Springer Proceedings in Mathematics & Statistics

Ram N. Mohapatra
Debasis Giri
P. K. Saxena
P. D. Srivastava *Editors*

Mathematics and Computing 2013

International Conference in Haldia, India

Springer

Springer Proceedings in Mathematics & Statistics

Volume 91

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

Springer Proceedings in Mathematics & Statistics

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

Ram N. Mohapatra · Debasis Giri
P. K. Saxena · P. D. Srivastava
Editors

Mathematics and Computing 2013

International Conference in Haldia, India

Springer

Editors
Ram N. Mohapatra
Department of Mathematics
University of Central Florida
Orlando, FL
USA

Debasis Giri
Department of Computer Science and Engineering
Haldia Institute of Technology
Haldia
India

P. K. Saxena
Scientific Analysis Group, Defence Research and Development Organisation
Delhi
India

P. D. Srivastava
Department of Mathematics
Indian Institute of Technology Kharagpur
Kharagpur, West Bengal
India

ISSN 2194-1009 ISSN 2194-1017 (electronic)
ISBN 978-81-322-1951-4 ISBN 978-81-322-1952-1 (eBook)
DOI 10.1007/978-81-322-1952-1
Springer New Delhi Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014945634

Printed on acid-free paper

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

Preface

With a view to have together at one place experts and professionals working in different aspects of research in Mathematics, the idea came to organize an event where people may deliberate upon theoretical as well as computational aspects of mathematics at a common forum. We are delighted to give shape to the idea and host this International Conference on Mathematics and Computing (ICMC 2013) at Haldia Institute of Technology, Haldia, in collaboration with Scientific Analysis Group (DRDO, Ministry of Defence) and Department of Mathematics, IIT Kharagpur.

Haldia is a city and a municipality in Purba Medinapur in the Indian state of West Bengal. Haldia Institute of Technology is one of the premier educational establishments in this part of the State and has had the privilege to have organized international conferences in the past as well.

With three tracks of presentations, contributory papers were called for at ICMC 2013 and 81 papers were submitted to the conference in response. The papers were reviewed on the basis of the significance, novelty, and technical quality. Of these, 22 papers were selected for presentation and publication in the Conference Proceedings.

The papers cover different topics including Cryptography, Algebra, Functional Analysis, Approximation Theory, Fluid Dynamics, etc. The event also included Tutorials on important topics of current thrust. It is expected that the conference will witness eminent personalities both from India and abroad (USA, Canada, Russia, Japan, Hong Kong, Turkey) delivering invited as well as tutorial talks. The prominent speakers from India were from Government R&D Organizations such as Defense Research and Development Organization, Industry and from Academic Institutions such as Indian Statistical Institute Kolkata, IIT Kharagpur, IMSc Chennai, IISc Bangalore, etc.

Three Tutorials were planned preceding the main conference to be given by Prof. Bimal Roy (ISI, Kolkata, India), Prof. M. L. Chaudhry (Royal Military College, Canada), Prof. Ram N. Mohapatra (University of Central Florida, USA), and Prof. A. Vasudevarao, IIT Kharagpur. There were seven invited talks delivered by experts like Prof. Ram N. Mohapatra (University of Central Florida, USA), Prof. R. Balasubramanian (IMSc, Chennai, India), Prof. V. A. Artamonov (Lomonosov Moscow State University, Russia), Prof. C. E. Venimadhavan (IISc, Bangalore, India), Prof. Duan Li (Chinese University of Hong Kong), Prof. Hiroshi Yanagihara (Yamaguchi University, Japan), Prof. Rifat Colak (Firat University, Turkey).

A conference of this kind would not have been possible without the support from different organizations and people across different committees. We are indebted to the Defense Research and Development Organization (DRDO), Ministry of Communication and Information Technology (MCIT), National Board for Higher Mathematics (NBHM), Cryptology Research Society of India (CRSI), Department of Science and Technology (DST), and the Council of Scientific and Industrial Research (CSIR) for sponsoring the event. Their support helped in significantly raising the profile of the conference.

All logistic and general organizational aspects were looked after locally by the Organizing Committee members from the Institute, who spent their time and energy in making the conference a success. The Technical Program Committee and External Reviewers helped in selecting the papers for presentations and working out the technical program. We acknowledge the support and help from all of them.

The organizers also express their hearty thanks to Springer for agreeing to publish the proceedings in its Mathematics and Statistics series.

Last but not the least; our sincere thanks go to all the authors who submitted papers to ICMC 2013 and to all speakers and participants.

We sincerely hope that the readers will find the proceedings stimulating and inspiring. Any constructive suggestions for improvement are welcome.

December 2013

Ram N. Mohapatra
Debasis Giri
P. K. Saxena
P. D. Srivastava

Organizing Committee

Patron

Lakshman Seth, Chairman, Haldia Institute of Technology, Haldia, India

General Co-Chairs

P. K. Saxena, SAG, DRDO, Delhi, India
P. D. Srivastava, IIT Kharagpur, India

Program Co-Chairs

Ram N. Mohapatra, University of Central Florida, USA
Debasis Giri, Haldia Institute of Technology, India

Technical Program Committee

Program Committee Members for the Track, Mathematics

Name	Affiliation
Anders Lindquist	Royal Institute of Technology, Sweden
Anirban Banerjee	IISER, Kolkata, India
Aoi Honda	Kyushu Institute of Technology, Japan
Arcadii Grinshpan	University of South Florida, USA
Ashis SenGupta	ISI Kolkata, India

(continued)

(continued)

Name	Affiliation
Biswa Datta	Northern Illinois University, USA
Dhananjoy Dey	DRDO, Delhi, India
Don Hong	Middle Tennessee State University, USA
G. P. Raja Sekhar	IIT Kharagpur, India
Hans Schönemann	Universität Kaiserslautern, Germany
Indivar Gupta	DRDO, Delhi, India
Jean-Guillaume Dumas	Laboratoire de Modèlisation et Calcul, France
Juan Jesús Barbarán Sánchez	University of Granada, Spain
Kalyan Chakraborty	Harish-Chandra Research Institute, India
Karmeshu	Jawaharlal Nehru University, India
Kinkar Chandra Das	Sungkyunkwan University, Korea
Leonid Bokut	South China Normal University and Sobolev Institute of Mathematics, Russia
Manoj Kumar	The Australian National University, Australia and Harish-Chandra Research Institute, India
M. L. Chaudhry	Royal Military College of Canada, Canada
M. P. Biswal	IIT Kharagpur, India
Narendra Govil	Auburn University, Alabama, USA
Neeraj Misra	IIT Kanpur, India
Partha Sarathi Dey	New York University, New York
Peeyush Chandra	IIT Kanpur, India
P. D. Srivastava	IIT Kharagpur, India
P. R. Mishra	DRDO, Delhi, India
Raazesh Sainudiin	University of Canterbury, New Zealand
Rahul Mazumder	Massachusetts Institute of Technology, USA
Rajen Kumar Sinha	IIT Guwahati, India
Ram N. Mohapatra	University of Central Florida, USA
Ram U. Verma	Texas A&M University-Kingsville, USA
Renzo Pinzani	Dipartimento di Sistemi e Informatica, Italy
Sartaj Hasan	DRDO, Delhi, India
Sasmita Barik	IIT Bhubaneswar, India
Saugata Basu	Purdue University, USA
Somesh Kumar	IIT Kharagpur, India
S. Ponnusamy	IIT Madras, India and ISI Chennai, India
S. Sundar	IIT Madras, India
Thomas Schoenemann	Institut für Sprache und Information, Germany
U. C. Gupta	IIT Kharagpur, India
Varadachariar Kannan	University of Hyderabad, India
Vilém Novák	University of Ostrava, Czech Republic
Wei Sun	Concordia University, Canada
Yulii D. Shikhmurzaev	University of Birmingham Edgbaston, UK

Program Committee Members for the Track, Computing

Name	Affiliation
Abhijit Das	IIT Kharagpur, India
Adam Krzyzak	Concordia University, Canada
Amitava Bhattacharya	Tata Institute of Fundamental Research, Mumbai, India
Anders Hast	Uppsala University, Sweden
Ashok Kumar Das	IIIT, Hyderabad, India
Bernardete Ribeiro	University of Coimbra, Portugal
Bidyut Patra	VTT Technical Research Centre, Finland
Bimal Roy	ISI Kolkata, India
B. K. Dass	University of Delhi, India
C. E. Veni Madhavan	IISC Bangalore, India
Chi-Yi Tsai	Tamkang University, Taiwan
Debasis Giri	Haldia Institute of Technology, India
Friedhelm Schwenker	University of Ulm, Germany
Genge Bela	Via E. Fermi, Ispra, Italy
Huaqun Guo	Institute for Infocomm Research (I2R), Singapore
Jiqiang Lu	Institute for Infocomm Research (I2R), Singapore
Jaydeb Bhoumik	Haldia Institute of Technology, India
Joan-Josep Climent	Universitat d'Alacant, Spain
Jörn-Marc Schmidt	Applied Information Processing and Communications, Austria
Junwei Cao	Tsinghua University, Beijing, China
Igor Vitalievich Kotenko	St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, Russia
Kamisetty R. Rao	University of Texas at Arlington, USA
K. K. Biswas	IIT Delhi, India
Krishnaiyan Thulasiraman	University of Oklahoma, USA
Krzysztof Szczypiorski	Warsaw University of Technology, Poland
Lars Bengtsson	Chalmers University of Technology, Gothenburg, Sweden
Laurent Heutte	Université de Rouen, France
Lipo Wang	Nanyang Technological University, Singapore
Lotfi A. Zadeh	University of California, Berkeley, USA
Madhumangal Pal	Vidyasagar University, India
Maiga Chang	Athabasca University, Canada
Manik Lal Das	DA-IICT, Gandhinagar, India
Marco Baldi	Università Politecnica delle Marche, Italy
Mohammed Chadli	Université de Picardie-Jules Verne, France
Narendra S. Chaudhari	IIT Indore, India
Neeli R. Prasad	Aalborg University, Denmark
Phalguni Gupta	IIT Kanpur, India
P. K. Saxena	DRDO, Delhi, India

(continued)

(continued)

Name	Affiliation
P. Nagabhushan	University of Mysore, India
Punam Bedi	University of Delhi, India
Rajesh Pillai	DRDO, Delhi, India
Raj Jain	Washington University in St. Louis, USA
Rashid Mehmood	University of Huddersfield, UK
Rahul Jain	National University of Singapore, Singapore
Ronald R. Yager	Machine Intelligence Institute, New Rochelle, NY, USA
Sanasam Ranbir Singh	IIT Guwahati, India
Sanjoy Saha	Jadavpur University, India
Shanta Laishram	ISI Delhi, India
Seong Han Shin	Research Institute for Secure Systems (RISEC), National Institute of Advanced Industrial Science and Technology (AIST), Japan
Shyam S. Chakraborty	Helsinki University of Technology, Finland
Soumen Maity	Indian Institute of Science Education and Research, Pune, India
Sriparna Saha	IIT Patna, India
S. K. Pal	DRDO, Delhi, India
Sugata Gangopadhyay	IIT Roorkee, India
Sukhendu Das	IIT Madras, India
Sukumar Ghosh	The University of Iowa, USA
Sung-Bae Cho	Yonsei University, Korea
Sushil Jajodia	George Mason University, USA
Sushmita Ruj	IIT Indore, India
Tai-hoon Kim	Sungshin W. University, South Korea
Takeshi Furuhashi	Nagoya University, Japan
Wang Jiangang	Institute for Infocomm Research (I2R), Singapore
Wei Zhong	University of South Carolina Upstate, USA
Weifa Liang	The Australian National University, Australia
Yunquan Zhang	The Chinese Academy of Sciences, Beijing, China
Yuriy S. Shmaliy	Universidad de Guanajuato, Mexico

Additional Reviewers

Name	Affiliation
Ameeya K. Nayak	IIT Roorkee, India
B. C. Tripathy	Institute of Advanced Study in Science and Technology, Guwahati, India
Debjani Chakraborty	IIT Kharagpur, India
Duan Li	The Chinese University of Hong Kong, Hong Kong
Ekrem Savas	Istanbul Commerce University, Istanbul
G. Athithan	DRDO, Delhi, Indial
Kamini Malhotra	DRDO, Delhi, India
Mikail Et	Firat University Elazig, Turkey
M. Mursaleen	Aligarh Muslim University, India
Mukund Mishra	University of Delhi, India
Neelam Verma	DRDO, Delhi, India
Pradipta Maji	ISI Kolkata, India
Pratibha Yadav	DRDO, Delhi, India
Rifat Colak	Firat University, Turkey
Sumanta Sarkar	University of Calgary, Canada
Youesh Kumar	DRDO, Delhi, India

Local Organizing Committee

A. K. Dey, Asish Lahiri, Anjan Mishra, M. K. Pandit, Debasis Giri (Organizing Chair), Sk. Sahnawaj, Soumen Paul, Tarun Ghosh, Apratim Mitra, Subhabrata Barman, Sourav Mandal, Subhankar Joardar, Susmit Maity, Palash Roy, Mrinmoy Sen, Arif Ahmed.

Message from the General Chairs

As we all are aware, Mathematics has always been a discipline of interest not only to theoreticians but also to all practitioners irrespective of their specific profession. Be it Science, Technology, Economics, Commerce, or even Sociology, new Mathematical principles and models have been emerging and helping in new research and in drawing inferences from practical data as well as through logic. The past few decades have seen enormous growth in applications of Mathematics in different areas multidisciplinary in nature. Cryptography and signal processing are such areas, which have got more focus recently due to the need for securing communication while connecting with others. With emerging computing facilities and speeds, a phenomenal growth has happened in the problem solving area. Earlier, some observations were made and conjectures were drawn which remained conjectures till somebody could either prove it theoretically or found counter examples. But today, we can write algorithms and use computers for long calculations, verifications, or for generation of huge amounts of data. With available computing capabilities, we can find factors of very large integers of the size of hundreds of digits; we can find inverses of very large size matrices and solve a large set of linear equations, and so on. Thus, Mathematics and Computations have become more integrated areas of research these days, and it was thought to organize an event where thoughts may be shared by researchers and new challenging problems could be deliberated for solving these.

Apart from many other interdisciplinary areas of research, cryptography has emerged as one of the most important areas of research with discrete mathematics as a base. Several research groups are actively pursuing the research on different aspects of cryptology not only in terms of new cryptoprimitives and algorithms but a whole lot of concepts related to authentication, integrity, and security proofs/protocols are being developed, many times with open and competitive evaluation mechanism to evolve standards.

As conferences, seminars, and workshops are the mechanisms to share knowledge and new research results giving us a chance to get new innovative ideas for futuristic needs as threats and computational capabilities of adversaries are ever increasing, it was thought appropriate to organize the present conference focused on Mathematics and Computations covering theoretical as well as practical aspects of research, Cryptography being one of these.

Eminent personalities working in Mathematical Sciences and related areas were invited from abroad as well as from within the country to deliver invited talks and Tutorials for participants. The talks by these speakers covered a wide spectrum, viz., Number Theoretic Concepts, Cryptography, Algebraic Concepts like Quasi Groups and applications, etc. The conference was spread over 4 days (December 26–29, 2013) with the first day dedicated to Tutorials. The main conference was planned with special talks by experts and paper presentations in each session.

We hope that the conference met the aspirations of the participants and its objective of ideas and current research being shared and new targets/problems identified in the domain of Coding theory, Cryptography, Computational number theory, Algebra, Frame theory, Optimizations, Stochastic Processes, Compressive Sensing, Functional analysis, Complex variables etc., so that the researchers and students would get new directions to pursue their future research.

December 2013

P. K. Saxena
P. D. Srivastava

Message from the Program Chairs

It is a great pleasure for us to organize the International Conference on Mathematics and Computing-2013 to be held from December 26 to 29 at the Haldia Institute of Technology, Purba Medinipur, West Bengal, India. Our main goal is to provide an opportunity to the participants to learn about contemporary research in Mathematics and Computing and exchange ideas. With this aim in mind, we carefully chose the invited speakers and the tutorial speakers. It is our sincere hope that the conference will help the participants in their research and training and open new avenues for work in Mathematics and Computing.

On 26 December 2013, there will be tutorials. The conference will begin after a formal opening ceremony on 27 December 2013. There will be seven invited one-hour talks and 22, contributed half-hour talks. Our speakers/contributors come from Austria, Canada, Hong Kong, India, Japan, Philippines, Russia, Turkey, and USA.

After an initial call for papers, 81 papers were submitted in the conference. All submitted papers were sent to referees and after refereeing, 22 papers were recommended for publication. The proceedings of the conference will be published by Springer (Mathematics and Statistics Series).

We are grateful to the speakers, participants, referees, organizers, sponsors, and funding agencies for their support and help without which it would have been impossible to organize the conference. We owe our gratitude to the volunteers who work behind the scene in taking care of the details in making this conference a success.

December 2013

Ram N. Mohapatra
Debasis Giri

Contents

About the Editors

Ram N. Mohapatra is Professor of Mathematics, University of Central Florida, Orlando, USA. He received his Ph.D. degree from the University of Jabalpur, India, in 1968. Earlier, he taught at Sambalpur University in India, American University in Beirut, Lebanon, University of Alberta, and York University, Canada, prior to coming to Orlando. His area of research is Mathematical Analysis and he is the author of two books, two edited monographs, and over 120 research papers. He referees articles for professional journals and serves as a member of the editorial board of a number of journals.

Debasis Giri is Professor in the Department of Computer Science and Engineering, Haldia Institute of Technology, India. His topics of interest include discrete mathematics, cryptography, information security, coding theory, advanced algorithms, design and analysis of algorithms, and formal languages and automata theory. His research interests include cryptography, network security, security in wireless sensor networks, and security in VANETs. Dr. Giri has delivered several talks and guest lectures at various universities and conferences. He is supervisor of three Ph.D. research scholars. Further, he guided many B.Tech. and M.Tech. students. He is associate editor of the *Journal of Security and Communication Networks* (Wiley), and the *Journal of Electrical and Computer Engineering Innovations*. Further, he is editorial board member and reviewer of many reputed international journals. He is also program committee member of many international conferences. He is life member of the Cryptology Research Society of India. He received his Ph.D. on "Cryptanalysis and Improvement of Protocols for digital signature, smart-card authentication and access control" from Indian Institute of Technology Kharagpur. He did both his M.Tech. and M.Sc. from Indian Institute of Technology Kharagpur. He secured tenth position in all India rank with percentile score of 98.42 in the Graduate Aptitude Test in Engineering (GATE) Examination in 1999. Dr. Giri has published more than 25 technical papers in several internal journals and proceedings.

P. K. Saxena is Director of one of the R&D labs of Defense Research Development Organization (DRDO) under the Ministry of Defense. Dr. Saxena, an outstanding scientist, has done his Ph.D. on "Radical theory of near rings" from Indian Institute of Technology Kanpur. Before joining DRDO, he taught

Mathematics at National Institute of Technology (NIT) Silchar and at National Defense Academy (NDA), Pune. Dr. Saxena has published about 62 research papers in several journals and conferences on interdisciplinary topics such as algebra, cryptology, fuzzy logic, artificial neural networks, and speech technology. He has led many important R&D projects and guided many students in important engineering projects. Three scholars have received the Ph.D. degree under his supervision and several others are registered. Dr. Saxena has delivered talks, guest lectures, keynote addresses at various forums apart from organizing international conferences as general chair. He has been in program committees of many international conferences. He has authored a book on *Cryptology* (in Hindi) which was awarded first prize by DRDO in 1997.

P. D. Srivastava is Professor of Mathematics at Indian Institute of Technology Kharagpur, India. During his 34 years of teaching career, he taught several courses such as functional analysis, topology, numerical analysis, measure theory, real analysis, complex analysis, and calculus to undergraduate and postgraduate students. Besides teaching, Professor Srivastava is equally devoted to research activities. He has approximately 51 papers to his credit published in several international journals. He has supervised 10 research scholars for the Ph.D. degree in Mathematics and one for PDF. Various universities have invited him for lectures and keynote addresses at their conferences. Various universities also invite him as an expert in the faculty selection as well as an expert to adjudicate Ph.D. theses. He is also reviewer for the *Mathematical Reviews* as well as paper referee for many journals. He did his Ph.D. from Indian Institute of Technology Kanpur and B.Sc., M.Sc. degrees from Kanpur University. Dr. Srivastava is not only an established researcher in his area but also a teacher par excellence. His style of lecture presentation and full command over the subject impresses students, which is reflected in the Students' Profile Forms (teaching assessment by students).

Contributors

Avishek Adhikari Department of Pure Mathematics, University of Calcutta, Kolkata, India

Anuradha University of Delhi, Delhi, India

Jerico B. Bacani Department of Mathematics and Computer Science, College of Science, University of the Philippines Baguio, Baguio, Philippines

A. D. Banik School of Basic Sciences, Indian Institute of Technology Bhubaneswar, Bhubaneswar, India

Dana Baxley Department of Mathematics, University of Central Florida, Orlando, FL, USA

Jaydeb Bhaumik Haldia Institute of Technology, Haldia, India

A. K. Bhurjee Indian Institute of Technology Kharagpur, Kharagpur, India

M. P. Biswal Department of Mathematics, Indian Institute of Technology, Kharagpur, India

Debjani Chakraborty Department of Mathematics, Indian Institute of Technology Kharagpur, Kharagpur, West Bengal, India

A. K. B. Chand Department of Mathematics, Indian Institute of Technology Madras, Chennai, India

M. L. Chaudhry Department of Mathematics and Computer Science, Royal Military College of Canada, Kingston, ON, Canada

A. Choudhary Indian Institute of Technology Ropar, Rupnagar, Punjab, India

Angsuman Das Department of Mathematics, St. Xavier's College, Kolkata, India

Satyajit Das Indian Institute of Technology, Patna, India

Manish Kant Dubey Scientific Analysis Group, Defence Research and Development Organisation, Delhi, India

Navneet Gaba Scientific Analysis Group, Defence Research and Development Organisation, Delhi, India

Debdas Ghosh Department of Mathematics, Indian Institute of Technology Kharagpur, Kharagpur, West Bengal, India

Debashree Guha Indian Institute of Technology, Patna, India

Indivar Gupta Scientific Analysis Group, Defence Research and Development Organisation, Delhi, India

Christopher Huff Department of Mathematics, University of Central Florida, Orlando, FL, USA

G. Icoz Department of Mathematics, Gazi University, Teknikokullar, Turkey

Nabakumar Jana Indian Institute of Technology Kharagpur, Kharagpur, India

Balasubramaniam Jayaram Department of Mathematics, Indian Institute of Technology Hyderabad, Yeddumailaram, India

Somesh Kumar Indian Institute of Technology Kharagpur, Kharagpur, India

Amit Maji Department of Mathematics, Indian Institute of Technology Kharagpur, Kharagpur, West Bengal, India

Sayantan Mandal Department of Mathematics, Indian Institute of Technology Hyderabad, Yeddumailaram, India

Atanu Manna Indian Institute of Technology Kharagpur, Kharagpur, India

S. C. Martha Indian Institute of Technology Ropar, Rupnagar, Punjab, India

Prasanna Raghaw Mishra Scientific Analysis Group, Defence Research and Development Organisation, Delhi, India

Neeraj Misra Indian Institute of Technology Kanpur, Kanpur, India

Ram N. Mohapatra Department of Mathematics, University of Central Florida, Orlando, FL, USA

C. Nahak Department of Mathematics, Indian Institute of Technology Kharagpur, Kharagpur, India

G. Panda Faculty of Mathematics Department, Indian Institute of Technology Kharagpur, Kharagpur, India

Gopinath Panda School of Basic Sciences, Indian Institute of Technology Bhubaneswar, Bhubaneswar, India

Saliha Pehlivan Department of Mathematics, University of Central Florida, Orlando, FL, USA

Gunther Peichl Institute for Mathematics and Scientific Computing, University of Graz, Graz, Austria

Avik Pradhan Department of Mathematics, Indian Institute of Technology, Kharagpur, India

Dipanwita Roy Chowdhury Indian Institute of Technology Kharagpur, Kharagpur, India

N. K. Sahu Department of Mathematics, Indian Institute of Technology Kharagpur, Kharagpur, India

P. D. Srivastava Department of Mathematics, Indian Institute of Technology Kharagpur, Kharagpur, West Bengal, India

A. Vasudevarao Department of Mathematics, Indian Institute of Technology Kharagpur, Kharagpur, India

P. Viswanathan Department of Mathematics, Indian Institute of Technology Madras, Chennai, India

Chapter 1
Propagation of Water Waves in the Presence of Thin Vertical Barrier on the Bottom Undulation

A. Choudhary and S. C. Martha

Abstract The problem involving diffraction of water waves by submerged thin vertical barrier over irregular bottom is examined using linearized theory of water waves. While formulating the problem mathematically, a mixed boundary value problem (BVP) occurs. The problem is solved using perturbation theory along with least-squares method and Green's integral theorem. The first order reflection and transmission coefficients are obtained in terms of integrals involving the shape function $c(x)$ representing the bottom undulation and the solution of the scattering problem by the submerged barrier. A special case of bottom undulation is considered to evaluate the first order reflection and transmission coefficients in detail. The numerical results of these coefficients are shown graphically.

Keywords Water wave scattering · Bottom undulation · Vertical barrier · Perturbation analysis · Least-squares method · Green's integral theorem · Reflection and transmission coefficients

1 Introduction

The interaction of water waves with vertical barriers has received considerable attention from many researchers. These problems are important due to their applications in ocean engineering such as breakwaters and wavemakers which protect a harbor or

A. Choudhary is grateful to the University Grants Commission (UGC), Government of India, for providing the Junior Research Fellowship for pursuing Ph.D. degree at the Indian Institute of Technology Ropar, India. S. C. Martha thanks the Indian Institute of Technology Ropar, India, for providing all necessary facilities.

A. Choudhary (✉) · S. C. Martha
Indian Institute of Technology Ropar, Nangal Road, Rupnagar 140001, Punjab, India
e-mail: arunc@iitrpr.ac.in

S. C. Martha
e-mail: scmartha@gmail.com

R. N. Mohapatra et al. (eds.), *Mathematics and Computing 2013*,
Springer Proceedings in Mathematics & Statistics 91,
DOI: 10.1007/978-81-322-1952-1_1, © Springer India 2014

marinas from the rough sea. Dean [3] obtained the linearized solution of water wave scattering by the submerged plane barrier which extended infinitely downwards in deep water. Ursell [14] derived the solution of the problem of diffracted waves by thin vertical barrier partially immersed in deep water using the singular integral equation approach based on Havelock's expansion. Porter [12] solved the problem involving wave transmission through a gap in a vertical barrier in deep water using the complex variable method along with Green's integral theorem. Losada et al. [4] used Least-squares is bounded assolution of the problem involving scattering of water waves by different thin barriers. Mandal and Dolai [7] and Porter and Evans [13] obtained the solution of this problem using Galerkin approximation method.

The problems involving diffraction of water waves by undulating bottom topography of a seabed are also interesting to study because of their significance in finding the effect of naturally occurring bottom undulation such as sand ripples on wave propagation. Miles [11] derived the reflection and transmission coefficients using the finite cosine transform technique when oblique waves are incident to a cylindrical obstacle. Davies [1] discussed the problem of the reflection of incident waves by irregular bottom using Fourier transform technique. Davies and Heathershaw [2] compared the theoretical results of [1] by conducting experiments in wave tank. Martha et al. [10] solved the problem involving water wave scattering by small undulation on seabed using Fourier transform method and residue theorem.

However, looking at the present situation, the vertical barrier submerged on the undulating seabed will serve as an effective breakwater for coastal engineering. The literature in this direction is very limitted. Mandal and Gayen [8] solved one such problem using multi-term Galerkin approximation and Green's integral theorem.

In this paper, we discuss the problem involving diffraction of water waves by undulating bed topography and submerged vertical barrier. A mixed boundary value problem occurs while formulating the problem mathematically. On applying the perturbation analysis involving small parameter ε which characterizes the smallness of bottom undulation, two boundary value problems, namely BVP-I (by equating the coefficients of the powers of ε^0) and BVP-II (by equating the coefficients of the powers of ε), are obtained. The BVP-I corresponds to the problem of scattering of water waves by vertical barrier in water of uniform finite depth. The solution of the BVP-I is obtained by least-squares method for which the error is minimum. The zeroth order reflection and transmission coefficients involved in BVP-I are also determined. The BVP-II which involves the solution of BVP-I, represents the radiation problem in water of uniform finite depth. Green's integral theorem is used to obtain the solution of BVP-II and the first order reflection and transmission coefficients. The numerical results for these reflection and transmission coefficients are shown graphically.

2 Mathematical Formulation

A right-handed rectangular Cartesian coordinate system is employed in which x-axis is the position of undisturbed free surface and y-axis is taken positive vertically downwards from the origin. The bottom of the sea has small undulation and is

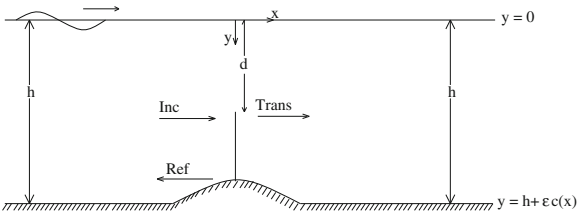


Fig. 1 Scattering of water waves by submerged vertical barrier with bottom undulation

described by $y = h+\varepsilon c(x)$, where $c(x)$ is a continuous bounded function describing the shape of the bottom undulation, h denotes the uniform finite depth of the ocean far to either side of the undulation of the bottom so that $c(x) \to 0$ as $|x| \to \infty$ and the nondimensional number $\varepsilon(\ll 1)$ gives the measure of smallness of the undulation. Consider a thin vertical barrier which is submerged on the bottom undulation, whose position is located at $x = 0$, $y \in L = [d, h]$ (Fig. 1).

It is assumed that the fluid is inviscid, incompressible and the motion is irrotational. If the motion is to be simple harmonic in time with angular frequency σ, then the velocity potential Φ which describes the fluid motion can be expressed as $\Phi(x, y, t) = Re\{\phi(x, y)e^{-i\sigma t}\}$. Then the complex valued potential ϕ satisfies the Laplace equation

$$\frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial y^2} = 0, \text{ in the fluid region,} \tag{1}$$

along with conditions:

$$\text{Free surface condition: } \frac{\partial\phi}{\partial y} + K\phi = 0, \quad \text{on } y = 0, \quad \left(\text{with } K = \frac{\sigma^2}{g}\right) \tag{2}$$

$$\text{Bottom condition: } \frac{\partial\phi}{\partial n} = 0, \quad \text{on } y = h + \varepsilon c(x), \tag{3}$$

$$\text{Condition on barrier: } \frac{\partial\phi}{\partial x} = 0, \quad \text{on } x = 0, y \in L, \tag{4}$$

$$\text{Condition across gap: } \frac{\partial\phi}{\partial x}|_{x=0^-} = \frac{\partial\phi}{\partial x}|_{x=0^+}, \quad \text{on } x = 0, y \in \bar{L}, \tag{5}$$

$$\phi|_{x=0^-} = \phi|_{x=0^+}, \quad \text{on } x = 0, y \in \bar{L}, \tag{6}$$

$$\text{Edge condition: } r^{1/2}\nabla\phi \text{ is bounded as } r \to 0, \tag{7}$$

$$\text{Far-field behavior: } \phi(x, y) \sim \begin{cases} \phi_{inc}(x, y) + R\phi_{inc}(-x, y) & \text{as } x \to -\infty \\ T\phi_{inc}(x, y) & \text{as } x \to \infty. \end{cases} \tag{8}$$

where $\dfrac{\partial}{\partial n}$ is the normal derivative at a point (x, y) on the bottom, g is gravitational constant, r is the distance from a submerged end of the barrier, R is the reflection coefficient, and T is transmission coefficient, $\bar{L} = 0 \leq y \leq d$ and ϕ_{inc} denotes the incident wave.

The incident wave in a finite depth of water h can be written as

$$\phi_{inc}(x, y) = \psi_0(y) e^{i\hat{k}_0 x}$$

where $\psi_0(y) = N_0^{-1} \cosh \hat{k}_0(h - y)$ with $N_0 = \left[\dfrac{4\hat{k}_0 h}{2\hat{k}_0 h + \sinh 2\hat{k}_0 h} \right]^{-1/2}$, $\hat{k}_0$ is the wave number of incident wave, the positive real root of the transcendental equation

$$K - k \tanh kh = 0. \tag{9}$$

3 Method of Solution

The bottom condition (3) can be approximated up to the first order of the small parameter ε as

$$\frac{\partial \phi}{\partial y} - \varepsilon \frac{\mathrm{d}}{\mathrm{d}x}\{c(x)\phi_x\} + O(\varepsilon^2) = 0 \quad \text{on } y = h. \tag{10}$$

The approximate boundary condition (10) suggests that ϕ, R, T can be expanded in terms of ε as given by

$$\begin{aligned} \phi(x, y) &= \phi_0 + \varepsilon\phi_1 + O(\varepsilon^2), \\ R &= R_0 + \varepsilon R_1 + O(\varepsilon^2), \\ T &= T_0 + \varepsilon T_1 + O(\varepsilon^2). \end{aligned} \tag{11}$$

Substituting the expressions of $\phi(x, y)$, R, and T from relation (11) into (1), (2), (4)–(8), and (10) and equating the coefficients of ε^0 and ε from both sides, the functions $\phi_0(x, y)$ and $\phi_1(x, y)$ satisfy the following BVPs:

BVP-I: The function $\phi_0(x, y)$ satisfies

$$\frac{\partial^2 \phi_0}{\partial x^2} + \frac{\partial^2 \phi_0}{\partial y^2} = 0 \quad \text{in the fluid region,} \tag{12}$$

$$\frac{\partial \phi_0}{\partial y} + K\phi_0 = 0, \quad \text{on } y = 0, \tag{13}$$

$$\frac{\partial \phi_0}{\partial y} = 0 \quad \text{on } y = h, \tag{14}$$

$$\frac{\partial \phi_0}{\partial x} = 0, \quad \text{on } x = 0, y \in L, \tag{15}$$

$$\frac{\partial \phi_0}{\partial x} \mid_{x=0^-} = \frac{\partial \phi_0}{\partial x} \mid_{x=0^+}, \quad \text{on } x = 0, y \in \bar{L}, \tag{16}$$

$$\phi_1 \mid_{x=0^-} = \phi_1 \mid_{x=0^+} \quad \text{on } x = 0, y \in \bar{L}, \tag{17}$$

$$r^{1/2}\nabla\phi_0 \quad \text{is bounded as } r \to 0, \tag{18}$$

$$\phi_0(x, y) \sim \begin{cases} (e^{i\hat{k_0}x} + R_0 e^{-i\hat{k_0}x})\psi_0(y) \text{ as } x \to -\infty \\ T_0 e^{i\hat{k_0}x}\psi_0(y) \qquad\qquad \text{as } x \to \infty. \end{cases} \tag{19}$$

BVP-II: The function $\phi_1(x, y)$ satisfies

$$\frac{\partial^2 \phi_1}{\partial x^2} + \frac{\partial^2 \phi_1}{\partial y^2} = 0 \quad \text{in the fluid region,} \tag{20}$$

$$\frac{\partial \phi_1}{\partial y} + K\phi_1 = 0, \quad \text{on } y = 0, \tag{21}$$

$$\frac{\partial \phi_1}{\partial y} = \frac{d}{dx}\{c(x)\frac{\partial \phi_0}{\partial x}\} = p(x) \text{ (say) on } y = h, \tag{22}$$

$$\frac{\partial \phi_1}{\partial x} = 0, \quad \text{on } x = 0, y \in L, \tag{23}$$

$$\frac{\partial \phi_1}{\partial x} \mid_{x=0^-} = \frac{\partial \phi_1}{\partial x} \mid_{x=0^+}, \quad \text{on } x = 0, y \in \bar{L}, \tag{24}$$

$$\phi_1 \mid_{x=0^-} = \phi_1 \mid_{x=0^+} \quad \text{on } x = 0, \ y \in \bar{L}, \tag{25}$$

$$r^{1/2}\nabla\phi_1 \text{ is bounded as } r \to 0, \tag{26}$$

$$\phi_1(x, y) \sim \begin{cases} R_1 e^{-i\hat{k_0}x}\psi_0(y) \text{ as } x \to -\infty \\ T_1 e^{i\hat{k_0}x}\psi_0(y) \quad \text{as } x \to \infty. \end{cases} \tag{27}$$

Here, the BVP-I represents the scattering of water waves by thin vertical barriers in water of finite depth h. The solution for ϕ_0 can be expressed as

$$\phi_0(x, y) = \begin{cases} (e^{i\hat{k_0}x} + R_0 e^{-i\hat{k_0}x})\psi_0(y) + \sum_{n=1}^{\infty} A_n e^{k_n x}\psi_n(y) \text{ as } x \to -\infty, \\ T_0 e^{i\hat{k_0}x}\psi_0(y) + \sum_{n=1}^{\infty} B_n e^{-k_n x}\psi_n(y) \qquad\qquad \text{as } x \to \infty, \end{cases} \tag{28}$$

where $\pm ik_n$, $(n = 1, 2, \ldots)$ are the purely imaginary roots of the transcendental Eq. (9), A_n, B_n, $(n = 1, 2, \ldots)$ are constants to be determined and $\psi_n(y) = N_n^{-1}$

$\cos k_n(h-y)$ with $N_n = \left[\dfrac{4k_nh}{2k_nh + \sin 2k_nh}\right]^{-1/2}$.

Now, using the boundary conditions (16) and (17), we obtain

$$R_0 + T_0 = 1 \text{ and } A_n = -B_n, \tag{29}$$

$$\text{and} \quad \frac{1}{h}\psi_0(y) + \frac{1}{h}\sum_{n=0}^{\infty} A_n\psi_n(y) = 0, \quad \text{on } x = 0, y \in \bar{L}, \tag{30}$$

where $A_0 = R_0 - 1, k_0 = -i\hat{k_0}$.
Again, using the boundary condition (15), we get

$$\sum_{n=0}^{\infty} A_n(k_nh)\frac{1}{h}\psi_n(y) = 0, \quad \text{on } x = 0, y \in L. \tag{31}$$

These two relations, (30) and (31), can be combined to make one fixed boundary condition which specifies the potential as given by

$$G(y) = 0, \quad 0 < y < h, \tag{32}$$

where

$$G(y) = \frac{1}{h}\psi_0(y) + \frac{1}{h}\sum_{n=0}^{\infty} A_n\psi_n(y), \quad \text{on } y \in \bar{L},$$

and

$$G(y) = \sum_{n=0}^{\infty} A_n(k_nh)\frac{1}{h}\psi_n(y), \quad \text{on } y \in L.$$

The relation (32) represents an overdetermined system of equations which can be solved by applying least-squares method which requires

$$\text{Error} = \left(\int_o^h |G(y)|^2 \mathrm{d}y\right)^{1/2} = \left(\int_{y \in \bar{L}} |G(y)|^2 \mathrm{d}y + \int_{y \in L} |G(y)|^2 \mathrm{d}y\right)^{1/2} \tag{33}$$

to be minimum.
This error will be minimum when

$$\int_{y \in \bar{L}} G^*(y)\frac{\partial G(y)}{\partial A_m}\mathrm{d}y + \int_{y \in L} G^*(y)\frac{\partial G(y)}{\partial A_m}\mathrm{d}y = 0, \quad \mathrm{m} = 0, 1, 2, \ldots \tag{34}$$

where $G^*(y)$ is the complex conjugate of $G(y)$.
Substituting the expressions of $G(y)$, $G^*(y)$ and their derivatives, we get

$$A_m - \sum_{n=0}^{\infty} A_n^* c_{nm} \left[1 - (k_m h)(k_n^* h)\right] = \delta_{0m} - c_{0m} \quad (\mathrm{m} = 0, 1, 2, \ldots), \tag{35}$$

where

$$c_{nm} = \frac{1}{h} \int_d^h \psi_n(y)\psi_m(y)\mathrm{d}y = \delta_{nm} - \frac{1}{h} \int_0^d \psi_n(y)\psi_m(y)\mathrm{d}y.$$

Truncating the series for n and m, the system given by relation (35) can be solved numerically for $N + 1$ equations with $N + 1$ unknowns A_n.

The BVP-II which involves $\phi_0(x, y)$ the solution of BVP-I, represents the radiation problem. On applying Green's integral theorem to the functions $\phi_0(x, y)$ and $\phi_1(x, y)$ on the region bounded by

$$y = 0, 0 < x \le X; x = 0^+, 0 \le y \le d; x = 0^-, 0 \le y \le d; y = 0, -X \le x < 0;$$

$$x = -X, 0 \le y \le h; y = h, -X \le x \le X; x = X, 0 \le y \le h;$$

where X is positive, large, and tends to infinity, we obtain

$$R_1 = \frac{1}{2i\hat{k}_0} \int_{-\infty}^{\infty} c(x) \Big(\frac{\partial \phi_0(x, h)}{\partial x} \Big)^2 \mathrm{d}x. \tag{36}$$

Similarly, applying Green's integral theorem to the functions $\phi_0(-x, y)$ and $\phi_1(x, y)$ in the same region, we have

$$T_1 = -\frac{1}{2i\hat{k}_0} \int_{-\infty}^{\infty} c(x) \Big(\frac{\partial \phi_0(x, h)}{\partial x} \Big) \Big(\frac{\partial \phi_0(-x, h)}{\partial x} \Big) \mathrm{d}x. \tag{37}$$

These R_1 and T_1 can be evaluated when the shape function $c(x)$ is known.

3.1 Particular Cases

Case (i): In the absence of vertical barrier, the problem assumed here will be the problem involving scattering of water waves by bottom undulation only, and in this

case, the first order reflection and transmission coefficients given by relations (36) and (37) are the same as the relations (31) and (32) of [10] and relations (3.5) and (3.6) of [5] when the surface tension is negligible and angle of incidence is zero.

Case (ii): In the absence of undulation at the bottom, the given problem reduces to the scattering of water waves by vertical barrier. In this situation, the results involving reflection and transmission coefficients exactly match with the results of [4, 7].
In the next section, we consider the special form of the shape function $c(x)$ to evaluate the coefficients given by relations (36) and (37).

4 Example of Bottom Undulation

Different examples for the shape functions as considered in [6, 9] can be taken here to evaluate the reflection and transmission coefficients given by relations (36) and (37). However, the shape function function $c(x)$ is considered in the form of a patch of sinusoidal ripples because the functional forms of the uneven bottom closely resemble some naturally occurring obstacles formed at bottom due to ripple growth of sands and sedimentation as reported by [1].
The patch of sinusoidal bottom ripples can be expressed as

$$c(x) = \begin{cases} c_0 \sin \lambda x, & \dfrac{-M\pi}{\lambda} \le x \le \dfrac{M\pi}{\lambda} \\ 0, & \text{otherwise} \end{cases} \tag{38}$$

where c_0 is amplitude and λ is wave number, of the sinusoidal undulation and M is a positive integer. For M sinusoidal ripples, T_1 vanishes and R_1 is given by

$$\begin{aligned} R_1 = {} & \frac{c_0\hat{k_0}(R_0-1)}{2N_0^2}\left\{\frac{\sin(\lambda-2\hat{k_0})l}{\lambda-2\hat{k_0}} - \frac{\sin(\lambda+2\hat{k_0})l}{\lambda+2\hat{k_0}}\right\} \\ & + \frac{ic_0\hat{k_0}R_0}{2N_0^2}\left\{\frac{2(1-\cos\lambda l)}{\lambda} - \frac{2\lambda}{\lambda^2-4\hat{k_0}^2} + \frac{\cos(\lambda-2\hat{k_0})l}{\lambda-2\hat{k_0}} - \frac{\cos(\lambda+2\hat{k_0})l}{\lambda+2\hat{k_0}}\right\} \\ & + \frac{ic_0}{N_0}\sum_{n=1}^{\infty}\Bigg[\frac{k_n}{(\lambda-\hat{k_0})^2+k_n^2} - \frac{k_n}{(\lambda+\hat{k_0})^2+k_n^2} \\ & + \Bigg\{\frac{(\lambda-\hat{k_0})\sin(\lambda-\hat{k_0})l - k_n\cos(\lambda-\hat{k_0})l}{(\lambda-\hat{k_0})^2+k_n^2} \\ & - \frac{(\lambda+\hat{k_0})\sin(\lambda+\hat{k_0})l - k_n\cos(\lambda+\hat{k_0})l}{(\lambda+\hat{k_0})^2+k_n^2}\Bigg\}e^{-k_nl}\Bigg]\frac{k_nA_n}{N_n}, \end{aligned} \tag{39}$$

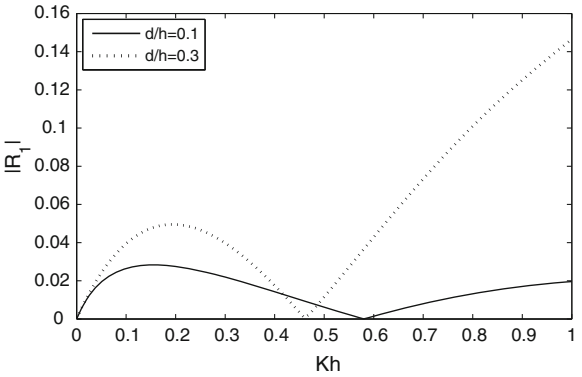


Fig. 2 $|R_1|$ for different water depths $d/h = 0.1, 0.3$ with $c_0/h = 0.1$, $M = 1$, $\lambda h = 1$

where $l = \frac{M\pi}{\lambda}$.

5 Numerical Evaluation and Discussions

The numerical computation is shown for the first order reflection coefficient R_1 given by Eq. (39). For computation of R_1, we need the values of R_0 and the constants A_n, $(n = 1, 2, \ldots, N)$ which are evaluated numerically from relation (35).

In Fig. 2, the reflection coefficient R_1 is plotted versus Kh for different water depths $d/h = 0.1, 0.3$ with $c_0/h = 0.1$, $M = 1$, $\lambda h = 1$. From this figure, it is observed that for fixed values of M, c_0/h and λh, the zeros of R_1 are shifted toward the left as the values of water depth d/h increases.

In Fig. 3, the first order reflection coefficient R_1 is depicted against Kh for different values of ripple amplitude $c_0/h = 0.1, 0.2$ with fixed values of $M = 1$, $d/h = 0.5$ and $\lambda h = 1$. From this graph, it is clear that the values of R_1 increase as the ripple amplitude c_0/h increases but the zeros of R_1 remain unchanged.

Different curves for R_1 against Kh are shown in Fig. 4 for different number of ripples $M = 2, 4, 6$ with fixed values of other parameters $d/h = 0.3$, $c_0/h = 0.1$ and $\lambda h = 1$. From this figure, it is found that the peak values of reflection coefficient R_1 are 0.1439, 0.3183, and 0.4583 corresponding to $M = 2, 4, 6$ respectively, i.e., it is clear that when the number of ripples M of undulation increases, the peak value of the reflection coefficient increases. It is also observed that the reflection coefficient becomes more oscillatory and the number of zeros increases with increasing number of ripples.

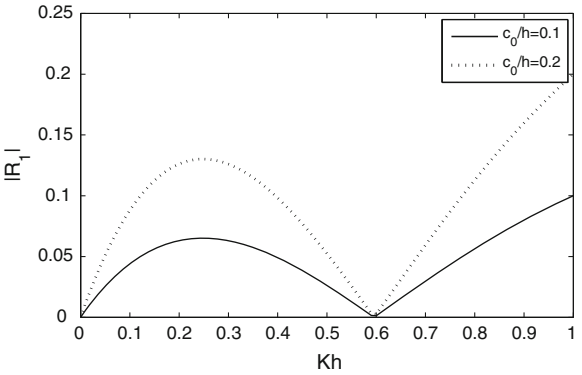


Fig. 3 $|R_1|$ for different ripple amplitude $c_0/h = 0.1, 0, 2.$ with $d/h = 0.5$, $M = 1$, $\lambda h = 1$

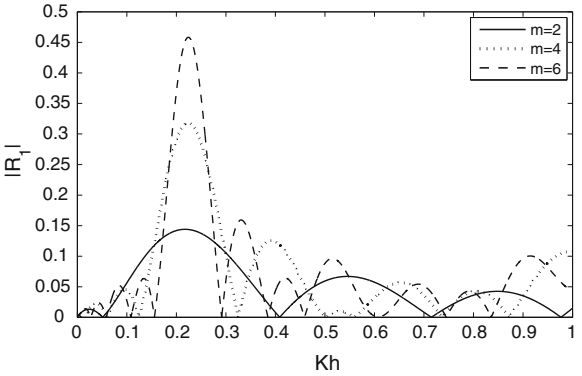


Fig. 4 $|R_1|$ for different number of ripples $M = 2, 4, 6$ with $d/h = 0.3$, $c_0/h = 0.1$, $\lambda h = 1$

6 Conclusion

Perturbation analysis is used to analyze the problem involving scattering of water waves by submerged thin vertical barrier over irregular bottom. The first order reflection and transmission coefficients are obtained in terms of integrals involving the shape function $c(x)$ describing the bottom undulation and the solution of the problem involving scattering by submerged barrier. A special case of bottom undulation as a patch of sinusoidal ripples is considered and the first order reflection coefficient is obtained and is shown in different figures.

References

1. Davies, A.G.: The reflection of wave energy by undulations on the seabed. Dyn. Atmos. Oceans **6**, 207–232 (1982)
2. Davies, A.G., Heathershaw, A.D.: Surface wave propagation over sinusoidally varying topography. J. Fluid Mech. **144**, 419–443 (1984)
3. Dean, W.R.: On the reflection of surface waves by a submerged plane barrier. Math. Proc. Camb. Phil. Soc. **41**, 231–238 (1945)
4. Losada, I.J., Losada, M.A., Roldan, A.J.: Propagation of oblique incident waves past rigid vertical thin barriers. Appl. Ocean Res. **14**, 191–199 (1992)
5. Mandal, B.N., Basu, U.: A note on oblique water-wave diffraction by a cylindrical deformation of the bottom in the presence of surface tension. Archive of Mech. **42**, 723–727 (1990)
6. Mandal, B.N., Basu, U.: Wave diffraction by a small elevation of the bottom of an ocean with an ice-cover. Archive of Appl. Mech. **73**, 812–822 (2004)
7. Mandal, B.N., Dolai, D.P.: Oblique water wave diffraction by thin vertical barriers in water of uniform finite depth. Appl. Ocean Res. **16**, 195–203 (1994)
8. Mandal, B.N., Gayen, R.: Water wave scattering by bottom undulations in the presence of a thin partially immersed barrier. Appl. Ocean Res. **28**, 113–119 (2006)
9. Martha, S.C., Bora, S.N.: Oblique surface wave propagation over a small undulation on the bottom of an ocean. Geophys. Astrophys. Fluid Dyn. **101**, 65–80 (2007)
10. Martha, S.C., Bora, S.N., Chakrabarti, A.: Interaction of surface water waves with small bottom undulation in sea-bed. J. Appl. Math. Inf. **27**, 1017–1031 (2009)
11. Miles, J.W.: Oblique surface-wave diffraction by a cylindrical obstacle. Dyn. Atmos. Oceans **6**, 121–123 (1981)
12. Porter, D.: The transmission of surface waves through a gap in a vertical berrier. Math. Proc. Cambridge Philos. Soc. **71**, 411–422 (1972)
13. Porter, R., Evans, D.V.: Complementary approximations to wave scattering by vertical barriers. J. Fluid Mech. **294**, 155–180 (1995)
14. Ursell, F.: The effect of a fixed barrier on surface waves in deep water. Math. Proc. Cambridge Philos. Soc. **43**, 374–382 (1947)

Chapter 2
Cryptanalysis of Multilanguage Encryption Techniques

Prasanna Raghaw Mishra, Indivar Gupta and Navneet Gaba

Abstract We present an analysis of an encryption scheme MUlti-Language Encryption Technique (MULET) proposed by G. Praveen Kumar et al. in the Seventh International Conference on Information Technology, ITNG 2010. Using our analysis, we have successfully recovered 80 % of the plaintext from the MULET ciphertext. We also give quantitative results in support of our findings.

1 Introduction

MUlti-Language Encryption Technique (MULET) proposed by Praveen Kumar et al. [11] is an encryption scheme designed to facilitate encryption/decryption for a range of languages supported by Unicode [14]. The authors have shown that the scheme is secure against brute-force attack only and have not discussed its security against cryptanalytic attacks. However, the scheme escaped the attention of cryptanalysts. Although Anoop Kumar et al. [1] indicated some of the flaws in the technique, no comprehensive cryptanalysis was presented. This motivated us to go for an in-depth cryptanalysis of the scheme. We have launched a ciphertext only attack [13]

The authors are grateful to Dr. P. K. Saxena, Director SAG for his support and encouragement. The authors would also like to express their sincere thanks to Dr. S. S. Bedi, Associate Director SAG for the valuable suggestions given by him during the course of the work.

P. R. Mishra (✉) · I. Gupta · N. Gaba
Scientific Analysis Group, Defence Research and Development Organisation,
Delhi 110054, India
e-mail: prasanna.r.mishra@gmail.com

I. Gupta
e-mail: indivar_gupta@yahoo.com; indivargupta@sag.drdo.in

N. Gaba
e-mail: navneetgaba2000@yahoo.com

R. N. Mohapatra et al. (eds.), *Mathematics and Computing 2013*,
Springer Proceedings in Mathematics & Statistics 91,
DOI: 10.1007/978-81-322-1952-1_2, © Springer India 2014

Table 1 Notations

M	Mapping constant/modulus
ch_map	A set of M-characters from the universal character set is considered as a mapping array
$chno$	A set of characters for universal character set is considered as a substitution array
Quo	Quotients required for decryption (key)
Enc	Ciphered text
Dec	Deciphered text

on MULET ciphertext and successfully recovered more than 80 % of the plaintext out of it. The organization of this paper is as follows.

In the Sect. 2, we describe MULET algorithm in brief. In Sect. 3, we describe our technique to recover plaintext out of MULET ciphertext. MULET is a double-layer encryption scheme and we have tried to remove layers in the reverse order to get back the plaintext. In Sect. 4, we give a step-by-step complexity analysis of the technique. In Sect. 5, we give the results of our attack applied on a ciphertext of English encrypted with MULET.

2 Description of MULET

In this section, we describe MULET encryption and decryption algorithm in brief [11]. Before describing the scheme, we first discuss the notations used in the scheme (see Table 1).

The Scheme

Key: *Secret Keys-* M, ch_{map}, $chno$, *Publicly Known-* Unicode

Encryption Algorithm
Input: Plaintext, Arrays ch_map and $cnno$, Modulus M
Output: Ciphertext, Array Quo

```
while (! End of plaintext) do
    Read a character from the original file and store the Unicode value in a variable n ;
    R := n%M
    Quo[i] := n/M
    Enc[i] := ch_map[R]
    Increment i ;
end while
while (! end of Enc) do
    while (Enc[i] == Enc[i + 1]) do
        Increment count;
        Increment i ;
    end while
    if (count>= 2) then
        Replace the repetitions with chno[count] in enc
        Reset count to zero
    end if
end while
```

Decryption Algorithm:
Input: Ciphertext, Array Quo
Output: Plaintext

while (! end of enc) **do**
  if (character is $chno[i]$) **then**
    Remove the character from enc and the character preceding $chno[i]$ in the cipher text is repeated 'i' number of times and store in dec
  end if
end while
while (!end of dec) **do**
  Compare the character with the mapping array ch_ map; Position of the character in ch_map is the required remainder R;
  $U := Quo[i] * M + R$;
  Convert U to the corresponding character;
end while

2.1 Example of MULET

Mapping Constant/Modulus $M = 16$
ch_map:

0	1	2	3	4	5	6	7
अ	आ	इ	ई	उ	ऊ	ऋ	ऌ
8	9	10	11	12	13	14	15
ऍ	ऐ	ए	ऐ	ऑ	ओ	ओ	औ

chno:

0	1	2	3	4	5	6	7	8	9
০	১	২	৩	৪	৫	৬	৭	৮	৯

Plaintext:

Cryptography is the science of secret writing

Key generated (in Hex):
724 707 6f7 726 706 796 692 207 206 637 656
636 206 666 732 636 657 207 727 746 6e6 a6

Ciphertext:

ईऍउॡआॅअईआईऍओऊऔअऊइउॡऍ२ॡ

3 Cryptanalysis

MULET encryption is done in two layers. The first layer makes use of secret data *ch_map* and the modulus M and the second layer uses secret data *chno*. We notice that because of the two layers of encryption, the existing attacks on classical ciphers [2, 13] are not applicable directly on MULET. We have devised a ciphertext only attack [13] on the scheme to recover various secret parameters and finally the plaintext. The only assumption we make is that the plaintext language is known. There are two main parameters whose knowledge leads to recovery of plaintext, viz., the modulus M and the useful portion of array *chno*. To start with, we first guess the modulus M.

3.1 Guessing the Modulus

A MULET ciphertext contains characters from both the arrays, viz. *ch_map* and *chno*. Let the number of distinct characters occurring in the ciphertext be d and the number of characters from *chno* occurring in the ciphertext be b_1. We observe that the index 0 and 1 of *chno* is never accessed. Similarly, an index higher than 5 corresponds to 6-graph or higher. For a reasonable size of modulus ($M \geq$ size of plaintext alphabet/2) occurrence of 6-graph or higher is extremely rare, therefore, index higher than 5 is rarely accessed. Thus, the wise choice of b_1 to start with is 4.

The maximum number of characters from *ch_map* that can occur in ciphertext is M. Thus, the bound on the number of distinct characters occurring in ciphertext is $M + b_1$. There is a possibility that all the 4 characters from *chno* may not be occurring in the ciphertext. Similarly, some characters from *ch_map* may also escape ciphertext. Let the maximum number of characters from the two arrays not occurring in the ciphertext be b_2. Now we have the following relation:

$$M - b_2 + b_1 \leq d \leq M + b_1$$

which implies that M can be found by trying b_2 values precisely $d - b_1, d - b_1 + 1, \ldots, d - b_1 + b_2$. The value of b_2 is relatively small (say ≤ 10, as it is less likely that more than 10 characters from ch_map are skipped) in most cases. Based on experimentation, the value of b_2 is taken as 6 in our case. As the value of d is determined from the ciphertext, modulus M can be guessed in much smaller number of trials.

3.2 *Segregating the* ch_map *and* chno *Characters in the Ciphertext*

Let us assume that the set of unicode values of the plaintext alphabets are $\mathcal{A} = \{v_i, i = 1, 2 \cdots, l\}$. and the expected frequency of $v \in \mathcal{A}$ in a meaningful text is f_v^e. The first layer of encryption process uses a function $\phi_M : \mathcal{A} \to \{0, 1, \cdots, M-1\}$ given as $\phi_M(v) = r, r \equiv v \pmod{M}, 0 \leq r < M$. Clearly, ϕ_M is many-one if $M < l$. $M \geq l$ makes the cipher merely a simple substitution [2, 8], and it can be analyzed using existing methods [2, 3, 5, 10].

We have only to consider the other case, i.e., $M < l$. The remainders are replaced by the corresponding *ch_map* array. Once the first layer encryption is over, an n consecutive occurrence of a character v is replaced by the couple of characters $ch\_map(r)chno(n)$. Here, we make an assumption that the second layer changes do not alter the relative frequency distribution of *ch_map* characters in the ciphertext. We find expected frequency distribution of ϕ_M after the first layer. The expected frequency R_r of a value r of ϕ_M can be given as $R_r = \sum_{v \in \phi_M^{-1}(r)} f_v^e$. Expected frequency of a *chno* character at index n is the same as frequency of n-ets (n consecutive occurrences of any character) in the first layer text. Let these frequencies lie in the interval $[0, b_3]$. We consider the set

$$S = \{c \text{ is a ciphertext character : frequency of } c \leq b_3\}$$

We select b_1 characters from S and arrange them in decreasing order of their frequencies. The highest frequent character corresponds to index 2, the next to 3, and so on. With this information, we remove the second layer changes and compute the frequency distribution of the changed ciphertext. We measure how close the resulting frequency distribution is to the expected one. To measure the closeness we use a metric similar as ℓ_1 metric. The distance between the guessed and the expected frequency distributions d with respect to our metric is given as $\frac{\sum_{i=1}^{M} |R_{r_i} - f_{c_i}^e|}{M}$ (meanings of the symbols used are described in the next section). We carry out trials on all possible values of b_1. There will be $\binom{o(S)}{b_1}$ trials for a given value of b_1. The selection giving the minimum distance will reveal the part of *chno* used in the second layer (it should be noted that minimum is calculated over all possible values of b_1 and M).

3.3 *Making Final Substitution*

We assume that upto this stage we have successfully undone the second layer changes. We rewrite the expected frequency distribution of ϕ_M as $(r_1, R_{r_1}), (r_2, R_{r_2}), \ldots, (r_M, R_{r_M})$ such that $1 \leq i < j \leq M \implies R_{r_i} \geq R_{r_j}$ (where $\{r_1, r_2, \ldots, r_M\} = \{0, 1, \ldots, k-1\}$). Let the frequency of ciphertext character c be f_c^o. The frequency distribution is $(c_1, f_{c_1}^o), (c_2, f_{c_2}^o), \cdots, (c_M, f_{c_M}^o)$ such that $1 \leq i < j \leq M \implies$

$f^o_{c_i} \geq f^o_{c_j}$ (where $c_1, c_2, \ldots, c_M$ are distinct letters in the intermediate ciphertext). From here we guess the map $u : \{c_1, c_2, \ldots, c_M\} \rightarrow \{0, 1, \ldots, M-1\}$ which establishes the relation between remainders and ciphertext characters as $u(c_i) = r_i,\ i = 1, 2, \ldots, M$. We finally replace the ciphertext character c by v_c^{mp} where v_c^{mp} is chosen such that $f^e_{v_c^{mp}} = \max_{v \in \phi_M^{-1}(u(c))}\{f^e_v\}$. v_c^{mp} is the most probable replacement for the ciphertext character c.

4 Complexity Analysis

The first step of attack, i.e., guessing the modulus requires at most d trials. While the removal of the second layer encryption requires $\binom{o(S)}{b_1}$ calculations of frequency distributions and distance calculations. For a ciphertext of length L, the total number of operations required to find frequency distribution will roughly take $2L$ operations. The Euclidean distance calculation will take at most $M \log_2 L$ operations. Therefore, the total number of trials are bounded above by the expression $\sum_{i=1}^{b_1} \sum_{M=1}^{d} \binom{o(S)}{i} \left(2L + M \log_2 L\right) \leq L \left(\sum_{i=1}^{b_1} \sum_{M=1}^{d} \binom{o(S)}{i} (M+2)\right)$. This shows that our attack is linear in the size of ciphertext.

5 Experimental Results

To verify our strategy, we encrypted an English text of 1,000 characters with MULET. The two arrays and the modulus we took were the same as taken in example 1 in [11]. After making trials on M and removing the first layer, we found that the modulus M is 16. We calculate the frequency of characters of the ciphertext. Table 2 gives the ciphertext characters and their percentage occurrence in the ciphertext in descending order.

We calculated the remainder r for each of the 52 English unicode characters. The characters giving the same remainder were grouped together. Table 3 lists the remainder r, its expected frequency (R_r) corresponding English character (v), and the most probable among them (v^{mp}) in descending order of R_r.

A comparison of Tables 2 and 3 suggests the possible mapping of ciphertext characters and remainders. We replace a ciphertext character with the most probable character corresponding to the possible remainder. In Table 4 we show the ciphertext characters, their possible remainders, and the possible replacement (the most probable character corresponding to the possible remainder).

Carrying out these replacements gives us the recovered plaintext. A comparison of the recovered plaintext from the original one reveals that the attack successfully recovers 83.36 % of the plaintext in our case. The point to note is that this recovered text may further be fed to text mining technique including pattern recognition and dictionary-based techniques [4, 6, 7, 9, 12] to further enhance the success rate.

Table 2 Frequency distribution of ciphertext character

S.N.	Ciphertext character	Percentage occurrence
1	0x090a	15.729
2	0x0909	13.022
3	0x090e	9.635
4	0x0908	9.461
5	0x0914	7.909
6	0x0906	7.798
7	0x0907	7.001
8	0x0913	6.833
9	0x090d	5.209
10	0x090c	4.456
11	0x0911	3.917
12	0x090b	3.276
13	0x0912	2.384
14	0x0905	2.291
15	0x0910	0.870
16	0x090f	0.207

Table 3 Pre-computed table for determination of most probable character corresponding to a given remainder

S.N.	Remainder (r)	Expected frequency of $r(R_r)$	Possible plaintext characters (v)	Most probable character (v^{mp})
1	5	15.6289	E U e u	e
2	4	13.1218	D T d t	t
3	9	9.6341	I Y i y	i
4	3	9.4623	C S c s	s
5	15	7.9191	O o	o
6	1	7.7885	A Q a q	a
7	2	7.0108	B R b r	r
8	14	6.8234	N n	n
9	8	5.2195	H X h x	h
10	7	4.4463	G W g w	g
11	12	3.9274	L l	l
12	6	3.2665	F V f v	f
13	13	2.3832	M m	m
14	0	2.2916	P p	p
15	11	0.8688	K k	k
16	10	0.2084	J Z j z	j

Table 4 Possible plaintext character corresponding to a ciphertext character

S.N.	Ciphertext character	Remainder	Possible replacement
1	0x090a	5	e
2	0x0909	4	t
3	0x090e	9	i
4	0x0908	3	s
5	0x0914	15	o
6	0x0906	1	a
7	0x0907	2	r
8	0x0913	14	n
9	0x090d	8	h
10	0x090c	7	g
11	0x0911	12	l
12	0x090b	6	f
13	0x0912	13	m
14	0x0905	0	p
15	0x0910	11	k
16	0x090f	10	j

The bounds we set were $b_1 = 4, b_2 = 6, b_3 = 5$ and $b_4 = 10$. For our case, the plaintext language was English. For $b_1 = 4, b_3 = 5$ we have $o(S) \leq 11$ (see Table 2). So, the worst case complexity will be $L\left(\sum_{i=1}^{b_1}\sum_{M=1}^{20}\binom{11}{i}(M+2)\right) = L\left(\sum_{i=1}^{4}\binom{11}{i}\sum_{M=1}^{20}(M+2)\right) = L \times 550 \times 250 = 137500L$. For $L = 1000$ the complexity is of order 2^{27}. It is to be noted that the calculations are made for the worst case and we have finished our attack in a much lesser time. Below are given the parts of plaintext, ciphertext, and the recovered.

Plaintext	:	P a g e o f Y O . . .
Unicode value	:	0x50, 0x61,0x67, 0x65, 0x6f, 0x66, 0x59, 0x4f, . . .
MULET encryption (Unicode value)	:	0x0905, 0x0906,0x090c, 0x090a, 0x0914, 0x090b, 0x090e, 0x0914, . . .
Recovered text	:	p a g e o f i o . . .

6 Conclusion

In this paper, we have presented a cryptanalysis of an encryption scheme named MUlti-Language Encryption Technique (MULET). Based on our analysis, we have launched a ciphertext-only attack and retrieved more than 80 % of the plaintext from

MULET ciphertext. We have also shown that the scheme can be broken in linear time, contrary to the claim of exponential complexity by the proposers of the scheme. There are many more schemes proposed on this philosophy. We hope that this analysis will be helpful to demonstrate the weaknesses of such schemes.

References

1. Anoop Kumar, S., Sanjeev S., Santosh, S.: MSMET: a modified and secure multilanguage encryption technique. Int. J. Comput. Sci. Eng. **4**(03), 402–405 (2012)
2. Bauer, F.L.: Decrypted Secret: Methods and Maxims of cryptology, 4th edn. Springer, Germany (2006)
3. Churchhouse, R.: Codes and Ciphers, Julius Caeser, the Enigma and the Internet. Cambridge University Press, Cambridge (2002)
4. Elder, J., Miner, G., Nisbet, B.: Practical Text Mining and Statistical Analysis for Non-structure Text Data Applications. Academic Press, Burlington (2012)
5. Harris, F.A., Tuck, S.: Solving Simple Substitution Ciphers. American Cryptogram Association, New York (1959)
6. Jakobsen, T.: A fast method for cryptanalysis of substitution ciphers. Cryptologia **19**(3), 265–274 (1995)
7. Lucks, M.: A constraint satisfaction algorithm for the automated decryption of simple substitution ciphers. In: Advance in Cryptology, CRYPTO'1988, LNCS, vol. 1462, pp. 132–144. Springer, Berlin (1998)
8. Menezes, A.J., van Oorschot, P.C., Vanstone, S.A.: Handbook of Applied Cryptography. CRC Press, New York (1997)
9. Olson, E.: Robust dictionary attack of short simple substitution ciphers. Cryptologia **31**(4), 332–342 (2007)
10. Peleg, S., Rosenfeld, A.: Breaking substitution ciphers using a relaxation algorithm. Commn. ACM **22**(11), 598–605 (1973)
11. Praveen Kumar, G., Arjun Kumar M., Parajuli, B., Choudhury, P.: MULET: a multilanguage encryption technique. In: Proceedings of Seventh International Conference on Information Technology 2010. IEEE Computer Society (2010)
12. Spillman, R., Janssen, M., Nelson, B., Kepner, M.: Use of a genetic algorithm in the cryptanalysis of simple substitution ciphers. Cryptologia. **XVII**(3), 31–44 (1993)
13. Stamp, M., Low, R.M.: Applied Cryptanalysis: Breaking Ciphers in the Real World. John Wiley & Sons, Hoboken (2007)
14. Unicode Character form. http://www.Unicode.org

Chapter 3
Signcryption with Delayed Identification

Angsuman Das and Avishek Adhikari

Abstract This paper introduces a novel cryptographic primitive called Signcryption with Delayed Identification (SCDI), where a sender signcrypts a message m such that the reciever can unsigncrypt it using his private key to recover m, but cannot get any information about the identity of the sender. The sender at a later point of time can claim the ownership of the message m by providing a "tag", which proves that the signcryptext was generated by the sender. As an application of the primitive, it is shown that it can be used for safe and anonymous contractual bidding, submission of papers in a journal or conference, etc. As regards security, formal definitions of security for the proposed primitive are given and at the end, a generic construction secure with respect to the proposed definitions is given.

Keywords Partial signature · Signcryption · Random oracle model

1 Introduction

Signcryption, introduced by Zheng [17] and formalized in [1, 2], has been an area of active research from the day of its inception. Signcryption is a primitive which encrypts as well as authenticates a message. The main objective in the study of

The authors thank Sumit Kumar Pandey of C. R. Rao Institute, India, and Partha Sarathi Roy and Sabyasachi Dutta of University of Calcutta, India for several fruitful discussions during the work.

A. Das (✉)
Department of Mathematics, St. Xavier's College, Kolkata 700016, India
e-mail: angsumandas054@gmail.com

A. Adhikari
Department of Pure Mathematics, University of Calcutta, Kolkata 700019, India
e-mail: avishek.adh@gmail.com

R. N. Mohapatra et al. (eds.), *Mathematics and Computing 2013*,
Springer Proceedings in Mathematics & Statistics 91,
DOI: 10.1007/978-81-322-1952-1_3,

signcryption scheme was twofold: to reduce the cost of signcryption than naive combination of encryption and signature and to achieve better security than component encryption and signature scheme (SS). Since then, depending on their applicability in various requirements, various signcryption schemes along with several variants like identity-based signcryption [13], ring signcryption, proxy signcryption, aggregate signcryption [8], threshold signcryption [12], heterogeneous signcryption[11], etc., and their corresponding security notions have evolved.

In this paper, we propose a new signcryption variant, called Signcryption with Delayed Identification (SCDI), which provides a layer of anonymity that can be revealed later. Consider the following scenario: a sender Alice signcrypts a message m such that the receiver Bob can unsigncrypt it using his private key to recover m and a "stub", but cannot get any information about the identity of the sender. Alice at a later point of time (not necessarily predetermined) can claim that the message m was generated by her by providing a "tag" corresponding to the "stub", which proves or convinces Bob that the signcryptext was actually sent by Alice and no one else.

We give some potential applications of this primitive SCDI as follows:

- **Anonymous Contractual Bidding**: Suppose the government of a country decides to call for tenders to construct a bridge. Various construction companies will be applying for it. Now, while selecting from their proposals, government officials not only look at the lowest quoted price but also at the quality of the proposal, e.g., quality of materials used, number of days needed to complete the task, etc. Now suppose Alice, one such bidder, does not have good relations with the government official, Bob, dealing with this tender. Hence she might fear that her proposal would be rejected regardless of its quality. Thus, she wants to signcrypt her bid in such a way that it will not reveal her identity, but if her proposal is accepted, then she can claim it to be hers. SCDI can be used in this scenario. In fact, SCDI is handy in any anonymous bidding or contractual agreement where "highest/lowest value wins" is not the case.
- **Anonymous Paper/Patent Submission**: Suppose Bob is the editor of a journal and he invites papers for a particular issue. Alice wants to submit a paper to the journal, but Bob has certain professional or personal conflicts with Alice. So, Alice wants to hide her identity from Bob until the selection procedure is over. Once the selection is done, Bob puts up the titles of the selected papers on the journal webpage. Now Alice would like to claim her ownership on her paper, only if it is accepted. If this can be done, then at least impartiality by Bob with respect to the identity of Alice is maintained.[1]
- **Disclosure of Secret and Sensitive Information**: Suppose the police department of a country has declared a prize money for anyone who gives proper information about the whereabouts of a notorious gangster X. Alice knows the hiding place of X and wants to inform the police about it to claim the prize. But Alice suspects that even some of the policemen might be involved with X. As a result, her life can be in danger if she delivers the information to such corrupted policemen. In such

[1] If the content of the paper itself leaks any direct or indirect information about Alice, then nothing can be done, as the editor cannot be made blind to the content of the paper.

a scenario, Alice can use SCDI to send the information such that the police does not know anything about her identity. But when X is arrested, she can identify herself to be the sender of the information. Although this scenario looks more like a movie plot, SCDI turns out to be a handy tool.

The need for SCDI (at least in the first two scenarios) rises from the fact that no absolute winning condition can be specified. Thus, the public verification of the fairness of the result is not possible. As a result, there is always a chance of unfairness on the basis of identity.

1.1 Related Work

The idea of revealing the identity of the sender Alice at a later point of time (to be determined by the sender, but not necessarily predetermined) was first noticed in [4]. The authors in [4] introduced a new signature primitive called *Partial Signature*, where Alice, given a message, can compute a "stub" which preserves her anonymity, yet later she, but nobody else, can complete the stub to a full and verifiable signature under her public key.

However, what is of note here is that in the above scenarios confidentiality is needed along with anonymity and nonrepudiation. For example, Alice would like to hide her bid or price quotation from other potential bidders. In fact, as we will show later (see Sect. 1.2) a partial signature scheme alone cannot ensure unbiasedness or safety, at least with respect to the identity of Alice. Our proposal, SCDI, can serve these specified goals.

1.2 Security Requirements

We now discuss informally the security requirements for SCDI, namely confidentiality, anonymity, unambiguity, unforgeability, and unlinkability.

- **Confidentiality**: None other than the receiver, Bob, should be able to get any information about the message m signcrypted within the ciphertext c. This is required as the sender, Alice, may want no one other than Bob to know anything about the content m of the submission. Though in most cases it may be enough to have IND-CPA security for confidentiality, there might be some scenarios where IND-CCA2 security is in demand.
- **Anonymity**: This means that the receiver, Bob cannot get any information about Alice's identity from the *stub* in the initial phase. To elaborate, Bob may have some prior information about the potential identity of the sender, say for example it belongs to a set S, which at worse can have only two elements. The protocol should be such that given the *stub*, the message m and the knowledge that is created

by either of the two members in S will give Bob only a negligible advantage in guessing the right identity.

- **Unambiguity**: Unambiguity demands that none other than Alice (including Bob) can generate a *tag* which matches with the ciphertext generated by Alice's secret key to claim authorship of the ciphertext. It requires that an adversary $\mathcal{A}$, given a ciphertext under the secret key of Alice, sk_A of a message m of his choice, is unable to create a *tag* (may be under a different verification key pk' other than pk_A) such that it verifies as a "valid" one. In fact, it will be better if this can be done publicly, i.e., once Alice reveals the message m, the *stub* and the *tag*, anyone (not only Bob) can verify the authorship of Alice on m. Unambiguity ensures that anonymity is not at the cost of authenticity.
- **Unforgeability**: Unforgeability claims that an adversary $\mathcal{A}$ cannot produce a valid ciphertext-tag pair, i.e., (c, tag) on a message m of his choice under the secret key sk_A of Alice. It should be noted here that unambiguity does not imply unforgeability. Unambiguity prevents forgery under an adversarially modified verification key pk' ($\neq pk_A$), whereas unforgeability prevents forgery on the target verification key pk_A itself.
- **Un/Linkability**: This feature, unlike others, is specific to the contractual bidding/paper submission protocol. Let us discuss it in detail: Suppose Alice submits a ciphertext to Bob in the initial submission phase. Bob, due to some technical reasons, may ask Alice to submit a slightly modified version of the ciphertext before the review is done. Even Alice herself may want to modify the submission before the deadline is over. In any case, as the submission is anonymous (even to Bob), how can Bob be sure about the fact that the revised submission is made by the actual author, i.e., Alice and not someone else (other than Bob).[2] Thus, Bob must be able to link the two submissions together, i.e., he must be convinced that both the submissions are done by the same person, Alice. In other words, no one other than Alice should be able to replace/revise her original submission.

Remark 1 It is worth mentioning here that unambiguity prevents an adversary from claiming the authorship of Alice's ciphertext once the selection has been made, whereas unforgeability prevents an adversary from submitting a ciphertext pretending to be Alice, such that Alice might face problems in the future, like plagiarism, etc. In other words, unforgeability is required in the submission phase, whereas unambiguity is required in the revealing stage.

Remark 2 Anonymous encryption [3] does not provide the required anonymity in our case. Bellare et al. [3] deals with receiver anonymity, but here sender anonymity is required.

Remark 3 It is to be noted that though the criteria of anonymity, unambiguity, and unforgeability were also discussed in [4], the issues of confidentiality and linkability were not discussed. However as explained above, they are two vital components for

[2] Here, we assume that Bob is not the adversary. The rationale behind this assumption is that as Bob is unaware of the identity of Alice, why would he try to replace Alice's original submission.

an SCDI and hence partial signature cannot suffice the need. In fact our definitions of anonymity, unambiguity, and unforgeability are not exactly the same as that in [4]. In some cases, our definitions demand weaker security guarantees, whereas in other cases ours is stronger than that of [4] depending on the application.

Remark 4 Linkability as a feature can also be omitted in certain cases. This can be done if we agree that, if the receiver/sender finds any need for revision or modification before the selection is done, he can simply ask for a fresh proposal and cancel the previous one. However, we insist on linkability, keeping in mind that most of the paper-submission softwares allow revised submission before the review is done or before the submission deadline is over.

1.2.1 Organization of the Paper

The rest of the paper is organized as follows: In Sect. 2, some definitions on basic cryptographic primitives and their security notions are discussed. The main construction is given in Sect. 3 and its security analysis is done in Sect. 4. Finally, we conclude with some open issues in Sect. 5.

2 Preliminaries and Definitions

We begin by formally defining the notions of *Randomness-Extractable Public-Key Encryption* (RE-PKE), Signature Scheme (SS), Commitment Scheme (CS), and *Signcryption with Delayed Identification* (SCDI) and their corresponding security notions.

2.1 Randomness-Extractable Public-Key Encryption

A Randomness-Extractable Public-Key Encryption (RE-PKE) is a tuple of probabilistic polynomial-time (ppt.) algorithms (Gen, Enc, Dec) such that:

1. The key generation algorithm, Gen, takes as input a security parameter 1^n and outputs a public-key/ private-key pair (pk, sk).
2. The encryption algorithm Enc takes as input a public key pk, a message m from the underlying plaintext space, and an ephemeral key r from the randomness space to output a ciphertext $c := \mathsf{Enc}(pk, m, r)$.
3. The decryption algorithm Dec takes as input a private key sk and a ciphertext c to output a plaintext m and an ephemeral key r.

It is required that there exists a negligible function negl such that for every n, every (pk, sk) and every message m in the corresponding plaintext space, it holds that $\mathsf{Pr}[\mathsf{Dec}(sk, \mathsf{Enc}(pk, m, r)) \neq (m, r)] \leq \mathsf{negl}(n)$.

Remark 5 Paillier encryption scheme [15], OAEP [5] and its variants are some of the existing examples of IND-CPA-secure RE-PKE, whereas constructions like [7, 9, 10, 14, 16] are examples of IND-CCA2-secure RE-PKE.

Remark 6 If we supress the decryption algorithm Dec to return only the plaintext m in the RE-PKE, we get a usual public-key encryption scheme.

2.1.1 Security Notions for Public-Key Encryption

Although there are various notions of security for public-key encryption scheme, only the relevant (IND-CPA and NM-CPA) ones are discussed here.

Indistinguishability against Chosen Plaintext Attack: Indistinguishability against chosen plaintext attack to a cryptosystem is defined as a game played between a challenger $\mathcal{C}$ and an adversary $\mathcal{A}$ in a public-key encryption scheme PKE as follows:

1. Given the security parameter, $\mathcal{C}$ generates a pair (pk, sk).
2. $\mathcal{A}$ is given the public-key pk. $\mathcal{A}$ outputs a pair of messages (m_0, m_1) from the plaintext space associated with pk with $|m_0| = |m_1|$.
3. $\mathcal{C}$ chooses $b \in_R \{0, 1\}$ and sends the ciphertext $c^* = \mathsf{Enc}_{pk}(m_b)$ to $\mathcal{A}$;
4. $\mathcal{A}$ finally outputs a bit b'.

The advantage $\mathbf{Adv}^{\text{cpa}}_{\mathcal{A},\text{PKE}}(n)$ is defined as $|Pr[b' = b] - 1/2|$. The scheme PKE is said to be secure against chosen plaintext attack if for all probabilistic polynomial-time adversaries $\mathcal{A}$, the advantage $\mathbf{Adv}^{\text{cpa}}_{\mathcal{A},\text{PKE}}(\cdot)$ is negligible.

Nonmalleability against Chosen Plaintext Attack: There have been so far various equivalent definitions of nonmalleability in the literature. In this work, we choose to work with a simplified version of relation-based definition of nonmalleability as follows: Nonmalleability against chosen plaintext attack to a public-key cryptosystem is defined as a game played between a challenger $\mathcal{C}$ and an adversary $\mathcal{A}$ in a public-key encryption scheme PKE as follows:

1. Given the security parameter, $\mathcal{C}$ generates a pair (pk, sk).
2. $\mathcal{A}$ is given the public-key pk.
3. $\mathcal{C}$ chooses a message m randomly from the plaintext space associated with pk and sends the ciphertext $c = \mathsf{Enc}_{pk}(m)$ to $\mathcal{A}$;
4. $\mathcal{A}$ finally outputs a ciphertext c' and a polynomial-time checkable relation ρ on the message space $\mathcal{M}$.

$\mathcal{A}$ wins the game if $c \neq c'$, $\mathsf{Dec}(sk, c') \neq \perp$ and $\rho(m, \mathsf{Dec}(sk, c'))$ holds.

If we wish to analyze a scheme PKE in the random oracle model [6], the hash functions are replaced by random oracle queries as appropriate, and both $\mathcal{C}$ and $\mathcal{A}$ are given access to the random oracle in the above attack game.

2.2 Signature Scheme

A Signature Scheme (SS) is a tuple of ppt. algorithms (Gen, Sign, Ver) such that

1. The key generation algorithm, Gen, takes as input a security parameter 1^n and outputs a signing-key/ verification-key pair (sk, pk).
2. The signing algorithm Sign takes as input a signing-key sk, a message m from the underlying plaintext space to output a signature $s := \mathsf{Sign}(sk, m)$.
3. The verification algorithm Ver takes as input a verification-key pk and a message-signature pair (m, s) to output 0 or 1.

It is required that there exists a negligible function negl such that for every n, every (pk, sk) and every message m in the corresponding plaintext space, it holds that $\mathsf{Pr}[\mathsf{Ver}(pk, m, \mathsf{Sign}(sk, m)) \neq 1] \leq \mathsf{negl}(n)$.

2.2.1 Security Notions for Signature Scheme

A Signature Scheme $\mathcal{SS} = (\mathsf{Gen}, \mathsf{Sign}, \mathsf{Ver})$ is said to achieve *existential unforgeability against chosen message attack* (UF-CMA) if any probabilistic polynomial-time adversary $\mathcal{A}$ has negligible chance of winning against a challenger $\mathcal{C}$ in the following game:

1. Given the security parameter, $\mathcal{C}$ generates a key pair (pk, sk) and returns pk to $\mathcal{A}$.
2. $\mathcal{A}$ is given oracle access to the signing oracle.
3. $\mathcal{A}$ outputs a message-signature pair (m^*, s^*).

$\mathcal{A}$ wins the game if s^* is a valid signature on m^* and if m^* was never queried to the signing oracle.

2.3 Commitment Scheme

A Commitment Scheme is a tuple of ppt. algorithms (CMT, DCMT) such that

1. The committing algorithm, CMT, takes as input a security parameter 1^n and a string $s \in \{0, 1\}^*$ and outputs a commitment–decommitment pair (c, d).
2. The decommiting algorithm DCMT takes as input a string s and the commitment–decommitment pair (c, d) to output 0 or 1.

It is required that there exists a negligible function negl such that for every n and every string s, it holds that $\mathsf{Pr}[\mathsf{DCMT}(s, \mathsf{CMT}(s)) \neq 1] \leq \mathsf{negl}(n)$.

2.3.1 Security Notions for Commitment Scheme

Hiding: A Commitment Scheme $\mathcal{CS} = (\mathsf{CMT}, \mathsf{DCMT})$ is said to be hiding if any probabilistic polynomial-time adversary $\mathcal{A}$ has negligible advantage against a challenger $\mathcal{C}$ in the following game:

1. Given the security parameter 1^n, $\mathcal{A}$ ouputs two strings s_0, s_1 of the same length with $s_0 \neq s_1$.
2. $\mathcal{C}$ chooses a bit $b \in_R \{0, 1\}$ and computes $(c, d) := \mathsf{CMT}(s_b)$. $\mathcal{C}$ outputs c.
3. $\mathcal{A}$ outputs a guess b' for b.

The advantage $\mathbf{Adv}^{\text{hiding}}_{\mathcal{A},\text{CS}}(n)$ is defined to be $|Pr[b' = b] - 1/2|$.

Binding: A Commitment Scheme (CS) is said to be binding if any probabilistic polynomial-time adversary $\mathcal{A}$ has negligible advantage in the following game: Given the security parameter 1^n, $\mathcal{A}$ ouputs two strings s_0, s_1 of the same length with $s_0 \neq s_1$ and two commitment-decommitment pair (c, d_0) and (c, d_1). The advantage $\mathbf{Adv}^{\text{binding}}_{\mathcal{A},\text{CS}}(n)$ is defined as

$$Pr[\mathsf{DCMT}(s_0, c, d_0) = 1 \text{ and } \mathsf{DCMT}(s_1, c, d_1) = 1].$$

2.4 Signcryption with Delayed Identification

Signcryption with Delayed Identification (SCDI) consists of six-tuple of ppt. algorithms (Setup, $\mathsf{Keygen_A}$, $\mathsf{Keygen_B}$, $\mathsf{Signcrypt}$, $\mathsf{Decrypt}$, Verify) such that

1. The setup algorithm, Setup, takes as input a security parameter 1^n and returns common parameter *par* required by the SCDI scheme.
2. The key generation algorithm for the sender A, $\mathsf{Keygen_A}$, takes as input the common parameters *par* and outputs a public-key/ private-key pair (pk_A, sk_A).
3. The key generation algorithm for the receiver B, $\mathsf{Keygen_B}$, takes as input the common parameters *par* and outputs a public-key/ private-key pair (pk_B, sk_B).
4. The signcryption algorithm $\mathsf{Signcrypt}$ takes as input common parameters *par*, sender secret key sk_A, receiver public key pk_B, and a message m to output a signcryptext $c := \mathsf{Signcrypt}(par, sk_A, pk_B, m)$ and a tag τ, corresponding to the message m from sender A.
5. The decryption algorithm $\mathsf{Decrypt}$ takes as input common parameter *par*, receiver secret key sk_B, a signcryptext c to output a message-stub pair $(m, \sigma) := \mathsf{Decrypt}(par, sk_B, c)$ or an error symbol $\perp$.
6. The verification algorithm, Verify, takes as input common parameter *par*, a message m, a stub σ and a tag τ to output 1 or 0, i.e., $\mathsf{Verify}(par, m, \sigma, \tau) = 1$ or 0.

In addition to these six algorithms, the SCDI may have two more algorithms $\mathsf{ReSigncrypt}$ and Link to enable submission of revised/modified version of a previously submitted ciphertext.

1. The re-submission algorithm ReSigncrypt takes as input common parameter par, a previously submitted signcryptext c_1, the sender secret key sk_A, the receiver public-key pk_B, and the revised submission m_2 to output another signcryptext c_2. *In some cases,* ReSigncrypt *may additionally take as input the internal random coins used while generating* c_1 *using* Signcrypt.
2. The linking algorithm Link, takes as input common parameters par, two signcryptexts c_1, c_2 and receiver secret key sk_B to output 1 or 0, i.e., $\mathsf{Link}(par, sk_B, c_1, c_2) = 1$ or 0.

Correctness:
It is required that for every n, every (pk_A, sk_A), (pk_B, sk_B), every message m in the corresponding plaintext space, it holds that

$$\mathsf{Decrypt}(sk_B, (\mathsf{Signcrypt}(sk_A, pk_B, m)) = (m, \sigma),$$

$$\mathsf{Verify}(\mathsf{Decrypt}(sk_B, \mathsf{Signcrypt}(sk_A, pk_B, m)), \tau) = 1 \text{ and}$$

$$\mathsf{Link}(sk_B, c_1, \mathsf{ReSigncrypt}(pk_B, sk_A, c_1)) = 1.$$

2.5 Security Notions for SCDI

2.5.1 Confidentiality

A Signcryption Scheme with Delayed Identification (SCDI) is said to achieve dynamic multi-user insider confidentiality in IND-SCDI-CCA2 sense if any probabilistic polynomial-time adversary $\mathcal{A}$ has negligible advantage against a challenger $\mathcal{C}$ in the following game:

1. Given the security parameter, $\mathcal{C}$ generates common parameter par and then with that generates a receiver key-pair (pk_B, sk_B) using $\mathsf{KeyGen_B}$.
2. $\mathcal{A}$ is given par, pk_B as well as oracle access to the decryption algorithm, $\mathsf{Decrypt}(par, sk_B, \cdot)$. $\mathcal{A}$ outputs a sender key-pair (pk_A, sk_A) and a pair of messages (m_0, m_1) from the associated plaintext space with $|m_0| \neq |m_1|$.
3. $\mathcal{C}$ chooses $b \in_R \{0, 1\}$, computes and sends the challenge signcryptext $c^* = \mathsf{Signcrypt}(par, sk_A, pk_B, m_b)$ to $\mathcal{A}$;
4. $\mathcal{A}$ continues to have oracle access to $\mathsf{Decrypt}(par, sk_B, \cdot)$ but with the restriction that it cannot query c^*; $\mathcal{A}$ outputs a bit b'.

The advantage $\mathbf{Adv}^{cca2}_{\mathcal{A},SCDI}(n)$ is defined as $|Pr[b' = b] - 1/2|$.

This game is defined analogous to that of dynamic multi-user insider security for confidentiality in a regular signcryption scheme. Although for our purpose outsider security is enough (as Alice knows her submission), we propose a stronger security requirement as our construction achieves it without much overhead.

2.5.2 Anonymity

A Signcryption Scheme with Delayed Identification (SCDI) is said to achieve anonymity if any probabilistic polynomial-time adversary $\mathcal{A}$ has negligible chance of winning against a challenger $\mathcal{C}$ in the following game:

1. Given the security parameter, $\mathcal{C}$ generates common parameter par and gives it to $\mathcal{A}$.
2. $\mathcal{A}$ with the help of par, outputs two sender-key pairs (pk_{A_0}, sk_{A_0}), (pk_{A_1}, sk_{A_1}), a receiver key-pair (pk_B, sk_B) and a message m.
3. $\mathcal{C}$ chooses $b \in_R \{0, 1\}$, computes and outputs $c^* = \mathsf{Signcrypt}(par, sk_{A_b}, pk_B, m)$ to $\mathcal{A}$.
4. $\mathcal{A}$ outputs a guess b' for b.

The advantage $\mathbf{Adv}^{anon}_{\mathcal{A},SCDI}(n)$ is defined to be $|Pr[b' = b] - 1/2|$.

2.5.3 Unambiguity

A Signcryption Scheme with Delayed Identification (SCDI) is said to achieve unambiguity if any probabilistic polynomial-time adversary $\mathcal{A}$ has negligible chance of winning against a challenger $\mathcal{C}$ in the following game:

1. Given the security parameter, $\mathcal{C}$ generates common parameter par and gives it to $\mathcal{A}$.
2. $\mathcal{A}$ outputs a receiver key-pair (pk_B, sk_B), two tags τ_0, τ_1 corresponding to different senders A_0 and A_1, and a signcryptext c.

$\mathcal{A}$ wins the game if $\mathsf{Verify}(par, \mathsf{Decrypt}(par, sk_B, c), \tau_b) = 1$ for both $b \in \{0, 1\}$.

2.5.4 Unforgeability

A Signcryption Scheme with Delayed Identification (SCDI) is said to achieve multi-user insider existential signcryptext unforgeability against chosen message attack in UF-SCDI-CMA sense if any probabilistic polynomial-time adversary $\mathcal{A}$ has negligible chance of winning against a challenger $\mathcal{C}$ in the following game:

1. Given the security parameter, $\mathcal{C}$ generates common parameter par and then with that generates a sender key-pair (pk_A, sk_A) using $\mathsf{KeyGen}_\mathsf{A}$.
2. $\mathcal{A}$ is given par, pk_A as well as access to the (flexible) signcryption oracle $\mathsf{Signcrypt}(par, sk_A, \cdot, \cdot)$. Each signcryption query consists of a pair $(pk_{B'}, m)$ where $pk_{B'}$ is a receiver public-key. The oracle answers it with $c = \mathsf{Signcrypt}(par, sk_A, pk_{B'}, m)$ and a corresponding tag τ, which verifies the authorship of A on c.
3. $\mathcal{A}$ outputs a receiver key pair (pk_B, sk_B), a signcryptext c^* and a tag τ^*.

$\mathcal{A}$ wins the game if $\mathsf{Verify}(par, \mathsf{Decrypt}(par, sk_B, c^*), \tau^*) = 1$ and if its underlying plaintext m^* was never submitted to the (flexible) signcryption oracle $\mathsf{Signcrypt}(par, sk_A, pk_B, \cdot)$.

2.5.5 Unlinkability

A Signcryption Scheme with Delayed Identification (SCDI) is said to achieve unlinkability if any probabilistic polynomial-time adversary $\mathcal{A}$ has negligible chance of winning against a challenger $\mathcal{C}$ in the following game:

1. Given the security parameter, $\mathcal{C}$ generates common parameter par, then with that generates a receiver key-pair (pk_B, sk_B) using $\mathsf{KeyGen_B}$ and feeds $\mathcal{A}$ with (par, pk_B).
2. $\mathcal{A}$ outputs a challenge sender-key pair (pk_A, sk_A) to $\mathcal{B}$.
3. $\mathcal{B}$ outputs a signcryptext c_1 from A to B (i.e., created under pk_B and sk_A).
4. $\mathcal{A}$ outputs a signcryptext c_2.

$\mathcal{A}$ wins the game if c_2 is a valid signcryptext under pk_B & $\mathsf{Link}(par, sk_B, c_1, c_2) = 1$.

3 The Proposed Construction

In this section, we propose a generic construction of a signcryption scheme with delayed identification (SCDI) from a randomness extractable public-key encryption (RE-PKE), a commitment scheme (CS), and a signature scheme (SS). The construction is based on "Sign-then-Commit-then-Encrypt" paradigm.

Let Π=(Gen, Enc, Dec) be an RE-PKE scheme with message space $\{0, 1\}^k$, CS=(CMT, DCMT) be a commitment scheme, and $\mathcal{SS}$=(Gen′, Sign, Ver) be a signature scheme. Let $l, t << k$ be two positive integers such that $2^{-l}, 2^{-t}$ are negligible. We construct a signcryption scheme with delayed identification (SCDI) given by (Setup, $\mathsf{Keygen_A}$, $\mathsf{Keygen_B}$, Signcrypt, Decrypt, Verify, ReSigncrypt, Link) with message space $\{0, 1\}^{k-t-l}$ as follows:

1. Setup:
   (a) $\mathsf{Setup}(1^n) \rightarrow par$. ($par$ denotes the common parameter required by the signcryption scheme.)
   (b) Choose a hash function $H : \{0, 1\}^* \rightarrow \{0, 1\}^l$.
   (c) Publish par, H globally.
2. $\mathsf{KeyGen_A}$: $\mathsf{Gen}'(par) \rightarrow (pk_A, sk_A)$
3. $\mathsf{KeyGen_B}$:
   (a) $\mathsf{Gen}(par) \rightarrow (pk_B, sk_B)$
   (b) B publishes pk_B and keeps sk_B as his decryption key.

4. Signcrypt: For a given message $m \in \{0,1\}^{k-t-l}$,
   (a) $s = \mathsf{Sign}(sk_A, m)$.
   (b) $(\sigma, \eta) := \mathsf{CMT}(s||pk_A)$ and set $\tau = (s||\eta||pk_A)$.
   (c) Choose $r \in_R \{0,1\}^t$ and compute $\alpha = H(m||r||\sigma||\tau)$.
   (d) Compute $c := \mathsf{Signcrypt}(pk_B, m, \sigma) = \mathsf{Enc}(pk_B, m||r||\sigma, \alpha)$
5. Decrypt: For a given signcryptext c,
   (a) $\mathsf{Dec}(sk_B, c) := (m||r||\sigma, \alpha)$ or $\perp$, if the signcryptext is invalid.
6. Verify: For a given $(m||r||\sigma, \alpha)$ and a tag τ,
   (a) Parse τ as s, η, pk_A.
   (b) If $\mathsf{Ver}(pk_A, m, s) = 1$, $\mathsf{DCMT}(s||pk_A, \sigma, \eta) = 1$ and $\alpha = H(m||r||\sigma||\tau)$, output 1, else output 0.
7. ReSigncrypt: For a given signcryptext c_1, the sender key-pair (pk_A, sk_A) and a revised message m_2,
   (a) Recollect $(m_1||r_1||\sigma_1, \alpha_1)$, the plaintext and randomness pair for c_1. (As Alice herself resigncrypts, she knows $(m_1||r_1||\sigma_1, \alpha_1)$.)
   (b) $s_2 = \mathsf{Sign}(sk_A, m_2)$.
   (c) $(\sigma_2, \eta_2) := \mathsf{CMT}(s_2||pk_A)$ and set $\tau_2 = (s_2||\eta_2||pk_A)$.
   (d) Compute $\alpha_2 = H(m_2||r_1||\sigma_2||\tau_2)$.
   (e) Compute $c_2 := \mathsf{Signcrypt}(pk_B, m_2, \sigma_2) = \mathsf{Enc}(pk_B, m_2||r_1||\sigma_2, \alpha_2)$
   (f) Output (c_1, c_2). [Note that the same randomness r_1 is used while computing α_2 and c_2.]
8. Link: For given signcryptexts c_1, c_2 and receiver secret key sk_B,
   (a) Compute $(m_i||r_i||\sigma_i, \alpha_i) = \mathsf{Dec}(sk_B, c_i)$ for $i = 1, 2$.
   (b) If $r_1 = r_2$, output 1, else output 0.

4 Security Analysis of Construction

Theorem 1 *$\mathcal{SCDI}$ is IND-SCDI-CCA2 secure in the sense of dynamic multi-user insider confidentiality in random oracle model if the underlying RE-PKE is IND-CPA secure.*

Proof We will construct an IND-CPA adversary $\mathcal{B}$ against RE-PKE using an IND-SCDI-CCA2 adversary $\mathcal{A}$ in dynamic multi-user insider model against $\mathcal{SCDI}$. As an input, $\mathcal{B}$ is fed with pk_B of RE-PKE, which $\mathcal{B}$ passes on to $\mathcal{A}$. Moreover, the H and Decrypt oracle queries are provided by $\mathcal{B}$.

Simulation of H-oracle: When $\mathcal{A}$ submits a H-query $m_i||r_i||\sigma_i||\tau_i$, $\mathcal{B}$ first checks whether (σ_i, τ_i) is a valid stub-tag pair for m_i or not, using the algorithm Ver and DCMT. (Note that $\tau_i = s_i||\eta_i||pk_{A_i}$) If it is valid, $\mathcal{B}$ chooses a random $\alpha_i \in \{0,1\}^l$

and returns α_i to $\mathcal{A}$. If it is not valid, $\mathcal{B}$ returns "invalid query". For each returned value, $\mathcal{B}$ maintains a list called H-list containing $(m_i||r_i||\sigma_i||\tau_i, \alpha_i)$

Simulation of decryption oracle ($\mathcal{O}$Decrypt): In decryption queries, when a query (c, pk'_A) is asked, $\mathcal{B}$ checks whether any previous query-answer history in H-list leads to c, i.e., $c = \mathsf{Enc}(pk_B, m_i||r_i||\sigma_i, \alpha_i)$ for any entry $(m_i||r_i||\sigma_i||\tau_i, \alpha_i)$ in H-list or not. If yes, then $\mathcal{B}$ parses that corresponding τ_i as $s_i||\eta_i||pk_{A_i}$ and checks if $pk'_A = pk_{A_i}$. If both the checks are cleared, then $\mathcal{B}$ returns that corresponding $(m_i||r_i||\sigma_i, \alpha_i)$ else return "invalid decryption query". *It should be noted here that while answering H-queries, $\mathcal{B}$ ensures that the answer α_i does not lead to a signcryptext which was previously declared as "invalid" by the decryption oracle.* This provides a perfect simulation since the probability of producing a valid query without making the corresponding H-query is zero.

Once the first query phase is over, $\mathcal{A}$ returns two plaintexts $m_0, m_1 \in \{0, 1\}^{k-t-l}$ and an attacked sender key-pair (pk_A, sk_A) to $\mathcal{B}$. $\mathcal{B}$ randomly chooses $r_0, r_1 \in_R \{0, 1\}^t$, computes (σ_0, τ_0), (σ_1, τ_1) as in Signcrypt and submits $m_0||r_0||\sigma_0, m_1||r_1||\sigma_1$ to the IND-CPA challenger $\mathcal{C}$. $\mathcal{C}$ randomly chooses a bit $b \in_R \{0, 1\}$, $\alpha \in_R \{0, 1\}^l$. $\mathcal{C}$ returns the challenge ciphertext $c^* = \mathsf{Enc}(pk_B, m_b||r_b||\sigma_b, \alpha)$ to $\mathcal{B}$ and $\mathcal{B}$ passes it on to $\mathcal{A}$.

In the second query phase, $\mathcal{A}$ is allowed to make H-queries as before and decryption queries other than the challenge ciphertext c^*. If $\mathcal{A}$ makes a *valid* H-query with $(m_0||r_0||\sigma_0||\tau_0)$ or $(m_1||r_1||\sigma_1||\tau_1)$, one stops the game and $\mathcal{B}$ returns failure. (The rationale behind this thought is discussed in the proof of Lemma 1.) If not, after the second query phase is over, $\mathcal{A}$ outputs a guess b' to $\mathcal{B}$ and $\mathcal{B}$ returns b'. The theorem now follows immediately from the following lemma.

Lemma 1 *If ϵ be the probability that given a valid signcryptext, $\mathcal{A}$ can correctly guess the bit b, then $\mathcal{B}$ can win the IND-CPA game with a probability greater or equal to $\epsilon - q_H/2^t - q_H/2^l$, where l denotes the length of the hash output, t denotes the length of randomness used in the signcrytion algorithm, and q_H denotes the number of hash queries.*

Proof If, in the second query phase, $\mathcal{A}$ asks a *valid* H-query $m_b||r_b||\sigma_b||\tau_b$ and gets the answer α', then c^* will be a valid ciphertext only if $c^* = \mathsf{Enc}_{pk}(m_b||r_b||\sigma_b, \alpha')$ i.e., $\alpha = \alpha'$. Therefore, in order to maintain the validity of c^*, $\mathcal{B}$ should respond to the query $(m_b||r_b||\sigma_b||\tau_b)$ with $\alpha' =$ randomness used in c^* by IND-CPA challenger $\mathcal{C}$. However, the probability of guessing the right α' for α is $1/2^l$, which is negligible. This is the reason that $\mathcal{B}$ aborts the game when a *valid* $(m_0||r_0||\sigma_0||\tau_0)$ or $(m_1||r_1||\sigma_1||\tau_1)$ is queried upon. However, the probability of abortion due to the above reason is $\leq q_H/2^t$, where q_H is the total number of H-queries.

Now, if $(m_0||r_0||\sigma_0||\tau_0)$ or $(m_1||r_1||\sigma_1||\tau_1)$ are not queried in this second query phase, it is valid to assume that $H(m_b||r_b||\sigma_b||\tau_b) = \alpha$, (where α is the randomness used by $\mathcal{C}$) except the case when there exists another previous entry of the form $(m||r||\sigma||\tau, \alpha)$ in the H-list. However, the probability that α have been the response to some valid H-query previously is $\leq q_H/2^l$.

Thus, c^* is a valid ciphertext except the case when $(m_0||r_0||\sigma_0||\tau_0)$ or $(m_1||r_1||\sigma_1||\tau_1)$ is queried to the H-oracle or α has been received as a response from the

H-oracle. Hence, $\mathcal{A}$ can guess the correct bit b, i.e., $\mathcal{B}$ can win the IND-CPA game with probability $\geq \epsilon - q_H/2^t - q_H/2^l$. □

Theorem 2 *$\mathcal{SCDI}$ is anonymous in random oracle model if the underlying commitment scheme CS is hiding.*

Proof We construct an adversary $\mathcal{B}$ against hiding property of CS using an adversary $\mathcal{A}$ against anonymity of $\mathcal{SCDI}$. $\mathcal{A}$ outputs two sender key-pairs (pk_{A0}, sk_{A0}), (pk_{A1}, sk_{A1}), a receiver key-pair (pk_B, sk_B) and a message m to $\mathcal{B}$. $\mathcal{B}$ simulates the H-oracle for $\mathcal{A}$ and computes $s_0 = \mathsf{Sign}(sk_{A0}, m)$ and $s_1 = \mathsf{Sign}(sk_{A1}, m)$ and output $(s_0||pk_{A0})$ and $(s_1||pk_{A1})$ to the challenger $\mathcal{C}$ of the hiding property of CS. $\mathcal{C}$ chooses a bit $b \in_R \{0, 1\}$, computes $(\sigma^*, \eta^*) = \mathsf{CMT}(s_b||pk_{Ab})$ and output σ^* to $\mathcal{B}$. B chooses $r \in_R \{0, 1\}^t$, $\alpha \in_R \{0, 1\}^l$ and returns $c^* = \mathsf{Enc}(pk_B, m||r||\sigma^*, \alpha)$ to $\mathcal{A}$. In the guess phase, $\mathcal{A}$ outputs a bit b' to $\mathcal{B}$ and $\mathcal{B}$ outputs the same b' to $\mathcal{C}$.

Simulation of H-oracle: When $\mathcal{A}$ submits an H-query $m'||r'||\sigma'||\tau'$, $\mathcal{B}$ parses $\tau' = s'||\eta'||pk_{A'}$ and checks if $\mathsf{DCMT}(s'||pk_{A'}, \sigma', \eta') = 1$. If not, then return "invalid query". If it is "valid," check if $m'||r'||\sigma' = m||r||\sigma^*$. If they are not equal, then choose $\alpha' \in_R \{0, 1\}^l$ such that $\alpha' \neq \alpha$ and return it to $\mathcal{A}$. If they are equal, return α to $\mathcal{A}$. For each returned value, $\mathcal{B}$ maintains a list called H-list containing $(m'||r'||\sigma'||\tau', \alpha')$.

Note that the simulation of H-oracle is perfect in $\mathcal{A}$'s view. Now the theorem follows from the fact that $Pr[\mathcal{B}_{CS}^{hiding}|b = b'] = Pr[\mathcal{A}_{SCDI}^{anon}|b = b']$. □

Theorem 3 *$\mathcal{SCDI}$ is unambiguous if the underlying commitment scheme CS is binding.*

Proof We construct an adversary $\mathcal{B}$ against binding property of CS using an adversary $\mathcal{A}$ against unambiguity of $\mathcal{SCDI}$. $\mathcal{A}$ outputs a receiver key-pair (pk_B, sk_B), a pair of tags τ_0, τ_1 corresponding to different senders A_0 and A_1, and a signcryptext c to $\mathcal{B}$. $\mathcal{B}$ decrypts c with sk_B to get $(m||r||\sigma, \alpha)$ and parses τ_0 as $(s_0||\eta_0||pk_{A0})$ and τ_1 as $(s_1||\eta_1||pk_{A1})$. $\mathcal{B}$ then outputs $(s_0||pk_{A0}, \sigma, \eta_0)$ and $(s_1||pk_{A1}, \sigma, \eta_1)$ to the challenger $\mathcal{C}$ against the binding property of CS.

Let us define the events $F_0, F_1, G_0, G_1, W_0, W_1$ as follows:
$F_0 =$ event $\mathsf{DCMT}(s_0||pk_{A0}, \sigma, \eta_0) = 1$, $F_1 =$ event $\mathsf{DCMT}(s_1||pk_{A1}, \sigma, \eta_1) = 1$,
$G_0 =$ event that $\mathsf{Ver}(pk_{A0}, m, s_0) = 1$ and $G_1 =$ event that $\mathsf{Ver}(pk_{A1}, m, s_1) = 1$.
$W_0 =$ event that $\alpha = H(m||r||\sigma||\tau_0)$ and $W_1 =$ event that $\alpha = H(m||r||\sigma||\tau_1)$.
Now, the theorem follows from the fact that $Pr[\mathcal{B}\ wins] = Pr[F_0 \cap F_1] \geq Pr[(F_0 \cap G_0 \cap W_0) \cap (F_1 \cap G_1 \cap W_1)] = Pr[\mathcal{A}\ wins]$. □

Theorem 4 *$\mathcal{SCDI}$ is multi-user insider existential signcryptext unforgeable against chosen message attack in UF-SCDI-CMA sense in standard model if the underlying signature scheme SS is UF-CMA secure.*

Proof We construct an UF-CMA adversary $\mathcal{B}$ against SS using an UF-SCDI-CMA adversary $\mathcal{A}$ against $\mathcal{SCDI}$. $\mathcal{B}$ takes as input the common parameter par, a sender public-key pk_A and a signing oracle $\mathcal{O}\mathsf{Sign}(sk_A, \cdot)$. $\mathcal{B}$ chooses a hash function

$H : \{0,1\}^* \rightarrow \{0,1\}^l$ and feeds $\mathcal{A}$ with par, pk_A and H. In the query phase, $\mathcal{A}$ submits a query for $(pk_{B'}, m_i)$, $\mathcal{B}$ queries the $\mathcal{O}\mathsf{Sign}(sk_A, \cdot)$ with m_i to get a response s_i, computes $(\sigma_i, \eta_i) = \mathsf{CMT}(s_i||pk_A)$ and $\tau_i = (s_i||\eta_i||pk_A)$. $\mathcal{B}$ then chooses $r_i \in_R \{0,1\}^t$ and computes $c_i = \mathsf{Enc}(pk_{B'}, m_i||r_i||\sigma_i, \alpha_i)$, where $\alpha_i = H(m_i||r_i||\sigma_i||\tau_i)$ and finally returns (c_i, τ_i) to $\mathcal{A}$. $\mathcal{B}$ also maintains a list, S-list, consisting of the queried messages, m_i's. Once the query phase is over, $\mathcal{A}$ outputs a receiver key pair (pk_B, sk_B), a signcryptext c^* and a corresponding tag τ^* to $\mathcal{B}$. $\mathcal{B}$ computes $(m^*||r^*||\sigma^*, \alpha^*) = \mathsf{Dec}(sk_B, c^*)$ and parses τ^* as $(s^*||\eta^*||pk_A)$ and returns (m^*, s^*) to the UF-CMA challenger $\mathcal{C}$.

Let U be event that s^* is a valid signature on m^*, i.e., $\mathsf{Ver}(pk_A, m^*, s^*) = 1$ and $m^* \notin S$-list and V be the event that (σ^*, η^*) is a valid commitment-decommitment pair for $s^*||pk_A$ i.e., $\mathsf{DCMT}(s^*||pk_A, \sigma^*, \eta^*) = 1$ and $\alpha^* = H(m^*||\ r^*||\sigma^*||\tau^*)$. Note that, if m^*, the underlying message of c^*, has not been submitted to the signcryption oracle $\mathcal{O}\mathsf{Signcrypt}(sk_A, pk_B, \cdot)$, then, as per the simulation, m^* have not been queried to the signing oracle $\mathcal{O}\mathsf{Sign}(sk_A, \cdot)$. Hence, we have $Pr[\mathcal{B}\ wins] = Pr[U] \geq Pr[U \cap V] = Pr[\mathcal{A}\ wins]$. □

Theorem 5 *$\mathcal{SCDI}$ is unlinkable in random oracle model if the underlying encryption scheme Π is NM-CPA secure.*

Proof We will construct an NM-CPA adversary $\mathcal{B}$ against Π using a linking adversary $\mathcal{A}$ against $\mathcal{SCDI}$. As an input, $\mathcal{B}$ is fed with common parameter par and pk_B of Π, which $\mathcal{B}$ passes on to $\mathcal{A}$. Moreover, the H-queries needed by $\mathcal{A}$ will be provided by $\mathcal{B}$.

Simulation of H-oracle: When $\mathcal{A}$ submits an H-query $m_i||r_i||\sigma_i||\tau_i$, $\mathcal{B}$ first checks whether (σ_i, τ_i) is a valid stub-tag pair for m_i or not, using the algorithm Ver and DCMT. (Note that $\tau_i = s_i||\eta_i||pk_{A_i}$) If it is valid, $\mathcal{B}$ chooses a random $\alpha_i \in \{0,1\}^l$ and returns α_i to $\mathcal{A}$. If it is not valid, $\mathcal{B}$ returns "invalid query." For each returned value, $\mathcal{B}$ maintains a list called H-list containing $(m_i||r_i||\sigma_i||\tau_i, \alpha_i)$.

In the challenge phase, $\mathcal{A}$ submits a target sender key pair (pk_A, sk_A) to $\mathcal{B}$. Now, $\mathcal{B}$ recieives a challenge ciphertext c_1 from the NM-CPA challenger $\mathcal{C}$ and passes it to $\mathcal{A}$ as the challenge signcryptext. $\mathcal{A}$ continues to have access to the H-oracle. *It should be noted here that while answering H-queries in second phase, $\mathcal{B}$ ensures that any answer α_i does not lead to* c_1, i.e., $c_1 \neq \mathsf{Enc}(pk_B, m_i||r_i||\sigma_i, \alpha_i)$ $\mathcal{A}$ finally outputs another signcryptext c_2 as the linked signcryptext. $\mathcal{B}$ outputs (c_2, ρ), where ρ is the relation between two ciphertexts if their corresponding plaintexts have the same $(k-t-l+1)$-th bit to $(k-l)$-th bit.

Now, let us consider $\mathcal{A}$'s view toward c_1. Let $c_1 = \mathsf{Enc}(pk_B, m_1||r_1||\sigma_1, \alpha_1)$. Let E_1 be the event that $(m_1||r_1||\sigma_1||\cdot)$ has been queried to the H-oracle in the first phase and E_2 be the event that α_1 has been received as response from the H-oracle in the first query phase. Thus, the simulation of H-oracle is perfect in $\mathcal{A}$'s view unless E_1 or E_2 occurs, i.e., the simulation is correct in $\mathcal{A}$'s view with probability $\mathsf{Pr}[(E_1 \cup E_2)^c] = \mathsf{Pr}[{E_1}^c \cap {E_2}^c] = \mathsf{Pr}[{E_1}^c] \cdot \mathsf{Pr}[{E_2}^c] = \left(1 - \frac{q_H}{2^k}\right) \cdot \left(1 - \frac{q_H}{2^l}\right)$ $\geq 1 - q_H/2^k - q_H/2^l$, where q_H denotes the total number of H-queries. Now, we define $\mathcal{B}_{\mathsf{win}}$ to be the event that $\mathcal{B}$ wins the NM-CPA game and $\mathcal{A}^{\mathsf{real}}_{\mathsf{win}}$ to be the event

that $\mathcal{A}$ wins the real unlinkability game. Thus, as per the simulation of $\mathcal{B}$, we have, $\mathsf{Pr}[\mathcal{B}_{\mathsf{win}}] \geq \mathsf{Pr}[\mathcal{A}^{\mathsf{real}}_{\mathsf{win}}] \cdot (1 - q_H/2^k - q_H/2^l)$. □

Remark 7 It is to be noted that an adversary can always mount a denial-of-service (DoS) attack on this primitive by submitting numerous *garbage* ciphertexts. As the sender's anonymity is maintained in the decryption phase, the correctness of the committed signature cannot be tested, i.e., before the sender decommits the signature, the receiver is not able to verify whether the received ciphertext is a proper output of the signcryption scheme. This is not a limitation of this construction, rather it is a price to be paid for using this primitive. On the other side, the risk of DoS attacks is also there if we use a normal signcryption scheme in the existing scenarios, i.e., an adversary can create multiple fake email-ids and submit numerous ciphertexts created against them.

5 Conclusion

In this paper, we have introduced a new primitive called Signcryption with Delayed Identification (SCDI) and discussed its application in various functionalities. We put forward proper security notions for this primitive and a generic construction of SCDI from basic cryptographic primitives secure under these notions. Similar to the goal of signcryption, as pointed out by Zheng [17], not only do we achieve the goal that length of ciphertext of SCDI is less than the length of individual encryption and partial signature, but also gain in terms of security from the component primitives. Further research in this direction could be related to its efficiency and hence hybrid SCDI can be an interesting open issue.

References

1. An, J.H., Dodis, Y., Rabin, T.: On the security of joint signature and encryption. In: Proceedings of the EUROCRYPT 2002. LNCS, vol. 2332, pp. 83–107. Springer (2002)
2. Baek, J., Steinfeld, R., Zheng, Y.: Formal proofs for the security of signcryption. J. Cryptol. **20**(2), 203–235 (2007)
3. Bellare, M., Boldyreva, A., Desai, A., Pointcheval, D.: Key-privacy in public-key encryption. In: Proceedings of the ASIACRYPT 2001. LNCS, vol. 2248, pp. 566–582. Springer (2001)
4. Bellare, M., Duan, S.: Partial signatures and their applications, Cryptology ePrint Archive, Report 2009/336. http://eprint.iacr.org/2009/336 (2009)
5. Bellare, M., Rogaway, P.: Optimal asymmetric encryption—how to encrypt with RSA. In: Proceedings of the EUROCRYPT '94. LNCS, vol. 950. Springer (1995)
6. Bellare, M., Rogaway, P.: Random oracles are practical: a paradigm for designing efficient protocols. In: Proceedings of the 1st CCS, pp. 62–73. ACM Press, New York (1993)
7. Coron, J.S., Handschuh, H., Joye, M., Paillier, P., Pointcheval, D., Tymen, C.: GEM: A generic chosen-ciphertext secure encryption method. In: Proceedings of the CT-RSA 2002. LNCS, vol. 2271, pp. 263–276. Springer (2002)

8. Deva Selvi, S.S., Vivek, S.S., Shiriam, J., Kalaivani, S., Pandu Rangan, C.: Identity based aggregate signcryption schemes. In: Proceedings of the INDOCRYPT 2009. LNCS, vol. 5922, pp. 378–397
9. Das, A., Adhikari, A.: An efficient IND-CCA2 secure Paillier-based cryptosystem. Inf. Process. Lett. **112**, 885–888 (2012) (Elsevier, 2012)
10. Fujisaki, E., Okamoto, T.: How to enhance the security of public-key encryption at minimum cost. In: Proceedings of the PKC '99. LNCS, vol. 1560, pp. 53–68. Springer (1999)
11. Hang, Q., Wong, D.S., Yang, G.: Heterogeneous signcryption with key privacy. Comput. J. **54** (4), 525–536 (2011)
12. Ma, C., Chen, K., Zheng, D., Liu, S.: Efficient and proactive threshold signcryption,. In: Proceedings of the ISC 2005. LNCS, vol. 3650, pp. 233–243. Springer (2005)
13. Malone-Lee, J.: Identity-Based Signcryption, Cryptology ePrint Archive, Report 2002/098. http://eprint.iacr.org/2002/098
14. Okamoto, T., Pointcheval, D.: REACT: Rapid enhanced-security asymmetric cryptosystem transform. In: Proceedings of the CT-RSA 2001. LNCS vol. 2020, pp. 159–174. Springer (2001)
15. Paillier, P.: Public-key cryptosystems based on composite degree residuosity classes. In: Proceedings of the EUROCRYPT '99. LNCS, vol. 1592, pp. 223–238. Springer (1999)
16. Pointcheval, D.: Chosen-ciphertext security for any one-way cryptosystem. In: Proceedings of the PKC 2000. LNCS, vol. 1751, pp. 29–146. Springer (2000)
17. Zheng, Y.: Digital signcryption or how to achieve cost (signature & encryption) $\ll$ cost(signature) + cost(encryption). In: Proceedings of the CRYPTO 97. LNCS, vol. 1294, pp. 165–179. Springer (1997)

Chapter 4
HDNM8: A Round-8 High Diffusion Block Cipher with Nonlinear Mixing Function

Jaydeb Bhaumik and Dipanwita Roy Chowdhury

Abstract Since 2001, AES-128 is accepted as the standard block cipher. Till date, full-round AES is secure against all existing attacks, but reduced-round versions are susceptible to several attacks. In this paper, diffusion of AES-like block cipher is improved by incorporating a 128-bit diffusion layer based on a maximum distance separable code. Moreover, a nonlinear vectorial Boolean function is employed for round key mixing, which improves the nonlinearity. Employing this high diffusion and improved nonlinearity, a new block cipher called 'HDNM8' is proposed. It is shown that HDNM8 is secure against several existing cryptographic attacks. HDNM8 has been implemented on an FPGA platform. It has been found that it requires reasonable hardware and provides an acceptable throughput.

