by selective context based compression. The proposed method

will prove very useful in high diagnostic medical image compression without losing any useful information. The results obtained in the proposed context based (CBDWT) compression method given in Tables 12.3 and 12.4 clearly show the improved performance in terms of CR, MSE, PSNR, CoC, and the visual quality of image at almost all compression ratios and bpp as compared to other methods given in [13, 15, 23]. The PSNR and bpp comparative results of different standard methods (EBCOT, Implicit and Maxshift) compared in [15] given in Table 12.5 and Fig. 12.20 show improved performance of the proposed method at all bit rates. It is found that the proposed method gives a significant improvement in the PSNR value in the bpp range of 0.0625–1.00 and the CR value of as high as from 8.0018:1 to 128.4705:1 with a better quality of the reconstructed medical image judged on the basis of the IQM and HVS. Thereby, the proposed method can maintain excellent rate distortion performance as well as high image fidelity. So, finally we can conclude that the proposed CBDWT method is very suitable for low bpp and high CR compression as well as can perform lossy and lossless CROI coding along with high PSNR, CoC, low MSE, and good visual quality of the reconstructed medical image. It can also maintain the high diagnostic quality of the compressed image data and hence can reduce heavily the transmission and the storage costs of the huge medical data generated everyday and will be well suited for telemedicine and teleradiology application over limited BW networks.

References

1. Strom J, Cosman PC (1997) Medical image compression with lossless regions of interest. Elsevier. Signal Processing 59(2):155–171
2. Ramin K (2004) Image compression in your PACS: should you do it? what are the issues? J Am Coll Radiol 1(10):780–781
3. Pennebaker WB, Mitchell JL (1993) JPEG: still image data compression standard. Van Nostrand Reinhold, New York
4. Ahmed N, Natarajan T, Rao KR (1974) Discrete cosine transform. IEEE Trans Comput 23:90–93
5. Mallat S (1989) A theory for multiresolution signal decomposition: The wavelet representation. IEEE Trans Pattern Anal Mach Int 11(7):674–693
6. Shapiro JM (1993) Embedded image coding using zerotrees of wavelet coefficients. IEEE Trans Signal Processing 41(12):3445–3462
7. Said A, Pearlman WA (1996) A new, fast and efficient image codec based on set partitioning in hierarchical trees. IEEE Trans Circ Syst Video Technol 6(3):243–250
8. Taubman D (2000) High performance scalable image compression with EBCOT. IEEE Trans Image Processing 9(7):1151–1170
9. ISO/IEC (2001) Final Draft International Standard 15444-2, ITU Recommendation T.801. JPEG 2000 Image Coding System: Part II Extensions, Aug 2001
10. ISO/IEC (2000) International Standard 15444-1, ITU Recommendation T.800. JPEG 2000 Image Coding System
11. Hsin HC, Li CC (2003) Image coding with modulated wavelets. Elsevier Sci: Pattern Recogn Lett 24:2391–2396
12. Atsumi E, Farvardin N (1998) Lossy/lossless region-of-interest image coding based on set partitioning in hierarchical trees. In: Proceedings of IEEE international conference of image processing, pp 87–91, Chicago, USA

13. Park K, Park HW (2002) Region-of-interest coding based on set partitioning in hierarchical trees. IEEE Trans Circ Syst Video Technol 12(2):106–113
14. Askelof J, Carlander M, Christopoulos C (2002) Region of interest coding in JPEG2000. Sig Process Image Commun 17:105–111
15. Yang H, Long M, Tai H-M (2005) Region-of-interest image coding based on EBCOT. In: IEEE Proceedings on visual image signal processing, vol 152, No. 5, Oct 2005
16. Antonini M, Barlaud M, Mathieu P, Daubechies I (1992) Image coding using wavelet transform. IEEE Trans Image Process 1:205–220
17. Christopoulos C, Askelof J, Larsson M (2000) Efficient methods for encoding regions of interest in the upcoming JPEG 2000 still image coding standard. IEEE Signal Process Lett 7(9):247–249
18. Zhang L, Yu X (2006) Multiple regions of interest image coding using compensation scheme and alternating shift. In: The 18th IEEE computer society international conference on pattern recognition (ICPR'06), vol 3, pp 758–761
19. Liu L, Fan G (2003) A new JPEG 2000 region of interest image coding method: Partial Significant Bitpanes Shift. IEEE Signal Process Lett 10(2):35–38
20. Signoroni A, Lazzaroni F, Leonardi R (2003) Exploitation and extension of the Region-of-Interest coding functionalities in JPEG2000. IEEE Trans Consum Electron 49 (4):818–823
21. Ansari MA, Anand RS (2008) A novel ROI based algorithm with DCT, wavelet transform and set partioning in hierarchial trees for medical image compression. Int J Sci Comput (IJSC) 2(1):7–22
22. Ansari MA, Anand RS (2008) Implementation of efficient medical image compression algorithms with JPEG. wavelet transform and SPIHT. Int J Comput Intell Res Appl (IJCIRA) 2(1):43–55
23. Yan X et al (2004) The coding technique of image with multiple ROIs using standard maxshift method. In: The 30th annual conference of the IEEE industrial electronics society, Busan, Korea, pp 2077–2080
24. Wang Y, Li H, Xuan J, Lo SCB, Mun SK (1997) Modeling of wavelet coefficients in medical image compression. In: Proceedings of international confonerence image processing, vol 1, pp 644–647
25. Tian DZ, Ha MH (2004) Applications of wavelet transform in medical image processing. In: IEEE Proceedings of international conference on machine learning and cybernetics, vol 3, pp 1816–1821
26. Chen Y-T, Tseng D-C, Chang P-C (2005) Wavelet-based medical image compression with adaptive prediction. In: IEEE Proceedings of international symposium on intelligent signal processing and communication systems (ISPACS'05), pp 825–828
27. Yelland MR, Aghdasi F (1999) Wavelet transform for medical image compression. IEEE AFRICON 1:303–308
28. Gonzalez RC, Woods RE (2002) Digital image processing, 2nd edn. Pearson Education, New Delhi
29. Sedig A, Balasubramanian Vittal, Morales Aldo (2007) Semi-automatic region of interest identification algorithm using wavelets. J Opt Eng 46(3):035003–035006
30. Vlahakis V, Kitney RI (1997) Wavelet-based, inhomogeneous, near-lossless compression of ultrasound images of the heart. In: Proceedings of the IEEE international conference on computers in cardiology, pp 549–552
31. Saint-Marc P, Chen J-S, Medioni G (1991) Adaptive smoothing: a general tool for early vision. IEEE Trans Pattern Anal Mach Intell 13(6):514–529
32. Crus DS, Ebrahimi T, Larsson M, Askelöf J, Christopoulos C (1999) Region of interest coding in JPEG2000 for interactive client/sever applications. In: Proceedings of IEEE 3rd workshop on multimedia signal process, pp 389–394

33. Zhang L, Yu X, Wang S (2006) New region of interest image coding based on multiple bitplanes up-down shift using improved SPECK algorithm. In: Proceedings of the 1st IEEE international conference on innovative computing, information and control (ICICIC '06), vol 3, pp 629–632
34. Doukas C, Ilias Maglogiannis (2007) Region of interest coding techniques for medical image compression. IEEE Eng Med Biol Mag 26(5):29–35
35. Taubman D, Marcellin M (2002) JPEG2000: image compression fundamentals, standards and practice. Kluwer, Norwell, MA

Chapter 13
Srinivasa Ramanujan: A Creative Genius

K. Srinivasa Rao

Abstract This is a brief review article about the life and work of the peerless mathematician Srinivasa Ramanujan and CD ROMs on his life and work.

Keywords Trigonometry · Hypergeometric series · Theta function

13.1 Introduction

Srinivasa Ramanujan, who left behind more than 4000 theorems as his indelible mathematical discoveries, is a creative genius, who has no peer in the world of mathematicians. His life and work is a great story, which had inspired generations of mathematicians the world over and his mathematical theorems are treasures to cherish.

Srinivasa Ramanujan was born at 6 p.m. on Thursday, December 22, 1887, at Erode, the parental home of his mother, Komalathammal. His father, K. Srinivasa Iyengar, was a 'gumasta,' or a clerk, to a cloth merchant in Kumabakonam. His mother believed that her first son was a gift of Goddess Namagiri of Namakkal, her family's deity, in the Lakshmi Narasimhar Temple. The name of the town Namakkal (360 km southwest of Chennai) derives from Namagiri, which is the name of a single rock formation at the center of the historic town. The day on which this first son of Komalathammal and Srinivasa Iyengar was born was a Thursday, coinciding with the day on which the Vaishnavite saint and founder of the Visishtadvaita philosophy, Ramanujachariar (1017–1137 A.D.) was also born. So, the name Ramanujan was an automatic choice for the newborn.

K. Srinivasa Rao (✉)
Institute of Mathematical Sciences, Chennai 600113, India
e-mail: ksrao18@gmail.com

K. Srinivasa Rao
Srinivasa Ramanujan Academy of Maths Talent, Sankaralayam, Second Floor,
66, Mayor Ramanathan (Spur Tank) Road, Chetpet, Chennai 600031, India

A.H. Siddiqi et al. (eds.), *Mathematical Models, Methods and Applications*,
Industrial and Applied Mathematics, DOI 10.1007/978-981-287-973-8_13

Ramanujan's mother could recite a thousand verses from the 'Nalaayira Divya Prabandham.' These were compositions by 12 saints, known as 'Azhwaars,' in praise of Lord Maha Vishnu and she took part in the group signing of these at the Sarangapani Temple, in Kumbakonam. This extraordinary memory of the mother was an inherited talent of Ramanujan. For, according to his mentor G.H. Hardy, Ramanujan could recall anyone of the more than 4000 theorems he discovered, at will, and he could provide not one but several proofs to them, if anyone asked him for a proof of an entry in his notebooks.

Precocious at School, Ramanujan won several prizes in his II, IV, and VI forms at Town High School, Kumbakonam, for 'Proficiency in Mathematics' and as a reward of merit and an incentive for further improvement. S.L. Loney's 'Trigonometry', was a book he mastered in form VI. Two boarders in his parental home, who were college students borrowed for him George Shoobridge Carr's: 'A Synopsis of Elementary Results', a book on Pure Mathematics, a compilation of propositions, formulae, and methods of analysis with abridged demonstrations, published in 1886—a useful book for all those writing the Mathematical Tripos examinations of the Cambridge University, at that time. These two books perhaps made an indelible impression on the mind of Ramanujan. He noted his discoveries in his notebooks without proofs. By one reckoning, 3254 entries are in his Notebooks, and generations of mathematicians had studied and continue to wonder in awe at how Ramanujan, without formal education, could discover such an incredibly vast number of profound theorems.

In a short life span of 32 years, 4 months and 4 days, Ramanujan published 37 research papers, of which 7 were in collaboration with his friend Professor G.H. Hardy of Trinity College, Cambridge. Ramanujan also proposed 59 Questions or Answers to Questions in the Journal of the Indian Mathematical Society (JIMS), in which he published his first 5 papers and in all 11 out of his 37 papers. This journal was started by V. Ramaswamy Iyer, the founder of the Indian Mathematical Society, in 1907, possibly with the intention of helping Ramanujan with a journal, suitable for the dissemination of his prolific research work.

Ramanujan's work in the area of 'Partitions' and on what he called as 'mock' theta functions opened up two new avenues for research. Chapter XII of the first notebook and Chapter X of the second notebook of Ramanujan are devoted to generalized hypergeometric series. What is amazing is that in this area he starts with an entry which gives the most general summation theorem in mathematics known till date, the ${}_7F_6(1)$ summation theorem and from it deduced all other known summation theorems including the Gauss summation theorem, discovered in 1812. This theorem is known today as the Dougall–Ramanujan summation theorem. Such a prodigious feat, of discovering all that was known at that time in the world on hypergeometric series, with just a hint in Carr's synopsis of the Gauss summation theorem is an unprecedented prodigious feat, unparalleled in the annals of mathematics.

After his successful schooling though he entered the Government Arts College in Kumbakonam, his lack of interest in the collegiate education, ended with his failure at the 1905 and 1907 examinations of the University of Madras. Thus, formally he is a failure in the First degree in Arts (F.A.) examinations.

During 1906–1912, Ramanujan was constantly in search of a benefactor and a job to eke out a livelihood. He tutored a few students in mathematics, in Kumbakonam, and later even sought employment as a Tutor in mathematics. Disappointed at the lack of recognition, during this trying period in his life, Ramanujan lamented to a friend of his that he was probably destined to die in poverty like Galileo! Only in the centenary year, the world came to know, through the revelation of Dr. S. Chandrasekhar, the Nobel Laureate, at his Ramanujan centenary address, that Ramanujan even attempted suicide by lying down on the train tracks in London; and he was rescued by the driver by bringing the train to a halt and handing Ramanujan over to the police. When Ramanujan revealed his connection with Professor Hardy, who was summoned to the Station, the policeman was generous to ignore the incident, since he did not want to harm the future of a young Indian mathematician.

Ramanujan received constant support and encouragement from his mother, though his father was not impressed with his preoccupation with mathematics always. By his own untiring efforts, Ramanujan garnered the support of the then Collector of Nellore, Diwan Bahadur B. Ramachandra Rao, when Ramanujan called on him with a classmate of his, and got an audience with him on his fifth attempt. The condescension and offer of Rs. 20 per month as a dole was quietly rejected by Ramanujan, who due to the efforts of his mentor, S. Narayana Iyer got a job at the Accountant General's office for a month, and later, a class-III, grade-IV clerical post in the Madras Post Trust, in 1913.

Ramanujan saw the book 'Orders of Infinity' of Godfrey Harold Hardy, at the Presidency College, Madras, when he called on Professor P.V. Seshu Iyer, his mathematics professor at Kumbakonam. Browsing through it, he read that "no definite expression has yet been found for the number of primes less than any given number." An excited Ramanujan told Seshu Iyer that he had that exact formula in his notebooks. Ramanujan was then asked to write to Professor Hardy and that historic (January 16, 1913) first letter of his started with "Dear Sir, I beg to introduce myself to you as a clerk in the Accounts Department of the Port Trust Office…". The attached 11 pages containing about a hundred theorems selected from his notebooks was sufficient to convince Hardy and his friend J.E. Littlewood, to put in a stupendous effort to make Ramanujan agree to go to Cambridge.

The formulae in the eleven pages attached with Ramanujan's first letter to G.H. Hardy, in January 1913, contained theorems *on a different level and obviously both deep and difficult* which even an exceptional but conventional mathematician like Hardy *had never seen anything in the least like them before*. Hardy stated[1] that *they defeated me completely*. Those gems are:

If $u = \frac{x}{1} + \frac{x^5}{1} + \frac{x^{10}}{1} + \frac{x^{15}}{1} + \cdots$ and $u = \frac{x^{1/5}}{1} + \frac{x}{1} + \frac{x^2}{1} + \frac{x^3}{1} + \cdots$

[1] *Rmanujan: Twelve lectures on subjects suggested by his life work, G.H. Hardy, Chelsea, N.Y. (1940) p. 9.*

then

$$v^5 = u\frac{1-2u+4u^2-3u^3+u^4}{1+3u+4u^2-2u^3+u^4} \tag{13.1}$$

$$\frac{1}{1}+\frac{e^{-2\pi}}{1}+\frac{e^{-4\pi}}{1}+\cdots=\left\{\sqrt{\left(\frac{5+\sqrt{5}}{2}\right)-\left(\frac{1+\sqrt{5}}{2}\right)}\right\}e^{2\pi/5}. \tag{13.2}$$

and

$$\frac{1}{1}+\frac{e^{-2\pi\sqrt{5}}}{1}+\frac{e^{-4\pi\sqrt{5}}}{1}+\cdots=\left\{\frac{\sqrt{5}}{1+\sqrt[5]{\left\{5^{3/4}\left(\frac{\sqrt{5}-1}{2}\right)^{5/2}-1\right\}}}-\left(\frac{1+\sqrt{5}}{2}\right)\right\}e^{2\pi/\sqrt{5}}. \tag{13.3}$$

Hardy stated that *a single look at them is enough to show that they could be written down by a mathematician of the highest class. They must be true because, if they are not true, no one would have had the imagination to invent them. Finally, (you must remember that I knew nothing whatever about Ramanujan, and I had to think of every possibility), the writer must be completely honest, because great mathematicians are commoner than thieves or humbugs of such incredible skill.* Hardy considered it *sufficiently marvelous that he had not even dreamed of problems such as these and Rogers and Watson found the proofs of the extremely difficult theorems, in the later years.*

The Madras University deserves kudos for rising time and again to provide all the financial support required: for Ramanujan's passage to England; for his stay for 5 years (1914–1919) in Cambridge; continuing support to him, after his return to India for a year; and for providing a pension to Janaki, Ramanujan's wife, till she died 74 years after Ramanujan's untimely death—due to undiagnosed, untreated, hepatic amoebiasis—on April 26, 1920.

13.2 CD ROMs on the Life and Work of Srinivasa Ramanujan

Ramanujan's work has inspired generations of mathematicians. There are three journals named after him—Journal of the Ramanujan Mathematical Society, The Hardy-Ramanujan Journal, and the Ramanujan Journal—and hundreds of papers have appeared and continue to appear based on Ramanujan's work. These facts reveal the enduring nature of his remarkably stupendous contributions to mathematics. The "Collected Papers of Srinivasa Ramanujan", edited by G.H. Hardy,

P.V. Seshu Aiyar, and B.M. Wilson, was originally published in 1927, and "Ramanujan: Twelve lectures on subjects suggested by his life and work", by Hardy was first published in 1940, which have both been reprinted in 1999 by AMS Chelsea Publishing, American Mathematical Society, Providence, Rhode Island, USA.

This clearly shows the relevance and importance of the original papers of Ramanujan in this millennium. The 125th birth anniversary of Ramanujan was inaugurated by the former Prime Minister, Dr. Manmohan Singh, who declared the year 2012 as the National Mathematics Year, released a commemorative stamp and the first day cover, and declared from December 22, 2012 onwards, the date of birth of Ramanujan will be celebrated as National Mathematics Day. A new edition of the renowned notebooks of Ramanujan was released. Mathematicians would hail these as significant contributions for the development of Mathematics.

Two CD ROMs were developed by The Institute of Mathematical Sciences (IMSc), Chennai and by the National Multimedia Resource Center (NMRC) of the Center for Development of Advanced Computing (C-DAC), Pune. This Project was sponsored by the Department of Science and Technology (DST), Government of India, with the concept and design of contents by Dr. K. Srinivasa Rao, Senior Professor (Retd.), IMSc., Chennai and at present the Director, Srinivasa Ramanujan Academy of Maths Talent. This was a two and a half-year (Two Million Rupees) project, at IMSc, between Dec. 2002 and April 2005. The first CD ROM, Part-1 contains all the available information about the life and work of Srinivasa Ramanujan, the very first digital scanned version of the original notebooks of Ramanujan, his 39 research publications and the Collected Papers of Srinivasa Ramanujan, as well as the 59 Questions or Answers to Questions of Ramanujan which appeared in the Journal of the Indian Mathematical Society. Dissemination of the CD ROMs was entrusted to: Director, Vigyan Prasar, C-24, Qutab International Area, New Delhi—110016. For further details, the following are the references to the books, on Ramanujan, by the author:

- Srinivasa Ramanujan: a Mathematical Genius,
  East West Books (1998 and revised 2004).
- Mathematical Genius Ramanujan,
  Allied Publishers Pvt. Ltd. (2005).
- A mathematical genius: Srinivasa Ramanujan,
  Srinivasa Ramanujan Academy of Maths Talent, Chennai (Dec. 2012).

Documentaries on Ramanujan have been made based on the life and work of Ramanujan by Nandan Kudhyadi of Nandan Kudhyadi Productions. The first of these was entitled, "The Enigma of Srinivasa Ramanujan," was made in the birth centenary year, 1987. The second, in which the author had a role to play as one of the resource persons for the producer–director, was entitled, "The Genius of Ramanujan" and released in March 2013 at the Indian Institute of Science Education and Research, Pune. A third Documentary is in the making and scheduled for Telecast on Rajya Sabha TV channel (on Dec. 22, 2014, at 1 P.M. and 5 P.M.,

duration 55 mins.) and it is entitled: “Conjectures: Ramanujan's spiritual realization in mathematics” and it will complete a trilogy of films by this reputed documentary film maker, who has made more than 65 documentary films including excellent ones on the Nobel Prizemen Dr. Sir C.V. Raman and Dr. S. Chandrasekhar. This quality conscious director’s other documentaries have been screened at International Film Festivals in Brussels, Hawaii, Paris, Tokyo, Vancouver, etc., and his ‘The Genius of Srinivasa Ramanujan’ was screened at the International Conference on Number Theory at New Delhi, Dec. 2012 and was also scheduled to be screened at the annual joint meeting of the American Mathematical Society and the Mathematical Association of America, at Baltimore, in January 2014.

Acknowledgments The author wishes to thank Professor A.H. Siddiqi for his kind invitation to deliver the keynote address and for bringing out the conference proceedings.

