Debasis Giri
Ram N. Mohapatra
Heinrich Begehr
Mohammad S. Obaidat (Eds.)

Communications in Computer and Information Science 655

Mathematics and Computing

Third International Conference, ICMC 2017
Haldia, India, January 17–21, 2017
Proceedings

Springer

Communications in Computer and Information Science 655

Commenced Publication in 2007
Founding and Former Series Editors:
Alfredo Cuzzocrea, Dominik Ślęzak, and Xiaokang Yang

Editorial Board

Simone Diniz Junqueira Barbosa
Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Rio de Janeiro, Brazil
Phoebe Chen
La Trobe University, Melbourne, Australia
Xiaoyong Du
Renmin University of China, Beijing, China
Joaquim Filipe
Polytechnic Institute of Setúbal, Setúbal, Portugal
Orhun Kara
TÜBİTAK BİLGEM and Middle East Technical University, Ankara, Turkey
Igor Kotenko
St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, St. Petersburg, Russia
Ting Liu
Harbin Institute of Technology (HIT), Harbin, China
Krishna M. Sivalingam
Indian Institute of Technology Madras, Chennai, India
Takashi Washio
Osaka University, Osaka, Japan

More information about this series at http://www.springer.com/series/7899

Debasis Giri · Ram N. Mohapatra
Heinrich Begehr · Mohammad S. Obaidat (Eds.)

Mathematics and Computing

Third International Conference, ICMC 2017
Haldia, India, January 17–21, 2017
Proceedings

Springer

Editors
Debasis Giri
Haldia Institute of Technology
Haldia
India

Heinrich Begehr
Freie Universität Berlin
Berlin
Germany

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

Mohammad S. Obaidat
Fordham University
Bronx, NY
USA

ISSN 1865-0929 ISSN 1865-0937 (electronic)
Communications in Computer and Information Science
ISBN 978-981-10-4641-4 ISBN 978-981-10-4642-1 (eBook)
DOI 10.1007/978-981-10-4642-1

Library of Congress Control Number: 2017937713

© Springer Nature Singapore Pte Ltd. 2017
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

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

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 an enormous growth in the applications of mathematics in different areas that are multidisciplinary in nature. Cryptography, security, and signal processing are such areas that are being focused on recently due to the need of securing communication while connecting with others. With emerging computing facilities and speeds, a phenomenal growth has occurred in problem solving area. Earlier, some observations were made and conjectures were drawn which remained conjectures until somebody either could prove it theoretically or find counter examples. Today, however, we can write algorithms and use computers for long calculations, verifications, or generation of huge amount of data. With available computing capabilities, we can find factors of very large integers in the size of hundreds of digits; we can find inverses of very large matrices and solve a large set of linear equations etc. Thus Mathematics and computation have become more integrated areas of research today and it was decided to organize an event where researchers can share ideas and deliberate on new challenging problems.

Apart from many other interdisciplinary areas of research, cryptography and security have emerged as one of the most important areas of research with discrete mathematics as a base. Several research groups are actively pursuing research on different aspects of cryptology not only in terms of new crypto-primitives and algorithms but also numerous concepts related to authentication, integrity and security proofs/protocols are being developed, often with open and competitive evaluation mechanisms to evolve standards.

As conferences, seminars, and workshops are the platforms for sharing knowledge and new research results giving us a chance to get new innovative ideas for future needs as the threats and computational capabilities of adversaries are ever increasing, it was thought appropriate to organize a conference focused on Mathematics and computations covering theoretical as well as practical aspects of research, with cryptography and security being one of these.

Eminent personalities working in Mathematical and computer science and related areas were invited to deliver invited talks and tutorials. The talks covered a wide spectrum, namely, number theoretic concepts, cryptography, algebraic concepts such as quasi groups and applications etc. The conference was spread over five days (January 17–21, 2017) with the first two days dedicated to tutorials. The main conference was planned with special talks by experts and paper presentations in each session.

I hope that the conference met the aspirations of the participants and its objective of sharing ideas and current research and identifying new targets/problems. We are confident that the young researchers and students found new directions to pursue in their future research.

We express our heartfelt thanks to the National Board for Higher Mathematics (NBHM), the Indian Space Research Organisation (ISRO), the Science and Engineering Research Board (Department of Science and Technology), the Council of Scientific and Industrial Research (CSIR), Defense Research and Development Organization (DRDO), the International Society for Analysis, Its Applications and Computation (ISAAC), the Indian National Science Academy (INSA), Haldia Institute of Technology (Haldia, India), and the University of Central Florida (USA).

We are also very much thankful to our fellow organizing chair, Prof. Debasis Giri, who is the founder of the International Conference on Mathematics and Computing (ICMC), for his effort to make the event a grand success. We extend our sincere thanks to all speakers, participants, referees, and organizers for their support.

March 2017

P.K. Saxena
P.D. Srivastava

Message from the Program Chairs

It was a great pleasure for us to organize the third International Conference on Mathematics and Computing held during January 17–21, 2017, at the Haldia Institute of Technology, Purba Medinipur, West Bengal, India. Our main goal was to provide an opportunity to the participants to learn about contemporary research in cryptography, security, mathematics, and computing and exchange ideas among themselves and with experts present in the conference as tutorial presenters and the plenary as well as invited speakers. With this aim in mind we carefully selected the invited speakers and the speakers for the tutorials. It is our sincere hope that the conference helped the participants in their research and training by opening new avenues for those who are either starting their research or are looking to extend their area of research to a different field in cryptography, security, mathematics, and computing.

During January 17–18, 2017, there were five tutorial talks by Prof. Dipanwita Roy Chowdhury (IIT, Kharagpur), Prof. Abhijit Das (IIT, Kharagpur), Dr. Avishek Adhikari (Calcutta University), Dr. Manish Kumar (Birla Institute of Technology and Science, Pilani), and Sweta Mishra (Indian Institute of Technology, Delhi).

The conference began after a formal opening ceremony on January 19. The program offered one 75-minute keynote talk by Prof. Mohammed S. Obaidat (Fordham University, USA) and 11 invited one-hour talks by Prof. Sudip Misra (IIT Kharagpur, India), Prof. Subhamoy Maitra (Indian Statistical Institute, Kolkata, India), Prof. Heinrich Begehr (Free University of Berlin, Germany), Prof. Ram N. Mohapatra (University of Central Florida, USA), Prof. S. Ponnusamy (Indian Statistical Institute, Chennai Centre, India), Prof. Maria A. Navascues (Universidad de Zaragoza, Spain), Prof. Margareta Heilmann (University of Wuppertal, Germany), Prof. Rifat Colak (Firat University, Turkey), Prof. Elena Berdysheva (Justus-Liebig-Universität, Giessen, Germany), Prof. W.M. Shah (Institute for Research in Mathematical Sciences, Srinagar, Kashmir, India), and Dr. Manish Kumar (Birla Institute of Technology & Science, Pilani, India). Our speakers/contributors were from Germany, Spain, Turkey, Bangladesh, India, Russia, and USA.

After an initial call for papers, 129 papers were submitted for presentation at the conference. All submitted papers were sent to external reviewers and after refereeing, 35 papers were recommended for publication in the conference proceedings published by Springer in their *Communications in Computer and Information Science* (CCIS) series.

We are grateful to the speakers, participants, reviewers, organizers, sponsors, and funding agencies for their support and help, without which it would have been impossible to organize the conference, the workshops, and the tutorials. We owe our gratitude to the volunteers who worked behind the scenes tirelessly taking care of the details to make this conference a success.

March 2017

Debasis Giri
Ram N. Mohapatra
Heinrich Begehr
Mohammad S. Obaidat

Preface

The Third International Conference on Mathematics and Computing (ICMC 2017) was held at the Haldia Institute of Technology, Haldia, during January 17–21, 2017. Haldia is a city and a municipality in Purba Medinipur in the Indian state of West Bengal, and Haldia Institute of Technology is a premier institution training engineers and computer scientists for the past several years. It has gained its reputation through its institutional dedication to teaching and research.

In response to the call for papers for ICMC 2017, 129 papers were submitted for presentation and inclusion in the proceedings of the conference. The papers were evaluated and ranked on the basis of their significance, novelty, and technical quality by at least two reviewers per paper. After a careful blind refereeing process, 35 papers were selected for inclusion in the conference proceedings. The papers cover current research in cryptography, security, abstract algebra, functional analysis, fluid dynamics, fuzzy modeling and optimization etc. ICMC 2017 had eminent personalities both from India and abroad (USA, Germany, Spain, China, and Turkey), who delivered invited addresses and tutorial talks. The speakers from India are recognized leaders in government, industry, and academic institutions such as the Indian Statistical Institute Kolkata, Indian Statistical Institute Chennai, IIT Kharagpur, Jammu and Kashmir Institute of Mathematical Sciences, Srinagar, Kashmir, Calcutta University, Birla Institute of Technology and Science, Pilani, and Indian Institute of Technology Delhi, etc. All of them are involved in research dealing with the current issues of interest related to the theme of the conference. The conference offered five tutorial talks by Prof. Dipanwita Roy Chowdhury (IIT, Kharagpur), Prof. Abhijit Das (IIT, Kharagpur), Dr. Avishek Adhikari (Calcutta University), Dr. Manish Kumar (Birla Institute of Technology and Science, Pilani), and Sweta Mishra (Indian Institute of Technology, Delhi). In addition to these the program included one keynote talk by Prof. Mohammed S. Obaidat (Fordham University, USA) and 11 invited talks by Prof. Sudip Misra (IIT Kharagpur, India), Prof. Subhamoy Maitra (Indian Statistical Institute, Kolkata, India), Prof. Heinrich Begehr (Free University of Berlin, Germany), Prof. Ram N. Mohapatra (University of Central Florida, USA), Prof. S. Ponnusamy (Indian Statistical Institute, Chennai Centre, India), Prof. Maria A. Navascues (Universidad de Zaragoza, Spain), Prof. Margareta Heilmann (University of Wuppertal, Germany), Prof. Rifat Colak (Firat University, Turkey), Prof. Elena Berdysheva (Justus-Liebig-Universität, Giessen, German), Prof. W.M. Shah (Institute for Research in Mathematical Sciences, Srinagar, Kashmir, India), and Dr. Manish Kumar (Birla Institute of Technology and Science, Pilani, India).

A conference of this kind would not be possible to organize without the full support from different people across different committees. All logistics and general organizational aspects were looked after by the Organizing Committee members, who spent their time and energy in making the conference a reality. We also thank all the Technical Program Committee members and external reviewers for thoroughly

reviewing the papers submitted for the conference and sending their constructive suggestions within the deadlines. Our hearty thanks to Springer for agreeing to publish the proceedings in its *Communications in Computer and Information Science* (CCIS) series.

We are indebted to the National Board for Higher Mathematics (NBHM), the Indian Space Research Organisation (ISRO), the Science and Engineering Research Board (Department of Science and Technology), the Council of Scientific and Industrial Research (CSIR), the Defense Research and Development Organization (DRDO), the International Society for Analysis, Its Applications and Computation (ISAAC), the Indian National Science Academy (INSA), Haldia Institute of Technology (Haldia, India), and the University of Central Florida (USA) for sponsoring the event. Their support has significantly helped raise the profile of the conference.

Last but not the least, our sincere thanks go to all authors who submitted papers to ICMC 2017 and to all speakers and participants. We sincerely hope that the readers will find the proceedings stimulating and inspiring.

March 2017

Debasis Giri
Ram N. Mohapatra
Heinrich Begehr
Mohammad S. Obaidat

Organization

Patron

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

General Co-chairs

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

Program Co-chairs

Heinrich Begehr	Free University of Berlin, Germany
Debasis Giri	Haldia Institute of Technology, India
M.S. Obaidat	Fordham University, USA
R.N. Mohapatra	University of Central Florida, USA

Tutorial Co-chairs

Debiao He	Wuhan University, Wuhan, China
Ekrem Savas	Istanbul Commerce University, Turkey

Organizing Chair

Debasis Giri	Haldia Institute of Technology, India

Technical Program Committee

Computing Track

Jaydeb Bhaumik	Haldia Institute of Technology, India
Christina Boura	Université de Versailles Saint-Quentin-en-Yvelines, France
Rajat Subhra Chakraborty	IIT Kharagpur, India
Shehzad Ashraf Chaudhry	International Islamic University Islamabad, Pakistan
Abhijit Das	IIT Kharagpur, India
Josep Domingo-Ferrer	Universitat Rovira i Virgili Tarragona, Spain
Debin Gao	Singapore Management University, Singapore
Debasis Giri	Haldia Institute of Technology, India
Debiao He	Wuhan University, China
Marko Hölbl	University of Maribor, Slovenia
Hafizul Islam	IIIT Kalyani, India

Qi Jiang	Xidian University, China
Sokratis Katsikas	University of Science and Technology, Norway
Neeraj Kumar	Thapar University, Patiala, India
Saru Kumari	Chaudhary Charan Singh University, India
Manik Lal Das	Dhirubhai Ambani Institute of Information and Communication Technology, India
Cheng-Chi Lee	Fu Jen Catholic University, Taiwan
Fagen Li	University of Electronic Science and Technology, China
J.K. Mandal	Kalyani University, India
Dheerendra Mishra	The LNM Institute of Information Technology, Raipur, India
Lu Leng Nanchang	Hangkong University, Nanchang, China
M.S. Obaidat	Fordham University, USA
Eiji Okamoto	University of Tsukuba, Japan
S.K. Pal	SAG, DRDO, Delhi, India
Gerardo Pelosi	Politecnico di Milano, Italy
Jun Peng	University of Texas - Rio Grande Valley, USA
Kostas Psannis	University of Macedonia, Greece
Indrakshi Ray	Colorado State University, USA
Bimal Roy	ISI Kolkata, India
Dipanwita Roychowdhury	IIT Kharagpur, India
Mohammad Sabzinejad	Farash Kharazmi University, Tehran, Iran
Kouichi Sakurai	Kyushu University, Japan
Somitra Sanadhya	IIIT Delhi, India
Nitesh Saxena	University of Alambama at Birmingham, USA
P.K. Saxena	SAG, DRDO, Delhi, India
Peter Schwabe	Radboud University, The Netherlands
M. Sethumadhavan	Amrita Vishwa Vidyapeetham, Coimbatore, India
C.E. Veni Madhavan	IISC Bangalore, India
Meng Yu	The University of Texas at San Antonio, USA
Sherali Zeadally	University of Kentucky, USA

Mathematics Track

Ravi P. Agarwal	Texas A&M University, Kingsville, USA
M. Amer Qazi	Tuskegee University, Alabama, USA
Muhammad Aslam Noor	COMSATS Institute of Information Technology, Pakistan
Valentina E. Balas Aurel	Vlaicu University of Arad, Romania
Feyzi Basar	Fatih University, Turkey
Heinrich Begehr	Free University of Berlin, Germany
Elena Berdysheva	Justus-Liebig-Universität, Giessen, Germany
Oscar Castillo	Tijuana Institute Technology, Mexico
Rifat Colak	Firat University, Elazig, Turkey
Ashok Kumar Das	IIIT Hyderabad, India

Gennadii Demidenko	Siberian Branch of Russian Academy of Sciences, Novosibirsk, Russia
Dhananjoy Dey	SAG, DRDO, Delhi, India
Sever Dragomir	Victoria University, Melbourne, Australia
Roland Duduchava	Ivane Javakhishvili Tbilisi State University, Georgia
Leopoldo Eduardo	Cárdenas-Barrón Tecnológico de Monterrey, Mexico
Esra Erkus	Duman Gazi University, Ankara, Turkey
Narendra Govil	Auburn University, Alabama, USA
Indiver Gupta	SAG, DRDO, Delhi, India
U.C. Gupta	IIT Kharagpur, India
Geni Gupur	Xinjiang University, Urumqi, China
Margareta Heilmann	University of Wuppertal, Germany
Tian-Xiao Heinrich	Illinois Wesleyan University, Illinois, USA
Don Hong	Middle Tennessee State University, Tennessee, USA
Rajeeva Karandikar	Chennai Mathematical Institute, India
Somesh Kumar	IIT Kharagpur, India
Shijun Liao	Shanghai Jiao Tong University, Shanghai, China
Dilip Kumar Maiti	Vidyasagar University, India
Manoranjan Maiti	Vidyasagar University, India
P.R. Mishra	SAG, DRDO, Delhi, India
Alip Mohammed	The Petroleum Institute, Abu Dhabi, United Arab Emirates
Ram N. Mohapatra	University of Central Florida, USA
Maria A. Navascues	University of Zaragoza, Spain
Ameeya Kumar Nayak	IIT Roorkee, India
Madhumangal Pal	Vidyasagar University, India
G.P. Raja Sekhar	IIT Kharagpur, India
Sujit Kumar Samanta	NIT Raipur, India
Ekrem Savas	Istanbul Commerce University, Turkey
Nita H. Shah	Gujarat University, Navrangpura, Ahmedabad, India
P.L. Sharma	Himachal Pradesh University, Shimla, India
Gopal Chandra Shit	Jadavpur University, Kolkata, India
Zhisheng Shuai	University of Central Florida, USA
P.D. Srivastava	IIT Kharagpur, India
Ram U. Verma	University of North Texas, Denton, USA

Additional Reviewers

Deepmala -	ISI Kolkata, India
Alrazi Abdeljabba	The Petroleum Institute, Abu Dhabi, United Arab Emirates
S. Abhishek Anand	University of Alabama at Birmingham, USA
Carles Anglés	Universitat Rovira i Virgili, Tarragona, Spain
Ali Aral	Kirikkale University, Turkey
Subhabrata Barman	Haldia Institute of Technology, India
B. Bhowmik	IIT Kharagpur, India

Chen Biwen	Wuhan University, China
Alberto Blanco-Justicia	Universitat Rovira i Virgili, Tarragona, Spain
Debjani Chakraborty	IIT Kharagpur, India
Pratish Datta	IIT Kharagpur, India
Basudeb Dhara	Belda College, India
Uday Dixit	IIT Guwahati, India
Manish Kant Dubey	DRDO, Delhi, India
Ratna Dutta	IIT Kharagpur, India
Zoltan Finta	University of Babeş-Bolyai, Romania
Rupanwita Gayen	IIT Kharagpur, India
N. Gnaneshwa	IIT Kharagpur, India
D.K. Gupta	IIT Kharagpur, India
Vijay Gupta	NSIT, Delhi, India
Biswapati Jana	Vidyasagar University, India
Dipak Kumar Jana	Haldia Institute of Technology, India
Rajesh Kannan	IIT Kharagpur, India
Sathya Babu Kora	National Institute of Technology, Rourkela, India
Manish Kumar	BITS, Pilani Hyderabad Campus, Telangana, India
Manoj Kumar	University of Delhi, India
Fagen Li	University of Electronic Science and Technology, China
Xiong Li	Hunan University of Science and Technology, Hunan Xiangtan, China
Chandrashekhar Meshram	R.D. University, Jabalpur, India
Girish Mishra	DRDO, Delhi, India
C. Nahak	IIT Kharagpur, India
Ajaya Neupane	University of Alabama at Birmingham, USA
Petros Nicopolitidis	Aristotle University of Thessaloniki, Greece
Bhawani Sankar Panda	IIT Kharagpur, India
Geetanjali Panda	IIT Kharagpur, India
Pratima Panigrahi	IIT Kharagpur, India
Kailash C. Patidar	University of the Western Cape, South Africa
Bidyut Kumar Patra	National Institute of Technology, Rourkela, India
Wang Qing	Aerodynamics Research and Development Center, Mianyang, China
Kiran Ramesh	University of Glasgow, UK
Ram Ratan	DRDO, Delhi, India
Jordi Ribes-González	Universitat Rovira i Virgili, Tarragona, Spain
Maliheh Shirvanian	University of Alabama at Birmingham, USA
Prakash Shrestha	University of Alabama at Birmingham, USA
Sanasam Ranbir Singh	IIT Guwahati, India
Konstantina Skouri	University of Ioannina, Greece
Binod Chandra Tripathy	Institute of Advanced Study in Science and Technology, India
Abdullahi Umar	The Petroleum Institute, Abu Dhabi, United Arab Emirates
B.B. Upadhyay	NIT Manipur, India

Odelu Vanga	Indian Institute of Information Technology Chittoor, Andhra Pradesh, India
P. Vijayakumar	University College of Engineering Tindivanam, Melpakkam, India
Ravi Vishwakarma	Rani Durgawati Vishwavidyalya, Jabalpur, India
Jing Wang	University of Illinois at Urbana Champaign, USA
Qian Weiqi	Aerodynamics Research and Development Center, Mianyang, China
Dirk Werner	Free University, Berlin, Germany
Fan Wu	Xiamen Institute of Technology, Huaqiao University, Xiamen, China
Xing Xie	Colorado State University, USA
Zhiyan Xu	Wuhan University, China
Xie Yong	Wuhan University, China
Yubo Zhang	Wuhan University, China

Local Organizing Committee

Asish Lahiri, Sayantan Seth, M.N. Bandyopadhyay, Anjan Mishra, Debasis Giri (Organizing Chair), Jaydeb Bhaumik, Sk. Sahnawaj, Soumen Paul, Sudipta Kumar Basu, Debasis Das, Apratim Mitra, Subhankar Joardar, Subhabrata Barman, Sourav Mandal, Susmit Maity, Tarun Kumar Ghosh, Sk. Arif Ahmed, Dipak Kumar Jana, Palash Ray, Mihir Baran Bera and Jayeeta Majumder.
All Organizing Committee members are from Haldia Institute of Technology, West Bengal, India.

Contents

Applied Mathematics

Pure Mathematics

Security and Privacy

Design of a Certificateless Designated Server Based Searchable Public Key Encryption Scheme

SK Hafizul Islam[1(✉)], Mohammad S. Obaidat[2,3], Varun Rajeev[4], and Ruhul Amin[5]

[1] Indian Institute of Information Technology, Kalyani 741235, West Bengal, India
hafi786@gmail.com, hafi786@iiitkalyani.edu.in
[2] University of Jordan, Amman, Jordan
[3] Fordham University, Bronx, NY, USA
[4] EdgeVerve Systems Limited, Bengaluru 560100, Karnataka, India
[5] Thapar University, Patiala 147004, Punjab, India

Abstract. In the present day, applications of cloud computing is growing exponentially and clients are inclined to use the cloud server to store sensitive data, which is indexed by important or related keyword(s) available in the data. Once the data is stored, the client supplies some keywords to the cloud server and requests the corresponding data. If the data is stored in plaintext form, data privacy will be violated. Thus the client encrypts the data along with the realted keywords, and then stores the ciphertext in the cloud server. Encryption of data maintains the confidentiality, but this makes keyword search difficult. To solve this issue, designated server based public key encryption with keyword search (dPEKS) scheme is used. In dPEKS scheme, to get the encrypted data, the client computes a trapdoor related to a relevant keyword, and sends it to the cloud server, which then gives the ciphertext to the client provided that the trapdoor is verified. Hence, the client gets the data from the ciphertext. However, an adversary will not get any information on the data or the keywords. A certificateless dPEKS (CL-dPEKS) scheme is proposed in this paper. It provides indistinguishability to the ciphertext and trapdoor, and resilience to off-line keyword guessing attack. The Computational Diffie-Hellman (CDH) problem and Bilinear Diffie-Hellman (BDH) problem keep the proposed scheme secure.

Keywords: CL-PKC · dPEKS · Bilinear pairing · Cloud server · Keyword guessing attack

1 Introduction

In 2004, Boneh et al. [1] introduced the notion of public key encryption with keyword search (PEKS) scheme, which is used for secure email access from a

M.S. Obaidat—Fellow of IEEE

© Springer Nature Singapore Pte Ltd. 2017
D. Giri et al. (Eds.): ICMC 2017, CCIS 655, pp. 3–15, 2017.
DOI: 10.1007/978-981-10-4642-1_1

email server containing a list of relevant keywords. A PEKS scheme can be used in an e-mail system as follows. A sender C selects an email data $m \in \{0,1\}^{\ell}$ and a list of keywords $\{w_1, w_2, \cdots, w_n\}$, which are contained in $m \in \{0,1\}^{\ell}$. C then employs the public key of the receiver R to generate a PEKS ciphertext $\{U, V, z_1, z_2, \cdots, z_n\}$ by encrypting $m \in \{0,1\}^{\ell}$ and $\{w_1, w_2, \cdots, w_n\}$. Following this, C delivers $\{U, V, z_1, z_2, \cdots, z_n\}$ to the local email server S of R (the receiver). To get the encrypted email from S, R selects a keyword w_j and computes its corresponding trapdoor Z_j by using his/her private key. Then Z_j is sent to his/her S (the server) to check whether $\{U, V, z_1, z_2, \cdots, z_n\}$ contains w_j, which is concealed in Z_j. S then prepares a ciphertext C_m using his/her private key provided that Z_j is verified. Then S sends $\{C_m, U, V\}$ to R, and he/she recovers the corresponding $m \in \{0,1\}^{\ell}$ using his/her own private key. Note that, S and an outsider $\mathcal{A}$ do not learn any information about the encrypted email and keywords.

In recent years, the popularity of sharing data on a public cloud has increased. The client-server storage service in the public cloud allow clients to store important data in the cloud server at cheap rates. However, the sharing of data must be done securely since data privacy is a major concern in today's world. Generally, a client encrypts the data and then uploads the ciphertext to the cloud server. The encrypted data uploaded by the client to a pubic cloud server, is indexed by one or more keyword(s), which are elements of the uploaded data. But searching for a keyword in an encrypted data is difficult and complex. For this purpose, PEKS scheme is popularly used in the cloud computing environment for secure data storage and access. Here, we aim to design a secure data storage and access mechanism in cloud environments. The proposed data storage and access mechanism is explained as follows. A client C encrypts data $m \in \{0,1\}^{\ell}$ and a list of relevant keywords $\{w_1, w_2, \cdots, w_n\}$, which are components of $m \in \{0,1\}^{\ell}$, by using the public key of the cloud server S to generate a PEKS ciphertext $\{U, V, z_1, z_2, \cdots, z_n\}$. Then, C stores $\{U, V, z_1, z_2, \cdots, z_n\}$ to S. To get $m \in \{0,1\}^{\ell}$ containing a particular w_j, C computes a trapdoor Z_j of w_j using his/her private key, and then sends it to S. Following this, S prepares a ciphertext C_m using his/her private key provided that Z_j is correct. Then S sends $\{C_m, U, V\}$ to C if Z_j is verified. Now, C extracts $m \in \{0,1\}^{\ell}$ from $\{C_m, U, V\}$ using his/her private key.

PEKS scheme proposed in [1] is useful for both email and client-server storage systems. However, the scheme in [1] is bound to use a secure channel between client/receiver and email server/cloud server [2]. To eliminate this requirement, Baek et al. [2] put forwarded the concept of designated server based PEKS (dPEKS) scheme. In dPEKS scheme, only the designated server is allowed to verify whether a keyword of the trapdoor is identical to any of the keywords associated with the data. Unfortunately, Rhee et al. [3] argued that the security model proposed in [2] provides limited capabilities to the adversary and the proposed dPEKS scheme is insecure. In this scheme, an attacker can perform off-line keyword guessing attack to guess the keyword from a given trapdoor. Accordingly, Rhee et al. revised the security model proposed in [2] and proposed

the concept of trapdoor indistinguishability. They also put forwarded a secure dPEKS scheme and analyze its security using the refined security model. According to the analysis made in [3], a dPEKS is secure against off-line keyword guessing attack if the scheme provides trapdoor indistinguishability property. Unfortunately, Hu et al. [4,5] found that the dPEKS scheme proposed in [3] is vulnerable to the off-line keyword guessing attack, which is performed by a malicious server. Then they proposed two improved dPEKS schemes in [4,5]. Unfortunately, Ni et al. [6] found that the schemes in [4,5] are vulnerable to the off-line keyword guessing attack performed by a malicious server and chosen keyword attack.

All the schemes proposed in [1–6] are designed using certificate based public key cryptography (CA-PKC). In these schemes, the certificate of the public key must be verified before using it to get assured that the public key actually belongs to the correct party. In CA-PKC, public key infrastructure (PKI) is required to mange the complex public key certificate management process to authenticate the public key, which decreases the applicability in real environments. To defeat these troubles, Shamir [7] introduces the idea of identity-based cryptography (IBC), which eliminates the use of public key certificate as needed in CA-PKC. In IBC, client's public key is calculated from the publicly known identity of the client, such as email identity, passport number, social security number, etc. and a trusted third party, called private key generator (PKG) is responsible to generate the corresponding private key of the client by binding client's identity and PKG's private key. Bones and Franklin [8] designed map-to-point hash function to proposed a practical identity-based encryption (IBE) scheme using elliptic curve [9,10] and bilinear pairing. Based on this IBE scheme, in 2013, Wu et al. [11] proposed a dPEKS scheme, called dIBEKS. However, Wu et al.'s dIBEKS scheme has a limitation due to the existing problem of IBC, called private key escrow problem. Certificateless public key cryptography (CL-PKC) is introduced in [12] by incorporating the merits of IBC and CA-PKC. Note that, CL-PKC abolishes the troubles of IBC and CA-PKC. In CL-PKC, the full private key of a client has two values, one is selected by the client himself/herself and the other is the identity-based private key, which is computed by the PKG. This ensures that the client does not have to put complete trust on PKG.

In 2014, Yanguo et al. [13] proposed a dPEKS scheme using CL-PKC, called CL-dPEKS. This scheme used the elliptic curve and bilinear pairing [8]. This scheme is proven to be probably secure in the random oracle model. However, the computation costs of the scheme is high. Thus, we propose a new CL-dPEKS scheme. The proposed CL-dPEKS scheme is robust and computation cost efficient than the scheme proposed in [13]. Our scheme also provides the indistinguishability to the ciphertext and trapdoor, and resilience to off-line keyword guessing attack. The proposed CL-dPEKS scheme is secure based on CDH and BDH problems.

This paper is arranged as follows. In Sect. 2, we discuss the preliminaries, which are necessary to understand our CL-dPEKS scheme. In Sect. 3, we provide a framework of CL-dPEKS scheme. Section 4 describes a concrete CL-dPEKS

scheme. Section 5 is devoted to the security analysis of our CL-dPEKS scheme. Section 6 is addresses the performance comparison of our scheme with other related scheme. We conclude the paper with some remarks in Sect. 7.

2 Preliminaries

2.1 Bilinear Pairing

Let p be a large prime number of length k bits, and F_p be the finite field over p. We define $E(F_p) : v^2 = u^3 + xu + y \pmod p$, where $(4x^3 + 27y^2) \neq 0 (\text{mod } p)$ over F_p be the elliptic curve, where $x, y \in F_p$. Let $\mathcal{O}$ denote the "point at infinity" [9,10]. Assume that P is the generator of the group $G_1 = E(F_p) \cup \{\mathcal{O}\}$ of order p, where $P \neq \mathcal{O}$. Here G_1 must be additive cyclic group of elliptic curve points. Assume that G_2 is a multiplicative cyclic group of order p. A bilinear pairing $e : G_1 \times G_1 \rightarrow G_2$ is a mapping, which satisfies the following properties [8]:

- **Bilinearity:** For any $P, Q \in G_1$ and $a, b \in Z_p^*$, $e(aP, bQ) = e(P, Q)^{ab}$ must hold.
- **Non-degeneracy:** If P is a generator of G_1, $e(P, P)$ is generator of G_2.
- **Computability:** An efficient polynomial time algorithm $\mathcal{C}$ must exist for the calculation of $e(P, Q)$, for all $P, Q \in G_1$.

A bilinear map e is called an admissible bilinear map if it satisfies above properties. The map e will be derived either from the modified Weil pairing or Tate pairing over F_p [8].

2.2 Bilinear Diffie-Hellman Parameter Generator (BDH-PG)

A BDH-PG $\mathcal{X}$ is a polynomial time bounded algorithm, which takes the security parameter 1^k as input and it then outputs a uniformly random tuple (p, e, G_1, G_2, P) of bilinear parameters.

2.3 Computational Diffie-Hellman (CDH) Problem

Given a random tuple $(P, aP, bP) \in G_1$ for any $a, b \in_R Z_p^*$ and $P \in G_1$, $\mathcal{C}$ cannot calculate abP with in polynomial time.

2.4 Bilinear Diffie-Hellman (BDH) Problem

Given a random tuple (P, aP, bP, cP), for any $a, b, c \in Z_p^*$ and $P \in G_1$, $\mathcal{C}$ cannot calculate $e(P, P)^{abc}$ with in polynomial time.

2.5 System Model

In cloud environments, a dPEKS scheme offers a secure client-server storage system. Our CL-dPEKS scheme is proposed to fulfill this objective. The proposed client-server storage system in public cloud environments is described briefly in Fig. 1. In our CL-dPEKS scheme, three entities are involved: (i) a private key generator (PKG), (ii) a cloud server S, which is identified by the identity ID_S, and (iii) a client C, which is identified by the identity ID_C. The PKG provides identity-based partial private key for C and S. C stores his/her important data in an encrypted form to S after encrypting the data with the public keys of C and S. The whole scenario can be described as follows. Assume that C wishes to upload a data $m \in \{0,1\}^\ell$, which contains n keywords $\{w_1, w_1, \cdots, w_n\}$. Then C encrypts $m \in \{0,1\}^\ell$ as $\{U, V\}$ and $\{w_1, w_1, \cdots, w_n\}$ as $\{z_1, z_1, \cdots, z_n\}$ using the full public keys pk_C of C and pk_S of S, respectively. Finally, C uploads $\{U, V, z_1, z_1, \cdots, z_n\}$ to the S using a public channel. Later on, if C wants to retrieve $m \in \{0,1\}^\ell$ from S, then C prepares a trapdoor Z_j on a keyword w_j using his/her full private key sk_C and sends it to S over a public channel. To search the encrypted $m \in \{0,1\}^\ell$ on the storage, S will prepare a ciphertext C_m using his/her full private key sk_S provided that the trapdoor Z_j is correct. Then S will sends $\{C_m, U, V\}$ to C over a public channel. Note that a third party including S will not learn the data $m \in \{0,1\}^\ell$ using any of the public information. After receiving $\{C_m, U, V\}$ from S, C recovers $m \in \{0,1\}^\ell$ from it using his/her full private key sk_C. List of notation used in this paper is listed in Table 1.

3 Framework of a CL-dPEKS Scheme

A CL-dPEKS scheme includes the following algorithms.

1. **CL-dPEKS-Setup:** The PKG executes this deterministic algorithm. As input, it takes 1^k and it generates a public parameter set Γ and a master secret key msk of PKG.
2. **CL-dPEKS-Gen-Secret-Key:** An entity ID_i $(i = C, S)$ executes this probabilistic polynomial time (PPT) algorithm. As input, it takes Γ and it outputs a secret key x_i and a public key P_i for ID_i.
3. **CL-dPEKS-Gen-Partial-Private-Key-Extract:** The PKG executes this PPT algorithm to generate an identity-based partial private key for ID_i $(i = C, S)$. As inputs, it takes Γ, msk of PKG, and an identity ID_i, P_i of ID_i, and then it returns an identity-based partial private key d_i and a public information T_i for ID_i.
4. **CL-dPEKS-Set-Private-Key:** The entity ID_i $(i = C, S)$ keeps $sk_i = (d_i, x_i)$, as his/her full private key.
5. **CL-dPEKS-Set-Public-Key:** The entity ID_i $(i = C, S)$ keeps $pk_i = (T_i, P_i)$, as his/her full public key.
6. **CL-dPEKS-Encrypt:** The client C performs the execution of this PPT algorithm, which takes the full public key pk_C of C, full public key pk_S of S, a data $m \in \{0,1\}^\ell$, a list of relevant keywords $\{w_1, w_2, \cdots, w_n\}$ as inputs, and then it outputs a ciphertext $\{U, V, z_1, z_2, \cdots, z_n\}$.

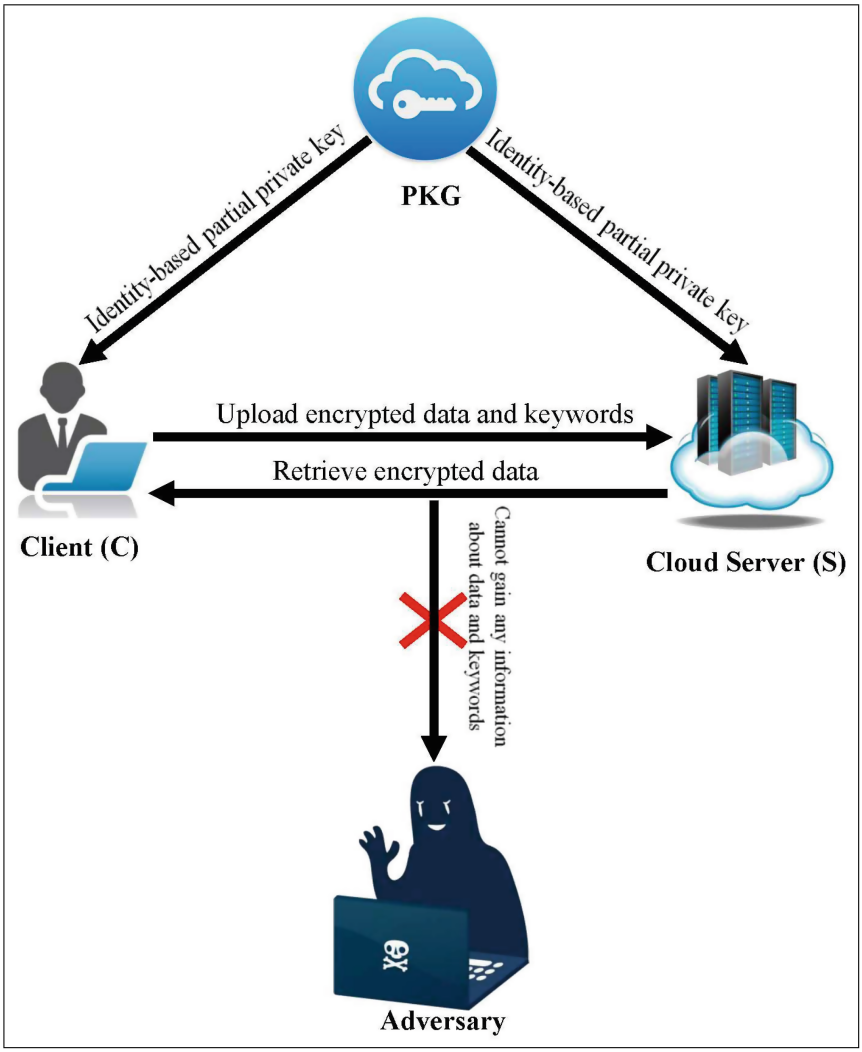


Fig. 1. Proposed client-server storage system in public cloud environments.