1 Introduction

Confusion and diffusion are two important cryptographic properties for the design of a secure block cipher. Each round function of a Substitution Permutation Networks (SPN) type block cipher consists of three layers: substitution layer, permutation layer, and round key mixing layer. The permutation layer dissipates the statistics of the plaintext in the statistics of the ciphertext; it is often referred to as the diffusion layer. The substitution layer creates confusion, i.e., it makes the relationship between the key and ciphertext as complex as possible. AES [10] is the most popular SPN-type block cipher and it has a wide range of applications. It has a block size of 128-bit and number of rounds 10/12/14. There are also some constructions, which use four

J. Bhaumik (✉)
Haldia Institute of Technology, Haldia 721657, India
e-mail: bhaumik.jaydeb@gmail.com

D. Roy Chowdhury
Indian Institute of Technology Kharagpur, Kharagpur 721302, India
e-mail: drc@cse.iitkgp.ernet.in

R. N. Mohapatra et al. (eds.), *Mathematics and Computing 2013*,
Springer Proceedings in Mathematics & Statistics 91,
DOI: 10.1007/978-81-322-1952-1_4,

rounds of AES as building block such as Pelican MAC [11], PC-MAC [17], and the stream cipher LEX [6].

Only a substitution layer, which is strong against differential cryptanalysis (DC) and linear cryptanalysis (LC), does not guarantee a secure SPN structure against DC and LC if a diffusion layer does not provide an avalanche effect. Hence, the role of the diffusion layer is very important for the design of secure block cipher. AES employs a 32-bit diffusion layer for a 128-bit block cipher. The AES diffusion layer is based on a Maximum Distance Separable (MDS) code and the distance between any two distinct codewords (called branch number [9]) is five. In AES, all plaintext bits diffuse completely after two rounds. Therefore, diffusion in AES is relatively slow. Junod and Vaudenay have presented perfect diffusion primitives for block ciphers by considering software implementations on various platforms [14]. Authors in [14] have constructed efficient (4×4) and (8×8) matrices over GF(2^8) for block cipher. Hence, for a 128-bit block cipher, multiple parallel modules are required and complete diffusion is not possible in a single round. Koo et al. proposed binary (16×16) and (32×32) matrices for SPN-type block ciphers in [15, 16]. Recently, a new (16×16) involutory MDS matrix for AES is proposed in [18]. In scheme [18], complete diffusion is possible after a single round, but the drawback of the proposed construction is the performance penalty. SHARK [20] is a 64-bit block cipher, which uses a Reed-Solomon (RS) code to construct its diffusion layer. It has branch number 9. Two other block ciphers Khazad [2] and Anubis [1] have been designed by Barreto and Rijmen. Khazad is a 64-bit, 8-round block cipher and it employs an MDS diffusion layer, which has branch number 9. It provides complete diffusion after one round. Anubis is a 128-bit, 12–18 rounds block cipher, but it has a slower, Rijndael-like 32-bit diffusion layer [5]. A diffusion layer with large branch number increases the security of cipher. A common feature exploited by several existing attacks on reduced-round AES is the slow diffusion via the combination of ShiftRows and MixColumns [18].

Boolean functions XOR and addition modulo 2^n are popularly used for round key mixing in several existing block ciphers. Two popular block ciphers DES and AES use XOR as a key mixing function because it is balanced, involutary, and efficient for implementation, although it is purely linear. In case of block ciphers like IDEA, MARS, RC6, FEAL, SEA, addition mod 2^n is used for round key mixing operation. But in case of modulo addition, the individual output bit as well as a linear combination of consecutive output bits has a high bias value $\frac{1}{4}$. Therefore, it will be advantageous if we replace XOR or modulo addition by a nonlinear function, which maintains properties like balancedness, reversibility, low hardware complexity, in addition to good nonlinearity. Introduction of nonlinear key mixing function enhances the overall nonlinearity of the round function of an SPN-type block cipher.

In this Chapter, Cellular Automata (CA)-based MDS codes for diffusion layer, nonlinear round key mixing function, and AES S-boxes are used to design a new SPN-type block cipher called 'HDNM8' (**H**igh **D**iffusion **N**onlinear key **M**ixing with **8** rounds). Also, the strength of proposed cipher against existing attacks is evaluated and it has been implemented on an FPGA platform. The proposed design is amenable for hardware implementation.

The rest of this Chapter is organized as follows. The next section discusses the CA-based diffusion layer. A description of 'Nmix' function is given in Sect. 3. Design and implementation of the proposed block cipher HDNM8 is discussed in Sect. 4. The diffusion property of full round and reduced round versions of HDNM8 are examined in Sect. 5. Strength of the proposed cipher against existing attacks is analyzed in Sect. 6 and finally the paper is concluded in Sect. 7.

2 Diffusion Layer Using CA-Based MDS Code

A diffusion layer does not allow to preserve some characteristics that result from a substitution layer. Several SPN-type block ciphers use MDS codes for the construction of diffusion layer. The main aim in the design of MDS codes-based diffusion layer is to reduce the computational cost by selecting an appropriate MDS matrix. One such diffusion layer based on CA is introduced in [4]. For the sake of completeness, a brief description of CA and CA-based diffusion layer is given in this section.

2.1 Cellular Automata

It consist of a number of cells arranged in a regular manner, where the state transitions of each cell depends on the state of its neighbors. Each cell consists of a D flip-flop and a combinational logic implementing the next-state function. An r-cell linear CA can be characterized by an $(r \times r)$ binary characteristic matrix T. The i-th row of the matrix T describes the neighborhood relation of the i-th cell.If an element T_{ij} (at row i and column j of matrix T) is 1, then the ith cell in the array has neighborhood dependence on the jth cell.The state S_{t+1} can be computed by multiplying S_t with T, where S_t and S_{t+1} represents the states of the CA at t-th and $(t+1)$-th instant, respectively. If the state transition graph of an r-cell CA consists of a single cycle containing all $L = 2^r - 1$ nonzero states, then the CA is called as maximum length CA. One characteristic matrix (T) of an 8-cell maximum length CA is as follows:

$$T = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

The characteristic polynomial is defined as determinant of $(T+x[I])$. The polynomial associated with T is $p(x) = x^8 + x^7 + x^6 + x^5 + x^4 + x^2 + 1$, which is one of the primitive polynomials of GF(2^8). In the rest of this Chapter, T will indicate characteristic matrix of the 8-cell maximum length CA, which is mentioned above.

2.2 CA-Based MDS Code

A (n, m, d) code that meets the Singleton bound, namely $d = n - m + 1$, is called an MDS code, where m is the number of data symbols, n is the number of symbols in a codeword, and d is the distance of separation between two distinct codewords. For an MDS code, the minimum number of nonzero symbols in any codeword is d. The generator matrix $G = [I|M]$ of a (n, m, d) MDS code over GF(2^8) is a $(m \times n)$ matrix, where elements of G are in GF(2^8), I is a $(m \times m)$ identity matrix and M is a $m \times (n - m)$ matrix. Sometimes, the matrix M is designed using Vandermonde's construction. In this case, each element of M is power of a primitive element of GF(2^8). In case of maximum length CA, a characteristic matrix T is equivalent to a primitive element α. Therefore, the matrix $M_{16\times16}$ of a (32, 16, 17) code can be constructed from characteristic matrix T, where each element of M is a power of T, and it is as follows:

$$\mathrm{M} = \begin{bmatrix} T & T^2 & T^3 & \cdots & T^{15} & T^{16} \\ T^2 & T^4 & T^6 & \cdots & T^{30} & T^{32} \\ T^3 & T^6 & T^9 & \cdots & T^{45} & T^{48} \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ T^{15} & T^{30} & T^{45} & \cdots & T^{225} & T^{240} \\ T^{16} & T^{32} & T^{48} & \cdots & T^{240} & T \end{bmatrix}$$

The linear code generated by the generator matrix $G = [I|M]$ is an MDS code, where $I_{16\times16}$ is an identity matrix and each element of I is an (8×8) binary matrix. The linear code has dimension 16, length 32, and the minimum distance of separation between two distinct codewords is 17. The matrix M is sometimes called MDS matrix.

For a 128-bit block cipher, a single 128-bit diffusion layer can be used in all rounds, and it is advantageous for a single round iterative architecture. Figure 1 gives an estimation for the minimum number of active S-boxes in a 4-round cipher. In Fig. 1, black square boxes indicate the active S-boxes. There are a total of 34 active S-boxes in a 128-bit four rounds cipher. In case of AES, minimum number of active S-boxes for a 4-round cipher is 25. The higher the number of active S-boxes are, the more secure will be the cipher against DC and LC.

Boolean function XOR is used for round key mixing in many block ciphers. XOR is linear, so nonlinearity of round function solely depends on S-Box. But nonlinear

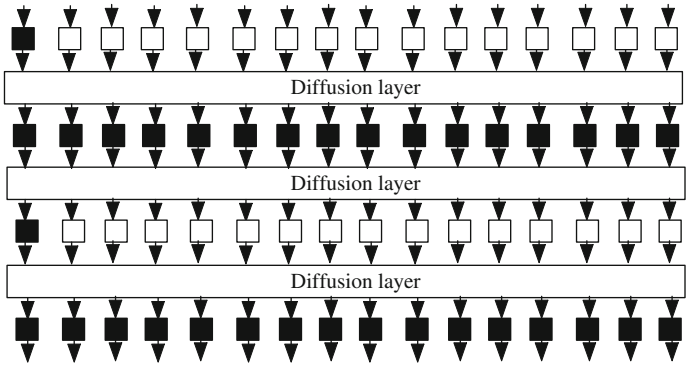


Fig. 1 Minimum number of active S-boxes

round key mixing function adds extra nonlinearity in the round function. Following section describes a nonlinear round key mixing function.

3 Function: *Nmix*

A nonlinear, reversible, balanced vectorial Boolean function *Nmix* is introduced in [3]. A brief description of *Nmix* is given in this section for the sake of completeness. *Nmix*: Assume that $X = (x_{n-1}\, x_{n-2}\, \ldots\, x_0)$ is an n-bit data, $K = (k_{n-1}\, k_{n-2}\, \ldots\, k_0)$ is an n-bit round-key and $Y = (y_{n-1}\, y_{n-2}\, \ldots\, y_0)$ is the n-bit output after mixing X with K. Then each output bit is related to the input bits by the following relationship:

$$y_i = x_i \oplus k_i \oplus c_{i-1}\ ;\ c_i = \bigoplus_{j=0}^{i} x_j k_j \oplus x_{i-1}x_i \oplus k_{i-1}k_i \tag{1}$$

where $0 \leq i < n$, $c_{-1} = 0$, $x_{-1} = 0$, $k_{-1} = 0$ and c_i is the carry term propagating from the i-th bit position to the $(i+1)$-th bit position. The end carry c_{n-1} is neglected. Each y_i is also balanced function for $0 \leq i < n$. It is shown in [3] that the bias for the best linear approximation of output bit y_i of *Nmix* is 2^{-i}, where $2 \leq i < n$. The bias of the best linear approximation for y_0 and y_1 are, respectively, $\frac{1}{2}$ and $\frac{1}{4}$. Further, the bias for the best linear approximation of $y_i \oplus y_{i+1}$ is 0.0625, where $2 \leq i < n$. In the following section, the function *Nmix* is employed for round-key mixing in a block cipher called 'HDNM8'. In case of *Nmix*, the round key is mixed with round input in byte by byte fashion, so that there is no carry propagation from lower significant byte to higher significant byte. Byte wise key mixing is used to minimize the carry propagation delay and to provide a reasonable amount of nonlinearity.

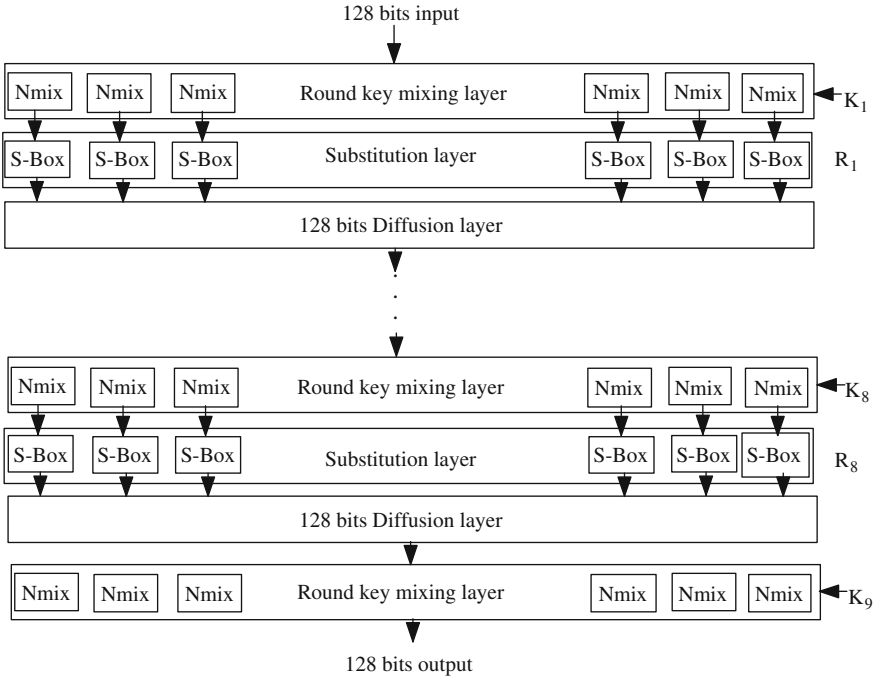


Fig. 2 Block diagram of HDNM8

4 Design and Implementation of HDNM8

In this section, a brief introduction of cipher HDNM8 is given first. It is a 128-bit SPN type block cipher with number of rounds is eight. Each round consists of three layers: nonlinear round key mixing layer, substitution layer having 16 AES S-boxes, and a single 128-bit diffusion layer. In this cipher, the ShiftRows and MixColumns operations of AES-like ciphers are replaced by a single 128-bit diffusion layer. There are sixteen *Nmix* modules in each key mixing layer. The block diagram of the cipher HDNM8 is shown in Fig. 2. Proposed cipher can operate in counter mode, output feedback mode, and cipher feedback mode, where only encryption module is required for both encryption and decryption. In Fig. 2, K_i's are round keys, where $1 \leq i \leq 9$, K_1 is the cipher key, and other round keys are generated from the cipher key using key schedule algorithm of AES. There are eight similar type of rounds (R_1 to R_8) and one final round key mixing. In the design of HDNM8, AES S-boxes are used in the substitution layer. The operation of AES S-box can be described as [10]

$$S(z) = L(z^{-1}) + c \tag{2}$$

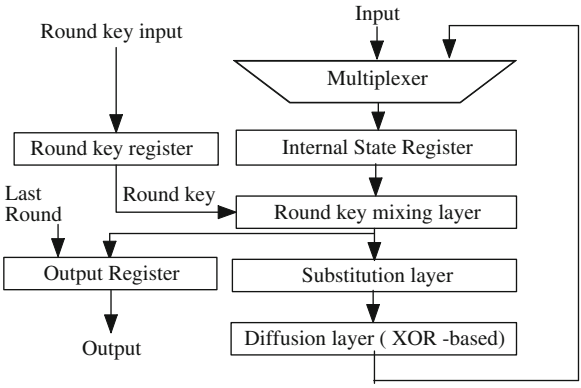


Fig. 3 Architecture of HDNM8 with 16 S-boxes

Table 1 Resources for HDNM8 with 16 S-boxes implementation

Resource type	HDNM8				AES	
	Round key mixing by					
	Nmix		XOR		XOR	
	#	%	#	%	#	%
Slices	1,736	31	1,613	29	1,028	18
Slice FFs	388	3	260	2	388	3
4-input LUTs	3,352	30	3,131	29	1,994	18
Frequency (MHz)	101.535		112.018		151.493	
Throughput (Gbps)	1.299		1.433		1.615	

where $z \in \text{GF}(2^8)$, z^{-1} is the multiplicative inverse of z and 0 is mapped to 0. L is a linear transformation in GF(2) and c is a constant. For efficient hardware implementation of AES S-box, composite field arithmetic is frequently used. In [7], the author has proposed a compact implementation of S-box using normal basis for each subfield. For implementation of HDNM8, Canright's scheme is employed for S-box design. The present section explains two architectures of HDNM8 for VLSI implementation. The first approach primarily focuses on latency (in terms of clock cycles) optimization, whereas the second approach addresses the area (in terms of gate count) optimization.

4.1 Latency Optimized Implementation

In the first architecture, there are sixteen S-boxes in a substitution layer. Each S-box is implemented using normal basis [7]. The diffusion layer is constructed using a single 128-bit layer and the corresponding MDS matrix M (explained in Sect. 2) has dimension (16×16), where each element is an (8×8) binary matrix. A (128×128) binary matrix is realized by substituting the values of all elements,

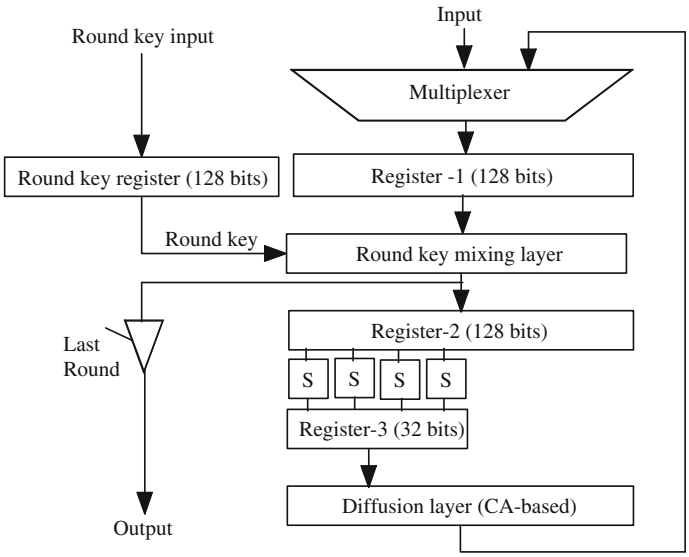


Fig. 4 Architecture of HDNM8 with four S-boxes

Table 2 Resources for HDNM8 with four S-boxes implementation

Resource type	Round key mixing by Nmix	
	#	%
Slices	748	13
Slice FFs	691	6
4-input LUTs	1,357	12
Frequency (MHz)	178.524	
Throughput (Mbps)	148.383	

which are power of T. The value of T is given in Sect. 2 and the other powers of T are obtained by matrix operation in GF(2). As a result, each output bit of the diffusion layer can be expressed as bitwise XOR of input bits. Round key mixing is done by *Nmix* function. In the proposed cipher, we use the AES key-schedule algorithm, and hence it is not implemented in this work. It is assumed that round-keys are available in round-key register during round operation. Table 1 shows the resources used for FPGA implementation of the architecture given in Fig. 3. Every architectural module has been implemented in Verilog and simulated using ModelSim XE III 6.0a. The design has been synthesized by Xilinx ISE 7.1i tool and the target FPGA device was Virtex 4vfx12sf363-12. In Table 1, resources required for key scheduling algorithm is not considered. Moreover, an iterative architecture of AES-128 is implemented using S-box architecture proposed in [7]. Table 1 shows a comparison of synthesis results of HDNM8 with AES. From Table 1, it is noted that the implementation of AES-128 in FPGA requires lesser amount of resources as compared to HDNM8. It is found that proposed cipher provides reasonable throughput.

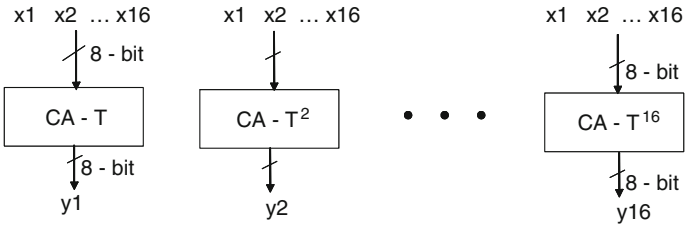


Fig. 5 Block diagram of 128-bit diffusion layer

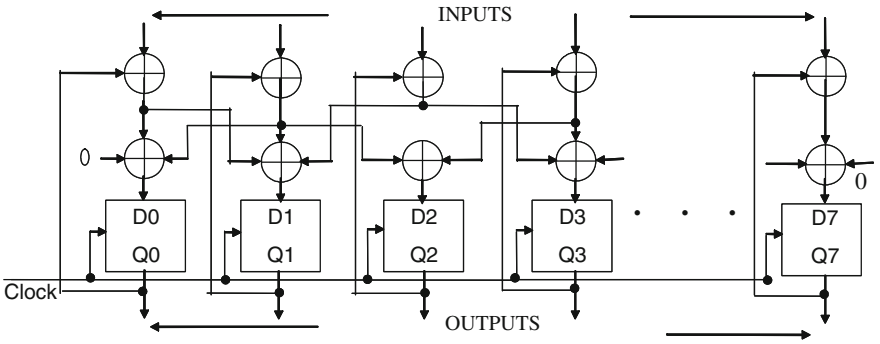


Fig. 6 Internal architecture of CA-T

4.2 Area Optimized Implementation

In this architecture, there are four S-boxes in the substitution layer as shown in Fig. 4. The diffusion layer is constructed by employing a single 128-bit layer and is implemented using CA. Figure 5 shows a 128-bit diffusion layer using an 8-bit maximum length CA. In Fig. 5, sixteen output bytes $y_1, y_2, ...y_{16}$ are computed by running CA-T, CA-T^2, ... CA-T^{16}, respectively, for 16 times, while sequentially feeding 16 input bytes (starting from x_1 up to x_{16}). The CA-based implementation of the diffusion layer has latency of 16-clock cycles. Figure 6 shows the internal architecture of CA-T, which represents an 8-bit CA having characteristic matrix T. The matrix T is given in Sect. 2. In Fig. 6, there are XOR gates in two levels. Two inputs XOR gates in the first level are used to add previous state output with input. XOR gates in the second level are used to connect neighboring cells. The internal architecture of CA-T^i is obtained by configuring the second level XOR gates according to T^i. Also there are 16 *Nmix* modules in each round. In this case, latency is high, but the implementation requires smaller amount of hardware compared to Latency Optimized Implementation. Table 2 shows the synthesis result of the architecture shown in Fig. 4. In Table 2, resources required for key scheduling algorithm is not considered. It is observed from Table 2 that the proposed architecture has a

Table 3 Dependence test results

Number of rounds	Average number of output bit changes	Degree of completeness	Degree of avalanche effect	Degree of strict avalanche effect
$1+\frac{1}{3}$	64.26	1.0	0.9947	0.9585
$2+\frac{1}{3}$	63.99	1.0	0.9989	0.9888
$3+\frac{1}{3}$	63.99	1.0	0.9990	0.9887
$4+\frac{1}{3}$	64.00	1.0	0.9990	0.9886
$5+\frac{1}{3}$	64.01	1.0	0.9991	0.9887
$6+\frac{1}{3}$	63.99	1.0	0.9991	0.9887
$7+\frac{1}{3}$	63.99	1.0	0.9990	0.9887
$8+\frac{1}{3}$	64.00	1.0	0.9991	0.9887

small area overhead. In the following section, the diffusion properties of full round and reduced round versions of HDNM8 are examined.

5 Dependence Tests

In this section, the average number of output bits will change when single input bit changes, the degree of completeness, the degree of avalanche effect, and the degree of strict avalanche criterion are determined for full round and reduced round versions of HDNM8. The terms degree of completeness (d_c), degree of avalanche effect (d_a), and degree of strict avalanche criterion (d_{sa}) have been defined in New European Schemes for Signatures, Integrity, and Encryption (NESSIE) report IST-1999-12324 [19]. For a function f, it is desirable to have $d_c = 1$, $d_{sa} \approx 1$ and $d_a \approx 1$. However, bad results would have been a strong indication of weakness in the algorithm. Examination on all four parameters is carried out by varying number of rounds from 1 round to 8 rounds. Also in every computation the last round key mixing ('$\frac{1}{3}$') is kept fixed. Here, dependence and distance matrices are computed by considering 5,000 randomly chosen inputs and single randomly chosen 128-bit key. Table 3 shows the variation of average number of output bit change, degree of completeness, degree of avalanche effect, degree of strict avalanche effect with the variation of number of rounds in HDNM8. From Table 3, it is observed that all four parameters approximately attain their desired values after one round.

Table 4 Comparison of EDP(DP) versus frequency of occurrence of DDT for S-box

$S^{\star}$-Box		AES S-Box	
EDP value	Number of occurrence	DP value	Number of occurrence
EDP = 0	877	0	33,150
$0.0001 \leq \text{EDP} \leq 0.001$	271		
$0.001 \leq \text{EDP} \leq 0.002$	1,806		
$0.002 \leq \text{EDP} \leq 0.003$	8,079		
$0.003 \leq \text{EDP} \leq 0.004$	2,5282		
$0.004 \leq \text{EDP} \leq 0.005$	2,2291		
$0.005 \leq \text{EDP} \leq 0.006$	5,379		
$0.006 \leq \text{EDP} \leq 0.007$	1,002		
$0.007 \leq \text{EDP} \leq 0.008$	485	0.0078	32,130
$0.008 \leq \text{EDP} \leq 0.009$	39		
$0.009 \leq \text{EDP} \leq 0.01$	10		
$0.01 \leq \text{EDP} \leq 0.013$	13		
EDP = 0.0156	1	0.0156	255
EDP = 1.0	1	1.0	1

6 Security Analysis

The robustness of the block cipher HDNM8 against several existing cryptographic attacks has been studied in this section. It is observed that it is secure against linear cryptanalysis, differential cryptanalysis, higher order differential attack, interpolation attack, algebraic attack, and integral attack.

6.1 Expected Differential Probability Value for Characteristic

The probability of the differential is a more accurate measure for the success rate of a differential attack. But in general, the probability of differential over multiple rounds of an SPN-type block cipher is difficult to compute. Therefore, in this paper the upper bound of expected differential probability (EDP) for characteristic is computed. The differential probability $\text{DP}_f(a, b)$ of a differential (a, b) with respect to $f(x)$ is defined in [12], and the expression is as follows:

$$\text{DP}_f(a, b) = 2^{-n} \#\{x \in F_2^n | f(x + a) = f(x) + b\} \tag{3}$$

If f is a function of parameter k, then expected differential probability (EDP) of a differential (a, b) is defined as the mean value of parameterized differential probability $\text{DP}[k](a, b)$ and expressed as

$$\mathrm{EDP}(a, b) = 2^{-|\kappa|} \Sigma_{k \in \kappa} \mathrm{DP}[k](a, b) \tag{4}$$

here k is assumed to be a uniformly distributed random variable taking values in κ, set of all keys of size $|\kappa|$ bits. In order to compute the maximum differential probability for single round, the AES S-box (S) is combined with the 8-bit *Nmix* function. The keyed substitution box is represented by $S^\star$-box. For $S^\star$-box, it is obtained that $\mathrm{EDP}(0, 0) = 1$ and $\mathrm{EDP}(a, 0) = \mathrm{EDP}(0, b) = 0$ for all $a, b \in F_2{}^n$. It is observed that maximum expected differential probability of $S^\star$ is 2^{-6} for nonzero input difference, i.e., $a \neq 0$. Table 4 describes the range of expected differential probabilities and corresponding frequency of occurrence in the difference distribution table of $S^\star$-box. It is observed from Table 4 that EDP values of $S^\star$-box are more evenly distributed compared to AES S-box. The 128-bit diffusion layer has branch number 17. There are at least 34 active S-boxes in the 4-rounds cipher. It assumed that the round keys are independent and random. Therefore, the best EDP value for characteristics of the 128-bit 2-round cipher is bounded by $(2^{-6})^{17} = 2^{-102}$. For a 4-round cipher the value is $(2^{-102})^2 = 2^{-204}$. Therefore, classical differential attack is not possible after four rounds.

6.2 Maximum Expected Probability for Linear Characteristic

According to Hong et al., the linear probability [13] of an S-box S_i is defined as follows:

$$\mathrm{LP}^{S_i}(\Gamma x \rightarrow \Gamma y) = \left(\frac{\#\{x \in Z_2^m | \Gamma x.x = \Gamma y.S_i(x)\}}{2^{m-1}} - 1 \right)^2$$
$$\mathrm{LP}^{S_i}_{\max} = \max_{\Gamma x, \Gamma y \neq 0} \mathrm{LP}^{S_i}(\Gamma x \rightarrow \Gamma y) \tag{5}$$

where Γx and Γy are input and output mask, respectively, and $1 \leq i \leq n$. It has been shown in [13] that the probability for each linear characteristic of Substitution, Diffusion, and Substitution (SDS) function is bounded by q^n, where $q = \mathrm{LP}^{S_i}_{\max}$ is the maximum linear probability of S-boxes in the substitution layer and $n + 1$ is a lower bound for the number of active S-boxes in two consecutive rounds of a linear approximation. For the computation of the expected linear probability (ELP), a modified S-box ($S^\star$) is considered, which is the combination of nmix and AES S-box. The linear probability of ($S^\star$) is computed for each 8-bit key and then the expected linear probability is computed by taking the average over all 8-bit keys. In case of ($S^\star$)-box, the value of $\mathrm{ELP}^S_{\max}$ is also 2^{-6}. But it is found that the value of $\mathrm{ELP}^S_{\max} = 2^{-6}$ appears 1,275 times in the linear approximation table of the AES S-box, while it appears only 5 times in case of $S^\star$-box. Hence, the expected probability for linear characteristic of $S^\star D S^\star$ function is bounded by $(2^{-6})^{16} = 2^{-96}$. So the maximum probability for linear characteristic of the four rounds cipher is

$(2^{-96})^2 = 2^{-192}$. Hence, four rounds of proposed construction is sufficient to resist classical linear attack.

6.3 Higher Order Differential Cryptanalysis

The S-Boxes (AES S-Box) of HDNM8 have algebraic degree 7. Each output bit of the S-Box can be regarded as a Boolean function with 8 input variables. Key mixing function *Nmix* has algebraic degree two except the first bit, which has algebraic degree one. Therefore, after one round algebraic degree of any intermediate bit becomes at least 13. Hence, after two rounds the algebraic degree of any intermediate bit becomes 13^2. Thus, the number of plaintexts needed for higher order differential attack using a two rounds distinguishers is greater than 2^{128}. So, the proposed cipher is secure against higher order differential attack.

6.4 Interpolation Attack

In interpolation attack, plaintext and ciphertext pairs are used to construct the relation between the input and output of the cipher. If the constructed polynomials have small algebraic degree, then small number of plaintext and ciphertext pairs are required to solve the coefficients of the polynomial. It is expected that interpolation attack is not possible just after few rounds because of the complicated expression of the S-Boxes, together with the effect of diffusion layer and nonlinear round key mixing used in HDNM8.

6.5 Algebraic Attack

The AES S-Box has been studied by Courtois and Pieprzyk [8]. They observed that there are 39 quadratic equations over F_2 of probability one and one additional quadratic equation of probability $\frac{255}{256}$ exit between input and output of the S-Box. Hence, this is also the case for the S-box of the proposed cipher HDNM8. It is an 8-round 128-bit cipher, therefore total $16 \times 8 + 4 \times 8 = 160$ number of S-Boxes are used for one encryption. Using these 40 equations for each S-Box, one can construct 6,400 quadratic equations of 1,280 unknown variables. The solution of these equations can be used to derive the value of the secret key used in the encryption. To the best of our knowledge, the time complexity to solve the above set of quadratic equations is unknown. Therefore, algebraic attack cannot be faster than exhaustive key search.

6.6 Integral Cryptanalysis

Integral cryptanalytic attack is particularly applicable to SPN-type block ciphers, which have strong word-like structure. Integral cryptanalysis uses sets or even multisets of chosen plaintexts of which part is held constant and another part varies through all possibilities. XOR sum of set of input plaintexts and XOR sum of corresponding ciphertexts are used to derive the secret key. In HDNM8, the operations in substitution layer is bytewise operation. But keymixing and diffusion layer are bitwise operation. Therefore, development and propagation of bytewise structure is disrupted by the bitwise operation in round key mixing and diffusion.

7 Conclusions

In this Chapter, a CA-based MDS code is employed to construct the diffusion layer of a 128-bit block cipher. A new block cipher called 'HDNM8' is introduced employing the the diffusion layer, the nonlinear key mixing function, and the AES S-boxes. The robustness of the proposed cipher against several existing cryptographic attacks has been shown. Also, HDNM8 has been implemented in hardware. Performance of HDNM8 on an FPGA platform is evaluated and compared with AES.

References

1. Barreto, P., Rijmen, V.: The Anubis block cipher. Submission to the NESSIE Project (2000a)
2. Barreto, P., Rijmen, V.: The Khazad legacy-level block cipher. Submission to the NESSIE Project (2000b)
3. Bhaumik, J., Roy Chowdhury, D.: Nmix: an ideal candidate for key mixing. In: Proceedings International Conference on Security and Cryptography, Italy, pp. 285–288 (2009)
4. Bhaumik, J., Roy Chowdhury, D.: CA-based diffusion layer for an SPN-type block cipher. In: Proceedings of the 17th International Workshop on Cellular Automata and Discrete Complex Systems Chile, pp. 243–251 (2011)
5. Biryukov, A.: Analysis of involutional ciphers Khazad and Anubis. In: Proceedings of the Fast Software Encryption, Sweden. LNCS, vol. 2887, pp. 45–53 (2003)
6. Biryukov, A.: The design of stream cipher LEX. In: Proceedings of the Selected areas in cryptography. Canada. LNCS, vol. 4356, pp. 67–75 (2007)
7. Canright, D.: A very compact S-box for AES. In: Prcoceedings of Cryptographic Hardware and Embedded Systems. UK. LNCS, vol. 3659, pp. 441–455 (2005)
8. Courtois, N.T., Pieprzyk, J.: Cryptanalysis of block ciphers with overdefined systems of equations. In: Proceedings of the ASIACRYPT, New Zealand. LNCS, vol. 2501, pp. 267–287 (2002)
9. Daemen, J.: Cipher and hash function design strategies based on linear and differential cryptanalysis. Doctoral Dissertation, K. U. Leuven (1995)
10. Daemen, J., Rijmen, V.: The Design of Rijndael-AES: The Advanced Encryption Standard. Springer, New York (2002)
11. Daemen, J., Rijmen, V.: The Pelican MAC function. In: Cryptology ePrint Archive. Report 2005/008. http://eprint.iacr.org/

12. Daemen, J., Lamberger, M., Pramstaller, N., Rijmen, V., Vercauteren, F.: Computatioal aspects of the expected differential probability of a 4-round AES and AES-like ciphers. J. Comput. **85**(1–2), 85–104 (2009)
13. Hong, S., Lee, S., Lim, J., Sung, J., Cheon, D., Cho, I.: Provable security against differential and linear cryptanalysis for the SPN structure. In: Proceedings of the Fast Software Encryption. LNCS, vol. 1978, pp. 273–283 (2000)
14. Junod, P., Vaudenay, S.: Perfect diffusion primitives for block ciphers building efficient MDS matrices. In: Proceedings of the Selected Areas in Cryptography. LNCS, vol. 3357, pp. 84–99 (2004)
15. Koo, B.W., Jang, H.S. Song, J.H.: Constructing and cryptanalysis of a 16×16 binary matrix as a diffusion layer. In: Proceedings of the WISA. LNCS, vol. 2908, pp. 489–503 (2003)
16. Koo, B.W., Jang, H.S. Song, J.H.: On constructing of a 32×32 binary matrix as a diffusion layer for a 256-bit block cipher. In: Proceedings of the International Conference on Information Security and Cryptology. LNCS, vol. 4296, pp. 51–64 (2006)
17. Minematsu, K., Tsunoo, Y.: Provable secure MACs from differentially-unifrom permutations and AES-based implementations. In: Proceedings of the Fast Software Encryption, Austria. LNCS, vol. 4047, pp. 226–241 (2006)
18. Nakahara Jr, J., Abrahao, E.: A New involutory MDS matrix for the AES. Int. J. Netw. Secur. **9**(2), 109–116 (2009)
19. Preneel, B., Bosselaers, A., Rijmen, V., Van Rompay, B., Granboulan, L., Stern, J., Murphy, S., Dichtl, M., Serf, P., Biham, E., Dunkelman, O., Furman, V., Koeune, F., Piret, G., Quisquater, J-J., Knudsen, L., Raddum, H.: Comments by the NESSIE Project on the AES Finalists (2000)
20. Rijmen, V., Daemen, J., Preneel, B., Bosselaers, A., De Win, E.: The cipher SHARK. In: Proceedings of the Fast Software Encryption. LNCS, vol. 1039, pp. 99–111 (1996)

Chapter 5
Frames and Erasures

Saliha Pehlivan

Abstract Frames have been useful in signal transmission due to the built in redundancy. In recent years, the erasure problem in data transmission has been the focus of considerable research in the case the error estimate is measured by operator (or matrix) norm. Sample results include the characterization of one-erasure optimal Parseval frames, the connection between two-erasure optimal Parseval frames and equiangular frames, and some characterization of optimal dual frames. If iterations are allowed in the reconstruction process of the signal vector, then spectral radius measurement for the error operators is more appropriate than the operator norm measurement. A complete characterization of spectrally one-uniform frames (i.e., one-erasure optimal frames with respect to the spectral radius measurement) in terms of the redundancy distribution of the frame is obtained. The characterization relies on the connection between spectrally optimal frames and the linear connectivity property of the frame. The linear connectivity property is equivalent to the intersection dependence property, and is also closely related to the concept of k-independent set.

1 Introduction

In the study of Hilbert spaces, an orthonormal basis, possessing some desirable properties, is one of the most important concepts. One such property is that each element in Hilbert space can be written uniquely as a linear combination of the elements in the basis. For instance, in the signal transmission, a signal is thought of as a vector in a Hilbert space that is represented as a linear combinations of orthogonal basis vectors. The signal is transmitted to a receiver by transmitting the sequence of coefficients that represents the signal. These coefficients can be computed by taking

S. Pehlivan (✉)
Department of Mathematics, University of Central Florida, Orlando, FL 32816, USA
e-mail: salihapehlivan@gmail.com

R. N. Mohapatra et al. (eds.), *Mathematics and Computing 2013*,
Springer Proceedings in Mathematics & Statistics 91,
DOI: 10.1007/978-81-322-1952-1_5, © Springer India 2014

some inner products. The receiver on the other side reconstructs the signal. However, if one of the coefficients is lost during the transmission, the receiver cannot recovers the signal. The orthogonality property of the basis is restrictive in this sense. This brings us the notion of frame that has redundancy so that if some pieces of information is lost, it is recovered with the other pieces that are received.

A vector in a Hilbert space can be represented by the elements of a frame but not necessarily uniquely as in the case of an orthonormal basis. Thus, frames are considered as a generalization of orthogonal basis. The redundancy property of frames makes it more robust than orthogonal basis in some applications such as signal processing, image processing, coding, and sampling. These applications have naturally led to the investigations of optimal frames or dual frames that yield better approximations to the original signals. Typically, there are two types of investigations on optimal dual frames: one of them is to find (characterize) and construct optimal frames among a class of frames. Examples of this kind include the known theory established for erasure optimal Parseval frames (i.e., frames that are erasure optimal in the class of all Parseval frames (c.f. [1–4, 6, 8, 10–12, 21]). The other kind is the investigation of optimal dual frames for a given frame. This case addresses the applications when a particular frame that models the nature of the application is preselected for encoding (decomposition of) the signal. In this case, the theory of optimal dual frames (for the purpose of better decoding) needs to be established (c.f. [13, 15–19]). When it comes to the terminology of optimal, we mean the reconstruction error is minimal with respect to some prescribed measurement.

2 Preliminaries

2.1 Frames in Hilbert Spaces

Let H be a Hilbert space with inner product $\langle\,,\,\rangle$, and norm $\|\cdot\|$. The formal definition of a frame which is valid in both finite and infinite dimensional Hilbert spaces is the following:

Definition 2.1 A collection $\{f_i\}_{i\in\mathbb{N}}$ of elements of a Hilbert space H is called a frame for H if there are positive constants $0 < A \leq B < \infty$ such that

$$A\,\|f\|^2 \leq \sum_{i\in\mathbb{N}} |\langle f, f_i\rangle|^2 \leq B\,\|f\|^2, \quad \text{for all } f \in H. \tag{1}$$

In the above definition, A and B are called lower and upper frame bounds, respectively.

A frame is called a tight frame if $A = B$, and if $A = B = 1$, it is called a Parseval frame. If the norm of frame vectors are equal, it is called a uniform frame and if additionally norm is one, it is called a unit norm frame.

Now let us see some frame examples on ℓ^2.

Example 2.1

(i) Standard orthonormal basis is a Parseval frame with $A = 1$.
(ii) $F = \{0, 0, 0, 0, 0, e_1, e_2, e_3, \ldots\}$ is a Parseval frame with $A = 1$.
(iii) $F = \{e_1, e_1, e_1, e_1, e_2, e_3, e_4, \ldots\}$ is a frame with bounds $A = 1$ and $B = 4$.
(iv) $F = \left\{e_1, \frac{1}{\sqrt{2}}e_2, \frac{1}{\sqrt{2}}e_2, \frac{1}{\sqrt{3}}e_3, \frac{1}{\sqrt{3}}e_3, \frac{1}{\sqrt{3}}e_3, \ldots\right\}$ is a Parseval frame.

The definition given in (1) is true for both finite and infinite dimensional Hilbert spaces. However, there is an alternative definition to frames in finite dimensional Hilbert spaces.

Theorem 2.1 *A family of elements $\{f_i\}_{i=1}^N$ in a finite dimensional Hilbert space H is a frame for H if and only if $\{f_i\}_{i=1}^N$ spans H; i.e., span$\{f_i\}_{i=1}^N = H$.*

Proof Assume that $H = \text{span}\{f_i\}_{i=1}^N$. We can find nonzero $h \in H$ with $\|h\| = 1$ such that

$$A = \sum_{i=1}^{N} |\langle h, f_i\rangle|^2 = \min\left\{\sum_{i=1}^{N} |\langle f, f_i\rangle|^2 \,:\, f \in H\,, \|f\| = 1\right\}, \tag{2}$$

where $\sum_i |\langle f, f_i\rangle|^2$ is a continuous function of f. We see that $A > 0$ and

$$\sum_{i=1}^{N} |\langle f, f_i\rangle|^2 = \sum_{i=1}^{N} \left|\left\langle \frac{f}{\|f\|}, f_i\right\rangle\right|^2 \|f\|^2 \geq A\,\|f\|^2\,. \tag{3}$$

Note that by Cauchy-Schwarz' inequality, we have

$$\sum_{i=1}^{N} |\langle f, f_i\rangle|^2 \leq \sum_{i=1}^{N} \|f_i\|^2\,\|f\|^2\,, \tag{4}$$

and since the sequence of vectors $\{f_i\}_i^N$ is finite, $B = \sum_{i=1}^{N} \|f_i\|^2 < \infty$. Hence, $\{f_i\}_i^N$ is a frame for H.

For the other direction, assume that F is a frame and $\{f_i\}_{i=1}^N$ does not span H. Then there exists a vector $f \in M^\perp$ where $M = \text{span}\{f_i\}_{i=1}^N$. Note that f is orthogonal to each f_i. Thus, $\sum_{i=1}^N |\langle f, f_i\rangle| = 0$. This implies that the lower frame bound is 0, which contradict the fact that F is a frame.

Note here that, particularly, this definition implies that every basis for a Hilbert space H is a frame for H. Moreover, a finite collection of vectors $\{f_i\}_i^N$ is a frame for its span, span$\{f_i\}_i^N$.

Propsition 2.1 *Let $\{f_i\}_{i=1}^N$ be a frame with a lower and upper frame bounds A and B, respectively. Then, $\|f_i\|^2 \leq B$ for all $i = 1, \ldots, N$. If $\|f_i\|^2 = B$ for all i, then f_i is orthogonal to every f_j for $j \neq i$. Moreover, if $\|f_i\|^2 < A$, then $f_i \in$ span$\{f_j\}_{j\neq i}^N$.*

Proof Let $\{f_i\}_{i=1}^N$ be a frame with bounds A and B. Then from the frame definition, for every $j \in \{1, \ldots, N\}$, we have $B \left\| f_j \right\|^2 \geq \sum_{i=1}^N \left| \langle f_j, f_i \rangle \right|^2 \geq \left| \langle f_j, f_j \rangle \right|^2 = \left\| f_j \right\|^4$. Thus, $\| f_i \|^2 \leq B$.

For the second part of the proposition assume that $\| f_i \|^2 = B$, then, from the definition of frame, $B \left\| f_j \right\|^2 \geq \sum_{i=1}^N \left| \langle f_j, f_i \rangle \right|^2 = \left| \langle f_j, f_j \rangle \right|^2 + \sum_{\substack{i=1 \\ i \neq j}}^N \left| \langle f_j, f_i \rangle \right|^2 = B^2 + \sum_{\substack{i=1 \\ i \neq j}}^N \left| \langle f_j, f_i \rangle \right|^2$, and, this implies that $\sum_{\substack{i=1 \\ i \neq j}}^N \left| \langle f_j, f_i \rangle \right|^2 \leq 0$. Therefore, $\langle f_j, f_i \rangle = 0$ for all $i \neq j$.

To show the last part of the proposition, suppose $\| f_i \|^2 < A$ for all i, and assume for a contradiction that there exist $j \in \{1, \ldots, N\}$ such that f_j is not in the span of $\{f_i\}_{i \neq j}$, in other words, $\langle f_i, f_j \rangle = 0$ for every $i \neq j$. Then, from the definition of frame, we have $A \left\| f_j \right\|^2 \leq \sum_{i=1}^N \left| \langle f_j, f_i \rangle \right|^2 = \left| \langle f_j, f_j \rangle \right|^2 + \sum_{\substack{i=1 \\ i \neq j}}^N \left| \langle f_j, f_i \rangle \right|^2 = \left\| f_j \right\|^4$, that is, $\left\| f_j \right\|^2 \geq A$. This contradicts with the assumption. Hence, $f_i \in \text{span}\{f_j\}_{j \neq i}^N$ for all $i \in \{1, \ldots, N\}$.

As particular cases of the proposition, we state the following two corollaries:

Corollary 2.1 *Let $\{f_i\}_{i=1}^N$ be a tight frame with frame bound A. Then, $\| f_i \|^2 \leq A$ for all $i = 1, \ldots, N$, and the inequality holds if and only if f_i is orthogonal to every f_j for $j \neq i$.*

Proof It is enough to show that if f_i is orthogonal to every f_j for $j \neq i$, then $\| f_i \|^2 = A$, the rest follows from the proof of the above proposition. In fact, assume that f_i is orthogonal to every f_j for $j \neq i$. Then, by the last part of the proposition, we have $\| f_i \|^2 \geq A$; moreover, we have $\| f_i \|^2 \leq A$ from the first part of the proposition. Thus, $\| f_i \|^2 = A$ for all $i \in \{1, \ldots, N\}$.

Corollary 2.2 *Let $\{f_i\}_{i=1}^N$ be a Parseval frame. Then, $\| f_i \|^2 \leq 1$ for all $i = 1, \ldots, N$, and the inequality holds if and only if f_i is orthogonal to every f_j for $j \neq i$.*

Propsition 2.2 *If one of the vectors f_j of a Parseval frame $\{f_i\}_{i=1}^N$ is removed, then the family of the vectors $\{f_i\}_{i \neq j}$ is either a frame or an incomplete set.*

Proof By Corollary 2.2, the norm of vectors of a Parseval frame is either one or less than one. If $\left\| f_j \right\| = 1$, then f_i is orthogonal to $\text{span}\{f_i\}_{i \neq j}$; thus, $\{f_i\}_{i \neq j}$ ceases to be a frame. On the other hand, when $\left\| f_j \right\| < 1$, $f_j \in \text{span}\{f_i\}_{i \neq j}$. Hence, $\{f_i\}_{i \neq j}$ spans the Hilbert space H, thus, $\{f_i\}_{i \neq j}$ is a frame for H.

2.2 Frame Operators

We have a reconstruction formula for frames similar to the reconstruction formula for orthonormal basis. To derive the formula, we first define the analysis and synthesis operators.

Definition 2.2 Let $\{f_i\}_{i\in I}$ be a frame for a Hilbert space H and $\{e_i\}_{i\in I}$ be the standard orthonormal basis. The analysis operator $\Theta : H \to \ell_2(I)$ is defined to be

$$\Theta(f) = \sum_{i\in I} \langle f, f_i\rangle e_i \quad \text{for all f } \in H. \tag{5}$$

The adjoint of the analysis operator is called the synthesis operator that is given by $\Theta^*(e_i) = f_i$. By composing synthesis operator Θ^* with its adjoint operator Θ, we get the frame operator S which is given by $Sf = \Theta^*\Theta f = \sum_{i\in I}\langle f, f_i\rangle f_i$.

Remark 2.1 That S is self-adjoint and positive operator follows from $S^* = (\Theta^*\Theta)^* =) \Theta^*\Theta = S$, and $\langle Sf, f\rangle = \left\langle \sum_{i\in I}\langle f, f_i\rangle f_i, f\right\rangle = \sum_{i\in I}\langle f, f_i\rangle\langle f_i, f\rangle = \sum_{i\in I}|\langle f, f_i\rangle|^2$, respectively.

Remark 2.2 By the definition (1) of frame and Remark 2.1, we have $A\|f\|^2 \le \langle Sf, f\rangle \le B\|f\|^2$ for all $f \in H$, or, $AI \le S \le BI$. If $\{f_i\}_{i\in I}$ is a tight frame; i.e., $A = B$, then $S = AI$, and if $\{f_i\}_{i\in I}$ is a Parseval frame; i.e., $A = B = 1$, then $S = I$.

Next, we give some properties of analysis operator.

Propsition 2.3 *Let Θ_{Tf} be an analysis operator for the set of vectors $\{Tf_i\}_{i\in I}$ where $T : H \to H$ is a linear operator. Then, $\Theta_{Tf}h = \Theta_f T^*h$.*

Proof Let $h \in H$. By the definition of analysis operator we have, $\Theta_{Tf}h = \sum_{i\in I}\langle h, Tf_i\rangle e_i = \sum_{i\in I}\langle T^*h, f_i\rangle e_i = \Theta_f T^*h$.

Propsition 2.4 *Let $\Theta_{\alpha f}$ be an analysis operator of the set of vectors $\{\alpha f_i\}_{i\in I}$ where α is a scalar. Then, $\Theta_{\alpha f} = \bar{\alpha}\Theta_f$.*

Proof Letting $T = \alpha I$ in Proposition 2.3, the result follows.

Following couple propositions show the relationship between frames and its corresponding analysis and frame operators, respectively. In other words, frames can be characterized by analysis and frame operators.

Propsition 2.5 *Let H be a finite, say n, dimensional Hilbert space. Then, $\{f_i\}_{i=1}^N$ is a frame for H if and only if the analysis operator Θ is one-to-one.*

Proof First, suppose that $\{f_i\}_{i=1}^N$ is a frame for H. And assume that $\Theta f = 0$ for some $f \in H$. Then, $\sum_{i=1}^N \langle f, f_i\rangle e_i = 0$, which means that $\langle f, f_i\rangle = 0$ for all $i = 1, \ldots, N$ because $\{e_i\}_{i=1}^n$ is the standard orthonormal basis. On the other hand, since $\{f_i\}_{i=1}^N$ is a frame, we can write every $f \in H$ as a linear combination of frame vectors such that $f = \sum_{i=1}^N c_i f_i$ for some constants c_i. Then $\langle f, f\rangle = \left\langle f, \sum_{i=1}^N c_i f_i\right\rangle = \sum_{i=1}^N \bar{c_i}\langle f, f_i\rangle = 0$. Hence $f = 0$, and f is one-to-one.

Now, suppose that Θ is one-to-one, and assume for a contradiction that $\{f_i\}_{i=1}^N$ is not a frame for H; i.e., $\{f_i\}_{i=1}^N$ does not span H. Then there exist nonzero $f \in H$ such that $\langle f, f_i\rangle = 0$ for all $i = 1, \ldots, N$. Thus, we have $\Theta f = \sum_{i=1}^{N} \langle f, f_i\rangle e_i = 0$. This contradicts with Θ being one-to-one. Hence, $\{f_i\}_{i=1}^N$ is a frame for H.

Propsition 2.6 *Let H be a finite dimensional Hilbert space. Then, $\{f_i\}_{i=1}^N$ is a frame for H if and only if the frame operator S is invertible.*

Proof First assume that $\{f_i\}_{i=1}^N$ is a frame for H. To show that S is one-to-one, assume further that $Sf = 0$. Then by Remark 2.1, we have $\sum_{i=1}^{N} |\langle f, f_i\rangle| = 0$. This implies that $\|f\| = 0$ by the definition of frame. Hence, f is one-to-one. Now, to show that S is onto, assume that there exist nonzero element f in the orthogonal complement of the range of S. Then $\langle Sg, f\rangle = 0$ for all $g \in H$. Thus, $\langle Sf, f\rangle = 0$. Again, from Remark 2.1 and the definition of frame, $f = 0$. Therefore, range of S is the entire space H.

To show the opposite direction, assume that S is invertible with the inverse operator S^{-1}. Then, for each $f \in H$, $f = SS^{-1}f = \sum_{i=1}^{N} \langle S^{-1}f, f_i\rangle f_i = \sum_{i=1}^{N} \langle f, S^{-1}f_i\rangle f_i$. This shows that $\{f_i\}_{i=1}^N$ spans H and, therefore, $\{f_i\}_{i=1}^N$ is a frame.

If the inverse S^{-1} of frame operator is applied to the frame vectors f_i for $i = 1, \ldots, N$, then the new collection of vectors $\{S^{-1}f_i\}_{i=1}^N$ is a frame and its frame bounds are characterized by the frame bounds of $\{f_i\}_{i=1}^N$.

Propsition 2.7 *If $\{f_i\}_{i=1}^N$ is a frame for a finite dimensional H with corresponding frame operator S and frame bounds A and B, then $\{S^{-1}f_i\}_{i=1}^N$ is also a frame for H with lower and upper frame bounds B^{-1} and A^{-1}, respectively. Moreover, the frame operator for $\{S^{-1}f_i\}_{i=1}^N$ is S^{-1}.*

Proof Recall from Remark 2.2 that $AI \le S \le BI$. Now, applying S^{-1} to each side, we have $S^{-1}A \le S^{-1}S = I \Rightarrow S^{-1} \le A^{-1}I$, and $I = S^{-1}S \le S^{-1}B \Rightarrow S^{-1} \ge B^{-1}I$, which is $B^{-1}I \le S^{-1} \le A^{-1}I$, or,

$$B^{-1}\|f\|^2 = \langle B^{-1}f, f\rangle \le \langle S^{-1}f, f\rangle \le \langle A^{-1}f, f\rangle = A^{-1}\|f\|^2 \quad \text{for all f} \in H. \tag{6}$$

On the other hand, $S^{-1}f = S^{-1}SS^{-1}f = S^{-1}\sum_{i=1}^{N}\langle S^{-1}f, f_i\rangle f_i = \sum_{i=1}^{N} \langle f, S^{-1}f_i\rangle S^{-1}f_i$. (Note that this shows that S^{-1} is the frame operator for $\{S^{-1}f_i\}_{i=1}^N$.) This implies that

$$\left\langle S^{-1}f, f\right\rangle = \left\langle \sum_{i=1}^{N} \langle f, S^{-1}f_i\rangle S^{-1}f_i, f\right\rangle = \sum_{i=1}^{N} \left|\langle f, S^{-1}f_i\rangle\right|^2. \tag{7}$$

From (6) and (7), we have $B^{-1}\|f\|^2 \leq \sum_{i=1}^{N} \left|\langle f, S^{-1}f_i\rangle\right|^2 \leq A^{-1}\|f\|^2$ for all f $\in H$. Therefore, $\{S^{-1}f_i\}_{i=1}^{N}$ is a frame with lower and upper frame bounds B^{-1} and A^{-1}, respectively.

Now, we shall show the relationship between frame bounds and the eigenvalues of frame operators.

Propsition 2.8 *Let $\{f_i\}_{i=1}^{N}$ be a frame with frame operator S for a finite dimensional H. Then the smallest and largest eigenvalues of S are a lower and an upper frame bounds, respectively, for $\{f_i\}_{i=1}^{N}$.*

Proof Assume that $\{f_i\}_{i=1}^{N}$ is a frame for H with frame operator S and n is the dimension of H. For any $f \in H$, we can write $f = \sum_{i=1}^{n}\langle f, e_i\rangle e_i$, where $\{e_i\}_{i=1}^{n}$ is the standard orthonormal basis. Then $Sf = \sum_{i=1}^{n}\langle f, e_i\rangle Se_i = \sum_{i=1}^{n}\lambda_i\langle f, e_i\rangle e_i$, where $\{\lambda_i\}_{i=1}^{n}$ are the eigenvalues for S corresponding to the eigenvectors $\{e_i\}_{i=1}^{n}$. And,

$$\langle Sf, f\rangle = \left\langle \sum_{i=1}^{n}\lambda_i\langle f, e_i\rangle e_i, f\right\rangle = \sum_{i=1}^{n}\lambda_i\langle f, e_i\rangle\langle e_i, f\rangle = \sum_{i=1}^{n}\lambda_i\left|\langle f, e_i\rangle\right|^2. \tag{8}$$

Note that in Remark 2.1, it is shown that $\langle Sf, f\rangle = \sum_{i=1}^{N}|\langle f, f_i\rangle|^2$, and we also have

$$\|f\|^2 = \langle f, f\rangle = \left\langle \sum_{i=1}^{n}\langle f, e_i\rangle e_i, f\right\rangle = \sum_{i=1}^{n}|\langle f, e_i\rangle|^2 \tag{9}$$

Thus, by (8) and (9),

$$\begin{aligned}\lambda_{\min}\|f\|^2 = \lambda_{\min}\sum_{i=1}^{n}|\langle f, e_i\rangle|^2 &\leq \sum_{i=1}^{n}\lambda_i\left|\langle f, e_i\rangle\right|^2 \\ = \sum_{i=1}^{N}|\langle f, f_i\rangle|^2 \leq \lambda_{\max}\sum_{i=1}^{n}|\langle f, e_i\rangle|^2 &= \lambda_{\max}\|f\|^2.\end{aligned}$$

2.3 Parseval Frames

In this section, we shall show that Parseval frames have the reconstruction property of orthonormal bases. For the rest of the paper, we assume that H is finite dimensional Hilbert space. Let us first make the following observation.

Remark 2.3 If the collection of vectors $\{f_i\}_{i=1}^{N}$ is a Parseval frame then the corresponding analysis operator Θ is an isometry; that is $\langle\Theta f, \Theta f\rangle = \langle\Theta^*\Theta f, f\rangle = \langle Sf, f\rangle = \sum_{i=1}^{N}|\langle f, f_i\rangle|^2 = \|f\|^2 = \langle f, f\rangle$ which follows from Remark 2.1 and the definition of Parseval frame ($A = B = 1$). Furthermore, Θ preserves inner products; i.e., $\langle\Theta f, \Theta g\rangle = \langle f, g\rangle$ for every $f, g \in H$.

Theorem 2.2 *A family of vectors $\{f_i\}_{i=1}^N$ is a Parseval frame if and only if it satisfies the reconstruction property, that is, for every $f \in H$, $f = \sum_{i=1}^N \langle f, f_i\rangle f_i$.*

Proof Assume that $\{f_i\}_{i=1}^N$ is a Parseval frame, and let $\{e_i\}_{i=1}^N$ be the standard orthonormal basis for $\mathbb{C}^N$ and $\{v_i\}_{i=1}^n$ be an orthonormal basis for H. Then, from the reconstruction property of orthonormal basis and Remark 2.3, we have

$$\begin{aligned} f &= \sum_{i=1}^n \langle f, v_i\rangle v_i = \sum_{i=1}^n \langle \Theta f, \Theta v_i\rangle v_i = \sum_{i=1}^n \Big\langle \sum_{j=1}^N \langle f, f_j\rangle e_j, \sum_{k=1}^N \langle v_i, f_k\rangle e_k \Big\rangle v_i \\ &= \sum_{i=1}^n \sum_{j=1}^N \sum_{k=1}^N \langle f, f_j\rangle \overline{\langle v_i, f_k\rangle} \langle e_j, e_k\rangle v_i = \sum_{i=1}^n \sum_{j=1}^N \langle f, f_j\rangle \overline{\langle v_i, f_j\rangle} v_i \\ &= \sum_{j=1}^N \langle f, f_j\rangle \sum_{i=1}^n \langle f_j, v_i\rangle v_i = \sum_{j=1}^N \langle f, f_j\rangle f_j. \end{aligned}$$

Thus, $\{f_i\}_{i=1}^N$ satisfies reconstruction property.

For the converse, assume that $f = \sum_{i=1}^N \langle f, f_i\rangle f_i$ holds true for the family of vectors $\{f_i\}_{i=1}^N$. Then $\|f\|^2 = \langle f, f\rangle = \left\langle f, \sum_{i=1}^N \langle f, f_i\rangle f_i \right\rangle = \sum_{i=1}^N \overline{\langle f, f_i\rangle}\langle f, f_i\rangle = \sum_{i=1}^N |\langle f, f_i\rangle|^2$. Therefore, $\{f_i\}_{i=1}^N$ is a Parseval frame.

Propsition 2.9 *If the collection of vectors $\{f_i\}_{i=1}^N$ in H is a frame for H with frame operator S, then $\{S^{-\frac{1}{2}} f_i\}_{i=1}^N$ is a Parseval frame for H.*

Note 1 The frame operator S being a positive invertible operator has a positive square root operator $S^{\frac{1}{2}}$. Similarly, since S^{-1} is positive operator, there is a corresponding positive square root operator $S^{-\frac{1}{2}}$. Both $S^{\frac{1}{2}}$ and $S^{-\frac{1}{2}}$ are self-adjoint operators.

Proof Let $\{f_i\}_{i=1}^N$ be a frame for H with frame operator S. Then, from Note 1, we have $f = S^{-\frac{1}{2}} S S^{-\frac{1}{2}} f = S^{-\frac{1}{2}} \sum_{i=1}^N \langle S^{-\frac{1}{2}} f, f_i\rangle f_i = \sum_{i=1}^N \langle S^{-\frac{1}{2}} f, f_i\rangle S^{-\frac{1}{2}} f_i = \sum_{i=1}^N \langle f, S^{-\frac{1}{2}} f_i\rangle S^{-\frac{1}{2}} f_i$, which means that $\left\{S^{-\frac{1}{2}} f_i\right\}_{i=1}^N$ satisfies reconstruction formula; hence, $\left\{S^{-\frac{1}{2}} f_i\right\}_{i=1}^N$ is a Parseval frame for H.

2.4 Dual Frames

For every frame, we have a general reconstruction formula similar to the reconstruction formula for Parseval frames. To define reconstruction formula, we need a new set of vectors called dual frames.

Definition 2.3 Let $\{f_i\}_{i=1}^N$ be a frame for a Hilbert space H. A set of vectors $\{g_i\}_{i=1}^N$ which satisfies the following formula

$$f = \sum_{i=1}^{N} \langle f, g_i \rangle f_i = \sum_{i=1}^{N} \langle f, f_i \rangle g_i, \quad \text{for all } f \in H \tag{10}$$

is called a dual frame for $\{f_i\}_{i=1}^N$. The set of vectors $\{S^{-1} f_i\}_{i=1}^N$ is a dual frame for $\{f_i\}_{i=1}^N$, and is called standard or canonical dual frame. If $\{g_i\}_{i=1}^N$ is not a standard dual, it is called an alternate dual frame.

Propsition 2.10 *Let $F = \{f_i\}_{i=1}^N$ be a frame. Then $\{S^{-1} f_i\}_{i=1}^N$ is a dual frame for F.*

Proof Recall that the frame operator S for a frame $\{f_i\}_{i=1}^N$ is given by

$$Sf = \sum_{i=1}^{N} \langle f, f_i \rangle f_i, \quad \text{for all } f \in H. \tag{11}$$

Since S is a positive and invertible operator, we can substitute S^{-1} for f in Eq. (11), and we get the reconstruction formula

$$f = S(S^{-1} f) = \sum_{i=1}^{N} \langle S^{-1} f, f_i \rangle f_i = \sum_{i=1}^{N} \langle f, S^{-1} f_i \rangle f_i. \tag{12}$$

using the fact that S^{-1} is self-adjoint. Similarly, if we apply S^{-1} to both sides of Eq. (11), we obtain the dual of reconstruction formula

$$f = S^{-1}(Sf) = S^{-1}\Big(\sum_{i=1}^{N} \langle f, f_i \rangle f_i \Big) = \sum_{i=1}^{N} \langle f, f_i \rangle S^{-1} f_i. \tag{13}$$

Thus, by (12) and (13), we conclude that $\{S^{-1} f_i\}_{i=1}^N$ is a dual frame for F.

Remark 2.4 Standard dual of a tight frame F is $A^{-1}F$. Indeed, using the fact that $S = AI$, the inverse of frame operator is $A^{-1}I$; thus, $S^{-1}F = A^{-1}F$. In particular, the standard dual of a Parseval frame F is itself because $S = I$ in Parseval case.

Remark 2.5 Standard dual of the frame $\{S^{-1} f_i\}_{i=1}^N$ is $\{f_i\}_{i=1}^N$ because of the fact that the frame operator for the frame $\{S^{-1} f_i\}_{i=1}^N$ is S^{-1}.

Definition 2.4 Let $F = \{f_i\}_{i=1}^N$ and $G = \{g_i\}_{i=1}^N$ be sequences in a Hilbert space H, and let Θ_F and Θ_G be the corresponding analysis operators for F and G, respectively. Then, if $\Theta_F \perp \Theta_G$, F and G are called orthogonal sequences. If these sequences F and G are frames, they are called orthogonal frames.

Propsition 2.11 *Let $F = \{f_i\}_{i=1}^N$ and $G = \{g_i\}_{i=1}^N$ be sequences in a Hilbert space H. Then F and G are orthogonal if and only if $\Theta_F^* \Theta_G = 0$, where Θ_F and Θ_G are the corresponding analysis operators for F and G, respectively.*

Proof Let F and G be sequences in H with analysis operators Θ_F and Θ_G, respectively. Then $\Theta_F^*\Theta_G = 0 \Leftrightarrow \langle \Theta_F f, \Theta_G g\rangle = \langle f, \Theta_F^*\Theta_G g\rangle = 0$, for all $f, g \in H \Leftrightarrow \Theta_F \perp \Theta_G$.

Now we shall show the relationship between standard dual and alternate dual by giving the characterization of duals.

Propsition 2.12 *Let $F = \{f_i\}_{i=1}^N$ be a frame with frame operator S. Then, $G = \{g_i\}_{i=1}^N$ is a dual frame of F if and only if there exists a sequence $H = \{h_i\}_{i=1}^N$ such that $\Theta_H^*\Theta_F = 0$ and $\{g_i\}_{i=1}^N = \{S^{-1}f_i + h_i\}_{i=1}^N$, where Θ_F and Θ_H are the corresponding analysis operators for F and H.*

Proof Assume that $G = \{g_i\}_{i=1}^N$ is a dual of $F = \{f_i\}_{i=1}^N$, and let $h_i = g_i - S^{-1}f_i$. Then $\sum_{i=1}^N \langle f, f_i\rangle h_i = \sum_{i=1}^N \langle f, f_i\rangle g_i - \sum_{i=1}^N \langle f, f_i\rangle S^{-1}f_i = f - f = 0$ This implies that, for every $f, h \in H$, $\left\langle \sum_{i=1}^N \langle f, f_i\rangle h_i, h\right\rangle = \sum_{i=1}^N \langle f, f_i\rangle\langle h_i, h\rangle = \langle \Theta_F f, \Theta_H h\rangle = \langle \Theta_H^*\Theta_F f, h\rangle = 0$. Therefore, $\Theta_H^*\Theta_F = 0$.

Conversely, assume that there exist a sequence $\{h_i\}_{i=1}^N$ such that $\{g_i\}_{i=1}^N = \{S^{-1}f_i + h_i\}_{i=1}^N$ with $\Theta_H^*\Theta_F = 0$. Then, for all $f, h \in H$, $\langle \Theta_H^*\Theta_F f, h\rangle = \langle \Theta_F f, \Theta_H h\rangle = \sum_{i=1}^N \langle f, f_i\rangle\langle h_i, h\rangle = 0$. This implies that $\sum_{i=1}^N \langle f, f_i\rangle h_i = 0$ for all f in H. Thus, $\sum_{i=1}^N \langle f, f_i\rangle g_i = \sum_{i=1}^N \langle f, f_i\rangle S^{-1}f_i + \sum_{i=1}^N \langle f, f_i\rangle h_i = f + 0 = f$, which implies that G is a dual of F.

2.5 Traces of Frame Operators

Theorem 2.3 *Let T be a linear operator on a Hilbert Space H, and n be the dimension of H. Assume that $k \geq n$ and $N \geq n$. If $\{v_i\}_{i=1}^k$ and $\{f_i\}_{i=1}^N$ are frames for H with corresponding dual frames $\{w_i\}_{i=1}^k$ and $\{g_i\}_{i=1}^N$, then*

$$\sum_{i=1}^k \langle T v_i, w_i\rangle = \sum_{i=1}^N \langle T f_i, g_i\rangle. \tag{14}$$

Proof

$$\begin{aligned}
\sum_{i=1}^k \langle T v_i, w_i\rangle &= \sum_{i=1}^k \left\langle \sum_{j=1}^N \langle T v_i, g_j\rangle f_j, w_i\right\rangle = \sum_{i=1}^k \sum_{j=1}^N \langle T v_i, g_j\rangle\langle f_j, w_i\rangle \\
&= \sum_{j=1}^N \sum_{i=1}^k \langle f_j, w_i\rangle\langle T v_i, g_j\rangle = \sum_{j=1}^N \left\langle \sum_{i=1}^k \langle f_j, w_i\rangle v_i, T^* g_j\right\rangle \\
&= \sum_{j=1}^N \langle f_j, T^* g_j\rangle = \sum_{j=1}^N \langle T f_j, g_j\rangle.
\end{aligned}$$

Corollary 2.3 *Let T be a linear operator and $\{f_i\}_{i=1}^N$ be a frame for H with dual frame $\{g_i\}_{i=1}^N$. Then $tr(T) = \sum_{i=1}^N \langle Tf_i, g_i\rangle$.*

Proof In Theorem 2.3, let $k = n$ and $\{v_i\}_{i=1}^n$ be the standard orthonormal basis; i.e., $\{e_i\}_{i=1}^n$. Since $tr(T) = \sum_{i=1}^{n} \langle Te_i, e_i\rangle$, the result follows from the Theorem.

Corollary 2.4 *Let $\{f_i\}_{i=1}^N$ be a frame of H with dual frame $\{g_i\}_{i=1}^N$. Then $n = \sum_{i=1}^N \langle f_i, g_i\rangle$.*

Proof In Corollary 2.3, let T be an Identity operator I_n. Then, the result is immediate.

Remark 2.6 As a special case of the above Corollary, for Parseval frames $\{f_i\}_{i=1}^N$, we have $n = \sum_{i=1}^N \langle f_i, f_i\rangle = \sum_{i=1}^N \|f_i\|^2$, that is, the dimension of Hilbert space H is the sum of the squares of the lengths of frame vectors.

Propsition 2.13 *If $\{f_i\}_{i=1}^N$ is a uniform Parseval frame, then $\|f_i\| = \sqrt{\frac{n}{N}}$ for all i, where n is the dimension of H.*

Proof Since the norm of vectors is uniform , for any $j \in \{1, \ldots, N\}$, we have $\|f_j\|^2 = \frac{1}{N}\sum_{i=1}^N \|f_i\|^2 = \frac{n}{N}$, where the last equality follows from Remark 2.6.

3 Erasures

3.1 The Erasure Problem

The property of frames that the number of vectors, N, greater than or equal to the dimension, n, of the Hilbert space has a great significance in applications. For instance, in coding theory, the information of a vector f is transmitted, or encoded, by the analysis operator $\Theta_F f$, that is, $\Theta_F f = \{\langle f, f_i\rangle\}_{i=1}^N$, where $\{f\}_{i=1}^N$ is a frame for a Hilbert space H. On the other side, the receiver reconstructs, or decodes, the vector f, by the help of synthesis operator, Θ_G^* of a dual $\{g_i\}_{i=1}^N$, $\Theta_G^*\Theta_F f$. If there is no erasure, the receiver is able to reconstruct f completely. If there is loss of data or any erasure, however, the receiver still may be able to reconstruct f perfectly with the help of redundancy property of frames, which is the quantity $\frac{N}{n}$.

To deal with the erasures, maximum errors for erasures are to be minimized. To minimize the maximal errors for erasures, two approaches are provided in [11] and [17]. One approach provided by Holmes and Paulsen in [11] is to select an optimal frame for erasures. On the other hand, second approach provided by Lopez and Han in [17] is to select optimal dual frames for erasures for a given frame. Second approach is motivated mainly, because of the limitations on optimal frames, to give more freedom to frames that are to be used in coding. To find optimal frame means to

find a best frame that minimizes the error on reconstructed vectors; however, to find an optimal dual frame for a given frame is to find a best dual frame that minimizes the error on reconstructed vectors.

To make the notion of optimal frames and optimal dual frames precise, let us first define the error operator E_Λ for erasures. Let D be an $N \times N$ diagonal matrix with m ones and $n - m$ zeros, and $\mathscr{D}_m$ be the set of all such diagonal matrices, D. For any frame pairs $F = \{f_i\}_{i=1}^N$ and $G = \{g_i\}_{i=1}^N$, where G is the dual frame of F, and Θ_F and Θ_G are the respective analysis operators for F and G, the error operator for m-erasure where $\Lambda = \{i_1, \ldots, i_m\}$ is defined by

$$E_\Lambda(f) = f - \sum_{i \notin \Lambda} \langle f, f_i \rangle g_i = \sum_{i \in \Lambda} \langle f, f_i \rangle g_i = \Theta_G^* D \Theta_F f, \tag{1}$$

and the maximum error when m-erasures occur is defined by $\max \left\{ \left\| \Theta_G^* D \Theta_F \right\| : D \in \mathscr{D}_m \right\}$, where $\|\cdot\|$ is a measurement for the error operator (it could be the usual matrix norm, Hilbert-Schmidt norm, or some other measurement). The goal is either to characterize the dual frame G that minimizes the maximum error if a frame F is preselected, or to characterize Parseval frames F such that $\max \left\{ \left\| \Theta_G^* D \Theta_F \right\| : D \in \mathscr{D}_m \right\}$ is minimal among all the Parseval frames. The similar setup can be used for other applications (e.g., optimal for sparsity, noise control). In the following sections, we will give precise definitions for optimal frames and optimal dual frames, and give some results.

3.2 Optimal Frames for Erasures

From now on, for a frame $F = \{f_i\}_{i=1}^N$ for a Hilbert space H of dimension n, we will call F an (N, n) frame, and we will let $\| \cdot \|$ be a matrix norm. Throughout this section, we let F be a Parseval frame.

A Parseval frame F' is called optimal frame for 1-erasure if it satisfies $\delta^1_{F'} = \min_F \max \left\{ \left\| \Theta_F^* D \Theta_F \right\| : D \in \mathscr{D}_1 \right\}$, and a Parseval frame F' is called optimal frame for any m-erasure if it is optimal for $(m-1)-$erasure and $\delta^m_{F'} = \min_F \max \left\{ \left\| \Theta_F^* D \Theta_F \right\| : D \in \mathscr{D}_m \right\}$. In other words, a Parseval frame that is optimal for $m-$erasures is optimal for m or less erasures.

One-erasure optimal Parseval frames are characterized in [11].

Propsition 3.1 *An (N, n) Parseval frame is 1-erasure optimal if and only if it is uniform. Moreover, minimum error, $\delta^1_{F'}$, is n/N.*

Definition 3.1 If F is an (N, n) uniform Parseval frame and $\|\Theta_F^* D \Theta_F\|$ is a constant for all D where D is a diagonal matrix with 2 ones and $N - 2$ zeros on the diagonal, and Θ_F^* and Θ_F are synthesis operator and analysis operator of F, respectively, then F is called 2-uniform Parseval frame.

The following Theorem provides an alternative definition for a 2-uniform Parseval frame that is given in [11].

Theorem 3.1 *Assume that F is a uniform (N, n) Parseval frame. Then, F is 2-uniform if and only if $\left|\langle f_i, f_j\rangle\right| = c$ is constant for all $i \neq j$ where*

$$c = \sqrt{\frac{n(N-n)}{N^2(N-1)}}. \tag{2}$$

The proof of the theorem in [11] implies that 2-uniform Parseval frames are 2-erasure optimal Parseval frames.

For a 2-uniform (N, n) Parseval frame, $\Theta_F \Theta_F^*$ can be written in the following way:

$$\Theta_F \Theta_F^* = \begin{bmatrix} \langle f_1, f_1\rangle & \langle f_2, f_1\rangle & \dots & \langle f_N, f_1\rangle \\ \langle f_1, f_2\rangle & \langle f_2, f_2\rangle & \dots & \langle f_N, f_2\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle f_1, f_N\rangle & \langle f_2, f_N\rangle & \dots & \langle f_N, f_N\rangle \end{bmatrix} = \begin{bmatrix} n/N & \pm c & \dots & \pm c \\ \pm c & n/N & \dots & \pm c \\ \vdots & \vdots & \ddots & \vdots \\ \pm c & \pm c & \dots & n/N \end{bmatrix}. \tag{3}$$

In other words, $\Theta_F \Theta_F^* = \frac{n}{N} I + cQ$ where $Q = (q_{ij})$ is a self-adjoint matrix with $q_{ii} = 0$ for all i and $\left|q_{ij}\right| = 1$ for all $i \neq j$.

Definition 3.2 Let F be a 2-uniform (N, n) Parseval frame. Then, the $(N \times N)$ matrix Q derived above is called signature matrix of F.

In [11], the characterization of 2-uniform Parseval frames is given in the following way:

Propsition 3.2 *Let Q be a signature matrix of a 2-uniform (N, n) Parseval frame F. Then $Q^2 = (N-1)I + \mu Q$, where $\mu = (N-2n)\sqrt{\frac{N-1}{n(N-n)}}$. Conversely, let Q be a signature matrix of the form $Q^2 = (N-1)I + \mu Q$, $\mu^2 \neq -4(N-1)$. Then, Q is a signature matrix of a 2-uniform (N, n) Parseval frame with $n = \frac{N}{2} - \frac{\mu N}{2\sqrt{4(N-1)+\mu^2}}$ and $\Theta_F \Theta_F^* = \frac{n}{N} I + cQ$.*

3.3 Optimal Dual Frames

Given a frame F, we search for a dual frame G of F, which makes the error of erasures minimum. Now, in the following subsections, we will look at the optimal dual frames with respect to matrix norm measurement and spectral radius measurement.

3.3.1 Optimality with Respect to Matrix Norm Measurement

Let a frame F be given. Then a dual frame G' for F is called optimal dual frame of F for 1-erasure if $\delta^{(1)}_{F,G'} = \min_G \max\left\{\left\|\Theta^*_G D\Theta_F\right\| : D \in \mathscr{D}_m\right\}$, and a dual frame G' for F is called optimal dual frame of F for any m-erasure if it is optimal for $(m-1)-$erasure and $\delta^{(m)}_{F,G'} = \min_G \max\left\{\left\|\Theta^*_G D\Theta_F\right\| : D \in \mathscr{D}_m\right\}$.

In [17], the condition in which the standard dual of a frame is the unique optimal dual frame for $m-$erasures is given.

Theorem 3.2 *Let $F = \{f_i\}_{i=1}^N$ be an (N, n) frame for a Hilbert space H. If $\|S^{-1} f_i\| \cdot \|f_i\|$ is constant for all i, then the standard dual is the unique optimal dual frame for $m-$erasure.*

In particular, the standard dual of a uniform tight frame is the optimal dual frame for $m-$erasures. In fact, because the frame operator S of a tight frame is of the form $S = AI$, where A is the frame bound, $\|S^{-1} f_i\| = \frac{1}{A}\|f_i\|$ for all i. Using the uniformness of the frame, we obtain the conditions of the Theorem.

The necessary and sufficient condition for the standard dual of a frame to be the 1-erasure optimal dual frame is given in [15]. Let F be an (N, n) frame and $c = \max\left\{\|S^{-1} f_i\| \cdot \|f_i\| : i \in \{1, \ldots, N\}\right\}$. Define $H_i = \text{span}\{f_i : i \in \Lambda_j\}$ for $j = 1, 2$, where $\Lambda_1 = \{i : |S^{-1} f_i\| \cdot \|f_i\| = c\}$, and $\Lambda_2 = \{1, \ldots, N\}\backslash\Lambda_1$.

Theorem 3.3 *The standard dual is the unique 1-erasure optimal dual if and only if $H_1 \cap H_2 = \{0\}$ and $\{f_i\}_{i\in\Lambda_2}$ is linearly independent set.*

Propsition 3.3 *For an (N, n) Parseval frame $F = \{f_i\}_{i=1}^N$ for H, the standard dual is the unique optimal dual frame for $m-$erasure if and only if $\|f_i\|$ is constant for all i.*

3.3.2 Optimality with Respect to Spectral Radius Measurement

Most of the research so far (c.f. [11, 15, 17]) have focused on measuring the error of the reconstructed vector by operator norm. For example, it is known that a Parseval frame is one-erasure optimal if and only if it is uniform, and it is 2-erasure optimal if it is equiangular (c.f. [11]). For the case when a frame F is preselected, optimal dual problems for erasures were studied for example in [15–17], optimal dual frame for sparsity was investigated in [13], and some other optimality was also studied for different purposes (c.f. [5, 7, 9, 14, 18, 19]).

Now consider the case when iterations are applied in the reconstruction process: Let $F = \{f_i\}_{i=1}^N$ be a frame and $G = \{g_i\}_{i=1}^N$ be a dual frame of F in a Hilbert space H with dimension n. For any $f \in H$, we have $f = \sum_{i=1}^N \langle f, g_i\rangle f_i = \sum_{i=1}^N a_i f_i$, where $\langle f, g_i\rangle = a_i$. Let $\Lambda = \{i : a_i \text{ is lost or erased}\}$ and $\Lambda^c = \{1, \ldots, N\}\backslash\Lambda$. Now, we can rewrite the reconstruction formula for f in the following way; $f = \sum_{i\in\Lambda}\langle f, g_i\rangle f_i + \sum_{i\in\Lambda^c}\langle f, g_i\rangle f_i$, or equivalently, $f = E_\Lambda f + R_\Lambda f$, where

$E_\Lambda f = \sum_{i\in\Lambda}\langle f, g_i\rangle f_i$ and $R_\Lambda f = \sum_{i\in\Lambda^c}\langle f, g_i\rangle f_i$. Note that $E_\Lambda + R_\Lambda = I$. This implies that the receiver knows both operators E_Λ and R_Λ. The first step approximation of f is given by $f^{(1)} = R_\Lambda f$. However, we can achieve higher approximation accuracy by employing the following iterations:

$$\begin{aligned} f^{(1)} &= R_\Lambda f \\ f^{(2)} &= E_\Lambda f^{(1)} + R_\Lambda f \\ f^{(3)} &= E_\Lambda f^{(2)} + R_\Lambda f \\ &\vdots \qquad\quad \vdots \\ f^{(n)} &= E_\Lambda f^{(n-1)} + R_\Lambda f. \end{aligned}$$

Then, the error of the reconstruction is $f - f^{(n)} = E_\Lambda f - E_\Lambda f^{(n-1)} = E_\Lambda(f - f^{(n-1)}) = E_\Lambda(E_\Lambda f - E_\Lambda f^{(n-2)}) = E_\Lambda^2(f - f^{(n-2)}) = E_\Lambda^{n-1}(f^{(1)} - f) = E_\Lambda^n f$. Thus, we have $\|f - f^{(n)}\| = \|E_\Lambda^n f\| \le \|E_\Lambda^n\|\|f\|$.

To measure the error, we need to look at the norm of E_Λ^n, $\|E_\Lambda^n\|$. It can be estimated by the spectral radius of E_Λ. Recall that $r(E_\Lambda) \le \|E_\Lambda\|$. In the case that E_Λ is positive or normal, $E_\Lambda^* E_\Lambda = E_\Lambda E_\Lambda^*$, we have $\|E_\Lambda\| = r(E_\Lambda)$, thus, $\|E_\Lambda\| = \max_i |\lambda_i|$, where λ_i is an eigenvalue of E_Λ. But, it could happen that $r(E_\Lambda) << \|E_\Lambda\|$. In this case, $\lim_{n\to\infty}\|E_\Lambda^n\|^{1/n} = r(E_\Lambda)$, where $r(E_\Lambda)$ is the spectral radius of E_Λ. Therefore, the spectral radius $r(E_\Lambda)$ of E_Λ satisfies $r(E_\Lambda) = \max\left\{|\lambda| : \lambda \in \sigma(E_\Lambda)\right\} = \lim_{n\to\infty}\|E_\Lambda^n\|^{1/n}$.

Definition 3.3 Let F be a frame and G be a dual frame of F. For each k, let $r_{F,G}^{(k)} = \max\{r(E_\Lambda) : |\Lambda| = k\}$ and $r_F^{(k)} = \min\{r_{F,G}^{(k)} : G \text{ is a dual frame of } F\}$, where $|\Lambda|$ denotes the cardinality of Λ. A dual frame G of F is called 1-erasure spectrally optimal if $r_{F,G}^{(1)} = r_F^{(1)}$. We say that G is k-erasure spectrally optimal if it is $(k-1)$-erasure spectrally optimal and $r_{F,G}^{(k)} = r_F^{(k)}$.

Clearly we have $r_{F,G}^{(k)} \le \delta_{F,G}^{(k)}$. In the iterated reconstruction introduced in this section, the reconstruction error of a signal f is dominated by $\|E_\Lambda^n\| \cdot \|f\|$. Therefore in order to completely recover f as $n \to \infty$, we need the necessary condition that $r_{F,G}^{(k)} < 1$ (or a more stronger $\delta_{F,G}^{(k)} < 1$). In this section, we present two conditions, mentioned in [20], to ensure this inequality. The first one is a necessary and sufficient condition on the frame F such that this happens for one of the dual frames G. The second one is a necessary and sufficient condition on the triple (N, n, k) such that there exists a dual frame pair (F, G) for H with the property that $r_{F,G}^{(k)} < 1$. Both results involve the standard dual frames.

Propsition 3.4 *Let $F = \{f_i\}_{i=1}^N$ be a frame for a Hilbert space H of dimension n. Assume that k represents the number of erased coefficients in the frame expansion, and S is the frame operator of F. Then the following are equivalent:*

(i) *Every* $(N-k)$ *vectors span the Hilbert space* H,
(ii) $\delta^{(k)}_{S^{-1/2}F,S^{-1/2}F} < 1$,
(iii) $r^{(k)}_{F,S^{-1}F} < 1$,
(iv) *There exists a dual frame* G *of* F *such that* $r^{(k)}_{F,G} < 1$.

Propsition 3.5 Let n be the dimension of H. Then the following are equivalent:

(i) $N-k \geq n$,
(ii) There exists a frame F such that $\delta^{(k)}_{S^{-1/2}F,S^{-1/2}F} < 1$,
(iii) There exists a frame F such that $r^{(k)}_{F,S^{-1}F} < 1$,
(iv) There exists dual pair (F, G) such that $r^{(k)}_{F,G} < 1$.

4 Spectrally One-Uniform Frames

In this section, we mention about a recent work [20] on one-uniform frames and one-erasure optimal dual frames under spectral radius measurement. We define and talk about some properties of spectrally one-uniform frames that admit one-erasure spectrally optimal dual frames.

4.1 Spectrally One-Uniform Frames

Recall that for the 1-erasure case, spectral radius of the error operator satisfies $r^{(1)}_{F,G} = \max\{|\langle g_i, f_i\rangle| : 1 \leq i \leq N\}$. Therefore, for one-erasure spectrally optimal dual frame we have $r^{(1)}_F \geq n/N$ since $\sum_{i=1}^N \langle g_i, f_i\rangle = n$. This leads to the question of characterizing all the frames F such that $r^{(1)}_F = n/N$, and the questions of how to compute $r^{(1)}_F$ and how to construct frames F and their duals G with prescribed maximal error $r^{(k)}_{F,G}$. It turns out that the answers to all these questions rely on an interesting connectivity property for finite sequences (or subset) of nonzero vectors in H. From application point of view we are only interested in frames consisting of nonzero vectors. So we will assume this property throughout the rest of the paper.

Definition 4.1 Let F be an (N, n) frame. Then F is called spectrally one-uniform frame if there exists a dual frame G of F such that $\langle g_i, f_i\rangle = c$ for all $i = 1, \ldots, N$ where $c = n/N$.

Theorem 4.1 *Let* F *be an* (N, n) *frame. Then* F *is spectrally one-uniform frame if and only if there exists a dual* G *such that* $r^{(1)}_F = r^{(1)}_{F,G} = n/N$.

4.2 Linearly Connected Sequences

In [20], three properties of frames: linear connectivity, intersection dependence, and k-independence properties, on which the characterization of spectrally one-uniform frames rely, are defined. And it is proved that the linear connectivity property is equivalent to the intersection dependence property, and is also closely related to the well-known concept of k-independent set.

Two vectors f and g in a sequence F of vectors are linearly F-connected (or simply, connected) if there exist vectors $\{u_1, \ldots, u_\ell\}$ from F such that $\{g, u_1, \ldots, u_\ell\}$ are linearly independent and $f = cg + \sum_{i=1}^{\ell} c_i u_i$ with c, c_i all nonzero. Clearly connectivity is reflexive and symmetric. It is shown in [20] that it is also transitive.

We use the notation $f \overset{F}{\leftrightarrow} g$ if f and g are F-connected.

Definition 4.2 Let $F = \{f_i\}_{i=1}^N$ be a finite sequence of nonzero vectors in H. We say that F

(i) is linearly connected if every two vectors in F are F-connected.
(ii) has the intersection dependent property if $H_\Lambda \cap H_{\Lambda^c} \neq \{0\}$ holds for every proper subset Λ of $\{1, \ldots, N\}$, where H_Λ is the subspace spanned by $\{f_i : i \in \Lambda\}$.
(iii) is k-independent if every k vectors in F are linearly independent.

The Theorem 4.2 states that all these three properties are closely related.

We note as a result of the transitivity property of connected vectors that adding up a vector that is in the span of a connected sequence forms a new sequence of vectors which is connected.

Corollary 4.1 *Let $F = \{f_1, \ldots, f_N\}$ be a connected sequence of H. Then for any nonzero vector $f \in \text{span}\{f_1, \ldots, f_N\}$, the sequence $\{f_1, \ldots, f_N, f\}$ is connected.*

Theorem 4.2 *Let $F = \{f_i\}_{i=1}^N$ be a sequence of H and let $\ell = \dim \text{span}\{f_i : 1 \leq i \leq N\}$. Then the following are equivalent:*

(i) *F is linearly connected.*
(ii) *F has the intersection dependent property.*
(iii) *F contains an $\ell-$independent subset of cardinality of at least $\ell + 1$.*

As a consequence of Theorem 4.2 we obtain the following partition of frames:

Corollary 4.2 *Let $F = \{f_i\}_{i=1}^N$ be a sequence of H. Then there exists a (unique up to permutations) partition $\{\Lambda_j\}_{j=1}^J$ of $\{1, 2, \ldots, N\}$ such that each $\{f_i\}_{i \in \Lambda_j}$ is linearly connected, and H is the direct sum of the subspaces $H_j = \text{span}\{f_i : i \in \Lambda_j\}$.*

4.3 Redundancy Distribution of a Frame

By Corollary 4.2, the redundancy distribution of a frame that helps to characterize and construct spectrally one-uniform frames, and to compute maximum erasure errors, $r_{F,G}^{(1)}$, is defined in the following way:

Definition 4.3 Let $F = \{f_i\}_{i=1}^N$ be a frame for H, and let H_j, Λ_j be as in Corollary 4.2. Then the redundancy distribution of F is defined to be $\left\{\frac{\dim H_j}{|\Lambda_j|}\right\}_{1\le j\le J}$. We say that F has the uniform redundancy distribution if $\frac{\dim H_j}{|\Lambda_j|}$ is a constant for all j.

Let $G = \{g_1, \ldots, g_N\}$ be a dual frame of a frame $F = \{f_1, \ldots, f_N\}$. Define $\Lambda_G = \left\{i : \langle g_i, f_i\rangle = n/N\right\}$ and $\Lambda_G^{\mathsf{c}} = \{1, 2, \ldots, N\}\backslash\Lambda_G$.

Lemma 4.1 *Let $|\Lambda_G^{\mathsf{c}}| \ge 1$ and $i_1, i_2 \in \Lambda_G^{\mathsf{c}}$. If f_{i_1} and f_{i_2} are F-connected, then there exists a dual G' such that $|\Lambda_{G'}| > |\Lambda_G|$.*

Remark 4.1 If $|\Lambda_G^{\mathsf{c}}| = 1$, then for all $i = 1, \ldots, N$ $\langle g_i, f_i\rangle = \frac{n}{N}$. Indeed, by assumption $N-1$ vectors, say $f_1, \ldots, f_{N-1}$, have the property $\langle g_i, f_i\rangle = \frac{n}{N}$. Because $\sum_{i=1}^N \langle g_i, f_i\rangle = n$, we also have $\langle g_N, f_N\rangle = \frac{n}{N}$.

Corollary 4.3 *If F is a connected frame, then there exists a dual $G' = \{g_i'\}_{i=1}^N$ such that $\langle g_i', f_i\rangle = n/N$ for all i.*

Example 4.1 The converse of Corollary 4.3 is not true. Consider the frame $F = \{e_1, e_1, e_1\} \cup \{e_2, e_2, e_2\}$ in $\mathbb{C}^2$. It has a dual $G = \{e_1/3, e_1/3, e_1/3\} \cup \{e_2/3, e_2/3, e_2/3\}$ with $\langle g_i, f_i\rangle = 1/3$ for $i = 1, 2, 3$. However, $f_3 = e_1$ and $f_4 = e_2$ are not F-connected.

Let $F = \{f_i\}_{i=1}^N = \cup_{j=1}^J F_j$ be a frame with a partition $\{\Lambda_j\}_{j=1}^J$ of $\{1, \ldots, N\}$ and $F_j = \{f_i : i \in \Lambda_j\}$ is linearly connected. Let $n_j = \dim H_j = \dim \operatorname{span}\{f_i : i \in \Lambda_j\}$, $N_j = |\Lambda_j|$ and $d_j = \frac{n_j}{N_j}$.

Propsition 4.1 *There exist a dual frame $G_j = \{g_i\}_{i\in\Lambda_j}$ of F_j with $\langle g_i, f_i\rangle = d_j$ for all $i \in \Lambda_j$. Moreover, there exist a dual frame G' of F such that $\langle g_i', f_i\rangle = d_j$ for all $i \in \Lambda_j$.*

The following lemma gives the precise value of $r_F^{(1)}$ for any given frame F.

Lemma 4.2 *Let F be a frame and $\{d_j\}_{j=1}^J$ be its redundancy distribution. Then $r_F^{(1)} = \max\{d_j : j = 1, \ldots, J\}$. In particular, $r_F^{(1)}$ only takes rational values.*

In the following theorem, the characterization of all the frames that admit dual frames, so that the maximal one-erasure reconstruction error is minimal, is given by Theorem 4.2.

Theorem 4.3 *Let $F = \{f_i\}_{i=1}^N$ be a frame for H. Then the following are equivalent:*

(i) *F is spectrally one-uniform frame;*
(ii) *F has the uniform redundancy distribution;*
(iii) *There exists a dual frame $G = \{g_i\}_{i=1}^N$ of F such that $\langle g_i, f_i\rangle = n/N$ for all i;*
(iv) *There exists a dual frame $G = \{g_i\}_{i=1}^N$ of F such that $|\langle g_i, f_i\rangle| = n/N$ for all i.*

Frames with any prescribed redundancy distributions can be easily constructed by the following theorem.

Theorem 4.4 *Let* $d_j = \frac{n_j}{N_j} \in (0, 1)$ *with the property that* $\sum_{j=1}^{J} n_j = n$ *and* $\sum_{j=1}^{J} N_j = N$. *Then there exists a frame* $F = \{f_i\}_{i=1}^{N}$ *such that its redundancy distribution is* $\{d_j\}_{j=1}^{J}$. *Moreover, such a frame* F *can be explicitly constructed out of any given basis of* H.

Corollary 4.4 (i) *If* F *is a uniform Parseval frame, then it has the uniform redundancy distribution.*

(ii) *Assume that* F *has the uniform redundancy distribution and* N *and* n *are coprime to each other. Then* F *is connected.*

References

1. Bodmann, B., Paulsen, V.I.: Frames, graphs and erasures. Linear Algebra Appl. **404**, 118–146 (2005)
2. Bodmann, B.: Optimal linear transmission by loss-insensitive packet encoding. Appl. Comput. Harmon. Anal. **22**, 274–285 (2007)
3. Bodmann, B., Paulsen, V., Tomforde, M.: Equiangular tight frames from complex Seidel matrices containing cube roots of unity. Linear Algebra Appl. **430**, 396–417 (2009)
4. Bodmann, B., Singh, P.: Burst erasures and the mean-square error for cyclic Parseval frames. IEEE Trans. Inform. Theor. **57**, 4622–4635 (2011)
5. Cahill, J., Fickus, M., Mixon, D.G., Poteet, M.J., Strawn, N.K.: Constucting finite frames of a given spectrum and set of lengths (submitted). arXiv:1106.0921v1
6. Casazza, P.G., Kovacević, J.: Equal-norm tight frames with erasures. Adv. Compt. Math. **18**, 387–430 (2003)
7. Casazza, P.G., Heinecke, A., Krahmer, F., Kutyniok, G.: Optimally sparse frames. IEEE Trans. Inform. Theor. **57**, 7279–7287 (2011)
8. Duncan, D.M., Hoffman, T.R., Solazzo, J.P.: Equiangular tight frames and fourth root Seidel matrices. Linear Algebra Appl. **432**, 2816–2823 (2010)
9. Goyal, V.K., Kovacević, J., Kelner, J.A.: Quantized frame expansions with erasures. Appl. Comput. Harmon. Anal. **10**, 203–233 (2001)
10. Hoffman, T., Solazzo, J.: Complex equiangular tight frames and erasures (submitted). arXiv:1107.2173v1
11. Holmes, R., Paulsen, V.: Optimal frames for erasures. Linear Algebra Appl. **377**, 31–51 (2004)
12. Kalra, D.: Complex equiangular cyclic frames and erasures. Linear Algebra Appl. **419**, 373–399 (2006)
13. Krahmer, F., Kutyniok, G., Lemvig, J.: Sparsity and spectral properties of dual frames (submitted). arXiv:1204.5062v1
14. Lemvig, J., Miller, C., Okoudjou, K.A.: Prime tight frames (submitted). arXiv:1202.6350v3
15. Leng, J., Han, D.: Optimal dual frames for erasures. Linear Algebra Appl. **435**, 1464–1472 (2011)
16. Leng, J., Han, D., Huang, T.: Optimal dual frames for communication coding with probabilistic erasures. IEEE Trans. Signal Process. **59**, 5380–5389 (2011)
17. Lopez, J., Han, D.: Optimal dual frames for erasures. Linear Algebra Appl. **432**, 471–482 (2010)
18. Massey, P., Ruiz, M., Stojanoff, D.: Optimal completions of a frame (submitted). arXiv:1206.3588v1

19. Massey, P., Ruiz, M., Stojanoff, D.: Optimal dual frames and frame completions for majorization. Appl. Comput. Harmon. Anal. (in press)
20. Pehlivan, S., Han, D., Mohapatra, R.: Linearly connected sequences and spectrally optimal dual frames for erasures. J. Funct. Anal. (to appear) (2013)
21. Strohmer, T., Heath, R.W.: Grassmannian frames with applications to coding and communication. Appl. Comp. Harmonic Anal. **14**, 257–275 (2003)

Chapter 6
Semi-inner Product: Application to Frame Theory and Numerical Range of Operators

N. K. Sahu and C. Nahak

Abstract This paper deals with the theory of semi-inner product, its generalizations, and applications to frame theory and numerical range of operators. The notion of frames is introduced in classical and generalized semi-inner product spaces. Numerical range of two operators is also studied in semi-inner product spaces.

1 Semi-inner Product

An inner product is a handy and powerful tool to study the geometrical properties of Hilbert space. It is difficult to build Hilbert space-like theory in Banach spaces because of the absence of inner product. A semi-inner product is a generalization of inner product. It was introduced by Lumer [11] for the purpose of extending Hilbert space-like arguments to Banach spaces. It plays a vital role in describing the geometry on Banach spaces. The formal definition of semi-inner product due to Lumer is as follows:

Definition 1.1 (Lumer [11])
Let X be a vector space over the real or complex field F. A semi-inner product $[.,.]$ on X is a real or complex valued functional defined on $X \times X$, which satisfies the following properties:

1. $[x + y, z] = [x, z] + [y, z]$
   $[\lambda x, y] = \lambda[x, y]$ for all $x, y, z \in X$ and $\lambda \in F$;

The authors are thankful to the referees for their valuable suggestions which improved the presentation of the paper. The first author is thankful to the Council of Scientific and Industrial Research (CSIR), Govt. of India, for the financial support.

N. K. Sahu · C. Nahak (✉)
Department of Mathematics, Indian Institute of Technology Kharagpur, Kharagpur 721302, India
e-mail: cnahak@maths.iitkgp.ernet.in

N. K. Sahu
e-mail: nabin@maths.iitkgp.ernet.in

R. N. Mohapatra et al. (eds.), *Mathematics and Computing 2013*,
Springer Proceedings in Mathematics & Statistics 91,
DOI: 10.1007/978-81-322-1952-1_6, © Springer India 2014

2. $[x, x] > 0$ for $x \neq 0$ for all $x \in X$;
3. $|[x, y]|^2 \leq [x, x][y, y]$ for all $x, y \in X$.

The vector space X endowed with $[.,.]$ is called a semi-inner product space.

Lumer proved that a semi-inner product space is a normed linear space with the norm $\|x\| = [x, x]^{\frac{1}{2}}$. Every normed linear space can be made into semi-inner product space in many ways. An inner product space is a semi-inner product space where the inner product plays the role of semi-inner product. Conversely, a semi-inner product is an inner product if and only if the norm induced by the semi-inner product obeys the parallelogram law. It was Giles [7] who put forward some decisive structural modifications to the notion of semi-inner product. He imposed the additional homogeneity property in the Definition 1.1 of Lumer semi-inner product. That is, $[x, \lambda y] = \overline{\lambda}[x, y]$ for all $\lambda \in F$, where $\overline{\lambda}$ denotes the conjugate of λ. The imposition of this property adds much convenience without causing any significant restriction. He proved that every normed linear space is a semi-inner product space with the homogeneity property.

Definition 1.2 (Giles [7])
A semi-inner product $[.,.]$ is continuous, if it satisfies

$$\lim_{\lambda \to 0} Re[y, x + \lambda y] \to Re[y, x] \text{ for all } x, y \in X \text{ and } \lambda \in \mathbb{R}.$$

The corresponding space X is called continuous semi-inner product space. If the involved limit is uniform, then it is called uniformly continuous semi-inner product space.

Giles also defined the orthogonality relation in semi-inner product space.

Definition 1.3 Let X be a semi-inner product space. For $x, y \in X$, x is said to be normal to y and y is said to be transversal to x if $[y, x] = 0$. A vector $x \in X$ is normal to a subspace S of X and S is transversal to x if x is normal to all vectors $y \in S$.

Definition 1.4 A normed linear space X is said to be G$\hat{a}$teaux differentiable or smooth if for all $x, y \in X$ and real λ, $\lim_{\lambda \to 0} \dfrac{\|x + \lambda y\| - \|x\|}{\lambda}$ exists.

Giles proved that the continuity restriction on the semi-inner product is equivalent to the G$\hat{a}$teaux differentiability of the norm.

To extend Hilbert space-type argument to the theory of the dual of a semi-inner product space, one has to impose more restriction on the semi-inner product to guarantee the existence of normals to closed vector subspaces. For that, one has to restrict the normed space.

Definition 1.5 A normed space X is strictly convex if whenever $\|x\| + \|y\| = \|x + y\|$, where $x, y \neq 0$, then $y = \lambda x$ for some real $\lambda > 0$.

Definition 1.6 A normed space X is uniformly convex if given $\varepsilon > 0$, there exists a $\delta(\varepsilon) > 0$ such that for $x, y \in X$ with $\|x\| = \|y\| = 1$, we have $\frac{\|x+y\|}{2} \leq 1 - \delta(\varepsilon)$ when $\|x - y\| > \varepsilon$.

It is true that uniform convexity implies strict convexity. It is also proved that a semi-inner product space is strictly convex if and only if the equality $[x, y] = \|x\|\|y\|$, where $x, y \neq 0$, implies that $y = \lambda x$ for some real $\lambda > 0$ (see Berkson [2]).

In Hilbert space, the representation theorem for continuous linear functionals sets up a natural correspondence between vectors and continuous linear functionals by means of the inner product. This correspondence was discovered by the famous mathematician Riesz and is known as the Riesz representation theorem. There is a similar representation theorem named as the generalized Riesz representation theorem in a continuous semi-inner product space which is a uniformly convex Banach space.

Theorem 1.1 [Generalized Riesz representation theorem] (Giles [7])
Let X be a continuous semi-inner product space which is uniformly convex and complete in its norm. Let X^* be the dual space of X. Then for every continuous linear functional $f \in X^*$ there exists a unique vector $y \in X$ such that $f(x) = [x, y]$ for all $x \in X$.

Definition 1.7 A uniform semi-inner product space is a uniformly continuous semi-inner product space where the induced normed space is uniformly convex and complete.

Theorem 1.2 (Giles [7])
If X is a uniform semi-inner product space, then the dual space X^* is also a uniform semi-inner product space with respect to the semi-inner product defined by $[f_x, f_y]_{X^*} = [y, x]$, where $[.,.]_{X^*}$ denotes the semi-inner product in X^*.

Giles also proved that every finite dimensional strictly convex, continuous semi-inner product space is a uniform semi-inner product space. We have the following examples of uniform semi-inner product spaces:

Example 1.1 The real Banach space $L_p(X, \rho, \mu)$ for $1 < p < \infty$ is a uniform semi-inner product space with the semi-inner product defined as

$$[y, x] = \frac{1}{\|x\|_p^{p-2}} \int_X y|x|^{p-1} \operatorname{sgn}(x) \mathrm{d}\mu.$$

Example 1.2 The real sequence space l^p for $1 < p < \infty$ is a uniform semi-inner product space with the semi-inner product defined as

$$[x, y] = \frac{1}{\|y\|_p^{p-2}} \sum_i x_i y_i |y_i|^{p-2}.$$

If the vector space is a uniformly convex smooth Banach space, then there is unique semi-inner product.

The notion of generalized adjoint of a bounded linear operator in a semi-inner product space was introduced by Koehler [10]. Let X be a uniformly convex smooth Banach space. If A is a bounded linear operator from X to itself, then the map $g_y : X \to F(\mathbb{R}\ or\ \mathbb{C})$, defined by $g_y(x) = [Ax, y]$ is a continuous linear functional. By the generalized Riesz representation theorem, it follows that there is a unique vector $A^\dagger(y)$ such that $[Ax, y] = [x, A^\dagger y]$ for all $x \in X$. The operator $A^\dagger$ is called the generalized adjoint of A. This generalized adjoint operator is not usually linear but still it has some interesting properties. The following properties are investigated by Koehler [10] for the generalized adjoint operator:

Theorem 1.3 *Let A and B be two bounded linear functionals on a uniformly convex smooth Banach space X and λ be a scalar. Then,*

1. $(\lambda A)^\dagger = \overline{\lambda} A^\dagger$;
2. $(AB)^\dagger = B^\dagger A^\dagger$;
3. $A^\dagger$ *is one-to-one if and only if the range of A is dense in X;*
4. *If the norm of X is strongly (Frechet) differentiable, then $A^\dagger$ is continuous.*

1.1 Semi-inner Product Space of Type (p)

Nath [13] generalized the concept of semi-inner product introduced by Lumer [11], by replacing the Schwarz's inequality with the Holder's inequality. The similar type of semi-inner product is called semi-inner product of type (p), and is defined as follows:

Definition 1.8 Let X be a vector space over the field F of real or complex numbers. The functional $[.,.] : X \times X \to F$ satisfying

1. $[x + y, z] = [x, z] + [y, z]$ for all $x, y, z \in X$;
2. $[\lambda x, y] = \lambda [x, y]$ for all $\lambda \in F$ and $x, y \in X$;
3. $[x, x] > 0$ for all $x \neq 0$;
4. $|[x, y]| \leq [x, x]^{\frac{1}{p}} [y, y]^{\frac{p-1}{p}}$ for all $x, y \in X$ and $1 < p < \infty$;

is called a semi-inner product of type (p) on X. The space equipped with $[.,.]_p$ is called the semi-inner product space of type (p).

The semi-inner product of type (p) induces a norm by setting $\|x\| = [x, x]^{\frac{1}{p}}$. Also, for every normed space we can construct semi-inner product of type (p) in many ways. Pap and Pavlovic [14] discovered the adjoint theorem for maps on semi-inner product spaces of type (p). They proved some properties of the generalized adjoint operator similar to the properties established by Koehler [10] in semi-inner product spaces. El-Sayyad and Khaleelulla [6] introduced the semi-inner product algebras of type (p). They found some interesting results on the generalized adjoint of an operator defined on this space.

Theorem 1.4 (El-Sayyad and Khaleelulla [6])
Let T be a bounded linear operator defined on a semi-inner product space of type (p) and $T^\dagger$ be its generalized adjoint. Then,

(i) $\|T\| = \|T^\dagger\|^{p-1}$,
(ii) $\|T^\dagger T\|^{p-1} = \|T\|^p$.

1.2 Generalized Semi-inner Product

With a view to study regularized learning in general Banach spaces, Zhang and Zhang [18] introduced the concept of generalized semi-inner product.

To define generalized semi-inner product, one has to know the notion of gauge function. A gauge function ϕ is a map $\phi : \mathbb{R}^+ \to \mathbb{R}^+$ such that ϕ is continuous, surjective, and strictly increasing with $\phi(0) = 0$ and $\lim_{t\to\infty} \phi(t) = +\infty$. The definition of generalized semi-inner product is as follows:

Definition 1.9 Let X be a vector space over the field F of real or complex numbers. Let ϕ and ψ be two gauge functions with $\phi(t)\psi(t) = t$ for all positive real numbers t. The map $[.,.]_\phi : X \times X \to F$ satisfying

1. $[\alpha x + \beta y, z]_\phi = \alpha[x, z]_\phi + \beta[y, z]_\phi$ for all $\alpha, \beta \in F$ and $x, y, z \in X$;
2. $[x, x]_\phi > 0$ for all $x \in X \setminus \{0\}$;
3. $|[x, y]_\phi| \leq \phi([x, x]_\phi)\ \psi([y, y]_\phi)$ for all $x, y \in X$ and the equality holds when $x = y$;

is called a generalized semi-inner product on X. The space X equipped with $[.,.]_\phi$ is called a generalized semi-inner product space.

When $\phi(t) = t^{\frac{1}{p}}$ and $\psi(t) = t^{\frac{1}{q}}$, $p, q \in (1, +\infty)$ with $\frac{1}{p} + \frac{1}{q} = 1$, the generalized semi-inner product reduces to the semi-inner product of type (p) introduced by Nath [13]. Again if $\phi(t) = t^{\frac{1}{2}}$ and $\psi(t) = t^{\frac{1}{2}}$ then the generalized semi-inner product reduces to the classical semi-inner product introduced by Lumer [11]. Zhang and Zhang [18] proved that if $[.,.]_\phi$ is a generalized semi-inner product on a vector space X then $\|x\| = \Phi([x, x]_\phi)$ defines a norm on X. Conversely, if Φ is surjective onto $\mathbb{R}^+$ then for any normed space X, there exists a generalized semi-inner product on it such that $\|x\| = \Phi([x, x]_\phi)$. The Riesz representation of continuous linear functionals is also true in this generalized semi-inner product space.

2 Bessel Sequence and Frame in Semi-inner Product Space

Frames are redundant signal representations having a wide range of applications in signal and image processing, wavelet analysis, data transmission with erasures, wireless communication, data transmission, and many more new applications arising every year. In this section, we define Bessel sequence and frame in Banach spaces

by using semi-inner product. The notion of frames was introduced by Duffin and Schaeffer [5] in 1952 while studying the nonharmonic Fourier series. Frames in L^p spaces and other Banach spaces are effective tools for modeling a variety of natural signals and images. There is a plethora of literature available for frames in Banach spaces. For classical frame theory in Banach spaces, one may refer to Casazza and Christensen [3], Christensen and Heil [4], Gröchenig [8], Kaushik [9], and Stoeva [15]. To smoothen the study of frames in Banach spaces, Zhang and Zhang [19] defined this notion by taking the help of semi-inner product.

Here we assume that X is a uniformly convex smooth Banach space. In particular, we concentrate on the spaces l^p and L^p, where $1 < p < \infty$. It is seen that those spaces are semi-inner product spaces with uniquely defined semi-inner product (Giles [7]). Our definition is completely different from those Banach space frames available in the literature. In the remainder of this section, we assume that X is a real uniformly convex smooth Banach space with norm $\|.\|_p$ and semi-inner product $[.,.]$.

Definition 2.1 A set of elements $f = \{f_i\}_{i=1}^{\infty} \subseteq X$ is called a Bessel sequence if there exists a constant $B > 0$, such that

$$\sum_{i=1}^{\infty} |[f_i, x]|^q \leq B(\|x\|_p)^q, \ \forall x \in X,$$

where $1 < p, q < \infty$ and $\frac{1}{p} + \frac{1}{q} = 1$. The number B is called Bessel bound.

Definition 2.2 A sequence of elements $\{f_i\}_{i=1}^{\infty}$ in X is called a frame if there exist positive constants A and B such that

$$A(\|x\|_p)^q \leq \sum_{i=1}^{\infty} |[f_i, x]|^q \leq B(\|x\|_p)^q, \quad \forall x \in X,$$

where $1 < p, q < \infty$ and $\frac{1}{p} + \frac{1}{q} = 1$. A and B are called lower and upper frame bound, respectively.

If $A = B$ then the frame is called a tight frame, and if $A = B = 1$ then the frame is called a Parseval frame. A frame is called a normalized frame if each frame element has unit norm. We have proved the following necessary and sufficient condition for a sequence of elements in X to be a Bessel sequence.

Theorem 2.1 *Let $f = \{f_i\}_{i=1}^{\infty}$ be a sequence in X. Then, the sequence f is a Bessel sequence if and only if $T : \{c_i\}_{i=1}^{\infty} \to \sum_{i=1}^{\infty} c_i f_i \frac{|[f_i, x]|^{q-2}}{\|\{[f_i, x]\}\|^{q-2}}$ is a well-defined and bounded operator from l^q into X.*

Our main focus is on Parseval frame and tight frame because the reconstruction formula naturally holds true without any assumptions. The following two results establish the reconstruction formulae for Parseval frames and tight frames:

Theorem 2.2 *A set of elements $\{f_i\}_{i=1}^{\infty}$ is a Parseval frame for X if and only if*

$$x = \sum_{i=1}^{\infty} \frac{|[f_i, x]|^{q-2}}{\|\{[f_i, x]\}\|^{q-2}} [f_i, x] f_i, \quad \forall x \in X. \tag{1}$$

Theorem 2.3 *A set of elements $\{f_i\}_{i=1}^{\infty}$ is a tight frame with bound A for X if and only if*

$$x = \sum_{i=1}^{\infty} \frac{1}{A^{\frac{2}{q}}} \frac{|[f_i, x]|^{q-2}}{\|\{[f_i, x]\}\|^{q-2}} [f_i, x] f_i \quad \forall x \in X. \tag{2}$$

Definition 2.3 A tight frame is said to be a normalized tight frame if each of its element has unit norm.

Definition 2.4 An operator T on X is said to be a co-isometry if its generalized adjoint is an isometry.

The following theorem tells about the invariance of frame under a co-isometry operator.

Theorem 2.4 (a) *Let $\{f_i\}_{i=1}^{\infty}$ be a frame for the space X and T be a co-isometry, then $\{Tf_i\}_{i=1}^{\infty}$ is a frame. Moreover, $\{Tf_i\}_{i=1}^{\infty}$ is a normalized tight frame if $\{f_i\}_{i=1}^{\infty}$ is a normalized tight frame.*

(b) *Let $\{f_i\}_{i=1}^{\infty}$ and $\{g_i\}_{i=1}^{\infty}$ be Parseval frames for X and T be a bounded linear operator defined by $Tg_i = f_i$. Then T is a co-isometry.*

3 Bessel Sequence and Frame in Generalized Semi-inner Product Space

Let X be a generalized semi-inner product space with generalized semi-inner product $[.,.]_\phi$ and norm $\|.\|_X$. Let X_d be an associated BK-space with norm $\|.\|_{X_d}$. Suppose that X^* and X_d^* are the dual spaces of X and X_d, respectively. We define X_d-Bessel sequence and X_d^*-Bessel sequence in a generalized semi-inner product space X, and prove that the space of all X_d^*-Bessel sequences form a Banach space.

Definition 3.1 A sequence of elements $\{f_j\} \subseteq X$ is called an X_d-Bessel sequence in X if $\{[x, f_j]_\phi\} \in X_d$, and there exists a positive real constant B such that

$$\|\{[x, f_j]_\phi\}\|_{X_d} \leq B\, \phi([x, x]_\phi), \ \forall x \in X,$$

where $\phi : (0, \infty) \to (0, \infty)$ is a continuous, nondecreasing function with $\phi(0) = 0$ and $\phi(t) \to \infty$ as $t \to \infty$.

Definition 3.2 Let $\{f_j\} \subseteq X$. Then $\{f_j^*\} \subseteq X^*$ is an X_d^*-Bessel sequence for X^* if $\{[f_j, x]_\phi\} \in X_d^*$, and there exists a positive real constant B such that

$$\|\{[f_j, x]_\phi\}\|_{X_d^*} \leq B\ \phi([x, x]_\phi),\ \forall x \in X.$$

We define X_d-frame and X_d^*-frame in this prospective.

Definition 3.3 Let X be a generalized semi-inner product space with compatible generalized semi-inner product $[.,.]_\phi$. The sequence $\{f_j\} \subseteq X$ is said to be an X_d-frame for X if $\{[f, f_j]_\phi\} \in X_d$ for all $x \in X$, and there exist two positive constants A, B such that

$$A\ \phi([f, f]_\phi) \leq \|\{[f, f_j]_\phi\}\|_{X_d} \leq B\ \phi([f, f]_\phi),\ \ \forall f \in X. \tag{3}$$

Definition 3.4 Let $\{f_j\} \subseteq X$. Then $\{f_j^*\}$ is an X_d^*-frame for X^* if $\{[f_j, f]_\phi\} \in X_d^*$ for all $f \in X$, and there exist two positive constants A, B such that

$$A\ \phi([f, f]_\phi) \leq \|\{[f_j, f]_\phi\}\|_{X_d^*} \leq B\ \phi([f, f]_\phi),\ \ \forall f \in X. \tag{4}$$

In this section, we also define Riesz basis in a generalized semi-inner product space. Likewise X_d-frame and X_d^*-frame, we have X_d-Riesz basis and X_d^*-Riesz basis.

Definition 3.5 A sequence of elements $\{f_j\} \subseteq X$ is an X_d-Riesz basis for X if $\overline{span\{f_j\}} = X$, $\sum_{j\in I} c_j f_j$ converges in X for all $c \in X_d$, and there exist positive finite real numbers $A,\ B$ with $A \leq B$, such that

$$A\phi([c, c]_{X_d}) \leq \Big\| \sum_{j\in I} c_j f_j \Big\|_X \leq B\phi([c, c]_{X_d}) \ \text{ for all } c \in X_d. \tag{5}$$

Definition 3.6 A sequence of elements $\{f_j^*\} \subseteq X^*$ is an X_d^*-Riesz basis for X^* if $\overline{span\{f_j^*\}} = X^*$, $\sum_{j\in I} d_j f_j^*$ converges in X^* for all $d \in X_d^*$, and there exist positive finite real numbers $A,\ B$ with $A \leq B$, such that

$$A\phi([d, d]_{X_d^*}) \leq \Big\| \sum_{j\in I} d_j f_j^* \Big\|_{X^*} \leq B\phi([d, d]_{X_d^*}) \ \text{ for all } d \in X_d^*. \tag{6}$$

We can show that Riesz basis automatically generates a frame for the dual space.