Chapter 14
Estimation of Longitudinal Diffusivity in Laminar/Turbulent Flow Through Curved Channels with Absorbing Boundaries Using Method of Moments

Sushil Kumar and Girija Jayaraman

Abstract Two-dimensional model for the flow hydrodynamics and mass transport is considered by Kalkwijk and de Vriend (J Hydraul Res 18(4):327–342, 1980) [20] in curved channels with absorbing boundaries. Longitudinal velocity dominates over the lateral flow for the case of laminar flow, while the transverse velocity is also considered with longitudinal velocity for turbulent flow. However, the transverse velocity is much less than the longitudinal velocity in the case of mildly curved channel flow. Longitudinal diffusivity is estimated using method of moments on the advection–diffusion equation governing the concentration of the diffusing substance, which gives substantial information about the concentration distribution of diffusing substance across the flow. It is observed that the effective dispersion coefficient is a function of the curvature parameter and absorbing parameter. It is found that the steady state of dispersion coefficient is achieved earlier in the case of turbulent flow than in the case of laminar flow. Effective dispersion coefficient incorporates the combined effects of wall curvature and absorption on boundaries.

Keywords Turbulent flow · Curved channel · Wall absorption · Method moments

S. Kumar (✉)
Department of Applied Mathematics, School of Vocational Studies and Applied Sciences, Gautam Buddha University, Greater Noida 201312, UP, India
e-mail: sushil.kumar@gbu.ac.in

G. Jayaraman
Centre for Atmospheric Sciences, Indian Institute of Technology, Delhi 110016, New Delhi, India

A.H. Siddiqi et al. (eds.), *Mathematical Models, Methods and Applications*, Industrial and Applied Mathematics, DOI 10.1007/978-981-287-973-8_14

14.1 Introduction

Dispersion in laminar/turbulent flow with boundary absorption as applied to environmental, industrial, chemical, navigation processes and prediction of water quality in rivers and channels has received considerable attention. Dispersion is the combined mechanism of advection and diffusion (laminar/turbulent). However, some other parameters also play a vital role for dispersion such as flow oscillation, chemical reactions, boundary irregularities, etc. Taylor [1, 2] initiated the pioneer study on dispersion of contaminants in laminar/turbulent flows and diffusivity $\frac{U^2a^2}{48D}$ is obtained by considering the transverse diffusion in the straight circular tube, where 'D' is the molecular diffusion coefficient and 'a' is the pipe radius. Subsequently, Aris [3] extended the work by considering the axial diffusion also using the method of moments and diffusivity $D + \frac{U^2a^2}{48D}$ is obtained as the total longitudinal diffusivity. Barton [4] modified the Aris [2] method of moments and the solution is obtained for all time. Smith [5] studied the boundary absorption on the longitudinal diffusion in shear flows. Barton [6] modified the Aris [2] method of moments to handle the case of reactive solutes and the solution was obtained for moderate and large time. Purnama [7] studied boundary retention effects upon contaminant dispersion in parallel flows. Various studies have been conducted on dispersion [8–11]. The effect of the irreversible boundary reaction on the dispersion of tracer in an annular region, in the presence of oscillatory flow, was studied by Sarkar and Jayaraman [12]. Kumar and Jayaraman [13] studied dispersion of a solute in a curved channel using method of moments for laminar dispersion in curved channels. Mondal and Mazumder [14] and Ng [15] studied the absorption/reaction/retention of the tracer at the boundary wall using the method of moments or some kind of averaging.

In this paper, the dispersion of a solute in turbulent flow in a curved channel with boundary absorption is studied. Dispersion process is described appropriately in terms of the rate of growth of variance with the apparent diffusion coefficient as a function of time. The mathematical formulation for flow in a curved channel viz., the coordinate system chosen and the velocity profile for a curved channel are given in Sect. 14.2. The methodology for dispersion in curved channel is discussed in Sect. 14.2. Section 14.3 discusses the numerical solution for the longitudinal dispersion for curved channel for the general moment equations for different curvature ratios. The results are discussed in Sect. 14.4 and salient conclusions are listed in Sect. 14.5.

14.2 Mathematical Formulation

Steady plane laminar/turbulent flow of a viscous incompressible fluid through a smooth curvilinear channel of constant width 2a, is shown in (Fig. 14.1). The complete three-dimensional hydrodynamic equations and equation of continuity for

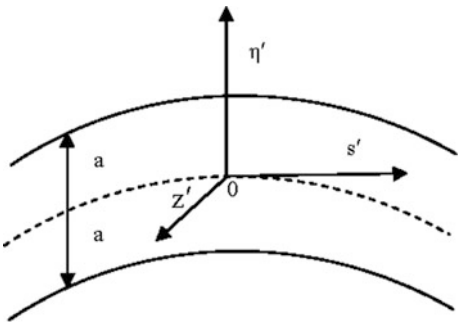

Fig. 14.1 Schematic diagram for coordinate system in curved channel

the velocity components (u', v', w) in (s', η', z') directions are given in Kumar and Jayaraman [16, 17]. The centre line is characterised by a constant curvature k'. If s', η' and z' are the directions along the flow, transverse to the flow and vertically upward at a centre line of a channel, respectively, and if $\vec{r}$ is the position vector of any point inside the channel, then

$$(\mathrm{d}\vec{r})^2 = (1 + k'\eta')^2(\mathrm{d}s')^2 + (\mathrm{d}\eta')^2 + (\mathrm{d}z')^2 \tag{14.1}$$

where $((1 + k'\eta'), 1, 1)$ are the scale factors.

14.2.1 Steady Laminar Flow in Curved Channel

The detailed analyses are given in Kumar [19],

$$u_0(\eta) = \frac{u(\eta)}{U_M} = 2(1-k)\frac{\left\{\left[\frac{(1+k)^2}{4k}\ln\left(\frac{1+k}{1-k}\right)\right]\left[\frac{1+k\eta}{1-k} - \frac{1-k}{1+k\eta}\right] - \left(\frac{1+k\eta}{1-k}\right)\ln\left(\frac{1+k\eta}{1-k}\right)\right\}}{\left\{1 - \frac{(1+k)^2(1-k)^2}{4k^2}\left[\ln\left(\frac{1+k}{1-k}\right)\right]^2\right\}} \tag{14.2}$$

This is the exact expression [16–18] for the steady velocity profile along the channel in terms of the non-dimensional variables.

U_M is the non-dimensional mean velocity in a curved channel. The same mean pressure gradient in the straight channel is assumed along the centre line of the curved channel as well. k is the curvature ratio. The centrifugal force is matched by the pressure gradient in the η direction and the other velocity component corresponding to this two-dimensional steady, laminar flow vanishes. Thus, with no lateral flow, the effect of curvature is purely geometrical.

14.2.2 Turbulent Flow in Shallow Curved Channel

A depth-averaged computation procedure is considered including the convective influence of the secondary flow to get more appropriate results in shallow curved channels. Apart from the other assumptions, longitudinal component of the velocity is considered more dominating than the other ones. Fully developed velocity profile is taken for computing the longitudinal diffusivity as it is independent of the axial coordinate to make applicable the method of moments. The detailed analyses are given in [19, 20].

14.2.3 Longitudinal Diffusivity Using Method of Moments

The pth integral moment of the concentration distribution is given as

$$C_p(t', \eta') = \int_{-\infty}^{\infty} s'^p \overline{C}(t', s', \eta') \mathrm{d}s' \tag{14.3}$$

The functions $C_p (p = 0, 1, 2 \ldots)$ describe the distribution of contaminant in a filament centred on $\eta' =$ constant. In particular, C_0 gives the total mass of the contaminant for $\eta' =$ constant, $\frac{C_1}{C_0}$ gives the position of the centre of gravity and it can be shown that C_2 is related to the variance σ^2 through $\sigma^2 = \frac{C_2}{C_0} - \frac{C_1^2}{C_0^2}$.

The moments, further averaged over the cross-section of the channel, are defined as

$$M_p(t') = \frac{1}{2a} \int_{-a}^{a} C_p(t', \eta') \mathrm{d}\eta' \tag{14.4}$$

which gives the corresponding information over the whole cloud.

Method of moments is applied for estimating the second order central moments for longitudinal diffusivity when fully developed flow is attained in both longitudinal and transverse velocity profiles. Moment equations for the concentration equation and the corresponding boundary conditions using Eqs. (14.3) and (14.4) are given as follows:

$$\begin{aligned} \frac{\partial h C_p}{\partial t'} - E_{\eta'} h \frac{\partial^2 C_p}{\partial \eta'^2} - \frac{E_{\eta'} h}{R(1 + \frac{\eta'}{R})} \frac{\partial C_p}{\partial \eta'} + \frac{\partial (\tilde{v}' C_p)}{\partial \eta'} \\ = \frac{p \tilde{u}'(\eta')}{(1 + \frac{\eta'}{R})} C_{p-1} + E_{s'} h \frac{p(p-1)}{(1 + \frac{\eta'}{R})^2} C_{p-2} \end{aligned} \tag{14.5}$$

$$C_p(0,\eta') = \int_{-\infty}^{\infty} s^p\ \delta(\eta')\mathrm{d}s' = \delta_p(\eta'), \tag{14.6}$$

$$\frac{\partial \overline{C}_p}{\partial \eta'} + \beta C_P = 0, \eta' = \pm a \tag{14.7}$$

$$\frac{h\mathrm{d}M_p}{\mathrm{d}t'} - \frac{1}{2a}\int_{-a}^{a} \frac{E_{\eta'}h}{R(1+\frac{\eta'}{R})}\frac{\partial C_p}{\partial \eta'}\mathrm{d}\eta' + \frac{1}{2a}\int_{-a}^{a}\frac{\partial(\bar{v}'C_p)}{\partial \eta'}\mathrm{d}\eta' = \frac{1}{2a}\int_{-a}^{a}\left[p\left(\frac{u'C_{p-1}}{(1+\frac{\eta'}{R})}\right) + p(p-1)\left(\frac{E_{s'}hC_{p-2}}{(1+\frac{\eta'}{R})^2}\right)\right]\mathrm{d}\eta' - \frac{1}{2}\beta\big(C_p(-1,t) + C_p(1,t)\big) \tag{14.8}$$

$$M_p(0) = \overline{\delta_p}(\eta') = \begin{matrix} 1, & p = 0, \\ 0, & p > 0. \end{matrix}$$

The moments about the mean are defined as

$$v_p = \frac{1}{2\,aM_0}\int_{-a}^{a}\int_{-\infty}^{\infty} \overline{C}\left(s' - \overline{s'_g}\right)^p \mathrm{d}s'\mathrm{d}\eta'$$

where the mean (first central moment) s_g is given as

$$s'_g = \frac{1}{2\,aM_0}\int_{-a}^{a}\int_{-\infty}^{\infty} s'\overline{C}\mathrm{d}s'\mathrm{d}\,\eta' = \frac{M_1}{M_0}$$

s'_g can be regarded as the centroid of the contaminant distribution, which measures the location of the centre of gravity of the cloud movement with the mean velocity of the fluid.

The variance related to the dispersion of the contaminants about the mean position is obtained as

$$v_2 = \frac{M_2}{M_0} - s'^2_g \tag{14.9}$$

Thus, the asymptotic longitudinal dispersion coefficient D_a is defined as

$$D_a(t) = \frac{1}{2}\frac{\mathrm{d}v_2}{\mathrm{d}t} \tag{14.10}$$

These equations along with their initial and boundary conditions can be solved numerically for different absorption parameters of (β). For the case of laminar flow, the detailed analysis of the governing equations are given in Kumar and Jayaraman [16, 17], Kumar [19].

14.3 Numerical Scheme

Longitudinal diffusivity was computed by solving Eqs. (14.5–14.8) numerically. The discretization scheme for Eq. (14.5) was based upon a finite difference Crank-Nicholson implicit scheme with truncation error of $O(\Delta t'^2) + O(\Delta \eta'^2)$. The derivatives and all other terms were written at the mesh point (i, j), where $i = 0$ corresponded to time $t = 0$ and $j = 0$, to the inner wall of the channel at $\eta' = -3$. The mesh point (i, j) indicates a point where $t'_i = i \times \Delta t'$ and $\eta'_j = -3 + j \times \Delta \eta'$. $\Delta t'$ and $\Delta \eta'$ are the increments of t' and η' respectively. The finite difference scheme leads to a system of linear algebraic equations with a tridiagonal coefficient matrix, which is solved by the method of Thomas algorithm using the initial and boundary conditions.

Initial and boundary conditions in finite difference form are given as

$$\overline{C}_p(0, j) = \left\{ \begin{array}{ll} 1, & \text{for } p = 0 \\ 0 & \text{for } p \geq 1. \end{array} \right\}$$

$\overline{C}_p(i, -3) = \overline{C}_p(i, 3)$, at the inner wall and $\overline{C}_p(i, M + 3) = \overline{C}_p(i, M - 3)$, at the outer wall of the channel for $p \geq 0$. M is the value of j at the outer wall. Moments M_p were calculated by solving Eq. (14.8) and with initial condition after applying Simpson's one-third rule for averaging $\overline{\tilde{u}\, C}_{p-1}$ and $\overline{C}_{p-2}$ with the known values of $\overline{C}_p$, $\overline{u'}(\eta')$ and $\overline{v'}(\eta')$ at the corresponding grid points. Longitudinal diffusivity, as defined by Eq. (14.10), requires only calculations of M_2 and M_0 and hence numerical calculations were carried out only for M_p, $p = 0, 1, 2$.

14.4 Results and Discussion

The objective of this paper is to study the combined effects of curvature and absorption parameter on the longitudinal diffusivity in turbulent flow. Numerical simulations were made for several cases of channel curvature parameter and boundary absorption parameter at various cross-sections of the channel. For the calculation of velocity distribution in both longitudinal and transverse, the detail analysis of the governing equations are given in Kumar [19] and Kalkwijk and De Vriend [20]'s hydraulic model. Radii of curvature ranged from 10 to 100 m, in

particular 10, 50 and 100 m. The longitudinal diffusion coefficients (Es' and $E\eta'$) were fixed to be $O(10^{-3})$ and friction coefficient of Chezy about 60 $m^{1/2}/s$.

Figure 14.2 shows the axial velocity profile in a curved channel at different cross-sections and variations of the longitudinal velocity are also plotted at different cross-sections. Main velocity is assumed to have the logarithmic distribution in the vertical direction. Longitudinal velocity in the channel decreases with increasing longitudinal distance of the channel. Velocity profiles shift towards the outer bend of the channel due to the effect of curvature. The effect of the secondary flow convection factor ($k_{s'\eta'}$) on the axial velocity at different locations of the channel is to increase the velocities in the outer bend and decrease in the inner bend and the longitudinal velocity distribution tends to be skewed outwards.

Figure 14.3 shows the variations of the transverse component of the velocity at different axial locations of the channel for radius of curvature $R = 50$. It is obvious that only the secondary flow due to the curvature of the bend is considered in the formulation of the transverse velocity profile. Magnitude of the transverse component of the velocity decreases with increasing the distance towards the downstream side. Transverse velocity contributes considerably for mildly curved channel flow also and the distribution is strongly skewed towards the outer bend of the channel.

14.4.1 Longitudinal Dispersion Coefficient D_a in Laminar Flow

The effective longitudinal dispersion coefficient D_{eff} as defined in Kumar and Jayaraman [18] depends on β, k and Pe. The variations of D_{eff} with time are

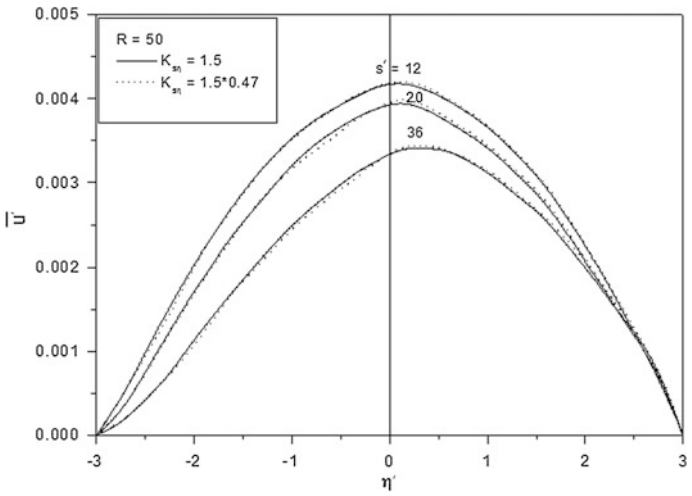

Fig. 14.2 The velocity profile u at different cross-section ratios through cross-section in curved channel

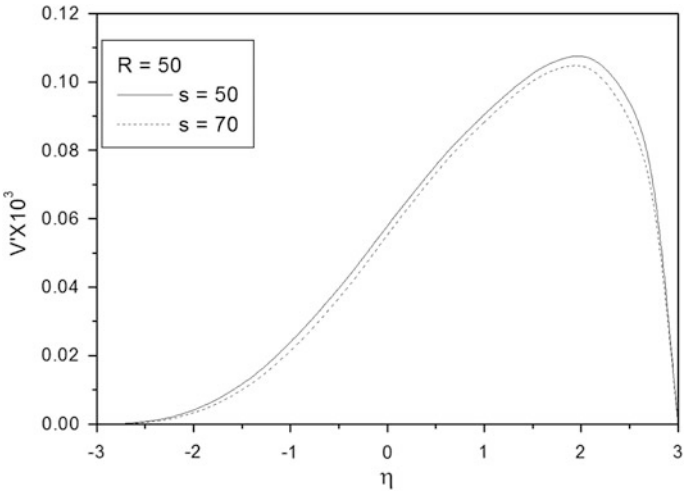

Fig. 14.3 Transverse component of velocity for different curvatures

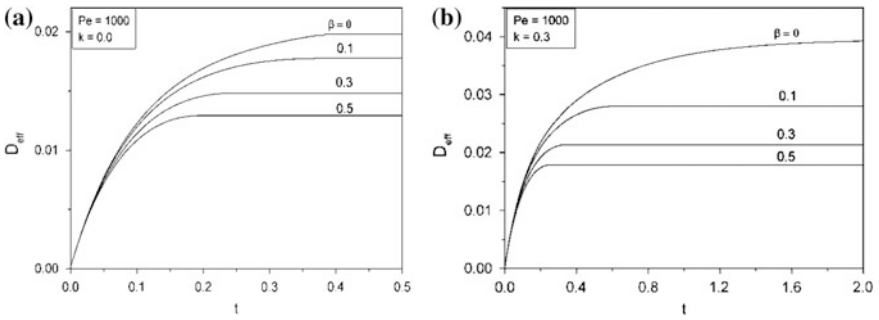

Fig. 14.4 Dispersion coefficient (D_{eff}) versus time (t) for different dimensionless absorption parameter (β) **a** $k = 0.0$, **b** $k = 0.3$

depicted for $k = 0.0, 0.3$ and for different values of β in Fig. 14.4a, b. It can be seen in Fig. 14.4a that for a straight channel ($k = 0$), D_{eff} initially increases and reaches asymptotically to a steady state value at large dimensionless dispersion time t for all β. For $\beta = 0$, it reaches the asymptotic value $D_a(t) = \frac{1}{\text{Pe}^2} + \frac{2}{105}$. The increase in absorption at the boundaries, i.e increase in β will change the amount of slug material across the channel and hence for a given time t, D_{eff} decreases with increase in absorption parameter β. D_{eff} is found to increase with increase in curvature ratios but it decreases with increase in boundary absorption parameter and finally reaches an asymptotic value in each case depending on the value of k and β.

14.4.2 Longitudinal Dispersion Coefficient D_a in Turbulent Flow

Longitudinal diffusivity is for fully developed velocity profiles using method of moments, however, lateral component occurring in the governing equation is considered significant for matter dispersion at the initial stage and stationary stage. The variations of D_a with time for turbulent flow are depicted for radius of curvature = 100, 50 and 10 with a constant absorption parameter 1.0 in Fig. 14.5a. D_a

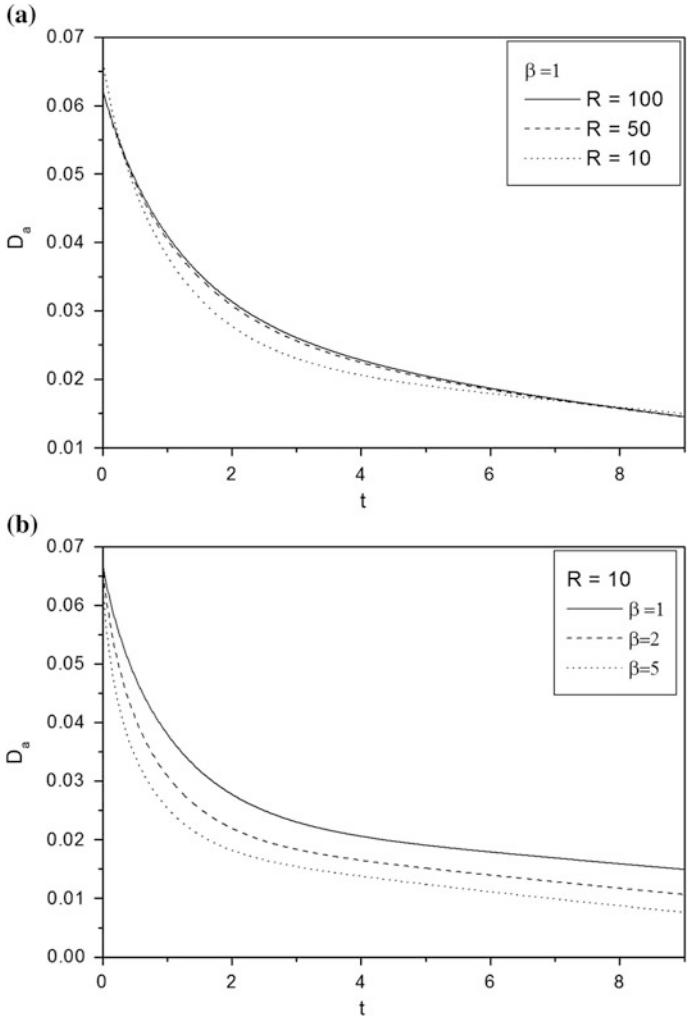

Fig. 14.5 **a** Dispersion coefficient (D_a) versus time (t) for turbulent flow for radius of curvature. **b** Dispersion coefficient (D_a) versus time (t) for turbulent flow for different absorption parameter

decreases with decreasing radius of curvature; it reaches a stationary state after a certain critical time, which depends on the radius of curvature and boundary absorption. It can be seen from Fig. 14.5b that as β increases, the diffusivity D_a decreases and the stationary state is delayed as curvature increases. But for a given radius of curvature (R), the stationary state is achieved earlier as absorption parameter β increases. It should be possible to link the transverse mixing rate either to the degree of curvature or to some measure of the strength of the secondary circulation, shear effects.