7. **CL-dPEKS-Gen-Trapdoor:** The client C executes this PPT algorithm, which takes the full private key sk_C of C, a keyword w_j as inputs and then it outputs a trapdoor Z_j.
8. **CL-dPEKS-Test-Trapdoor:** The cloud server S executes this deterministic algorithm. As inputs, it takes a trapdoor Z_j, a ciphertext $\{U, V, z_1, z_2, \cdots, z_n\}$, full private key $sk_S = (d_S, x_S)$ of S and full public key $pk_C = (T_C, P_C)$ of C. It outputs `True` if Z_j is correct, i.e., w_j is matched with any of $\{w_1, w_2, \cdots, w_n\}$, else outputs `False`.

Table 1. List of notations used in the proposed protocol.

Notation	Description
p	A large prime number of k-bit
F_p	A finite field of order p
$E(F_p)$	A set of elliptic curve points
Z_p	$Z_p = \{0, 1, \cdots, p-1\}$
Z_p^*	$Z_p^* = Z_p \backslash \{0\}$
$x \in_R A$	An element x is randomly selected from the set A
G_1	Additive cyclic group of order p
G_2	Multiplicative cyclic group of order p
P	Generator of G_1, $P \neq \mathcal{O}$
PKG	Private key generator
s	Master key of PKG
e	An admissible bilinear map, $e : G_1 \times G_1 \rightarrow G_2$
P_0	Public key of PKG, $P_0 = sP$
C	A client
S	A cloud server
ID_i	Identity of the entity i, $i = C, S$
d_i	Partial private key of ID_i, $i = C, S$
x_i	Secret value of ID_i, $i = C, S$
sk_i	Full private key of ID_i, $i = C, S$
pk_i	Full Public key of ID_i, $i = C, S$
m	The data, where $m \in \{0, 1\}^\ell$
w_i	i-th keyword, $i = 1, 2 \cdots, n$
Z_j	Trapdoor of the keyword w_j
$h(\cdot)$	One-way general hash function
$H(\cdot)$	Map-to-point hash function
$\oplus$	Bitwise exclusive-or operation

9. **CL-dPEKS-Decrypt-Ciphertext:** The cloud server S executes this deterministic algorithm. As input, it takes a trapdoor Z_j and the full private key sk_S of S. If the output of **CL-dPEKS-Test-Trapdoor** algorithm is `True`, then S computes a ciphertext C_m and returns $\{C_m, U, V\}$ to C.
10. **CL-dPEKS-Data-Recovery:** The client C executes this deterministic algorithm. As inputs, it takes $\{C_m, U, V\}$ and the full private key pk_C of C and outputs the original data $m \in \{0, 1\}^\ell$.

4 Proposed CL-dPEKS Scheme

Our CL-dPEKS scheme includes the following polynomial time bounded algorithms: (i) CL-dPEKS-Setup, (ii) CL-dPEKS-Gen-Secret-Key, (iii) CL-dPEKS-Gen-Partial-Private-Key, (iv) CL-dPEKS-Set-Private-Key, (v) CL-dPEKS-Set-Public-Key, (vi) CL-dPEKS-Encrypt, (vii) CL-dPEKS-Gen-Trapdoor, (viii) CL-dPEKS-Test-Trapdoor, (ix) CL-dPEKS-Decrypt-Ciphertext and (x) CL-dPEKS-Data-Recovery.

1. **CL-dPEKS-Setup:** PKG takes the 1^k as inputs and then does as follows.
   (a) Choose a large prime number p of size k bits.
   (b) Choose a tuple $\{F_p, E(F_p), G_1, G_2, e, P\}$.
   (c) Select $s \in Z_p^*$ as the master key. The public key will be calculated as $P_0 = sP$.
   (d) Select the map-to-point hash function $H(\cdot) : \{0,1\}^* \rightarrow G_1$ [8], and a general one-way hash function $h(\cdot) : \{0,1\}^* \rightarrow \{0,1\}^\ell$, where ℓ depends on the specific hash function. For example, if $h(\cdot)$ is considered as SHA-512, then $\ell = 512$ bits.
   (e) Publish $\Gamma = \{F_p, E(F_p), G_1, G_2, e, P, P_0, H(\cdot), h(\cdot)\}$.
2. **CL-dPEKS-Gen-Secret-key:** The entity i $(i = C, S)$ with identity ID_i selects $x_i \in_R Z_p$ as his/her secret key and then calculates the corresponding public key as $P_i = x_i P$.
3. **CL-dPEKS-Gen-Partial-Private-Key-Extract:** The entity ID_i $(i = C, S)$ delivers $\{ID_i, P_i\}$ to the PKG over a private channel. Then PKG does as follows:
   (a) Choose $t_i \in_R Z_p^*$ and calculates $T_i = t_i P$.
   (b) Calculate $l_i = h(ID_i, T_i, P_i)$ and $d_i = (t_i + sl_i) \bmod p$.
   Now PKG sends the tuple (d_i, T_i) to ID_i through a secure channel. Here the partial private key of ID_i is d_i and $Q_i = d_i P$ will serve as the partial public key of ID_i. The private key d_i is considered legitimate if $d_i P = T_i + h(ID_i, T_i, P_i)P_0 = Q_i$ holds. Since we have,

$$\begin{aligned} Q_i &= T_i + h(ID_i, T_i, P_i)P_0 \\ &= t_i P + l_i P_0 \\ &= t_i P + l_i sP \\ &= (t_i + sl_i)P \\ &= d_i P \end{aligned}$$

4. **CL-dPEKS-Set-Private-Key:** The entity ID_i $(i = C, S)$ considers $sk_i = (d_i, x_i)$ as full private key.
5. **CL-dPEKS-Set-Public-Key:** The entity ID_i $(i = C, S)$ considers $pk_i = (P_i, T_i)$ as full public key.
6. **CL-dPEKS-Encrypt:** Given an identity ID_C of C, an identity ID_S of S, full public key $pk_C = (T_C, P_C)$ of C, full public key $pk_S = (P_S, T_S)$ of S, and a list of keywords $\{w_1, w_2, \cdots, w_n\}$, C runs this algorithm to generate a ciphertext $\{U, V, z_1, z_2, \cdots, z_n\}$ as follows

(a) Select $r \in_R Z_p^*$ and calculate $U = rP$.
(b) Select $m \in \{0,1\}^\ell$ and calculate $V = m \oplus h[e(rP_0, P_C + P_S + T_C + T_S + (l_C + l_S)P_0)]$, where $l_C = h(ID_C, T_C, P_C)$ and $l_S = h(ID_S, T_S, P_S)$.
(c) Calculate $z_i = r[H(w_i) + P_S + T_S + l_S P_0]$, for $i = 1, 2, \cdots, n$.
C sends $\{U, V, z_1, z_2, \cdots, z_n\}$ to S over a public channel.

7. **CL-dPEKS-Gen-Trapdoor:** Given the identity ID_C of C, full private key $sk_C = (d_C, x_C)$ of C, and a keyword w_j, C runs this algorithm to generate a trapdoor $Z_j = (x_C + d_C)H(w_j)$. C then sends Z_j to S over a public channel.
8. **CL-dPEKS-Test-Trapdoor:** Given a tuple $\{U, V, z_1, z_2, \cdots, z_n\}$, full private key $sk_S = (d_S, x_S)$ of S, full public key $pk_C = (T_C, P_C)$ of C, and a trapdoor Z_j, then S runs this algorithm to check whether $e(Z_j + (x_S + d_S)(P_C + T_C + l_C P_0), U) = e(z_i, P_C + T_C + l_C P_0)$, for $i = 1, 2, \cdots, n$. If the justification of one of the equations is correct, S returns `True`, it means that the keyword w_j of included in Z_j is identical to one of the keywords $\{w_1, w_2, \cdots, w_n\}$. Otherwise, S returns `False` and terminates the process. Suppose that, $j = i$ for some i, then we have

$$\begin{aligned}
& e(Z_j + (x_S + d_S)(P_C + T_C + l_C P_0), U) \\
&= e((x_C + d_C)H(w_j) + (x_S + d_S)(x_C + d_C)P, rP) \\
&= e((x_C + d_C)H(w_j), rP)e((x_S + d_S)(x_C + d_C)P, rP) \\
&= e(rH(w_j), (x_C + d_C)P)e(r(x_S + d_S)P, (x_C + d_C)P) \\
&= e(rH(w_j) + r(x_S + d_S)P, (x_C + d_C)P) \\
&= e(r[H(w_j) + (x_S + d_S)P], (x_C + d_C)P) \\
&= e(r[H(w_j) + P_S + T_S + l_S P_0], P_C + T_C + l_C P_0) \\
&= e(z_j, P_C + T_C + l_C P_0)
\end{aligned}$$

9. **CL-dPEKS-Decrypt-Ciphertext:** If the algorithm **CL-dPEKS-Test-Trapdoor** outputs `True` for Z_j, then S run this algorithm and computes $C_m = e(U, (x_S + d_S)P)$. Now, S returns $\{C_m, U, V\}$ to C over a public channel.
10. **CL-dPEKS-Data-Recovery:** After receiving $\{C_m, U, V\}$ from S, C recover the original data $m \in \{0,1\}^\ell$ by executing $V \oplus h[e(U, (x_C + d_C)P_0)C_m]$. Since, we have

$$\begin{aligned}
& V \oplus h[e(U, (x_C + d_C)P_0)C_m] \\
&= V \oplus h[e(rP, (x_C + d_C)P_0)e(U, (x_S + d_S)P)] \\
&= V \oplus h[e(rP_0, (x_C + d_C)P)e(U, (x_S + d_S)P)] \\
&= V \oplus h[e(rP_0, (x_C + d_C)P)e(rP_0, (x_S + d_S)P)] \\
&= V \oplus h[e(rP_0, (x_C + d_C)P + (x_S + d_S)P)] \\
&= V \oplus h[e(rP_0, x_C P + d_C P + x_S P + d_S P)] \\
&= m \oplus h[e(rP_0, P_C + P_S + T_C + T_S + (l_C + l_S)P_0)] \\
&\quad \oplus h[e(rP_0, P_C + T_C + l_C P_0 + P_S + T_S + l_S P_0)] \\
&= m
\end{aligned}$$

The proposed CL-dPEKS scheme is further described in the Fig. 2.

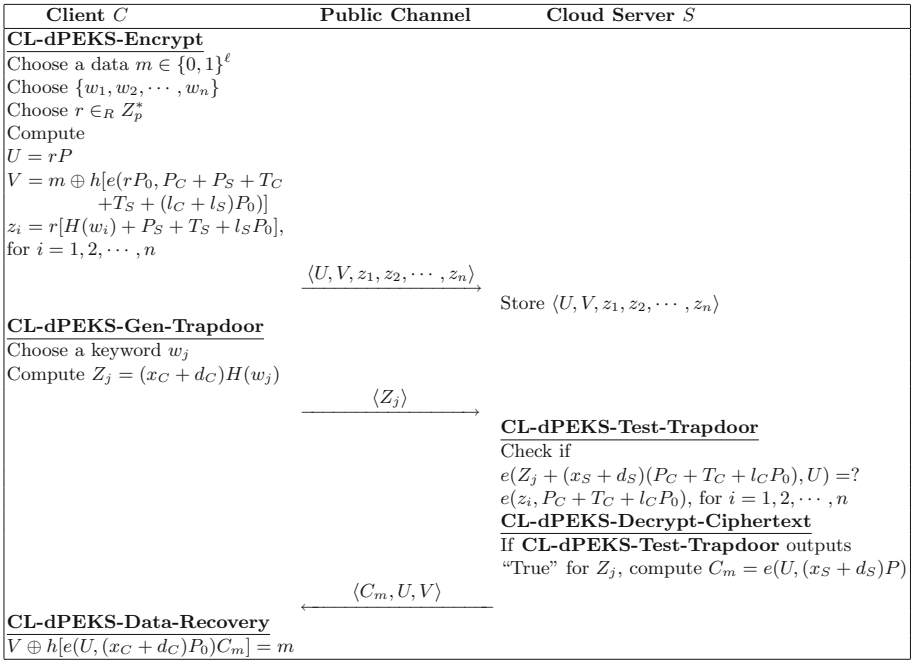


Fig. 2. Proposed CL-dPEKS scheme

5 Security Analysis

The following adversaries are considered for CL-PKC system [12].

1. **Type I adversary** $\mathcal{A}_I$: The master key $msk = s$ of the PKG cannot be accessed by $\mathcal{A}_I$, but may get the public keys P_C and P_S of C and S, and can replace these public keys with the new public keys P'_C and P'_S chosen by him/her.
2. **Type II adversary** $\mathcal{A}_{II}$: The master key $msk = s$ of PKG can be accessed by $\mathcal{A}_{II}$. But, $\mathcal{A}_{II}$ is not permitted to change the public keys P_C and P_S of C and S.

A CL-dPEKS scheme must provide the following security requirements.

– **Ciphertext Indistinguishability:** In our CL-dPEKS scheme, C encrypts a data $m \in \{0,1\}^{\ell}$ and a list of keywords $\{w_1, w_2, \cdots, w_n\}$ using **CL-dPEKS-Encrypt** algorithm, and the ciphertext $\{U, V, z_1, z_2, \cdots, z_n\}$ is delivered to S over a public channel, where $U = rP$, $V = m \oplus h[e(rP_0, P_C + P_S + T_C + T_S + (l_C + l_S)P_0)]$ and $z_i = r[H(w_i) + P_S + T_S + l_S P_0]$, for $i = 1, 2, \cdots, n$. Assume that an adversary $\mathcal{A} \in \{\mathcal{A}_I, \mathcal{A}_{II}\}$ captures the ciphertext. Note that r is a random bit string, unknown to $\mathcal{A}$ and it will change despite the same data and list of keywords getting encrypted every time. The probability of guessing

r is $\frac{1}{2^k}$, where length of r is k bits. However, $\mathcal{A}$ may try to compute $H(w_i)$ from z_i, but since r is unknown, $\mathcal{A}$ is unable to compute any of $\{w_1, w_2, \cdots, w_n\}$ even if he/she knows the public keys pk_C and pk_S. Furthermore, V is still protected under the BDH problem. Accordingly, $\mathcal{A}$ can not compute r from $U = rP$, since it is protected under the CDH problem. Therefore, under the CDH and the BDH problems, our CL-dPEKS scheme provides the ciphertext indistinguishability.

- **Trapdoor Indistinguishability:** In our CL-dPEKS scheme, C sends a trapdoor Z_j of the keyword w_j to S over a public channel to get the encrypted data. For w_j, C computes $Z_j = (x_C + d_C)H(w_j)$ using the **CL-dPEKS-Gen-Trapdoor** algorithm and then sends it to S over a public channel. Assume that $\mathcal{A} \in \{\mathcal{A}_I, \mathcal{A}_{II}\}$ captures Z_j. $\mathcal{A}_I$ can calculate d_C, but cannot compute x_C. Therefore, $\mathcal{A}_I$ cannot compute $H(w_j)$ from Z_j within polynomial time due to CDH problem. Accordingly, our CL-dPEKS scheme provides trapdoor indistinguishability.
- **Off-line Keyword Guessing Attack:** From the security requirements of trapdoor indistinguishability and ciphertext indistinguishability, $\mathcal{A} \in \{\mathcal{A}_I, \mathcal{A}_{II}\}$ can not derive the hashed keywords $H(w_i)$ and $H(w_j)$ from z_i, and Z_j, respectively. Therefore, according to analysis provided in [3], we conclude that our CL-dPEKS scheme is not susceptible to the off-line keyword guessing attack.

6 Performance Evaluation

Here, we have included a computation cost comparison of our CL-dPEKS scheme with the scheme proposed by Yanguo et al. [13]. We define T_M, T_H and T_P as the computation costs of elliptic curve scalar point multiplication, map-to-point hashing operation, and bilinear pairing operation, respectively. According to the result obtained in [14], we know that $T_M \approx 29T_m$, $T_P \approx 87T_m$ and $T_H \approx 29T_m$, where T_m is the time needed for the execution of a modular multiplication

Table 2. Computation cost comparison

Phase	Yanguo et al. [13]	Proposed
CL-dPEKS-Encrypt	$3nT_M + (n+2)T_H + (3n+4)T_P$	$(n+3)T_M + T_P$
CL-dPEKS-Gen-Trapdoor	$4T_M + T_H$	T_M
CL-dPEKS-Test-Trapdoor	$2nT_P$	$3nT_M + 2nT_P$
CL-dPEKS-Decrypt-Ciphertext	Not proposed	$T_M + T_P$
CL-dPEKS-Data-Recovery	Not proposed	$T_M + T_P$
Overall computation cost	$(3n+4)T_M + (n+3)T_H + (5n+4)T_P \approx 551(n+1)T_m$	$(4n+6)T_M + (2n+3)T_P \approx 290(n+2)T_m$

operation. The computation cost comparison is given in Table 2. The overall computation cost of our CL-dPEKS scheme is lower compared to the scheme proposed in [13].

7 Conclusion

A new CL-dPEKS scheme is proposed in this paper for secure client-server storage service in public cloud environments. A client of our scheme is allowed to deliver a trapdoor to the cloud server over a public channel. The proposed CL-dPEKS scheme is compared with the scheme proposed in [13] and found that our scheme is more computation-cost-effective. We also found out that our CL-dPEKS scheme offers ciphertext indistinguishability and trapdoor indistinguishability, and resists off-line keyword guessing attack.

References

1. Boneh, D., Di Crescenzo, G., Ostrovsky, R., Persiano, G.: Public key encryption with keyword search. In: Cachin, C., Camenisch, J.L. (eds.) EUROCRYPT 2004. LNCS, vol. 3027, pp. 506–522. Springer, Heidelberg (2004). doi:10.1007/978-3-540-24676-3_30
2. Baek, J., Safavi-Naini, R., Susilo, W.: Public key encryption with keyword search revisited. In: Gervasi, O., Murgante, B., Laganà, A., Taniar, D., Mun, Y., Gavrilova, M.L. (eds.) ICCSA 2008. LNCS, vol. 5072, pp. 1249–1259. Springer, Heidelberg (2008). doi:10.1007/978-3-540-69839-5_96
3. Rhee, H.S., Park, J.H., Susilo, W., Lee, D.H.: Improved searchable public key encryption with designated tester. In: Proceedings of the 4th International Symposium on Information, Computer, and Communications Security (ASIACCS 2009), pp. 376–379 (2009)
4. Hu, C., Liu, P.: An enhanced searchable public key encryption scheme with a designated tester and its extensions. J. Comput. **7**(3), 716–723 (2012)
5. Hu, C., Liu, P.: A secure searchable public key encryption scheme with a designated tester against keyword guessing attacks and its extension. In: Lin, S., Huang, X. (eds.) CSEE 2011. CCIS, vol. 215, pp. 131–136. Springer, Heidelberg (2011). doi:10.1007/978-3-642-23324-1_23
6. Ni, J., Yu, Y., Xia, Q., Niu, L.: Cryptanalysis of two searchable public key encryption schemes with a designated tester. J. Inf. Comput. Sci. **9**(16), 4819–4825 (2012)
7. Shamir, A.: Identity-based cryptosystems and signature schemes. In: Blakley, G.R., Chaum, D. (eds.) CRYPTO 1984. LNCS, vol. 196, pp. 47–53. Springer, Heidelberg (1985). doi:10.1007/3-540-39568-7_5
8. Boneh, D., Franklin, M.K.: Identity based encryption from the Weil Pairing. SIAM J. Comput. **32**(3), 586–615 (2003)
9. Miller, V.S.: Use of elliptic curves in cryptography. In: Williams, H.C. (ed.) CRYPTO 1985. LNCS, vol. 218, pp. 417–426. Springer, Heidelberg (1986). doi:10.1007/3-540-39799-X_31
10. Koblitz, N.: Elliptic curve cryptosystem. J. Math. Comput. **48**(177), 203–209 (1987)
11. Wu, T.-Y., Tsai, T.-T., Tseng, Y.-M.: Efficient searchable ID-based encryption with a designated server. Ann. Telecommun. **69**(7–8), 391–402 (2014)

12. Al-Riyami, S.S., Paterson, K.G.: Certificateless public key cryptography. In: Laih, C.-S. (ed.) ASIACRYPT 2003. LNCS, vol. 2894, pp. 452–473. Springer, Heidelberg (2003). doi:10.1007/978-3-540-40061-5_29
13. Yanguo, P., Jiangtao, C., Changgen, P., Zuobin, Y.: Certificateless public key encryption with keyword search. China Commun. **11**(11), 100–103 (2014)
14. Islam, S.H., Khan, M.K., Al-Khouri, A.M.: Anonymous and provably secure certificateless multireceiver encryption without bilinear pairing. Secur. Commun. Netw. **8**, 2214–2231 (2015)

On the Security of a Searchable Anonymous Attribute Based Encryption

Payal Chaudhari[1,2] and Manik Lal Das[1(✉)]

[1] DA-IICT, Gandhinagar, India
payal.ldrp@gmail.com, maniklal_das@daiict.ac.in
[2] LDRP, Gandhinagar, India

Abstract. Ciphertext Policy Attribute-based Encryption(CP-ABE) is a public key primitive in which a user is able to decrypt a ciphertext if the attributes associated with secret key and the access policy connected with ciphertext matches. Although CP-ABE provides both confidentiality and fine-grained access control to the data stored in public cloud, anonymous CP-ABE adds interesting feature of sender and/or receiver anonymity. In this paper, we discuss a recent work on anonymous CP-ABE [1], which aims to provide secure and efficient data retrieval anonymously. We show that the scheme has major security weakness and does not ensure anonymity feature, which is the main claim of the scheme. We then present an improved scheme for mitigating the weaknesses of the scheme. The improved scheme retains the security claims of the original scheme [1] without adding any computation and communication overhead.

Keywords: Attribute based encryption · Anonymity · Confidentiality · Access structure

1 Introduction

Cloud computing is a comprehensive model, which provides on-demand computing resources such as storage, network, applications and services. Many enterprizes and individuals outsource their data to the cloud storage servers in order to reduce the cost for resource management. While making this flexibility to manage data in third party server, the security and privacy of data are major concerns. The outsourced data may contain sensitive information, such as Electronic Health Records(EHRs), financial details, personal photos etc. Therefore, data must be protected in the cloud storage server, so that unauthorized data access and data privacy protection need to be handled appropriately based on application requirement. There have been several approaches to securing data in cloud server. However, data encryption is a widely used primitive for securing data from authorized users. Before storing the data in cloud server, the data owner can encrypt the data so that the cloud server cannot learn anything from the stored data. Once the encrypted data are stored in the cloud server, two requirements become apparent for user convenience - *Access control* and

© Springer Nature Singapore Pte Ltd. 2017
D. Giri et al. (Eds.): ICMC 2017, CCIS 655, pp. 16–25, 2017.
DOI: 10.1007/978-981-10-4642-1_2

Search over encrypted data. To provide a solution for secure and fine-grained data access, Sahai and Waters introduced the concept of attribute-based encryption (ABE) [2]. Ciphertext-Policy ABE(CP-ABE) [3] enables data encryption as per the access policy, where the access policy describes the combination of required attributes. User's secret key contains the attribute values which the user possesses. If the user's key matches with the access policy then he can decrypt the documents.

Although ABE scheme supports fine-grained access control [4], it discloses sender and/or receiver identity by which an adversary can guess the meaning or purpose of the message by seeing the attributes. Therefore, protecting sender and/or receiver identity while using ABE has been found a challenging research problem. In order to address this problem, anonymous ABE (AABE) schemes have been proposed in literature [5–9]. In anonymous CP-ABE, access policy is concealed inside the ciphertext components. A user tries to decrypt a ciphertext using the secret key made up with his attributes. If his attributes fulfills the access policy, then the decryption operation is successful. If the attributes included in the secret key do not match with the access policy, then the user can neither decrypt the ciphertext nor he can uncover the access policy hidden inside the ciphertext.

In 2013, Koo *et al.* [1] have proposed a searchable anonymous ABE scheme, where search on encrypted data is done on data owner's identity and data retriever's attributes. The scheme claimed that a user in the system can search on encrypted data stored in cloud with preserving sender and receiver anonymity. In this paper, we show that Koo *et al.*'s scheme fails to achieve the receiver anonymity [10]. We then propose an improved scheme, which mitigates the security flaw and retains the claimed security strength without adding any overhead.

The remaining of the paper is organized as follows. In Sect. 2, we give some preliminaries. In Sect. 3, we discuss Koo *et al*'s scheme. In Sect. 4, we show the security weaknesses of Koo *et al.*'s scheme. In Sect. 5, we present an improved scheme and provide its analysis in Sect. 6. We conclude the paper in Sect. 7.

2 Preliminaries

2.1 Bilinear Mapping

Let G_1 and G_2 be two multiplicative cyclic groups of prime order p. Let g be a generator of G_1 and e be a bilinear map, $e : G_0 \times G_0 \rightarrow G_1$. The bilinear map e has the following properties:

- Bilinearity: For all $u,v \in G_0$ and a, $b \in Z_p^*$, we have $e(u^a, v^b) = e(u, v)^{ab}$.
- Non-degeneracy: $e(g, g) \neq 1$.
- Efficiency: The function e is efficiently computable.

We say that G_0 is a bilinear group if the group operation in G_0 and the bilinear map e: $G_0 \times G_0 \rightarrow G_1$ are both efficiently computable.

2.2 Access Tree

Access structure is represented in form of an access tree T. Each non-leaf node of the tree behaves as a threshold gate. It is defined as a tuple of its children and a threshold value. Let num_x denotes the number of children of a node x and k_x represents the threshold value of the node x, then $0 < k_x \leq num_x$. For an OR gate $k_x = 1$, and for an AND gate, $k_x = num_x$. Each leaf node x represents an attribute and threshold value $k_x = 1$. Each child of a parent will have unique index number from set [1,*num*] in an ordered fashion. To assist in traversing the access trees in cryptographic operations, following functions are being used.

- $parent(x)$ = parent of the node x in the tree.
- $att(x)$ = attribute associated with the leaf node x.
- $index(x)$ = index number of node x as a child of its parent node. The value will be between 1 to *num*.

The encryption algorithm first chooses a polynomial q_x for each node x (including the leaves) in the tree $\mathcal{T}$. The polynomial is chosen in a top-to-bottom fashion, initiating from the root node R. For each node x in the tree, the degree d_x of the polynomial $q_x = k_x - 1$, that is d_x is one less than the threshold value k_x of that node. For the root node R, the algorithm selects a random $s \in Z_p$ and sets $q_R(0) = s$. Then, it picks d_R number of random points to define the polynomial q_R. For every other node x of access tree, it computes $q_x(0) = q_{parent}(x)(index(x))$ and selects d_x number of random points randomly to define a polynomial q_x.

3 Koo *et al.*'s Scheme

Koo *et al.* [1] proposed a scheme for secure and efficient data retrieval using anonymous attribute based encryption. The scheme works with the four entities as follows.

- Trusted Authority (TA), who generates user specific secret keys.
- Cloud service provider(CSP) is a semi-trusted entity where the users stored their data in encrypted form.
- Data owner/encryptor, who encrypts and stores the data in CSP.
- Data retriever/receiver, who issues queries to the CSP to access encrypted data from the cloud storage and retrieves the data only if his attributes satisfies the access policy specified by the data owner.

The scheme consists of five phases - System Setup, Key Generation, Encryption, Data Access and Decryption.

3.1 System Setup

The TA performs the setup. It chooses a bilinear group G of prime order p with generator g. It picks two random exponents α, β from Z_p and also selects a cryptographic hash function H: $\{0,1\}^* \rightarrow G$. TA computes the public parameter PK and master secret MK for the system as: $PK = \langle G, g, \omega = e(g,g)^{\alpha}, h = g^{\beta} \rangle$, $MK = \langle g^{\alpha}, \beta \rangle$.

3.2 Key Generation

Each data owner gets a secret key A_O from TA in which data owner identity is hidden. Each receiver gets a secret key SK from TA for decryption operation.

- For the data owner having identity ID_0, TA computes and returns him the anonymous key, $A_O = H(ID_0)^{\beta}$.
- The TA chooses a random $r \in Z_p$ for each individual user u_i in the system and $r_j \in Z_p$ for each attribute $\lambda_j \in \Lambda_i$. Here Λ_i is the set of attributes that belongs to user u_i. The private key SK is computed as

$$SK = \langle D = g^{\frac{(\alpha+r)}{\beta}}, \{D_j = g^r H(\lambda_j)^{r_j}, \\ D'_j = g^{r_j}, D''_j = H(\lambda_j)^{\beta}\}_{\lambda_j \in \Lambda_i} \rangle$$

3.3 Encryption

Before uploading data content to cloud storage, the data owner having the identity ID_O computes his pseudonym as $P_O = H(ID_O)^t$. Here t is the random value selected by the data owner from Z_p. The data owner publicizes his pseudonym. To encrypt data M, the data owner runs **Encrypt** algorithm, as explained below. The encryption algorithm inputs the public parameter PK, its pseudonym P_O, a message M to be encrypted under the access tree $\mathcal{T}$, and outputs the ciphertext CT_0. After that, the attribute scrambling procedure, **AttrScm**, is applied to the ciphertext CT_0 for generating new ciphertext CT to be located in the cloud storage.

Data Encryption(Encrypt). This algorithm chooses a polynomial q_x for each node x (including the leaves) in a top-down manner, starting from the root node R in the tree $\mathcal{T}$. For each node x in the tree, set the degree d_x of the polynomial q_x as $k_x - 1$. The algorithm chooses a random $s \in Z_p$ and sets one point for polynomial q_R as $(0, s)$. Rest of the d_R points are chosen randomly to completely define the polynomial q_R. For every other node x, the algorithm fixes $q_x(0) = q_{parent(x)}(index(x))$ and selects d_x number of random points to completely define a polynomial q_x. Let Y be the set of leaf nodes in $\mathcal{T}$. The ciphertext is built upon the basis of the access tree $\mathcal{T}$ as $CT' = (\mathcal{T}, \tilde{C} = M\omega^s, C = h^s, C'' = P_O, \{C_y = g^{q_y(0)}, C'_y = H(attr_y)^{q_y(0)}\}_{y \in Y})$.

Attribute Scrambling(AttrScm). In this phase the data owner garbles each attribute value included in $\mathcal{T}$ and obtains a new access tree $\mathcal{T}'$ by running **AttrScm**(CT_0,A_O,S). S is the set of attributes which are included in the access policy of CT_0. $S = \{\lambda_i, \cdots, \lambda_k | 1 \leq i \leq k \leq |L|\}$. For each attribute included in S, the data owner computes

$$\begin{aligned} K_{O,S} &= \{e(A_O^t, H(\lambda_j))\}_{\lambda_j \in S} \\ &= \{e(H(ID_O)^{\beta t}, H(\lambda_j))\}_{\lambda_j \in S} \\ &= \{e(H(ID_O), H(\lambda_j))^{\beta t}\}_{\lambda_j \in S} \end{aligned}$$

and replaces the value of λ_x of every leaf node x related to $attr_x$ in $\mathcal{T}$ with the value of $scm_{att_x} \in K_{O,S}$. This results in the access tree $\mathcal{T}'$. The output of this algorithm is $CT = \langle \mathcal{T}', \tilde{C}, C, C'', \{C_y, C'_y\}_{y \in Y} \rangle$

At the end of this phase, the data owner uploads the CT on the cloud storage.

3.4 Data Access

This phase facilitates the retrieval of encrypted data from CSP.

- **Data query.** In the initial phase, a retriever can first gets a pseudonym list of data owners either from the CSP or directly from the data owners. Once the retriever determines to retrieve the data with $C'' = P_O$ from the cloud storage, it can generate cryptographic index terms for the attributes included in his secret key SK as follows.

$$\begin{aligned} K_{O,\Lambda_i} &= \{e(D''_j, C'')\}_{j \in \Lambda_i} \\ &= \{e(H(ID_O)^t, H(\lambda_j)^\beta)\}_{j \in \Lambda_i} \\ &= \{e(H(ID_O), H(\lambda_j))^{\beta t}\}_{j \in \Lambda_i} \end{aligned}$$

  After that, the retriever submits his data request query in the form of a subset of these scrambled index information $K_{O,\Lambda'_i} \subseteq K_{O,\Lambda_i}$ to the CSP.
- **Data Retrieval.** After receiving search query in form of scrambled index terms K_{O,Λ'_i}, the CSP searches in his database if the requested item is present in the storage and if it is present then whether it is satisfied by the requested index attributes. This is done by the algorithm $\mathcal{C}(\mathcal{T}, K_{O,\Lambda'_i})$. The algorithm returns *true* or *false*.

  Let T_x be a subtree of T with root node x and $X' = \{x' \in Y_x$ and parent(x') $= x\}$. $\mathcal{C}(\mathcal{T},K_{O,\Lambda'_i})$ is computed recursively as follows. If x is a leaf node, $\mathcal{C}(\mathcal{T}_x,K_{O,\Lambda'_i})$ returns true if and only if $attr_x \in K_{O,\Lambda'_i}$. If x is a non-leaf node in $\mathcal{T}$, $\mathcal{C}(\mathcal{T},K_{O,\Lambda'_i})$ returns true if and only if at least k_x children return true. For each ciphertext CT_i, where $0 \leq i \leq m$, the CSP simply follows the access tree T^i and determines whether $\mathcal{C}(\mathcal{T}^i,K_{O,\Lambda'_i})$ returns *true* or not. The CSP sends the ciphertexts to the retriever for which the algorithm $\mathcal{C}(\mathcal{T}^i,K_{O,\Lambda'_i})$ returns true.

3.5 Decryption

When a retriever receives the requested content from CSP in encrypted form, then he applies the decryption algorithm DecryptNode on that encrypted content to obtain the plaintext.

DecryptNode(*CT, SK, S*). For a leaf node x in access tree the algorithm computes as follows: If $i\ (= attr_x) \in S$ then

$$\begin{aligned} DecryptNode(CT, SK, S) &= \frac{e(D_i, C_x)}{e(D'_i, C'_x)} \\ &= \frac{e(g^r \cdot H(i)^{r_i}, g^{q_x(0)})}{e(g^{r_i}, H(i)^{q_x(0)})} \\ &= e(g,g)^{rq_x(0)} \\ &= F_x \end{aligned}$$

If x is a nonleaf node then the algorithm proceeds as follows : $\{\forall z \in$ children of $x\}$, it invokes the **DecryptNode(*CT, SK, z*)** and stores the output as F_z. Let S_x is the arbitrary k_x sized set of child nodes z such that $F_z \neq \perp$, then next step is computed as

$$\begin{aligned} F_x &= \prod_{z \in S_x} F_z^{\Delta_{i,s'_x(0)}} \\ &= \prod \left(e(g,g)^{rq_z(0)}\right)^{\Delta_{i,s'_x(0)}} \\ &= \prod \left(e(g,g)^{rq_{parent(z)(index(z))}}\right)^{\Delta_{i,s'_x(0)}} \\ &= \prod \left(e(g,g)^{rq_z(i)}\right)^{\Delta_{i,s'_x(0)}} \\ &= e(g,g)^{rq_z(0)} \end{aligned}$$

(Here, Δ is Lagrange coefficient).

The decryption result becomes $F_R = e(g,g)^{rq_R(0)} = e(g,g)^{rs}$.

From this, the algorithm can decrypt the ciphertext and restore the original data content M by computing

$$\begin{aligned} \frac{\tilde{C}}{e(C,D)/F_R} &= \frac{M\omega^s}{e(h^s, g^{(\alpha+r)/\beta})/e(g,g)^{rs}} \\ &= \frac{Me(g,g)^{\alpha s}}{e(g^{\beta s}, g^{(\alpha+r)/\beta})/e(g,g)^{rs}} \\ &= M \end{aligned}$$

4 Weaknesses in Koo *et al.*'s scheme

In the scheme [1], the attributes in the access policy are scrambled with a pseudonym computed by the data owner. The pseudonym hides the data owner(encryptor)'s identity. To fetch the documents from CSP the receiver requires the pseudonym. The receiver gets the pseudonym in either of these two ways:

1. a pseudonym directly from data owner.
2. a list of pseudonyms from the CSP.

If the receiver gets a pseudonym directly from the data owner then the receiver is knowing the data owner. The receiver scrambles his attributes using the pseudonym and retrieves the documents from the cloud as described in the **Retrieve** procedure. This compromises the anonymity of the sender. If the receiver gets a list of pseudonyms from the CSP, then following two cases arise.

(i) The receiver does not know which pseudonym refers to which data owner. Therefore, sender and receiver anonymity is preserved. However, concealing the sender identity from the receiver leads an attack as described later in this section.
(ii) It creates an operational overhead for the receiver when he gets a list of pseudonyms from the CSP and the receiver does not know which pseudonym refers to which data owner. The receiver can scramble his attributes either with all pseudonyms one-by-one and send them to the CSP or the receiver can select a subset of pseudonyms, scramble his attributes with each of the pseudonym from subset one-by-one and send the queries to the CSP.

The scheme requires every user to get an anonymous encryption key A_O from trusted authority. Then the user is able to encrypt and upload the documents on CSP. However, we show that a user who knows the public parameters can generate a pseudonym, encrypt a message and upload the document on CSP. A user who has the knowledge of the public parameters $PK = (G,\ g,\ h = g^{\beta},\ \omega = e(g,g)^{\alpha})$ chooses a random element $t \in Z_p$, generates his pseudonym g^t and publishes it. The user scrambles the attributes included in $\mathcal{T}$ as $e(h, H(\lambda_j)^t) = e(g, H(\lambda_j))^{t\beta}\ \forall\ \lambda_j \in \mathcal{T}$. Now, this ciphertext can be uploaded to the CSP. Next, we show that the CSP can break the receiver anonymity, if he has the knowledge of the public parameter and attributes in the system. The CSP performs following steps to identify the attributes of a receiver who has submitted a search query to CSP.