4 Numerical Range of Two Operators in Semi-inner Product Spaces

Quadratic forms are quite useful in linear algebra. The numerical range is a natural extension of quadratic forms in vector spaces. Like the spectrum, the numerical range of a linear operator is a subset of the scalar field. It is structured in such a way that

it is related to both algebraic as well as norm structures of the operator. Whereas the spectrum of an operator is related only to algebraic structure of the operator. One can extract much information about the operator through numerical range.

Lumer [11] discussed the numerical range for a linear operator in a Banach space by using semi-inner product. Williams [17] studied the spectra of products of two linear operators and their numerical ranges. To study the generalized eigenvalue problem $Tx = \lambda Ax$, Amelin [1] introduced the concept of numerical range for two linear operators in Hilbert space. The numerical range of two nonlinear operators in a semi-inner product space was defined by Nanda [12].

4.1 Numerical Range of Two Linear Operators

Let X be a uniformly convex smooth Banach space equipped with norm $\|.\|$ and semi-inner product $[.,.]$. Let T and A be two linear operators defined on X.

Definition 4.1 The numerical range $W(T, A)$ of the two linear operators T and A is defined as $W(T, A) := \{[Tx, Ax] : \|Ax\| = 1,\ x \in \mathrm{D}(T) \cap \mathrm{D}(A)\}$, where $D(T)$ and $D(A)$ are denoted as the domain of T and the domain of A, respectively. The numerical radius $w(T, A)$ is defined as $w(T, A) = \sup\{|\lambda| : \lambda \in W(T, A)\}$.

Definition 4.2 The coupled numerical range $W_A(T)$ of T with respect to A is defined as

$$W_A(T) := \left\{ \frac{[ATx, x]}{[Ax, x]} : \|x\| = 1,\ [Ax, x] \neq 0 \right\}. \tag{7}$$

In the above definition, we have assumed that $\mathrm{Dom}(A) \cap \mathrm{Range}(T) \neq \phi$. We can easily prove the following properties of the numerical range of two linear operators:

Theorem 4.1 *Let T_1, T_2, T, A be linear operators and α, μ, λ be scalars. Then,*

(i) $W(T_1 + T_2, A) \subseteq W(T_1, A) + W(T_2, A)$,
(ii) $W(\alpha T, A) = \alpha W(T, A)$,
(iii) $W(T, \mu A) = \overline{\mu} W(T, A)$,
(iv) $W(T - \lambda A, A) = W(T, A) - \{\lambda\}$,
(v) $w(T_1 + T_2, A) \leq w(T_1, A) + w(T_2, A)$,
(vi) $w(\lambda T, A) = |\lambda|\, w(T, A)$.

Theorem 4.2 *Let T_1, T_2, T, A be linear operators and α be a scalar. Then,*

(i) $W_A(T_1 + T_2) \subseteq W_A(T_1) + W_A(T_2)$,
(ii) $W_A(\alpha T) = \alpha W_A(T)$,
(iii) $W_{\alpha A}(T) = W_A(T)$.

Proof (i) Let $x, y \in \text{Dom}(T_1) \cap \text{Dom}(T_2)$. We assume that $\text{Dom}(A) \cap \text{Range}(T_1) \neq \phi$ and $\text{Dom}(A) \cap \text{Range}(T_2) \neq \phi$.
Then

$$\frac{[A(T_1 + T_2)x, x]}{[Ax, x]} = \frac{[AT_1x + AT_2x, x]}{[Ax, x]} = \frac{[AT_1x, x]}{[Ax, x]} + \frac{[AT_2x, x]}{[Ax, x]}.$$

Therefore $W_A(T_1 + T_2) \subseteq W_A(T_1) + W_A(T_2)$.
(ii) If $\text{Dom}(A) \cap \text{Range}(T) \neq \phi$, then

$$\frac{[A(\alpha T)x, x]}{[Ax, x]} = \frac{[\alpha ATx, x]}{[Ax, x]} = \alpha \frac{[ATx, x]}{[Ax, x]}.$$

Hence $W_A(\alpha T) = \alpha W_A(T)$.
(iii) If $\text{Dom}(A) \cap \text{Range}(T) \neq \phi$, then

$$\frac{[(\alpha A)Tx, x]}{[\alpha Ax, x]} = \frac{\alpha[ATx, x]}{\alpha[Ax, x]} = \frac{[ATx, x]}{[Ax, x]}.$$

As a result $W_{\alpha A}(T) = W_A(T)$.

Definition 4.3 The spectrum $\sigma(T, A)$ of the two linear operators T and A is defined as

$$\sigma(T, A) := \{\lambda \in \mathbb{C} : (T - \lambda A) \text{ is not invertible}\}. \tag{8}$$

The spectral radius $r(T, A)$ is defined as $r(T, A) = \sup\{|\lambda| \ : \ \lambda \in \sigma(T, A)\}$.

Definition 4.4 The eigen spectrum or point spectrum $e(T, A)$ of two linear operators T and A is defined as

$$e(T, A) := \{\lambda \in \mathbb{C} : Tx = \lambda Ax \ \text{ for } x \neq 0\}. \tag{9}$$

Definition 4.5 The approximate point spectrum $\pi(T, A)$ of two linear operators T and A is defined as
$\pi(T, A) := \{\lambda \in \mathbb{C}$ such that there exists a sequence $x_n \in X$ with $\|Ax_n\| = 1$ and $\|Tx_n - \lambda Ax_n\| \to 0$ as $n \to \infty\}$.

Definition 4.6 The compression spectrum $\sigma_0(T, A)$ of two linear operators T and A is defined as

$$\sigma_0(T, A) := \{\lambda \in \mathbb{C} : \text{Range}(T - \lambda A) \text{ is not dense in } X\}. \tag{10}$$

One can establish the inclusion relations among spectrum, eigen spectrum, compression spectrum, approximate point spectrum, and numerical range of two linear operators.

4.2 Numerical Range of Two Nonlinear Operators

Let X be a normed space, and T be an operator defined on X. Then T is said to be Lipschitz if there exists a constant $M > 0$ such that $\|Tx - Ty\| \leq M\|x - y\|$ for all $x, y \in X$. Let $Lip(X)$ denote the set of all Lipschitz operators on X. Suppose that $T \in Lip(X)$, and $x, y \in \mathrm{Dom}(T)$ with $x \neq y$. The generalized Lipschitz norm $\|T\|_L$ of a nonlinear operator T on a Banach space X is defined as $\|T\|_L = \|T\| + \|T\|_l$, where $\|T\| = \sup_x \dfrac{\|Tx\|}{\|x\|}$ and $\|T\|_l = \sup_{x \neq y} \dfrac{\|Tx - Ty\|}{\|x - y\|}$. If there exists a finite constant M such that $\|T\|_L < M$, then the operator T is called the generalized Lipschitz operator (see Verma [16]). Let $G_L(X)$ be the class of all generalized Lipschitz operators.

Definition 4.7 The numerical range $V_L(T, A)$ of two nonlinear operators T and A is defined as

$$V_L(T, A) := \left\{ \frac{[Tx, Ax] + [Tx - Ty, Ax - Ay]}{\|Ax\|^2 + \|Ax - Ay\|^2} : \ x, y \in D(T) \cap D(A), \ x \neq y \right\}, \tag{11}$$

where $D(T)$ and $D(A)$ are the domains of the operators T and A, respectively. The numerical radius $w_L(T, A)$ is defined as $w_L(T, A) = \{\sup |\lambda| : \ \lambda \in V_L(T, A)\}$.

We have the following elementary properties for the numerical range of two nonlinear operators.

Theorem 4.3 *Let X be a Banach space over $\mathbb{C}$. If T, A, T_1, T_2 be nonlinear operators defined on X and λ, μ be scalars, then*

(i) $V_L(\lambda T, A) = \lambda V_L(T, A)$,
(ii) $V_L(T, \mu A) = \frac{1}{\mu} V_L(T, A)$,
(iii) $V_L(T_1 + T_2, A) \subseteq V_L(T_1, A) + V_L(T_2, A)$,
(iv) $V_L(T - \lambda A, A) = V_L(T, A) - \{\lambda\}$.

Proof (i) We see that for any $x, y \in X$,

$$\frac{[\lambda Tx, Ax] + [\lambda Tx - \lambda Ty, Ax - Ay]}{\|Ax\|^2 + \|Ax - Ay\|^2} = \lambda \frac{[Tx, Ax] + [Tx - Ty, Ax - Ay]}{\|Ax\|^2 + \|Ax - Ay\|^2}.$$

Hence $V_L(\lambda T, A) = \lambda V_L(T, A)$.

(ii) For any $x, y \in X$,

$$\begin{aligned} \frac{[Tx, \mu Ax] + [Tx - Ty, \mu Ax - \mu Ay]}{\|\mu Ax\|^2 + \|\mu Ax - \mu Ay\|^2} &= \frac{\overline{\mu}[Tx, Ax] + \overline{\mu}[Tx - Ty, \mu Ax - \mu Ay]}{|\mu|^2(\|Ax\|^2 + \|Ax - Ay\|^2)} \\ &= \frac{\overline{\mu}}{|\mu|^2} \ \frac{[Tx, Ax] + [Tx - Ty, Ax - Ay]}{\|Ax\|^2 + \|Ax - Ay\|^2} \\ &= \frac{1}{\mu} \ \frac{[Tx, Ax] + [Tx - Ty, Ax - Ay]}{\|Ax\|^2 + \|Ax - Ay\|^2}. \end{aligned}$$

Hence $V_L(T, \mu A) = \frac{1}{\mu} V_L(T, A)$.

(iii) Let $x, y \in Dom(T_1) \cap Dom(T_2)$.
Then

$$\begin{aligned}
&\frac{[(T_1 + T_2)x, Ax] + [(T_1 + T_2)x - (T_1 + T_2)y, Ax - Ay]}{\|Ax\|^2 + \|Ax - Ay\|^2} \\
&= \frac{[T_1x, Ax] + [T_2x, Ax] + [T_1x - T_1y, Ax - Ay] + [T_2x - T_2y, Ax - Ay]}{\|Ax\|^2 + \|Ax - Ay\|^2} \\
&= \frac{[T_1x, Ax] + [T_1x - T_1y, Ax - Ay]}{\|Ax\|^2 + \|Ax - Ay\|^2} + \frac{[T_2x, Ax] + [T_2x - T_2y, Ax - Ay]}{\|Ax\|^2 + \|Ax - Ay\|^2}.
\end{aligned}$$

Therefore $V_L(T_1 + T_2, A) \subseteq V_L(T_1, A) + V_L(T_2, A)$. Thus (iii) is proved.

(iv) For any $x, y \in X$,

$$\begin{aligned}
&\frac{[(T - \lambda A)x, Ax] + [(T - \lambda A)x - (T - \lambda A)y, Ax - Ay]}{\|Ax\|^2 + \|Ax - Ay\|^2} \\
&= \frac{[Tx, Ax] - \lambda\|Ax\|^2 + [Tx - Ty, Ax - Ay] - \lambda\|Ax - Ay\|^2}{\|Ax\|^2 + \|Ax - Ay\|^2} \\
&= \frac{[Tx, Ax] + [Tx - Ty, Ax - Ay]}{\|Ax\|^2 + \|Ax - Ay\|^2} - \lambda.
\end{aligned}$$

This implies that $V_L(T - \lambda A, A) = V_L(T, A) - \{\lambda\}$.

We give example of two nonlinear operators in a semi-inner product space and compute their numerical range and numerical radius.

Example 4.1 Consider the real sequence space l^p, $1 < p < \infty$.
Let $x = (x_1, x_2, ...)$, $y = (y_1, y_2, ...) \in l^p$. Consider the two nonlinear operators $T, A : l^p \to l^p$ defined by $Tx = (\|x\|, x_1, x_2, ...)$ and $Ax = (\|x\|, 0, 0, ...)$. The unique semi-inner product on the real sequence space l^p is defined as

$$[x, y] = \frac{1}{\|y\|^{p-2}} \sum_{n=1}^{\infty} |y_n|^{p-2} y_n x_n, \ \forall x = \{x_n\}, y = \{y_n\} \in l^p.$$

One can easily compute that $\|Ax\| = \|x\|$, $\|Ax - Ay\| = |\|x\| - \|y\||$,
$[Tx, Ax] = \|x\|^2$ and

$$\begin{aligned}
[Tx - Ty, Ax - Ay] &= \frac{1}{\|Ax - Ay\|^{p-2}} \{|(\|x\| - \|y\|)|^{p-2} (\|x\| - \|y\|)^2\} \\
&= \frac{1}{|(\|x\| - \|y\|)|^{p-2}} |(\|x\| - \|y\|)|^p = |(\|x\| - \|y\|)|^2.
\end{aligned}$$

One can calculate that

$$\frac{[Tx, Ax] + [Tx - Ty, Ax - Ay]}{\|Ax\|^2 + \|Ax - Ay\|^2} = \frac{\|x\|^2 + |(\|x\| - \|y\|)|^2}{\|x\|^2 + |(\|x\| - \|y\|)|^2} = 1, \ \forall x, y \in l^p.$$

Therefore $V_L(T, A) = \{1\}$ and $w_L(T, A) = 1$.

5 Conclusion

Researchers usually take the help of bounded linear functionals to establish Hilbert space-like theory in Banach spaces. Without using arbitrary bounded linear functionals, we have taken the help of semi-inner product to study frames and numerical range of operators in Banach spaces. The main benefits of this approach are threefold. It is computationally easy. We can avoid the inconvenience of using arbitrary bounded linear functionals. It helps in constructing concrete examples.

References

1. Amelin, C.F.: A numerical range for two linear operators. Pac. J. Math. **48**, 335–345 (1973)
2. Berkson, E.: Some types of Banach spaces, Hermitian operators and Bade functionals. Trans. Am. Math. Soc. **116**, 376–385 (1965)
3. Casazza, P.G., Christensen, O.: The reconstruction property in Banach spaces and perturbation theorem. Canad. Math. Bull. **51**, 348–358 (2008)
4. Christensen, O., Heil, C.: Perturbations of Banach frames and atomic decompositions. Math. Nachr. **158**, 33–47 (1997)
5. Duffin, R.J., Schaeffer, A.C.: A class of nonharmonic Fourier series. Trans. Am. Math. Soc. **72**, 341–366 (1952)
6. El-Sayyad, S.G., Khaleelulla, S.M.: Semi-inner product algebras of type (p). Zb. Rad. Prirod.-Mat. Fak. Ser. Mat. **23**, 175–187 (1993)
7. Giles, J.R.: Classes of semi-inner product spaces. Trans. Am. Math. Soc. **129**, 436–446 (1967)
8. Gröchenig, K.: Localization of frames, Banach frames, and the invertibility of the frame operator. J. Four. Anal. Appl. **10**, 105–132 (2004)
9. Kaushik, S.K.: A generalization of frames in Banach spaces. J. Contemp. Math. Anal. **44**, 212–218 (2009)
10. Koehler, D.O.: A note on some operator theory in certain semi-inner product spaces. Proc. Am. Math. Soc. **30**, 363–366 (1971)
11. Lumer, G.: Semi-inner product spaces. Trans. Am. Math. Soc. **100**, 29–43 (1961)
12. Nanda, S.: Numerical range for two non-linear operators in semi-inner product space. J. Nat. Acad. Math. **17**, 16–20 (2003)
13. Nath, B.: On generalization of semi-inner product spaces. Math. J. Okayama Univ. **15**, 1–6 (1971)
14. Pap, E., Pavlovic, R.: Adjoint theorem on semi-inner product spaces of type (p). Zb. Rad. Prirod.-Mat. Fak. Ser. Mat. **25**, 39–46 (1995)
15. Stoeva, D.T.: On p-frames and reconstruction series in separable Banach spaces. Integr. Transf. Spec. Funct. **17**, 127–133 (2006)

16. Verma, R.U.: The numerical range of nonlinear Banach space operators. Acta Math. Hungar. **63**, 305–312 (1994)
17. Williams, J.P.: Spectra of products and numerical ranges. J. Math. Anal. Appl. **17**, 214–220 (1967)
18. Zhang, H., Zhang, J.: Generalized semi-inner products with applications to regularized lerning. J. Math. Anal. Appl. **372**, 181–196 (2010)
19. Zhang, H., Zhang, J.: Frames, Riesz bases, and sampling expansions in Banach spaces via semi-inner products. Appl. Comput. Harmon. Anal. **31**, 1–25 (2011)

Chapter 7
Multi-level Nonlinear Programming Problem with Some Multi-choice Parameter

Avik Pradhan and M. P. Biswal

Abstract Decentralized planning is important for modeling a real-life decision-making problem. Multi-level programming is a very powerful tool for modeling such type of decentralized planning problems. In a multi-level programming problem, the decision is taken by several decision makers who are in different levels. In this paper, we studied a multi-level nonlinear programming problem where some (or all) of the coefficients of the objectives and the constraints are multi-choice type. We propose a suitable solution procedure to solve the stated multi-level programming problem. To solve these type of problems, first we tackle each multi-choice parameter of the multi-level programming problem by using interpolating polynomial and obtain a multi-level mixed integer nonlinear programming problem. Then we use the concept of tolerance membership function for the objectives and the control variables of the decision makers and formulate a fuzzy max–min type decision model to obtain a Pareto optimal solution of the transformed multi-level programming problem. We present a numerical example to illustrate the solution procedure of the stated problem.

1 Introduction

Modeling a large and complex systems invariably needs to decompose the system into a number of smaller subsystems, each with its own goals and constraints. The interconnections among the subsystems may take on many forms, but one of the most

A. Pradhan (✉) · M. P. Biswal
Department of Mathematics, Indian Institute of Technology, Kharagpur 721302, India
e-mail: avikiitm@gmail.com

M. P. Biswal
e-mail: mpbiswal@maths.iitkgp.ernet.in

R. N. Mohapatra et al. (eds.), *Mathematics and Computing 2013*,
Springer Proceedings in Mathematics & Statistics 91,
DOI: 10.1007/978-81-322-1952-1_7, © Springer India 2014

common form is the hierarchical organization in which a particular level decision maker controls or co-ordinates his/her lower level decision makers. These types of decomposed system is called multi-level system [1]. In a hierarchical decision-making system, if two or more levels are present with one decision maker in each level to take decision, then the problem is called a multi-level programming problem. Multi-level programming can be considered as an extension of Stackelberg games for solving decentralized planning problems with multiple decision makers in a hierarchical organization. Multi-level programming problem (MLPP) can also be defined as a k-person, nonzero sum game with perfect information in which each player moves from top-down sequentially. In a multi-level programming problem, if there are only two levels, then the problem is called a bi-level programming problem. The formal mathematical formulation of bi-level programming problem was studied by Fortuni-Amat and McCarl [13] in 1981 and Candler and Townsley [9] in 1982. After the formulation of the bi-level programming problem it has been extended to formulate a multi-level programming problem. Multi-level programming problem has the following properties (see Shih et al. [21]; Lai [15]):

(i) the decision makers take decision interactively within a predominantly hierarchical structure;
(ii) the decisions are made sequentially from upper level to lower level;
(iii) each level decision maker independently maximize its own net benefits, but their actions are affected by the action of the other decision maker (DM);
(iv) these affects can be reflected in both the objective and the feasible space.

Most of the real-life case study in this area can be found in export-import business, agriculture, government policy, economic systems, finance, warfare, transportation, network designs, and is especially suitable for conflict resolution.

In real-life situations, there exist some decision-making problems with different structures. These type of problems cannot be solved using standard decision-making approaches. For example, suppose in a multi-level nonlinear programming problem all the DM have multiple number of choices for the cost coefficients of their objectives, or for the resource level, then classical method like vertex enumeration, KKT transformation, fuzzy programming, or fuzzy goal programming can not be used directly to solve the problem. To deal with these type of problems, we need to construct a new mathematical model. In this paper, we transform these type of problem into a mathematical model such that these type of situation can be handled. We formulate a model for a multi-level nonlinear programming problem whose parametric space contained some multi-choice type parameter. We use some standard mathematical methods to tackle these multi-choice parameter. Then we solve the problem by standard optimization methods. After presenting the Introduction, some literature reviews on the work are presented in the next Section. Then Mathematical formulation of the problem, Proposed methodology, Numerical example, and conclusions are presented in the subsequent Sections.

2 Literature Review

In last three decades several development has been done in the field of hierarchial optimization. The simplest form of a multi-level programming problem is bi-level programming problem. After the formation of bi-level programming problem by Fortuni-Amat and McCarl [13], Candler and Townsley [9] in several direction research has been done on this topic. Candler and Townsley [9] has proposed an implicit search algorithm which generates an enumerating bases from lower level activities, but no progress has been made for a large system. Then Bialas and Karwan [5, 6] has proposed two methods to solve a bi-level programming problem, they are known as vertex enumeration method and k-th best method. In 1988 Anandalingam [1] discussed multi-level programming problem (MLPP) as well as bi-level decentralized programming problem based on Stackelberg solution procedure. He used KKT transformation technique to solve these multi-level and bi-level decentralized problems. It was Lai [15] who used the concept of fuzzy set theory to solve multi-level programming problem. After Lai's fuzzy set theory concept for multi-level programming Shih et al. [21], Shih and Lee [22] extended his concept by introducing noncompensatory max-min aggregation operator and compensatory fuzzy operator respectively for MLPP. After the development of fuzzy programming approach for hierarchial optimization problem fuzzy goal programming approach has been developed by Pal and Moitra [18] and Pramanik and Roy [19]. Baky [3] has used fuzzy goal programming approach and proposed two algorithm to solve multi-level multiobjective linear programming problem.

In case of multi-choice programming problem, it was Healey [14] who originated the problem.The problem belongs to a class of combinatorial optimization problems with a requirement to choose a value from a number of choice, and to find a combination which optimize an objective function subject to a set of constraints. In practice, MCP can be extended as an application of generalized assignment problems, multiple choice knapsack problems, sales resource allocation, multi-item scheduling, timetabling, etc. Chang [10, 11] has proposed the formulation of multi-choice goal programming(MCGP), in this problem the DM allowed to set multi-choice aspiration levels (MCAL) for each goal. Paksoy and Chang [17] have applied the revised multi-choice goal programming approach of Chang [11] to deal with the multi-choice parameters and solved a supply chain network design problem. Liao [16] follows the method of Chang [10] to solve the multi-segment goal programming problem. Then Biswal and Acharya [7] has extended Chang's method to transform multi-choice programming problem to a deterministic model where right-hand side parameter of linear programming problem is multi-choice type. Further Biswal and Acharya [7] used interpolating polynomial to remove multi-choiceness of the right- hand side parameter of constraints and formulated a mixed integer nonlinear programming problem. Chang et al. [12] have studied multi-coefficient goal programming in their paper and use Chang's [10] transformation technique to deal with multi-choice parameter. There are no decision-making model in the OR literature for solving

multi-level nonlinear programming problem containing some multi-choice parameter. In the next section, we present the mathematical model for the multi-choice multi-level nonlinear programming (MLNLP) problem.

3 Mathematical Formulation of the Problem

Let us consider a multi-level nonlinear programming (MLNLP) problem with K number of decision maker at different levels. Let $X \in \mathbb{R}^n$ be the decision vector of the problem. The decision vector is controlled by the DM and it is partitioned among the DMs. Each level DM control at least one variable of the DV. Suppose $X_k \in \mathbb{R}^{n_k}$ $(k = 1, 2, ..., K)$ is controlled by the k-th level decision maker and $n_1 + n_2 + \cdots + n_K = n$. Also, each level decision makers have their own objective function. We consider the MLNLP problem where the parameters of the objective functions and the constraints are multi-choice type. Let $\bar{F_k} : \mathbb{R}^n \to \mathbb{R}$ $(k = 1, 2, ..., K)$ be the objective function of the k-th level DM, where the parameter of the function are multi-choice type. Hence this type of problem can be presented as:
find $X = (x_1, x_2, ..., x_n)^T$ so as to

$$\begin{aligned} &\max_{X_1} : \bar{F_1}(X) \\ &\max_{X_2} : \bar{F_2}(X) \\ &\vdots \\ &\max_{X_K} : \bar{F_K}(X) \end{aligned} \tag{1}$$

subject to

$$S = \{(x_1, x_2, ..., x_n) | \bar{g_i}(X) \leq 0,\ i = 1, 2, ..., m;\ x_j \geq 0,\ j = 1, 2, ..., n.\} \tag{2}$$

where $X_1 \cup X_2 \cup \ldots \cup X_K = X,\ X_i \cap X_j = \phi,\ i \neq j\ \forall i, j$ and $\bar{g_i} : \mathbb{R}^n \to \mathbb{R}$ $(i = 1, 2, ..., m)$ are the real-valued function where the parameters are multi-choice type.

In the above model, for each multi-choice parameter in $\bar{g_i}$ the feasible region will be different. Also, for each multi-choice parameter in the objective functions $\bar{F_k}$ the value of the functions will be different. Let us consider a multi-level multi-choice linear programming problem where only two parameters are multi-choice type. Suppose there are p alternative choices for first parameter and q alternative choices for the second parameter. Then we have to solve pq different MLPP to obtain the optimal solution. Due to the presence of the multi-choice parameters, the problem cannot be solved directly. Further, the MLPP has to be solved for K different decision makers. Hence, we present a suitable methodology to solve these type of problem.

Table 1 Data table for multi-choice coefficient c_{kj}

u_{kj}	0	1	2	$\cdots$	$s_{kj}-1$
$f_{c_{kj}}(u_{kj})$	$c_{kj}^{(1)}$	$c_{kj}^{(2)}$	$c_{kj}^{(3)}$	$\cdots$	$c_{kj}^{(s_{kj})}$

4 Proposed Methodology

The main difficulties in solving the model problem occurs due to the presence of several multi-choice parameter. To overcome these difficulties, first we tackle these multi-choice parameters. Chang [10, 11], Liao[16], Biswal and Acharya [7, 8] has introduced methods to transform the problem containing multi-choice parameter to a mixed integer programming problem. We formulate the interpolating polynomials for each of the multi-choice parameters present in the problem.

4.1 Formulation of Interpolating Polynomial

For the discussion, let us consider that $C_k(=(c_{k1}, c_{k2}, ...c_{kj}, ...))$ be the vector of multi-choice parameters present in the k level DM's objective function. The set of alternative choices for the parameter c_{kj} be $\{c_{kj}^{(1)}, c_{kj}^{(2)}, ..., c_{kj}^{(s_{kj})}\}$, i.e., there are s_{kj} number of alternative choices available for the parameter c_{kj} out of which we have to select one to obtain optimal solution. Hence, to tackle the multi-choice parameter c_{kj} we introduce an integer variable u_{kj} which takes s_{kj} number of values. We formulate a Lagrange interpolating polynomial $f_{c_{kj}}(u_{kj})$ which passes through all the s_{kj} number of points given by Table 1.

Following Lagrange's formula [2] we get the interpolating polynomial for the multi-choice parameter c_{kj} as:

$$\begin{aligned} f_{c_{kj}}(u_{kj}) = & \frac{(u_{kj}-1)(u_{kj}-2)\cdots(u_{kj}-s_{kj}+1)}{(-1)^{(s_{kj}-1)}(s_{kj}-1)!}c_{kj}^{(1)} \\ & + \frac{u_{kj}(u_{kj}-2)\cdots(u_{kj}-s_{kj}+1)}{(-1)^{(s_{kj}-2)}(s_{kj}-2)!}c_{kj}^{(2)} \\ & + \frac{u_{kj}(u_{kj}-2)(u_{kj}-3)\cdots(u_{kj}-s_{kj}+1)}{(-1)^{(s_{kj}-3)}2!(s_{kj}-3)!}c_{kj}^{(3)} + \cdots \\ & + \frac{u_{kj}(u_{kj}-1)(u_{kj}-2)\cdots(u_{kj}-s_{kj}+2)}{(s_{kj}-1)!}c_{kj}^{(s_{kj})}. \ k=1,2,...,K \quad (3) \end{aligned}$$

Similarly, let us consider that the vector of the multi-choice parameters presents in the i-th constraint be $A_i(=(a_{i1}, a_{i2}, ..., a_{ij}...))$ $(i = 1, 2, ..., m)$. Suppose for the multi-choice parameter a_{ij} there are p_{ij} number of alternative choice. To tackle the multi-choice parameter a_{ij} we introduce an integer variable w_{ij} which takes p_{ij} number of different values, and construct an interpolating polynomial $f_{a_{ij}}(w_{ij})$

Table 2 Data table for multi-choice coefficient a_{ij}

w_{ij}	0	1	2	$\cdots$	$p_{ij}-1$
$f_{a_{ij}}(w_{ij})$	$a_{ij}^{(1)}$	$a_{ij}^{(2)}$	$a_{ij}^{(3)}$	$\cdots$	$a_{ij}^{(p_{ij})}$

following the Lagrange's formula. The interpolating polynomial $f_{a_{ij}}(w_{ij})$ passes through all the p_{ij} number of points which are given by Table 2. The interpolating polynomial can be written as:

$$\begin{aligned} f_{a_{ij}}(w_{ij}) = & \frac{(w_{ij}-1)(w_{ij}-2)\cdots(w_{ij}-p_{ij}+1)}{(-1)^{(p_{ij}-1)}(p_{ij}-1)!}a_{ij}^{(1)} \\ & + \frac{w_{ij}(w_{ij}-2)\cdots(w_{ij}-p_{ij}+1)}{(-1)^{(p_{ij}-2)}(p_{ij}-2)!}a_{ij}^{(2)} \\ & + \frac{w_{ij}(w_{ij}-2)(w_{ij}-3)\cdots(w_{ij}-p_{ij}+1)}{(-1)^{(p_{ij}-3)}2!(p_{ij}-3)!}a_{ij}^{(3)} + \cdots \\ & + \frac{w_{ij}(w_{ij}-1)(w_{ij}-2)\cdots(w_{ij}-p_{ij}+2)}{(p_{ij}-1)!}a_{ij}^{(p_{ij})}, \\ & \qquad i = 1, 2, ..., m. \end{aligned} \tag{4}$$

After transforming all the multi-choice parameters c_{kj}, a_{ij} $(k = 1, 2, ..., K;\ i = 1, 2, ..., m)$ by introducing interpolating polynomial, we obtain a multi-level mixed integer nonlinear programming problem. Hence the transformed multi-level programming model for the problem (1–2) is given by:

$$\begin{aligned} & \max_{X_1} : F_1(X, U_1) \\ & \max_{X_2} : F_2(X, U_2) \\ & \vdots \\ & \max_{X_K} : F_K(X, U_K) \end{aligned} \tag{5}$$

subject to

$$S = \{(x_1, x_2, ..., x_n) | g_i(X, w_{ij}) \leq 0,\ i = 1, 2, ..., m;\ x_j \geq 0,\ j = 1, 2, ..., n.\} \tag{6}$$

$$\begin{aligned} & 0 \leq u_{kj} \leq s_{kj} - 1 \\ & 0 \leq w_{ij} \leq p_{ij} - 1 \\ & u_{kj},\ w_{ij} \in \mathbb{N}_0 \quad k = 1, 2, ..., K;\ i = 1, 2, ..., m. \end{aligned} \tag{7}$$

where $\mathbb{N}_0$ is the set of nonnegative integers and $U_k = (u_{k1}, u_{k2}, ..., u_{kj}, ...)$ $(k = 1, 2, ..., K)$.

4.2 Fuzzy Programming Approach to Solve the Transformed Problem

In the previous Section, we have derived transformed multi-level mixed integer nonlinear programming problem in which there are no more multi-choice parameters. We use fuzzy programming approach to obtain a Pareto optimal solution of the transformed model. Since the feasible region are same for the original and for the transformed model, the obtained solution will be a Pareto optimal solution for the original problem also. In the above MLP problems, K number of different decision makers are there. The objective function for all the decision makers are different and conflicting in nature. So, we have to find a solution which will satisfy all the criteria of the decision makers. To fulfill this demand we have to find a compromise solution which can be achieved by using fuzzy programming approach. To solve the problem (5–7), the first-level DM provides his/her preferred ranges for the objective function Z_{41} and control variable X_1 to the second-level DM and asked for his solution. The second-level DM solves his/her problem in isolation with the additional preference information from the first-level DM and submit the solution to the first-level DM. Then first-level DM will modify the solution under the overall benefit of the organization; this process will continue until DMs get a satisfactory solution. By following the satisfactory solution, both DMs individually rebuild/build the revised ranges for their objective functions and control variable which become the additional constraints of the third-level DM. The solution of the third-level DM is proposed to the upper levels. If any upper levels are not satisfied with this proposal, the third-level DM will then solve a new problem with new information from the upper level DMs until a satisfactory solution is reached. This procedure continues until the K-th level DM's solution satisfies all DMs and the final solution will be a satisfactory solution for the problem (5–7). Fuzzy membership functions have been introduced to represent the ranges given by each level DM for their objective function and the control variable. To formulate the fuzzy programming model for the k-th level DM of the problem we construct the membership function for all the DM.

4.2.1 Construction of the Fuzzy Membership Function

In order to construct the membership function for the DM's objective function, we solve all the DM's problem with their own objective function individually over the same feasible region. After solving all the problems we obtain the optimal solution for the k-th level DM as $(F_k^*; X^{*k}; U_k^*)$, where $U_k = (u_{k1}, u_{k2}, ..., u_{kj}, ...))$ and $X^{*k} (\in \mathbb{R}^n)$ is the optimal solution for the k-th level DM $(k = 1, 2, ..., K)$. To set the minimum tolerance value for the membership function of the objective function of the DM, we construct the $pay\text{-}off$ matrix. Set the minimum tolerance value for the k-th level DM's objective as the minimum value for the objective in the $pay\text{-}off$ matrix, and denote them as F_k' for k-th level DM. Since the objective function are conflicting in nature, $F_k^* \neq F_k'$. The k-th level DM control the decision variable X_k,

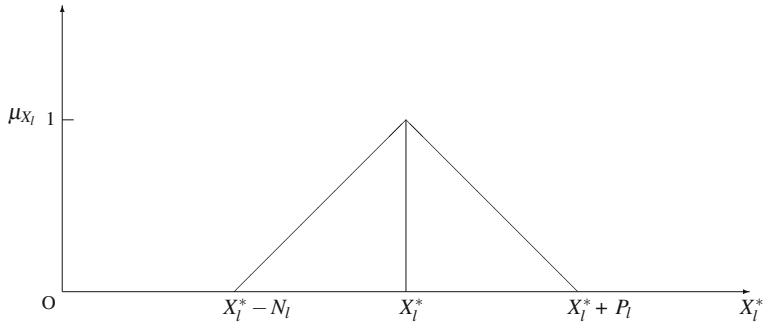


Fig. 1 Triangular membership function for l-th level decision Variable

$k = 1, 2, ..., K$; to get a compromise solution, all the $k-1$ DM have to give a range for their controlled decision variables. Let the positive and negative deviation for X_l be P_l and N_l respectively, where $P_l, N_l \in \mathbb{R}^{n_l}$, $l = 1, 2, ..., k-1$, and they need not be same. Hence we construct the fuzzy membership function for F_k,and X_l as:

$$\mu_{F_k}(X, U_k) = \begin{cases} 1, & F_k \geq F_k^* \\ \frac{F_k - F_k^{'}}{F_k^* - F_k^{'}}, & F_k^{'} < F_k < F_k^* \\ 0, & F_k \leq F_k^{'} \end{cases} \tag{8}$$

$$\mu_{X_l} = \begin{cases} \frac{(X_l^* + P_l) - X_l}{P_l}, & X_l^* \leq X_l \leq X_l^* + P_l \\ \frac{X_l - (X_l^* - N_l)}{N_l}, & X_l^* - N_l \leq X_l \leq X_l^* \end{cases} \tag{9}$$

where $X_l^* \subseteq X^{*l}, l = 1, 2, ..., k-1$. Note that, the membership function corresponding to the objective functions are nonlinear but for the decision variable membership function is linear. The membership function corresponding to the decision variable is shown in Fig. 1.

4.2.2 Fuzzy Programming Model

We define all the membership functions to establish the fuzzy programming model for the k-level problem. Let α_k be the minimum acceptable degree of satisfaction for the objective F_k. Then we have $\mu_{F_k} \geq \alpha_k$. Let $\bar{\gamma_l}$ be the minimum acceptable degree of satisfaction of the decision variable X_l, then we have $\mu_{X_l} \geq \bar{\gamma_l}$. Let us set $\lambda_1 = \min\{\alpha_1, \alpha_2, ..., \alpha_k, \bar{\gamma}_1, \bar{\gamma}_2, ..., \bar{\gamma}_{k-1}\}$. Now we apply Bellman and Zadeh's [4] max-min operator to construct the fuzzy programming model for the problem. The fuzzy programming model for k-th level DM is given by:

$$\max : \lambda_1 \tag{10}$$

subject to,

$$\begin{aligned}
&F_q - \lambda_1(F_q^* - F_q') \geq F_q' \\
&(X_l^* + P_l) - X_l \geq \lambda_1 P_l \\
&X_l - (X_l^* - N_l) \geq \lambda_1 N_l \\
&S = \{(x_1, x_2, ..., x_n) | g_i(X, w_{ij}) \leq 0,\ i = 1, 2, ..., m;\ x_j \geq 0,\ j = 1, 2, ..., n.\} \\
&0 \leq u_{qj} \leq s_{qj} - 1 \\
&0 \leq w_{ij} \leq p_{ij} - 1 \\
&u_{qj}, w_{ij} \in \mathbb{N}_0 \quad q = 1, 2, ..., k;\ i = 1, 2, 3, ..., m.
\end{aligned} \tag{11}$$

where $\mathbb{N}_0$ is the set of nonnegative integer. Also $I_l \in \mathbb{R}^{n_l}$ and all elements of it are 1. This is treated as an mixed integer nonlinear programming problem. Using any nonlinear programming solver, we solve the problem. After obtaining a satisfactory solution for all the k DM, we have to solve the problem for $(k+1)$-th level DM. If any one of the $(k-1)$ upper level DM is not satisfied with the solution, then we rebuild the membership function for the objectives and the decision variables, and again have to solve the reformulated fuzzy programming model for the k-th level DM. This procedure will continue until all the $(K-1)$-th level DMs are satisfied by the K-th level DM's proposal.

5 Numerical Example

In this Section, We present a numerical example to illustrate the solution procedure for a multi-level multi-choice nonlinear programming problem. We consider a tri-level nonlinear programming problem where some of the parameters of the problem are multi-choice type. We consider the following multi-level example:

$$\max_{x_1} : F_1 = \{5, 6, 7\}x_1 + \{2, 4\}x_2 + \{1, 4, 5\}x_3 - x_1^2 - \{5, 7\}x_2^2 - \{2, 3, 4\}x_3^2 \tag{12}$$

$$\max_{x_2} : F_2 = \{2, 4, 6\}x_1 + \{1, 2, 3\}x_2 + \{3, 4\}x_3 - \{1, 2\}x_1^2 - \{1, 3, 5\}x_2^2 - x_3^2 \tag{13}$$

$$\max_{x_3} : F_3 = \{3, 5\}x_1 + 3x_2 + \{2, 3, 4, 6\}x_3 - x_1^2 - \{2, 4\}x_2^2 - x_3^2 \tag{14}$$

subject to,

$$\{1, 3\}x_1 + \{1, 4\}x_2 + x_3 \leq \{7, 9, 10, 12\} \tag{15}$$

$$\{1, 2\}x_1 + 2x_2 + \{2, 4, 5\}x_3 \leq \{8, 9, 11, 12, 15\} \tag{16}$$

$$\{3, 4, 5\}x_1 + \{2, 3, 4, 5\}x_2 + \{3, 4\}x_3 \leq \{20, 21, 23, 24\} \tag{17}$$

$$x_i \geq 0 \qquad i = 1, 2, 3. \tag{18}$$

We formulate interpolating polynomial for each of the discrete parameter present in the problem. We replace those parameters by corresponding interpolating polynomial to obtain a multi-level mixed integer nonlinear programming problem. After solving the transformed model, we obtain the best individual for the DMs as: $F_1^* = 16.175$ at $X^{*1} = (3.5, 0.4, 1.25)$ for the first-level DM, $F_2^* = 15.25$ at $X^{*2} = (3, 1.5, 2)$ for the second-level DM and $F_3^* = 16.375$ at $X^{*3} = (2.5, .75, 3)$ for the third-level DM. From the $pay\text{-}off$ matrix of the objective functions value, we set the minimum tolerance values for all the DM's objective functions as $F_1' = 8.4375$, $F_2' = 13.2275$ and $F_3' = 12.0675$. With these tolerance limits we formulate nonlinear membership functions for all the DM's objective functions (using the formula (8)) and apply fuzzy programming approach to obtain the Pareto optimal solution of the problem. First we solve the transformed for the second level DM.

For second level DM's problem, the first-level DM give the positive and negative deviation for his/her control variable x_1 as $p_1 = 0.5$, $n_1 = 0.3$ and we construct the membership function for x_1 with the help of (9). Hence we construct the fuzzy programming model for second-level DM with the help of (10–11) (with k=2) and solve it using LINGO 11.0 [20]. Then we obtain the solution as $\lambda = 0.7310467$, $x_1 = 3.419314$, $x_2 = 0.9355070$, $x_3 = 1.777556$, $f_1 = 14.17802$; $f_2 = 14.706042$. Suppose, this is a satisfactory solution for first-level DM, then we can proceed to third level.

We keep the bounds for the objective functions same as previous case and let the first-level DM decide to give tolerance limit for $x_1 = 3.419314$ as 0.3 (negative) and 0.5 (positive). Similarly, the second-level DM chooses the tolerance limit for $x_2 = 0.9355070$ as 0.3 (negative) and 0.5 (positive). Hence by solving the fuzzy programming model for the third-level DM, we obtain the satisfactory solution as $x_1 = 3.256412, x_2 = 0.7726045, x_3 = 2.530299$ with $\lambda = 0.4569917$ where $f_1 = 12.143165$; $f_2 = 14.373931$; $f_3 = 15.5812$. Suppose this solution satisfy all the upper level DM then this solution will be a Pareto optimal solution for the problem.

6 Conclusions

In this paper we have presented a multi-level nonlinear programming problem, where some of the parameters are multi-choice type. Interpolating polynomials are used to replace those multi-choice parameters. The transformed problem becomes a multi-level mixed integer nonlinear programming problem which can be solved by using fuzzy programming approach directly. Instead of using interpolating polynomials, if we use auxiliary binary variables for the transformation, the size of the transformed problem will become larger due to the presence of more number of auxiliary binary variables however it takes more number of iteration compare to the previous case. On the basis of the proposed method any DM can use other membership functions (such as piecewise, exponential, hyperbolic functions) to establish the fuzzy programming model.

References

1. Anandalingam, G.: A mathematical programming model of decentralized multi-level systems. J. Oper. Res. Soc. **39**, 1021–1033 (1988)
2. Atkinson, K.E.: An Introduction to Numerical Analysis. Wiley, New York (2009)
3. Baky, I.A.: Solving multi-level multi-objective linear programming problems through fuzzy goal programming approach. Appl. Math. Model. **34**, 2377–2387 (2010)
4. Bellman, R., Zadeh, L.A.: Decision-making in a fuzzy environment. Manage. Sci. **17**, B141–164 (1970)
5. Bialas, W.F., Karwan, M.H.: On two-level optimization. IEEE Trans. Autom. Control **AC-27**, 211–214 (1982)
6. Bialas, W.F., Karwan, M.H.: Two-level linear programming. Manage. Sci. **30**, 1004–1020 (1984)
7. BIswal, M.P., Acharya, S.: Transformation of multi-choice linear programming problem. Appl. Math. Comput. **210**, 182–188 (2009)
8. Biswal, M.P., Acharya, S.: Solving multi-choice linear programming problems by interpolating polynomials. Math. Comput. Model. **54**, 1405–1412 (2011)
9. Candler, W., Townsley, R.: A linear bilevel programming problem. Comput. Oper. Res. **9**, 59–76 (1982)
10. Chang, Ching-Ter: Multi-choice goal programming. Omega Int. J. Manage. Sci. **35**, 389–396 (2007)
11. Chang, Ching-Ter: Revised multi-choice goal programming. Appl. Math. Model. **32**, 2587–2595 (2008)
12. Chang, C.-T., Chen, H.-M., Zhuang, Z.-Y.: Multi-coefficients goal programming. Comput. Ind. Eng. **62**, 616–623 (2012)
13. Fortuni-Amat, J., McCarl, B.: A representation and economic interpretation of a two-level programming problem. J. Oper. Res. Soc. **32**, 783–792 (1981)
14. Healey, W.C.: Multiple choice programming. Oper. Res. **12**, 122–138 (1964)
15. Lai, Y.J.: Hierarchical optimization: a satisfactory solution. Fuzzy Sets Syst **77**, 321–335 (1996)
16. Liao, Chin-Nung: Formulating the multi-segment goal programming. Comput. Ind. Eng. **56**, 138–141 (2009)
17. Paksoy, T., Chang, C.T.: Revised multi-choice goal programming for multi-period, multi-stage inventory controlled supply chain model with popup stores in Guerrilla marketing. Appl. Math. Model. **34**, 3586–3598 (2010)
18. Pal, B.B., Moitra, B.N.: A fuzzy goal programming procedure for solving quadratic bilevel programming problems. Int. J. Intell. Syst. **14**, 89–98 (2003)
19. Pramanik, S., Roy, T.K.: Fuzzy goal programming approach to multi-level programming problem. Eur. J. Oper. Res. **176**, 1151–1166 (2007)
20. Schrage, L.: LINGO release 11.0. LINDO System Inc (2008)
21. Shih, H.-S., Lai, Y.-J., Stanley Lee, E.: Fuzzy approach for multi-level programming problems. Comput. Oper. Res. **23**(1), 73–91 (1996)
22. Shih, H.S., Lee, E.S.: Compensatory fuzzy multiple level decision making. Fuzzy Sets Syst **114**, 71–87 (2000)

Chapter 8
A New Class of Rational Cubic Fractal Splines for Univariate Interpolation

P. Viswanathan and A. K. B. Chand

Abstract Fractal interpolation functions that share smoothness or nonsmoothness property of the prescribed interpolation data provide a novel method of interpolation. The present paper proposes a new type of rational cubic spline fractal interpolation function which involves two families of free shape parameters and which does not require derivatives at knots for its construction. The scaling factors inherent with the structure facilitate the proposed rational fractal interpolation function to recapture a classical rational cubic spline studied earlier in the literature as a special case. In addition, the scaling factors are the key ingredients that provide fractality to the derivative of the constructed interpolant. Thus, in contrast to the classical nonrecursive rational splines, the proposed rational cubic fractal spline can produce interpolants whose derivatives have irregularity in finite or dense subsets of the interpolation intervals depending on the nature of the problem. Assuming that the original data defining function belongs to the smooth class $\mathscr{C}^2$, an upper bound for the interpolation error with respect to the L_∞-norm is obtained and the uniform convergence of the rational cubic fractal interpolant is deduced. The developed rational fractal interpolation scheme is illustrated with numerical examples and some possible extensions are exposed.

1 Introduction

The theory of interpolation, in particular, the spline theory, has evolved beyond its mathematical framework and has become a powerful tool in the applied sciences as well as engineering, computer aided geometric design for instance. The kinds of

P. Viswanathan (✉) · A. K. B. Chand
Department of Mathematics, Indian Institute of Technology Madras, Chennai 600036, India
e-mail: amritaviswa@gmail.com

A. K. B. Chand
e-mail: chand@iitm.ac.in

R. N. Mohapatra et al. (eds.), *Mathematics and Computing 2013*,
Springer Proceedings in Mathematics & Statistics 91,
DOI: 10.1007/978-81-322-1952-1_8,

splines that are widely applied are polynomial splines. For a given set of interpolation data, in general, the polynomial spline is unique, and consequently, local modification of the interpolating curve is impossible. On the other hand, shape modification is a crucial requirement in geometric design environment.

In recent years, rational splines with parameters, where the free parameters can be adjusted so as to yield a variety of interpolating curves for a prescribed data set, have received considerable attention in the literature (see, for instance, [7, 10] and references therein). Wide applicability of rational interpolants may be attributed to their: (i) ability to receive free parameters within the spline structure, (ii) ability to accommodate a wider range of shapes than the polynomial family, (iii) excellent asymptotic properties, (iv) capability to model complicated structures, (v) better interpolation properties, and (vi) excellent extrapolating powers.

These traditional nonrecursive polynomial and rational splines are differentiable indefinite number of times except possibly at a finite number of points in the interpolation interval. Consequently, these techniques do not work satisfactorily for interpolating a dataset wherein variable representing the derivative of a suitable order has to be modeled with a function having irregularity in a dense subset of the interpolation interval. On the other hand, such datasets appear naturally and abundantly in nonlinear and nonequilibrium phenomena, for instance, in electromechanical systems (e.g., a pendulum-cart system [14]) and in fluid dynamics (e.g., falling sphere experiments in polymeric/wormlike micellar fluids [13]).

Using theory of Iterated Function System (IFS), Barnsley [1] proposed Fractal Interpolation Function (FIF), which offers an alternative to the traditional interpolation techniques. FIFs aim mainly at data that exhibit an irregular, nonsmooth structure which cannot be conveniently described using functions occurring in traditional interpolation and approximation theory. Such data arise in the study of real-world sampled signals such as financial series, seismic data, speech signals, and bioelectric recordings. The main differences of a FIF with the traditional interpolants include: (i) the construction via IFS theory that implies a self- similarity in small scales, (ii) the construction by iteration of the functional equation instead of using an analytic formula, (iii) the usage of free parameters termed scaling factors, which offer flexibility in the choice of interpolant in contrast to the unicity of a typical traditional interpolant, and which determine the fractal dimension of the graph of the corresponding interpolant.

We shall supply more particulars—of a technical nature—concerning the notion of fractal interpolation in the next section, soon after we finish discussing this general introduction.

Although fractal interpolants were introduced originally for modeling nonsmooth signals, a little later, Barnsley and Harrington [2] observed that by appropriate choice of the elements in the IFS, smooth FIFs can also be constructed. This observation initiated a striking relationship between the classical splines and the fractal functions. A fractal spline is not a typical fractal, but the adjective *fractal* is retained because (i) of the flavor of the scaling in its definition, (ii) a certain derivative of this function is typically a fractal, and (iii) the graph of a fractal spline is a union of transformed copies of itself. An alternative name could be self-referential splines to alert us to

the fact that graph of such a spline is a union of the transformed copies of itself. Fractal splines constitute an advance in the techniques of approximation, since the classical methods of real-data interpolation can be generalized by means of smooth fractal techniques (see, for instance, [3, 11]). For a fractal spline $f \in \mathscr{C}^p(I)$, where $\mathscr{C}^p(I)$ denotes the space of p-times differentiable real-valued functions defined on a real compact interval I with continuous p-th derivatives, the function $f^{(p)}$ may be nondifferentiable in a finite or dense subset of the interpolation interval. Further, if the experimental data are approximated by a $\mathscr{C}^p$-FIF f, then the fractal dimension of the graph of $f^{(p)}$ provides a quantitative parameter for the analysis of the data, allowing to compare and discriminate the experimental processes [12].

The most widely studied fractal splines so far in the literature are obtained through polynomial IFSs. Recently, by establishing some constrained aspects of cubic Hermite FIFs, the authors have demonstrated that the polynomial FIFs can be explored in the field of constrained interpolation (see [5]). To exploit the advantage of rational functions over polynomials and the versatility of fractal splines, Chand et al. [6, 15] recently constructed rational fractal splines and studied their shape preserving properties. The aforementioned construction requires derivatives at knots as input. Unfortunately, in some manufacturing processes, the derivatives are difficult to obtain.

In the present paper, we introduce a new class of rational FIFs that depend only on function values at knots. For suitable values of the scaling factors, the constructed rational FIFs recover a standard rational cubic spline studied in [8]. To demonstrate the effectiveness of the constructed rational cubic fractal spline f in approximation of a function, an upper bound for the L_∞-norm of the interpolation error $\Phi - f$ for an original data defining function $\Phi \in \mathscr{C}^2$ is obtained. As a consequence, uniform convergence of the constructed fractal spline is deduced. In our test examples, we have compared the plots obtained by the present fractal interpolation scheme and its traditional nonrecursive counterpart. The result is encouraging for the spline class treated now, especially when the data arise from a smooth function whose first derivative has varying irregularity (from smooth to nowhere differentiable).

2 Preliminaries

In this section, we shall briefly recall some basic facts on fractal interpolation. We shall uncover these preludatory materials in three subsections. For a complete and rigorous treatment, the reader is referred to the well-known treatises [1, 2].

2.1 Iterated Function Systems

Definition 1 Let (X, d) be a complete metric space and $M \in \mathbb{N}$. If $f_m : X \to X$, $m = 1, 2, \ldots, M$ are continuous mappings, then $\mathscr{I} = \{X; f_1, f_2, \ldots, f_M\}$ is called an *Iterated Function System* (IFS). If, in addition, there exists a constant $0 \leq c < 1$ such that for all $f_i \in \mathscr{I}$,

$$d\big(f_i(x), f_i(y)\big) \leq cd(x, y) \ \ \forall\, x, y \in I,$$

then $\mathscr{I}$ is called a *hyperbolic* (*contractive*) IFS. The constant c is referred to as the contraction factor of the IFS $\mathscr{I}$.

Associated with the IFS $\mathscr{I}$, there is a set valued map W from the hyperspace $\mathscr{H}(X)$ of nonempty compact subsets of (X, d) into itself. More precisely,

$$W : \mathscr{H}(X) \longrightarrow \mathscr{H}(X), \quad W(E) := \bigcup_{m=1}^{M} f_m(E) \text{ for } E \in \mathscr{H}(X).$$

There exists a natural metric h on $\mathscr{H}(X)$, called the Hausdorff metric, which completes $\mathscr{H}(X)$. This metric h is defined as follows:

$$h(A, B) = \max\Big\{\max_{a\in A}\min_{b\in B} d(a, b), \max_{b\in B}\min_{a\in A} d(b, a)\Big\}.$$

When $\mathscr{I}$ is a contractive IFS with contraction factor c, it is well-known that the collage map W is a contraction on the complete metric space $\big(\mathscr{H}(X), h\big)$ with the same contraction factor c. A basic result in the theory of IFS is the following:

Theorem 1 (Barnsley [1]) *Given a contractive IFS $\mathscr{I}$ on a complete metric space (X, d) and any set $A_0 \in \mathscr{H}(X)$, there exists a unique set A, called the attractor of the hyperbolic IFS, such that*

$$A = \lim_{n\to\infty} W^n(A_0) \ \text{ and } \ W(A) = A.$$

Here the limit is taken in the Hausdorff metric and W^n denotes the n-fold composition of W with itself.

Next the question of how to obtain functions whose graphs are attractors of suitable IFSs is investigated.

2.2 Fractal Interpolation Functions

Let $N \in \mathbb{N}$, $N > 2$. Let $x_1 < x_2 < \cdots < x_N$ be real numbers and a set of data points $\{(x_n, y_n) \in I \times \mathbb{R} : \ n = 1, 2, \ldots, N\}$ be given. Set $I = [x_1, x_N] = [a, b]$ and $I_n = [x_n, x_{n+1}]$ for $n \in J = \{1, 2, \ldots, N-1\}$. Suppose $L_n : I \to I_n$ are contraction homeomorphisms such that

$$L_n(x_1) = x_n, \ L_n(x_N) = x_{n+1}. \tag{1}$$

For instance, if $L_n(x) = a_n x + b_n$, then the prescription in (1) yields

$$a_n = \frac{x_{n+1} - x_n}{x_N - x_1}, \ \ b_n = \frac{x_N x_n - x_1 x_{n+1}}{x_N - x_1}, \ \ n \in J. \tag{2}$$

Let $0 < r_n < 1, n \in J$, and $K = I \times D$, where D is a large enough compactum, i.e., a compact connected subset of $\mathbb{R}$. Let $N-1$ continuous mappings $F_n : K \to \mathbb{R}$ be given satisfying:

$$F_n(x_1, y_1) = y_n, \quad F_n(x_N, y_N) = y_{n+1}, \quad |F_n(x, y) - F_n(x, y^*)| \leq r_n |y - y^*|. \tag{3}$$

Now define functions $w_n : K \to K, \ w_n(x, y) = \big(L_n(x), F_n(x, y)\big) \ \forall \ n \in J$.

Proposition 1 (Barnsley [1]) *The IFS* $\{K; w_n : n \in J\}$ *defined above admits a unique attractor* G, *and* G *is the graph of a continuous function* $f : I \to \mathbb{R}$ *which obeys* $f(x_n) = y_n$ *for* $n = 1, 2, \ldots, N$.

Definition 2 The function f whose graph is the attractor of an IFS as described in the above proposition is called a *Fractal Interpolation Function* (FIF) corresponding to the IFS $\{K; \ w_n : n \in J\}$.

Let us provide some excerpts from the proof of the above proposition, which is needed in the sequel.
Let $\mathscr{F}$ be the set of continuous functions $g : I \to \mathbb{R}$ such that $g(x_1) = y_1$ and $g(x_N) = y_N$. Then $\mathscr{F}$ endowed with the uniform metric $d_\infty(g, h) = \max\{|g(x) - h(x)| : x \in I\}$ is a complete metric space. Define the Read-Bajraktarević operator $T : \mathscr{F} \to \mathscr{F}$ by

$$(Tg)(x) = F_n\left(L_n^{-1}(x), g \circ L_n^{-1}(x)\right) \ \forall \ x \in I_n, \ n \in J.$$

Then, T is a contraction mapping on $(\mathscr{F}, d_\infty)$, i.e.,

$$d_\infty(Tg, Tg^*) \leq |r|_\infty \, d_\infty(g, g^*),$$

where $|r|_\infty := \max\{r_n : n \in J\} < 1$. Hence, by the Banach fixed point theorem, T possesses a unique fixed point on $\mathscr{F}$, that is to say, there is a unique $f \in \mathscr{F}$ such that $(Tf)(x) = f(x) \ \forall \ x \in I$. The aforementioned function f is the FIF corresponding to the IFS $\{K; \ w_n : n \in J\}$, and it satisfies the functional equation:

$$f(x) = F_n\left(L_n^{-1}(x), f \circ L_n^{-1}(x)\right), \ x \in I_n, \ n \in J.$$

The most extensively studied FIFs in theory and applications so far are defined by the iterated mappings:

$$L_n(x) = a_n x + b_n, \ F_n(x, y) = \lambda_n y + R_n(x), \ n \in J. \tag{4}$$

Here a_n and b_n are given by (2), $-1 < \lambda_n < 1$ and $R_n : I \to \mathbb{R}$ are suitable continuous functions such that the conditions specified in (3) are satisfied. The parameter λ_n is called a scaling factor of the transformation w_n, and $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_{N-1})$ is the scale vector corresponding to the IFS. The properties such as smoothness, shape,

and fractal dimension of the FIF depends heavily on the scaling factors $\lambda_n, n \in J$. The function f obtained through the IFS (4) is, in general, nonsmooth and possesses noninteger Hausdorff dimension.

2.3 Differentiable FIFs (Fractal Splines)

For a prescribed set of data, a FIF with $\mathscr{C}^p$-continuity is obtained as the fixed point of IFS (4), where the scaling factors λ_n and the functions R_n are chosen according to the following proposition.

Proposition 2 (Barnsley and Harrington [2]) *Let $x_1 < x_2 < \cdots < x_N$ and $L_n(x) = a_n x + b_n, n \in J$, satisfy (1). Let $F_n(x, y) = \lambda_n y + R_n(x), n \in J$, satisfy (3). Suppose that for some integer $p \geq 0$, $|\lambda_n| \leq \kappa a_n^p$, $0 < \kappa < 1$, and $R_n \in \mathscr{C}^p(I), n \in J$. Let*

$$F_{n,k}(x, y) = \frac{\lambda_n y + R_n^{(k)}(x)}{a_n^k}, \quad y_{1,k} = \frac{R_1^{(k)}(x_1)}{a_1^k - \lambda_1}, \quad y_{N,k} = \frac{R_{N-1}^{(k)}(x_N)}{a_{N-1}^k - \lambda_{N-1}}, \quad k = 1, 2, \ldots, p.$$

If $F_{n-1,k}(x_N, y_{N,k}) = F_{n,k}(x_1, y_{1,k})$ for $n = 2, 3, \ldots, N-1$ and $k = 1, 2, \ldots, p$, then the IFS $\left\{I \times \mathbb{R}; \left(L_n(x), F_n(x, y)\right) : n \in J\right\}$ determines a FIF $f \in \mathscr{C}^p[x_1, x_N]$, and $f^{(k)}$ is the FIF determined by the IFS $\left\{I \times \mathbb{R}; \left(L_n(x), F_{n,k}(x, y)\right) : n \in J\right\}$ for $k = 1, 2, \ldots, p$.

Armed with these requisite general material, we proceed to the next section wherein we develop a new class of rational cubic spline FIFs.

3 Construction of Rational Cubic Fractal Spline with Linear Denominator

Given a set of interpolation data $\{(x_n, y_n) : n = 1, 2, \ldots, N\}$, we consider the IFS given in (4) with $R_n(x) = \frac{P_n(x)}{Q_n(x)}$. Here, for $n \in J$, $P_n(x)$ are cubic polynomials whose coefficients are to be determined using interpolation conditions and $Q_n(x)$ are preassigned linear polynomials. The construction of corresponding FIF $f \in \mathscr{C}^1(I)$ is enunciated in the following theorem.
For convenience in writing the formulas which enter into the theorem, let us denote: $h_n = x_{n+1} - x_n$, $\Delta_n = \frac{y_{n+1} - y_n}{h_n}$, and $|\lambda|_\infty = \max\{|\lambda_n| : n \in J\}$.

Theorem 2 *Let $\{(x_n, y_n) : n = 1, 2, \ldots, N+1\}$ be a given set of data points and $\{(x_n, y_n) : n = 1, 2, \ldots, N\}$ be the set of interpolation points, where $x_1 < x_2 < \cdots < x_{N+1}$. Consider the rational IFS $\{I \times \mathbb{R}; \left(L_n(x), F_n(x, y)\right) : n \in J\}$, where $L_n(x) = a_n x + b_n$ and $F_n(x, y) = \lambda_n y + R_n(x)$, $|\lambda_n| \leq \kappa a_n$, $0 < \kappa < 1$, $n \in J$. Further, let $R_n(x) = \frac{P_n(x)}{Q_n(x)}$, where $P_n(x)$ is a cubic polynomial and $Q_n(x) \neq 0$ (for all $x \in I = [x_1, x_N]$) is a preassigned linear polynomial involving shape parameters*

such that $F_n(x_1, y_1) = y_n$, $F_n(x_N, y_N) = y_{n+1}$ *are satisfied. With* $F_{n,1}(x, y) = \frac{\lambda_n y + R_n^{(1)}(x)}{a_n}$, *let* $F_{n,1}(x_1, \Delta_1) = \Delta_n$ *and* $F_{n,1}(x_N, \Delta_N) = \Delta_{n+1}$. *Then a* $\mathscr{C}^1$*-rational cubic spline FIF* f *satisfying* $f(x_n) = y_n$, $f^{(1)}(x_n) = \Delta_n$, $n = 1, 2, \ldots, N$ *exists, and it is unique for a fixed choice of the shape parameters and the scaling factors.*

Proof Consider the IFS $\mathscr{I} = \big\{I \times \mathbb{R}; \big(L_n(x), F_n(x, y)\big) : n \in J\big\}$, where $L_n(x) = a_n x + b_n$ satisfy the prescription in (1), and $F_n(x, y) = \lambda_n y + R_n(x)$ fulfill the conditions $F_n(x_1, y_1) = y_n$, $F_n(x_N, y_N) = y_{n+1}$. Let $R_n(x) \equiv R_n\big(x_1 + \theta(x_N - x_1)\big) = \frac{P_n(x)}{Q_n(x)}$, where

$$\begin{aligned} P_n(x) &\equiv P_n(x_1 + \theta(x_N - x_1)) \\ &= A_{1n}(1-\theta)^3 + A_{2n}\theta(1-\theta)^2 + A_{3n}\theta^2(1-\theta) + A_{4n}\theta^3, \\ Q_n(x) &\equiv Q_n(x_1 + \theta(x_N - x_1)) = \alpha_n(1-\theta) + \beta_n\theta. \end{aligned}$$

The constants α_n and β_n are free parameters that can be utilized for shape modification and shape control of the fractal interpolant. We impose the conditions $\alpha_n > 0$ and $\beta_n > 0$ so as to ensure strict positivity of Q_n, which in turn avoid any singularity of the rational expression R_n. It is worthwhile to mention that our strategy of preassigning the denominator polynomial Q_n, and determining only the numerator polynomial P_n via interpolation conditions avoids the possibility of nonlinearity in the system governing the coefficients of the rational expression.

Consider $\mathscr{F} := \{g \in \mathscr{C}(I) |\ g(x_1) = y_1 \text{ and } g(x_N) = y_N\}$ equipped with the uniform metric d_∞. The IFS $\mathscr{I}$ induces a contraction map $T : \mathscr{F} \to \mathscr{F}$, $g \mapsto Tg$, $(Tg)\big(L_n(x)\big) := F_n\big(x, g(x)\big)$, $x \in I$, whose contraction factor is $|\lambda|_\infty$. The contraction map T has a unique fixed point $f \in \mathscr{F}$, which obeys:

$$\begin{aligned} f\big(L_n(x)\big) &= F_n\big(x, f(x)\big), \\ &= \lambda_n f(x) + \frac{A_{1n}(1-\theta)^3 + A_{2n}\theta(1-\theta)^2 + A_{3n}\theta^2(1-\theta) + A_{4n}\theta^3}{\alpha_n(1-\theta) + \beta_n\theta}. \end{aligned} \tag{5}$$

The conditions $F_n(x_1, y_1) = y_n$, $F_n(x_N, y_N) = y_{n+1}$ can be reformulated as the interpolation and continuity conditions $f(x_n) = y_n$, $f(x_{n+1}) = y_{n+1}$, $n \in J$. In view of (1), the functional Eq. (5) with $x = x_1$ yields

$$f\big(L_n(x_1)\big) = \lambda_n f(x_1) + \frac{P_n(x_1)}{Q_n(x_1)} \Longrightarrow y_n = \lambda_n y_1 + \frac{A_{1n}}{\alpha_n} \Longrightarrow A_{1n} = (y_n - \lambda_n y_1)\alpha_n.$$

Similarly, substituting $x = x_N$ in (5) and using (1), we obtain

$$A_{4n} = (y_{n+1} - \lambda_n y_N)\beta_n.$$

Now we make $f \in \mathscr{C}^1(I)$ by imposing the conditions prescribed in Barnsley-Harrington theorem (see Proposition 2).

By hypothesis, $|\lambda_n| \leq \kappa a_n$, $n \in J$, where $0 < \kappa < 1$. We also have $R_n \in \mathscr{C}^1(I)$. Adhering to the notation of Proposition 2, for $n \in J$, we let:

$$F_{n,1}(x, y) = \frac{\lambda_n y + R_n^{(1)}(x)}{a_n},$$
$$y_{1,1} = \Delta_1,\ y_{N,1} = \Delta_N,\ F_{n,1}(x_1, \Delta_1) = \Delta_n,\ F_{n,1}(x_N, \Delta_N) = \Delta_{n+1}.$$

Then by Proposition 2, the FIF $f \in \mathscr{C}^1(I)$. Further, $f^{(1)}$ is the fractal function determined by the IFS $\mathscr{I}^* \equiv \big\{I \times \mathbb{R}; \big(L_n(x), F_{n,1}(x, y)\big) : n \in J\big\}$. Consider $\mathscr{F}^* := \{g \in \mathscr{C}(I) : g(x_1) = \Delta_1 \text{ and } g(x_N) = \Delta_N\}$ endowed with the uniform metric. The IFS $\mathscr{I}^*$ induces a contraction map $T^* : \mathscr{F}^* \to \mathscr{F}^*$ defined by $(T^*g^*)\big(L_n(x)\big) = F_{n,1}\big(x, g^*(x)\big),\ x \in I$. By Proposition 2, the fixed point of T^* is $f^{(1)}$. Consequently, $f^{(1)}$ obeys the functional equation:

$$f^{(1)}\big(L_n(x)\big) = F_{n,1}\big(x, f^{(1)}(x)\big) = \frac{\lambda_n f^{(1)}(x) + R_n^{(1)}(x)}{a_n}. \tag{6}$$

The conditions $F_{n,1}(x_1, \Delta_1) = \Delta_n$ and $F_{n,1}(x_N, \Delta_N) = \Delta_{n+1}$ can be reformulated as follows: $f^{(1)}(x_n) = \Delta_n$ and $f^{(1)}(x_{n+1}) = \Delta_{n+1}$, $n \in J$. Now from (6) and (1), we have

$$f^{(1)}(L_n(x_1)) = \frac{\lambda_n f^{(1)}(x_1) + \frac{Q_n(x_1)P_n^{(1)}(x_1) - P_n(x_1)Q_n^{(1)}(x_1)}{Q_n^2(x_1)}}{a_n}$$
$$\Longrightarrow A_{2n} = (2\alpha_n + \beta_n)y_n + \alpha_n h_n \Delta_n - \lambda_n\big[(2\alpha_n + \beta_n)y_1 + \alpha_n(x_N - x_1)\Delta_1\big].$$

Finally, substituting $x = x_N$ in the functional Eq. (6) and using (1), we get

$$A_{3n} = (2\beta_n + \alpha_n)y_{n+1} - h_n\beta_n\Delta_{n+1} - \lambda_n\big[(2\beta_n + \alpha_n)y_N - \beta_n(x_N - x_1)\Delta_N\big].$$

Therefore, with $\theta := \frac{x - x_1}{x_N - x_1}$, the desired rational cubic spline FIF receives the form:

$$f\big(L_n(x)\big) = \lambda_n f(x) + \frac{P_n(x)}{Q_n(x)}, \quad x \in I, n \in J, \tag{7}$$
$$\begin{aligned} P_n(x) = {} & (y_n - \lambda_n y_1)\alpha_n(1-\theta)^3 + (y_{n+1} - \lambda_n y_N)\beta_n\theta^3 + \big\{(2\alpha_n + \beta_n)y_n + \alpha_n h_n \Delta_n \\ & - \lambda_n[(2\alpha_n + \beta_n)y_1 + \alpha_n(x_N - x_1)\Delta_1]\big\}\theta(1-\theta)^2 + \big\{(2\beta_n + \alpha_n)y_{n+1} \\ & - h_n\beta_n\Delta_{n+1} - \lambda_n[(2\beta_n + \alpha_n)y_N - \beta_n(x_N - x_1)\Delta_N]\big\}\theta^2(1-\theta), \\ Q_n(x) = {} & \alpha_n(1-\theta) + \beta_n\theta. \end{aligned}$$

Since the FIF f in (7) is obtained as a solution of the fixed point equation $Tg = g$, it is unique for a fixed choice of the scaling factors and the shape parameters. □

Remark 1 If $\lambda_n = 0$ for all $n \in J$, then the rational spline FIF defined above reduces to the classical rational cubic interpolant C introduced by Qi Duan et al. [8], which can be defined on $[x_n, x_{n+1}]$ as follows:

$$C(x) = \frac{U_n(\theta)}{V_n(\theta)}, \quad \theta = \frac{x - x_n}{x_{n+1} - x_n}, \tag{8}$$
$$U_n(\theta) = (1-\theta)^3\alpha_n y_n + \theta(1-\theta)^2[(2\alpha_n + \beta_n)y_n + \alpha_n h_n \Delta_n] + \theta^2(1-\theta)[(2\beta_n + \alpha_n)y_{n+1} - h_n\beta_n\Delta_{n+1}] + \theta^3\beta_n y_{n+1},$$
$$V_i(\theta) = \alpha_n(1-\theta) + \beta_n\theta.$$

4 Convergence Analysis

Among various considerations in the analysis of an interpolation method, accuracy of the interpolant is of major importance. Effectiveness of the rational cubic spline FIF in approximation of a function can be explained by studying its convergence properties. This section is devoted to establish a uniform error bound for the rational cubic spline FIF constructed in the foregoing section. We shall demonstrate that for suitable choices of parameters (scaling factors and shape parameters) in the rational IFS, our rational cubic spline FIF converges to the data generating function $\Phi \in \mathscr{C}^2(I)$ at least as rapidly as the square of the mesh norm approaches zero. The equality of the knot spacing is supposed here for the sake of simplicity. As a first step towards establishing the desired error bound for the rational cubic spline FIF, we have the following error bound for its classical counterpart C:

Lemma 1 (Qi. Duan et al. [8]) *Let $\Phi \in \mathscr{C}^2(I)$ and C be the piecewise rational cubic interpolant [cf. (8)] for the data $\{(x_n, y_n) : n = 1, 2, \ldots, N\}$ where $y_n = \Phi(x_n)$. Then for $x \in [x_n, x_{n+1}]$,*

$$|\Phi(x) - C(x)| \le \frac{c_n h_n^2}{2}\|\Phi^{(2)}\|_\infty,$$

where $c_n = \max\limits_{0\le\theta\le1} \frac{w(\alpha_n,\beta_n,\theta)}{v(\alpha_n,\beta_n,\theta)}$,

$$w(\alpha_n, \beta_n, \theta) = \theta\Big[(\alpha_n\beta_n - 3\beta_n^2)\theta^4 + (7\beta_n^2 - 5\alpha_n\beta_n + \alpha_n^2)\theta^3 - (4\beta_n^2 - 7\alpha_n\beta_n + 3\alpha_n^2)\theta^2 + 3(\alpha_n^2 - \alpha_n\beta_n)\theta - \alpha_n^2\Big],$$
$$v(\alpha_n, \beta_n, \theta) = \big[(1-\theta)\alpha_n + \beta_n\theta\big]\big[\beta_n\theta^2 + (\alpha_n - 2\beta_n)\theta - \alpha_n\big].$$

Moreover, for any $\alpha_n > 0$, $\beta_n > 0$, c_n is bounded and $\frac{1}{4} \le c_n \le \max\limits_{0\le\theta\le1} \frac{3\theta^3 - 7\theta^2 + 4\theta}{2-\theta} =$ 0.42330428...

This next lemma establishes an upper bound for the uniform distance between the rational cubic spline FIF f and its classical counterpart C.

Lemma 2 *Let f and C, respectively, be the rational cubic spline FIF and the classical cubic spline interpolant for the data set $\{(x_n, y_n) : n = 1, 2, \ldots, N\}$. Let the*

rational function $R_n(\lambda_n, x) = \frac{P_n(\lambda_n,x)}{Q_n(\lambda_n,x)}$ *appearing in the functional equation of the FIF* f *satisfy* $\left|\frac{\partial R_n}{\partial \lambda_n}(\tau_n, x)\right| \leq D_0 \ \forall\ (\tau_n, x) \in (-a_n, a_n) \times I_n$ *and* $n \in J$. *Then,* $\|f - C\|_\infty \leq \frac{|\lambda|_\infty}{1-|\lambda|_\infty}\big(\|C\|_\infty + D_0\big)$.

Proof The rational cubic spline FIF $f \in \mathscr{C}^1(I)$ is the fixed point of the Read-Bajraktarević operator T_λ defined on the space $\mathscr{F} = \{g \in \mathscr{C}^1(I) \mid g(x_1) = y_1, g(x_N) = y_N, g^{(1)}(x_1) = \Delta_1, \text{and } g^{(1)}(x_N) = \Delta_N\}$ such that

$$T_\lambda g(x) = \lambda_n g\big(L_n^{-1}(x)\big) + R_n\big(\lambda_n, L_n^{-1}(x)\big), \tag{9}$$

where $R_n\big(\lambda_n, L_n^{-1}(x)\big) = \frac{P_n(\lambda_n,\theta)}{Q_n(\theta)}$, $\theta = \frac{L_n^{-1}(x)-x_1}{x_N-x_1} = \frac{x-x_n}{h_n}$, $x \in I_n$. Here we write $R_n(\theta) \equiv R_n(\lambda_n, \theta)$ in order to put in evidence that the coefficients of the rational function depend on the scaling factor λ_n. Let $\Lambda = [-\kappa a_1, \kappa a_1] \times [-\kappa a_2, \kappa a_2] \times \ldots \times [-\kappa a_{N-1}, \kappa a_{N-1}]$, $0 < \kappa < 1$. For a given $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_{N-1}) \in \Lambda \subset \mathbb{R}^{N-1}$ with at least one $\lambda_n \neq 0$, the rational cubic spline FIF f is the fixed point of T_λ, and for $\mathbf{0} = (0, 0, ..., 0) \in \Lambda$, the classical rational quadratic C is the fixed point of $T_{\mathbf{0}}$. For $x \in [x_n, x_{n+1}]$, we obtain

$$\begin{aligned}\Big|T_\lambda C(x) - T_{\mathbf{0}} C(x)\Big| &= \Big|\lambda_n C\big(L_n^{-1}(x)\big) + R_n\big(\lambda_n, L_n^{-1}(x)\big) - R_n\big(0, L_n^{-1}(x)\big)\Big|,\\ &\leq |\lambda_n|\|C\|_\infty + |\lambda_n|\left|\frac{\partial R_n(\tau_n, L_n^{-1}(x))}{\partial \lambda_n}\right|,\\ &\leq |\lambda_n|\big(\|C\|_\infty + D_0\big).\end{aligned}$$

The first step in the preceding analysis used definition of the map T, the second step followed from the mean value theorem, and the last step was plain due to the definition of D_0. Validity of the above inequality for all $x \in I_n$, $n \in J$, asserts

$$\|T_\lambda C - T_{\mathbf{0}} C\|_\infty \leq |\lambda|_\infty\big(\|C\|_\infty + D_0\big). \tag{10}$$

Inasmuch as T_λ is a contraction map, we have

$$\|T_\lambda f - T_\lambda C\|_\infty \leq |\lambda|_\infty \|f - C\|_\infty. \tag{11}$$

In view of (10)–(11),

$$\begin{aligned}\|f - C\|_\infty = \|T_\lambda f - T_{\mathbf{0}} C\|_\infty &\leq \|T_\lambda f - T_\lambda C\|_\infty + \|T_\lambda C - T_{\mathbf{0}} C\|_\infty,\\ &\leq |\lambda|_\infty \|f - C\|_\infty + |\lambda|_\infty\big(\|C\|_\infty + D_0\big),\end{aligned}$$

which on simplification yields

$$\|f - C\|_\infty \leq \frac{|\lambda|_\infty\big(\|C\|_\infty + D_0\big)}{1 - |\lambda|_\infty}.$$

This completes the proof. □

Theorem 3 *Let $\Phi \in \mathscr{C}^2(I)$ and let f be the rational cubic spline FIF such that $f(x_n) = \Phi(x_n) = y_n$ and $f^{(1)}(x_n) = \Delta_n, n = 1, 2, \ldots, N$. Then*

$$\|\Phi - f\|_\infty \leq \frac{ch^2}{2}\|\Phi^{(2)}\|_\infty + \frac{|\lambda|_\infty}{1 - |\lambda|_\infty}\big(\|C\|_\infty + D_0\big),$$

where $c = \max\{c_n : n \in J\}$, c_n, $n \in J$, are as in Lemma (1)*, and D_0 is prescribed in Lemma* (2)*.*

Proof By triangle inequality:

$$\|\Phi - f\|_\infty \leq \|\Phi - C\|_\infty + \|C - f\|_\infty. \tag{12}$$

First summand in the right-hand side of the above inequality can be bounded using Lemma 1 as follows.

$$\|\Phi - C\|_\infty \leq \frac{ch^2}{2}\|\Phi^{(2)}\|_\infty. \tag{13}$$

Similarly, the uniform distance between the rational cubic spline FIF f and its classical counterpart C present in the second summand can be bounded by Lemma 2:

$$\|f - C\|_\infty \leq \frac{|\lambda|_\infty}{1 - |\lambda|_\infty}(\|C\|_\infty + D_0). \tag{14}$$

Substitution of (13)–(14) in (12) completes the proof. □

As consequences of the previous theorem we have the following convergence results.

4.1 Convergence Results

(a) For the scaling factors satisfying $|\lambda_n| < a_n = \frac{h_n}{x_N - x_1}$, the rational cubic spline FIF f uniformly converges to the data defining function $\Phi \in \mathscr{C}^2(I)$ as the mesh norm approaches zero, i.e., $\|\Phi - f\|_\infty = O(h)$.

(b) If the scaling factors are selected such that $|\lambda_n| < a_n^2$, then we have $|\lambda|_\infty < \frac{h^2}{|I|^2}$. Consequently, the estimate of the uniform error bound for the rational cubic spline FIF obtained in the above theorem provides $\|\Phi - f\|_\infty = O(h^2)$. Thus, for suitable values of the scaling factors, the rational cubic spline FIF f has the same order of convergence as that of its classical counterpart C.

Remark 2 Since the rational cubic spline FIF does not possess a closed-form expression, the standard methods such as Taylor series analysis, Cauchy remainder formula, and Peano kernel theorem cannot be easily adapted for analyzing its convergence.

Instead, we have used the error bound for the classical rational cubic spline via the triangle inequality $\|\Phi - f\|_\infty \leq \|\Phi - C\|_\infty + \|C - f\|_\infty$ to establish that f has the same order of convergence as that of its classical counterpart C. As consequence, it is possible to approximate any regular function by using a rational cubic spline FIF with arbitrary accuracy. The application of triangle inequality to obtain the convergence does not imply that the error committed by the rational cubic spline FIF in approximating a smooth function will be always greater than the error committed by its classical counterpart. Furthermore, we feel that any possible loss of precision is to be counterbalanced with the generality offered by the method.

5 Numerical Examples and Discussion

Consider the data set $\{(1, 24), (2, 2.5), (4, 41), (5, 4), (7, 57), (8, 5), (9, 0.5), (10, 2.5)\}$, where it is required to interpolate the subset $\{(1, 24), (2, 2.5), (4, 41), (5, 4), (7, 57), (8, 5), (9, 0.5)\}$. Due to the principle of construction of a $\mathscr{C}^1$-FIF, we take $|\lambda_n| < a_n$ for $n = 1, 2, 3, \ldots, 6$. The graph of the desired $\mathscr{C}^1$-rational cubic spline FIF f is obtained as the fixed point of the IFS:

$$\{I \times \mathbb{R}; w_n(x, y) = \big(L_n(x), F_n(x, y)\big) : n = 1, 2, \ldots, 6\}, \tag{15}$$

where with $\theta := \frac{x - x_1}{x_N - x_1}$, the component maps are given by $L_n(x) = x_n(1-\theta) + x_{n+1}\theta$ and

$$\begin{aligned} F_n(x, y) = \big(\alpha_n(1 - \theta) + \beta_n\theta\big)^{-1} \times \Big[& (y_n - \lambda_n y_1)\alpha_n(1 - \theta)^3 + (y_{n+1} \\ & - \lambda_n y_N)\beta_n\theta^3 + \big\{(2\alpha_n + \beta_n)y_n + \alpha_n h_n \Delta_n - \lambda_n[(2\alpha_n + \beta_n)y_1 \\ & + \alpha_n(x_N - x_1)\Delta_1]\big\} \times \theta(1 - \theta)^2 + \big\{(2\beta_n + \alpha_n)y_{n+1} - h_n\beta_n\Delta_{n+1} \\ & - \lambda_n[(2\beta_n + \alpha_n)y_N - \beta_n(x_N - x_1)\Delta_N]\big\}\theta^2(1 - \theta)\Big] + \lambda_n y. \end{aligned}$$

We have written a simple computer program in MatLab for plotting the graphs of FIFs. One inputs the data points, scaling factors, and shape parameters whereupon points on the graph are recursively generated. Theoretically, to obtain the actual fractal interpolant, one needs to continue the iterations indefinitely. However, in practice, computation is very fast; a good view of the whole function is quickly obtained and may be printed with a graphics printer.

Due to the implicit and recursive nature of the FIF, perturbation in the scaling factor or shape parameters in a particular subinterval may influence the entire configuration. However, through various test examples, we observed that the portions of the interpolating curve pertaining to other subintervals are not extremely sensitive toward changes in the parameters of a particular subinterval.

To illustrate this, we take the rational cubic spline FIF in Fig. 1, as a reference curve, and analyze the effects of perturbing the parameters of a particular segment

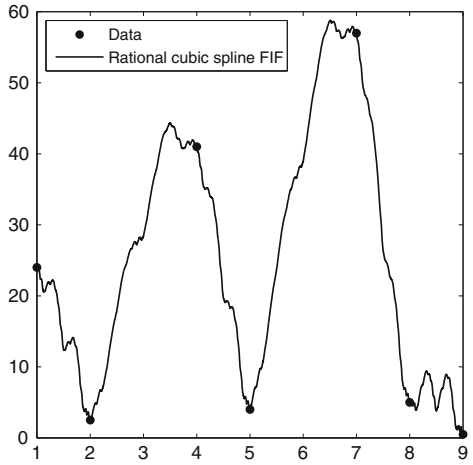


Fig. 1 A rational cubic spline FIF

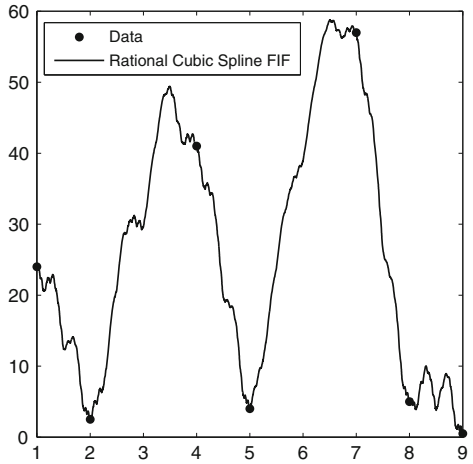


Fig. 2 Effects of λ_2

of this curve. Our reference curve is obtained by iterating the IFS (15), where the scaling factors are $\lambda_n = 0.12$ and the shape parameters are $\alpha_n = \beta_n = 0.5$ for $n = 1, 2, \ldots, 6$. By changing λ_2 to 0.2 and keeping other parameters as in Fig. 1, we obtain the FIF displayed in Fig. 2. It can be observed that the perturbation in λ_2 affects the rational fractal interpolant in the interval $[x_2, x_3]$, whereas there are no perceptible changes in other subintervals. In Fig. 3 we display the fractal spline with spline parameters same as in our reference curve except for $\lambda_3 = 0$. Next we change λ_4 to 0.05 with respect to the reference curve and iterate the IFS code to obtain Fig. 4. By comparing Fig. 4 with Fig. 1, it can be observed that the changes in λ_4 also produce local effects.

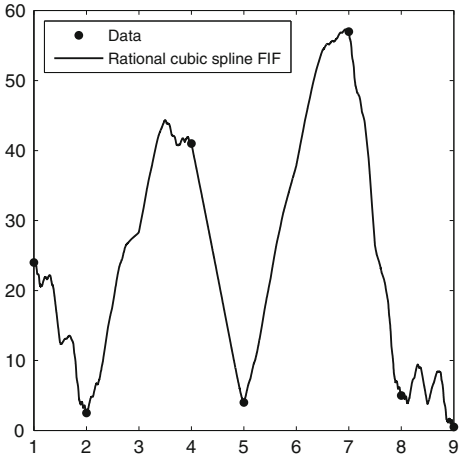


Fig. 3 Effects of λ_3

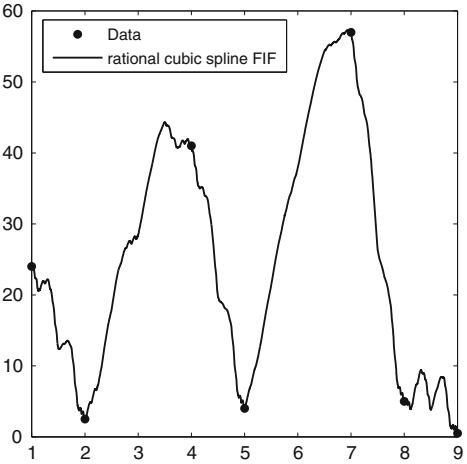


Fig. 4 Effects of λ_4

Similar experiments can be conducted by changing other parameters, and it can be observed that the changes in the parameters pertaining to a particular subinterval do not produce considerable effects in the other subintervals. To be more precise, since the completely local classical rational spline C is a particular case of the proposed rational FIF (obtained when scaling in each subinterval is taken as zero), the fractal scheme is local or global depending on the magnitude of the scaling in each subinterval.

We recover a classical rational cubic spline C in Fig. 5 by iterating the functional Eq. (7) with the scaling factor in each subinterval as zero and the shape parameters as

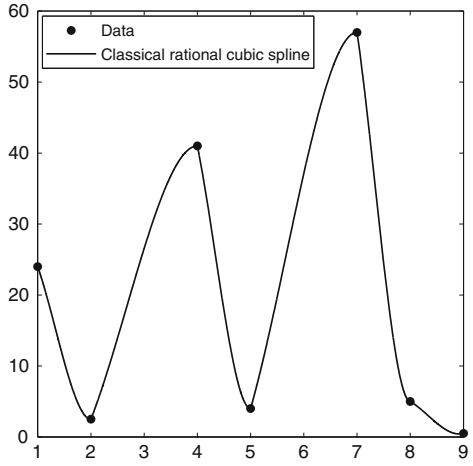


Fig. 5 Classical rational cubic spline interpolant

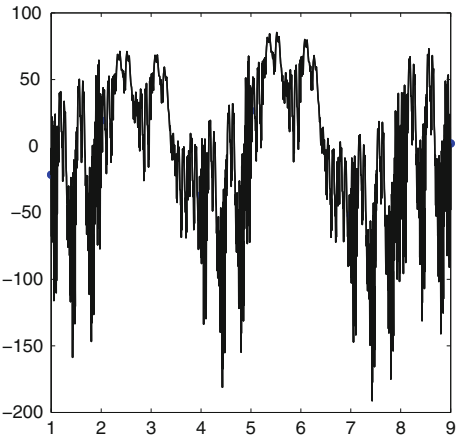


Fig. 6 Derivative of the rational cubic spline FIF in Fig. 1

in Fig. 1. The derivatives of the rational cubic FIF in Figs. 1 and 3, and the classical cubic spline in Fig. 5 are displayed in Figs. 6–8 respectively.

It can be observed that (see Fig. 8) the derivative of the classical rational cubic spline is smooth except possibly at knots. Further, the presence of the shape parameters in the classical rational cubic spline C does not help to produce an interpolant with nonsmooth derivative. On the other hand, the proposed cubic spline FIF with nonzero scaling factors can produce interpolant whose derivative is irregular in a dense set of points in the interpolation interval (see Figs. 6 and 7). By comparing Figs. 3 and 5 it can be seen that rational spline FIF f may agree with classical rational spline C in specified subintervals, by taking corresponding scaling factors to be zero. In this way, the fractality of the derivative $f^{(1)}$ can be confined in a piece of the

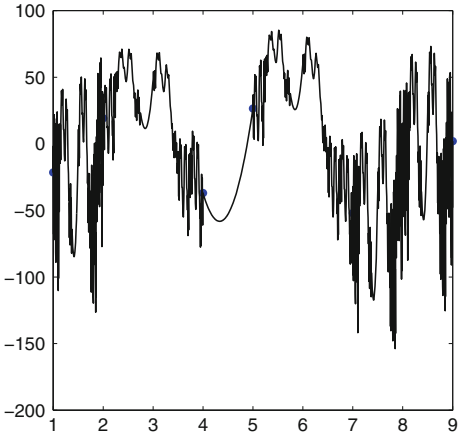


Fig. 7 Derivative of the rational cubic spline FIF in Fig. 3

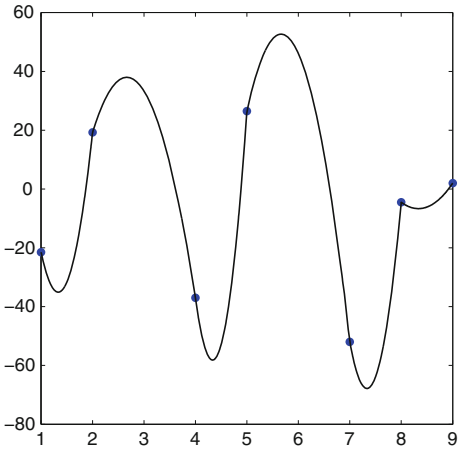


Fig. 8 Derivative of the classical rational cubic spline in Fig. 5

domain, if in this part the possible signal displays some complex disturbance (see Figs. 7 and 8, where the derivatives coincide in the subinterval [4, 5].

Due to the fractal nature of the corresponding derivative functions $f^{(1)}$, the graphs of f depicted in Figs. 1–4 themselves have some artifacts when compared with the classical counterpart in Fig. 5. However, this is to be counteracted with the generality offered by the proposed $\mathscr{C}^1$-continuous fractal interpolant and its suitability in representing a function $\Phi \in \mathscr{C}(I)$ whose derivative $\Phi^{(1)}$ is continuous but nondifferentiable in a dense subset of I. Furthermore, among various desirable properties of an interpolant, if the *fairness* (i.e., visual pleasantness) of the graph of f is of very high concern in a particular problem, then we can take the scaling factor in each subinterval to be very close to zero. On the other hand, if nonsmoothness of the derivative function has high priority, then we may choose the scaling factors

with larger magnitudes. Thus, the parameters can be adjusted to find an interpolant satisfying chosen properties such as smoothness, localness, fairness, and fractality in the derivative.

6 Concluding Remarks and Possible Extensions

In order to combine the advantages of rational splines involving free parameters over the polynomial splines and the versatility of fractal splines over the traditional interpolating splines, a new class of rational cubic spline fractal interpolation functions is introduced in the present work. For zero scaling in each subinterval, the proposed rational cubic fractal spline recaptures the traditional cubic spline (cf. [8]).

We may compare and contrast the fractal spline f introduced herein with its traditional nonrecursive counterpart C as follows. The presence of additional parameters, namely, the scaling factors in the proposed rational cubic spline FIF f offer an additional flexibility in modifying the shape of the interpolant. Furthermore, the scaling factors are the key ingredient that provide fractality to the derivative $f^{(1)}$ of the proposed rational spline f. That is, the present fractal spline f allows varying irregularity for the first derivative $f^{(1)}$, and larger the value of $|\lambda|_\infty$ with respect to the interpolation step, more pronounced is the irregularity in $f^{(1)}$. The irregularity in $f^{(1)}$ can be quantified by means of its fractal dimension, which provides an index for the analysis of the underlying experimental process [12]. These are to be compared with the fact that the classical counter part C of f, which is studied in [8], has a derivative $C^{(1)}$ that is smooth except possibly at the knots. In spite of this flexibility and versatility, for relatively mild conditions on the scaling factors, the proposed scheme possesses the same convergence property as that of its classical counterpart. A possible objection to fractal splines may be that in contrast to the classical nonrecursive interpolants, evaluation of a fractal interpolant at a point is not straightforward and requires a recursive procedure. However, the widespread availability of high-speed computing machines has reduced the importance of this disadvantage.

As far as ability to construct a smooth interpolant with fractality in the derivative of the interpolant is concerned, perhaps a closest competitor for the fractal interpolation scheme is the interpolatory subdivision scheme. For obtaining representative results of subdivision scheme, the reader may consult the nice survey article [9]. In what follows, we shall supply a brief comparison between the two traditions—interpolatory subdivision and fractal interpolation. In both these methods, interpolants are constructed iteratively in contrast to the analytic formulae employed in traditional nonrecursive interpolation techniques. Both these methods can produce an interpolant whose derivative of a suitable order has varying irregularity (from smooth to nowhere differentiable). The fractality in the derivative of fractal interpolant can be measured and controlled with the help of scaling factors, whereas, upto our knowledge, a quantification of irregularity in terms of the parameters involved in the subdivision scheme is not available. For further details on this comparison, the reader is invited to refer to the recent article [5] by the authors.

Identifying appropriate values for the parameters of the rational IFS presented herein so that corresponding FIFs can be applied in shape preserving interpolation can be considered for a future research work. Extension of the proposed rational fractal spline to shape preserving bivariate interpolation can also be considered.

References

1. Barnsley, M.F.: Fractal functions and interpolation. Constr. Approx. **2**(4), 303–329 (1986)
2. Barnsley, M.F., Harrington, A.N.: The calculus of fractal functions. J. Approx. Theor. **57**(1), 14–34 (1989)
3. Chand, A.K.B., Kapoor, G.P.: Generalized cubic spline fractal interpolation functions. SIAM J. Numer. Anal. **44**(2), 655–676 (2006)
4. Chand, A.K.B., Navascués, M.A.: Generalized Hermite fractal interpolation. Rev. R. Acad. de ciencias. Zaragoza **64**(2), 107–120 (2009)
5. Chand, A.K.B., Viswanathan, P.: A constructive approach to cubic Hermite fractal interpolation function and its constrained aspects. BIT Numer. Math. **53**, 841–865 (2013). doi:10.1007/s10543-013-0442-4
6. Chand, A.K.B., Vijender, N., Navascués, M.A.: Shape preservation of scientific data through rational fractal splines. Calcolo **51**(2), 329–362 (2013). doi:10.1007/s10092-013-0088-2
7. Delbourgo, R., Gregory, J.A.: Shape preserving piecewise rational interpolation. SIAM J. Stat. Comput. **6**(4), 967–976 (1985)
8. Duan, Q., Djidjeli, K., Price, W.G., Twizell, E.H.: A rational cubic spline based on function values. Comput. Graph. **22**(4), 479–486 (1998)
9. Dyn, N., Levin, D.: A rational cubic spline based on function values. Acta Numerica **1**(72), 73–144 (2002)
10. Han, X.: Convexity-preserving piecewise rational quartic interpolation. SIAM. J. Numer. Anal. **46**(2), 920–929 (2008)
11. Navascués, M.A., Sebastián, M.V.: Generalization of Hermite functions by fractal interpolation. J. Approx. Theor. **131**(1), 19–29 (2004)
12. Navascués, M.A., Sebastián, M.V.: A relation between fractal dimension and Fourier transform-electroencephalographic study using spectral and fractal parameters. Int. J. Comp. Math. **85**, 657–665 (2008)
13. Rajagopalan, D., Arigo, M.T., McKinley, G.H.: Quantitative experimental and numerical studies of the settling motion of a sphere in highly elastic liquids: part II transient motion. J. Non-Newtonian Fluid Mech. **65**, 17–46 (1996)
14. Roberge, J. K.: The mechanical seal. Bachelor's thesis, Massachusetts Institute of Technology (May 1960)
15. Viswanathan, P., Chand, A.K.B., Agarwal, R.P.: Preserving convexity through rational cubic spline fractal interpolation function. J. Comput. Appl. Math. **263**, 262–276 (2014)

Chapter 9
Applications of Compressive Sensing to Surveillance Problems

Christopher Huff and Ram N. Mohapatra

Abstract In many surveillance scenarios, one concern that arises is how to construct an imager that is capable of capturing the scene with high fidelity. This could be problematic for two reasons: First, the optics and electronics in the camera may have difficulty in dealing with so much information; second, bandwidth constraints may pose difficulty in transmitting information from the imager to the user efficiently for reconstruction or realization. This paper is a study of the application of various compressive sensing methods to surveillance problems. It is based largely on the work of [7], with theory and algorithms presented in the same manner. We explore two of the seminal works in compressive sensing and present the key theorems and definitions from these two papers. We then survey three different surveillance scenarios and their respective compressive sensing solutions. The original contribution of this paper is the development of a distributed compressive sensing model.

1 Introduction

Recent advances in technology have brought with them a great capacity for storing large amounts of data. With data sets becoming increasingly large, it is becoming difficult to analyze the data in order to make use of it. As an example, consider a network of surveillance cameras monitoring a particular area. If the number of cameras is large, it would be difficult to have a small group of people to monitor them carefully. To remedy this situation, one may want to have a computer program to monitor the data and tell the user when a particular event of interest is happening in the scene. An immediate issue that one would encounter in such a scenario (in

C. Huff · R. N. Mohapatra (✉)
Department of Mathematics, University of Central Florida, Orlando, FL 32816, USA
e-mail: ram.mohapatra@ucf.edu

C. Huff
e-mail: ChrisHuffMath@gmail.com

R. N. Mohapatra et al. (eds.), *Mathematics and Computing 2013*,
Springer Proceedings in Mathematics & Statistics 91,
DOI: 10.1007/978-81-322-1952-1_9, © Springer India 2014

addition to many computer vision-related obstacles) would be that of the program parsing the large amount of video data quickly enough so as to alert the user of an event in a timely manner.

Another situation in which a great amount of data is difficult to manage can be found in signal transmission. Suppose one wanted to construct UAV (unmanned aerial vehicle) with the capability of being able to capture (very) high-resolution video of the events happening on the ground below it. Assuming the UAV is able to have such a sensor attached to it (this is not a trivial consideration) the data collected by the UAV must be transmitted in order to be of use. This transmission may not always be possible, since the transmission channel will have limited bandwidth.

These two examples are among the many where a large amount of data are needed for a task, but the amount is too great to manage. This motivates one to ask the questions: Is there any structure in the data set that I am interested in? If so, may I exploit that structure to make the data easier to use?

Depending on the data one is interested in, the answers to the above questions will vary. Recently, there has been a great deal of work in dealing with data sets which exhibit a characteristic, now known as sparsity. We say that a data set is sparse if most of the values in that data set are zero, or so close to zero so as to have little contribution to the overall information of the data. As a frivolous example, consider the vector

$$[1\,0\,0\,0\,1\,1\,1\,0\,0\,1\,0\,0\,0\,1\,0\,0] \tag{1}$$

Suppose we were interested in sensing this vector so that we may transmit it to a user. The vector has 16 entries, but only 6 of the entries are nonzero. This means that the information contained in the vector only depends on the location of the 6 nonzero entries, not the values in all 16 entries. This suggests that one may want to sense all 16 entries and then transmit the locations of the nonzero vectors. The problem with this approach is that it requires one to sense all of the data, parse all of it, and then determine the locations of the nonzero entries. This process involves many calculations, which is not desirable. This begs the following question: If we knew a priori that the vector of interest was sparse, could we take a small number of measurements and then transmit them to the user in a way such that the user could reconstruct the vector from the measurements provided? This would mean that the UAV would not be tasked with the computations described earlier.

This question has been answered affirmatively using a technique known as compressive sensing [9]. The idea behind compressive sensing is as follows: given a signal $x \in \mathbb{R}^n$, one may capture $m << n$ linear measurements $y \in \mathbb{R}^m$ of x and then accurately reconstruct the original signal from y. There are conditions that must be levied on the measurement process and the signal of interest must be sparse; but with these two requirements met, compressive sensing allows one to sense and compress the signal simultaneously.

In this work, we will be interested in dealing with digital images and video. It has been known for some time that natural images and videos are compressible (we will define this precisely later). This essentially means that images and videos may

be represented sparsely on some basis. With the knowledge of this sparsity in hand, one needs only to devise a sensing scheme which is consistent with the theory of compressed sensing in order to enable accurate reconstruction from dramatically undersampled data.

The work that follows is organized as follows: first we review some of the mathematical results that provide support for much of the literature dealing will compressive sensing. Second, we look at some applications of compressive sensing to surveillance problems. This part of the work will demonstrate different ways in which one may find sparsity in a problem and different algorithms that may be applied to a given problem. The third and final section will conclude this work with further research questions and possible directions to their solutions.

2 The Mathematics of Compressive Sensing

In this section, we will survey two of the most important works which developed compressive sensing into a rigorous mathematical theory. The first work we present is entitled *Stable Signal Recovery from Incomplete and Inaccurate Measurements* [9], while the second is entitled *Near Optimal Signal Recovery From Random Projections: Universal Encoding Strategies?* [3] The former used the restricted isometry property to prove that measurements with additive noise could still be used to recover the original signal with reasonable error. The latter established the fact that compressible signals could be recovered from compressive measurements efficiently.

2.1 Recovery from Noisy Measurements

Suppose we wish to recover a sparse vector $x_o \in \mathbb{R}^m$ from incomplete measurements $y \in \mathbb{R}^n$, $n << m$, which are subject to additive noise, e, such that $||e||_2 \leq \epsilon$. That is, $y = Ax_o + e$, where A is a matrix whose columns are the codes against which x_o is inner produced to produce linear measurements/observations of x_o. The above problem is considered in the paper *Stable Signal Recovery from Incomplete and Inaccurate Measurements*.

The key contributions of the paper are twofold: first, that paper was among the first to introduce an error model into the sparse recovery problem. Second, that paper contains a theorem which bounds the error of the recovery by a multiple of the l^2 norm of e. Before we state the major result of that paper, we need to develop the concept of the restricted isometry property.

Let $T \subset \{1, \ldots, m\}$. Let A_T be the $n \times |T|$ submatrix of A obtained by keeping only the columns of A which correspond to the indices in T. Then, we may define the S-restricted isometry constant δ_S for A which is the smallest quantity such that

$$(1 - \delta_S)||c||_2^2 \leq ||A_T c||_2^2 \leq (1 + \delta_S)||c||_2^2 \tag{2}$$

for all subsets T with $|T| \leq S$ and vectors $c \in \mathbb{R}^T$. We say that matrices which have an associated restricted isometry constant exhibit the restricted isometry property (RIP). With these definitions and notation in mind, we may now state the major result from [9]:

Theorem 1 *Let S be such that* $\delta_{3S} + 3\delta_{4S} < 2$. *Then for any signal* x_o *with sparsity less than s and any perturbation e with* $||e||_2 \leq \epsilon$, *the solution* $x^{\#}$ *to the minimization problem is*

$$\min ||x||_1 \text{subject to } ||\text{Ax} - \text{y}||_2 \leq \epsilon \tag{3}$$

obeys

$$||x^{\#} - x_o||_2 \leq C_S \cdot \epsilon, \tag{4}$$

where the constant C_S *may only depend on* δ_{4S}.

This theorem is important due to the stability and error estimate provided for robustly recovering a sparse signal with additive noise.

2.2 Recovering a Compressible Signal

In the above theorem, we have assumed that the signal of interest x_o was sparse in the canonical basis. This is not a reasonable assumption for many signals such as natural images. To appeal to compressive sensing in the context of image acquisition, we will make use of transform coding.

Suppose $I \in \mathbb{R}^m$ denotes a vectorized natural image. Then, we may represent I as a sparse linear linear combination of appropriately chosen vectors. That is,

$$I = \Psi x, \tag{5}$$

where x is S-sparse. This representation introduces a sparse vector, but it is still not clear how to apply the results of compressive sensing. A reasonable question that one may ask is, for what matrix A of test vectors can we use so that the product $A\Psi = \Theta$ exhibits the RIP? If we had such a matrix, then we would have that the solution to the minimization problem

$$\min ||x||_1 \text{subject to } \Theta x - y = 0 \tag{6}$$

is the sparsest solution. We could then recover the original image via $I = \Psi x$. The answer to the above question was addressed in the paper *Near Optimal Signal Recovery From Random Projections: Universal Encoding Strategies?*

That work addressed signals whose coefficients decay like a power law in some basis. That is, if $\Psi = (\psi_j)_{j=1,\ldots,N}$ is an orthonormal basis and $I \in \mathbb{R}^N$ is the signal of interest. Let $x_j =< I, \psi_j >$ and let us sort the vector x according to the magnitude

of its elements so that $|x_k| \geq |x_{k-1}| \geq \cdots \geq |x_1|$. We say that x decays like a power law if there exists $C > 0$ such that

$$|x_k| \leq C \cdot k^{-1/p}, \quad 1 \leq k \leq N \tag{7}$$

If p is sufficiently small ($0 \leq p \leq 1$) then we say that I is compressible.

With this type of signal in mind, we are then introduced to two principles which the measurement matrix (the role assumed by A) is to obey: the uniform uncertainty principle (UUP) and the exact reconstruction principle (ERP). Suppose that $1 \leq k \leq N$ and $\Omega = \{1, \ldots, k\}$. Then we suppose that the measurement matrix $A = A_\Omega$ is a random matrix of dimension $|\Omega|$ by N. Let the number of measurements $|\Omega|$ be a random variable and denote the expected value of $|\Omega|$ by K. Further still, let R_T denote the restriction map from $\mathbb{R}^N$ to a set $T \subset \mathbb{R}^N$. Then, we may define $R_T^* : T \to \mathbb{R}^N$ as the function which inserts zeros outside of T (if $x \in \mathbb{R}^N$, then $\text{supp}(R_T^* x) \subset T$). Let $A_{\Omega T} := A_\Omega R_T^*$. Then $A_{\Omega T}$ is an $|\Omega|$ by $|T|$ matrix obtained by extracting $|T|$ columns from A_Ω , where the jth column is chosen if $j \in T$.

The Uniform Uncertainty Principle (UUP) [3]: We say that the measurement matrix A obeys the uniform uncertainty principle with oversampling factor λ if for every sufficiently small $\alpha > 0$, the following statement is true with probability greater than or equal to $1 - O(N^{-\rho/\alpha})$ for some fixed $\rho > 0$: for all subsets T such that

$$|T| \leq \alpha \cdot K/\lambda, \tag{8}$$

the matrix A obeys the bounds

$$1/2 \cdot K/N \leq \lambda_{\min}(A_{\Omega T} * A_{\Omega T}) \leq \lambda_{\max}(A_{\Omega T} * A_{\Omega T}) \leq 3/2 \cdot K/N. \tag{9}$$

The Exact Reconstruction Principle (ERP) [3]: We say that the measurement matrix A obeys the exact reconstruction principle with oversampling factor if for all sufficiently small $\alpha > 0$, each fixed subset T obeying (equation number) and each sign vector σ defined on T, $|\sigma(t)| = 1$ if $t \in T$, there exists with probability greater than $1 - O(N^{-\rho/\alpha})$ a vector $P \in \mathbb{R}^N$ with the following properties:

1. $P(t) = \rho(t)$, for all $t \in T$.
2. P is a linear combination of rows from A.
3. $|P(t)| \leq 1/2$ for all $t \in T^c$.

Now that we may describe a measurement matrix A with the UUP and ERP, we may formally state the theorem which will allow us to use sparse representation to recover compressible signals.

Theorem 2 [3] *Let F be a measurement matrix such that the UUP and ERP hold with oversampling factors λ_1 and λ_2, respectively. Let $\lambda = \max(\lambda_1, \lambda_2)$ and assume that $K \geq \lambda$. Suppose I is a signal satisfying the compression inequality for some fixed $0 < p < 1$, and let $r := 1/p - 1/2$. Then for any sufficiently small $\alpha > 0$, any minimizer $x^\#$ to the problem (4) will obey,*

$$||x - x^{\#}||_2 \leq C_{p,\alpha} \cdot R \cdot (K/\lambda)^{-r} \tag{10}$$

with probability $1 - O(N^{-\rho/\alpha})$.

This theorem, together with the fact that Gaussian measurement matrices obey the UUP and ERP with $\lambda = \log(N)$, enables us to consider the reconstruction of large classes of signals which are sparse in some orthonormal bases. This includes the class of natural images which are sparse in a wavelet basis. Videos may be regarded as sequences of images, and hence these results enable us to address problems of capturing videos as well.

3 Applications to Surveillance Problems

This section is primarily concerned with application of compressive sensing to different types of surveillance problems. The first scenario deals with a rather typical surveillance task; monitoring a parking lot. The second scenario will address the need to track motion in a video sequence. The third situation is one in which we are concerned about reconstructing a photograph of a very large land area.

The acquisition and transmission of high-resolution video signal is often problematic due to the limitations of the ability of the camera to capture sufficient amounts of data and the transmission channel's bandwidth, which limits the amount of information that can be transmitted once the data are acquired. This motivates the need to develop a framework by which a scene can be sampled at a relatively low rate and then reconstructed in a way such that the video is of high quality.

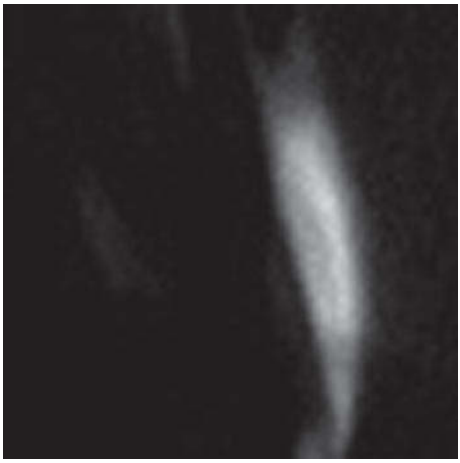
There are many different types of scenes that one might capture. The type of motion in the video, the amount of the viewing area being consumed by the motion, lighting conditions, etc. For our purpose, we will assume that we want to reconstruct a video in which most of the scene is static, and the lighting conditions are constant. This may seem rather restrictive at first glance, but such scenes naturally arise in the area of surveillance (traffic cameras, UAVs, etc.). From hereon we proceed with these types of surveillance applications in mind (Fig. 1).

The first portion of this section deals with a stationary camera capturing a dynamic scene. The second section also involves a stationary camera, but explores the idea of using compressive sensing to capture purely motion information from a scene. In the third and final section, we deal with the problem of surveying a large piece of land. The contents of this final section are largely taken from a recent work written by the author of this paper which appeared in the proceedings of an SPIE conference [6].

3.1 Video Reconstruction Using LDS Model

One potential solution for compressive sensing of such video sequences was offered in *Compressive Acquisition of Dynamic Scenes* [1]. In the paper, the authors modeled the compressive sensing of a scene in time as a linear dynamical system. The basic

Fig. 1 Frame 30 ground truth



model of a linear dynamical system is as follows: let $\{I_t, t = 0, \ldots, T\}$ be a sequence of frames indexed by time t. Then we may model each frame of video $I_t \in \mathbb{R}^{\mathbb{N}}$ as

$$x_t = Cz_t,$$

where $C \in \mathbb{R}^{N \times d}$ is the observation matrix and $z_t \in \mathbb{R}^d$ is the hidden state vector. Let y_t denote the compressive measurement of x_t. That is,

$$y_t = \Phi_t x_t. \tag{11}$$

where Φ_t is the sensing matrix to be used at time t. At each time instance we encode the static portions of the scene as well as the dynamic portions. Let $\check{y}_t$ and $\tilde{y}_t$ denote the static and dynamic measurements, respectively. Let $\check{\Phi}$ and $\tilde{\Phi}_t$ denote the measurement matrices for the static and dynamic portions of the scene, respectively (Fig. 2).

Then at each time instant t, we take the following measurements:

$$y_t = \begin{pmatrix} \check{y}_t \\ \tilde{y}_t \end{pmatrix} = \begin{bmatrix} \check{\Phi} \\ \tilde{\Phi}_t \end{bmatrix} I_t = \Phi_t y_t, \tag{12}$$

where $\check{y}_t \in \mathbb{R}^{\check{M}}$ denotes the constant measurements associated with the constant sensing matrix $\check{\Phi}$ (essentially encoding the constant motion from the scene), and $\tilde{y}_t$ denotes the dynamic measurements associated with the matrix $\tilde{\Phi}_t$.

To recover the video sequence $[x_t]$ via the LDS model, we first solve for the state sequence $[z_t]$ and then solve for the observation matrix C (the notation $[x_t]$ denotes the matrix with columns equal to $x_t, t = 1, \ldots, N$). To solve for the state sequence, we make the following observation: if $[x]$ lies in the column span of C, then $\left[\check{y}_t\right]$ lies in the column span of $\check{\Phi}C$. This implies that the SVD of $\left[\check{y}_t\right]$ will render an approximation of the state sequence $\left[\hat{z}\right]$. More precisely, if $\check{M} \geq d$, and

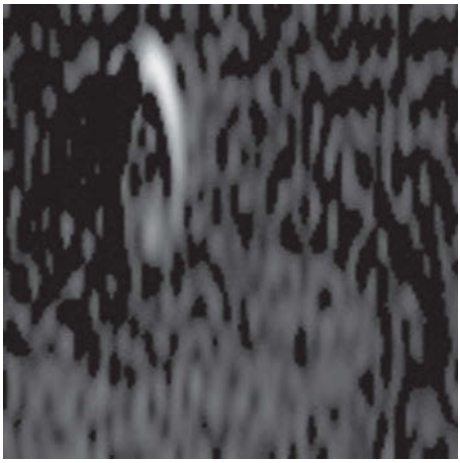


Fig. 2 Frame 30 reconstruction with 74 frames

$\left[\check{y}_t\right] = USV^T$, then $\left[\hat{z}_t\right] = S_d V_d^T$, where S_d is the $d \times d$ principal submatrix of S and V_d is the $T \times d$ matrix formed by columns of V corresponding to the singular values in S_d. We have that this estimate of the state sequence is reasonably accurate when x_t is compressible [ref].

Once the estimated state sequence $[\hat{x}_t]$ has been constructed, we can recover C by solving the following problem:

$$\min \sum_{k=1}^{d} ||\Psi^T c_k||_1 \quad \text{subject to} \quad ||\Phi_t C\hat{z} - y_t||_2 \leq \epsilon, \forall t. \tag{13}$$

Rather than solving this problem directly, we may use a modified CoSAMP algorithm in order to take advantage of the redundancy in the common measurements. The pseudocode for this algorithm is provided below:

This version of the CoSAMP algorithm can be interpreted as a special case of the model-based CoSAMP algorithm developed in [ref]. This interpretation offers the advantage of allowing the calculation of the number of measurements required for stable recovery by simply looking at the model sparsity of the signal. Specifically, if the sparsity of the signal (in our case $\hat{C}$) is s, then results of model-based CoSAMP guarantee that $\mathcal{O}(s\log(Nd))$ are needed. The results in [ref] show that if the columns of C are K-sparse, then the sparsity of $\hat{C}$ is equal to dK. Thus, we need $M = \mathcal{O}(s\log(Nd))$ measurements at each time instant in order to guarantee that the recovery will be accurate. That is, $M = \mathcal{O}(dK\log(Nd)/T)$. This implies that as the number of frames increases, the number of measurements needed decreases.

Algorithm 1: LDS CoSAMP

input : $\Phi_t, \Psi y_t, \hat{z}_t, K$
output: $\hat{C}$
$\Theta_t \leftarrow \Phi_t \Psi$;
$v_t \leftarrow 0$;
$\Omega \leftarrow 0$;
while *stopping criteria not met* **do**
  $R \leftarrow \sum_t \Theta_t^T v_t \hat{z}_t$;
  $\forall k \in \{1, \ldots, N\}, r(k) \leftarrow \sum_{i=1}^{d} R^2(k, i)$;
  $\Omega \leftarrow \Omega \cup r_{2K}$;
  $A \leftarrow \text{argmin} \sum_t ||y_t - (\Theta_t)_{\cdot,\Omega} A \hat{z}_t||_2$;
  $B_{\Omega,\cdot} \leftarrow A$;
  $B_{\Omega^c,\cdot} \leftarrow 0$;
  $\forall k \in \{1, \ldots, N\}, b(k) \leftarrow \sum_{i=1}^{d} B^2(k, i)$;
  $\Omega \leftarrow b_K$;
  $S_{\Omega,\cdot} \leftarrow B_{\Omega,\cdot}$;
  $S_{\Omega^c,\cdot} \leftarrow 0$;
  $\hat{C} \leftarrow \Psi B$;
  $v_t = y_t - \Theta_t S \hat{z}_t$;

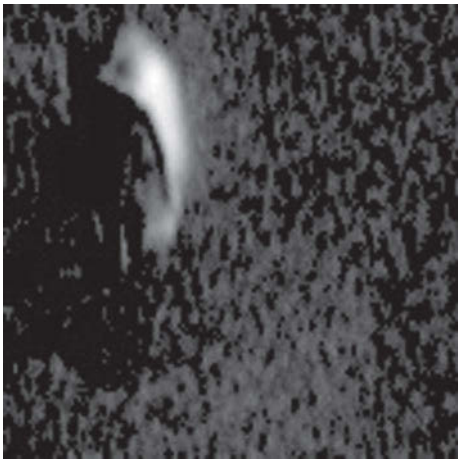


Fig. 3 Frame 30 reconstruction with 200 frames

3.2 Experiments with the LDS Model

The original paper which used the CS-LDS model focused mainly on scenes that resemble changing textures. One such scene is one that contains a flame from a lighter. To show how well this model works with such a scene, we present results of different reconstructions below. In each reconstruction, we vary the number of frames used. This illustrates the model's ability to allow very few measurements per frame to be used, so long as enough frames are used (Fig. 3).