14.5 Conclusion

In this paper we have reported the combined approach of curvature and boundary absorption for solving the moment equations, in order to estimate the effective longitudinal dispersion coefficient in a curved channel in turbulent flow. It is seen that transverse velocity contributes remarkably in turbulent flow. The flow curvature was effective on change of lateral profile of flow and on generation of the lateral component of flow. The effect of secondary flow convection is basically through the transverse component of velocity. Dispersion of a solute with absorbing boundaries has been carried out for mildly curved channel to understand the role of absorbing boundaries in transporting the solute in turbulent flow. Dispersion coefficient is found to increase with decrease in radius of curvature, but decreases with increase in boundary absorption parameter. It is seen that the effective dispersion coefficient reaches asymptotically to a steady-state value at large dimensionless dispersion time for all β and R. The study can be extended to include sinusoidal curvature effects on the dispersion in turbulent flows.

References

1. Taylor GI (1953) Dispersion of soluble matter in solvent flowing slowly through a tube. Proc Roy Soc Lond A 219:86–203
2. Taylor GI (1954) The dispersion of matter in turbulent flow through pipe. Proc Roy Soc Lond A 223:446–468
3. Aris R (1956) On the dispersion of a solute in a fluid flowing through a tube. Proc Roy Soc Lond A 235:67–77
4. Barton NG (1983) On the method of moments for solute dispersion. J Fluid Mech 126:205–218
5. Smith R (1983) Effect of boundary absorption upon longitudinal dispersion in shear flows. J Fluid Mech 134:161–177
6. Barton NG (1984) An asymptotic theory for dispersion of reactive contaminants in parallel flow. Austral Math Soc Ser B 25:287–310
7. Purnama A (1988) Boundary retention effects upon contaminant dispersion in parallel flows. J Fluid Mech 195:393–412

8. Aris R (1960) On the dispersion of a solute in pulsating flow through a tube. Proc Roy Soc Lond A 259:370–376
9. Chatwin PC (1975) On the longitudinal dispersion of passive contaminant in oscillatory flows in tubes. J Fluid Mech 71:513–527
10. Smith R (1982) Contaminant dispersion in oscillatory flows. J Fluid Mech 114:379–398
11. Yasuda H (1982) Longitudinal dispersion due to the boundary layer in an oscillatory current: Theoretical analysis in the case of an instantaneous line source. J Ocean Soc of Japan 38:385–394
12. Sarkar A, Jayaraman G (2004) The effect of wall absorption on dispersion in oscillatory flow in an annulus-application to catheterised artery. Acta Mech 172:151–167
13. Kumar S, Jayaraman G (2005) Method of moments for laminar dispersion in curved channels. In: Proceedings of international conference on environmental. Fluid Mechanics (ICEFM'05) Allied Publishers Pvt. Ltd. IIT Guwahati, pp 291–297
14. Mondal KK, Mazumder BS (2005) On solute dispersion in pulsatile flow through a channel with absorbing walls. Int J Non-linear Mech 40:69–81
15. Ng CO (2006) Dispersion in steady and oscillatory flows through a tube with reversible and irreversible wall reactions. Proc Roy Soc A 462:481–515
16. Kumar S, Jayaraman G (2008) Method of moments for laminar dispersion in an oscillatory flow through curved channels with absorbing walls. Heat Mass Transfer 44:1323–1336
17. Kumar S, Jayaraman G (2012) Method of moments for estimating two dimensional laminar dispersion in curved channels. Indian J Ind Appl Math 3(1):116–133
18. Goldstein S (1965) Modern developments in fluid dynamics. 1 Dover Publications, New-York
19. Kumar S (2008) Method of moments for laminar/turbulent dispersion in curved channel flows, Ph.D thesis, CAS, IIT Delhi
20. Kalkwijk JPTh, de Vriend HJ (1980) Computation of flow in shallow river bends. J Hydraul Res 18(4):327–342

Chapter 15
Recent Advances in Compressive Sensing

Noore Zahra

Abstract Compressive sensing is an efficient way to represent signal with less number of samples. Shannon's theorem which states that the sampling rate must be at least twice the maximum frequency present in the signal (the so-called Nyquist rate) is a common practice and conventional approach to sampling signals or images. Compressive sensing reveals that signals can be sensed or recovered from lesser data than required by Shannon's theorem. This paper presents a brief historical background, mathematical foundation, and a theory behind compressive sensing and its emerging applications with a special emphasis on communication, network design, signal processing, and image processing.

Keywords Sampling theorem · Compressive sampling · Sparsity · Incoherence

15.1 Introduction

We can understand compressive as compressed and sensing as sampling. As the technology advances, the concept of "doing more with less" becomes essential. To achieve this objective, researchers have been involved for a decade to improve and reinvent new techniques for digital signal processing, image and video processing, speech processing, sensors, digital data acquisition system, digital communication system etc. International Data Corporation reveals that the amount of data generated worldwide is nearly 1.8 trillion gigabytes in 2011, 2.8 trillion in 2012, and by 2020 it will be approx 40 trillion GBs or 40 zettabytes. As we see data are growing very fast per year. In contrast, the growth rate of memory storage is slower. So there is big gap between data production and data storage. As a consequence, this gap is

N. Zahra (✉)
School of Engineering and Technology, Sharda University, Noida, India
e-mail: noor_zahra_india@yahoo.co.in

A.H. Siddiqi et al. (eds.), *Mathematical Models, Methods and Applications*,
Industrial and Applied Mathematics, DOI 10.1007/978-981-287-973-8_15

going to be exponentially widened for data production and computational power and at the same time it will effect communication rates as well. With the introduction of compressive sensing in 2004 by Emmanuel Candès, Terence Tao, and David Donoho an exponentially growth of new techniques the sensor data has taken place. Data deluge is a major problem today and compressive sensing gives a promising solution to it.

CS mainly relies on two principles **sparsity** and **incoherence**. Sparsity pertains to the signals of interest and **incoherence** to the sensing modality.

Sparsity enables the signals to store information in few samples. It is an inherent property of CS by which reconstruction can be done accurately and by the virtue of this property it can be applied in diverse field. When we apply compressive sensing for digital images, it will overcome the problems associated with the huge memory storage, processing time, and cost of computational process.

15.2 Historical Background

In 1795, Prony [1] developed a method known as Prony's analysis to extract useful information from a uniformly sampled signal which further builds a series of sinusoids or damped complex exponentials. It is used for the parameter estimation of the signal frequency components like estimation of frequency, amplitude, phase, and damping components of a signal. It represents the sampled data as a linear combination of exponentials and was initiated because it can estimate the frequency [1]. This method also works well with nonlinear equation that utilizes the linear equations. In order to solve for the different exponential components, the square error of approximation must be calculated and it must attain minimal error. CS was used in seismology in 1970 when Claerbout and Muir gave use of absolute value error criteria in place of least square data modeling. An example can be seen of this stability in averaging by median rather than arithmetic mean.

One of the famous theories of signal processing is the Nyquist/Shannon sampling theory [2]. This principle states that a signal/image can be represented and reconstructed if sampling frequency f_s is greater than or equal to twice of the highest frequency f_m in the signal; $f_s \geq 2f_m$

Candès and Donoho introduced an emerging theory which goes by the name of "compressive sampling" or "compressed sensing," which says that this conventional theory is inaccurate. They discovered important results on the minimum amount of data needed to reconstruct an image even though the amount of data would be deemed insufficient by the Nyquist–Shannon criterion [3, 4]. They explain that signal and images can be reconstructed from far fewer data than what we usually do as a conventional and common practice which follows Shannon–Nyquist density sampling theory. Compressive sensing builds upon the fundamental fact that we can represent many signals using only a few nonzero coefficients in a suitable basis or dictionary. Nonlinear optimization can then enable recovery of such signals from very few measurements. The theoretical foundation of this revolution is the

pioneering work of Kotelnikov, Nyquist, Shannon, and Whittaker on sampling continuous-time band-limited signals [2, 5–7]. Their results demonstrate that signals, images, videos, and other data can be exactly recovered from a set of uniformly spaced samples taken at the so-called Nyquist rate of twice the highest frequency present in the signal of interest. After this discovery, much of signal processing has moved from the analog to the digital domain. Digitization has enabled the creation of sensing and processing systems that are more robust, cheaper and, consequently, more widely used than their analog counterparts.

Candès and Donoho stated that it is possible to reconstruct images or signals accurately from a number of samples which is far smaller than the desired resolution of the image/signal, e.g., the number of pixels in the image. The field of CS grew out of the work of Candès, Romberg, Tao and of Donoho, who showed that a finite-dimensional signal having a sparse or compressible representation can be recovered from a smaller set of linear, non-adaptive measurements [3, 4, 8–10].

In 1990 George, Gorodnitsky, and Rao studied sparsity in biomagnetic imaging and other contexts [11–13]. Simultaneously, Bresler, Feng, and Venkataramani proposed a sampling scheme for acquiring certain classes of signals consisting of k components with nonzero bandwidth (as opposed to pure sinusoids) under restrictions on the possible spectral supports, although exact recovery was not guaranteed in general [14–16]. In the early 2000s Blu, Marziliano, and Vetterli developed sampling methods for certain classes of parametric signals that are governed by uniform sampling [10].

15.3 Methodology

With the help of few sensors, only super-resolution signals can be obtained in compressive sensing. So, a new data acquisition protocol can be possible which converts analog signal to digital signal with less number of sensors. This sampling theory gives the fundamental methods for sampling and compressing data concurrently.

Figure 15.1a shows bulky data acquisition and Fig. 15.1b shows that the dispensable data can be rejected. Compressing of the dispensable data can be rejected. Compressing the signal which has some structure can be accomplished without intuitive a loss. This work can be carried out in the following steps:

a. Obtaining the whole signal.
b. Computation of transform coefficients.
c. Encoding the highest coefficients and throwing away all other coefficients.

This process of massive data acquisition followed by compression is extremely uneconomical (e.g. JPEG 2000) [3, 4, 8]. Instead of acquiring the data followed by compression, one can acquire the data that is already compressed, so there is no

Fig. 15.1 **a** Raw: 15 MB **b** JPEG: 150 KB. *Source* First EU-US Frontiers of Engineering Symposium, Cambridge, September 2010

need to dispose anything [9]. Compressive sampling suggests ways to economically translate analog data into compressed digital form [17, 18].

Some of the key insights underlying this new theory are given below.

Sparsity From a general viewpoint, sparsity and compressibility have played and continue to play a fundamental role in many fields of science. Sparsity leads to **efficient estimations**; for example, the quality of estimation by thresholding or shrinkage algorithms depends on the sparsity of the signal we wish to estimate. Sparsity leads to **efficient compression**; for example, the precision of a transform coder depends on the sparsity of the signal we wish to encode [19].

Mathematically, we have a vector $f \in R^n$ (such as the n-pixel image in Fig. 15.2) which we expand in an orthonormal basis (such as a wavelet basis)

$\psi = [\psi_1, \psi_2, \ldots \psi_n]$ as follows:

$$f(t) = \sum_{i=1}^{n} x_f \psi_f(t) \tag{15.1}$$

where x_i is the coefficient sequence of f, $x_f = \langle f, \psi_f \rangle$. It will be convenient to express f as ψ_x (where ψ is the $n \times n$ matrix with $\psi_1, \ldots, \psi_n$ as columns). The implication of sparsity is now clear: when a signal has a sparse expansion, one can discard the small coefficients without much perceptual loss.

Signal can be expanded as a superposition of spikes, sinusoids, B-splines, wavelets, curvelets, shearlets, alpha-molecule, and so on.

In the above figure, original megapixel image with pixel values in the range [0, 255] is taken and its wavelet transform coefficients are arranged in random order for enhanced visibility. Relatively few wavelet coefficients capture most of the signal energy; many such images are highly compressible. (c) The reconstruction obtained by zeroing out all the coefficients in the wavelet expansion but the 25,000 largest (pixel values are threshold to the range [0, 255]). The difference with the original picture is hardly noticeable. As described in "Under sampling and Sparse

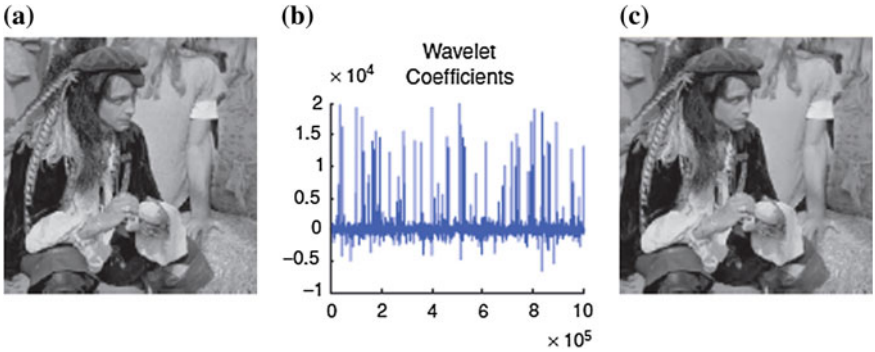

Fig. 15.2 **a** Original megapixel image with pixel values in the range [0, 255]. **b** Its wavelet transform coefficients. **c** The reconstruction obtained by zeroing out all the coefficients. *Source* Emmanuel et al., IEEE signal processing, page 23, March 2008

Signal Recovery," by Candès [4], this image can be perfectly recovered from just 96,000 incoherent measurements.

Incoherence extends the duality between time and frequency and expresses the idea that objects having a sparse representation in ψ must be spread out in the domain in which they are acquired, just as a Dirac or a spike in the time domain is spread out in the frequency domain. Put differently, incoherence says that unlike the signal of interest, the sampling/sensing waveforms have an extremely dense representation in ψ. The observation is that one can design efficient sensing or sampling protocols that capture the useful information content embedded in a sparse signal and condense it into a small amount of data. These protocols are nonadaptive and simply require correlating the signal with a small number of fixed waveforms that are incoherent with the sparse basis. The most remarkable thing about these sampling protocols is that they allow a sensor to very efficiently capture the information in a sparse signal without trying to comprehend that signal. Further, there is a way to use numerical optimization to reconstruct the full-length signal from the small amount of collected data. In other words, CS is a very simple and efficient signal acquisition protocol which samples—in a signal independent fashion—at a low rate and later uses computational power for reconstruction from what appears to be an incomplete set of measurements [4, 13].

To address the logistical and computational challenges involved in dealing with such high-dimensional data, we often depend on compression, which aims attending the most concise representation of a signal that is able to achieve a target level of acceptable distortion. One of the most popular techniques for signal compression is known as **transform coding**, and typically relies on finding a basis or frame that provides sparse or compressible representations for signals in a class of interest [16, 20–26]. Compressed sensing (CS) has emerged as a new framework for signal acquisition and sensor design. CS enables a potentially large reduction in

the sampling and computation costs for sensing signals that have a sparse or compressible representation. While the Nyquist–Shannon sampling theorem states that a certain minimum number of samples is required in order to perfectly capture an arbitrary band-limited signal, when the signal is sparse in a known basis, we can vastly reduce the number of measurements that need to be stored. Consequently, when sensing sparse signals we might be able to do better than suggested by classical results.

Restricted Isometric Property (RIP) is a famous tool for analyzing the performance of CS in acquisition and reconstruction. RIP in acquisition although similar to conventional method but its sensing paradigm makes the difference.

If $\not{X}$ is the signal to be sensed then

$$Y = \Phi \not{X} \tag{15.2}$$

where $\not{X} \in R^n$, Φ is m by n measurement matrix and $Y \epsilon R^m$ is measurement vector. In conventional sensing paradigm, m must be equal to n whereas in CS m can be far less than n.

Some advances using CS in communication, signal, and video processing are discussed below

1. **Block compressed sensing**

Gan [27] in block compressed sensing for natural images proposed image acquisition through block by block manner. Here, original image is divided into small blocks (basically $B \times B$) and each block is sampled independently which leads to faster speed. CS was recommended in order to overcome the conventional imaging system which samples the original images into digital format at a higher rate which was impossible. In case of devices with low power and image resolution, when considering $I_r \times I_c$ where N = IrIc with n measurements, the image in the block CS which is divided into small blocks is sampled the block CS shows that it can be implemented as a random 2D filter bank [27] (Fig. 15.3).

The sensors are pseudoinverse, i.e., they are sensitive to noise. In filter bank implementation of Block CS, the input image of each FIR filter goes through a rectangular decimation matrix M which gives the CS samples as output.

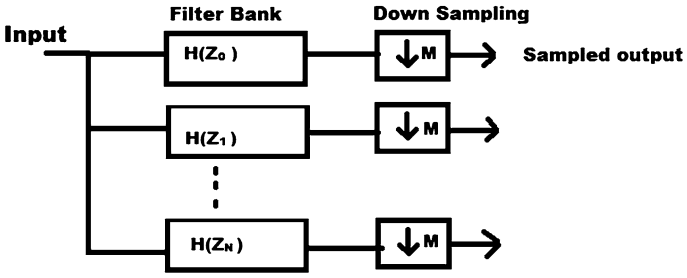

Fig. 15.3 Filter bank implementation of block CS

Nonlinear signal reconstruction method is used to improve the quality of reconstructed images, it uses a process of hard thresholding to remove the Gaussian noise which is based merely on three functions describing the whole procedure transmitting the input signal to yield a certain coefficient, that largest coefficient is kept and rest are set to zero, later the inverse transformed is applied to obtain the reconstructed signal. Weiner filter exploits the statistical properties of the image and can be used to restore images in the presence of blur as well as noise, the minimum squared error (M.S.E.) is minimized by using the orthogonality condition. In minimizing the M.S.E., the Weiner filter tends to smooth the image more than what the human eye would prefer. The reason being that M.S.E. weighs all the errors equally regardless of their location in the image and after that hard thresholding are applied which cuts down the Gaussian noise. Frame Expansions are used in order to recover natural images. For better reconstruction results, two-frame expansions are used. UWT which stands for undecimated wavelet transform which provides full description of the image, and OLT, i.e., oversampled lapped transform as the name suggests it represents the structures by overlapped block by block processing [27].

2. Acquisition and reconstruction using FIR (Finite Impulse Response) filter

Tropp et al. [28] proposed a technique for acquiring and reconstruction of signals using FIR filter. The main purpose of compressed sensing is to develop a linear measurement operator, $\mathbf{\Phi}$: $R^d \rightarrow R^n$, nonlinear reconstruction algorithm, A: $R^n \rightarrow R^d$ (to recover sparse signals) [27].

Each of the signals in CSA represented are in 2 m real numbers where $m < d$ where m = no. of sparse signals and n = no. of measurement of signals and d = length of the signals. The given equation shows that the reconstruction process is stable

$$||A(\Phi_S + v)||_2 \;=\; C||v||_2 \tag{15.3}$$

The compressive sampling acquisition (CSA) functions are designed only for finite length signals and the reconstruction process requires too much of space and time. Reduction of the size of audio signals is done by sampling at Nyquist rate (done to determine the stability of the feedback systems) before applying the lossy compressed equations.

Random filtering process was proposed to think beyond the Shannon–Nyquist sampling where the compressed version of digital signal is acquired which is applicable to analog signals and had a huge impact on analog/digital converters.

Random filters helps in acquiring compressed version of digital data which includes following processes like convolution of the signals and down sampling and helps in implementation in analog hardware in analog/digital converters. Random filtering contributes well in measurement process of analog signals and captures all sparse and compressible signals as compared to CSA where analog signals can perform limited functions which includes (1) filtering, (2) modulation, (3) sampling, and where the measurement process is not casual.