CSP generates a fake pseudonym say $P_O = g^t$ for $1 \leq i \leq n$, where t is chosen randomly from Z_p. Using this fake pseudonym, CSP computes and prepares a list of scrambled attributes for each of the attribute in the system as follows. $\{e(h, H(\lambda_j))^t\}_{\lambda_j \in L} = \{e(g, H(\lambda_j))^{\beta t}\}_{\lambda_j \in L}\ for\ 1 \leq i \leq n.$

This list of values he stores in a set T'. When a data retriever $\mathcal{U}$ wants a list of pseudonyms from the CSP, then the CSP submits this list of pseudonyms in which the fake pseudonym generated by the CSP is also included. The data retriever $\mathcal{U}$ will not be able to detect if there is any fake pseudonym, as all pseudonyms are random values. Let us denote the list as L. The $\mathcal{U}$ chooses a subset L' of L, where $L' \subseteq L$. Then $\mathcal{U}$ scrambles his attributes using each of the pseudonym present in the L' as $K_{O_l,\Lambda_i} = \{\mathrm{e}(P_{Ol}, D'')\}_{j \in \Lambda_i} = \{\mathrm{e}(g^{t_l}, H(\lambda_j)^{\beta})\}_{j \in \Lambda_i} = \{\mathrm{e}(g, H(\lambda_j))^{\beta t_l}\}_{j \in \Lambda_i}$ using each pseudonym $P_{Ol} = g^{t_l}$ present in the set L'. $\mathcal{U}$ then submits these different sets of scrambled attributes $\langle K_{O_l,\Lambda_i}$ for each $P_{Ol} \in L' \rangle$ to the CSP. The CSP needs to compare each set

K_{O_l,Λ_i} with the set of pre-computed values T' that he has. Whenever he finds $K_{O_l,\Lambda_i} \subseteq T'$, then CSP identifies which attributes the $\mathcal{U}$ possesses. Once the CSP identifies the attributes of $\mathcal{U}$, then by comparing the remaining sets of scrambled attributes K_{O_l,Λ_i} with the stored access policies of other encrypted documents, the CSP can either uncover the hidden access policies of other encrypted documents. Therefore, the receiver anonymity of a ciphertext is revealed.

5 Improved Scheme

The security flaws in scheme [1] occur because of the use of pseudonym. We propose an improvement without using pseudonym, which retains the security claims of the scheme without increasing any overhead. The improved scheme has the following phases.

5.1 System Setup

The System setup phase is same as described in Sect. 3.1.

5.2 Key Generation

The Key Generation phase remains same as explained in Sect. 3.2. In addition to the Key Generation algorithm, the trusted authority publicizes a list of IDs and the mapping of IDs with the data owners owing that ID. We note that the secret parameter β scrambles the attributes in access policy, so the mapping of IDs with the data owners do not reveal any information about the sender and receiver of the encrypted documents stored in CSP.

5.3 Encryption

The data owner encrypts data M as per the access policy $\mathcal{T}$ by running the **Encrypt** algorithm as mentioned in Sect. 3.3. After that, the attribute scrambling algorithm, **AttrScm**, is applied to the ciphertext CT_0 for generating the ciphertext CT to be located in the cloud storage. We propose a modification in the **AttrScm** algorithm by removing the use of random value t. The data owner can use his secret encryption key for attribute scrambling as described below. S is the set of attributes to be included in access tree. For each attribute from set $S = \{\lambda_i, \cdots, \lambda_k \mid 1 \leq i \leq k \leq |L|\}$, the data owner calculates

$$\begin{aligned} K_{O,S} &= \{e(A_O, H(\lambda_j))\}_{\lambda_j \in S} \\ &= \{e(H(ID_O)^{\beta}, H(\lambda_j))\}_{\lambda_j \in S} \\ &= \{e(H(ID_O), H(\lambda_j))^{\beta}\}_{\lambda_j \in S} \end{aligned}$$

and assigns $scm_{att_x} \in K_{O,S}$ to leaf node x in $\mathcal{T}$ instead of λ_x corresponding to $attr_x$. This results in the access tree $\mathcal{T}'$. The new encrypted content CT to be stored is made as $CT = (\mathcal{T}', \tilde{C}, C, C'', \{C_y, C'_y\}_{y \in Y})$. After this phase, the data owner uploads CT to the storage managed by the CSP.

5.4 Data Access

Data query (Query). In this phase there is no need for a retriever to acquire a pseudonym of any data owner. When the retriever determines to retrieve a data with identity ID_O from the CSP then the retriever generates cryptographic index terms for corresponding attributes as

$$\begin{aligned} K_{O,\Lambda_i} &= \{e(D''_j, H(ID_O))\}_{j\in\Lambda_i} \\ &= \{e(H(ID_O), H(\lambda_j)^{\beta})\}_{j\in\Lambda_i} \\ &= \{e(H(ID_O), H(\lambda_j))^{\beta}\}_{j\in\Lambda_i} \end{aligned}$$

After that, the retriever submits the data request query in form of $K_{O,\Lambda'_i} \subseteq K_{O,\Lambda_i}$ to the CSP.

Data Retrieval (Retrieve). It is same as described in Sect. 3.4.

5.5 Decrypt

The decrypt operation is same as explained in Sect. 3.5.

6 Analysis

Theorem 1. The improved scheme provides sender and receiver anonymity.

Proof. We prove that the CSP or any other unintended receiver can not learn the sender or receiver identity. To break the sender and receiver anonymity the adversary needs to find out the value of sender's ID ID_O and λ_j from the scrambled attribute value $\{e(H(ID_O), H(\lambda_j)^{\beta})\}_{j\in\Lambda_i}$. For each of the attribute λ_j in the system and senders' identities $ID_{O,i}$, the following computed results are stored in CSP.

$$\{\{e(H(ID_{O,i}), H(\lambda_j))\}_{j\in\Lambda_i}\}_{1\leq i\leq n}.$$

Here, n is the number of users in the system and it is assumed that every user possesses a unique identity and a set of attributes. To compare the scrambled attributes stored along with the ciphertext the adversary needs to get the value of β, where β is the master key of the system which the adversary can not get. The use of β prevents any unintended retriever to generate the scrambled attributes index terms for which he has not got the private key. The complexity of getting the value of β from the public parameter $h = g^{\beta}$ is equivalent to that of solving the discrete logarithm problem, which is an intractable problem. Therefore, the adversary can not learn the sender or receiver identity from the hidden access policy or from the search query because of scrambled attributes.□

In addition to the security strengths of the improved scheme, the scheme reduces the communication and computation overheads, as the receiver neither requires a pseudonym from data owner or from CSP nor uses it in attributes scrambling procedure.

7 Conclusion

Anonymous attributes based schemes provide interesting features such as sender and/or receiver anonymity, privacy-preserved data access and unlinkability. We discussed a recently proposed anonymous CP-ABE scheme, which claims secure and efficient data retrieval with sender and receiver anonymity. We showed that the scheme suffers from security weaknesses, lacks sender and receiver anonymity. We proposed an improved scheme by removing the use of pseudonym that mitigates the weaknesses of the scheme and retains the claimed security features intact without adding any communication and computation overhead.

References

1. Koo, D., Hur, J., Yoon, H.: Secure and efficient data retrieval over encrypted data using attribute-based encryption in cloud storage. Comput. Electr. Eng. **39**, 34–46 (2013)
2. Sahai, A., Waters, B.: Fuzzy identity-based encryption. In: Cramer, R. (ed.) EUROCRYPT 2005. LNCS, vol. 3494, pp. 457–473. Springer, Heidelberg (2005). doi:10.1007/11426639_27
3. Bethencourt, J., Sahai, A., Waters, B.: Ciphertext-policy attribute-based encryption. In: Proceedings of IEEE Symposium on Security and Privacy (2007)
4. Goyal, V., Pandey, O., Sahai, A., Waters, B.: Attribute-based encryption for fine-grained access control of encrypted data. In: Proceedings of the ACM Conference on Computer and Communications Security, pp. 89–98 (2006)
5. Zhang, Y., Chen, X., Li, J., Wong, D.S., Li, H.: Anonymous attribute-based encryption supporting efficient decryption test. In: Proceedings of the ACM SIGSAC Symposium on Information, Computer and Communications Security, pp. 511–516 (2013)
6. Kapadia, A., Tsang, P.P., Smith, S.W.: Attribute-based publishing with hidden credentials and hidden policies. In: Proceedings of Network and Distributed System Security Symposium, pp. 179–192 (2007)
7. Yu, S., Ren, K., Lou, W.: Attribute-based content distribution with hidden policy. In: Proceedings of the IEEE Workshop on Secure Network Protocols, pp. 39–44 (2008)
8. Nishide, T., Yoneyama, K., Ohta, K.: Attribute-based encryption with partially hidden encryptor-specified access structures. In: Proceedings of Applied Cryptography and Network Security, pp. 111–129 (2008)
9. Li, J., Ren, K., Zhu, B., Wan, Z.: Privacy-aware attribute-based encryption with user accountability. In: Samarati, P., Yung, M., Martinelli, F., Ardagna, C.A. (eds.) ISC 2009. LNCS, vol. 5735, pp. 347–362. Springer, Heidelberg (2009). doi:10.1007/978-3-642-04474-8_28
10. Chaudhari, P., Das, M.L.: Cryptanalysis of searchable anonymous attribute based encryption. IACR Cryptology ePrint Archive 2016: 347 (2016)

Security on "Secure Remote Login Scheme with Password and Smart Card Update Facilities"

Marimuthu Karuppiah[1(✉)], Akshat Pradhan[2], Saru Kumari[3], Ruhul Amin[4], S. Rajkumar[1], and Rahul Kumar[5]

[1] School of Computer Science and Engineering, VIT University, Vellore 632014, Tamilnadu, India
marimuthume@gmail.com, rajkumars@vit.ac.in
[2] School of Information Technology and Engineering, VIT University, Vellore 632014, Tamilnadu, India
akat1296@gmail.com
[3] Department of Mathematics, CCS University, Meerut 632014, India
saryusiirohi@gmail.com
[4] Department of Computing Science and Engineering, Thapar University, Patiala 632014, Punjab, India
ruhul_amin@live.com
[5] SSV Degree College, Hapur 245101, Ghaziabad, India
ujjwalrahul@gmail.com

Abstract. Several password authentication schemes utilizing smart cards have been proposed in the literature. Recently Kumar et al. proposed a new authentication scheme to access remote server over insecure channels. They also claimed that their scheme is able to resist various attacks. However, in this paper we demonstrate that Kumar et al. scheme is still vulnerable to various malicious attacks and is aslo unable to provide several essential security properties.

Keywords: Denial of Service (DoS) · User anonymity · Password guessing · Forgery attack · Forward secrecy

1 Introduction

Several authentication scheme have been proposed for remote user authentication in the traditional client-server scenario. Lamport [1] was the first to propose remote authentication using one-way hash function. However, Lamport's scheme was found to be vulnerable to stolen verifier attack. Later several authentication schemes were proposed [2–15]. Regrettably, many of these schemes [16–18] are often found to be vulnerable. Karupiah and Saravanan [19] analyzed the scheme in [20] and showed that it is vulnerable to several attacks. They then proposed a new scheme. Wang et al. in [21] proved that the schemes in [18,22] are vulnerable to several malicious attacks. They then presented an enhanced scheme

© Springer Nature Singapore Pte Ltd. 2017
D. Giri et al. (Eds.): ICMC 2017, CCIS 655, pp. 26–33, 2017.
DOI: 10.1007/978-981-10-4642-1_3

to mitigate the vulnerabilities of [18,22]. Wang et al. [23] analyzed Yang et al. [24] and Hsieh-Leu [16] schemes and found that their schemes were susceptible to smart card loss attack. Wang et al. [23] then, proposed an efficient scheme to overcome Yang et al. [24] and Hsieh-Leu [16] schemes vulnerabilities. Ruhul et al. [7] illustrated the weaknesses of the scheme in [21] and also proposed an improved scheme.

Kumar et al. [25], in 2016, for mobile networking scenario proposed a scheme and claimed that their scheme could resist several attacks and provide various security properties. However, after careful analysis we discerned that Kumar et al. scheme is vulnerable to various attacks such as replay attack, offline password guessing attack, Denial of Service (DoS) attack and session key disclosure attack. Moreover, the scheme was unable to provide essential security properties such as forward secrecy and user anonymity.

Roadmap of the paper: The rest of the paper is organised as follows. Section 2 provides a brief overview of Kumar et al. scheme. In Sect. 3 we cryptanalyze Kumar et al. scheme. In Sect. 4, we make the conclusion.

2 Overview of the Scheme in [25]

This section briefs the scheme in [25]. It is divided into five parts. These parts are explained extensively as follows. The nomenclature of this paper is summarized in Table 1.

2.1 Nomenclature

Table 1. Nomenclature

Notations	Descriptions
Pw_i	User password
Id_i	User identity
S	Server
UT	User table
x_s	Master key of S
r_s	Random nonce of S
r_i	Random nonce of user
$\oplus$	Bitwise XOR operation
$h(\cdot)$	Hash function
$\|\|$	Concatanation

2.2 Registration Phase

The steps for the registration phase are mentioned below.

1. User selects an identity Id_i and submits $\{Id_i, MNUM\}$ with c_i to S, where c_i is an individual credential data and $MNUM$ is the legal mobile number of the user.
2. When S receives $\{Id_i, MNUM\}$, it computes $REG_i = h(Id_i||x_s)$ and sends it to $MNUM$ securely. Then, S keeps an user table (as shown in Table 2) which is comprising elements $\{Id_i, MNUM\}$ and c_i.
3. When the user receives REG_i, he computes $a_i = h(Id_i||Pw_i||REG_i)$, $b_i = REG_i \oplus Pw_i)$. Then, he stores $\{Id_i, a_i, b_i\}$ into smart card and sends $\{Id_i, b_i, CUN_i\}$ to S, where CUN_i is the unique number of the smart card.
4. S receives $\{Id_i, b_i, CUN_i\}$, then checks $Id_i \in UT$ or not. If the condition holds, S inserts $\{CUN_i, b_i\}$ into UT.

Table 2. User table of the server S

User ID	Variable	$MNUM$	CUN	Credential
Id_1	b_1	9894567	CUN_1	c_1
Id_2	b_2	9904558	CUN_2	c_2
Id_3	b_3	9704956	CUN_3	c_3
...	...	...	...	...
...	...	...	...	...
Id_n	b_n	9774511	CUN_n	c_n

2.3 Login and Authentication Phase

The detailed steps of this phase are as follows

1. At card reader machine, user inserts the card and keys Pw_i. After receiving the password, the reader calculates $REG_i^* = b_i \oplus Pw_i$, $a_i^* = h(Id_i||Pw_i||REG_i^*)$ and verifies $a_i^* \stackrel{?}{=} a_i$. If not, the login process is ended. Else, the user has entered the correct password. Then, reader generates a nonce r_i and finds $c_i = h(Id_i||r_i||Pw_i)$ and $d_i = r_i \oplus Pw_i$. The reader then transmits $\{Id_i, c_i, d_i\}$ to the server.
2. When S receives $\{Id_i, c_i, d_i\}$, it checks $Id_i \in UT$ or not. If not, session is ended. Else, S sends a OTP (one-time password) to $MNUM$ of the User. After the user receives the OTP, he sends it to S. Then, S checks the OTP verification. If the verification is not true, the session is ended. Else, it computes $REG_i^* = h(Id_i||x_s)$, $Pw_i^* = b_i \oplus REG_i^*$, $r_i^* = d_i \oplus Pw_i$, $c_i^* = h(Id_i||r_i^*||Pw_i^*)$ and verifies $c_i^* \stackrel{?}{=} c_i$. If not, the session is ended.
3. S now generates a nonce r_s and finds $g_i = r_i^* \oplus r_s$, $f_i = h(Id_i||r_i^*||r_s||REG_i^*)$. It then sends $\{f_i, g_i\}$ to the user.

4. After receiving $\{f_i, g_i\}$, the user derives $r_s^* = g_i \oplus r_i$, $f_i^* = h(Id_i||r_i||r_s^*||REG_i)$ and verifies whether $f_i^* \stackrel{?}{=} f_i$. If the verification fails, it ends the session. Otherwise, both compute the session key $SK = h(Id_i||REG_i||r_i||r_s)$ and start the secure session.

2.4 Password Update Phase

The steps for the password change process are mentioned below.

1. At card reader machine, user inserts the card and keys Pw_i. After receiving Pw_i, the reader finds $REG_i = b_i \oplus Pw_i$, $a_i^* = h(Id_i||Pw_i||REG_i^*)$ and checks whether $a_i^* \stackrel{?}{=} a_i$. If true, the user is prompted to enter the new password Pw_i^{new} to the user.
2. After receiving Pw_i^{new}, the reader computes $b_i^{new} = REG_i \oplus Pw_i^{new})$, $a_i^{new} = h(Id_i||Pw_i^{new}||REG_i)$ and substitutes b_i, a_i with b_i^{new}, a_i^{new} in the smart card. Hence, the password has been updated successfully.

2.5 Forgot Password Recover Phase

The steps for recovering the user's password are as follows.

1. User submits $\{Id_i, MNUM\}$ to S.
2. When receiving $\{Id_i, MNUM\}$, S verifies $Id_i \in UT$ and $MNUM \in UT$ or not. If true, S finds $REG_i = h(Id_i||x_s)$ and $Pw_i = b_i \oplus REG_i$. Otherwise, the request is terminated.
3. S then sends Pw_i to the user's $MNUM$ securely.

2.6 Smart Card Revocation Phase

The steps to acquire a new smart card without re-registration are as follows.

1. The user submits $\{Id_i, MNUM\}$ and his personal credentials to the server.
2. After receiving $\{Id_i, MNUM\}$, S verifies the validity of the user on the basis of the personal credentials and $\{Id_i, MNUM\}$. If the check holds, S computes $REG_i = h(Id_i||x_s)$ and transmits it to $MNUM$ of the user securely.
3. After receiving REG_i, the user computes $a_i^{new} = h(Id_i||Pw_i^{new}||REG_i)$ and $b_i^{new} = REG_i \oplus Pw_i^{new})$.
4. The user then acquires a new smart card containing $\{Id_i, a_i^{new}, b_i^{new}\}$. The user then sends $\{Id_i, b_i^{new}, CUN_i\}$ to the server.
5. After receiving $\{Id_i, b_i^{new}, CUN_i\}$, S checks whether the Id_i exists in the user table. If it does, S further enters b_i^{new} and $CUN_i\}$ into the table.

3 Cryptanalysis of the Scheme in [25]

This section proves that the scheme in [25] is susceptible to various types of attacks. We establish the following two assumptions. Note that the assumptions are relatively reasonable and have also been used in recent related works [26–31].

1. Adversary has absolute control on the insecure public medium. Therefore, he can modify, inject, delete and block messages transmitted over the public channel [32].
2. The secret stored data may be extricated by the adversary from the lost/stolen smart card via side channel attacks [33–35].

Thus the adversary can extricate the security credentials $\{Id_i, a_i, b_i\}$ from the lost/stolen smart card. He can also trap the messages such as m_1=$\{Id_i, c_i, d_i\}$ and m_2=$\{f_i, g_i\}$ between S and user.

3.1 Lack of user anonymity

In scheme [25], the Id_i is sent as a palindrome in m_1=$\{Id_i, c_i, d_i\}$ to S. Therefore, an adversary can identify a particular user and track his login history. Hence, the scheme of Kumar et al. is not confering the user anonymity feature.

3.2 Incorrect Password Change Process

In scheme [25], the credentials associated with Pw_i kept in the smart card are b_i and a_i. After a successful password update, the smart card is updated with the parameters b_i^{new} and a_i^{new}. However, recall that the parameter b_i is also stored in the User table at the server side. Moreover, there is no updation message sent to the remote server. Therefore, the user is denied services permanently if he tries to login after a successful password change process. Hence, we show that Kumar et al's incorrect password change phase culminates into a Denial of Service (DoS) attack.

3.3 Susceptible to Off-Line Password Guessing Attack

We assume that the login request message m_1=$\{Id_i, c_i, d_i\}$ is intercepted by the adversary during any login and authentication session. The, the user's password can be acquired as follows.

1. Adversary guesses the password Pw_a.
2. Computes $r_i^{'} = d_i \oplus Pw_a$
3. Computes $c_i^{'} = h(Id_i||r_i^{'}||Pw_a)$
4. Verify $c_i \stackrel{?}{=} c_i^{'}$. If verification does not hold, reiterate steps 1–4 till the correct password is found.

If $c_i \stackrel{?}{=} c_i^{'}$ is true, then this implies that $Pw_i = Pw_a$ and hence, the adversary has successfully obtained the user's password. Thus, we prove that the scheme in [25] is susceptible to offline password guessing attack.

3.4 Disclosure of Session Key

As discussed in Sect. 3.3, the adversary can discover the random number r_i and password Pw_i of user. Furthermore, he has $\{Id_i, a_i, b_i\}$ in accordance with our assumption 2. Thus, he can derive $REG_i = b_i \oplus Pw_i = REG_i \oplus Pw_i \oplus Pw_i$ where $b_i = REG_i \oplus Pw_i$. Moreover, he has trapped the message m_2=$\{f_i, g_i\}$ and therefore further derives $r_s = g_i \oplus r_i$ where $g_i = r_i \oplus r_s$. Thus, the adversary can deduce $SK = h(Id_i||REG_i||r_i||r_s)$.

3.5 Absence of Perfect Forward Secrecy

As discussed in Sect. 3.4, the session key is disclosed for i^{th} session. Note that Pw_i, Id_i as well as REG_i are static parameters for all the sessions. We assume that the attacker has intercepted the messages $\{Id_i, c_{i+1}, d_{i+1}\}$ as well as $\{f_{i+1}, g_{i+1}\}$ for the $i + 1^{th}$ session. He then computes the user's random number r_{i+1} (random value of $i + 1^{th}$ session) as discussed in Sect. 3.3. He further derives the server's random number r_s for the $i+1^{th}$ session as discussed in Sect. 3.4. Hence, the attacker deduces $SK = h(Id_i||REG_i||r_{i+1}||r_s)$ for the $i + 1^{th}$ session. Thus, the scheme in [25] does not confer the property of forward secrecy.

3.6 Replay Attack

It is clear that there is no mechanism for the remote server to verify the freshness of data in the user's login request message m_1=$\{Id_i, c_i, d_i\}$. Hence, any previously legitimate login request can be replayed by the attacker to get login as a valid user, and remote server cannot detect this malicious behavior and will respond to user (actually Attacker) as usual. Therefore, the scheme of Kumar et al. is susceptible to replay attack.

4 Conclusion

In this paper we analyzed the scheme coined by Kumar et al. scheme for remote login and pointed out that Kumar et al. scheme is susceptible to several malicious attacks like offline password guessing attack, Denial of Service (DoS) attack and replay attack. Furthermore, our analysis revealed that Kumar et al. scheme is unable to provide crucial security features such as perfect forward secrecy and user anonymity. Moreover, their password change process was found to be inefficient and thus Kumar et al. scheme is unsuitable for practical applications.

References

1. Lamport, L.: Password authentication with insecure communication. Commun. ACM **24**(11), 770–772 (1981)
2. Amin, R.: Cryptanalysis and an efficient secure id-based remote user authentication scheme using smart card. IJCA 75, 1149–1157. Citeseer (2013)

3. Amin, R., Biswas, G.P.: Anonymity preserving secure hash function based authentication scheme for consumer USB mass storage device. In: IEEE 2015 Third International Conference on Computer, Communication, Control and Information Technology (C3IT), pp. 1–6 (2015)
4. Amin, R., Biswas, G.P.: Design and analysis of bilinear pairing based mutual authentication and key agreement protocol usable in multi-server environment. Wirel. Pers. Commun. **84**, 439–462 (2015)
5. Amin, R., Biswas, G.P.: A novel user authentication and key agreement protocol for accessing multi-medical server usable in TMIS. J. Med. Syst. **39**(3), 1–17 (2015)
6. Amin, R., Biswas, G.P.: Remote access control mechanism using rabin public key cryptosystem. In: Mandal, J., Satapathy, S., Kumar Sanyal, M., Sarkar, P., Mukhopadhyay, A. (eds.) Information Systems Design and Intelligent Applications, pp. 525–533. Springer, Heidelberg (2015)
7. Amin, R., Maitra, T., Rana, S.P.: An improvement of wang. et. al'.s remote user authentication scheme against smart card security breach. Int. J. Comput. Appl. **75**(13), 37–42 (2013)
8. Giri, D., Maitra, T., Amin, R., Srivastava, P.: An efficient and robust rsa-based remote user authentication for telecare medical information systems. J. Med. Syst. **39**(1), 1–9 (2015)
9. He, D., Kumar, N., Chilamkurti, N.: A secure temporal-credential-based mutual authentication and key agreement scheme with pseudo identity for wireless sensor networks. Inf. Sci. **321**, 263–277 (2015)
10. He, D., Kumar, N., Chilamkurti, N., Lee, J.H.: Lightweight ECC based RFID authentication integrated with an ID verifier transfer protocol. J. Med. Syst. **38**(10), 1–16 (2014)
11. Islam, S.H.: A provably secure ID-based mutual authentication and key agreement scheme for mobile multi-server environment without ESL attack. Wirel. Pers. Commun. **79**(3), 1975–1991 (2014)
12. Islam, S.H.: Design and analysis of a three party password-based authenticated key exchange protocol using extended chaotic maps. Inf. Sci. **312**, 104–130 (2015)
13. Islam, S., Biswas, G.P., Choo, K.K.R.: Cryptanalysis of an improved smartcard-based remote password authentication scheme. Inf. Sci. Lett. **3**(1), 35–40 (2014)
14. Islam, S., Khan, M.K., Obaidat, M., Muhaya, F.: Provably secure and anonymous password authentication protocol for roaming service in global mobility networks using extended chaotic maps. Wirel. Pers. Commun. **84**, 2013–2034 (2015)
15. Kumari, S., Khan, M.K.: Cryptanalysis and improvement of a robust smart-card-based remote user password authentication scheme. Int. J. Commun. Syst. **27**, 3939–3955 (2013). doi:10.1002/dac.2590
16. Hsieh, W.B., Leu, J.S.: Exploiting hash functions to intensify the remote user authentication scheme. Comput. Secur. **31**(6), 791–798 (2012)
17. Kumari, S., Khan, M.K., Li, X.: An improved remote user authentication scheme with key agreement. Comput. Electr. Eng. **40**(6), 1997–2012 (2014)
18. Ku, W.C., Chen, S.M.: Weakness and improvement of an efficient password based remote user authentication scheme using smart cards. IEEE Trans. Consum. Electron. **50**(1), 204–207 (2004)
19. Karuppiah, M., Saravanan, R.: A secure remote user mutual authentication scheme using smart cards. J. Inf. Secur. Appl. **19**(4–5), 282–294 (2014). doi:10.1016/j.jisa.2014.09.006
20. Ramasamy, R., Muniyandi, A.P.: New remote mutual authentication scheme using smart cards. Trans. Data Priv. **2**(2), 141–152 (2009)

21. Wang, X.M., Zhang, W.F., Zhang, J.S., Khan, M.K.: Cryptanalysis and improvement on two efficient remote user authentication scheme using smart cards. Comput. Stan. Interfaces **29**(5), 507–512 (2007)
22. Yoon, E.J., Ryu, E.K., Yoo, K.Y.: Further improvement of an efficient password based remote user authentication scheme using smart cards. IEEE Trans. Consum. Electron. **50**(2), 612–614 (2004)
23. Wang, D., Ma, C.G., Zhang, Q.M., Zhao, S.: Secure password-based remote user authentication scheme against smart card security breach. J. Netw. **8**(1), 148–155 (2013)
24. Yang, G., Wong, D.S., Wang, H., Deng, X.: Two-factor mutual authentication based on smart cards and passwords. J. Comput. Syst. Sci. **74**(7), 1160–1172 (2008)
25. Kumar, R., Amin, R., Karati, A., Biswas, G.P.: Secure remote login scheme with password and smart card update facilities. In: Das, S., Pal, T., Kar, S., Satapathy, S., Mandal, J. (eds.) Proceedings of the 4th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA). Advances in Intelligent Systems and Computing (AISC), vol. 404, pp. 495–505. Springer, Heidelberg (2015). doi:10.1007/978-81-322-2695-6-41
26. Karuppiah, M., Saravanan, R.: A secure authentication scheme with user anonymity for roaming service in global mobility networks. Wirel. Pers. Commun. **84**(3), 2055–2078 (2015)
27. Karuppiah, M., Saravanan, R.: Cryptanalysis and an Improvement of New Remote Mutual Authentication Scheme using Smart Cards. Journal of Discrete Mathematical Sciences and Cryptography **18**(5), 623–649 (2015)
28. Wu, F., Xu, L., Kumari, S., Li, X., Das, A.K., Khan, M.K., Karuppiah, M., Baliyan, R.: A novel and provably secure authentication and key agreement scheme with user anonymity for global mobility networks. Netw. Secur. Commun. **9**, 3527–3542 (2016). doi:10.1002/sec.1558
29. Kumari, S., Karuppiah, M., Li, X., Wu, F., Das, A.K., Odelu, V.: A Secure Trust-Extended Authentication Mechanism for VANETs. Security and Communication Network (2016)
30. Karuppiah, M., Kumari, S., Das, A.K., Li, X., Wu, F., Basu, S.A.: A secure lightweight authentication scheme with user anonymity for roaming service in ubiquitous networks. Secur. Commun. Netw. **9**, 4192–4209 (2016)
31. Karuppiah, M.: Remote user authentication scheme using smart card: a review. Int. J. Internet Protoc. Technol. **9**, 107–120 (2016)
32. Xu, J., Zhu, W.T., Feng, D.G.: An improved smart card based password authentication scheme with provable security. Comput. Stand. Interfaces **31**(4), 723–728 (2009)
33. Kocher, P., Jaffe, J., Jun, B.: Differential power analysis. In: Wiener, M. (ed.) Advances in Cryptology – CRYPTO' 99. LNCS, vol. 1666, pp. 388–397. Springer, Heidelberg (1999). doi:10.1007/3-540-48405-1_25
34. Messerges, T.S., Dabbish, E.A., Sloan, R.H.: Examining smart-card security under the threat of power analysis attacks. IEEE Trans. Comput. **51**(5), 541–552 (2002)
35. Ma, C.G., Wang, D., Zhao, S.D.: Security flaws in two improved remote user authentication schemes using smart cards. Int. J. Commun. Syst. **27**, 2215–2227 (2012). doi:10.1002/dac.2468

Design of Secure and Efficient Electronic Payment System for Mobile Users

Prerna Mohit[1(✉)], Ruhul Amin[2], and G.P. Biswas[1]

[1] Indian Institute of Technology (ISM), Dhanbad 826004, India
prernamohit@outlook.com
[2] Thapar University, Patiala 147004, India

Abstract. The recent advancement in smart phones and its widespread popularity switches the users of traditional computing to mobile computing. In addition, to facilitate users, hand held devices such as a mobile phone application for the payment method should be accepted for practical implementation. Recently, Yang et al. proposed an electronic payment protocol using payment gateway and claims that this scheme is suitable for cloud computing, where payment gateway is placed in the cloud area and all the communication between user, merchant, bank is performed via the payment gateway. However, it is known that cloud server is not considered as completely secured entity. Hence, by putting payment gateway on cloud server the author is endangering the security of system as a consequence it is not suitable for cloud environment. In this paper, we propose an efficient electronic payment protocol for mobile environment where mobile users can directly communicate with the merchant. It has been shown that our protocol has better security performance in terms of different attacks.

Keywords: e-payment · Mobile commerce · Security · Symmetric key

1 Introduction

With the rapid development of online shopping, the demand of secure payment system is imperative and increased with time. In electronic transaction application, people mainly use mobile device to deal with the transaction due to user friendly services. One of the objectives of electronic payment transactions is to provide security to the customer during the process of the transaction as it is performed over a public channel. In order to protect the data from a malicious action, that may cause loss and theft of the customers money. An efficient electronic payment protocol is proposed. The use of online payment systems was mainly among banking institutions. During the same time, credit cards and ATM's were first introduced to customers. The exponential growth of the Internet has helped the development of online payment systems and has changed the way consumers do business. The electronic payment system is considered as an integral part of any E-commerce system and categorized as Business-to-Business (B2B), Business-to-Consumer (B2C), Consumer-to-Business (C2B), and Consumer-to-Consumer (C2C) transaction.

© Springer Nature Singapore Pte Ltd. 2017
D. Giri et al. (Eds.): ICMC 2017, CCIS 655, pp. 34–43, 2017.
DOI: 10.1007/978-981-10-4642-1_4

1.1 Study and Review on Existing Research

For the protection of online payment transactions, there are different type of electronic payment systems have been suggested by researchers and scientist in [1–3,13]. In 2001, Chari et al. [4] shows that mobile communication is different from electronic communication as the underlying technologies are different. Therefore, the idea of security in mobile commerce should be different from that of the electronic commerce. However, there are some of well-known existing protocols for secure electronic payment exists such as secure electronic transaction (SET) [5], Internet Key Protocol (iKP) [6] and these protocol are successfully implemented over Internet. However, Kungpisdan et al. [7] justified that SET and iKP payment protocols are not suitable for mobile communication payment transaction and only be suitable for electronic communication for payment transaction. Then, Tellez et al. in [8] also supports in [7] that existing SET and iKP payment protocols are inapplicable for mobile payment transaction in wireless network due to their heavy computational and communication operations and proposed an improved protocol. Then, Kungpisdan et al. [9] proposed an enhanced version of [7]. In 2008 Fun et al. [10] discusses a new protocol for personal mobile payment, which is based on a client centric model using symmetric key [10] and also claims that the protocol achieves privacy protection for the pair of communication entity. Isaac et al. [11] proposes a secure payment transaction protocol using payment gateway, where the client and merchant always communicate via a payment gateway in order to exchange message. Later on, Yang et al. in [12] shows that Isaac et al.'s in [11] scheme does not provides non-repudiation and suffers from the key management problem. Hence, Yang et al. [12] proposed a new mobile payment protocol and claimed that it is suitable for cloud computing environment.

1.2 Organization of the Paper

Section 2 gives the preliminary for the protocol, which also includes review of Yang et al. protocol and its weaknesses. Section 3 presents the proposed protocol for e-payment system. The security and performance evaluation of our protocol are given in Sect. 4. Finally, we conclude the paper in Sect. 5.

2 Preliminary

This section explains some of the concepts used in order to understand our protocol.

2.1 RSA Digital Signature

1. Key Generation: Randomly select two prime numbers p, q and compute $n = p * q$, $\phi(n) = (p-1) \times (q-1)$.
   Choose e such that $\gcd(\phi(n), e) = 1$.
   Compute $d = e^{-1} mod(\phi(n))$

2. Signature: Compute $Sig = M^d mod\ n$
   Send Sig, M
3. Verification: $M' = Sig^e mod\ n$
   check $M = M'$; if correct accept; otherwise reject

2.2 Roles

The proposed scheme, consists of five entities: Client (C), Merchant (M), Payment Gateway (PG), Issuer (I) and Acquirer (A). They are introduced as follows.

- Merchant: A person or company, who is selling its goods.
- Client: A person or organization using the services of merchant.
- Payment gateway: Use in the payment transaction between the bank and merchant/client.
- Issuer: The client's bank.
- Acquirer: The merchant's bank.

Table 1. List of the symbols used in Yang et al.'s scheme

Notation	Meaning
NID_C	The temporary identity of the client
ID_i	The identity of the participant i
$TInfo$	The transaction information includes time, date, and the serial number
$Price$	The amount of the payment
m	The payment information computed by $m = \langle NID_C, TInfo, Price \rangle$
$SRequest$	The signature request
TS_i	The timestamp generated by the participant i
$Issuer_{ID}$	The identity of the issuer
$Acquirer_{ID}$	The identity of the acquirer
Stt	The state of a transaction
KS_{A-B}	The session key shared between A and B

2.3 Review of Yang et al. e-payment System [12]

We briefly review Yang et al.'s e-payment protocol, where all the transactions are performed via payment gateway. The detail of the scheme is described below. The list of notations used in this paper is given in Table 1.

Step 1 C → PG: NID_C, A
PG → M: NID_C, A
M → PG: $TInfo$
PG → C: $TInfo$

Step 2 C → PG: $SRequest = (h(TInfo), ID_C, NID_C, h(m), Price, TS_C)_{KS_{C-I}}$
PG → I: $SRequest$
I → PG: $(S)_{KS_{C-I}}$
Decrypt $[SRequest]$; check TS_C
$S = h(m)^d mod\ n$
PG → C: $(S)_{KS_{C-I}}$

Step 3 C → PG: $(S, m, h(TInfo), TS_C, Issure_ID)_{KS_{C-PG}}$
Decrypt and get $(S, m, h(TInfo), TS_C, Issure_{ID})$
PG → M: $(S, m, h(TInfo), TS_C, Issuer_{ID})_{KS_{M-PG}}$

Step 4 M → PG: $(S, m, h(TInfo), ID_M, TS_M, Issuer_{ID}, Acquirer_{ID})_{KS_{M-PG}}$

Step 5 PG → I: $(S, NID_C, ID_M, h(TInfo), Price, Acquirer_{ID})$ using private network
PG → A: $(h(TInfo), Price, ID_M, Issuer_{ID})$
Step 6 I → PG: $PResponse, h(TInfo)$
A → PG: $Stt, h(TInfo)$

Step 7 PG → C: $PResponse, h(TInfo)$
PG → M: $Stt, h(TInfo)$

2.4 Weakness of Yang et al.

We found that Yang et al.'s scheme is not suitable for cloud computing, as it was claimed by Yang that her scheme provides anonymity for cloud client. The details are discussed below.

- It is assumed that the Payment Gateway is in the area of cloud and due to this the protocol can be implemented in cloud environment.
- The payment gateway plays very important role as all the entities communicate through the payment gateway for payment related request. Moreover, the client cannot communicate directly with the merchant to process the Payment request.
- In short, the security of Yang et al.'s scheme directly depends on the security of Payment Gateway.

However, it is known that the cloud servers are not considered as secure [14,15]. So, by putting the Payment gateway in cloud the security of transaction is becoming more dangerous.

3 Proposed Protocol for e-payment System

In this section, a new payment scheme is proposed for online transaction system. The proposed scheme consists of two phases, namely the set up phase and transaction phase. The detailed descriptions of each phase are given below. Table 2 introduces the notations used in our protocol.

Table 2. List of the symbols used

Notation	Meaning
NID_C	The temporary identity of the client
ID_i	The identity of the i'th participant
ID_P	The identity of the product
TID	The transaction information includes transaction time, date, and the serial number
PI	The payment information computed by $PI = h(TID \parallel Price \parallel h(OI))$
OI	The order information computed by $OI = h(ID_P \parallel h(Price) \parallel TID)$
T_i	The timestamp generated by i
$h()$	One-way hash function
Stt	The state of a transaction
K_{AB}	Secret key between A and B
VS	The value- subtraction
$PResponse$	The product response
$PRequest$	The product request
$VCRequest$	The value claim request
$VCResponse$	The value claim response

3.1 Proposed Architecture and Discussion

In Fig. 1, we have provided the architecture of e-payment, which consists of five entities, namely Client (C), Merchant (M), Payment Gateway (PG), Acquirer (A) and Issuer (I). The client requests for the product by looking on the merchant's web site. Additionally, the merchant provides product detail, including serial number, price, date, time to client. Now, the mobile client asks for the product request including the value need to be subtracted by bank and forwards it to merchant, where M keeps the product request and forwards the value claim request to payment gateway. The gateway performs some verification steps and forwards value substraction request to issuer. At the same time, Payment Gateway forwards some encrypted message to Acquirer. On receiving the value substraction request, issuer verifies it and sends value substraction respond to payment gateway and acknowledgement for payment gateway to A. Then, A forwards it to payment gateway. The payment gateway computes value claimed response and forwards it to M, where merchant verifies it and generates product response which is acceptable after verification.