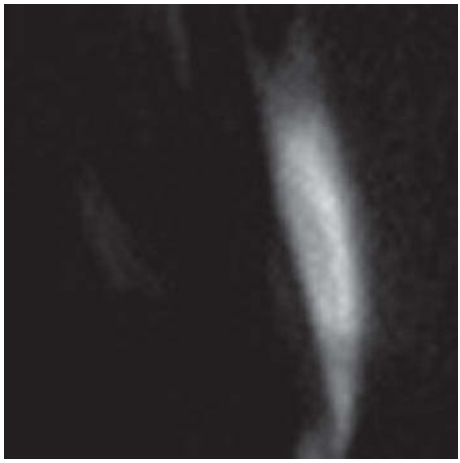


Fig. 4 Frame 30 reconstruction with 560 frames

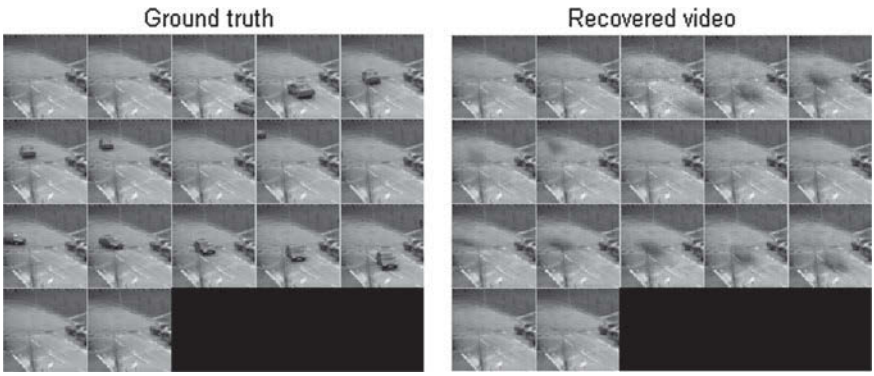


Fig. 5 Using the CSLDS-mean model

For our next experiment, we consider a portion of a video which captures a car passing through a static background (Fig. 4).

One notices that the static portion of the scene is reconstructed accurately, but the dynamic portion of the scene is hardly reconstructed at all. In fact, the cars driving by are reconstructed as faint spectres. Their positions can be gathered from the reconstruction, but their features are completely gone. In the next experiment, we consider a scene with people walking around. The first example considers only a portion of the scene where the people are pacing in the same small area, turning and walking a very small distance. The second example is of the same general scene, except that now we have a person walking a significant distance through the scene (Fig. 5).

Looking at these results, we notice that the appearance of the figures which pace around, but stay entirely within the frame, are well recovered while the person who walks off frame is poorly reconstructed (see Figs. 6 and 7), with their features being

Fig. 6 Pedestrians with little motion

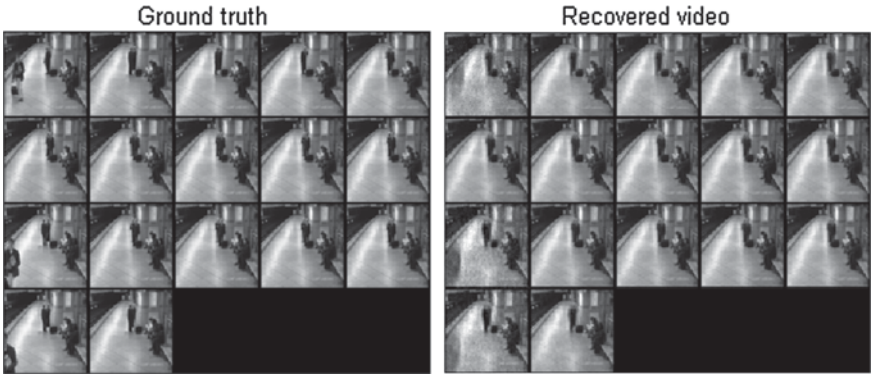


Fig. 7 Pedestrians with significant motion

dissolved in the same way as the features of the moving cars in the preceding experiment. This begs the following question: Why does this model reconstruct persistently visible objects well, while failing to reconstruct objects that are not always within the scene? A rigorous answer to this question is a great opportunity for further research, as this answer may lead to a better model which will be more robust to a variety of scenes.

3.3 Monitoring Motion in a Scene

In certain scenarios, the user might not be interested in what the scene looks like, but rather, what is happening in the scene. For example, one might want to know when there are moving objects in the scene and the nature of their motion, rather than

the look of the scene itself. To address this surveillance concern, we demonstrate a method developed in *Compressive Sensing for Background Subtraction* [5]. In this work, the authors make use of the following observation: given a scene with a static background and a changing foreground, the difference image from the two adjacent frames will have a higher degree of sparsity than the frames themselves.

To be more precise, let us introduce some notation. Let x_b denote the background image, x_c the current frame, and x_d the difference image, with $x_d = x_c - x_b$. Let $\mathcal{S}_d$ denote the support of the difference image. Then by parsing $\mathcal{S}_d$ one may determine the overall shape and location of motion in the frame. A conventional imaging scheme would sense x_b and x_c and then directly construct x_d. Since we are not concerned with the actual appearance of the scheme, the work needed to capture x_b and x_c is excessive. We instead seek a way to use compressive sensing to reconstruct the difference image in a way such that we never need to reconstruct x_b or x_c.

Indeed, let us observe the following:

$$y_b = \Phi x_b, \tag{14}$$

and,

$$y_c = \Phi x_c. \tag{15}$$

Therefore,

$$y_b - y_c = \Phi(x_b - x_c), \tag{16}$$

or,

$$y_d = \Phi x_d, \tag{17}$$

where $y_d = y_b - y_c$ denotes the difference compressive measurements. This simple idea give us a way by which to reconstruct the difference image by requiring that we only compressively sense the background and current images. Further still, when one looks at a difference image, one notices that it is mostly black. This suggests that x_d should be sparser than x_b and x_c.

Indeed, let us suppose that the sparsity of the x_b and x_c is K (it is reasonable to make this assumption because of the similarities between the two images). Let K_d denote the sparsity of the difference image, x_d. Because much of the difference image will be empty, but for any motion, we may conclude that the wavelet coefficients used to represent the information contained in the static portion of the scene may be discarded. Hence, $K_d \leq K$. This means that we should be able to take few compressive measurements of x_b and x_c and still be able to reconstruct the difference image at the level of the quality it would have been seen at if we took all K measurements of x_b and x_c. We demonstrate this point empirically in the next section.

Fig. 8 The ground truth difference image

Fig. 9 Frame 30 reconstruction using 5 % of the data

3.4 Experiments with Motion Tracking

In this section, we present numerical experiments which demonstrate that a reasonable difference image may be reconstructed from the compressive measurements of the scene, and that the number of measurements required to reconstruct the difference scene is much less than the number required to reconstruct the scene itself. We use the Coiflet wavelet basis as the sparsifying basis for each frame of video. We will recover images via the NESTA algorithm.

In our first experiment, our objective is to reconstruct a scene of a parking lot with a car driving past. The field-of-view is 64 by 64 pixels. We reconstruct the difference images in two ways: first we sense the video in the traditional manner and construct

Fig. 10 Frame 30 reconstruction using 10 % of the data

the difference images from the actual image sequence, providing the ground truth. Second, we compressively sense the scene and construct the difference image from the difference of compressive measurements.

The results of the first experiment are presented in Figs. 7, 8, 9 and 10. At first glance, one may look at these results and be left feeling that the compressive sensing scheme does not offer much of a benefit. The amount of data the sensor needs to process is far lower with the compressive sensing scheme, but in exchange the reconstruction quality is far worse, both in terms of appearance of the reconstruction and the error measured in terms of the L2-norm. However, when one looks closely at the reconstructed difference image, one notices that the outline of a car is clearly visible and distinct from the noise. Also, the noise looks like noise. To be exact, it is clear that the errors in the reconstruction are extraneous. This gives reason to believe that a filtering process may be performed on the reconstructed difference images in order to produce more accurate results.

As can be seen from Figs. 11, 12, 13 and 14, even a very naive thresholding technique can dramatically improve the quality of the reconstructed image. In particular, the portion of the scene with the moving car peaks high enough so that its motion is sensed correctly in every frame of video. This means that the motion sensing problem may in fact be resolved via a compressive sensing approach.

3.5 Using a Compressive Background Model for Object Detection

Often, it is the case that there is no new object in the scene. This implies that there is nothing of interest taking place in the scene. The above model calls for an l_1 minimization for each and every difference image. This is computationally taxing, and so it is worthwhile to investigate whether or not the minimization step really needs to be performed at every time instance. This section proposes a way of determining

Fig. 11 Frame 30 reconstruction using 20 % of the data

Fig. 12 Frame 30 reconstruction using 30 % of the data

whether or not the scene is changing. To do this, we develop a statistical model for the compressive background measurements and then use the compressive measurements directly to determine if a new object has entered the scene (Fig. 15).

Suppose we have a collection of compressive measurements of the background images. Let $y_{bi} \in \mathbb{R}^M$ denote the ith compressive measurement vector of the background of the scene with $i = 1, \ldots B$. Let y_b denote the mean of the background images. Let us consider the distribution of l_2 distances of the background images about their mean:

$$||y_{bi} - y_b||_2^2 = \sigma^2 \sum_{k=1}^{M} \left(\frac{y_{bi}(k) - y_b(k)}{\sigma} \right)^2 \tag{18}$$

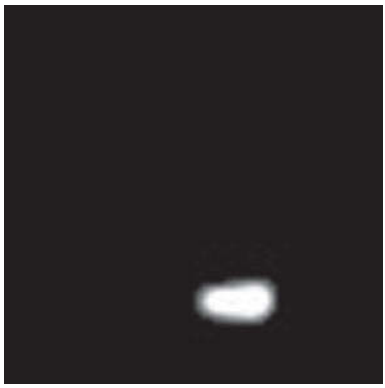


Fig. 13 Frame 30 filtered reconstruction using 5 % of the data

Fig. 14 Frame 30 filtered reconstruction using 10 % of the data

Fig. 15 Frame 30 filtered reconstruction using 20 % of the data

If we take $M > 30$, then the central limit theorem gives us that the distribution of l_2 distances may be approximated by a Gaussian distribution. That is,

$$||y_{bi} - y_b||_2^2 \sim \mathcal{N}(M\sigma^2, 2M\sigma^4). \tag{19}$$

Now suppose that we are comparing a test image to the mean background. Then we may derive the following distribution:

$$||y_t - y_b||_2^2 \sim \mathcal{N}(M\sigma^2 + ||\mu_d||_2^2, 2M\sigma^2 + 4\sigma^4||\mu_d||_2^2). \tag{20}$$

We can simplify matters by considering the logarithms of the l_2 distances. Using this approach, we may write that

$$\log ||y_{bi} - y_b||_2^2 \sim \mathcal{N}(\mu_b, \sigma_b^2). \tag{21}$$

and

$$||y_t - y_b||_2^2 \sim \mathcal{N}(\mu_t, \sigma_t^2). \tag{22}$$

Our goal is to use these statistics to determine if a new object has entered a scene without having to perform a costly l_1 minimization to reconstruct the difference image. Toward this end, we learn the parameters in (11) via maximum likelihood estimates. With μ_b and σ_b known, we have that if σ_t^2 is sufficiently different from σ_b^2, then a simple two-sided threshold test is optimal for discriminating between another background image and an image with a new object in it [10]. Thus, we say that there is a new object in the scene if

$$|\log ||y_t - y_b||_2^2 - \mu_b| \geq a\sigma_b, \tag{23}$$

where a is a constant to be chosen by the user.

3.6 Monitoring a Large Track of Land

High-resolution imaging sensors used in observing terrestrial activities over a very wide field-of-view will be required to produce gigapixel images at standard video rates. This data deluge affects not just the sensor but all of the processing, communication, and exploitation systems downstream. A key challenge is to achieve the resolution needed to observe and make inferences regarding events and objects of interest while maintaining the area coverage, and minimizing the cost, size, and power of the sensor system. One particularly promising approach to the data deluge problem is to apply the theory of compressive sensing, which enables one to collect fewer, information-rich measurements, rather than the many information-poor measurements from a traditional pixel-based imager (Fig. 16).

For the wide field-of-view imaging application, Muise [8] designed a compressive imaging algorithm with associated measurement kernels and has simulated results based upon a field-of-view multiplexing sensor described by Mahalanobis et al. [2]. These works show a viable concept for wide area imaging at high resolution. In this section, we explore concepts of collecting measurements of a wide area through

Fig. 16 Frame 30 filtered reconstruction using 30 % of the data

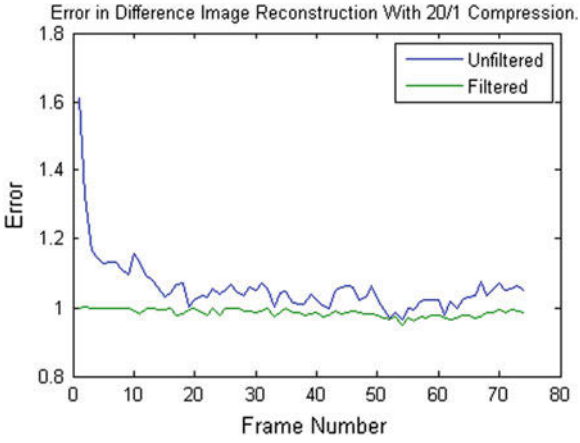


Fig. 17 Using 5 % of the data

multiple cameras and reconstructing the entire wide area image. This process is known as distributed compressive imaging (DCI).

Consider an N-pixel area to be sensed with multiple cameras and suppose we have limited bandwidth for communications. The bandwidth restriction precludes us from allowing for intra-camera communication. Compressive sensing theory tells us that $M = \beta \log N/K$ measurements are sufficient to guarantee an accurate signal recovery (here K denotes the sparsity of the area of interest). Suppose we have α cameras at our disposal and that these cameras have overlapping fields-of-view. Then, assuming the cameras end up covering the entire area in aggregate, each camera need only take M/α compressive measurements in order to facilitate accurate signal reconstruction. The clear benefit here is that as the number of cameras increases, the amount of information each camera is responsible for acquiring decreases (Fig. 17).

3.7 DCI Model

Here we propose a simple extension to the traditional compressive sensing model to make use of a camera ensemble. The naive DCI model is

$$\mathbf{Y} = \mathcal{P}\mathbf{B}x + \epsilon, \tag{24}$$

where $\mathcal{P}$ is a concatenation of the random Gaussian sensing matrices of each of the α cameras in our ensemble and $\mathbf{B}$ is the sparsity basis for the scene, x. That is,

$$\mathcal{P} = [\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_\alpha]^T = [p_1^1, p_2^1, \ldots, p_{k/\alpha}^1, p_1^2, \ldots, p_{k/\alpha}^2, \ldots p_1^\alpha, \ldots, p_{k/\alpha}^\alpha]^T.$$

Each entry of $\mathbf{Y}$ is an inner product of the image with a random projection vector p_j^i and so its form is

$$\mathbf{Y} = [\langle p_1^1, \mathbf{B}x\rangle, \langle p_2^1, \mathbf{B}x\rangle, \ldots, \langle p_{k/\alpha}^1, \mathbf{B}x\rangle, \ldots, \langle p_1^\alpha, \mathbf{B}x\rangle, \ldots, \langle p_{k/\alpha}^\alpha, \mathbf{B}x\rangle]^T + \epsilon.$$

Our interest lies in having multiple cameras, all surveying a large region from different perspectives. As such, if we take an absolute coordinate system for the entire region we model the differences in perspectives with an operator $\mathcal{O}_i$ so that $\mathcal{O}_i(\mathbf{B}\alpha)$ generates the underlying scene $\mathbf{B}\alpha$ from the point of view of the ith camera. With this idea, we may rewrite the observed measurements as

$$\begin{aligned}\mathbf{Y} = [&\langle p_1^1, \mathcal{O}_1(\mathbf{B}x)\rangle, \langle p_2^1, \mathcal{O}_1(\mathbf{B}x)\rangle, \ldots, \langle p_{k/\alpha}^1, \mathcal{O}_1(\mathbf{B}x)\rangle, \ldots, \langle p_1^\alpha, \mathcal{O}_\alpha(\mathbf{B}x)\rangle,\\ &\ldots, \langle p_{k/\alpha}^\alpha, \mathcal{O}_\alpha(\mathbf{B}x)\rangle]^T.\end{aligned}$$

For a particular perspective operator, $\mathcal{O}_i$, we wish to derive the adjoint (for lack of a better term), $\mathcal{O}^*$, so that

$$\langle h, \mathcal{O}_i(y)\rangle \;=\; \langle \mathcal{O}_i^*(h), y\rangle, \text{ for all } h, y.$$

For example, if $\mathcal{O}_i$ translates an image by $[a, b]$ pixels, then $\mathcal{O}_i^*$ would translate the measurement mask by $[-a, -b]$ pixels for an equivalent inner product. With this idea in mind, we may once again rewrite our observation vector, $\mathbf{Y}$, as

$$\begin{aligned}\mathbf{Y} &= [\langle \mathcal{O}_1^*(p_1^1), \mathbf{B}x\rangle, \langle \mathcal{O}_1^*(p_2^1), \mathbf{B}x\rangle, \ldots, \langle \mathcal{O}_1^*(p_{k/\alpha}^1), \mathbf{B}x\rangle, \ldots, \langle \mathcal{O}_\alpha^*(p_1^\alpha), \mathbf{B}x\rangle,\\ &\quad \ldots, \langle \mathcal{O}_\alpha^*(p_{k/\alpha}^\alpha), \mathbf{B}x\rangle]^T + \epsilon\\ &= \mathcal{P}^*\mathbf{B}x + \epsilon.\end{aligned}$$

Thus we take as the general DCI model

$$\mathbf{Y} = \mathcal{P}^*\mathbf{B}x + \epsilon, \tag{25}$$

where ϵ is included to take into consideration additive error in the sensing process. Also, unlike (1), this accounts for different camera perspectives. Thus we will be solving

$$\min_{\hat{\mathcal{P}},x} ||x||_1 \text{ subject to } ||\mathbf{Y} - \hat{\mathcal{P}}\mathbf{B}x||_2 \leq c \tag{26}$$

where $\hat{\mathcal{P}} = \mathcal{P}^* + \mathcal{P}^e$, the ideal perspective operator plus an error. This alters our model for the observations to

$$\begin{aligned}\mathbf{Y} &= \hat{\mathcal{P}}\mathbf{B}x + \epsilon \\ &= (\mathcal{P}^* + \mathcal{P}^e)\mathbf{B}x + \epsilon \\ &= \mathcal{P}^*\mathbf{B}x + \mathcal{P}^e\mathbf{B}x + \epsilon \\ &= \mathcal{P}^*\mathbf{B}x + \epsilon'\end{aligned}$$

where our new error term is bounded by

$$\begin{aligned}||\epsilon'||_2^2 &= ||\mathcal{P}^e\mathbf{B}x + \epsilon||_2^2 \\ &\leq ||\mathcal{P}^e||_2^2||\mathbf{B}x||_2^2 + ||\epsilon||_2^2 \\ &\leq E||\mathcal{P}^e|| + c,\end{aligned}$$

where E is the overall energy in the image. Appealing to the result from Candes, Romberg, and Tao, we can solve (10) for $x^\sharp$ with the guarantee that

$$||x_{\text{true}} - x^\sharp||_2 \leq \mathcal{O}(E||\mathcal{P}^e|| + c).$$

Although the behavior of $E||\mathcal{P}^e||$ is difficult to characterize, there are several observations:

- When the ideal perspective estimates are known, $\mathcal{P}^\rceil = 0$ and thus $E||\mathcal{P}^e||$ is a minimum, and Eq. (4) distils down to the case studied by Candes, Romberg, and Tao.
- An iteration of (5) while perturbing the perspective estimates should generate a surface which has a global minimum when $\hat{\mathcal{P}} = \mathcal{P}^*$.

Hence, we are left with a procedure and an optimality criterion which theoretically should give us estimates for x and the camera perspectives by minimizing the l_1 norm of x while fitting the observed data (Fig. 18).

3.8 Experiments with Large Area Monitoring

Given a wide field-of-regard image, we wish to collect image projections from multiple cameras and rebuild the scene with minimal data being transmitted. Assuming

Fig. 18 Using 10 % of the data

that we know the perspective parameters for the multiple cameras, we have a sequence of cameras depicted by Fig. 1.

We assume that the bandwidth of the data-link can only afford to send down 0.2 % of the image over the support of its field-of-view. For example, if a camera generated a 128×128 image, then the amount of information transmitted for reconstruction would be approximately 24 numbers. The reconstruction from the noncompressed sensing is accomplished by observing the image, calculating the compression coefficients assuming a DCT basis set, and sending the top 0.2 % of the coefficients to the reconstruction algorithm. Under this paradigm, the results of the scene reconstruction are shown in Fig. 2.

For a distributed compressive imaging scenario, we assume the entire scene of interest can be compactly represented in a DCT basis and each individual camera would sample an image projection of a limited FOV of the scene. The projection masks should be randomized (to guarantee incoherence with the DCT basis) but should also have a notion of random sampling (as this is optimally incoherent with the DCT basis). We choose a methodology of projection mask construction as the following:

1. Randomly generate a size and location for the pixel sampling.
2. Iterate until roughly $1/4$ of the pixels are contributing to the projection (this will ensure an SNR advantage through multiplexing).

An example of a projection mask used for this experiment is given in Fig. 3.

With the camera perspective parameters assumed to be known, we calculate the projection mask in terms of the underlying scene coordinate system. This results in calculating the rows of the projection matrix $\mathcal{P}$. Two of these example perspective masks are given in Fig. 4.

With this calculation of the projection masks into the underlying scene coordinate system, we use the STOMP [4] as our compressive sensing reconstruction algorithm

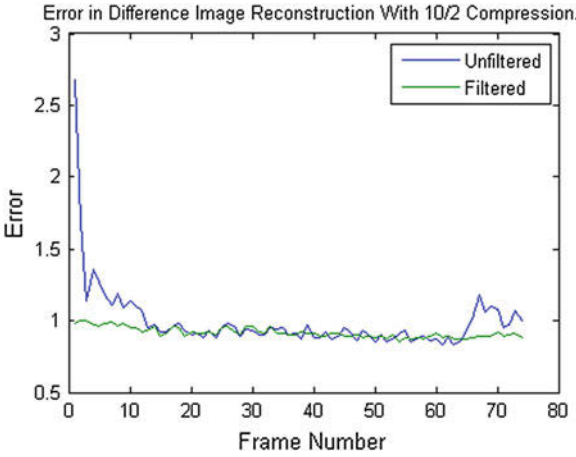


Fig. 19 Using 20 % of the data

with less than 0.2 % of the underlying dimension of each camera's FOV. The results are shown in Fig. 5.

One notices that while new information is collected and transmitted, all of the areas of the underlying scene experience higher fidelity reconstruction. The final reconstruction with and without DCI is shown in Fig. 6.

One notices a very low frequency image from the standard compression which results from only 0.2 % of the information being transmitted. The overall shape of the larger buildings is successfully reconstructed as well as the general large road network. With DCI, the reconstruction contains far more high-frequency content with many smaller buildings visible and the texture and shape of the trees on the right being of higher quality (Fig. 19).

3.9 Multi-camera Registration Issues

The above experiment was conducted under the assumption that the camera perspectives were known. Such information is not generally known and an image registration step would be required. For standard video cameras, this registration can be nontrivial, but solvable with standard tie-point correlation and re-sampling, or other techniques. For DCI, the imagery is unavailable to calculate correlations and we are required to register the imagery without access to the images. This problem was solved with manifold lifting techniques by Wakin [11], while we wish to test whether Eq. (3) gives us a general optimization criterion for estimating the camera perspectives from the image projection data stream (Fig. 20).

Again, Eq. (3) suggests that the same criteria used to estimate the nonzero coefficients of a sparse model can be used to iteratively estimate the perspective parameters of our distributed compressive imaging system. To see the intuition, imagine that our

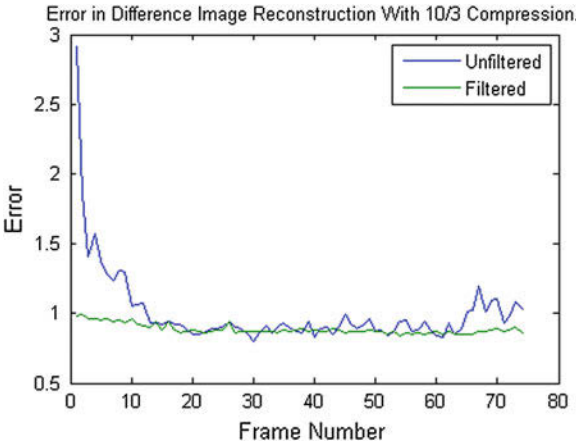


Fig. 20 Using 30 % of the data

perspective estimates for the cameras are incorrect. It should take more coefficients to reconstruct the incorrect scene than it would take to reconstruct the correctly registered information. Thus, finding the sparsest solution (or equivalently, the minimum l_1 solution) over all possible perspective parameters should lead to the correct perspective estimates. We test this through a nine camera DCI test with the Lena image as described in the next section.

3.10 A DCI Model with Unknown Registration

In this experiment, we treat the image of Lena as the field-of-regard and we have nine cameras surveying the image, each of which has a limited field-of-view. While no two cameras share the same field-of-view, each camera's field-of-view overlaps with at least one other camera's (Fig. 21).

The registration of the center camera's position is assumed to be unknown; and for the purposes of this experiment, all other camera perspective parameters are assumed to be known. Also, although the results of our experiments should generalize to most camera perspective parameters, we test only unknown x, y translation.

With real surveillance applications in mind, it is reasonable to assume that one will have approximate camera registration parameters available. These approximate values will serve as an initial guess. Our experiment takes in the measurements from all nine cameras (the only unknown is the position of the center camera, denoted as γ), then takes in the estimate for γ, calculates the projection masks in terms of the underlying scene coordinate system, recovers an image through l_1 minimization, and saves the associated l_1 norm of the scene coefficients. We then make another estimate for γ and repeat this process, always saving the l_1 norms associated with each reconstruction. This process is meant to visualize the function from Eq. (3), which

Fig. 21 A wide field-of-regard image being sensed with *multiple* cameras

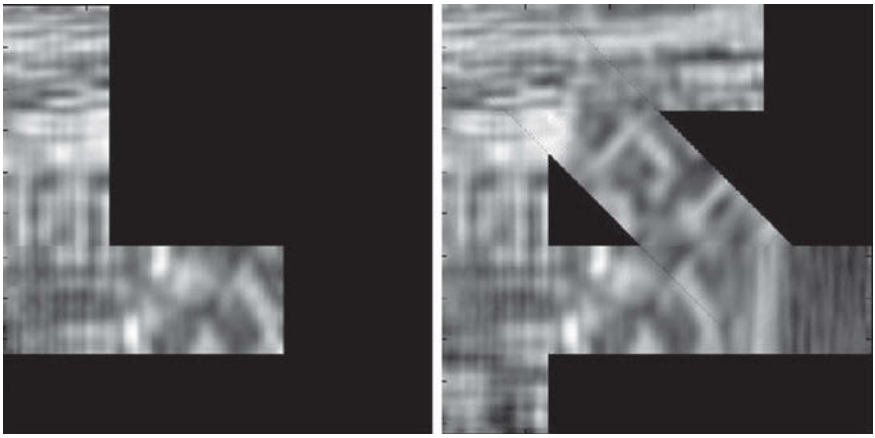


Fig. 22 The scene being surveyed by traditional cameras with reconstruction via traditional compression

should give us an optimality criteria for estimating x and the camera registration parameters (which are embedded in the estimate for x) (Fig. 22).

Graphing the l_1 norms for each of the reconstructions as a function of the unknown parameter γ we have the following result in Fig. 8.

The noisy nature of this surface suggests that determining the optimal camera parameters based on the l_1 norm would be a difficult task. However, if we smooth this function by convolving it with a Gaussian mask of size 7×7 we gain insight into the nature of how this function behaves.

This graph suggests that this function is locally quadratic. This offers one the intuition that one can find the global minimum of the function by taking a smoothed version of $||x||_1$ as our optimality criteria. To this end, for each test value of γ we solve

Fig. 23 A typical projection mask

Fig. 24 Projection masks placed in the scene's coordinate system

(4) for several values close to γ and take the average value of $||x||_1$ as our objective function. The gradient of this new surface (represented in Fig. 9) should now be relatively continuous and should give us insight into the possible convergence of a gradient descent algorithm. These gradients were calculated for the raw and smoothed versions of our objective function and are shown in Fig. 10 as arrows overlaid on an image of our objective function. The ideal perspective estimates correspond to the center of each image (Fig. 23).

The results of this experiment are promising and lend support to the argument that the minimum l_1 norm taken over different image reconstructions is minimized when the projection masks are correctly positioned within the scene's coordinate system (Fig. 24).

The analysis and experiment, coupled with the calculations in 3 bode well for the concept of distributed compressive imaging. Randomized projections of limited field-of-view images seem to contain enough information to not only recover the underlying large area image, but also estimate the viewing geometry of each individual camera. This conclusion is also supported by the experiments conducted by Wakin [11] (Fig. 25).

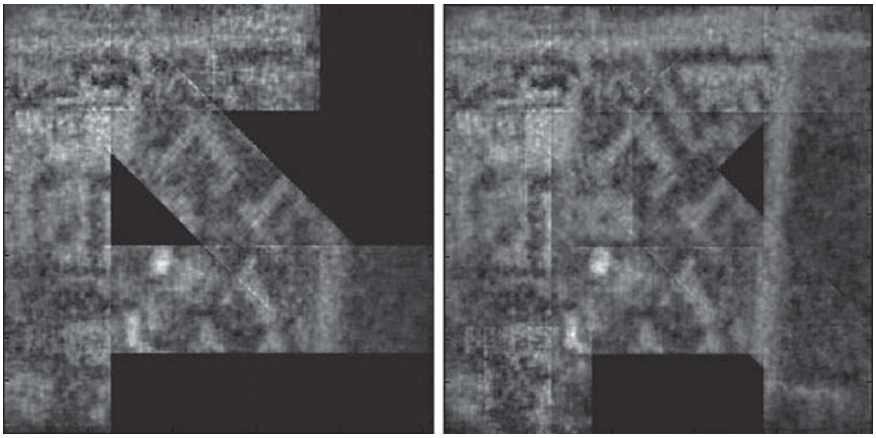


Fig. 25 The scene being surveyed by compressive sensing cameras

4 Conclusions and Further Research

In this work we looked at different surveillance problems and the results that compressive sensing approaches can deliver. The LDS method is capable of reconstructing certain types of surveillance scenes with a high degree of accuracy. This model also enjoys the ability to reduce the number of measurements needed for each frame of video, so long as there is a sufficiently large number of frames available. The major drawback of this model is that it fails to reconstruct the features of dynamics that are not present in each frame. This drawback presents us with an opportunity for future research, with questions of why this model fails in these instances and whether or not it can be generalized to allow it to reconstruct additional classes of video (Fig. 26).

In the context of motion sensing, we have presented results that show that motion information can be sensed directly by a compressive imager. The results were noisy, but the silhouettes of the moving objects were preserved. Further, we demonstrated that even a very naive filtering method could get rid of most of the noise. There are limitations to this method, however. In the scene, we observed the object of interest was fairly large relative to the field-of-view. If the object(s) of interest was smaller, say a group of pedestrians from far above, then the pedestrian silhouettes may look like noise. As such, our filtering technique may disregard valuable motion information. One potential solution might be to use optical flow data. If one looks at the optical flow of the reconstructed sequence, surely one will observe mostly erratic motion vectors. However, the (small) portions of the scene that are actually representative of motion should still exhibit stable motion vectors. The portions of the scene associated with the smoothly changing motion vectors could be weighted heavily in a new filtering process. This will help prevent legitimate motion from being regarded as noise (Fig. 27).

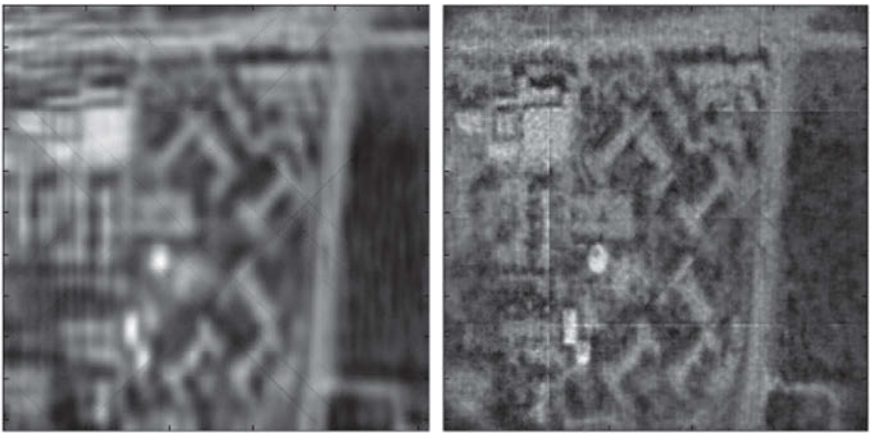


Fig. 26 On the *left* is the complete reconstruction via traditional imaging. On the *right* is the reconstruction via compressive sensing

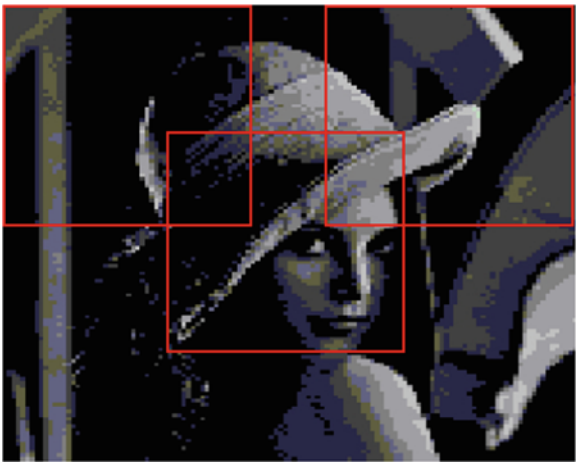


Fig. 27 How the Lena image is being sensed

The third problem we looked at was that of wide-area surveillance. We have shown through analysis and simulation that there is significant benefit in distributed compressive imaging (DCI) to sense a very large area with significant benefits when there are severe bandwidth transmission restrictions. We have shown that the same criteria which allows compressive sensing to work (namely minimizing the L1-norm of the reconstruction coefficients) is also a viable criteria to estimate the registration parameters of the multiple cameras. It is particularly beneficial that one can take advantage of the redundancy of multiple cameras without intra-camera communications (something unattainable with traditional compression). A topic for further research is some combination of the manifold lifting algorithm developed by Wakin [11] with the L1-minimization techniques outlined in this paper. This might lead to a faster

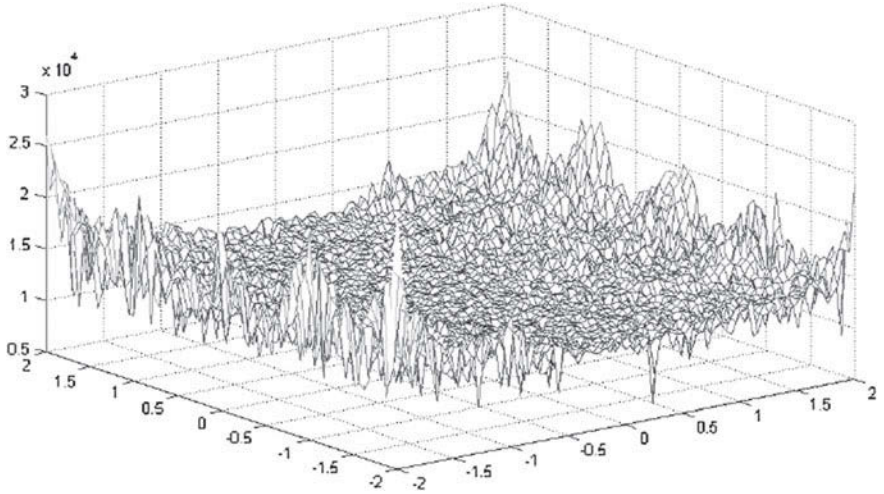


Fig. 28 The x and y axes represent the guess for γ, while the z *axis* represents the l_1 norm of the reconstruction given γ. There are 81 nodes in both the x and y directions

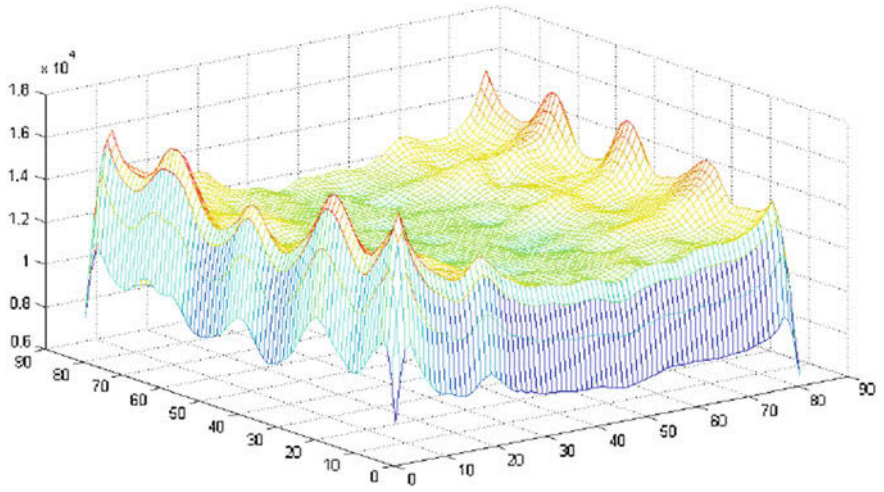


Fig. 29 The smoothed graph of the l_1 norms as a function of γ

method by which to accurately estimate the camera registration parameters (Figs. 27, 28, 29, 30 and 31).

Another topic for further research would be to determine an effective way to incorporate prior information about a scene into the model. This information should be used in a way that would increase the sparsity of the system (so that fewer measurements need to be taken) and/or decrease the number of iterations needed to converge to an accurate solution to the system. As an example, consider the wide-area surveillance application we discussed. Suppose that a low-resolution photo of the

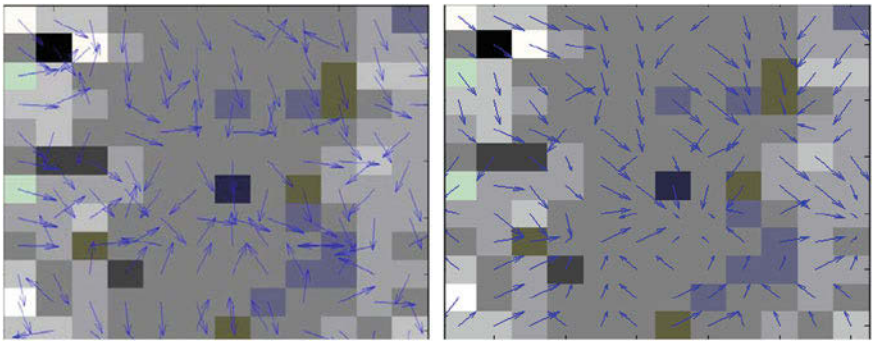


Fig. 30 The graph on the left displays the gradient of $||x||_1$ before being smoothed. There is no indication that a gradient descent search will converge to the correct solution. The graph on the *right* displays the *gradient* of the smoothed l_1 function. There is a clear convergence to a point which is very close to the ideal perspective estimates

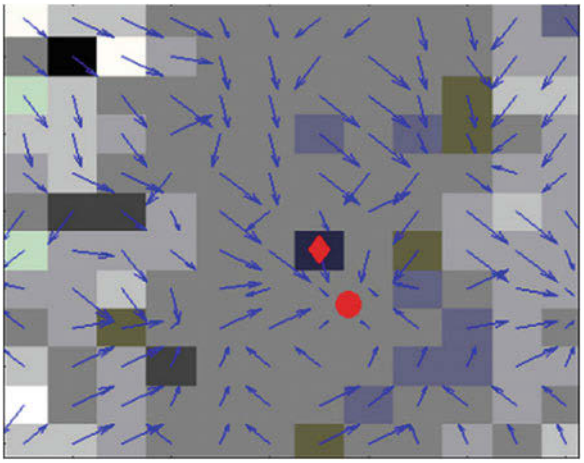


Fig. 31 The *red* diamond displays the location of the true camera registration. The *circle* displays the location of the point that the graph's gradient converges to

entire track of land was available (this could be thought of as being given by a satellite with typical optics, without need for high-resolution capabilities). The resolution would be relatively poor, but the overall shape of the image could be captured. A good question to ask, then, would be if one could use this information to speed up the reconstruction process. The CoSAMP algorithm uses a support pruning procedure. Could knowing roughly what the scene should look like help to more efficiently hone in on what the correct support of the sparse solution is? The ability to use such information would make for a novel algorithm and would contribute greatly in the applicability of compressive sensing to surveillance and imaging problems.

References

1. Baraniuk, R., Sankaranarayanan, A., Turaga, P., Chellappa, R.: Compressive acquisition of dynamic scenes. In: Proceedings of the 11th European Conference on Computer Vision: Part I, ECCV'10, pp. 129–142. Springer, Berlin (2010)
2. Bhagavatula, V.K., Haberfelde, T., Mahalanobis, A., Neifeld, M., Brady, D.: Off-axis sparse aperture imaging using phase optimization techniques for application in wide-area imaging systems. Appl. Opt. **48**(28), 5212–5224 (2009)
3. Candes, E., Tao, T.: Near-optimal signal recovery from random projections: universal encoding strategies? IEEE Trans. Inf. Theor. **54**(12), 5406–5425 (2006)
4. Drori, I., Donoho, D., Tsaig, Y., Starck, J.: Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit. Technical report (2006)
5. Duarte, M., Reddy, D., Baraniuk, R., Cevher, V., Sankaranarayanan, A., Chellappa, R.: Compressive sensing for background subtraction. In: Proceedings of the European Conference on Computer Vision (ECCV) (2008)
6. Huff, C., Muise, R.: Wide-area surveillance with multiple cameras using distributed compressive imaging. In: Proceedings of the SPIE (2011)
7. Huff, C.: Applications of compressive sensing to surveillance problems. Master's thesis, University of Central Florida, Orlando (2012)
8. Muise, R.: Compressive imaging: an application. SIAM J. Imaging Sci. **2**(4), 1255–1276 (2009)
9. Romberg, J., Candes, E., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. **59**(8), 1207–1223 (2006)
10. Van Trees, H.: Detection, Estimation, and Modulation Theory. Part I. Wiley, New York (1968)
11. Wakin, M.: A manifold lifting algorithm for multi-view compressive imaging. In: Proceedings of the Picture Coding Symposium (2009)

Chapter 10
Region of Variability for Some Subclasses of Univalent Functions

A. Vasudevarao

Abstract Let $\mathscr{A}$ denote the class of analytic functions f in the unit disk $\mathbb{D}$ with $f(0) = 0$ and $f'(0) = 1$. Let $\mathscr{S}$ denote the class of univalent functions in $\mathscr{A}$. Let $\widetilde{\mathscr{F}}$ (for example, class of starlike, convex, close-to-convex, spiralike, etc.) be any arbitrary subfamily of $\mathscr{S}$ and $z_0 \in \mathbb{D}$ then upper and lower estimates of $|f(z_0)|$, $|f'(z_0)|$ and $\operatorname{Arg} f'(z_0)$ for all $f \in \widetilde{\mathscr{F}}$ are respectively called a growth theorem, a distortion theorem and a rotation theorem at z_0 for $\widetilde{\mathscr{F}}$. These estimates deal only with absolute values of $f(z_0)$ and $f'(z_0)$ or with the argument of $f'(z_0)$. The aim of this paper is to give a survey on regions of variability of $f(z_0)$ or $f'(z_0)$ or $\log f'(z_0)$ when f ranges over some well-known subclasses of $\mathscr{S}$. As a consequence, we present the sharp Pre-Schwarzian norm and Block semi-norm for some of the subclasses of $\mathscr{S}$. We also graphically illustrate the region of variability for several sets of parameters.

1 Introduction

Let $\mathbb{D} := \{z : |z| < 1\}$ be the unit disk in the complex plane $\mathbb{C}$ and $\mathscr{H}$ denote the space of all analytic functions on $\mathbb{D}$. Here we think of $\mathscr{H}$ as a topological vector space endowed with the topology of uniform convergence over compact subsets of $\mathbb{D}$. Further, let $\mathscr{A} := \{f \in \mathscr{H} : f(0) = 0 = f'(0) - 1\}$. If $f \in \mathscr{A}$ then $f(z)$ has the following representation

$$f(z) = z + \sum_{n=2}^{\infty} a_n z^n.$$

The author thank SRIC, IIT Kharagpur (ref. IIT/SRIC/MA/SUA/2012-13/144) for the financial support.

A. Vasudevarao (✉)
Department of Mathematics, Indian Institute of Technology Kharagpur,
Kharagpur 721302, India
e-mail: alluvasu@maths.iitkgp.ernet.in

R. N. Mohapatra et al. (eds.), *Mathematics and Computing 2013*,
Springer Proceedings in Mathematics & Statistics 91,
DOI: 10.1007/978-81-322-1952-1_10, © Springer India 2014

A single-valued function f is said to be univalent in a domain $\Omega \subset \mathbb{C}$ if it is one-to-one in Ω. Let $\mathscr{S}$ denote the class of univalent functions in $\mathscr{A}$. A function $f \in \mathscr{A}$ is called starlike if $f(\mathbb{D})$ is a starlike domain with respect to the origin, and the class of univalent starlike functions is denoted by $\mathscr{S}^*$. Each univalent starlike function f is characterized by

$$\operatorname{Re}\left(\frac{zf'(z)}{f(z)}\right) > 0 \quad \text{for } z \in \mathbb{D}.$$

A function $f \in \mathscr{A}$ is called convex if $f(\mathbb{D})$ is a convex domain. We denote the class of univalent convex functions in $\mathscr{A}$ by $\mathscr{C}$. It is known that a function $f \in \mathscr{A}$ is in $\mathscr{C}$ if and only if

$$\operatorname{Re}\left(1 + \frac{zf''(z)}{f'(z)}\right) > 0 \quad \text{for } z \in \mathbb{D}.$$

It is geometrically evident that a convex domain is starlike with respect to each of its points. Hence, $\mathscr{C} \subsetneq \mathscr{S}^*$. The Koebe function $k(z) = z/(1-z)^2$ shows that this containment is proper. For each $f \in \mathscr{S}^*$ the Alexander transformation [3] defined by

$$g(z) = \int_0^z \frac{f(t)}{t}\,\mathrm{d}t$$

is convex. This transformation provides a nice bridge between functions in $\mathscr{C}$ and $\mathscr{S}^*$. A function $f \in \mathscr{A}$ is called close-to-convex (see [13]) if there exists a convex (univalent) function g and a number $\phi \in \mathbb{R}$ such that

$$\operatorname{Re}\left(\mathrm{e}^{i\phi}\frac{f'(z)}{g'(z)}\right) > 0 \quad \text{for } z \in \mathbb{D}.$$

We denote the class of close-to-convex functions in $\mathscr{A}$ by $\mathscr{K}$. Also, it is known that every close-to-convex function is univalent in $\mathbb{D}$. There is another natural generalization of starlike functions, namely spirallike functions which again leads to a useful criterion for univalence. A function $f \in \mathscr{A}$ is α-spirallike if for some real constant α ($|\alpha| < \pi/2$),

$$\operatorname{Re}\left(\mathrm{e}^{i\alpha}\frac{zf'(z)}{f(z)}\right) > 0 \quad \text{for } z \in \mathbb{D}.$$

A complete characterization of spirallike functions by means of subordination has been investigated by Ruscheweyh [34]. Also, we refer to the survey by Ahuja and Silverman [2] for several other important properties and characterizations of spirallike functions.

Although, the class of starlike functions (with respect to an interior point) has been studied extensively among many other subclasses, not much is known about starlike functions with respect to a boundary point until the work of Robertson (see

[33]). Motivated by the work in [33] and characterizations of this class of functions, some advancement in this direction has taken place (see [9, 14, 16, 35]). On the other hand, there does not seem to be any development on spirallike functions with respect to a boundary point until the recent work of Elin et al. [8] (see also [9]). More recently, Aharonov et al. [1] provided a natural geometric approach for discussing spirallike functions with respect to a boundary point.

Let $\mathscr{F}_\mu$ denote the class of functions $f \in \mathscr{H}$, which is non-vanishing in $\mathbb{D}$ with $f(0) = 1$, and for $\mu \in \mathbb{C}$, such that $\operatorname{Re} \mu > 0$, satisfying

$$\operatorname{Re}\left(\frac{2\pi}{\mu}\frac{zf'(z)}{f(z)} + \frac{1+z}{1-z}\right) > 0 \quad \text{for } z \in \mathbb{D}.$$

Functions in the class $\mathscr{F}_\mu$ are called spirallike functions with respect to a boundary point. The basic properties and a number of equivalent characterizations of the class $\mathscr{F}_\mu$ have been studied in [1]. In particular, if $\mu = \pi$ the class $\mathscr{F}_\pi$ coincides with the class of starlike functions with respect to a boundary point introduced by Robertson (see [33]) This has led to considerable research in this class and associated classes. It is also known that functions in $\mathscr{F}_\pi$ are either close-to-convex or just the constant 1.

An analytic univalent function f in $\mathbb{D}$ is called exponentially convex if $\mathrm{e}^{f(z)}$ maps $\mathbb{D}$ onto a convex domain. For $\alpha \in \mathbb{C}\setminus\{0\}$, the family $\mathscr{E}(\alpha)$ of α-exponential functions was introduced in [4]. A function $f \in \mathscr{S}$ is said to be in $\mathscr{E}(\alpha)$ if $F(\mathbb{D})$ is a convex domain, where $F(z) = \mathrm{e}^{\alpha f(z)}$. Although Arango et al. [4] studied exponentially convex functions in 1997, no attempt has been made until the recent work [23] on the region of variability for this class.

Let f and g be analytic functions in the unit disk $\mathbb{D}$. The function f is said to be subordinate to g, written as $f \prec g$ or $f(z) \prec g(z)$, if there exists a function ω analytic in $\mathbb{D}$, with $\omega(0) = 0$ and $|\omega| < 1$, and such that $f(z) = g(\omega(z))$. If g is univalent, then $f \prec g$ if and only if $f(0) = g(0)$ and $f(\mathbb{D}) \subset g(\mathbb{D})$. For a detailed study on differential subordination, we refer to the monograph of Miller and Mocanu [17].

The class of univalent functions is preserved under a number of elementary transformations. The preservation of $\mathscr{S}$ under the disk automorphism (also called the *Koebe transform*) leads to the study of the behaviour of the pre-Schwarzian norm of f given by

$$\|f\| := \sup_{z\in\mathbb{D}}(1-|z|^2)\left|\frac{f''(z)}{f'(z)}\right|,$$

where f is locally univalent function in $\mathbb{D}$. A function f is called a Bloch function (see [28, p. 72]) if it is analytic in $\mathbb{D}$ and

$$\|f\|_{\mathscr{B}} := \sup_{z\in\mathbb{D}}(1-|z|^2)|f'(z)| < \infty.$$

This defines a semi-norm, and the class of Bloch functions forms a complex Banach space $\mathscr{B}$ with respect to the Bloch norm $|f(0)| + \|f\|_{\mathscr{B}}$. It is well known that

$\|f\|_{\mathscr{B}}$ is conformally invariant. That is, if h is a conformal automorphism of $\mathbb{D}$, then $\|f \circ h\|_{\mathscr{B}} = \|f\|_{\mathscr{B}}$. There is a close connection between Bloch functions and univalent functions, in particular with their derivatives (see [28]). That is if f maps $\mathbb{D}$ conformally into $\mathbb{C}$ then $\|\log(f-a)\|_{\mathscr{B}} \leq 4$ for $a \notin f(\mathbb{D})$ and $\|\log f'\|_{\mathscr{B}} \leq 6$. Conversely, if $\|g\|_{\mathscr{B}} \leq 1$ then $g = \log f'$ for some conformal map f. We refer to the monographs by Duren [7], Goodman [11] and Pommerenke [28] for a detailed study of analytic univalent functions.

We need to recall the following lemma which plays a vital role in proving our main results on regions of variability. For a positive integer p, let

$$(\mathscr{S}^*)^p = \{f = f_0^p : f_0 \in \mathscr{S}^*\}.$$

Lemma 1 *Let f be an analytic function in $\mathbb{D}$ with $f(z) = z^p + \cdots$. If*

$$\operatorname{Re}\left(z\frac{f''(z)}{f'(z)}\right) > -1, \quad z \in \mathbb{D},$$

then $f \in (\mathscr{S}^*)^p$.

Although we could not find any historical reference for a proof of Lemma 1, it might be well-known (see [11, 12]). For an analytic proof of Lemma 1 we refer to [38].

Let $\widetilde{\mathscr{F}} \subset \mathscr{A}$ and $z_0 \in \mathbb{D}$. Then upper and lower estimates of the form

$$K_1 \leq |f(z_0)| \leq K_2, \quad M_1 \leq |f'(z_0)| \leq M_2, \quad m_1 \leq \operatorname{Arg} f'(z_0) \leq m_2 \quad \text{for all } f \in \widetilde{\mathscr{F}}$$

are respectively called a growth theorem, a distortion theorem and a rotation theorem at z_0 for $\widetilde{\mathscr{F}}$, where K_i, M_i and m_i $(i = 1, 2)$ are some non-negative constants. These estimates deal only with absolute values of $f(z_0)$ and $f'(z_0)$ or with the argument of $f'(z_0)$. If one wants to study the exact value of $f(z_0)$ or $f'(z_0)$, then it is necessary to consider the region of variability of $f(z_0)$ or $f'(z_0)$ when f ranges over the class $\widetilde{\mathscr{F}}$. For example it is known that for each fixed $z_0 \in \mathbb{D}$,

$$\left\{\log\left(\frac{f(z_0)}{z_0}\right) : f \in \mathscr{S}\right\}$$

is precisely a closed disk, and $\{\log \phi'(z_0) : \phi \in \mathscr{C}\}$ is the set $\{\log(1-z)^{-2} : |z| \leq |z_0|\}$ (see also [7, Exercises 10, 11 and 13 in Chap. 2]).

2 Main Results

In 2005, the region of variability for functions of bounded derivative and of positive real part has been discussed in [38]. Also, the region of variability of $\log f'(z_0)$ when f ranges over the class of convex functions f with $f''(0) = 2\lambda$ (where $\lambda \in \overline{\mathbb{D}}$) has

been investigated in [39]. The aim of this paper is to provide region of variability for well-known subclasses of the class of univalent functions.

In [20] the authors considered the following two subclasses $\mathscr{F}_1$ and $\mathscr{F}_2$ of $\mathscr{S}$ to determine the region of variability. More precisely, Let $\mathscr{F}_1$ ($\mathscr{F}_2$ respectively) denote the subclass of locally univalent normalized functions $f \in \mathscr{A}$ such that

$$\operatorname{Re} P_f(z) < \frac{3}{2} \left(\operatorname{Re} P_f(z) > -\frac{1}{2} \text{ respectively} \right), \quad z \in \mathbb{D},$$

where

$$P_f(z) = 1 + \frac{zf''(z)}{f'(z)}, \quad z \in \mathbb{D}.$$

It is well-known that (see [29, Eq. (16)] and [30]) $f \in \mathscr{F}_1$ implies

$$\left| \frac{zf'(z)}{f(z)} - \frac{2}{3} \right| < \frac{2}{3},$$

which implies that

$$\operatorname{Re} \left(\frac{zf'(z)}{f(z)} \right) > 0, \quad \text{for } z \in \mathbb{D}.$$

Hence $\mathscr{F}_1 \subset \mathscr{S}^*$. Also $\mathscr{F}_2 \subset \mathscr{K}$. For $f \in \mathscr{F}_j$ $(j = 1, 2)$, we denote by $\log f'$ the single-valued branch of the logarithm of f' with $\log f'(0) = 0$. Using the Herglotz representation for analytic function with positive real part in $\mathbb{D}$, we can write that if $f \in \mathscr{F}_1$, then there exists a unique positive unit measure μ on $(-\pi, \pi]$ such that

$$1 - 2\frac{zf''(z)}{f'(z)} = \int_{-\pi}^{\pi} \frac{1 + ze^{-it}}{1 - ze^{-it}} \, d\mu(t).$$

This easily gives

$$\log f'(z) = \int_{-\pi}^{\pi} \log \left(1 - ze^{-it} \right) d\mu(t).$$

It follows that for each fixed $z_0 \in \mathbb{D}$ the region of variability

$$\left\{ \log f'(z_0) : \ f \in \mathscr{F}_1 \right\}$$

coincides with the set $\{\log(1 - z) : \ |z| \leq |z_0|\}$. Similarly if $f \in \mathscr{F}_2$ then by applying the Herglotz formula we obtain

$$1+\frac{zf''(z)}{f'(z)}=-\frac{1}{2}+\frac{3}{2}\int\limits_{-\pi}^{\pi}\frac{1+ze^{-it}}{1-ze^{-it}}\,d\mu(t)$$

from which we can easily deduce that

$$\log f'(z)=3\int\limits_{-\pi}^{\pi}\log\left(\frac{1}{1-ze^{-it}}\right)d\mu(t)$$

and so for each fixed $z_0\in\mathbb{D}$ the region of variability

$$\left\{\log f'(z_0):\ f\in\mathscr{F}_2\right\}$$

coincides with the set $\{-3\log(1-z):\ |z|\le|z_0|\}$. Although one may question the significance of the classes $\mathscr{F}_1$ and $\mathscr{F}_2$ but on the positive side, we give a precise description of the region of variability of $\log f'(z_0)$ which always is a nice feature. To make this point precise, for $\lambda\in\overline{\mathbb{D}}$ and for $z_0\in\mathbb{D}$ fixed, we define

$$\begin{aligned}\mathscr{C}_1(\lambda)&=\{f\in\mathscr{F}_1:\ f''(0)=-\lambda\}\\ \mathscr{C}_2(\lambda)&=\{f\in\mathscr{F}_2:\ f''(0)=3\lambda\}\\ V_j(z_0,\lambda)&=\{\log f'(z_0):\ f\in\mathscr{C}_j(\lambda)\},\quad\text{for } j=1,2.\end{aligned}$$

The basic properties of the set $V_j(z_0,\lambda)$, $j=1,2$ are:

Corollary 1 *We have*

1. *For each $j=1,2$, $V_j(z_0,\lambda)$ is a compact subset of $\mathbb{C}$.*
2. *For each $j=1,2$, $V_j(z_0,\lambda)$ is a convex subset of $\mathbb{C}$.*
3. *If $|\lambda|=1$ or $z_0=0$, then*

$$V_1(z_0,\lambda)=\{\log(1-\lambda z_0)\}\ \text{and}\ V_2(z_0,\lambda)=\{-3\log(1-\lambda z_0)\}.$$

4. *For $|\lambda|<1$ and $z_0\neq 0$, the set $V_1(z_0,\lambda)$ has $\log(1-\lambda z_0)$ as an interior point, whereas the set $V_2(z_0,\lambda)$ has $-3\log(1-\lambda z_0)$ as an interior point.*
5. *For each $j=1,2$ $V_j(e^{i\theta}z_0,\lambda)=V_j(z_0,e^{i\theta}\lambda)$ for $\theta\in\mathbb{R}$.*

In view of the property (5) in Corollary 1, it is sufficient to determine $V_j(z_0,\lambda)$ $(j=1,2)$ for $0\le\lambda<1$ and $z_0\in\mathbb{D}$. For $z,a\in\mathbb{D}$, we define

$$\delta(z,a)=\frac{z+a}{1+\overline{a}z}.$$

Let $a\in\overline{\mathbb{D}}$, $\lambda\in[0,1)$. If $f\in\mathscr{C}_1(\lambda)$ then we introduce

$$F_{a,\lambda}(z) = \int_0^z \exp\left\{ \int_0^{\zeta_2} \frac{\delta(a\zeta_1, \lambda)}{\zeta_1 \delta(a\zeta_1, \lambda) - 1} \, \mathrm{d}\zeta_1 \right\} \mathrm{d}\zeta_2, \quad z \in \mathbb{D}, \tag{1}$$

and for $f \in \mathscr{C}_2(\lambda)$ we put

$$G_{a,\lambda}(z) = \int_0^z \exp\left\{ \int_0^{\zeta_2} \frac{3\delta(a\zeta_1, \lambda)}{1 - \zeta_1 \delta(a\zeta_1, \lambda)} \, \mathrm{d}\zeta_1 \right\} \mathrm{d}\zeta_2, \quad z \in \mathbb{D}. \tag{2}$$

It is not difficult to see that $F_{a,\lambda} \in \mathscr{C}_1(\lambda)$ and $G_{a,\lambda} \in \mathscr{C}_2(\lambda)$.

The following results give the precise description of regions of variability for the classes $\mathscr{C}_j(\lambda)$ for $j = 1, 2$.

Theorem 1 *For $0 \le \lambda < 1$ and $z_0 \in \mathbb{D} \setminus \{0\}$, the boundary $\partial V_1(z_0, \lambda)$ is the Jordan curve given by*

$$(-\pi, \pi] \ni \theta \mapsto \log F'_{\mathrm{e}^{i\theta},\lambda}(z_0) = \int_0^{z_0} \frac{\delta(\mathrm{e}^{i\theta} z, \lambda)}{z\delta(\mathrm{e}^{i\theta} z, \lambda) - 1} \, \mathrm{d}z.$$

If $\log f'(z_0) = \log F'_{\mathrm{e}^{i\theta},\lambda}(z_0)$ for some $f \in \mathscr{C}_1(\lambda)$ and $\theta \in (-\pi, \pi]$, then $f(z) = F_{\mathrm{e}^{i\theta},\lambda}(z)$. Here $F_{\mathrm{e}^{i\theta},\lambda}(z)$ is given by (1).

Theorem 2 *For $0 \le \lambda < 1$ and $z_0 \in \mathbb{D} \setminus \{0\}$, the boundary $\partial V_2(z_0, \lambda)$ is the Jordan curve given by*

$$(-\pi, \pi] \ni \theta \mapsto \log G'_{\mathrm{e}^{i\theta},\lambda}(z_0) = \int_0^{z_0} \frac{3\delta(\mathrm{e}^{i\theta} z, \lambda)}{1 - z\delta(\mathrm{e}^{i\theta} z, \lambda)} \, \mathrm{d}z.$$

If $\log f'(z_0) = \log G'_{\mathrm{e}^{i\theta},\lambda}(z_0)$ for some $f \in \mathscr{C}_2(\lambda)$ and $\theta \in (-\pi, \pi]$, then $f(z) = G_{\mathrm{e}^{i\theta},\lambda}(z)$. Here $G_{\mathrm{e}^{i\theta},\lambda}(z)$ is given by (2).

Another class of our interest is $\mathscr{S}_\alpha$. More precisely, for $-\pi/2 < \alpha < \pi/2$, we say that $f \in \mathscr{S}_\alpha$ provided $f \in \mathscr{A}$ is locally univalent in $\mathbb{D}$ and

$$\operatorname{Re} \mathrm{e}^{i\alpha} \left(1 + \frac{zf''(z)}{f'(z)} \right) > 0, \quad z \in \mathbb{D}.$$

It is easy to see that $f \in \mathscr{S}_\alpha$ if and only if there exists a function $g \in \mathscr{S}^*$ such that

$$f'(z) = \left(\frac{g(z)}{z} \right)^{(\cos\alpha)\exp(-i\alpha)}.$$

Also, we observe that the above conditions are precisely the conditions for the function $zf'(z)$ to belong the class of spirallike functions. The class $\mathcal{S}_0$ consists of the normalized convex functions. For a general value of α ($|\alpha| < \pi/2$), a function in $\mathcal{S}_\alpha$ need not be univalent in $\mathbb{D}$. For example, the function $f(z) = i(1-z)^i - i$ is known to belong to $\mathcal{S}_{\pi/4} \setminus \mathcal{S}$. In 1968, Robertson [32] showed that $f \in \mathcal{S}_\alpha$ is univalent if $0 < \cos\alpha \leq 0.2315\cdots$ and showed that there are non-univalent functions $f \in \mathcal{S}_\alpha$ for each α, $1/2 < \alpha < 1$. Subsequently Libera and Zeigler [15] improved the range of univalency of $f \in \mathcal{F}_\alpha$ to $0 < \cos\alpha \leq 0.2564\cdots$. In 1975, Chichra [6] has improved the range still further to $0 < \cos\alpha \leq 0.2588\cdots$ and indicated that his result is the best possible one obtainable solely from an application of Nehari's test for univalence [18]. In the same year Pfaltzgraff [19] has shown that $f \in \mathcal{S}_\alpha$ is univalent whenever $0 < \cos\alpha \leq 1/2$. This settles the improvement of range of α for which $f \in \mathcal{S}_\alpha$ is univalent. On the other hand, Singh [36] has shown that functions in $\mathcal{S}_\alpha$ which satisfy $f''(0) = 0$ are univalent for all real values of α with $|\alpha| < \pi/2$.

For $f \in \mathcal{S}_\alpha$, $\log f'(z)$ denotes the single-valued branch of the logarithm of $f'(z)$ with $\log f'(0) = 0$. The Herglotz representation for analytic function with positive real part in $\mathbb{D}$ shows that if $f \in \mathcal{S}_\alpha$, then there exists a unique positive unit measure μ on $(-\pi, \pi]$ such that

$$1 + \frac{zf''(z)}{f'(z)} = \mathrm{e}^{-i\alpha}\left[\cos\alpha \int_{-\pi}^{\pi} \frac{1 + z\mathrm{e}^{-it}}{1 - z\mathrm{e}^{-it}}\,\mathrm{d}\mu(t) + i\sin\alpha\right]$$

from which we obtain that

$$\log f'(z) = 2\mathrm{e}^{-i\alpha}\cos\alpha \int_{-\pi}^{\pi} \log\left(\frac{1}{1 - z\mathrm{e}^{-it}}\right)\mathrm{d}\mu(t).$$

In view of this formula, for a fixed $z_0 \in \mathbb{D}$, the region of variability of

$$\left\{\log f'(z_0) : f \in \mathcal{S}_\alpha\right\}$$

coincides with the set

$$\left\{-\left(2\mathrm{e}^{-i\alpha}\cos\alpha\right)\log(1-z) : |z| \leq |z_0|\right\}.$$

Let $\mathcal{B}_0$ be the class of analytic functions ω in $\mathbb{D}$ such that $|\omega(z)| \leq 1$ in $\mathbb{D}$ and $\omega(0) = 0$. Functions in $\mathcal{B}_0$ are called Schwarz functions. It is easy to see that for each $f \in \mathcal{S}_\alpha$ there exists an $\omega_f \in \mathcal{B}_0$ such that

$$\omega_f(z) = \frac{\mathrm{e}^{i\alpha}\left(1 + \frac{zf''(z)}{f'(z)}\right) - \mathrm{e}^{i\alpha}}{\mathrm{e}^{i\alpha}\left(1 + \frac{zf''(z)}{f'(z)}\right) + \mathrm{e}^{-i\alpha}}, \quad z \in \mathbb{D}.$$

Further, if $f \in \mathscr{S}_\alpha$ then a simple computation shows that

$$e^{-i\alpha}\frac{d}{dz}\, e^{i\alpha}\left(1+\frac{zf''(z)}{f'(z)}\right)\Big|_{z=0} = f''(0) = \left(2e^{-i\alpha}\cos\alpha\right)\omega_f'(0).$$

Since $\omega_f \in \mathscr{B}_0$, the Schwarz lemma then gives that $|f''(0)| \le 2\cos\alpha$. Now for $\lambda \in \overline{\mathbb{D}}$ and for $z_0 \in \mathbb{D}$ fixed, we introduce

$$\begin{aligned}\mathscr{S}_\alpha(\lambda) &= \{f \in \mathscr{S}_\alpha : f''(0) = 2\lambda e^{-i\alpha}\cos\alpha\}, \text{ and}\\ V_3(z_0,\lambda) &= \{\log f'(z_0) : f \in \mathscr{S}_\alpha(\lambda)\}.\end{aligned}$$

In [21] the explicit region of variability $V_3(z_0,\lambda)$ of $\log f'(z_0)$ when f ranges over the class $\mathscr{S}_\alpha(\lambda)$ has been investigated.

Proposition 1 *For $f \in \mathscr{S}_\alpha(\lambda)$ we have*

$$\left|\frac{f''(z)}{f'(z)} - c_1(z,\lambda)\right| \le r_1(z,\lambda), \quad z \in \mathbb{D},$$

where

$$\begin{aligned}c_1(z,\lambda) &= \frac{(2e^{-i\alpha}\cos\alpha)\left\{\lambda\left(1-|z|^2\right)+\overline{z}\left(|z|^2-\lambda^2\right)\right\}}{\left(1-|z|^2\right)\left(1-2\lambda\,\mathrm{Re}\,z+|z|^2\right)}, \quad \textit{and}\\ r_1(z,\lambda) &= \frac{2(1-\lambda^2)|z|\cos\alpha}{\left(1-|z|^2\right)\left(1-2\lambda\,\mathrm{Re}\,z+|z|^2\right)}.\end{aligned}$$

For each $z \in \mathbb{D}\setminus\{0\}$, equality holds if and only if $f = H_{e^{i\theta},\lambda}$ for some $\theta \in \mathbb{R}$ where

$$H_{e^{i\theta},\lambda}(z) = \int_0^z \exp\left\{\int_0^{\zeta_2} \frac{(2e^{-i\alpha}\cos\alpha)\delta(e^{i\theta}\zeta_1,\lambda)}{1-\zeta_1\delta(e^{i\theta}\zeta_1,\lambda)}\,d\zeta_1\right\} d\zeta_2, \quad z \in \mathbb{D}.$$

The case $\lambda = 0$ of Proposition 1 gives the following interesting result.

Corollary 2 *Let $f \in \mathscr{S}_\alpha(0)$. Then we have*

$$\left|\frac{f''(z)}{f'(z)} - \frac{(2e^{-i\alpha}\cos\alpha)\overline{z}|z|^2}{1-|z|^4}\right| \le \frac{2|z|\cos\alpha}{1-|z|^4}, \quad z \in \mathbb{D}.$$

In particular,

$$(1-|z|^2)\left|\frac{f''(z)}{f'(z)}\right| \le 2|z|\cos\alpha, \quad z \in \mathbb{D}.$$

On the other hand, the case $\alpha = 0$ of Corollary 2 shows that if f is convex with $f''(0) = 0$, then we have the sharp Pre-Schwarzian estimate $\|f\| \le 2$. The convex function

$$f(z) = \frac{1}{2}\log\left(\frac{1+z}{1-z}\right)$$

shows that number 2 cannot be replaced by a smaller number. Moreover

$$\|f\| \leq 2\cos\alpha \quad \text{if} \quad f \in \mathscr{S}_\alpha(0).$$

For $z_0 \in \mathbb{D}\setminus\{0\}$, $\lambda \in \mathbb{D}$ and $f \in \mathscr{S}_\alpha(\lambda)$ one can prove that $V_3(e^{i\theta}z_0, \lambda) = V_3(z_0, e^{i\theta}\lambda)$. In deed, this is a consequence of the fact $e^{-i\theta}f(e^{i\theta}z) \in \mathscr{S}_\alpha(e^{i\theta}\lambda)$ if and only if $f \in \mathscr{S}_\alpha(\lambda)$. In view of this fact it suffices to consider $V_3(z_0, \lambda)$ for $\lambda \in [0, 1)$.

Theorem 3 *For $0 \leq \lambda < 1$ and $z_0 \in \mathbb{D}\setminus\{0\}$, the boundary $\partial V_3(z_0, \lambda)$ is the Jordan curve given by*

$$(-\pi, \pi] \ni \theta \mapsto \log H'_{e^{i\theta},\lambda}(z_0) = \int_0^{z_0} \frac{(2e^{-i\alpha}\cos\alpha)\delta(e^{i\theta}z, \lambda)}{1 - z\delta(e^{i\theta}z, \lambda)}\,dz.$$

If $\log f'(z_0) = \log H'_{e^{i\theta},\lambda}(z_0)$ for some $f \in \mathscr{S}_\alpha(\lambda)$ and $\theta \in (-\pi, \pi]$, then $f(z)$ coincides with $H_{e^{i\theta},\lambda}(z)$.

One of the important subclasses of $\mathscr{S}$ is $\mathscr{K}$. In 2008 (see [22, 25]), the authors considered the region of variability for close-to-convex functions with fixed derivative. More precisely, let α be a complex number for which $\operatorname{Re}\alpha > 0$ and $\phi \in \mathscr{C}$. Let $\mathscr{K}_\phi(\alpha)$ denote the class of functions $f \in \mathscr{H}$ with $f(0) = 0$, $f'(0)/\phi'(0) = \alpha$ and

$$\operatorname{Re}\left(\frac{f'(z)}{\phi'(z)}\right) > 0, \quad z \in \mathbb{D}$$

(with $\phi \in \mathscr{C}$ fixed). For each fixed $z_0 \in \mathbb{D}$, we denote the class $V_\phi(z_0, \alpha)$ by

$$V_\phi(z_0, \alpha) = \{f(z_0) : f \in \mathscr{K}_\phi(\alpha)\}.$$

The basic properties of $V_\phi(z_0, \alpha)$ are:

Proposition 2 *For $f \in \mathscr{K}_\phi(\alpha)$ we have*

1. *$V_\phi(z_0, \alpha)$ is a compact subset of $\mathbb{C}$.*
2. *$V_\phi(z_0, \alpha)$ is a convex subset of $\mathbb{C}$.*
3. *If $z_0 = 0$ then $V_\phi(z_0, \alpha) = \{0\}$.*
4. *For $|c| < 1$ and $z_0 \in \mathbb{D}\setminus\{0\}$, $V_\phi(z_0, \alpha)$ has*

$$\int_0^{z_0} \left(\frac{\alpha + \overline{\alpha}c\zeta}{1 - c\zeta}\right)\phi'(\zeta)\,d\zeta$$

as an interior point.

Theorem 4 *Let $z_0 \in \mathbb{D}$ and $\operatorname{Re}\alpha > 0$. If $z_0 = 0$, then $V_\phi(z_0, \alpha) = \{0\}$. If $z_0 \neq 0$, then $V_\phi(z_0, \alpha)$ is the convex closed Jordan domain surrounded by the simple closed curve $\partial\mathbb{D} \ni c \mapsto f_c(z_0)$, where*

$$f_c(z) = \int_0^z \left(\frac{\alpha + \overline{\alpha} c\zeta}{1 - c\zeta}\right) \phi'(\zeta)\, \mathrm{d}\zeta, \quad z \in \mathbb{D}.$$

Furthermore if $f(z_0) = f_c(z_0)$ for some $f \in \mathcal{K}_\phi(\alpha)$ and $c \in \partial\mathbb{D}$, then $f = f_c$.

The following growth result is a simple consequence of Theorem 4.

Corollary 3 *For $z_0 \in \mathbb{D}\backslash\{0\}$ and $f \in \mathcal{K}_\phi(\alpha)$, we have*

$$\operatorname{Re}\left\{\frac{f(z_0)}{\phi(z_0)}\right\} \leq (\operatorname{Re}\alpha)\left(\frac{1 + |z_0|}{1 - |z_0|}\right) - (\operatorname{Re}\alpha)\int_0^1 \operatorname{Re}\left(\frac{2c}{(1 - cz_0t)^2}\frac{z_0\phi(z_0t)}{\phi(z_0)}\right) \mathrm{d}t.$$

Also we have

$$\operatorname{Re}\left\{\frac{f(z_0)}{\phi(z_0)}\right\} \geq (\operatorname{Re}\alpha)\left(\frac{1 - |z_0|}{1 + |z_0|}\right) - (\operatorname{Re}\alpha)\int_0^1 \operatorname{Re}\left(\frac{2c}{(1 - cz_0t)^2}\frac{z_0\phi(z_0t)}{\phi(z_0)}\right) \mathrm{d}t.$$

By fixing the convex functions $\phi(z)$ by

$$-\log(1 - z), \ \frac{1}{2}\log\left(\frac{1 + z}{1 - z}\right), \ \frac{z}{1 - z}$$

in Theorem 4 we can obtain precise regions of variability of $f(z_0)$ when $f \in \mathcal{K}_\phi(\alpha)$ (see [22, 25]).

Although, the class of starlike functions (with respect to an interior point) has been studied extensively among many other subclasses, little was known about starlike functions with respect to a boundary point until the work of Robertson [33]. Motivated by the work in [33] and characterizations of this class of functions, some advancement in this direction has taken place (see [9, 14, 16, 35]). On the other hand, there does not seem to be any development on spirallike functions with respect to a boundary point until the recent work of Elin et al. [8] (see also [9]). More recently, Aharonov et al. [1] provide a natural geometric approach to discuss spirallike functions with respect to a boundary point and the Ref. [1] contains the result of others.

Let $\mathcal{F}_\mu$ denote the class of functions $f \in \mathcal{H}$, and non-vanishing in $\mathbb{D}$ with $f(0) = 1$, and for $\mu \in \mathbb{C}$, such that $\operatorname{Re}\mu > 0$ satisfying

$$\mathrm{Re}\left(\frac{2\pi}{\mu}\frac{zf'(z)}{f(z)}+\frac{1+z}{1-z}\right)>0 \quad \text{for } z\in\mathbb{D}.$$

Basic properties and a number of equivalent characterizations of the class $\mathscr{F}_\mu$ are formulated in [1]. The case $\mu=\pi$ coincides with the class starlike functions with respect to a boundary point introduced by Robertson [33] who has generated interest on this class. Let $f(z)$ be analytic in $\mathbb{D}$ with $f(0)=1$. Then $f\in\mathscr{F}_\pi$ if and only if there exists a function $S(z)\in\mathscr{S}^*(1/2)$ such that

$$f(z)=(1-z)\left(\frac{S(z)}{z}\right).$$

In 2006, Elin [10] obtained the growth theorem and distortion theorem for functions in the class $\mathscr{F}_\pi$. For $\lambda\in\overline{\mathbb{D}}$ and for $z_0\in\mathbb{D}$ fixed we introduce

$$\mathscr{F}_\mu(\lambda)=\left\{f\in\mathscr{F}_\mu:\ f'(0)=\frac{\mu}{\pi}(\lambda-1)\right\}$$
$$V_4(z_0,\lambda)=\{\log f(z_0):\ f\in\mathscr{F}_\mu(\lambda)\}$$

In [23] the region of variability $V_4(z_0,\lambda)$ for $\log f(z_0)$ when f ranges over the class $\mathscr{F}_\mu(\lambda)$ has been investigated. Some of the basic properties of $V_4(z_0,\lambda)$ are:

Proposition 3 *For $f\in\mathscr{F}_\mu(\lambda)$ we have*

1. *$V_4(z_0,\lambda)$ is a compact and convex subset of $\mathbb{C}$.*
2. *For $|\lambda|=1$ or $z_0=0$,*

$$V_4(z_0,\lambda)=\left\{\frac{\mu}{\pi}\log\left(\frac{1-z_0}{1-\lambda z_0}\right)\right\}.$$

3. *For $|\lambda|<1$ and $z_0\in\mathbb{D}\setminus\{0\}$, $V_4(z_0,\lambda)$ has $(\mu/\pi)\log\left(\frac{1-z_0}{1-\lambda z_0}\right)$ as an interior point.*

The precise geometric description of the set $V_4(z_0,\lambda)$ is:

Theorem 5 *For $\lambda\in\mathbb{D}$ and $z_0\in\mathbb{D}\setminus\{0\}$, the boundary $\partial V_4(z_0,\lambda)$ is the Jordan curve given by*

$$(-\pi,\pi]\ni\theta\mapsto\log J_{e^{i\theta},\lambda}(z_0)=\frac{\mu}{\pi}\int_0^{z_0}\frac{\delta(e^{i\theta}\zeta,\lambda)-1}{(1-\delta(e^{i\theta}\zeta,\lambda)\zeta)(1-\zeta)}\,d\zeta. \tag{3}$$

If $\log f(z_0)=\log J_{e^{i\theta},\lambda}(z_0)$ for some $f\in\mathscr{F}_\mu(\lambda)$ and $\theta\in(-\pi,\pi]$, then $f(z)=J_{e^{i\theta},\lambda}(z)$ where

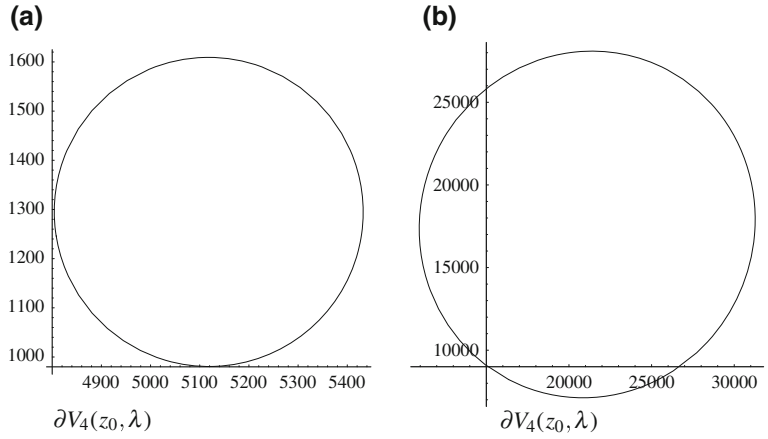


Fig. 1 Region of variability of $\log f(z_0)$ when $f \in \mathcal{F}_\mu(\lambda)$. **a** $z_0 = -0.173777 + 0.0869191i$, $\lambda = -0.196029 + 0.480913i$, $\mu = 32796 + 64560.2i$. **b** $z_0 = -0.713811 - 0.0997298i$, $\lambda = -0.225338 + 0.323073i$, $\mu = 69097.4 + 83886.6i$

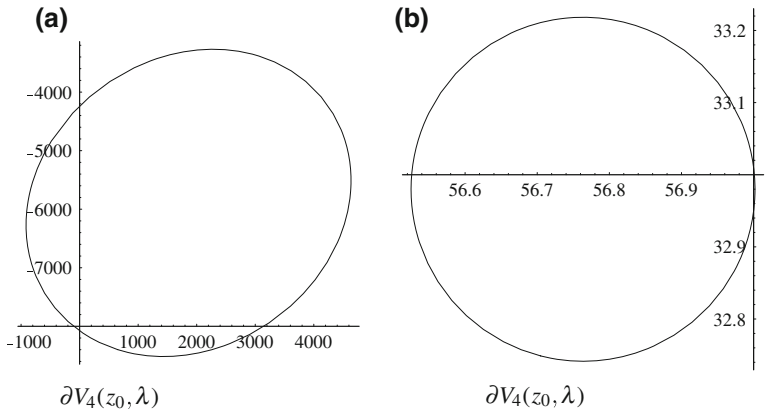


Fig. 2 Region of variability of $\log f(z_0)$ when $f \in \mathcal{F}_\mu(\lambda)$. **a** $z_0 = 0.737135 + 0.496542i$, $\lambda = -0.00646307 - 0.0167039i$, $\mu = 14038.5 + 9544.66i$. **b** $z_0 = -0.00588894 - 0.00496324i$, $\lambda = -0.0472837 + 0.0970889i$, $\mu = 25447.1 - 2011.7i$

$$J_{e^{i\theta},\lambda}(z) = \exp\left(\frac{\mu}{\pi}\int_0^z \frac{\delta(e^{i\theta}\zeta, \lambda) - 1}{(1 - \delta(e^{i\theta}\zeta, \lambda)\zeta)(1 - \zeta)}\, d\zeta\right), \quad z \in \mathbb{D}.$$

It is clearly evident from Proposition 3 that the region bounded by the curve $\partial V_4(z_0, \lambda)$ (see Figs. 1, 2, 3, 4) is compact and convex.

In univalent function theory, there are several subclasses of $\mathcal{S}$ having analytic characterizations involving the positive real part of an appropriate quantity. In [26], the authors considered one of the subclasses $\mathcal{P}_{\gamma,\beta}$ of $\mathcal{S}$. More precisely, let $\mathcal{P}_{\gamma,\beta}$

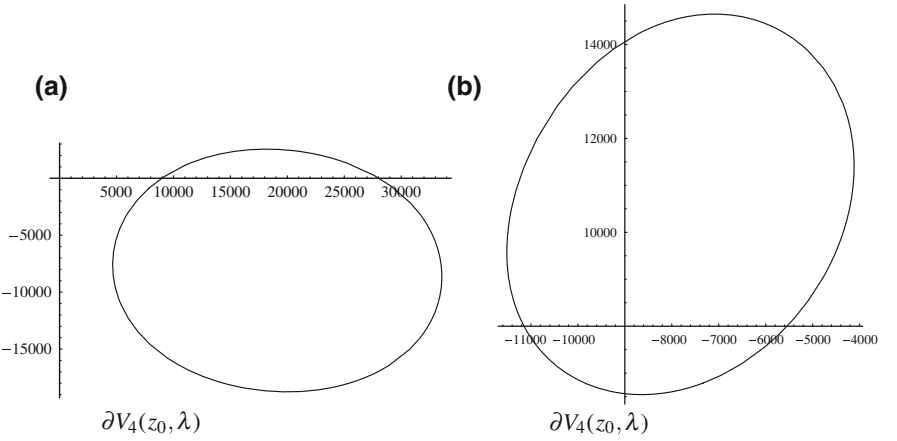


Fig. 3 Region of variability of $\log f(z_0)$ when $f \in \mathscr{F}_\mu(\lambda)$. **a** $z_0 = -0.734426 + 0.61942i$, $\lambda = -0.0564481 - 0.00656122i$, $\mu = 54025 - 5108.28i$. **b** $z_0 = -0.69693 - 0.601351i$, $\lambda = -0.0416728 - 0.683999i$, $\mu = 23944.2 + 50613.5i$

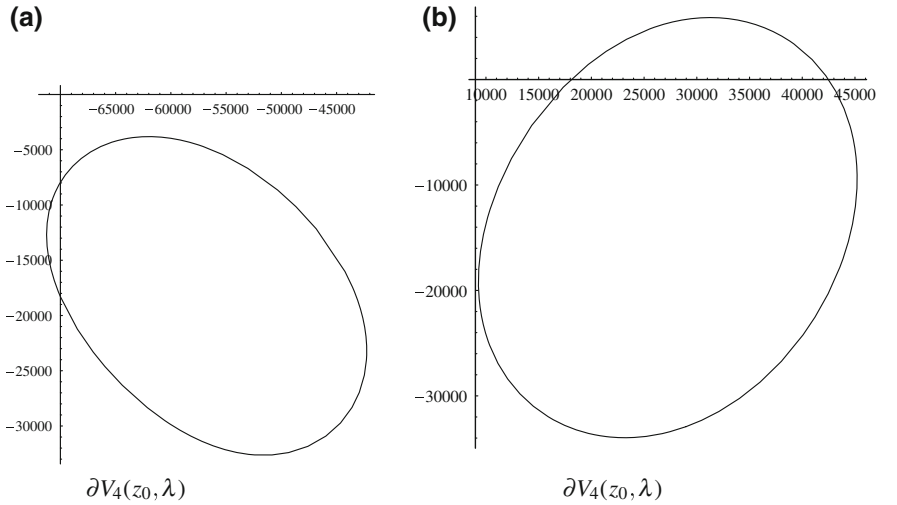


Fig. 4 Region of variability of $\log f(z_0)$ when $f \in \mathscr{F}_\mu(\lambda)$. **a** $z_0 = 0.80351 + 0.549035i$, $\lambda = -0.55886 + 0.0419296i$, $\mu = 83278.8 - 90464.3i$. **b** $z_0 = 0.691568 + 0.644823i$, $\lambda = 0.126172 + 0.137643i$, $\mu = 47178.4 + 83497.8i$

denote the class of functions $P \in \mathscr{H}$ with $P(0) = 1$ and

$$\text{Re}\left(e^{i\gamma} P(z)\right) > \beta \cos\gamma \quad \text{in } \mathbb{D},$$

for some β with $\beta < 1$ and $\gamma \in \mathbb{C}$ with $|\gamma| < \pi/2$. If we choose $P(z) = \frac{zf'(z)}{f(z)}$ and $\beta = 0$ then the class $\mathscr{P}_{\gamma,\beta}$ reduces to

$$\mathscr{S}^{\gamma}(0) = \left\{ f \in \mathscr{A} : \operatorname{Re}\left(e^{i\gamma} \frac{zf'(z)}{f(z)} \right) > 0 \quad \text{in } \mathbb{D} \right\}$$

for some γ with $|\gamma| < \pi/2$. Functions in $\mathscr{S}^{\gamma}(0)$ are known to be univalent in $\mathbb{D}$ and $\mathscr{S}^{0}(0) \equiv \mathscr{S}^{*}$. Functions in $\mathscr{S}^{\gamma}(0)$ are called spirallike functions (see [37]).

Herglotz representation for analytic functions with positive real part in $\mathbb{D}$ shows that if $P \in \mathscr{P}_{\gamma,\beta}$, then there exists a unique positive unit measure μ on $(-\pi, \pi]$ such that

$$P(z) = \int_{-\pi}^{\pi} \frac{1 + [1 - 2\beta e^{-i\gamma} \cos\gamma] z e^{-it}}{1 - z e^{-it}} \, d\mu(t).$$

Then it is a simple exercise to see that for each $P \in \mathscr{P}_{\gamma,\beta}$ there exists an $\omega_P \in \mathscr{B}_0$ such that

$$\omega_P(z) = \frac{e^{i\gamma} P(z) - e^{i\gamma}}{e^{i\gamma} P(z) - (2\beta \cos\gamma - e^{-i\gamma})}, \quad z \in \mathbb{D}, \tag{4}$$

and conversely. A simple computation of (4) gives

$$P'(0) = 2e^{-i\gamma} \omega_P'(0)(1 - \beta) \cos\gamma.$$

Suppose that $P \in \mathscr{P}_{\gamma,\beta}$. Then, because $|\omega_P'(0)| \leq 1$, by the classical Schwarz lemma we let

$$P'(0) = 2\lambda e^{-i\gamma}(1 - \beta) \cos\gamma$$

for some $\lambda \in \overline{\mathbb{D}}$, with $\omega_P'(0) = \lambda$. Now for $\lambda \in \mathbb{D}$ and $a \in \overline{\mathbb{D}}$ we define

$$\widetilde{H}_{a,\lambda}(z) = 1 + 2(1 - \beta) e^{-i\gamma} \cos\gamma \frac{\delta(az, \lambda) z}{1 - \delta(az, \lambda) z}.$$

Obviously $\widetilde{H}_{a,\lambda}(0) = 1$. Since $\delta(az, \lambda)$ lies in the unit disk $\mathbb{D}$ and $\varphi(w) = w/(1 - w)$ maps $|w| < 1$ onto $\operatorname{Re} \varphi(w) > -1/2$, we obtain that

$$\operatorname{Re}\left(e^{i\gamma} \widetilde{H}_{a,\lambda}(z) \right) > \beta \cos\gamma \text{ in } \mathbb{D}.$$

For $\lambda \in \overline{\mathbb{D}}$ and $z_0 \in \mathbb{D}$ fixed, it is natural to introduce (for convenience with the notation $\mathscr{P}(\lambda)$ instead of $\mathscr{P}_{\gamma,\beta}(\lambda)$)

$$\mathscr{P}(\lambda) := \mathscr{P}_{\gamma,\beta}(\lambda) = \left\{ P \in \mathscr{P}_{\gamma,\beta} : \ P'(0) = 2(1-\beta)\mathrm{e}^{-i\gamma}\lambda\cos\gamma \right\}$$

$$V_{\mathscr{P}}(z_0,\lambda) = \left\{ \int_0^{z_0} P(\zeta)\,\mathrm{d}\zeta : \ P \in \mathscr{P}(\lambda) \right\}.$$

The main aim of this paper is to provide explicitly the region of variability of $V_{\mathscr{P}}(z_0,\lambda)$ for $\int_0^{z_0} P(\zeta)\,\mathrm{d}\zeta$ when P ranges over the class $\mathscr{P}(\lambda)$.

Proposition 4 *For $f \in \mathscr{P}_{\gamma,\beta}$ we have*

1. *$V_{\mathscr{P}}(z_0,\lambda)$ is a compact and convex subset of $\mathbb{C}$.*
2. *For $|\lambda| = 1$ or $z_0 = 0$,*

$$V_{\mathscr{P}}(z_0,\lambda) = \left\{ z_0 - 2(1-\beta)\mathrm{e}^{-i\gamma}\cos\gamma\left(z_0 + \frac{1}{\lambda}\log(1-\lambda z_0)\right)\right\}.$$

3. *For $|\lambda| < 1$ and $z_0 \in \mathbb{D}\setminus\{0\}$, $V_{\mathscr{P}}(z_0,\lambda)$ has*

$$z_0 - 2(1-\beta)\mathrm{e}^{-i\gamma}\cos\gamma\left(z_0 + \frac{1}{\lambda}\log(1-\lambda z_0)\right)$$

as an interior point.

Proposition 5 *For $P \in \mathscr{P}(\lambda)$ with $\lambda \in \mathbb{D}$, we have*

$$|P(z) - c(z,\lambda)| \le r(z,\lambda), \quad z \in \mathbb{D},$$

where

$$\begin{aligned} c(z,\lambda) &= \frac{(1+\lambda z(\mathrm{e}^{-i\gamma} - 2\beta\cos\gamma)\mathrm{e}^{-i\gamma})(1-\overline{\lambda}\overline{z})}{\left(1-|z|^2\right)\left(1+|z|^2-2\mathrm{Re}\,(\lambda z)\right)} \\ &\quad + \frac{|z|^2(\overline{z}-\lambda)\left(\overline{\lambda} + z(\mathrm{e}^{-i\gamma} - 2\beta\cos\gamma)\mathrm{e}^{-i\gamma}\right)}{\left(1-|z|^2\right)\left(1+|z|^2-2\mathrm{Re}\,(\lambda z)\right)}, \\ r(z,\lambda) &= \frac{2(1-|\lambda|^2)(1-\beta)|z|^2\cos\gamma}{\left(1-|z|^2\right)\left(1+|z|^2-2\mathrm{Re}\,(\lambda z)\right)}. \end{aligned}$$

For each $z \in \mathbb{D}\setminus\{0\}$, equality holds if and only if $P = \widetilde{H}_{\mathrm{e}^{i\theta},\lambda}$ for some $\theta \in \mathbb{R}$.

The choice of $\lambda = 0$ in Proposition 5 gives the following interesting result.

Corollary 4 *For $P \in \mathscr{P}(0)$ we have*

$$\left| P(z) - \frac{1+(1-2\beta)|z|^4}{1-|z|^4} \right| \le \frac{2(1-\beta)|z|^2}{1-|z|^4}, \quad z \in \mathbb{D}. \tag{5}$$

For each $z \in \mathbb{D}\setminus\{0\}$, equality holds if and only if $P = \widetilde{H}_{\mathrm{e}^{i\theta},0}$ for some $\theta \in \mathbb{R}$.

Theorem 6 *For $\lambda \in \mathbb{D}$ and $z_0 \in \mathbb{D} \setminus \{0\}$, the boundary $\partial V_{\mathscr{P}}(z_0, \lambda)$ is the Jordan curve given by*

$$(-\pi, \pi] \ni \theta \mapsto \int_0^{z_0} \widetilde{H}_{e^{i\theta},\lambda}(\zeta)\, d\zeta = \int_0^{z_0} \frac{1 + [2(1-\beta)(\cos\gamma)e^{-i\gamma} - 1]\delta(e^{i\theta}\zeta, \lambda)\zeta}{1 - \delta(e^{i\theta}\zeta, \lambda)\zeta}\, d\zeta.$$

If

$$\int_0^{z_0} P(\zeta)\, d\zeta = \int_0^{z_0} \widetilde{H}_{e^{i\theta},\lambda}(\zeta)\, d\zeta$$

for some $P \in \mathscr{P}(\lambda)$ and $\theta \in (-\pi, \pi]$, then $P(z) = \widetilde{H}_{e^{i\theta},\lambda}(z)$.

As a special case, we consider $P = f'$ and $\gamma = 0$ in the class $\mathscr{P}_{\gamma,\beta}$. Thus, $\mathscr{P}_{\gamma,\beta}$ reduces to $\mathscr{R}_\beta$, where

$$\mathscr{R}_\beta = \{f \in \mathscr{A} : \operatorname{Re} f'(z) > \beta \quad \text{in } \mathbb{D}\}.$$

Then $\mathscr{R}_\beta \subset \mathscr{S}$ for $0 \le \beta < 1$. For $\lambda \in \overline{\mathbb{D}}$ and $z_0 \in \mathbb{D}$ being fixed, we introduce

$$\begin{aligned} \mathscr{R}(\lambda) &= \big\{f \in \mathscr{R}_\beta : f''(0) = 2(1-\beta)\lambda\big\}, \\ V_{\mathscr{R}}(z_0, \lambda) &= \{f(z_0) : f \in \mathscr{R}(\lambda)\}. \end{aligned}$$

For $P = f'$, a computation shows that the extremal function $\widetilde{H}_{e^{i\theta},\lambda}(z)$ for the class $\mathscr{R}(\lambda)$ takes the form

$$\widetilde{H}_{e^{i\theta},\lambda}(z) = z_0 + 2(1-\beta)\int_0^{z_0} \frac{(e^{i\theta}\zeta + \lambda)\zeta}{1 + \overline{\lambda}e^{i\theta}\zeta - (e^{i\theta}\zeta + \lambda)\zeta}\, d\zeta. \tag{6}$$

One can easily obtain the following result which is the analog of Theorem 6 for the class $\mathscr{R}(\lambda)$.

Corollary 5 *For $\lambda \in \mathbb{D}$ and $z_0 \in \mathbb{D} \setminus \{0\}$, the boundary $\partial V_{\mathscr{R}}(z_0, \lambda)$ is the Jordan curve given by*

$$(-\pi, \pi] \ni \theta \mapsto \widetilde{H}_{e^{i\theta},\lambda}(z_0) = z_0 + 2(1-\beta)\int_0^{z_0} \frac{(e^{i\theta}\zeta + \lambda)\zeta}{1 + \overline{\lambda}e^{i\theta}\zeta - (e^{i\theta}\zeta + \lambda)\zeta}\, d\zeta.$$

If $f(z_0) = \widetilde{H}_{e^{i\theta},\lambda}(z_0)$ for some $f \in \mathscr{R}(\lambda)$ and $\theta \in (-\pi, \pi]$, then $f(z) = \widetilde{H}_{e^{i\theta},\lambda}(z)$.

For $0 \le \beta < 1$ and $\lambda = 0$, set

$$\mathscr{R}(0) = \{f \in \mathscr{A} : f''(0) = 0 \text{ and } \operatorname{Re} f'(z) > \beta \text{ in } \mathbb{D}\} \subset \mathscr{R}_\beta.$$

In particular, the choices $\gamma = 0$ and $P(z) = f'(z)$ in Corollary 4 give the following: if $f \in \mathscr{R}(0) \subset \mathscr{R}_\beta$ for some $0 \leq \beta < 1/2$, then by (5) we obtain

$$|f'(z)| \leq \frac{1 + (1-2\beta)|z|^4 + 2(1-\beta)|z|^2}{1-|z|^4} = \frac{1 + (1-2\beta)|z|^2}{1-|z|^2}, \quad z \in \mathbb{D},$$

so that

$$\|f\|_{\mathscr{B}} := \sup_{z\in\mathbb{D}}(1-|z|^2)|f'(z)| \leq 2(1-\beta).$$

Equality holds for

$$f(z) = \beta z + \frac{(1-\beta)}{2}\log\left(\frac{1+z}{1-z}\right), \quad z \in \mathbb{D}.$$

Let ω be a simply connected domain in the right half plane $\mathbb{H}^+ = \{w \in \mathbb{C} : \operatorname{Re} w > 0\}$ with $1 \in \omega$. Let P_ω be the conformal mappings of the unit disk $\mathbb{D}$ onto Ω with $P_\Omega(0) = 1$ and $P'_\Omega(0) > 0$. Let $\mathscr{CV}_\Omega$ denotes the class of functions $f \in \mathscr{A}$ such that

$$1 + \frac{zf''(z)}{f'(z)} \in \Omega, \quad z \in \mathbb{D}.$$

Clearly $\mathscr{CV}_\Omega$ is a subclass of $\mathscr{C}$. Let

$$H_\Omega(z) = \int_0^z \frac{P_\Omega(\zeta) - 1}{\zeta}\,\mathrm{d}\zeta \quad \text{and} \quad F_\Omega(z) = \int_0^z \mathrm{e}^{H_\Omega(\zeta)}\,\mathrm{d}\zeta \quad \text{for } z \in \mathbb{D}.$$

In [40], the author proved the following result:

Theorem 7 *Let $z_0 \in \mathbb{D} \setminus \{0\}$. If Ω is starlike with respect to 1 and*

$$\operatorname{Re}\left(\frac{zP'_\Omega(z)}{P_\Omega(z)-1} + P_\Omega(z) - 1\right) > 0 \tag{7}$$

holds in $\mathbb{D}$, then the variability region $V_\Omega(z_0)$ is the convex and closed Jordan domain bounded by the simple closed curve $\partial\mathbb{D} \ni \varepsilon \mapsto \varepsilon^{-1}F_\Omega(\varepsilon z_0)$, and

$$V_\Omega(z_0) = \{\varepsilon^{-1}F_\Omega(\varepsilon z_0) : |\varepsilon| \leq 1\}$$

holds. Furthermore $f(z_0) = \varepsilon^{-1}F_\Omega(\varepsilon z_0)$ holds for some $f \in \mathscr{CV}_\Omega$ and $|\varepsilon| = 1$ if and only if $f(z) \equiv \varepsilon^{-1}F_\Omega(\varepsilon z)$.

From Theorem 7 it follows the following result on subordination.

Corollary 6 *If Ω is starlike with respect to 1 and (7) holds in $\mathbb{D}$, then for $f \in \mathscr{CV}_\Omega$ a subordination relation*

$$\frac{f(z)}{z} \prec \frac{F_{\Omega}(z)}{z}$$

holds in $\mathbb{D}$.

The region of variability for certain families of harmonic univalent mappings has been investigated by Ponnusamy et al. [31]. Region of variability for concave univalent functions has been studied in [5] (also see the references therein [24], [27]). This is a rich field of current research interest and this survey is an attempt to introduce new researchers into this field.

References

1. Aharonov, D., Elin, M., Shoikhet, D.: Spiral-like functions with respect to a boundary point. J. Math. Anal. Appl. **280**, 17–29 (2003)
2. Ahuja, O.P., Silverman, H.: A survey on spiral-like and related function classes. Math. Chron. **20**, 39–66 (1991)
3. Alexander, J.W.: Functions which map the interior of the unit circle upon simple regions. Ann. Math. **17**(1), 12–22 (1915)
4. Arango, J.H., Mejia, D., Ruscheweyh, St.: Exponentially convex univalent functions. Complex Var. Elliptic Equ. **33**(1), 33–50 (1997)
5. Bhowmik, B., Ponnusamy, S.: Region of variability for concave univalent functions. Analysis (Munich) **28**(3), 333–344 (2008)
6. Chichra, P.N.: Regular functions $f(z)$ for which $zf'(z)$ is α-spiral. Proc. Am. Math. Soc. **49**, 151–160 (1975)
7. Duren, P.L.: Univalent Functions (Grundlehren der mathematischen Wissenschaften), vol. 259. Springer, New York (1983)
8. Elin, E., Reich, S., Shoikhet, D.: Holomorphically accretive mappings and spiral-shaped functions of proper contractions. Nonlinear Anal. Forum **5**, 149–161 (2000)
9. Elin, M., Reich, S., Shoikhet, D.: Dynamics of inequalities in geometric function theory. J. Inequal. Appl. **6**, 651–664 (2001)
10. Elin, M.: Covering and distortion theorems for spirallike functions with respect to a boundary point. Int. J. Pure Appl. Math. **28**(3), 387–400 (2006)
11. Goodman, A.W.: Univalent Functions, vols. I–II. Mariner Publishing Co., Tampa (1983)
12. Hallenbeck, D.J., Livingston, A.E.: Applications of extreme point theory to classes of multivalent functions. Trans. Am. Math. Soc. **221**, 339–359 (1976)
13. Kaplan, W.: Close-to-convex schlicht functions. Mich. Math. J. **1**, 169–185 (1952)
14. Lecko, A.: On the class of functions starlike with respect to the boundary point. J. Math. Anal. Appl. **261**, 649–664 (2001)
15. Libera, R.J., Ziegler, M.R.: Regular functions $f(z)$ for which $zf'(z)$ is α-spiral. Trans. Am. Math. Soc. **166**, 361–370 (1972)
16. Lyzzaik, A.: On a conjecture of M.S. Robertson. Proc. Am. Math. Soc. **91**, 108–110 (1984)
17. Miller, S.S., Mocanu, P.T.: Differential Subordinations. Theory and Applications, **225**. Marcel Dekker Inc, New York, Basel (2000)
18. Nehari, Z.: The Schwarzian derivative and schlicht functions. Bull. Am. Math. Soc. **55**, 545–551 (1949)
19. Pfaltzgraff, J.A.: Univalence of the integral of $f'(z)^{\lambda}$. Bull. Lond. Math. Soc. **7**, 254–256 (1975)
20. Ponnusamy, S., Vasudevarao, A.: Region of variability of two subclasses of univalent functions. J. Math. Anal. Appl. **332**(2), 1323–1334 (2007)

21. Ponnusamy, S., Vasudevarao, A., Yanagihara, H.: Region of variability of univalent functions $f(z)$ for which $zf'(z)$ is spirallike. Houston J. Math. **34**(4), 1037–1048 (2008a)
22. Ponnusamy, S., Vasudevarao, A., Yanagihara, H.: Region of variability for close-to-convex functions. Complex Var. Elliptic Equ. **53**(8), 709–716 (2008b)
23. Ponnusamy, S., Vasudevarao, A., Vuorinen, M.: Region of variability for spirallike functions with respect to a boundary point. Colloq. Math. **116**(1), 31–46 (2009a)
24. Ponnusamy, S., Vasudevarao, A., Vuorinen, M.: Region of variability for certain classes of univalent functions satisfying differential inequalities. Complex Var. Elliptic Equ. **54**(10), 899–922 (2009b)
25. Ponnusamy, S., Vasudevarao, A., Yanagihara, H.: Region of variability for close-to-convex functions-II. Appl. Math. Comp. **215**(3), 901–915 (2009c)
26. Ponnusamy, S., Vasudevarao, A.: Region of variability for functions with positive real part. Ann. Polon. Math. **99**(3), 225–245 (2010)
27. Ponnusamy, S., Vasudevarao, A., Vuorinen, M.: Region of variability for exponentially convex functions. Complex Anal. Oper. Theory **5**(3), 955–966 (2011)
28. Pommerenke, Ch.: Boundary behaviour of conformal maps. Springer, Berlin (1992)
29. Ponnusamy, S., Rajasekaran, S.: New sufficient conditions for starlike and univalent functions. Soochow J. Math. **21**, 193–201 (1995)
30. Ponnusamy, S., Singh, V.: Univalence of certain integral transforms. Glas. Mat. Ser. III **31**(2), 253–262 (1996)
31. Ponnusamy, S., Yamamoto, H., Yanagihara, H.: Variability regions for certain families of harmonic univalent mappings. Complex Var. Elliptic Equ. **58**(1), 23–34 (2013)
32. Robertson, M.S.: Univalent functions $f(z)$ for which $zf'(z)$ is spirallike. Mich. Math. J. **16**, 97–101 (1969)
33. Robertson, M.S.: Univalent functions starlike with respect to a boundary point. J. Math. Anal. Appl. **81**, 327–345 (1981)
34. Ruscheweyh, St.: A subordination theorem for Φ-like functions. J. Lond. Math. Soc. **13**, 275–280 (1976)
35. Silverman, H., Silvia, E.M.: Subclasses of univalent functions starlike with respect to a boundary point. Houston J. Math. **16**(2), 289–299 (1990)
36. Singh, V., Chichra, P.N.: Univalent functions $f(z)$ for which $zf'(z)$ is α-spirallike. Indian J. Pure Appl. Math. **8**, 253–259 (1977)
37. Špaček, L.: Contribution à la théorie des fonctions univalentes (in Czech), Časop Pěst. Mat. Fys. **62**, 12–19 (1933)
38. Yanagihara, H.: Regions of variability for functions of bounded derivatives. Kodai Math. J. **28**, 452–462 (2005)
39. Yanagihara, H.: Regions of variability for convex function. Math. Nachr. **279**, 1723–1730 (2006)
40. Yanagihara, H.: Variability regions for families of convex function. Comput. Methods Funct. Theory **10**(1), 291–302 (2010)

Chapter 11
Ideal Cone: A New Method to Generate Complete Pareto Set of Multi-criteria Optimization Problems

Debdas Ghosh and Debjani Chakraborty

Abstract In this paper, a new classical method, entitled ideal cone (IC), is presented to generate complete Pareto set of multi-criteria optimization problems (MOP). Systematically changing a parameter, which is independent of decision maker's (DM) preferences, the method seeks Pareto optimal solutions sequentially. Parameter of the proposed classical method is independent of objective functions of the problem. Formulated method is a non-gradient direction-based technique. Directions of the method essentially lie on k-dimensional unit sphere for k-criteria problems. Though proposed method is a direction-based method, it bears necessary and sufficient condition for globally weak Pareto optimality. It is shown that a simple modification of the presented method can attain D-Pareto optimal points of the problem, where D is any pointed convex cone. Thus, formulated technique not only can generate Pareto set, but also obtain general D-Pareto set. A brief comparison of the proposed method with the existing similar classical methods is also made. Developed method is supported by several numerical and pictorial illustrations.

Keywords Multiple objective programming · Pareto set · Direction-based Pareto set generation algorithm · Ideal cone method

1 Introduction

Most of the optimization problems stemming from engineering design or other complex decision situations are characterized by the existence of multiple criteria with some complicated additional constraints. In practice, the decision maker (DM) has

D. Ghosh (✉) · D. Chakraborty
Department of Mathematics, Indian Institute of Technology Kharagpur,
Kharagpur 721302, West Bengal, India
e-mail: debdas.email@gmail.com

D. Chakraborty
e-mail: debjani@maths.iitkgp.ernet.in

R. N. Mohapatra et al. (eds.), *Mathematics and Computing 2013*,
Springer Proceedings in Mathematics & Statistics 91,
DOI: 10.1007/978-81-322-1952-1_11, © Springer India 2014

to reconcile those dissimilar and conflicting criteria to obtain optimum design or solution. Mathematically, to get the optimum solution, DM has to optimize multiple criteria simultaneously with respect to the constraints. This optimization problem is called multi-criteria optimization problem (MOP). In general, a unique solution of such problem may not exist since otherwise there is no conflict between the objectives. Thus, usually there are many optimal solutions of a MOP. This optimal solution concept leads to Pareto optimality. A Pareto optimum solution is the feasible solution where any improvement in one criterion can only take place through worsening of at least one another criterion. The concept of Pareto optimality is of primordial importance to recognize the conflicting nature of the criteria, and hence, to capture the trade-off between the criteria. All the Pareto points are equally acceptable as solution of the MOP. Usually, by using some additional requirements, DM selects one point from Pareto sets as final solution. These additional requirements may be subjective, and practically depends on DM's preference.

Solving MOP effectively means to generate complete Pareto set. There exist various methods [(classical and evolutionary (see [5] for comparison)] in the existing literature to generate Pareto set. The approaches of all the classical methods can be mainly categorized into two parts: aggregation of objective function (AOF) methods and Pareto surface generation methods [2, 12, 14–16, 18, 20].

In the first category, the approaches involve forming an AOF with respect to the constraints. Here, the objective is to find optimal solution of the scalar optimization problem with a suitable AOF as the objective function, and subject to the constraints in the considered MOP.

In the second category, methods are purposed to find complete Pareto set or set of non-dominated solutions. However, each of the existing methods have one or more of the three deficiencies (see [24]): (1) the method is not necessary and sufficient for Pareto optimality, (2) the method cannot generate complete Pareto surface, or (3) the method requires significant knowledge about the physical properties of the criteria. Thus, a need arises to find a procedure having ability to generate the entire Pareto surface, and which does not carry any of the above-mentioned deficiencies. In this paper, an attempt for the same is made. The proposed method in this paper belongs to the second category.

In this paper, a new classical method for MOP—hereby named ideal cone (IC)—is proposed. The main idea for the methodology is confined under the simple and well-known fact that "in the criterion space, if the intersecting set of the feasible region and the translated nonpositive orthant having vertex at a feasible point contains only the vertex of the translated cone, then that feasible point must be Pareto optimal solution and vice versa". To implement this idea for generating complete Pareto set of a MOP, in general, the whole feasible region will be translated on the nonnegative orthant first and then the cone of nonpositive orthant will be moved along all possible directions in the nonnegative orthant to test whether the above-mentioned fact holds true. The direction is the only parameter of the problem. The method seeks Pareto optimal solutions one after another by systematically changing this parameter, which is, obviously, independent of DM's preferences. Detail of the method is demonstrated in the Sect. 3. The paper is organized as follows.

In the next section, preliminaries of MOP and the notations, which are used throughout this paper, are given. In the Sect. 3, the proposed method and corresponding useful results are demonstrated. Numerical explorations and efficiencies the proposed method is discussed in the Sect. 4. A brief comparison of the proposed method with the existing methods is given in the Sect. 5. In the last section, which is Sect. 6, contribution and future scope of this paper on MOP is drawn.

2 Preliminaries and Notations

In mathematical notions, MOPs are defined in the following way

$$\min_{x \in \mathscr{X}} f(x) = \big(f_1(x), f_2(x), \ldots, f_k(x)\big)^{\mathrm{T}}, \quad k \geq 2, \tag{1}$$

where $\mathscr{X} = \{x \in \mathbb{R}^n\colon\ g(x) \leq 0,\ h(x) = 0,\ a \leq x \leq b\}$ is the feasible set; $g\colon \mathbb{R}^n \to \mathbb{R}^r$ and $h\colon \mathbb{R}^n \to \mathbb{R}^s$ are vector valued functions; the constant vectors $a \in (\mathbb{R} \cup \{-\infty\})^n$ and $b \in (\mathbb{R} \cup \{\infty\})^n$ are, respectively, lower and upper bound of the decision vector $x = (x_1, x_2, \ldots, x_n)^{\mathrm{T}}$.

We denote the image of the feasible set $\mathscr{X}$ under the vector mapping f by $\mathscr{Y} := f(\mathscr{X})$. Therefore, $\mathscr{Y}$ is the feasible set in the criterion space. If for each individual $i \in \{1, 2, \ldots, k\}$, x_i^* is the point of global minima of the objective function f_i, the point $y_i^* := f(x_i^*)$, $i = 1, 2, \ldots, k$ in the criterion space is said to be an *anchor point*. Again, the point $y^I = (y_1^I, y_2^I, \ldots, y_k^I)^{\mathrm{T}}$ given by $y_i^I := \min\limits_{x \in \mathscr{X}} f_i(x) = \min\limits_{y \in \mathscr{Y}} y_i$ is called as *ideal point* or *utopia point*. As in general y^I is not attainable, notion of Pareto optimality is being introduced as follows. The definitions of weak Pareto optimality and proper Pareto optimality are also given subsequently.

Definition of Pareto optimality depends on a dominance structure or componentwise order in the space $\mathbb{R}^k$. To represent dominance structure on $\mathbb{R}^k$, the following subsets are usually used. The nonnegative orthant of $\mathbb{R}^k$ is represented by $\mathbb{R}^k_{\geqq} := \{y \in \mathbb{R}^k\colon y \geqq 0\}$; $y = (y_1, y_2, \ldots, y_k)^{\mathrm{T}}$. The notation $y \geqq 0$ implies $y_i \geq 0$ for each $i = 1, 2, \ldots, k$. The set $\mathbb{R}^k_{\geq}$ is defined by $\{y \in \mathbb{R}^k\colon y \geq 0\}$ where $y \geq 0$ means $y \geqq 0$ but $y \neq 0$. The notation $\mathbb{R}^k_{>} := \{y \in \mathbb{R}^k\colon y > 0\}$ indicates the positive orthant of $\mathbb{R}^k$. Here, $y > 0$ stands for $y_i > 0$ for each $i = 1, 2, \ldots, k$. The relations '$\leqq$', '$\leq$' and '$<$' are similarly defined. For $\hat{x},\ \bar{x} \in \mathscr{X}$, the vector $f(\hat{x})$ is said to dominate another vector $f(\bar{x})$ if $f(\hat{x}) \leq f(\bar{x})$.

Definition 1 (*Pareto optimality* [7]). A feasible solution $\hat{x} \in \mathscr{X}$ is called efficient or Pareto optimal, if there is no other $x \in \mathscr{X}$ such that $f(x) \leq f(\hat{x})$. If $\hat{x}$ is efficient, $f(\hat{x})$ is called non-dominated. The set of all efficient points is denoted by $\mathscr{X}_{\mathrm{E}}$. The collection of all non-dominated points is denoted by $\mathscr{Y}_{\mathrm{N}}$.

Definition 2 (*Weak Pareto optimality* [7]). A feasible solution $\hat{x} \in \mathscr{X}$ is called weakly Pareto optimal if there is no $x \in \mathscr{X}$ such that $f(x) < f(\hat{x})$. The point

$\hat{y} = f(\hat{x})$ is then called weakly non-dominated and $\hat{x}$ is called weakly Pareto optimal point. The set of all weakly efficient points is denoted by $\mathscr{X}_{\mathrm{wE}}$. The collection of all non-dominated points is denoted by $\mathscr{Y}_{\mathrm{wN}}$.

Definition 3 (*Proper Pareto optimality* [7]). A feasible solution $\hat{x}$ which is a Pareto optimal point is said to be properly Pareto optimal if there exists a positive real number M such that for any $i \in \{1, 2, \ldots, k\}$ and $x \in \mathscr{X}$ satisfying $f_i(x) < f_i(\hat{x})$ there exists an index j such that $f_j(\hat{x}) < f_j(x)$ such that $\frac{f_i(\hat{x}) - f_i(x)}{f_j(x) - f_j(\hat{x})} \leq M$. The point $\hat{y} = f(\hat{x})$ is then called properly non-dominated. The set of all proper Pareto optimal solutions is denoted by $\mathscr{X}_{\mathrm{pE}}$.

It can be easily perceived that a feasible point $\hat{x} \in \mathscr{X}$ belongs to $\mathscr{X}_{\mathrm{E}}$ if and only if $(f(\hat{x}) - \mathbb{R}^k_{\underline{\underline{\geq}}}) \bigcap f(\mathscr{X}) = \{f(\hat{x})\}$. Similarly, a feasible point $\hat{x} \in \mathscr{X}$ belongs to $\mathscr{X}_{\mathrm{wE}}$ if and only if $(f(\hat{x}) - \mathbb{R}^k_{>}) \bigcap f(\mathscr{X}) = \emptyset$.

In a more general sense, if the objective space $\mathbb{R}^k$ is partially ordered by a pointed convex cone D say, a decision point $\hat{x} \in \mathscr{X}$ is efficient with respect to D when $(f(\hat{x}) - D) \bigcap f(\mathscr{X}) = \{f(\hat{x})\}$. Analogously, a point $\hat{x} \in \mathscr{X}$ is weakly efficient with respect to D when $(f(\hat{x}) - \mathrm{int}(D)) \bigcap f(\mathscr{X}) = \emptyset$, where $\mathrm{int}(D)$ represents the interior of D. If D is taken as $\mathbb{R}^k_{\varepsilon} := \{y \in \mathbb{R}^k : \mathrm{dist}(y, \mathbb{R}^k_{>}) \leq \varepsilon \|y\|\}$, then a (weakly) Pareto optimal point with respect to D is said to be (weakly) ε-Pareto optimal point. Since, at any ε-Pareto optimal point, trade-off between any two criteria are bounded by ε and $1/\varepsilon$, any ε-Pareto optimal point is properly Pareto optimal point.

We observe that the nonpositive orthant $-\mathbb{R}^k_{\underline{\underline{\geq}}}$ and its interior are two convex cones, which can be used to examine whether a feasible point is Pareto optimal or weakly Pareto optimal or none of them. Thus, the closed-convex cone $-\mathbb{R}^k_{\underline{\underline{\geq}}}$ may be imagined as an IC to test Pareto optimality. Application of the same is implemented while forming constraint inequalities of the proposed method—hence, the method is named as *IC method*. In the next section, detailed construction of the IC method is studied.

3 Ideal Cone Method and Results

We describe IC method and its mathematical perspective in the following three subsections.

3.1 Mathematical Description of the Method

Ideal cone is a classical method to generate Pareto set ($\mathscr{X}_{\mathrm{E}}$), and hence, complete non-dominated set ($\mathscr{Y}_{\mathrm{N}}$) of a MOP. How the method generates $\mathscr{X}_{\mathrm{E}}$ is discussed in the following. Central idea behind IC method lies in the following two facts—first,

a feasible point $x^* \in \mathscr{X}$ is a Pareto optimal solution of the MOP (1) if

$$\left(f(x^*) - \mathbb{R}^k_{\leqq}\right) \bigcap f(\mathscr{X}) = \{f(x^*)\},$$

second, the non-dominated set $\mathscr{Y}_N$ is a subset of the boundary of the criterion feasible region, i.e.,

$$\mathscr{Y}_N \subset bd(\mathscr{Y}).$$

From the geometrical point of view, first fact means that—in the criterion space, if the criterion feasible region and the translated nonpositive orthant whose vertex is being shifted from origin to the point $f(x^*)$ have intersection the single point $f(x^*)$ only, then the feasible point x^* is a Pereto optimal solution of the considered MOP. If x^* is Pareto optimal solution, the point $y^* = f(x^*)$ must be a non-dominated solution. So, to get a non-dominated solution, we may translate the cone of non-positive orthant of the criterion space along a particular direction $\hat{\beta} \in \mathbb{R}^k_{\leqq}$ till this cone does not touch the criteria feasible region. Translation of the cone $-\mathbb{R}^k_{\leqq}$ along a particular direction $\hat{\beta} \in \mathbb{R}^k_{\leqq}$ means that the vertex of the cone is retained on the line $z\hat{\beta}$, $z \in \mathbb{R}$. Now, if the cone $-\mathbb{R}^k_{\leqq}$ is being translated along $\hat{\beta} \in \mathbb{R}^k_{\leqq}$, then it can touch the boundary of the criterion feasible region $\mathscr{Y}$ ($bd(\mathscr{Y})$) in two possible ways: either the vertex of the cone touches first or one (or more) boundary plane(s) of the cone touches first. Once the first case happens, the point where the cone touches the criterion feasible region is certainly be a globally non-dominated point. If the latter case happens, it is possible in two different ways: touching portion is either a single point or a set of points. In the first case, though the touching point is a Pareto optimal point, but not a proper Pareto optimal solution. In the second case, it can be easily perceived that all the points except the extreme points of the touching portion are weakly Pareto optimal solutions.