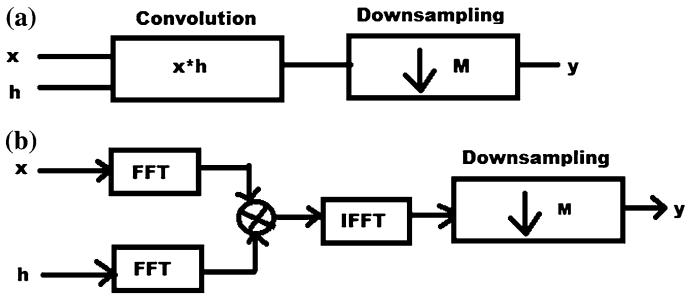

Fig. 15.4 **a** Signal acquisition using convolution. **b** Signal acquisition using FFT

In the random filtering process of compressed sensing, signal x is measured by:

(a) Convolution of the signal x with an FIR filter h has random taps which is then down sampled to obtain a compressed data y (Fig. 15.4).
(b) Using FFT/IFFT.

Random filters also accelerates the reconstruction algorithm and measurement algorithm, can trade longer filters for fewer measurements, easily implementable in software and hardware, and focuses on continuous-time signals [28].

3. Temporal Compressive Sensing

Yuan et al. [29] has introduced adaptive temporal compressive sensing for video. Here video compressive sensing has been developed to capture high-speed videos at low frame rate by means of temporal compression. The proposed method is to determine the temporal compression ratio N_F based on the motion of the scene which is to be sensed (Fig. 15.5).

They have estimated the motion of the objects within the scene by doing partitioning of frame A (e.g., previous frame) into $P \times P$ (pixels) blocks then took a predefined window size $M \times M$ (pixels) searching all the $P \times P$ blocks in the $M \times M$ windows in frame B (e.g., current frame) around the selected block in frame A and finally find the best-matching block in the window according to some metric (e.g., mean squared error), and use this to compute the block motion.

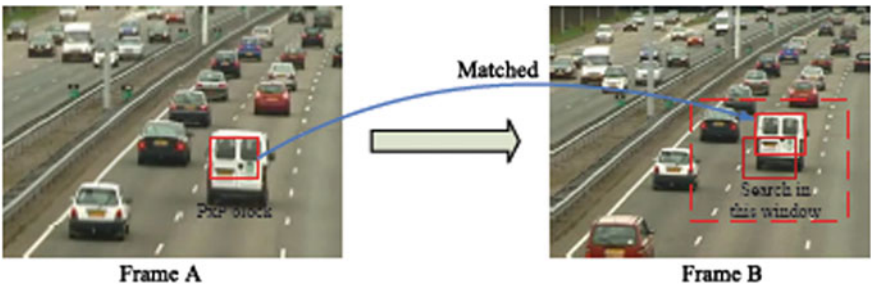

Fig. 15.5 Block matching of $P \times P$ block in frame B with best-matched block in frame A. *Source* Xin Yuang et al., ICIP 2013, IEEE [29]

15.4 Applications

1. Data acquisition: Compressive sensing reduces the number of measurements. Single pixel camera which is invented at Rice University makes a drastic change in imaging system.
2. Medical imaging: CS is widely used in medical imaging, particularly in magnetic resonance imaging. MR images have sparsity properties in Fourier, wavelet, curvelet, and shearlet domain. With the introduction of CS-based techniques, it is easy to take advantage of their implicit sparsity and reduction in the number of measurements without hampering the image quality.
3. Channel estimation: Compressed channel-based estimation uses nonlinear reconstruction algorithm and gives better result.
4. Wireless sensor network: Large number of sensors perform the task of data gathering for wireless sensor network.
5. Video scrambling: Block-based CS sampling is used on quantized coefficient. It provides security improvement and coding efficiency.

Other applications of compressed sensing include coding and information theory, machine learning, hyperspectral imaging, seismic imaging, cognitive radio networks, geophysical data analysis, computational biology, network traffic, remote sensing, radar analysis, robotics and control, A/D conversion, and many more.

Acknowledgments I take this opportunity to thank Prof. A.H. Siddiqi who introduced to me the exciting and the most useful theme of wavelet methods in signal and image processing and now the compressive sensing.

References

1. Grattan-Guinness I (1990) Convolution in French mathematics, 1800–1840, vol 1. Birkhauser, Boston, p 177
2. Shannon C (1949) Communication in the presence of noise. Proc Inst Radio Eng 37(1):10{21}
3. Candès EJ (2006) Compressive sampling. Candès EJ (ed) Proceedings of the international congress of mathematicians. European Mathematical Society, Madrid, Spain
4. Candès EJ, Wakin MB (2008) An introduction to compressive sampling. IEEE Signal Process Mag 25:21–30
5. Kotelnikov V (1933) On the carrying capacity of the ether and wire in telecommunications. In: Izd. Red. Upr. Svyazi RKKA, Moscow, Russia
6. Nyquist H (1928) Certain topics in telegraph transmission theory. Trans AIEE 47:617{644}
7. Whittaker E (1915) On the functions which are represented by the expansions of the interpolation theory. Proc. Royal Soc. Edinburgh Sec. A 35:181{194}
8. Baraniuk R (2007) Compressive sensing. IEEE Signal Process Mag 24(4):118{120, 124}
9. Candès E (2006) Compressive sampling. In: Proceedings of the international Congress of mathematicians, Madrid, Spain
10. Davenport MA, Duarte MF, Eldar YC, Kutyniok G (2010) Introduction to compressed sensing. Chapter 1, 2010

11. Gorodnitsky I, Rao B, George J (1992) Source localization in magnetoencephalagraphy using an iterative weighted minimum norm algorithm. In: Proceedings of the asilomar conference on signals, systems, and computers, Paci_c Grove, CA
12. Gorodnitsky I, George J, Rao B (1995) Neuromagnetic source imaging with FOCUSS: a recursive weighted minimum norm algorithm. Electroencephalogr Clin Neurophysiol 95(4):231{251}
13. Rao B (1998) Signal processing with the sparseness constraint. In: Proceedings of the IEEE international conference on acoustic, speech, and signal processing (ICASSP), Seattle, WA
14. Feng P, Bresler Y (1996) Spectrum-blind minimum-rate sampling and reconstruction of multiband signals. In: Proceedings of the IEEE international conference on acoustic, speech, and signal processing (ICASSP), Atlanta, GA
15. Feng P (1997) Universal spectrum blind minimum rate sampling and reconstruction of multiband signals. PhD thesis, University of Illinois at Urbana-Champaign
16. Vasanawala S, Alley M, Barth R, Hargreaves B, Pauly J, Lustig M (2009) Faster pediatric MRI via compressed sensing. In: Proceedings of annual meeting society pediatric radiology (SPR), Carlsbad, CA
17. Candès, EJ, Tao T (2004) Near-optimal signal recovery from random projections and universal encoding strategies. IEEE Trans. Inform. Theory 52:5406–5425
18. Donoho DL (2004) Compressed sensing. Technical report, Stanford University
19. Donoho DL, Vetterli M, DeVore RA, Daubechies I (1998) Data compression and harmonic analysis. IEEE Trans. Inform. Theory 44:2435–2476
20. Bruckstein AM, Donoho DL, Elad M (2009) From sparse solutions of systems of equations to sparse modeling of signals and images. SIAM Rev 51(1):34{81}
21. DeVore R (1998) Nonlinear approximation. Acta Numer 7:51{150}
22. Elad M (2010) Sparse and redundant representations: from theory to applications in signal and image processing. Springer, New York
23. Gedalyahu K, Eldar YC (2010) Time-delay estimation from low-rate samples: a union of subspaces approach. IEEE Trans Signal Process 58(6):3017{3031}
24. Gedalyahu KK, Tur R, Eldar YC (2011) Multichannel sampling of pulse streams at the rate of innovation. IEEE Trans Signal Process 50:153–156
25. Lustig M, Donoho D, Pauly J (2006) Rapid MR imaging with compressed sensing and randomly under-sampled 3DFT trajectories. Proceedings of the annual meeting of ISMRM, Seattle, WA
26. Mishali M, Eldar YC (2009) Blind multi-band signal reconstruction: compressed sensing for analog signals. IEEE Trans Signal Process 57(3):993{1009}
27. Gan L (2007) Block compressed sensing of natural images. In: Conference on digital signal processing (DSP), Cardiff, UK
28. Tropp JM, Duarte WM, Baron D, Baraniuk R (2006) Random filters for compressive sampling and reconstruction. In: Proceedings of the IEEE international conference on acoustics, speech, and signal processing (ICASSP), Toulouse, France
29. Yuan X, Yang J, Llull P, Liao X, Sapiro G, Brady DJ, Carin L (2013) Adaptive temporal compressive sensing for video. In: Proceeding of the IEEE international conference on image processing (ICIP), Melbourne, Australia

Chapter 16
Emergence of Shearlets and Its Applications

Ruchira Aneja

Abstract In recent years, serious effort has been made to design directional representation system for images such as curvelets, ridgelets and shearlets and corresponding transforms. Amongst these transforms the shearlet transforms seems quite interesting since it stems from a square integrable group representations and has the corresponding useful mathematical properties. As we know wavelets are associated with Besov spaces via atomic decompositions, shearlets correspond to certain function spaces known as shearlet co-orbit spaces. Shearlets provide an optimally sparse approximation in the class of piecewise smooth functions with C^2 singularity curves namely,

$$\|f - f_N\|_{L2}^2 \leq C_N^{-2} (\log N)^3 \text{ as } N \to \infty$$

where f_N is the non-linear shearlet approximation of a function The main objective of this review paper is to introduce basic elements of shearlet along with our own result regarding denoising of MRI images using shearlet.

Keywords Bandlimited shearlets · Compactly supported shearlets · Continuous shearlet transform · Discrete shearlet transform · Fast finite shearlet transform

16.1 Introduction

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 certain class if we have an understanding of its dominant features, which are associated with their geometric properties. For instance, edges in natural

R. Aneja (✉)
SET, Sharda University, Greater Noida, India
e-mail: aneja_ruchi@rediffmail.com

A.H. Siddiqi et al. (eds.), *Mathematical Models, Methods and Applications*,
Industrial and Applied Mathematics, DOI 10.1007/978-981-287-973-8_16

images. One major goal of applied harmonic analysis is constructing classes of analyzing elements which capture the most relevant information in a certain data class.

16.1.1 Wavelets and Beyond

Shearlets emerged as a part of an extensive research activity during the last 10 years, which allows encoding of several classes of multivariate data through its ability to represent anisotropic features such as singularities for example: edges in natural images. For higher dimensional data analysis, it is of fundamental importance to understand these geometric structures which go beyond the limitations of Fourier, wavelet and curvelet systems. Shearlets provide a unified treatment of continuum models as well as digital models; it allows a precise resolution of wave front sets, optimally sparse representations of cartoon like images and fast decomposition algorithm.

The emergence of wavelets [1] was a great success as it has the ability to provide optimally sparse approximations of a large class of frequently occurring signals, fast algorithmic implementations compared to traditional Fourier methods, rich mathematical structure which allows one to design families of wavelets. As a consequence of all these properties, wavelets have revolutionized image and signal processing area with wide range of applications ranging from denoising, enhancement, feature extraction, classification etc.

Despite their success, wavelets are not very effective when dealing with multivariate data as wavelet representation is not sparse, that is many wavelet coefficients are needed to accurately represent the edges. Wavelet representations are optimal for approximating data with pointwise singularities only but cannot handle singularities along curves. This limitation of wavelets prompted the mathematicians, engineers and scientists to introduce some form of directional sensitivity, and "directional" wavelets were introduced such as steerable pyramids by Simoncelli, directional filter banks by Bamberger and Smith and 2D directional wavelets by Antoine.

The breakthrough occurred with the introduction of curvelets by Candés and Donoho in 2004 with a pyramid of analyzing functions defined not only at various scales and locations as wavelets, but also at various orientations. Their supports are highly anisotropic and increasingly elongated at finer scales. Curvelets have excellent adaptive representation system to sparsely approximate image with edges. However curvelets need to be band limited and can only have limited spatial localization. Construction of curvelets involves rotations and these operators don't preserve the digital lattice.

In 2005, Do and Vitterli introduced a discrete filter bank version of curvelet framework-contourlet. Contourlet [2, 3] has a tree structured filter bank implementation similar to standard wavelet systems. It can achieve great efficiency in terms of redundancy and good spatial localization. But the limitation of contourlet

is that there is a limit in the number of directions used for shearing. This limitation was overcome by a new class of affine systems dealing efficiently with multivariate data viz shearlets.

Shearlets was introduced by Guo, Kutyniok, Labate, Lim and Weiss in 2005, derived from composite Wavelets. In contrast to rotation used by curvelets, shearlets [4] makes use of shearing to control directional selectivity.

The important features of shearlets include:

- Spatial localization
- High directional sensitivity
- Fast algorithmic implementations
- Optimally sparse approximations of anisotropic features in multivariate data
- Parabolic scaling
- Compactly supported analyzing elements.

16.2 Geometric Transformations

Geometric Transformations [5] are broadly classified into:

1. Euclidean Transformation
2. Affine Transformation

16.2.1 Euclidean Transformation

It is either a translation, rotation or reflection. Translation: Suppose a point (x, y) in the xy plane gets translated to a new point $x'y'$ where

$$x' = x + h \text{ and } y' = y + k$$

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & h \\ 0 & 1 & k \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{1}$$

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & -h \\ 0 & 1 & -k \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} \tag{2}$$

Rotation: If a point (x, y) is rotated by an angle θ about the origin to become a new point

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} \tag{3}$$

Reflection: To reflect a vector about a line that goes through the origin, let $\vec{l} = (l_x, l_y)$ be a vector in the direction of the line:

$$A = \frac{1}{\left\|\vec{l}\right\|^2} \begin{bmatrix} l_x^2 - l_y^2 & 2l_x l_y \\ 2l_x l_y & l_y^2 - l_x^2 \end{bmatrix} \tag{4}$$

16.2.2 Affine Transformation

Affine [6] is the combined effect of translation, rotation, scaling and shear.

$$T = \begin{bmatrix} t_{11} & t_{12} & t_{13} \\ t_{21} & t_{22} & t_{23} \\ 0 & 0 & 1 \end{bmatrix} \tag{5}$$

Scaling: Scaling transformation shrink/stretch an object which implies change in length and angle. Scaling simply means x co-ordinate is enlarged c_1 times and y co-ordinate is enlarged c_2 times.

$$x' = x\,c_1 \text{ and } y' = yc_2$$

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} 1/c_1 & 0 & 0 \\ 0 & 1/c_2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{6}$$

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} c_1 & 0 & 0 \\ 0 & c_2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} \tag{7}$$

Shear: Shear transformation has the effect of pulling/stretching an object in a direction parallel to the co-ordinate axis. Shear factor gives an indication of the amount of pulling (stretching). This factor can be positive or negative and can be applied to x-axis and y-axis independently.

Shear transformation in x-direction

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & -s_1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{8}$$

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & s_1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} \tag{9}$$

(Figs. 16.1 and 16.2).

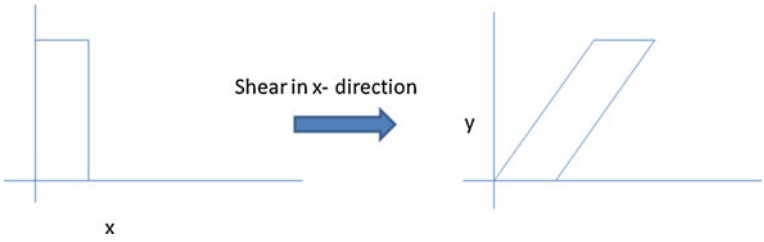

Fig. 16.1 Original axes and sheared axis with shearing in x-direction

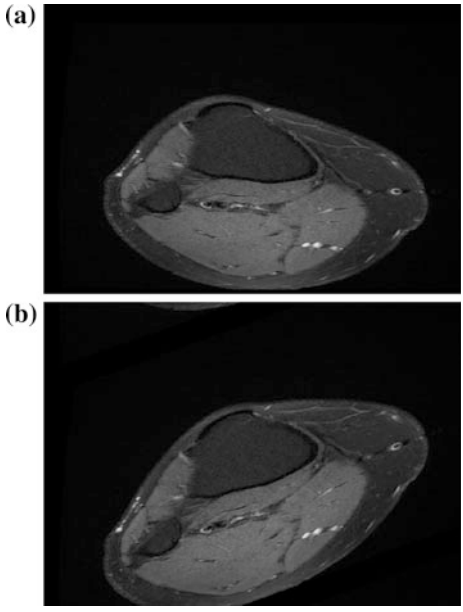

Fig. 16.2 **a** MRI brain image. **b** MRI brain sheared image at a slope = 0.5

16.3 Continuous Shearlet Transform

16.3.1 Introduction to Continuous Shearlet Transform

The traditional wavelet theory was a one-dimensional theory and can be extended for multivariate data but has some limitations. Due to its isotropic nature, Continuous wavelet transform is unable to provide information about the geometry of the set of singularities of a function. Though it has the advantage of simplicity, but it lacks directional sensitivity and the ability to detect the geometry of function.

Shearlet system is a special case of composite wavelet systems [7] which provide optimally sparse representation for a large class of bivariate functions and the continuous Shearlet transform is derived from it.

The continuous shearlets [8] depend on three parameters, the scaling parameter $a > 0$, the shear parameter $s \in R$ and the translation parameter $t \in R^2$, and they are defined by

$$\Psi_{a,s,t}(x) = a^{-3/4}\psi((D_{a,s}^{-1}(x-t))) \text{ where } D_{a,s} = [a, -a^{1/2}s; 0, a^{1/2}] \tag{10}$$

The mother shearlet function ψ is defined almost like a tensor product by

$$\Psi(\xi_1, \xi_2) = \Psi_1(\xi_1)\Psi_2(\xi_2/\xi_1) \tag{11}$$

where ψ_1 is a wavelet and ψ_2 is a bump function.

The associated continuous shearlet transform again depends on the scaling parameter a, the shear parameter s and the translation parameter t, and is defined by:

$$\mathrm{SH}_f(a, s, t) = <f, \Psi_{a,s,t}> \tag{12}$$

The scaling matrix A_a makes use of parabolic scaling, and is represented by $\mathrm{diag}(a, a^{\alpha})$ where the parameter α controls the degree of anisotropy. The shearing matrices S_s gives an idea of the orientations using the variable s associated with slopes rather than angles.

By sampling the continuous shearlet transform on an appropriate discrete set of the scaling, shear, and translation parameters, it is possible to obtain a frame or even a Parseval frame for $L^2(R)$. To obtain the discrete shearlets, we sample the three parameters as $a_j = 2^j(j\varepsilon Z), s_{j,k} = k\, a_j^{1/2} = k\, 2^{j/2}(k\varepsilon Z) t_{j,k,m} = D_{a_j,s_j,k}(m\varepsilon Z^2)$. We choose the mother shearlet function ψ in a similar fashion as in the continuous case, i.e., we now choose ψ_1 to be a discrete wavelet and ψ_2 to be bump function with certain weak additional properties. The tiling of the frequency plane is illustrated in Fig. 16.3a. This system forms a Parseval frame for $L^2(R)$, and they are optimally sparse. Furthermore, they are associated with a generalized MRA-structure, where the scaling space is not only translation invariant but also invariant under the shear operator.

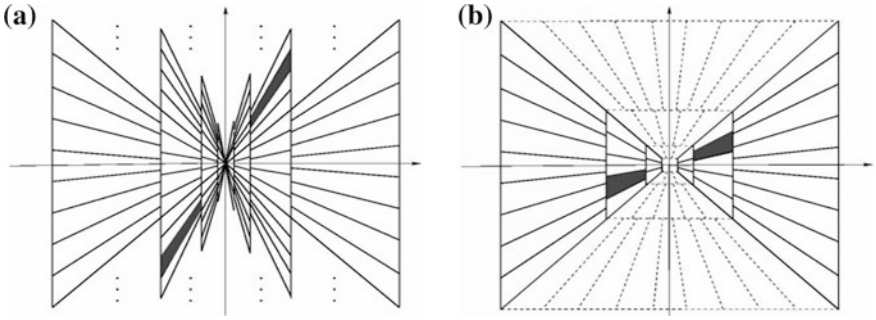

Fig. 16.3 **a** Frequency domain support of several elements of the shearlet system. **b** Tiling of frequency plane in cone adapted continuous shearlet system

The discrete shearlets on the cone, whose tiling of the frequency plane is shown in this Fig. 16.3b have the advantage that all directions are treated equally. Also each scale is associated with a finite number of shear parameters. This fact has certain advantages for numerical implementation.

16.3.2 Properties of Continuous Shearlet Transform

Continuous shearlet transform is able to identify not only the singular support of a distribution f but also the orientation of distributed singularities along curves. Decay properties of continuous shearlet transform as $a \to 0$ characterize the wavefront set of f with the translation parameter detecting the location and shear parameter detecting the orientation of a singularity. It is related to a compact group-shearlet group, which has a rich mathematical structure, uncertainty principle gives the accuracy of the transform and co-orbit theory is used to study smoothness spaces called the co-orbit spaces.

(1) Localization of shearlets: Analyzing elements of Continuous shearlets [9] decay rapidly as

$$|x| \to \infty \text{ that is } \Psi_{a,s,t}(x) = O(|x|^{-k}) \text{ as } |x| \to \infty \text{ for every } k \geq 0. \tag{13}$$

Rate of decay of Continuous shearlet transform exactly describes the location and orientation of singularities.