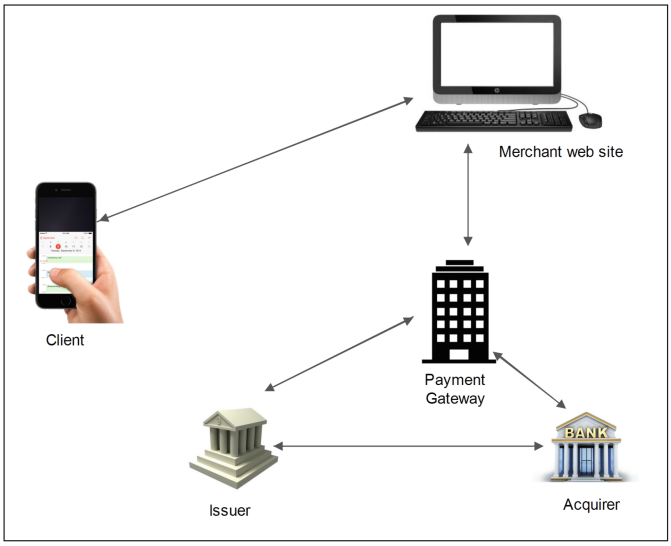


Fig. 1. Proposed model of e-payment mechanism

3.2 Setup Phase

In Setup phase, each entity the client, merchant, issuer, acquirer registers with the payment gateway to establish their secret key with gateway K_{CP}, K_{MP}, K_{IP}, K_{AP} respectively. Secret key is required to perform secure communication. In addition, the client and merchant also establish a secret key K_{CM} between them self.

The issuer as well as client use RSA signature to perform digital signature on the document using the private key. Note that the public key pair has been certified by a certificate authority.

3.3 Transaction Phase

Client starts the transaction by sending its temporary identity to the merchant. In the whole transaction process, the client can directly communicate with merchant while for communication with bank, the merchant as well as the client required the payment gateway to make the communication more simple. Detail description of the protocol is given below, where the symbol $A \rightarrow B : C$ means a message C is sent to B by A. The detail, description is shown in Fig. 2.

Step 1 C $\rightarrow$ M: NID_C
M $\rightarrow$ C: $\{ID_P, TID, Price\}_{K_{CM}}$

Step 2 C $\rightarrow$ M: $PRequest = \{TID, NID_C, OI, T_{C2}, VS\}_{K_{CM}}$
$VS = \{Sig, PI, ID_C, T_{c1}\}_{K_{CP}}$
$MD = h(PI)$, $Sig = MD^d mod\ n$

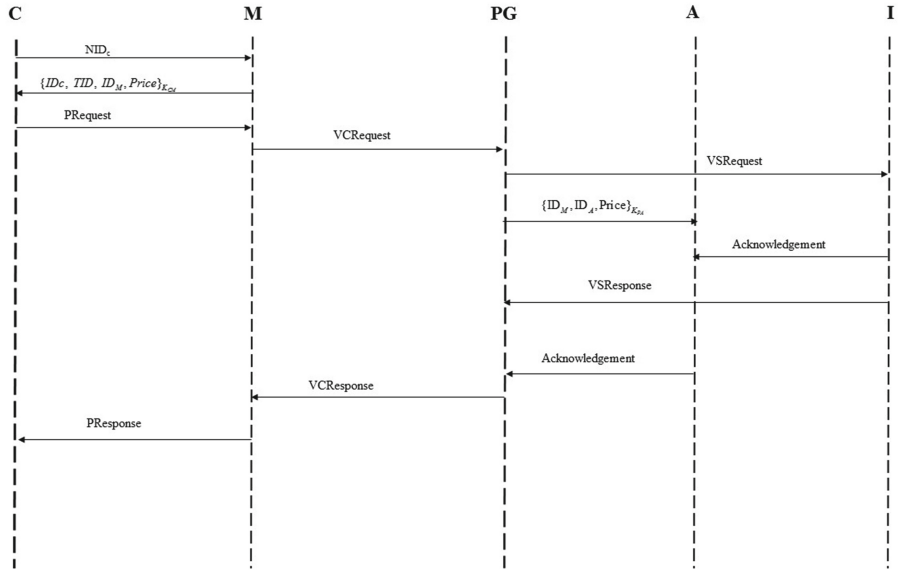


Fig. 2. The steps of the proposed protocol

Step 3 M → PG: $VCRequest = \{VS, ID_M, Price, TID, T_M\}_{K_{MP}}$
Decrypt $PRequest$; Check $|T_M - T_{C2}| \leq \Delta T$

Step 4 PG → I: $VSRequest = \{PI, Sig, Price, ID_C, ID_A\}_{K_{IP}}$
Decrypt $VCRequest$; Check $|T_{PG} - T_M| \leq \Delta T$
Decrypt VS; Check $|T_{PG} - T_{C2}| \leq \Delta T$ and $h(PI) = ?sig^e mod\ n$
PG → A: $\{ID_M, ID_A, Price\}_{K_{PA}}$

Step 5 I → A: *Acknowledgement*
Decrypt $(VSRequest)$; Check $Sig^e mod\ n = ?h(PI)$
Check ID_C in its database and find MN of client
I → C: OTP Request
C → I: OTP Response
A → PG: *Acknowledgement*

Step 6 I → PG: $VSResponse = \{stt, h(price), Sig_I\}$
$AD = h(ID_C \parallel MN)$
$MD_1 = h(AD)$; $Sig_I = MD_1^d mod\ n$

Step 7 PG → M: $VCRespond = \{stt, h(OI), Sig_I\}$
Check $h(OI) = ?$ Stored $h(OI)$
M → C: $PRespond = Sig_I$
verify $Sig_I^e mod\ n = ?h(h(ID_C \parallel MN))$

4 Security Analysis

This section discusses various types of attacks to analyze the security of the proposed protocol. The detail is described below

1. **Confidentiality**: In this scheme we always encrypt data before transferring it to the other communicating party. If adversary $\mathcal{A}$ interrupts between communication. $\mathcal{A}$ get the encrypted message which can not be decrypted without the key. Hence, confidentiality is always achieved.
2. **Non-repudiation**: The Issuer uses the client's signature to ensure that the legal person send the request to deduct the money from its account. The client also can verify the issuer signature. If there are some problems, the client as well as the issuer can not deny from the fact the signature is performed by them. Thus, non-repudiation is achieved.
3. **Replay attack**: We use timestamps, which is checked by the receiving party if the time stamps is not legal and not showing the valid time interval. For example, when the merchant receives $PRequest = \{TID, NID_C, OI, T_{C2}, VS\}_{K_{CM}}$. The merchant first decrypt it and check $|T_M - T_{C2}| \leq \Delta T$ i.e. if it is larger than mention time, then the merchant will discover that message will send by attacker $\mathcal{A}$. Therefore, the protocol can defend against replay attack.
4. **Insider attack**: As the communicated message are encrypted by the session key between sending and receiving party. So, only the two can see the message. For instance, let us consider that if merchant want to know the original identity of client ID_C, contain in the message VS, it is impossible for merchant as it is encrypted by client-issuer key.
5. **Anonymity**: The client identity ID_C is always kept secret during the communication and client use temporary identity NID_C which is session dependent for communication. Thus, it prevents client's Anonymity.
6. **Impersonation attack**: If attacker $\mathcal{A}$, interrupts the message of the client and trying to be like client by modifying its message $PRequest$ which contain VS where further contain Sig signature of the client. Which cannot be performed by $\mathcal{A}$. Thus, the protocol protects impersonation attack.

4.1 Performance Analysis

This section gives the computation cost comparison of our scheme with related scheme used in online transactions [11,12] as shown in Table 4. It is found that our scheme has less computation cost, then [11] but more the [12]. Moreover, Yang et al. uses private channel in order to communicate with bank. We do not consider concatenation, hash operation, as its computation is very less than symmetric encryption/decryption. The notation T_S refers to symmetric encryption/decryption (Table 3).

Table 3. Security comparison of proposed scheme with related schemes

Schemes	Isaac et al. [11]	Yang et al. [12]	Our protocol
Provide confidentiality	Yes	Yes	Yes
Provide integrity	Yes	Yes	Yes
Provide non-repudiation	No	Yes	Yes
Resist anonymity	No	Yes	Yes
Resist replay attack	Yes	Yes	Yes
Resist insider attack	Yes	Yes	Yes
Resist impersonation attack	No	No	Yes

Table 4. Computation cost comparison of the proposed scheme with related schemes

Schemes	Isaac et al. [11]	Yang et al. [12]	Our protocol
Client	4 T_S	3 T_S	2 T_S
Merchant	5 T_S	2 T_S	2 T_S
Gateway	3 T_S	2 T_S	4 T_S

5 Conclusion

This paper presents a new method for electronic payment system, which is the improvement of Yang et al. e-payment system. Our protocol withstands the security weaknesses found in Yang et al.'s scheme. In our implementation, the payment gateway acts as a proxy to communicate between bank and client/merchant. The security analysis shows that the proposed scheme can resist against various type of attacks.

References

1. Sun, P.-C., Liu, Y.-L., Luo, J.-J.: Perceived risk and trust in online group buying context. In: 2010 3rd International Conference on Information Management, Innovation Management and Industrial Engineering, vol. 3, pp. 660–663. IEEE (2010)
2. Tsai, M.-T., Cheng, N.-C., Chen, K.-S.: Understanding online group buying intention: the roles of sense of virtual community and technology acceptance factors. Total Qual. Manage. Bus. Excellence **22**(10), 1091–1104 (2011)
3. Buccafurri, F., Lax, G.: Implementing disposable credit card numbers by mobile phones. Electron. Commer. Res. **11**(3), 271–296 (2011)
4. Chari, S., Kermani, P., Smith, S., Tassiulas, L.: Security issues in M–commerce: a usage–based taxonomy. In: Liu, J., Ye, Y. (eds.) E-Commerce Agents. LNCS, vol. 2033, pp. 264–282. Springer, Heidelberg (2001). doi:10.1007/3-540-45370-9_16
5. Lu, S., Smolka, S.A.: Model checking the secure electronic transaction (set) protocol. In: Proceedings of the 7th International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems, pp. 358–364. IEEE (1999)

6. Harkins, D., Carrel, D.: The internet key exchange (ike). Technical report (1998)
7. Kungpisdan, S., Srinivasan, B., Le, P.D.: Lightweight mobile credit-card payment protocol. In: Johansson, T., Maitra, S. (eds.) INDOCRYPT 2003. LNCS, vol. 2904, pp. 295–308. Springer, Heidelberg (2003). doi:10.1007/978-3-540-24582-7_22
8. Isaac, J.T., Cámara, J.S.: Anonymous payment in a client centric model for digital ecosystems. In: 2007 Inaugural IEEE-IES Digital EcoSystems and Technologies Conference, pp. 422–427. IEEE (2007)
9. Kungpisdan, S., Srinivasan, B., Le, P.D.: A secure account-based mobile payment protocol. In: Proceedings of the International Conference on Information Technology: Coding and Computing, ITCC 2004, vol. 1, pp. 35–39. IEEE (2004)
10. Fun, T.S., Beng, L.Y., Roslan, R., Habeeb, H.S.: Privacy in new mobile payment protocol. In: Proceedings of World Academy of Science, Engineering and Technology, vol. 30, pp. 443–447. Citeseer (2008)
11. Isaac, J.T., Zeadally, S.: An anonymous secure payment protocol in a payment gateway centric model. Procedia Comput. Sci. **10**, 758–765 (2012)
12. Yang, J.-H., Lin, P.-Y.: A mobile payment mechanism with anonymity for cloud computing. J. Syst. Softw. **116**, 69–74 (2016)
13. HafizulIslam, S.K., Amin, R., Biswas, G.P., Obaidat, M.S., Khan, M.K.: Provably secure pairing-free identity-based partially blind signature scheme and its application in online e-cash system. Arab. J. Sci. Eng. **41**(8), 3163–3176 (2016)
14. Kandukuri, B.R., Rakshit, A., et al.: Cloud security issues. In: IEEE International Conference on Services Computing, SCC 2009, pp. 517–520. IEEE (2009)
15. Krutz, R.L., Vines, R.D.: Cloud Security: A Comprehensive Guide to Secure Cloud Computing. Wiley, Indianapolis (2010)

A Deep Learning Based Artificial Neural Network Approach for Intrusion Detection

Sanjiban Sekhar Roy[1(✉)], Abhinav Mallik[1], Rishab Gulati[1], Mohammad S. Obaidat[2,3], and P.V. Krishna[4]

[1] School of Computer Science and Engineering, VIT University, Vellore, India
sanjibanroy09@gmail.com, gulati.rishab5@gmail.com, abhinavmalik94@yahoo.com
[2] Fordham University, New York, USA
m.s.obaidat@ieee.org
[3] University of Jordan, Amman, Jordan
[4] Department of Computer Science, Sri Padmavati Mahila Visvavidyalayam, Tirupati, India
dr.krishna@ieee.org

Abstract. Security of data is considered to be one of the most important concerns in today's world. Data is vulnerable to various types of intrusion attacks that may reduce the utility of any network or systems. Constantly changing and the complicated nature of intrusion activities on computer networks cannot be dealt with IDSs that are currently operational. Identifying and preventing such attacks is one of the most challenging tasks. Deep Learning is one of the most effective machine learning techniques which is getting popular recently. This paper checks the potential capability of Deep Neural Network as a classifier for the different types of intrusion attacks. A comparative study has also been carried out with Support Vector Machine (SVM). The experimental results show that the accuracy of intrusion detection using Deep Neural Network is satisfactory.

Keywords: Security · Intrusions · Deep Neural Network · Support Vector Machine

1 Introduction

Intrusion Detection System [1, 2] is a type of security management system for computers and networks. It gathers and analyzes information from various areas within a computer or a network to identify possible security breaches, which include both intrusions (attacks from outside the organization) and misuse (attacks from within the organization). ID uses vulnerability assessment, developed to assess the security of a computer system or network. Data is considered to be the most important aspect of any organization. If the organization's data is secure, only then it can successfully carry out its operations. However, data have always been under a constant threat from external attacks. The hackers and crackers come up with new ways every day to destroy or steal the data that every organization holds so precious. In this paper, we have analyzed a

© Springer Nature Singapore Pte Ltd. 2017
D. Giri et al. (Eds.): ICMC 2017, CCIS 655, pp. 44–53, 2017.
DOI: 10.1007/978-981-10-4642-1_5

dataset containing information about the various attacks that have been carried out by the hackers and based on the parameters, an attempt to predict the kind of attack that will be used by the hacker, is carried out. The data set has been obtained from UCI machine learning repository. The data set is related to intrusion detection system (IDS) and in this work, a Deep learning [3] approach based on neural network has been adopted to predict different types of IDS attacks.

An Intrusion Detection System, popularly known as IDS, is a software that monitors the network for malicious activities or violations of policies regarding cybercrime and produces a report to the management system. IDS is related to network security just like a firewall, it differs from a firewall in the manner of looking for intrusions. The firewall looks at the outward intrusions in order to prevent them and limits the access between networks to prevent intrusion. On the other hand, IDS evaluates an intrusion that has already taken place and then sends an alarm signal. A lot of predictions has been accomplished using machine learning [4, 5, 12, 13, 15]. Also, several intrusion detection systems were proposed by several authors using roughest theory and other methods [7]. In this paper, we have used a multilayer feed forward network to represent a deep learning concept for IDS. The feed forward network includes input layers, about 400 hidden layer neurons and output neurons. The activation functions used are rectifier activation function and softmax activation function.

Deep learning has been used in this paper. It is a branch of machine learning that attempts to model higher level abstractions in data by using model architectures with non-linear transformations [6]. It is chosen since it focuses on computational models for information representation. It is implemented in such a way that it is able to display classification invariance with respect to a wide range of transformations and distortions. It enables us to train a network having a large set of observations and excerpt signals from this network. The deep learning algorithms use simple features in the lower layers and more complex features in the higher layers. Here, each hidden layer has statistical knowledge about the lower layers while higher layer representations are more complex. The network is trained using greedy layer-wise training which involves the training of the hidden layers one at a time in a bottom-up fashion. Deep learning has a myriad of applications. It is used in the medical field where robotics surgery is becoming a common trend, which relies extensively on tactile equipment. Deep learning is utilized for developing the robotic equipment. This may enable the doctors to move to a precision of a millimeter. Also, we can see the application of deep learning in the field of automotive in terms of self-driving cars, which apply the concepts of deep learning to emulate the senses of sight and hearing. It is also used in military forces in a country where a large number of military drones utilize the concept of deep learning to follow a moving target. Much research is required in this field as it is not yet fully functional. Currently, Google Brain is a technology used by Google that uses neural networks to recognize high level inputs only from watching unlabeled images from YouTube.

IDS set has been used in the Support Vector Machine (SVM) as well and the result is juxtaposed with the one obtained by using the Neural Network. The results obtained from the Support Vector Machine are complimentary to the ones obtained by using Neural Networks. Thus, it confirms that the results obtained are satisfactory.

2 Deep Neural Network

The neural network used is a multilayer feed forward neural network. In this network, the information moves in only one direction, forward, from the input nodes, through the hidden nodes (if any) and to the output nodes [8]. There are no cycles or loops in the network. Each neuron in one layer has direct connections to the neurons in the subsequent layers. It contains an input layer, a number of hidden layers and an output layer. The back propagation method is used for learning the weights of the network. The input layer has an identity function as its activation function. The output layer and the hidden layers may have rectifier or softmax activation function. Also, a multilayer neuron does not have a linear activation function in all its neurons. Some of its neurons might have a nonlinear activation function (Fig. 1).

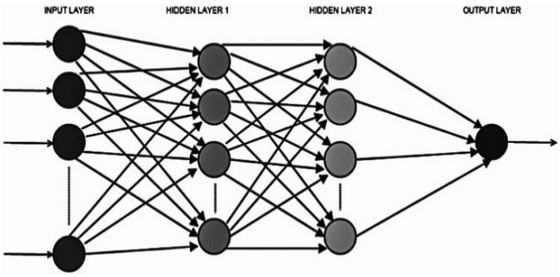


Fig. 1. Feed forward neural network [18]

Feed forward neural network is popular due to 2 factors:

(i) It has the ability to give very closely related approximations for complex multivariate nonlinear function directly from input values.
(ii) It has a strong modelling capability for a large class of natural and artificial phenomena.

However, in most of the practical scenarios, all parameters of a feed forward network need to be adjusted in a backward way which leads to creation of dependencies among various neurons in various layers.

Mean squared error (MSE) measures the average of the squares of the "errors", that is, the difference between the estimator and what is being estimated [9].

The mean square error is calculated in the following way:

$$\begin{aligned} \text{MSE} &= \text{RSS}/\text{N} \\ &\text{where MSE} - \text{Mean Squared Error} \\ &\text{RSS} - \text{Residual Sum of Squares} \\ &\text{N} - \text{Population Size} \end{aligned}$$

RSS is also known as Sum of Squared Residuals (SSR) and Sum of Squared Error (SSE). It is given by [9, 10],

$$\mathrm{RSS} = \sum (y_i - f(x_i))^2 \tag{1}$$

So, MSE is given by,

$$\frac{1}{N}\sum (y_i - f(x_i))^2 \tag{2}$$

The value of R^2 denotes how close the obtained result is to the expected regression line. R^2 can have a value within the range [0,1]. The higher value of R^2, the more accurate the obtained result is. It can be computed in the following way:

$$\begin{aligned} R^2 &= SS_R/SS_T \text{ where,} \\ SS_T &= \sum (y_i - f(y_i))^2 \end{aligned} \tag{3}$$

$$SS_R = \sum (\hat{y}_l - \overline{y})^2 \tag{4}$$

In some of the research experiments, another class of neural network is used which is known as deep belief network and is composed of Restricted Boltzmann Machines (RBMs) and uses a greedy layer by layer learning algorithm. However, the type of architecture used in this paper has a better approach since it provides discriminating powers for pattern classification by characterizing the posterior distributions of classes conditioned on the data. The following table contains definitions of the terms used here (Table 1).

3 Experimental Results and Outcome

The data set used in the experiment is the KDD Cup 1999 dataset which is a collection of simulated raw TCP dump data over an epoch of 9 weeks on a LAN. The training data has about 5 million connection records from seven weeks of network traffic and two weeks of testing data yielded around 2 million connection records. The training data have 22 of the total 29 attacks present in the test data. The known attack types are present in the training set while the novel attacks are additional attacks that are present in the test data set and not in the training data set. The attack types are grouped into 4 categories:

- DOS – Denial of Service (DoS) attack – e.g. syn flooding
- Probing – Surveillance and other probing – e.g. port scanning
- U2R – Unauthorized access to the root user privileges. e.g. Buffer overflow attacks
- R2L – Unauthorized access from a remote machine, e.g. password guessing.
- The training set has about 494,021 records from which 97,277 are normal, 391,458

are DOS attacks, 4107 are Probe, 1126 are R2L and 52 are U2R connections. Each connection has about 41 attributes describing different features of connection and a label assigned to each either as an attack type or normal. This data set was used originally in The Third International Knowledge Discovery and Data Mining Tools Competition, which was held in conjunction with KDD-99 The Fifth International Conference on Knowledge Discovery and Data Mining. This database contains a standard set of data to be audited, which includes a wide variety of intrusions simulated in a military network environment.

Table 1. Basics terminology [11, 16, 17]

Terminology	Meaning
Deep Learning	It is a class of machine learning techniques, based on a set of algorithms that use multiple layers with complex structures composed of non-linear transformations to model high level data
Deep Belief Networks	It is a probabilistic generative model composed of multiple layers of stochastic, hidden variables. The top two layers have undirected, symmetric connections. The lower layers have direct connections from above and as such receive top-down
Boltzmann Machine	It is a network of neuron like units that are symmetrically connected. They are concerned with making stochastic decisions about whether to be on or off
Restricted Boltzmann Machine	It consists of a layer of visible units and a layer of hidden units with no visible-visible and hidden-hidden connections
Deep Boltzmann Machine	It is a special kind of BM where hidden neurons are arranged in a deep layered manner. There exist no visible-visible or hidden-hidden connections within the same layer. This involves a connection between only the adjacent layers
Deep Neural Network	It is a multilayer network with many hidden layers. The weights in these networks are fully connected and pre-trained
Deep Auto Encoder	It is a special kind of deep neural network where the output target is the input itself. Deep Belief Networks or distorted training data are used to train the network
Distributed Representation	It is the representation of the data in such a way that it appears to be generated by interaction of various hidden factors. They form a basis for deep learning

3.1 Simulation Results

The data set that was used had response values in column 42 with losses being set as Cross Entropy in order to get classification model (Table 2). The input data set has been divided into two parts - training frame and validation frame. 75% of the data set has been assigned as the training frame and 25% of the data set has been assigned as the validation frame. Upon running the algorithm, a scoring history in the form a graph was obtained as shown below. The graph produced is between training and validation frame as x axis and epochs as the y axis. It depicts the similarity between the training and validation frame and that the model that has been created is correct (Fig. 2).

Table 2. Model parameters

Parameter	Value	Description
Response column	C42	Response column
Hidden	200,200	Hidden layer sizes (e.g. 100,100)
Seed	7069314529076090000	Seed for random numbers (affects sampling) - Note: only reproducible when running single threaded
Loss	Cross Entropy	Loss Function

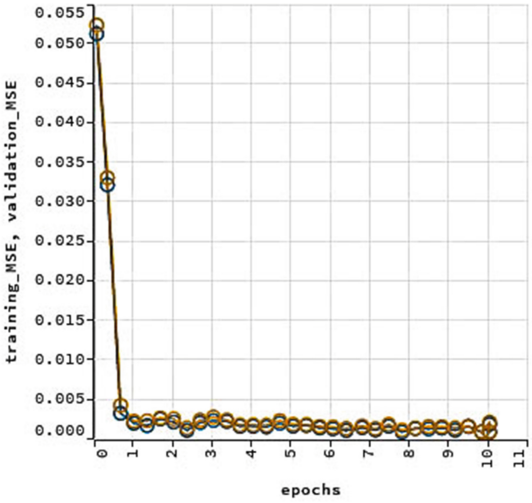


Fig. 2. Training and validation error of deep learning neural network

3.1.1 Experimental Outcome of Deep Neural Network

The activation functions used are rectifier activation function and softmax activation function (Table 3).

Table 3. Status of neurons

A	B	C	D	E	F	G	H	I	J	K	L
1	119	Input	0	0	-	-	-	-	-	-	-
2	200	Rectifier	0	0	0.6364	0.4589	0	0.0015	0.1133	0.4745	0.1081
3	200	Rectifier	0	0	0.6957	0.4432	0	0.0028	0.0984	0.9853	0.0676
4	23	Softmax	0	0	0.9427	0.2252	0	0.3050	0.4532	−0.2707	0.0619

A – Layer, B – Union, C – Type, D – L1, E – L2, F – Mean Rate, G – rate_RMS, H – Momentum, I – Mean Weight, J – Weight RMS, K – Mean Bias, L – Bias RMS.

The rectifier is an activation function defined as,

$$f(x) = \max(0, x) \tag{5}$$

Where x is the input.
It can also be expanded to include Gaussian noise given as,

$$f(x) = \max(0,\ x + N(0, \sigma(x))) \tag{6}$$

Softmax function is a generalization of logistic function that squashes a M-dimensional vector z of arbitrary real values to a M dimensional vector $\sigma(z)$ of real values in the range (0,1) that add up to 1. The function is given by,

$$P(y = j|x) = \frac{e^{x^T w_j}}{\sum e^{x^T w_k}} \tag{7}$$

3.1.2 Output - Training Metrics

This includes the output obtained from the training set. The following training metrics depict the efficacy of the implementation (Table 4).

Table 4. Output training metrics

Parameters	Values
Description	Metrics reported on temporary training frame with 9910 samples
Model_category	Multinomial
Scoring Time	1442054607700
MSE	0.000961
R^2	0.999944
Logloss	0.012146

The Mean Square Error is approximately 0.09%. The value of R^2 is 0.999944 which means that it is more than 99% similar to the expected result. Log loss function maps the variables to the real numbers which represent the cost associated. Hit Ratio is the number of times a correct prediction was made over total predictions. Top 10 hit ratios are used for the prediction and that has been given in the following Table 5.

3.1.3 Output - Validation Metrics

Output Validation metrics depict the output of the testing set. The following output metrics help in determining the efficacy of the model (Table 6).

Here as well, the MSE value is 0.09%. The R^2 value is more than 99%, which means the predicted value is 99% correct. The hit ratio is given in the following Table 7.

Table 5. Hit ratio for training set

K (Number of hits)	Hit ratio
1	0.9989
2	1.0
3	1.0
4	1.0
5	1.0
6	1.0
7	1.0
8	1.0
9	1.0
10	1.0

Table 6. Output validation metrics

Name of the parameter	Outcomes
Description	Metrics reported on full validation frame
Model_category	Mutinomial
MSE	0.000970
R^2	0.999944
Logos	0.011482

Table 7. Hit ratio for validation metrics

K	Hit ratio
1	0.9989
2	0.9997
3	0.9998
4	0.9998
5	0.9999
6	0.9999
7	0.9999
8	0.9999
9	0.9999
10	0.9999

4 Comparison with Support Vector Machine (SVM)

Support vector machines are supervised learning models that are used in machine learning that utilize learning algorithms to analyze and recognize patterns for classification [14]. It's training algorithm creates a model that assigns new examples into one category or the other and thus is a non-probabilistic binary linear classifier. It is a representation in terms of points in space such that there exists a clear gap in between various kinds of points grouped together. New data are predicted and classified based on how much it is closer to one particular group than the other.

4.1 Simulation Results for SVM

SV type: C-svc (classification)
Parameter: cost C = 5, Gaussian Radial Basis kernel function.
Hyperparameter: sigma = 0.05
Number of Support Vectors: 16860
Objective Function Value:
−2.0098 −11.4563 −31.787 −98.428 −50.5466 −1.999 −22.3287 −1.999 −1.7028 −1.8817 −1.9603 −1.9239 −1 −1.8357 −1 −8.426 −10.269 −9.3452 −1.7028 −7.5755 −1.7028 −24.0647 −19.539 −1.8817 −13.127 −1.8817 −33.4674 −1.9603 −18.2219 −1.9603 −1.924 −15.5029 −1.924 −1.8357 −1 −1.8357
Training error: 0.15365
Cross validation error: 0.00435

As we can see, the cross validation error is very low. Hence the model is accurate.

Comparison between the neural network and SVM can be tabulated as follows (Table 8):

Table 8. Comparison between deep neural network & SVM

Deep neural network	SVM
Error: 0.000961	Error: 0.15365
Accuracy: 0.999944	Accuracy: 0.84635

5 Conclusion

In this work, the training and validation models have a very high R^2 value. This high value has indicated that the adopted model is highly accurate. Application of the deep learning algorithm to the Intrusion detection System has enabled us to produce a detailed confusion matrix for the training set, as well as for the validation set. The result is supported along with a precise MSE graph. With the loss being set as Cross Entropy, we get a classification model that can be used to detect future intrusion attacks. The results obtained by Deep Neural Network are compared with the results obtained by Support Vector Machine.

References

1. Portnoy, L., Eskin, E., Stolfo, S.: Intrusion detection with unlabeled data using clustering. In: Proceedings of ACM CSS Workshop on Data Mining Applied to Security (DMSA 2001) (2001)
2. Mukkamala, S., Janoski, G., Sung, A.: Intrusion detection using neural networks and support vector machines. In: Proceedings of the 2002 International Joint Conference on Neural Networks, IJCNN 2002, vol. 2, pp. 1702–1707. IEEE (2002)

3. Ouyang, W., Wang, X.: Joint deep learning for pedestrian detection. In: 2013 IEEE International Conference on Computer Vision (ICCV), pp. 2056–2063. IEEE, December 2013
4. Roy, S.S., Mittal, D., Basu, A., Abraham, A.: Stock market forecasting using LASSO linear regression model. In: Abraham, A., Krömer, P., Snasel, V. (eds.) Afro-European Conference for Industrial Advancement. AISC, vol. 334, pp. 371–381. Springer, Cham (2015). doi:10.1007/978-3-319-13572-4_31
5. Basu, A., Roy, S.S., Abraham, A.: A novel diagnostic approach based on support vector machine with linear kernel for classifying the Erythemato-Squamous disease. In: 2015 International Conference on Computing, Communication Control and Automation (ICCUBEA), pp. 343–347. IEEE, February 2015
6. Arel, I., Rose, D., Coop, R.: DeSTIN: a scalable deep learning architecture with application to high-dimensional robust pattern recognition. In: AAAI Fall Symposium: Biologically Inspired Cognitive Architectures, November 2009
7. Roy, S.S., Viswanatham, V.M., Krishna, P.V., Saraf, N., Gupta, A., Mishra, R.: Applicability of rough set technique for data investigation and optimization of intrusion detection system. In: Singh, K., Awasthi, A.K. (eds.) QSHINE 2013. LNICST, vol. 115, pp. 479–484. Springer, Heidelberg (2013). doi:10.1007/978-3-642-37949-9_42
8. Wang, S., Jiang, Y., Chung, F.L., Qian, P.: Feedforward kernel neural networks, generalized least learning machine, and its deep learning with application to image classification. Appl. Soft Comput. **37**, 125–141 (2015)
9. Wackerly, D., Mendenhall, W., Scheaffer, R.: Mathematical Statistics with Applications. Cengage Learning (2007)
10. Draper, N.R., Smith, H., Pownell, E.: Applied Regression Analysis, vol. 3. Wiley, New York (1966)
11. Deng, L.: Three classes of deep learning architectures and their applications: a tutorial survey. APSIPA Trans. Sig. Inf. Process. (2012)
12. Roy, S.S., Viswanatham, V.M.: Classifying spam emails using artificial intelligent techniques. Int. J. Eng. Res. Africa **22**, 152–161 (2016)
13. Roy, S.S., Viswanatham, V.M., Krishna, P.V.: Spam detection using hybrid model of rough set and decorate ensemble. Int. J. Comput. Syst. Eng. **2**(3), 139–147 (2016)
14. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
15. Mittal, D., Gaurav, D., Roy, S.S.: An effective hybridized classifier for breast cancer diagnosis. In: 2015 IEEE International Conference on Advanced Intelligent Mechatronics (AIM), pp. 1026–1031. IEEE, July 2015
16. Bengio, Y.: Learning deep architectures for AI. Found. Trends® Mach. Learn. **2**(1), 1–127 (2009)
17. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. Neural Comput. **18**(7), 1527–1554 (2006)
18. Hansen, L.K., Salamon, P.: Neural network ensembles. IEEE Trans. Patt. Anal. Mach. Intell. **12**, 993–1001 (1990)

Computing

A Note on the Optimal Immunity of Boolean Functions Against Fast Algebraic Attacks

Jing Shen[1] and Yusong Du[2(✉)]

[1] Guangdong College of Industry and Commerce, Guangzhou 510510, China
[2] School of Information Management, Sun Yat-sen University, Guangzhou 510006, China
duyusong@mail.sysu.edu.cn

Abstract. The immunity of Boolean functions against fast algebraic attacks is an important cryptographic property. When deciding the optimal immunity of an n-variable Boolean function against fast algebraic attacks, one may need to compute the ranks of a series of matrices of size $\sum_{i=d+1}^{n}\binom{n}{i}\times\sum_{i=0}^{e}\binom{n}{i}$ over binary field $\mathbb{F}_2$ for each positive integer e less than $\lceil\frac{n}{2}\rceil$ and corresponding d. In this paper, for an n-variable balanced Boolean function, exploiting the combinatorial properties of the binomial coefficients, when n is odd, we show that the optimal immunity is only determined by the ranks of those matrices such that $\sum_{i=0}^{e}\binom{n}{i}$ is even. When n is even but not the power of 2, we show that the optimal immunity is only determined by the ranks of those matrices such that $\sum_{i=0}^{e}\binom{n}{i}$ is even or such that both $\sum_{i=0}^{e}\binom{n}{i}$ and $\sum_{i=0}^{e+1}\binom{n}{i}$ are odd.

Keywords: Boolean function · Fast algebraic attack · Algebraic immunity

1 Introduction

Boolean functions play a vital role in coding theory and in symmetric cryptography [8]. Various criteria related to cryptographically desirable Boolean functions have been proposed.

Boolean functions used in stream ciphers, especially in the filer and combination generators of stream ciphers based on linear feedback shift registers, should have large algebraic immunity, in order to help resist algebraic attacks [3,6,14]. Moreover, Boolean functions should also have the resistance against a variant of the algebraic attack, called the *fast algebraic attack* (FAA) [1,5,7]. To a certain degree the algebraic immunity can be covered by the immunity of Boolean functions against fast algebraic attacks (FAA's). Algebraic immunity, as well as

This work is supported by National Natural Science Foundations of China (Grant No. 61309028, Grant No. 61472457, Grant No. 61502113), Science and Technology Planning Project of Guangdong Province, China (Grant No. 2014A010103017), and Natural Science Foundation of Guangdong Province, China (Grant No. 2016A030313298).

© Springer Nature Singapore Pte Ltd. 2017
D. Giri et al. (Eds.): ICMC 2017, CCIS 655, pp. 57–67, 2017.
DOI: 10.1007/978-981-10-4642-1_6

the immunity against FAA's, has been considered as a important cryptographic property for Boolean functions used in stream ciphers [10,11,15,16,19].

Studies show that a good immunity for an n-variable function f against FAA's is that $\deg(fg) > d$ for any nonzero n-variable Boolean function g of algebraic degree at most e, where $1 \leq e < \lceil \frac{n}{2} \rceil$ and d is as large as possible but less than $n-e$, such as $d = n-e-1$, $d = n-e-2$ or $d = n-e-3$ [2,10,13,15]. In particular, if $\deg(fg) \geq n-e$ for any nonzero n-variable Boolean function g of degree at most e and any positive integer $e < \lceil n/2 \rceil$, then we say that Boolean function f has the *optimal immunity against fast algebraic attacks*.

When considering the immunity of n-variable Boolean function f against FAA's, we may need to determine whether $\deg(fg) > d$ for any nonzero n-variable Boolean function g of degree at most e. Clearly, if it is true for each integer $e = 1, 2, \cdots, \lceil \frac{n}{2} \rceil - 1$ and $d = n-e-1$, then f has the optimal immunity. This problem is then converted into determining the ranks of a series of matrices of size

$$\sum_{i=d+1}^{n} \binom{n}{i} \times \sum_{i=0}^{e} \binom{n}{i}$$

over $\mathbb{F}_2$, denoted by $W(f; e, d)$, for each integer positive e less than $\lceil \frac{n}{2} \rceil$ and corresponding d. More precisely, $\deg(fg) > d$ for a given nonzero n-variable Boolean function g of degree at most e if and only if $W(f; e, d)$ has full column rank [10,12].

A class of n-variable balanced Boolean functions [4], called Carlet-Feng functions, denoted by ϕ_{CF}, was proved to satisfy $\deg(\phi_{CF} \cdot g) \geq n-e-1$ and even satisfy $\deg(\phi_{CF} \cdot g) \geq n-e$ when $n = 2^s + 1$ with positive integer s, for any nonzero n-variable Boolean function g of degree at most e and any positive integer $e < \lceil n/2 \rceil$ [12]. Another class of even n-variable balanced Boolean functions [17], called Tang-Carlet functions, denoted by τ_{CF}, may also have good immunity, i.e., it was proved that $\deg(\tau_{CF} \cdot g) \geq n-e-2$ for all possible functions g and integers e [13].