Let us illustrate how the above said touching portion of $bd(\mathscr{Y})$ and the cone $z\hat{\beta} - \mathbb{R}^k_{\leqq}$, for a particular direction $\hat{\beta} \in \mathbb{R}^k_{\leqq}$, can be found. To demonstrate, let us begin with a graphical perspective of the IC method for a simple bi-objective optimization problem. Figure 1 portrays the criterion feasible region $\mathscr{Y} = f(\mathscr{X})$ for a generic bi-objective problem and the cone $z\hat{\beta} - \mathbb{R}^k_{\leqq}$ for three different values of z, namely z_1, z_2 and z_3 corresponding to the points A, B, and C respectively. Here $OA = z_1$, $OB = z_2$ and $OC = z_3$. Let us now consider the set $\left\{y: z\hat{\beta} \geqq f(x),\ y = f(x),\ x \in \mathscr{X}\right\}$, $z \in \mathbb{R}$. For each specific value of $z \in \mathbb{R}$, this set is either an empty set or a subset of $\mathscr{Y}$. For example, for $z = z_1$ the set is empty; for $z = z_2$ the set is the singleton set $\{B\}$, $B \equiv z_2\hat{\beta}$; for $z = z_3$ the set is the shaded region $CDBEC$. We note that for a fixed $z \in \mathbb{R}$, the set $\left\{y: z\hat{\beta} \geqq f(x),\ y = f(x),\ x \in \mathscr{X}\right\}$ represents the intersecting region of $(z\hat{\beta} - \mathbb{R}^k_{\leqq})$ and $f(\mathscr{X})$. Now, for generic $z \in \mathbb{R}$ let us try to minimize the intersecting region between $(z\hat{\beta} - \mathbb{R}^k_{\leqq})$ and $f(\mathscr{X})$ by translating the cone $(z\hat{\beta} - \mathbb{R}^k_{\leqq})$

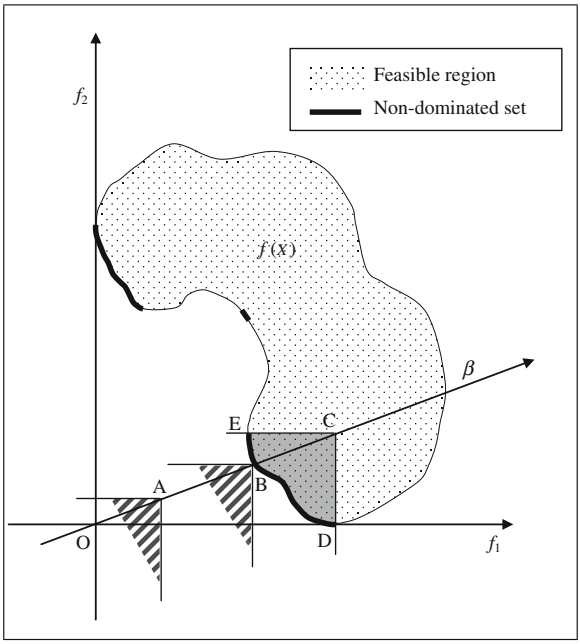


Fig. 1 Explanation of IC method

along $\hat{\beta}$ such a way that the cone does not leave $f(\mathscr{X})$. In the optimum situation if the intersection $(z\hat{\beta} - \mathbb{R}^k_{\geqq}) \bigcap f(\mathscr{X})$ contains only one point, then that singleton point indeed be a non-dominated point. We note that minimizing the intersecting region $(z\hat{\beta} - \mathbb{R}^k_{\geqq}) \bigcap f(\mathscr{X})$ eventually involve minimizing the value of z with the constraints $z\hat{\beta} \geqq f(x), x \in \mathscr{X}$. It is worthy to note that the earlier discussions do not depend on the number of criteria. Therefore, to get a non-dominated solution of the MOP (1) we can solve the following minimization problem:

$$\text{IC}(\beta) \quad \begin{cases} \min & z \\ \text{subject to} & z\hat{\beta} \geqq f(x), \\ & x \in \mathscr{X}. \end{cases} \tag{2}$$

Solving this problem for various values of $\hat{\beta}$ in $\mathbb{R}^k_{\geqq}$, whole non-dominated set of the considered MOP can be generated. It is to notice that if $\hat{\beta}$ is replaced by any vector, β say, parallel to $\hat{\beta}$, then solution of the subproblem will remain the same with a little modification to the value of the objective function z. That is why subproblem (2) has been referred as IC(β). Solution of IC(β) may be represented by $(x^*, z^*, \hat{u})$, where x^* is the solution of (2) with objective value z^* and $\hat{u} = \frac{f(x^*)}{\|f(x^*)\|}$. Since the utopia

point is assumed to be infeasible solution, the unit vector $\hat{u}$ is always well-defined. Here, we note that the unit vector $\hat{u}$ may not always be identical to $\hat{\beta}$.

3.2 Theoretical Results on IC Method

We note that IC method is intended to generate non-dominated solutions. Thus, one natural question may arise: whether solution of IC(β) for any $\beta \in \mathbb{R}^k_{\geqq}$ is Pareto optimal or not, and conversely whether each non-dominated solution of the considered MOP is attainable by the IC subproblem or not? We have made attempt for the same in the following two theorems.

Theorem 1 *Solution of IC(β) for any $\beta \in \mathbb{R}^k_{\geqq}$ is weakly Pareto optimal.*

Proof Suppose $(x^*, z^*, \hat{u})$ is solution of IC(β). Let $x^* \notin \mathscr{X}_{\mathrm{wE}}$. So, there must exists some $\bar{x} \in \mathscr{X}$ such that $f(\bar{x}) < f(x^*)$. Now the constraints of the IC(β) directly show that optimal value of the objective function z in IC(β) must be less than z^*. A contradiction arises. Hence, x^* must be weakly Pareto optimal.

Theorem 2 *Let $x^* \in \mathscr{X}_{\mathrm{E}}$. Then there exists some $\beta \in \mathbb{R}^k_{\geqq}$ such that* IC(β) *has optimum solution at x^*.*

Proof Let us choose $\beta = f(x^*)$. As $x^* \in \mathscr{X}_{\mathrm{E}}$, $(\beta - \mathbb{R}^k_{\geqq}) \bigcap f(\mathscr{X}) = \{\beta\}$. Therefore, optimal solution of IC(β) must be x^*. Hence the result follows.

Due to Theorems 1 and 2, IC method bears necessary and sufficient condition for weakly Pereto optimality. Once solving all possible IC($\hat{\beta}$) is accomplished for all $\hat{\beta}$, from the solution set itself we can easily detect the points, which lie on $\mathscr{X}_{\mathrm{wE}} \setminus \mathscr{X}_{\mathrm{E}}$ as follows. Suppose $(x^*, z^*, \hat{u})$ be a solution of a particular IC subproblem and S be the set of all such solution points. If there exists $(x^*, z_1^*, \hat{u}_1) \in S$, then solution of IC($\hat{u}$) is all the points lie on the line segment joining x^* and $f^{-1}(z^*\hat{u})$. Therefore, solution of IC($\hat{u}$) is not unique and x^* must be weakly Pareto optimal point. Thus, though IC method may generate some points on $\mathscr{X}_{\mathrm{wE}} \setminus \mathscr{X}_{\mathrm{E}}$, they can be easily detected. So, IC method bears necessary and sufficient condition for weak Pareto optimality and if the solution of the IC(β) is unique, that solution must be globally Pareto optimal.

Here, from the easy geometrical visualization of the IC method we obtain the following two results. Some more theoretical aspects of IC method can be obtained in our further research on mathematical perspective of IC method.

Lemma 1 *Let us suppose β_1, β_2 are two non-parallel vectors in $\mathbb{R}^k_{\geq}$. If* IC(β_1) *and* IC(β_2) *have solutions $(x^*, z_1^*, \hat{\beta}_2)$ and $(x^*, z_2^*, \hat{\beta}_2)$ respectively, then*

$$\left(z_2^*\hat{\beta}_2 - \mathbb{R}^k_{\leqq}\right) \subset \left(z_1^*\hat{\beta}_1 - \mathbb{R}^k_{\geq}\right),$$
$$\left(z_2^*\hat{\beta}_2 - \mathbb{R}^k_{\leqq}\right) \not\subset \left(z_1^*\hat{\beta}_1 - \mathbb{R}^k_{>}\right) \text{ and}$$
$$z_2^* < z_1^*.$$

Lemma 2 *Let us suppose $\beta_1, \beta_2 \in \mathbb{R}^k_{\geq}$ are two non-parallel vectors and bd($\mathscr{Y}$) is smooth. If* IC(β_1) *and* IC(β_2) *have solutions* $(x^*, z_1^*, \hat{\beta}_2)$ *and* $(x^*, z_2^*, \hat{\beta}_2)$ *respectively, then for each* $t \in (0, 1)$, IC(β_t) *must have solution* $(x^*, z_t^*, \hat{\beta}_2)$ *for some* $z_t^* \in \mathbb{R}_>$, *where* $\beta_t = z_1^*\hat{\beta}_1 + t(z_2^*\hat{\beta}_2 - z_1^*\hat{\beta}_1)$. *Furthermore, if* $0 < t_1 < t_2 < 1$, *then* $z_{t_2}^* < z_{t_1}^*$.

Theorem 3 *Let* $(x^*, z^*, \hat{\beta}_1)$ *is solution of* IC(β_1). *If there does not exist any other vector* $\beta_2 \in \mathbb{R}^k_{\geq}$ *which is not parallel to* β_1 *such that IC*(β_2) *has solution* $(x^*, z^*, \hat{\beta}_1)$, *then either* x^* *is an anchor point or* $x^* \in \mathscr{X}_{\text{pE}}$.

Proof If possible let $x^* \notin \mathscr{X}_{\text{pE}}$.
Then, there must exist some $\delta > 0$ and one $j \in \{1, 2, \ldots, k\}$ such that

$$B(f(x^*), \delta) \bigcap f(\mathscr{X}) = \{f_j(x) \colon f_j(x) \geq f_j(x^*)\} \bigcap B(f(x^*), \delta) \bigcap f(\mathscr{X}). \tag{3}$$

Since otherwise there exists $\varepsilon > 0$ such that x^* is an ε-Pareto optimal point.
This implies trade-offs between all the objectives are bounded by ε and $1/\varepsilon$, and hence, $x^* \in \mathscr{X}_{\text{pE}}$. A contradiction arises. Therefore, (3) holds true, and we note that (3) clearly implies the theorem.

Proposition 1 *Let us suppose* $\hat{\beta}_1, \hat{\beta}_2 \in \mathbb{R}^k_{\leqq} \bigcap \mathbb{S}^{k-1}$ *are two non-parallel vectors and bd($\mathscr{Y}$) is smooth. If* CM($\hat{\beta}_1$) *and* CM($\hat{\beta}_2$) *have solutions* $(x^*, z_1^*, \hat{\beta}_3)$ *and* $(x^*, z_2^*, \hat{\beta}_3)$ *respectively, then there must exist two criteria whose trade-off is unbounded at* x^*.

Let us now try to solve the IC subproblems efficiently under which generation of the complete efficient set of the considered MOP is confined.

3.3 Algorithmic Implementation of the Ideal Cone Method

Earlier discussion and results show that IC(β) subproblems are to be solved for each unit vector $\hat{\beta} \in \mathbb{R}^k_{\leqq}$ to obtain complete Pareto and weakly Pareto sets of a MOP. In practice algorithmic implementation of IC method, a uniform discretization of the set $\mathbb{S}^{k-1}_{\geq}$ $\mathbb{S}^{k-1} \bigcap \mathbb{R}^k_{\geq}$ would be considered to get required $\hat{\beta}$s, where $\mathbb{S}^{k-1}$ is the k-dimensional unit sphere. Solving IC($\hat{\beta}$) for more and more $\hat{\beta}$s, number of obtained Pareto points will be increased and gradually the entire Pareto set will be captured.

Let us note that any $\hat{\beta} \in \mathbb{S}^{k-1}$ can be expressed by

$$\left(\cos\phi_1, \cos\phi_2 \sin\phi_1, \cos\phi_3 \sin\phi_2 \sin\phi_1, \ldots, \cos\phi_{k-1} \prod_{i=1}^{k-2} \sin\phi_i, \prod_{i=1}^{k-1} \sin\phi_i\right),$$

for $\phi_i \in [0, \frac{\pi}{2}]$, $i = 1, 2, \ldots, (k-1)$. This is well know spherical discretization technique. However, if we discretize each ϕ_i to equal number of subintervals, then set of discretized points will be much congested near the point $(1, 0, 0, \ldots, 0)$. Thus, to get a uniform discretized points on $\mathbb{S}^{k-1}$, let us attempt to divide ϕ_1 by m number of points and ϕ_i by round$(m \prod_{l=1}^{i} \sin\phi_i)$ number of points, for $i = 2, 3, \ldots, k-1$. Here *round* is the rounding function to the nearest integer.

Following Algorithm 1 provides a sequential procedure to obtain complete Pareto set of a tri-criteria problem. In tri-criteria problem, we need to run 2 for loops for each ϕ_i, $i = 1, 2$. For k-criteria problem, we only have to run $k-1$ for loops for each ϕ_i, $i = 1, 2, \ldots, k-1$.

Algorithm 1 Algorithm to generate complete non-dominated set

Require: Given MOP:

$$\begin{cases} \min\ f(x) \\ \text{subject to } x \in \mathscr{X}. \end{cases}$$

Final output $\mathscr{Y}_{\mathrm{N}}$ of the algorithm is the complete non-dominated set of the problem.

1: Initialize ϕ_1 and ϕ_2 to 0.
2: Initialize $\mathscr{Y}_{\mathrm{N}} \leftarrow \emptyset$.
3: Give m (total number of grid points for ϕ_1).
4: **for** $\phi_1 = 0$ to $\frac{\pi}{2}$ with step length $\frac{\pi}{2m}$ **do**
5: Find $m_2 = \text{round}(m \sin\phi_1)$
6: **for** $\phi_2 = 0$ to $\frac{\pi}{2}$ with step length $\frac{\pi}{2m_2}$ **do**
7: Find $\hat{\beta} = (\cos\phi_1, \cos\phi_2 \sin\phi_1, \sin\phi_2 \sin\phi_1)$
8: Find x^*_β where $(x^*_\beta, z^*_\beta, \hat{u})$ is the solution of the following problem for $\hat{\beta}$:
9:

$$\mathrm{IC}(\hat{\beta}) \begin{cases} \min\ z \\ \text{subject to } z\hat{\beta} \geqq f(x), \\ x \in \mathscr{X}. \end{cases}$$

10: Set $\mathscr{Y}_{\mathrm{N}} \leftarrow \mathscr{Y}_{\mathrm{N}} \bigcup f(x^*_\beta)$.
11: **end for**
12: **end for**

The above, discretization of $\mathbb{S}^{k-1}_{\geq}$ is done aiming to get evenly spaced $\hat{\beta}$s over $\mathbb{S}^{k-1}_{\geq}$. Since with the generated $\hat{\beta}$s thereby, if the solutions of IC$(\hat{\beta})$ subproblems are accumulated, it is sure that IC method has not missed any portion of the Pareto surface to seek Pareto points.

4 Examples and Discussions

In this section, a couple of test problems are considered to compare the proposed IC method with the existing classical methods. Advantages and efficiencies of IC method are also discussed. The considered problems are either studied extensively or used as benchmark. In the literature Pareto set generating methods, or existing classical methods have been failed to obtain their entire Pareto set efficiently. In comparison, direct search domain (DSD) [9] method has not been considered since DSD method—originally being a modification of the physical programming (PP) [19] method—bears all the deficiencies of the PP method and highly depended on the method's shrinking angle [9]. Nonetheless, DSD subproblems can attain locally Pareto optimal points as their solutions. It is to be noted that only varying the parameter β over the first orthant $\mathbb{R}^k_{\geq}$, IC method can efficiently obtain well-diversified Pareto optimal points. In all the following examples, uniform spherical discretization of $\mathbb{S}^{k-1}_{\geq}$ is taken to get $\hat{\beta}$s of IC method.

Example 1 This test problem is a simple bi-objective optimization problem studied in [19] to compare performance of PP, WS, and CP methods. The problem is stated as follows:

$$\min \begin{pmatrix} f_1(\theta) \\ f_2(\theta) \end{pmatrix}$$
$$\text{subject to } 0.5326 \leq \theta \leq 1.2532$$

where $f_1(\theta) = \sin\theta$, $f_2(\theta) = 1 - \sin^7\theta$.

Messac and Mattson [19] mentioned that here performance of PP method is superior than WS and CP methods.

In Figs. 2, 3, 4, 5, performance of PP method and IC method for 50 and 200 evaluations are explored. For PP method, used so-called pseudo-preferences P_i $(i = 1, 2)$ are $P_i = (f_{i1}, f_{i2}, f_{i3}, f_{i4}, f_{i5})^{\mathrm{T}} = f_i^{(0)}(1, 1, 1, 1, 1)^{\mathrm{T}} + \delta_i(0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1)^{\mathrm{T}}$ where $\delta_i = (f_{i,\max} - f_{i,\min})/n_d$ and $f_i^{(0)}$ is a free parameter. The parameter n_d defines PP method's search box size δ_i [19, 24]. Here $f_i^{(0)}$ is chosen as $\alpha_{ij} f_{i,\min} + (1 - \alpha_{ij}) f_{i,\max}$, $i = 1, 2$ where $\alpha_{1j} + \alpha_{2j} = 1$, $\alpha_{ij} \in [0, 1]$. The subscript j corresponds of the number on discretized points of the interval $[0, 1]$. Here anchor points are $(0.5078, 0.9913)$ and $(0.95, 0.3017)$. Results on the Fig. 2 obtained for $n_d = 50$ and for the Fig. 3, $n_d = 200$. We observe that PP subproblems are incapable to generate a significant portion of the Pareto frontier and obtained solutions are not well-distributed. But solutions of IC subproblems, as observed in Figs. 4 and 5, are well-diversified over the entire Pareto set and no portion is left out to generate.

Example 2 In this example, we have considered an engineering design problem—a three-bar truss problem. This problem and its variations are broadly used [1, 4, 13, 17–19] as benchmark to recognize efficiency of Pareto frontier generation methods.

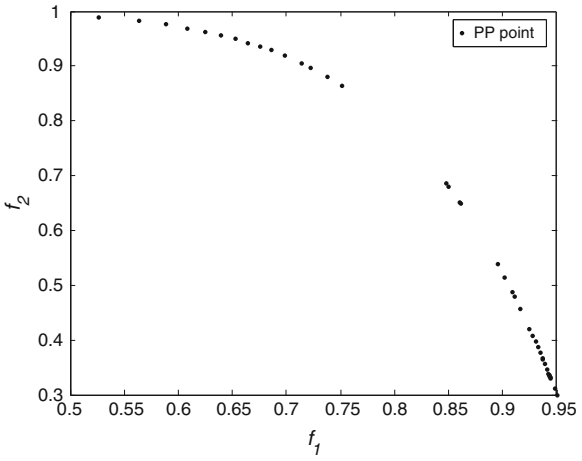


Fig. 2 Performance of PP method on Example 1 for 50 evaluations

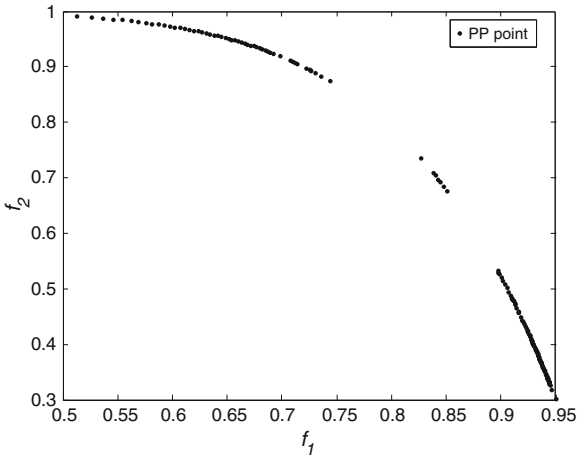


Fig. 3 Performance of PP method on Example 1 for 200 evaluations

The three-bar truss under static loading is shown in the Fig. 6. In the problem, total volume of the truss and linear combination of the horizontal and vertical displacements of the node P for a small deformation of the truss are to be minimized simultaneously. The design variables are cross section of the bars: a_1, a_2 and a_3 say. All of them are bounded by 0.1 and 2 cm^2. Here, the subscripts 1, 2 and 3 are used to refer left, middle, and right bar, respectively. Different numerical data of the problem are as follows:

$F = 20\,\text{kN}$, $L = 1\,\text{m}$, Young modulus of the bars $E = 200\,\text{GPa}$, maximum stress accepted in each bar is $\sigma = 200\,\text{MPa}$.

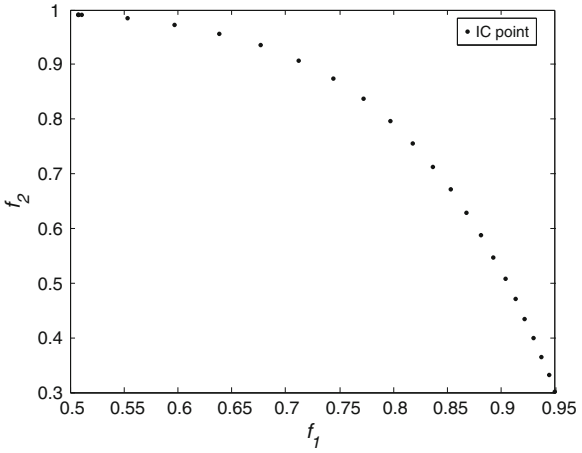


Fig. 4 Performance of IC method on Example 1 for 50 evaluations

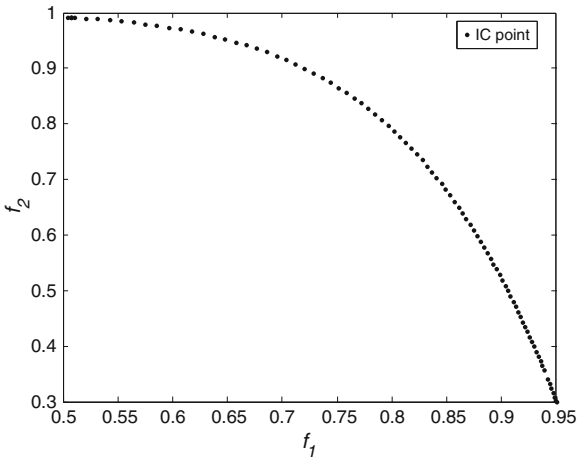


Fig. 5 Performance of IC method on Example 1 for 200 evaluations

Thus, the MOP can be described as:

$$\begin{aligned} &\min \begin{pmatrix} \delta(a_1, a_2, a_3) \\ V(a_1, a_2, a_3) \end{pmatrix} \\ &\text{subject to } \frac{|T_i|}{a_i} \leq \sigma \\ &\qquad 0.1 \times 10^{-4} \leq x_i \leq 2 \times 10^{-4} \\ &\qquad i = 1, 2, 3. \end{aligned}$$

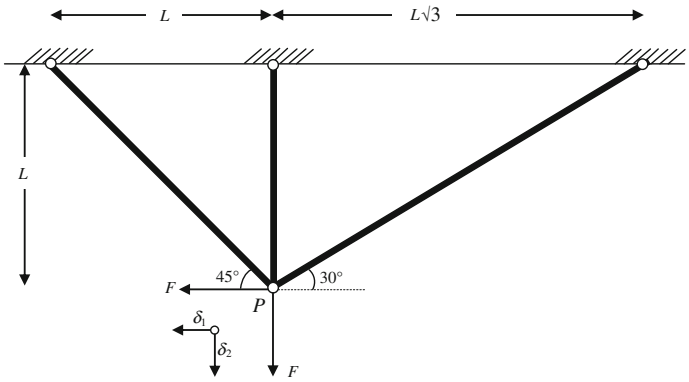


Fig. 6 Three-bar truss under static loading

where T_is are tension of the bars. They can be calculated as:

$$T_1 = \frac{a_1 E}{2L}(\delta_1 - \delta_2),$$

$$T_2 = \frac{a_2 E}{L}\delta_2,$$

$$T_3 = \frac{a_3 E}{4L}\left(\delta_1 + \sqrt{3}\delta_2\right).$$

The objective functions are:

$$f_1 \equiv \delta(a_1, a_2, a_3) = \frac{\delta_1}{4} + \frac{3\delta_2}{4},$$

$$f_2 \equiv V(a_1, a_2, a_3) = L\left(\sqrt{2}a_1 + a_2 + 2a_3\right).$$

The displacements δ_1 and δ_2 can be determined from the expression of of T_is and the force balance equations:

$$\text{vertical: } F = T_2 + \frac{T_1}{\sqrt{2}} + \frac{T_3}{2},$$

$$\text{horizontal: } F = \frac{\sqrt{3}T_3}{2} - \frac{T_1}{\sqrt{2}}.$$

A discrete approximation of the feasible set of this problem is shown in the Fig. 7. As noticed, the arcs *AB* and *CD* of the boundary of the feasible region contain Pareto optimal points. However, the boundary also contains the arc *BC* including non-Pareto optimal points. It is mentioned and illustrated in [18] that to obtain all the globally Pareto optimal points through NBI and NC method one needs to apply Pareto filter algorithm; WS method performs quite poorly—it can offer only two Pareto optimal points; to apply CP method, several iterations are needed to find appropriate scale of weights. Here, we observe in the Fig. 8 that IC method works significantly well and

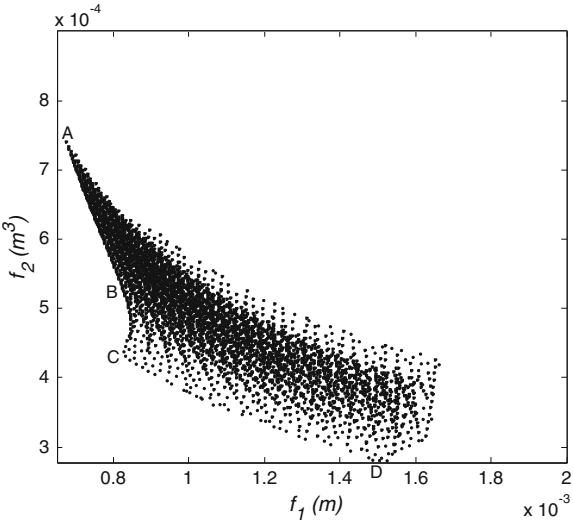


Fig. 7 Feasible region of Example 2

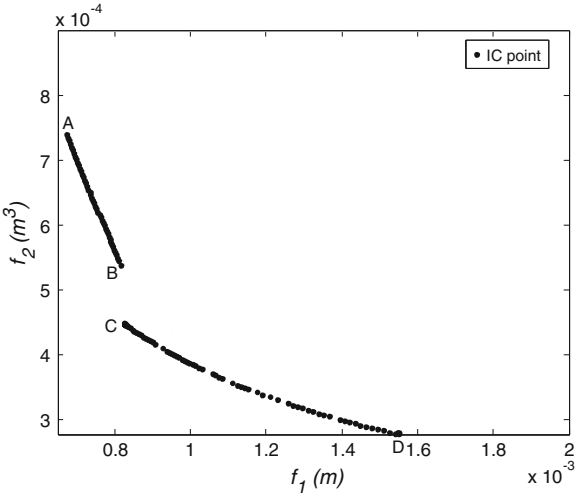


Fig. 8 Performance of IC method on Example 2 for 200 evaluations

generates only globally Pareto optimal points—they are also well-diversified over the Pareto frontier.

Example 3 In this test problem, well-known DTLZ5 problem [6] is considered. This problem, although a three criteria optimization problem, has two-dimensional efficient frontier. As mentioned in [22], all existing classical methods fail to capture the efficient frontier of this problem. This problem has been stated as follows:

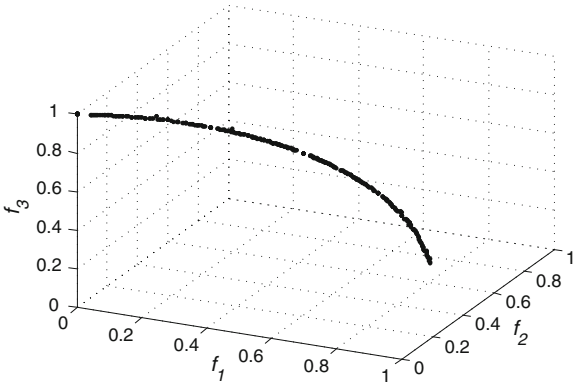


Fig. 9 Performance of IC method on Example 3 (10,000 evaluations)

$$\min \begin{pmatrix} f_1(x) \\ f_2(x) \\ f_3(x) \end{pmatrix}$$
$$\text{subject to } 0 \le x_i \le 1, \quad i = 1, 2, 3.$$

where $x = (x_1,\ x_2,\ x_3)^{\mathrm{T}}$ and
$f_1(x) = (1 + g(x_3)) \cos(\theta_1(x)) \cos(\theta_2(x))$,
$f_2(x) = (1 + g(x_3)) \cos(\theta_1(x)) \sin(\theta_2(x))$,
$f_3(x) = (1 + g(x_3)) \sin(\theta_1(x))$,
$g(x_3) = \left(x_3 - \frac{1}{2}\right)^2$,
$\theta_1(x) = \frac{\pi}{2} x_1$,
$\theta_2(x) = \frac{\pi}{4} \frac{1+2g(x_3)x_2}{1+g\left(\sqrt{x_1^2+x_2^2+x_3^2}\right)}$.

This problem has Pareto optimal curve [22]: $f_3^2 = 1 - f_1^2 - f_2^2$ with $f_1 = f_2 \in [0, \frac{1}{\sqrt{2}}]$. As shown in [9], DSD method can yield all the Pareto points, but parameters of the method should be choosen very tactfully and this leads to significant knowledge about the solutions to be found. However, proposed IC method can also generate the entire Pareto optimal frontier—depicted in the Fig. 9—without requiring prior knowledge about the problem.

Thus, we observe that direction-based IC method successfully obtain global Pareto optimal points of the problems.

- *Finding D-Pareto points*: It is worthy to mention here that IC method can not only efficiently obtain global Pareto optimal points, but also obtain ε-Pareto optimal points and more generally Pareto optimal points with respect to any ordering (pointed convex) cone D. To get D-Pareto optimal points the constraint inequality in IC(β) has to be taken as $z\hat{\beta} - f(x) \in D$ instead of $z\hat{\beta} \geqq f(x)$. Similarly,

to obtain ε-Pareto optimal points, one needs to take the constraint inequality as $z\hat{\beta} - f(x) \in \mathbb{R}^k_{\varepsilon}$, or $\frac{\text{dist}\left(z\hat{\beta} - f(x),\ \mathbb{R}^k_{>}\right)}{\|z\hat{\beta} - f(x)\|} \leq \varepsilon$. Quite often identifying ε-Pareto optimal points are of due importance since at an ε-Pareto optimal point trade-off of all the objective functions are bounded by ε and $1/\varepsilon$. Thus, ε-Pareto optimal points are properly Pareto optimal. Owing to this fact, DM always try to choose an ε-Pareto optimal point as most preferable solution of the MOP because DM usually is not willing to improve one unit of an objective function at the cost of an infinite loss of another objective function.

- *Finding knee regions*: Let by knee points/regions we refer the points of local minima of distances between ideal point and Pareto points. Finding knee regions of Pareto set usually facilitate DM's final selection of solution from the Pareto set, since, DM ideally wants to obtain ideal point, but it is not attainable by criteria feasible set, and thus, DM may like to obtain a point which has smallest possible deviation from ideal point. Let us note that if $(x^*, z^*, \hat{u})$ is solution of an IC subproblems, then z^* essentially measures the distance between ideal point and the Pareto point $f(x^*)$. Thus, local minimum of values of z^* of IC subproblems offer knee regions of the Pareto set with respect to the ideal point.

5 Comparison

In this section, let us compare the proposed IC method with the existing other similar methods. First, let us see the subproblem formulation of each of those existing techniques. Discussion of all the problems are made with respect to the pointed, closed, convex cone $K = \mathbb{R}^k_{\geqq}$ and used f^* is the ideal point. The matrix Φ has the meaning as described in [3].

Pascoletti–Serafini Scalarization [21]:

$$\text{SP}(a, r) \quad \begin{cases} \min & t \\ \text{subject to} & a + tr - f(x) \in K \\ & x \in \mathscr{X}. \end{cases}$$

Normal Boundary Intersection [3]:

$$\text{NBI}(\beta) \quad \begin{cases} \min & t \\ \text{subject to} & \Phi\beta + t\hat{n} - f(x) + f^* = 0 \\ & x \in \mathscr{X}. \end{cases}$$

Proposed IC:

$$\text{IC}(\beta) \begin{cases} \min & t \\ \text{subject to} & t\hat{\beta} - f(x) + f^* \in K \\ & x \in \mathscr{X}. \end{cases}$$

5.1 Pascoletti–Serafini Scalarization

- Solutions of original Pascoletti–Serafini scalarization are not always Pareto/ weakly-Pareto points.
- Helbig [11] proved that if Pareto set of the MOP in nonempty, then for $(a, r) \in \mathbb{R}^k \times \text{int}(K)$, solution of $\text{PS}(a, r)$ is weakly Pareto optimal.
- Observing that Helbig's restriction considers potentially an unbounded set of point in $\mathbb{R}^k \times \text{int}(K)$, Eichfelder [8] tried to find more stricter condition on the parameters (a, r).
- For a bi-objective problem, Eichfelder considered $r = -\hat{n}$, where $\hat{n}$ is the unit normal direction on the so-called CHIM (Convex Hull of Individual Minima), and allowed the reference point a to vary on the projection set of $f(\mathscr{X})$ on a line parallel to CHIM. This line must lie beneath $f(\mathscr{X})$. Mathematically, Eichfelder showed that in this way complete Pareto set of bi-objective problem can be generated. However, the projection method is no-longer applicable for more than two objective functions (see [8]).
- Eichfelder's method though efficiently capture Pareto set of bi-objective problems, but it does not give any other information to facilitate DM's final selection of solution. Like, it does not give positions of weak Pareto points or proper Pareto points or knee regions of the Pareto set.
- We also note that how to choose parameter a over the projection set is not given properly, and thus, diversity of generated solution over the entire Pareto set is questionable.
- Parameter restriction of Eichfelder's approach needs information about the the criteria feasible set $f(\mathscr{X})$, and thus, this parameter set changes for every MOP.

5.2 Normal Boundary Intersection

- NBI technique is a restricted case of $\text{PS}(a, r)$ with $a = \Phi\beta$, $r = -\hat{n}$ and the cone K must have empty interior.
- Similarly to Eichfelder's approach, to obtain Pareto set, NBI method also considered r to be a fixed direction and a to be a variable reference point restricted to lie on CHIM.
- As K is restricted to have empty interior, outcome solution of the NBI subproblem may be non-Pareto optimal. Nonetheless, NBI cannot capture the entire Pareto set, cannot work for non-convex problems, and it has several other deficiencies [3].
- NBI also does not give positions of weak Pareto points or proper Pareto points or knee regions of the Pareto set.
- Shukla [23] proposed a modification of NBI such that solution of modified NBI must be weakly-Pareto optimal. However, being originated from NBI, the modified NBI method bears all the deficiencies of the NBI method, but less likely.

5.3 Ideal Cone

- Solution of each IC subproblem is shown to be weakly-Pareto optimal and each Pareto point is attainable by an IC subproblem. In contrast, NBI or SP method does not guaranty weak-Pareto optimality of the outcome solutions.
- To obtain entire Pareto set of MOP, in the proposed technique, the reference point a is taken as a fixed point and $r = \hat{\beta}$ is considered to vary over the unit sphere $\mathbb{S}^{k-1}$ on the first hyperoctant. But in all other methods, r is taken as fixed and a is considered to vary. This variable reference point considerably changes the subproblems, and hence, needs extra computational cost than the IC method. Nevertheless, NBI and SP method cannot generate complete Pareto set.
- A simple uniform-discretization of $\mathbb{S}^{k-1} \bigcap \mathbb{R}^k_{\geqq}$ (a bounded set, unlike Helbig's unbounded set) will give a discrete approximation of parameter set of IC subproblems.
- Approach of the IC method can easily detect and separate weakly-Pareto point (please refer to the paragraph after Theorem 2). Moreover, through solution set of the IC method we can easily detect the region of the Pareto set where objectives can have unbounded trade-offs (Proposition 1).
- IC method's solution can also capture knee regions of the Pareto set with respect to the ideal point. Here, we note that no such extra information can be obtained from the solution set of the NBI or SP method.
- We also note that MNBI left a significant portion of the Pareto set to obtain, Eichfelder's techniques gives several redundant solutions, but IC method captures the entire Pareto set without such drawbacks.
- More importantly, parameter restriction of Eichfelder's approach needs information about the the criteria feasible set $f(\mathscr{X})$, and thus, this parameter set changes for every MOP. For MNBI and NBI also we need to compute the set $\Phi\beta$, which changes to every MOP. But proposed method's parameter $\hat{\beta}$ does not depend on $f(\mathscr{X})$ and the parameter set does not change corresponding to each MOPs.
- We note that IC method is searching Pareto points in each and every possible directions form the ideal point, and thus, the generated Pareto set obviously maintains a diversity throughout the Pareto surface. By contrast, other methods start from a reference point, which is restricted to lay on a plane, and then search Pareto points along the normal to the considered plane, i.e., those methods search Pareto points along a particular direction. Therefore, unless the Pareto surface is approximately parallel to that plane, the generated Pareto set by those methods are trivially not diversified over the entire Pareto set.

6 Conclusion

In the presented study, a Pareto set generation method has been developed. To solve the IC subproblems (2) efficiently, a uniform discretization for $\mathbb{S}^{k-1}_{\geqq}$ is used. Proposed IC method bears necessary and sufficient condition for global Pareto optimality if

$\mathscr{Y}_N = \mathscr{Y}_{wN}$. A little modification of the constraint inequality, which is mentioned in the foregoing section, ensures general D-Pareto optimality of the outcome solutions by IC method. Thus, the proposed method not only captures global Pareto points, but also obtains D-Pareto points and, more importantly, ε-Pareto optimal points. We have shown that IC method though intended to obtain only Pareto points, may also attain weakly Pareto points. A simple procedure to identify weak Pareto optimal points attained by IC method has been mentioned. Similarly, from the IC solution points itself we can easily detect the position of the Pareto surface where objective functions may have unbounded trade-offs. This information may facilitate DM to choose the best preferable solution or best design or best decision of the MOP. Identification of the points of unbounded trade-offs of two criteria eventually mean finding the Pareto points, which are not proper Pareto optimal points. This identification of proper Pareto optimal points from IC solution points using Lemmas 1 and 2 will be done in our further research on IC method. More details on mathematical perspective and advantage of IC method over the existing classical methods to generate Pareto set can be also obtained in future.

It is important to mention here that the approach of the proposed IC method also can efficiently generate complete fuzzy non-dominated set for fuzzy multi-objective optimization problems. This work on capturing complete fuzzy non-dominated set can be found in [10].

Acknowledgments Authors are grateful to the anonymous reviewers for their constructive comments and valuable suggestions. First, the author gratefully acknowledges a research scholarship awarded by the Council of Scientific and Industrial Research, Government of India (award no. 09/081(1054)/2010-EMR-I). Second, the author acknowledges the financial support given by the Department of Science and Technology, Government of India (SR/S4/M:497/07).

References

1. Athan, T., Panos, Y.: A note on weighted criteria methods for compromise solutions in multi-objective optimization. Eng. Optim. **27**(2), 155–176 (1996)
2. Cohon, J.L.: Multiobjective Programming and Planning, vol. 140. Dover Publications, New York (2004)
3. Das, I., Dennis, J.E.: A closer look at drawbacks of minimizing weighted sums of objectives for pareto set generation in multicriteria optimization problems. Struct. Multidisc. Optim. **14**(1), 63–69 (1997)
4. Das, I., Dennis, J.E.: Normal-boundary intersection: a new method for generating the pareto surface in nonlinear multicriteria optimization problems. SIAM J. Optim. **8**(3), 631–657 (1998)
5. Deb, K.: Multi-objective Optimization Using Evolutionary Algorithms, vol. 16. Wiley, New York (2001)
6. Deb, K., Thiele, L., Laumanns, M., Zitzler, E.: Scalable multi-objective optimization test problems. In: Proceedings of the Congress on Evolutionary Computation-2002, pp. 825–830 (2002)
7. Ehrgott, M.: Multicriteria Optimization, vol. 491. Springer, Berlin (2005)
8. Eichfelder, G.: Adaptive Scalarization Methods in Multiobjective Optimization. Springer, Berlin (2008)
9. Erfani, T., Utyuzhnikov, S.: Directed search domain: a method for even generation of the pareto frontier in multiobjective optimization. Eng. Optim. **43**(5), 467–484 (2011)

10. Ghosh, D., Chakraborty, D.: Fuzzy ideal cone: a method to obtain complete fuzzy non-dominated set of fuzzy multi-criteria optimization problems with fuzzy parameters. In: IEEE International Conference on Fuzzy Systems (FUZZ), pp. 1–8 (2013)
11. Helbig, S.: An interactive algorithm for nonlinear vector optimization. Appl. Math. Optim. **22**(2), 147–151 (1990)
12. Hwang, C., Masud, A., Paidy, S.: Multiple Objective Decision Making, Methods and Applications: A State-of-the-art Survey. Springer, Berlin (1979)
13. Koski, J.: Defectiveness of weighting method in multicriterion optimization of structures. Commun. Appl. Numer. Methods **1**(6), 333–337 (1985)
14. Li, D., Yang, J.B., Biswal, M.P.: Quantitative parametric connections between methods for generating noninferior solutions in multiobjective optimization. Eur. J. Oper. Res. **117**(1), 84–99 (1999)
15. Lighter, M.R., Director, S.W.: Multiple criterion optimization for the design of electronic circuits. IEEE Trans. Circuits Syst. **28**(3), 169–179 (1981)
16. Marler, R., Arora, J.: Survey of multi-objective optimization methods for engineering. Struct. Multidisc. Optim. **26**(6), 369–395 (2004)
17. Martínez, M., Sanchis, J., Blasco, X.: Global and well-distributed pareto frontier by modified normalized normal constraint methods for bicriterion problems. Struct. Multidisc. Optim. **34**(3), 197–209 (2007)
18. Messac, A., Ismail-Yahaya, A., Mattson, C.A.: The normalized normal constraint method for generating the pareto frontier. Struct. Multidisc. Optim. **25**(2), 86–98 (2003)
19. Messac, A., Mattson, C.: Generating well-distributed sets of pareto points for engineering design using physical programming. Optim. Eng. **3**(4), 431–450 (2002)
20. Miettinen, K.: Nonlinear Multiobjective Optimization, vol. 12. Springer, Berlin (1999)
21. Pascoletti, A., Serafini, P.: Scalarizing vector optimization problems. J. Optim. Theory Appl. **42**(4), 499–524 (1984)
22. Shukla, P., Deb, K.: On finding multiple pareto-optimal solutions using classical and evolutionary generating methods. Eur. J. Oper. Res. **181**(3), 1630–1652 (2007)
23. Shukla, P.K.: On the normal boundary intersection method for generation of efficient front. In: Proceedings of the International Conference on Computational Science-2007, LNCS, vol. 4487. Springer, pp. 310–317 (2007)
24. Utyuzhnikov, S., Fantini, P., Guenov, M.: A method for generating a well-distributed pareto set in nonlinear multiobjective optimization. J. Comput. Appl. Math. **223**(2), 820–841 (2009)

Chapter 12
Fractional Programming Problem with Bounded Parameters

A. K. Bhurjee and G. Panda

Abstract In this paper, existence of the solution of a nonlinear fractional programming problem with parameters varying in some bounds, is studied. A general nonlinear programming problem, which is free from uncertain parameters, is formulated using the uncertain parameters of the original problem. Relation between the solution of the original problem and the transformed problem is established. The theoretical developments are justified in a numerical example.

Keywords Efficient solution · Fractional programming problem · Parametric optimization problem · Interval valued function

1 Introduction

In a general optimization problem, the parameters are usually considered as real numbers. But there are many real-life situations where parameters are not fixed due to several type of uncertainties associated with the data set. If these parameters vary in between some lower and upper bounds (i.e., the parameters lie in closed intervals), then the corresponding optimization problem is an interval optimization problem. Readers may refer [1, 4–7, 10–15, 19] for some major contributions in the area of optimization problems with interval parameters during the last 2 decades. If the objective function of an interval optimization problem is the ratio of two interval

The authors thank the referee whose justified critical remarks on the original version led to an essential reworking of the paper.

A. K. Bhurjee · G. Panda (✉)
Indian Institute of Technology Kharagpur, Kharagpur 721302, India
e-mail: ajaybhurji@gmail.com

G. Panda
e-mail: geetanjali@maths.iitkgp.ernet.in

R. N. Mohapatra et al. (eds.), *Mathematics and Computing 2013*,
Springer Proceedings in Mathematics & Statistics 91,
DOI: 10.1007/978-81-322-1952-1_12, © Springer India 2014

valued function, then we call this as interval fractional programming problem and denote by (IFP). Existence of the solution of general fractional programming problem is studied by many authors (see Refs. [3, 9, 17, 18, 20]) in several directions. For the first time, Hladik [4] focused on interval fractional programming, whose objective function is the ratio of two linear interval valued functions. Nonlinear interval fractional programming problem has not been studied yet.

In this paper, a fractional programming problem (IFP) is considered in which the objective function is a ratio of two nonlinear interval valued functions and the constraints have linear/nonlinear interval valued functions. Section 2 provides some preliminaries on interval analysis. In Sect. 3 a new interval optimization problem ($\text{IFP}^{\boldsymbol{\lambda}}$) is constructed using the interval valued functions in the numerator and denominator of (IFP), to get rid of the denominator function. Next, ($\text{IFP}^{\boldsymbol{\lambda}}$) is transformed to a deterministic nonlinear programming problem ($\text{IFP}^{\boldsymbol{\lambda}}_w$), which is free from all uncertain parameters. Relations between the solutions of these three optimization problems are established in this section. Finally, it is proved that solution of ($\text{IFP}^{\boldsymbol{\lambda}}_w$) is an efficient solution of (IFP). These results are illustrated with a numerical example in Sect. 4.

Throughout the paper, the following notations are used. Bold capital letters denote closed intervals; $I(R)$ = The set of all closed intervals in R; $(I(R))^k$ = The product space $\underbrace{I(R) \times I(R) \times \cdots \times I(R)}_{(k \text{ times})}$; $\mathbf{C}^k_v = k$-dimensional column whose elements are intervals; $\mathbf{C}^k_v \in (I(R))^k, \mathbf{C}^k_v = (\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_k)^{\mathrm{T}}, \mathbf{C}_j = [c^L_j, c^R_j],\ j \in \Lambda_k,\ \Lambda_k = \{1, 2, \ldots, k\}$.

2 Preliminaries

Let $* \in \{+, -, \cdot, /\}$ be a binary operation on the set of real numbers. The binary operation $\circledast$ between two intervals $\mathbf{A} = [a^L, a^R]$ and $\mathbf{B} = [b^L, b^R]$ in $I(R)$, denoted by $\mathbf{A} \circledast \mathbf{B}$ is the set $\{a * b : a \in \mathbf{A}, b \in \mathbf{B}\}$. In the case of division, $(\mathbf{A} \oslash \mathbf{B})$, it is assumed that $0 \notin \mathbf{B}$. These interval operations can also be expressed in terms of parameters. Any point in $\mathbf{A}$ may be expressed as $a(t) = a^L + t(a^R - a^L),\ t \in [0, 1]$. An interval $\mathbf{A}$ is said to be a positive interval if $a(t)$ is positive $\forall t$. Algebraic operations of intervals can also be explained in parametric form as follows:

$$\mathbf{A} \circledast \mathbf{B} = \{a(t_1) * b(t_2) |\ t_1, t_2 \in [0, 1]\} \tag{1}$$

An interval vector $\mathbf{C}^k_v \in (I(R))^k,\ \mathbf{C}^k_v = (\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_k)^{\mathrm{T}}$, can be expressed in terms of parameters as

$$\mathbf{C}_v^k = \Big\{c(t)|c(t) = (c_1(t_1), c_2(t_2), \ldots, c_k(t_k))^{\mathrm{T}}, t = (t_1, t_2, \ldots, t_k)^{\mathrm{T}},$$
$$c_j(t_j) \in \mathbf{C}_j, c_j(t_j) = c_j^L + t(c_j^R - c_j^L), t_j \in [0, 1], j \in \Lambda_k\Big\}$$

The set of intervals, $I(R)$ is not a totally order set. Several partial ordering in $I(R)$ exist in the literature. Interval valued function is defined in several ways by many authors (see Refs. [6, 8, 16]). We accept the following partial ordering and express an interval valued function as follows:

Definition 2.1 [2] *For* $\mathbf{A}, \mathbf{B} \in I(R)$,

$$\mathbf{A} \preceq \mathbf{B} \text{ if } a(t) \leq b(t), \quad \forall t \in [0, 1], \text{ and } \mathbf{A} \prec \mathbf{B} \text{ if } a(t) < b(t), \quad \forall t \in [0, 1] \quad (2)$$

Definition 2.2 [2] *For* $c(t) \in \mathbf{C}_v^k$, *let* $f_{c(t)}\colon R^n \to R$. *Then for a given interval vector* $\mathbf{C}_v^k$, *we define an interval valued function* $\mathbf{F}_{\mathbf{C}_v^k}\colon R^n \to I(R)$ *by*

$$\mathbf{F}_{\mathbf{C}_v^k}(x) = \left\{ f_{c(t)}(x) \;\middle|\; f_{c(t)}\colon R^n \to R, c(t) \in \mathbf{C}_v^k \right\}$$

For every fixed x, if $f_{c(t)}(x)$ is continuous in t then $\min\limits_{t\in[0,1]^k} f_{c(t)}(x)$ and $\max\limits_{t\in[0,1]^k} f_{c(t)}(x)$, exist. In that case

$$\mathbf{F}_{\mathbf{C}_v^k}(x) = \left[\min_{t\in[0,1]^k} f_{c(t)}(x), \max_{t\in[0,1]^k} f_{c(t)}(x) \right]$$

If $f_{c(t)}(x)$ is linear in t then $\min\limits_{t\in[0,1]^k} f_{c(t)}(x)$ and $\max\limits_{t\in[0,1]^k} f_{c(t)}(x)$ exist in the set of vertices of $\mathbf{C}_v^k$. If $f_{c(t)}(x)$ is monotonically increasing in t then $\mathbf{F}_{\mathbf{C}_v^k}(x) = [f_{c(0)}(x), f_{c(1)}(x)]$.

3 Existence of Solution of (IFP)

In this section we propose single objective fractional programming problem whose parameters lie in intervals as follows:

$$\text{(IFP): min} \quad \frac{\mathbf{F}_{\mathbf{C}_v^k}(x)}{\mathbf{G}_{\mathbf{D}_v^l}(x)}$$
$$\text{subject to } \mathbf{H}^j_{\mathbf{B}_v^{m_j}}(x) \preceq \mathbf{A}_j, \quad j \in \Lambda_p, \quad (3)$$

where $\mathbf{F}_{\mathbf{C}_v^k}, \mathbf{G}_{\mathbf{D}_v^l}, \mathbf{H}^j_{\mathbf{B}_v^{m_j}}\colon R^n \to I(R)$, $\mathbf{G}_{\mathbf{D}_v^l}(x) \succ \mathbf{0}$, $\mathbf{A}_j \in I(R)$, $\mathbf{A}_j = [a_j^L, a_j^R]$ and $j \in \Lambda_p$. Using Definition 2.2, the objective function can be express as

$$\frac{\mathbf{F}_{\mathbf{C}_v^k}(x)}{\mathbf{G}_{\mathbf{D}_v^l}(x)} = \left\{ \frac{f_{c(t)}(x)}{g_{d(t')}(x)} \middle| \, c(t) \in \mathbf{C}_v^k, d(t') \in \mathbf{D}_v^l, g_{d(t')}(x) > 0 \right\}$$

Using Definition 2.1 in inequality (3), the constraints of (IFP) can be expressed as

$$\left\{ x \in R^n | \mathbf{H}^j_{\mathbf{B}_v^{m_j}}(x) \preceq \mathbf{A}_j, \ j \in \Lambda_q \right\} \equiv \left\{ x \in R^n | h^j_{b_j(t_j'')}(x) \leq a(t_j'') \forall t_j'' \in [0,1], j \in \Lambda_q \right\}$$

Throughout this section, we consider $t = (t_1, t_2, \ldots, t_k)^{\mathrm{T}}$, $t_i \in [0,1]$, $i \in \Lambda_k$, $t' = (t_1', t_2', \ldots, t_l')^{\mathrm{T}}$, $t_q' \in [0,1]$, $q \in \Lambda_l$, $t_j'' \in [0,1]$, $j \in \Lambda_p$.
The feasible set for (IFP) can be expressed as the set

$$\begin{aligned} S &= \left\{ x \in R^n : \mathbf{H}^j_{\mathbf{B}_v^{m_j}}(x) \preceq \mathbf{A}_j, j \in \Lambda_p \right\} \\ &= \bigcap_{j \in \Lambda_p} \left\{ x \in R^n : h^j_{b_j\left(t_j''\right)}(x) \leq a_j\left(t_j''\right), a_j\left(t_j''\right) \in \mathbf{A}_j \right\} \end{aligned}$$

Using Definition 2.2, (IFP) can be rewritten as

$$\min_{x \in S} \frac{\mathbf{F}_{\mathbf{C}_v^k}(x)}{\mathbf{G}_{\mathbf{D}_v^l}(x)} = \min_{x \in S} \left\{ \frac{f_{c(t)}(x)}{g_{d(t')}(x)} \middle| \, g_{d(t')}(x) > 0, c(t) \in \mathbf{C}_v^k, d(t') \in \mathbf{D}_v^l \right\} \tag{4}$$

Here the objective function $\frac{\mathbf{F}_{\mathbf{C}_v^k}(x)}{\mathbf{G}_{\mathbf{D}_v^l}(x)}$ is an interval valued mapping. So minimum of this function should be obtained using a partial ordering. For this reason the exact minimum of (4) does exist. Since for different pairs (t, t'), $\frac{f_{c(t)}(x)}{g_{d(t')}(x)}$ represents different functions of x, so minimum solution of (4) can be considered as an efficient solution. Assuming that for every pair $(c(t), d(t'))$, the optimization problem $\min_{x \in S} \frac{f_{c(t)}(x)}{g_{d(t')}(x)}$ has a solution, we define the solution of (IFP) in the light of solution of set optimization problem as follows:

Definition 3.1 *$x^* \in S$ is called an efficient solution of* (IFP) *if there is no $x \in S$ with*

$$\frac{f_{c(t)}(x)}{g_{d(t')}(x)} \leq \frac{f_{c(t)}(x^*)}{g_{d(t')}(x^*)} \ \forall (t, t') \text{ and for at least one } (\mathbf{t}, \mathbf{t}') \neq (t, t'), \ \frac{f_{c(\mathbf{t})}(x)}{g_{d(\mathbf{t}')}(x)} < \frac{f_{c(\mathbf{t})}(x^*)}{g_{d(\mathbf{t}')}(x^*)}$$

Denote $\boldsymbol{\lambda} = [\lambda^L, \lambda^R] = \{\lambda | \lambda \in [\lambda^L, \lambda^R]\}$, where for fixed $x \in S$, $\lambda^L = \min_{t,t'} \frac{f_{c(t)}(x)}{h_{e(t')}(x)}$ and $\lambda^R = \max_{t,t'} \frac{f_{c(t)}(x)}{h_{e(t')}(x)}$. Consider the following parametric problem

$$(\mathrm{IFP}^{\boldsymbol{\lambda}}): \min_{x \in S} \left[\mathbf{F}_{\mathbf{C}_v^k}(x) \ominus \left(\boldsymbol{\lambda} \otimes \mathbf{G}_{\mathbf{D}_v^l}(x) \right) \right] \tag{5}$$

Equation (5) is equivalent to

$$\min_{x\in S}\left\{f_{c(t)}(x)-\lambda g_{d(t')}(x)|\ c(t)\in \mathbf{C}_v^k, d(t')\in \mathbf{D}_v^l, \lambda\in\left[\lambda^L,\lambda^R\right]\right\} \tag{6}$$

Denote $\varphi_{t,t'}(\lambda,x)=f_{c(t)}(x)-\lambda g_{d(t')}(x)$ and $\Phi(\boldsymbol{\lambda})=\min_{x\in S}[\mathbf{F}_{\mathbf{C}_v^k}(x)\ominus(\boldsymbol{\lambda}\otimes\mathbf{G}_{\mathbf{D}_v^l}(x))]$.

The efficient solution of (IFP$^{\boldsymbol{\lambda}}$) can be defined in the light of Definition 3.1 as follows:

Definition 3.2 *$x^*\in S$ is called an efficient solution of* (IFP$^{\boldsymbol{\lambda}}$) *if there is no $x\in S$ with $\varphi_{t,t'}(\lambda,x)\le\varphi_{t,t'}(\lambda,x^*)\ \forall(t,t')$ and for at least one $(\mathbf{t},\mathbf{t}')\ne(t,t')$, $\varphi_{\mathbf{t},\mathbf{t}'}(\lambda,x)<\varphi_{\mathbf{t},\mathbf{t}'}(\lambda,x^*)$.*

Consider a weight function w: $[0,1]^k\times[0,1]^l\to R_+$, so that $w(t,t')\varphi_{t,t'}(\lambda,x)$ is integrable and construct an optimization problem

$$(\mathrm{IFP}_w^{\boldsymbol{\lambda}}):\ \min_{\substack{\lambda^L\le\lambda\le\lambda^R\\ x\in S}}\int_{k+l} w(t,t')\varphi_{t,t'}(\lambda,x)\,\mathrm{d}t\,\mathrm{d}t',$$

where $\int_{k+l}=\underbrace{\int_0^1\int_0^1\cdots\int_0^1}_{(k+l\text{ times})}$, $t=(t_1,t_2,\ldots,t_k)^{\mathrm{T}}$, $t'=(t_1',t_2',\ldots,t_l')^{\mathrm{T}}$, $\mathrm{d}t=\mathrm{d}t_1\mathrm{d}t_2\ldots\mathrm{d}t_k$, $\mathrm{d}t'=\mathrm{d}t_1'\mathrm{d}t_2'\ldots\mathrm{d}t_l'$.

Note: $w(t,t')$ may be treated as a preference weight function, which has to be provided by the decision maker. Different preference functions can be provided to estimate the Pareto optimal value of the model. For every t,t', $w(t,t')=1$ indicates that the investor's natural attitude is to estimate the mean. If $\int_{k+l}w(t,t')\,\mathrm{d}t'=1$ then the investor's inclination is to estimate in between the optimistic and pessimistic optimal value.

Theorem 3.1 *If (x^*,λ^*) is an optimal solution of* (IFP$_w^{\boldsymbol{\lambda}}$) *then x^* is an efficient solution of* (IFP$^{\boldsymbol{\lambda}}$).

Proof Let (x^*,λ^*) be an optimal solution of (IFP$_w^{\boldsymbol{\lambda}}$) and x^* is not an efficient solution of (IFP$^{\boldsymbol{\lambda}}$). Then by Definition 3.2, there is some $x\in S$ with

$$\begin{aligned} f_{c(t)}(x)-\lambda^* g_{d(t')}(x)&\le f_{c(t)}(x^*)-\lambda^* g_{d(t')}(x^*)\ \forall(t,t')\\ &\quad\text{and for at least one } (\mathbf{t},\mathbf{t}')\ne(t,t'),\\ f_{c(\mathbf{t})}(x)-\lambda^* g_{d(\mathbf{t}')}(x)&< f_{c(\mathbf{t})}(x^*)-\lambda^* g_{d(\mathbf{t}')}(x^*)\end{aligned}$$

Hence for a weight function w: $[0,1]^k\times[0,1]^l\to R_+$, there exists $x\in S$, such that

$$w\left(t,t'\right)\left(f_{c(t)}(x)-\lambda^{*}g_{d(t')}(x)\right)\leq w\left(t,t'\right)\left(f_{c(t)}(x^{*})-\lambda^{*}g_{d(t')}(x^{*})\right)\ \forall\left(t,t'\right)$$
$$\text{and for at least one } \left(\mathbf{t},\mathbf{t}'\right)\neq\left(t,t'\right),$$
$$w\left(\mathbf{t},\mathbf{t}'\right)\left(f_{c(\mathbf{t})}(x)-\lambda^{*}g_{d(\mathbf{t}')}(x)\right)<w\left(\mathbf{t},\mathbf{t}'\right)\left(f_{c(\mathbf{t})}(x^{*})-\lambda^{*}g_{d(\mathbf{t}')}(x^{*})\right)$$

Integrating with respect to t, t', the above relation implies that for some x in S

$$\int_{k+l} w(t,t')\varphi_{t,t'}(\lambda^{*},x)\,\mathrm{d}t\mathrm{d}t' < \int_{k+l} w(t,t')\varphi_{t,t'}(\lambda^{*},x^{*})\,\mathrm{d}t\mathrm{d}t',$$

which is impossible since (x^{*},λ^{*}) is the optimal solution of $(\mathrm{IFP}_{w}^{\boldsymbol{\lambda}})$. Hence x^{*} is an efficient solution of (IFP). □

Proceeding in a similar way the relationship between the solution of the problems (IFP) and $(\mathrm{IFP}^{\boldsymbol{\lambda}})$ can be studied in the following theorem.

Theorem 3.2 *$x^{*}\in S$ is an efficient solution of* (IFP) *if and only if x^{*} is an efficient solution of* $(\mathrm{IFP}^{\boldsymbol{\lambda}})$ *and* $0\in\Phi(\boldsymbol{\lambda})$.

Proof Let x^{*} be an efficient solution of (IFP) then there is no $x\in S$ such that

$$\frac{f_{c(t)}(x)}{g_{d(t')}(x)}\leq\frac{f_{c(t)}(x^{*})}{g_{d(t')}(x^{*})}\ \forall\left(t,t'\right)\text{ and for at least one }\left(\mathbf{t},\mathbf{t}'\right)\neq\left(t,t'\right),\ \frac{f_{c(\mathbf{t})}(x)}{g_{d(\mathbf{t}')}(x)}<\frac{f_{c(\mathbf{t})}(x^{*})}{g_{d(\mathbf{t}')}(x^{*})}$$

For fixed (t,t'), $(\mathbf{t},\mathbf{t}')$, there exist λ, λ' such that $\frac{f_{c(t)}(x^{*})}{g_{d(t')}(x^{*})}=\lambda$, $\frac{f_{c(\mathbf{t})}(x^{*})}{g_{d(\mathbf{t}')}(x^{*})}=\lambda'\leq\lambda$. This implies that

$$f_{c(t)}(x)-\lambda g_{d(t')}(x)\leq f_{c(t)}(x^{*})-\lambda g_{d(t')}(x^{*})\ \forall\left(t,t'\right)$$

and for at least one $(\mathbf{t},\mathbf{t}')\neq(t,t')$, $f_{c(\mathbf{t})}(x)-\lambda g_{d(\mathbf{t}')}(x)<f_{c(\mathbf{t})}(x^{*})-\lambda g_{d(\mathbf{t}')}(x^{*})$. From Definition 3.2 for (IFP^{λ}), we obtain that x^{*} is an efficient solution for the problem (6) and $\varphi_{t,t'}(\lambda,x^{*})=0$, so $0\in\Phi(\boldsymbol{\lambda})$.

Suppose there is a λ such that x^{*} is an efficient solution of problem (IFP^{λ}) then there exist no $x\in S$ with $f_{c(t)}(x)-\lambda g_{d(t')}(x)\leq f_{c(t)}(x^{*})-\lambda g_{d(t')}(x^{*})$, and for at least one $(\mathbf{t},\mathbf{t}')\neq(t,t')$, $f_{c(\mathbf{t})}(x)-\lambda g_{d(\mathbf{t}')}(x)<f_{c(\mathbf{t})}(x^{*})-\lambda g_{d(\mathbf{t}')}(x^{*})$. Since $0\in\Phi(\boldsymbol{\lambda})$, so for fixed (t,t') and $(\mathbf{t},\mathbf{t}')$ there exists λ such that $f_{c(t)}(x^{*})-\lambda g_{d(t')}(x^{*})=0$ and $f_{c(\mathbf{t})}(x^{*})-\lambda g_{d(\mathbf{t}')}(x^{*})=0$. From the above discussion, there is no $x\in S$ such that

$$\frac{f_{c(t)}(x)}{g_{d(t')}(x)}\leq\frac{f_{c(t)}(x^{*})}{g_{d(t')}(x^{*})}\ \forall\left(t,t'\right)\text{ and for at least one }\left(\mathbf{t},\mathbf{t}'\right)\neq\left(t,t'\right),\ \frac{f_{c(\mathbf{t})}(x)}{g_{d(\mathbf{t}')}(x)}<\frac{f_{c(\mathbf{t})}(x^{*})}{g_{d(\mathbf{t}')}(x^{*})}$$

From Definition 3.1, x^{*} is an efficient solution of (IFP). □

Using the result in Theorems 3.1 and 3.2, we may conclude that (x^*, λ^*) is an optimal solution of $(\text{IFP}_w^{\boldsymbol{\lambda}})$ then x^* is an efficient solution of (IFP).

4 A Numerical Example

Some results of the previous section can be verified in the following example. Consider the following interval fractional quadratic programming problem as

$$\begin{aligned}(\text{IFP}):\ \min\ & \frac{[-10,-6]x_1 \oplus [2,3]x_2 \oplus [4,10]x_1^2 \oplus [-1,1]x_1x_2 \oplus [10,20]x_2^2}{[-5,-3]x_1 \oplus [1,2]x_2 \oplus [1,1]\left(2x_1^2 - 2x_1x_2 + 2x_2^2\right)}\\ \text{subject to}\ & [1,2]x_1 \oplus [3,3]x_2 \succeq [1,10],\ [-2,8]x_1 \oplus [4,6]x_2 \succeq [4,6],\\ & [-5,-3]x_1 \oplus [1,2]x_2 \oplus [1,1]\left(2x_1^2 - 2x_1x_2 + 2x_2^2\right) \succ \mathbf{0}, x_1, x_2 \geq 0.\end{aligned}$$

Denote

$$\begin{aligned}&\mathbf{F}_{\mathbf{C}_v^5}(x_1,x_2) = [-10,-6]x_1 \oplus [2,3]x_2 \oplus [4,10]x_1^2 \oplus [-1,1]x_1x_2 \oplus [10,20]x_2^2,\\ &\mathbf{G}_{\mathbf{D}_v^2}(x_1,x_2) = [-5,-3]x_1 \oplus [1,2]x_2 \oplus [1,1]\left(2x_1^2 - 2x_1x_2 + 2x_2^2\right),\\ &\mathbf{H}^1_{\mathbf{B}_v^2}(x_1,x_2) = [1,2]x_1 \oplus [3,3]x_2, \text{ and } \mathbf{H}^2_{\mathbf{B}_v^2}(x_1,x_2) = [-2,8]x_1 \oplus [4,6]x_2.\end{aligned}$$

Then $f_{c(t)}(x_1,x_2) = (-10+4t_1)x_1 + (2+t_2)x_2 + (4+6t_3)x_1^2 + (-1+2t_4)x_1x_2 + (10+10t_5)x_2^2$ and $g_{d(t')}(x_1,x_2) = (-5+2t_1')x_1 + (1+t_2')x_2 + 2x_1^2 - 2x_1x_2 + 2x_2^2$, where $t = (t_1,t_2,\ldots,t_5)^{\mathrm{T}}$, $t \in [0,1]^5$, $t' = (t_1',t_2')^{\mathrm{T}}$, $t' \in [0,1]^2$. Using Definition 2.1, the parametric form of $\mathbf{H}^1_{\mathbf{B}_v^2}(x_1,x_2) \succeq [1,10]$, $\mathbf{H}^2_{\mathbf{B}_v^2}(x_1,x_2) \succeq [4,6]$, and $\mathbf{G}_{\mathbf{D}_v^2}(x_1,x_2) \succ \mathbf{0}$ can be written as $h^1_{b_1(t_1'')}(x_1,x_2) \geq (1+9t_1'')\ \forall t_1'' \in [0,1]$, $h^2_{b_2(t_2'')}(x_1,x_2) \geq (4+2t_2'')\ \forall t_2'' \in [0,1]$, and $g_{d(t_3'')}(x_1,x_2) > 0\ \forall t_3'' \in [0,1]$, respectively, where $h^1_{b_1(t_1'')}(x_1,x_2) = (1+t_1'')x_1 + 3x_2$, $h^2_{b_2(t_2'')}(x_1,x_2) = (-2+10t_2'')x_1 + (4+2t_2'')x_2$ and $g_{d(t_3'')}(x_1,x_2) = (-5+2t_3'')x_1 + (1+t_3'')x_2 + 2x_1^2 - 2x_1x_2 + 2x_2^2$. Hence

$$\begin{aligned}S &= \Big\{(x_1,x_2) \in R^2 |\ h^1_{b_1(t_1'')}(x_1,x_2) \geq (1+9t_1''), h^2_{b_2(t_2'')}(x_1,x_2) \geq (4+2t_2''),\\ &\quad g_{d(t_3'')}(x_1,x_2) > 0, x_1 \geq 0, x_2 \geq 0, t_1'', t_2'', t_3'' \in [0,1]\Big\}\\ &= \Big\{(x_1,x_2) \in R^2 |\ x_1 + 3x_2 \geq 1, 2x_1 + 3x_2 \geq 10, -2x_1 + 4x_2 \geq 4, 8x_1 + 6x_2 \geq 6,\\ &\quad -5x_1 + x_2 + 2x_1^2 - 2x_1x_2 + 2x_2^2 > 0, x_1 \geq 0, x_2 \geq 0\Big\}\end{aligned}$$

Here $\boldsymbol{\lambda} = [4, 5]$. Consider weight function w: $[0, 1]^7 \to R_+$ define as $w(t, t') = t_1 + t_3$ then the deterministic problem (IFP_w^{λ}) becomes

$$\min_{\substack{4\le\lambda\le5\\(x_1,x_2)\in S}} \left((-23/3 + 4\lambda)x_1 + (5/2 - (3/2)\lambda)x_2 + (15/2 - 2\lambda)x_1^2 + 2\lambda x_1 x_2 + (15 - 2\lambda)x_2^2\right)$$

Using LINGO the optimal solution of the above problem is found as $(x_1^*, x_2^*, \lambda^*) = (0, 3.3333, 5)$. Hence from Theorems 3.1 and 3.2, $(0, 3.3333)$ is an efficient solution for (IFP) and the optimal value of $(\mathbf{IFP^{\lambda}})$ is $[-26.687, 119.382]$.

5 Conclusion

In this paper, a fractional programming problem with interval parameters and interval parametric optimization problem are discussed. The interval parametric optimization problem is converted to a general optimization problem, which is free from uncertain parameters. It is proved that the solution to this transformed problem is an efficient solution of the original problem. This development may be used to discuss the existence of the solution of multi-objective fractional programming problem, which is the future research scope of the present work.

References

1. Bhurjee, A.K., G. Panda: Nonlinear fractional programming problem with inexact parameter. J. Appl. Math. Inform. **31**, 853–867 (2013)
2. Bhurjee, A.K., Panda, G.: Efficient solution of interval optimization problem. Math. Methods Oper. Res. **76**(3), 273–288 (2012)
3. Dinkelbach, Werner: On nonlinear fractional programming. Manag. Sci. **13**(7), 492–498 (1967)
4. Hladik, M.: Generalized linear fractional programming under interval uncertainty. Eur. J. Oper. Res. **205**, 42–46 (2010)
5. Hladik, Milan: Optimal value bounds in nonlinear programming with interval data. TOP **19**(1), 93–106 (2011)
6. Hsien-Chung, Wu: On interval-valued nonlinear programming problems. J. Math. Anal. Appl. **338**(1), 299–316 (2008)
7. Hsien-Chung, Wu: Duality theory in interval-valued linear programming problems. J. Optim. Theory Appl. **150**, 298–316 (2011)
8. Ishibuchi, Hisao, Tanaka, Hideo: Multiobjective programming in optimization of the interval objective function. Eur. J. Oper. Res. **48**(2), 219–225 (1990)
9. Jagannathan, R.: On some properties of programming problems in parametric form pertaining to fractional programming. INFORMS Manag. Sci. **12**(7), 609–615 (1966)
10. Jayswal, A., Stancu-Minasian I., Ahmad, I.: On sufficiency and duality for a class of interval-valued programming problems. Appl. Math. Comput. **21**8(8), 4119–4127, (2011)
11. Jeyakumar, V., Li, G.Y.: Robust duality for fractional programming problems with constraint-wise data uncertainty. Eur. J. Oper. Res. **151**(2), 292–303 (2011)

12. Levin, V.I.: Nonlinear optimization under interval uncertainty. Cybern. Syst. Anal. **35**(2), 297–306 (1999)
13. Li, W., Tian, X.: Numerical solution method for general interval quadratic programming. Appl. Math. Comput. **202**(2), 589–595 (2008)
14. Liu, Shiang-Tai, Wang, Rong-Tsu: A numerical solution method to interval quadratic programming. Appl. Math. Comput. **189**(2), 1274–1281 (2007)
15. Liu, G., Jiang, C., Han, X., Liu, G.: A nonlinear interval number programming method for uncertain optimization problems. Eur. J. Oper. Res. **188**(1), 1–13 (2008)
16. Moore, R.E.: Interval Analysis. Prentice-Hall, Englewood Cliffs (1966)
17. Schaible, Siegfried: Fractional programming. I Duality. Manag. Sci. **22**(8), 858–867 (1976)
18. Schaible, Siegfried, Ibaraki, Toshidide: Fractional programming. Eur. J. Oper. Res. **12**(4), 325–338 (1983)
19. Shaocheng, T.: Interval number and fuzzy number linear programmings. Fuzzy Sets Syst. **66**(3), 301–306 (1994)
20. Wolf, Hartmut: A parametric method for solving the linear fractional programming problem. INFORMS Manag. Sci. **33**(4), 835–841 (1985)

Chapter 13
Approximation Properties of Linear Positive Operators with the Help of Biorthogonal Polynomials

G. Icoz

Abstract In this paper we introduce Konhauser polynomials, Kantorovich type modification of Konhauser polynomials, and q-Laguerre polynomials. Approximation properties of these operators are obtained with the help of the Korovkin theorem. The order of convergence of these operators is computed by means of modulus continuity, Peetre's K-functional, the elements of the Lipschitz class, and the second order modulus of smoothness. Also we introduce the r-th order generalization of these operators and we evaluate their generalizations. Finally we give some applications to differential equations for operators which include Konhauser polynomials.

1 Introduction

In 1960, the Meyer-König and Zeller operators were introduced by Meyer-König and Zeller in [28] as

$$M_n(f;x)=\sum_{k=0}^{\infty} f\left(\frac{k}{k+n+1}\right)\binom{n+k}{k}x^k(1-x)^{n+1}$$

where $0 \le x < 1$.

In order to obtain the monotonicity properties, Cheney and Sharma [7] modified these operators by

$$M_n^*(f;x)=\sum_{k=0}^{\infty} f\left(\frac{k}{k+n}\right)\binom{n+k}{k}x^k(1-x)^{n+1}$$

where $0 \le x < 1$.

G. Icoz (✉)
Department of Mathematics, Gazi University, Teknikokullar 06500, Turkey
e-mail: gurhanicoz@gazi.edu.tr

R. N. Mohapatra et al. (eds.), *Mathematics and Computing 2013*,
Springer Proceedings in Mathematics & Statistics 91,
DOI: 10.1007/978-81-322-1952-1_13, © Springer India 2014

In [7], Cheney and Sharma also introduced the operators

$$P_n(f;x)=\exp\left(\frac{tx}{1-x}\right)\sum_{k=0}^{\infty}f\left(\frac{k}{k+n}\right)L_k^{(n)}(t)\,x^k(1-x)^{n+1}$$

where $0\le x<1$, $-\infty<t\le 0$ and $L_k^{(n)}(t)$ are the Laguerre polynomials. Since $L_k^{(n)}(0)=\binom{n+k}{k}$, the operator $M_n^*(f;x)$ is a special case of the operators $P_n(f;x)$.

Before proceeding further, we recall the following Konhauser's polynomials introduced by Konhauser in [23] with $k\in\mathbb{Z}^+$ as

$$Y_\upsilon^n(t;k)=\frac{1}{\upsilon!}\sum_{i=0}^{\upsilon}\frac{t^i}{i!}\sum_{j=0}^{i}(-1)^j\binom{i}{j}\left(\frac{j+n+1}{k}\right)_\upsilon.$$

Detailed properties of these polynomials can be found in [23].

If we choose $k=1$ in $Y_\upsilon^n(t;k)$, then we obtain $L_k^{(n)}(t)$, the well-known Laguerre polynomials.

We consider the sequence of linear positive operators for $x\in[0,1)$, $t\in(-\infty,0]$ as

$$L_n(f;x)=\frac{1}{F_n(x,t)}\sum_{\upsilon=0}^{\infty}f\left(\frac{\upsilon k}{k(\upsilon-1)+n+1}\right)Y_\upsilon^n(t;k)\,x^\upsilon \tag{1}$$

where $\{F_n(x,t)\}_{n\in N}$ are the generating functions for the sequence of functions $\left\{Y_\upsilon^n(t;k)\right\}_{\upsilon\in N_0}$, $N_0\equiv N\cup\{0\}$, is given by Carlitz [6] in the form

$$F_n(x,t)=\sum_{\upsilon=0}^{\infty}Y_\upsilon^n(t;k)\,x^\upsilon$$

and

$$F_n(x,t)=(1-x)^{-\frac{n+1}{k}}\exp\left\{-t\left[(1-x)^{-\frac{1}{k}}-1\right]\right\}$$

also $Y_\upsilon^n(t;k)\ge 0$ for $t\in(-\infty,0]$. This recurrence relation was given by Srivastava in [33] as

$$tY_{\upsilon-1}^{n+1}(t;k)=(k(\upsilon-1)+n+1)\,Y_{\upsilon-1}^n(t;k)-k\upsilon Y_\upsilon^n(t;k)$$

where $Y_\upsilon^n(t;k)=0$ for $\upsilon\in\mathbb{Z}^-$.

Assume that the following condition holds:

$$\max\{\upsilon,n\}\le k(\upsilon-1)+n+1.$$

Remark 1.1 If we choose $k = 1$ in (1), then L_n turns out to be P_n which was introduced by Cheney and Sharma in [7].

Remark 1.2 If we choose $k = 1$ and $t = 0$ in (1), then L_n reduces to M_n^*. These operators are called as Bernstein power series by Cheney and Sharma in [7].

First, we introduce the Kantorovich type generalization of the operators L_n. Let us denote by $M[0, b]$ $(0 < b < 1)$, the class of measurable functions on $[0, b]$.

We modify the operators L_n (similarly in [3]) by replacing $f\left(\frac{u}{k(\upsilon-1)+n+1}\right)$ in (1) with an integral mean of $f(x)$ over a small interval $\left[\frac{\upsilon}{k(\upsilon-1)+n+1}, \frac{\upsilon+1}{k\upsilon+n+1}\right]$ as follows:

$$\left(L_n^* f\right)(x, t; k) = \frac{1}{F_n(x;t)} \sum_{\upsilon=0}^{\infty} \frac{k\upsilon+n+1}{n} Y_\upsilon^{(n)}(t; k)\, x^\upsilon \int_{\upsilon k}^{\upsilon k+\frac{n}{k\upsilon+n+1}} f\left(\frac{u}{k(\upsilon-1)+n+1}\right) du \tag{2}$$

where $f \in M[0, b]$ $(0 < b < 1)$, $x \in [0, 1)$, $t \in (-\infty, 0]$ and $k < n + 1$.

Remark 1.3 Notice that this modification involves the Kantorovich type generalization of the operators P_n and M_n.

The q-type generalization of the linear positive operators was initiated by Phillips in [31]. He introduced the q-type generalization of the classical Bernstein operators and obtained the rate of convergence and the Voronovskaja type asymptotic formula for these operators.

q-Laguerre polynomials were defined by Hahn [15], p. 29; Jackson [19], p. 57; Moak [29], p. 21, Eq. 23 as

$$L_n^{(\alpha)}(x; q) = \frac{\left(q^{\alpha+1}; q\right)_n}{(q; q)_n} \sum_{k=0}^{n} \frac{\left(q^{-n}; q\right)_k q^{\binom{k}{2}} (1-q)^k \left(q^{n+\alpha+1} x\right)^k}{\left(q^{\alpha+1}; q\right)_k (q; q)_k}.$$

Moak gave the following recurrence relation ([29], p. 29, eq. 4.14) and the generating function ([29], p. 29, eq. 4.17) for the q-Laguerre polynomials by

$$t L_{k-1}^{(\alpha+1)}(t; q) = [k+\alpha]\, q^{-\alpha-k} L_{k-1}^{(\alpha)}(t; q) - [k]\, q^{-\alpha-k} L_k^{(\alpha)}(t; q)$$

$$(Re\alpha > -1, k = 1, 2, \ldots),$$

and

$$F_\alpha(x, t) = \frac{\left(xq^{\alpha+1}; q\right)_\infty}{(x; q)_\infty} \sum_{m=0}^{\infty} \frac{q^{m^2+\alpha m} [-(1-q)\, xt]^m}{(q; q)_m \left(xq^{\alpha+1}; q\right)_m}$$

$$= \sum_{k=0}^{\infty} L_k^{(\alpha)}(t; q)\, x^k \quad (Re\alpha > 1). \tag{3}$$

Trif [34] defined the Meyer-König and Zeller operators based on q-integers as follows:

$$M_{n,q}(f;x)=\prod_{j=0}^{n}\left(1-q^{j}x\right)\sum_{k=0}^{\infty}f\left(\frac{[k]}{[k+n]}\right)\left[\begin{matrix}n+k\\k\end{matrix}\right]x^{k}$$

where $0\leq x<1$.

In [30], Özarslan defined the q-analog of $P_n f$ as follows:

$$P_{n,q}(f;x)=\frac{1}{F_n(x,t)}\sum_{k=0}^{\infty}f\left(\frac{[k]}{[k+n]}\right)L_k^{(n)}(t;q)\,x^{k}$$

where $x\in[0,1]$, $t\in(-\infty,0]$, $q\in(0,1]$ and $\{F_n(x,t)\}_{n\in\mathbb{N}}$ is the generating functions for the q-Laguerre polynomials. Since $L_k^{(n)}(0;q)=\left[\begin{matrix}n+k\\k\end{matrix}\right]$ and $F_n(x,0)=\prod_{j=0}^{n}\left(1-q^{j}x\right)$, then $M_{n,q}(f;x)$ is the special case of the operators $P_{n,q}(f;x)$.

Let us recall the concepts of q-differential and q-derivative, respectively.

For an arbitrary function $f(x)$, the q-differential is given as

$$d_q f(x)=f(qx)-f(x).$$

For an arbitrary function $f(x)$, the q-derivative is defined as

$$D_q f(x)=\frac{d_q f(x)}{d_q x}=\frac{f(qx)-f(x)}{(q-1)\,x}.$$

We mention some notations for q-calculus. Throughout the present article q will be a real number satisfying the inequality $0<q\leq 1$. For $n\in\mathbb{N}$,

$$[n]_q=[n]:=\begin{cases}(1-q^{n})/(1-q), & q\neq 1\\ n, & q=1\end{cases},$$

$$[n]_q!=[n]!:=\begin{cases}[n][n-1]\ldots[1], & n\geq 1\\ 1, & n=0\end{cases},$$

$$(\alpha;q)_n=\begin{cases}1 & ;n=0\\ (1-\alpha)(1-\alpha q)\ldots\left(1-\alpha q^{n-1}\right) & ;n\in\mathbb{N},\alpha\in\mathbb{C}\end{cases},$$

and

$$(a;q)_\infty=\prod_{j=0}^{\infty}\left(1-aq^{j}\right),\ (a\in\mathbb{C}).$$

For the integers $n, k, n \geq k \geq 0$, the q-polynomial coefficients are defined by

$$\begin{bmatrix} n \\ k \end{bmatrix} = \frac{[n]!}{[k]!\,[n-k]!}.$$

Now suppose that $0 < a < b$, $0 < q \leq 1$ and f is a real-valued function. The q-Jackson integral of f over the interval $[0, b]$ and a general interval $[a, b]$ are defined as (see [20])

$$\int_0^a f(t)\, d_q t = (1-q)\, a \sum_{j=0}^{\infty} f\left(q^j a\right) q^j$$

and

$$\int_a^b f(t)\, d_q t = \int_0^b f(t)\, d_q t - \int_0^a f(t)\, d_q t$$

respectively.

It is clear that q-Jackson integral of f over an interval $[a, b]$ contains two infinite sums, so some problems are encountered in deriving the q-analogs of some well-known integral inequalities which are used to compute the order of approximation of the linear positive operators containing the q-Jackson integral. In order to overcome these problems Gauchman [13] and Marinković et al. [26] introduced a new type of q-integral. This new q-integral is called Riemann type q-integral and is defined as

$$\int_a^b f(t)\, d_q^R t = (1-q)\,(b-a) \sum_{j=0}^{\infty} f\left(a + (b-a)\, q^j\right) q^j$$

where a, b and q are some real numbers such that $0 < a < b$ and $0 < q < 1$. Contrary to the classical definition of q-integral, this definition includes only points within the interval of integration.

Now, we describe a Kantorovich type generalization of operators P_n, M_n^*, $M_{n,q}$ and $P_{n,q}$. This Kantorovich type generalization was studied by Dalmanoğlu [9]; Radu [32] and etc. We consider the sequence of Kantorovich type linear positive operators as follows:

$$\left(K_{n,q} f\right)(x,t) = \frac{1}{F_{n,q}(x,t)} \sum_{k=0}^{\infty} \left(\int_{[k]/[n+k]}^{[k+1]/[n+k]} f(t)\, d_q^R t \right) q^{-k} [n+k]\, L_k^{(n)}(t;q)\, x^k \tag{4}$$

where $x \in [0,1]$, $t \in (-\infty, 0]$, $q \in (0,1]$, $n > 1$ and $\{F_n(x,t)\}_{n\in\mathbb{N}}$ is the generating functions for the q-Laguerre polynomials which was given in (3).

2 Approximation Properties of These Operators

Let b be a real number in the interval $(0, 1)$, $x \in [0, 1)$ and $t \in (-\infty, 0]$. We have the following theorem for the convergence of the sequence operators $\{L_n\}$.

Theorem 2.1 (Icoz, Tasdelen and Varma [16])
If f is continuous on $[0, b]$, $\frac{|t|}{n} \to 0$ then $\{L_n f\}$ converges to f uniformly on $[0, b]$.

Now, we state the following theorem for the convergence of $K_{n,q} f$ operators.

Theorem 2.2 (Icoz, Varma and Tasdelen [18])
Let $q := q_n$ be a sequence satisfying $\lim_n q_n = 1$ and $0 < q_n < 1$. If $f \in C[0, 1]$ and $\frac{|t|}{[n]} \to 0$ $(n \to \infty)$ then $\left\{K_{n,q} f\right\}$ converges to f uniformly on $[0, b]$ $(0 < b < 1)$.

Finally, in order to obtain uniform convergence of the linear operators L_n^*, we will use the classical Bohman-Korovkin theorem (see [5] and [25]).

Theorem 2.3 (Icoz, Tasdelen and Dogru [17])
If f is continuous on $[0, b]$ and $\frac{|t|}{n} \to 0$ then $\left\{L_n^*\right\}$ converges to f uniformly on $[0, b]$.

3 Rates of Convergence

In this section, we compute the rates of convergence for these operators by means of the modulus of continuity, Peetre's K-functional, elements of Lipschitz class, and the second order modulus of smoothness.

Let $f \in C[0, b]$. The modulus of continuity of f denotes by $\omega(f, \delta)$, is defined to be

$$\omega(f, \delta) = \sup_{\substack{s,x \in [0,b] \\ |s-x| < \delta}} |f(s) - f(x)|.$$

It is well known that a necessary and sufficient condition for a function $f \in C[0, b]$ is

$$\lim_{\delta \to 0} \omega(f, \delta) = 0.$$

It is also well known that for any $\delta > 0$ and each $s \in [0, b]$

$$|f(s) - f(x)| \leq \omega(f, \delta) \left(1 + \frac{|s - x|}{\delta}\right).$$

The following theorem gives the rate of convergence of the operator $L_n f$ to the function f by means of modulus of continuity.

Theorem 3.1 (Icoz, Tasdelen and Varma [16])
For all $f \in C[0, b]$, we have

$$\|L_n(f;x) - f(x)\|_{C[0,b]} \leq \left(1 + (3\gamma)^{\frac{1}{2}}\right) \omega\left(f;\delta_n^{**}\right) \tag{5}$$

where

$$\delta_n^{**} = \frac{1}{\sqrt{n}} \text{ and } \gamma = \max\left\{kb, k|t|b, 3|t|b^2\right\}. \tag{6}$$

Our next theorem gives the rates of convergence of the sequence $\left\{L_n^* f\right\}$ to f by means of the modulus continuity of f.

Theorem 3.2 (Icoz, Tasdelen and Dogru [17])
If $f \in C[0,b]$ $(0 < b < 1)$ *and* $\frac{|t|}{n} \to 0$ $(n \to \infty)$ *then we have*

$$\left\|\left(L_n^* f\right)(\cdot, t;k) - f(\cdot)\right\|_{C[0,b]} \leq 2\omega\left(f, \delta_n^*\right) \tag{7}$$

where

$$\delta_n^* = \sqrt{\frac{|t|b}{n(1-b)}\left(k + 3b + \frac{1}{n}\right) + (k+2)\frac{b}{n} + \frac{1}{3n^2}}. \tag{8}$$

Before mentioning the theorem on the rate of convergence of the operator $K_{n,q} f$ to f, let us recall an inequality on $C[0,b]$ ([18], page 92):

$$\left\|\left(K_{n,q}(e_1 - x)^2\right)(x,t)\right\|_{C[0,b]} \leq \frac{|t|\left(3b^2 + b\right)}{[n]\left(1 - bq^{n+1}\right)} + \frac{2|t|b}{[2][n]^2\left(1 - bq^{n+1}\right)} + \left(1 + \frac{4}{[2]}\right)\frac{b}{[n]} + \frac{1}{[3][n]^2}. \tag{9}$$

The following theorem gives the rate of convergence of the operator $K_{n,q} f$ to the function f by means of modulus of continuity:

Theorem 3.3 (Icoz, Varma and Tasdelen [18])
Let $q := q_n$ *be a sequence satisfying* $\lim_n q_n = 1$ *and* $0 < q_n < 1$. *For all* $f \in C[0,b]$ *and* $\frac{|t|}{[n]} \to 0$ $(n \to \infty)$

$$\left\|\left(K_{n,q} f\right)(x,t) - f(x)\right\|_{C[0,b]} \leq 2\omega(f, \delta_n) \tag{10}$$

where

$$\delta_n = \left[\frac{|t|\left(3b^2 + b\right)}{[n]\left(1 - bq^{n+1}\right)} + \frac{2|t|b}{[2][n]^2\left(1 - bq^{n+1}\right)} + \left(1 + \frac{4}{[2]}\right)\frac{b}{[n]} + \frac{1}{[3][n]^2}\right]^{1/2}. \tag{11}$$

Let $f \in C[0,b]$ and $0 < \alpha \leq 1$. We recall that f belongs to $Lip_M(\alpha)$ if the inequality

$$|f(\zeta) - f(\eta)| \leq M|\zeta - \eta|^{\alpha}; \zeta, \eta \in [0, b]$$

holds.

We will now study the rate of convergence of the positive linear operators L_n means of the Lipschitz class $Lip_M(\alpha)$, for $0 < \alpha \leq 1$.

Theorem 3.4 (Icoz, Tasdelen and Varma [16])
For all $f \in Lip_M(\alpha)$, we have

$$\|L_n(f;x) - f(x)\|_{C[0,b]} \leq M(3\gamma)^{\frac{\alpha}{2}} \left(\delta_n^{**}\right)^{\alpha} \tag{12}$$

where γ and δ_n^{**} are given by (6).

Next, we mention the approximation order of operator $K_{n,q}f$ in term of the elements of the usual Lipschitz class.

Theorem 3.5 (Icoz, Varma and Tasdelen [18])
Let $q := q_n$ be a sequence satisfying $\lim_n q_n = 1$ and $0 < q_n < 1$. For all $f \in Lip_M(\alpha)$ and $\frac{|t|}{[n]} \to 0$ $(n \to \infty)$

$$\left\|\left(K_{n,q}f\right)(x,t) - f(x)\right\|_{C[0,b]} \leq M\delta_n^{\alpha} \tag{13}$$

where δ_n is given by (11).

Next, we obtain the rates of convergence of the sequence L_n^*f to f by means of the elements of the Lipschitz class $Lip_M(\alpha)$, for $0 < \alpha \leq 1$.

Theorem 3.6 (Icoz, Tasdelen and Dogru [17])
If $f \in Lip_M(\alpha)$ $(0 < \alpha < 1)$ and $\frac{|t|}{n} \to 0$ $(n \to \infty)$ then we have

$$\left\|\left(L_n^*f\right)(x,t;k) - f(x)\right\|_{C[0,b]} \leq M\left(\delta_n^*\right)^{\alpha} \tag{14}$$

where δ_n^* is given by (8).

Let $C^2[0,b] := \left\{g \in C[0,b] : g', g'' \in C[0,b]\right\}$. Similarly in [4], we define the following norm in the space $C^2[0,b]$:

$$\|f\|_{C^2[0,b]} := \|f\|_{C[0,b]} + \left\|f'\right\|_{C[0,b]} + \left\|f''\right\|_{C[0,b]}.$$

For any $\delta > 0$, the Peetre's K-functional is defined by

$$K_2(\varphi;\delta) = \inf_{g \in C^2[0,b]} \left\{\|\varphi - g\| + \delta\left\|g''\right\|\right\}$$

where $\|.\|$ is the uniform norm on $C[0,b]$ (see [14]). From [10] (p.177, Theorem 2.4), there exists an absolute constant $C > 0$ such that

$$K_2(f;\delta) \leq C\omega_2\left(f;\sqrt{\delta}\right)$$

where the second order modulus of smoothness of $f \in C[0,b]$ is denoted by

$$\omega_2(f;\delta) = \sup_{0<h\leq\delta} \sup_{x,x+2h\in[0,b]} |f(x+2h) - 2f(x+h) + f(x)|.$$

We recall the usual modulus of continuity of $f \in C[0,b]$ by

$$\omega(f;\delta) = \sup_{0<h\leq\delta} \sup_{x,x+h\in[0,b]} |f(x+h) - f(x)|.$$

We have the following result:

Theorem 3.7 (Icoz, Tasdelen and Varma [16])
If $f \in C[0,b]$ then we have

$$\|L_n(f;x) - f(x)\|_{C[0,b]} \leq 2K\left(f, \varepsilon_n^{**}\right) \tag{15}$$

where the operators L_n are defined by (1) and

$$\varepsilon_n^{**} = \frac{2|t|\,b + kb + k|t|\,b + 3b^2|t|}{4n}. \tag{16}$$

Note that for each fixed t, $\varepsilon_n^{**} \to 0$, when $n \to \infty$.

Next, let us consider the following operator:

$$\left(L_{n,q}f\right)(x,t) = \left(K_{n,q}f\right)(x,t) - f\left(x - \frac{tx}{[n]\left(1-bq^{n+1}\right)} + \frac{1}{[2][n]}\right) + f(x) \tag{17}$$

for $x \in [0,1]$. We need the following lemma to prove Theorem 3.8.

Lemma 3.1 (Icoz, Varma and Tasdelen [18])
Let $g \in C^2[0,1]$. Then we have

$$\left|\left(L_{n,q}g\right)(x,t) - g(x)\right| \leq \left\{\frac{-t\left(3x^2+x\right)}{[n]\left(1-bq^{n+1}\right)} - \frac{2tx}{[2][n]^2\left(1-bq^{n+1}\right)} + \left(1+\frac{4}{[2]}\right)\frac{x}{[n]} + \frac{1}{[3][n]^2} + \left(\frac{-tx}{[n]\left(1-bq^{n+1}\right)} + \frac{1}{[2][n]}\right)^2\right\}\|g''\|. \tag{18}$$

The next result establishes a local approximation theorem for the operator $K_{n,q}f$.

Theorem 3.8 (Icoz, Varma and Tasdelen [18])
Let $q := q_n$ be a sequence satisfying $\lim_n q_n = 1$ and $0 < q_n < 1$. For each $f \in C[0,1]$ and $x \in [0,1]$, we have

$$\left|\left(K_{n,q}f\right)(x,t)-f(x)\right|\leq C\omega_2\left(f;\sqrt{\varepsilon_n(x)}\right)+\omega\left(f;\left|\frac{-tx}{[n]\left(1-bq^{n+1}\right)}+\frac{1}{[2][n]}\right|\right) \tag{19}$$

where

$$\varepsilon_n(x)=\frac{-t\left(3x^2+x\right)}{[n]\left(1-bq^{n+1}\right)}-\frac{2tx}{[2][n]^2\left(1-bq^{n+1}\right)}+\left(1+\frac{4}{[2]}\right)\frac{x}{[n]}$$
$$+\frac{1}{[3][n]^2}+\left(\frac{-tx}{[n]\left(1-bq^{n+1}\right)}+\frac{1}{[2][n]}\right)^2 \tag{20}$$

and C is a positive constant.

Now, we compute the rate of convergence of the sequence L_n^*f to f by means of the Peetre's K-functional.

Theorem 3.9 (Icoz, Tasdelen and Dogru [17])
If $f\in C[0,b]$ and $\frac{|t|}{n}\to 0\,(n\to\infty)$ then we have

$$\left\|\left(L_n^*f\right)(\cdot,t;k)-f(\cdot)\right\|_{C[0,b]}\leq 2K\left(f;\varepsilon_n^*\right) \tag{21}$$

where

$$\varepsilon_n^*=\tfrac{3|t|b^2}{2n(1-b)}+\tfrac{k|t|b}{2n(1-b)}+\tfrac{|t|b}{2n^2(1-b)}+\tfrac{(k+1)b}{2n}+\tfrac{b}{2n}+\tfrac{1}{6n^2}. \tag{22}$$

4 A Generalization of r-th Order for These Operators

By $C^r[0,b]$ $(0<b<1, r=0,1,2,...)$ we denote the set of functions f having continuous $r-th$ derivatives $f^{(r)}\left(f^{(0)}(x)=f(x)\right)$ on the segment $[0,b]$.

We consider the following generalization of the positive linear operators L_n:

$$L_n^{[r]}(f;x)=\frac{1}{F_n(x;t)}\sum_{\upsilon=0}^{\infty}\sum_{i=0}^{r}f^{(i)}\left(\frac{\upsilon k}{k(\upsilon-1)+n+1}\right)\frac{\left(x-\frac{\upsilon k}{k(\upsilon-1)+n+1}\right)^i}{i!}Y_\upsilon^n(t;k)\,x^\upsilon \tag{23}$$

where $f\in C^r[0,b]$, $r=0,1,2,...$ and $n\in\mathbb{N}$. We call the operators above the r-th order of the operators L_n, (for instance, [21, 22]). Note that when $r=0$, we get the sequence of operators $\{L_n\}$.

Theorem 4.1 (Icoz, Tasdelen and Varma [16])
If $f^{(r)}\in Lip_M(\alpha)$ and $f\in C^r[0,b]$ then we have

$$\left\|L_n^{[r]}(f;x)-f(x)\right\|_{C[0,b]}\leq\frac{M}{(r-1)!}\frac{\alpha}{\alpha+r}B(\alpha,r)\left\|L_n\left(|s-x|^{\alpha+r};x\right)\right\|_{C[0,b]} \tag{24}$$

where $B(\alpha, r)$ is the beta function and $r, n \in \mathbb{N}$.

Now, consider the function $g \in C[0, b]$ defined as

$$g(s) = |s - x|^{\alpha + r}. \tag{25}$$

Since $g(x) = 0$, Theorem 2.1 yields

$$\lim_n \|L_n(g;x)\|_{C[0,b]} = 0. \tag{26}$$

So, it follows from Theorem 4.1 that, for all $f \in C^r[0, b]$ such that $f^{(r)} \in Lip_M(\alpha)$, we have

$$\lim_n \left\|L_n^{[r]}(f;x) - f(x)\right\|_{C[0,b]} = 0. \tag{27}$$

Finally, taking into consideration Theorems 3.1 and 3.4 with $M = b^r$ and observing $g \in Lip_{b^r}(\alpha)$ one can deduce the following results from Theorem 4.1 immediately.

Corollary 4.1 (Icoz, Tasdelen and Varma [16])
For all $f \in C^r[0, b]$ such that $f^{(r)} \in Lip_M(\alpha)$, we have

$$\left\|L_n^{[r]}(f;x) - f(x)\right\|_{C[0,b]} \leq \frac{M}{(r-1)!}\frac{\alpha}{\alpha + r} B(\alpha, r)\left(1 + (3\gamma)^{\frac{1}{2}}\right)\omega\left(g;\delta_n^{**}\right) \tag{28}$$

where δ_n^{**} and γ are the same as in Theorem 3.1 and g is defined by (25).

Corollary 4.2 (Icoz, Tasdelen and Varma [16])
For all $f \in C^r[0, b]$ such that $f^{(r)} \in Lip_M(\alpha)$, we have

$$\left\|L_n^{[r]}(f;x) - f(x)\right\|_{C[0,b]} \leq \frac{Mb^r}{(r-1)!}\frac{\alpha}{\alpha + r} B(\alpha, r)(3\gamma)^{\frac{\alpha}{2}}\left(\delta_n^{**}\right)^{\alpha} \tag{29}$$

where δ_n^{**} and γ are the same as in Theorem 3.1.

The last two results give us the rates of convergence of the sequence $\left\{L_n^{[r]} f\right\}$ to f by means of the modulus of continuity and the elements of the Lipschitz class $Lip_M(\alpha)$, respectively.

Similarly in [21] (see also [1, 11, 22]), we have considered the following generalization of the positive linear operators L_n^* defined by (2)

$$\begin{aligned}\left(L_{n,r}^* f\right)(x,t;k) = {} & \frac{1}{F_n(xt)} \sum_{\upsilon=0}^{\infty} \frac{k\upsilon+n+1}{n} Y_\upsilon^{(n)}(t;k)\, x^\upsilon \\ & \times \int_{\upsilon k}^{\upsilon k + \frac{n}{k\upsilon+n+1}} \sum_{j=0}^{r} f^{(j)}\left(\frac{u}{k(\upsilon-1)+n+1}\right) \frac{\left(x - \frac{u}{k(\upsilon-1)+n+1}\right)^j}{j!} du\end{aligned} \tag{30}$$

where $f \in C^r[0,b]$, $r = 0, 1, 2, \ldots$ and $n \in \mathbb{N}$. Note that taking $r = 0$, we get a result for L_n^* defined in (2).

Theorem 4.2 (Icoz, Tasdelen and Dogru [17])
If $f \in C^r[0,b]$ and $f^{(r)} \in Lip_M(\alpha)$, then we have

$$\left\| \left(L_{n,r}^* f\right)(x,t;k) - f(x) \right\|_{C[0,b]} \leq \frac{M}{(r-1)!} \frac{\alpha}{\alpha + r} B(\alpha, r) \times \left\| \left(L_{n,r}^* |s-x|^{\alpha+r}\right)(x,t;k) \right\|_{C[0,b]} \quad (31)$$

where $B(\alpha, r)$ is the Beta function and $r, n \in \mathbb{N}$.

Now, consider the function $g \in C[0,b]$ defined by (25). Since $g(x) = 0$, Theorem 2.1 yields $\lim_{n\to\infty} \|(L_n g)(x,t;k)\|_{C[0,b]} = 0$. So, from Theorem 4.2, for all $f \in C^r[0,b]$ such that $f^{(r)} \in Lip_M(\alpha)$, we have

$$\lim_{n\to\infty} \left\| \left(L_{n,r}^* f\right)(\cdot,t;k) - f(\cdot) \right\|_{C[0,b]} = 0. \quad (32)$$

5 An Application to Differential Equations

In this section, we mention a result on functional differential equation for $L_n(f;x)$ defined in (1). This equation seems to be fundamental for the investigation of many kinds of linear positive operators. In May [27], Volkov [35] and Alkemade [2], there are equations similar to the equation mentioned in Theorem 5.1

Theorem 5.1 (Icoz, Tasdelen and Varma [16])
Let $g(s) = \frac{s}{1-s}$. For each $x \in [0,b]$ $(0 < b < 1)$ and $f \in C[0,b]$, $L_n(f;x)$ as defined in (1), satisfies the functional differential equation

$$x \frac{d}{dx} L_n(f;x) = -x \frac{n + 1 - t(1-x)^{-\frac{1}{k}}}{k(1-x)} L_n(f;x) + \frac{n+1-k}{k} L_n(fg;x). \quad (33)$$

Remark 5.1 These results show how using q-calculus we can obtain many new operators and study their degree of convergence.

References

1. Agratini, O.: Korovkin type error estimates for Meyer-König and Zeller operators. Math. Inequal. Appl. **4**(1), 119–126 (2001)
2. Alkemade, J.A.H.: The second moment for the Meyer-König and Zeller operators. J. Approx. Theor. **40**, 261–273 (1984)

3. Altın, A., Doğru, O., Özarslan, M.A.: Kantorovich type generalization of certain class of positive linear operators. WSEAS Trans. Math. **3**(3), 607–610 (2004)
4. Bleimann, G., Butzer, P.L., Hahn, L.: A Bernstein-type operator approximating continuous functions on the semi-axis. Nederl. Akad. Wetensch Indag. Math. **42**, 255–262 (1980)
5. Bohman, H.: On approximation of continuous and of analytic functions. Ark. Mat. **2**, 43–56 (1952)
6. Carlitz, L.: A note on certain biorthogonal polynomials. Pacific J. Math. **24**, 425–430 (1968)
7. Cheney, E.W., Sharma, A.: Bernstein power series. Canad. J. Math. **16**, 241–252 (1964)
8. Dalmanoğlu, Ö., Doğru, O.: Statistical approximation properties of Kantorovich type q-MKZ operators. Creative Math. Inf. **1**(19), 15–24 (2010)
9. Dalmanoğlu, Ö., Doğru, O.: On statistical approximation properties of Kantorovich type q-Bernstein operators. Math. Comput. Model. **52**, 760–771 (2010)
10. DeVore, R.A., Lorentz, G.G.: Constr. Approx. Springer, Berlin (1993)
11. Doğru, O.: Approximation order and asymptotic approximation for generalized Meyer-König and Zeller operators. Math. Balkanica. **12**(3–4), 359–368 (1998)
12. Doğru, O., Özarslan, M.A., Taşdelen, F.: On positive operators involving a certain class of generating functions. Stud. Sci. Math. Hung. **41**(4), 415–429 (2004)
13. Gauchman, H.: Integral inequalities in q-calculus. Comput. Math. Appl. **47**, 281–300 (2004)
14. Gupta, V., Finta, Z.: On certain q-Durrmeyer type operators. Appl. Math. Comput. **209**, 415–420 (2009)
15. Hahn, W.: Über orthogonal polynome die q-differenzengleichungen genügen. Math. Nach. **2**, 4–34 (1949)
16. İçöz, G., Taşdelen, F., Varma, S.: On linear positive operators involving biorthogonal polynomial. Ars Combinatoria. **105**, 319–331 (2012)
17. İçöz, G., Taşdelen, F., Doğru, O.: Kantorovich process of linear positive operators via biorthogonal polynomials. J. Inequal. Appl. Spec. Funct. **3**(4), 77–84 (2012)
18. İçöz, G., Varma, S., Taşdelen, F.: Integral type modification for q-Laguerre polynomials. Bull. Math. Anal. Appl. **4**(3), 87–98 (2012)
19. Jackson, F.H.: Basic double hypergeometric functions (II). Quart. J. Math. Oxf. **15**, 49–51 (1944)
20. Kac, V.G., Cheung, P.: Quantum Calculus. Universitext. Springer, New York (2002)
21. Kirov, G.H., Popova, L.: A generalization of the linear positive operators. Math. Balkanica. **7**, 149–162 (1993)
22. Kirov, G.H.: Approximation with Quasi-Splines. Inst. Physics Publ., Bristol, New York (1992)
23. Konhauser, J.D.E.: Some properties of biorthogonal polynomials. J. Math. Anal. Appl. **11**(1–3), 242–260 (1965)
24. Konhauser, J.D.E.: Biorthogonal polynomials suggested by the Laguerre polynomials. Pacific J. Math. **21**, 303–314 (1967)
25. Korovkin, P.P.: On convergence of linear positive operators in the space of continuous functions. Doklady Akad. Nauk SSSR. **90**, 961–964 (1953)
26. Marinković, S., Rajković, P., Stanković, M.: The inequalities for some types of q-integrals. Comput. Math. Appl. **56**, 2490–2498 (2008)
27. May, C.P.: Saturation and inverse theorems for combinations of a class of exponential-type operators. Canad. J. Math. **28**, 1224–1250 (1976)
28. Meyer-König, W., Zeller, K.: Bernsteinsche potenzreihen. Stud. Math. **19**, 89–94 (1960)
29. Moak, D.S.: The q-analogue of the Laguerre polynomials. J. Math. Anal. Appl. **81**, 20–47 (1981)
30. Özarslan, M.A.: q-Laguerre type linear positive operators. Stud. Sci. Math. Hung. **44**(1), 65–80 (2007)
31. Phillips, G.M.: On generalized Bernstein polynomials. In: Watson, G.A. (ed.) Numerical Analysis (A. R. Mitchell 75th Birthday Volume), pp. 263–269. World Science, Singapore (1996)
32. Radu, C.: Statistical approximation properties of Kantorovich operators based on q-integers. Creative Math. Inf. **17**(2), 75–84 (2008)

33. Srivastava, H.M.: Some biorthogonal polynomials suggested by the Laguerre polynomials. Pac. J. Math. **98**(1), 235–250 (1982)
34. Trif, T.: Meyer-König and Zeller operators based on the q-integers. Rev. Anal. Numer. Theor. Approx. **2**(29), 221–229 (2000)
35. Volkov, Y.: Certain positive linear operators. Mat. Zametki. **23**, 363–368 (1978)

Chapter 14
Similarity-Based Reasoning Fuzzy Systems and Universal Approximation

Sayantan Mandal and Balasubramaniam Jayaram

Abstract In this work, we show that fuzzy inference systems (FIS) based on similarity-based reasoning (SBR), where the modification function is a fuzzy implication, is a universal approximator under suitable conditions on the other components of the fuzzy system.

Keywords Similarity-based reasoning · Fuzzy implications · Universal approximation

1 Introduction

The term *approximate reasoning* (AR) refers to methods and methodologies that enable reasoning with imprecise inputs to obtain meaningful outputs [5]. AR schemes involving fuzzy sets are one of the best known applications of fuzzy logic in the wider sense. Fuzzy inference systems (FIS) have many degrees of freedom, viz., the underlying fuzzy partition of the input and output spaces, the fuzzy logic operations employed, the fuzzification and defuzzification mechanism used, etc. This freedom gives rise to a variety of FIS with differing capabilities. One of the important factors considered while employing an FIS is its approximation capability. Many studies have appeared on this topic and due to space constraints, we only refer the readers to the following exceptional review on this topic [14], or the recent work dealing with the approximation capabilities of implicative models of fuzzy relational inference mechanisms [11] and the references therein.

S. Mandal · B. Jayaram (✉)
Department of Mathematics, Indian Institute of Technology Hyderabad,
Yeddumailaram 502205, India
e-mail: jbala@iith.ac.in

S. Mandal
e-mail: ma10p002@iith.ac.in; saayaantaan17@gmail.com

R. N. Mohapatra et al. (eds.), *Mathematics and Computing 2013*,
Springer Proceedings in Mathematics & Statistics 91,
DOI: 10.1007/978-81-322-1952-1_14,

In this work, we consider a similarity-based reasoning (SBR) FIS where similarity between the inputs and the antecedents is used to subsequently modify the consequents to obtain a final output. Such inference schemes are also known as plausible reasoning scheme [6]. After detailing the inference mechanism in an SBR, we show that when the modification functions are modeled based on fuzzy implications, under suitable conditions on the other components of an SBR, the FIS based on SBR becomes a universal approximator, i.e., can approximate a continuous function over a compact set to arbitrary accuracy. Also, we deal only with single variable functions, alternately where the rule base consists of single input–single output (SISO) rules.

2 Preliminaries

We assume that the reader is familiar with the classical results concerning fuzzy set theory and basic fuzzy logic connectives, but to make this work more self-contained, we introduce some notations, concepts, and results employed in the rest of the work.

2.1 Fuzzy Sets

If X is a nonempty set we denote by $\mathcal{F}(X)$ the fuzzy power set of X, i.e., $\mathcal{F}(X) = \{A | A : X \to [0, 1]\}$.

Definition 1 A fuzzy set A is said to be

- *normal* if there exists an $x \in X$ such that $A(x) = 1$,
- *convex* if X is a linear space and for any $\lambda \in [0, 1]$, $x, y \in X$, $A(\lambda x + (1-\lambda)y) \geq \min\{A(x), A(y)\}$.

Definition 2 For an $A \in \mathcal{F}(X)$, the *Support*, *Height*, *Kernel* and *Ceiling* of A are denoted, respectively, as Supp A, Hgt A, Ker A and Ceil A and are defined as

$$\begin{aligned} \text{Supp } A &= \{x \in X | A(x) > 0\}, \\ \text{Hgt } A &= \sup\{A(x) | x \in X\}, \\ \text{Ker } A &= \{x \in X | A(x) = 1\}, \\ \text{Ceil } A &= \{x \in X | A(x) = \text{Hgt } A\}. \end{aligned}$$

A is said to be *bounded* if Supp A is a bounded set. Note that for a normal fuzzy set Ker A = Ceil A.

We denote the space of fuzzy sets which are bounded, normal, convex, and continuous as $\mathcal{F}_{BNCC}(X)$. Clearly, $\mathcal{F}_{BNCC}(X) \subseteq \mathcal{F}(X)$.

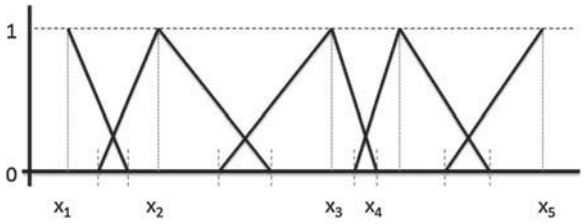


Fig. 1 An illustrative example for $\frac{1}{3}$-type partition in Definition 5

Definition 3 Let $\mathcal{P}$ be an arbitrary collection of fuzzy sets of X, i.e., $\mathcal{P} = \{A_k\}_{k=1}^n \subseteq \mathcal{F}(X)$. $\mathcal{P}$ is said to form a *fuzzy partition* on X if

$$X \subseteq \bigcup_{k=1}^{n} \text{Supp } A_k.$$

In the literature, a partition $\mathcal{P}$ of X as defined above is also called a **complete** partition.

Definition 4 A fuzzy partition $\mathcal{P} = \{A_k\}_{k=1}^n \subseteq \mathcal{F}(X)$ is said to be

- **consistent** if $A_k(x) = 1$ then $A_j(x) = 0$ for any $j \neq k$.
- **Ruspini partition** if $\sum_{k=1}^{n} A_k(x) = 1$ for every $x \in X$.

Definition 5 Let $\{x_k\}_{k=1}^n$ be a classical partition of X, i.e., $X = \bigcup_{k=1}^{n-2}[x_k, x_{k+1}) \cup [x_{n-1}, x_n]$. If $\mathcal{P} = \{A_k\}_{k=1}^n$ is a fuzzy partition of the space X in such a way that

- each A_k is normal at $x_k \in X$, i.e., $A_k(x_k) = 1$,
- Supp $A_k = \left(x_{k-1} + \frac{x_k - x_{k-1}}{3}, x_{k+1} - \frac{x_{k+1} - x_k}{3}\right)$ for $k = 2, \ldots, n-1$, while Supp $A_1 = \left(x_1, x_2 - \frac{x_2 - x_1}{3}\right)$ and Supp $A_n = \left(x_{n-1} + \frac{x_n - x_{n-1}}{3}, x_n\right)$,

we call this type of partition as $\frac{1}{3}$**-type partition**.

For instance, see Fig. 1 for $n = 5$.

2.2 Defuzzification

Often there is a need to convert a fuzzy set into a crisp value, a process which is called *Defuzzification*. This process of defuzzification can be seen as a mapping $g : \mathcal{F}(X) \longrightarrow X$. There are many types of defuzzification techniques available in the literature, see [13] for a good overview. In this work, we use the following defuzzifier extensively.

Example 1 For an $A \in \mathcal{F}(X)$, the *First of Maxima* (FOM) defuzzifier gives as output the smallest of all those values in X with the highest membership value that can be mathematically expressed as

$$\text{FOM}(A) = \min \left\{ x | A(x) = \max_{w} A(w) \right\}. \tag{1}$$

Similarly the *Last of Maxima* (LOM) defuzzifier is defined as

$$\text{LOM}(A) = \max \left\{ x | A(x) = \max_{w} A(w) \right\}. \tag{2}$$

2.3 Fuzzy Logic Connectives

Definition 6 ([7]) A binary operation $T : [0, 1]^2 \to [0, 1]$ is called a *t-norm*, if it is increasing in both variables, commutative, associative and has 1 as the neutral element.

Definition 7 ([1]) A function $I : [0, 1]^2 \to [0, 1]$ is called a *fuzzy implication* if it is decreasing in the first variable, increasing in the second variable, and $I(0, 0) = 1$, $I(1, 1) = 1$, $I(1, 0) = 0$. The set of all fuzzy implications are denoted by $\mathcal{I}$.

Definition 8 ([1]) A fuzzy implication $I : [0, 1]^2 \to [0, 1]$ is said to

- satisfy the *ordering property*, if

$$I(x, y) = 1 \iff x \leq y, \quad x, y \in [0, 1]. \tag{OP}$$

- be a *positive fuzzy implication* if $I(x, y) > 0$, for all $x, y \in (0, 1)$.

3 Fuzzy Inference Mechanism

Given two nonempty classical sets $X, Y \subsetneq \mathbb{R}$, a fuzzy single input–single output (SISO) IF-THEN rule is of the form:

$$\textbf{IF } \tilde{x} \text{ is } A \textbf{ THEN } \tilde{y} \text{ is } B, \tag{3}$$

where $\tilde{x}$, $\tilde{y}$ are the linguistic variables and $A \in \mathcal{F}(X)$, $B \in \mathcal{F}(Y)$ are the linguistic values taken by the linguistic variables. A knowledge base consists of a collection of such rules. Hence, we consider a rule base of n SISO rules which is of the form:

$$\textbf{IF } \tilde{x} \text{ is } A_i \textbf{ THEN } \tilde{y} \text{ is } B_i, \tag{4}$$

where $\tilde{x}$, $\tilde{y}$, and $A_i \in \mathcal{F}(X)$, $B_i \in \mathcal{F}(Y)$, $i = 1, 2, \ldots n$ are as mentioned above.

As an example, consider the rule

IF *Temperature* is *High* **THEN** *Fanspeed* is *Medium*.

Here *Temperature* and *Fanspeed* are the linguistic variables and *High*, *Medium* are the linguistic values taken by the linguistic variables in a suitable domain. Now given a single SISO rule (3) or a rule base (4) and given any input " $\tilde{x}$ is A'", the main objective of an inference mechanism is to find B' such that " $\tilde{y}$ is B' ". Many types of inference mechanisms are available to us in [2, 10, 17], etc. Here we consider only the case of similarity based reasoning.

4 Similarity-Based Reasoning

Consider the fuzzy if-then rule (3). Let the given input be $\tilde{x}$ is A'. Inference in SBR schemes in AR is based on the calculation of a measure of compatibility or similarity $M(A, A')$ of the input A' to the antecedent A of the rule, and the use of a modification function J to modify the consequent B, according to the value of $M(A, A')$.