(2) Point singularities: If we substitute $t = 0$ in Eq. (10) we have

$$\mathrm{SH}_{\Psi}\delta(a, s, t) \approx \mathrm{a}^{-3/4} \tag{14}$$

In all other cases, $\mathrm{SH}_{\Psi}\delta(a, s, t)$ decays rapidly as $a \to 0$

(3) Linear singularities: We consider the linear delta distribution $v_p(x1, x2) = \delta(x1 + px2), p\,\varepsilon\, R$ defined by:

$$\langle v_p, f \rangle = \int_R f(-px2, x2)\mathrm{d}x_2 \tag{15}$$

Continuous Shearlet transform [10] determines both the position and orientation of linear singularity, in the sense that transform $\mathrm{SH}\,\psi\, v_p(a, s, t)$ always decays rapidly as $a \to 0$ except when t is on the singularity and $s = p$ i.e. direction perpendicular to the singularity or in other words, in which the singularity occurs.

If $t1 = -pt_2$ and $s = p$, $\mathrm{SH}_\Psi v_p(a, s, t) \approx \mathrm{a}^{-3/4}$as $a \to 0$. In all other cases, $\mathrm{SH}_\Psi v_p(a, s, t)$ decays rapidly as $a \to 0$ (Fig. 16.4).

For piecewise smooth boundary ∂S, let $B = \chi_S$, where $S \subset R^2$ and its boundary ∂S is a piecewise smooth curve.

1. If $t \notin \partial S$, then $\mathrm{SH}_\psi B(a, s, t)$ has rapid asymptotic decay as $a \to 0$, for each $s\,\varepsilon\, R$.
2. If $t\,\varepsilon\, \partial S$ and ∂S is smooth near t, then $\mathrm{SH}_\psi B(a, s, t)$ has rapid asymptotic decay as $a \to 0$, for each $s\,\varepsilon\, R$ unless $s = s_0$ is the normal orientation to ∂S at p. In this case $\mathrm{SH}_\psi B(a, s_0, t) \approx a^{3/4}$ as $a \to 0$.
3. If t is a corner point of ∂S and $s = s_0$, $s = s_1$ are the normal orientation to ∂S at t, then $\mathrm{SH}_\psi B(a, s_0, t)$, $\mathrm{SH}_\psi B(a, s_1, t) \approx a^{3/4}$ as $a \to 0$. For all other orientations the asymptotic decay of $\mathrm{SH}_\psi B(a, s, t)$ is faster (Fig. 16.5).

(4) Polygonal singularities: We consider the characteristic function χ_v of the cone $V = \{(x1, x2 : x1 \geq 0,\ qx1 \leq x2 \leq px1\},$ where $0 < q \leq p < \infty$

For $t = 0$, if $s = -\frac{1}{p}$ or $s = -\frac{1}{q}$, $\mathrm{SH}_\psi \chi_v(a, s, t) \approx a^{3/4}$ as $a \to 0$ and

$$\text{if } s \neq -\frac{1}{p}, s \neq -\frac{1}{q}, \mathrm{SH}_\psi \chi_v(a, s, t) \approx a^{5/4} \text{ as } a \to 0 \tag{16}$$

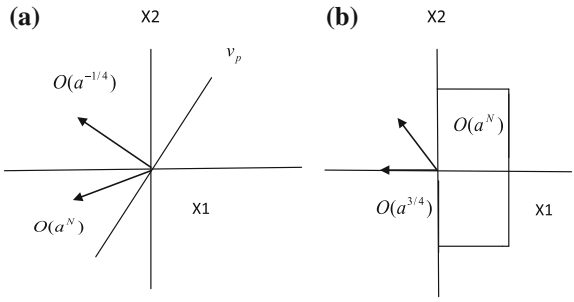

Fig. 16.4 **a** Continuous shearlet transform of linear delta distribution. **b** Continuous shearlet transform of heavyside function

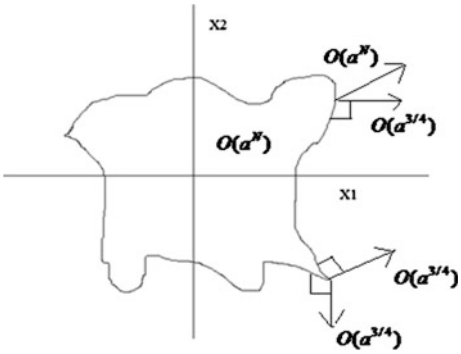

Fig. 16.5 Region S with piecewise smooth boundary ∂S

For $t \neq 0$, if $s = -\frac{1}{p}$ or $s = -\frac{1}{q}$ we have $SH_{\psi}\chi_{v}(a,s,t) \approx a^{3/4}$ as $a \to 0$. In all other cases, $\mathrm{SH}_{\psi}\,\chi_{\mathrm{v}}(a,s,t)$ decays rapidly as $a \to 0$.
The decay of Continuous shearlet transform $\mathrm{SH}_{\psi}\chi_{\mathrm{v}}(a,s,t)$ exactly identifies the location and orientation of the singularities. The orientation of linear singularities can be detected considering only the point singularity at the origin.

16.3.3 3D Continuous Shearlet Transform

We use separate Shearlet systems defined in different subregions of the frequency space. This leads to three pyramid based systems, associated with the pyramidal regions.

$$
\begin{aligned}
P_1 &= \left\{(\xi_1,\xi_2,\xi_3)\varepsilon R^3 : |\xi_1| \geq 2, \left|\frac{\xi_2}{\xi_1}\right| \leq 1 \text{ and } \left|\frac{\xi_3}{\xi_1}\right| \leq 1\right\},\\
P_2 &= \left\{(\xi_1,\xi_2,\xi_3)\varepsilon R^3 : |\xi_1| \geq 2, \left|\frac{\xi_2}{\xi_1}\right| > 1 \text{ and } \left|\frac{\xi_3}{\xi_1}\right| \leq 1\right\} \qquad (17)\\
P_2 &= \left\{(\xi_1,\xi_2,\xi_3)\varepsilon R^3 : |\xi_1| \geq 2, \left|\frac{\xi_2}{\xi_1}\right| \leq 1 \text{ and } \left|\frac{\xi_3}{\xi_1}\right| > 1\right\}
\end{aligned}
$$

For $\xi = (\xi_1, \xi_2, \xi_3)\varepsilon\, R^3, \xi_1 \neq 0,$ let $\psi^{(d)}, d = 1, 2, 3$ be defined

$$\begin{aligned}
\hat{\Psi}^{(1)}(\xi) &= \hat{\Psi}^{(1)}(\xi_1, \xi_2, \xi_3)\hat{\Psi}_1(\xi_1)\hat{\Psi}_2(\frac{\xi_2}{\xi_1})\hat{\Psi}_2(\frac{\xi_3}{\xi_1}), s \\
\hat{\Psi}^{(2)}(\xi) &= \hat{\Psi}^{(2)}(\xi_1, \xi_2, \xi_3)\hat{\Psi}_1(\xi_1)\hat{\Psi}_2(\frac{\xi_1}{\xi_2})\hat{\Psi}_2(\frac{\xi_3}{\xi_2}), \\
\hat{\Psi}^{(3)}(\xi) &= \hat{\Psi}^{(3)}(\xi_1, \xi_2, \xi_3)\hat{\Psi}_1(\xi_3)\hat{\Psi}_2(\frac{\xi_2}{\xi_3})\hat{\Psi}_2(\frac{\xi_1}{\xi_3}),
\end{aligned} \tag{18}$$

where ψ_1, ψ_2 satisfy the same assumptions as in 2D case. Hence for $d = 1, 2, 3$, the 3D pyramid based continuous shearlet systems for $L^2(P_d)^V$ are the systems

$$\left\{\psi^{(d)}_{a,s_{11},s_2,t} : 0 \leq \mathrm{a} \leq 1/4, -\frac{3}{2} \leq s_1 \leq \frac{3}{2}, -\frac{3}{2} \leq s_2 \leq \frac{3}{2}, t \in \mathrm{R}^3\right\}$$

where

$$\begin{aligned}
&\psi^{(d)}_{a,s_1,s_2,\mathrm{t}}(x) = \left|\det M^{(\mathrm{d})}_{as_1s_2}\right|^{\frac{1}{2}}\psi^{(d)}((M^{(d)}_{as_1,s_2})^{-1}(x-t)), \text{ and} \\
&M^{(1)}_{as_1s_2} = \begin{pmatrix} a & -a^{1/2}s_1 & -a^{1/2}s_2 \\ 0 & a^{1/2} & 0 \\ 0 & 0 & a^{1/2} \end{pmatrix}, \; M^{(2)}_{as_1s_2} = \begin{pmatrix} a^{1/2} & 0 & 0 \\ -a^{1/2} & a & -a^{1/2}s_2 \\ 0 & 0 & a^{1/2} \end{pmatrix}, \\
&M^{(3)}_{as_1s_2} = \begin{pmatrix} a^{1/2} & 0 & 0 \\ 0 & a^{1/2} & 0 \\ -a^{1/2}s_1 & -a^{1/2}s_2 & a \end{pmatrix}
\end{aligned} \tag{19}$$

The elements of shearlet systems $\psi^{(d)}_{a,s_1,s_2,t}$ are well localized waveforms associated with various scales, controlled by a, various orientations controlled by two shear variables s_1, s_2 and variables, controlled by t.

For $f\,\varepsilon L^2(R^3)$ we define the 3D pyramid based continuous shearlet transform [11] $f \to \mathrm{SH}\,\psi f(a, s_1, s_2, t),$ for $a > 0, \mathrm{s}_1, s_2\,\varepsilon R, t\,\varepsilon R^3$ by

$$\begin{aligned}
\mathrm{SH}\psi f(a, s_1, \mathrm{s}_2, t) = &<f, \psi^{(1)}_{\mathrm{a},\mathrm{s}_1,s_2,t}> \quad \text{if } |s_1|, |s_2| \leq 1, \\
&<f, \psi^{(2)}_{a,1/s_1,s_2/s_1,t}> \quad \text{if } |s_1| > 1,\ |s_2| \leq |s_1| \\
&<f, \psi^{(3)}_{a,\frac{s_1}{s_2},\frac{1}{s_2},t}> \quad \text{if } |s_2| > 1, |s_2| > |s_1|
\end{aligned} \tag{20}$$

Thus 3D continuous shearlet transform corresponds to a specific pyramid based shearlet system.

16.4 Discrete Shearlet Transform

Shearlets are broadly classified into Bandlimited and Compactly supported shearlets.

16.4.1 Bandlimited Shearlets (Cartesian Grid)

Bandlimited shearlets have the advantage that it allows a high localization in frequency domain which is important for handling seismic data. They admit a digitization of the continuum theory. But they have the drawback that they have a higher computational complexity due to the fact that the windowing takes place in the frequency domain.

Shearlet transform can be computed by using Cartesian grid. The steps in computing Shearlet transform are:

(1) Generate shearing filters for each scale j and shear parameter k. Windowing is done using Meyer based window function given by:

$$y = 35x^4 - 84x^5 + 70x^6 - 20x^7 \tag{21}$$

(2) Compute the norm of shearlets for each scale and direction with inputs as Laplacian pyramid filter, cell array of directional shearing filters and size of the input image.
(3) Compute translation invariant shearlet transform with inputs as input image, Filter for non-subsampled laplacian pyramid and cell array of directional shearing filters. A-trous decomposition decomposes the input image into sub-bands of scales $j = 1, 2, 3, \ldots$ level. Then apply directional shearing filters to decompose images for each scale j.

Shearlet transform can also be computed on pseudopolar grid. The steps in computing *fast shearlet transform* are:

(1) PPFT: Pseudopolar Fourier transform with oversampling factor in the radial direction.
(2) Weighting: Multiplication by density compensation style weights.
(3) Windowing: Decomposing the pseudo-polar grid into rectangular subband windows followed by 2D inverse FFT.

These have been implemented in Shearlab [9].

16.4.2 Bandlimited Shearlets (Pseudopolar Grid)

16.4.2.1 Pseudopolar Fourier Transform

Pseudopolar Fourier transform makes use of pseudopolar grid [12] in contrast to polar grid in case of Polar fourier transform. The polar grid points sit at the intersection between linearly growing concentric circles and angularly equispaced rays. The pseudopolar points sit at the intersection between linearly growing concentric squares and a specific choice of angularly non-equispaced rays. Pseudopolar grid gives a denser sampling near origin, enabling better interpolation performance (Fig. 16.6).

The problem with pseudopolar Fourier transform [13] is highly non-uniform arrangement of points on the pseudopolar grid. Therefore, it is required to down weight points in regions of very high density by using weights that corresponds to density compensation weights underlying the continuous change of variables. Oversampling of pseudopolar grid is done by introducing an oversampling rate R in the radial direction.

Fast modified PPFT is obtained by substituting $(w1, w2) = \left(\frac{-2n}{R} \cdot \frac{2l}{n}, \frac{2\mathrm{n}}{\mathrm{R}}\right)$ in

$$\hat{I}(w1,\, w2) = \sum_{u,v=-N/2}^{N/2-1} I(u,v)\, e^{\frac{2\pi i(uw1 + vw2)}{m0}}$$

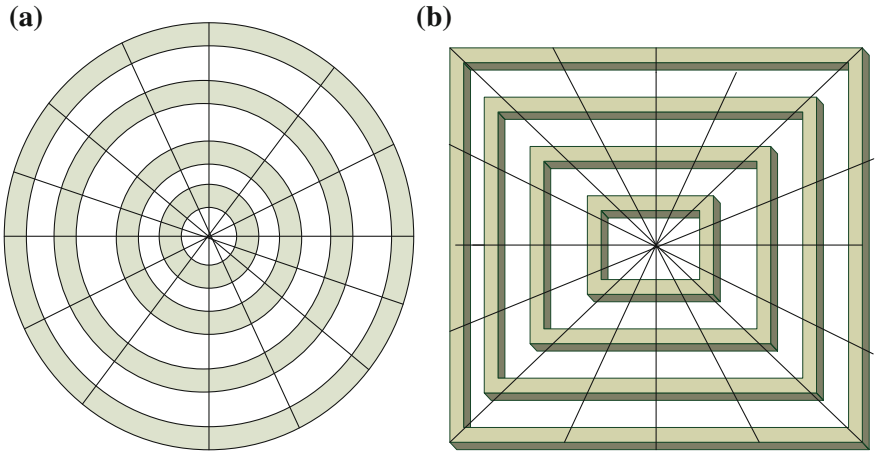

Fig. 16.6 **a** Polar grid with intersection of 8 concentric circles and 16 equispaced rays. **b** Pseudopolar grid with 8 concentric squares and equispaced rays in angle

where n = pseudo angle and l = pseudo radius

$$\begin{aligned}\hat{I}(w1,\, w2) &= \sum_{u=-N/2}^{N/2-1} \sum_{v=-N/2}^{N/2-1} I(u,v)\, e^{\frac{-2\pi iu\left(\frac{-4ul}{RN}\right)+v\left(\frac{2N}{R}\right)}{m0}} \\ &= \sum_{u=-N/2}^{N/2-1} \left(\sum_{v=-N/2}^{N/2-1} I(u,v)\, e^{\frac{-2\pi ivn}{Rn+1}} \right) e^{-2\pi iul\left(\frac{-2n}{(RN+1)N}\right)}\end{aligned} \tag{22}$$

The pseudopolar Fourier transform $\hat{I}$ on I on Ω_R^1 can be obtained by performing the 1D FFT on I along direction v and then applying a fractional Fourier transform along direction u.

16.4.2.2 Weighting

Weights are chosen such that the modified PPFT becomes an isometry i.e.

$$\sum_{u,v=-N/2}^{N/2-1} |I(u,v)|^2 = \sum\nolimits_{(w1,w2)\varepsilon\,\Omega_R} w(w1,w2) \cdot \left|\hat{I}(w1,w2)\right|^2 \tag{23}$$

We first choose a set of basis functions $w1$, …, $wn0$: $\Omega_R \to R^+$ such that

$$\sum_{j=1}^{n0} w_j(w1,w2) \neq 0 \,\text{for all}\, (w1,\, w2)\varepsilon\,\Omega_R \tag{24}$$

We then represent the weight functions w: $\Omega_R \to R^+$ by

$$w := \sum_{j=1}^{n0} c_j w_j \text{ with } c1, c2, \ldots \text{cn0 being non-negative constants} \tag{25}$$

Let $J := \hat{I} : \Omega_R \to C$ be the pseudopolar Fourier transform of $N \times N$ image I and w: $\Omega_R \to R^+$ be any suitable weight function on Ω_R the values

$$J_w(w1,w2) = J(w1,w2) \cdot \sqrt{w(w1,w2)} \text{ for all } (w1,w2)\ \varepsilon\ \Omega_\mathrm{R} \text{ is to be computed.} \tag{26}$$

Square root of weight is taken so that the image can be reconstructed from its weighted pseudopolar Fourier transform by applying the adjoint of the weighted pseudopolar Fourier transform.

16.4.2.3 Windowing

The final step of FDST is decomposing the data on the points of the pseudopolar grid into rectangular subband windows followed by 2D-ifft.

Given J_w, the set of digital shearlet coefficients

$$\begin{aligned} c_{n_0}^{l_0} &:= \left\langle J_W, \Phi_{n_0}^{l_0} \right\rangle \Omega_R \quad \text{for all } l_0, n_0 \text{ and} \\ c_{j,k,m}^{l} &:= \left\langle J_w, \sigma_{j,k,m}^{l} \right\rangle \quad \text{for all } j,k,m,l \end{aligned} \tag{27}$$

is computed followed by application of 2D-invFFT to each windowed image $J_w \Phi_0^{l0}$and $J_w \sigma_{j,k,m}^{l}$ respectively (Fig. 16.7).

16.4.3 Introduction to Compactly Supported Shearlets

Shearlets can be regarded as wavelets associated with an anisotropic scale matrix A_{2^j}, when the shear parameter k is fixed. This observation allows us to apply the wavelet transform to compute the shearlet coefficients, once the shear operation is computed for each shear parameter k. This approach is used in the digital formulation of compactly supported shearlet transform.

All constructions of shearlets are bandlimited functions which have unbounded support in space domain. In order to capture the local features of a given image efficiently, representation elements need to be compactly supported in the space domain.

But there are 2 problems associated with compactly supported shearlets.

1. Compactly supported shearlets don't form a tight frame which prevents utilization of adjoint as inverse transform.
2. There doesn't exist a natural hierarchical structure, mainly due to the application of a shear matrix.

A shearlet frame is defined as:

Let $s_j = \left[a^{j(q-1)/2}\right]$ and $c \in R^+$ be the sampling constant. For $\psi_0^i, \psi_1^1, \ldots, \psi_1^L \in L^2(R^2)$ and $\phi \in L^2(R^2)$ we define

$$\Psi^0 = \left\{ \psi_{jkm}^i : j,k \in Z, m \in Z^2, i = 1,2,\ldots L \right\}$$

and $\Psi = \left\{ T_{cm\phi} : m \in Z^2 \right\} \cup \left\{ \psi_{jkm}^i : j \geq 0,\ -s_j \leq k \leq s_j, m \in Z^2, i = 1,2,\ldots L \right\} \cup \left\{ \tilde{\psi}_{jkm}^i : j \geq 0, -s_j \leq k \leq s_j, m \in Z^2, i = 1,2,\ldots L \right\}$ where $\psi_{jkm}^i = D_{A_0^{-j} B_0^{-k}} T_{cm} \psi_0^i$ and

$$\tilde{\psi}_{jkm}^i = D_{A_1^{-j} B_1^{-k}} T_{cm} \psi_1^i \tag{28}$$

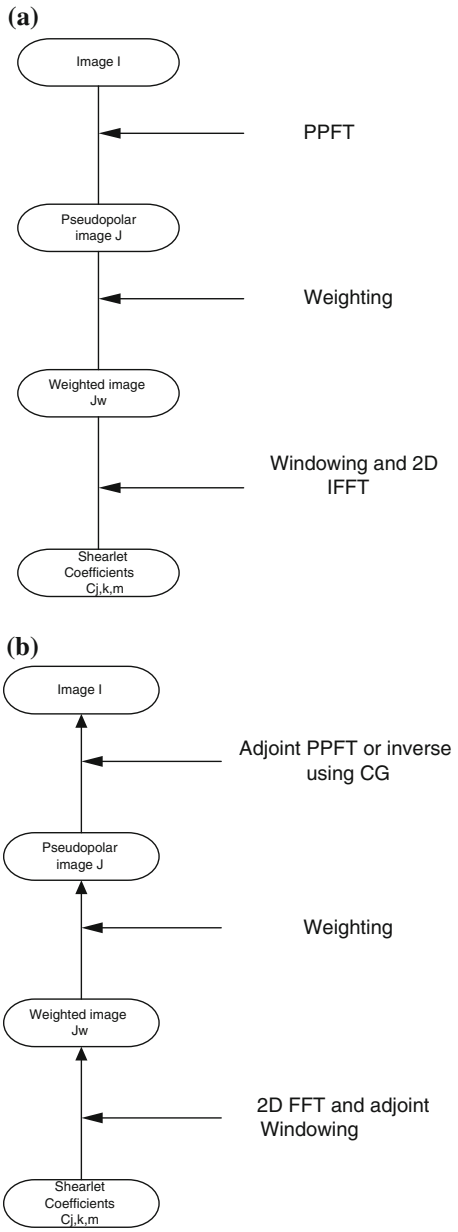

Fig. 16.7 **a** Fast discrete shearlet transform. **b** Inverse fast discrete shearlet transform

If Ψ or (Ψ^0) is a frame for $L^2(R^2)$, then we call the function ψ^i_{jkm} and $\tilde{\psi}^i_{jkm}$ in the system Ψ or (Ψ^0) shearlets.
Properties of Shearlets:

1. Frame property: It leads to a stable reconstruction of a given image [14].
2. Localization: Each of the shearlet frame element needs to be well localized in both space and frequency domain.
3. Efficient implementation: The discrete implementation needs to be derived from the construction of shearlets so that it inherits the nice properties from the corresponding shearlet systems.
4. Sparse approximation: It provides sparse approximation comparable with bandlimited shearlets.