In this paper, we further discuss the generic method of deciding the immunity of Boolean functions against FAA's by observing the combinatorial properties of $W(f; e, d)$ matrix. For an n-variable *balanced* Boolean function f, when n is odd, we show that the optimal immunity can be determined only by the ranks of those $W(f; e, d)$ matrices such that $\sum_{i=0}^{e} \binom{n}{i}$ is even; when n is even but not the power of 2, we show that the optimal immunity can be determined only by the ranks of those $W(f; e, d)$ matrices such that $\sum_{i=0}^{e} \binom{n}{i}$ is even or such that both $\sum_{i=0}^{e} \binom{n}{i}$ and $\sum_{i=0}^{e+1} \binom{n}{i}$ are odd. This result may help us better study the optimal immunity of balanced Boolean functions against FAA's, and shorten the actual time of deciding the optimal immunity of a Boolean function, because the number of matrices, whose ranks that we need to compute, may be smaller.

2 Preliminaries

Let n be a positive integer. An n-variable Boolean function f is viewed as a mapping from vector space $\mathbb{F}_2^n$ to binary field $\mathbb{F}_2$ and has a unique n-variable polynomial representation over

$$\mathbb{F}_2[x_1, x_2, \cdots, x_n]/(x_1^2 - x_1, x_2^2 - x_2, \cdots, x_n^2 - x_n),$$

called the *algebraic normal form* (ANF) of f,

$$f(x_1, x_2, \cdots, x_n) = a_0 + \sum_{1 \le i \le n} a_i x_i + \sum_{1 \le i < j \le n} a_{ij} x_i x_j + \cdots + a_{12\cdots n} x_1 x_2 \cdots x_n,$$

where $a_0, a_i, a_{ij}, \ldots, a_{12\cdots n}$ belong to $\mathbb{F}_2$. For simplicity, an n-variable Boolean function $f(x)$ sometimes is written as $f(x) = \sum_{c \in \mathbb{F}_2^n} f_c x^c$, where

$$x^c = x_1^{c_1} x_2^{c_2} \cdots x_n^{c_n}$$

and $f_c \in \mathbb{F}_2$. We denote by $\mathbb{B}_n$ the set of all the n-variable Boolean functions.

For $f \in \mathbb{B}_n$, the set of $x = (x_1, x_2, \cdots, x_n) \in \mathbb{F}_2^n$ for which $f(x) = 1$ is called the support of the function, denoted by $\mathrm{supp}(f)$. The Hamming weight of f is the cardinality of $\mathrm{supp}(f)$, denoted by $wt(f)$. Boolean function f is called balanced if $wt(f) = 2^{n-1}$. The algebraic degree of Boolean function f, denoted by $\deg(f)$, is the degree of its ANF. It is well-known that the algebraic degree of a balanced n-variable Boolean function is less than n, i.e., the coefficient of term $x_1 x_2 \cdots x_2$ in its ANF must be zero.

A Boolean function $g \in \mathbb{B}_n$ is called an *annihilator* of $f \in \mathbb{B}_n$ if $fg = 0$. The lowest algebraic degree of all the nonzero annihilators of f and $1 + f$ is called *algebraic immunity* of f or $1 + f$, denoted by $\mathcal{AI}_n(f)$, and it has been proved that $\mathcal{AI}_n(f) \le \lceil \frac{n}{2} \rceil$ for a given $f \in \mathbb{B}_n$. A Boolean function $f \in \mathbb{B}_n$ has the *maximum algebraic immunity* if $\mathcal{AI}_n(f) = \lceil \frac{n}{2} \rceil$.

Definition 1. *An n-variable Boolean function f has the optimal immunity against FAA's if $\deg(fg) \ge n - e$ for any nonzero n-variable Boolean function g of degree at most e and for any positive integer $e < \lceil n/2 \rceil$.*

It is not hard to see that n-variable Boolean functions with the optimal immunity against FAA's have the maximum algebraic immunity. Also, Boolean functions with the optimal immunity against FAA's were said to be perfect algebraic immune functions in [12].

When studying the immunity of Boolean Functions against FAA's, the following two sets of vectors and a matrix over $\mathbb{F}_2$ are useful.

For $x = (x_1, x_2, \cdots, x_n) \in \mathbb{F}_2^n$, let $wt_2(x)$ be the number of its nonzero coordinates. Denote $\mathcal{W}_e$ by the set $\{x \in \mathbb{F}_2^n \mid wt_2(x) \le e\}$ in lexicographic order and $\overline{\mathcal{W}}_d$ by the set $\{x \in \mathbb{F}_2^n \mid wt_2(x) \ge d+1\}$ in reverse lexicographic order where $1 \le e < \lceil \frac{n}{2} \rceil$ and $d < n$. As a matter of fact, the orderings of $\mathcal{W}_e$ and $\overline{\mathcal{W}}_d$ do not essentially affect the deciding results on the immunity of Boolean functions, but good orderings may be good for observing and computing [9].

Let $\bar{x} = (x_1 + 1, x_2 + 1, \cdots, x_n + 1)$. It is clear that if x is the i-th element in $\mathcal{W}_e$ and $\bar{x} \in \overline{\mathcal{W}}_d$ then $\bar{x}$ is the i the element in $\overline{\mathcal{W}}_d$. In particular, $\mathbf{1}_n = (11 \cdots 1)$ and $\mathbf{0}_n = (00 \cdots 0)$ are the first elements in $\overline{\mathcal{W}}_d$ and $\mathcal{W}_e$ respectively.

For $y, z \in \mathbb{F}_2^n$, let $z \subset y$ be an abbreviation for $\mathrm{supp}(z) \subset \mathrm{supp}(y)$, where $\mathrm{supp}(x) = \{i \mid x_i = 1\}$; and let $y \cap z = (y_1 \wedge z_1, y_2 \wedge z_2, \cdots, y_n \wedge z_n)$, where $\wedge$ is the bit AND operation.

Denote $W(f;e,d)$ by a matrix over $\mathbb{F}_2$ related to function $f \in \mathbb{B}_n$, which has been mentioned in Sect. 1. It is a

$$\sum_{i=d+1}^{n} \binom{n}{i} \times \sum_{i=0}^{e} \binom{n}{i}$$

matrix with ij-th element equal to

$$w_{ij} = w_{yz} = f_{y \cap \bar{z}},$$

where y is the i-th element in $\overline{\mathcal{W}}_d$ and z is j-th element in $\mathcal{W}_e$. It was shown that one can determine the (optimal) immunity against FAA's through computing the rank of matrix $W(f;e,d)$.

Theorem 1 ([10,12]). *Let $f \in \mathbb{B}_n$. There exists no non-zero function g of degree at most e such that the product fg has degree at most d, i.e.,* $\deg(fg) \geq d+1$*, if and only if $W(f;e,d)$ has full column rank.*

According to Theorem 1, if $W(f;e,n-e-1)$ has full column rank then $\deg(fg) \geq n-e$ for any nonzero n-variable Boolean function g of degree at most e. Then from Theorem 1 we have a sufficient condition such that an n-variable Boolean function having the optimal immunity against FAA's.

Corollary 1. *An n-variable Boolean function has the optimal immunity against fast algebraic attacks if $W(f;e,n-e-1)$ has full column rank for each integer $e = 1, 2, \cdots, \lceil \frac{n}{2} \rceil - 1$.*

3 Deciding the Immunity of Balanced Boolean Functions in Odd Variables Against Fast Algebraic Attacks

Balanced Boolean functions are more interesting for cryptography. From this section, we focus on the optimal immunity of n-variable *balanced* Boolean functions against FAA's.

It is clear that $W(f;e,n-e-1)$ is a symmetric matrix of size $\sum_{i=0}^{e}\binom{n}{i} \times \sum_{i=0}^{e}\binom{n}{i}$. For simplicity, we denote $W(f;e,n-e-1)$ by $W(f;e)$. Then the immunity of function f against FAA's is related to the problem whether matrix $W(f;e)$ has nonzero determinant over $\mathbb{F}_2$. It was also noted that $W(f;e)$ has an interesting property about its determinant.

Lemma 1 ([12]). *If $w_{\mathbf{11}} = \sum_{i=0}^{e}\binom{n}{i} + 1 \bmod 2$ then $\det(W(f;e)) = 0$, and if $w_{\mathbf{11}} = \sum_{i=0}^{e}\binom{n}{i} \bmod 2$ then*

$$\det(W(f;e)) = \det(W(f;e)^{(1,1)}),$$

where $W(f;e)^{(1,1)}$ is the matrix that results from $W(f;e)$ by removing the first row and the first column. In particular, when $w_{\mathbf{11}} = 0$, $\det(W(f;e)) = 1$ only if $\sum_{i=0}^{e}\binom{n}{i}$ is even.

For balanced Boolean functions, entry $w_{\mathbf{11}}(= f_{\mathbf{1}_n})$ in Lemma 1 is always zero. Then it was further proved in [12] that an n-variable balanced Boolean function has the optimal immunity against FAA's only if $n = 2^s+1$ with positive integer s. More precisely, it was proved that $\sum_{i=0}^{e}\binom{n}{i}$ are all even for each integer $e = 1, 2, \cdots, \lceil \frac{n}{2} \rceil - 1$ only if $n = 2^s + 1$ with positive integer s. This means that $\det(W(f; e)) = 0$ and $\deg(fg) \geq n - e$ may never hold for some n and e. For example, if $n = 7$ and $e = 2$, then $\sum_{i=0}^{e}\binom{n}{i} = 29$ is odd, and $\det(W(f; e)) = 0$. In this case, we can only determine whether $\deg(fg) \geq n - e - 1$. That is to say, it may be the best case for an n-variable balanced function f against FAA's that $\deg(fg) \geq n - e$ when $\sum_{i=0}^{e}\binom{n}{i}$ is even and $\deg(fg) \geq n - e - 1$ when $\sum_{i=0}^{e}\binom{n}{i}$ is odd. The Carlet-Feng functions [4], denoted by ϕ_{CF}, was proved to satisfy $\deg(\phi_{CF} \cdot g) \geq n - e$ when $\sum_{i=0}^{e}\binom{n}{i}$ is even and $\deg(\phi_{CF} \cdot g) \geq n - e - 1$ when $\sum_{i=0}^{e}\binom{n}{i}$ is odd, for any nonzero n-variable Boolean function g of degree at most e and any positive integer $e < \lceil n/2 \rceil$ [12]. We say that balanced functions like the Carlet-Feng functions have the optimal immunity against FAA's.

Definition 2. *Let f be an n-variable balanced Boolean function. The function f has the optimal immunity against fast algebraic immunity if* $\deg(fg) \geq n - e$ *when* $\sum_{i=0}^{e}\binom{n}{i}$ *is even and* $\deg(fg) \geq n - e - 1$ *when* $\sum_{i=0}^{e}\binom{n}{i}$ *is odd for any nonzero n-variable Boolean function g of degree at most e and for any positive integer $e < \lceil n/2 \rceil$.*

According to Theorem 1 again, if $W(f; e, n - e - 2)$ has full column rank then $\deg(fg) \geq n-e-1$ for any nonzero n-variable Boolean function g with $\deg(g) \leq e$. This implies that one can determine the optimal immunity by computing the rank of $W(f; e, n - e - 1) = W(f; e)$ or $W(f; e, n - e - 2)$ for all the possible e. The following corollary provides a generic method of deciding the optimal immunity of balanced Boolean functions against FAA's.

Corollary 2. *An n-variable balanced Boolean function has the optimal immunity against fast algebraic attacks if the following two conditions hold for each positive integer e less than $\frac{n}{2}$:*

1. $\det(W(f; e)) = 1$ *when* $\sum_{i=0}^{e}\binom{n}{i}$ *is even;*
2. $W(f; e, n - e - 2)$ *has full column rank when* $\sum_{i=0}^{e}\binom{n}{i}$ *is odd.*

For balanced Boolean functions in odd number of variables, we give a simplified sufficient condition, compared to Corollary 2, such that they have the optimal immunity against FAA's. More precisely, we prove that the optimal immunity is determined only by the determinant (rank) of $W(f; e)$ over $\mathbb{F}_2$ such that $\sum_{i=0}^{e}\binom{n}{i}$ is even. This observation is mainly based on the following combinatoric property.

Lemma 2. *Let n be odd and e be integers with $1 < e < n$. If* $\sum_{i=0}^{e}\binom{n}{i}$ *is odd, then both e and* $\sum_{i=0}^{e+1}\binom{n}{i}$ *are even. Moreover,* $\sum_{i=0}^{\frac{n-1}{2}}\binom{n}{i}$ *is even.*

Proof. Note that

$$\sum_{i=0}^{e}\binom{n}{i}=\sum_{i=1}^{e}\binom{n-1}{i}+\sum_{i=1}^{e}\binom{n-1}{i-1}+\binom{n}{0}\equiv\binom{n-1}{e}\bmod 2.$$

According to Lucas' theorem about the binomial coefficient, for positive integers n and e, the congruence relation holds

$$\binom{n}{e}\equiv\prod_{k=0}^{s-1}\binom{n_k}{e_k}\bmod 2,$$

where $n=\sum_{k=0}^{s-1}n_k2^k$ and $e=\sum_{k=0}^{s-1}e_k2^k$ are the binary expansions of n and e respectively. Since n is odd it follows that $n-1=\sum_{k=1}^{s-1}n_k2^k$. We have

$$\binom{n-1}{e}\equiv\binom{n_{s-1}}{e_{s-1}}\binom{n_{s-2}}{e_{s-2}}\cdots\binom{n_1}{e_1}\binom{0}{e_0}\bmod 2,$$

where $n_{s-1},\cdots,n_1$ are not all zero. If $e>1$ and $\binom{n-1}{e}$ is odd then $e_0=0$, which means that e is even. Then we have

$$\sum_{i=0}^{e+1}\binom{n}{i}\equiv\binom{n-1}{e+1}\equiv\binom{n_{s-1}}{e_{s-1}}\binom{n_{s-2}}{e_{s-2}}\cdots\binom{n_1}{e_1}\binom{0}{1}\equiv 0\bmod 2,$$

i.e., $\sum_{i=0}^{e+1}\binom{n}{i}$ is even. Moreover, we also have

$$\sum_{i=0}^{\frac{n-1}{2}}\binom{n}{i}\equiv\binom{n-1}{\frac{n-1}{2}}\equiv\binom{n_{s-1}}{0}\binom{n_{s-2}}{n_{s-1}}\cdots\binom{n_2}{n_3}\binom{n_1}{n_2}\binom{0}{n_1}\equiv 0\bmod 2,$$

This implies that $\binom{n-1}{\frac{n-1}{2}}$ must be even, otherwise $n_1=n_2=\cdots=n_{s-1}=0$, which is a contradiction. This completes the proof. □

Theorem 2. *Let n be odd and e be integers with $1\le e<\lceil\frac{n}{2}\rceil$. Let f be an n-variable balanced Boolean function. If $\det(W(f;e))=1$ for each integer e such that $\sum_{i=0}^{e}\binom{n}{i}$ is even, then f has the optimal immunity against fast algebraic attacks.*

Proof. Function f satisfies the first condition in Corollary 2. When $\sum_{i=0}^{e}\binom{n}{i}$ is odd we need to check the rank of $W(f;e,n-e-2)$, which is a

$$\sum_{i=n-e-1}^{n}\binom{n}{i}\times\sum_{i=0}^{e}\binom{n}{i}$$

matrix. But this happens only when $2\le e\le\lceil\frac{n}{2}\rceil-2$ because $\sum_{i=0}^{1}\binom{n}{i}$ and $\sum_{i=0}^{\frac{n-1}{2}}\binom{n}{i}$ are always even by Lemma 2. Note that $\sum_{i=n-e-1}^{n}\binom{n}{i}=\sum_{i=0}^{e+1}\binom{n}{i}$

and then matrix $W(f; e, n-e-2)$ consists of the first $\sum_{i=0}^{e}\binom{n}{i}$ columns of $W(f; e+1, n-e-2)$, which is a square matrix of size

$$\sum_{i=0}^{e+1}\binom{n}{i} \times \sum_{i=0}^{e+1}\binom{n}{i}.$$

According to Lemma 2, if $\sum_{i=0}^{e}\binom{n}{i}$ is odd then $\sum_{i=0}^{e+1}\binom{n}{i}$ must be even. We have $\det(W(f; e+1, n-e-2)) = \det(W(f; e+1)) = 1$, hence $W(f; e, n-e-2)$ has full column rank for integer e such that $\sum_{i=0}^{e}\binom{n}{i}$ is odd. This means that f also satisfies the second condition in Corollary 2. Finally, for the maximum $e = \lceil \frac{n}{2} \rceil - 1 = \frac{n-1}{2}$ we have $W(f; e)$ has full rank because $\sum_{i=0}^{\frac{n-1}{2}}\binom{n}{i}$ must be even according to Lemma 2. Therefore, f has the optimal immunity. □

As an example of using Theorem 2, when $n = 13$ we have the sequence $(\sum_{i=0}^{1}\binom{13}{i}, \cdots, \sum_{i=0}^{6}\binom{13}{i}) = (14, 92, 378, 1093, 2380, 4096)$. By the method from Corollary 2, we may need to compute the determinants of 5 square matrices, and the rank of one matrix of size 2380×1093. It accounts for the vast majority of the total computational cost to compute the determinants of the square matrix of order 4096. However, using the method from Theorem 2, we do not need to compute the rank of the matrix of size 2380×1093 anymore, though the computational complexity is unchanged.

We randomly choose 100 balanced Boolean functions in 13 variables and determine if each of them has the optimal immunity by the method from Corollary 2 and by the method from Theorem 2 respectively. We implement the test by using g++ compiler and Number Theory Library (NTL, a C++ library for doing number theory) on a laptop computer (Intel Core i7-6820hq at 2.7 Ghz, 8 GB RAM, Ubuntu 16.04). The results show that the time of deciding the optimal immunity of a balanced Boolean function in 13 variables can fall by 23% on average.

Similarly, when $n = 15$ we have the sequence $(\sum_{i=0}^{1}\binom{15}{i}, \cdots, \sum_{i=0}^{7}\binom{15}{i}) = (16, 121, 576, 1941, 4994, 9949, 16384)$. Using the method from Theorem 2, we do not need to compute the rank of the matrix of size 16384×9949. We randomly choose 100 balanced Boolean functions in 15 variables for the test. The results show that the the time can fall by nearly 35% on average.

There is a special case of Theorem 2 when $n = 2^s + 1$ with positive s. In this case, $\sum_{i=0}^{e}\binom{n}{i}$ is even for each integer e less than $n/2$. The theorem still holds, but it is the same as Corollary 2.

4 Deciding the Immunity of Balanced Boolean Functions in Even Variables Against Fast Algebraic Attacks

In this section, for balanced Boolean functions in even number of variables, similarly, we give a reduced sufficient condition, compared to Corollary 2, such that they have the optimal immunity against FAA's.

Lemma 3. *Let $m > 1$ be odd. If $n = 2^t \cdot m$ with positive integer t, then $\sum_{i=0}^{n/2-1} \binom{n}{i}$ must be even.*

Proof. As in the proof of Lemma 2. we have $\sum_{i=0}^{n/2-1} \binom{n}{i} \equiv \binom{n-1}{n/2-1} \bmod 2$. Since $n = 2^t \cdot m$ with odd $m > 1$ and positive t it follows that

$$n-1 = \sum_{k=1}^{s-1} m_k 2^{t+k} + (2^t - 1) \quad \text{and} \quad n/2 - 1 = \sum_{k=1}^{s-1} m_k 2^{t+k-1} + (2^{t-1} - 1),$$

where $m = \sum_{k=0}^{s-1} m_k 2^k$ is the binary expansions of m. According to Lucas' theorem about the binomial coefficient, we have

$$\binom{n-1}{n/2-1} \equiv \binom{m_{s-1}}{0}\binom{m_{s-2}}{m_{s-1}} \cdots \binom{m_2}{m_3}\binom{m_1}{m_2}\binom{0}{m_1} \underbrace{\binom{1}{0}\binom{1}{1} \cdots \binom{1}{1}}_{t} \bmod 2.$$

In particular, when $t = 1$ we have

$$\binom{n-1}{n/2-1} = \binom{n-1}{m-1} \equiv \binom{m_{s-1}}{0}\binom{m_{s-2}}{m_{s-1}} \cdots \binom{m_2}{m_3}\binom{m_1}{m_2}\binom{0}{m_1}\binom{1}{0} \bmod 2.$$

This implies that $\binom{n-1}{n/2-1}$ must be even, otherwise $m_1 = m_2 = \cdots = m_{s-1} = 0$, which is a contradiction. This completes the proof. □

Theorem 3. *Let n be even but not the power of 2 and e be integers with $1 \le e < \lceil \frac{n}{2} \rceil$. Let f be an n-variable balanced Boolean function. If $W(f; e)$ has full rank over $\mathbb{F}_2$ for each integer e such that $\sum_{i=0}^{e} \binom{n}{i}$ is even, and $W(f; e, n-e-2)$ has full column rank over $\mathbb{F}_2$ for each integer e such that both $\sum_{i=0}^{e} \binom{n}{i}$ and $\sum_{i=0}^{e+1} \binom{n}{i}$ are odd, then f has the optimal immunity against fast algebraic attacks.*

Proof. For each integer e from 1 to $n/2-1$, if $\sum_{i=0}^{e} \binom{n}{i}$ is even, then $W(f; e)$ have full rank. This means that f satisfies the first condition of Corollary 2. If $\sum_{i=0}^{e} \binom{n}{i}$ is odd but $\sum_{i=0}^{e+1} \binom{n}{i}$ is even, then $W(f; e+1)$ has full rank. Note that $W(f; e, n-e-2)$ consists of the first $\sum_{i=0}^{e} \binom{n}{i}$ columns of $W(f; e+1, n-e-2) = W(f; e+1)$. It follows that $W(f; e, n-e-2)$ has full column rank. This means that f satisfies the second condition of Corollary 2. If both $\sum_{i=0}^{e} \binom{n}{i}$ and $\sum_{i=0}^{e+1} \binom{n}{i}$ are odd, then we have $W(f; e, n-e-2)$ has full column rank, which directly satisfies the second condition of Corollary 2. Finally, for the maximum $e = n/2-1$ we have $W(f; n/2-1)$ has full rank because $\sum_{i=0}^{n/2-1} \binom{n}{i}$ must be even according to Lemma 3. This means that f satisfies the first condition of Corollary 2 for the maximum $e = n/2-1$. This complete the proof. □

As an example of using Theorem 3, when $n = 14$ we have the sequence $(\sum_{i=0}^{1} \binom{14}{i}, \cdots, \sum_{i=0}^{6} \binom{14}{i}) = (15, 106, 470, 1471, 3473, 6476)$. Using the method from Theorem 3, we do not need to compute the rank of the matrix of size 6476×3473. The experiment shows that the time of deciding the optimal immunity of a balanced Boolean function in 14 variables can fall by 38% on average.

The conditions given by Theorem 3 can be further reduced for an n-variable Boolean function f and $n = 2m$ with odd $m > 1$, if we only want to decide whether $\deg(fg) \geq n - e - 2$ for any nonzero n-variable Boolean function g of degree at most e and for any positive integer $e < n/2$. In this case, f can be also considered as a boolean function with *almost optimal immunity* against FAA's. As mentioned in Sect. 1, Tang-Carlet functions, denoted by τ_{CF}, were proved to satisfy $\deg(\tau_{CF} \cdot g) \geq n - e - 2$ for any nonzero n-variable Boolean function g of degree at most e and for any positive integer $e < n/2$ [13].

Lemma 4. *Let $n = 2m$ with odd $m > 1$. If both $\sum_{i=0}^{e} \binom{n}{i}$ and $\sum_{i=0}^{e+1} \binom{n}{i}$ are odd, then $\sum_{i=0}^{e+2} \binom{n}{i}$ must be even.*

Proof. As in the proof of Lemma 2, we have $\sum_{i=0}^{e} \binom{n}{i} \equiv \binom{n-1}{e} \bmod 2$. Since m is odd it follows that $n - 1 = 2m - 1 = 1 + \sum_{k=1}^{s-1} m_k 2^{k+1}$, where $m = \sum_{k=0}^{s-1} m_k 2^k$ is the binary expansion of m. According to Lucas' theorem about the binomial coefficient, for positive integers m and e, we have

$$\binom{n-1}{e} \equiv \binom{m_{s-1}}{0}\binom{m_{s-2}}{e_{s-1}} \cdots \binom{m_2}{e_3}\binom{m_1}{e_2}\binom{0}{e_1}\binom{1}{e_0} \bmod 2,$$

where $e = \sum_{k=0}^{s-1} e_k 2^k$ is the binary expansion of e and $m_{s-1}, \cdots, m_1$ are not all zero. If $e > 1$ and $\binom{n-1}{e}$ is odd then $e_1 = 0$, which also means that

$$e \equiv 0 \bmod 4 \quad \text{or} \quad e \equiv 1 \bmod 4.$$

This implies that $\sum_{i=0}^{e+2} \binom{n}{i}$ must be even if both $\sum_{i=0}^{e} \binom{n}{i}$ and $\sum_{i=0}^{e+1} \binom{n}{i}$ are odd. □

Corollary 3. *Let $n = 2m$ with odd $m > 1$ and e be integers with $1 \leq e \leq m-1$. Let f be an n-variable balanced Boolean function. If $W(f; e)$ has full rank for each integer e such that $\sum_{i=0}^{e} \binom{n}{i}$ is even, then $\deg(fg) \geq n - e - 2$ for any nonzero n-variable Boolean function g of degree at most e and for any positive integer $e < n/2$.*

Proof. If $\sum_{i=0}^{e} \binom{n}{i}$ is even or $\sum_{i=0}^{e} \binom{n}{i}$ is odd but $\sum_{i=0}^{e+1} \binom{n}{i}$ is even, as in the proof of Theorem 3, we have $\deg(fg) \geq n - e$ or $\deg(fg) \geq n - e - 1$ respectively for any nonzero n-variable Boolean function g of degree at most e. If both $\sum_{i=0}^{e} \binom{n}{i}$ and $\sum_{i=0}^{e+1} \binom{n}{i}$ are odd, then we have $\sum_{i=0}^{e+2} \binom{n}{i}$ is even by Lemma 4 and then $W(f; e+2, n-e-3) = W(f; e+2)$ has full rank. Note that $W(f; e, n-e-3)$ consists of the first $\sum_{i=0}^{e} \binom{n}{i}$ columns of $W(f; e+2, n-e-3)$. Therefore, in this case, $W(f; e, n-e-3)$ has full column rank. This means that $\deg(fg) \geq n-e-2$ for any nonzero n-variable Boolean function g of degree at most e. □

When $n = 2^s$ with positive s, i.e., when n is the power of 2, Theorem 3 is no longer applicable. In this case, it is not hard to see that $\sum_{i=0}^{e} \binom{n}{i}$ is odd for each integer e less than $n/2$. Therefore, we may need to compute the rank of matrix $W(f; e, n - e - 2)$ for each integer e less than $n/2$ according to Corollary 2.

5 Conclusion

In this paper, we further discuss the sufficient conditions of deciding the optimal immunity of balanced Boolean functions against FAA's. By exploiting the combinatorial properties of $W(f; e, d)$ matrix, we give two reduced conditions such that balanced Boolean functions have the optimal immunity against FAA's. This result may help us better study the immunity of Boolean functions against FAA's, and decrease the actual time of deciding the optimal immunity of balanced Boolean functions against FAA's.

References

1. Armknecht, F.: Improving fast algebraic attacks. In: Roy, B., Meier, W. (eds.) FSE 2004. LNCS, vol. 3017, pp. 65–82. Springer, Heidelberg (2004). doi:10.1007/978-3-540-25937-4_5
2. Armknecht, F., Carlet, C., Gaborit, P., Künzli, S., Meier, W., Ruatta, O.: Efficient computation of algebraic immunity for algebraic and fast algebraic attacks. In: Vaudenay, S. (ed.) EUROCRYPT 2006. LNCS, vol. 4004, pp. 147–164. Springer, Heidelberg (2006). doi:10.1007/11761679_10
3. Carlet, C., Dalai, D.K., Gupta, K.C., Maitra, S.: Algebraic immunity for cryptographically significant boolean functions: analysis and construction. IEEE Trans. Inform. Theory **52**(7), 3105–3121 (2006)
4. Carlet, C., Feng, K.: An infinite class of balanced functions with optimal algebraic immunity, good immunity to fast algebraic attacks and good nonlinearity. In: Pieprzyk, J. (ed.) ASIACRYPT 2008. LNCS, vol. 5350, pp. 425–440. Springer, Heidelberg (2008). doi:10.1007/978-3-540-89255-7_26
5. Courtois, N.T.: Fast algebraic attacks on stream ciphers with linear feedback. In: Boneh, D. (ed.) CRYPTO 2003. LNCS, vol. 2729, pp. 176–194. Springer, Heidelberg (2003). doi:10.1007/978-3-540-45146-4_11
6. Courtois, N.T., Meier, W.: Algebraic attacks on stream ciphers with linear feedback. In: Biham, E. (ed.) EUROCRYPT 2003. LNCS, vol. 2656, pp. 345–359. Springer, Heidelberg (2003). doi:10.1007/3-540-39200-9_21
7. Courtois, N.T.: Cryptanalysis of sfinks. In: Won, D.H., Kim, S. (eds.) ICISC 2005. LNCS, vol. 3935, pp. 261–269. Springer, Heidelberg (2006). doi:10.1007/11734727_22
8. Crama, Y., Hammer, P.: Boolean Models and Methods in Mathematics, Computer Science, and Engineering, Encyclopedia of Mathematics and its Applications. Cambridge University Press, Cambridge (2010)
9. Dalai, D.K.: Computing the rank of incidence matrix and algebraic immunity of Boolean functions. http://eprint.iacr.org/2013/273.pdf
10. Du, Y., Zhang, F., Liu, M.: On the resistance of boolean functions against fast algebraic attacks. In: Kim, H. (ed.) ICISC 2011. LNCS, vol. 7259, pp. 261–274. Springer, Heidelberg (2012). doi:10.1007/978-3-642-31912-9_18
11. Liu, M., Lin, D.: Fast algebraic attacks and decomposition of symmetric boolean functions. IEEE Trans. Inform. Theory **57**(7), 4817–4821 (2011)
12. Liu, M., Zhang, Y., Lin, D.: Perfect algebraic immune functions. In: Wang, X., Sako, K. (eds.) ASIACRYPT 2012. LNCS, vol. 7658, pp. 172–189. Springer, Heidelberg (2012). doi:10.1007/978-3-642-34961-4_12

13. Liu, M., Lin, D.: Almost perfect algebraic immune functions with good nonlinearity. In: International Symposium on Information Theory, ISIT 2014, pp. 1837–1841. IEEE, New York (2014)
14. Meier, W., Pasalic, E., Carlet, C.: Algebraic attacks and decomposition of boolean functions. In: Cachin, C., Camenisch, J.L. (eds.) EUROCRYPT 2004. LNCS, vol. 3027, pp. 474–491. Springer, Heidelberg (2004). doi:10.1007/978-3-540-24676-3_28
15. Pasalic, E.: Almost fully optimized infinite classes of boolean functions resistant to (Fast) algebraic cryptanalysis. In: Lee, P.J., Cheon, J.H. (eds.) ICISC 2008. LNCS, vol. 5461, pp. 399–414. Springer, Heidelberg (2009). doi:10.1007/978-3-642-00730-9_25
16. Rizomiliotis, P.: On the resistance of boolean functions against algebraic attacks using univariate polynomial representation. IEEE Trans. Inform. Theory **56**(8), 4014–4024 (2010)
17. Tang, D., Carlet, C., Tang, X.: Highly nonlinear boolean functions with optimal algebraic immunity and good behavior against fast algebraic attacks. IEEE Trans. Inform. Theory **59**(1), 653–664 (2013)
18. Wang, W., Liu, M., Zhang, Y.: Comments on "A design of boolean functions resistant to (Fast) algebraic cryptanalysis with efficient implementation". Crypt. Commun. **5**(1), 1–6 (2013)
19. Zhang, Y., Liu, M., Lin, D.: On the immunity of rotation symmetric Boolean functions against fast algebraic attacks. Discrete Appl. Math. **162**(1), 17–27 (2014)

Study of Five-Neighborhood Linear Hybrid Cellular Automata and Their Synthesis

Swapan Maiti(✉) and Dipanwita Roy Chowdhury(✉)

Department of Computer Science and Engineering,
Indian Institute of Technology Kharagpur, Kharagpur, India
swapankumar_maiti@yahoo.co.in, drc@cse.iitkgp.ernet.in

Abstract. Cellular automata (CA) is universally known as very good pseudorandom sequence generator. It has wide applications in several fields like VLSI design, error-correcting codes, test pattern generation, cryptography etc. Most of these applications use 3-neighborhood one dimensional CA. Cellular automata have been chosen as a better crypto-primitives for providing very good pseudorandom sequences and their high diffusion property. The randomness and diffusion properties can be increased with the increase of the size of neighborhood radius of the CA cell. In this work, we study a class of 5-neighborhood null boundary linear CA. We present an algorithm for synthesizing 5-neighborhood linear CA from its characteristic polynomial by assuming that some of the CA sub-polynomials are available.

Keywords: Cellular automata · 5-neighborhood linear rules · CA synthesis algorithm

1 Introduction

Cellular Automata (CA) have long been of interest to researchers for their theoretical properties and practical applications. It was initiated in the early 1950's by John von Neumann [12] and Stan Ulam as a general framework for modeling complex structures capable of self-reproduction and self-repair. In 1986, Wolfram first applied CA in pseudorandom number generation [15]. CA has made understanding of many occurrences in nature easier. The simple and regular structure of CA has attracted researchers and practitioners of different fields. In the last two decades, one-dimensional (1-D) CA based Pseudorandom Number Generators (PRNGs) have been extensively studied [2,5,10,11]. Though the recent interest is more focused on two-dimensional (2-D) CA PRNGs [9,13] since it seems that their randomness is much better than that of 1- D CA PRNGs, but considering the design complexity and computation efficiency, it is quite difficult to conclude which one is better. Compared to 2-D CA PRNGs, 1-D CA PRNGs are easier to be implemented in a large scale [3,8,14]. Random bit generators play an important role in different computer simulation methods such as Monte Carlo techniques, Browmian dynamics, stochastic optimization, computer-based

© Springer Nature Singapore Pte Ltd. 2017
D. Giri et al. (Eds.): ICMC 2017, CCIS 655, pp. 68–83, 2017.
DOI: 10.1007/978-981-10-4642-1_7

gaming, test pattern generation for VSLI circuit test, error-correcting codes, image processing, neural networks and cryptography etc. Most of these works are devoted to the study of cellular automata as pseudorandom bit generators. A central problem in any stream cipher scheme is to generate long, unpredictable random key sequences and Cellular Automata resolves this problem.

In most of all these applications, 1-D elementary cellular automata (i.e. three-neighborhood CA) are used. There are also some applications [6,9,13] of five or more neighborhood 2-D CA but that need more hardware complexity. In [7], it has been shown a 4-neighborhood nonlinear 1-D CA as a better cryptographic primitive. The randomness and diffusion properties of the CA can be developed with the increase of the size of neighborhood radius of the CA cell. More diffusion property of CA can make fast initialization of a stream cipher. In this paper, we study 5-neighborhood linear 1-D CA for providing very good pseudorandom sequences and high diffusion. We present an algorithm for synthesizing the CA.

This paper is organized as follows. Following the introduction, the basics of CA are presented in Sect. 2. In Sect. 3, we present 5-neighborhood Linear Hybrid Cellular Automata with the CA transition matrix and the characteristic polynomial. A recurrence relation is introduced for determining the characteristic polynomial and a CA synthesis algorithm is presented. We also present the randomness and diffusion properties of 5-neighborhood CA rule vectors and the comparison of their properties with 3/4 neighborhood CA. Finally, the paper is concluded in Sect. 4.

2 Basics of Linear Cellular Automata

Cellular Automata are studied as mathematical model for self organizing statistical systems [12]. CA can be one-dimensional or multi-dimensional. One-dimensional CA random number generators have been extensively studied in the past [4,11,15]. In one-dimensional CA, they can be considered as an array of cells where each cell is a one bit memory element. The neighbor set N(i) is defined as the set of cells on which the state transition function of the i-th cell is dependent on each iteration. In three-neighborhood CA, each cell evolves in every time step based on some combinatorial logic on the cell itself and its two nearest neighbors. More formally, for a three-neighborhood CA, the neighbor set of i-th cell is defined as $N(i) = \{s_{i-1}, s_i, s_{i+1}\}$. The state transition function of is i-th cell of 3-neighborhood CA is as follows:

$$s_i^{t+1} = f_i(s_{i-1}^t, s_i^t, s_{i+1}^t)$$

where, s_i^t denotes the current state of the i-th cell at time step t and s_i^{t+1} denotes the next state of the i-th cell at time step t+1 and f_i denotes some combinatorial logic for i-th cell. The set of all feedback functions is considered as ruleset for the CA. Since, a three-neighborhood CA having two states (0 or 1) in each cell, can have $2^3 = 8$ possible binary states, there are total $2^{2^3} = 256$ possible boolean functions, called rules. Each rule can be represented as

an decimal integer from 0 to 255. If the combinatorial logic for the rules have only Boolean XOR operation, then it is called linear or additive rule. Some of the three-neighborhood additive CA rules are 0, 60, 90, 102, 150 etc. Moreover, if the combinatorial logic contains AND/OR operations, then it is called non-linear rule. An n cell CA with cells $\{s_1, s_2, \cdots, s_n\}$ is called null boundary CA if $s_{n+1} = 0$ and $s_0 = 0$. Similarly for a periodic boundary CA $s_{n+1} = s_1$. A CA is called uniform, if all its cells follow the same rule. Otherwise, it is called non-uniform or hybrid CA. If all the ruleset of a hybrid CA are linear, then we call the CA a linear one. However, out of all possible Boolean functions, called rules, only two are of prime interest i.e. Rule 90 and 150 (ascertained from the decimal value of their position in the truth table). The state of the i-th cell at time instant t can be expressed as:

$$s_i^{t+1} = s_{i-1}^t \oplus d_i.s_i^t \oplus s_{i+1}^t, d_i = \begin{cases} 0, \text{ if } d_i \rightarrow \text{Rule 90} \\ 1, \text{ if } d_i \rightarrow \text{Rule 150} \end{cases}$$

Thus, an LHCA can be completely specified by a combination of Rule 90 and 150, denoted as an n-tuple $[d_1, d_2, \cdots, d_n]$. An example of a 5-cell CA $\mathcal{L}$ can be found in Fig. 1, specified by the rule vector $[1, 1, 1, 1, 0]$. Further details of CA can be found in [4].