Some of the well-known examples of SBR are compatibility modification inference (CMI) [4], "Approximate Analogical Reasoning Scheme" (AARS) in [15] and "Consequent Dilation Rule" (CDR) in [12], Smets and Magrez [10], Chen [3], etc. In this section, we detail the typical inferencing mechanism in SBR, but only in the case of SISO fuzzy rule bases.

4.1 Matching Function M

Given two fuzzy sets, say A, A', on the same domain, a matching function M compares them to get a degree of similarity, which is expressed as a real in the $[0, 1]$ interval. We refer to M as the matching function in the sequel. Formally, $M : \mathcal{F}(X) \times \mathcal{F}(X) \to [0, 1]$.

Example 2 Let X be a nonempty set and $A, A' \in \mathcal{F}(X)$. Below we list a few of the matching functions employed in the literature.

- Zadeh [18]: $M_{\mathbf{Z}}(A, A') = \max_{x \in X} \min(A(x), A'(x))$.
- Magrez–Smets [10]: Given a fuzzy negation N,

$$M_{\mathbf{M}}(A, A') = \max_{x \in X} \min(N(A(x)), A'(x)).$$

- Measure of Subsethood [12]: For an $I \in \mathcal{FI}$,

$$M_{\mathbf{S}}(A, A') = \min_{x \in X} I(A'(x), A(x)).$$

Definition 9 Let $\mathcal{F}^* \subseteq \mathcal{F}(X)$ be an arbitrary collection (not necessarily a fuzzy partition) of fuzzy sets on X. M is said to be **consistent with** $\mathcal{F}^*$ if for any $A \in \mathcal{F}^*$,

$$M(A, A) = 1. \tag{MCF}$$

Definition 10 Let $\mathcal{P} = \{A_k\}_{k=1}^n \subseteq \mathcal{F}^*$ be the given fuzzy partition of X. Let $A' \in \mathcal{F}^*$. M is said to be **consistent with** $\mathcal{P}$ (and $\mathcal{F}^*$) if

$$\sum_{k=1}^{n} M(A', A_k) \leq 1. \tag{MCP}$$

Definition 11 The matching function M is said to be **Strong** if

$$\text{Ker } A \subseteq \text{Ker } B \text{ or Ker } B \subseteq \text{Ker } A \Longrightarrow M(A, B) = 1 \tag{MS}$$

Example 3 Let $X \subsetneq \mathbb{R}$ be any bounded interval and $\mathcal{F}^* = \mathcal{F}_{BNCC}(X)$. For a given fuzzy partition $\mathcal{P} = \{A_k\}_{k=1}^n \subseteq \mathcal{F}_{BNCC}(X)$, we define a matching function as,

$$M_{\mathcal{P}}(A_k, A') = \frac{Area(A' \cap A_k)}{Area(A')}, \quad A' \in \mathcal{F}_{BNCC}(X). \tag{5}$$

Clearly M satisfies (MCF), (MCP), and (MS).

Example 4 Let $X \subsetneq \mathbb{R}$ be any bounded interval. Let the antecedent fuzzy sets $\{A_k\}_{k=1}^n = \mathcal{P}_X \subseteq \mathcal{F}^*(X)$ partition the input space X such that it forms a partition of the type defined in Definition 5.

Now, if $x' \in X$ is the input let $A' \in \mathcal{F}(X)$ be the fuzzified input such that A' attains normality at x', i.e., $A'(x') = 1$. Then the matching function defined as $M(A', A) = A(x')$ for any $A \in \mathcal{F}(X)$ has the property (MCP).

4.2 Modification Function J

Let A' be the fuzzy input and $s = M(A, A') \in [0, 1]$, a measure of the compatibility of A' to A.

The modification function J is again a function from $[0, 1]^2$ to $[0, 1]$ and, given the rule (3), modifies $B \in \mathcal{F}(Y)$ to $B' \in \mathcal{F}(Y)$ based on s, i.e., the consequent in SBR, using the modification function J, is given by

$$B'(y) = J(s, B(y)) = J(M(A, A'), B(y)), \quad y \in Y.$$

In AARS [15] the following modification operators have been used:

(i) $J_{\mathbf{ML}}(s, B) = B'(x) = \min\{1, B(x)/s\}, \ x \in X;$

(ii) $J_{\mathbf{MVR}}(s, B) = B'(x) = s \cdot B(x), x \in X.$

In CMI [4] and CDR [12] J is taken to be a fuzzy implication operator. In fact, $J_{\mathbf{ML}}(s, B) = I_{\mathbf{GG}}(s, B)$, where $I_{\mathbf{GG}}$ is the Goguen implication [1].

4.3 Aggregation Function G

In the case of multiple rules

$$R_i : \text{IF } \tilde{x} \text{ is } A_i \text{THEN } \tilde{y} \text{ is } B_i, \quad i = 1, 2, \ldots, m,$$

we infer the final output by aggregating over the rules, using an associative operator $G\colon [0, 1]^2 \to [0, 1]$ as follows:

$$B'(y) = G_{i=1}^{m}\Big(J\big(M(A_i, A'), B_i(y)\big)\Big), \quad y \in Y. \tag{6}$$

Usually, G is a t-norm, t-conorm, or a uninorm [7].

5 Fuzzy Systems $\mathcal{F}$ Based on SBR

An SBR fuzzy inference system can be represented by the hexatuple $\mathbb{F} = \{\mathcal{R}(A_i, B_j), f, M, J, G, g\}$ where

- $\mathcal{R}$ is the fuzzy if-then rule base formed from the fuzzy partitions $\{A_i\}, \{B_j\}$ on X, Y, respectively,
- $f : X \longrightarrow \mathcal{F}(X)$ is called the fuzzification mapping that maps an element $x \in X$ to a fuzzy set of $\mathcal{F}(X)$,
- M is matching function,
- J is modification function,
- G is aggregation function, and
- $g : \mathcal{F}(Y) \to Y$ is any defuzzifier, that converts the output fuzzy set into a crisp value $y \in Y$.

We consider $\mathbb{F}$ with the following assumptions on the different components/elements.

5.1 The Fuzzy Partitions A_i, B_j

Let $X, Y \subsetneq \mathbb{R}$ be arbitrary but fixed and let $\mathcal{F}^*(Z) = \mathcal{F}_{BNCC}(Z)$, where $Z = X$ or Y.

Let the antecedent fuzzy sets $\{A_k\}_{k=1}^n = \mathcal{P}_X \subseteq \mathcal{F}^*(X)$ partition the input space X such that it forms a partition of the type defined in Definition 5, which also implies it is complete.

Similarly, let the consequent fuzzy sets $\{B_j\}_{j=1}^m = \mathcal{P}_Y \subseteq \mathcal{F}^*(Y)$ form a complete and Ruspini partition of the output space Y.

5.2 The Fuzzified Input A'

Let us consider a fuzzification $f : X \longrightarrow \mathcal{F}^*(X)$ that maps $x' \in X$ to a fuzzy set of $A' \in \mathcal{F}^*(X) = \mathcal{F}_{BNCC}(X)$ such that

$$\text{Supp}\left(f(x') = A'\right) \cap \text{Supp } A_k \neq \emptyset,$$

for some $A_k \in \mathcal{P}_X$. Moreover, it is assumed that A' intersects only two of the adjacent fuzzy sets A_k, i.e., Supp $A' \cap$ Supp $A_k \neq \emptyset$ if and only if $k = m, m+1$ for some $m \in \mathbb{N}_{n-1}$.

Note that it is with this fuzzified input A' the antecedents A_i of the different rules are matched against.

Example 5 Let $\{x_k\}_{k=1}^n$ be a crisp partition of X. Let$\{A_k\}_{k=1}^n$ partitioning the input space X be such that $A_k \in \mathcal{P}_X$ and forms a fuzzy partition of the type defined in Definition 5. Then if we take

$$|\text{Supp } A'| \leq \frac{1}{3} \cdot \min_{i=1}^{l} \left\{|x_{i+1} - x_i|\right\},$$

then A' intersects at most two of the adjacent fuzzy sets A_k.

5.3 The Operations M, J, G

We choose a matching function M such that M is Consistent w.r.to the partition $\mathcal{P}_X$ given in Sect. 5.1, i.e., M satisfies both (MCP) and (MS).

We choose the modification function J to be a fuzzy implication, i.e., $J \in \mathcal{FI}$. For notational convenience we will denote it by " $\longrightarrow$" in the sequel.

The aggregation function G is any t-norm T.

5.4 The Fuzzy Output B'

With the above assumptions, the output fuzzy set B' for a given crisp input x' (or fuzzy input A') takes the form as given in the following lemma:

Lemma 1 *With the operations of the SBR FIS* (6) *as in* Sects. 5.1–5.3 *the fuzzy output of the SBR FIS* (6)*, for a given input* $x' \in X$ *is given by*

$$B'(y) = T\left[s_m \longrightarrow B_m(y), s_{m+1} \longrightarrow B_{m+1}(y)\right], \tag{7}$$

where $s_m = M\left(A', A_m\right)$ *and* $s_{m+1} = M\left(A', A_{m+1}\right)$.

Proof With the above operations M, J, G the fuzzy output for a given input $x' \in X$ is given by (6) as follows:

$$\begin{aligned} B'(y) = T\big[&M(A', A_1) \longrightarrow B_1(y)), M(A', A_2) \longrightarrow B_2(y), \ldots, \\ &\ldots, M(A', A_n) \longrightarrow B_n(y)\big]. \end{aligned}$$

We can write the above as

$$B'(y) = T_{k=1}^{n}[M(A', A_k) \longrightarrow B_k(y)]. \tag{8}$$

By the choice of our fuzzification based on our above notations on A', A_k, viz., that A' intersects only two adjacent fuzzy sets among the $\{A_k\}$, say A_m, A_{m+1}, we have that $M(A', A_k) = 0$ for all $k \neq m, m+1$. Note also that $I(0, y) = 0 \longrightarrow y = 1$ for any $y \in [0, 1]$. Now, the fuzzy output $B'(y)$ for any $y \in Y$ which is given by (8) becomes

$$\begin{aligned} B'(y) &= T_{k=1}^{n}\Big[M(A', A_k) \longrightarrow B_k(y)\Big], \\ &= T\Big[T_{k\neq m,m+1}\big(M(A', A_k) \longrightarrow B_k(y)\big), \\ &\qquad M\left(A', A_m\right) \longrightarrow B_m(y), M\left(A', A_{m+1}\right) \longrightarrow B_{m+1}(y)\Big] \\ &= T\Big[M\left(A', A_m\right) \longrightarrow B_m(y), M\left(A', A_{m+1}\right) \longrightarrow B_{m+1}(y)\Big] \\ &= T\left[s_m \longrightarrow B_m(y), s_{m+1} \longrightarrow B_{m+1}(y)\right] = (1.7). \end{aligned}$$

5.5 The Defuzzified Output $g(x')$

We have chosen the modification function J to be a fuzzy implication, i.e., $J = I \in \mathcal{FI}$. Assuming that the considered modification function J has (OP), we define the defuzzification function g appropriately so that g is continuous. In the following, we

discuss the explicit formulae for g. Note that g is also known as the system function of the fuzzy system $\mathbb{F}$ [8, 9].

6 SBR Fuzzy Systems and Universal Approximation

In this section, we show that $\mathbb{F} = \{\mathcal{R}(A_i, B_i), M, J, G, g\}$ such that the fuzzy partitions $\{A_k\}, \{B_k\}$ and the operations M, J, G, g as given in Sects. 5.1–5.5 are universal approximators, i.e., they can approximate any continuous function over a compact set to arbitrary accuracy.

Theorem 1 *For any continuous function $h : [a, b] \to \mathbb{R}$ over a closed interval and an arbitrary given $\epsilon > 0$, there is an SBR fuzzy system $\mathbb{F} = \{\mathcal{R}(A_i, B_i), M, J, G, g\}$ with M having the property* (MCP) *w.r.to $\mathcal{P}_X = \{A_i\}$, J having* (OP)*, G being a t-norm and g as given in* (11) *or* (12) *such that* $\max_{x \in [a,b]} |h(x) - g(x)| < \epsilon$.

Proof We prove this result in the following steps.

Step I : Choosing the points of normality

Since h is continuous over a closed interval $[a, b]$, h is uniformly continuous on $[a, b]$. Thus for a given $\epsilon > 0$ there exists $\delta > 0$ such that

$$\left|w - w'\right| < \delta \Longrightarrow \left|h(w) - h(w')\right| < \frac{\epsilon}{2}.$$

Step I *(a): A Coarse Initial Partition*

With the $\delta = \delta(\epsilon)$ defined above and taking $l = \left\lceil \frac{b-a}{\delta} \right\rceil$ we now choose $w_i \in X, i = 1, 2, \ldots l$, such that $|w_i - w_{i+1}| < \delta$.

Let $z_i = h(w_i)$, the value h takes at the above chosen w_i, for $i = 1, 2, \ldots l$. We call these points w_i and z_i the points of normality on the input space and the output space respectively.

In Fig. 2, *the points $w_1, w_2, \ldots, w_{11}$ and the points $z_1, z_2, \ldots z_8$ are the points of normality in the input and the output spaces, respectively.*

Step I *(b): Redundancy Removal and Reordering*

Let us choose the distinct z_i's from the above and sort them in ascending order. Let $\sigma : \mathbb{N}_l \longrightarrow \mathbb{N}_k$ denote the above permutation map such that $z_i = u_{\sigma(i)}$, for $i = 1, 2, \ldots l$ and u_j, $j = 1, 2, \ldots, k$ are in ascending order.

By rearranging the z_i's in ascending order and renaming them we have obtained: $u_1 = z_1, u_2 = z_8, u_3 = z_6, u_4 = z_5, u_5 = z_7, u_6 = z_2, u_7 = z_4, u_8 = z_3$.

Step I *(c): Refinement of the input space partition:*

Thus for each $i = 1, 2, \ldots, l$ we have $h(w_i) = z_i = u_{\sigma(i)}$. However, note that consecutive points of normality w_i, w_{i+1} in the input space need not be mapped to consecutive points of normality $u_{\sigma(i)}, u_{\sigma(i)+1}$ or $u_{\sigma(i)}, u_{\sigma(i)-1}$.

In Fig. 2, $h(w_1) = u_1$ *and* $h(w_2) = u_6$. *Thus for the consecutive points* w_1 and w_2 *the function values are* u_1 and u_6, *which are not consecutive.*

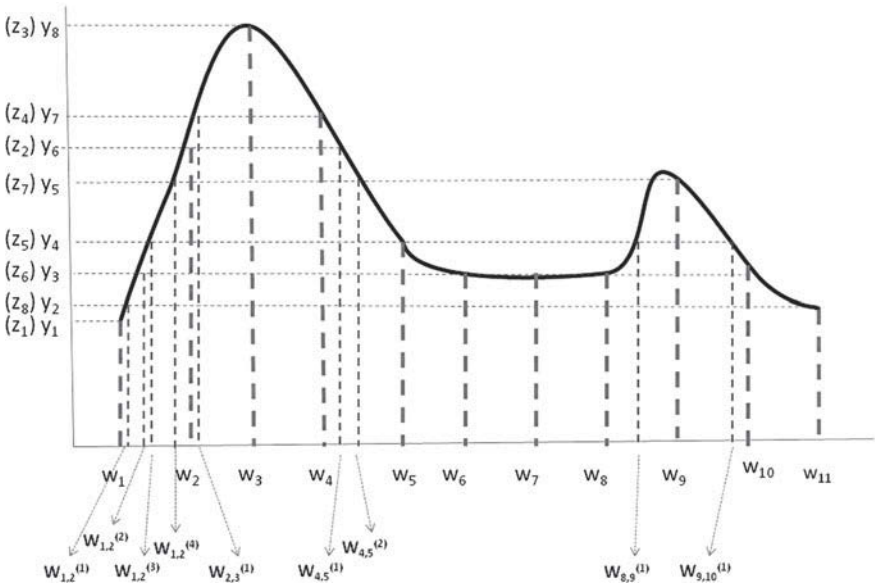


Fig. 2 An illustrative example for **Step I** in the proof of Theorem 1

To ensure the above, we further refine the input space partition. To this end, we refine every sub-interval $[w_i, w_{i+1}]$, for $i = 1, 2, \ldots l-1$ as follows. Note that $h(w_{i+1}) = u_{\sigma(i+1)}$.

Refinement Procedure:
For every $i = 1, 2, \ldots l-1$ do the following:

(i) If $u_{\sigma(i+1)} = u_{\sigma(i)+1}$ or $u_{\sigma(i)-1}$ then we do nothing.
(ii) Let $u_{\sigma(i+1)} = u_{\sigma(i)+p}$, where $p \geq 2$. For every $u \in \{u_{\sigma(i)+1}, u_{\sigma(i)+2}, \ldots, u_{\sigma(i)+p-1}\}$ we find a point $v \in [w_i, w_{i+1}]$ such that $h(v) = u$. Note that the existence of such a $v \in [w_i, w_{i+1}]$ is guaranteed by the continuity—essentially the ontoness—of the function h. If $u = u_{\sigma(i)+q}$, for some $1 \leq q \leq p-1$, then we denote the point v as $w_{i,i+1}^{(q)}$.
(iii) Similarly, let $u_{\sigma(i+1)} = u_{\sigma(i)-p}$, where $p \geq 2$. For every $u \in \{u_{\sigma(i)-1}, u_{\sigma(i)-2}, \ldots, u_{\sigma(i)-p+1}\}$ we find a $v \in [w_i, w_{i+1}]$ such that $h(v) = u$. Once again, if $u = u_{\sigma(i)-q}$, for some $1 \leq q \leq p-1$, then we denote v as $w_{i,i+1}^{(q)}$.

From Fig. 2, *it can be seen that we have inserted points* $w_{1,2}^1, w_{1,2}^2, w_{1,2}^3, w_{1,2}^4 \in [w_1, w_2]$. *Proceeding similarly, the following sub-intervals, shown in* Fig. 2, *have been refined:* $[w_2, w_3]$, $[w_4, w_5]$, $[w_8, w_9]$, *and* $[w_9, w_{10}]$.

Step I *(d): Final Points of Normality:*

Once the above process is done, we again rename the points of normality $w_{i,i+1}^{(q)}$ in the input space X in ascending order as $x_1, x_2, \ldots, x_n (n \geq l)$ and the $u_{\sigma(i)}$'s of the output space as $y_1, y_2, \ldots y_k$.

Step II : Construction of the Fuzzy Partitions

In the next step, we construct fuzzy sets on both the input and output spaces with the above obtained x_i's and y_j's as the points of normality, as given below.

Step II *(a): Fuzzy Partition on the input space:* We construct n fuzzy sets such that

- each A_i is centered at x_i,
- Supp $A_i = \left(x_{i-1} + \frac{x_i - x_{i-1}}{3}, x_{i+1} - \frac{x_{i+1} - x_i}{3}\right)$ for $i = 2, \ldots, n-1$, while Supp $A_1 = \left(x_1, x_2 - \frac{x_2 - x_1}{3}\right)$ and Supp $A_n = \left(x_{n-1} + \frac{x_n - x_{n-1}}{3}, x_n\right)$,
- each A_i is normal at x_i, i.e., $A_i(x_i) = 1$,
- each A_i is a continuous convex fuzzy set,
- $\{A_i\}_{i=1}^n$ form a partition as defined in Definition 5.

For instance, if each of the A_i's $(i = 2, \ldots, n-1)$ *is a triangular fuzzy set and* A_1, A_n *are half-triangular with all of them attaining normality at* x_i *then clearly we can construct* $\{A_i\}_{i=1}^n$*'s partitioning the input space* X *as in* Definition 5 *and are continuous, convex, of finite support and* $A_i(x_i) = 1$.

Step II *(b): Fuzzy Partition on the output space*

Now we have the output space partition points as $y_1, y_2, \ldots y_k$. We partition the output space such that $B_1, B_2, \ldots B_k$ form a Ruspini partition (as above) with $B_j(y_j) = 1, \quad j = 1, 2, \ldots k$. Here obviously,

$$|y_j - y_{j-1}| < \frac{\epsilon}{2}, \quad j = 1, 2, \ldots k.$$

Further, let the fuzzy sets $\{B_j\}_{j=1}^k$ be continuous, convex, and of finite support along the same lines as the A_i's above, i.e., Supp $B_1 = (y_1, y_2)$, Supp $B_j = (y_{j-1}, y_{j+1})$, $j = 2, 3, \ldots k-1$, Supp $B_k = (y_{k-1}, y_k)$.

Step III: Construction of the *smooth* rule base

We construct the rule base with l rules of the following form:

$$\text{IF } x \text{ is } A_i \text{ THEN } y \text{ is } B_i, \quad i = 1, 2, \ldots n, \tag{9}$$

where the consequent B_i in the i-th rule is chosen such that $i = j$ is the index of that $y_j = h(x_i)$, where x_i is the point at which A_i attains normality.

Note that, since h is continuous, by the above assignment of the rules, we have that rules whose antecedents are adjacent also have adjacent consequents, i.e., for any $i = 1, 2, \ldots n-1$ we have Supp $B_i \cap$ Supp $B_{i+1} \neq \emptyset$. Thus, the constructed rule base is smooth as defined in [16].

Step IV : Approximation capability of the output

Now we consider an SBR fuzzy system with multiple SISO rules of the form (9). Let $x' \in X$ be the given input. Clearly, $x' \in [x_m, x_{m+1}]$ for some $m \in \mathbb{N}_n$. Now as in Sect. 5.2, we fuzzify x' in such a way that the fuzzified input A' (with $A'(x') = 1$) intersects at most two of the A_i's, say, A_m, A_{m+1}.

For instance, one could take A' Example 5.

So we have the following:

$$\begin{aligned} B'(y) &= T\big[M\left(A', A_m\right) \longrightarrow B_m(y), \\ &\qquad M\left(A', A_{m+1}\right) \longrightarrow B_{m+1}(y)\big] \\ &= T\big[s_m \longrightarrow B_m(y), s_{m+1} \longrightarrow B_{m+1}(y)\big], \end{aligned}$$

where $s_m = M\left(A', A_m\right)$ and $s_{m+1} = M\left(A', A_{m+1}\right)$. Note that by our assumption on M, we have that $s_m + s_{m+1} \leq 1$.

The output fuzzy set B' is given by (7). We consider the kernel of B', i.e., Ker $B' = \{y : B'(y) = 1\}$. We choose the defuzzified output y' such that it belongs to Ker B' (Fig. 3).

Since T is a t-norm, we know that $T(p, q) = 1$ if and only if $p = 1$ and $q = 1$. Noting that J has (OP), i.e., $p \longrightarrow q = 1 \Leftrightarrow p \leq q$ and $s_m + s_{m+1} \leq 1$, we have

$$\begin{aligned} \text{Ker } B' &= \left\{y : B'(y) = 1\right\} \\ &= \{y : s_m \longrightarrow B_m(y) = 1\} \bigcap \{y : s_{m+1} \longrightarrow B_{m+1}(y) = 1\} \\ &= \{y : s_m \leq B_m(y)\} \bigcap \{y : s_{m+1} \leq B_{m+1}(y)\}. \end{aligned}$$

Let $\alpha_m = \min\{\alpha : s_m \longrightarrow \alpha = 1\}$ and $\beta_{m+1} = \min\{\beta : s_{m+1} \longrightarrow \beta = 1\}$. Since J has (OP), clearly $\alpha_m = s_m$ and $\beta_{m+1} = s_{m+1}$.

By the continuity and convexity of B_m, B_{m+1} there exist $a_m, b_m, a_{m+1}, b_{m+1}$ such that $B_m(a_m) = B_m(b_m) = s_m$ and $B_{m+1}(a_{m+1}) = B_{m+1}(b_{M=1}) = s_{m+1}$. By the monotonicity of the implication in the second variable, for every $y \in [a_m, b_m]$ we have that $s_m \rightarrow B_m(y) = 1$ and for every $y \in [a_{m+1}, b_{m+1}]$ we have that $s_{m+1} \rightarrow B_{m+1}(y) = 1$. Thus,

$$\begin{aligned} \{y : s_m \leq B_m(y)\} &= [a_m, b_m], \quad \text{and} \\ \{y : s_{m+1} \leq B_{m+1}(y)\} &= [a_{m+1}, b_{m+1}]. \\ \text{Hence, Ker } B' = \{y : B'(y) = 1\} &= [a_m, b_m] \bigcap [a_{m+1}, b_{m+1}]. \end{aligned} \tag{10}$$

Claim: Ker $B' = [a_{m+1}, b_m] \neq \emptyset$.

First, note that for any $s_m \in [0, 1]$ by the normality of B_m we have that $B_m(y_m) = 1$ and hence $y_m \in \{y : s_m \leq B_m(y)\} = y_m \in [a_m, b_m] \neq \emptyset$. Similarly, $y_{m+1} \in [a_{m+1}, b_{m+1}] \neq \emptyset$. it suffices to show that $a_{m+1} \leq b_m$ from whence Ker $B' = [a_{m+1}, b_m]$.

Note that since $m < m+1$, $y_m < y_{m+1}$ and from $a_{m+1} \in$ Supp B_{m+1} we have that $y_m \leq a_{m+1} \leq y_{m+1}$. Similarly, $y_m \leq b_m \leq y_{m+1}$. Hence, $y_m \leq a_{m+1}, b_m \leq y_{m+1}$.

Since B_{m+1} is monotonic on $[y_m, y_{m+1}]$,

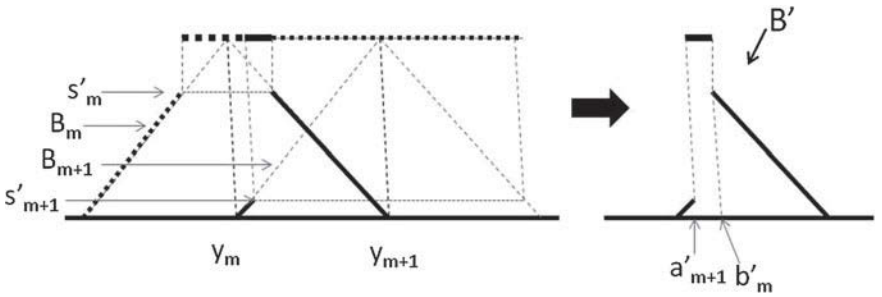


Fig. 3 The output fuzzy set B'

$$\begin{aligned} a_{m+1} > b_m \text{ implies } & B_{m+1}(a_{m+1}) \geq B_{m+1}(b_m) \\ \text{implies } & s_{m+1} \geq 1 - B_m(b_m) \\ \text{implies } & s_{m+1} \geq 1 - s_m \\ \text{implies } & s_m + s_{m+1} \geq 1. \end{aligned}$$

Since M satisfies (MCP), $s_m + s_{m+1} \leq 1$ and hence $s_m + s_{m+1} = 1$. Now,

$$\begin{aligned} s_m + s_{m+1} = 1 \text{ implies} & B_{m+1}(a_{m+1}) + B_m(b_m) = 1 \\ \text{implies } & B_{m+1}(a_{m+1}) = 1 - B_m(b_m) \\ \text{implies } & B_{m+1}(a_{m+1}) = B_{m+1}(b_m) \\ \text{implies } & b_m \in [a_{m+1}, b_{m+1}], \text{ i.e., } a_{m+1} \leq b_m. \end{aligned}$$

Now, we define $g(x')$ as either of the following—(11) or (12):

$$y' = g(x') = FOM(B'(y)) = a_{m+1} \tag{11}$$

$$y' = g(x') = LOM(B'(y)) = b_m \tag{12}$$

Now from the above we have the system function as, $y' = g(x') = a_{m+1}$ or b_m. Now clearly, $a_{m+1}, b_m \in [y_m, y_{m+1}]$ and hence,

$$\left|y_m - g(x')\right| < \frac{\epsilon}{2} \quad \text{or} \quad \left|y_{m+1} - g(x')\right| < \frac{\epsilon}{2}.$$

WLOG, let $|y_m - g(x')| < \frac{\epsilon}{2}$ i.e., $|y_m - y'| < \frac{\epsilon}{2}$. Now since $x' \in [x_m, x_{m+1}]$, we have $|h(x') - y_m| < \frac{\epsilon}{2}$. Finally we have the following:

$$\begin{aligned} \left|g(x') - h(x')\right| &= \left|y' - h(x')\right| \\ &\leq \left|y' - y_m\right| + |y_m - h(x')| \\ &< \frac{\epsilon}{2} + \frac{\epsilon}{2} < \epsilon. \end{aligned}$$

Since x' is arbitrary we have, $\max_{x\in[a,b]} |h(x) - g(x)| < \epsilon$.

Remark 1 Note that with g as in (11) or (12) and since M satisfies (MS), if $x' = x_k \in X$ we have $M(A', A_k) = 1$ and we obtain $B' = B_k$, i.e., $g(x') = y_k$ and the interpolativity of the inference is preserved.

7 Conclusion

In this work, we provided a constructive proof of the universal approximation properties of SBR when the modification function taken in the inference is a fuzzy implication.

References

1. Baczyński, M., Jayaram, B.: Fuzzy Implications, Studies in Fuzziness and Soft Computing, vol. 231. Springer, Berlin (2008)
2. Bandler, W., Kohout, L.J.: Semantics of implication operators and fuzzy relational products. Int. J. Man Mach. Stud. **12**(1), 89–116 (1980)
3. Chen, S.M.: A new approach to handling fuzzy decision-making problems. In: IEEE Trans. Sys. Man Cybern. **18**(6), 1012–1016 (1988)
4. Cross, V., Sudkamp, T.: Fuzzy implication and compatibility modification. In: IEEE International Conference on Fuzzy Systems, vol. 1, pp. 219–224 (1993)
5. Driankov, D., Hellendoorn, H., Reinfrank, M.: An Introduction to Fuzzy Control, 2nd edn. Springer, UK (1996)
6. Dubois, D., Prade, H.: The generalized modus ponens under sup-min composition—a theoretical study. Gupta, M.M., Kandel, A., Bandler, W., Kiszka, J.B. (Eds.), Approximate Reasoning in Expert Systems, Elsevier, North Holland, pp. 157–166 (1985)
7. Klement, E.P., Mesiar, R., Pap, E.: Triangular Norms, Trends in Logic, vol. 8. Kluwer Academic Publishers, Dordrecht (2000)
8. Li, Y.M., Shi, Z.K., Li, Z.H.: Approximation theory of fuzzy systems based upon genuine many-valued implications: MIMO cases. Fuzzy Sets Syst. **130**, 159–174 (2002)
9. Li, Y.M., Shi, Z.K., Li, Z.H.: Approximation theory of fuzzy systems based upon genuine many-valued implications: SISO cases. Fuzzy Sets Syst. **130**, 147–157 (2002)
10. Magrez, P., Smets, P.: Fuzzy modus ponens: A new model suitable for applications in knowledge-based systems. Int. J. Intell. Syst. **4**(2), 181–200 (1989). doi:10.1002/int.4550040205
11. Mandal, S., Jayaram, B.: Approximation capability of siso fuzzy relational inference systems based on fuzzy implications. In: IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2013 (2013). doi:10.1109/FUZZ-IEEE.2013.6622438
12. Morsi, N.N., Fahmy, A.A.: On generalized modus ponens with multiple rules and a residuated implication. Fuzzy Sets Syst. **129**(2), 267–274 (2002)
13. Roychowdhury, S., Pedrycz, W.: A survey of defuzzification strategies. Int. J. Intell. Syst. **16**(6), 679–695 (2001)
14. Tikk, D., Kóczy, L.T., Gedeon, T.D.: A survey on universal approximation and its limits in soft computing techniques. Int. J. Approx. Reasoning **33**(2), 185–202 (2003)
15. Turksen, I., Zhong, Z.: An approximate analogical reasoning approach based on similarity measures. IEEE Trans. Syst. Man Cybern. **18**(6), 1049–1056 (1988). doi:10.1109/21.23107

16. Štěpnička, M., Baets, B.D.: Monotonicity of implicative fuzzy models. In: IEEE International Conference on Fuzzy Systems (2010). doi:10.1109/FUZZY.2010.5584142
17. Zadeh, L.A.: Outline of a new approach to the analysis of complex systems and decision processes. IEEE Trans. Syst. Man Cybern. SMC-**3**(1), 28–44 (1973)
18. Zadeh, L.A.: The concept of a linguistic variable and its application to approximate reasoning-III. Inf. Sci. **9**, 43–80 (1975)

Chapter 15
Similarity Measure of Intuitionistic Fuzzy Numbers by the Centroid Point

Satyajit Das and Debashree Guha

Abstract The aim of the paper is to introduce a new similarity measure between intuitionistic fuzzy numbers (IFNs). The proposed method is based on the centroid point of IFNs. It is also proved that the proposed measure satisfies the properties of similarity measure. Examples are considered to compare the proposed similarity measure with the existing similarity measures. The similarity results show that the new similarity measure can overcome the faults of the existing similarity measures.

Keywords Intuitionistic fuzzy number · Centroid point · Similarity measure

1 Introduction

As Zadeh proposed fuzzy sets [1], many researchers concentrated on computing similarity of fuzzy sets and they have applied them in several fields such as pattern recognition [2, 3], approximate reasoning [4], decision making [5], etc. Similarity measures between fuzzy numbers have also been derived by researchers [6–13].

Fuzzy set was further generalized and out of several higher order fuzzy sets, intuitionistic fuzzy set (IFS), introduced by Atanassov [14] became very useful to deal with uncertainty present in real-world situations. Different similarity measures between IFSs have also been proposed in the literature. Guha and Chakraborty [15] developed a theoretical-based similarity measure between IFSs. Based on Housedroff distance, a similarity measure of IFSs was proposed by Hung and Yang [16]. Szmidt and Kacprzyk [17] proposed a similarity measure based on Hamming and Euclidean distance measures. In 2007, Li et al. [18] gave a comparative analysis of similarity

S. Das · D. Guha (✉)
Indian Institute of Technology, Patna 800013, India
e-mail: debashree@iitp.ac.in

S. Das
e-mail: satyajit@iitp.ac.in

R. N. Mohapatra et al. (eds.), *Mathematics and Computing 2013*,
Springer Proceedings in Mathematics & Statistics 91,
DOI: 10.1007/978-81-322-1952-1_15,

measure between IFSs. Recently with the universe as the real line, research on the concept of intuitionistic fuzzy numbers (IFNs) has received attention from many scholars. However, till now little research has been done on computing similarity measure between IFNs. In 2011, Ye [19] proposed a multi-criteria group decision-making method using vector similarity measure for trapezoidal intuitionistic fuzzy numbers (TrIFNs). Furthermore, using the Hamming distance and Euclidean distance between TrIFNs, a similarity measure was proposed by Ye [20]. In 2013, Farhadinia and Ban [21] developed a new similarity measure of generalized IFNs and generalized interval-valued fuzzy numbers.

However, after studying the above similarity measures it can be observed that they fail to compute the similarity measure properly for some cases. Under these situations, experts may not be able to implement the comparison in a proper manner. This creates problems in case of practical applications. With this point of view, to overcome the shortcomings of existing similarity measures, a new similarity measure of IFNs by utilizing distance between centroid point of IFNs has been proposed in this paper. Furthermore, analysis and comparison of existing similarity measures of IFNs have been described with the help of a set of examples.

This paper has been arranged as follows. In Sect. 2, definitions of IFNs have been studied. A brief description of existing similarity measures is given in Sect. 3. Section 4 describes the proposed similarity measure. In Sect. 5, some examples are given to compare the proposed measure with the existing similarity measures. The conclusions are drawn in Sect. 6.

2 Preliminaries

In this section, definitions and arithmetic operations of IFNs are analyzed.

Definition 1 [22] Let $A = [(a, b, c, d)], w_A; (a', b, c, d'), u_A]$ be a generalized TrIFN (GTrIFN) and its membership and non-membership functions are, respectively, defined as follows:

$$\mu_A(x) = \begin{cases} \dfrac{(x-a)w_A}{(b-a)}, & \text{for } a \le x < b, \\ w_A, & \text{for } b \le x \le c, \\ \dfrac{(d-x)w_A}{(d-c)}, & \text{for } c < x \le d, \\ 0, & \text{for } x < a, x > d. \end{cases} \tag{1}$$

and

$$\nu_A(x) = \begin{cases} \dfrac{(b-x)+\left(x-a'\right)u_A}{(b-a')}, & \text{for } a' \le x < b, \\ u_A, & \text{for } b \le x \le c, \\ \dfrac{(x-c)+\left(d'-x\right)u_A}{(d'-c)}, & \text{for } c < x \le d', \\ 1, & \text{for } x < a', x > d'. \end{cases} \tag{2}$$

where $0 \le w_A, u_A \le 1; w_A + u_A \le 1; a, b, c, d, a', d' \in R$. For the sake of simplicity, throughout this paper we have considered $a = a'$ and $d = d'$. Symbolically, then GTrIFN can be represented as $A = [(a, b, c, d); w_A, u_A]$. If $b = c$ then GTrIFN transforms to generalized triangular intuitionistic fuzzy number (GTIFN).

Definition 2 [23] Let $A_1 = [(a_1, b_1, c_1, d_1); w_1, u_1]$ and $A_2 = [(a_2, b_2, c_2, d_2); w_2, u_2]$ be two TrIFNs and $K \ge 0$ be a scalar, then

(1) $A_1 \oplus A_2 = [(a_1 + a_2, b_1 + b_2, c_1 + c_2, d_1 + d_2); w_1 + w_2 - w_1 w_2, u_1 u_2]$
(2) $A_1 \otimes A_2 = [(a_1 a_2, b_1 b_2, c_1 c_2, d_1 d_2); w_1 w_2, u_1 + u_2 - u_1 u_2]$
(3) $KA_1 = [(Ka_1, Kb_1, Kc_1, Kd_1); 1 - (1 - w_1)^K, u_1^K]$
(4) $A_1^K = [(a_1^K, b_1^K, c_1^K, d_1^K); w_1^K, 1 - (1 - u_1)^K]$

3 Existing Similarity Measures of IFNs

In the literature, there are very few existing similarity measures of IFNs. A brief description of these methods is given below.

3.1 Vector Cosine Similarity Measure

For two TrIFNs $A = [(a_1, a_2, a_3, a_4), 1; (b_1, b_2, b_3, b_4), 0]$ and $B = [(a_1', a_2', a_3', a_4'), 1; (b_1', b_2', b_3', b_4'), 0]$, vector cosine similarity measure is given as [19]

$$\cos(A, B) = \frac{\sum_{i=1}^{4} a_i a_i' + \sum_{i=1}^{4} b_i b_i'}{\sqrt{\sum_{i=1}^{4} \left(a_i^2\right) + \sum_{i=1}^{4} \left(b_i^2\right)} \sqrt{\sum_{i=1}^{4} \left(a_i'\right)^2 + \sum_{i=1}^{4} \left(b_i'\right)^2}} \tag{3}$$

3.2 Similarity Measure Using Distance Measure

The Hamming distance and Euclidean distance-based similarity measures were proposed in [20]. For two TrIFNs $A = [(a_1, a_2, a_3, a_4), 1; (b_1, b_2, b_3, b_4), 0]$ and $B = [(a_1', a_2', a_3', a_4'), 1; (b_1', b_2', b_3', b_4'), 0]$,

Hamming distance:

$$d_H(A, B) = \frac{1}{8}\left(\sum_{i=1}^{4} |a_i - a_i'| + \sum_{i=1}^{4} |b_i - b_i'|\right) \tag{4}$$

and Euclidean distance:

$$d_E(A, B) = \sqrt{\frac{1}{8}\left(\sum_{i=1}^{4} \left(a_i - a_i'\right)^2 + \sum_{i=1}^{4} \left(b_i - b_i'\right)^2\right)} \tag{5}$$

Then similarity measures are given as

$$S_H(A, B) = 1 - \frac{1}{8}\left(\sum_{i=1}^{4} |a_i - a_i'| + \sum_{i=1}^{4} |b_i - b_i'|\right) \tag{6}$$

$$S_E(A, B) = 1 - \sqrt{\frac{1}{8}\left(\sum_{i=1}^{4} \left(a_i - a_i'\right)^2 + \sum_{i=1}^{4} \left(b_i - b_i'\right)^2\right)} \tag{7}$$

The bigger the value of $S_H(A, B)$ or $S_E(A, B)$, the more the similarity between A and B.

3.3 Farhadinia and Ban's Process

The similarity measure between two GTIFNs $A = [(a_1, a_2, a_3),\ w_A; (b_1, b_2, b_3),\ u_A]$ and $B = [(a_1', a_2', a_3'),\ w_B; (b_1', b_2', b_3'),\ u_B]$ is given as [21]

$$S_F(A, B) = \sigma_L^p \cdot \sigma_U^q, \text{ where } p + q = 1. \tag{8}$$

Here

$$\begin{aligned}
\sigma_L &= S_F\left((\Phi(A))^L, (\Phi(B))^L\right) \\
&= \left(1 - \frac{\sum_{i=1}^{3} |a_i - a_i'|}{3}\right) \times \frac{\min\left(P_A^L, P_B^L\right) + \min\left(w_A, w_B\right)}{\max\left(P_A^L, P_B^L\right) + \max\left(w_A, w_B\right)} \\
\sigma_U &= S_F\left((\Phi(A))^U, (\Phi(B))^U\right) \\
&= \left(1 - \frac{\sum_{i=1}^{3} |b_i - b_i'|}{3}\right) \times \frac{1 + \min\left(P_A^U, P_B^U\right) - \max(u_A, u_B)}{1 + \max\left(P_A^U, P_B^U\right) - \min(u_A, u_B)}
\end{aligned}$$

$$P_A^L = P_e\left((\Phi(A))^L\right) = \mathrm{e}^{\sqrt{(a_1-a_2)^2+(w_A)^2}+\sqrt{(a_2-a_3)^2+(w_A)^2}+a_3-a_1}$$

and

$$P_A^U = P_e\left((\Phi(A))^U\right) = \mathrm{e}^{\sqrt{(b_1-b_2)^2+(w_A)^2}+\sqrt{(b_2-b_3)^2+(w_A)^2}+b_3-b_1}$$

where Φ is a mapping from GTIFN to generalized interval-valued triangular fuzzy number and is defined as

$\Phi(A) = \left[(\Phi(A))^L, (\Phi(A))^U\right] = [((a_1, a_2, a_3), w_A), ((b_1, b_2, b_3), 1-u_A)]$ and $P_e(\Phi(A))$ denotes the exponential function value of perimeter of $\Phi(A)$. Similarly, P_B^L, P_B^U and $\Phi(B)$ can be determined.

4 New Concept of Similarity Measure for IFNs

In this section, a new similarity measure between IFNs has been constructed on the basis of distance between centroid points of IFNs. For this purpose, the centroid point of IFNs has been introduced first.

4.1 Centroid Point of IFN

Let $A = [(a, b, c, d; w, u]$ be a GTrIFNs. The centroid point of A according to [24] can be given as

$$X_A = \frac{x1}{x2},$$

where

$$x1 = \int_a^{\frac{aw-au+b}{w-u+1}} x g_L \mathrm{d}x + \int_{\frac{aw-au+b}{w-u+1}}^{b} x f_L \mathrm{d}x + \int_b^c x w \mathrm{d}x + \int_c^{\frac{dw-du+c}{w-u+1}} x f_R \mathrm{d}x + \int_{\frac{dw-du+c}{w-u+1}}^{c} x g_R \mathrm{d}x \tag{9}$$

$$x2 = \int_a^{\frac{aw-au+b}{w-u+1}} g_L \mathrm{d}x + \int_{\frac{aw-au+b}{w-u+1}}^{b} f_L \mathrm{d}x + \int_b^c w \mathrm{d}x + \int_c^{\frac{\mathrm{d}w-\mathrm{d}u+c}{w-u+1}} f_R \mathrm{d}x + \int_{\frac{\mathrm{d}w-\mathrm{d}u+c}{w-u+1}}^{c} g_R \mathrm{d}x \tag{10}$$

and

$$Y_A = \frac{y1}{y2},$$

where

$$y1 = \int_0^w y(h_R - h_L)\mathrm{d}y + \left[\int_0^1 yd \cdot \mathrm{d}y - \int_0^{\frac{w}{w-u+1}} yh_R\mathrm{d}y - \int_{\frac{w}{w-u+1}}^1 yk_R\mathrm{d}y\right] + \left[\int_0^{\frac{w}{w-u+1}} yh_L\mathrm{d}y + \int_{\frac{w}{w-u+1}}^1 yk_L\mathrm{d}y - \int_0^1 ay\mathrm{d}y\right] \tag{11}$$

$$y2 = \int_0^w (h_R - h_L)\mathrm{d}y + \left[\int_0^1 d \cdot \mathrm{d}y - \int_0^{\frac{w}{w-u+1}} h_R\mathrm{d}y - \int_{\frac{w}{w-u+1}}^1 k_R\mathrm{d}y\right] + \left[\int_0^{\frac{w}{w-u+1}} h_L\mathrm{d}y + \int_{\frac{w}{w-u+1}}^1 k_L\mathrm{d}y - \int_0^1 a\mathrm{d}y\right] \tag{12}$$

where f_L and f_R are the left and right parts of membership function and g_L and g_R are the left and right parts of non-membership function of GTrIFN A defined in Eqs. (1) and (2), respectively (See Fig. 1). The inverse functions of f_L and f_R are h_L and h_R, respectively, k_L and k_R are the inverse functions of g_L and g_R, respectively (See Fig. 2). The inverse functions h_L, h_R, k_L and k_R can be computed by utilizing Eqs. (1) and (2). See the more detailed argumentation in [24].

4.2 Distance and Similarity Measure

Let us consider two GTrIFNs $A = [(a_1, a_2, a_3, a_4); w_A, u_A]$ and $B = [(b_1, b_2, b_3, b_4); w_B, u_B]$. The centroid point of these two numbers can be determined by Eqs. (9)–(12) and denoted by (X_A, Y_A) and (X_B, Y_B), respectively. The distance between the centroid point of two GTrIFNs A and B is denoted by $D(A, B)$ and defined by

$$D(A, B) = \sqrt{(X_A - X_B)^2 + (Y_A - Y_B)^2} \tag{13}$$

The distance $D(A, B)$ between the centroid point of two GTrIFNs A and B defined in Eq. (13) satisfies all the metric properties.

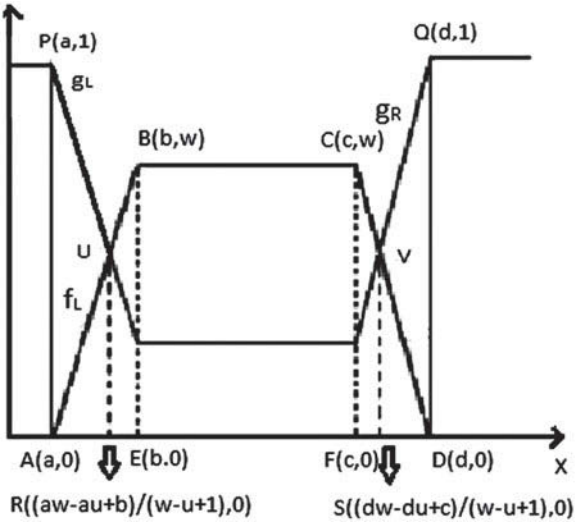


Fig. 1 GTrIFN A

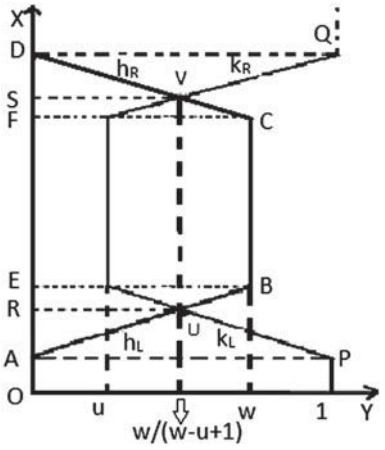


Fig. 2 Inverse of GTrIFN A

Similarity measure is such a function that calculates the degree of similarity between two classes. It is well known that distance measure and similarity measures are dual concepts and so similarity measures can be obtained from distance measures. For two GTrIFNs $A = [(a_1, a_2, a_3, a_4); w_A, u_A]$ and $B = [(b_1, b_2, b_3, b_4); w_B, u_B]$, similarity measure can be defined as follows:

$$S(A, B) = \frac{1}{1 + D(A, B)} = \frac{1}{1 + \sqrt{(X_A - X_B)^2 + (Y_A - Y_B)^2}} \tag{14}$$

4.3 Properties of the Proposed Similarity Measure

The proposed similarity measure between GTrIFNs satisfies the following properties:

$(P1)$ $0 < S(A, B) \leq 1$

Proof: We have $D(A, B) = \sqrt{(X_A - X_B)^2 + (Y_A - Y_B)^2}$

which is greater than zero.

$\therefore 1 + D(A, B) = 1 + \sqrt{(X_A - X_B)^2 + (Y_A - Y_B)^2} \geq 1$

$\Rightarrow \frac{1}{1+D(A,B)} = \frac{1}{1+\sqrt{(X_A-X_B)^2+(Y_A-Y_B)^2}} \leq 1$

Also, $\frac{1}{1+D(A,B)} > 0$

$\therefore 0 < S(A, B) \leq 1$

$(P2)$ $S(A, B) = 1 \Longleftrightarrow A = B$

Proof: Since $A = B$ implies that $X_A = X_B$ and $Y_A = Y_B$.

Hence $D(A, B) = 0$.

$\therefore S(A, B) = \frac{1}{1+0} = 1$.

Again if $S(A, B) = 1$, then

$\frac{1}{1+\sqrt{(X_A-X_B)^2+(Y_A-Y_B)^2}} = 1$

$\Rightarrow 1 = 1 + \sqrt{(X_A - X_B)^2 + (Y_A - Y_B)^2}$

$\Rightarrow \sqrt{(X_A - X_B)^2 + (Y_A - Y_B)^2} = 0$

$\Rightarrow X_A - X_B = 0$ and $Y_A - Y_B = 0$

$\therefore X_A = X_B$ and $Y_A = Y_B \Rightarrow A = B$.

$(P3)$ $S(A, B) = S(B, A)$

Proof: We have

$$S(A, B) = \frac{1}{1+\sqrt{(X_A-X_B)^2+(Y_A-Y_B)^2}} = \frac{1}{1+\sqrt{(X_B-X_A)^2+(Y_B-Y_A)^2}} = S(B, A)$$

$(P4)$ Let A, B, and C be three TrIFNs. If B is more similar to A than C then $S(A, B) > S(B, C)$.

For example, let us consider three TrIFNs

$A = [(1, 3, 4, 5); 0.8, 0.1]$

$B = [(2, 3, 5, 6); 0.7, 0.2]$

$C = [(4, 6, 7, 8); 1, 0]$

The centroid point of these three TrIFNs are given as

$X_A = 2.9947, Y_A = 0.3721$

$X_B = 4.0000, Y_B = 0.3539$

$X_C = 6.0385, Y_C = 0.4231$

Then $D(A, B) = 1.0054$ and $D(B, C) = 2.0396$.

$\therefore \quad S(A, B) = 0.4986$ and $S(B, C) = 0.3289$.

We see that $S(A, B) > S(B, C)$.

Also, the desirable result is B is more similar to A.

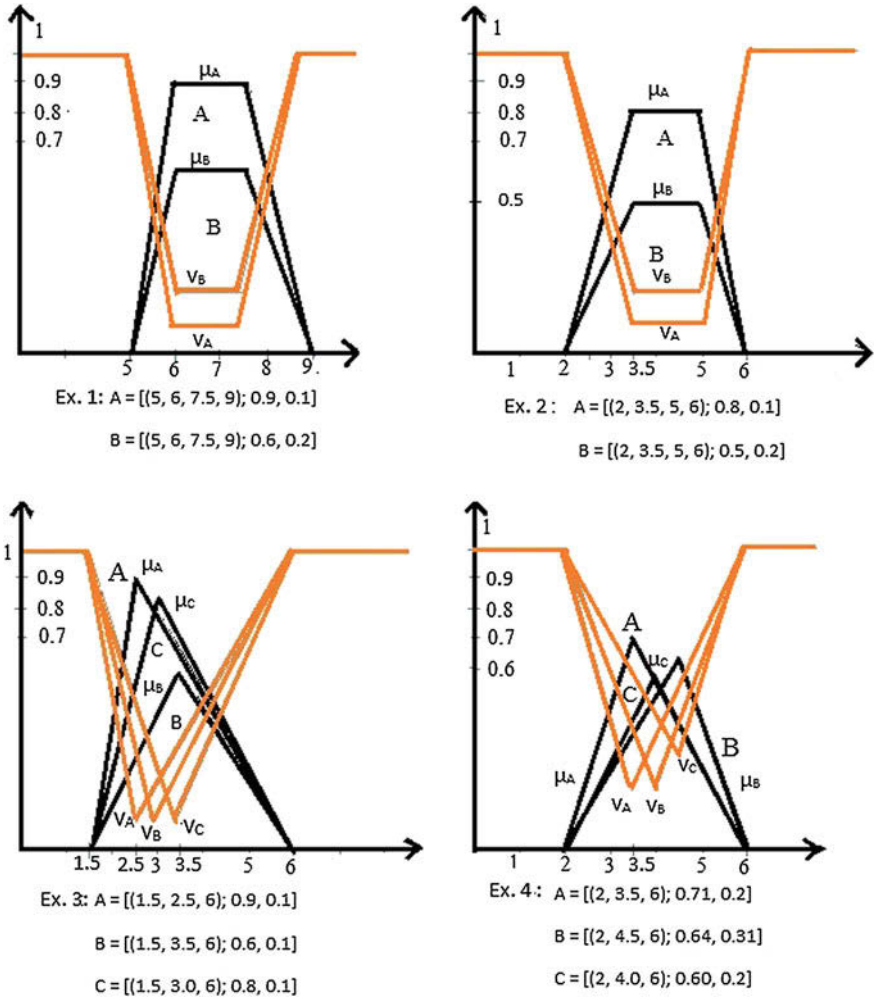


Fig. 3 Sets of IFNs

5 Comparison of Similarity Measures with Existing Methods

In this section, a set of examples of IFNs have been constructed (Fig. 3) for giving a comparative analysis of the proposed similarity measure and existing similarity measures [19–21]. In Table 1, some shortcomings of the existing measures as well as advantages of the proposed measure are shown and are described below.

- From Examples 1 and 2 (see Fig. 3), in both cases similarity between two TrIFNs A and B by [19, 20] is $S(A, B) = 1$, although A and B have different membership

Table 1 A comparison of the proposed similarity measure with the existing methods

Existing similarity measures	Examples set	The proposed method
Ye's process [19]	Example 1	
$\cos(A, B) = \frac{\sum_{p=1}^{4} a_{1p}a_{2p} + \sum_{q=1}^{4} b_{1q}b_{2q}}{\sqrt{\sum_{p=1}^{4}(a_{1p}^2) + \sum_{q=1}^{4}(b_{1q}^2)}\sqrt{\sum_{p=1}^{4}(a_{2p})^2 + \sum_{q=1}^{4}(b_{2q}^2)}}$	$A = [(5, 6, 7.5, 9); 0.9, 0.1]$	$D(A, B) = 0.0877$
	$B = [(5, 6, 7.5, 9); 0.6, 0.2]$	$S(A, B) = 0.9194$
	$\cos(A, B) = 1$	
Ye's process [20]	Example 2	
$S_H(A, B) = 1 - \frac{1}{8}(\sum_{p=1}^{4} \lvert a_{1p} - a_{2p}\rvert + \sum_{q=1}^{4} \lvert b_{1q} - b_{2q}\rvert)$	$A = [(2, 3.5, 5, 6); 0.8, 0.1]$	$D(A, B) = 0.0819$
$S_E(A, B) = 1 - \sqrt{\frac{1}{8}(\sum_{p=1}^{4}(a_{1p} - a_{2p})^2 + \sum_{q=1}^{4}(b_{1q} - b_{2q})^2)}$	$B = [(2, 3.5, 5, 6); 0.5, 0.2]$	$S(A, B) = 0.9243$
	$S_H(A, B) = 1$	
	$S_E(A, B) = 1$	
Ye's process [20]	Example 3	
$S_H(A, B) = 1 - \frac{1}{8}(\sum_{p=1}^{4} \lvert a_{1p} - a_{2p}\rvert + \sum_{q=1}^{4} \lvert b_{1q} - b_{2q}\rvert)$	$A = [(1.5, 2.5, 6); 0.9, 0.1]$	$D(A, C) = 0.016177$
$S_E(A, B) = 1 - \sqrt{\frac{1}{8}(\sum_{p=1}^{4}(a_{1p} - a_{2p})^2 + \sum_{q=1}^{4}(b_{1q} - b_{2q})^2)}$	$B = [(1.5, 3.5, 6); 0.6, 0.1]$	$D(B, C) = 0.0261$
	$C = [(1.5, 3, 6); 0.8, 0.1]$	$S(A, C) = 0.9840$
	$S_H(A, C) = 0.75$	$S(B, C) = 0.9745$
	$S_H(B, C) = 0.75$	$S(A, C) > S(B, C)$
	$S_E(A, C) = 0.6465$	
	$S_E(B, C) = 0.6465$	
	$S_H(A, C) = S_H(B, C)$	
	$S_E(A, C) = S_E(B, C)$	
Farhadinia and Ban's process [21]	Example 4	
$S_F(A, B) = \sigma_L^p \cdot \sigma_U^q$	$A = [(2, 3.5, 6); 0.71, 0.2]$	$D(A, C) = 0.0493$
where $p + q = 1$	$B = [(2, 4.5, 6); 0.64, 0.31]$	$D(B, C) = 0.0599$
	$C = [(2, 4, 6); 0.6, 0.2]$	$S(A, C) = 0.953$
	$S(A, C) = 0.7931$	$S(B, C) = 0.943$
	$S(B, C) = 0.7931$	$S(A, C) > S(B, C)$
	$S(A, C) = S(B, C)$	

values. But by the proposed similarity method, the similarity results for Examples 1 and 2 are $S(A, B) = 0.9194$ and $S(A, B) = 0.9243$, respectively.

- From Example 3 (see Fig. 3), it can be easily observed that the TrIFN C is more similar to A than B . By Ye's method [20] C has the same similarity as A and B. By utilizing the proposed method, the similarity result is $S(A, C) > S(B, C)$ which implies that C is more similar to A than B as expected.
- From Example 4 (see Fig. 3), by Farhadinia and Ban's method [21] C has the same similarity with A and B as $S(A, C) = S(B, C)$. But according to the proposed method C is more similar to A than B as $S(A, C) > S(B, C)$.

Therefore, from Table 1, it can be observed that the proposed similarity measure determines the similarity correctly and overcomes the shortcomings of existing methods.

6 Conclusion

In this paper, a new process of similarity measure of IFNs is proposed by utilizing distance measure. Here distance measure is the distance between centroid point of IFNs. After calculating distance measure, a new similarity measure has been introduced to calculate the degree of similarity between IFNs. In order to compare the proposed similarity measure with the existing measures some examples have been shown and we see that the proposed similarity measure can get over the faults of the existing measures. The proposed similarity measure may be applicable to the multi-criteria decision-making problem, pattern recognition, risk analysis, and many other fields which will be our future research topic.

References

1. Zadeh, L.A.: Fuzzy sets. Inf. Control **8**, 338–353 (1965)
2. Dengfeng, L., Chuntian, C.: New similarity measure of intuitionistic fuzzy sets and application to pattern recognitions. Pattern Recogn. Lett. **23**, 221–225 (2002)
3. Mitchell, H.B.: On the Dengfeng–Chuntian similarity measure and its application to pattern recognitions. Pattern Recogn. Lett. **24**, 3101–3104 (2003)
4. Li, Y., Zhongxian, C., Degin, Y.: Similarity measures between vague sets and vague entropy. J. Comput. Sci. **29**, 129–132 (2002) (in chinese)
5. Chen, S.M.: A new approach to handling fuzzy decision making problems. IEEE Trans. Syst. Man Cybern. **18**, 1012–1016 (1988)
6. Chen, S.M.: New methods for subjective mental workload assessment and fuzzy risk analysis. Cybern. Syst. **27**, 449–472 (1996)
7. Hsieh, C.H., Chen, S.H.: Similarity of generalized fuzzy numbers with graded mean integration representation. In: Proceedings of the 8th International Fuzzy Systems Association World Congress, pp. 551–555. Taipei (1999)
8. Chen, S.J., Chen, S.M.: Fuzzy risk analysis based on similarity measures of generalized fuzzy numbers. IEEE Trans. Fuzzy Syst. **11**, 45–56 (2003)

9. Hsu, H.M., Chen, C.T.: Aggregation of fuzzy opinions under group decision making. Fuzzy Sets Syst. **79**, 279–285 (1996)
10. Lee, H.S.: An optimal aggregation method for fuzzy opinions of group decision. In: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics **3**, 314–319 (1999)
11. Chen, S.M.: New methods for subjective mental workload assessment and fuzzy risk analysis. Int. J. Cybern. Syst. **27**, 449–472 (1996)
12. Kangari, R., Riggs, L.S.: Construction risk assessment by linguistics. IEEE Trans. Eng. Manag. **36**, 126–131 (1989)
13. Schmucker, K.J.: Fuzzy Sets, Natural Language Computations, and Risk Analysis. Computer Science, Rockville (1984)
14. Atanassov, K.: Intuitionistic fuzzy sets. Fuzzy Sets Syst. **20**, 87–96 (1986)
15. Guha, D., Chakraborty, D.: A theoretical development of similarity measure for intuitionistic fuzzy sets and its applications in multiple attribute decision making. J. Fuzzy Math. **18**, 391–402 (2010)
16. Hung, W., Yang, M.: Similarity measures of intuitionistic fuzzy sets based on Hausdroff distance. Pattern Recogn. Lett. **25**, 1603–1611 (2004)
17. Szmidt, E., Kacprzyk, J.: Distances between intuitionistic fuzzy sets. Fuzzy Sets Syst. **114**, 505–518 (2000)
18. Li, Y., Olson, D.L., Qin, Z.: Similarity measures between intuitionistic fuzzy (vague) sets: a comparative analysis. Pattern Recogn. Lett. **28**, 278–285 (2007)
19. Ye, J.: Multicriteria group decision-making method using vector similarity measure for trapezoidal intuitionistic fuzzy numbers. Group Decis. Negot. **21**, 519–530 (2012)
20. Ye, J.: Multicriteria group decision-making method using distance-based similarity measure for intuitionistic trapezoidal fuzzy numbers. Int. J. Gen. Syst. **41**, 729–739 (2012)
21. Farhadinia, B., Ban, A.I.: Developing new similarity measures of generalized intuitionistic fuzzy numbers and generalized interval-valued fuzzy numbers from similarity measures of generalized fuzzy numbers. Math. Comput. Model. **57**, 812–825 (2013)
22. Wang, J.Q., Zhang, Z.: Multi-criteria decision making method with incomplete certain information based on Intuitionistic fuzzy number. Control Decis. **24**, 226–230 (2009)
23. Wu, J., Cao, Q.: Same families of geometric aggregation operators with intuitionistic trapezoidal fuzzy numbers. Appl. Math. Model. **37**, 318–327 (2013)
24. Das, S., Guha, D.: Ranking of intuitionistic fuzzy number by centroid point. J. Ind. Intel. Inf. **1**, 107–110 (2013)

Chapter 16
Classification Rules for Exponential Populations Under Order Restrictions on Parameters

Nabakumar Jana, Somesh Kumar and Neeraj Misra

Abstract Classification procedures of an observation into one of two exponential populations is considered. Assuming a known order between the population parameters, a class of classification rules is proposed. Our study shows that each classification rule in the class is better than the likelihood ratio based classification rule. Comparison of these classification rules with respect to correct probability of classification has been done by extensive simulations.

Keywords Exponential population · Classification rule · Probability of correct classification

1 Introduction

The exponential distribution is one of the most widely used distributions. It has extensive applications in reliability and life testing problems (see for example, Rausand and Høyland [9] and Pal et al. [8]). The problem of classification of an observation into one of two exponential populations also arises frequently. For example, suppose a certain machine works properly if all components connected through series work properly. The lifetime of two different components in this series is exponentially distributed with different means. Suppose the system fails after a certain time. It

N. Jana (✉) · S. Kumar
Indian Institute of Technology Kharagpur, Kharagpur 721302, India
e-mail: nabakumar652@gmail.com; nabakumarj9@gmail.com

S. Kumar
e-mail: smsh@maths.iitkgp.ernet.in

N. Misra
Indian Institute of Technology Kanpur, Kanpur 208016, India
e-mail: neeraj@iitk.ac.in

R. N. Mohapatra et al. (eds.), *Mathematics and Computing 2013*,
Springer Proceedings in Mathematics & Statistics 91,
DOI: 10.1007/978-81-322-1952-1_16, © Springer India 2014

is desirable to know from the observed lifetime which particular component had actually failed.

The problem of classification of an observation into one of two or more normal populations is quite old and has been studied by many authors. One may refer to McLachlan [7] and Wakaki and Aoshima [11] for a detailed review of the literature on this problem. The problem of classification for one parameter exponential populations was first studied by Basu and Gupta [3]. They proposed classification rules assuming population parameters are unknown. For this problem Adegboye [1] numerically compared the probabilities of misclassification by taking two different classification rules. Basu and Gupta [4] investigated in detail the problem of classification for two parameter exponential populations. They considered the cases when some or all parameters may be known or unknown. The case of censored samples was also studied by them.

The performance of the classification rules can be improved if some additional information is available on unknown parameters. Long and Gupta [6] first proposed ordered classification rules for two univariate normal population assuming the mean of the first population is greater than that of the second population. For one parameter exponential populations, the classification rule is proposed by Conde et al. [5] assuming order restrictions on parameters. They used a mixed estimator of Vijayasree and Singh [10] to estimate parameters involved in the classification rule and showed that the new rule is better than the earlier rule proposed by Basu and Gupta [3]. In our study, we propose a class of classification rules based on a class of mixed estimators. We investigate the performance of these rules in terms of probabilities of correct classification.

The article is organized as follows. In Sect. 2, we derive the class of ordered classification rules and prove that each ordered classification rule is better than the usual classification rule with respect to the expected probability of correct classification. In Sect. 3, we compare between the rules with respect to each type of probability of correct classification through extensive simulations. We give an application of the proposed ordered classification rules in Sect. 4.

2 Classification Rules

Let Π_1 and Π_2 be two independent exponential populations. The probability density function associated with the population Π_i is

$$f_i(x) = \frac{1}{\sigma_i} \exp\left(-\frac{x}{\sigma_i}\right), \quad \sigma_i > 0, \ x > 0, \ i = 1, 2. \tag{1}$$

The parameters σ_i's are unknown but it is known a priori that $\sigma_1 \leq \sigma_2$. In other words, the expected value of the second population is larger than that of the first population. This type of situation arises for example, when it is known a priori that due to improved design, the mean life of the new system is more than that of

an old system. From each population training sample is available to estimate the parameters. Let $X = (X_1, X_2, \cdots, X_{n_1})$ be a training sample from population Π_1 and $Y = (Y_1, Y_2, \cdots, Y_{n_2})$ be a training sample from population Π_2. Without any restrictions on parameters the usual maximum likelihood estimators (MLEs) of σ_1 and σ_2 are $\hat{\sigma_1} = \bar{X} = \sum_{i=1}^{n_1} X_i/n_1$ and $\hat{\sigma_2} = \bar{Y} = \sum_{i=1}^{n_2} Y_i/n_2$ respectively. Note that $\bar{X}$ and $\bar{Y}$ are independently distributed with respective probability density functions

$$g_1(\bar{x}, \sigma_1) = \frac{n_1^{n_1}}{\Gamma n_1 \sigma_1^{n_1}} e^{-\frac{\bar{x} n_1}{\sigma_1}} \bar{x}^{n_1 - 1}, \qquad \bar{x} > 0,\ \sigma_1 > 0,$$

and

$$g_2(\bar{y}, \sigma_2) = \frac{n_2^{n_2}}{\Gamma n_2 \sigma_2^{n_2}} e^{-\frac{\bar{y} n_2}{\sigma_2}} \bar{y}^{n_2 - 1}, \qquad \bar{y} > 0,\ \sigma_2 > 0.$$

Under the considered ordered restrictions, Vijayasree and Singh [10] proposed the mixed estimators

$$\tilde{\sigma_1} = \min(\bar{X}, \alpha\bar{X} + (1-\alpha)\bar{Y}),\ 0 \le \alpha < 1,$$

and

$$\tilde{\sigma_2} = \max(\bar{Y}, \alpha\bar{Y} + (1-\alpha)\bar{X}),\ 0 \le \alpha < 1$$

for σ_1 and σ_2 respectively. For $\alpha = n_1/(n_1 + n_2)$, $\tilde{\sigma_1}, \tilde{\sigma_2}$ give the MLE of σ_1, σ_2 respectively. To ensure $\tilde{\sigma_1} \le \tilde{\sigma_2}$, we must have $0 \le \alpha < \frac{1}{2}$.

Suppose x is an observation which comes from one of these two populations but the exact population is unknown. We consider x as realization of a random variable X whose distribution is either f_1 or f_2. Assume that the populations are equally likely and the costs of misclassifications are equal. The usual classification rule R_U for these exponential population is given by Basu and Gupta [3]:

$$\begin{aligned} &\text{Classify } x \text{ into } \Pi_1 \text{ iff } (x - \hat{x_0})(\hat{\sigma_2} - \hat{\sigma_1}) \le 0, \\ &\text{Classify } x \text{ into } \Pi_2 \text{ iff } (x - \hat{x_0})(\hat{\sigma_2} - \hat{\sigma_1}) > 0, \end{aligned} \tag{2}$$

where

$$\hat{x_0} = \frac{\hat{\sigma_1}\hat{\sigma_2}\log(\hat{\sigma_1}/\hat{\sigma_2})}{\hat{\sigma_1} - \hat{\sigma_2}}.$$

We propose the ordered classification rule R_O^{α} based on the mixed estimators $\tilde{\sigma_1}$ and $\tilde{\sigma_2}$. The classification rule is:

$$\begin{aligned} &\text{Classify } x \text{ into } \Pi_1 \text{ iff } (x - \tilde{x_0}) \le 0, \\ &\text{Classify } x \text{ into } \Pi_2 \text{ iff } (x - \tilde{x_0}) > 0, \end{aligned} \tag{3}$$

where

$$\tilde{x_0} = \frac{\tilde{\sigma_1}\tilde{\sigma_2}\log(\tilde{\sigma_1}/\tilde{\sigma_2})}{\tilde{\sigma_1} - \tilde{\sigma_2}}.$$

For $\alpha = 0$, $\tilde{\sigma_1} = \min(\bar{X}, \bar{Y}), \tilde{\sigma_2} = \max(\bar{X}, \bar{Y})$ and the corresponding ordered classification rule was studied by Conde et al. [5].

Let $P_\star(i|j)$ denote the probability that an observation coming from population j is classified in population i under the rule $R_\star$. Then the global probability of misclassification under rule $R_\star$ is $P_\star(MC) = \frac{1}{2}(P_\star(1|2) + P_\star(2|1))$.

Theorem 1 *Let R_O^α and R_U be the classification rules defined in (3) and (2), respectively. Then $P_O(MC) \leq P_U(MC)$ for any $\sigma_2 \geq \sigma_1 > 0$ and $0 \leq \alpha < \frac{1}{2}$.*

Proof The probability of correct classification for the usual rule R_U is

$$\begin{aligned} 1 - P_U(MC) &= \frac{1}{2}[P_U(1|1) + P_U(2|2)] \\ &= \frac{1}{2}[P_{\sigma_1}(X < \hat{x_0}, \hat{\sigma_1} \leq \hat{\sigma_2}) + P_{\sigma_1}(X \geq \hat{x_0}, \hat{\sigma_1} > \hat{\sigma_2})] \\ &\quad + \frac{1}{2}[P_{\sigma_2}(X > \hat{x_0}, \hat{\sigma_1} \leq \hat{\sigma_2}) + P_{\sigma_2}(X \leq \hat{x_0}, \hat{\sigma_1} > \hat{\sigma_2})], \end{aligned}$$

where P_{σ_1} refers to probability when X is from Π_1 and P_{σ_2} refers to probability when X is from Π_2. For any fixed α $(0 \leq \alpha \leq \frac{1}{2})$, the probability of correct classification for the ordered rule R_O^α is

$$\begin{aligned} 1 - P_O(MC) &= \frac{1}{2}[P_O(1|1) + P_O(2|2)] \\ &= \frac{1}{2}[P_{\sigma_1}(X \leq \tilde{x_0}) + P_{\sigma_2}(X \geq \tilde{x_0})] \\ &= \frac{1}{2}[P_{\sigma_1}(X \leq \tilde{x_0}, \hat{\sigma_1} \leq \hat{\sigma_2}) + P_{\sigma_1}(X \leq \tilde{x_0}, \hat{\sigma_1} > \hat{\sigma_2})] \\ &\quad + \frac{1}{2}[P_{\sigma_2}(X \geq \tilde{x_0}, \hat{\sigma_1} \leq \hat{\sigma_2}) + P_{\sigma_2}(X \geq \tilde{x_0}, \hat{\sigma_1} > \hat{\sigma_2})] \end{aligned}$$

We note that under the condition $\hat{\sigma_1} \leq \hat{\sigma_2}$, $\tilde{\sigma_1} = \hat{\sigma_1}$ and $\tilde{\sigma_2} = \hat{\sigma_2}$, which implies $\tilde{x_0} = \hat{x_0}$. The condition $1 - P_U(MC) \leq 1 - P_O(MC)$ is then equivalent to

$$\begin{aligned} &P_{\sigma_1}(X \leq \tilde{x_0}, \hat{\sigma_1} > \hat{\sigma_2}) + P_{\sigma_2}(X \geq \tilde{x_0}, \hat{\sigma_1} > \hat{\sigma_2}) \\ &\geq P_{\sigma_1}(X \geq \hat{x_0}, \hat{\sigma_1} > \hat{\sigma_2}) + P_{\sigma_2}(X \leq \hat{x_0}, \hat{\sigma_1} > \hat{\sigma_2}). \end{aligned} \tag{4}$$

We can write

$$
\begin{aligned}
&P_{\sigma_1}(X \le \tilde{x_0}, \hat{\sigma_1} > \hat{\sigma_2}) = P_{\sigma_1}(X \le \tilde{x_0}, \bar{X} > \bar{Y}) \\
&= \int\limits_{\bar{x}>\bar{y}} \int P_{\sigma_1}(X \le \tilde{x_0}) g_1(\bar{x}, \sigma_1) g_2(\bar{y}, \sigma_2)\, d\bar{x}\, d\bar{y} \\
&= \int\limits_{\bar{x}>\bar{y}} \int \left(1 - \exp(-\frac{\tilde{x_0}}{\sigma_1})\right) g_1(\bar{x}, \sigma_1) g_2(\bar{y}, \sigma_2)\, d\bar{x}\, d\bar{y}.
\end{aligned}
$$

Proceeding in a similar way, expressions for other terms in the inequality (4) can be evaluated and we obtain

$$
\begin{aligned}
&P_{\sigma_1}(X \le \tilde{x_0}, \hat{\sigma_1} > \hat{\sigma_2}) + P_{\sigma_2}(X \ge \tilde{x_0}, \hat{\sigma_1} > \hat{\sigma_2}) \\
&- P_{\sigma_1}(X \ge \hat{x_0}, \hat{\sigma_1} > \hat{\sigma_2}) - P_{\sigma_2}(X \le \hat{x_0}, \hat{\sigma_1} > \hat{\sigma_2}) \\
&= \int\limits_{\bar{x}>\bar{y}} \int \left(\exp(-\frac{\tilde{x_0}}{\sigma_2}) - \exp(-\frac{\tilde{x_0}}{\sigma_1}) + \exp(-\frac{\hat{x_0}}{\sigma_2}) - \exp(-\frac{\hat{x_0}}{\sigma_1})\right) \\
&\quad \times g_1(\bar{x}, \sigma_1) g_2(\bar{y}, \sigma_2)\, d\bar{x} d\bar{y},
\end{aligned}
$$

which is nonnegative since $\sigma_1 \le \sigma_2$. Hence ordered classification rule R_O^{α} is better than the usual classification rule R_U for all $\alpha \in [0, \frac{1}{2})$.

For each of the ordered classification rules R_O^{α}, we can also compare between individual probabilities of the correct classification. In fact for $\alpha = 0$, Conde et al. [5] proved that for $n_1 = n_2 = n$,

(i) if $n > 1$, $P_O(1|1) > P_U(1|1)$ for all $\sigma_2 \ge \sigma_1 > 0$.
(ii) if $n = 1$, $P_O(2|2) \ge P_U(2|2)$ for all $\sigma_2 \ge \sigma_1 > 0$.
(iii) if $n = 1$, there are δ_0 in (0,1) such that

$$P_O(1|1) \ge P_U(1|1) \text{ for all } \sigma_1, \sigma_2 > 0, \quad 0 < \rho \le \delta_0,$$

$$P_O(1|1) < P_U(1|1) \text{ for all } \sigma_1, \sigma_2 > 0, \quad \delta_0 < \rho < 1,$$

(iv) if $n > 1$, there are δ_0' and δ_1' in (0,1) such that

$$P_O(2|2) \ge P_U(2|2) \text{ for all } \sigma_1, \sigma_2 > 0, \quad 0 < \rho \le \delta_0',$$

$$P_O(2|2) < P_U(2|2) \text{ for all } \sigma_1, \sigma_2 > 0, \quad \delta_1' < \rho < 1.$$

where $\rho = \frac{\sigma_1}{\sigma_2}$.

For $\alpha \in (0, \frac{1}{2})$, we show using a Monte Carlo study that the above conclusions (i–iv) hold true.

Table 1 The probability of correct classification for $n_1 = n_2 = 1$

σ_1/σ_2	R_U	$R_O(0)$	$R_O(0.1)$	$R_O(0.2)$	$R_O(0.3)$	$R_O(0.4)$
0.05	0.749700	0.759750	0.760850	0.761750	0.762500	0.762650
	0.846200	0.888050	0.887900	0.887750	0.887600	0.887600
0.1	0.694600	0.707750	0.710850	0.712350	0.713200	0.713900
	0.761900	0.831450	0.830650	0.830250	0.830050	0.829850
0.2	0.622700	0.642450	0.648750	0.652200	0.654300	0.655350
	0.651700	0.757050	0.754650	0.753100	0.752200	0.751500
0.3	0.579150	0.599850	0.609450	0.613450	0.616550	0.618050
	0.592000	0.708300	0.704150	0.701300	0.699500	0.698400
0.4	0.551000	0.564350	0.577250	0.583150	0.587400	0.588900
	0.551250	0.672500	0.665450	0.661450	0.658400	0.657050
0.5	0.530650	0.537500	0.553100	0.560400	0.565100	0.567050
	0.525650	0.642400	0.632550	0.626850	0.622050	0.619400
0.99	0.501200	0.448450	0.476050	0.489250	0.496150	0.500250
	0.493400	0.546850	0.520050	0.506750	0.497900	0.493900

Table 2 The probability of correct classification for $n_1 = n_2 = 5$

σ_1/σ_2	R_U	$R_O(0)$	$R_O(0.1)$	$R_O(0.2)$	$R_O(0.3)$	$R_O(0.4)$
0.05	0.920750	0.920800	0.920800	0.920800	0.920800	0.920800
	0.862250	0.862300	0.862300	0.862300	0.862300	0.862300
0.1	0.879550	0.880050	0.880050	0.880050	0.880050	0.880050
	0.787150	0.787450	0.787450	0.787450	0.787450	0.787450
0.2	0.815000	0.819350	0.819350	0.819350	0.819450	0.819450
	0.680400	0.684650	0.684650	0.684650	0.684650	0.684650
0.3	0.753150	0.770750	0.770900	0.770950	0.771100	0.771100
	0.608000	0.619400	0.619250	0.619150	0.619150	0.619150
0.4	0.697150	0.732450	0.732650	0.732700	0.732850	0.732850
	0.552300	0.569600	0.569300	0.568950	0.568700	0.568700
0.5	0.653350	0.702450	0.703150	0.703450	0.704100	0.704300
	0.512700	0.528900	0.528550	0.528150	0.527900	0.567600
0.99	0.503700	0.591250	0.595900	0.599550	0.601650	0.602850
	0.492600	0.405150	0.400500	0.397600	0.396000	0.394500

3 Numerical Studies

We now investigate performances of the proposed class of classification rules with respect to the probabilities of correct classification. We use Monte Carlo simulation technique taking 20,000 replications. In Tables 1 and 2, we have taken the size of the training samples (1,1) and (5,5) respectively. First and second row of each pair of these tables represent $P(1|1)$ and $P(2|2)$ respectively. The following conclusions are drawn from the simulation study.

(a) When the size of the training samples are equal, the expected probability of correct classification corresponding to the rule R_O^α with $\alpha = 0$ is higher than that of R_O^α for all $\alpha > 0$.
(b) The ordered classification rules R_O^α are better than the usual classification rule R_U for all $\alpha \in [0, \frac{1}{2})$ with respect to expected probability of the correct classification.
(c) The behavior of the ordered classification rules is similar to that of the usual classification rule when the parameters σ_1 and σ_2 are sufficiently different.
(d) For $n_1 = n_2 = 1$, $P_O(2|2)$ is greater than or equal to $P_U(2|2)$ for all $0 < \sigma_1 \leq \sigma_2$ and for all $\alpha \in [0, \frac{1}{2})$.
(e) For $n_1 = n_2 > 1$, $P_O(1|1)$ is strictly greater than $P_U(1|1)$.
(f) As α decreases the corresponding classification rule is better than the usual classification rule in terms of expected probabilities of correct classification.
(g) When $n_1 = n_2 = n > 1$, then $P_O(1|1)$ is higher than $P_U(1|1)$.
(h) In case the population parameters σ_1 and σ_2 are very close, then for small size training samples as α $(0 \leq \alpha < \frac{1}{2})$ increases, $P_O(1|1)$ increases but $P_O(2|2)$ decreases.
(i) When σ_1 and σ_2 are far from each other then for the small sizes samples the performance of the usual classification rule is similar to each of the ordered classification rule.

Similar observations are made for various other values of n_1, n_2 and σ_1/σ_2.

4 Example

In this section, we consider an illustrative example to show the usefulness of the ordered classification rules proposed in this paper. Consider the example of Barlow et al. [2], (see page 270), where the operating times between successive failures of air conditioning equipment in two aircraft (plane 7916 and plane 7907) are given in terms of samples of size five.

- For plane 7916, sample values are 50, 254, 5, 283, 35.
- For plane 7907, sample values are 194, 15, 41, 29, 33.

The sample values shown to fit exponential distributions. Let us assume that due to a design change, the expected lifetime of the air conditioners of plane 7916 is less than the expected lifetime of the air conditioners used in plane 7907. Suppose an observation belonging to either of plane 7916 or that of 7907 is available. However, due to an error in record keeping, it is not clear whether the observation corresponds to plane 7916 or 7907. We use the proposed classification rules to classify the observations 12 and 181 into one of these two populations. The observation 12 is correctly classified to plane 7916 by the proposed ordered classification rule R_O^α for $\alpha \in [0, \frac{1}{2})$ but it will be misclassified under the usual classification rule. The observation 181 is correctly classified to plane 7907 by the proposed ordered classification rule R_O^α for $\alpha \in [0, \frac{1}{2})$ but it will be misclassified if the usual classification rule is used.

5 Conclusion

We have proposed classification procedures for exponential populations which are based on improved estimators of scale parameters when a priori information is available on ordering. It is shown that the new procedure is better than the classical classification rule in terms of expected probability of correct classification. A simulation study is also carried out to show numerically the superiority of new procedures.

References

1. Adegboye, O.S.: The optimal classification rule for exponential populations. Austral. J. Stat. **35**, 185–194 (1993)
2. Barlow, R.E., Bartholomew, D.J., Bremnner, J.M., Brunk, H.D.: Statistical Inference under Order Restrictions. The Theory and Application of Isotonic Regression. John Wiley, New York (1972)
3. Basu, A.P., Gupta, A.K.: Classification rules for exponential populations. In: Proschan F., Serfling R.J. (eds.) Reliability and Biometry: Statistical Analysis of Life Length, pp. 637–650. SIAM, Philadelphia (1974)
4. Basu, A.P., Gupta, A.K.: Classification Rules for Exponential Populations: Two Parameter Case. DTIC Document, North-Holland (1976)
5. Conde, D., Fernández, M.A., Salvador, B.: A classification rule for ordered exponential populations. J. Statist. Plann. Infer. **135**(2), 339–356 (2005)
6. Long, T., Gupta, R.D.: Alternative linear classification rules under order restrictions. Comm. Statist. Theory Methods **27**(3), 559–575 (1998)
7. McLachlan, G.: Discriminant Analysis and Statistical Pattern Recognition. John Wiley, New York (2004)
8. Pal, N., Jin, C., Lim, W.K.: Handbook of Exponential and Related Distributions for Engineers and Scientists. Chapman & Hall/CRC, New York (2006)
9. Rausand, M., Høyland, A.: System Reliability Theory: Models, Statistical Methods, and Applications. John Wiley, New York (2004)
10. Vijayasree, G., Singh, H.: Mixed estimators of two ordered exponential means. J. Statist. Plann. Infer. **35**(1), 47–53 (1993)
11. Wakaki, H., Aoshima, M.: Optimal discriminant functions for normal populations. J. Multivar. Anal. **100**(1), 58–69 (2009)

Chapter 17
Solving the Exterior Bernoulli Problem Using the Shape Derivative Approach

Jerico B. Bacani and Gunther Peichl

Abstract In this paper, we are interested in solving the exterior Bernoulli free boundary problem by minimizing a particular cost functional J over a class of admissible domains subject to two well-posed PDE constraints: a Dirichlet boundary value problem and a Neumann boundary value problem. The main result for this paper is the thorough computation of the first-order shape derivative of J using the shape derivatives of the state variables. At first, the material derivatives of the states are rigorously justified. Then the equation and the boundary conditions satisfied by the corresponding shape derivatives are derived directly from the definition of the shape derivative and the variational equation for the material derivative. It becomes apparent that the analysis of the shape derivatives of the states requires more regular domains. Finally, it is noted that the shape gradient agrees with the structure predicted by the Hadamard structure theorem.

Keywords Shape optimization · Free boundary problem · Overdetermined boundary value problem · Material derivative · Shape derivative

AMS Subject Classifications: 35R35, 35N25, 49K20, 49Q12

The results in this paper are made possible through the support of ÖAD—Austrian Agency for International Cooperation in Education and Research for the Technologiestipendien Südostasien (Doktorat) scholarship in the frame of the ASEA-UNINET. The results are first presented in [3]. The presentation in ICMC 2013 and publication of this paper are realized through the research dissemination grants given by University of the Philippines Baguio and University of the Philippines System. Special thanks to Prof. Gilbert Peralta for his helpful suggestions and to the referees of this manuscript.

J. B. Bacani (✉)
Department of Mathematics and Computer Science, College of Science, University of the Philippines Baguio, Governor Pack Road, 2600 Baguio, Philippines
e-mail: jicderivative@yahoo.com

G. Peichl
Institute for Mathematics and Scientific Computing, University of Graz, Heinrichstrasse 36, 8010 Graz, Austria
e-mail: gunther.peichl@uni-graz.at

R. N. Mohapatra et al. (eds.), *Mathematics and Computing 2013*, Springer Proceedings in Mathematics & Statistics 91,
DOI: 10.1007/978-81-322-1952-1_17, © Springer India 2014

1 Introduction

The paper deals with sensitivity analysis in a particular shape optimization formulation of exterior Bernoulli problems, which can be considered to be prototype of free boundary problems (FBPs). They represent mathematical models describing phenomena in different areas of physics, chemistry, medicine, industry such as phase transitions, flows through porous media, tumor growth, electrochemical machining, among others [1, 5, 6, 9, 10, 22]. Their common feature is the fact that equations are solved in domains which are not known a-priori; that is, the shape of domains is one of the unknowns. For this reason, numerical realization of FBPs is much more complicated compared with standard boundary value problems. Shape optimization approach is a possible technique which can be used for this purpose. The idea is simple: one of the boundary conditions (Dirichlet or Neumann) which have to be satisfied simultaneously on the free boundary in Bernoulli problems is removed from the system and is satisfied by minimizing an appropriate cost functional. After a discretization, we obtain a nonlinear mathematical programming problem. Minimization is usually carried out by using a gradient type method for which gradient information is needed. Therefore, sensitivity analysis is an integral part of any optimization problem. The main goal of this work is to perform sensitivity analysis for the continuous setting of the shape optimization formulation of the Bernoulli problem with the Kohn-Vogelius cost functional. In this approach, we first justify the existence of material derivatives of the states. Then the equation and the boundary conditions satisfied by their corresponding shape derivatives are derived directly from the definition of the shape derivative and the variational equation for the material derivative. This method is more tedious compared to differentiating directly the variational equation for the states but it is rigorous and avoids the formal arguments one often finds in the literature, for instance in the work of Fuji [11], Masanao and Fuji [18], and Simon [19].

The paper is outlined as follows. In Sect. 2, the exterior Bernoulli FBP is defined and reformulated as a shape optimization problem. Section 3 is a survey of tools in shape optimization which are used in subsequent parts of this paper. Shape variations are realized by deformations of admissible domains using a perturbation of identity mapping that has an appropriate regularity. Basic properties of such mapping are recalled. Further, the material and shape derivative of state variables are mentioned as well as results on differentiation of domain and boundary integrals with respect to the shape. Section 4 is devoted to the rigorous derivation of the first-order shape derivative of the above-mentioned cost functional. Conclusion is given in Sect. 5.

2 Shape Optimization Formulation of the Bernoulli Problem

The exterior Bernoulli free boundary problem can be reformulated as follows: Given a bounded and connected domain $A \subset \mathbb{R}^2$ with a fixed boundary $\Gamma := \partial A$ and a constant $\lambda < 0$, one needs to find a bounded connected domain $B \subset \mathbb{R}^2$ with a

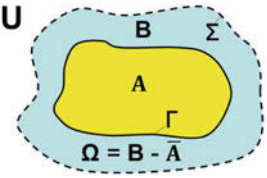


Fig. 1 The domain Ω for the exterior Bernoulli problem

free boundary Σ and containing the closure of A, and an associated state function $u : \Omega \to \mathbb{R}$, where $\Omega = B \backslash \bar{A}$, such that the overdetermined conditions are satisfied:

$$\begin{cases} -\Delta u = 0 \quad \text{in } \Omega, \\ u = 1 \quad \text{on } \Gamma, \\ u = 0 \quad \text{on } \Sigma, \\ \dfrac{\partial u}{\partial \mathbf{n}} = \lambda \quad \text{on } \Sigma. \end{cases} \tag{1}$$

Here, $\mathbf{n}$ is the outward unit normal vector to Σ. The domain for the exterior Bernoulli free boundary problem is illustrated in Fig. 1.

This boundary problem is ill-posed because of the presence of overdetermined conditions on the free boundary Σ. To overcome the difficulty of solving the Bernoulli problem, we reformulate it as a shape optimization problem:

$$\min_{\Omega} J(\Omega) \equiv \min_{\Omega} \frac{1}{2} \int_{\Omega} |\nabla (u_D - u_N)|^2 \, \mathrm{d}x \tag{2}$$

over all admissible domains Ω, where the state function u_D is the solution to the Dirichlet problem:

$$\begin{cases} -\Delta u_D = 0 \text{ in } \Omega, \\ u_D = 1 \quad \text{on } \Gamma, \\ u_D = 0 \quad \text{on } \Sigma, \end{cases} \tag{3}$$

and the state function u_N is the solution to the Neumann problem:

$$\begin{cases} -\Delta u_N = 0 \text{ in } \Omega, \\ u_N = 1 \quad \text{on } \Gamma, \\ \dfrac{\partial u_N}{\partial \mathbf{n}} = \lambda \quad \text{on } \Sigma. \end{cases} \tag{4}$$

The functional J is introduced by Kohn and Vogelius in the context of inverse problems, and so this functional is named after them [16]. Notice that if we find Ω such that u_N happens to be identical to u_D (that is, $u_N = u_D = u$), then the overdetermined conditions (1) are satisfied.

3 Tools from Shape Calculus

In this work, we are interested in $C^{k,1}$-domains or domains with $C^{k,1}$ boundaries, where $k \geq 0$. Aside from being $C^{k,1}$, we also assume that these are bounded and connected subsets of a bigger set U called the *hold-all domain.*

The reference domain under consideration is a bounded, connected annulus with disjoint boundaries Γ and Σ. This domain is perturbed via the perturbation of identity operator

$$T_t : \bar{U} \to \mathbb{R}^2, \qquad T_t(x) = x + t\mathbf{V}(x), \quad x \in \bar{U}, \tag{5}$$

where $\mathbf{V}$ belongs to Θ, which is defined as

$$\Theta = \left\{ \mathbf{V} \in C^{1,1}(\bar{U}, \mathbb{R}^2) : \mathbf{V}|_{\Gamma \cup \partial U} = 0 \right\}. \tag{6}$$

One can show that the reference Ω and the perturbed domain Ω_t have the same topological structure and regularity under the transformation T_t. These properties are given as a theorem and a corollary and can be seen in [4].

Theorem 1 *Let Ω and U be nonempty bounded open connected subsets of $\mathbb{R}^2$ with Lipschitz continuous boundaries, such that $\bar{\Omega} \subseteq U$, and $\partial\Omega$ is the union of two disjoint boundaries Γ and Σ. Let T_t be defined as in (5) where $\mathbf{V}$ belongs to Θ, defined as (6). Then for sufficiently small t,*

1. *$T_t : \bar{U} \to \bar{U}$ is a homeomorphism,*
2. *$T_t : U \to U$ is a $C^{1,1}$ diffeomorphism, and in particular, $T_t : \Omega \to \Omega_t$ is a $C^{1,1}$ diffeomorphism,*
3. *$\Gamma_t = T_t(\Gamma) = \Gamma$, and*
4. *$\partial\Omega_t = \Gamma \cup T_t(\Sigma)$.*

Corollary 1 *Let Ω and U be two domains in $\mathbb{R}^2$ with $C^{1,1}$ boundary. Then for sufficiently small t,the perturbed domain $\Omega_t := T_t(\Omega)$ is also of class $C^{1,1}$.*

In this paper, we use the following notations:

$$\begin{cases} I_t(x) = \det DT_t(x), & x \in \bar{U}, \\ M_t(x) = (DT_t(x))^{-T}, & x \in \bar{U}, \\ A_t(x) = I_t(x) M_t^T(x) M_t(x), & x \in \bar{U}, \\ w_t(x) = I_t(x) |(DT_t(x))^{-T} \mathbf{n}(x)|, & x \in \Sigma. \end{cases} \tag{7}$$

We now enumerate several properties of T_t that are used in the analysis.

Lemma 1 [12, 15] *Consider the transformation T_t, where the fixed vector field $\mathbf{V}$ belongs to Θ, defined in (6). Then there exists $t_V > 0$ such that T_t and the functions in (7) restricted to the interval $I_V = (-t_V, t_V)$ have the following regularity and properties:*

1. $t \mapsto T_t \in C^1(I_V, C^{1,1}(\bar{U}, \mathbb{R}^2))$.
2. $t \mapsto I_t \in C^1(I_V, C^{0,1}(\bar{U}))$.
3. $t \mapsto T_t^{-1} \in C(I_V, C^1(\bar{U}, \mathbb{R}^2))$
4. $t \mapsto w_t \in C^1(I_V, C(\Sigma))$.
5. $t \mapsto A_t \in C(I_V, C(\bar{U}, \mathbb{R}^{2\times 2}))$.
6. *There is $\beta > 0$ such that $A_t(x) \geq \beta I$ for $x \in U$.*
7. $\dfrac{\mathrm{d}}{\mathrm{d}t} T_t|_{t=0} = \mathbf{V}$.
8. $\dfrac{\mathrm{d}}{\mathrm{d}t} T_t^{-1}|_{t=0} = -\mathbf{V}$.
9. $\dfrac{\mathrm{d}}{\mathrm{d}t} DT_t|_{t=0} = D\mathbf{V}$.
10. $\dfrac{\mathrm{d}}{\mathrm{d}t} (DT_t)^{-1}|_{t=0} = -D\mathbf{V}$.
11. $\dfrac{\mathrm{d}}{\mathrm{d}t} I_t|_{t=0} = \mathrm{div}\mathbf{V}$.
12. $\dfrac{\mathrm{d}}{\mathrm{d}t} A_t|_{t=0} = (\mathrm{div}\mathbf{V})I - (D\mathbf{V} + (D\mathbf{V})^T) \equiv A$.
13. $lim_{t\to 0} w_t = 1$.
14. $\dfrac{\mathrm{d}}{\mathrm{d}t} w_t|_{t=0} = \mathrm{div}_\Sigma \mathbf{V}$, *where* div_Σ *is defined by*

$$\mathrm{div}_\Sigma \mathbf{V} = \mathrm{div}\mathbf{V}|_\Sigma - (D\mathbf{V}\mathbf{n}) \cdot \mathbf{n}.$$

3.1 Material and Shape Derivatives

The material and shape derivatives of state variables are defined as follows [13, 21]:

Definition 1 Let u be defined in $[0, t_V] \times U$. An element $\dot{u} \in H^k(\Omega)$, called the *material derivative* of u, is defined as

$$\dot{u}(x) := \dot{u}(\Omega; \mathbf{V}) = \lim_{t\to 0^+} \frac{u(t, T_t(x)) - u(0, x)}{t} = \frac{\mathrm{d}}{\mathrm{d}t} u(t, x + t\mathbf{V}(x))\Big|_{t=0}$$

if the limit exists in $(H^k(\Omega))$.

Remark 1 The material derivative can be written as

$$\dot{u}(x) = \lim_{t\to 0^+} \frac{u_t \circ T_t(x) - u(x)}{t} = \frac{\mathrm{d}}{\mathrm{d}t} (u_t \circ T_t(x))\Big|_{t=0}. \tag{8}$$

It characterizes the behavior of the function u at $x \in \Omega \subset U$ in the direction $\mathbf{V}(x)$.

Definition 2 Let u be defined in $[0, t_V] \times U$. An element $u' \in H^k(\Omega)$ is called the *shape derivative* of u at Ω in the direction $\mathbf{V}$, if the following limit exists in $H^k(\Omega)$:

$$u'(x) := u'(\Omega; \mathbf{V}) = \lim_{t \to 0^+} \frac{u(t, x) - u(0, x)}{t}. \tag{9}$$

Remark 2 If $\dot{u}$ and $\nabla u \cdot V$ exist in $H^k(\Omega)$ then the shape derivative can be written as

$$u'(x) = \dot{u}(x) - (\nabla u \cdot \mathbf{V})(x). \tag{10}$$

In general, if $\dot{u}(x)$ and $\nabla u \cdot \mathbf{V}(x)$ both exist in $W^{m,p}(\Omega)$, then $u'(x)$ also exists in that space.

Remark 3 Definitions 1 and 2 are still valid if Ω is replaced by $\partial\Omega$.

3.2 Domain and Boundary Transformations

Lemma 2 [20]

1. *Let $\varphi_t \in L^1(\Omega_t)$. Then $\varphi_t \circ T_t \in L^1(\Omega)$ and*

$$\int_{\Omega_t} \varphi_t \, \mathrm{d}x_t = \int_{\Omega} \varphi_t \circ T_t I_t \, \mathrm{d}x$$

2. *Let $\varphi_t \in L^1(\partial\Omega_t)$. Then $\varphi_t \circ T_t \in L^1(\partial\Omega)$ and*

$$\int_{\partial\Omega_t} \varphi_t \, \mathrm{d}s_t = \int_{\partial\Omega} \varphi_t \circ T_t w_t \, \mathrm{d}s$$

where I_t and w_t are defined in (7).

Proofs can be found in [15, 20].

Remark 4 The function $u_t : \Omega_t \to \mathbb{R}$ can be referred to the reference domain by composing u_t with T_t; that is,

$$u^t = u_t \circ T_t : \Omega \to \mathbb{R}$$

and by chain rule of differentiation, we get

$$(\nabla u_t) \circ T_t = (DT_t)^{-T} \nabla u^t = M_t \nabla u^t. \tag{11}$$

3.3 Some Tools in Tangential Calculus

Here are some properties of tangential differential operators which we used in this work (cf. [2, 7, 14, 20]). Let Γ be a boundary of a bounded domain $\Omega \subset \mathbb{R}^n$.

Definition 3 The *tangential gradient* of $f \in C^1(\Gamma)$ is given by

$$\nabla_\Gamma f := \nabla F|_\Gamma - \frac{\partial F}{\partial \mathbf{n}}\mathbf{n} \in C(\Gamma, \mathbb{R}^n), \tag{12}$$

where F is any C^1 the extension of f into a neighborhood of Γ.

Definition 4 The *tangential Jacobian matrix* of a vector function $\mathbf{v} \in C^1(\Gamma, \mathbb{R}^n)$ is given by

$$D_\Gamma \mathbf{v} = D\mathbf{V}|_\Gamma - (D\mathbf{V}\mathbf{n})\mathbf{n}^T \in C(\Gamma, \mathbb{R}^{n\times n}), \tag{13}$$

where $\mathbf{V}$ is any C^1 the extension of $\mathbf{v}$ into a neighborhood of Γ.

Definition 5 For a vector function $\mathbf{v} \in C^1(\Gamma, \mathbb{R}^n)$, its *tangential divergence* on Γ is given by

$$\mathrm{div}_\Gamma \mathbf{v} = \mathrm{div}\mathbf{V}|_\Gamma - D\mathbf{V}\mathbf{n}\cdot\mathbf{n} \ \in C(\Gamma), = tr[D\mathbf{V}|_\Gamma - (D\mathbf{V}\mathbf{n})\mathbf{n}^T] \tag{14}$$

where $\mathbf{V}$ is any C^1 the extension of $\mathbf{v}$ into a neighborhood of Γ.

Remark 5 The details of the existence of the extension F and $\mathbf{V}$ can be found in [7, pp. 361–366]. We note that Definitions 3, 4, and 5 do not depend on the choice of the extension, cf. [20, pp. 82–83].

We now provide some useful identities in tangential calculus.

Lemma 3 [20] *Consider a C^2 domain Ω with boundary $\Gamma := \partial\Omega$. Then for $u \in H^1(\Gamma)$ and $\mathbf{V} \in C^1(\Gamma, \mathbb{R}^n)$ the following identities hold:*

$$(1) \quad \mathrm{div}_\Gamma(u\mathbf{V}) = \nabla_\Gamma u \cdot \mathbf{V} + u\mathrm{div}_\Gamma \mathbf{V} \tag{15}$$

$$(2) \quad \int_\Gamma \mathrm{div}_\Gamma \mathbf{V}\, ds = \int_\Gamma \kappa \mathbf{V}\cdot\mathbf{n}\, ds \tag{16}$$

$$(3) \quad \int_\Gamma (u\mathrm{div}_\Gamma \mathbf{V} + \nabla_\Gamma u \cdot \mathbf{V})\, ds = \int_\Gamma \kappa u \mathbf{V}\cdot\mathbf{n}\, ds \tag{17}$$

$$(4) \quad \int_\Gamma \nabla_\Gamma u \cdot \mathbf{V}\, ds = -\int_\Gamma u\mathrm{div}_\Gamma \mathbf{V}\, ds, \ \text{where } \mathbf{V}\cdot\mathbf{n} = 0. \tag{18}$$

Remark 6 In Lemma 3, the first identity is the *tangential divergence formula*, the second is the *tangential Stoke's formula*, and the third is the *tangential Green's formula*. This lemma can also be shown to be true for $C^{1,1}$ domains.

3.4 Domain and Boundary Differentiation

The following are the formulas for the derivatives of integrals with respect to the domain of integration. For the first theorem, it is sufficient to have at least $C^{0,1}$ domains while the second theorem requires at least $C^{1,1}$ domains. For proof, see [20].

Theorem 2 (Domain Differentiation Formula) *Let $u \in C(I_V, W^{1,1}(U))$ and suppose $\dot{u}(0,\cdot) := \frac{d}{dt}u(t, T_t(\cdot))\big|_{t=0}$ exists in $L^1(U)$. Then*

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\Omega_t} u(t,x)\,\mathrm{d}x\bigg|_{t=0} = \int_{\Omega} u'(0,x)\,\mathrm{d}x + \int_{\Sigma} u(0,s)\mathbf{V}\cdot\mathbf{n}\,\mathrm{d}s \tag{19}$$

Theorem 3 (Boundary Differentiation Formula) *Let u be defined in a neighborhood of Σ. If $u \in C(I_V, W^{2,1}(U))$ and $\dot{u}(0,\cdot) \in W^{1,1}(U)$, then*

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\Sigma_t} u(t,s)\,\mathrm{d}s\bigg|_{t=0} = \int_{\Sigma} u'(0,s)\,\mathrm{d}s + \int_{\Sigma}\Big(\frac{\partial u}{\partial \mathbf{n}} + u(0,s)\kappa\Big)\mathbf{V}\cdot\mathbf{n}\,\mathrm{d}s, \tag{20}$$

where κ is the mean curvature of the free boundary Σ.

3.5 The First-Order Eulerian Derivative

Definition 6 The *Eulerian derivative* of the shape functional $J : \Omega \to \mathbb{R}$ defined in (2) at the domain Ω in the direction of the deformation field $\mathbf{V} \in \Theta$ is given by

$$\mathrm{d}J(\Omega;\mathbf{V}) := \lim_{t\to 0^+}\frac{J(\Omega_t) - J(\Omega)}{t}, \tag{21}$$

if the limit exists.

Remark 7 J is said to be shape differentiable at Ω if $dJ(\Omega;\mathbf{V})$ exists for all $\mathbf{V} \in \Theta$ and is linear and continuous with respect to $\mathbf{V}$.

4 Main Result

In this section, we present a rigorous derivation of the Eulerian shape derivative of the Kohn-Vogelius functional J by employing the shape derivatives of the states. First, due to its high importance in this study, we recall our result in [4] regarding the higher regularity of solutions to the PDEs (3) and (4):

Theorem 4 *Let Ω be a bounded domain with boundary of class $C^{1,1}$. Let $u_D, u_N \in H^1(\Omega)$ be weak solutions of the BVPs (3) and (4), respectively. Then u_D and u_N also belong to $H^2(\Omega)$. More generally, if Ω is of class $C^{k+1,1}$ then u_D and u_N are elements of $H^{k+2}(\Omega)$.*

Since, the existence of the shape derivatives of the states is based on the existence of the material derivatives, the latter is proven first.

4.1 Material Derivatives of States

4.1.1 Material Derivative of u_D

We first show the existence of the material derivative of u_D. One can show that $y^t = u_D^t - u_D \in H_0^1(\Omega)$ is a unique solution to

$$(A_t \nabla y^t, \nabla \varphi)_\Omega = -(A_t \nabla u_D, \nabla \varphi)_\Omega, \qquad \forall \varphi \in H_0^1(\Omega), \tag{22}$$

where A_t is given by (7) and u_D^t is the unique solution of the following variational equation:

$$(A_t \nabla u_D^t, \nabla \varphi)_\Omega = 0 \tag{23}$$

for all $\varphi \in H_0^1(\Omega)$, $u_D^t = 1$ on Γ and $u_D^t = 0$ on Σ. The bilinear form $b_t(y^t, \varphi) = \int_\Omega A_t \nabla y^t \cdot \nabla \varphi$ for all $y^t, \varphi \in H_0^1(\Omega)$ is continuous and coercive. With these characteristics, one can show that

$$|y^t|_{H_0^1(\Omega)} \leq 2|A_t|_\infty |u_D|_{H^1(\Omega)}.$$

Therefore, the set $\{y^t = u_D^t - u_D : t \in (0, t_V)\}$ is bounded in $H_0^1(\Omega)$ for sufficiently small t_V.

Note that the variational form of the Dirichlet problem (3) is given by: Find $u_D \in H^1(\Omega)$ such that

$$\begin{cases} (\nabla u_D, \nabla \psi)_\Omega = 0 & \forall \psi \in H_0^1(\Omega) \\ u_D = 1 & \text{on } \Gamma, \\ u_D = 0 & \text{on } \Sigma. \end{cases} \tag{24}$$

Using (24) we write (22) as

$$((A_t - I)\nabla y^t, \nabla \varphi)_\Omega + (\nabla y^t, \nabla \varphi)_\Omega = -((A_t - I)\nabla u_D, \nabla \varphi)_\Omega$$

Define $z^t = \frac{1}{t} y^t$ which also belongs to $H_0^1(\Omega)$. Then we have

$$(\nabla z^t, \nabla \varphi)_\Omega = -\left(\frac{1}{t}(A_t - I)\nabla y^t, \nabla \varphi\right)_\Omega - \left(\frac{1}{t}(A_t - I)\nabla u_D, \nabla \varphi\right)_\Omega. \tag{25}$$

Now we choose a sequence $\{t_n\}$ such that $\lim_{n\to\infty} t_n = 0$, and we want to show that $\lim_{n\to\infty} z^{t_n}$ exists.

Lemma 1, together with the boundedness of y^{t_n} in $H_0^1(\Omega)$ implies that ∇z^{t_n} is bounded in $L^2(\Omega; \mathbb{R}^2)$, equivalently that z^{t_n} is bounded in $H_0^1(\Omega)$. Thus there is a subsequence, which we still denote by t_n with $t_n \to 0$ and an element $z \in H_0^1(\Omega)$ such that $z^{t_n} \rightharpoonup z$ weakly in $H_0^1(\Omega)$. Since $\nabla u_D^{t_n} \to \nabla u_D$ in $L^2(\Omega; \mathbb{R}^2)$, $\lim_{t_n\to 0} A_{t_n} = I$ uniformly on $\bar{\Omega}$, and using property 12 of Lemma 1 we get

$$(\nabla z, \nabla \varphi)_\Omega = -(A\nabla u_D, \nabla \varphi)_\Omega \qquad \varphi \in H_0^1(\Omega). \tag{26}$$

Since this equation has a unique solution, we deduce that $z^{t_n} \rightharpoonup z$ for any sequence $\{t_n\}$. In order to show strong convergence we show $\lim_{n\to\infty} |z^{t_n}|_{H_0^1(\Omega)} = |z|_{H_0^1(\Omega)}$. This follows from:

$$\begin{aligned}\lim_{t_n\to 0} |z^{t_n}|^2_{H_0^1(\Omega)} &= -\lim_{t_n\to 0}\left(\frac{1}{t_n}(A_{t_n} - I)\nabla(u_D^{t_n} - u_D), \nabla z^{t_n}\right)_\Omega \\ &\quad - \lim_{t_n\to 0}\frac{1}{t_n}\left((A_{t_n} - I)\nabla u_D, \nabla z^{t_n}\right)_\Omega \\ &= -(A\nabla u_D, \nabla z)_\Omega = (\nabla z, \nabla z)_\Omega = |z|^2_{H_0^1(\Omega)}.\end{aligned}$$

This, together with the weak convergence, implies that z^{t_n} strongly converges to z in $H_0^1(\Omega)$.

4.1.2 Material Derivative of u_N

The variational form of the Neumann problem (4) is formulated as follows.

Find $u_N \in H^1(\Omega)$ such that

$$\begin{cases} (\nabla u_N, \nabla \varphi)_\Omega - (\lambda, \varphi)_\Sigma = 0 & \forall \varphi \in H^1_{\Gamma,0}(\Omega) \\ u_N = 1 & \text{on } \Gamma. \end{cases} \tag{27}$$

It is well-known that (27) has a unique solution $u_N \in H^1_{\Gamma,1}(\Omega)$, where the space $H^1_{\Gamma,1}(\Omega)$ is defined as

$$H^1_{\Gamma,1}(\Omega) = \left\{\varphi \in H^1(\Omega) : \varphi|_\Gamma = 1\right\}.$$

Also, the state $u_N^t \in H^1_{\Gamma,1}(\Omega)$ uniquely solves the variational equation

$$(A_t \nabla u_N^t, \nabla \phi)_\Omega - (w_t \lambda, \phi)_\Sigma = 0, \qquad \forall \phi \in H^1_{\Gamma,0}(\Omega), \tag{28}$$

where $u_N^t = 1$ on Γ. Subtracting (27) from (28) for all $\varphi \in H^1_{\Gamma,0}(\Omega)$, we have

$$\begin{aligned} 0 =& (A_t \nabla u_N^t, \nabla \varphi)_\Omega - (w_t \lambda, \varphi)_\Sigma - (\nabla u_N, \nabla \varphi)_\Omega + (\lambda, \varphi)_\Sigma \\ =& (A_t \nabla u_N^t - \nabla u_N^t + \nabla u_N^t - \nabla u_N, \varphi)_\Omega + (w_t \lambda - \lambda, \varphi)_\Sigma. \end{aligned} \tag{29}$$

Hence we have a unique solution $u_N^t - u_N \in H^1_{\Gamma,0}(\Omega)$ to

$$(\nabla(u_N^t - u_N), \nabla \varphi)_\Omega = -((A_t - I)\nabla u_N^t, \nabla \varphi)_\Omega + \lambda(w_t - 1, \varphi)_\Sigma \qquad \forall \varphi \in H^1_{\Gamma,0}(\Omega). \tag{30}$$

One can show that ∇u_N^t is uniformly bounded in $L^2(\Omega; \mathbb{R}^2)$ and that $\nabla u_N^t \to \nabla u_N$ in that space. Defining $y^t = \frac{1}{t}(u_N^t - u_N)$ we find that y^t satisfies

$$(\nabla y^t, \nabla \varphi)_\Omega = -\left(\left(\frac{A_t - I}{t}\right) \nabla u_N^t, \nabla \varphi\right)_\Omega + \lambda \left(\frac{w_t - 1}{t}, \varphi\right)_\Sigma \quad \forall \varphi \in H^1_{\Gamma,0}(\Omega). \tag{31}$$

Choose a sequence $\{t_n\}$ with $\lim_{n\to\infty} t_n = 0$. As in the Dirichlet case, we want to show here that $\lim_{n\to\infty} y^{t_n}$ exists. Since $\frac{1}{t_n}(A_{t_n} - I)$ and $\frac{1}{t_n}(w_{t_n} - 1)$ are bounded in L^∞, $\nabla u_N^{t_n}$ is bounded in $L^2(\Omega; \mathbb{R}^2)$, and we deduce that y^{t_n} is bounded in $H^1_{\Gamma,0}(\Omega)$. Hence there exists a subsequence, which we still denote as $\{t_n\}$, and this tends to zero. Furthermore, there is an element $y \in H^1_{\Gamma,0}(\Omega)$ such that $y^{t_n} \rightharpoonup y$ weakly in $H^1_{\Gamma,0}(\Omega)$. Considering $\nabla u_N^{t_n} \to \nabla u_N$ in $L^2(\Omega; \mathbb{R}^2)$, and applying properties 12 and 14 of Lemma 1 we obtain

$$(\nabla y, \nabla \varphi)_\Omega = -(A \nabla u_N, \nabla \varphi)_\Omega + \lambda(\mathrm{div}_\Sigma \mathbf{V}, \varphi)_\Sigma. \tag{32}$$

Since this equation has a unique solution, we deduce that $y^{t_n} \rightharpoonup y$ for any sequence $\{t_n\}$. This implies that y^{t_n} converges strongly to y in $L^2(\Sigma)$. Now choosing $\varphi = y^{t_n} \in H^1_{\Gamma,0}(\Omega)$ yields the following:

$$\begin{aligned} \lim_{t\to 0} |y^{t_n}|^2_{H^1(\Omega)} &= -\lim_{t_n \to 0} \left(\left(\frac{A_{t_n} - I}{t_n}\right) \nabla u_N^{t_n}, \nabla y^{t_n}\right)_\Omega + \lambda \lim_{t_n \to 0} \left(\frac{w_{t_n} - 1}{t_n}, y^{t_n}\right)_\Sigma \\ &= -(A \nabla u_N, \nabla y)_\Omega + \lambda(\mathrm{div}_\Sigma \mathbf{V}, y)_\Sigma = (\nabla y, \nabla y)_\Omega = |y|^2_{H^1(\Omega)}. \end{aligned}$$

The convergence in norm and the weak convergence of y^{t_n} in $H^1_{\Gamma,0}(\Omega)$ justifies the strong convergence of y^{t_n} to y in that space.

4.2 Shape Derivatives of States

The boundary value problems satisfied by the shape derivatives of the state variables can be derived in a rigorous manner. The approach that is presented here does not utilize the domain and boundary differentiation formulas which are usually employed when the derivation is done formally.

4.2.1 Shape Derivative of u_D

Theorem 5 *Let Ω be a $C^{2,1}$ bounded domain. The shape derivative of the state variable $u_D \in H^3(\Omega)$ satisfying the pure Dirichlet problem (3) is a solution to the following nonhomogeneous Dirichlet boundary value problem:*

$$\begin{cases} -\Delta u_D' = 0 & \text{in } \Omega, \\ u_D' = 0 & \text{on } \Gamma, \\ u_D' = -\dfrac{\partial u_D}{\partial \mathbf{n}} \mathbf{V} \cdot \mathbf{n} & \text{on } \Sigma. \end{cases} \tag{33}$$

Proof We have shown in Sect. 4.1 that $\dot{u}_D := z$ exists in $H_0^1(\Omega)$ and satisfies

$$(\nabla z, \nabla \varphi)_\Omega = -(A \nabla u_D, \nabla \varphi)_\Omega, \qquad \varphi \in H_0^1(\Omega), \tag{34}$$

where u_D satisfies (3) and A is given by property 12 of Lemma 1. By (10), we can write u_D' as $u_D' = \dot{u}_D - \nabla u_D \cdot \mathbf{V}$. Hence $u_D' = \dot{u}_D - \dfrac{\partial u_D}{\partial \mathbf{n}} \mathbf{V} \cdot \mathbf{n}$. Since $\mathbf{V}$ vanishes on Γ, $u_D' = 0$ on Γ. On the free boundary, $\dot{u}_D = 0$, thus $u_D' = -\dfrac{\partial u_D}{\partial \mathbf{n}} \mathbf{V} \cdot \mathbf{n}$.

Now we determine the variational equation satisfied by u_D'. Using the relationship between the material and shape derivatives, we write

$$(\nabla z, \nabla \varphi)_\Omega = (\nabla u_D', \nabla \varphi)_\Omega + (\nabla(\nabla u_D \cdot \mathbf{V}), \nabla \varphi)_\Omega, \tag{35}$$

which is valid for all $\varphi \in H_0^1(\Omega)$. Applying (34) we get

$$-(A \nabla u_D, \nabla \varphi)_\Omega = (\nabla u_D', \nabla \varphi)_\Omega + (\nabla(\nabla u_D \cdot \mathbf{V}), \nabla \varphi)_\Omega, \qquad \forall \varphi \in H_0^1(\Omega). \tag{36}$$

At first we choose $\varphi \in H^2(\Omega) \cap H_0^1(\Omega)$. Using the following identity (from [4]):

$$\int_\Omega A\nabla u\cdot\nabla v=\int_\Omega \Delta u(\mathbf{V}\cdot\nabla v)+\int_\Omega \Delta v(\mathbf{V}\cdot\nabla u)-\int_\Sigma \frac{\partial v}{\partial \mathbf{n}}(\mathbf{V}\cdot\nabla u)$$
$$-\int_\Sigma \frac{\partial u}{\partial \mathbf{n}}(\mathbf{V}\cdot\nabla v)+\int_\Sigma (\nabla u\cdot\nabla v)\mathbf{V}\cdot\mathbf{n}.$$

wherein we replace u by u_D and v by φ, we obtain

$$-\int_\Omega A\nabla u_D\cdot\nabla\varphi=-\int_\Omega \Delta\varphi(\mathbf{V}\cdot\nabla u_D)+\int_\Sigma \frac{\partial\varphi}{\partial\mathbf{n}}(\mathbf{V}\cdot\nabla u_D),\quad \forall\varphi\in H^2(\Omega)\cap H_0^1(\Omega).$$

By applying Green's formula we obtain

$$-\int_\Omega A\nabla u_D\cdot\nabla\varphi=\int_\Omega \nabla\varphi\cdot\nabla(\mathbf{V}\cdot\nabla u_D),\quad \varphi\in H^2(\Omega)\cap H_0^1(\Omega). \tag{37}$$

Substituting (37) into (36), we obtain $(\nabla u'_D,\nabla\varphi)_\Omega=0$, where $\varphi\in H^2(\Omega)\cap H_0^1(\Omega)$. Using Green's formula, we have $(-\Delta u'_D,\varphi)_\Omega=0$. But the functions in $H^2(\Omega)\cap H_0^1(\Omega)$ are dense in $L^2(\Omega)$. Therefore $-\Delta u'_D=0$ in Ω. In summary, we have shown that u'_D satisfies the boundary value problem (33).

Theorem 4 tells us that the solution to (3) indeed belongs to $H^2(\Omega)$. However this regularity of the solution is not sufficient to justify the existence of the shape derivative of u_D satisfying (33). We need higher regularity of the solution. So we take $C^{2,1}$ bounded domains and by the same theorem, u_D belongs to $H^3(\Omega)$.