16.4.3.1 Digital Separable Shearlet Transform (DSST)

Let shearlet coefficients $<f, \psi_{j,k,m}>$ be associated with shearlets $\psi_{j,k,m}$ belonging to $\Psi(\psi; c)$. Similarly shearlet coefficients associated with shearlets $\tilde{\psi}_{j,k,m} \in \tilde{\Psi}(\tilde{\psi}; c)$ are computed.

To construct a separable shearlet generator [15] $\psi \in L^2(R^2)$ and an associated scaling function $\phi \in L^2(R^2)$. Let $\phi \in L^2(R^2)$ be a compactly supported 1D scaling function satisfying

$$\phi_1(x_1) = \sum_{n_1 \in Z} h(n_1)\sqrt{2}\,\phi_1(2x_1 - n_1) \tag{29}$$

An associated compactly supported 1D wavelet $\psi_1 \in L^2(R^2)$ can be defined by:

$$\psi_1(x_1) = \sum_{n_1 \in Z} g(n_1)\sqrt{2}\phi_1(2x_1 - n_1) \tag{30}$$

where h and g are filter coefficients which are chosen so that ψ satisfies certain decay condition to guarantee a stable reconstruction from the shearlet coefficients. The selected shearlet generator is then defined to be

$$\psi(x_1, x_2) = \psi_1(x_1)\phi_1(x_2) \tag{31}$$

and scaling function by

$$\phi(x_1, x_2) = \phi_1(x_1)\phi_1(x_2) \tag{32}$$

For the signal $f \in L^2(R^2)$ to be analyzed we assume that for $J > 0$, f is of the form

$$f(x) = \sum_{n \in Z^2} f_J(n)\, 2^J \phi(2^J x_1 - n_1, 2^J x_2 - n_2) \tag{33}$$

An assumption for digital implementation is that the scaling coefficients can be viewed as sample values of f-in fact $f_J(n) = f(2^{-J}n)$ with appropriately chosen ϕ. Then for faithful digitization of shearlet coefficients $<f, \psi_{j,k,m}> = <f(S_{2^{-j/2}k}(.)), \psi_{j,0,m}(.)>$ and we assume that $j/2$ is an integer. We can digitize the shearlet coefficients $<f, \psi_{j,k,m}>$ by applying discrete separable wavelet transform associated with the anisotropic sampling matrix A_{2^j} to the sheared version of the data $f(S_{2^{-j/2}k}(.))$. Comparing we see that $f(S_{2^{-j/2}k}(.))$ is contained in the scaling space $V_j = \left\{2^J \phi\,(\,2^J. - n_1, 2^J. - n_2 \colon (n_1, n_2) \in Z^2\right\}$. If shear parameter $2^{-j/2}k$ is non-integer, then shear matrix $S_{2^{-j/2}k}$ doesn't preserve the regular grid $2^{-J}Z^2$ in V_J. So the new scaling space is obtained by refining the regular grid along x_1 axis by a factor of $2^{j/2}$.

The steps involved in DSST are:

1. For given input data f_J, apply the 1D up sampling operator by a factor of $2^{j/2}$ at the finest scale $j = J$.
2. Apply 1D convolution to the upsampled input data f_J with 1D low pass filter $h_{j/2}$ at the finest scale $j = J$. This gives $\tilde{f}_J$.
3. Resample $\tilde{f}_J$ to obtain $\tilde{f}_J(S_k(n)$ according to the shear sampling matrix S_k at the finest scale $j = J$.
4. Apply 1D convolution to $\tilde{f}_J(S_k(n)$ with $\tilde{h}_{j/2}$ followed by 1D down sampling by a factor of $2^{j/2}$ at the finest scale $j = J$.
5. Apply the separable wavelet transform $W_{J-1,\,J-j/2}$ across scales $j = 0, 1, \ldots, s-1$.

Features of DSST:

1. Digital Realization of Directionality: Rotation and shearing provides directionality. Rotation is a convenient tool to provide directionality that it preserves important geometric information such as lengths, angles and parallelism. But it doesn't preserve the integer lattice which causes severe problems of digitization. In contrast shear matrix S_k does not only provide directionality but also preserves the integer lattice when shear parameter k is an integer. Thus directionality can be discretized by using a shear matrix S_k.
2. Redundancy: To quantify redundancy of DST, we assume that the input data f is a linear combination of translates of a 2D scaling function ϕ at scale J as follows:

$$f(x) = \sum_{n_1=0}^{2^j-1} \sum_{n_2=0}^{2^J-1} d_n \phi(2^J x - n) \tag{34}$$

Redundancy is the number of shearlet elements necessary to represent f.
Redundancy of DSST is $(4/3)(1/c_1c_2)$

3. Computational Complexity: Computational Complexity of DSST is $O(2^{\log_2(1/2(L/2-1))}L \cdot N)$

16.4.3.2 Digital Non-separable Shearlet Transform (DNST)

The following drawbacks of DSST led to the introduction of Digital non-Separable shearlet transform:

(1) DSST is not time variant so another approach is needed to incorporate time variance.
(2) Since DSST is not based on tight frame, so it is difficult to approximate the inverse of shearlet transform.
(3) Computation of interpolated sampling values is also a problem.

We define shearlets generated by non-separable generator functions Ψ_j^{non} for each scale index $j \geq 0$ by setting:

$$\psi_{j,k,m}^{\text{non}}(x) = 2^{3/4j}\psi^{\text{non}}(S_c A_{2^j} x - M_{c_j} m) \text{ where } M_{c_j} \text{ is a sampling matrix} \qquad (35)$$

$M_{c_j} = \text{diag}(c_1^j, c_2^j)$ where c_1^j and c_2^j are sampling constants for translation.

The non-separable shearlet generator $\psi_{j,k,m}^{\text{non}}$ have high directional selectivity in frequency domain has compared to separable shearlet $\psi_{j,k,m}$. The details of DNST and its inverse can be found in [16].

16.5 Fast Finite Shearlet Transform

Discrete shearlet transform can be computed efficiently using Fast fourier transform (FFT) which gives rise to FFST. To compute FFST, we discretize the involved parameters a, s and t, but also consider only a finite number of discrete translations t.

16.5.1 Finite Discrete Shearlets

Let $j_0 = \lfloor 1/2 \log_2 N \rfloor$ be the number of considered scales. To obtain a discrete shearlet transform, we discretize the scaling, shear and translation parameters as

$$\begin{aligned} a_j &= 2^{-2j} = \frac{1}{4^j}, j = 0, \ldots, j_0 - 1, \\ s_{j,k} &= k2^{-j}, -2^j \le k \le 2^j, \\ t_m &= \left(\frac{m_1}{M}, \frac{m_2}{N}\right), m \in I \end{aligned} \tag{36}$$

With these parameters, shearlets can be written as:

$$\psi_{j,k,m}(x) = \psi_{a_j, s_{j,k}, t_m}(x) = \psi(A_{a_j}^{-1} S_{s_{j,k}}^{-1}(x - t_m)) \tag{37}$$

To obtain complete shearlets at the seam lines, we combine the three parts together, thus we define for $|k| = 2^j$, a sum of shearlets

$$\hat{\psi}_{j,k,m}^{h\times \mathrm{v}} := \hat{\psi}_{j,k,m}^{h} + \hat{\psi}_{j,k,m}^{v} + \hat{\psi}_{j,k,m}^{X} \tag{38}$$

The discrete shearlet transform can be defined as:

$$\mathrm{SH}(f)(\kappa, j, k, m) := \begin{cases} \langle f, \phi_{\mathrm{m}} \rangle & \text{for } \kappa = 0, \\ \left\langle f, \hat{\psi}_{j,k,m}^{\kappa} \right\rangle & \text{for } \kappa = \{h, v\}, \\ \left\langle f, \hat{\psi}_{j,k,m}^{h\times v} \right\rangle & \text{for } \kappa = \times, |k| = 2^j \end{cases}$$

where

$$j = 0, \ldots, j_0 - 1,\ -2^j + 1 \le k \le 2^j - 1 \text{ and } m \in \mathrm{I} \tag{39}$$

The shearlet transform can be efficiently realized by applying fft2 and inverse fft2 which compute the discrete Fourier transforms.

The complete shearlet transform is derived in [17] and is given by:

$$\mathrm{SH}(f)(\kappa, j, k, m) = \begin{cases} \text{ifft } 2(\hat{\varphi}(\omega 1, \omega 2)\hat{f}(\omega 1, \omega 2)) & \text{for } \kappa = 0 \\ \text{ifft } 2(\hat{\psi}(4^{-j}\omega 1,\ 4^{-j}k\omega 1 + 2^{-j}\omega 2)\hat{f}(\omega 1, \omega 2)) & \text{for } \kappa = h,\ |k| \le 2^j - 1 \\ \text{ifft } 2(\hat{\psi}(4^{-j}\omega 2,\ 4^{-j}k\omega 2 + 2^{-j}\omega 1)\hat{f}(\omega 1, \omega 2)) & \text{for } \kappa = v,\ |k| \le 2^j - 1 \\ \text{ifft } 2(\hat{\psi}^{h\times v}(4^{-j}\omega 1,\ 4^{-j}k\omega 1 + 2^{-j}\omega 2)\hat{f}(\omega 1, \omega 2)) & \text{for } \kappa \ne 0, |k| \le 2^j \end{cases} \tag{40}$$

The software implementation of FFST is available at:
http://www.mathematik.uni-kl.de/∼haeuser/FFST/

The MRI image of lungs is taken and the obtained shearlet coefficients along with shearlet representation and reconstructed image using FFST software in MATLAB is shown below (Fig. 16.8).

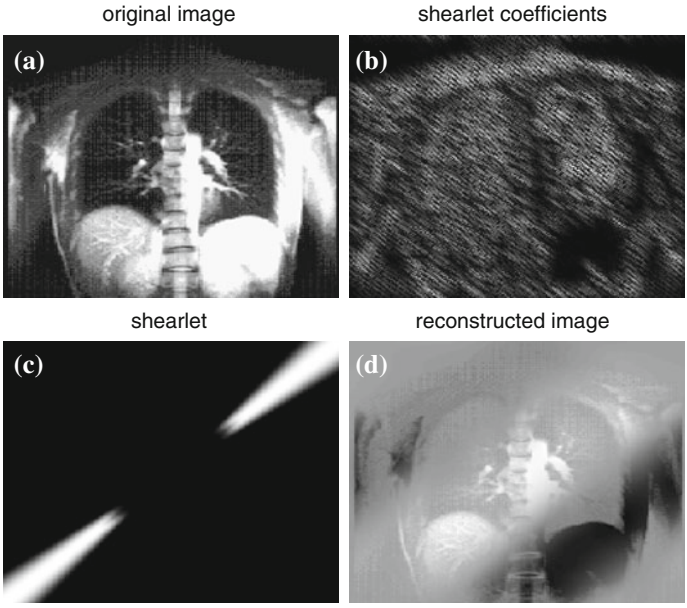

Fig. 16.8 **a** MRI lungs original image. **b** Shearlet coefficients of the image. **c** Shearlet representation of the image. **d** Reconstructed image after applying FFST

16.6 Applications of Shearlets

16.6.1 Shearlets for Biomedical Images

Shearlets are widely used in biomedical applications because of its geometric properties. The Shearlet transform of biomedical images of MRI (Brain) and X-ray (Breast) obtained in MATLAB are shown (Figs. 16.9 and 16.10).

16.6.2 Shearlets for Image Processing

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:

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.

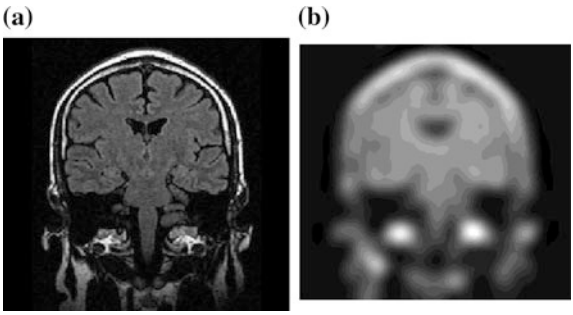

Fig. 16.9 **a** MRI brain image. **b** Shearlet transform of the image

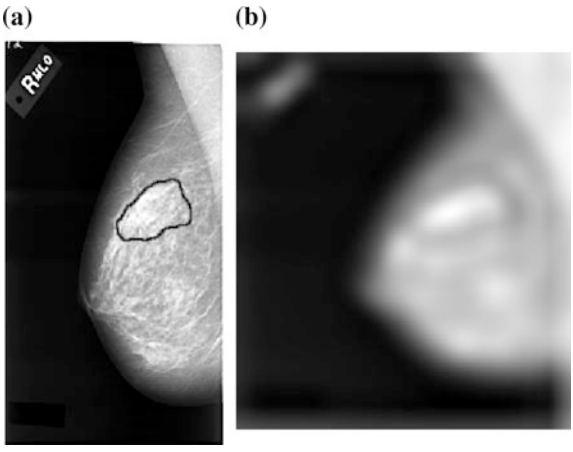

Fig. 16.10 **a** X-ray breast image. **b** Shearlet transform of the image

Various quantitative measures such as PSNR, MSE, SSIM index, ISNR, BSNR are used to determine the quality of the image.

Image Enhancement It is particularly used in medical imaging, where it is of prime interest to explore certain regions which include the essential features for medical diagnostic. Example enhancement of mammography images to improve the visibility of tumours for early detection. Shearlet transform is effectively used for enhancement because of its geometric features, the magnitude of the obtained shearlet coefficients enhance the necessary features for medical diagnosis. This is done by amplifying the weak edges while keeping the strong edges intact.

Image Separation is the process of decomposing a signal into its morphologically different components. Morphological Component Analysis (MCA) [18] has been recently proposed and the basic idea of this approach is to choose two frames $\Phi 1$ and $\Phi 2$ adapted to the two components to be separated in such a way that the two frames provide a sparse representation for each of the components. Searching for the sparsest representation of the signal in the combined dictionary $[\Phi 1|\Phi 2]$ would provide desired separation. For example: in neurobiological imaging, it would be desirable to separate 'spines' (point like objects) from 'dendrites' (curve like objects) in order to analyze them independently aiming to detect characteristics of

Alzheimer disease. Also, in astronomical imaging, astronomers would often like to separate stars from filaments for further analysis, hence again separating point-from curve like structures.
Wavelets along with curvelets separate point and curve like objects from the single image. But some part of the curve is missed and appears in the pointlike part. In contrast to this, compactly supported shearlets offer much better spatial localization than band-limited curvelets, which positively affects the capturing of localized features of the curve so Wavelets along with shearlets can be used efficiently for image separation as explained in [19].

Image Inpainting refers to the filling-in of missing data in digital images based on the information available in the observed region. Mathematically speaking, inpainting is essentially an interpolation problem, and thus directly overlaps with many other important tasks in computer vision and image processing, including image replacement, disocclusion, zooming etc. Inpainting in wavelet domain or using a sparse representation is a completely different problem since there are no well defined inpainting regions in the pixel domain.

Edge Detection shearlets play an important role in finding edges. The difficulty of *Edge detection* is particularly prominent in the presence of noise and when several edges are close together or cross each other. Wavelet suffers from the drawback of the inability to distinguish close edges and has poor angular accuracy. Wavelets have a limited capability in dealing with directional information. In order to overcome these difficulties, one has to account for the anisotropic nature of edge lines and curves. Shearlets 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 of edges. As a result, the shearlets framework provides highly competitive algorithms, for detecting both the location and orientation of edges [20] and for extracting and classifying basic edge features such as corners and junctions. Shearlet transform provides improved accuracy in the detection of edge orientation by using anisotropic dilations and multiple orientations. Various edge detection methods like Prewitt's, Canny's, Sobel are used, amongst all these Shearlet give an accurate estimate of the edges. Various quantitative measures are Prewitt's figure of merit etc. can be used to judge how effectively an edge is detected. Higher the value of Figure of merit, greater is the probability of detection of edge.

16.6.3 Shearlets for Biometric Applications

Face recognition [21] is popular nowadays because of its applications in areas like biometric security, image search engines. Each face image is described by a subset of band filtered images containing shearlet coefficients and form compact and

meaningful feature vectors using statistical measures. A face recognition system consists of two main stages-training stage and Classification stage. In training stage each facial image is decomposed along horizontal and vertical directions to extract facial features by using shearlet transform. Since process is repeated for all images in the training database and a feature matrix is constructed from these shearlet coefficients. In testing phase some steps are repeated to obtain shearlet coefficients. This feature matrix is then used in classification stage to classify the unknown test face image. A no. of feature evaluation measures can be used such as probabilistic distance measures-Euclidean distance, Bhattacharya distance etc.

Various facial databases such as Yale, ORL, Pie can be used for training. The directional information of the Shearlets makes it useful for design of face recognition systems. The features of shearlets make it successful for face recognition.

16.6.4 Shearlets for Inverse Problems

Inverse problem is one in which given the effect, one wants to recover the cause. Let X and Y be spaces having app structure, say a Banach space or a Hilbert space.

Direct problem: Given $x \in X$ and $A : X \rightarrow Y$. Find AX such that $Y = AX$

Inverse problem: Given an observed output Y, find an input X that produces it. $x \in X \Rightarrow Y = AX \in Y$.

For example: In medical X-ray tomography, direct problem would be to find out what kind of X-ray projection images would we get from a patient whose internal organs we know precisely. The corresponding inverse problem will be to reconstruct the 3D structure of the patient's insides given a collection of X-ray images taken from different directions. Here the patient is the cause and the collection of X-ray images is the effect.

Shearlet transform is used for Radon transform inversion as described in [22]. The Radon transform is the mathematical framework for Computerized tomography used in medical diagnosis.

Another example of inverse problem is image deblurring in which direct problem is finding out how a given sharp photograph would look if blurring is introduced by camera motion or by noise introduced by the electronics of the system. The inverse problem will be deblurring i.e. finding the sharp photograph which is done by deconvolution and is known to be an ill posed inverse problem. To regularize the ill posed problem, the sparse representation properties of Shearlets can be utilized. ISNR and BSNR are used to measure the effectiveness of the deblurred image using Shearlets.

16.7 Proposed Work

1. Acquire medical images of MRI scan and CT scan in DICOM (Digital Imaging in Communication and Medicine) format i.e. .dcm files in MATLAB.

2. Add Gaussian noise to the acquired images with different values of standard deviation with an interval of 5 viz $\sigma = 5, 10, 15, 20, 25$ and 30
3. Compute the value of shear parameter using shearing filters with Meyer based window function

$$y = 35x^4 - 84x^5 + 70x^6 - 20x^7$$

4. Compute the norm of Shearlets for each scale and direction with inputs as laplacian pyramid filter [23], cell array of directional shearing filters and size of the input image.
5. Compute translation invariant shearlet transform [24] with inputs as input image, Filter for non-subsampled LP and cell array of directional shearing filters. A-trous decomposition decomposes the input image into sub-bands of scales $j = 1, 2, 3, \ldots$ level. Then apply directional Shearing filters to decompose images $y\{j\}$ for each $y\{j\}$ scale j.
6. The obtained shearlet coefficients are thresholded using hard thresholding rule. For each shearlet coefficient $d\{j\}(n1, n2, k)$,

   (1) set $d\{j\}(n1, n2, k) = 0$ if $|d\{j\}(n1, n2, k)|/E(j, k) <$ sc(j)*lambda
   (2) keep d{j}(n1,n2,k) if $|d\{j\}(n1, n2, k)|/E(j, k)>$ = sc(j)*lambda

   Here, $E(j, k)$ is l^2 norm of shearlet for each scale j and shear parameter k. Then inverse shearlet transform is applied on the thresholded shearlet coefficients with shear parameter and Laplacian pyramid filter for non-subsampled LP. Apply directional Shearlet filters to decomposed images for each scale j followed by atrous recomposition using the same filter used for decomposition to reconstruct the denoised image.
7. The denoised/reconstructed image is compared with the original image and various quantitative measures such as PSNR, MSE are used to find the effectiveness of the denoised image.