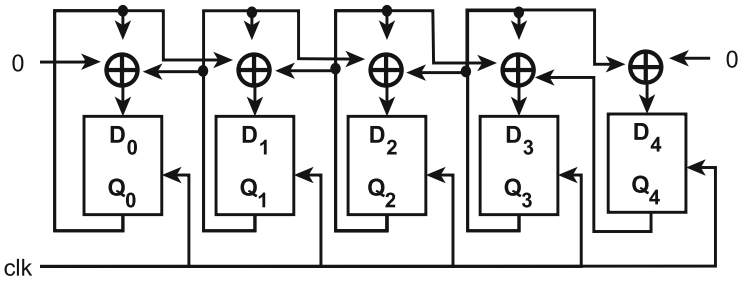


Fig. 1. 3-neighborhood null boundary LHCA $\mathcal{L}$ with rule vector [1, 1, 1, 1, 0]

3 5-Neighborhood Linear Cellular Automata

In the previous section, we have studied 1D elementary CA (i.e. 3-neighborhood CA) [4,11]. In this section, we consider a 5-neighborhood null boundary n-cell Linear Hybrid CA (LHCA) denoted by $\{s_1, s_2, \cdots, s_n\}$, where the state of a cell at a given instant is updated based upon its five neighboring cells including itself and because of null boundary $s_{-1} = s_0 = 0$, $s_{n+1} = s_{n+2} = 0$. More formally, for a five-neighborhood CA, the neighbor set of i-th cell is defined as $N(i) = \{s_{i-2}, s_{i-1}, s_i, s_{i+1}, s_{i+2}\}$. The state transition function of is i-th cell of 5-neighborhood CA is as follows:

$$s_i^{t+1} = f_i(s_{i-2}^t, s_{i-1}^t, s_i^t, s_{i+1}^t, s_{i+2}^t)$$

Table 1. Linear rules of 5-neighborhood LHCA

Rules	State transition function of i^{th} cell
$Rule_0$	$s_i^{t+1} = s_{i-2}^t \oplus s_{i+2}^t$
$Rule_1$	$s_i^{t+1} = s_{i-2}^t \oplus s_{i+1}^t \oplus s_{i+2}^t$
$Rule_2$	$s_i^{t+1} = s_{i-2}^t \oplus s_i^t \oplus s_{i+2}^t$
$Rule_3$	$s_i^{t+1} = s_{i-2}^t \oplus s_i^t \oplus s_{i+1}^t \oplus s_{i+2}^t$
$Rule_4$	$s_i^{t+1} = s_{i-2}^t \oplus s_{i-1}^t \oplus s_{i+2}^t$
$Rule_5$	$s_i^{t+1} = s_{i-2}^t \oplus s_{i-1}^t \oplus s_{i+1}^t \oplus s_{i+2}^t$
$Rule_6$	$s_i^{t+1} = s_{i-2}^t \oplus s_{i-1}^t \oplus s_i^t \oplus s_{i+2}^t$
$Rule_7$	$s_i^{t+1} = s_{i-2}^t \oplus s_{i-1}^t \oplus s_i^t \oplus s_{i+1}^t \oplus s_{i+2}^t$

where, s_i^t denotes the current state of the i-th cell at time step t and s_i^{t+1} denotes the next state of the i-th cell at time step t+1 and f_i denotes some combinatorial logic for i-th cell. Since, a 5-neighborhood CA having two states (0 or 1) in each cell, can have $2^5 = 32$ possible binary states, there are total $2^{2^5} = 2^{32}$ possible boolean functions. Out of all possible Boolean functions, called rules, there are total $2^5 = 32$ possible linear rules. Based on neighborhood radius exactly 5, there are only $2^3 = 8$ liner rules shown in Table 1.

Table 2. Counting rule vectors of max. period 5-bit 5-neighborhood CA

	$Rule_0$	$Rule_1$	$Rule_2$	$Rule_3$	$Rule_4$	$Rule_5$	$Rule_6$
$Rule_1$	2						
$Rule_2$	0	2					
$Rule_3$	2	6	2				
$Rule_4$	2	2	2	4			
$Rule_5$	6	2	4	5	2		
$Rule_6$	2	4	2	2	6	5	
$Rule_7$	4	5	6	2	5	8	2

For all possible pair of these 8 linear rules, maximum period 5-neighborhood CA rule vectors can be obtained. Table 2 shows the number of rule vectors obtained for maximum period 5-bit 5-neighborhood CA against each pair of the linear rules shown in Table 1. From Table 2, we see that only the pair of rule combinations, ($Rule_5$, $Rule_7$), provides largest number of rule vectors (i.e. 8). Therefore, we consider these two linear rules (i.e. $Rule_5$, $Rule_7$), denoted as R_0 and R_1, respectively, to design 5-neighborhood LHCA. These two linear rules can again be specified as follows:

$$R_0 : s_i^{t+1} = s_{i-2}^t \oplus s_{i-1}^t \oplus s_{i+1}^t \oplus s_{i+2}^t$$
$$R_1 : s_i^{t+1} = s_{i-2}^t \oplus s_{i-1}^t \oplus s_i^t \oplus s_{i+1}^t \oplus s_{i+2}^t$$

where, s_i^t is the current state and s_i^{t+1} is the next state of the i-th cell of the CA. Thus, the state transition function of i-th cell of the CA can be expressed as:

$$s_i^{t+1} = s_{i-2}^t \oplus s_{i-1}^t \oplus d_i.s_i^t \oplus s_{i+1}^t \oplus s_{i+2}^t,\ d_i = \begin{cases} 0, & \text{if } i^{th} \text{ cell follows rule } R_0 \\ 1, & \text{if } i^{th} \text{ cell follows rule } R_1 \end{cases}$$

Thus, a five-neighborhood n-cell LHCA $\mathcal{L}$ denoted by $\{s_1, s_2, \cdots, s_n\}$, can be completely specified by a combination of these two rules R_0 and R_1, denoted as an n-tuple $[d_1, d_2, \cdots, d_n]$, called the rule vector of the CA. An example of a 5-cell null boundary 5-neighborhood CA can be found in Fig. 2, specified by the rule vector $[1, 1, 1, 0, 0]$.

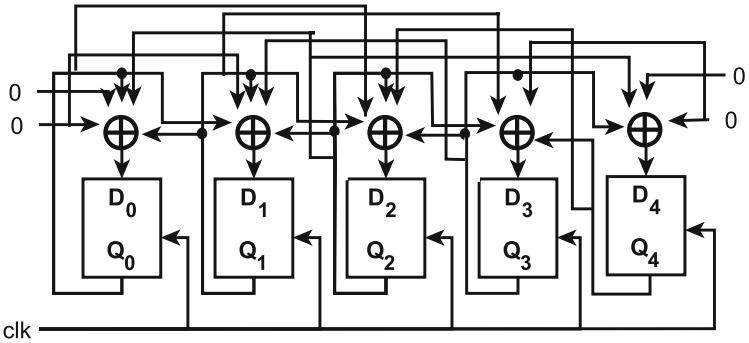


Fig. 2. 5-neighborhood null boundary LHCA $\mathcal{L}$ with rule vector [1, 1, 1, 0, 0]

A five-neighborhood n-cell LHCA $\mathcal{L}$ can be characterised by an $n \times n$ matrix, called characteristic matrix. The characteristic matrix A for the n-cell CA rule vector $[d_1, d_2, \cdots, d_n]$ is as follows:

$$A = \begin{bmatrix} d_1 & 1 & 1 & 0 & 0 & \cdots & \cdots & \cdots & 0 & 0 \\ 1 & d_2 & 1 & 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 1 & 1 & d_3 & 1 & 1 & \cdots & \cdots & \cdots & \cdots & \vdots \\ 0 & 1 & 1 & d_4 & 1 & \cdots & \cdots & \cdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \cdots & \cdots & \cdots & \cdots & 1 & d_{n-3} & 1 & 1 & 0 \\ \vdots & \cdots & \cdots & \cdots & \cdots & 1 & 1 & d_{n-2} & 1 & 1 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 1 & 1 & d_{n-1} & 1 \\ 0 & 0 & \cdots & \cdots & \cdots & 0 & 0 & 1 & 1 & d_n \end{bmatrix}$$

The state of a CA at time step t is an n-tuple formed from the states of the individual cells. The CA state is expressed in matrix form as follows

$$S^t = [s_1^t, \cdots, s_n^t]$$

The next state of the CA is denoted as

$$S^{t+1} = [s_1^{t+1}, \cdots, s_n^{t+1}]$$

The next-state of the CA, S^{t+1}, is computed as

$$(S^{t+1})^T = A \cdot (S^t)^T$$
$$\text{or,} \qquad S^{t+1} = ((S^{t+1})^T)^T$$

where, A is the CA transition matrix and $(S^t)^T = [s_1^t, \cdots, s_n^t]^T$ (the superscript T represents the transpose of the vector) and the product is a matrix-vector multiplication over GF(2). It has been shown that $A \cdot (S^t)^T$ is indeed the next state of the CA. Therefore, the next state of the i^{th} cell is computed as the product of the i^{th} row of A and $(S^t)^T$ as follows:

$$\begin{aligned} s_i^{t+1} &= [0, \cdots, 0, 1, 1, d_i, 1, 1, 0, \cdots, 0] \\ &\quad \cdot [s_1^t, \cdots, s_{i-2}^t, s_{i-1}^t, s_i^t, s_{i+1}^t, s_{i+2}^t \cdots, s_n^t]^T \\ &= s_{i-2}^t + s_{i-1}^t + d_i \cdot s_i^t + s_{i+1}^t + s_{i+2}^t \end{aligned}$$

The characteristic polynomial Δ_n of the n-cell CA is defined by

$$\Delta_n = |x\mathcal{I} - A|$$

where, x is an indeterminate, $\mathcal{I}$ is the identity matrix of order n, and A is the CA transition matrix. The matrix $x\mathcal{I} - A$ is called the characteristic matrix of the CA. The characteristic polynomial is a degree n polynomial in x.

The following example clearly illustrates how the characteristic polynomial of a 5-neighborhood linear CA can be computed using the characteristic matrix of the CA.

Example 1: Let us consider a 5-cell null boundary 5-neighborhood linear CA with the rule vector [1, 1, 1, 0, 0]. We have $[d_1, d_2, d_3, d_4, d_5] = [1, 1, 1, 0, 0]$. The transition matrix A is as follows:

$$A = \begin{bmatrix} d_1 & 1 & 1 & 0 & 0 \\ 1 & d_2 & 1 & 1 & 0 \\ 1 & 1 & d_3 & 1 & 1 \\ 0 & 1 & 1 & d_4 & 1 \\ 0 & 0 & 1 & 1 & d_5 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

The corresponding characteristic matrix is as follows:

$$x\mathcal{I} - A = \begin{bmatrix} x+d_1 & 1 & 1 & 0 & 0 \\ 1 & x+d_2 & 1 & 1 & 0 \\ 1 & 1 & x+d_3 & 1 & 1 \\ 0 & 1 & 1 & x+d_4 & 1 \\ 0 & 0 & 1 & 1 & x+d_5 \end{bmatrix} = \begin{bmatrix} x+1 & 1 & 1 & 0 & 0 \\ 1 & x+1 & 1 & 1 & 0 \\ 1 & 1 & x+1 & 1 & 1 \\ 0 & 1 & 1 & x & 1 \\ 0 & 0 & 1 & 1 & x \end{bmatrix}$$

where, x is an indeterminate, $\mathcal{I}$ is the identity matrix with dimension 5, and A is the CA transition matrix shown above. The characteristic polynomial Δ_5 of the 5-cell CA is defined as follows:

$$\Delta_5 = |x\mathcal{I} - A|$$

$$\Delta_5 = \begin{vmatrix} x+1 & 1 & 1 & 0 & 0 \\ 1 & x+1 & 1 & 1 & 0 \\ 1 & 1 & x+1 & 1 & 1 \\ 0 & 1 & 1 & x & 1 \\ 0 & 0 & 1 & 1 & x \end{vmatrix} = x^5 + x^4 + x^2 + x + 1$$

Theorem 1. Let Δ_n be the characteristic polynomial of a n-cell null boundary 5-neighborhood Linear CA with rule vector $[d_1, d_2, \cdots, d_n]$. Δ_n satisfies the following recurrence relation:

$$\Delta_{-3} = 0, \quad \Delta_{-2} = 0, \quad \Delta_{-1} = 0, \quad \Delta_0 = 1$$
$$\Delta_n = (x + d_n)\Delta_{n-1} + \Delta_{n-2} + (x + d_{n-1})\Delta_{n-3} + \Delta_{n-4},\ n > 0 \qquad (1)$$

Proof: Consider the transition matrix A for the n-cell null boundary 5-neighborhood Linear CA with rule vector $[d_1, d_2, \cdots, d_n]$

$$A = \begin{bmatrix} d_1 & 1 & 1 & 0 & 0 & \cdots & \cdots & \cdots & 0 & 0 \\ 1 & d_2 & 1 & 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 1 & 1 & d_3 & 1 & 1 & \cdots & \cdots & \cdots & \cdots & \vdots \\ 0 & 1 & 1 & d_4 & 1 & \cdots & \cdots & \cdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \cdots & \cdots & \cdots & \cdots & 1 & d_{n-3} & 1 & 1 & 0 \\ \vdots & \cdots & \cdots & \cdots & \cdots & 1 & 1 & d_{n-2} & 1 & 1 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 1 & 1 & d_{n-1} & 1 \\ 0 & 0 & \cdots & \cdots & \cdots & 0 & 0 & 1 & 1 & d_n \end{bmatrix}$$

The characteristic polynomial Δ_n of the CA is defined by

$$\Delta_n = |x\mathcal{I} - A|$$

$$\Delta_n = \begin{vmatrix} x+d_1 & 1 & 1 & 0 & 0 & \cdots & \cdots & \cdots & 0 & 0 \\ 1 & x+d_2 & 1 & 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 1 & 1 & x+d_3 & 1 & 1 & \cdots & \cdots & \cdots & \cdots & \vdots \\ 0 & 1 & 1 & x+d_4 & 1 & \cdots & \cdots & \cdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \cdots & \cdots & \cdots & \cdots & 1 & x+d_{n-3} & 1 & 1 & 0 \\ \vdots & \cdots & \cdots & \cdots & \cdots & 1 & 1 & x+d_{n-2} & 1 & 1 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 1 & 1 & x+d_{n-1} & 1 \\ 0 & 0 & \cdots & \cdots & \cdots & 0 & 0 & 1 & 1 & x+d_n \end{vmatrix}$$

By expanding the determinant shown above with respect to the last row, we can compute Δ_n as follows: $\Delta_n = (x + d_n) * \Delta_{n-1} + 1 * B + 1 * C$, where B and C with dimension $(n-1) \times (n-1)$ are as follows:

$$B = \begin{vmatrix} x+d_1 & 1 & 1 & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ 1 & x+d_2 & 1 & 1 & 0 & \cdots & \cdots & \cdots & 0 \\ 1 & 1 & x+d_3 & 1 & 1 & \cdots & \cdots & \cdots & \vdots \\ 0 & 1 & 1 & x+d_4 & 1 & \cdots & \cdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \cdots & \cdots & \cdots & \cdots & 1 & x+d_{n-3} & 1 & 0 \\ \vdots & \cdots & \cdots & \cdots & \cdots & 1 & 1 & x+d_{n-2} & 1 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 1 & 1 & 1 \end{vmatrix}$$

and

$$C = \begin{vmatrix} x+d_1 & 1 & 1 & 0 & 0 & \cdots & \cdots & 0 & 0 \\ 1 & x+d_2 & 1 & 1 & 0 & \cdots & \cdots & \cdots & 0 \\ 1 & 1 & x+d_3 & 1 & 1 & \cdots & \cdots & \cdots & \vdots \\ 0 & 1 & 1 & x+d_4 & 1 & \cdots & \cdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \cdots & \cdots & \cdots & \cdots & 1 & x+d_{n-3} & 1 & 0 \\ \vdots & \cdots & & \cdots & \cdots & 1 & 1 & 1 & 1 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 1 & x+d_{n-1} & 1 \end{vmatrix}$$

By expanding the determinant B with respect to the last column, we can compute B as follows: $B = \Delta_{n-2} + D$, where D with dimension $(n-2) \times (n-2)$ is as follows:

$$D = \begin{vmatrix} x+d_1 & 1 & 1 & 0 & 0 & \cdots & \cdots & 0 \\ 1 & x+d_2 & 1 & 1 & 0 & \cdots & \cdots & \cdots \\ 1 & 1 & x+d_3 & 1 & 1 & \cdots & \cdots & \cdots \\ 0 & 1 & 1 & x+d_4 & 1 & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \cdots & \cdots & \cdots & \cdots & 1 & x+d_{n-3} & 1 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 1 & 1 \end{vmatrix}$$

By expanding the determinant C with respect to the last column, we can compute C as follows: $C = E + F$, where E and F with dimension $(n-2) \times (n-2)$ are as follows:

$$E = \begin{vmatrix} x+d_1 & 1 & 1 & 0 & 0 & \cdots\cdots & \cdots & 0 \\ 1 & x+d_2 & 1 & 1 & 0 & \cdots\cdots & \cdots & \cdots \\ 1 & 1 & x+d_3 & 1 & 1 & \cdots\cdots & \cdots & \cdots \\ 0 & 1 & 1 & x+d_4 & 1 & \cdots\cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\ \vdots\ \vdots & \vdots & \vdots \\ \vdots & \cdots & \cdots & \cdots & \cdots\cdots & 1 & x+d_{n-3} & 1 \\ 0 & \cdots & \cdots & \cdots & \cdots\cdots & 1 & 1 & 1 \end{vmatrix}$$

and

$$F = \begin{vmatrix} x+d_1 & 1 & 1 & 0 & 0 & \cdots\cdots & \cdots & 0 \\ 1 & x+d_2 & 1 & 1 & 0 & \cdots\cdots & \cdots & \cdots \\ 1 & 1 & x+d_3 & 1 & 1 & \cdots\cdots & \cdots & \cdots \\ 0 & 1 & 1 & x+d_4 & 1 & \cdots\cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\ \vdots\ \vdots & \vdots & \vdots \\ \vdots & \cdots & \cdots & \cdots & \cdots\cdots & 1 & x+d_{n-3} & 1 \\ 0 & \cdots & \cdots & \cdots & \cdots\cdots & 0 & 1 & x+d_{n-1} \end{vmatrix}$$

By expanding the determinant F with respect to the last column, we can compute F as follows:

$$F = (x + d_{n-1}) * \Delta_{n-3} + \Delta_{n-4}$$

Note that the determinant E can be easily found by changing rows into columns and columns into rows of the determinant D, therefore, D and E determines the same polynomial and so, D+E determines zero in GF(2). Finally, we have

$$\begin{aligned} \Delta_n &= (x + d_n) * \Delta_{n-1} + 1 * B + 1 * C \\ &= (x + d_n) * \Delta_{n-1} + (\Delta_{n-2} + D) + (E + F) \\ &= (x + d_n) * \Delta_{n-1} + \Delta_{n-2} + F \\ &= (x + d_n) * \Delta_{n-1} + \Delta_{n-2} + (x + d_{n-1}) * \Delta_{n-3} + \Delta_{n-4} \end{aligned}$$

Theorem 1 provides an efficient algorithm to compute the Characteristic polynomial of a CA. Initially, Δ_{-3}, Δ_{-2}, Δ_{-1} are all set to zero and Δ_0 is set to one. Equation (1) is applied to obtain Δ_1. It is then reapplied to calculate Δ_2 from Δ_{-2} to Δ_1, Continuing, the polynomials $\Delta_3, \Delta_4, \cdots, \Delta_n$ are computed.

The following example clearly illustrates how the characteristic polynomial of a 5-neighborhood linear CA can be computed using the recurrence relation shown above. Table 3 shows characteristic polynomials of a 5-cell null boundary 5-neighborhood linear CA.

Example 2: Let us consider a 5-cell null boundary 5-neighborhood linear CA with the rule vector [1, 1, 1, 0, 0]. We have, $[d_1, d_2, d_3, d_4, d_5] = [1, 1, 1, 0, 0]$

$$
\begin{aligned}
\Delta_{-3} &= 0, \quad \Delta_{-2} = 0, \quad \Delta_{-1} = 0, \quad \Delta_0 = 1 \\
\Delta_1 &= (x + d_1)\Delta_0 + \Delta_{-1} + (x + d_0)\Delta_{-2} + \Delta_{-3} \\
&= (x + 1).1 + 0 + 0 + 0 = x + 1 \\
\Delta_2 &= (x + d_2)\Delta_1 + \Delta_0 + (x + d_1)\Delta_{-1} + \Delta_{-2} \\
&= (x + 1)(x + 1) + 1 + 0 + 0 = x^2 \\
\Delta_3 &= (x + d_3)\Delta_2 + \Delta_1 + (x + d_2)\Delta_0 + \Delta_{-1} \\
&= (x + 1)x^2 + (x + 1) + (x + 1) + 0 \\
&= x^3 + x^2 \\
\Delta_4 &= (x + d_4)\Delta_3 + \Delta_2 + (x + d_3)\Delta_1 + \Delta_0 \\
&= (x + 0)(x^3 + x^2) + x^2 + (x + 1)(x + 1) + 1 \\
&= x^4 + x^3 + x^2 + x^2 + 1 + 1 \\
&= x^4 + x^3 \\
\Delta_5 &= (x + d_5)\Delta_4 + \Delta_3 + (x + d_4)\Delta_2 + \Delta_1 \\
&= (x + 0)(x^4 + x^3) + (x^3 + x^2) + (x + 0)(x^2) + (x + 1) \\
&= x^5 + x^4 + x^3 + x^2 + x^3 + x + 1 \\
&= x^5 + x^4 + x^2 + x + 1
\end{aligned}
$$

3.1 Synthesis of 5-Neighborhood Linear CA

In this section, we present an algorithm Algorithm 1 for synthesizing 5-neighborhood CA from its characteristic polynomial.

Algorithm 1. Synthesis Algotithm

Input: The characteristic polynomial of an n-cell CA, Δ_n
Output: 5-neighborhood rule vector $[d_1, d_2, \cdots, d_n]$
Suppose, Δ_{n-1}, Δ_{n-2} and Δ_{n-3} are known and $\Delta_{-3} = \Delta_{-2} = \Delta_{-1} = 0, \Delta_0 = 1$.
Here, all operations are done in $GF(2)$.

1. Consider $\Delta_n = (x + d_n)\Delta_{n-1} + \Delta_{n-2} + (x + d_{n-1})\Delta_{n-3} + \Delta_{n-4}$
2. Compute $x + d_n$ using Division Algorithm
3. For k=n downto 3
4. Consider $\Delta_k = (x + d_k)\Delta_{k-1} + \Delta_{k-2} + (x + d_{k-1})\Delta_{k-3} + \Delta_{k-4}$
5. Compute $x + d_{k-1}$ *and* Δ_{k-4} using Division Algorithm
   End for
6. Consider $\Delta_1 = (x + d_1)\Delta_0$
7. Compute $x + d_1$
8. **Return** $[d_1, d_2, \cdots, d_n]$

Explanation: Suppose, Δ_{n-1}, Δ_{n-2} and Δ_{n-3} are known. Here, all operations are done in $GF(2)$. We consider the recurrence relation:

$$\Delta_n = (x + d_n)\Delta_{n-1} + \Delta_{n-2} + (x + d_{n-1})\Delta_{n-3} + \Delta_{n-4}$$

Table 3. Characteristic polynomials of null boundary 5-neighborhood LHCA

Sl No.	Rule vector	Characteristic polynomial	Primitive polynomial
1	00000	$x^5 + x^3 + x$	NO
2	00001	$x^5 + x^4 + x^3 + x^2 + x$	NO
3	00010	$x^5 + x^4 + x^3 + x + 1$	YES
4	00011	$x^5 + x^2 + 1$	YES
5	00100	$x^5 + x^4 + x^3 + x^2 + x + 1$	NO
6	00101	$x^5 + x + 1$	NO
7	00110	$x^5 + x^2$	NO
8	00111	$x^5 + x^4 + x^2 + x + 1$	YES
9	01000	$x^5 + x^4 + x^3 + x + 1$	YES
10	01001	$x^5 + x^2 + x + 1$	NO
11	01010	$x^5 + x$	NO
12	01011	$x^5 + x^4 + 1$	NO
13	01100	$x^5 + x^2$	NO
14	01101	$x^5 + x^4 + x^2$	NO
15	01110	$x^5 + x^4 + x + 1$	NO
16	01111	$x^5 + x^3 + x + 1$	NO
17	10000	$x^5 + x^4 + x^3 + x^2 + x$	NO
18	10001	x^5	NO
19	10010	$x^5 + x^2 + x + 1$	NO
20	10011	$x^5 + x^4 + x^2 + x$	NO
21	10100	$x^5 + x + 1$	NO
22	10101	$x^5 + x^4$	NO
23	10110	$x^5 + x^4 + x^2$	NO
24	10111	$x^5 + x^3 + x^2 + x + 1$	YES
25	11000	$x^5 + x^2 + 1$	YES
26	11001	$x^5 + x^4 + x^2 + x$	NO
27	11010	$x^5 + x^4 + 1$	NO
28	11011	$x^5 + x^3 + x$	NO
29	11100	$x^5 + x^4 + x^2 + x + 1$	YES
30	11101	$x^5 + x^3 + x^2 + x + 1$	YES
31	11110	$x^5 + x^3 + x + 1$	NO
32	11111	$x^5 + x^4 + x^3 + x^2 + x + 1$	NO

0-Rule R_0; 1-Rule R_1

Now, we follow the Table 4. In the step 1, Δ_n and Δ_{n-1} are known. By the polynomial division algorithm, considering Δ_n as dividend and Δ_{n-1} as divisor, the degree 1 quotient polynomial $(x + d_n)$ is uniquely determined and easily

calculated; since, the remainder polynomial in the relation (i.e. $\Delta_{n-2} + (x + d_{n-1})\Delta_{n-3} + \Delta_{n-4}$) is of degree less than $n-1$. In the step 2, Δ_n, Δ_{n-1}, Δ_{n-2} and Δ_{n-3} are known. In the above relation, the polynomial $\Delta_n + (x+d_n)\Delta_{n-1} + \Delta_{n-2}$ is of degree $n-2$. Now, if the polynomial division algorithm is again applied considering $\Delta_n + (x + d_n)\Delta_{n-1} + \Delta_{n-2}$ as dividend and Δ_{n-3} as divisor then, it will calculate $(x + d_{n-1})$ as quotient and Δ_{n-4} as remainder from the above relation. In the step 3, we consider the relation:

$$\Delta_{n-1} = (x + d_{n-1})\Delta_{n-2} + \Delta_{n-3} + (x + d_{n-2})\Delta_{n-4} + \Delta_{n-5}$$

Now, Δ_{n-1}, Δ_{n-2}, Δ_{n-3} and Δ_{n-4} are known and $(x+d_{n-1})$ is also known as it is computed in the previous step. If we apply the division algorithm considering $\Delta_{n-1} + (x+d_{n-1})\Delta_{n-2} + \Delta_{n-3}$ as dividend and Δ_{n-4} as divisor, it can calculate $(x + d_{n-2})$ as quotient and Δ_{n-5} as remainder from the above relation. In this way, if we proceed for n steps, then we get the sequence of degree 1 quotient polynomials as follows:

$$[(x + d_n), (x + d_{n-1}), (x + d_{n-2}), \cdots, (x + d_2), (x + d_1)]$$

where $d_k (1 \leq k \leq n)$ is either 0 or 1. By taking the constant terms of these quotient polynomials and reversing, we get the rule vector $[d_1, d_2, \cdots, d_n]$ for a 5-neighborhood LHCA with the characteristic polynomial Δ_n. The total number of polynomial divisions performed is O(n), where, n is degree of the characteristic polynomial Δ_n of n-bit CA. Each polynomial division needs $O(n^2)$ time. Therefore, the required time complexity for this algorithm is $O(n^3)$.

3.2 Randomness of 5-Neighborhood Linear CA Rule Vectors

A statistical test suite is developed by National Institute of Standards and Technology (NIST) that is known as NIST-statistical test suite [1]. The NIST Test Suite is a statistical package consisting of 15 tests that were developed to test the randomness of (arbitrarily long) binary sequences produced by either hardware or software based cryptographic random or pseudorandom number generators. To test the randomness of 5-neighborhood linear CA rule vectors, we consider a 24-bit 5-neighborhood maximum period LHCA with rule vector

$$[1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 1]$$

where $d_i = 0$ in the rulevector $[d_0, \cdots, d_{23}]$ represents that i^{th} cell of the CA follows rule R_0 and $d_i = 1$ in the rulevector $[d_0, \cdots, d_{23}]$ represents that i^{th} cell of the CA follows rule R_1. 100 bit-streams with each stream of 1,00,000 bits are generated from the middle cell ($12^{th} cell$) of this 24-bit LHCA and stored in a data file, and then the data file is fed to NIST test suite. The generated bit-streams show high randomness property as depicted in Table 5.

3.3 Diffusion Property of 5-Neighborhood Linear CA Rule Vectors

To test the diffusion property of 5-neighborhood linear CA rule vectors, we consider a 24-bit 5-neighborhood maximum period LHCA $[s_0, \cdots, s_{23}]$ with the same rule vector

Table 4. Synthesis of 5-neighborhood linear CA

Step	Known quotient	Known poly, subpoly	Relation used	Evaluated quotient	Evaluated sub-poly
1	—	Δ_n, Δ_{n-1}	$\Delta_n = (x+d_n)\Delta_{n-1} + \Delta_{n-2} + (x+d_{n-1})\Delta_{n-3} + \Delta_{n-4}$	$x+d_n$	—
2	$x+d_n$	$\Delta_n, \Delta_{n-1}, \Delta_{n-2}, \Delta_{n-3}$	$\Delta_n = (x+d_n)\Delta_{n-1} + \Delta_{n-2} + (x+d_{n-1})\Delta_{n-3} + \Delta_{n-4}$	$x+d_{n-1}$	Δ_{n-4}
3	$x+d_{n-1}$	$\Delta_{n-1}, \Delta_{n-2}, \Delta_{n-3}, \Delta_{n-4}$	$\Delta_{n-1} = (x+d_{n-1})\Delta_{n-2} + \Delta_{n-3} + (x+d_{n-2})\Delta_{n-4} + \Delta_{n-5}$	$x+d_{n-2}$	Δ_{n-5}
4	$x+d_{n-2}$	$\Delta_{n-2}, \Delta_{n-3}, \Delta_{n-4}, \Delta_{n-5}$	$\Delta_{n-2} = (x+d_{n-2})\Delta_{n-3} + \Delta_{n-4} + (x+d_{n-3})\Delta_{n-5} + \Delta_{n-6}$	$x+d_{n-3}$	Δ_{n-6}
⋮	⋮	⋮	⋮	⋮	⋮
n-3	$x+d_5$	$\Delta_5, \Delta_4, \Delta_3, \Delta_2$	$\Delta_5 = (x+d_5)\Delta_4 + \Delta_3 + (x+d_4)\Delta_2 + \Delta_1$	$x+d_4$	Δ_1
n-2	$x+d_4$	$\Delta_4, \Delta_3, \Delta_2, \Delta_1, \Delta_0$	$\Delta_4 = (x+d_4)\Delta_3 + \Delta_2 + (x+d_3)\Delta_1 + \Delta_0$	$x+d_3$	—
n-1	$x+d_3$	$\Delta_3, \Delta_2, \Delta_1, \Delta_0, \Delta_{-1}$	$\Delta_3 = (x+d_3)\Delta_2 + \Delta_1 + (x+d_2)\Delta_0 + \Delta_{-1}$	$x+d_2$	—
n	—	Δ_1, Δ_0	$\Delta_1 = (x+d_1)\Delta_0$	$x+d_1$	—

Table 5. Results of NIST-statistical test suite

Sl. No	Test name	P-value	Status
1	Frequency test	0.883171	Pass
2	BlockFrequency (block len.=128)	0.851383	Pass
3	Cumulative sums	0.574903	Pass
4	Runs	0.383827	Pass
5	Longest run	0.867692	Pass
6	FFT	0.401199	Pass
7	Non-OverlappingTemplate (block len.=9)	0.474986	Pass
8	OverlappingTemplate (block len.=9)	0.066882	Pass
9	ApproximateEntropy (block len.=10)	0.798139	Pass
10	Random excursions test	0.350485	Pass
11	Random excursions variant Test	0.534146	Pass

$$[1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,1,1,0,1]$$

as considered in the previous section, and some CA initial values and we notice the status of the CA cells in some clock cycles. The result of the CA states for some clock cycles is depicted in Table 6. The result shows that the diffusion rate of CA cell contents is 2 times faster than 3-neighborhood CA. For the sake of simplicity, the rule value of the CA is given in hexadecimal notation i.e. a CA

rule value $0xA5$ denotes the rule vector $[1, 0, 1, 0, 0, 1, 0, 1]$ and a CA initial value $0xA5$ denotes the CA value [10100101].

Table 6. Diffusion of 5-neighborhood LHCA rule vector

	CA initial (in Hex)	Remarks
Average case	000800	12^{th} cell bit is diffused to MSB/LSB in 6/7 clock cycles, respectively.
	001000	11^{th} cell bit is diffused to MSB/LSB in 11/6 clock cycles, respectively.
Worst case	800000	0^{th} cell bit is diffused to LSB in 16 clock cycles
	000001	23^{rd} cell bit is diffused to MSB in 16 clock cycles

Table 7. Comparison of 5-neighborhood linear CA with 3/4 neighborhood CA

Properties	3-neighborhood LHCA	4-neighborhood LHCA	5-neighborhood LHCA
State transition function of i^{th} cell	[a] $s_i^{t+1} = f_i(s_{i-1}^t, s_i^t, s_{i+1}^t)$	$s_i^{t+1} = f_i(s_{i-1}^t, s_i^t, s_{i+1}^t, s_{i+2}^t)$ or $s_i^{t+1} = f_i(s_{i-2}^t, s_{i-1}^t, s_i^t, s_{i+1}^t)$	$s_i^{t+1} = f_i(s_{i-2}^t, s_{i-1}^t, s_i^t, s_{i+1}^t, s_{i+2}^t)$
# of linear rules (neighborhood radius at most r, r=3,4,5)	$2^3 = 8$	$2^4 = 16, 2^4 = 16$	$2^5 = 32$
# of linear rules (neighborhood radius exactly r, r=3,4,5)	$2^1 = 2$	$2^2 = 4, 2^2 = 4$	$2^3 = 8$
Rules combinations (with largest no. of max period CA rule vectors)	< Rule 90, Rule 150 >	$< s_i^{t+1} = s_{i-1}^t \oplus s_{i+1}^t \oplus s_{i+2}^t$, $s_i^{t+1} = s_{i-1}^t \oplus s_i^t \oplus s_{i+1}^t \oplus s_{i+2}^t >$ or $< s_i^{t+1} = s_{i-2}^t \oplus s_{i-1}^t \oplus s_{i+1}^t$, $s_i^{t+1} = s_{i-2}^t \oplus s_{i-1}^t \oplus s_i^t \oplus s_{i+1}^t >$	$< R_0, R_1 >$ [b]
Diffusion rate of n-bit CA (Average case)	At least $n/2$ clock cycles	At least $n/4$ clock cycles	At least $n/4$ clock cycles
Diffusion rate of n-bit CA (Worst case)	At most $(n-1)$ clock cycles	At most $(n-1)$ clock cycles	At most $3n/4$ clock cycles

a s_i^{t+1} denotes the state of the i-th cell at time step t+1
b Rules R_0, R_1 are defined in Sect. 3.

3.4 Comparison of Properties of 5-Neighborhood Linear CA with 3/4 Neighborhood Linear CA

In this section, we study the comparison of properties of 5-neighborhood linear CA with 3/4 neighborhood linear CA, shown in Table 7. Delay will obviously increase for 5-neighborhood CA with respect to 3-neighborhood CA. On the other hand, one clock cycle period is at least the time period required for one time CA evolving and the average diffusion rate for 5-neighborhood CA is 2 times faster than 3-neighborhood CA. Therefore, because of high diffusion rate, 5-neighborhood CA is also suitable for high speed application.

4 Conclusion

In this paper, we have studied 5-neighborhood null boundary linear CA with two linear rules. The characteristic polynomial has been realized from 5-neighborhood rule vector of the CA. We have presented an algorithm for synthesizing the 5-neighborhood CA from its characteristic polynomial by assuming some CA sub-polynomials. We have shown the randomness and diffusion properties of the 5-neighborhood CA rule vectors and the comparison of their properties with 3/4 neighborhood CA. At present, we are working on how the CA can be synthesized from its characteristic polynomial without the knowledge of CA sub-polynomials.

References

1. NIST SP 800-22: A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications. U.S. Department of Commerce (2010)
2. Bardell, P.H.: Analysis of cellular automata used as pseudorandom pattern generators. In: Proceedings IEEE International Test Conference 1990, Washington, D.C., USA, 10–14 September 1990, pp. 762–768 (1990)
3. Bouganim, L., Guo, Y.: Database Encryption in Encyclopedia of Cryptography and Security, 2nd edn. Springer, New York (2010)
4. Chaudhuri, P.P., Roy Chowdhury, D., Nandi, S., Chattopadhyay, S.: Additive Cellular Automata: Theory and Applications. IEEE Computer Socity Press, New York (1997)
5. la Guaz-Martinez, D., Fuster-Sabater, A.: Cryptographic design based on cellular automata In: Proceedings of IEEE International Symposium on Information Theory, p. 180 (1997)
6. Ikenaga, T., Ogura, T.: Real-time morphology processing using highly parallel 2-D cellular automata. IEEE Trans. Image Process. **9**(12), 2018–2026 (2000)
7. Jose, J., Roy Chowdhury, D.: Four neighbourhood cellular automata as better cryptographic primitives. IACR Cryptology ePrint Archive 2015, 700 (2015)
8. Kumar, K.J.J., Sudharsan, S., Karthick, V.: FPGA implementation of cellular automata based stream cipher: Yugam-128. IJAREEIE **3** (2014)
9. Tomassini, M., Sipper, M., Perrenoud, M.: On the generation of high-quality random numbers by two-dimensional cellular automata. IEEE Trans. Comput. **49**, 1146–1151 (2000)

10. Matsumoto, M.: Simple cellular automata as pseudorandom m-sequence generators for built-in self-test. ACM Trans. Model. Comput. Simul. **8**(1), 31–42 (1998)
11. Nandi, S., Kar, B.K., Chaudhuri, P.P.: Theory and applications of cellular automata in cryptography. IEEE Trans. Comput. **43**(12), 1346–1357 (1994)
12. Neumann, J.V.: The Theory of Self- Reproducing Automata. University of Illinois Press Urbana (1966). (Edited by Burks, A.W.)
13. Roy Chowdhury, D., Sengupta, I., Chaudhuri, P.P.: A class of two-dimensional cellular automata and their applications in random pattern testing. J. Electron. Test. **5**(1), 67–82 (1994)
14. Sudhakar, P., Chinnarao, B., Latha, D.M.M.: Optimization of 1D and 2D cellular automata for pseudo random number generator. IOSR J. VLSI Sig. Proc. (IOSR-JVSP) **4**, 28–33 (2014)
15. Wolfram, S.: Theory and Applications of Cellular Automata (Including Selected Papers 1983–1986). World Scientific Pub. Co., Inc., River Edge (1986)

Cache Optimized Solution for Sparse Linear System over Large Order Finite Field

A.K. Bhateja[1] and Vaishnavi Kannan[2(✉)]

[1] Defence Research and Development Organisation, New Delhi, India
akbhateja@gmail.com
[2] Delhi Technological University, Rohini, India
kannan.vaishnavi25@gmail.com

Abstract. Many mathematical, engineering and cryptographic applications require the solution of sparse linear equations over large order finite fields. The Gaussian elimination is a standard algorithm used for the above. However, its use remains limited because of its implementation difficulty for large matrices. For large and sparse linear systems the iterative Lanczos and Wiedemann are the most efficient techniques. However, the computation intensive matrix vector multiplications in these algorithms make them unsuitable for large systems, increasing the computation time due to constant accesses to the RAM and hard disk for fetching and storing data. In this paper we present a cache optimized implementation of the Lanczos and Wiedemann algorithm that can be used for very large matrices even when there is not sufficient cache to store all the non zero matrix elements. Our algorithm makes optimal use of the cache, decreases the number of memory accesses and therefore reduces the time taken for the algorithms to provide a solution. The results show an improvement of 16% in Lanczos and 13% in Wiedemann in the execution time, with number of equations as 105 and same numbers of variables over the field of order 529 bits.