4.2.2 Shape Derivative of u_N

Theorem 6 *Let Ω be a bounded $C^{2,1}$ domain. The shape derivative of the state variable $u_N\in H^3(\Omega)$ satisfying the Neumann problem (4) is a solution to the following mixed boundary value problem:*

$$\begin{cases}-\Delta u'_N=0 & \text{in } \Omega,\\ u'_N=0 & \text{on } \Gamma,\\ \dfrac{\partial u'_N}{\partial\mathbf{n}}=\operatorname{div}_\Sigma(\mathbf{V}\cdot\mathbf{n}\nabla_\Sigma u_N)+\kappa\lambda\mathbf{V}\cdot\mathbf{n} & \text{on } \Sigma.\end{cases} \tag{38}$$

Proof It was shown in Sect. 4.1 that the material derivative $w:=\dot{u}_N\in H^1_{\Gamma,0}(\Omega)$ satisfies

$$(\nabla w,\nabla\varphi)_\Omega=-(A\nabla u_N,\nabla\varphi)+\lambda\int_\Sigma \varphi\operatorname{div}_\Sigma\mathbf{V},\qquad \varphi\in H^1_{\Gamma,0}(\Omega). \tag{39}$$

First we note that u_N and u_N^t are both functions in $H^1_{\Gamma,1}(\Omega)$. Hence $\frac{1}{t}(u_N^t - u_N)$ belongs to $H^1_{\Gamma,0}(\Omega)$ for sufficiently small t. Thus $\dot{u}_N = 0$ on Γ.

Applying Green's formula to (39), we get

$$(-\Delta w, \varphi)_\Omega + \int_\Sigma \frac{\partial w}{\partial \mathbf{n}}\varphi = (\mathrm{div}(A\nabla u_N), \varphi)_\Omega - \int_\Sigma \varphi A\nabla u_N \cdot \mathbf{n} + \lambda \int_\Sigma \varphi \mathrm{div}_\Sigma \mathbf{V}$$

At first, we choose $\varphi \in H^1_0(\Omega)$, so we get

$$(-\Delta w, \varphi)_\Omega = (\mathrm{div}(A\nabla u_N), \varphi)_\Omega.$$

Since $H^1_0(\Omega)$ is dense in $L^2(\Omega)$ we have $-\Delta w = \mathrm{div}(A\nabla u_N)$ in Ω. Then we choose $\varphi \in H^1_{\Gamma,0}(\Omega)$ such that φ is arbitrary on Σ, to get

$$\int_\Sigma \frac{\partial w}{\partial \mathbf{n}}\varphi = -\int_\Sigma \varphi A\nabla u_N \cdot \mathbf{n} + \lambda \int_\Sigma \varphi \mathrm{div}_\Sigma \mathbf{V}.$$

Since the traces of functions $\varphi \in H^1_{\Gamma,0}(\Omega)$ are dense in $L^2(\Sigma)$ we obtain

$$\frac{\partial w}{\partial \mathbf{n}} = -A\nabla u_N \cdot \mathbf{n} + \lambda \mathrm{div}_\Sigma \mathbf{V}.$$

Therefore, w satisfies the following boundary value problem:

$$\begin{cases} -\Delta w = \mathrm{div}(A\nabla u_N) & \text{in } \Omega, \\ w = 0 & \text{on } \Gamma, \\ \dfrac{\partial w}{\partial \mathbf{n}} = -A\nabla u_N \cdot \mathbf{n} + \lambda \mathrm{div}_\Sigma \mathbf{V} & \text{on } \Sigma. \end{cases} \tag{40}$$

Next we consider $\varphi \in H^2(\Omega)$. This time we also consider $u_N \in H^2(\Omega)$. Applying the identity (37) for u_N and φ we obtain:

$$\int_\Omega A\nabla u_N \cdot \nabla\varphi = \int_\Omega \Delta u_N (\mathbf{V}\cdot\nabla\varphi) + \int_\Omega \Delta\varphi(\mathbf{V}\cdot\nabla u_N) - \int_\Sigma \frac{\partial\varphi}{\partial\mathbf{n}}(\mathbf{V}\cdot\nabla u_N) - \int_\Sigma \frac{\partial u_N}{\partial\mathbf{n}}(\mathbf{V}\cdot\nabla\varphi) + \int_\Sigma (\nabla u_N\cdot\nabla\varphi)\mathbf{V}\cdot\mathbf{n}.$$

Since $-\Delta u_N = 0$ in Ω and applying Green's theorem we obtain

$$-\int_{\Omega} A\nabla u_N\cdot\nabla\varphi=\int_{\Omega}\nabla\varphi\cdot\nabla(\mathbf{V}\cdot\nabla u_N)+\int_{\Sigma}\frac{\partial u_N}{\partial\mathbf{n}}\mathbf{V}\cdot\nabla\varphi-\int_{\Sigma}(\nabla u_N\cdot\nabla\varphi)\mathbf{V}\cdot\mathbf{n},\quad(41)$$

for all $\varphi\in H^2(\Omega)$ Note however that by definition of the shape derivative we have

$$(\nabla w,\nabla\varphi)_\Omega=(\nabla u'_N,\nabla\varphi)_\Omega+(\nabla(\mathbf{V}\cdot\nabla u_N),\nabla\varphi)_\Omega,\qquad\varphi\in H^1_{\Gamma,0}(\Omega).\quad(42)$$

Combining (39) and (42), we get

$$(\nabla u'_N,\nabla\varphi)_\Omega+(\nabla(\mathbf{V}\cdot\nabla u_N),\nabla\varphi)_\Omega=-(A\nabla u_N,\nabla\varphi)_\Omega+\lambda\int_{\Sigma}\varphi\mathrm{div}_\Sigma\mathbf{V},\quad(43)$$

for $\varphi\in H^1_{\Gamma,0}(\Omega)$. Applying Green's formula on the left hand side of (43) , and using (41) for the right hand side we get

$$(-\Delta u'_N,\varphi)_\Omega+\int_{\Sigma}\frac{\partial u'_N}{\partial\mathbf{n}}\varphi=\int_{\Sigma}\frac{\partial u_N}{\partial\mathbf{n}}\mathbf{V}\cdot\nabla\varphi-\int_{\Sigma}(\nabla u_N\cdot\nabla\varphi)\mathbf{V}\cdot\mathbf{n}+\lambda\int_{\Sigma}\varphi\mathrm{div}_\Sigma\mathbf{V}.$$

Choosing $\varphi\in C_0^\infty(\Omega)$, we obtain $-\Delta u'_N=0$ in Ω. This implies

$$\begin{aligned}\int_{\Sigma}\frac{\partial u'_N}{\partial\mathbf{n}}\varphi&=\int_{\Sigma}\frac{\partial u_N}{\partial\mathbf{n}}\mathbf{V}\cdot\nabla\varphi-\int_{\Sigma}\nabla u_N\cdot\nabla\varphi\mathbf{V}\cdot\mathbf{n}+\lambda\int_{\Sigma}\varphi\mathrm{div}_\Sigma\mathbf{V}\\&=\int_{\Sigma}(\lambda\mathbf{V}-\nabla u_N\mathbf{V}\cdot\mathbf{n})\cdot\nabla\varphi+\int_{\Sigma}\lambda\varphi\mathrm{div}_\Sigma\mathbf{V}\qquad\forall\varphi\in H^2(\Omega).\end{aligned}$$

Since $(\lambda\mathbf{V}-\nabla u_N\mathbf{V}\cdot\mathbf{n})\cdot\mathbf{n}=0$ one can replace $\nabla\varphi|_\Sigma$ by $\nabla_\Sigma\varphi$ which leads to

$$\int_{\Sigma}\frac{\partial u'_N}{\partial\mathbf{n}}\varphi=\int_{\Sigma}(\lambda\mathbf{V}-\nabla u_N\mathbf{V}\cdot\mathbf{n})\cdot\nabla_\Sigma\varphi+\int_{\Sigma}\lambda\varphi\mathrm{div}_\Sigma\mathbf{V}.$$

Applying (18) one finds

$$\int_{\Sigma}\frac{\partial u'_N}{\partial\mathbf{n}}\varphi=\int_{\Sigma}\varphi\mathrm{div}_\Sigma(\nabla u_N\mathbf{V}\cdot\mathbf{n}).$$

Applying (17), and (18) (noting that $\mathbf{V}\cdot\mathbf{n}\nabla_\Sigma u_N\cdot\mathbf{n}=0$), we have

$$\begin{aligned}\int_\Sigma \frac{\partial u_N'}{\partial \mathbf{n}}\varphi &= \int_\Sigma \kappa\varphi(\nabla u_N \mathbf{V}\cdot\mathbf{n})\cdot\mathbf{n} - \int_\Sigma \nabla_\Sigma\varphi\cdot\nabla u_N\mathbf{V}\cdot\mathbf{n}\\ &= \int_\Sigma \varphi\kappa\frac{\partial u_N}{\partial\mathbf{n}}\mathbf{V}\cdot\mathbf{n} - \int_\Sigma \nabla_\Sigma\varphi\cdot\nabla_\Sigma u_N\mathbf{V}\cdot\mathbf{n}\\ &= \int_\Sigma \varphi\kappa\lambda\mathbf{V}\cdot\mathbf{n} + \int_\Sigma \varphi\,\mathrm{div}_\Sigma(\mathbf{V}\cdot\mathbf{n}\nabla_\Sigma u_N).\end{aligned}$$

Since the trace of functions in $H^2(\Omega)$ is dense in $L^2(\Sigma)$, we deduce the boundary condition on Σ for u_N', which is

$$\frac{\partial u_N'}{\partial\mathbf{n}} = \mathrm{div}_\Sigma(\mathbf{V}\cdot\mathbf{n}\nabla_\Sigma u_N) + \kappa\lambda\mathbf{V}\cdot\mathbf{n}.$$

Therefore, the shape derivative of u_N satisfies the boundary value problem (38). Using similar arguments as in the previous theorem, we take $C^{2,1}$ bounded domains and by Theorem 4, we consider u_N belonging to $H^3(\Omega)$.

4.3 The Shape Derivative of J

We now prove the following result.

Theorem 7 *Let Ω be a $C^{2,1}$ bounded domain. The shape derivative of the Kohn-Vogelius cost functional*

$$J(\Omega) = \frac{1}{2}\int_\Omega |\nabla(u_D - u_N)|^2\,\mathrm{d}x$$

in the direction of a perturbation field $\mathbf{V}\in\Theta$, where Θ is defined by (6) and the state functions u_D and u_N satisfy the Dirichlet problem (3) and the Neumann problem (4), respectively, is given by

$$dJ(\Omega;\mathbf{V}) = \frac{1}{2}\int_\Sigma (\lambda^2 - (\nabla u_D\cdot\mathbf{n})^2 + 2\lambda\kappa u_N - (\nabla u_N\cdot\tau)^2)\mathbf{V}\cdot\mathbf{n}\,\mathrm{d}s.$$

Here, $\mathbf{n}$ is the unit exterior normal vector to Σ, τ is a unit tangent vector to Σ, and κ is the mean curvature of Σ.

Proof In this approach, we need $C^{2,1}$ domains Ω, which by elliptic regularity theory implies $u_D, u_N \in H^3(\Omega)$. Using the domain differentiation formula (19) we obtain

$$\mathrm{d}J(\Omega;\mathbf{V}) = \int_{\Omega} \nabla(u'_D - u'_N)\cdot\nabla(u_D - u_N)\,\mathrm{d}x + \frac{1}{2}\int_{\Sigma} |\nabla(u_D - u_N)|^2 v_n\,\mathrm{d}s \quad (44)$$

where the shape derivatives u'_D and u'_N satisfy (33) and (38), respectively, and v_n refers to the normal component of $\mathbf{V}$ on Σ.

We simplify each integral in (44). For the first integral, we use the Green's formula and the boundary conditions for u_D and u_N and their shape derivatives, to obtain

$$\begin{aligned}\int_{\Omega} \nabla(u'_D - u'_N)\cdot\nabla(u_D - u_N)\,\mathrm{d}x &= -\int_{\Sigma} \frac{\partial u_D}{\partial \mathbf{n}}\left(\frac{\partial u_D}{\partial \mathbf{n}} - \lambda\right)\mathbf{V}\cdot\mathbf{n} \\ &\quad - \int_{\Sigma} (\mathrm{div}_{\Sigma}(\mathbf{V}\cdot\mathbf{n}\nabla_{\Sigma}u_N) + \mathbf{V}\cdot\mathbf{n}\kappa\lambda)(u_D - u_N)\,\mathrm{d}s \\ &= -\int_{\Sigma}\left(\left(\frac{\partial u_D}{\partial \mathbf{n}}\right)^2 - \lambda\frac{\partial u_D}{\partial \mathbf{n}}\right)v_n\,\mathrm{d}s \\ &\quad + \int_{\Sigma} \mathrm{div}_{\Sigma}(v_n\nabla_{\Sigma}u_N)u_N\,\mathrm{d}s + \int_{\Sigma}\kappa\lambda u_N v_n\,\mathrm{d}s.\end{aligned}$$

Since $v_n\nabla_{\Sigma}u_N\cdot\mathbf{n} = 0$, one can apply (18) to get

$$\begin{aligned}\int_{\Sigma} \mathrm{div}_{\Sigma}(v_n\nabla_{\Sigma}u_N)u_N\,\mathrm{d}s &= -\int_{\Sigma}\nabla_{\Sigma}u_N\cdot(\nabla_{\Sigma}u_N)v_n\,\mathrm{d}s - \int_{\Sigma}|\nabla_{\Sigma}u_N|^2 v_n\,\mathrm{d}s \\ &= -\int_{\Sigma}(\nabla u_N\cdot\tau)^2 v_n. \quad (45)\end{aligned}$$

Therefore, (45) can be simplified as

$$\begin{aligned}\int_{\Omega}(\nabla(u'_D - u'_N)\cdot\nabla(u_D - u_N)\,\mathrm{d}x &= -\int_{\Sigma}\left(\left(\frac{\partial u_D}{\partial \mathbf{n}}\right)^2 - \lambda\frac{\partial u_D}{\partial \mathbf{n}}\right)v_n\,\mathrm{d}s \\ &\quad - \int_{\Sigma}(\nabla u_N\cdot\tau)^2 v_n + \int_{\Sigma}\kappa\lambda u_N v_n\,\mathrm{d}s. \quad (46)\end{aligned}$$

The second integral in (44) is simplified as follows:

$$\begin{aligned}\frac{1}{2}\int_{\Sigma}|\nabla(u_D - u_N)|^2\mathbf{V}\cdot\mathbf{n}\,\mathrm{d}s &= \frac{1}{2}\int_{\Sigma}(|\nabla u_D|^2 - 2\nabla u_D\nabla u_N + |\nabla u_N|^2)v_n\,\mathrm{d}s \\ &= \frac{1}{2}\int_{\Sigma}\left(\left(\frac{\partial u_D}{\partial \mathbf{n}}\right)^2 - 2\frac{\partial u_D}{\partial \mathbf{n}}\lambda + \lambda^2 + (\nabla u_N\cdot\tau)^2\right)v_n\,\mathrm{d}s. \\ &\quad (47)\end{aligned}$$

Combining the integrals (46) and (47), we get the desired result; that is,

$$dJ(\Omega;\mathbf{V}) = \frac{1}{2}\int_\Sigma (\lambda^2 - (\nabla u_D \cdot \mathbf{n})^2 + 2\lambda\kappa u_N - (\nabla u_N \cdot \tau)^2)\mathbf{V}\cdot\mathbf{n}\,\mathrm{d}s. \tag{48}$$

We rewrite (48) as $dJ(\Omega;\mathbf{V}) = \int_\Sigma F\mathbf{V}\cdot\mathbf{n}$, where

$$F = \frac{1}{2}\Big(-(\nabla u_N \cdot \tau)^2 - (\nabla u_D \cdot \mathbf{n})^2 + \lambda^2 + 2\lambda u_N\kappa\Big). \tag{49}$$

We conclude that J is shape differentiable at Ω because $dJ(\Omega;\mathbf{V})$ exists for all $\mathbf{V}\in\Theta$ and the mapping $\mathbf{V}\mapsto dJ(\Omega;\mathbf{V})$ is linear and continuous with respect to $\mathbf{V}\in\Theta$ since

$$|dJ(\Omega;\mathbf{V})| \le |F|_{L^1(\Sigma)}|\mathbf{V}|_{C(\Sigma)} \le |F|_{L^1(\Sigma)}|\mathbf{V}|_{C^{1,1}(\bar{U})}.$$

5 Conclusion

First, we consider the solution $u_D \in H^1(\Omega)$ to the Dirichlet problem (3) and we originally consider $C^{1,1}$ domains. By elliptic regularity theory, this solution is indeed an element of $H^2(\Omega)$. By definition, the shape derivative of u_D exists in $H^1(\Omega)$ because the material derivative is in $H^1_0(\Omega)$ and the term $\nabla u_D\cdot\mathbf{V}$ belongs to $H^1(\Omega)$. Then we determine the boundary value problem that is satisfied by this shape derivative. It turns out that the regularity of u_D is not sufficient to justify the boundary value problem satisfied by its shape derivative. The same is true for the state u_N. Thus in this approach, we require more regular domains; that is, $C^{2,1}$-domains, and we have to consider more regular solutions; that is, $u_D, u_N \in H^3(\Omega)$. This is in contrast to an approach that bypasses shape derivatives of states, where $C^{1,1}$ regularity of domains is sufficient to justify the Eulerian derivative of J (cf. [4]).

The explicit form of the shape derivative of J is determined. We observe that neither derivatives of the state variables nor the adjoint states appear in the final form. Also, the explicit form obeys the Hadamard structure theorem [8, 17]; that is, there is a function F defined on the free boundary Σ such that

$$dJ(\Omega;\mathbf{V}) = \int_\Sigma F\mathbf{V}\cdot\mathbf{n}\,\mathrm{d}s.$$

If we perturb the domain in the direction $\mathbf{V}|_\Sigma = -F\mathbf{n}$, then we are sure that the value of the functional J decreases.

References

1. Abda, B., Bouchon, F., Peichl, G., Sayeh, M., Touzani, R.: A Dirichlet-Neumann cost functional approach for the Bernoulli problem. J. Eng. Math. **81**, 157–176 (2013)
2. Afraites, L., Dambrine, M., Kateb, D.: On second-order shape optimization methods for electrical impedance tomography. Preprint, HAL-00140211, version 1, pp. 1–28 (2007)
3. Bacani, J.B.: Methods of shape optimization in free boundary problems. Ph.D. Thesis, Karl-Franzens-Universitaet Graz (2013)
4. Bacani, J.B., Peichl, G.: On the first-order shape derivative of the Kohn-Vogeliuscost functional of the Bernoulli problem. Abstr. Appl. Anal. **2013**, 19 (2013). Article ID 384320. doi:10.1155/2013/384320
5. Caffarelli, L.A., Salsa, S.: A Geometric Approach to Free Boundary Problems. American Mathematical Society, Providence (2005)
6. Crank, J.: Free and Moving Boundary Problems. Oxford University Press Inc., New York (1984)
7. Delfour, M.C., Zolesio, J.P.: Shapes and Geometries. SIAM, Philadelphia (2001)
8. Delfour, M.C., Zolesio, J.P.: Anatomy of the shape Hessian. Annali di Matematica pura ed applicata **159**, 315–339 (1991)
9. Flucher, M., Rumpf, M.: Bernoulli's free-boundary problem, qualitative theory and numerical approximation. J. Reine Angew. Math. **486**, 165–204 (1997)
10. Friedman, A.: Free boundary problems in science and technology. Not. AMS **47**, 854–861 (2000)
11. Fujii, N.: Second variation and its application in domain optimization problem, control of distributed parameter systems. In: Proceedings of the 4th IFAC Symposium, vol. 24, pp. 346–360. Pergamon Press (1986)
12. Haslinger, J., Ito, K., Kozubek, T., Kunisch, K., Peichl, G.: On the shape derivative for problems of Bernoulli type. Interfaces Free Boundaries **1**, 317–330 (2009)
13. Haslinger, J., Mäkinen, R.A.E.: Introduction to Shape Optimization (Theory, Approximation, and Computation). SIAM Advances and Control, Philadelphia (2003)
14. Henrot, A., Pierre, M.: Variation et Optimisation de Formes. Springer, Berlin (2005)
15. Ito, K., Kunisch, K., Peichl, G.: Variational approach to shape derivatives for a class of Bernoulli problems. J. Math. Anal. Appl. **314**, 126–149 (2006)
16. Kohn, R., Vogelius, M.: Determining conductivity by boundary measurements. Commun. Pure Appl. Math. **37**, 289–298 (1984)
17. Lamboley, J., Pierre, M.: Structure of shape derivatives around irregular do- mains and applications. eprint arXiv:math/0609526, pp. 1–14 (2006)
18. Masanao, T., Fujii, N.: Second-order necessary conditions for domain optimization problems in elastic structures, Part 1: surface traction given as a field. J. Optim. Theor. Appl. **72**, 355–382 (1992)
19. Simon, J.: Second variations for domain optimization problems. Int. Ser. Numer. Math. **91**, 361–378 (1989)
20. Sokolowski, J., Zolesio, J.: Introduction to Shape Optimization. Springer, Berlin (1991)
21. Tiihonen, T.: Shape optimization and trial methods for free boundary problems. RAIRO Modelisation mathematique et analyse numerique **31**, 805–825 (1997)
22. Toivanen, J.I., Haslinger, J., Mäkinen, R.A.E.: Shape optimization of systems governed by Bernoulli free boundary problems. Comput. Methods Appl. Mech. Eng. **197**, 3803–3815 (2008)

Chapter 18
Applications of the Hausdorff Measure of Noncompactness on the Space $l_p(r, s, t; B^{(m)}), 1 \leq p < \infty$

Amit Maji and P. D. Srivastava

Abstract In this paper, we have introduced a sequence space $l_p(r, s, t; B^{(m)})$, $1 \leq p < \infty$ and proved that the space is a complete normed linear space. We have also shown that the space $l_p(r, s, t; B^{(m)})$ is linearly isomorphic to l_p for $1 \leq p < \infty$. Further, we have established some identities or estimates for the operator norms and the Hausdorff measure of noncompactness of certain matrix operators on this space. Finally, we have characterized some classes of compact operators on this space.

Keywords Difference operator · Sequence space · Hausdorff measure of noncompactness · Compact operators

2010 Mathematics Subject Classification 46A45 · 46B15 · 46B50 · 40A05

1 Introduction and Preliminaries

Let w be the space of all real or complex sequences $x = (x_n), n \in \mathbb{N}_0 = \{0, 1, 2, \cdots\}$. We denote by l_∞, c, c_0 and $l_p (1 \leq p < \infty)$ for the space of all bounded, convergent, null sequences and absolutely p-summable sequences respectively. Moreover, bs, cs stand for the sequence spaces of all bounded and convergent series respectively. We

The authors are thankful to the referees for their valuable comments and suggestions which improved the presentation of the paper. The first author is grateful to CSIR, New Delhi, Govt. of India for the financial support award no. 09/081(1120)/2011-EMR-I.

A. Maji (✉) · P. D. Srivastava
Department of Mathematics, Indian Institute of Technology Kharagpur, Kharagpur 721302, West Bengal, India
e-mail: amit.iitm07@gmail.com

P. D. Srivastava
e-mail: pds@maths.iitkgp.ernet.in

R. N. Mohapatra et al. (eds.), *Mathematics and Computing 2013*, Springer Proceedings in Mathematics & Statistics 91,
DOI: 10.1007/978-81-322-1952-1_18,

denote by $e = (1, 1, \cdots)$ and e_n for the sequence whose n-th term is 1 and others are zero. A sequence space X is called a BK space if it is a Banach space with continuous coordinates $p_n : X \to \mathbb{K}(n \in \mathbb{N}_0)$, where $\mathbb{K}$ denotes the real or complex field and $p_n(x) = x_n$ for all $x = (x_k) \in X$ and every $n \in \mathbb{N}_0$. For an infinite matrix A and a sequence space λ, the matrix domain of A denoted by λ_A and is defined as $\lambda_A = \{x \in w : Ax \in \lambda\}$ [9]. An infinite matrix $T = (t_{nk})$ is called a triangle if $t_{nn} \neq 0$ and $t_{nk} = 0$ for all $k > n$ $(n \in \mathbb{N}_0)$, and if X is a BK space then X_T is also a BK space.

In recent times, there is an approach of forming a new sequence space by using a suitable matrix domain and give a characterization of some class of compact operators on this space by applying the Hausdorff measure of noncompactness, which was first introduced and studied by Goldenstein, Gohberg, and Markus in 1957. Recently, several authors, namely Djolović [2], Djolović et al. [3], Mursaleen and Noman [7], Kara and Başarir [4], etc., have established some identities or estimates for the operator norms and the Hausdorff measure of noncompactness of matrix operators from an arbitrary BK space to arbitrary BK space.

In this paper, our aim is to introduce a sequence space $l_p(r, s, t; B^{(m)})$ for $1 \leq p < \infty$. We have proved that the space is a complete normed linear space and linearly isomorphic to l_p. Moreover, we have obtained some identities or estimates for the operator norms and for the Hausdorff measure of noncompactness of matrix operators on this space and also characterized some classes of compact operators.

2 Difference Sequence Space $l_p(r, s, t; B^{(m)})$ for $1 \leq p < \infty$

In 2011, Mursaleen and Noman [8] introduced the notion of generalized means. Let $\mathscr{U}$ and $\mathscr{U}_0$ be the sets defined by

$$\mathscr{U} = \left\{u = (u_n)_{n=0}^{\infty} \in w : u_n \neq 0 \text{ for all } n\right\} \text{ and } \mathscr{U}_0 = \left\{u = (u_n)_{n=0}^{\infty} \in w : u_0 \neq 0\right\}.$$

Let $r = (r_n), t = (t_n) \in \mathscr{U}$ and $s = (s_n) \in \mathscr{U}_0$. The sequence $y = (y_n)$ of generalized means of a sequence $x = (x_n)$ is defined by

$$y_n = \frac{1}{r_n} \sum_{k=0}^{n} s_{n-k} t_k x_k \qquad (n \in \mathbb{N}_0).$$

The infinite matrix $A(r, s, t)$ of generalized means is defined by

$$(A(r, s, t))_{nk} = \begin{cases} \frac{s_{n-k} t_k}{r_n}, & 0 \leq k \leq n \\ 0, & k > n. \end{cases}$$

The inverse of $A(r,s,t)$ is the triangle $B=(b_{nk})_{n,k}$, which is defined as

$$b_{nk}=\begin{cases}(-1)^{n-k}\,\frac{D_{n-k}^{(s)}}{t_n}r_k, & 0\le k\le n\\ 0, & k>n,\end{cases}$$

where $D_0^{(s)}=\frac{1}{s_0}$ and

$$D_n^{(s)}=\frac{1}{s_0^{n+1}}\begin{vmatrix} s_1 & s_0 & 0 & 0\cdots & 0\\ s_2 & s_1 & s_0 & 0\cdots & 0\\ \vdots & \vdots & \vdots & \vdots & \\ s_{n-1} & s_{n-2} & s_{n-3} & s_{n-4}\cdots & s_0\\ s_n & s_{n-1} & s_{n-2} & s_{n-3}\cdots & s_1\end{vmatrix}\qquad \text{for } n=1,2,3,\cdots$$

The generalized difference matrix of order m denoted as $B^{(m)}=B^{(m)}(u,v)=(b_{nk}^{(m)})$, $u,v\neq 0$ (see [1]) is defined as

$$b_{nk}^{(m)}=\begin{cases}\binom{m}{n-k}u^{m-n+k}v^{n-k}, & \text{if } \max\{0,n-m\}\le k\le n\\ 0, & \text{if } 0\le k<\max\{0,n-m\}\\ 0 & \text{if } k>n.\end{cases}$$

In particular, if $u=1$, $v=-1$ then the matrix $B^{(m)}$ reduces to $\Delta^{(m)}$, a difference operator of order m.

Now combining the generalized means and the operator $B^{(m)}$, we introduce a sequence space $l_p(r,s,t;B^{(m)})$ for $1\le p<\infty$ as

$$l_p(r,s,t;B^{(m)})=\left\{x=(x_n)\in w:(((A(r,s,t).B^{(m)})x)_n)\in l_p\right\}.$$

By using matrix domain, we can write $l_p(r,s,t;B^{(m)})=(l_p)_{A(r,s,t;B^{(m)})}=\{x\in w:A(r,s,t;B^{(m)})x\in l_p\}$, where $A(r,s,t;B^{(m)})=A(r,s,t).B^{(m)}$, product of two triangles $A(r,s,t)$ and $B^{(m)}$. The sequence $y=(y_n)$ is $A(r,s,t).B^{(m)}$-transform of a sequence $x=(x_n)$, i.e., for each $n\in\mathbb{N}_0$

$$y_n=\sum_{j=0}^{n}\left(\sum_{i=j}^{n}\binom{m}{i-j}\frac{s_{n-i}t_i}{r_n}u^{m+j-i}v^{i-j}\right)x_j.$$

3 Main Results

Theorem 1 *The sequence space $l_p(r,s,t;B^{(m)})$ for $1\le p<\infty$ is a complete normed linear space under the norm defined by*

$$\|x\|_{l_p(r,s,t;B^{(m)})} = \Big(\sum_{n=0}^{\infty}\Big|\sum_{j=0}^{n}\Big(\sum_{i=j}^{n}\binom{m}{i-j}\frac{s_{n-i}t_i}{r_n}u^{m+j-i}v^{i-j}\Big)x_j\Big|^p\Big)^{\frac{1}{p}}$$
$$= \Big(\sum_{n=0}^{\infty}|(A(r,s,t;B^{(m)})x)_n|^p\Big)^{\frac{1}{p}}.$$

Proof Since $B^{(m)}$ is a linear operator, it is easy to show that $l_p(r,s,t;B^{(m)})$ is a linear space and the functional $\|.\|_{l_p(r,s,t;B^{(m)})}$ defined above gives a norm on the linear space $l_p(r,s,t;B^{(m)})$.
To show completeness, let (x^k) be a Cauchy sequence in $l_p(r,s,t;B^{(m)})$, where $x^k = (x_j^k) = (x_0^k, x_1^k, x_2^k, \ldots) \in l_p(r,s,t;B^{(m)})$ for each $k \in \mathbb{N}_0$. Then for every $\epsilon > 0$, there exists $k_0 \in \mathbb{N}$ such that

$$\|x^k - x^l\|_{l_p(r,s,t;B^{(m)})} < \frac{\epsilon}{2} \text{ for } k, l \geq k_0,$$

i.e.,

$$\Big(\sum_{n=0}^{\infty}|(A(r,s,t;B^{(m)})x^k)_n - (A(r,s,t;B^{(m)})x^l)_n|^p\Big)^{\frac{1}{p}} < \frac{\epsilon}{2} \text{ for all } k, l \geq k_0, \quad (1)$$

This shows that the sequence $((A(r,s,t).B^{(m)})x^k)_n$ is a Cauchy sequence of scalars for each $n \in \mathbb{N}_0$ and hence $((A(r,s,t).B^{(m)})x^k)_n$ converges for each n. We write

$$\lim_{k\to\infty}((A(r,s,t).B^{(m)})x^k)_n = ((A(r,s,t).B^{(m)})x)_n \quad \text{for each } n \in \mathbb{N}_0.$$

On taking $l \to \infty$ in (1), we obtain

$$\Big(\sum_{n=0}^{\infty}|(A(r,s,t;B^{(m)})x^k)_n - (A(r,s,t;B^{(m)})x)_n|^p\Big)^{\frac{1}{p}} < \epsilon \text{ for all } k \geq k_0.$$

Hence $\|x^k - x\|_{l_p(r,s,t;B^{(m)})} < \epsilon$ for all $k \geq k_0$. This implies that the sequence (x^k) converges to x in $l_p(r,s,t;B^{(m)})$. Next we show that $x \in l_p(r,s,t;B^{(m)})$.

Since $x^k \in l_p(r,s,t;B^{(m)})$ for $k \geq k_0$, we have

$$\|x\|_{l_p(r,s,t;B^{(m)})} \leq \|x^{k_0}\|_{l_p(r,s,t;B^{(m)})} + \|x^{k_0} - x\|_{l_p(r,s,t;B^{(m)})},$$

which is finite. Hence $x \in l_p(r,s,t;B^{(m)})$. This completes the proof.

Theorem 2 *The sequence space $l_p(r,s,t;B^{(m)})$ is linearly isomorphic to the space l_p, i.e., $l_p(r,s,t;B^{(m)}) \cong l_p$ for $1 \leq p < \infty$.*

Proof We define a map $T : l_p(r,s,t;B^{(m)}) \to l_p$ by $x \longmapsto Tx = y = (y_n)$, where

$$y_n = \sum_{j=0}^{n} \Big(\sum_{i=j}^{n} \binom{m}{i-j} \frac{s_{n-i} t_i}{r_n} u^{m+j-i} v^{i-j} \Big) x_j.$$

Since $B^{(m)}$ is a linear operator, so the linearity of T is trivial. It is clear from the definition that $Tx = 0$ implies $x = 0$. Thus T is injective. To prove T is surjective, let $y = (y_n) \in l_p$. Since $y = (A(r, s, t).B^{(m)})x$, so we have

$$x = (A(r, s, t).B^{(m)})^{-1} y = (B^{(m)})^{-1}.A(r, s, t)^{-1} y.$$

So we can get a sequence $x = (x_n)$ as

$$x_n = \sum_{l=0}^{n} \sum_{j=l}^{n} (-1)^{j-l} \binom{m+n-j-1}{n-j} \frac{(-v)^{n-j}}{u^{m+n-j}} \frac{D_{j-l}^{(s)}}{t_j} r_l y_l, \qquad n \in \mathbb{N}_0.$$

Then

$$\begin{aligned} \|x\|_{l_p(r,s,t;B^{(m)})} &= \Big(\sum_{n=0}^{\infty} \Big| \sum_{j=0}^{n} \Big(\sum_{i=j}^{n} \binom{m}{i-j} \frac{s_{n-i} t_i}{r_n} u^{m+j-i} v^{i-j} \Big) x_j \Big|^p \Big)^{\frac{1}{p}} \\ &= \Big(\sum_{n=0}^{\infty} |y_n|^p \Big)^{\frac{1}{p}} = \|y\|_p < \infty. \end{aligned}$$

Thus $x \in l_p(r, s, t; B^{(m)})$ and this shows that T is surjective. Hence T is a linear bijection from $l_p(r, s, t; B^{(m)})$ to l_p. Also T is norm preserving. So $l_p(r, s, t; B^{(m)}) \cong l_p$. This completes the proof.

Remark 1 Since $l_p(r, s, t; B^{(m)}) \cong l_p$, the Schauder basis of the sequence space $l_p(r, s, t; B^{(m)})$ is the inverse image of the basis of l_p. Hence the space $l_p(r, s, t; B^{(m)})$ for $1 \le p < \infty$ is separable.

4 Compact Operators on the Space $l_p(r, s, t; B^{(m)})$

In this section, we apply the Hausdorff measure of noncompactness to establish necessary and sufficient conditions for an infinite matrix to be a compact operator on the space $l_p(r, s, t; B^{(m)})$.

Let X and Y be two Banach spaces. We denote by $\mathscr{B}(X, Y)$, the set of all bounded (continuous) linear operators $L : X \to Y$, which is also a Banach space with the operator norm given by

$$\|L\| = \sup_{x \in S_X} \|L(x)\|_Y \text{ for all } L \in \mathscr{B}(X, Y),$$

where S_X denotes the unit sphere, i.e., $S_X = \{x \in X : \|x\| = 1\}$. A linear operator $L : X \to Y$ is said to be compact if the domain of L is X and for every bounded sequence $(x_n) \subset X$, the sequence $(L(x_n))$ has a subsequence which is convergent in Y. We denote by $\mathscr{C}(X, Y)$, the class of all compact operators in $\mathscr{B}(X, Y)$. An operator $L \in \mathscr{B}(X, Y)$ is said to be finite rank if $\dim R(L) < \infty$, where $R(L)$ is the range space of L. If X is a BK space and $a = (a_k) \in w$, then we consider

$$\|a\|_X^* = \sup_{x \in S_X} \left| \sum_{k=0}^{\infty} a_k x_k \right|, \tag{2}$$

provided the expression on the right side exists and is finite which is the case whenever $a \in X^\beta$ [7].

Let (X, d) be a metric space and $\mathscr{M}_X$ be the class of all bounded subsets of X. Let $B(x, r) = \{y \in X : d(x, y) < r\}$ denotes the open ball of radius $r > 0$ with center at x. The Hausdorff measure of noncompactness of a set $Q \in \mathscr{M}_X$, denoted by $\chi(Q)$, is defined by

$$\chi(Q) = \inf \left\{ \epsilon > 0 : Q \subset \bigcup_{i=0}^{n} B(x_i, r_i), x_i \in X, r_i < \epsilon, n \in \mathbb{N} \right\}.$$

The function $\chi : \mathscr{M}_X \to [0, \infty)$ is called the Hausdorff measure of noncompactness. The basic properties of the Hausdorff measure of noncompactness can be found in [5, 7]. For example, if Q, Q_1 and Q_2 are bounded subsets of a metric space (X, d) then

$$\begin{aligned} &\chi(Q) = 0 \text{ if and only if } Q \text{ is totally bounded and} \\ &\text{if } Q_1 \subset Q_2 \text{ then } \chi(Q_1) \leq \chi(Q_2). \end{aligned}$$

In addition, if X is a normed linear space, then the function χ has some additional properties due to linear structure, namely,

$$\begin{aligned} &\chi(Q_1 + Q_2) \leq \chi(Q_1) + \chi(Q_2) \\ &\chi(\alpha Q) = |\alpha| \chi(Q) \text{ for all } \alpha \in \mathbb{K}. \end{aligned}$$

Let ϕ denotes the set of all finite sequences, i.e., of sequences that terminate with only zeros. Let p' be the conjugate of p, i.e., $p' = \frac{p}{p-1}$ for $1 < p < \infty$ and $p' = \infty$ for $p = 1$.

Lemma 1 ([5], Theorem 1.29) *Let X denotes any of the spaces c_0, c, l_∞ or l_p. Then $X^\beta = l_1$ ($X^\beta = l_{p'}$ for $X = l_p$) and $\|a\|_X^* = \|a\|_{l_1}$ for all $a \in l_1$ ($\|a\|_X^* = \|a\|_{l_{p'}}$ for $X = l_p$).*

Lemma 2 [7] *Let $X \supset \phi$ and Y be BK spaces. Then $(X, Y) \subset \mathscr{B}(X, Y)$, i.e., every matrix $A \in (X, Y)$ defines an operator $L_A \in \mathscr{B}(X, Y)$, where $L_A(x) = Ax$ for all $x \in X$.*

Lemma 3 [2] *Let $X \supset \phi$ be a BK space and Y be any of the spaces c_0, c or l_∞. If $A \in (X, Y)$, then*

$$\|L_A\| = \|A\|_{(X,l_\infty)} = \sup_n \|A_n\|_X^* < \infty.$$

Lemma 4 [5] *Let Q be a bounded subset of the normed space X, where $X = l_p$ for $1 \leq p < \infty$ and $X = c_0$ for $p = \infty$. If $P_l : X \to X$ is an operator defined by $P_l(x) = (x_0, x_1, \cdots, x_l, 0, 0, \cdots)$ for all $x = (x_k) \in X$, then*

$$\chi(Q) = \lim_{l\to\infty} \left(\sup_{x\in Q} \|(I - P_l)(x)\| \right),$$

where I is the identity operator on X.

Lemma 5 [5] *Let X, Y be two Banach spaces and $L \in \mathscr{B}(X, Y)$. Then*

$$\|L\|_\chi = \chi(L(S_X))$$

and

$$L \in \mathscr{C}(X, Y) \text{ if and only if } \|L\|_\chi = 0.$$

We establish the following lemmas which are required for our study.

Lemma 6 *If $a = (a_k) \in [l_p(r, s, t; B^{(m)})]^\beta$ then $\tilde{a} = (\tilde{a}_k) \in l_p^\beta = l_{p'}$ and the equality*

$$\sum_{k=0}^{\infty} a_k x_k = \sum_{k=0}^{\infty} \tilde{a}_k y_k$$

holds for every $x = (x_k) \in l_p(r, s, t; B^{(m)})$ and $y = (y_k) \in l_p$, where $y = (A(r, s, t).B^{(m)})x$. In addition,

$$\tilde{a}_k = r_k \Bigg[\frac{a_k}{s_0 t_k u^m} + \sum_{i=k}^{k+1} (-1)^{i-k} \frac{D_{i-k}^{(s)}}{t_i} \sum_{j=k+1}^{\infty} \binom{m+j-i-1}{j-i} \frac{(-v)^{j-i}}{u^{j-i+m}} a_j + \sum_{i=k+2}^{\infty} (-1)^{i-k} \frac{D_{i-k}^{(s)}}{t_i} \sum_{j=i}^{\infty} \binom{m+j-i-1}{j-i} \frac{(-v)^{j-i}}{u^{j-i+m}} a_j \Bigg]. \quad (3)$$

Proof Let $a = (a_k) \in [l_p(r, s, t; B^{(m)})]^\beta$. Then by ([6], Theorem 3.2), we have $R(a) = (R_k(a)) \in l_p^\beta = l_{p'}$ and also

$$\sum_{k=0}^{\infty} a_k x_k = \sum_{k=0}^{\infty} R_k(a) T_k(x) \quad \forall\, x \in l_p(r,s,t;B^{(m)}),$$

where

$$\begin{aligned} R_k(a) &= \sum_{j=k}^{\infty}\sum_{i=k}^{j}(-1)^{i-k}\binom{m+j-i-1}{j-i}\frac{D_{i-k}^{(s)}}{t_i}\frac{(-v)^{j-i}}{u^{j-i+m}} r_k a_j \\ &= r_k\Bigg[\frac{a_k}{s_0 t_k u^m} + \sum_{i=k}^{k+1}(-1)^{i-k}\frac{D_{i-k}^{(s)}}{t_i}\sum_{j=k+1}^{\infty}\binom{m+j-i-1}{j-i}\frac{(-v)^{j-i}}{u^{j-i+m}}a_j \\ &\qquad + \sum_{i=k+2}^{\infty}(-1)^{i-k}\frac{D_{i-k}^{(s)}}{t_i}\sum_{j=i}^{\infty}\binom{m+j-i-1}{j-i}\frac{(-v)^{j-i}}{u^{j-i+m}}a_j\Bigg] = \tilde{a}_k \end{aligned}$$

and $y = T(x) = (A(r,s,t).B^{(m)})x$. This completes the proof.

Lemma 7 *Let $1 \le p < \infty$. Then we have*

$$\|a\|^*_{l_p(r,s,t;B^{(m)})} = \|\tilde{a}\|_{l_{p'}} = \begin{cases} \Big(\sum_{k=0}^{\infty}|\tilde{a}_k|^{p'}\Big)^{\frac{1}{p'}}, & 1 < p < \infty \\ \sup_k |\tilde{a}_k|, & p = 1 \end{cases}$$

for all $a = (a_k) \in [l_p(r,s,t;B^{(m)})]^{\beta}$, where $\tilde{a} = (\tilde{a}_k)$ is defined in (3).

Proof Let $a = (a_k) \in [l_p(r,s,t;B^{(m)})]^{\beta}$. Then from Lemma 6, we have $\tilde{a} = (\tilde{a}_k) \in l_{p'}$. Also $x \in S_{l_p(r,s,t;B^{(m)})}$ if and only if $y = T(x) \in S_{l_p}$ as $\|x\|_{l_p(r,s,t;B^{(m)})} = \|y\|_{l_p}$. From (2), we have

$$\|a\|^*_{l_p(r,s,t;B^{(m)})} = \sup_{x \in S_{l_p(r,s,t;B^{(m)})}} \Big|\sum_{k=0}^{\infty} a_k x_k\Big| = \sup_{y \in S_{l_p}} \Big|\sum_{k=0}^{\infty}\tilde{a}_k y_k\Big| = \|\tilde{a}\|^*_{l_p}.$$

Using Lemma 1, we have $\|a\|^*_{l_p(r,s,t;B^{(m)})} = \|\tilde{a}\|^*_{l_p} = \|\tilde{a}\|_{l_{p'}}$, which is finite as $\tilde{a} \in l_{p'}$. This completes the proof.

Lemma 8 *Let Y be any sequence space, $A = (a_{nk})_{n,k}$ be an infinite matrix and $1 \le p < \infty$. If $A \in (l_p(r,s,t;B^{(m)}), Y)$ then $\tilde{A} \in (l_p, Y)$ such that $Ax = \tilde{A}y$ for all $x \in l_p(r,s,t;B^{(m)})$ and $y \in l_p$, which are connected by the relation $y = (A(r,s,t).B^{(m)})x$ and $\tilde{A} = (\tilde{a}_{nk})_{n,k}$ is given by*

$$\tilde{a}_{nk} = r_k\Bigg[\frac{a_{nk}}{s_0 t_k u^m} + \sum_{i=k}^{k+1}(-1)^{i-k}\frac{D_{i-k}^{(s)}}{t_i}\sum_{j=k+1}^{\infty}\binom{m+j-i-1}{j-i}\frac{(-v)^{j-i}}{u^{j-i+m}}a_{nj}$$

$$+\sum_{i=k+2}^{\infty}(-1)^{i-k}\frac{D_{i-k}^{(s)}}{t_i}\sum_{j=i}^{\infty}\binom{m+j-i-1}{j-i}\frac{(-v)^{j-i}}{u^{j-i+m}}a_{nj}\Bigg], \quad (4)$$

provided the series on the right side converges for all n, k.

Proof We assume that $A \in (l_p(r,s,t;B^{(m)}), Y)$, then $A_n \in [l_p(r,s,t;B^{(m)})]^\beta$ for all n. Thus it follows from Lemma 6 that $\tilde{A}_n \in l_p^\beta = l_{p'}$ for all n and $Ax = \tilde{A}y$ holds for every $x \in l_p(r,s,t;B^{(m)})$, $y \in l_p$, which are connected by the relation $y = (A(r,s,t).B^{(m)})x$. Hence $\tilde{A}y \in Y$. Since $x = (B^{(m)})^{-1}(A(r,s,t))^{-1}y$, for every $y \in l_p$, we get some $x \in l_p(r,s,t;B^{(m)})$ and hence $\tilde{A} \in (l_p, Y)$. This completes the proof.

Lemma 9 *Let* $1 < p < \infty$, $A = (a_{nk})_{n,k}$ *be an infinite matrix and* $\tilde{A} = (\tilde{a}_{nk})_{n,k}$ *be the associate matrix defined in* (4). *If* $A \in (l_p(r,s,t;B^{(m)}), Y)$, *where* $Y \in \{c_0, c, l_\infty\}$, *then*

$$\|L_A\| = \|A\|_{(l_p(r,s,t;B^{(m)}),l_\infty)} = \sup_n\Bigg(\sum_{k=0}^{\infty}|\tilde{a}_{nk}|^{p'}\Bigg)^{\frac{1}{p'}} < \infty.$$

Proof The proof follows from Lemmas 3 and 7.

Now we state and prove the main result of this section.

Theorem 3 *Let* $1 < p < \infty$. *We have*
(*a*) *if* $A \in (l_p(r,s,t;B^{(m)}), c_0)$ *then*

$$\|L_A\|_\chi = \limsup_{n\to\infty}\Bigg(\sum_{k=0}^{\infty}|\tilde{a}_{nk}|^{p'}\Bigg)^{\frac{1}{p'}} \quad (5)$$

(*b*) *if* $A \in (l_p(r,s,t;B^{(m)}), l_\infty)$ *then*

$$0 \le \|L_A\|_\chi \le \limsup_{n\to\infty}\Bigg(\sum_{k=0}^{\infty}|\tilde{a}_{nk}|^{p'}\Bigg)^{\frac{1}{p'}}. \quad (6)$$

Proof (*a*) Clearly the expressions in (5) and in (6) exist by Lemma 9. We write $S = S_{l_p(r,s,t;B^{(m)})}$ in short. Then by Lemma 5, we have $\|L_A\|_\chi = \chi(AS)$. Since $l_p(r,s,t;B^{(m)})$ and c_0 are BK spaces, A induces a continuous map L_A from $l_p(r,s,t;B^{(m)})$ to c_0 by Lemma 2. Thus AS is bounded in c_0, i.e., $AS \in \mathscr{M}_{c_0}$. Now by Lemma 4,

$$\chi(AS) = \lim_{l\to\infty}\Big(\sup_{x\in S}\|(I-P_l)(Ax)\|_\infty\Big),$$

where the operator $P_l : c_0 \to c_0$ is defined by $P_l(\xi) = (\xi_0, \xi_1, \cdots, \xi_l, 0, 0, \cdots)$ for all $\xi = (\xi_k) \in c_0$ and $l \in \mathbb{N}_0$. Therefore, $\|(I - P_l)(Ax)\|_\infty = \sup_{n>l} |A_n(x)|$ for all $x \in l_p(r, s, t; B^{(m)})$. Using (2) and Lemma 7, we have

$$\begin{aligned}\sup_{x\in S} \|(I - P_l)(Ax)\|_\infty &= \sup_{n>l} \|A_n\|^*_{l_p(r,s,t;B^{(m)})} \\ &= \sup_{n>l} \|\tilde{A}_n\|_{l_{p'}}\end{aligned}$$

Therefore, $\chi(AS) = \lim_{l\to\infty} \left(\sup_{n>l} \|\tilde{A}_n\|_{l_{p'}}\right) = \limsup_{n\to\infty} \|\tilde{A}_n\|_{l_{p'}} = \limsup_{n\to\infty} \left(\sum_{k=0}^{\infty} |\tilde{a}_{nk}|^{p'}\right)^{\frac{1}{p'}}$. This completes the proof.

(*b*) We first define an operator $P_l : l_\infty \to l_\infty$ by $P_l(\xi) = (\xi_0, \xi_1, \cdots, \xi_l, 0, 0, \cdots)$ for all $\xi = (\xi_k) \in l_\infty$ and $l \in \mathbb{N}_0$. We have

$$AS \subset P_l(AS) + (I - P_l)(AS).$$

By the property of χ, we have

$$\begin{aligned}0 \le \chi(AS) &\le \chi(P_l(AS)) + \chi((I - P_l)(AS)) \\ &= \chi((I - P_l)(AS)) \\ &\le \sup_{x\in S} \|(I - P_l)(Ax)\|_\infty \\ &= \sup_{n>l} \|\tilde{A}_n\|_{l_{p'}}.\end{aligned}$$

Hence

$$0 \le \chi(AS) \le \limsup_{n\to\infty} \|\tilde{A}_n\|_{l_{p'}} = \limsup_{n\to\infty} \left(\sum_{k=0}^{\infty} |\tilde{a}_{nk}|^{p'}\right)^{\frac{1}{p'}}.$$

This completes the proof.

Corollary 1 *Let* $1 < p < \infty$.

(*a*) *If* $A \in (l_p(r, s, t; B^{(m)}), c_0)$, *then* L_A *is compact if and only if*

$$\lim_{n\to\infty} \left(\sum_{k=0}^{\infty} |\tilde{a}_{nk}|^{p'}\right)^{\frac{1}{p'}} = 0.$$

(*b*) *If* $A \in (l_p(r, s, t, B^{(m)}), l_\infty)$, *then* L_A *is compact if* $\lim_{n\to\infty} \left(\sum_{k=0}^{\infty} |\tilde{a}_{nk}|^{p'}\right)^{\frac{1}{p'}} = 0$.

Proof The proof is immediate from Theorem 3.

5 Conclusion

Here we have defined a new sequence space combining the generalized means and difference operator, which is more general than the previous classes of sequences. We have also shown that the new space is a complete normed linear space and also a BK space having Schauder basis. We have characterized some classes of compact operators on this new space using the Hausdorff measure of noncompactness. We have also obtained some identities and estimates for the operator norms and the Hausdorff measure of noncompactness of certain matrix operators.

References

1. Başarir, M., Kayikçi, M.: On the generalized $B^{(m)}$- Riesz difference sequence space and β-property. J. Inequal. Appl. **2009**(385029) 18 (2009)
2. Djolović, I.: On the space of bounded Euler difference sequences and some classes of compact operators. Appl. Math. Comput. **182**(2), 1803–1811 (2006)
3. Djolović, I. Malkowsky, E.: A note on compact operators on matrix domains. J. Math. Anal. Appl. **340**(1), 291–303 (2008)
4. Kara, E.E., Başarir, M.: On compact operators and some Euler $B^{(m)}$-difference sequence spaces. J. Math. Anal. Appl. **379**, 499–511 (2011)
5. Malkowsky, E., Rakočević, V.: An introduction into the theory of sequence spaces and measure of noncompactness. Zb. Rad. (Beogr.) **9**(17), 143–234 (2000)
6. Malkowsky, E., Rakočević, V.: On matrix domains of triangles. Appl. Math. Comput. **189**, 1146–1163 (2007)
7. Mursaleen, M., Noman, A.K.: Applications of the Hausdorff measure of noncompactness in some sequence spaces of weighted means. Comput. Math. Appl. **60**(5), 1245–1258 (2010)
8. Mursaleen, M., Noman, A.k.: On generalized means and some related sequence spaces. Comput. Math. Appl. **61**(4), 988–999 (2011)
9. Wilansky, A.: Summability Through Functional Analysis, vol. 85. North-Holland Mathematics Studies, Elsevier Science Publishers, Amsterdam (1984)

Chapter 19
Some Geometric Properties of Generalized Cesàro–Musielak–Orlicz Sequence Spaces

Atanu Manna and P. D. Srivastava

Abstract A generalized Cesàro–Musielak–Orlicz sequence space $Ces_{\Phi}(q)$ equipped with the Luxemberg norm is introduced. It is proved that $Ces_{\Phi}(q)$ is a Banach space and also criteria for the coordinatewise uniformly Kadec–Klee property and the uniform Opial property are obtained.

Keywords Musielak–Orlicz function · Riesz weighted mean · Luxemberg norm · Coordinatewise Kadec–Klee property · Uniform Opial property

Mathematics Subject Classification (2010): 46B20, 46B45, 46A45, 46A80, 46E30

1 Introduction

In fixed point theory, geometrical properties of Banach space, such as Kadec–Klee property, Opial property, and their several generalizations play fundamental role. In particular, the Opial property of a Banach space has its applications in differential equations and integral equations, etc. On the other hand the Kadec–Klee property has several applications in Ergodic theory and many other branches of analysis [22].

The authors are very much thankful to the anonymous reviewers for their valuable comments, which improved the presentation of the paper. First author is grateful to CSIR, New Delhi, Government of India for the research fellowship with award no. 09/081(0988)/2009-EMR-I during this work.

A. Manna (✉) · P. D. Srivastava
Indian Institute of Technology Kharagpur, Kharagpur 721302, India
e-mail: atanumanna@maths.iitkgp.ernet.in

P. D. Srivastava
e-mail: pds@maths.iitkgp.ernet.in

R. N. Mohapatra et al. (eds.), *Mathematics and Computing 2013*,
Springer Proceedings in Mathematics & Statistics 91,
DOI: 10.1007/978-81-322-1952-1_19,

In recent times, the theory of Cesàro–Orlicz sequence spaces and Musielak–Orlicz sequence spaces and their geometric properties has been studied extensively. Some topological properties like absolute continuity, order continuity, separability, completeness, and relations between norm and modular as well as some geometrical properties like Fatou property, monotonicity, Kadec–Klee property, uniform Opial property, rotundity, local rotundity, property-β etc. are studied in [2–4, 6, 8, 13, 20, 21]. Recently, Khan (see [15, 16]) introduced Riesz–Musielak–Orlicz sequence spaces and studied some geometric properties of this space. Quite recently, Mongkolkeha, and Kumam [17] studied (H)-property and uniform Opial property of generalized Cesàro sequence spaces. Some topological properties of sequence spaces defined by using Orlicz function are also studied in [1, 5, 25]. This motivated us to introduce generalized Cesàro–Musielak–Orlicz sequence spaces, which include the well known Cesàro, generalized Cesàro [24], Cesàro-Orlicz, Cesàro–Musielak–Orlicz sequence spaces etc. in particular cases. In this paper, we have made an attempt to study some of the geometric properties in generalized Cesàro–Musielak–Orlicz sequence spaces.

Throughout the paper, we denote $\mathbb{N}$, $\mathbb{R}$ and $\mathbb{R}^+$ as the set of natural numbers, real numbers, and nonnegative real numbers, respectively. Let $(X, \|.\|)$ be a Banach space and l^0 be the space of all real sequences $x = (x(i))_{i=1}^{\infty}$. Let $S(X)$ and $B(X)$ denote the unit sphere and closed unit ball, respectively. A sequence $(x_l) \subset X$ is said to be ε-separated sequence if separation of the sequence (x_l) denoted by $sep(x_l) = \inf\{\|x_l - x_m\| : l \neq m\} > \varepsilon$ for some $\varepsilon > 0$ [11].

A Banach space X is said to have the *Kadec–Klee property*, denoted by (H), if weakly convergent sequence on the unit sphere is strongly convergent, i.e., convergent in norm [12]. A Banach space X is said to possess *coordinatewise Kadec–Klee property*, denoted by (H_c) [7], if $x \in X$ and every sequence $(x_l) \subset X$ such that

$$\|x_l\| \to \|x\| \text{ and } x_l(i) \to x(i) \text{ for each } i, \text{ then } \|x_l - x\| \to 0.$$

It is known that $X \in (H_c)$ implies $X \in (H)$, because weak convergence in X implies the coordinatewise convergence. A Banach space X has the *coordinatewise uniformly Kadec–Klee property*, denoted by (UKK_c) [27], if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$(x_l) \subset B(X), sep(x_l) \geq \varepsilon, \|x_l\| \to \|x\| \text{ and } x_l(i) \to x(i) \text{ for each } i \text{ implies } \|x\| \leq 1 - \delta.$$

It is known that the property (UKK_c) implies property (H_c).

A Banach space X is said to have the *Opial property* [23] if for every weakly null sequence $(x_l) \subset X$ and every nonzero $x \in X$, we have

$$\liminf_{l\to\infty} \|x_l\| < \liminf_{l\to\infty} \|x_l + x\|.$$

A Banach space X is said to have the *uniform Opial property* [23] if for each $\varepsilon > 0$ there exists $\mu > 0$ such that for any weakly null sequence (x_l) in $S(X)$ and $x \in X$ with $\|x\| \geq \varepsilon$ the following inequality hold:

$$1 + \mu \leq \liminf_{l \to \infty} \|x_l + x\|.$$

In any Banach space X an *Opial property* is important because it ensures that X has a weak fixed point property [9]. Opial in [19] has shown that the space $L_p[0, 2\pi]$ $(p \neq 2, 1 < p < \infty)$ does not have this property, but the Lebesgue sequence space $l_p (1 < p < \infty)$ has.

A map $\varphi : \mathbb{R} \to [0, \infty]$ is said to be an Orlicz function if it is an even, convex, left continuous on $[0, \infty)$, $\varphi(0) = 0$, not identically zero and $\varphi(u) \to \infty$ as $u \to \infty$. A sequence $\Phi = (\varphi_n)$ of Orlicz functions φ_n is called Musielak–Orlicz function [18]. For a Musielak–Orlicz function Φ, the complementary function $\Psi = (\psi_n)$ of Φ is defined in the sense of Young as

$$\psi_n(u) = \sup_{v \geq 0} \{ |u|v - \varphi_n(v)\} \quad \text{for all } u \in \mathbb{R} \text{ and } n \in \mathbb{N}.$$

Given any Musielak–Orlicz function Φ and $x = (x(n))_{n=1}^{\infty} \in l^0$, a convex modular $I_\Phi : l^0 \to [0, \infty]$ is defined by

$$I_\Phi(x) = \sum_{n=1}^{\infty} \varphi_n\Big(|x(n)|\Big) \text{ and}$$

the linear space $l_\Phi = \{x \in l^0 : I_\Phi(rx) < \infty \text{ for some } r > 0\}$ is called Musielak–Orlicz sequence space. The space l_Φ equipped with functional $|||x|||_\Phi^L$ defined by

$$|||x|||_\Phi^L = \inf \left\{ r > 0 : I_\Phi\Big(\frac{x}{r}\Big) \leq 1 \right\}$$

becomes a Banach space. This functional $|||x|||_\Phi^L$ is called Luxemberg norm and the corresponding Musielak–Orlicz sequence space is denoted by l_Φ^L. For the details about Musielak–Orlicz sequence spaces and their geometric properties we refer to the articles [3, 10, 13, 18]. The subspace of l_Φ defined as

$$\left\{ x = (x(n)) \in l^0 : \forall r > 0 \ \exists n_r \in \mathbb{N} \text{ such that } \sum_{n=n_r}^{\infty} \varphi_n\Big(r|x(n)|\Big) < \infty \right\},$$

equipped with the Luxemberg norm induced from l_Φ is denoted by h_Φ^L.

A Musielak–Orlicz function Φ is said to satisfy the δ_2^0-condition denoted by $\Phi \in \delta_2^0$ if there are positive constants a, K, a natural m and a sequence (c_n) of positive numbers such that $(c_n)_{n=m}^{\infty} \in l_1$ and the inequality

$$\varphi_n(2u) \leq K\varphi_n(u) + c_n \tag{1}$$

holds for every $n \in \mathbb{N}$ whenever $\varphi_n(u) \leq a$. If a Musielak–Orlicz function Φ satisfies δ_2^0-condition with $m = 1$, then Φ is said to satisfy δ_2-condition [10, 18].

For any Musielak–Orlicz function Φ, h_Φ coincides with l_Φ if and only if Φ satisfies δ_2^0-condition [10].

A Musielak–Orlicz function $\Phi = (\varphi_n)_{n=1}^{\infty}$ satisfies the condition $(*)$ [13] if for any $\varepsilon \in (0, 1)$ there is a $\delta > 0$ such that

$$\varphi_n(u) < 1 - \varepsilon \text{ implies } \varphi_n((1+\delta)u) \leq 1 \text{ for all } n \in \mathbb{N} \text{ and } u \geq 0. \tag{2}$$

A Musielak–Orlicz function Φ is to said to vanishes only at zero, which is denoted by $\Phi > 0$ if $\varphi_n(u) > 0$ for any $n \in \mathbb{N}$ and $u > 0$.

2 Class $Ces_{\Phi}(q)$

Let $q = (q_n)_{n=1}^{\infty}$ be a sequence of real numbers with $q_k \geq 1$ for $k \in \mathbb{N}$, and $Q_n = \sum_{k=1}^{n} q_k$. We introduce the *Riesz weighted mean* map R^q on l^0 as $R^q : l^0 \to [0, \infty)$ such that $x \to R^q x$, where

$$R^q x = (R^q x(n))_{n=1}^{\infty}, \text{ with } R^q x(n) = \frac{1}{Q_n} \sum_{k=1}^{n} q_k |x(k)| \text{ for each } n = 1, 2, \ldots$$

and $x \in l^0$.

Using this *Riesz weighted mean* map and a Musielak–Orlicz function $\Phi = (\varphi_n)$, we define on l^0 a functional $\sigma_\Phi(x)$ by

$$\sigma_\Phi(x) = I_\Phi(R^q x) = \sum_{n=1}^{\infty} \varphi_n\Big(\frac{1}{Q_n} \sum_{k=1}^{n} q_k |x(k)|\Big).$$

Since Φ is convex, so it is easy to verify that $\sigma_\Phi(x)$ is a convex modular on l^0(for definition see [18]), i.e., it satisfies $\sigma_\Phi(x) = 0$ if and only if $x = 0, \sigma_\Phi(-x) = \sigma_\Phi(x)$, $\sigma_\Phi(\gamma x + \delta y) \leq \gamma \sigma_\Phi(x) + \delta \sigma_\Phi(y)$ whenever $x, y \in l^0$ and $\gamma, \delta \geq 0$ with $\gamma + \delta = 1$.

We now introduce the space $Ces_\Phi(q)$ as follows:

$$Ces_\Phi(q) = \{x \in l^0 : R^q x \in l_\Phi\} = \{x \in l^0 : \sigma_\Phi(rx) < \infty \text{ for some } r > 0\}.$$

Clearly, it is a linear space and also forms a normed linear space under the norm $\|x\|_\Phi^L = |||R^q x|||_\Phi^L$ introduced with the help of the norm on l_Φ. We call $Ces_\Phi(q)$ as the generalized Cesàro–Musielak–Orlicz sequence space.

The generalized class $Ces_\Phi(q)$ include the following classes in particular cases:

(i) When $q_n = 1, n = 1, 2, \ldots$, the $Ces_\Phi(q)$ reduces to the Cesàro–Musielak –Orlicz sequence space ces_Φ studied by Wangkeeree [26], where

$$ces_\Phi = \left\{x \in l^0 : \sum_{n=1}^{\infty} \varphi_n\left(\frac{r}{n}\sum_{k=1}^{n}|x(k)|\right) < \infty \text{ for some } r > 0\right\},$$

(ii) For $\varphi_n = \varphi$, $\forall n$ the ces_Φ becomes well-known Cesàro–Orlicz sequence space ces_φ studied recently by Cui et al. [2], Foralewski et al. [6], Petrot and Suantai [20],

(iii) For $\varphi_n(x) = |x|^{p_n}$, $p_n \geq 1$ $\forall n$ the $Ces_\Phi(q)$ reduces to the sequence space $Ces_{(p)}(q)$ studied by Mongkolkeha and Kumam [17] and when $\varphi_n(x) = |x|^{p_n}$ with $p_n = p \geq 1$ $\forall n$ then $Ces_\Phi(q)$ reduces to the sequence space $Ces_p(q)$ studied by Khan [14].

We consider the subspace $(Ces_\Phi^L(q))_a$ of $Ces_\Phi(q)$ as

$$(Ces_\Phi(q))_a = \left\{x \in Ces_\Phi(q) : \forall r > 0 \ \exists n_r \text{ such that } \sum_{n=n_r}^{\infty} \varphi_n\left(\frac{r}{Q_n}\sum_{k=1}^{n} q_k|x(k)|\right) < \infty\right\}.$$

In this article, we have introduced the generalized Cesàro–Musielak–Orlicz sequence space and have established the completeness property of the space and also obtained criteria for some geometric properties like coordinatewise Uniform Kadec–Klee property, uniform Opial property with respect to the Luxemberg norm.

Notations:

For any $x \in l^0$ and $i \in \mathbb{N}$, throughout the paper we use the following notations:

$x|_i = (x(1), x(2), x(3), \ldots, x(i), 0, 0, \ldots)$, called the truncation of x at i,

$x|_{\mathbb{N}-i} = (0, 0, 0, \ldots, 0, x(i+1), x(i+2), \ldots)$,

$x|_I = \{x = (x(i)) \in l^0 : x(i) \neq 0 \text{ for all } i \in I \subseteq \mathbb{N} \text{ and } x(i) = 0 \text{ for all } i \in \mathbb{N} \setminus I\}$,

For simplifying notations, we write $Ces_\Phi^L(q) = (Ces_\Phi(q), \|.\|_\Phi^L)$.

3 Main Results

This section contains main results of our work.

Theorem 1 *Let Φ be a Musielak–Orlicz function. Then the following statements are true:*

(i) $(Ces_\Phi(q), \|.\|_\Phi^L)$ *is a Banach space,*
(ii) $(Ces_\Phi^L(q))_a$ *is a closed subspace of* $Ces_\Phi^L(q)$,
(iii) if Φ *satisfies* δ_2*-condition then* $(Ces_\Phi^L(q))_a = Ces_\Phi^L(q)$.

Proof Let $(x^s)_{s=1}^{\infty}$ be a Cauchy sequence in $Ces_{\Phi}^{L}(q)$, where $x^s = (x^s(k))_{k=1}^{\infty}$ and $\varepsilon > 0$ be given. Then there exists a natural number T such that for every $\varepsilon > 0$ one can find r_ε with $r_\varepsilon < \varepsilon$, we have

$$\sigma_{\Phi}\left(\frac{x^s - x^t}{r_\varepsilon}\right) \leq 1 \text{ for all } s, t \geq T.$$

By definition of σ_Φ for each $l \in \mathbb{N}$, we have

$$\sum_{n=1}^{l} \varphi_n\left(\frac{1}{r_\varepsilon Q_n}\sum_{k=1}^{n} q_k|x^s(k) - x^t(k)|\right) \leq 1 \text{ for all } s, t \geq T, \tag{3}$$

which implies that for each $l \geq n \geq 1$

$$\varphi_n\left(\frac{1}{r_\varepsilon Q_n}\sum_{k=1}^{n} q_k|x^s(k) - x^t(k)|\right) \leq 1 \text{ for all } s, t \geq T. \tag{4}$$

Let p_n be the corresponding kernel of the Orlicz function φ_n for each n. We choose a constant $s_0 > 0$ and $\gamma > 1$ such that $\gamma \frac{s_0}{2} p_n(\frac{s_0}{2}) \geq 1$, for each $n \in \mathbb{N}$ (which is follows from $\varphi_n(\frac{s_0}{2}) = \int_0^{\frac{s_0}{2}} p_n(t)\mathrm{d}t$ and $s_0 > 0$).

By the integral representation of φ_n for each n, we have

$$\frac{1}{r_\varepsilon Q_n}\sum_{k=1}^{n} q_k|x^s(k) - x^t(k)| \leq \gamma s_0 \text{ for each } n \in \mathbb{N} \text{ and for all } s, t \geq T. \tag{5}$$

Otherwise, one can find a natural n with $\frac{1}{r_\varepsilon Q_n}\sum_{k=1}^{n} q_k|x^s(k) - x^t(k)| > \gamma s_0$ such that

$$\varphi_n\left(\sum_{k=1}^{n}\frac{q_k|x^s(k) - x^t(k)|}{r_\varepsilon Q_n}\right) \geq \int_{\frac{\gamma s_0}{2}}^{\sum_{k=1}^{n}\frac{q_k|x^s(k) - x^t(k)|}{r_\varepsilon Q_n}} p_n(t)\mathrm{d}t > \frac{\gamma s_0}{2} p_n(\frac{s_0}{2}),$$

which contradicts (4). Hence from (5), we have $(x^s(k))_{s=1}^{\infty}$ is a Cauchy sequence of real numbers for each k and hence converges for each k. Suppose for each $k \in \mathbb{N}$, $\lim_{t\to\infty} x^t(k) = x(k)$. Taking $t \to \infty$ in (3), we obtain for each $l \in \mathbb{N}$

$$\sum_{n=1}^{l} \varphi_n\left(\frac{1}{r_\varepsilon Q_n}\sum_{k=1}^{n} q_k|x^s(k) - x(k)|\right) \leq 1 \text{ for all } s \geq T,$$

which implies that $\sigma_\Phi\left(\frac{x^s-x}{r_\varepsilon}\right) \leq 1$ for all $s \geq T$, i.e., $\|x^s - x\|_\Phi^L \leq r_\varepsilon < \varepsilon$ for all $s \geq T$. Therefore $x^s \to x$ in $\|.\|_\Phi^L$ as $s \to \infty$. We omit the verification of $x \in Ces_\Phi^L(q)$ as it is easy to obtain. This finishes the proof of part (i).

(ii) Clearly $(Ces_\Phi^L(q))_a$ is a subspace $Ces_\Phi^L(q)$. It is sufficient to show that $(Ces_\Phi^L(q))_a$ is a closed subspace of $Ces_\Phi^L(q)$. For this, let $x_i = (x_i(k))_{k=1}^\infty \in (Ces_\Phi^L(q))_a$ for each $i \in \mathbb{N}$ and $\|x - x_i\|_\Phi^L \to 0$ as $i \to \infty$ and $x \in Ces_\Phi^L(q)$. We show that $x \in (Ces_\Phi^L(q))_a$. By the equivalent definition of norm and modular convergence, we have $\sigma_\Phi(r(x-x_i)) \to 0$ as $i \to \infty$ for all $r > 0$. So for all $r > 0$ there exists $J \in \mathbb{N}$ such that $\sigma_\Phi(2r(x-x_J)) < 1$. Since $x_J \in (Ces_\Phi^L(q))_a$ so there exists n_J such that $\sum_{n=n_J}^\infty \varphi_n\Big(\frac{2r}{Q_n}\sum_{k=1}^n |q_k x_J(k)|\Big) < \infty \ \forall r > 0$. We choose $n_r = n_J$, then we have

$$\begin{aligned}
&\sum_{n=n_J}^\infty \varphi_n\Big(\frac{r}{Q_n}\sum_{k=1}^n q_k|x(k)|\Big)\\
&\leq \sum_{n=n_J}^\infty \varphi_n\Big(\frac{r}{2Q_n}\sum_{k=1}^n 2q_k|x(k)-x_J(k)| + \frac{r}{2Q_n}\sum_{k=1}^n 2q_k|x_J(k)|\Big)\\
&\leq \frac{1}{2}\sum_{n=n_J}^\infty \varphi_n\Big(\frac{2r}{Q_n}\sum_{k=1}^n q_k|x(k)-x_J(k)|\Big) + \frac{1}{2}\sum_{n=n_J}^\infty \varphi_n\Big(\frac{2r}{Q_n}\sum_{k=1}^n q_k|x_J(k)|\Big)\\
&\leq \frac{1}{2}\sigma_\Phi(2r(x-x_J)) + \frac{1}{2}\sum_{n=n_J}^\infty \varphi_n\Big(\frac{2r}{Q_n}\sum_{k=1}^n q_k|x_J(k)|\Big) < \infty.
\end{aligned}$$

Since r is arbitrary, we have $x \in (Ces_\Phi^L(q))_a$. This completes the proof.

(iii) We need to show here only the inclusion $Ces_\Phi^L(q) \subset (Ces_\Phi^L(q))_a$. Let $x \in Ces_\Phi^L(q)$. Then for some $t > 0, \sigma_\Phi(tx) < \infty$, i.e., $\sum_{n=1}^\infty \varphi_n\Big(\frac{t}{Q_n}\sum_{k=1}^n q_k|x(k)|\Big) < \infty$. We show that for any $r > 0$ there exists a $n_r \in \mathbb{N}$ such that

$$\sum_{n=n_r}^\infty \varphi_n\Big(\frac{r}{Q_n}\sum_{k=1}^n q_k|x(k)|\Big) < \infty.$$

If $r \in [0, t]$ then it is easily follows from

$$\sum_{n=n_r}^\infty \varphi_n\Big(\frac{r}{Q_n}\sum_{k=1}^n q_k|x(k)|\Big) \leq \sum_{n=n_r}^\infty \varphi_n\Big(\frac{t}{Q_n}\sum_{k=1}^n q_k|x(k)|\Big) < \infty.$$

Now, we fix t and choose $r > t$. Since $x \in Ces_\Phi^L(q)$, i.e., for some $t > 0, \sigma_\Phi(tx) < \infty$, so there exists n_r and a constant a such that

$$\sum_{n=n_r}^{\infty} \varphi_n\Big(\frac{t}{Q_n}\sum_{k=1}^{n} q_k|x(k)|\Big) < \frac{a}{2}.$$

Therefore for each $n \geq n_r$, we have

$$\varphi_n\Big(\frac{t}{Q_n}\sum_{k=1}^{n} q_k|x(k)|\Big) < \frac{a}{2}.$$

Choose a sequence $(c_n)_{n=1}^{\infty}$ of positive real numbers such that $\sum_{n=1}^{\infty} c_n < \infty$. So for a given $\varepsilon > 0$, there exists a n_r such that $\sum_{n=n_r}^{\infty} c_n < \frac{\varepsilon}{2}$. Let $u = \frac{t}{Q_n}\sum_{k=1}^{n} q_k|x(k)|$, $K > 0$ be a constant and a is chosen above. Since $r > t$ so there is a $l \in \mathbb{N}$ such that $r \leq 2^l t$. Applying δ_2-condition for all $n \geq n_r$, we have

$$\varphi_n\Big(\frac{r}{Q_n}\sum_{k=1}^{n} q_k|x(k)|\Big) \leq \varphi_n\Big(\frac{2^l t}{Q_n}\sum_{k=1}^{n} q_k|x(k)|\Big) \leq K^l \varphi_n\Big(\frac{t}{Q_n}\sum_{k=1}^{n} q_k|x(k)|\Big) + \Big(\sum_{i=0}^{l-1} K^i\Big)c_n$$

Taking summation on both sides over $n \geq n_r$, we obtain

$$\sum_{n=n_r}^{\infty} \varphi_n\Big(\frac{r}{Q_n}\sum_{k=1}^{n} q_k|x(k)|\Big) \leq K^l \sum_{n=n_r}^{\infty} \varphi_n\Big(\frac{t}{Q_n}\sum_{k=1}^{n} q_k|x(k)|\Big) + \Big(\sum_{i=0}^{l-1} K^i\Big)\sum_{n=n_r}^{\infty} c_n < \infty.$$

Hence $x \in (Ces_{\Phi}^{L}(q))_a$.

We assume in the rest of this work that Musielak–Orlicz function $\Phi = (\varphi_n)$ with all φ_n being finitely valued. The following known lemmas are useful in the sequel:

Lemma 1 *Let $x \in (Ces_{\Phi}^{L}(q))_a$ be an arbitrary element. Then $\|x\|_{\Phi}^{L} = 1$ if and only if $\sigma_{\Phi}(x) = 1$.*

Proof The proof will run on the parallel lines of the proof of Lemma 2.1 in [2].

Lemma 2 *Suppose $\Phi \in \delta_2$ and $\Phi > 0$. Then for any sequence (x_l) in $Ces_{\Phi}^{L}(q)$, $\|x_l\|_{\Phi}^{L} \to 0$ if and only if $\sigma_{\Phi}(x_l) \to 0$.*

Proof For the proof of this lemma see [7, 13].

Lemma 3 *If $\Phi \in \delta_2$, i.e., (1), then for any $x \in Ces_{\Phi}^{L}(q)$,*

$$\|x\|_{\Phi}^{L} = 1 \text{ if and only if } \sigma_{\Phi}(x) = 1.$$

Proof Since $\Phi \in \delta_2$ implies $Ces_\Phi^L(q) = (Ces_\Phi^L(q))_a$. The proof follows from Lemma 1.

Lemma 4 *Let $\Phi \in \delta_2$, i.e., (1) and satisfies the condition $(*)$, i.e., (2). Then for any $x \in Ces_\Phi^L(q)$ and every $\varepsilon \in (0,1)$ there exists $\delta(\varepsilon) \in (0,1)$ such that $\sigma_\Phi(x) \leq 1-\varepsilon$ implies $\|x\|_\Phi^L \leq 1-\delta$.*

Proof The proof of this lemma will be in a way similar to that of the proof of Lemma 9 in [13].

Lemma 5 [13] *Let $(X, \|.\|)$ be normed space. If $f : X \to \mathbb{R}$ is a convex function in the set $K(0,1) = \{x \in X : \|x\| \leq 1\}$ and $|f(x)| \leq M$ for all $x \in K(0,1)$ and some $M > 0$ then f is almost uniformly continuous in $K(0,1)$; i.e., for all $d \in (0,1)$ and $\varepsilon > 0$ there exists a $\delta > 0$ such that $\|y\| \leq d$ and $\|x-y\| < \delta$ implies $|f(x)-f(y)| < \varepsilon$ for all $x, y \in K(0,1)$.*

Lemma 6 *Let $\Phi \in \delta_2$, i.e., (1), $\Phi > 0$ and satisfies the condition $(*)$, i.e., (2). Then for each $d \in (0,1)$ and $\varepsilon > 0$ there exists $\delta = \delta(d,\varepsilon) > 0$ such that $\sigma_\Phi(x) \leq d$, $\sigma_\Phi(y) \leq \delta$ imply*

$$|\sigma_\Phi(x+y) - \sigma_\Phi(x)| < \varepsilon \text{ for any } x, y \in Ces_\Phi^L(q). \tag{6}$$

Proof Since $\Phi \in \delta_2$ and satisfies condition $(*)$, so by Lemma 4, there exists $d_1 \in (0,1)$ such that $\|x\|_\Phi^L \leq d_1$. Also by Lemma 2, we find a $\delta > 0$ such that for every $\delta_1 > 0$, $\sigma_\Phi(y) \leq \delta$ implies $\|y\|_\Phi^L \leq \delta_1$ for any $y \in Ces_\Phi^L(q)$. So, if $\sigma_\Phi(x) \leq d$ and $\sigma_\Phi(y) \leq \delta$ then $\|x\|_\Phi^L \leq d_1$ and $\|y\|_\Phi^L \leq \delta_1$. Hence by Lemma 5, we have $|\sigma_\Phi(x+y) - \sigma_\Phi(x)| < \varepsilon$ because the functional σ_Φ satisfies all the assumptions of f defined in Lemma 5.

Lemma 7 *Let $\Phi \in \delta_2$, i.e., (1) and satisfies the condition $(*)$, i.e., (2) and $\Phi > 0$. Then for any $x \in Ces_\Phi^L(q)$ and any $\varepsilon > 0$ there exists $\delta = \delta(\varepsilon) > 0$ such that $\sigma_\Phi(x) \geq 1+\varepsilon$ implies $\|x\|_\Phi^L \geq 1+\delta$.*

Proof The proof of this lemma is parallel to the proof of the Lemma 4 in [3].

Theorem 2 *Let $\Phi > 0$ be a Musielak–Orlicz function satisfying condition δ_2, i.e., (1) and $(*)$, i.e., (2). Then sequence space $Ces_\Phi^L(q)$ has the UKK_c-property.*

Proof Since $\Phi > 0$ and it satisfies the condition δ_2, so by Lemma 2, for a given $\varepsilon > 0$ there exist a $\eta > 0$, we have

$$\|x\|_\Phi^L \geq \frac{\varepsilon}{4} \Rightarrow \sigma_\Phi(x) \geq \eta. \tag{7}$$

With this $\eta > 0$, by Lemma 4, one can find a $\delta \in (0,1)$ such that

$$\|x\|_\Phi^L > 1-\delta \Rightarrow \sigma_\Phi(x) > 1-\eta. \tag{8}$$

Let $(x_l) \subset B(Ces_\Phi^L(q))$, $\|x_l\|_\Phi^L \to \|x\|_\Phi^L$, $x_l(i) \to x(i)$ for all $i \in \mathbb{N}$ and $sep(x_l) \geq \varepsilon$. We show that there exists a $\delta > 0$ such that $\|x\|_\Phi^L \leq 1 - \delta$. If possible, let $\|x\|_\Phi^L > 1 - \delta$. Then one can select a finite set $I = \{1, 2, \ldots, N-1\}$ on which $\|x|_I\|_\Phi^L > 1 - \delta$. Since $x_l(i) \to x(i)$ for each $i \in \mathbb{N}$, so we obtain $x_l \to x$ uniformly on I. Consequently, by assumption $\|x_l\|_\Phi^L \to \|x\|_\Phi^L$ there exists $l_N \in \mathbb{N}$ such that

$$\|x_l|_I\|_\Phi^L > 1 - \delta \text{ and } \|(x_l - x_m)|_I\|_\Phi^L \leq \frac{\varepsilon}{2} \text{ for all } l, m \geq l_N.$$

Using Eq. (8), first one of the above inequalities implies that $\sigma_\Phi(x_l|_I) > 1 - \eta$ for $l \geq l_N$. Since $sep(x_l) \geq \varepsilon$, i.e., $\|x_l - x_m\|_\Phi^L \geq \varepsilon$, so second one of the above inequalities implies that $\|(x_l - x_m)|_{\mathbb{N}-I}\|_\Phi^L \geq \frac{\varepsilon}{2}$ for $l, m \geq l_N, l \neq m$. Hence for $N \in \mathbb{N}$ there exists a l_N such that $\|x_{l_N}|_{\mathbb{N}-I}\|_\Phi^L \geq \frac{\varepsilon}{4}$. Without loss of generality, we assume that $\|x_l|_{\mathbb{N}-I}\|_\Phi^L \geq \frac{\varepsilon}{4}$ for all $l, N \in \mathbb{N}$. Therefore by (7), we have $\sigma_\Phi(x_l|_{\mathbb{N}-I}) \geq \eta$.

By the integral representation of Musielak–Orlicz function Φ, we have $\varphi_n(u+v) \geq \varphi_n(u) + \varphi_n(v)$ for each n and all $u, v \in \mathbb{R}^+$. Using this, we obtain $\sigma_\Phi(x_l|_I) + \sigma_\Phi(x_l|_{\mathbb{N}-I}) \leq \sigma_\Phi(x_l) \leq 1$. This implies that $\sigma_\Phi(x_l|_{\mathbb{N}-I}) \leq 1 - \sigma_\Phi(x_l|_I) < 1 - (1 - \eta) = \eta$, i.e., $\sigma_\Phi(x_l|_{\mathbb{N}-I}) < \eta$, which contradicts to the fact that $\sigma_\Phi(x_l|_{\mathbb{N}-I}) \geq \eta$. This finishes the proof.

Theorem 3 *Let $\Phi > 0$ be a Musielak–Orlicz function satisfying condition δ_2, i.e., (1) and (∗), i.e., (2). Then $Ces_\Phi^L(q)$ has the uniform Opial property.*

Proof Let $(x_l) \subset S(Ces_\Phi^L(q))$ be any weakly null sequence and $\varepsilon > 0$ be given. We show that for any $\varepsilon > 0$ there is a $\mu > 0$ such that

$$\liminf_{l \to \infty} \|x_l + x\|_\Phi^L \geq 1 + \mu,$$

for each $x \in Ces_\Phi^L(q)$ satisfying $\|x\|_\Phi^L \geq \varepsilon$. Since $\Phi \in \delta_2$ and $\Phi > 0$, so by Lemma 2, for each $\varepsilon > 0$ there is a number $\delta \in (0, 1)$ such that for each $x \in Ces_\Phi^L(q)$, we have $\sigma_\Phi(x) \geq \delta$. Since Φ (> 0) satisfies the condition δ_2, and the condition (∗), so by Lemma 6 for any $\varepsilon > 0$, there exists $\delta_1 \in (0, \delta)$ such that $\sigma_\Phi(u) \leq 1, \sigma_\Phi(v) \leq \delta_1$ imply

$$|\sigma_\Phi(u + v) - \sigma_\Phi(u)| < \frac{\delta}{6} \text{ for any } u, v \in Ces_\Phi^L(q). \tag{9}$$

Since $\sigma_\Phi(x) < \infty$, so there is a number $n_0 \in \mathbb{N}$ such that

$$\sum_{n=n_0+1}^{\infty} \varphi_n\Big(\frac{1}{Q_n} \sum_{k=1}^{n} q_k |x(k)|\Big) \leq \frac{\delta_1}{6}. \tag{10}$$

From Eq. (10) it follows that

$$\delta \le \sum_{n=1}^{n_0} \varphi_n\Big(\frac{1}{Q_n}\sum_{k=1}^{n} q_k|x(k)|\Big) + \sum_{n=n_0+1}^{\infty} \varphi_n\Big(\frac{1}{Q_n}\sum_{k=1}^{n} q_k|x(k)|\Big)$$
$$\le \sum_{n=1}^{n_0} \varphi_n\Big(\frac{1}{Q_n}\sum_{k=1}^{n} q_k|x(k)|\Big) + \frac{\delta_1}{6},$$

which implies $\sum_{n=1}^{n_0} \varphi_n\Big(\frac{1}{Q_n}\sum_{k=1}^{n} q_k|x(k)|\Big) \ge \delta - \frac{\delta_1}{6} > \delta - \frac{\delta}{6} = \frac{5\delta}{6}$. Since $x_l \to 0$ weakly, i.e., $x_l(i) \to 0$ for each i, so there exists a l_0 such that for all $l \ge l_0$, the last inequality yields

$$\sum_{n=1}^{n_0} \varphi_n\Big(\frac{1}{Q_n}\sum_{k=1}^{n} q_k|x_l(k) + x(k)|\Big) \ge \frac{5\delta}{6}. \tag{11}$$

Also by $x_l \to 0$ weakly, we can choose an n_0 such that $\sigma_\Phi(x_l|_{n_0}) \to 0$ as $l \to \infty$. So there exists a $l_1 > l_0$ such that $\sigma_\Phi(x_l|_{n_0}) \le \delta_1$ for all $l \ge l_1$. Since $(x_l) \subset S(Ces_\Phi^L(q))$, i.e., $\|x_l\|_\Phi^L = 1$, so by Lemma 3, we have $\sigma_\Phi(x_l) = 1$, which implies that there exists n_0 such that $\sigma_\Phi(x_l|_{\mathbb{N}-n_0}) \le 1$. Now choose $u = x_l|_{\mathbb{N}-n_0}$ and $v = x_l|_{n_0}$. Then $u, v \in Ces_\Phi^L(q)$, $\sigma_\Phi(u) \le 1$, $\sigma_\Phi(v) \le \delta_1$. So from (9), for all $l \ge l_1$ we have

$$\big|\sigma_\Phi(x_l|_{\mathbb{N}-n_0} + x_l|_{n_0}) - \sigma_\Phi\big(x_l|_{\mathbb{N}-n_0}\big)\big| < \frac{\delta}{6},$$

which implies that $\sigma_\Phi(x_l) - \frac{\delta}{6} < \sigma_\Phi\big(x_l|_{\mathbb{N}-n_0}\big)$ for all $l \ge l_1$, i.e.,

$\sum_{n=n_0+1}^{\infty} \varphi_n\Big(\frac{1}{Q_n}\sum_{k=1}^{n} q_k|x_l(k)|\Big) > 1 - \frac{\delta}{6}$ for all $l \ge l_1$. Again, since $\sigma_\Phi\big(x_l|_{\mathbb{N}-n_0}\big) \le 1$ and $\sigma_\Phi\big(x|_{\mathbb{N}-n_0}\big) \le \frac{\delta_1}{6} < \delta_1$, so from the Eqs. (9) and (11), we obtain

$$\begin{aligned}
\sigma_\Phi(x_l + x) &= \sum_{n=1}^{n_0} \varphi_n\Big(\frac{1}{Q_n}\sum_{k=1}^{n} q_k|x_l(k) + x(k)|\Big) \\
&\quad + \sum_{n=n_0+1}^{\infty} \varphi_n\Big(\frac{1}{Q_n}\sum_{k=1}^{n} q_k|x_l(k) + x(k)|\Big) \\
&> \sum_{n=1}^{n_0} \varphi_n\Big(\frac{1}{Q_n}\sum_{k=1}^{n} q_k|x_l(k) + x(k)|\Big) \\
&\quad + \sum_{n=n_0+1}^{\infty} \varphi_n\Big(\frac{1}{Q_n}\sum_{k=1}^{n} q_k|x_l(k)|\Big) - \frac{\delta}{6} \\
&> \frac{5\delta}{6} + \Big(1 - \frac{\delta}{6}\Big) - \frac{\delta}{6} = 1 + \frac{\delta}{2}.
\end{aligned}$$

Since $\Phi \in \delta_2$ and satisfies the condition $(*)$ and $\Phi > 0$, so by Lemma 7 there is a $\mu > 0$ depending only on δ such that $\|x_l + x\|_\Phi^L > 1 + \mu$. Hence $\liminf_{l\to\infty} \|x_l + x\|_\Phi^L \geq 1 + \mu$. This completes the proof.

Corollary 1 *(i) If $\varphi_n = \varphi$, $q_n = 1$ $\forall n$ and $\Phi \in \delta_2$, then Cesàro–Orlicz sequence space ces_φ^L [20] has the uniform Opial property.*

(ii) Suppose $q_n = 1$, $n = 1, 2, \ldots$ and $\varphi_n(u) = |u|^{p_n}$ for all $u \in \mathbb{R}$, $1 < p_n < \infty$ $\forall n$. Then it is easy to verify that $\Phi \in \delta_2$ if and only if $\limsup_{n\to\infty} p_n < \infty$. Therefore $ces_{(p)}^L$ [21] has the uniform Opial property.

(iii) If $\varphi_n(u) = |u|^{p_n}$, $1 \leq p_n < \infty$ $\forall n$ and $\limsup_{n\to\infty} p_n < \infty$, then $Ces_{(p)}^L(q)$ has the uniform Opial property [17].

4 Conclusion

In this study, we have obtained geometric properties such as coordinatewise uniformly Kadec–Klee property and uniform Opial property in the generalized Cesàro–Musielak–Orlicz sequence spaces, which include the well known Cesàro [24], generalized Cesàro [21], Cesàro–Orlicz [2], Cesàro–Musielak–Orlicz [26] classes of sequences in particular cases with respect to the Luxemberg norm. In future, our plan is to obtain these results for a more generalized class of sequences with respect to both the Luxemberg and Amemiya norm.

References

1. Altin, Y., Et, M., Tripathy, B.C.: The sequence space $|\bar{N}_p|(M, r, q, s)$ on seminormed spaces. Appl. Math. Comput. **154**, 423–430 (2004)
2. Cui, Y., Hudzik, H., Petrot, N., Suantai, S., Szymaszkiewicz, A.: Basic topological and geometric properties of Cesàro-Orlicz spaces. Proc. Indian Acad. Sci. (Math. Sci.) **115**(4), 461–476 (2005)
3. Cui, Y., Hudzik, H.: Maluta's coefficient and Opial's properties in Musielak-Orlicz sequence spaces equipped with the Luxemberg norm. Nonlinear Anal. **35**, 475–485 (1999)
4. Cui, Y., Hudzik, H.: On the uniform opial property in some modular sequence spaces. Func. Approx. Comment. **26**, 93–102 (1998)
5. Et, M., Altin, Y., Choudhary, B., Tripathy, B.C.: On some classes of sequences defined by sequences of Orlicz functions. Math. Ineq. Appl. **9**(2), 335–342 (2006)
6. Foralewski, P., Hudzik, H., Szymaszkiewicz, A.: Some remarks on Cesàro-Orlicz spaces. Math. Ineq. Appl. **13**(2), 363–386 (2010)
7. Foralewski, P., Hudzik, H.: On some geometrical and topological properties of generalized Calderón -Lozanovskiĭ sequence spaces. Houston J. Math. **25**(3), 523–542 (1999)
8. Foralewski, P., Hudzik, H., Szymaszkiewicz, A.: Local rotundity structure of Cesàro-Orlicz sequence spaces. J. Math. Anal. Appl. **345**(1), 410–419 (2008)
9. Gossez, J.P., Lami Dozo, E.: Some geometric properties related to fixed point theory for non expansive mappings. Pac. J. Math. **40**, 565–573 (1972)

10. Hudzik, H., Ye, Y.N.: Support functionals and smoothness in Musielak-Orlicz sequence spaces endowed with Luxemberg norm. Comment. Math. Univ. Carolinae **31**(4), 661–684 (1990)
11. Huff, R.: Banach spaces which are nearly uniformly convex. Rockey Mt. J. Math. **10**, 743–749 (1980)
12. Kadets, M.I.: The relation between some properties of convexity of the unit ball of a Banach space. Funct. Anal. Appl. **16**(3), 204–206 (1982) (translation)
13. Kaminska, A.: Uniform rotundity of Musielak-Orlicz sequence spaces. J. Approximation Theor. **47**(4), 302–322 (1986)
14. Khan, V.A.: Some geometric properties of generalized Cesaro sequence spaces. Acta Math. Univ. Comenianae **79**(1), 1–8 (2010)
15. Khan, V.A.: On Riesz-Musielak-Orlicz sequence spaces. Numer. Funct. Anal. Optim. **28**(7–8), 883–895 (2007)
16. Khan, V.A.: Some geometric properties of Riesz-Musielak-Orlicz sequence spaces. Thai J. Math. **8**(3), 565–574 (2010)
17. Mongkolkeha, C., Kumam, P.: On H-property and Uniform Opial Property of generalized Cesaro sequence spaces. J. Inequalities Appl. **2012**, 76 (2012)
18. Musielak, J.: Orlicz Spaces and Modular Spaces. Springer Lecture Notes in Mathematics, vol. 1034. Springer, Berlin (1983)
19. Opial, Z.: Weak convergence of the sequence of successive approximations for non expansive mappings. Bull. Amer. Math. Soc. **73**, 591–597 (1967)
20. Petrot, N., Suantai, S.: Some geometric properties in Cesàro-Orlicz Spaces. Sci. Asia **31**, 173–177 (2005)
21. Petrot, N., Suantai, S.: Uniform Opial properties in generalized Cesàro sequence spaces. Nonlinear Anal. **63**, 1116–1125 (2005)
22. Prus, S.: Geometrical background of metric fixed point theory. In: Kirk, W.A., Sims, B. (eds.) Handbook of Metric Fixed Point Theory, pp. 93–132. Kluwer Academic Publishers, Dordrecht (2001)
23. Prus, S.: Banach spaces with uniform Opial property. Nonlinear Anal. Theory Appl. **18**(8), 697–704 (1992)
24. Shiue, J.S.: Cesàro sequence spaces. Tamkang J. Math. **1**, 19–25 (1970)
25. Tripathy, B.C., Altin, Y., Et, M.: Generalized difference sequence spaces on seminormed spaces defined by Orlicz functions. Math. Slovaca **58**(3), 315–324 (2008)
26. Wangkeeree, R.: On property $(k\text{-}NUC)$ in Cesàro-Musielak-Orlicz sequence spaces. Thai J. Math. **1**, 119–130 (2003)
27. Zhang, T.: The coordinatewise uniformly Kadec-Klee property in some Banach spaces. Siberian Math. J. **44**(2), 363–365 (2003)

Chapter 20
Inverting the Transforms Arising in the $GI/M/1$ Risk Process Using Roots

Gopinath Panda, A. D. Banik and M. L. Chaudhry

Abstract We consider an insurance risk model for which the claim arrival process is a renewal process and the sizes of claims occur an exponentially distributed random variable. For this risk process, we give an explicit expression for the distribution of probability of ultimate ruin, the expected time to ruin and the distribution of deficit at the time of ruin, using Padé-Laplace method. We have derived results about ultimate ruin probability and the time to ruin in the renewal risk model from its dual queueing model. Also, we derive the bounds for the moments of recovery time. Finally, some numerical results have been presented in the form of tables which compare these results with some of the existing results available in the literature.

Keywords Risk process · Ruin probability · Recovery time · Expected time to ruin · M/G/1 and GI/M/1 queue · Roots · Padé-Laplace method

The second author received partial financial support from the Department of Science and Technology, New Delhi, India research grant SR/FTP/MS-003/2012. The third author's research work was partially supported by NSERC.

G. Panda · A. D. Banik (✉)
School of Basic Sciences, Indian Institute of Technology Bhubaneswar,
Bhubaneswar 751007, India
e-mail: banikad@gmail.com; adattabanik@iitbbs.ac.in

G. Panda
e-mail: gopinath.panda@gmail.com

M. L. Chaudhry
Department of Mathematics and Computer Science, Royal Military College of Canada,
Kingston, ON K7K 7B4, Canada
e-mail: chaudhry-ml@rmc.ca

R. N. Mohapatra et al. (eds.), *Mathematics and Computing 2013*,
Springer Proceedings in Mathematics & Statistics 91,
DOI: 10.1007/978-81-322-1952-1_20, © Springer India 2014

1 Introduction

The classical risk model serves as a skeleton for more realistic risk models. Much of the literature on ruin theory is concentrated on the classical risk model, in which claims arrival process is a Poisson process. In 1957, Sparre Andersen introduced a mathematical model for the risk theory based on the assumption that the claim arrival process is an ordinary renewal process. The Sparre Andersen model in which the inter-claim times follow an Erlang or a generalized Erlang distribution has been studied extensively. The connection between risk theory and other applied probability areas has been initiated by Prabhu [8] in queueing theory context. Some of the main approaches to find the probability of ultimate ruin include Laplace transform inversion, matrix-analytic methods and differential and integral equations (see Asmussen and Albrecher [2] for a survey on risk theory). Frostig [6] studied the time to ruin, deficit at the time of ruin, and recovery times for $M/G/1$ and $GI/M/1$ risk processes. Frostig [6] established the relation between the time to ruin and busy-period of some queueing system and derived the bounds for expected value of time to ruin and deficit at the time of ruin. Thampi and Jacob [9] used duality results between the queueing theory and risk processes to derive explicit expressions for the ultimate ruin probability and moments of time to ruin in renewal risk model.

In this paper, we investigate $GI/M/1$ risk process using roots of certain characteristic equation in case of phase-type as well as nonphase-type inter-claim arrivals. We are interested in finding explicit expressions for the distributions of the ultimate ruin probability, time to ruin, deficit at the time of ruin, and the distribution of recovery time using Padé-Laplace method.

2 Mathematical Description of Risk Process

The ruin problem of insurance risk theory is closely related to the problem of single-server queue. Suppose the amount of capital at time t in one portfolio of an insurance company is denoted by $R(t)$. Initially $R(0) = u\ (> 0)$. During each unit of time, the portfolio receives an amount in premiums with a rate $p\ (> 0)$. At random times claims are made against the insurance company, which must pay the amount $X_n (> 0)$ to settle the nth claim. In risk theory, a risk reserve process $\{R(t),\ t \geq 0\}$ at time t is a model for the time evolution of the reserves of an insurance company and is given by the following expression:

$$R(t) = u + pt - \sum_{k=1}^{N(t)} X_k, \tag{1}$$

where u is the initial reserve, p is the rate at which the premiums are received, $N(t)$ is the total number of claims in $[0, t]$ and X_k is the size of the kth claim. For renewal risk model, $\{N(t),\ t \geq 0\}$ is an ordinary renewal process. The renewal

process $\{N(t),\ t \geq 0\}$ is independent of the claim sizes $\{X_k,\ k \geq 1\}$. We denote the inter-renewal times of claims by $T_i,\ i \geq 0$, where T_1 is the time of the first renewal, and T_i for $i = 2, 3, \ldots$, the time between the $(i-1)$th renewal and the ith renewal. Then, $\{T_i\}_{i=1}^{\infty}$ is an independent and identically distributed (i.i.d.) sequence of generally distributed random variables (r.v.'s), with distribution $G(t)$. Let the claim sizes X_k be i.i.d. r.v.'s distributed exponentially with rate δ. The model we consider will have the property $\lim_{t\to\infty} \sum_{i=1}^{N(t)} X_i/t \to \rho$, where ρ represents the average amount of claim per unit time. Another quantity of interest in ruin theory is the safety loading factor $\boldsymbol{\eta}$, defined as the relative amount by which the premium rate p exceeds ρ, i.e., $\boldsymbol{\eta} = (p-\rho)/\rho$. The insurance company should try to ensure that $\boldsymbol{\eta} > 0$. Another process in risk theory which is more convenient to work than the risk reserve process $\{R(t),\ t \geq 0\}$, is the claim surplus process $\{S(t),\ t \geq 0\}$, where

$$S(t) = u - R(t) = \sum_{k=1}^{N(t)} X_k - pt. \tag{2}$$

A crucial quantity in risk theory is the infinite-time ruin probability (probability of ultimate ruin) $\psi(u)$, the probability that the reserve ever drops below zero,

$$\psi(u) = P(\inf_{t\geq 0} R(t) < 0 | R(0) = u).$$

The probability of ruin before time $\bar{T}$ is $\psi(u, \bar{T}) = P\left(\inf_{0\leq t\leq \bar{T}} R(t) < 0 | R(0) = u\right)$. Let $\tau(u)$ be the time to ruin, i.e., the time for which the risk reserve of the insurance company becomes negative for the first time, then

$$\tau(u) = \inf\{t \geq 0 : R(t) < 0\} = \inf\{t \geq 0 : S(t) > u\}.$$

3 GI/M/1 Risk Process and Duality

Consider a risk process wherein the arrival of claims follow a renewal process. The inter-arrival times of claims $T_1, T_2, \ldots$ are i.i.d. r.v.'s, with distribution $B(t)$, density $b(t)$, moment generating function (mgf) $m_B(\theta) = E(\mathrm{e}^{\theta T})$ and Laplace-Stieltjes transform (LST) $\mathscr{L}_T(s) = \int_0^{\infty} \mathrm{e}^{-st}\, \mathrm{d}B(t) = E(\mathrm{e}^{-sT})$, where T is the generic of inter-arrival times. The claim sizes $X_1, X_2, \ldots$ are i.i.d. r.v.'s and are exponentially distributed with rate δ. The process of successive claim amounts is independent of the claim number process, i.e., the claim sizes X_i are independent of the inter-arrival times T_i. We assume through out that the premium income process has a constant rate $p = 1$ per unit time. We call this risk process a GI/M/1 risk process. The average amount of claim per unit time ρ in risk process is equal to $1/\delta E(T)$. Let $\mathscr{N}(u)$

denote the number of claim arrivals up to time of ruin. Then $\mathcal{N}(u)$ is

$$\mathcal{N}(u) = \inf\left\{n : u + \sum_{i=1}^{n} T_i - \sum_{i=1}^{n} X_i < 0\right\}.$$

and the time to ruin and the deficit at the time of ruin are, respectively, given by

$$\tau(u) = \sum_{i=1}^{\mathcal{N}(u)} T_i \quad \text{and} \quad \zeta(u) = \sum_{i=1}^{\mathcal{N}(u)} X_i - \sum_{i=1}^{\mathcal{N}(u)} T_i - u = \sum_{i=1}^{\mathcal{N}(u)} X_i - \tau(u) - u. \tag{3}$$

Let $\zeta(0) = \zeta$. The dual queueing system of $GI/M/1$ risk process is the $M/G/1$ queueing system, with an infinite waiting room and a single-server. Let the individual claims arrive at time epochs $0 = t_0, t_1, t_2, \ldots, t_n, \ldots$ following an exponential distribution with rate δ, i.e., with distribution $A(x) = 1 - \mathrm{e}^{-\delta x}$, density $a(x) = \delta \mathrm{e}^{\delta x}$ and LST $\mathscr{L}_X(s) = \delta/(\delta + s)$. The service times T_n, $(n = 1, 2, \ldots,)$ are i.i.d. r.v.'s with distribution $B(t)$, probability density function (pdf) $b(t), t \geq 0$, LST $\mathscr{L}_T(s)$ and mean service time $E(T) = \int_0^\infty t.b(t)\,\mathrm{d}t = -\frac{\mathrm{d}}{\mathrm{d}s}\mathscr{L}_T(s)|_{s=0}$. The customers are served by a single-server. The service discipline is first-come first-served (FCFS). The traffic intensity of the queueing system $\rho^* = 1/\rho = \delta E(T)$ is assumed to be less than unity, i.e., the condition of stability is $\rho^* < 1$. The service time of the customer that initiates the busy-period is $u + T_1$. Let $B(u)$ be the length of this busy-period and $\mathcal{N}_q(u)$, the number served during this busy-period. From the definition of $\mathcal{N}(u)$, it follows that $\mathcal{N}(u) = \mathcal{N}_q(u)$. So $B(u) = T_1 + u + \sum_{i=2}^{\mathcal{N}(u)} T_i = u + \sum_{i=1}^{\mathcal{N}(u)} T_i = \tau(u) + u$. Hence this busy-period of the dual queue ($B(u)$) is distributed as $\tau(u) + u$. Let $I(u)$ be the idle period that follows the busy-period $B(u)$. Then $I(u) = \sum_{i=1}^{\mathcal{N}_q(u)} X_i - B(u) = \sum_{i=1}^{\mathcal{N}_q(u)} X_i - \tau(u) - u = \zeta(u)$. Hence $\zeta(u)$ is distributed as $I(u)$. When $u = 0$, $B(0) = B$ and $I(0) = I$ are respectively, distributed as a regular busy-period and regular idle period. So $\tau(0)$ is distributed as B and $\zeta(0)$ is distributed as I.

4 Measures of Risk Process

A risk process is characterized by the probability of ultimate ruin, time to ruin, deficit at the time of ruin, recovery time after a ruin, and so on. We discuss some measures of these variables in subsequent analysis.

4.1 Time to Ruin and Deficit at the Time of Ruin

The ruin time, $(\tau(u))$, is the first time that the reserve becomes negative. Let $\zeta(u)$ be the deficit at the time of ruin when the initial reserve is u. We will study the distribution of time to ruin and deficit at the time of ruin for different values of $\rho\ (= 1/\delta E(T))$.

Case 1. $(\rho > 1)$ In this case, ruin occurs with probability 1 and all the moments of $\tau(u)$ exist. The busy-period B of the dual queue $M/G/1$ is finite, since the traffic intensity of the dual queue $\rho^*(= 1/\rho) < 1$. We will study the busy-period distribution of dual queue using the method of roots. The literature on queueing theory shows that distributions having Laplace-Stieltjes transform as a rational function cover a wide range of distributions that arise in applications, see Botta et al. [4]. In view of this, we consider those distributions that have rational Laplace-Stieltjes transform of the form $h(s) = P(s)/Q(s)$, where degree of the polynomial $Q(s)$ is n and that of the polynomial $P(s)$ is at most n. Here $\mathscr{L}_T(s) = B_1(s)/B_2(s)$ is a rational function satisfying the properties stated above. The LST of busy-period distribution, $B^*(s)$, of $M/G/1$ queue satisfies the functional equation

$$B^*(s) = \mathscr{L}_T(s + \delta - \delta B^*(s)) \tag{4}$$

We get the k-th moment $E(B^k)$ of busy-period by differentiating both sides of Eq. (4), k times w.r.t. s and then evaluate at $s = 0$. The first and second moments obtained are, respectively, $E(B) = \frac{E(T)}{1-\delta E(T)}$ and $E(B^2) = \frac{(1+\delta E(B))^2 E(T^2)}{1-\delta E(T)}$. After getting $E(B^k)$, $k = 1, 2, \ldots$, the LST of busy-period can be obtained as $B^*(s) = \sum_{k=0}^{\infty} (-1)^k E(B^k) \frac{s^k}{k!}$. We use Padé-Laplace method (see Akar and Arikan [1]) to approximate $B^*(s)$ by a rational function of order (m, n), defined as $R_{m,n}(s) = \frac{P_m(s)}{Q_n(s)} = \sum_{i=0}^{m} p_i s^i / \{1 + \sum_{i=1}^{n} q_i s^i\}$. Now replacing m by $n - 1$ and making partial fractions of $B^*(s)$, we obtain

$$B^*(s) = \frac{U(s)}{V(s)} = \sum_{i=1}^{k} \frac{A_i}{s - \gamma_i}, \tag{5}$$

where $B^*(s) = \frac{U(s)}{V(s)}$ is rational as stated earlier, with degree of $V(s)$, k and degree of $U(s)$ is at most k and $\gamma_i, i = 1, 2, \ldots, k$ are the roots (assumed distinct) of $V(s) = 0$ with $Re(s) < 0$. The moments of B can also be found from (5), $E(B^j) = (-1)^j B^{*j}(0) = (-1)^{j-1} \sum_{i=1}^{k} \frac{j! A_i}{\gamma_i^{j+1}}$, where $B^{*j}(0)$ is the jth differentiation of $B^*(s)$ w.r.t. s, evaluated at $s = 0$. The mean busy-period is $E(B) = -B^{*\prime}(0) = \sum_{i=1}^{k} \frac{A_i}{\gamma_i^2}$. The LST of the busy-period $B(u)$, initiated by the arrival of the first claim having service time $T_1 + u$ (service time of all other claims are i.i.d., distributed as T) is

$$\begin{aligned} B_u^*(s) &= \mathscr{L}_{T+u}(s+\delta(1-B^*(s))) = E\left(\mathrm{e}^{-(T+u)(s+\delta(1-B^*(s)))}\right) \\ &= E\left(\mathrm{e}^{-u(s+\delta(1-B^*(s)))}\right) E\left(\mathrm{e}^{-(s+\delta(1-B^*(s)))T}\right) \\ &= \mathrm{e}^{-u(s+\delta(1-B^*(s)))}\mathscr{L}_T(s+\delta-\delta B^*(s)) \\ &= \mathrm{e}^{-u(s+\delta(1-B^*(s)))}B^*(s) \end{aligned} \tag{6}$$

The moments of the busy-period can be found from the relation $E(B_u^j) = (-1)^j B_u^{*j}(0)$. For the sake of completeness, the first two moments are given as $E(B_u) = -B_u^{*\prime}(0) = E(B) + u(1+\delta E(B))$ and $E(B_u^2) = (1+u\delta)E(B^2) + 2uE(B)(1+\delta E(B)) + u^2(1+\delta E(B))^2$. The variance of $B_u^*(s)$ is $var(B_u) = E(B_u^2) - E(B_u)^2 = (1+u\delta)E(B^2) - E(B)^2$. Using the duality relation, $B(u) = \tau(u) + u$ and Eq. (6), we have the LST of the time to ruin

$$\begin{aligned} \tau_u^*(s) &= E(\mathrm{e}^{-s\tau(u)}) = E(\mathrm{e}^{-s(B_u-u)}) \\ &= \mathrm{e}^{us}\, {B_u}^*(s) = \mathrm{e}^{-u\delta(1-B^*(s))}B^*(s). \end{aligned} \tag{7}$$

The direct inversion of $\tau_u^*(s)$ is not possible, as it is not in rational form due to the presence of $\mathrm{e}^{-u\delta(1-B^*(s))}$. To bring $\tau_u^*(s)$ into rational form, so that we can obtain the distribution of $\tau(u)$ for a particular value of u, we use Padé-Laplace method as used above. Using this method, we approximate $\mathrm{e}^{-u\delta(1-B^*(s))}$ by a rational function $R_{m,n}(s)$, as defined earlier.

$$\tau_u^*(s) = \frac{P_{n-1}(s)}{Q_n(s)}B^*(s) = \frac{P_{n-1}(s)}{Q_n(s)}\frac{U(s)}{V(s)}. \tag{8}$$

As the degree of the denominator is strictly greater than the degree of the numerator of $\tau_u^*(s)$, making partial fractions, we get

$$\tau_u^*(s) = \sum_{i=1}^{n}\frac{C_i}{s-\beta_i} + \sum_{i=1}^{k}\frac{D_i}{s-\gamma_i} \tag{9}$$

where $\beta_i, i = 1, 2, \ldots, n$ and $\gamma_i, i = 1, 2, \ldots, k$ are, respectively, the distinct roots of $Q_n(s)$ and $V(s)$ with $Re(s) < 0$. If the roots are repeated, we can also obtain the coefficients using partial fraction method. The unknown coefficients can be found to be

$$C_i = \frac{P_{n-1}(\beta_i)}{Q_n'(\beta_i)}\frac{U(\beta_i)}{V'(\beta_i)},\ i = 1, 2, \ldots, n \quad \text{and} \quad D_i = \frac{P_{n-1}(\gamma_i)}{Q_n'(\gamma_i)}\frac{U(\gamma_i)}{V'(\gamma_i)},\ i = 1, 2, \ldots, k.$$

The density and distribution function of $\tau(u)$ are, respectively,

$$f_\tau(t) = \sum_{i=1}^{n} C_i e^{\beta_i t} + \sum_{i=1}^{k} D_i e^{\gamma_i t}, \quad \text{and} \quad F_\tau(t) = 1 + \sum_{i=1}^{n} \frac{C_i}{\beta_i} e^{\beta_i t} + \sum_{i=1}^{k} \frac{D_i}{\gamma_i} e^{\gamma_i t}. \tag{10}$$

The moments of the time to ruin for a particular value of u can be found from Eq. (10), as

$$E(\tau^j(u)) = \int_0^\infty t^j f_\tau(t) dt = (-1)^{j+1} \left\{ \sum_{i=1}^{n} C_i \frac{j!}{\beta_i^{j+1}} + \sum_{i=1}^{k} D_i \frac{j!}{\gamma_i{}^{j+1}} \right\}.$$

The mean and variance of the time to ruin are, respectively,

$$E(\tau(u)) = \sum_{i=1}^{n} \frac{C_i}{\beta_i^2} + \sum_{i=1}^{k} \frac{D_i}{\gamma_i^2} \tag{11}$$

and

$$var(\tau(u)) = -2 \left[\sum_{i=1}^{n} \frac{C_i}{\beta_i^3} + \sum_{i=1}^{k} \frac{D_i}{\gamma_i^3} \right] - \left[\sum_{i=1}^{n} \frac{C_i}{\beta_i^2} + \sum_{i=1}^{k} \frac{D_i}{\gamma_i^2} \right]^2. \tag{12}$$

The deficit, $\zeta(u)$, at the time of ruin is distributed as the idle period of the dual $M/G/1$ queue and using the lack of memory property of the exponential distribution, it is independent of the time to ruin $\tau(u)$. Because of the memoryless property, the idle periods I, (time from the end of a busy-period to the start of the next one) follow the same distribution, $I \sim Exp(\delta)$, thence $E(I) = 1/\delta$. So $\zeta(u)$ is exponentially distributed with rate δ giving $E(\zeta(u)) = 1/\delta$.

Remark 1 From Eq. (7), one can also get the moments of the time to ruin as $E(\tau^k(u)) = (-1)^k \tau_u^{*k}(0)$ with mean and second moment,

$$E(\tau(u)) = (1 + u\delta)E(B) \tag{13}$$

and $E(\tau^2(u)) = (1 + u\delta)E(B^2) + u\delta(2 + u\delta)(E(B))^2$, respectively. The variance of the time to ruin is $var(\tau(u)) = E(\tau_u^2) - E(\tau_u)^2 = (1 + u\delta)E(B^2) - E(B)^2 = var(B(u))$.

Case 2. ($\rho < 1$) In this case, the ruin probability is less than 1 and the expected time to ruin is infinite. For the dual queue $\rho^*(= 1/\rho) > 1$, the busy-period might be infinite as well. Thus, to obtain $E(\tau(u) : \tau(u) < \infty)$, we use the technique of change of measure via the exponential family (see Asumssen and Albrecher [2], p. 82). Consider another renewal risk process with claim renewal density, $b_\theta(t) = \frac{e^{-\theta t}}{\mathscr{L}_T(\theta)} b(t)$ and exponential claim size with rate $\delta_\theta = \frac{\delta}{\mathscr{L}_X(-\theta)} = \delta - \theta$. When $\theta = 0$, we get the original renewal risk process. To use the change of measure, we follow the definition stated below:

Definition 1 Let P be the probability measure induced by a renewal process where inter-arrival of claims distributed as T are served with rate δ . Define P_θ to be the probability measure governing the renewal risk process with claim arrival density b_θ and claim size distribution with rate δ_θ. The corresponding expectation operator is E_θ. The probability measure P_θ is the Esscher transform of P.

Define the Lundberg coefficient γ as the smallest positive root of the equation $\mathscr{L}_X(-\theta)\mathscr{L}_T(\theta) = 1$. As $\mathscr{L}_T(s) > 0$, and $\frac{d}{ds}\mathscr{L}_T(s)|_{s=0} = -E(T) < -1/\delta$, so $\mathscr{L}_T(s)$ is convex and monotone decreasing function in s. Thus, there exists a solution, γ, to the equation $\mathscr{L}_X(-\theta)\mathscr{L}_T(\theta) = 1$, such that $\gamma < \delta$. The changed risk process governed by P_γ is a risk process in which the claim inter-arrival and claim size density are, respectively,

$$b_\gamma(t) = \frac{\mathrm{e}^{-\gamma t}}{\mathscr{L}_T(\gamma)} b(t) \quad \text{and} \quad a_\gamma(x) = \frac{\mathrm{e}^{\gamma x}}{\mathscr{L}_X(-\gamma)} a(x) = (\delta - \gamma)\mathrm{e}^{-(\delta-\gamma)x}.$$

From Asmussen and Albrecher [2, p. 86] the changed risk process has $E(X_\gamma) > E(T_\gamma)$, which then gives $\rho_\gamma = 1/(\delta_\gamma E(T_\gamma)) > 1$. This is the condition we have in case 1. The traffic intensity of the dual queue $\rho^*_\gamma (= 1/\rho_\gamma) < 1$. Consider the dual queueing process of changed renewal risk process, where inter-arrival times are exponentially distributed as X_γ with rate $\delta_\gamma = \delta - \gamma$ and service times are i.i.d. r.v.'s distributed as T_γ with density $b_\gamma(t)$ and LST $\mathscr{L}_{T_\gamma}(s) = E(\mathrm{e}^{-sT_\gamma}) = \mathscr{L}_T(s+\gamma)/\mathscr{L}_T(\gamma)$. Following the analysis similar to that in Case 1, the busy-period $B_{p\gamma}$ of the dual queue can be derived. Using the analysis similar to that followed in Case 1, we obtain the busy-period distribution. For the sake of completeness, the LST of $B_{p\gamma}$ is

$$B^*_{p\gamma}(s) = \sum_{i=1}^{n} \frac{\bar{A}_i}{s - \alpha_i}, \tag{14}$$

where $B^*_{p\gamma}(s) = \tilde{U}(s)/\tilde{V}(s)$ is rational with degree of $\tilde{V}(s)$ equal to l and $\alpha_i, i = 1, 2, \ldots, l$ are the roots (assumed distinct) of $\tilde{V}(s) = 0$ with $Re(s) < 0$. The unknown constants are found to be, $\bar{A}_i = \frac{\tilde{U}(\alpha_i)}{\tilde{V}'(\alpha_i)}$, $i = 1, 2, \ldots, n$. From Eq. (14), the distribution and density functions of busy-period can be found easily in terms of the roots α_i. Let $B_\gamma(u)$, the busy-period initiated by the arrival of the first claim, having sevice time $u + T_\gamma$ and using the results obtained in Case 1, the LST of $B_\gamma(u)$ is

$$B^*_\gamma(s) = E_\gamma\left(\mathrm{e}^{-s(\tau(u)+u)}\right) = \mathrm{e}^{-u(s+(\delta-\gamma)\left(1-B^*_{p\gamma}(s)\right))} B^*_{p\gamma}(s). \tag{15}$$

Let $\zeta_\gamma(u)$ and $\tau_\gamma(u)$ be the deficit at the time of ruin and the time to ruin in the changed risk process with respect to the measure P_γ. Using the analysis similar to Case 1 and the lack of memory property of the exponential distribution implies that $\zeta_\gamma(u)$ is exponentially distributed with parameter $\delta - \gamma$ and is independent of $\tau_\gamma(u)$.