16.8 Implementation Steps

1. Acquire MRI (Brain) image and add Gaussian noise with $\sigma = 10$ (Fig. 16.11).
2. Compute shearing filter using Meyer based window function and compute the norm of Shearlets for each scale j and shear parameter k.
3. Compute translation invariant shearlet transform of the dicom image where d is the cell array of the shearlet coefficients.
   With N by N input image and shearing filters 'shear' obtained as above, we have
   $d\{0\}$: low frequency part N by N
   $d\{1\}$: 2^(n(1)) + 2 arrays (N by N) of the shearlet coefficients for shear parameters
   $k = -2$^(n(1) − 1)…2^(n(1) − 1) and scale $j = 1$ (coarse scale).

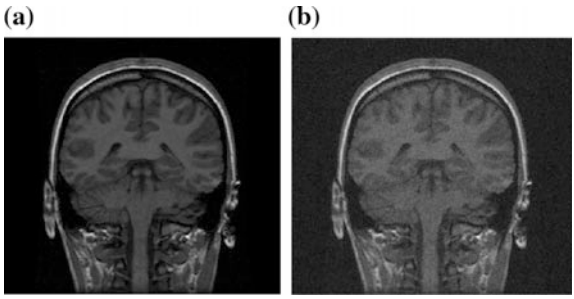

Fig. 16.11 **a** MRI (brain) image. **b** MRI (brain) noisy image

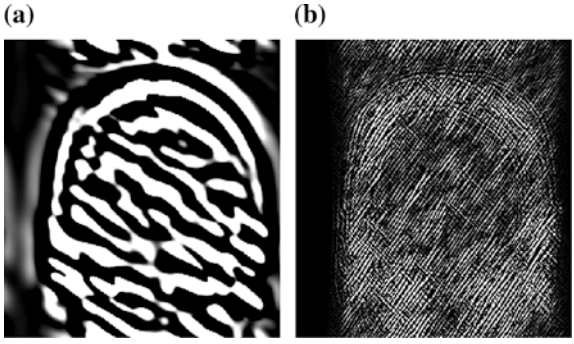

Fig. 16.12 **a** Approximation image. **b** One of the detailed images

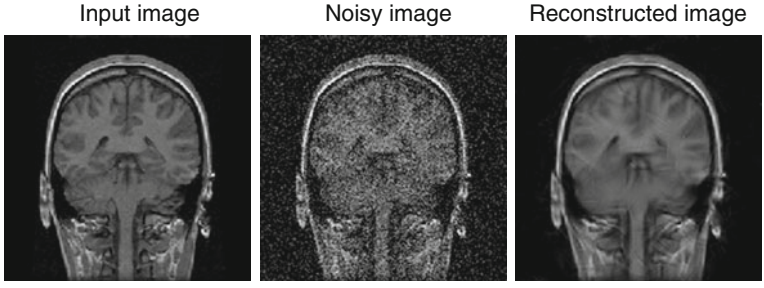

Fig. 16.13 Original MRI brain image followed by noisy image and reconstructed image using shearlets obtained in MATLAB

$d\{L\}$: $2\text{^}(n(L))$ + 2 arrays (N by N) of the shearlet coefficients for shear parameters
$k = -2\text{^}(n(L) - 1)\ldots 2\text{^}(n(L) - 1)$ and scale $j = L$ (fine scale).
Here, each entry of d is given by $d\{j\}(n1, n2, k) \rightarrow j$: scale, k: shear parameter, and $n1\&n2$: translation (Fig. 16.12).

4. Hard thresholding is done on the obtained shearlet coefficients. Most of the shearlet coefficients are set to zero.
5. Then inverse shearlet transform is applied on the thresholded shearlet coefficients with shear parameter and laplacian pyramid filter for non-subsampled LP (Fig. 16.13).

16.9 Results

Different images of MRI Brain and CT brain are taken with added Gaussian noise for different values of standard deviation. Peak Signal to noise ratio and Mean Square error are used to measure the effectiveness of the denoised image (Table 16.1, Fig. 16.14).

Table 16.1 Noisy PSNR, PSNR and MSE for MRI brain and CT brain images for different values of standard deviation

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
30	18.61	29.82	91.11
CT brain image	Gaussian noise		
SIGMA	NOISY PSNR	PSNR	MSE
10	28.15	34.56	22.77
15	24.64	32.23	38.93
20	22.13	30.53	57.50
25	20.16	29.37	75.24
30	18.55	28.35	95.15

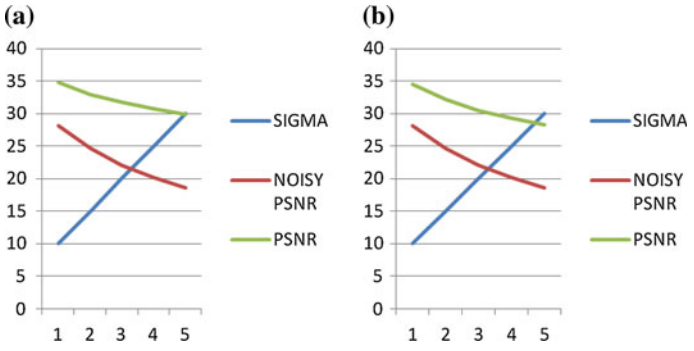

Fig. 16.14 **a** Comparative graph for MRI brain image for different values of standard deviation. **b** Comparative graph for CT brain image for different values of standard deviation

16.10 Conclusion

The aim of this review paper is to introduce shearlets, which go beyond the limitations of wavelets, curvelets etc. The applications of shearlets in image processing, biometric applications and inverse problem are introduced. Image denoising is done using discrete shearlet transform as explained in the proposed algorithm. From the obtained results it is found that with high value of sigma, improvement in PSNR from the noisy PSNR is considerably large. For sigma = 30, improvement in PSNR of original and reconstructed image with the PSNR of original and noisy image is of the order of 10 db. As the value of sigma decreases, improvement in PSNR is of the order of 5–8 db. MRI brain image and CT brain images are tested for denoising with added gaussian noise. Both images are effectively denoised, but MRI image performs slightly superior to the CT scan image. Wavelets denoising results in improvement of near about 2 db, Curvelets and contourlets give an improvement better than wavelets, but less than shearlets. Shearlets outperforms all the existing techniques in literature.

Acknowledgments I would like to thank the authors of ShearLab, and Wavelet Toolbox for making their codes available.

References

1. Soman KP, RamaChandran KI, Resmi NG (2010) Insight into wavelets from theory to practice. PHI Learning Pvt ltd., New Delhi
2. Arthur CL, Jianping Z, Do MN (2006) Nonsubsampled contourlet transform: theory, design, and applications. IEEE Trans Image Process 15(10):3089–3101
3. Arthur CL, Jianping Z, Do MN (2005) Nonsubsampled contourlet transform: filter design and application in image denoising. In: IEEE international conference on image processing, Genoa, Italy
4. Siddiqi AH (2012) Emerging applications of wavelet methods, American Institute of Physics. In: AIP conference proceedings, Meluille, New York
5. Sanjay S (2011) Digital image processing. S.K. Kataria and Sons, New Delhi
6. Guo K, Lim W, Labate D, Weiss G, Wilson E (2006) Wavelets with composite dilations and their MRA properties. Appl Comput Harmon Anal 20:231–249
7. Labate D, Weiss G (2008) Continuous and discrete reproducing systems that arise from translations, theory and applications of composite wavelets. Birkhauser, Four Short Courses on Harmonic Analysis, pp 1–48
8. http://www.shearlet.org/theory.html
9. Donoho DL, Kutyniok G, Shahram M, Zhuang X, A rationally designed digital shearlet transform-Shearlab. www.ShearLab.org
10. Dahlke S, Hauser S, Kutyniok G, Teschke G (2012) Continuous shearlet transform and shearlet Co-orbit spaces. Mathematics and Image Analysis, Paris
11. Kutyniok G, Labate D, Shearlets (2012) Multiscale analysis for multivariate data. In: Applied and numerical harmonic analysis, Springer, New York
12. Averbuch A, Coifman RR, Donoho DL, Israeli M, Shkolnisky Y (2008) A framework for discrete integral transformations I—the pseudopolar Fourier transform. SIAM J Sci Comput 30:764–784

13. Direct Exact Inverse Pseudopolar FFT and Radon transform using Orthogonalizing weights, Summer School at Inzell www.fim.uni-passau.de/…/Inzell2012/Inzell2012_Ofer_Levi.ppt
14. Easley G, Labate D, Lim W (2006) Optimally sparse image representations using shearlets. In: Proceedings of the 40th asilomar conference on signals, systems and computers, Monterey
15. Lim WQ (2010) The discrete shearlet transform: a new directional transform and compactly supported shearlet frames. IEEE Trans Image Process 19:1166–1180
16. Lim WQ (2013) Nonseparable shearlet transforms. IEEE Trans Image Process 22(5): 2056–2065
17. Hauser S (2012) Fast finite shearlet transform: a tutorial
18. Bobin J, Starck JL, Fadili J, Moudden Y, Donoho DL (2007) Morphological component analysis: an adaptive thresholding strategy. IEEE Trans Image Process 16(11):2675–2681
19. Donoho DL, Kutyniok G (2009) Geometric Separation using a wavelet-shearlet dictionary, SampTA-09. Marseille, France
20. Yi S, Labate D, Easley GR, Krim H (2008) Edge detection and processing using shearlets. In: Proceedings of the IEEE international conference on image processing, San Diego
21. Danti A, Poornima KM (2012) Face recognition using shearlets. In: IEEE international conference on industrial and information systems, Chennai
22. Flavia C, Glenn E, Kanghui G, Demetrio Labate (2009) Radon transform inversion using the shearlet representation. Appl Comput Harmonic Anal 29(2):232–250
23. Peter BJ, Edward Adelson H (1983) The Laplacian pyramid as a compact image code. IEEE Trans Comm 31(4):532–540
24. Easley G, Labate D, Lim W (2008) Sparse directional image representations using the discrete shearlet transform. Appl Comput Harmon Anal 25:25–46
25. Kutyniok G, Shahram M, Zhuang X (2011) Shearlab: a rational design of a digital parabolic scaling algorithm. SIAM J Multiscale Model, Simul

Chapter 17
Application of Wavelets in Numerical Evaluation of Hankel Transform Arising in Seismology

Nagma Irfan and A.H. Siddiqi

Abstract The computation of electromagnetic (EM) fields for 1-D layered earth model requires evaluation of Hankel transform. In this paper we propose a stable algorithm for the first time that is quite accurate and fast for numerical evaluation of the Hankel transform using wavelets arising in seismology. We have projected an approach depending on separating the integrand $tf(t)J_v(pt)$ into two components; the slowly varying components $tf(t)$ and the rapidly oscillating component $J_v(pt)$. Then either $tf(t)$ is expanded into wavelet series using wavelets orthonormal basis and truncating the series at an optimal level or approximating $tf(t)$ by a quadratic over the subinterval using the Filon quadrature philosophy. The solutions obtained by proposed wavelet method applied on three test functions indicate that the approach is easy to implement and computationally very attractive. We have supported a new efficient and stable technique based on compactly supported orthonormal wavelet bases.

Keywords Hankel transform wavelets · Bessel functions · Fourier Bessel series · Seismology

Mathematics Subject Classifications 44A15 65R10 65T60

17.1 Introduction

Electromagnetic (EM) depth sounding is, under favorable conditions, extremely useful in petroleum exploration, groundwater exploration, permafrost thickness determination exploration of geothermal resources, and foundation engineering

N. Irfan (✉) · A.H. Siddiqi
School of Basic Sciences and Research, Sharda University,
Knowledge Park III, Greater Noida, Delhi (NCR)-201306, India
e-mail: nagmairfanmath@gmail.com

A.H. Siddiqi et al. (eds.), *Mathematical Models, Methods and Applications*,
Industrial and Applied Mathematics, DOI 10.1007/978-981-287-973-8_17

problems. However, for data interpretation one needs fast and efficient computations of geoelectromagnetic anomaly equations. These equations appear as Hankel Transform (HT) (also known as Bessel Transform).

17.1.1 Hankel Transform

The efficient and accurate evaluation of the Hankel transform is required in a number of applications. This paper reviews a number of algorithms that have only recently been exposed in the literature. It is found that the performance of all algorithms depends on the type of function to be transformed. The wavelet based methods provide acceptable accuracy with better efficiency than numerical quadrature.

17.1.2 Mathematical Background

The general Hankel transform pair with the kernel being J_v is defined as [1]

$$F_v(p) = \int_0^\infty tf(t)J_v(pt)\mathrm{d}t, \tag{17.1}$$

and Hankel transform being self reciprocal, its inverse is given by

$$f(t) = \int_0^\infty pF_v(p)J_v(pt)\mathrm{d}p, \tag{17.2}$$

where J_v is the vth-order Bessel function of first kind. Due to oscillatory behaviour of $J_v(pt)$, standard quadrature methods applied to these integrals can be slow to convergence or may fail if the integral is divergent. It is only recently, mainly in the last 20 years, that attention has been turned to discovering algorithms useful for numerical evaluation of the Hankel transform. In this time a variety of algorithms of various strengths, weakness, and applicability's have been reported. As sometimes happens, the relevant literature is distributed through a number of journals, some of it in apparent ignorance of other research.

We believe it is now timely to bring this literature together, giving a review of the main methods available and providing pointers to some of the less efficient, but nevertheless elegant, methods.

17.1.3 Historical Background of Numerical Transforms Techniques

The literature concerning numerical Hankel transform techniques is very sparse from Longman until the late seventies when a flurry of papers were published on the topic. The various algorithms that have been published during and since the seventies can be filled into a few general categories. They are:

(1) Numerical quadrature
(2) Logarithmic change of variables
(3) Asymptotic expansion of the Bessel function
(4) Projection-slice/back projection method

Numerical evaluation of Hankel transforms is ubiquitous in the mathematical treatment of physical problems involving cylindrical symmetry, optics, electromagnetism and seismology. Many different types of algorithms and software have been developed to evaluate numerically hankel transform integrals in Geophysics [2, 3]. The ubiquity of these integrals in EM geophysics motivates the need for accurate and efficient numerical integral techniques.

17.1.4 Motivation of Present Work

(A) The Hankel transform arises naturally in the discussion of problems posed in cylindrical coordinates (with axial symmetry) and hence, as a result of separation of variables involving Bessel functions.
(B) Analytical evaluations are rare and hence numerical methods become important. The usual classical methods like Trapezoidal rule, cotes rule etc. connected with replacing the integrand by sequence of polynomials have high accuracy if integrand is smooth. But $tf(t)\ J_v(pt)$ and $pF_v(p)\ J_v(pt)$ are rapidly oscillating functions for large t and p, respectively.

To overcome these difficulties, various different techniques are available in the literature.

(1) *Fast Hankel Transform* Here, by substitution and scaling, the problem is transformed in the space of the logarithmic co-ordinates and the fast Fourier transform in that space.
(2) *Filon quadrature philosophy* In Filon quadrature philosophy, the integrand is separated into the product of an (assumed) slowly varying component and a rapidly oscillating component. In the context of the Hankel transform, the former is $tf(t)$ and the latter is $J_v(pt)$. This method works quite well for computing $F_0(p)$, for $p \geq 1$, but the calculation of inverse Hankel transform is more difficult, as $F_0(p)$ is no longer a smooth function but a rapidly oscillating one. Moreover the error is appreciable between $0 < p < 1$.

Several papers have been written to the numerical evaluation of the HT in general and the zeroth-order in particular [4–12]. There are two general methods of the effective calculation in this area. The first is the fast Hankel transform [13, 14]. The specification of that method is transforming the function to the logarithmical space and fast Fourier transform in that space. This method needs a smoothing of the function in log space. The second method is based on the separation of the integrand into product of slowly varying component and a rapidly oscillating Bessel function [15]. But it needs the smoothness of the slow component for its approximation by lower-order polynomials.

17.2 Preliminaries

17.2.1 Wavelets

Wavelets are a class of function constructed from dilation and translation of a single function called the mother wavelet. When the dilation and translation parameters a and b vary continuously, the following family of continuous wavelets are obtained

$$\psi_{a,b}(t) = |a|^{-\frac{1}{2}}\left(\frac{t-b}{a}\right), \quad a, b \in R, a \neq 0.$$

When the parameters a and b are restricted to discreet values as $a = 2^{-k}$, $b = n2^{-k}$,

Then, we have the following family of discrete wavelets

$$\psi_{kn}(t) = 2^{\frac{k}{2}}\psi\left(2^k t - n\right), \quad k, n \in Z,$$

where the function ψ, the mother wavelet, satisfies $\int_R \psi(t)\mathrm{d}t = 0$.

We are interested in the case where ψ_{kn} constitutes and orthonormal basis of $L^2(R)$. A systematic way to do this is by means of multiresolution analysis (MRA).

In 1910, Haar [16] constructed the first orthonormal basis of compactly supported wavelets for $L^2(R)$. It has the form $\{2^{\frac{j}{2}}\psi(2^j t - k) : j, k \in Z\}$ where the fundamental wavelet ψ is constructed as follows:

Construct a compactly supported scaling function ϕ by the two-scale scaling relation $\phi(t) = \phi(2t) + \phi(2t-1)$ together with the normalization constraint $\int \varphi(t)\mathrm{d}t = 1$. A solution of this recursion that represents φ in $L^2(\mathbf{R})$ is $\chi_{[0,1)}$.

Then $\psi(t) = \varphi(2t) - \varphi(2t-1)$. The Haar wavelets are piecewise continuous and have discontinuities at certain dyadic rational numbers.

In seminal papers; Daubechies [17, 18], constructed the first orthonormal basis of continuous compactly supported wavelets for $L^2(\mathbf{R})$. They have led to a significant literature and development, both in theoretical and applied arenas.

Later in 1989, Mallat [19] studied the properties of multiresolution approximation and proved that it is characterized by a 2π-periodic function. From any MRA, one can derive a function $\psi(t)$ called a wavelet such that $\left\{2^{\frac{j}{2}}\psi(2^j t - k) : j, k \in Z\right\}$ is an orthonormal basis of $L^2(\mathbf{R})$. The MRA showed the full computational power that this new basis for $L^2(\mathbf{R})$ possessed. In the same year, Mallat [20] applied MRA for analysing the information content of the images.

Note that a system $\{\varphi_k : k \in Z\}$ is called a Riesz basis if it is obtained from an orthonormal basis by means of a bounded invertible operator.

Definition The increasing sequence $\{V_k\}_{k\in Z}$ of closed subspaces of $L^2(\mathbf{R})$ with scaling function $\varphi \in V_0$ is called MRA if

(i) $\bigcup_k V_k$ is dense in $L^2(\mathbf{R})$ and $\bigcap_k V_k = \{0\}$,
(ii) $f(t) \in V_k$ iff $f(2^{-k}t) \in V_0$,
(iii) $\{\varphi(t-n)\}_{n\in Z}$ is a Riesz basis for V_0.

Note that (iii) implies that the sequence $\{2^{k/2}\varphi(2^k t - n)\}_{n\in Z}$ is an orthonormal basis for V_k. Let $\psi(t)$ be the mother wavelet, then $\psi(t) = \sum_{n\in Z} a_n \varphi(2t - n)$ and $\left\{2^{k/2}\psi\left(2^k t - n\right)\right\}_{k,n\in Z}$ forms an orthonormal basis for $L^2(\mathbf{R})$ under suitable conditions [21–24].

CAS Wavelets $\psi_{nm}(t) = \psi(k, n, m, t)$ involve four arguments n, k, m and t, where $n = 0, 1, \ldots, 2^k - 1, k$ is assumed any nonnegative integer, m is any integer and t is normalized time. CAS wavelets are defined as [25]

$$\psi_{nm}(t) = \begin{cases} 2^{\frac{1}{2}}\mathrm{CAS}_m(2^k t - n), & \text{for } \frac{n}{2} \le t < \frac{n+1}{2^k}, \\ 0, & \text{otherwise}, \end{cases} \tag{17.3}$$

where

$$\mathrm{CAS}_m(t) = \cos(2m\pi t) + \sin(2m\pi t). \tag{17.4}$$

It is clear that the set of CAS wavelets also forms and orthonormal basis for $L^2([0, 1])$.

17.3 Function Approximation

The function $f(t)$ representing physical fields are either zero or have an infinitely long decaying tail outside a disk of finite radius R. Hence, in most practical applications either the signal $f(t)$ has a compact support or for a given $\varepsilon > 0$ there exists a $R > 0$ such that $\left|\int_R^\infty tf(t)J_v(pt)\,\mathrm{d}t\right| < \varepsilon$.

Therefore, in either case,

$$\hat{F}_v(p) = \int_0^R tf(t)\, J_v(pt)\, \mathrm{d}t$$
$$= \int_0^1 tf(t)\, J_v(pt)\, \mathrm{d}t, (\text{by scaling}) \tag{17.5}$$

known as the finite Hankel transform (FHT) is a good approximation of the HT as given by (17.1). Writing $tf(t) = g(t)$ in Eq. (17.5), we get

$$\hat{F}_v(p) = \int_0^1 g(t) J_v(pt) \mathrm{d}t. \tag{17.6}$$

We may expand $g(t)$ as follows

$$g(t) = \sum_{m=0}^{\infty} \sum_{n=0}^{2^k-1} c_{nm} \psi_{nm}(t), \tag{17.7}$$

where $c_{nm} = \langle g(t),\ \psi_{nm}(t) \rangle$.

with (.,.) denoting the inner product.