Keywords: Sparse matrices · Finite field · Cache · Lanczos algorithm · Wiedemann algorithm

1 Introduction

Public key cryptosystems like the RSA and the Diffie Hellman key exchange rely on the integer factorization and the discrete logarithmic problem (Odlyzko 1984). Factorizing integers and computing the discrete transform are difficult tasks and often involve the solution to a system of large and sparse linear equations over finite fields GF(p). For small systems the Gaussian elimination works perfectly. However as the size and sparsity increases the Gaussian elimination becomes inefficient because of its fill-in problem. The Gaussian elimination can turn a sparse system to a dense one, to find a solution. The iterative methods that use successive approximation to obtain accurate solutions, like Lanczos algorithm (Lanczos 1952) and Wiedemann algorithm (Wiedemann 1986) work well for such systems.

© Springer Nature Singapore Pte Ltd. 2017
D. Giri et al. (Eds.): ICMC 2017, CCIS 655, pp. 84–95, 2017.
DOI: 10.1007/978-981-10-4642-1_8

But the problem with these algorithms is the computation intensive matrix-vector multiplications. This component increases the timing results of the program for large sparse systems where the data required for computation may be so large that it cannot fit into the cache. The access to the data would require constant calls to the hard disk and the ROM, increasing the time required for finding the solution. In this paper, we give a cache optimized implementation of the Lanczos and Wiedemann Algorithm for improvement in performance by the reduction in memory accesses, specifically for systems with low memory. Researchers have developed techniques for the cache efficient programs to solve compute intensive problems by optimally utilizing cache. The Cache efficient matrix transposition suggested in (Chatterjee & Sen 2000) studies the contributions of the data cache, the translation look aside buffer, register tiling, and the array layout function to the overall running time of the algorithms. Peter D. Sulatycke and Kanad Ghose suggested multithreaded fast multiplication of sparse matrices (Sulatycke & Ghose 1998). An efficient implementation of IIR and FIR filters by fitting the filter parameters in processor's cache (Ilmonen & Lokki 2006). This addresses the problem of a slower algorithm due to the inability of fitting the whole data into cache. The algorithm in (Zoican 2007) works by rearranging the filter equations to achieve a better cache hit rate. The convolution problem is broken down into a sum of sub-convolutions and several elements are computed together to achieve cache optimization. The efficient binary-mesh partitioning algorithm (Tchiboukdjian, Danjean, & Raffin 2010), and multiple string matching (Tan, Liu, & Liu 2008) aims to obtain efficient cache utilization for automata based algorithms by reducing the space requirements and by improving the cache locality for table-lookup-based algorithms. For solving dense system parallel solution for solving linear equations using Newton's iterative method (Pan & Reif 1989) was developed by choosing initial approximate inverse of the matrix. Preconditioned iterative method (Reif 1998) to find the approximate solution of sparse linear systems of equations was developed in which the condition number was obtained using algebraic and combinatorial methods. To solve matrix equations of the form $A_1 \times B_1 = F_1$ and $A_2 \times B_2 = F_2$, iterative approach (Ding, Liu, & Ding 2010) was designed by Ding et al. using iterative approach. They did not give any idea to select step size.

This paper is described in six sections. Section 2 provides an overview of the iterative Lanczos and Wiedemann algorithm. Section 3 gives an introduction to the CRS form for the storage of sparse matrices. Section 4 describes Cache optimized solution for sparse linear system over large order finite field. Section 5 describes the results obtained and Sect. 6 concludes.

2 Lanczos and Wiedemann Algorithms

Lanczos (Lanczos 1952) and Wiedemann (Wiedemann 1986) algorithms are the most common algorithms for finding the solution of linear system of equations over finite field. These algorithms use an iterative approach to obtain their solutions. In this section we discuss these two algorithms.

2.1 Lanczos Algorithm

Lanczos algorithm (Lanczos 1952) was invented for solving linear systems over real number field. LaMacchia and Odlyzko (LaMacchia & Odlyzko 1990) modified for solving linear system over finite field.

Let the system of linear equations is given by

$$Ax = w \tag{1}$$

where A is an $n*n$ square symmetric matrix and x and w are $n*1$ column matrices, over the finite field F.

The algorithm is given as follows-

Let the initial vector

$$w_0 = w. \tag{2}$$

Calculate

$$v_1 = Aw_0 \tag{3}$$

and

$$w_1 = v - \frac{(v_1, v_1)}{(v_0, v_1)} w_0 \tag{4}$$

For further values of i i.e. for $i >= 1$ define,

$$v_i + 1 = Aw_i, \tag{5}$$

$$w_{i+1} = v_{i+1} - \frac{(v_{i+1}, v_{i+1})}{(w_i, v_{i+1})} w_i - \frac{(v_{i+1}, v_i)}{(w_{i-1}, v_i)} w_{i-1} \tag{6}$$

and

$$H_i = \frac{(w_i, w)}{(w_i, v_{i+1})} \tag{7}$$

And the algorithm terminates when the condition $(w_k, Aw_k) = 0$ is satisfied because this gives a vector orthogonal to a set of n orthogonal vectors is a space of dimension of n. This happens for some $(k <= n)$. The solution is given by

$$x = \sum_{i=0}^{j-1} (H_i w_i) \tag{8}$$

However, in general the matrices to be solved are asymmetric and hence, the Lanczos needs to be modified to work for such matrices. Consider an asymmetric $m \times n\,(m \geq n)$ matrix B such that the system is given by

$$Bx{=}w' \tag{9}$$

A symmetric matrix A can then be formed as

$$A = B^T D^2 B \tag{10}$$

and

$$w = B^T D^2 w \tag{11}$$

where D is a diagonal matrix of the order $m \times m$ whose elements belong to F\{0}.

A solution to $Ax{=}b$ will then be a solution to (9).

However we do not need to calculate the matrix A to compute w_i. The vector matrix multiplication Aw_i can be computed as

$$B^T D^2 (BXw_i) \tag{12}$$

Let the number of non zeros be given as *nonz*. Let the cost of addition and multiplication be s_1 and s_2. The cost of computing (8) can then be given as $2 * nonz * s_1 + n * s_2$. Also each inner product costs about $n * s_1 + n * s_2$. The total cost of each iteration is then given by

$$T_i(n) = 2 * nonz \times s_1 + 4 * n * s_1 + 5 * n * s_2 \tag{13}$$

And the total cost for the running of the algorithm is then given by $n * T_i(n)$ for n iterations.

2.2 Wiedemann Algorithm

Wiedemann Algorithm (Wiedemann 1986) doesn't require the matrix A to be symmetric or positive-definite. Let $\mu_A(x)$ be the minimal polynomial of the matrix A. Wiedemann starts by probabilistically determining $\mu_A(x)$. Let

$$\mu_A(x) = x^d - C_{d-1}x^{d-1} - \ldots C_1 x - C_0 \tag{14}$$

where

$$d = deg(\mu_A(x)) <= n \tag{15}$$

Since

$$\mu_A(A) = 0, \tag{16}$$

from the Cayley Hamilton Algorithm, we have

$$A^k v - C_{d-1}A^{k-1}v - \ldots C_0 A^{k-d}v = 0 \tag{17}$$

Let v_k be the element of $A^k v$ at some particular position. The sequence v_k for $k >= 0$, satisfies the recurrence relation

$$v_k = C_{d-1}v_{k-1} + \ldots + C_1 v_1 + C_0 v_0 \tag{18}$$

For all $k >= d$. The minimal polynomial *C(x)* with degree $d' <= d$ can be calculated using the Berlekamp Massey Algorithm (Berlekamp 2015).

Put $k = d$ and $v = b$ in (17) to get

$$A(A^{d-1}b - C_{d-1}A^{d-2}b - c_1 Ab = c_0 b \tag{19}$$

If $C_0 \neq 0$, it becomes:

$$x = (C_0)^{-1}(A^{d-1}b - C_{d-1}A^{d-2}b - \ldots C_1 Ab \tag{20}$$

which is a solution to $Ax = b$.

The time consuming task in both Lanczos and Wiedemann is the computation of the matrix vector products $A^i b$ for $i = 0, 1, 2....$, which requires $O(n^2)$.

A. The calculation of matrix-vector multiplications will be needed to compute and check for the correct minimum polynomial in Wiedemann and for the computation of v_{i+1} values in Lanczos. This step involves the costliest matrix-vector multiplication. Also, in Wiedemann, the range of the loop variable i is twice the dimension of matrix A.
B. The calculation of the solution vector also, involves the computation intensive matrix-vector multiplication in addition to scalar-vector and vector-vector multiplication, vector addition and subtraction. This step, in comparison to the first, is less costlier as the range of the loop variable i is equal to the dimension of matrix A ($viz.$ $i = 0, 1, \ldots, n-1$) that is half of the range in the first step, for Wiedemann, though it is as costly as the first step considering a single iteration.

To improve the execution time of these algorithms we need to pay attention to optimally utilize cache so that the same element should not be fetched again and again memory.

3 Representation of Matrices

In RSA cryptanalysis and for finding the discrete log over a field of high order, the number of equations to be solved reaches the order of 10^5 or more. Also, the coefficient matrix A will have a majority of its elements as zero. To store such sparse matrices various methods are available, like the Compressed Column Storage Format (CCS) and Compressed Row Storage Format (CRS) (Bai, Demmel, Dongarra, Ruhe, & Vorst 2000), Jagged Diagonal Format (Saad 1989), Compressed Diagonal Storage Format (Bai et al. 2000) and linked list representation (Horowitz & Sahni 1983) are some methods used for the purpose. The CRS and CCS are the most efficient storage schemes due to their low memory requirements. The CRS maintains three arrays namely the value array, the column index and the row pointer. The value array stores the non- zero elements of the sparse matrix. The column index stores the column number corresponding to the non-zero elements stored in the value array and the row pointer array stores the beginning of each row. We used this scheme to store the sparse matrix for implementation. CRS for a matrix A method is described in Fig. 1.

Compressed Column Storage Format (CCS) is very similar to CRS. The only difference is, while storing in this format we are moving across a column first, storing the row number in the second array and the cumulative number of elements in a column in the final array.

$$A = \begin{bmatrix} 1\ 0\ 0\ 3 \\ 2\ 0\ 7\ 0 \\ 0\ 0\ 0\ 1 \\ 5\ 0\ 0\ 0 \end{bmatrix}$$

$$Value = \begin{bmatrix} 1\ 3\ 2\ 7\ 1\ 5 \end{bmatrix}$$

$$Col\_Index = \begin{bmatrix} 0\ 3\ 0\ 2\ 3\ 0 \end{bmatrix}$$

$$Row\_ptr = \begin{bmatrix} 0\ 2\ 4\ 5 \end{bmatrix}$$

Fig. 1. Representation of matrix A in CRS form.

For an $m \times n$ matrix, the CRS representation requires a total of $2 \times nonz + m$ instead of an initial $m \times n$ space for the storage of the matrix, where $nonz$ is the number of non zeros in the matrix A and m is the number of rows in the matrix.

In a similar way compressed column storage (CCS) can also be as shown in Fig. 2.

$$Value = \begin{bmatrix} 1\ 3\ 2\ 7\ 1\ 5 \end{bmatrix}$$

$$Row\_Index = \begin{bmatrix} 0\ 0\ 1\ 1\ 2\ 3 \end{bmatrix}$$

$$Col\_ptr = \begin{bmatrix} 0\ 2\ 4\ 5 \end{bmatrix}$$

Fig. 2. Representation of matrix A in CCS form.

4 Cache Optimized Solution for Sparse Linear System

Both Lanczos and Wiedemann algorithm include the compute intensive matrix-vector multiplication that makes a significant contribution to the running time of the algorithms. We have developed an algorithm for matrix vector multiplication with CRS representation by optimally utilizing the cache. Our algorithm for matrix-vector multiplication reduces the time taken by this matrix-vector multiplication, by making effective use of the available cache. Even though the CRS representation reduces the space required for the storage, the number of non- zeros themselves can increase to a limit where their storage may cause difficulty and insufficient memory, when the field order is large say 512 bits. The standard implementation of Lanczos and Wiedemann algorithm require that all the three arrays *value, column index* and *row pointer*. Since the cache cannot accommodate the

entire data due to insufficient space, the matrix vector multiplication works by making accesses to RAM or hard disk (if RAM is also not sufficient to store these arrays). The same set of data may be accessed and brought back to the cache for computation more than once. This increases the running time of the algorithms.

In our implementation, we retrieve an optimum number of non-zero values that can be completely accommodated in the cache, utilizing the complete capacity of cache, instead of retrieving the complete *value* and *column index* arrays. The required data is moved into the cache and processing is done on the retrieved data. This ensures that the computations are done within the cache. After the data in cache is processed, a new block of data is moved into the cache. The same set of non-zero values is not retrieved again. This method reduces the number of memory accesses for file read and write, reducing the time required to read files. Our results show a significant increase in the efficiency by the implementation of the above method. The program could also show an increase in the running time if the number of elements retrieved at a time are less than what the cache can be made to accommodate, thus increasing the number of avoidable cache accesses. Thus the method relies on a proper sensing of the available cache. The cache memory can be thought of as a buffer between the CPU registers of limited memory but high speed, and a comparatively slower but bigger main system memory (RAM). The similar operating speed of the Cache and the CPU prevents the CPU from waiting for the data. The configuration of the cache is such, that when data is to be read from RAM, the system first checks for the presence of data in the cache. If data is found in the cache, it is retrieved quickly to be used by the CPU.

If the data however, is not cached, the data is read from the RAM and transferred to the CPU. It is also cached for future references. If the CPU needs the same bit of data (a value from the same address), it will automatically look in cache first, which is much faster than RAM. The importance of the above mechanism also comes from the fact that all of this is done transparently with respect to the CPU so that the only difference is in the amount of time taken for the data to be retrieved. Transfer rate is not the only problem. Latency also reduces the CPU performance.

The other important reasons for the effectiveness of the cache are attributed to the exhibition of two forms of locality.

A. Spatial locality:- data within a block are likely to be fetched together.
B. Temporal locality:- data that has been recently used is likely to be used again in a short period of time.

The above suggest that benefits can be gained by implementing quickly accessible memory (temporal) and storing relevant information in small blocks (spatial) as efficiently as possible. When the dataset is large, it cannot fit into the small cache and needs to be stored in RAM/hard disk. Conclusively, we need to access RAM/hard disk for retrieving and storing the resultant and newly generated data. The access pattern to RAM/hard disk should be such as to minimize the number of accesses. Also, it would be preferable to store maximum amount of data in the cache to reduce the access time. However, since all of the data cannot be accommodated in the cache, we have retrieved only a small amount of data equal to the

available capacity of cache. Also, since data is stored sequentially i.e. in order of being retrieved, we benefit greatly from the spatial and temporal locality of cache.

Other improvements in the implementation were done as follows:-

A. For the generation of $A^i v$ iteratively for $i = 0, 1, \ldots, 2n-1$, we compute $A^i v$ in the i^{th} iteration by multiplying the matrix A and vector $A^{i-1}v$ (which has already been computed in the previous iteration). We maintain a single vector that stores only the previously computed vector $A^{i-1}v$. However, the rest of the previously computed $A^{i-1}v$ are stored in a file.
B. Only the matrix B was stored in CRS format. Matrix B^t is not stored explicitly. The same data used to store B is used for the computations with B^t. Corresponding adjustments are made in the program to use the data for B^t. As said before, the same data is not retrieved again. All the required computations using a set of data are done at once when the data is bought to the cache from the hard disk. This reduces the number of calls to the hard disk.

4.1 Our Algorithm

Let S be the size of available cache, *total_nonzeros* be the total number of non zeros in the matrix A. For every batch of non zeros to be processed, we need the entire row pointer vector, and non zero values from the *value* array and their corresponding column indices from the *column* array. Say, N non zeros are retrieved from the value array during one batch computation, then the N corresponding column indices also need to be brought into the cache. Hence, the total number of values bought to the cache are $2 \times N + rows$, which is equal to the total number of non-zeros that can be accommodated in the cache, *rem* be the remaining number of non zeros left to be brought to the cache and undergo computation. *col* array stores the column index, *val* array stores the non zero values and *row* array store the number of non zeros in a row, *rows* is the number of rows, the array C stores the result of matrix-vector multiplication.

Algorithm:-

int *rem = total_nonzeros*, *R = row[0]*, *r = 1*,*loop = 0;*
int *val[N]*, *col[N]*, *row[rows]*, *x[rows]*;
while $(rem > 0)$
    if $(rem >= N)$
        Retrieve next N values of *val* array and *col* array
        y = N;
        loop++;
        $rem = rem - N$;
    else if $(rem < N)$
        Retrieve next *rem* values of *val* array and *col* array
        y = rem;
        loop++;
        rem = rem-N;
    for *i =1* to N

Calculate the row j and col k of $val[loop \times N + i]$.
Multiply the element with $x[j]$ i.e. $C[j][k] = val[k] * x[j]$.

end

5 Experimental Results

The implementation was done on a 64-bit Intel core i3-2348M processor on Ubuntu 14.04 LTS operating system and a RAM of 3.6 GB. For arithmetic

Table 1. Improvement in execution time by optimally utilizing cache

Equations(Matrix Size)	Field Size (in bits)	Sparsity	% Improvement in Lanczos	% Improvement in Wiedemann
$10^3 \times 10^3$	131	1%	7.82	5.2
$10^3 \times 10^3$	131	2%	8.40	5.74
$10^3 \times 10^3$	131	3%	8.70	5.83
$10^3 \times 10^3$	263	1%	8.20	5.21
$10^3 \times 10^3$	263	2%	8.90	5.68
$10^3 \times 10^3$	263	3%	9.20	6.01
$10^3 \times 10^3$	529	1%	10.35	6.11
$10^3 \times 10^3$	529	2%	10.83	6.26
$10^3 \times 10^3$	529	3%	11.4	6.83
$10^4 \times 10^4$	131	1%	12.85	8.03
$10^4 \times 10^4$	131	2%	13.02	9.06
$10^4 \times 10^4$	131	3%	13.46	9.26
$10^4 \times 10^4$	263	1%	12.87	8.67
$10^4 \times 10^4$	263	2%	13.5	9.00
$10^4 \times 10^4$	263	3%	14.22	9.58
$10^4 \times 10^4$	529	1%	13.62	10.35
$10^4 \times 10^4$	529	2%	13.98	10.48
$10^4 \times 10^4$	529	3%	14.60	11.01
$10^5 \times 10^5$	131	1%	15.00	11.13
$10^5 \times 10^5$	131	2%	15.3	11.52
$10^5 \times 10^5$	131	3%	15.72	11.74
$10^5 \times 10^5$	263	1%	15.63	11.48
$10^5 \times 10^5$	263	2%	15.98	12.05
$10^5 \times 10^5$	263	3%	16.11	12.86
$10^5 \times 10^5$	529	1%	15.656	13.01
$10^5 \times 10^5$	529	2%	16.321	13.67
$10^5 \times 10^5$	529	3%	16.50	13.93

with large integers we used the GNU/MP library (Granlund 1991). The implementation was done in C language and compiled using GCC. Table 1 shows the results obtained from the implementation of Lanczos, Cache optimized Lanczos, Wiedemann and Cache optimized Wiedemann for solving $n \times n$ linear sparse system over finite field of size 131, 263 and 529 bits with n = $10^3, 10^4$ and 10^5.

The results are plotted and shown in the Figs. 3, 4, 5, 6, 7 and 8. The results show improvement in the timing of cache optimized Lanczos compared to the standard Lanczos and of Cache Optimized Wiedemann in comparison to Wiedemann. For the field of order 529 bits, the improvement in execution time is 16% in Lanczos and 13% in Wiedemann, with number of equations as 10^5. In general, an increase in the field size and the sparsity lead to a proportionate increase in the time taken for the algorithm to give a solution.

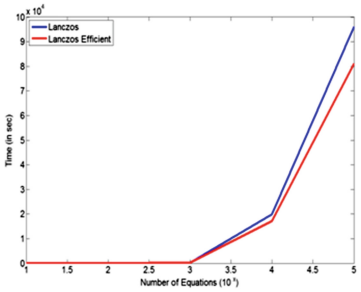


Fig. 3. Timing results of Lanczos and cache optimized Lanczos, (variation with number of equations), with sparsity 1% and field order 529 bits

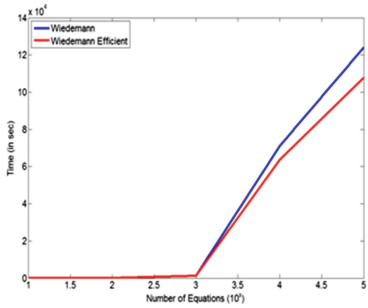


Fig. 4. Timing results of Wiedemann and cache optimized Wiedemann (variation with number of equations) with sparsity 1% and field order 529 bits

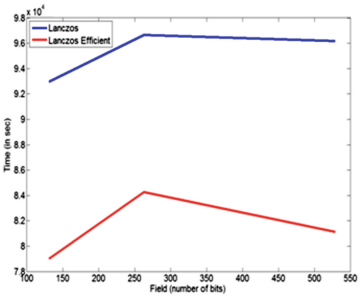


Fig. 5. Timing results of Lanczos and cache optimized Lanczos (variation with field order) with sparsity 1% and number of equations 10^5

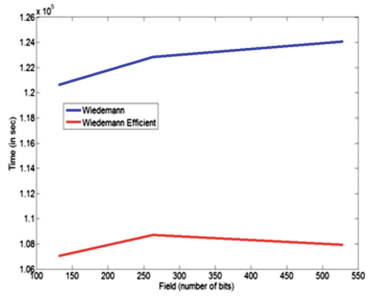


Fig. 6. Timing results of Wiedemann and cache optimized Wiedemann (variation with field order) with sparsity 1% and number of equations 10^5

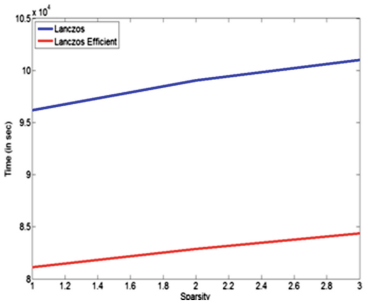


Fig. 7. Timing results of Lanczos and cache optimized Lanczos (variation with sparsity) with field order 529 bits and number of equations 10^5

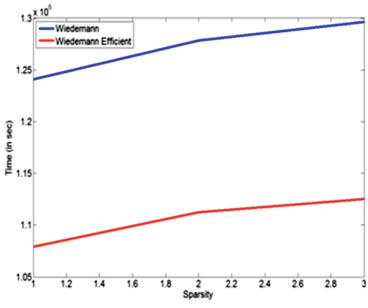


Fig. 8. Timing results of Wiedemann and cache optimized Wiedemann (variation with sparsity) with field order 529 bits and number of equations 10^5

6 Conclusion

In this paper, we have considered a cache optimized implementation of the Lanczos and Wiedemann algorithms with CRS implementation. Our program senses the cache to retrieve an optimal amount of data that can completely occupy the cache. The data retrieved is required only once. The processing is done with data available in the cache. Once, the required processing has been done, the data is removed from the cache and the next block of data is brought to the cache for further computations. The results show an improvement in execution time 16% in Lanczos and 13% in Wiedemann, with 10^5 number of equations over a field of order 529 bits.

References

Bai, Z., Demmel, J., Dongarra, J., Ruhe, A., van der Vorst, H.: Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide. SIAM, Philadelphia (2000)

Berlekamp, E.R.: Algebraic Coding Theory. World Scientific, Singapore (2015). (Revised ed.)

Chatterjee, S., Sen, S.: Cache-efficient matrix transposition. In: Paper Presented at the IEEE Sixth International Symposium on High-Performance Computer Architecture (HPCA-6) (2000)

Ding, J., Liu, Y., Ding, F.: Iterative solutions to matrix equations of the form Ai × Bi= Fi. Comput. Math. Appl. **59**(11), 3500–3507 (2010)

Granlund, T.: GMP, the GNU multiple precision arithmetic library (1991). Accessed http://gmplib.org/

Horowitz, E., Sahni, S.: Fundamentals of Data Structures. Pitman, London (1983). vol. 04; QA76. D35, H6

Ilmonen, T., Lokki, T.: Extreme filters-cache-efficient implementation of long IIR and FIR filters. IEEE Sig. Process. Lett. **13**(7), 401–404 (2006)

LaMacchia, B.A., Odlyzko, A.M.: Solving large sparse linear systems over finite fields. In: Menezes, A.J., Vanstone, S.A. (eds.) CRYPTO 1990. LNCS, vol. 537, pp. 109–133. Springer, Heidelberg (1991). doi:10.1007/3-540-38424-3_8

Lanczos, C.: Solution of systems of linear equations by minimized iterations. J. Res. Natl Bur. Stan. **49**(1) (1952)

Odlyzko, A.M.: Discrete logarithms in finite fields and their cryptographic significance. In: Beth, T., Cot, N., Ingemarsson, I. (eds.) EUROCRYPT 1984. LNCS, vol. 209, pp. 224–314. Springer, Heidelberg (1985). doi:10.1007/3-540-39757-4_20

Pan, V., Reif, J.: Fast and efficient parallel solution of dense linear systems. Comput. Math. Appl. **17**(11), 1481–1491 (1989)

Reif, J.H.: Efficient approximate solution of sparse linear systems. Comput. Math. Appl. **36**(9), 37–58 (1998)

Saad, Y.: Krylov subspace methods on supercomputers. SIAM J. Sci. Stat. Comput. **10**(6), 1200–1232 (1989)

Sulatycke, P.D., Ghose, K.: Caching-efficient multithreaded fast multiplication of sparse matrices. In: Paper Presented at the Parallel Processing Symposium 1998, IPPS/SPDP 1998 (1998)

Tan, J., Liu, Y., Liu, P.: Accelerating multiple string matching by using cache-efficient strategy. In: Paper Presented at the the Ninth International Conference on Web-Age Information Management 2008, WAIM 2008. IEEE(2008)

Tchiboukdjian, M., Danjean, V., Raffin, B.: Binary mesh partitioning for cache-efficient visualization. IEEE Trans. Vis. Comput. Graph. **16**(5), 815–828 (2010)

Wiedemann, D.H.: Solving sparse linear equations over finite fields. IEEE Trans. Inf. Theory **32**(1), 54–62 (1986)

Zoican, S.: Cache-efficient implementation of FIR filters using the Blackfin microcomputer. In: Paper Presented at the IEEE 8th International Conference on Telecommunications in Modern Satellite, Cable and Broadcasting Services 2007 (TELSIKS 2007) (2007)

Connected Fair Domination in Graphs

Angsuman Das[1(✉)] and Wyatt J. Desormeaux[2]

[1] Department of Mathematics, St. Xavier's College, Kolkata, India
angsumandas@sxccal.edu
[2] Department of Mathematics, University of Johannesburg, Auckland Park, South Africa
wjdesormeaux@gmail.com

Abstract. In this paper, we introduce the notion of connected fair domination in graphs. A connected fair dominating set in a graph G (or CFD-set) is a dominating set S such that $\langle S \rangle$ is connected in G and all vertices not in S are dominated by the same number of vertices from S, i.e., every two vertices not in S has the same number of neighbours in S. The connected fair domination number of G ($\mathsf{cfd}(G)$) is the minimum cardinality of a CFD-set in G. Apart from finding $\mathsf{cfd}(G)$ for some standard graphs G, we proved various bounds on $\mathsf{cfd}(G)$ in terms of order and some other graph parameters of G.

Keywords: Fair domination · Connected domination · Diagonal ramsey numbers

1 Introduction

The theory of domination in graphs has been an active area of research from the time of its inception. Two domination books [3,4] provide a comprehensive report of the vastness of research in the area of domination in graphs and its relation to other graph parameters. The notion of connected domination introduced in [5] gained a lot of attention due to its application in connectivity of networks, virtual backbone etc. In a simply modelled telecommunications network, the model consists of a central core of nodes and endnodes. The endnodes are client locations and the core nodes are interconnected and have switching ability. Naturally the core nodes are costly and one would want to minimize the number of core nodes while still maintaining their interconnectedness for example see [6]. This is ideally modelled by finding the connected domination number of the graph modelling your network and locating your core nodes at the vertices that form a minimum connected dominating set. The trouble is, you also must maintain fairness. In order to keep clients from feeling that they are not getting their fair share of resources, and that everyone has "equal" access to the network, it would be ideal for each endnode (client location) to have access (be adjacent) to the same number of core nodes. It is with this additional constraint on the connected domination number in mind, that we initiate in this paper the study of connected fair domination in graphs, which is an extension

© Springer Nature Singapore Pte Ltd. 2017
D. Giri et al. (Eds.): ICMC 2017, CCIS 655, pp. 96–102, 2017.
DOI: 10.1007/978-981-10-4642-1_9

of fair domination introduced by Caro *et.al.* in [1] and in [2]. For notation and graph theory terminology, we in general follow [3,7].

Let $G = (V, E)$ be a simple connected undirected n vertex graph and γ_c be its connected domination number. A set $S \subset V$ is said to be a connected k-fair dominating set, in short kCFD-set, if

1. S dominates G,
2. $\langle S \rangle$ is connected and
3. $\forall v \in V \setminus S, |N(v) \cap S| = k$.

Clearly for a connected graph G, V itself is a kCFD-set. The connected k-fair domination number of G, denoted by $\mathsf{cfd}_k(G)$, is the minimum cardinality of a kCFD-set in G. A kCFD-set of cardinality $\mathsf{cfd}_k(G)$ is called a $\mathsf{cfd}_k(G)$-set.

A connected fair dominating set, abbreviated as CFD-set, in G is a kCFD-set for some integer $k \geq 1$. The connected fair domination number, denoted by $\mathsf{cfd}(G)$, of a graph G is the minimum cardinality of a CFD-set in G, i.e., $\mathsf{cfd}(G) = \min\{\mathsf{cfd}_k(G)\}$, where the minimum is taken over all integers k where $1 \leq k \leq n-1$. A CFD-set of G of cardinality $\mathsf{cfd}(G)$ is called a $\mathsf{cfd}(G)$-set.

We start with some observations and inequalities involving $\mathsf{cfd}(G)$ for some standard graphs.

Observations

1. $\mathsf{cfd}(G) \leq n$ where n is the number of vertices in G.
2. $\gamma_c(G) \leq \mathsf{cfd}(G)$ where $\gamma_c(G)$ is the connected domination number of G.
3. $\mathsf{fd}(G) \leq \mathsf{cfd}(G)$, where $\mathsf{fd}(G)$ is the fair domination number of G. (See [1], for definition of $\mathsf{fd}(G)$).
4. $\mathsf{cfd}(P_n) = \mathsf{cfd}(C_n) = n - 2$, where P_n and C_n denote path and cycle on n vertices respectively.
5. $\mathsf{cfd}(K_n) = 1$ and $\mathsf{cfd}(K_{m,n}) = 2$, where K_n and $K_{m,n}$ denote the complete graph and complete bipartite graph.

2 Bounds on $\mathsf{cfd}(G)$ in Terms of Order of G

Theorem 1. *For any connected graph G with n vertices,* $\mathsf{cfd}(G) \leq n - 1$.

Proof: If $n = 2$, then $G = K_2$ and hence $\mathsf{cfd}(G) \leq 1$. Let $n \geq 3$. Then $\gamma_c(G) \leq n - 2$. Let D be a γ_c-set of G and $|D| \leq n - 2$. We choose $u \in V \setminus D$ and set $C = V \setminus \{u\}$. Clearly, C dominates G (as $D \subset C$) and 3rd condition also holds for C. Only thing remains to be shown is that $\langle C \rangle$ is connected. Let $a, b \in C$. Since a and b are either in D or adjacent to some vertices in D and $\langle D \rangle$ is connected, there exists a path from joining a and b in $\langle C \rangle$ and hence $\langle C \rangle$ is connected. Thus, C is a CFD-set in G with $n - 1$ vertices and thereby proving $\mathsf{cfd}(G) \leq n - 1$. □

Remark 1. The bound in Theorem 1 is tight. Consider the graphs in Fig. 1. They have $\mathsf{cfd}(G) = 4$ and 5 respectively. It is easy to check that there does not exist

any fair connected dominating set of size 3 or less for the first one (in left). For the other one (in right), we present a formal proof.

Consider the graph G on 6 vertices given in Fig. 1 (right). We prove that $\mathsf{cfd}(G) = 5$. If possible, let $\mathsf{cfd}(G) \leq 4$. Observe that a is a pendant vertex and b is a support vertex in G. Thus any one of them should be in any dominating set of G. As we are looking for connected dominating sets (CDS), b should be there. Now b dominates all the vertices in G except e. Thus to dominate e, either d or f should be in the CDS along with b.

Case 1: If $b, d \in$ CDS, to maintain connectedness of CDS, atleast one of c or f should be in CDS.

Case 1a: If $b, c, d \in$ CDS, then a, e, f are atleast dominated $1, 1, 3$ times respectively. So $\{b, c, d\}$ is not a CFD-set in G. Now if we include exactly one vertex in CDS other than a, then a will be dominated once by CDS and the last vertex will be dominated atleast twice. Thus only way to keep a outside CDS is to take all other vertices in CDS. That gives a CFD-set of size 5. On the other hand if we include a in CDS, i.e., $a, b, c, d \in$ CDS, f is dominated thrice and e is dominated only once. Thus, we need to include either e or f in CDS, thereby making it a CFD-set of size 5.

Case 1b: If $b, d, f \in$ CDS, then $\{b, d, f\}$ is not a CFD-set in G as e is dominated twice and a is dominated once. Similar to that in Case 1a, only way to keep a outside CDS is to take all other vertices in CDS. That gives a CFD-set of size 5. On the other hand if we include a in CDS, i.e., $a, b, d, f \in$ CDS, c is dominated thrice and e is dominated twice. Thus, the only option is to include either of c or e in CDS, thereby making it a CFD-set of size 5.

Case 2: If $b, f \in$ CDS, as already d is dominated twice and a can be dominated at most once, we need to either include all the vertices except a in CDS or we need to include a in CDS. As in the first case, we get a CFD-set of size 5, we include a in CDS, i.e., $a, b, f \in$ CDS. Now, c, d are dominated twice and e is dominated once. So, we need to include more vertices in CDS.

Case 2a: $a, b, c, f \in$ CDS. In this case d is dominated twice and e is dominated once. Thus we need to include one more vertex making it a CFD-set of size 5.

Case 2b: $a, b, d, f \in$ CDS. In this case c is dominated thrice and e is dominated twice. Thus we need to include one more vertex making it a CFD-set of size 5.

Case 2c: $a, b, e, f \in$ CDS. In this case c is dominated twice and d is dominated thrice. Thus we need to include one more vertex making it a CFD-set of size 5.

Thus, combining all cases, we get $\mathsf{cfd}(G) = 5$ and thereby proving the result. □

Remark 2. The gap $\mathsf{cfd}(G) - \gamma_c(G)$ can be arbitrarily large. Consider the following graph from [1]. For $n \geq 3$, define a graph on $2n$ vertices as follows: $V(G) = X \cup Y$ where $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_n\}$. Edges are defined as follows: $x_i \sim y_j$ if and only if $i \geq j$. For $i, j > 1$, $x_i \sim x_j$.

Clearly, $\{x_n, y_1\}$ is a connected dominating set and $\gamma_c(G) = 2$. It was proved in [1] that $\mathsf{fd}(G) = 2n - 2$. Since $\mathsf{cfd}(G) \geq \mathsf{fd}(G)$, we have $\mathsf{cfd}(G) - \gamma_c(G) \geq 2n - 2 - 2 = 2n - 4$. Thus, as n increases, $\mathsf{cfd}(G) - \gamma_c(G)$ increases arbitrarily. □

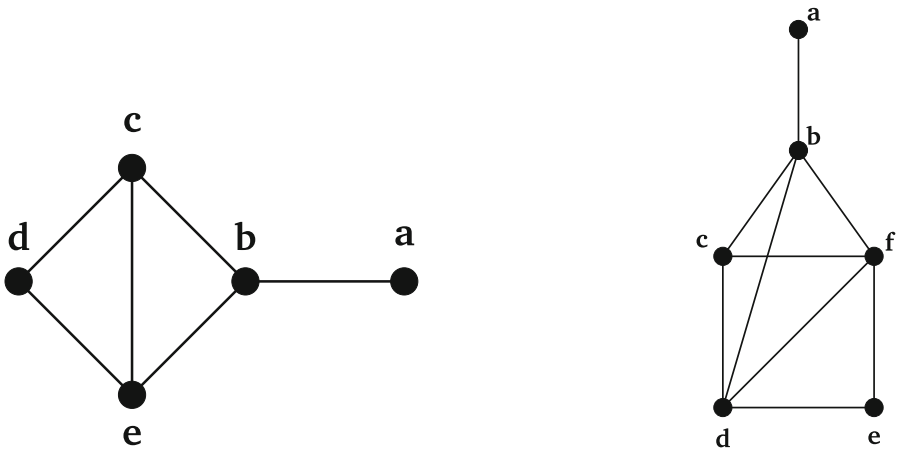


Fig. 1. Sharpness of upper bound

Theorem 2. *Let T be a tree with $n \geq 3$ vertices out of which l are pendant vertices. Then* $\mathsf{cfd}(T) = n - l$.