Also, $\zeta_\gamma^*(s) = \frac{\delta-\gamma}{s+\delta-\gamma}$. The expected value of $\zeta_\gamma(u)$ is $E(\zeta_\gamma) = 1/(\delta-\gamma)$. Under the probability measure P_γ, $\tau_\gamma(u)+u$ is distributed as the busy-period $B_\gamma(u)$ in the $M/G/1$ queue, where arrivals are according to a Poisson process with rate $\delta-\gamma$ and the service time of the first customer in the busy-period is distributed as $T_\gamma + u$, while the service times of all the other customers are distributed as T_γ. The LST of the time to ruin $\tau_\gamma(u)$, of the changed renewal risk process is given by,

$$\tau_\gamma^*(s) = E_\gamma(\mathrm{e}^{-s\tau(u)}) = \mathrm{e}^{-u(\delta-\gamma)(1-B_{p\gamma}^*(s))} B_{p\gamma}^*(s). \tag{16}$$

From Frostig [6], the LST of the time to ruin of the original risk process is

$$\begin{aligned}\tau^*(s) = \mathrm{e}^{-\gamma u}\zeta_\gamma^*(\gamma)\tau_\gamma^*(s) &= \frac{\delta-\gamma}{\delta}\mathrm{e}^{-\gamma u}\mathrm{e}^{-u(\delta-\gamma)(1-B_{p\gamma}^*(s))} B_{p\gamma}^*(s) \\ &= \frac{\delta-\gamma}{\delta}\mathrm{e}^{-u\delta(1-B_{p\gamma}^*(s))} B_{p\gamma}^*(s).\end{aligned} \tag{17}$$

The distribution and moments of $\tau(u)$ can be found for different values of u by using Padé-Laplace method similar to Case 1. For the sake of completeness, we have given

$$\tau^*(s) = \sum_{i=1}^{n} \frac{\bar{C}_i}{s-\theta_i} + \sum_{i=1}^{l} \frac{\bar{D}_i}{s-\alpha_i} \tag{18}$$

where $\theta_i, i = 1, 2, \ldots, n$ are the roots (assumed distinct) of $\tilde{Q}_n(s)$ with $Re(s) < 0$. The unknown coefficients are, given by, $\bar{C}_i = \frac{\tilde{P}_{n-1}(\theta_i)}{\tilde{Q}_n'(\theta_i)} \frac{\tilde{U}(\theta_i)}{\tilde{V}'(\theta_i)}$, $i = 1, 2, \ldots, n$, and $\bar{D}_i = \frac{\tilde{P}_{n-1}(\alpha_i)}{\tilde{Q}_n'(\alpha_i)} \frac{\tilde{U}(\alpha_i)}{\tilde{V}'(\alpha_i)}$, $i = 1, 2, \ldots, l$. The density and distribution function of $\tau(u)$ are, respectively,

$$f_\tau(t) = \sum_{i=1}^{n} \bar{C}_i \mathrm{e}^{\theta_i t} + \sum_{i=1}^{l} \bar{D}_i \mathrm{e}^{\alpha_i t} \quad \text{and} \quad F_\tau(t) = 1 + \sum_{i=1}^{n} \frac{\bar{C}_i}{\theta_i}\mathrm{e}^{\theta_i t} + \sum_{i=1}^{l} \frac{\bar{D}_i}{\alpha_i}\mathrm{e}^{\alpha_i t}. \tag{19}$$

The moments of the time to ruin for a fixed u, are given by

$$E(\tau^k(u)) = (-1)^{k+1}\left\{\sum_{i=1}^{n} \bar{C}_i \frac{k!}{\theta_i^{k+1}} + \sum_{i=1}^{l} \bar{D}_i \frac{k!}{{\alpha_i}^{k+1}}\right\}.$$

The expected time to ruin is

$$E(\tau(u)) = \sum_{i=1}^{n} \frac{\bar{C}_i}{{\theta_i}^2} + \sum_{i=1}^{l} \frac{\bar{D}_i}{{\alpha_i}^2}. \tag{20}$$

Also, the moments of the time to ruin can be obtained by differentiating Eq. (17) w.r.t. s and then substituting $s = 0$ as $E(\tau^k(u)) = (-1)^k \tau_u^{*k}(0)$. The first two moments are, respectively,

$$E(\tau(u)) = \frac{\delta - \gamma}{\delta}[1 + u(\delta - \gamma)]\mathrm{e}^{-\gamma u} E(B_{p\gamma}), \tag{21}$$

and $E(\tau^2(u)) = \frac{\delta-\gamma}{\delta}\mathrm{e}^{-\gamma u}\left\{[1 + u(\delta - \gamma)]E(B_{p\gamma}^2) + u(\delta - \gamma)[2 + u(\delta - \gamma)]E(B_{p\gamma})^2\right\}$.

4.2 Probability of Ultimate Ruin

The ultimate ruin probability of a risk process is $\psi(u) = P\{\tau(u) < \infty\}$. It is an important measure in the study of risk process. Different methods are used for calculation of ruin probability. Asmussen and Rolski [3] consider the probability of ultimate ruin for the Sparre Andersen model when individual claim amounts are distributed as phase type and they present an iterative method of evaluating this probability. Dickson and Hipp [5] consider a risk process in which inter-arrival of claims have an Erlang(2) distribution. They obtain explicit solution for the Laplace transform of the ruin probability by solving a second-order integro-differential equation. We follow Frostig [6] to find out the probability of ultimate ruin using the laplace transform method.

Case 1. ($\rho > 1$) In this case, the probabilty of ultimate ruin of the $GI/M/1$ risk process can be obtained from (7), by substituting $s = 0$. Then, $\psi(u) = 1$. The following proposition can be found in Asmussen and Albrecher [2, p. 3].

Proposition 1 *If $\eta < 0$ then $M = \infty$ and hence $\psi(u) = 1$, $\forall\ u$. If $\eta > 0$ then $M < \infty$ and hence $\psi(u) < 1$ for all sufficiently large u.*

where $M = \sup_{0 \le t < \infty} S(t)$ and $\eta < 0 \equiv \rho > 1$.

Case 2. ($\rho < 1$) In this case, the probabilty of ultimate ruin of the $GI/M/1$ risk process can be obtained from (17), by substituting $s = 0$. Then,

$$\psi(u) = \frac{\delta - \gamma}{\delta}\mathrm{e}^{-\gamma u}.$$

The same result is also found in Asmussen and Albrecher [2, p. 156].

4.3 Recovery Time

When ruin occurs, the surplus process will temporarily stay below the zero level and after some time the process will again come to the zero level; if not, then the company is declared insolvent or liquidated. The recovery time is the time interval

during which the surplus is negative. The duration of this negative surplus will depend on the severity of ruin and probability of ruin. The recovery time is distributed like a busy-period in a $GI/M/1$ queue.

Case 1. ($\rho > 1$) Since $\rho > 1$ for the above renewal risk process, the associated $GI/M/1$ queue is unstable and the busy-period of this $GI/M/1$ queue can be infifnite. As the exact expression for the recovery time is not tractable through queueing parameters, we provide the bounds for their moments. To derive the bounds for the moments, we use the technique of change of measure via the exponential family as explained in Sect. 4.1. Define $\kappa(\theta) = \delta[\mathscr{L}_T(-\theta) - 1] - \theta$. The Lundberg coefficient $\gamma > 0$ is the smallest positive root of the Lundberg equation, $\kappa(\theta) = 0$. The LSTs of claim arrivals and claim sizes of the changed risk process are, respectively, $\mathscr{L}_{T_\gamma}(s) = \frac{\mathscr{L}_T(s-\gamma)}{\mathscr{L}_T(-\gamma)}$ and $\mathscr{L}_{X_\gamma}(s) = \frac{\delta\mathscr{L}_T(-\gamma)}{\delta\mathscr{L}_T(-\gamma)+s}$. From $\kappa(\gamma) = 0$, we get $\mathscr{L}_T(-\gamma) = 1 + \gamma/\delta$. The associated queueing model for this changed risk process is the $GI/M/1$ queue, where the service times are exponentially distributed with rate $\delta + \gamma$ and inter-arrival times are distributed as T_γ. As the traffic intensity ρ_γ becomes less than 1, the stability condition holds. From Komota et al. [7], the LST of busy-period of $GI/M/1$ queue is given by,

$$B^*_{ch}(s) = 1 - \frac{a_2(s) - a_1(s)}{s - \omega} = 1 - \frac{s}{s - \omega}, \tag{22}$$

where $\mathscr{L}_{X_\gamma}(s) = a_1(s)/a_2(s) = \frac{\delta+\gamma}{\delta+\gamma+s}$ and ω is the unique root of the characteristic equation $1 = \mathscr{L}_{X_\gamma}(s)\mathscr{L}_{T_\gamma}(-s)$ with $Re(s) < 0$. The density and distribution functions of busy-period can be found to be $f(t) = -\omega e^{\omega t}$ and $F(t) = 1 - e^{\omega t}$, respectively. The moments of this busy-period are given by, $E(B^k_{ch}) = (-1)^k B^{*\,k}_{ch}(0) = \frac{(-1)^k k!}{\omega^k}$, with expected value $E(B_{ch}) = -1/\omega$. Since the service rate $(\delta + \gamma)$ of the changed queue, obtained after applying the exponential change of measure technique, is greater than the service rate (δ) of the original queue, the values of the moments of busy-period of the changed queue must be greater than or equal to the corresponding moment values of the original queue. Therefore, the moments of recovery time V_{rec} is bounded by the moments $E(B^k_{ch})$.

Case 2. ($\rho < 1$) When ruin has occured, the recovery time is distributed like a busy-period in a $GI/M/1$ queue. Since $\rho < 1$ for the above renewal risk process, the associated $GI/M/1$ queue is stable and the busy-period of this $GI/M/1$ queue is finite. The LST of busy-period of the stable $GI/M/1$ queue is equivalent to the LST of the recovery time V_{rec} of the risk process and is given by,

$$V^*_{rec}(s) = 1 - \frac{s}{s - \alpha} = \frac{-\alpha}{s - \alpha}, \tag{23}$$

where α is the unique root of the characteristic equation $1 = \mathscr{L}_X(s)\mathscr{L}_T(-s)$ with $Re(s) < 0$ and $\mathscr{L}_X(s) = \delta/(\delta + s)$. The distribution and density function of V_{rec} are, respectively,

$$f_{rec}(t) = -\alpha e^{\alpha t} \quad \text{and} \quad F_{rec}(t) = 1 - e^{\alpha t}. \tag{24}$$

The moments of recovery times are given by

$$E(V_{rec}^k) = (-1)^k V_{rec}^{*\,k}(0) = (-1)^k k!/\alpha^k.$$

The first and second moments of the recovery time V_{rec} are, respectively, given by $E(V_{rec}) = -V_{rec}^{*\prime}(0) = 1/-\alpha$ and $E(V_{rec}^2) = V_{rec}^{*\prime\prime}(0) = \frac{2}{\alpha^2}$. The variance of V_{rec} is $var(V_{rec}) = \frac{1}{\alpha^2}$.

5 Numerical Results and Discussion

Phase-type claims arrival: We consider $PH/M/1$ risk process with inter-claim arrivals following a phase-type distribution with representation (α, S) and claim sizes are exponentially distributed with rate $\delta = 1.5$, where $\alpha = (0.1, 0.6, 0, 0.3)$, $S = \begin{pmatrix} -3 & 1 & 0 & 1 \\ 1 & -5 & 1 & 0 \\ 0 & 2 & -4 & 2 \\ 1 & 0 & 1 & -4 \end{pmatrix}$, $E(T) = 0.565$ and $E(X) = 0.66$. So $\rho = \frac{1}{\delta E(T)} = 1.18 > 1$. The dual queueing model of the above risk process is the queue $M/PH/1$, where the inter-arrival times are exponentially distributed with rate δ and service times follow a phase-type distribution with representation (α, S). For this queueing model the offered load $\rho^* = 1/\rho = 0.848 < 1$.

In our numerical computations using Maple 12 and Eq. (5), we find the mean and variance of busy-period $E(B_p) = 3.72$ and $var(B_p) = 192.45$ respectively. Then the busy-period initiated by the service of first arrival in $M/PH/1$ queue with mean $E(\tilde{B}_u) = 6.58u + 3.72$ and variance $var(\tilde{B}_u) = 309.42u + 192.45$, are derived from (6). These values match exactly with Frostig [6]. The mean and variance of $\tau(u)$ are $E(\tau(u)) = E(\tilde{B}_u) - u = 5.58u + 3.72$ and $var(\tau(u)) = 309.42u + 192.45$, respectively (see Table 5).

Consider the above $PH/M/1$ risk process with claim size rate $\delta = 6$. For this risk process $\rho = 0.295 < 1$. So we can't get a stable dual queue. The parameters of the changed risk model are $\delta_\gamma = 2.07$, $E_\gamma(T) = 0.15$ and $\rho_\gamma = 3.16 > 1$. The dual queue of the change risk model has $\rho_\gamma^* = 1/\rho_\gamma = 0.316 < 1$. The expected values of $\tau(u)$ for different values of u are presented in Table 1. The probability of ultimate ruin is $\psi(u) = 0.346e^{-3.93u}$ with $\psi(0) = 0.346$. The expected value of recovery time is $E(V) = 0.25$. The expected values of the recovery times for different values of $\rho(< 1)$ are presented in Table 3. Also for $\rho(> 1)$, the upper bounds for the expected values of the recovery time are presented in Table 2.

Non-phase type claims arrival: We consider $ME/M/1$ risk process, where claim arrivals follow a matrix exponential distribution with density $f(t) = (1 + \frac{1}{4\pi^2})(1 - cos(2\pi t))e^{-t}$ and LST $f^*(s) = \frac{1+4\pi^2}{(s+1)[(s+1)^2+4\pi^2]}$ and claim sizes are exponentially

Table 1 Expected time to ruin for $PH/M/1$ risk process with parameters: $E(T) = 0.565$, $\delta = 6$ and $\rho = 0.295 < 1$

u	$E(\tau(u))$ from (20)	Frostig's $E(\tau(u))$	$E(\tau(u))$ from (21)
0.0	0.076994	0.076994	0.076994
0.1	0.062768	0.062768	0.062768
0.2	0.049661	0.049661	0.049661
0.3	0.038448	0.038448	0.038448
0.4	0.029280	0.029280	0.029280
0.5	0.022011	0.022011	0.022011
0.8	0.008846	0.008846	0.008846
1.0	0.004662	0.004662	0.004662
1.5	0.000875	0.000875	0.000875
2.0	0.000154	0.000154	0.000154
2.5	0.000026	0.000026	0.000026
3.0	0.000004	0.000004	0.000004
3.5	0.000000	0.000000	0.000000
4.0	0.000000	0.000000	0.000000
⋮	⋮	⋮	⋮

Table 2 Upper bounds for expected recovery time of $PH/M/1$ risk process for different ρ value

ρ	Frostig's $E(V)$	Upper bounds $E(B_{ch})$
1.18	3.574474	4.237768
1.22	2.913513	3.576978
1.31	2.064983	2.728797
1.47	1.351542	2.015896
1.77	0.833313	1.498423
2.21	0.529036	1.194944
2.95	0.328886	0.995637
4.42	0.187223	0.854865
5.90	0.130857	0.798965
8.84	0.081678	0.750263
11.79	0.059366	0.728196
17.69	0.038392	0.707469
35.38	0.018637	0.687966
⋮	⋮	⋮

distributed with rate $\delta = 0.5$. For this risk process, $E(T) = 1.05$, $E(X) = 2.0$ and $\rho = 1.9 > 1$. The offered load of the dual queue is $\rho^* = 1/\rho = 0.524 < 1$. So steady state solutions exist.

In our numerical computations using Maple 12 and Eq. (5), we find the mean and variance of busy-period $E(B_p) = 2.208$ and $var(B_p) = 14.258$ respectively. Then the busy-period initiated by the service of first arrival in $M/PH/1$ queue with mean

Table 3 Expected value of recovery time of $PH/M/1$ risk process for different values of ρ

ρ	Frostig's $E(V)$	Our $E(V_{rec})$
0.295	0.254656	0.254656
0.253	0.204603	0.204603
0.221	0.170805	0.170805
0.197	0.146482	0.146482
0.177	0.128158	0.128158
0.161	0.113867	0.113867
0.147	0.102416	0.102416
0.136	0.093039	0.093039
0.126	0.085222	0.085222
0.118	0.078608	0.078608
0.111	0.072939	0.072939
0.104	0.068028	0.068028
0.098	0.063733	0.063733
⋮	⋮	⋮

Table 4 Expected time to ruin for $ME/M/1$ risk process with parameters: $E(T) = 1.05$, $\delta = 0.5$ and $\rho = 1.9 > 1$

u	$E(\tau(u))$ from (11)	Frostig's $E(\tau(u))$	$E(\tau(u))$ from (13)
0	2.207909	2.207909	2.207909
1	3.311863	3.311863	3.311863
2	4.415817	4.415817	4.415817
3	5.519772	5.519772	5.519772
4	6.623726	6.623726	6.623726
5	7.727681	7.727681	7.727681
10	13.247452	13.247452	13.247452
15	18.767244	18.767244	18.767244
20	24.286996	24.286996	24.286996
30	35.326540	35.326540	35.326540
50	57.405628	57.405628	57.405628
100	112.603347	112.603347	112.603347
⋮	⋮	⋮	⋮

$E(\tilde{B_u}) = 2.104u + 2.208$ and variance $var(\tilde{B_u}) = 9.566u + 14.258$, are derived from (6). Both the mean and variance match exactly with Frostig [6]. The LST of the time to ruin is then obtained from Eq. (7). The first two moments of $\tau(u)$ are $E(\tau(u)) = E(\tilde{B_u}) - u = 1.104u + 2.208$ and $var(\tau(u)) = 9.566u + 14.258$, respectively (see Tables 4 and 6).

Consider the above $ME/M/1$ risk process with claim size rate $\delta = 5$. For this risk process $\rho = 0.19 < 1$. Applying the change of measure technique, the parameters of the changed risk model are $\delta_\gamma = 0.53$, $E_\gamma(T) = 0.34$ and $\rho_\gamma = 5.5 > 1$. The

Table 5 Variance of the time to ruin for $PH/M/1$ risk process with parameters: $E(T) = 0.565$, $\delta = 1.5$ and $\rho = 1.18 > 1$

u	Frostig's $var(\tau(u))$	$var(\tau(u))$ from (12)
0	192.449928	192.449928
1	501.874589	501.874589
2	811.299249	811.299249
3	1120.723911	1120.723911
4	1430.148572	1430.148572
5	1739.573233	1739.573233
6	2048.997894	2048.997894
7	2358.422555	2358.422555
8	2667.847216	2667.847216
9	2977.271877	2977.271877
10	3286.696538	3286.696538
⋮	⋮	⋮

Table 6 Variance of the time to ruin for $ME/M/1$ risk process with parameters: $E(T) = 1.05$, $\delta = 0.5$ and $\rho = 1.9 > 1$

u	Frostig's $var(\tau(u))$	$var(\tau(u))$ from (12)
0	14.257609	14.257609
1	23.823844	23.823844
2	33.390078	33.390078
3	42.956313	42.956313
4	52.522548	52.522548
5	62.088783	62.088783
6	71.655018	71.655018
7	81.221253	81.221253
8	90.787488	90.787488
9	100.353723	100.353723
10	109.919958	109.919958
⋮	⋮	⋮

dual queue of the change risk model has $\rho_\gamma^* = 1/\rho_\gamma = 0.18 < 1$. In this case we have derived the same risk measures and are match exactly with Frostig [6]. The probability of ultimate ruin is $\psi(u) = 0.346e^{-3.93u}$ with $\psi(0) = 0.346$. The expected value of recovery time is $E(V) = 0.22$.

6 Conclusion and Future Scope

In this paper, we have carried out the analysis of $GI/M/1$ risk process for different cases of the average amount of claim per unit time. We presented the distributions of time to ruin and recovery time for both cases. Also we derived the distributions for

the deficit at the time of ruin. We obtained different bounds for the expected values of the recovery time after ruin has happened. This model can be extended to include batch of claims arriving to the risk process at a particular time, using a dividend barrier and force of interest and are left for future investigations.

References

1. Akar, N., Arikan, E.: A numerically efficient method for the $MAP/D/1/K$ queue via rational approximations. Queueing Syst. **22**, 97–120 (1996)
2. Asmussen, S., Albrecher, H.: Ruin probabilities, vol. 14. World Scientific, New Jersey (2010)
3. Asmussen, S., Rolski, T.: Computational methods in risk theory: a matrix-algorithmic approach. Insur. Math. Econ. **10**(4), 259–274 (1992)
4. Botta, R.F., Harris, C.M., Marchal, W.G.: Characterization of generalized hyperexponential distribution functions. Stoch. Models **3**(1), 115–148 (1987)
5. Dickson, D., Hipp, C.: Ruin probabilities for erlang (2) risk processes. Insur. Math. Econ. **22**(3), 251–262 (1998)
6. Frostig, E.: Upper bounds on the expected time to ruin and on the expected recovery time. Adv. Appl. Probab. **36**, 377–397 (2004)
7. Komota, Y., Nogami, S., Hoshiko, Y.: Analysis of the $GI/G/1$ queue by the supplementary variable approach. Electron. Commun. Jpn. **66**(A-5), 10–19 (1983)
8. Prabhu, N.U.: On the ruin problem of collective risk theory. The Ann. Math. Stat. **32**(3), 757–764 (1961)
9. Thampi, K., Jacob, M.: On a class of renewal queueing and risk processes. J. Risk Finance **11**(2), 204–220 (2010)

Chapter 21
On Quasi-ideals in Ternary Semirings

Manish Kant Dubey and Anuradha

Abstract In this paper, we study the concept of minimal quasi-ideals in ternary semiring and prove some standard results analogous to ring theory. We also introduced the concept of a Q-simple ternary semiring and 0-Q-simple ternary semiring and characterize 0-minimal quasi-ideals in terms of Q-simple ternary semiring.

1 Introduction and Preliminaries

Lehmer [6] initiated the concept of ternary algebraic systems called triplexes in 1932. After that several authors have generalized the concept in many ways. In 2003, Dutta and Kar [1] have introduced the notion of ternary semiring which is generalization of ternary rings introduced by Lister [7]. Kar [4] have generalized the notion of quasi-ideal in ternary semirings and gave some properties of quasi-ideals and bi-ideals in ternary semirings. Steinfeld [8] have studied widely the notion of quasi-ideals in rings and semigroups. In this paper, we generalize the results of quasi-ideals in ternary semirings. Recall ([1, 4]) the following:

Definition 1.1 A nonempty set S together with a binary operation, called addition and ternary multiplication, denoted by juxtaposition, is said to be a ternary semiring if S is an additive commutative semigroup satisfying the following conditions:

(i) $(abc)de = a(bcd)e = ab(cde)$,
(ii) $(a + b)cd = acd + bcd$,

M. K. Dubey (✉)
Scientific Analysis Group, Defence Research and Development Organisation,
Delhi 110054, India
e-mail: kantmanish@yahoo.com

Anuradha
University of Delhi, Delhi 110007, India
e-mail: anuvikaspuri@yahoo.co.in

R. N. Mohapatra et al. (eds.), *Mathematics and Computing 2013*,
Springer Proceedings in Mathematics & Statistics 91,
DOI: 10.1007/978-81-322-1952-1_21, © Springer India 2014

(iii) $a(b+c)d = abd + acd$,
(iv) $ab(c+d) = abc + abd$, for all $a, b, c, d, e \in S$.

Definition 1.2 Let S be a ternary semiring. If there exists an element $0 \in S$ such that $0 + x = x$ and $0xy = x0y = xy0 = 0$ for all $x, y \in S$ then $'0'$ is called the zero element or simply the zero of the ternary semiring S. In this case we say that S is a ternary semiring with zero.

Throughout this paper, S will always denote a ternary semiring with zero, unless stated otherwise a ternary semiring means a ternary semiring with zero. Let A, B, C be three subsets of S. Then by ABC, we mean the set of all finite sums of the form $\sum a_i b_i c_i$, with $a_i \in A, b_i \in B, c_i \in C$.

Definition 1.3 An additive subsemigroup T of S is called a ternary subsemiring if $t_1 t_2 t_3 \in T$ for all $t_1, t_2, t_3 \in T$.

Definition 1.4 An additive subsemigroup I of S is called a left (right, lateral) ideal of S if $s_1 s_2 i$ (respectively $i s_1 s_2$, $s_1 i s_2$) $\in I$ for all $s_1, s_2 \in S$ and $i \in I$. If I is a left, a right, a lateral ideal of S then I is called an ideal of S.

Definition 1.5 An element a in a ternary semiring S is called regular if there exists an element x in S such that $axa = a$. A ternary semiring is called regular if all of its elements are regular.

Definition 1.6 A ternary semiring S with $|S| \geq 2$ is called a ternary division semiring if for any nonzero element a of S, there exists a nonzero element b in S such that $abx = bax = xab = xba = x$ for all $x \in S$.

Definition 1.7 An additive subsemigroup Q of a ternary semiring S is called a quasi-ideal of S if $QSS \cap (SQS + SSQSS) \cap SSQ \subseteq Q$. A ternary subsemiring B of a ternary semiring S is called a bi-ideal of S if $BSBSB \subseteq B$.

Definition 1.8 A proper ideal P of a ternary semiring S is called a semiprime if $I^3 \subseteq P$ implies $I \subseteq P$.

2 Minimal Quasi-ideals in Ternary Semiring

Steinfeld [8] had given many characterizations of minimal quasi-ideals in rings and semigroups. In this section, we proceed with the study of minimal quasi-ideals of ternary semiring which are analogous to ring theory.

Definition 2.1 A nonzero quasi-ideal Q of a ternary semiring S is called minimal if Q does not properly contain any nonzero quasi-ideal.

Theorem 2.1 *The intersection of a minimal right ideal R, a minimal lateral ideal M, and a minimal left ideal L of a ternary semiring S is either* 0 *or a minimal quasi-ideal of S.*

Proof Proof is similar to Theorem 3.9 [4].

Converse of above theorem is true if ternary semiring is semiprime. A ternary semiring S is called semiprime if (0) is a semiprime ideal of S.

Theorem 2.2 *Let S be a semiprime ternary semiring. Then each minimal quasi-ideal Q of S is the intersection of a minimal right ideal R, a minimal lateral ideal M, and a minimal left ideal L of S.*

Proof Since Q is a quasi-ideal of S, therefore $QSS \cap (SQS + SSQSS) \cap SSQ \subseteq Q$. Also Q is minimal, therefore either $QSS \cap (SQS + SSQSS) \cap SSQ = 0$ or $QSS \cap (SQS + SSQSS) \cap SSQ = Q$. Suppose that $QSS \cap (SQS + SSQSS) \cap SSQ = 0$. Then either $QSS = 0$ or $QSS \neq 0$. If $QSS = 0$ then Q would be a nonzero right ideal of S satisfying $Q^3 = 0$. This contradicts our assumption. If $QSS \neq 0$, then $QSSQSSQ \subseteq QSS \cap (SQS + SSQSS) \cap SSQ = 0$. This implies $(QSS)^3 = 0$ which contradicts our assumption that (0) is a semiprime ideal of S. Therefore, $QSS \cap (SQS + SSQSS) \cap SSQ = Q$. Now, we show that QSS is a minimal right ideal of S. Suppose that there exist a nonzero right ideal R' of S such that $R' \subseteq QSS$. Then $R'SS \cap (SQS + SSQSS) \cap SSQ$ is a quasi-ideal of S such that $R'SS \cap (SQS + SSQSS) \cap SSQ \subseteq Q$. Since Q is minimal, therefore either $R'SS \cap (SQS + SSQSS) \cap SSQ = 0$ or $R'SS \cap (SQS + SSQSS) \cap SSQ = Q$. Suppose $R'SS \cap (SQS + SSQSS) \cap SSQ = 0$. Then $R'QSSQ \subseteq R'SS \cap (SQS + SSQSS) \cap SSQ = 0$. Now $R' \subseteq QSS$ implies $R'^3 \subseteq (R'QSSQ)SS = 0$. This contradicts the condition that (0) is a semiprime ideal of S. Therefore, $R'SS \cap (SQS + SSQSS) \cap SSQ = Q$. This implies $Q \subseteq R'SS \subseteq R'$. Thus, $QSS \subseteq R'SS \subseteq R'$. Hence $R' = QSS$ is a minimal right ideal of S. Similarly, we can prove that $SQS + SSQSS$ is a minimal lateral ideal of S and SSQ is a minimal left ideal of S.

Theorem 2.3 *Let S be a ternary semiring. If S is a ternary division semiring, then S has no nonzero proper quasi-ideals of S.*

Proof Let S be a ternary division semiring and Q be a nonzero quasi-ideal of S. Let $0 \neq q \in Q$. Then there exists $0 \neq s \in S$ such that $qsx = sqx = xqs = xsq = x$ for all $x \in S$. This implies $S = QSS = SQS = SSQ$. Also $S = SQS = (SSQ)Q(QSS) \subseteq SSQSS$. Now, $S \subseteq QSS \cap (SQS + SSQSS) \cap SSQ \subseteq Q$. Consequently, $Q = S$. Hence S has no nonzero proper quasi-ideals.

Theorem 2.4 *Let S be a ternary semiring. If a quasi-ideal Q of S is a ternary division subsemiring of S, then Q is a minimal quasi-ideal of S.*

Proof Proof is trivial.

Remark 1 The following example shows that the above result does not hold for a quasi-ideal which is a zero ternary semiring that is a ternary semiring in which $abc = 0$ for all $a, b, c \in S$. Let $S = M_3(Z_0^-)$ be the ternary semiring of the set of

all 3×3 lower triangular square matrices over Z_0^-. Let $Q = \left\{ \begin{pmatrix} 0 & 0 & 0 \\ a & 0 & 0 \\ 0 & a & 0 \end{pmatrix} : a \in Z_0^- \right\}$ and $S = \left\{ \begin{pmatrix} 0 & 0 & 0 \\ x & 0 & 0 \\ y & z & 0 \end{pmatrix} : x, y, z \in Z_0^- \right\}$. It is easy to show that Q is a quasi-ideal of S such that $Q^3 = 0$. Clearly, Q is not a minimal quasi-ideal of S. Let $Q' = \left\{ \begin{pmatrix} 0 & 0 & 0 \\ a & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} : a \in Z_0^- \right\} \subseteq Q$. Then Q' is a quasi-ideal of S such that $Q' \subseteq Q$. Hence Q is not a minimal quasi-ideal of S.

3 Q-Simple and 0-Q-Simple Ternary Semirings

In this section, we study the concept of a Q-simple ternary semiring and 0-Q-simple ternary semiring.

Definition 3.1 A ternary semiring S without zero is called Q-simple if it has no proper quasi-ideals.

Definition 3.2 A ternary semiring S with zero is called 0-Q-simple if it has no nonzero proper quasi-ideals and $S^3 \neq \{0\}$.

Proposition 3.1 *Let S be a ternary semiring and A be any nonempty subset of S. Then the principal quasi-ideal generated by a is given by*
$\langle a \rangle_q = \left\{ aSS \cap (SaS + SSaSS) \cap SSa + na : n \in Z_0^+ \right\}$.

Proposition 3.2 *Let S be a ternary semiring and $a \in S$. Then the principal bi-ideal generated by a is given by $\langle a \rangle_b = \left\{ aSaSa + na + ma^3 : n, m \in Z_0^+ \right\}$.*

Proposition 3.3 *Let S be a ternary semiring. Then the set $aSS \cap (SaS + SSaSS) \cap SSa$ is a quasi-ideal of S for all $a \in S$.*

Proof It is straight forward.

Lemma 3.1 *Let S be a ternary semiring without zero. Then the following statements are equivalent:*

(i) *S is Q-simple,*
(ii) *$aSS \cap (SaS + SSaSS) \cap SSa = S$ for all $a \in S$.*
(iii) *$\langle a \rangle_q = S$ for all $a \in S$.*

Proof (i)⇒(ii): By above Proposition, $aSS \cap (SaS + SSaSS) \cap SSa$ is a quasi-ideal of S for all $a \in S$. Since S is Q-simple, therefore $aSS \cap (SaS + SSaSS) \cap SSa = S$ for all $a \in S$.
(ii)⇒(iii): It is clear by Proposition (3.1).
(iii)⇒(i): Let Q be a quasi-ideal of S and let $a \in Q$. Then $S = \langle a \rangle_q \subseteq Q$. Therefore $Q = S$. Hence S is Q-simple.

Lemma 3.2 *Let S be a ternary semiring. Then the following statements hold:*

(i) *If S is 0-Q-simple, then $\langle a\rangle_q = S$ for all $a \in S \setminus \{0\}$.*
(ii) *If $\langle a\rangle_q = S$ for all $a \in S \setminus \{0\}$, then either $S^3 = \{0\}$ or S is 0-Q-simple.*

Proof (i) Proof is straight forward by definition of 0-Q-simple.
(ii) Suppose $\langle a\rangle_q = S$ for all $a \in S \setminus \{0\}$ and $S^3 \neq \{0\}$. Let Q be a nonzero quasi-ideal of S. Let $a \in Q \setminus \{0\}$. Then $S = \langle a\rangle_q \subseteq Q$. Therefore $Q = S$. Hence S is 0-Q-simple.

Lemma 3.3 *Let Q be a quasi-ideal of a ternary semiring S and T be a ternary subsemiring of S. Then the following statements hold:*

(i) *If T is Q-simple such that $T \cap Q \neq \emptyset$, then $T \subseteq Q$.*
(ii) *If T is 0-Q-simple such that $(T \setminus \{0\}) \cap Q \neq \emptyset$, then $T \subseteq Q$.*

Proof (i) Suppose T is Q-simple such that $T \cap Q \neq \emptyset$. Let $a \in T \cap Q$. Then by Proposition 3.3, $aTT \cap (TaT + TTaTT) \cap TTa$ is a quasi-ideal of T for all $a \in T$. By Proposition 3.6 [4], $\{aTT \cap (TaT + TTaTT) \cap TTa\} \cap T$ is a quasi-ideal of T. Since T is Q-simple, therefore $\{aTT \cap (TaT + TTaTT) \cap TTa\} \cap T = T$. Thus, $T \subseteq aTT \cap (TaT + TTaTT) \cap TTa \subseteq QSS \cap (SQS + SSQSS) \cap SSQ \subseteq Q$. Hence $T \subseteq Q$.
(ii) Suppose T is 0-Q-simple such that $(T \setminus \{0\}) \cap Q \neq \emptyset$. Let $a \in (T \setminus \{0\}) \cap Q$. Then by Lemma (3.2)(i), we have

$$\begin{aligned} T &= \langle a\rangle_{qT} \\ &= [aTT \cap (TaT + TTaTT) \cap TTa] + Z_0^+ a \text{(by Proposition 3.1)} \\ &\subseteq [aSS \cap (SaS + SSaSS) \cap SSa] + Z_0^+ a = \langle a\rangle_q \subseteq Q. \end{aligned}$$

Therefore $T \subseteq Q$.

Theorem 3.1 *Let S be a ternary semiring without zero and Q be a quasi-ideal of S. Then the following statements hold:*

(i) *If Q is a minimal quasi-ideal without zero of S and Q is an ideal of S, then either there exists a quasi-ideal A of Q such that $AQQ \cap (QAQ + QQAQQ) \cap QQA = \emptyset$ and $QAQ \neq QQAQQ$ or Q is Q-simple.*
(ii) *If Q is Q-simple, then Q is a minimal quasi-ideal of S.*
(iii) *If Q is a minimal quasi-ideal with zero of S and Q is an ideal of S, then either there exists an nonzero quasi-ideal A of Q such that $AQQ \cap (QAQ + QQAQQ) \cap QQA = \{0\}$ and $QAQ \neq QQAQQ$ or Q is 0-Q-simple.*

Proof (i) Suppose an ideal Q is a minimal quasi-ideal without zero of S. Let A be a quasi-ideal of Q such that $AQQ \cap (QAQ + QQAQQ) \cap QQA \neq \emptyset$ and $QAQ = QQAQQ$. Clearly, Q is a ternary subsemiring of S. Now, $\emptyset \neq AQQ \cap (QAQ + QQAQQ) \cap QQA \subseteq A$. Define $H = \{h \in A : h \in AQQ \cap (QAQ + QQAQQ) \cap QQA\}$. Clearly H is non empty and $H \subseteq A \subseteq Q$. Now we show that H

is a quasi-ideal of S. Let $h_1, h_2, h_3 \in H$. Then $h_1 \in AQQ \cap (QAQ + QQAQQ) \cap QQA$. This implies $h_1 \in AQQ$, $h_1 \in QAQ + QQAQQ$ and $h_1 \in QQA$. Thus $h_1 = \sum a_i q_i p_i$, $h_1 = \sum (q_j a_j p_j + r_j s_j b_j u_j v_j)$, $h_1 = \sum q_k p_k a_k$ for all $a_i, a_j, b_j, a_k \in A$ and $q_i, p_i, q_j, p_j, r_j, s_j, u_j, v_j, q_k, p_k \in Q$. Similarly, we can define h_2 and h_3, respectively. It is easy to verify that H is a ternary subsemiring of S. Now we show that H is a quasi-ideal of S. Let $x \in HSS \cap (SHS + SSHSS) \cap SSH$. This implies $x \in HSS$, $x \in SHS + SSHSS$ and $x \in SSH$. That is, $x = h_1 SS$, $x = Sh_1 S + SSh_1 SS$ and $x = SSh_1$ for some $h_1 \in H$. Now, since Q is an ideal of S, $x = h_1 SS = (\sum a_i q_i p_i) SS \in AQQ$,

$$\begin{aligned} x = Sh_1 S + SSh_1 SS &= S\Big(\sum (q_j a_j p_j + r_j s_j b_j u_j v_j)\Big) S \\ &\quad + SS\Big(\sum (q_j a_j p_j + r_j s_j b_j u_j v_j)\Big) SS \\ &\subseteq QAQ + QQAQQ \end{aligned}$$

and $x = SS\Big(\sum q_k p_k a_k\Big) \in QQA$. Therefore $x \in AQQ \cap (QAQ + QQAQQ) \cap QQA$. Thus $x \in H$. Therefore $HSS \cap (SHS + SSHSS) \cap SSH \subseteq H$. Hence H is a quasi-ideal of S. Since Q is a minimal quasi-ideal of S, therefore $H = Q$. Thus $A = Q$. Hence Q is Q-simple.
(ii) Suppose Q is Q-simple. Let A be a quasi-ideal of S such that $A \subseteq Q$. Then $A \cap Q \neq \emptyset$. By Lemma 3.3(i), it follows that $Q \subseteq A$. Therefore $A = Q$. Hence Q is a minimal quasi-ideal of S.
(iii) Similar to (i).

Theorem 3.2 *Let S be ternary semiring without zero element having a proper quasi-ideal. Then every proper quasi-ideal of S is minimal if and only if the intersection of any two distinct proper quasi-ideals is empty.*

Proof Proof is trivial.

4 Quasi-ideals and Regular Ternary Semiring

In this section, we characterize the concept of quasi-ideals in terms of regular ternary semirings.

Theorem 4.1 *Let S be a ternary semiring with zero and let R, M, and L be a minimal right, a minimal lateral, and a minimal left ideals of S, respectively. Then RML is either $\{0\}$ or intersection of a minimal right, a minimal lateral, and a minimal left quasi-ideal of S satisfying $RML = R \cap M \cap L$.*

Proof Suppose $RML \neq \{0\}$. Then $RML \subseteq R \cap M \cap L = Q$ (by Theorem 3.2) where Q is a quasi-ideal of S. Now to show that RML is a quasi-ideal of S. That is $RMLSS \cap (SRMLS + SSRMLSS) \cap SSRML \subseteq RML$. If $RMLSS = \{0\}$

or $SRMLS + SSRMLSS = \{0\}$ or $SSRML = \{0\}$, then trivially $RMLSS \cap (SRMLS + SSRMLSS) \cap SSRML \subseteq RML$. Now let $RMLSS \neq \{0\}$, $SRMLS + SSRMLSS \neq \{0\}$ and $SSRML \neq \{0\}$. Since R, M, and L are minimal, therefore $RMLSS = R$, $SRMLS + SSRMLSS = M$ and $SSRML = L$. Now $RMLSS \neq \{0\}$, therefore there exists $0 \neq x \in RML$ with $xSS \neq \{0\}$. Since R is minimal, we have $xSS = R$. Similarly, $SxS + SSxSS = M$ and $SSx = L$. Thus

$$\begin{aligned} 0 \neq x \in RML = (xSS)(SxS + SSxSS)(SSx) &\subseteq xSxSx + xSSxSSx \\ &\subseteq xSSSx + xSSSSSx \\ &\subseteq xSx. \end{aligned}$$

Therefore, x is regular in S. Consequently, $RML = R \cap M \cap L$ (by Theorem 3.4 [4]). Now

$$\begin{aligned} &RMLSS \cap (SRMLS + SSRMLSS) \cap SSRML \\ &\quad \subseteq RSS \cap (SMS + SSMSS) \cap SSL \\ &\quad \subseteq R \cap M \cap L = RML. \end{aligned}$$

Hence RML is a quasi-ideal of S.

Theorem 4.2 *A ternary subsemiring Q of a regular ternary semiring S is a quasi-ideal of S if and only if $Q = QSQ$.*

Proof Since S is regular ternary semiring, therefore every bi-ideal of S is a quasi-ideal of S. Hence, result follows by Theorem 3.28 [4].

Theorem 4.3 *Let S be a ternary semiring. Then the following conditions are equivalent:*

(1) *Each right ideal R, lateral ideal M, and left ideal L of S satisfy $R \cap M \cap L = RML \subseteq LRM \cap MLR$.*
(2) *The set Q of all quasi-ideals of S is an idempotent ternary semiring with respect to the "product" $Q_1Q_2Q_3$.*
(3) *Each quasi-ideal Q of S satisfies $Q = Q^3$.*

Proof (1)⇒(2) The equality of condition (1) yields that S is regular (by Theorem 3.4 [2]). Now the set Q of all quasi-ideals of S is a ternary semiring with respect to the product $Q_1Q_2Q_3$ (by Corollary 3.32 [4]). Now we show that Q is an idempotent. We have

$$
\begin{aligned}
Q &= QSQSQ \quad \text{(by Theorem 3.28 [5])} \\
&= (QSQS)(QSQ)(SQS)(QSQ)(S)(QSQ)(SQS)(QSQ)(SQ) \\
&\subseteq (QSQS)\{(QSS)(SQS)(SSQ)\}(S)\{(QSS)(SQS)(SSQ)\}(SQ) \\
&\subseteq (QSQS)\{(SSQ)(QSS)(SQS)\}(S)\{(SQS)(SSQ)(QSS)\}(SQ) \\
&\qquad\qquad \text{(since RML} \subseteq \text{LRM} \cap \text{MLR)} \\
&= [(QSQS)SSQ][(QSS)(SQS)S(SQS)(SSQ)][(QSS)SQ] \\
&\subseteq (QSQSQ)(QSQSQSQ)(QSQ) \\
&= (QSQSQ)(QSQSQ)(QSQSQ) \quad (\text{since} Q = QSQ) \\
&= Q^3.
\end{aligned}
$$

Hence $Q = Q^3$.
(2)⇒(3) Straight forward.
(3)⇒(1) Let R, M and L be right, lateral and left ideal of S respectively. By Theorem 3.8 [4], the intersection $R \cap M \cap L$ is a quasi-ideal of S.

Therefore, condition (3) implies

$$RML = R \cap M \cap L = (R \cap M \cap L)^3 \subseteq LRM \cap MLR.$$

Theorem 4.4 *Let S be a regular ternary semiring. Then the following assertions hold:*

(1) *Every ideal of S is an idempotent.*
(2) *Every bi-ideal of any lateral ideal of S is a quasi-ideal of S.*

Proof (1) Straightforward by Theorem 3.4 [4].
(2) By Lemma 4.2 [1] every lateral ideal of a regular ternary semiring S is a regular ternary semiring. Therefore result follows by Theorem 3.30 [4].

Proposition 4.1 [4] Let S be a ternary semiring and $a \in S$. Then the principal left ideal generated by a is given by $< a >_l = \left\{\sum r_i s_i a + na : r_i, s_i \in S : n \in Z_0^+\right\}$, right ideal generated by a is given by $< a >_r = \left\{\sum a r_i s_i + na : r_i, s_i \in S : n \in Z_0^+\right\}$ and lateral ideal generated by a is given by $< a >_m = \left\{\sum r_i a s_i + \sum p_j q_j a r_j s_j + na : r_i, s_i, p_j, q_j, r_j, s_j \in S : n \in Z^+0\right\}$ where $\sum$ denote the finite sum and Z_0^+ is the set of all positive integer with zero.

Theorem 4.5 *Let S be a ternary semiring. Then the element a is regular in S if and only if the principal quasi-ideal $\langle a\rangle_q$ of S satisfies $\langle a\rangle_q = \langle a\rangle_q S\langle a\rangle_q S\langle a\rangle_q$.*

Proof Suppose a is regular in S. Clearly, $\langle a\rangle_q \subseteq \langle a\rangle_q S\langle a\rangle_q S\langle a\rangle_q$. Using Theorem (3.8) [4], it is easy to show that $\langle a\rangle_q = \langle a\rangle_r\langle a\rangle_m\langle a\rangle_l$. Now Let $x \in \langle a\rangle_q S\langle a\rangle_q S\langle a\rangle_q \subseteq \langle a\rangle_r S\langle a\rangle_m S\langle a\rangle_l = \{na + aSS\}S\{na + SaS + SSaSS\}S\{na + SSa\} \subseteq aSa = a \in \langle a\rangle_q$.

Conversely, suppose that $a \in \langle a\rangle_q = \langle a\rangle_q S\langle a\rangle_q S\langle a\rangle_q \subseteq \langle a\rangle_r S\langle a\rangle_m S\langle a\rangle_l \subseteq aSa$. Hence a is regular in S.

Theorem 4.6 *A ternary semiring S is regular if and only if for every bi-ideal B, every lateral ideal M and every quasi-ideal Q, we have $B \cap M \cap Q \subseteq BMQ$.*

Proof Suppose S is regular. Let $a \in B \cap M \cap Q$. Since S is regular, therefore for $a \in S$ there exist $x \in S$ such that $a = axa = axaxaxa = (axaxa)(xax)(a) \in (BSBSB)(SMS)(Q) \subseteq BMQ$. Conversely, suppose that $B \cap M \cap Q \subseteq BMQ$. Let $a \in S$. Consider the bi-ideal $\langle a\rangle_b$ of S generated by a, the lateral ideal $\langle a\rangle_m$ of S generated by a and the quasi-ideal $\langle a\rangle_q$ of S generated by a. Then

$$\begin{aligned} a \in \langle a\rangle_b \cap \langle a\rangle_m \cap \langle a\rangle_q &\subseteq \langle a\rangle_b \langle a\rangle_m \langle a\rangle_q \\ &\subseteq \{na + ma^3 + aSaSa\}\{pa + SaS + SSaSS\} \\ &\quad \{qa + aSS \cap (SaS + SSaSS) \cap SSa\} \\ &\qquad \text{for } n, m, p, q \in Z_0^+ \\ &\subseteq aSa. \end{aligned}$$

Hence there exists an element $x \in S$ such that $a = axa$. This implies that a is regular and hence S is regular.

Remark 1 Every left ideal is a quasi-ideal (by Lemma 3.3 [4]) and every quasi-ideal is a bi-ideal (by Lemma 3.15 [4]). Taking a left ideal L instead of a quasi-ideal Q in Theorem 4.5, we get the following theorem.

Theorem 4.7 *A ternary semiring S is regular if and only if for every bi-ideal B, every lateral ideal M and every left ideal L, we have $B \cap M \cap L \subseteq BML$.*

Theorem 4.8 *A ternary semiring S is regular if and only if for every right ideal R, every left ideal L and every quasi-ideal Q, we have $R \cap Q \cap L \subseteq RSQSL$.*

Proof Suppose S is regular. Let $a \in R \cap Q \cap L$. Since S is regular, therefore for $a \in S$ there exists $x \in S$ such that $a = axa = axaxaxa = (axa)xax(axa) \in (RSS)SQS(SSL) \subseteq RSQSL$. Conversely, suppose that $R \cap Q \cap L \subseteq RSQSL$. Let $a \in S$. Consider the right ideal $\langle a\rangle_r$ of S, the quasi ideal $\langle a\rangle_q$ of S and the left ideal $\langle a\rangle_l$ of S generated by a respectively. Then

$$\begin{aligned} a \in \langle a\rangle_r \cap \langle a\rangle_q \cap \langle a\rangle_l &\subseteq \langle a\rangle_r S \langle a\rangle_q S \langle a\rangle_l \\ &\subseteq \langle a\rangle_r SSS \langle a\rangle_l \subseteq aSa. \end{aligned}$$

Hence, there exists an element $x \in S$ such that $a = axa$. This implies that a is regular and hence S is regular.

5 Conclusion

In this paper, we have generalized the results of quasi-ideals in ternary semirings which are analogous to ring theory. We also find the relation between ternary semiring and left ideal, right ideal, quasi-ideal of ternary semirings. We have also introduced the concept of a Q-simple ternary semiring and 0-Q-simple ternary semiring, which can be useful for the study of various algebraic systems.

References

1. Dutta, T.K., Kar, S.: On regular ternary semirings, advances in Algebra. In: Proceedings of the ICM Satellite Conference in Algebra and Related Topics, pp. 343–355. World Scientific, New Jersey (2003)
2. Dutta, T.K., Kar, S.: A note on regular ternary semiring. Kyungpook J. Math. **46**, 357–365 (2006)
3. Iampan, A.: Characterization of ordered Quasi-ideals of ordered Γ-semigroups. Kragujevac J. Math. **35**(1), 13–23 (2011)
4. Kar, S.: On quasi-ideals and bi-ideals in ternary semirings, Int. J. Math. Sci. **18**, 3015–3023 (2005)
5. Lee, S.K., Kang, S.G.: Characterizations of regular po-semigroups. Comm. Korean Math. Soc. **14**(4), 687–691 (1999)
6. Lehmer, D.H.: A ternary analogue of an Abelian groups. Am. J. Math. **59**, 329 (1932)
7. Lister, W.G.: Ternary rings. Trans. Am. Math. Soc. **154**, 37 (1971)
8. Steinfeld, O.: Quasi-ideals in Rings and Semigroups. Akademiai Kiado, Budapest (1978)

Chapter 22
Epidemiological Models: A Study of Two Retroviruses, HIV and HTLV-I

Dana Baxley, N. K. Sahu and Ram N. Mohapatra

Abstract HIV is an example of a disease where the pathogen mutates so that it is not recognized by the immune system. In this paper, we have studied several models and two retroviruses, viz., HIV and Human T-lymphotropic virus (HTLV-I). We have used SIMULINK to draw graphs and study the associated modeling problems.

1 Introduction

Disease has played an important part throughout the history of mankind. Diseases have influenced the growth or decline of a population and have impact on the economy. It causes more deaths than any other source, including war and natural disasters. The manner in which diseases infect and invade a population has perplexed doctors and scientist for many years. A branch of science called epidemiology was developed in order to help analyze and understand the spread of disease.

Aristotle and Hippocrates of Cos started studying the transmission of diseases during 300–400 BC. Later on, germ theory was first studied by Jacob Henle in 1840 and was later developed by Robert Koch, Joseph Lister, and Louis Pasteur. Modern mathematics was first used in the study of diseases in 1873 by P. D. En'ko. Sir R. A. Ross, W. H. Hamer, A. G. Mckendrick, and W. O. Kermack laid the foundation of mathematics in epidemiology between 1900 and 1935 (see [18]). The study

D. Baxley · R. N. Mohapatra (✉)
Department of Mathematics, University of Central Florida, Orlando, FL 32816, USA
e-mail: ram.mohapatra@ucf.edu

D. Baxley
e-mail: dbaxley28@gmail.com

N. K. Sahu
Department of Mathematics, IIT Kharagpur, Kharagpur 721302, India
e-mail: nabin@maths.iitkgp.ernet.in

R. N. Mohapatra et al. (eds.), *Mathematics and Computing 2013*,
Springer Proceedings in Mathematics & Statistics 91,
DOI: 10.1007/978-81-322-1952-1_22, © Springer India 2014

of epidemiology has grown tremendously since and most known communicable diseases have been modeled and analyzed.

Epidemiology not only helps us to understand disease transmission, but also to know how to control the spread of a particular disease. It is not a static science and is constantly changing. Infectious diseases are constantly evolving and changing, making them harder to control. New strains, which are immune to antibiotics, are discovered everyday. HIV and Human T-lymphotropic virus (HTLV-I) are two new viruses which were first discovered in the 1980s. These viruses have no known cure but doctors are working with epidemiologists, mathematicians, and scientists to find a cure and limit its transmission. We will use mathematical models to help us understand the spread of these viruses in the human body and the progression of these viruses to disease.

The memory immune responses enable humans and animals to rapidly clear, or even prevent altogether, infection by pathogen with which they have previously been infected. As an example, one typically contracts chicken pox at the most once in a lifetime. One cause is the effectiveness of the memory response and the vaccines designed around the knowledge that our immune system will efficiently fight foreign invaders if already exposed to something similar. As a result, many pathogens use the strategy of disguise to survive in the host population. With enough mutation, a pathogen will ultimately be unrecognizable to the immune system of a host that previously has been infected with one of its ancestors. This ability to mutate allows the pathogens to escape partially the host immunity acquired from previous infections. In influenza A and Canine parvovirus, new antigenic variants arise continually affecting the epidemiology of the disease. In this study, we concentrate on the study of two retroviruses and use SIMULINK for analysis.

1.1 Introduction to HIV

In 1981, the Center for Disease Control reported an unusual collection of homosexual males that had Pneumocyctiscarinii pneumonia and Kaposi's sarcoma. These men were previously healthy individuals. This was a new retroviral disease later to be named AIDS or Acquired Immunodeficiency Syndrome, a disease for which there is still no cure and is the fourth leading cause of death worldwide. The etiologic agent of this new epidemic is the human immunodeficiency virus or HIV, which will be studied in detail in this paper. HIV is the retrovirus which causes AIDS. This virus slowly destroys the immune system over many years. Once the immune system is depleted, AIDS occurs. For more on HIV and AIDS, one may refer to [14, 23].

AIDS was first discovered in the United States but now affects the entire world and is considered the new "plague." It has killed more than 25 million people worldwide and is considered the most destructive epidemic in the recorded history. AIDS is now found in more than 163 countries with the most being in Africa and the Caribbean being the second. Sub-Saharan Africa is considered to be the global epicenter of the HIV epidemic (see [5]). Ninety percent of the individuals infected with HIV are

in developing countries and 40 Individuals in the 15–24 age group are the fastest growing segment who are being infected with HIV.

HIV can be transmitted in different ways. The virus is present in bodily fluids, specifically in blood, therefore, any activity that results in the transfer of bodily fluid can potentially result in the transfer of HIV. Intimate sexual contact is one of the modes in which fluid transfer occurs. Intravenous drug use is another mode in which HIV is transmitted between individuals because many drug users share needles. Two other modes which are not as common due to medical advances and new antiretroviral drugs are mother to child transmission and transmission through blood transfusion. Mother to child transmission can occur during the birthing process or through breast feeding. Although the rate of mother to child transmission has dropped in many developing countries, it is still prevalent in the sub-Saharan regions of Africa. Transmission of HIV through blood transfusions is rarely seen today due to examination of the blood from donors for presence of HIV prior to saving them in the blood bank for patient use. The U.S. blood supply is very safe due to the extensive questioning of blood donors and the extensive testing of donated blood.

HIV is characterized by immunosuppression, neurologic involvement, and secondary tumors. HIV attacks the CD4+ T cells, which are responsible for the immune system. The nature of this attack and how it occurs is modeled mathematically in Sect. 3 in order to help us understand and predict the course of the disease. Many graphs developed from the mathematical model help demonstrate the progression to AIDS. The graphs were produced using Simulink and match those produced by Stilianakis and Schenzle in Fortran.

1.2 Introduction to HTLV-I

HTLV-I was the first retrovirus to be discovered. This virus was discovered in Japan in 1980. HTLV-I is a virus which lies latent for many years before causing other diseases to proliferate. This virus is the predominant cause of two diseases. The first one is Adult T Cell leukemia/lymphoma or ATL, which is a T cell non-Hodgkin's lymphoma with a leukemic phase of circulating CD4+ T cells (see [4]). The progression from HTLV-I to ATL is mathematically modeled and studied in Sect. 4. The discovery of HTLV-I provided scientists with a clear proof of a relationship between viruses and cancer. The second disease that is caused by HTLV-I is myelopathy (HAM) which is also known as tropical spastic para paresis (TSP). Usually, this virus does not produce disease until approximately 20 years after initial infection. HTLV-I can also cause autoimmune or chronic inflammatory disorders such as arthropathy, Sjogren's syndrome, and facial nerve palsy. Identification of the HTLV-I virus facilitated the discovery and isolation of HIV.

HTLV-I infects 10–20 million people worldwide but only produces disease in approximately 5 % of infected individuals. Women are twice as likely to contract HTLV-I as men. The HTLV-I infection is thought to occur in geographical clusters which are located in southern Japan, the Caribbean, parts of Africa, the Middle East,

South America, Pacific Melanesian islands, and Papua New Guinea. The virus is also found in southeastern United States in certain immigrant groups.

HTLV-I is transmitted in the same way as HIV, through bodily fluid transfer. Unlike HIV, the main transmission is through breast feeding. The HTLV-I antigen is found in the infected mother's milk and is transmitted most likely through lymphocytes in the milk. The prevalence of this vertical transmission through breast feeding has caused a clustering of cases in familial or geographically discrete groups (see [17]). Other modes are sexual transmission, infection from blood transfusion, and sharing needles among drug users. For more details on HTLV-I such as disease associations, diagnosis, and treatment, one may refer to [16].

Many people can be infected with HTLV-I and will never develop a disease from this virus. Section 4 will feature a mathematical model of the HTLV-I infection of CD4+ T cells and the eventual progression to ATL. The stability analysis will illustrate two different steady states. One steady state when the virus will not progress to ATL, and another steady state when the virus will progress to ATL. A proposition for asymptotical stability is studied and a graph was produced using Simulink. Even after rigorous analysis, this graph does not match the graph presented by the authors and further work may be needed to explore the difference.

2 Mathematical Models

A mathematical model is a mathematical description of a real-world system or event (see [15]). Epidemiologists will use mathematical models to understand and predict the course of an infection or disease. A well-formulated model can help an epidemiologist to determine where resources need to be allocated and how those resources ca help control or eradication of the disease. In order to formulate a model for an infectious disease, an individual must first collect an abundance of empirical data through clinical testing. Once these data are collected and analyzed, the modelers develop a model using the following steps. First, they note all the relevant assumptions and then determine the relationship between the variables and parameters used in the model and, finally, analyze any specific patterns that are found. Deciding which parameters and variables will be used in the model and how much importance should be given depend on the characteristics of the disease under study and the intention of the model (see [2]). Once the model is formulated and analyzed, it will help the scientists to draw inferences from a set of hypotheses in order to determine the course of the disease in an individual or in a population. For mathematical study of malaria models, one may refer to [12].

Epidemiological models are usually formed using the general *MSEIR* model. This model places individuals from a constant population into certain groups within the model and describes the transition rate between each group. Each letter represents a different class or group. M represents the temporary immunity that a mother can pass on to her child through the placenta. The S describes the susceptibles, which are the members of the population who are at risk for contracting the disease. E stands

for exposed and describes the individuals from the population who are infected by the disease but are not infectious due to a latent period of a disease. The I group is the infectious group or the individuals from the population who have the ability to pass the disease to other members of the population. R represents the group of individuals who have recovered from the disease, whether temporary or permanent, and also possess some type of immunity. For a detailed mathematical study on HIV and HTLV-I, one may refer to [1].

2.1 Basic SIR Model

The first model to consider is the basic *SIR* model. It is a simple epidemic model developed by Kermack and McKendrick in 1927 to predict the behavior of many historical epidemics such as cholera, influenza, and the Great Plague. This model is used by many epidemiologists because it ca help to predict the behavior and progress of different diseases. This model is also a building block for many of the other more complicated models. The *SIR* model considers a population that remains constant. The population is divided into three classes: first S, the individuals who are susceptible to the disease, second I, the individuals being exposed and infected by the disease, and third R, the individuals who will recover from the infection and gain immunity to the disease. This model does not consider any latent period of the disease. Once an individual is infected, he is automatically moved into the infectious classification. The progress of the individuals from class to class can be demonstrated by

$$\mathbf{S} \longrightarrow \mathbf{I} \longrightarrow \mathbf{R}.$$

Some models only consider the S and I classes. Other models consider a fourth class E, which takes in account a latent period of the disease in which the virus is present in the host but has not infected the host. When modeling a disease like AIDS, it is better to use a model which includes this class.

This model makes many assumptions. We must first assume the collection of individuals in each class is a differentiable function of time. This is reasonable as long as there are enough people in each class. Next, the model is deterministic. This means that the behavior of the model is determined by the past behavior of diseases. A stochastic model would be more effective if the model described classes with small populations. Third, this model does not include a latent phase of the disease, which means that once a susceptible becomes infected, the individual is automatically placed into the infected class. Fourth, the model assumes that an infected individual makes contact significant enough to transmit the disease at the contact rate β.

If $\beta N \frac{S}{N} I = \beta SI$, new cases will occur when N is the total number in the population, S is the number of susceptibles and I is the number of infecteds. The fifth assumption is that the model has a mass action principle, which means every individual within the population has an equal chance to have contact with every other individual in the population. This information implies that β, the contact rate is the

ratio of rate of contact to the population size. Another assumption is that the recovery rate is proportional to the number of infecteds, and is represented by aI, where a is the removal rate. The last assumption is that there is no entry or exit from the population except through death. This occurs when the progression to disease is so quick that birth and death rates can be ignored. This assumption can be changed in certain models.

Based on these assumptions, the classic Kermack and McKendrick model is:

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\beta SI \tag{1}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \beta SI - aI \tag{2}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = aI. \tag{3}$$

Note that only nonnegative solutions for *S, I* and *R* are of interest. Also remember the total population is constant and is embedded in the model. If we add Eqs. (1)–(3), we will get:

$$\frac{\mathrm{d}S}{\mathrm{d}t} + \frac{\mathrm{d}I}{\mathrm{d}t} + \frac{\mathrm{d}R}{\mathrm{d}t} = 0. \tag{4}$$

Solving this differential equation we get

$$S(t) + I(t) + R(t) = N, \tag{5}$$

where N is the population size.

We also have the following initial conditions

$$S(0) = S_0, I(0) = I_0, R(0) = R_0, \tag{6}$$

where $S_0 > 0$ and $I_0 > 0$.

The population is constant, therefore, R can be determined if S and I are known. For this reason, Eq. (3) can be dropped and the system can be reduced to only two equations. This system is not possible to solve analytically but the equations can be analyzed using a qualitative approach. Note that $S' < 0$ and $I' > 0$ if $S_0 > \frac{a}{\beta}$. Since S is decreasing, I will initially increase but then will decrease to zero. The possibility of I increasing is what indicates an epidemic because I represents the infected individuals. If $S_0 < \frac{a}{\beta}$, then I will go to zero and there is no epidemic. If $S_0 > \frac{a}{\beta}$, the number of infected individuals will first increase to $S = \frac{a}{\beta}$, and then decrease to zero. From this, we see a threshold parameter. The behavior of the disease will depend on the threshold quantity, $\frac{S_0\beta}{a}$. This number defines the reproduction number. The reproduction number R_0 of the system is defined as the

number of secondary infections produced by one primary infection in the population of susceptibles. Therefore we have

$$R_0 = \frac{S_0\beta}{a}. \tag{7}$$

This number measures how fast the infection will spread. If $R_0 < 1$, the infection will not continue and the disease will disappear. If $R_0 = 1$, the infection will remain stable in transmission. If $R_0 > 1$, an epidemic will occur (see [3]). To find the trajectories in the phase plane, we first divide the two equations of the model and get

$$\frac{\mathrm{d}I}{\mathrm{d}S} = \frac{(\beta S - a)I}{-\beta SI} = -1 + \frac{a}{\beta S}. \tag{8}$$

Separation of variables and integration yields

$$I = -S + \frac{a}{\beta}\log S + c, \tag{9}$$

where c is an arbitrary constant of integration. Equation (9) can be defined as the following quantity:

$$J(S, I) = S + I - \frac{a}{\beta}\log S, \tag{10}$$

where $J(S, I) = c$. Different constants will give different trajectories and this constant can be obtained by knowing the initial values of S, I, S_0 and I_0. Now we have

$$J(S_0, I_0) = S_0 + I_0 - \frac{a}{\beta}\log S_0 = c. \tag{11}$$

If we assume a population of size K and introduce a small number of infecteds into the population, that is $S_0 \approx K$ and $I_0 \approx 0$, we can determine $R_0 = \frac{\beta K}{a}$ from Eq. (7). Taking the fact that $\lim_{t\to\infty} I(t) = 0$ and $\lim_{t\to\infty} S(t) = S_\infty$, we can find $J(S_0, I_0) = J(S_\infty, 0)$. This will yield

$$K - \frac{a}{\beta}\log S_0 = S_\infty - \frac{a}{\beta}\log S_\infty. \tag{12}$$

This helps to determine the reproduction number because it will give an expression for $\frac{\beta}{a}$ that can be determined by

$$K - S_\infty = \frac{a}{\beta}\log\frac{S_0}{S_\infty} \tag{13}$$

$$\frac{\beta}{a} = \frac{\log\frac{S_0}{S_\infty}}{K - S_\infty}. \tag{14}$$

Note that $S_0 > S_\infty$ because the initial number of susceptibles will be greater than the number of susceptibles who will become infected. This will occur because there are some who will not come into contact with the disease.

2.2 Basic SIS Model

The *SIS* model is another type of model to study infectious diseases. In this model, the infected will return to the susceptible class after recovery. This model is more effective to use when studying sexually transmitted diseases. The simplest model, which was given by Kermack and McKendrick is

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\beta SI + aI \tag{15}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \beta SI - aI. \tag{16}$$

This model is different from the *SIR* model in that the recovered members will return to the susceptible class at a rate of aI instead of moving to a recovered class. Just as in the SIR model the total population is constant, since $(S + I)' = 0$. Again, let the constant population be represented by K. If $K = S + I$, we can replace S by K-I and reduce the model to a single differential equation. This equation is

$$\begin{aligned}\frac{\mathrm{d}I}{\mathrm{d}t} &= \beta I(K - I) - aI \\ &= (\beta K - a)I - \beta I^2 \\ &= (\beta K - a)I\left(1 - \frac{I}{K - \frac{a}{\beta}}\right).\end{aligned} \tag{17}$$

This is a logistic equation with a growth rate of $\beta K - a$ and carrying capacity of $K - \frac{a}{\beta}$. An analysis of this will show that if $\beta K - a < 0$ or $\frac{\beta K}{a} < 1$, then for any $I_0 > 0$, we see that $\lim_{t\to\infty} I(t) = 0$ and $\lim_{t\to\infty} S(t) = K$. If $\frac{\beta K}{a} > 1$, then for any $I_0 > 0$, we will see that $\lim_{t\to\infty} I(t) = K - \frac{a}{\beta}$ and $\lim_{t\to\infty} S(t) = \frac{a}{\beta}$. As seen here, there is a single limiting value for I and this limiting value is determined by the quantity $\frac{\beta K}{a}$, regardless of the initial rate of infection. The infection will disappear or the number of infected will approach zero when $\frac{\beta K}{a} < 1$. Hence the equilibrium $I = 0$ and $S = K$ is considered the disease free equilibrium. If $\frac{\beta K}{a} > 1$, the infection will continue. The equilibrium $I = K - \frac{\beta}{a}$ which corresponds to $S = \frac{a}{\beta}$ is defined as the endemic equilibrium.

The dimensionless quantity $\frac{\beta K}{a}$ is the reproduction number for our system, noted as $R_0 = \frac{\beta K}{a}$. In Sect. 2.1, we discussed that the value of R_0 was the threshold

parameter. We also defined R_0, as the number of secondary infections produced by one primary infection in the population of susceptible. The reproduction number helps to determine the path which the disease will take. If $R_0 = \frac{\beta K}{a}$, where βK is the number of contacts made by an average infected per unit of time and $\frac{1}{a}$ is the mean infected period, we can clearly see if $R_0 < 1$, the infection will disappear and if $R_0 > 1$, the infection will persist.

3 Intrahost Dynamics of HIV

To understand how HIV destroys the immune system, we first must understand how the immune system works. When a foreign substance or antigen enters the body, the body will initiate an immune response. This immune response starts with macrophages and monocytes. These cells are the body's first defense against the antigen. They will seek out the antigen surround it and overtake it. This process is known as phagocytosis. The macrophages will then analyze the content of the antigen and pass this information along to the CD4+ T lymphocytes, also called CD4+ T cells (see [11]). The CD4+ T cells will call for the production of more CD4+ T cells or will call for the production of types of T cells such as the CD+8-T cells. Another weapon used by the body's defense system is the B lymphocytes or B cells. These cell produce antibodies specifically engineered to destroy the pathogen detected by the macrophages (see [7]).

HIV is considered a lentivirus, meaning slow virus, which is a subclass of the retrovirus. In general, a virus will insert its own DNA into the host cell. When the host cell replicates its DNA, the virus' DNA is also replicated. A retrovirus like HIV will insert RNA rather than DNA into the host. Retroviruses have a unique enzyme named reverse transcriptase (see [5]). This enzyme will prepare a DNA copy of the RNA genome into the host. This DNA copy is eventually inserted into the genome of the host cell where the virus will persist for years and is impossible to eradicate (see [21]). The HIV DNA will get copied every time the host cell divides.

On the cellular level, the HIV particles target the CD4+ T lymphocyte. It attracts the CD4+ T lymphocyte through a glycoprotein called gp120. The protein enzyme, gp120 is located on the surface of the HIV particle and is attracted to the CD4 protein on the surface of the T cells, macrophages, and monocytes. The CD4+ T cell attaches itself to the virus and is infected.

The HIV infection can typically be divided into three phases. The first phase is the primary infection. During this initial phase, the virus is present in the host and replicates in the manner describe previously. Three to six weeks after the infection, 50–75 % of the patients develop an acute viral syndrome (see [21]). There is also a significant reduction in CD4+ T cells. The second phase of HIV infection is the longest phase. It is the phase in which there is a long asymptomatic period and latency occurs. There are two major features of this phase. The first feature is the permanent viral replication in the lymphatic tissue and lymphoid organs. The second feature is

Table 1 Variables used in the simple model

Variable	
X	Total number of susceptible CD4+ T cells
Y	Total number of productively infected CD4+ T cells
V	Total number of HIV particles
K	Factor that describes the increase of the CD4+ T cell infection rate

the gradual decline of the CD4+ T cells. The final phase of the HIV infection shows a sharp decline in CD4+ T cells and the emergence of clinical immunodeficiency and progression to AIDS. The period of time from initial infection to the formation of AIDS can vary from person to person. The median estimate is 8–11 years without treatment and even longer with treatment (see [19]).

3.1 HIV Simple Model

Stilianakis and Schenzle developed this basic model to describe the long-term dynamics of HIV progression through the body and the eventual development of AIDS. The basic biomedical assumption of this model is the genetic variation of HIV. It is assumed that the infection rate is the major source for the increase and selection of the HIV mutants (Table 1).
The model consists of the following nonlinear differential equations:

$$\frac{\mathrm{d}X}{\mathrm{d}t} = \Lambda - \mu X - \kappa_0 KVX \tag{18}$$

$$\frac{\mathrm{d}Y}{\mathrm{d}t} = \kappa_0 KVX - \delta Y \tag{19}$$

$$\frac{\mathrm{d}V}{\mathrm{d}t} = \beta Y - \gamma V \tag{20}$$

$$\frac{\mathrm{d}K}{\mathrm{d}t} = \omega_K V(K_{\max} - K). \tag{21}$$

The biological representation of each term in each equation will now be discussed to provide a better understanding of the system. In the first Eq. (18), Λ represents the constant rate at which new CD4+ T cells are produced. These newly produced CD4+ T cells are considered to be susceptible. The term μX is the rate at which susceptible cells die. The last-term $\kappa_0 KVX$ is considered a mass action term which describes the rate at which susceptible cells are infected by the HIV particles. This

Table 2 Variables used in the extended model

Variable		Initial values
X	Total number of nonsusceptible CD4+ T cells	$X(0) = X_0 = 0.7 \times 2.5 \times 10^{11}$
S	Total number of susceptible CD4+ T cells	$S(0) = S_0 = 0.3 \times 2.5 \times 10^{11}$
Y	Total number of productively infected CD4+ T cells	$Y(0) = Y_0 = 0$
V	Total number of HIV particles	$V(0) = V_0 = 1$
Z	Anti-HIV activity of the immune system	$Z(0) = Z_0 = 10^{-6}$
P	Fraction of new CD4+ T cells entering the pool of susceptible CD4+ T cells	$P(0) = P_0 = 0.3$
K	Factor that describes the increase of the CD4+ T cell infection rate	$K(0) = K_0 = 1.0$
N	Total number of uninfected CD4+ T cells	$N(0) = N_0 = X_0 + S_0 = 2.5 \times 10^{11}$

mass action term is also seen in the first term of Eq. (19). The second term in Eq. (19) is δY. This term describes the death rate of the infected CD4+ T cells.

The first term in Eq. (20) is βY. This term represents the rate in which infectious viral particles infect the CD4+ T cells. γV represents the rate at which virus particles are cleared. The last Eq. (21) represents how fast the virus can reproduce within the host and the maximum amount of virus particles that can be seen within the host at any particular time within the evolutionary process.

The rate at which the virus reproduces is called the virus reproduction number. In this model, it is a dynamic quantity and it changes over time. The virus reproduction number is

$$R_0(t) = \frac{\beta\kappa_0 K(t) X_0}{\delta\gamma}. \tag{22}$$

This reproduction number will increase monotonically toward

$$R_0^* = \frac{\beta\kappa_0 K_{\max} X_0}{\delta\gamma}. \tag{23}$$

3.2 HIV Extended Model

The following model is an extension of the original basic model. The extended model takes into account the total number of susceptible CD4+ T cells, and how fast new CD4+ T cells become susceptible to the HIV infection (Table 2).

The model consists of the following nonlinear differential equations:

$$\frac{dX}{dt} = \alpha(1 - P) - \mu X \tag{24}$$

$$\frac{dS}{dt} = \alpha P - \mu S - \kappa_0 K V \frac{S}{(P + d)} \tag{25}$$

$$\frac{dY}{dt} = \kappa_0 K V \frac{S}{(P + d)} - (\mu_Y + \delta_Y Z)Y \tag{26}$$

$$\frac{dV}{dt} = \beta Y - (\mu_V + \delta_V Z)V \tag{27}$$

$$\frac{dZ}{dt} = \theta g(V) + \rho[f(S + X)Z_{\max} - Z] \tag{28}$$

$$\frac{dP}{dt} = \omega_P V(P_{\max} - P) \tag{29}$$

$$\frac{dK}{dt} = \omega_K V(K_{\max} - K) \tag{30}$$

where

$$f(N) = \frac{1 + b^c}{1 + (\frac{bN_0}{N})^c} \quad \text{and} \quad g(V) = \frac{V}{a + V}. \tag{31}$$

N can be divided into the number of nonsusceptible T cells X, and the number of susceptible T cells S. Therefore, $N = X + S$.

To understand the system, an understanding of what each term biologically represents must be presented. In Eq. (24), P is the fraction of new CD4+ T cells that enter the susceptible and $1 - P$ is the fraction of new CD4+ T cells that remain unsusceptible to the HIV virus. The first term in Eq. (24) is $\alpha(1 - P)$, where α is the T cell production rate. This term represents the immigration rate of new nonsusceptible T cells. The second term is μX in which μ is the natural death rate of the unsusceptible cells. Therefore, this term represents how many nonsusceptible CD4+ T cells die.

Equation (25) represents the dynamics of the susceptible cells in the system. The first term αP, describes the immigration rate of the susceptible CD4+ T cells. The second-term μS, represents the natural death rate of the susceptible cells. The last term in Eq. (25) is $\kappa_0 K V \frac{S}{(P+d)}$. This is a mass action term which describes the infection process between cells and viruses. In particular, $\frac{S}{(P+d)}$ describes the dynamics changes in the susceptible cells. The variable P, in this term is very important in helping determine the course of the infection and the progression of the disease. In fact, P shows that more cells can be attacked and infected by the virus than the immune system can combat.

Equation (26) has many terms and this equation determines how many productively infected cells are in the blood. The first term is the same mass action term that

is seen in Eq. (25). The second term in Eq. (26) is $(\mu_Y + \delta_Y Z)Y$. μ_Y represents the death rate of productively infected cells and $\delta_Y Z$ represents how fast these dead cells are removed from the system. Equation (27) describes the number of HIV particles that are produced and destroyed. The first term in Eq. (27) is βY. In this term, β describes the rate at which HIV particle cells are produced from infected cells. The second-term $(\mu_V + \delta_V Z)V$, describes the rate at which HIV particles are cleared and eliminated. The term μ_V is the rate in which virus particles are cleared and $\delta_V Z$ represents the anti-HIV activity and elimination.

Equation (28) is the most complicated equation within the model because not much is known about the dynamics of the HIV-specific immune response. Therefore, a general equation is used to model this response. The equation shows the coupling of a time-dependent decline of the CD4+ T cells and the intrinsic features of the immune response. The variable ρ in Eq. (28), represents the HIV-specific immune response. This response occurs independently of the number of HIV particles that are present in the body. The function g(V), models how the immune response is activated depending on the quantity of the virus. The term $\rho[f(S + X)Z_{\max}]$ is the rate once primary infection occurs in which HIV will start producing specific antibodies and the cytotoxic cells will start multiplying. Once this occurs, the immune system will eventually become independent of the number of HIV particles and infected cells. In Eq. (28), the function $f(N)$ describes how the activity of the immune system is related to the number of available uninfected cells. This function also takes account of the immune system's ability to combat HIV when the number of CD4+ T cells is not sufficiently high.

Equation (29) describes the increase in the rate of the fraction of new cells coming from the pool of susceptible cells and how they correspond to the generation and selection of HIV mutants. Equation (29) describes the rate at which the HIV infection increases due to the reproduction of each virus particle.

The virus reproduction number is also an important value to discuss. The reproduction number represents how quickly the virus is reproducing. The HIV reproduction number must be above one in order to show a persistent infection. The virus reproduction number for this model is

$$\bar{R}_0 = \frac{\beta\kappa_0 \bar{K}\,\bar{S}}{(\mu_Y + \delta_Y \bar{Z})(\mu_V + \delta_V \bar{Z})(\bar{P} + d)}. \tag{32}$$

If the values S, Z, K and P could be held at fixed values $\bar{S}$, $\bar{Z}$, $\bar{K}$ and $\bar{P}$, the biological interpretation would be that one HIV particle will generate $\bar{R}_0$ secondary particles into the host. At initial HIV infection with time $t = 0$, the virus reproduction number has a value of 10 and is represented by the following equation:

$$R_0 = \frac{\beta\kappa_0 S_0}{(\mu_Y \mu_V)(P_0 + d)}. \tag{33}$$

This is the initial reproduction number with no anti-HIV activity. A reproduction number, which represents the presence of a fully activated anti-HIV activity with a

Table 3 Parameter values used in the extended model

Parameter		Values
α	CD4+ T cell production rate	5×10^9 per day
μ	Natural death rate of uninfected cells	0.02 per day
κ_0	Initial rate at which a HIV particle transforms a susceptible CD4+ T cell to a productively infected cell	1.0×10^{-12} particles per day
μ_Y	Death rate of productively infected cells	0.6 per day
δ_Y	Maximum additional elimination rate of productively infected cell through anti-HIV activity	0.6 per day
β	HIV production rate from infected cells	150 particles per cell per day
μ_V	Clearance rate of infectious virus particles	6 per day
δ_V	Maximum additional elimination rate of virus particles through the anti-HIV activity	5 per day
θ	HIV dependent immune activation rate	10^{-6}
ρ	Autonomous immune activation rate	0.1 per day
ω_P	Rate of increase of the fraction of susceptible cells by generation and selection of HIV mutants	1.4×10^{-14} particles per day
ω_K	Rate of increase of reproduction per virus particle	1.1×10^{-15}
a	Constant	10^3
b	Constant	0.2
c	Constant	2.0
d	Constant	10^{-2}
$Z_{\max}$	Maximum ant-HIV activity	1.0
$P_{\max}$	Maximum fraction of susceptible cells	1.0
$K_{\max}$	Maximum infection rate of susceptible cells per infected cell	20

maximum number of susceptible cells can also be found. The reproduction number with maximum anti-HIV activity is represented by the following equation:

$$R' = \frac{\beta \kappa_0 S_0}{(\mu_Y + \delta_Y)(\mu_V + \delta_V)(P_0 + d)}. \tag{34}$$

In this equation, Z and K are held at fixed values, $Z = Z_{\max} = 1$ and $K = 1$. If R' is greater tha one, the infection will persist and cannot be cured. The calculated value of R' is 2.75. This value confirms that a patient with HIV will not be able to overcome the infection.

The HIV extended model is very complex and a full mathematical analysis is not possible. However, this model is also more realistic and applicable because it takes into account the difference between susceptible and nonsusceptible CD4+ T cells. Modeling with specific parameters will help explaining the system better. Most of the parameters used were found through clinical and experimental data (see [21]). The parameter values are described in Table 3.

3.3 HIV Extended Model Graphs and Biological Interpretation

The numerical results of the model using the parameter values from Sect. 3.2 were used to make the following graphs. Figure 1 represents the number of CD4+ T cells. Figure 2 represents the number of HIV particles. Figure 3 represents the anti-HIV activity. Each of the graphs represents the initial phase of the HIV infection within the first 6 months and supports the model predictions.

During primary infection, there are a large number of virus particles which enter the body and start infecting the CD4+ T cells. At the start of the infection, the number of HIV particles grows exponentially. The HIV viremia causes a temporary reduction of CD4+ T cells which then recover and remain at a lower level than before the infection. Notice in Figs. 1 and 2 the increase and decrease of HIV particles and CD4+ T cells occur at the same time around 15 days. Right after the initial infection, the anti-HIV activity mounts an attack against the invading virus particles and we see a resurgence of CD4+ T cells. The anti-HIV activity increases rapidly and then reaches its maximum. The anti-HIV activity is not the only reason the viremia starts to break down. Note there are only a certain number of available CD4+ T cells to infect.

During the second phase of the infection for about 10 years, the virus is slightly suppressed and increases slowly. This is the latent period of the infection. The model shows the immune system will hold to 50 % of the normal value for about 10 years but will drop significantly during the two years following. After about 12 years, the CD4+ T cells will drop below 20 % which is the definition of disease progression to AIDS (see [21]). We see a decline of anti-HIV activity. The HIV virus particles replicate freely and reach a higher concentration than that of the primary infection. At this point, the immune system cannot control other infections (Figs. 4, 5 and 6).

Surprisingly, the model predicts that the initial dose of HIV particles introduced into the host does not play an important role in progression to disease. A highly activated CD4+ T cell pool is one of the main determinants for infection and disease progression. If an individual is unhealthy, their CD4+ T cell pool would be larger than normal and would favor CD4+ T cell infection by the HIV virus. If an individual has an initial value of 1,200 per mm^3 CD4+ T cells, then the progression to disease

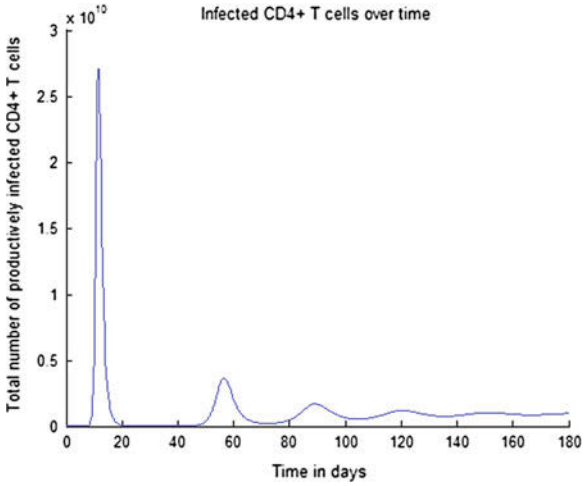


Fig. 1 Decline of CD4+ T cells over first 6 months after initial infection

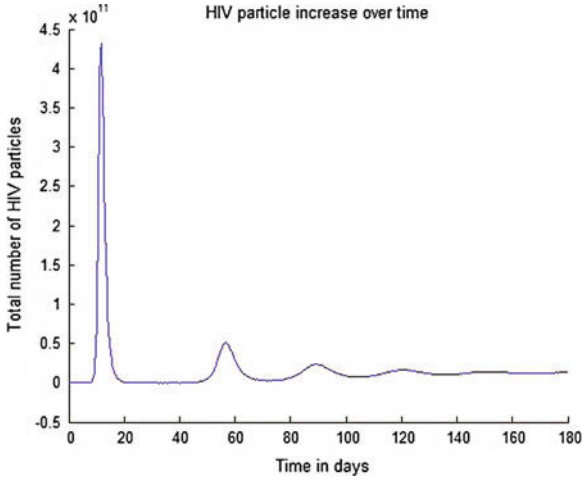


Fig. 2 Increase in HIV particles within the first 6 months of infection

occurs much faster. If the initial value was 800 per mm^3 CD4+ T cells a much smoother progression occurs. The following graph shows the impact of initial cell count on the infection process (Fig. 7).

This model also looks at the dynamics of the susceptible and nonsusceptible cells. The variable P, in the model represents the proportion of new CD4+ T cells which are becoming susceptible. The higher the amount of activated CD4+ T cells, the faster the virus progresses to disease. The initial value of P is important to the dynamics of this model. If the initial value of P is small, the immune system will hold at

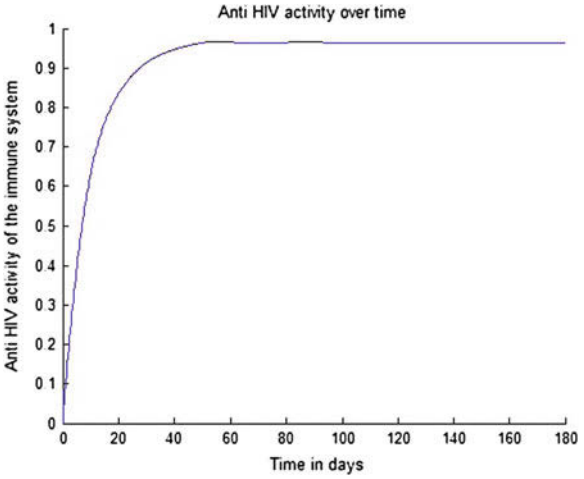


Fig. 3 Decline of anti-HIV activity within the first 6 months of initial infection

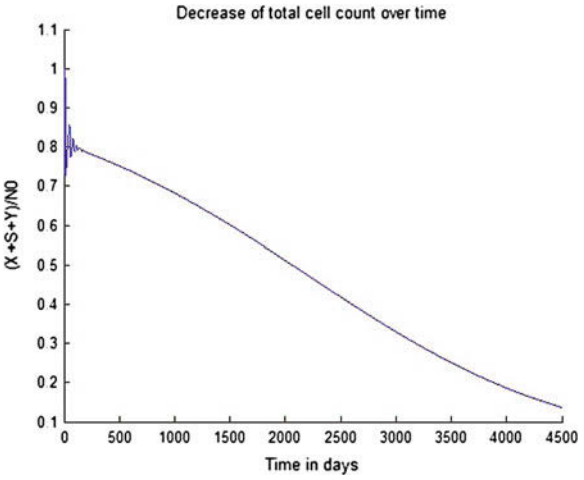


Fig. 4 Total cell count after 12 years. Progression to AIDS occurs at $y = 0.2$

50 % for about 12 years. However, if the initial value of P is large, the progression to disease is much faster. This means through the generation and selection of HIV mutants, the HIV virus will increase the range of CD4+ T cells tropism over more and more CD4+ T cell clones, until after 12 years almost all of the clones are equally susceptible to be infected by the HIV virus (see [21]). The variable K represents the infection rate at which the CD4+ T cells increase by the generation and selection of HIV mutants (Fig. 8).

The speed at which P and K change are measured by ω_P and ω_K. These values also play an important role in the model and in the disease progression. If the value of

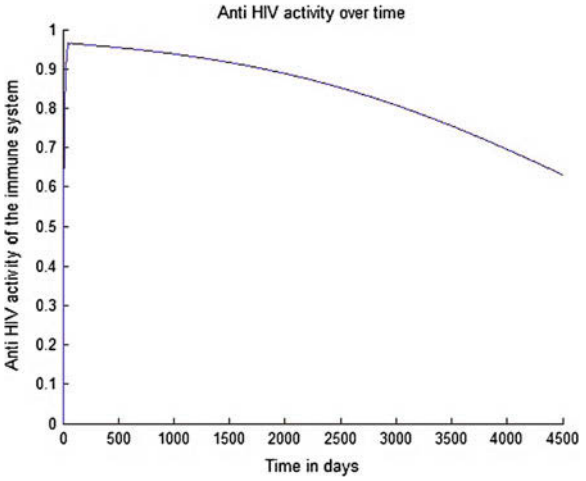


Fig. 5 Anti-HIV activity after 12 years

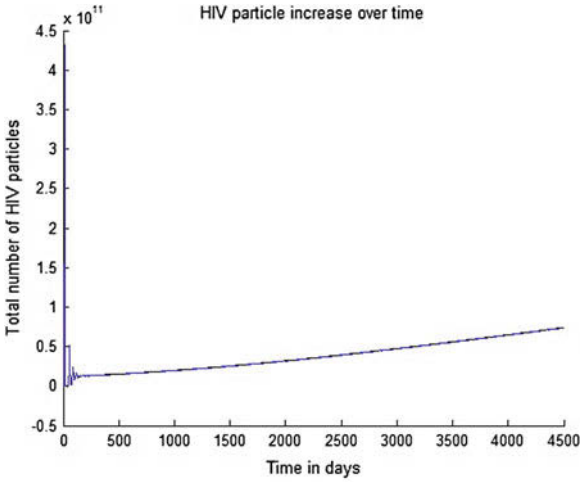


Fig. 6 HIV particle increase over 12 years

ω_K was increased or decreased by a factor of five, the reduction rate of the CD4+ T cells would look similar but the end result may be different. If ω_K was decreased by a factor of five, the model predicts the individual's life span would increase by two years. If ω_K was increased by a factor of five, the model predicts a faster progression to disease around 8 years. There is a stark difference when the value of ω_P is changed by a factor of five. If ω_P is increased by the factor, progression to disease occurs after 6 years (Figs. 9 and 10).

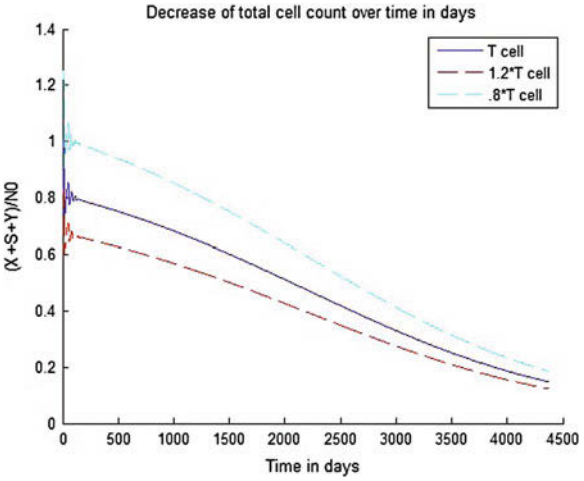


Fig. 7 Total CD4+ T cell count after a 20% reduction of CD4+ T cell count (*aqua*), normal reduction (*blue*), and initial increase by 20% of CD4+ T cell count (*red*)

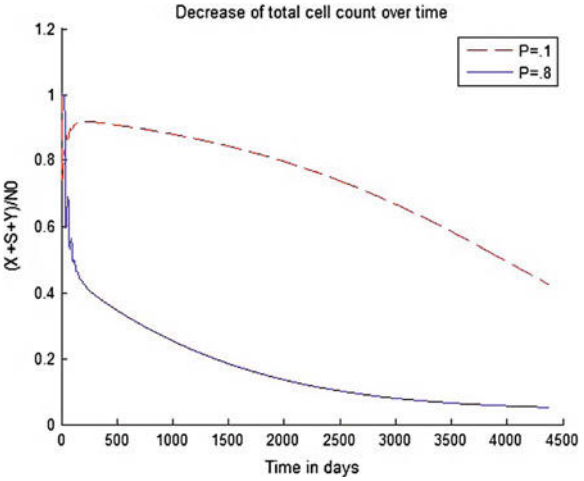


Fig. 8 Changes in the P value and the impact on the total cell count

As seen through the graphs, P, K, ω_P and ω_K are very important to the intrahost dynamics of HIV. The effect of the rate of the fraction of susceptible cells by generation and selection of HIV mutants is important in determining the progression to disease (see [21]).

HIV will affect many people in many different ways. Studies on various aspects of this disease are ongoing. Some recent articles of interest are [8, 10, 13].

This model helps to predict the course that HIV will take during the three stages of the disease. The model may lose applicability for the late part of the last stage of the

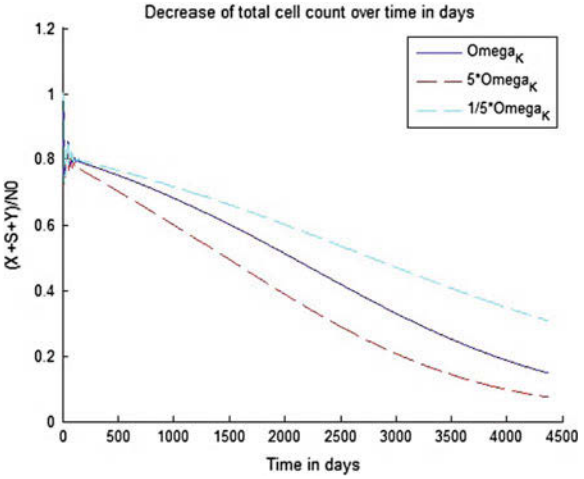


Fig. 9 Impact on CD4+ T cells when the value of ω_K is varied

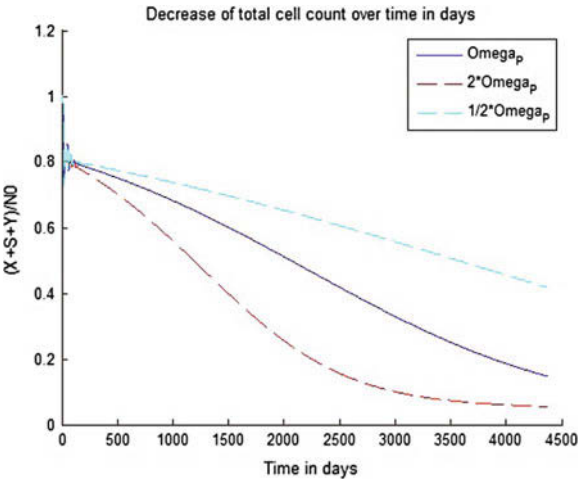


Fig. 10 Impact on CD4+ T cells when the value of ω_P is varied

disease because of many other extreme pathological conditions. In the latter stage of the disease, the immune system is completely compromised and is no longer able to fight infection. When this happens, a simple cold could cause death. Understanding the course of this disease through the model presented ca help doctors and scientist find a cure for this epidemic.

4 HTLV-I Virus and Adult T Cell Leukemia

As we have seen in the previous section, the HIV infection takes place through cell to cell contact with infected CD4+ T cells and eventually takes over the immune system. A virus that is similar and related to HIV is the first form of a human T-lymphotropic virus or HTLV. Just as HIV can lead to the AIDS virus, HTLV-I can lead to many diseases, including adult T cell leukemia/lymphoma. Actively infected T cells can infect other T cells and can eventually convert to ATL cells. This process typically happens during the latent phase of the virus.

HTLV-I shares many similarities with HIV except in the range of diseases that it causes and how it causes these diseases. There are two major virologic differences between HIV and HTLV-I. One difference is that HTLV-I does not destroy the CD4+ T cells but in fact, causes cell proliferation and transformation. The other is that HTLV-I has a low replication rate but a high fidelity of replication, which results in a low viral burden and high genetic stability. This reduces the possibility of immune escape (see [17]). HTLV-I is an enveloped double stranded RNA retrovirus which attacks the CD4+ T cells. Transmission of HTLV-I is mainly associated with the cells. The cells receive this virus through a glucose transporter called glut-1. Once received, the virus inserts a DNA copy into the host cell. The virus replicates with each mitotic cell division. As cells continue to divide, the virus spreads. HTLV-I will remain latent for many years before the virus causes Adult T cell leukemia to manifest. The latently infected cells contain the virus but do not produce DNA. Therefore, the cells are incapable of contagion. This section examines a mathematical model which examines the process of how HTLV-I causes ATL.

Adult T cell leukemia or lymphoma is a non-Hodgkins lymphoma. Adult T cell leukemia occurs first, which is a cancer of the cells. Lymphona also occurs and is a cancer which attacks the B lymphocytes and the lymphatic system. There are four distinct clinical forms of ATL. The disease can be classified as acute ATL, chronic ATL, lymphoma, and smoldering ATL. Once ATL develops, most individuals will survive for only a year or two (see [22]). The median survival rate for the acute and lymphoma subtypes is less than 1 year. Individuals with acute or smoldering ATL may survive longer (see [9]). Standard chemotherapy is not effective against ATL.

4.1 Mathematical Model of HTLV-I Infection to ATL

Stilianakis and Seydel produced a basic mathematical model that describes the T cell dynamics of the HTLV-I infection and the development of ATL (Table 4).

This model consists of the following nonlinear differential equations:

$$T' = \Lambda - \mu_T T - \kappa T_A T \tag{35}$$

$$T_L' = \kappa T_A T - (\mu_L + \alpha) T_L \tag{36}$$

Table 4 Variables used in the Stilianakis and Sydel model

Variable	
T	Number of susceptible CD4+ T cells
T_L	Number of latently infected CD4+ T cells
T_A	Number of actively infected CD4+ T cells
T_M	Number of leukemia T cells
Λ	Constant rate at which new CD4+ T cells are produced (assumed to be susceptible)
κ	Rate at which CD4+ T cells come into contact with actively infected cells.
α	Transmission rate in which latent cells become actively infected cells
ρ	Transmission rate in which actively infected cells convert to ATL cells
β	ATL proliferation rate of a classical logistic growth model
μ_T	Removal or death rate of susceptible CD4+ T cells
μ_L	Removal or death rate of latently infected CD4+ T cells
μ_A	Removal or death rate of actively infected CD4+ T cells

$$T_A' = \alpha T_L - (\mu_A + \rho)T_A \tag{37}$$

$$T_M' = \rho T_A + \beta T_M\left(1 - \frac{T_M}{T_{M_{\max}}}\right) - \mu_M T_M. \tag{38}$$

The terms in this model each have a biological meaning. The first term in Eq. (35) is Λ. This term is the rate in which the new CD4+ T cells are produced. Each cell that is produced is assumed to be susceptible to the virus. The second term in Eq. (35) represents the rate at which all CD4+ T cells die. The last term in Eq. (35) is $\kappa T_A T$ and is considered the mass action term. This term represents the infection process of susceptible cells which come into contact with actively infected CD4+ T cells.

Equation (36) starts with the same mass action term that is seen in Eq. (35). The second term is $(\mu_L + \alpha)T_L$. Let us break this term up into two terms, $\mu_L T_L$ and αT_L and explain them separately. The term $\mu_L T_L$ describes how fast the latently infected cells are dying, and αT_L describes how fast the latently infected cells become actively infected cells. In general, the whole term describes the dynamics of the latently infected cells.

The first term in Eq. (37) is αT_L, which represents how fast the latently infected cells become actively infected cells. The next term is $(\mu_A + \rho)T_A$. Again, lets break this up into two terms, $\mu_A T_A$ and ρT_A. The term $\mu_A T_A$ describes the death rate of the actively infected cell, and ρT_A describes how fast the actively infected cells become ATL cells. The terms in Eq. (37) represent the dynamics of the actively infected cells and how they change to ATL cells.

Equation (38) is the equation which represents the growth of the leukemia cells, which follows the classical logistical growth function. This equation begins with ρT_A. This term was also seen in Eq. (37) and it describes the speed at which actively infected cells become ATL cells. The second term is $\beta T_M(1 - \frac{T_M}{T_{M_{\max}}})$. This term describes the growth of the ATL cells, where β is the speed for which the saturation

level for leukemia cells is reached and $T_{M_{\max}}$ is the maximum amount of ATL cells that can be attained. The last-term $\mu_M T_M$, describes the death rate of the ATL cells. This equation illustrates the dynamics of the ATL or leukemia cells in the body. The virus reproduction number for this model is

$$R_0 = \frac{\alpha\kappa T_0}{(\mu_L + \alpha)(\mu_A + \rho)}. \tag{39}$$

This number helps to determine how fast the disease will spread throughout the body. R_0 represents the number of secondary infections caused by one primary infected cell introduced into the pool of susceptible CD4+ T cells during the infection period (see [20]). If $R_0 > 1$, a chronic infection is seen. This is typical in most HTLV-I infections. If $R_0 \leq 1$, the virus cannot reproduce enough to sustain an infection. The reproduction number will play an important role in determining the stability of the system.

4.2 Stability of the System

To analyze the stability of this system, we must first find the equilibrium points. In order to find the equilibrium points, we set Eqs. (35)–(38) equal to 0 and solve them. The system has two possible solutions or steady states. This system can have an uninfected steady state and a positively infected steady state. For the uninfected steady state, the T cell population will have the following value:

$$T_0 = \frac{\Lambda}{\mu_T}. \tag{40}$$

The initial conditions would then be $T(0) = T_0$, $T_L(0) = 0$, $T_A(0) = 0$, and $T_M(0) = 0$. Therefore the uninfected steady state would be $E_0 = (T_0, 0, 0, 0)$. The positive infected steady state would be $\bar{E} = (\bar{T}, \bar{T}_L, \bar{T}_A, \bar{T}_M)$, where

$$\begin{aligned}
\bar{T} &= \frac{(\mu_L + \alpha)(\mu_A + \rho)}{\alpha\kappa} \\
\bar{T}_L &= \frac{\Lambda\alpha\kappa - \mu_T(\mu_L + \alpha)(\mu_A + \rho)}{\alpha\kappa(\mu_L + \alpha)} \\
\bar{T}_A &= \frac{\Lambda\alpha\kappa - \mu_T(\mu_L + \alpha)(\mu_A + \rho)}{\kappa(\mu_L + \alpha)(\mu_A + \rho)} \\
\bar{T}_M^2 &- \frac{(\beta - \mu_M)T_{M_{\max}}}{\beta}\bar{T}_M - \frac{\rho\bar{T}_A T_{M_{\max}}}{\beta} = 0.
\end{aligned}$$

First, we examine the stability of the uninfected steady state. For this state, the values yield the following Jacobian matrix associated with Eqs. (35)–(38):

$$J = \begin{pmatrix} -\mu_T - \kappa T_A & 0 & -\kappa T & 0 \\ \kappa T_A & -\alpha - \mu_L & \kappa T & 0 \\ 0 & \alpha & -\mu_A - \rho & 0 \\ 0 & 0 & \rho & \beta\left(1 - 2\frac{T_M}{T_{M\max}}\right) - \mu_M \end{pmatrix}. \tag{41}$$

In the uninfected steady state, the characteristic polynomial is found by taking the determinant of the Jacobian or $\det(J - \lambda I)$. The characteristic polynomial is

$$p(\lambda) = (\beta - \mu_M - \lambda)(\mu_T - \lambda)(\lambda^2 + \lambda(\mu_L + \alpha + \mu_A + \rho) + (\mu_L + \alpha)(\mu_A + \rho) - \alpha\kappa. \tag{42}$$

The eigenvalues are

$$\begin{aligned} \lambda_1 &= \beta - \mu_M \\ \lambda_2 &= -\mu_T \\ \lambda_{3,4} &= \frac{-(\mu_L+\alpha+\mu_A+\rho)}{2} \pm \frac{\sqrt{(\mu_L+\mu_A+\alpha+\rho)^2 - 4(\mu_L+\alpha)(\mu_A+\rho) - \alpha\kappa\frac{\Lambda}{\mu_T}}}{2}. \end{aligned} \tag{43}$$

The eigenvalues help in determining the stability of the steady state. If $\lambda_1 = \beta - \mu_M > 0$, then the proliferation rate of the abnormal cells are greater than the death rate and the infection increases. If $\lambda_1 = \beta - \mu_M < 0$, then the death rate of the ATL cells is greater than the proliferation rate and the stability will actually depend on the other eigenvalues λ_3, λ_4. These eigenvalues are either real or complex conjugates. In both cases, the real parts are negative if and only if the reproduction number is less than or equal to one, that is

$$R_0 = \frac{\alpha\kappa T_0}{(\mu_L + \alpha)(\mu_A + \rho)} \leq 1. \tag{44}$$

If we assume that the ATL cells grow at an uncontrollable rate, then $\lambda_1 > 0$ and the point $E_0 = (T_0, 0, 0, 0)$, where $T_0 = \frac{\Lambda}{\mu_T}$, is an unstable saddle point. If $\lambda_1 < 0$, the reproduction number will determine the next steady state. If $R_0 \leq 1$ the uninfected steady state is the only state and it is stable. The system will move to the endemically infected steady state when $R_0 > 1$ and this represents a chronic infection. When this occurs, E_0 will become unstable and $\bar{E}$ will exist.

For the endemically infected steady state, the Jacobian and the determinant of Eqs. (35)–(38) will give the following characteristic equation:

$$\lambda^3 + \lambda^2 A + \lambda B + C = 0, \tag{45}$$

where

$$\begin{aligned} A &= \mu_T + \mu_L + \mu_A + \rho + \alpha + \kappa \bar{T}_A \\ B &= \mu_T \mu_L + \alpha \mu_T + \mu_T \rho + \kappa \mu_L \bar{T}_A + \kappa \mu_A \bar{T}_A + \kappa \rho \bar{T}_A + \alpha \kappa \bar{T}_A \\ C &= \kappa \mu_L \mu_A \bar{T}_A + \kappa \alpha \mu_A \bar{T}_A + \kappa \rho \mu_L \bar{T}_A + \alpha \kappa \rho \bar{T}_A. \end{aligned} \tag{46}$$

We must use the Routh–Hurwitz condition in order to determine the stability of the system. Note that $A > 0$, $B > 0$, and $C > 0$. By the Routh–Hurwitz condition, the eigenvalues of Eq. (45) will have negative real parts if and only if $A > 0$, $C > 0$ and $AB - C > 0$. We have already noted that $A > 0$ and $C > 0$. One can also see that $AB - C > 0$. Therefore, we can determine that the eigenvalues are always negative. When the eigenvalues are negative, we can show that steady state is stable and the infection is chronic.

4.3 Katri and Ruan Model and the Stability of the System

In 2004, Katri and Ruan developed a similar model which takes into account the difference between contact with the virus and infection by the virus. This is denoted by using κ_1 in certain equations. Remember that κ represents the rate at which uninfected cells are contacted by actively infected cells. In this model, κ_1 represents the rate of infection of the T cells by the actively infected T cells. The equations for the Katri and Ruan model are the same as the original model but in Eq. (36), κ is replaced with κ_1 and the new model is

$$T' = \Lambda - \mu_T T - \kappa T_A T \tag{47}$$

$$T_L' = \kappa_1 T_A T - (\mu_L + \alpha) T_L \tag{48}$$

$$T_A' = \alpha T_L - (\mu_A + \rho) T_A \tag{49}$$

$$T_M' = \rho T_A + \beta T_M \left(1 - \frac{T_M}{T_{M_{\max}}}\right) - \mu_M T_M. \tag{50}$$

This small change in the model changes the reproduction number

$$R_0 = \frac{\alpha \kappa_1 T_0}{(\mu_L + \alpha)(\mu_A + \rho)}. \tag{51}$$

The uninfected steady state and stability analysis remains the same as the Stilianakis and Seydel model; however, the new positive infected steady state would be $\bar{E} = (\bar{T}, \bar{T}_L, \bar{T}_A, \bar{T}_M)$, where

$$\bar{T} = \frac{(\mu_L + \alpha)(\mu_A + \rho)}{\alpha\kappa_1}$$
$$\bar{T}_L = \frac{\Lambda\alpha\kappa_1 - \mu_T(\mu_L + \alpha)(\mu_A + \rho)}{\alpha\kappa(\mu_L + \alpha)}$$
$$\bar{T}_A = \frac{\Lambda\alpha\kappa_1 - \mu_T(\mu_L + \alpha)(\mu_A + \rho)}{\kappa(\mu_L + \alpha)(\mu_A + \rho)}$$
$$\bar{T}_M^2 - \frac{(\beta - \mu_M)T_{M_{\max}}}{\beta}\bar{T}_M - \frac{\rho\bar{T}_A T_{M_{\max}}}{\beta} = 0.$$

For this state, the values yield the following Jacobian matrix associated with Eqs. (47)–(50)

$$J = \begin{pmatrix} -\mu_T - \kappa\bar{T}_A & 0 & -\kappa\bar{T} & 0 \\ \kappa_1\bar{T}_A & -\alpha - \mu_L & \kappa_1\bar{T} & 0 \\ 0 & \alpha & -\mu_A - \rho & 0 \\ 0 & 0 & \rho & \beta\left(1 - 2\frac{\bar{T}_M}{T_{M_{\max}}}\right) - \mu_M \end{pmatrix}. \tag{52}$$

We will denote

$$M' = \beta\left(1 - 2\frac{T_M}{T_{M_{\max}}}\right) - \mu_M. \tag{53}$$

The eigenvalues of this Jacobianare M' will always be negative since $\bar{T}_M > T_{M_{\max}}$ when the infection is chronic. The Jacobian will yield the following characteristic equation:

$$\lambda^3 + \lambda^2 a_1 + \lambda(a_2 + a_4) + (a_3 + a_5) = 0, \tag{54}$$

where

$$\begin{aligned} a_1 &= \kappa^2\bar{T}_A + \kappa\mu_L + \kappa\rho + \kappa\mu_T + \alpha\kappa + \kappa\mu_A \\ a_2 &= \kappa^2\bar{T}_A\mu_L + \kappa^2\bar{T}_A\alpha + \mu_T\kappa\mu_L + \kappa^2\bar{T}_A\mu_A + \kappa^2\bar{T}_A\rho \\ &\quad + \mu_A\kappa\mu_L + \mu_T\kappa\alpha + \mu_A\kappa\mu_T + \kappa\mu_L\rho + \kappa\alpha\mu_A + \kappa\alpha\rho + \mu_T\kappa\rho \\ a_3 &= \mu_T\kappa\alpha\rho + \kappa^2\bar{T}_A\alpha\rho + \kappa^2\bar{T}_A\mu_A\alpha + \mu_T\mu_A\kappa\mu_L \\ &\quad + \mu_T\kappa\mu_L\rho + \mu_T\kappa\alpha\mu_A + \kappa^2\bar{T}_A(\mu_L\rho + \mu_L\mu_A) \\ a_4 &= -\kappa_1\alpha\rho - \kappa_1\mu_L\rho - \kappa_1\alpha\mu_A - \kappa_1\mu_L\mu_A \\ a_5 &= -(\mu_T\kappa_1\alpha\rho + \mu_T\mu_A\kappa_1\mu_L + \mu_T\kappa_1\mu_L\rho + \mu_T\kappa_1\alpha\mu_A). \end{aligned}$$

Again, we must use the Routh–Hurwitz condition to further determine the stability of the system. According to the Routh–Hurwitz condition, the eigenvalues will have negative real parts if and only if

$$a_1 > 0,\ (a_3 + a_5) > 0 \quad \text{and} \quad a_1(a_2 + a_4) - (a_3 + a_5) > 0. \tag{55}$$

Table 5 Variables and parameter values for contagion used in the model

Parameter		Values
T	Uninfected CD4+ T cell population size	1,000/mm^3
T_L	Latently infected CD4+ T cell density	250/mm^3
T_A	Actively infected CD4+ T cell density	1.5/mm^3
T_M	Leukemic CD4+ T cell density	0
μ_T	Natural death rate of CD4+ T cells	0.6 mm^3 per day
μ_L	Blanket death rate of latently infected CD4+ T cells	0.006 per day
μ_A	Blanket death rate of actively infected cells	0.05 per day
μ_M	Death rate of leukemic cells	0.0005 per day
κ_1	Rate uninfected CD4+ T cells become latently infected	varies
κ	Rate infected cells are contacted	varies
β	Growth rate of leukemic CD4+ T cell population	0.0003 per day
α	Rate latently infected cells become actively infected	0.0004 per day
ρ	Rate actively infected cells become leukemic	0.00004 per day
$T_{M_{\max}}$	Maximal population level of leukemic CD4+ T cells	2,200/mm^3
λ	Source term for uninfected CD4+ T cells	6 per day
T_0	Derived quantity which represents the CD4+ T cell population for HTLV-I negative persons	1,000/mm^3

Proposition 41 (see [6]) *The infected steady state $\bar{E}$ is asympototically stable if $R_0 > 1$ and the inequalities in (55) are satisfied. This occurs if (a) $\kappa > \kappa_1$, or (b) $\kappa = \kappa_1$.*

To check that this proposition is valid, we will use the following parameters and values estimated by Stilianakis and Seydel (see [20]).

We can use these parameter values and the estimated values of κ and κ_1, given by Stilianakis and Seydel on the basis of parameter values from the HIV infection process, to find that $R_0 = 1.25$ if $\kappa_1 = 0.1$. If we take $\kappa = \kappa_1 = 0.1$, we find that the inequalities in Eq. (55) are satisfied and part (b) of Proposition 41 is true. Furthermore, the steady state, $\bar{E} = (800, 187.5, 1.5, 1.3)$, would be asymptotically stable. If we take $\kappa = 0.5$ and $\kappa_1 = 0.1$, we will again find that the inequalities in Eq. (55) are satisfied and part (a) of Proposition 41 is also true. The steady state would be $\bar{E} = (800, 37.38, 0.3, 0.6)$, which is also asymptotically stable. The following graph was created using the parameter values given in Table 5. The numerical simulation shows the number of healthy CD4+ T cells decreases dramatically while the latently infected cells increase, and then remain steady (Fig. 11).

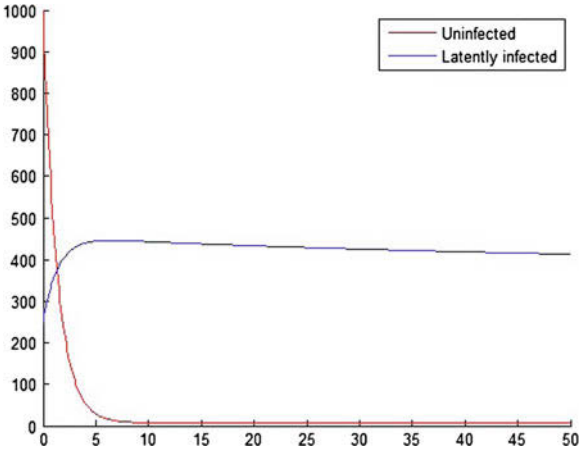


Fig. 11 Latently infected CD4+ T cells versus Uninfected CD4+ T cells

HTLV-I is a virus in which only 5 % of infected individuals will ever develop any disease such as Adult T cell Leukemia. We have shown through stability analysis that we can predict when the infection will persist and become ATL.

5 Conclusion

The interaction between HIV and the immune system is a dynamic process. Mathematical models are used to understand this dynamism to ascertain which biological mechanisms cause disease progression. Although at the moment there is no definite cure or vaccine for HIV, the treatment regimes used by physicians are able to extend the lives of HIV patients. Researchers are also working to understand the pathogen, its behavior, and transmission capability to find a vaccine. The future work should be focused on finding the optimal treatment schedule in order to prolong the life of patients and hopefully, find a permanent cure.

It is unclear why some HTLV-I carriers progress to disease while the majority of them do not. It is also not known why some infected individuals develop ATL and others develop HAM/TSP (see [9]). Further studies should focus on finding the mechanism that causes the virus to progress to disease and finding the genetic markers that will determine which disease the virus will trigger. HTLV-I also has no known cure or vaccine. However, a vaccine to prevent infection is currently being explored.

Graphs obtained from the model help us to predict what factors are important for HIV to progress to AIDS. In Sect. 4, we have mentioned two different models of HTLV-I progression to ATL. Our objective is to use SIMULINK to study the mathematical models with the hope to refine them in future as new information

becomes available to predict the course of the infection. The model can be improved by adding other factors which influence the disease and study the relative sensitivity of different factors.

References

1. Baxley, D.: A mathematical study of two retroviruses, HIV and HTLV-I. MS Thesis, University of Central Florida, Orlando Fl (2007)
2. Diekmann, O., Heesterbeek, J.A.P.: Mathematical Epidemiology of Infectious Diseases: Model Building Analysis and Interpretation. Wiley, New York (2002)
3. Diekmann, O., Heesterbeek, J.A.P., Metz, J.A.J.: On the definition and computation of the basic reproductive ratio in models for infectious diseases in heterogeneous population. J. Math. Biol. **28**, 365–382 (1990)
4. Freedman, A.S., Harris, N.L.: Clinical and pathologic features of adult T cell lymphoma/leukemia. UpToDate. http://www.utdol.com/utd/content/topic.do?topicKey=lymphoma/13762 (2007)
5. Kartikeyan, S., Bharmal, R.N., Tiwari, R.P., Bisen, P.S.: HIV and AIDS: Basic Elements and Priorities. Springer, The Netherlands (2007)
6. Katri, P., Ruan, S.: Dynamics of human T-cell lymphotropic virus I (HTLV-I) infection of CD4+ T cells. C. R. Biol. **327**(11), 1009–1016 (2004)
7. Kirschner, D.: Using Mathematics to understand HIV immune dynamics. Not. AMS **43**(2), 191–202 (1996)
8. Kitayimbwa, J.M., Mugisha, J.Y.T., Saenz, R.A.: The role of backward mutations on the within-host dynamics of HIV-1. J. Math. Biol. **67**(5), 1111–1139 (2013)
9. Manns, A., Hisada, M., Grenade, L.: Human T-lymphotropic virus type I infection. Lancet **353**, 1951–1958 (1999)
10. Mao, Y., Wang, L., Gu, C., Herschhorn, A., Dsormeaux, A., Xiang, S. H., Sodroski, J. G.: Molecular architecture of the uncleaved HIV-1 envelope glycoprotein trimer. Proc. Natl.Acad. Sci. USA **110**(30), 12438–12443 (2013)
11. Pantaleo, G., Graziosi, C., Fauci, A.S.: The immunopathogenesis of human immunodeficiency virus infection. N. Engl. J. Med. **328**, 327–335 (1993)
12. Plemmons, W.R.: A mathematical study of Malaria models of Ross and Ngwa. Masters Thesis, University of Central Florida, Orlando Fl (2006)
13. Poppvic, M., Sarangadhara, M.G., Read, E., Gallo, R.C.: Detection, isolation, and continous production of cytopathic retroviruses (HTLV-III) from patients with AIDS and pre-AIDS. Science **224**, 497–500 (1984)
14. Pratt, R.J.: HIV and AIDS: A Foundation for Nursing and Healthcare Practice. Arnold, London (2003)
15. Rodin, E.Y., Murthy, D.N.P., Page, N.W.: Mathematical Modeling: A Tool for Problem Solving in Engineering, Physical, Biological and Social Sciences. Pergamon Press, Oxford (1989)
16. Scadden, D.T., Freedman, A.R., Robertson, P.: Human T-lymphotropic virus type I: disease associations, diagnosis, and treatment. UpToDate. http://www.uptodate.com/contents/human-t-lymphotropic-virus-type-i-disease-associations-diagnosis-and-treatment (2007)
17. Scadden, D.T., Freedman, A.R., Robertson, P.: Human T-lymphotropic virus type I: virology, pathogenesis, and epidemiology. UpToDate. http://www.uptodate.com/contents/human-t-lymphotropic-virus-type-i-virology-pathogenesis-and-epidemiology (2007)
18. Singh, N.: Epidemiological models for mutating pathogen with temporary immunity. Doctoral Thesis, University of Central Florida, Orlando Fl (2006)
19. Stilianakis, N.I., Dietz, K., Schenzle, D.: Analysis of a model for the pathogenesis of AIDS. Math. Biosci. **145**(1), 27–46 (1997)

20. Stilianakis, N.I., Seydel, J.: Modeling the T-cell dynamics and pathogenesis of HTLV-I infection. Bull. Math. Biol. **61**(5), 935–947 (1999)
21. Stilianakis, N.I., Schenzle, D.: On the intra-host dynamics of HIV-1 infections. Math. Biosci. **199**, 1–25 (2006)
22. Wang, L., Li, M.Y., Kirschner, D.: Mathematical analysis of the global dynamics of a model for HTLV-I infection and ATL progression. Math. Biosci. **179**(2), 207–217 (2002)
23. Wessner, D.: HIV and AIDS. Pearson Benjamin/Cummings, San Francisco (2006)