By truncating the infinite series (17.7) at levels $m = 2L$ and $n = 2^k - 1$, we obtain an approximate representation for $g(t)$ as

$$g(t) \approx \sum_{m=0}^{2L} \sum_{n=0}^{2^k-1} c_{nm} \psi_{nm}(t) = C^T \psi(t), \tag{17.8}$$

where the matrices C and $\psi(t)$ are $2^k(2L+1) \times 1$ matrices given by

$$C = [c_{0,0}, c_{0,1}, , \ldots, c_{0,2L-1}, c_{10}, \ldots, c_{1,2L}, \ldots, c_{2^k-1,0}, \ldots, c_{2^k-1,2L}]^T \tag{17.9}$$

and

$$\psi(t) = [\psi_{0,0}(t), \psi_{0,1}(t), \ldots, \psi_{0,2L}(t), \psi_{1,0}(t), \ldots, \psi_{1,2L}(t), \psi_{2^k-1,0}(t), \ldots, \psi_{2^k-1,2L}(t)]^T. \tag{17.10}$$

Substituting (17.8) in (17.6), we get

$$\hat{F}_v(p) \approx C^T \int_0^1 \psi(r) J_v(pr) \mathrm{d}r. \tag{17.11}$$

Now (17.11) reduces to

$$\widehat{F}_v(p) \approx C^{\mathrm{T}} \left[\int_0^1 \psi_{0,0}(r) J_v(pr)\mathrm{d}r, \int_0^1 \psi_{0,1}(r) J_v(pr)\mathrm{d}r, \int_0^1 \psi_{0,2}(r) J_v(pr)\mathrm{d}r, \ldots, \int_0^1 \psi_{1.0}(r) J_v(pr)\mathrm{d}r, \ldots, \int_0^1 \psi_{1,4}(r) J_v(pr)\mathrm{d}r \right]^{\mathrm{T}} \tag{17.12}$$

where $\psi_{0,0}, \psi_{0,1} \ldots\ldots\ldots \psi_{1,4}$ are defined through Eq. (17.3). We re-label and write (17.12) as

$$\widehat{F}_v(p) \approx [c_{0,0}, c_{0,1}, \ldots, c_{1,4}][I_n^0, I_n^1, \ldots, I_n^{10}]^{\mathrm{T}}, \tag{17.13}$$

where I_v^l's are the lth place integral in Eq. (17.12).

The integrals arising in Eq. (17.12) are evaluated by using the following formulae [26].

$$\int_0^a J_v(t)\mathrm{d}t = 2 \lim_{N\to\infty} \sum_{z=0}^{N} J_{v+2z+1}(a), \mathrm{Re}v > -1 \tag{17.14}$$

and is calculated with the help of Simpson's one third rule, Simpson's three eight rule.

17.4 Numerical Implementation

Since it is always desirable to test the behaviour of a numerical scheme using simulated data, for which the exact results are known and thus making a comparison between the chosen well known test functions which are widely used by researchers in the area to validate the reliability of proposed method. Here we consider three examples for the numerical solutions on the prescribed method, in order to check the accuracy of our scheme. The simplicity and accuracy of sine-cosine wavelet method is illustrated by computing the absolute error graphically.

$$E\hat{F}_v(p) = F_v(p) - \hat{F}_v(p)$$

In this section, we test the proposed algorithm (17.13) by evaluating the approximate Hankel transforms of 2 well known test function with known analytical Hankel transforms. Note that in all the examples the truncation is done at level $m = 2L$ and $L = 2$, we observed that the accuracy of the method is very high even at such a low level of truncation. Note that the various graphs in the examples are plotted and sample points are chosen as $p = 0.01(0.01)N$, where $N = 60$ in all the figures.

Example 1 Let $f(r) = r^{v} \sin(\frac{\pi r^2}{4}), 0 \le r < 1$, then

$$F_v(p) = \frac{1}{\sqrt{2}} \left(\frac{\pi}{2}\right)^{-v-1} p^{v} \left[U_{v+1}\left(\frac{\pi}{2}, p\right) - U_{v+2}\left(\frac{\pi}{2}, p\right) \right]$$

(obtained from (p. 34, (16), [26] by putting $a = \frac{\pi}{4}, \quad b = 1$), where $U_v(w,\ p)$ is a Lommel's function of two variables,

$$= \frac{1}{\sqrt{2}p} \left[\sum_{\eta=0}^{L} \left[(-1)^{\eta} \left(\frac{\pi}{2p}\right)^{2\eta} \left(J_{v+2\eta+1}(p) - \frac{\pi}{2p} J_{v+2\eta+2}(p) \right) \right] \right] \text{ as } L \to \infty \tag{17.15}$$

The comparison of the approximation $Hv(p)$ (dotted line) with the exact Hankel transform $Fv(p)$ (solid line) is shown in Figs. 17.1 and 17.3 and the error $E(p) = Hv(p) - Fv(p)$ in Figs. 17.2 and 17.4.

Simpson's one third rule

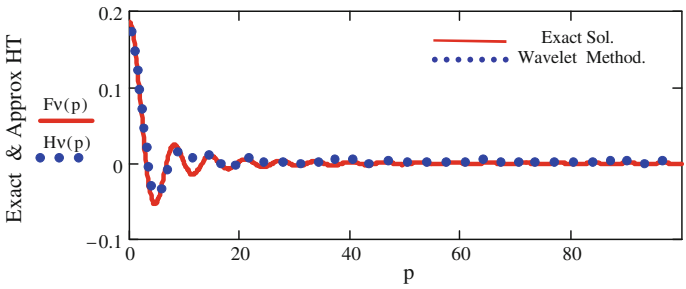

Fig. 17.1 The exact transform, $Fv(p)$ (*solid line*) and the approximate transform, $Hv(p)$ (*dotted-line*) where $v = 0$

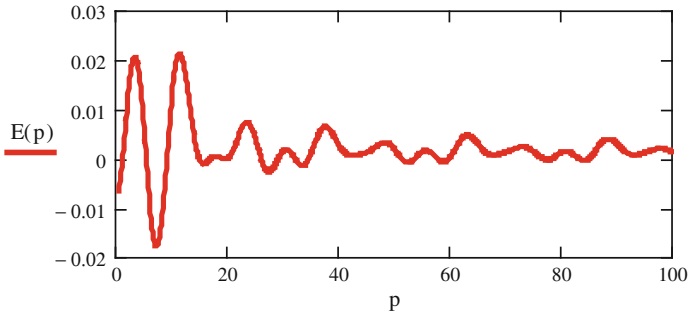

Fig. 17.2 Comparison of the errors

Simpson's three eight rule

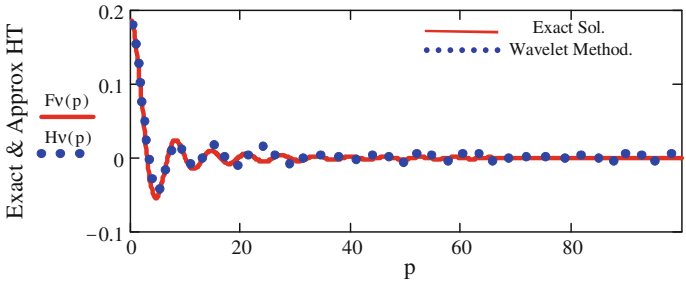

Fig. 17.3 The exact transform, $Fv(p)$ (*solid line*) and the approximate transform, $Hv(p)$ (*dotted-line*) where $v = 0$

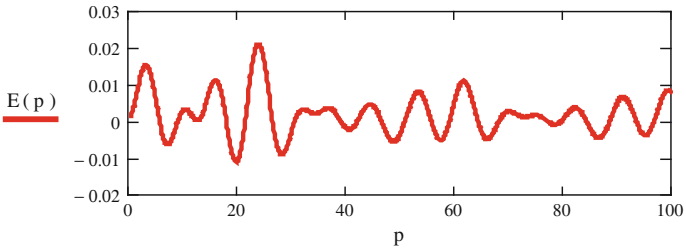

Fig. 17.4 Comparison of the errors

Example 2 In this example, we choose as a test function the generalized version of the top-hat function, given as

$f(r) = r^v\,[H(r) - H(r-a)],\ a > 0$ and $H(r)$ is the step function given by

$$H(r) = \begin{cases} 1, & r \geq 0 \\ 0, & r < 0 \end{cases}.$$

Then,

$$F_v(p) = \frac{J_{v+1}(p)}{p}. \tag{17.16}$$

Guizar-Sicairos [27], took $a = 1$ and $v = 4$ for numerical calculations. We take $a = 1$, $v = 0$, and observe that the error is quite small as shown in Fig. 17.5 and 17.7. The comparison of the approximate with exact transform is shown in Figs. 17.6 and 17.8.

Simpson's one third rule

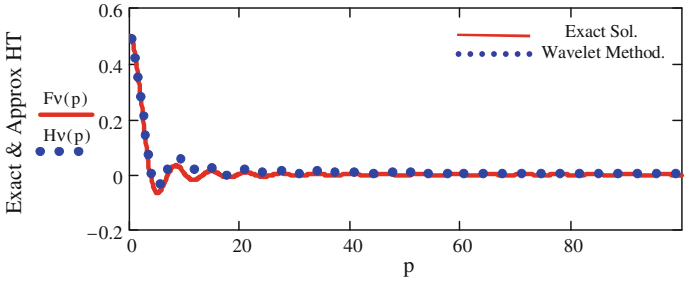

Fig. 17.5 The exact transform, *Fv*(*p*) (*solid line*) and the approximate transform, *Hv*(*p*) (*dotted-line*)

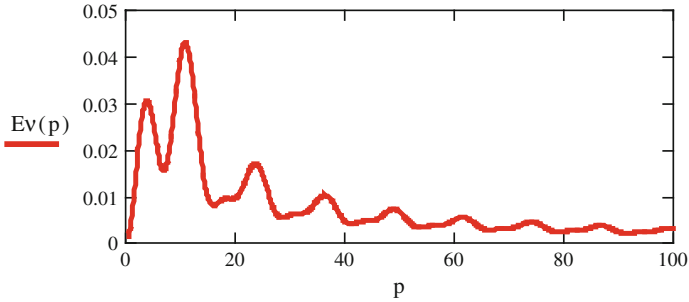

Fig. 17.6 Comparison of the errors

Simpson's three eight rule

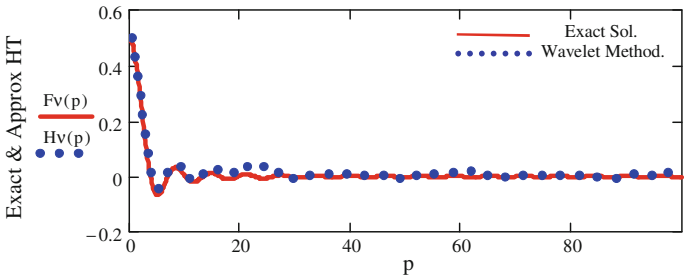

Fig. 17.7 The exact transform, *Fv*(*p*) (*solid line*) and the approximate transform, *Hv* (*p*) (*dotted-line*)

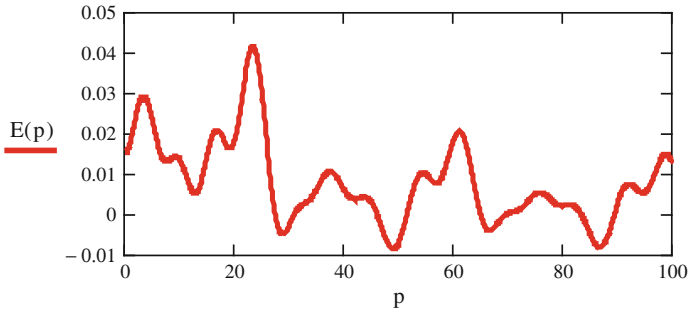

Fig. 17.8 Comparison of the errors

Example 3 Let $f(r) = (1 - r^2)^{1/2}$, $0 \le r \le 1$, then,

$$F_1(p) = \begin{cases} \pi \frac{J_1^2(p/2)}{2p}, & 0 < p < \infty \\ 0, & p = 0 \end{cases}. \tag{17.17}$$

Barakat et al., evaluated $F_1(p)$ numerically using Filon quadrature philosophy but again the associated error is appreciable for $p < 1$; whereas our method give almost zero error in that range. The comparison of the approximation $F(p)$ (dotted line) with the exact Hankel transform $F1(p)$ (solid line) is shown in Figs. 17.9 and 17.11 and the error $E(p) = F(p) - F1(p)$ in Figs. 17.10 and 17.12.

Simpson's one third rule

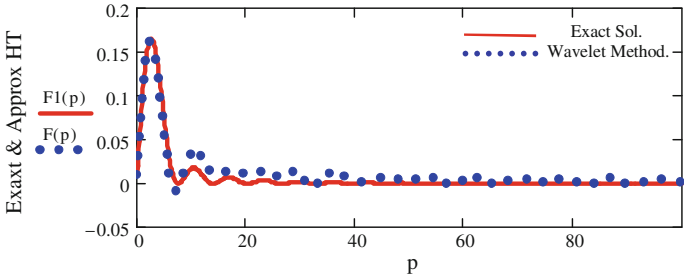

Fig. 17.9 The exact transform, $F1(p)$ (*solid line*) and the approximate transform, $F(p)$ (*dotted-line*)

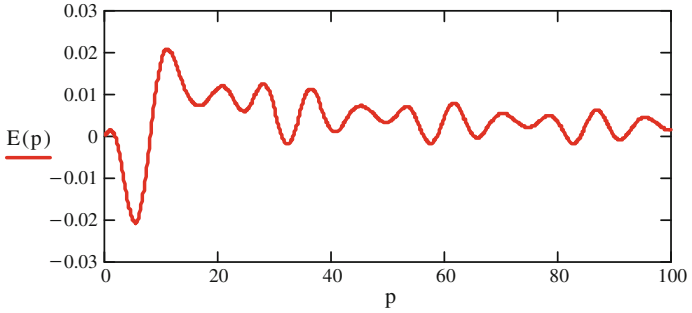

Fig. 17.10 Comparison of the errors

Simpson's three eight rule

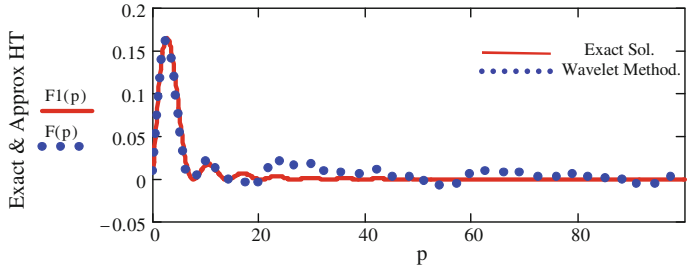

Fig. 17.11 The exact transform, $F1(p)$ (*solid line*) and the approximate transform, $F(p)$ (*dotted-line*)

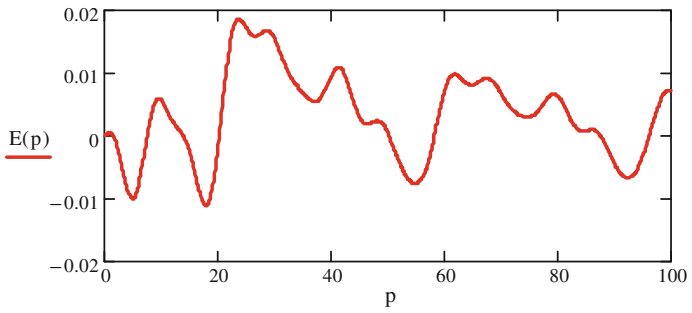

Fig. 17.12 Comparison of the errors

17.5 Summary and Conclusion

Since the basis functions used to construct the wavelets are orthogonal and have compact support, it makes them more useful and simple in actual computations. Also, since the numbers of mother wavelet's components are restricted to one, so they do not lead to the growth of complexity of calculations.

Wavelet method is very simple and attractive [27]. The implementation of current approach in analogy to existed methods is more convenient and the accuracy is high. The numerical example and the compared results support our claim. The difference between the exact and approximate solutions for each example plotted graphically to determine the accuracy of numerical solutions.

17.5.1 Future Work

Since computational work is fully supportive of compatibility of proposed algorithm and hence the same may be extended to other physical problems also. A very high level of accuracy explicitly reflects the reliability of this scheme for such problems. We would like to stress that the approximate solution includes not only time information but also frequency information due to the localization property of wavelet basis; with some change we can apply this method with the help of other wavelet basis.

References

1. Sneddon IN (1972) The use of Integral Transforms. McGraw-Hill
2. Key K (2012) Is the fast Hankel transform faster than quadrature. Geophysics 77(3). http://software.seg.org/2012/0003
3. Anderson WL (1989) A hybrid fast Hankel transform algorithm for electromagnetic modeling. Geophysics 54(2):263–266
4. Agnesi A, Reali GC, Patrini G, Tomaselli A (1993) Numerical evaluation of the Hankel transform: remarks. J Opt Soc Am A 10:1872–1874. doi:10.1364/JOSAA.10.001872
5. Secada D (1999) Numerical evaluation of the Hankel transform. Comput Phys Commun 116:278–294. doi:10.1016/S0010-4655(98)00108-8
6. Cavanagh EC, Cook BD (1979) Numerical evaluation of Hankel Transform via Gaussian-Laguerre polynomial expressions. IEEE Trans Acoust Speech Signal Process. ASSP-27:361–366. doi:10.1109/TASSP.1979.1163253
7. Cree MJ, Bones PJ (1993) Algorithms to numerically evaluate the Hankel transform. Comput Math Appl 26:59–72. doi:10.1016/0898-1221(93)90081-6
8. Murphy PK, Gallagher NC (2003) Fast algorithm for computation of zero-order Hankel transforms. J Opt Soc Am 73:1130–1137. doi:10.1364/JOSA.73.001130
9. Barakat R, Parshall E (1996) Numerical evaluation of the zero-order Hankel transforms using Filon quadrature philosophy. J Appl Math 5:21–26. doi:10.1016/0893-9659(96)00067-5

10. Markham J, Conchello JA (2003) Numerical evaluation of Hankel transform for oscillating function. J Opt Soc Am A 20(4):621–630. doi:10.1364/JOSAA.20.000621
11. Knockaert L (2000) Fast Hankel transform by fast sine and cosine transform: the Mellin connection. IEEE Trans Signal Process 48:1695–1701. doi:10.1109/78.845927
12. Singh VK, Singh OP, Pandey RK (2008) Efficient algorithms to compute Hankel transforms using wavelets. Comput Phys Commun 179:812–818. doi:10.1016/j.cpc.2008.07.005
13. Singh VK, Singh OP, Pandey RK (2008) Numerical evaluation of the Hankel transform by using linear Legendre multi-wavelets. Comput Phys Commun 179:424–429. doi:10.1016/j.cpc.2008.04.006
14. Eldabe NT, Shahed M, Shawkey M (2004) An extension of the finite Hankel transform. Appl Math Comput 151:713–717
15. Barakat R, Sandler BH (1996) Evaluation of first-order Hankel transforms using Filon quadrature philosophy. Appl Math Lett 11:127–131. doi:10.1016/S0893-9659(97)00145-6
16. Haar A (1910) zru theorie der orthogonalen funktionensysteme. Math Ann 69:331–371
17. Daubechies I (1988) Orthonormal bases of compactly supported wavelets. Commun Pure Appl Math 41:909–996
18. Daubechies I (1992) ten lectures on wavelets, CBMS-NSF
19. Mallat S (1989) Multiresolution approximation and wavelet orthonormal bases of L^2 (R). Trans Am Math Soc 315:69–87
20. Mallat S (1989) A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans Pattern Recogn Mach Intell 11:674
21. Postnikov EB (2003) About calculation of the Hankel transform using preliminary wavelet transform. J Appl Math 6:319–325
22. Irfan N, Kapoor S (2011) Quick glance on different wavelets and their operational matrix properties. Int J Res Rev Appl Sci 8(1):65–78. doi:10.1515/nleng-2015-0026
23. Siddiqi AH (2004) Applied functional analysis, numerical methods wavelets and image processing. Marcel Dekker, New York
24. Siddiqi AH (2004) Lead Technical Editor, Theme issues: wavelet and fractal methods in science and engineering. Arab J Sci Eng 29(2C), 28(1C), (Part I and Part II)
25. Irfan N, Siddiqi AH (2015) An application of wavelet technique in Numerical evaluation of Hankel transfroms. Int J Nonlinear Sci Numer Simul 16(6):293–299. doi:10.1515/ijnsns-2015-0031
26. Erdelyi A (ed) (1954) Tables of integral transforms. McGraw-Hill, New York
27. Guizar-Sicairos M, Gutierrez-vega JC (2004) Computation of quasi discrete Hankel transforms of integer order for propagating optical wave fields. J Opt Soc Am A, Col-21, (1):53. doi:10.1364/JOSAA.21.000053
28. Irfan N, Siddiqi AH (2015) A wavelet algorithm for Fourier-Bessel transform arising in optics. Int J Eng Math Article ID 789675, 9p. doi:10.1155/2015/789675