Proof: Let T be a tree with n vertices out of which l are pendant vertices. Then the set N of non-pendant vertices of T is a CFD-set of size $n - l$. Thus, $\mathsf{cfd}(T) \leq n - l$.

If possible, let $\mathsf{cfd}(T) < n - l$ and let S be a $\mathsf{cfd}(T)$-set. Then S does not contain atleast one non-pendant vertex, say u, and $deg(u) \geq 2$. Let v, w be two neighbours of u. If both $v, w \in S$, then S is not a CDS as there is no path joining v and w in $\langle S \rangle$ (only path joining v and w in G is $v \sim u \sim w$). If both $v, w \notin S$, then S contains a neighbour of u, other than v and w, say x, which dominates u. Again there exists some neighbour of v, say y (other than u), in S which dominates v. Now $\langle S \rangle$ being connected, there should be a path between x and y. As T is a tree, there exists a unique path in T, namely $x \sim u \sim v \sim y$ joining x and y. Since $u, v \notin S$, this path does not exist in $\langle S \rangle$, thereby making it disconnected. Thus, the only possibility left is when u has exactly two neighbours v, w, i.e., $deg(u) = 2$ and exactly one of them is in S. Let $v \in S$ and $w \notin S$. Since, $w \notin S$, there exists a neighbour of w, say z (other than u), in S which dominates w. Now $\langle S \rangle$ being connected, there should be a path between v and z. As T is a tree, there exists a unique path in T, namely $v \sim u \sim w \sim z$ joining v and z. Since $u, w \notin S$, this path does not exist in $\langle S \rangle$, thereby making it disconnected.

Thus, combining all the cases, it follows that all non-pendant vertices of T must be in S, i.e., $\mathsf{cfd}(T) \geq n - l$ and hence $\mathsf{cfd}(T) = n - l$. □

Corollary 1. *Let T be a tree with $n \geq 3$ vertices. Then* $\mathsf{cfd}(T) = \gamma_c(T)$. □

Corollary 2. *For any connected graph G with n vertices out of which l are pendant vertices,* $\mathsf{cfd}(G) \leq n - l$. □

Theorem 3. *For a connected regular graph G on $n \geq 3$ vertices,* $\mathsf{cfd}(G) \leq n-2$.

Proof: Let G be a connected r-regular graph. Let D be the minimum connected dominating set of G. Then $|D| \leq n-2$. Choose $u, v \in V \setminus D$ and set $S = V \setminus \{u, v\}$. As $D \subset S$, S dominates G. Let $a, b \in S$. Since a and b are either in D or adjacent to some vertices in D and $\langle D \rangle$ is connected, there exists a path from joining a and b in $\langle S \rangle$ and hence $\langle S \rangle$ is connected. Now, two cases may arise:

Case 1: $u \sim v$ in G. As G is r-regular, both u and v are adjacent to exactly $r-1$ vertices in S. Note that as $n \geq 3$, $r \geq 2$.

Case 1: $u \not\sim v$ in G. Similarly, in this case, both u and v are adjacent to exactly r vertices in S.

Thus, S is a CFD-set in G with $n-2$ vertices and thereby proving $\mathsf{cfd}(G) \leq n-2$. □

Theorem 4. *Let G be a graph on n vertices with diameter 2, maximum degree Δ and minimum degree δ. If $\Delta + \delta < n-1$, then* $\mathsf{cfd}(G) \leq n-2$.

Proof: Let v be a vertex of degree δ. As $diam(G) = 2$, $\langle N[v] \rangle$ is a connected dominating set of G. Consider the $n-\delta-1$ vertices in $V \setminus N[v]$. They have degrees lying between 1 and Δ. If $n - \delta - 1 > \Delta$ i.e., $\Delta + \delta < n-1$, by Pigeon-hole Principle, there exists at least two vertices $u_1, u_2 \in V \setminus N[v]$ with same degree k where $1 \leq k \leq \Delta$. Let $C = V \setminus \{u_1, u_2\}$. As $C \supset N[v]$, C dominates G.

For two vertices $u', u'' \in C$, either they are adjacent, or they are adjacent to same vertex in $N[v]$ or they are connected by a path u', v_1, v, v_2, u'' where $v_1, v_2 \in N[v]$. Thus $\langle C \rangle$ is connected.

Now, if $u_1 \sim u_2$ in G, then both u_1 and u_2 are adjacent to $k-1$ vertices in C. If $u_1 \not\sim u_2$ in G, then both are adjacent to k vertices in C. In any case, C is a connected fair dominating set in G. Thus $\mathsf{cfd}(G) \leq n-2$. □

Theorem 5. *Let G be a 3-connected graph on n vertices. Then* $\mathsf{cfd}(G) \leq n-2$.

Proof: Since G is 3-connected, we have $\delta \geq 3$. Now, there exists at least two vertices u, v in G such that $deg(u) = deg(v) = k$ where $3 \leq k \leq n-1$. Let $C = V \setminus \{u, v\}$. Since G is 3-connected, $\langle C \rangle$ is connected. Now, according as u and v are adjacent or non-adjacent in G, then u and v are adjacent to $k-1$ or k vertices in C. Thus C is a connected fair dominating set in G and hence $\mathsf{cfd}(G) \leq n-2$. □

Theorem 6. *Let G be a regular connected graph on n vertices with connected domination number γ_c. Moreover, let k be the highest positive integer such that $n - \gamma_c \geq R(k,k)$ where $R(k,k)$ is the diagonal Ramsey number. Then* $\mathsf{cfd}(G) \leq n-k$.

Proof: Let C be a γ_c-set of an r-regular graph G. Then $|V \setminus C| = n - \gamma_c$. Thus $\langle V \setminus C \rangle$ contains an independent set of size k or a clique of size k.

Case 1: Let $D \subseteq V \setminus C$ be an independent set of size k in $\langle V \setminus C \rangle$. Consider $V \setminus D$. We have $|V \setminus D| = n - k$. Since $C \subset V \setminus D$, $V \setminus D$ is a connected dominating set in G (using arguments similar to that used in proof of Theorem 1). Since $\langle D \rangle$ is an edgeless graph in $V \setminus C$, vertices in D are adjacent to exactly r vertices in $V \setminus D$. Thus $V \setminus D$ is a CFD-set in G and hence $\mathsf{cfd}(G) \leq n - k$.

Case 2: Let $D \subseteq V \setminus C$ be a clique of size k in $\langle V \setminus C \rangle$. Similar to that of Case 1, $V \setminus D$ is a connected dominating set of size $n - k$ in G. Since $\langle D \rangle$ is a clique in $V \setminus C$, vertices in D are adjacent to exactly $r - k + 1$ vertices in $V \setminus D$. Thus $V \setminus D$ is a CFD-set in G and hence $\mathsf{cfd}(G) \leq n - k$. □

Theorem 7. *For a connected graph G with n vertices and m edges,*

$$\frac{n}{\Delta(G) + 1} \leq \mathsf{cfd}(G) \leq m - 1.$$

Proof: The lower bound follows from that fact that $\gamma_c(G) \geq \frac{n}{\Delta(G)+1}$ and $\gamma_c(G) \leq \mathsf{cfd}(G)$. For the upper bound, first note that for a connected graph $m \geq n - 1$.

If $m = n - 1$ in a connected graph G, then G is a tree and in that case $\mathsf{cfd}(G) = n - l = m + 1 - l$, where l is the number of pendant vertices. Now as a tree contain atleast 2 pendant vertices, $\mathsf{cfd}(G) \leq m - 1$.

If $m \geq n$, then $\mathsf{cfd}(G) \leq n - 1 \leq m - 1$. □

3 Bounds on $\mathsf{cfd}(G)$ in terms of other graph parameters

An *outer-connected out-regular set*, abbreviated as OCOR-set is a subset Q of vertices in V such that $\langle V \setminus Q \rangle$ is connected and $|N(u) \cap (V \setminus Q)| = |N(v) \cap (V \setminus Q)| > 0$ for all $u, v \in Q$.

Let G be a connected graph and C be a γ_c-set of G. Then $|C| \leq n - 2$. Choose $u \in V \setminus C$ and set $Q = \{u\}$. Following the line of proof of Theorem 1, it can be shown that Q is an OCOR-set of G. Hence, every connected graph has a non-empty OCOR-set. The *outer-connected out-regular number* of a connected graph G, denoted by $\xi_{ocor}(G)$ is the maximum cardinality of an OCOR-set in G. An OCOR-set of size $\xi_{ocor}(G)$ is called a $\xi_{ocor}(G)$-set of G. It trivially follows that $\xi_{ocor}(G) \geq 1$ for any connected graph G.

Theorem 8. *For every connected graph G on n vertices, $\mathsf{cfd}(G) + \xi_{ocor}(G) = n$.*

Proof: Let D be a $\mathsf{cfd}(G)$-set. Then, by Theorem 1, $|D| \leq n - 1$. Let $Q = V \setminus D$. Then Q is an OCOR-set in G and hence $\xi_{ocor}(G) \geq |Q| = n - \mathsf{cfd}(G)$, i.e., $\mathsf{cfd}(G) + \xi_{ocor}(G) \geq n$.

On the other hand, let Q be an $\xi_{ocor}(G)$-set. Then $|Q| \geq 1$. Let $D = V \setminus Q$. Then D is a CFD-set in G and hence $\mathsf{cfd}(G) \leq |D| = n - \xi_{ocor}(G)$, i.e., $\mathsf{cfd}(G) + \xi_{ocor}(G) \leq n$. Thus we have the desired result. □

Theorem 9. *If G is a connected graph such that $\overline{G}$ is connected and $\overline{G}$ has an efficient dominating set, then $\mathsf{cfd}(G) = \mathsf{fd}(G) = \gamma(\overline{G})$.*

Proof: Let S be an efficient dominating set of $\overline{G}$ of size $\gamma_e(\overline{G})$. Then $|S| \geq 2$, because $|S| = 1$ implies that $\overline{G}$ has a universal vertex v (i.e., adjacent to all other vertices), which in turn implies that v is an isolated vertex in G contradicting that G is connected. Thus S dominates $\overline{G}$, S is an independent set in $\overline{G}$ and every vertex in $V(\overline{G}) \setminus S$ is adjacent to exactly one vertex of S in $\overline{G}$.

Hence, in G, every vertex in $V(G) \setminus S$ is adjacent to exactly $|S| - 1$ vertices of S and thus S is also a dominating set of G. Moreover, as S is an independent set in $\overline{G}$, $\langle S \rangle$ is a complete subgraph in G and hence connected in G. Thus, S is a connected $(|S| - 1)$-fair dominating set in G and hence $\mathsf{cfd}(G) \leq \gamma_e(\overline{G})$.

Now, we note that for any graph G' with an efficient dominating set, we have $\gamma_e(G') = \gamma(G')$. Thus, we have

$$\begin{aligned} \gamma_e(\overline{G}) = \gamma(\overline{G}) &= \mathsf{fd}(\overline{G}) && \text{(by Observation 2 in [1])} \\ &= \mathsf{fd}(G) && \text{(by Theorem 4(a) in [1])} \\ &\leq \mathsf{cfd}(G) && \text{(by Observation 3 in this paper)} \end{aligned}$$

Hence, we conclude that $\mathsf{cfd}(G) = \mathsf{fd}(G) = \gamma(\overline{G})$. □

4 Concluding Remarks

In this paper, we introduce the notion of connected fair domination number $\mathsf{cfd}(G)$ of a connected graph G and proved various bounds on $\mathsf{cfd}(G)$ in terms of the number of vertices and some other graph parameters. However, relationship of $\mathsf{cfd}(G)$ with respect to other graph parameters, still remain unexplored and can be an interesting topic for further investigation.

Acknowledgement. The research is partially funded by NBHM Research Project Grant, (Sanction No. 2/48(10)/2013/ NBHM(R.P.)/R&D II/695), Government of India.

References

1. Caro, Y., Hansberg, A., Henning, M.: Fair domination in graphs. Discrete Math. **312**, 2905–2914 (2012)
2. Hansberg, A.: Reviewing some results on fair domination in graphs. Electron. Notes Discrete Math. **43**, 367–373 (2013)
3. Haynes, T.W., Hedetniemi, S.T., Slater, P.J.: Fundamentals of Domination in Graphs. Marcel Dekker Inc., New York (1998)
4. Haynes, T.W., Hedetniemi, S.T., Slater, P.J. (eds.): Domination in Graphs: Advanced Topics. Marcel Dekker Inc., New York (1998)
5. Sampathkumar, E., Walikar, H.B.: The connected domination number of a graph. J. Math. Phy. Sci. **13**(6), 607–613 (1979)
6. Swamy, C., Kumar, A.: Primal-dual algorithms for connected facility location problems. Algorithmica **40**(4), 245–269 (2004)
7. West, D.B.: Introduction to Graph Theory. Prentice Hall, Upper Saddle River (2001)

Coordinating Particle Swarm Optimization, Ant Colony Optimization and K-Opt Algorithm for Traveling Salesman Problem

Indadul Khan[1(✉)], Manas Kumar Maiti[2], and Manoranjan Maiti[3]

[1] Department of Computer Science, Chandrakon Vidyasagar Mahavidyalaya, Paschim-Medinipur 721201, West Bengal, India
indadulkhan@gmail.com
[2] Department of Mathematics, Mahishadal Raj College, Mahishadal, Purba-Medinipur 721628, West Bengal, India
manasmaiti@yahoo.co.in
[3] Department of Applied Mathematics with Oceanology and Computer Programming, Vidyasagar University, Paschim-Medinipur, West Bengal, India

Abstract. In this paper combining the features of swap sequence and swap operation based Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO) and K-Opt operation a hybrid algorithm is proposed to solve well known Traveling Salesman Problem (TSP). Interchange of two cities of a path of a TSP is known as swap operation and a sequence of such operations is called swap sequence. Using swap operation and swap sequence PSO operations are redefined to solve TSP. Here ACO is used a small number of iterations to generate initial swarm of PSO. Then PSO operations are made on this swarm a sufficient number of times to find optimal path. During PSO iterations if a particle does not change its position for a predefined number of iterations then K-Opt operation is made on it a finite number of times to improve its position. The algorithm is tested with bench mark test problems from TSPLIB and it is observed that algorithm is more efficient with respect to accuracy as well as execution time to solve standard TSPs (Symmetric as well as Asymmetric) compared to existing algorithms. Details of the proposed algorithm along with swap operation, swap sequence and K-opt operation for the algorithm are elaborately discussed for the readers.

Keywords: Traveling salesmen problem · Ant colony optimization · Particle swarm optimization · Swap sequence · Swap operation · K-Opt

1 Introduction

The Traveling Salesmen Problem(TSP) is one of the standard combinatorial discrete optimization problem. The problem consists of a set of n vertices (node/cities) where distance between any two vertices is known. A salesman starts from a vertex, visits all the vertices exactly once and returned to the

© Springer Nature Singapore Pte Ltd. 2017
D. Giri et al. (Eds.): ICMC 2017, CCIS 655, pp. 103–119, 2017.
DOI: 10.1007/978-981-10-4642-1_10

starting vertex in such a way that the total distance traveled is a minimum. So the goal of the problem is to find a shortest possible tour through the set of vertices in such a way that each vertex is visited exactly once except for the starting vertex. It is also well-known NP-hard problem, can't be solved exactly in polynomial time [25,27]. Generally there are two approaches to solve a TSP exact methods and heuristic methods. The exact methods are required huge time for larger n, thus heuristic methods are typically used to solve a TSP. The exact methods include cutting plane [32], LP relaxation [6], branch and bound [39], branch and cut [36], etc. Only small size TSPs can be solved by exact methods in a reasonable time. On the other hand, several problems have been solved using heuristic or soft computing based techniques such as Ant Colony Optimization [9], local search [18], hybrid algorithm [12] and genetic algorithm [34]. In a TSP, when distance between vertices (node/cities) x_i and x_j is equal to the distance between vertex x_j and x_i then the problem is called Symmetric Traveling Salesmen Problem (STSP). Changdar et al. [4] solved a multi-objective solid TSP under fuzziness. In TSP with precedence constraint [33] there exists an order in which the vertices are to be visited. On the other hand, if the distance between vertices (node/city) x_i and x_j is not equal to the distance between vertices (node/cite) x_j and x_i, then the problem is called Asymmetric Traveling Salesmen problem (ATSP). Majumder and Bhunia [28] solved a ATSP with imprecise travel times using a genetic algorithm. In the TSP with time windows [12], each vertex is visited within a specified time window. In double TSP [38], the targets can be reached by two sales persons operating in parallel. Combing features of PSO, ACO and 3-Opt a hybrid algorithm PSO-ACO-3-Opt is presented by Mahi et al. [30] to solve TSP. Shi et al. [41] presented a PSO based algorithm for TSP. Geng et al. [13] proposed an effective local search algorithm based on Simulated Annealing (SA) and greedy search technique to solve the TSP. Jolai & Ghanbari [20] presented an improved Artificial Neural Network (ANN) approach for TSP. Dorigo et al. [9] proposed an Ant System to solve TSP. Dorigo & Gambardella [8] described an artificial ant colony (ACO) capable of solving the TSP. Karaboga & Gorkemli [21] proposed a new Artificial Bee Colony (ABC) algorithm called Combinatorial ABC for TSP. Bontoux & Feillet [3] proposed a hybrid algorithm to solve TSP. Beam-ACO algorithm [24] which is a hybrid method combining ACO with beam search was used to solve TSP. Gunduz et al. [16] presented a new heuristic method based on swarm intelligence algorithms for solving TSP. Tsai et al. [42] presented a meta-heuristic approach called ACOMAC algorithm for solving TSP.

From the above discussion it can be concluded that heuristic approaches are more powerful to solve TSP in a feasible time period. Since 1995, PSO has been proven to succeed in continuous optimization problems and much work has been done effectively in this area. But it can be used to solve TSP also. Using the concept of swap operator and swap sequence and redefining some operators of PSO on the basis of them, Wang et al. [41] proposed a special PSO to solve TSP. Akhand et al. [1] improved this algorithm to find solution of TSP and named it velocity tentative PSO. On the other hand ACO is a well established

technique to solve TSP [8,24,29]. Both the algorithms PSO and ACO sometimes converge to local optimal path(tour). K-Opt is a technique which can be apply on a tour(path) to overcome this convergence. In fact local search with K-exchange neighborhoods, K-Opt, is the most widely used heuristic method for the TSP. It works like as a mutation function. K-Opt is a tour improvement algorithm, where in each step K links of the current tour are replaced by K links in such a way that a shorter tour is achieved [17].

In this paper, combining the features of swap sequence and swap operation based PSO [44], ACO and K-Opt operation a hybrid algorithm is proposed to solved STSP as well as ATSP. In proposed method ACO is used a small number of iterations to generate initial solution set(swarm) of PSO. PSO operations are made on this swarm to find optimal path of a TSP. During PSO iterations if a particle does not change its position for a predefined number of iterations then K-Opt operation is made on it a finite number of times to improve its position. Here actually 3-Opt operation is used for this purpose and it is found that it acts better than 2-Opt operation for large size TSPs. The proposed algorithm is tested with bench mark test problems from TSPLIB and it is observed that algorithm is more efficient with respect to accuracy as well as execution time to solve standard TSPs (STSP as well as ATSP) compared to existing algorithms.

The rest of the paper is organized as follows: in Sect. 2, mathematical formulation of the problem is presented. In Sect. 3, some features of swap sequence based PSO (SSPSO) is discussed. Features of ACO are discussed in Sect. 4. K-Opt (Local Search) algorithm is presented in Sect. 5. Proposed algorithm is presented in Sect. 6. Experimental results are discussed in Sect. 7. A brief conclusion is drawn on Sect. 8.

2 Model Formulation

A TSP can be represented by a graph $G = (V, E)$, where $V = 1, 2, ...N$ is the set of vertices or nodes and E is the set of edges. Here each node represents a city and each edge represents path between two cities. Each edge associated with a distance which represents the distance between the cities associated with it. A salesman travels distances to visiting N number of cities (or nodes) cyclically. In one tour he visits each city exactly once, and finishes up where he started with a minimum travel distance. Let d_{jk} be the distance between j-th city and k-th city. Then the model is mathematically formulated as [6], Determine x_{jk}, $j = 1, 2, ...N$, $k = 1, 2, ...N$, to

$$\left.\begin{aligned} \text{Minimize} \quad & Z = \sum_{j=1}^{N} \sum_{k=1}^{N} x_{jk} d_{jk} \\ \text{subject to} \quad & \sum_{j=1}^{N} x_{jk} = 1, \ \text{for } k = 1, 2, ..., N \\ & \sum_{k=1}^{N} x_{jk} = 1, \ \text{for } j = 1, 2, ..., N \end{aligned}\right\} \tag{1}$$

where $x_{jk} = 1$ if the salesman travels from city-j to city-k, otherwise $x_{jk} = 0$.

Let $(x_1, x_2..., x_N, x_1)$ be a complete tour of a salesman, where $x_j \in \{1, 2, ..., N\}$ for $j = 1, 2, ..., N$ and all x_j's are distinct, i.e., $(x_1, x_2..., x_N, x_1)$ is the sequence of cities in which the salesman travels the cities. Then the above model reduces to [23],

$$\left.\begin{array}{c}\text{Determine a complete tour } (x_1, x_2, ...x_N, x_1) \\ \text{to minimize } \quad Z = \sum_{j=1}^{N-1} d_{x_j x_{j+1}} + d_{x_N x_1}\end{array}\right\} \tag{2}$$

3 Swap Sequence Based Particle Swarm Optimization (SSPSO) for TSP

PSOs are exhaustive search algorithms based on the emergent motion of a flock of birds searching for food [10,22] and has been extensively used/modified to solve complex decision making problems in different field of science and technology ([2,11,14,15]). A PSO normally starts with a set of potential solutions (called swarm) of the decision making problem under consideration. Individual solutions are called particles and food is analogous to optimal solution. In simple terms the particles are flown through a multi-dimensional search space, where the position of each particle is adjusted according to its own experience and that of its neighbors. Each particle i has a position vector $(X_i(t))$, a velocity vector $(V_i(t))$, the position at which the best fitness $(X_{pbesti}(t))$ encountered by the particle so far, and the best position of all particles $(X_{gbest}(t))$ in current generation t. In generation $(t + 1)$, the position and velocity of the particle are changed to $X_i(t + 1)$ and $V_i(t + 1)$ using following rules:

$$V_i(t+1) = wV_i(t) + c_1 r_1(X_{pbesti}(t) - X_i(t)) + c_2 r_2(X_{gbest}(t) - X_i(t)) \tag{3}$$

$$X_i(t+1) = X_i(t) + V_i(t+1) \tag{4}$$

The parameters c_1 and c_2 are set to constant values, which are normally taken as 2, r_1 and r_2 are two random values, uniformly distributed in [0, 1], $w(0 < w < 1)$ is inertia weight which controls the influence of previous velocity on the new velocity. It is mainly used to solve continuous optimization problems. It is also used to solve TSPs where swap sequence and swap operations are used to find velocity of a particle and its updating ([26,43,44]). A PSO that uses swap sequence and swap operation is called SSPSO. As discussed in Sect. 2, in a TSP a potential solution is represented by a sequence of nodes. In SSPSO, swap operations on different nodes are used to update a solution. A swap sequence represents a sequence of swap operations used to transform a solution to another solution. Basic operations of SSPSO are briefly presented below:

Swap Operator
Consider a normal solution sequence of TSP with n nodes, $X = (x_1, x_2, ...x_n, x_1)$, where $x_i \in \{1, 2, ...n\}$ and each x_i are distinct. Here swap operator, $SO(i, j)$ is defined as exchange of node x_i and node x_j in solution sequence

X. Then we define $X' = X + SO(i,j)$ as a new sequence on operating operator $SO(i,j)$ on X. So the plus sign '+', above has its new meaning. It can be given a concrete example: suppose there is a TSP problem with six nodes, and $X = (x_1, x_2, x_3, x_4, x_5, x_6) = (1,3,5,2,4,6)$ be a sequence. The swap operator is $SO(2,4)$, then $X' = X + SO(2,4) = (1,3,5,2,4,6) + SO(2,4) = (1,2,5,3,4,6)$, i.e., nodes of position 2 and position 4 are exchanged.

Swap Sequence
A swap sequence SS is made up of one or more swap operators. Let $SS = (SO_1, SO_2, ..., SO_n)$, where $SO_1, SO_2, ..., SO_n$ are swap operators. swap sequence acting on a solution means all the swap operators of the swap sequence act on the solution in order. This can be described by the following formula:

$$X' = X + SS = X + (SO_1, SO_2, ..., SO_n) = (((X + SO_1) + SO_2)... + SO_n)$$

Different swap sequences acting on the same solution may produce the same new solution. All these swap sequences are named the equivalent set of swap sequences. In the equivalent set, the sequence which has the least number of swap operators is called Basic Swap Sequence of the set or Basic Swap Sequence (BSS) in short.

Several swap sequences can be merged into a new swap sequence. Here the operator $\oplus$ is defined as merging two swap sequences into a new swap sequence. Suppose there is two swap sequences, $SS1$ and $SS2$ act on one solution X in order, namely $SS1$ first, $SS2$ second and a new solution X' is obtained. Let there is another swap sequence SS' acting on the same solution X and get the solution X', then SS' is called merging of $SS1$ and $SS2$ and it is represented as:

$$SS' = SS1 \oplus SS2$$

Here, SS' and $SS1 \oplus SS2$ are in the same equivalent set.

The Construction of Basic Swap Sequence
Suppose there is two solutions, A and B, and our task is to construct a Basic Swap Sequence SS which can act on B to get solution A. We define $SS = A - B$ (Here the sign $-$ also has its new meaning). We can swap the nodes in B according to A from left to right to get SS. So there must be an equation $A = B + SS$. For example, consider two solutions:

$$A = (1,2,3,4,5), B = (2,3,1,5,4)$$

Here $A(1) = B(3) = 1$. So the first swap operator is $SO(1,3)$. Let $B1 = B + SO(1,3)$ then we get the following result:

$$B1 : (1,3,2,5,4)$$

Again $A(2) = B1(3) = 2$, so the second operator is $SO(2,3)$ and $B2 = B1 + SO(2,3)$. The third operator is $SO(4,5)$, and $B3 = B2 + SO(4,5)$. Finally we get the Basic swap sequence $SS = A - B = (SO(1,3), SO(2,3), SO(4,5))$.

The Transformation of the Particle Updating Formulas
For solving TSP formulas (3) and (4) of PSO have to transformed using swap sequences and swap operations as follows:

$$V_i(t+1) = V_i(t) \oplus r_1 \odot (X_{pbesti}(t) - X_i(t)) \oplus r_2 \odot (X_{gbest}(t) - X_i(t)) \quad (5)$$
$$X_i(t+1) = X_i(t) + V_i(t+1) \quad (6)$$

Here r_1, r_2 are random numbers between 0 and 1. Velocity $V_i(t)$ represent a swap sequence. $r_1 \odot (X_{pbesti}(t) - X_i(t))$ means all swap operators in BSS $(X_{pbesti}(t) - X_i(t))$ should be maintained with the probability of r_1, i.e., each swap operator in BSS $(X_{pbesti}(t) - X_i(t))$ should be selected with probability r_1. The same meaning is for the expression $r_2 \odot (X_{gbest}(t) - X_i(t))$. From here it is seen that the bigger the value of r_1 the greater the influence of $X_{pbesti}(t)$ is, for more swap operators in $(X_{pbest}(t) - X_i(t))$ will be maintained, it is also the same as $r_2 \odot (X_{gbest}(t) - X_i(t))$.

4 Ant Colony Optimization (ACO)

The ACO algorithm was developed by Dorigo et al. [7] as inspired by actual ant colony behaviors to solve TSP. Ant algorithm are multi-agent system in which the behavior for each single agent, called artificial ant or ant, is inspired by the behavior of real ants. As discussed earlier a TSP consists of a set of N vertices (node/cites) where distance between two vertices is known. The goal of the problem is to find a shortest possible tour(path) from starting node s to destination node D. In ACO, a special variable τ_{ij}, called artificial pheromone trail, which associated with any two vertices i and j is defined. The ant used this pheromone in a stochastic way to decide which node to move to next. At the beginning of the search process a constant amount of pheromone are assigned to all the edges. When ants visit each node for creating a possible tour(path), the pheromone would be updated by ants. Maximum pheromone is available on the path through which maximum ants travel. An ant k is currently located at node i, selects the next node j, based on the following transition probability:

$$P_{ij}^k(t) = \begin{cases} \frac{\tau_{ij}^{\alpha}(t)\eta_{ij}^{\beta}(t)}{\sum_{u \in N_i^k} \tau_{iu}^{\alpha}(t)\eta_{iu}^{\beta}(t)} & \text{if } j \in N_i^k(t) \\ 0 & \text{if } j \notin N_i^k(t). \end{cases} \quad (7)$$

where τ_{ij} represents the pheromone value and η_{ij} represents the heuristic value of the move from node i to j at time step t. $N_i^k(t)$ represent the set of nodes which are not yet visited by ant k (when it is at node i). α and β are positive real parameters whose values determine the relative importance of pheromone versus heuristic information. η_{ij} is calculated by following equation,

$$\eta_{ij} = \frac{1}{d_{ij}} \quad (8)$$

where d_{ij} is the distance (cost) between the node i and j. During visit of nodes by the ants small amount of pheromone would be evaporated from each edge and some pheromone are deposited on the edges through which the ants move. For each edge(i, j), evaporation takes place using the following rule:

$$\tau_{ij}(t) \longleftarrow (1-\rho)\tau_{ij}(t) \tag{9}$$

with $\rho \in [0, 1]$. ρ is the constant, that specifies the rate at which pheromone evaporate. The more evaporate pheromone, the more random the search, that is $\rho = 1$, the search is completely random. After completion of a tour(path) from s to D by each ant, the pheromone on each edge(i, j) is updated (due to deposit of pheromone) as

$$\tau_{ij}(t+1) = \tau_{ij}(t) + \sum_{k=1}^{n_k} \triangle\tau_{ij}^{k}(t) \tag{10}$$

where $\triangle\tau_{ij}^{k}(t)$ is the amount of pheromone deposited by ant k on edge(i, j) and k at time step t and here $\triangle\tau_{ij}^{k}(t)$ is taken as

$$\triangle\tau_{ij}^{k}(t) = \begin{cases} \frac{1}{f(X_k)} & \text{if k-th ant passes through edge (i,j)} \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

5 K-Opt Operation

K-Opt is a local search algorithm based on exchange of K parts (sub-tours) and their reverses (reverse sub-tours) of a tour(path) of a TSP to find a better tour. It has been proven to be very successful for TSPs and similar problems. While breaking (removes) K edges in a tour, there are $(K-1)!2^{K-1}$ ways to reconnect it (including the initial tour) to form a valid tour [40]. Each new combination gives a new tour. Among these tours one may be better than the original tour and can be taken as an improvement. In the case of 2-Opt algorithm removes two edge form the tour, and reconnects the all combination of sub-tours and their reverses (Fig. 1). Continue this process until no 2-Opt improvements can be found. Similarly in the case of 3-Opt, breaking 3 edges in a tour there are total 8 cases of reconnection (Fig. 2). If a tour is 3-optimal it is also 2-optimal [40]. Continue break (remove) edges form tour i.e. $K = 1, 2, 3..., n$ and get new algorithm, like 2-Opt, 3-Opt, 4-Opt and so on. But increase of K increases time complexity. Due to this, here 3-Opt operation is used and it is found that it acts better than 2-Opt operation for large size TSPs.

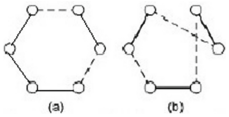


Fig. 1. All combinations of sub_toures for k = 2

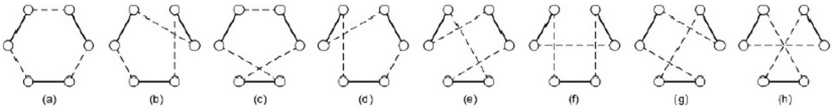


Fig. 2. All combinations of sub_toures for k = 3

6 Proposed Algorithm

A Hybrid Algorithm Based on ACO, PSO and K-Opt Algorithm for Solving TSP

It is assumed that problem involves n nodes, d_{ij} represent distance between node i and node j. In the algorithm a one dimensional array of size n, $X_k(t)$ is used to represent k-th solution in iteration t, i.e., path of k-th ant, which is again k-th particle of the swarm. N is node set and t is iteration counter. n_k is swarm size. $Maxit1$, $Maxit2$, $Maxit3$ represent number of iterations of ACO part, PSO part and K-Opt part of the algorithm respectively. $f(X_k(t))$ represent total length of the path $X_k(t)$. Other notations in the algorithm are same as previously discussed.

// ACO Operations

1. **Start Algorithm**
2. Set values of $Maxit1$, $Maxit2$, $Maxit3$, α, β, ρ. Set $t = 0$.
3. Set $\tau_{ij}(t)$ = $\eta_{ij}(t)$=$\frac{1}{(d_{ij})^r}$, for $i = 1, 2, ..., n$; $j = 1, 2, ..., n$, where r is positive real number.
4. **repeat**

   for $k = 1$ to n_k, **do**

   i=a random node from the node set $N = \{1, 2, ..., n\}$.

   $l = 1$

   $X_k(t)[l] = i$ //*Construct a path* $X_k(t)$.

   $N_i^k = N - \{i\}$

   repeat

   Select next node j from N_i^k based on the transition probability defined in Eq. (7).

   $l = l + 1$

   $X_k(t)[l] = j$

   $i = j$

   $N_i^k = N_i^k - \{i\}$

   until $N_i^k = \emptyset$ //$\emptyset$ is null set

Calculate the path length $f(X_k(t))$
end for
for $i = 1$ to n step 1 **do**
for $j = 1$ to n step 1 **do**
//*Pheromone evaporation.*
Reduce the pheromone, $\tau_{ij}(t)$ using Eq. (9)
end for
end for
for $i = 1$ to n **do**
for $j = 1$ to n **do**
$\tau_{ij}(t+1) = \tau_{ij}(t)$
Update $\tau_{ij}(t+1)$ using Eq. (10).
end for
end for
$t = t + 1$
Until $(t > Maxit1)$

//PSO Operations

5. **for** $k = 1$ to n **do**
$X_k(0) = X_k(t-1)$
$X_{pbestk}(0) = X_k(0)$
$V_k(0)$=SO(i, j) where i, j are randomly generated from the set N and
$i \neq j$
end for
6. t=1
7. X_{gbest}= solution having minimum path length from the solution set $\{X_1(0), X_2(0), ...X_{n_k}(0)\}$
8. **repeat**
for $k = 1$ to n_k, *do*
Determine $V_k(t)$ using Eq. (5)
Determine $X_k(t)$ using Eq. (6)
If $f(X_{pbestk}(t-1)) > f(X_k(t))$
$X_{pbestk}(t) = X_k(t)$
else
$X_{pbestk}(t) = X_{pbestk}(t-1)$
end if
If $f(X_{gbest}) > f(X_k(t))$
$X_{gbest} = X_k(t)$
end if
If $(X_{gbest} = X_k(t))$ holds for a predefined consecutive number of iterations then apply
K-Opt operation on $X_k(t)$ to improve its position (Sect. 6.1)
end for

9. $t = t + 1$.
   Until $(t > Maxit2)$

10. **Output** X_{gbest}
11. End of Algorithm.

6.1 K-Opt Operation on a Complete Tour $X_k(t)$

Detailed algorithm of k-opt operation for $K = 3$ is presented below. In the algorithm a one dimensional array of size n, $X_{tem_k}(t)$, is used to represented temporary k-th solution in iteration t, i.e., k-th particle of the swarm. $X_{ki}(t)$ and $X^r_{ki}(t)$, $i = 1, 2, 3$ are one dimensional arrays used to represent sub-tour and revers_sub-tour of the original tour $X_k(t)$.

for i=1 to $Maxit3$ **do**

Remove 3 edges(randomly selected) from tour $X_k(t)$, it makes 3 sub-tours $X_{ki}(t)$, $i = 1, 2, 3$.
Reverses of the contains of these sub-tours are called revers_sub-tours, represented by $X^r_{ki}(t)$, $i = 1, 2, 3$, i.e., $X^r_{k1}(t)$= revers_sub-tour$(X_{k1}(t))$, $X^r_{k2}(t)$= revers_sub-tour$(X_{k2}(t))$, $X^r_{k3}(t)$= revers_sub-tour$(X_{k3}(t))$.
Now combing the sub-tours $\{X_{k1}(t), X_{k2}(t), X_{k3}(t)\}$, $\{X^r_{k1}(t), X^r_{k2}(t), X^r_{k3}(t)\}$ new tours can be formed in following 8 combinations:

i $\{X_{k1}(t), X_{k2}(t), X_{k3}(t)\}$
ii $\{X_{k1}(t), X^r_{k2}(t), X_{k3}(t)\}$
iii $\{X_{k1}(t), X_{k2}(t), X^r_{k3}(t)\}$
iv $\{X_{k1}(t), X^r_{k3}(t), X^r_{k2}(t)\}$
v $\{X_{k1}(t), X_{k3}(t), X^r_{k2}(t)\}$
vi $\{X_{k1}(t), X^r_{k3}(t), X_{k2}(t)\}$
vii $\{X_{k1}(t), X^r_{k2}(t), X^r_{k3}(t)\}$
viii $\{X_{k1}(t), X^r_{k3}(t), X_{k2}(t)\}$

for each combination do
  Create a complete tour from the combination and let it be $X_{tem_k}(t)$
  if $f(X_{tem_k}(t) < f(X_k(t))$
    $X_k(t) = X_{tem_k}(t)$
  end if
end for

end for

7 Experimental Results

All computational experiments are conducted with Dev C++ 5.8.3, core $i3$ CPU @ 2.10 GH$_z$, Windows 8.1 Operating System and 4 GB RAM. Performance of the

proposed algorithm is tested using different size standard TSPs from TSPLIB. From each problem algorithm is tested by running the program 20 times for different seeds of random number generator and the best solution obtained, the average value of the solutions, Standard deviation (SD) value and percentage of relative error (Error(%)) according to optimal solution are calculated. The percentage of relative **Error(%)** is calculated using the following equation.

$$\textbf{Error}(\%) = \frac{average\ solution - optimal\ solution}{optimal\ solution} \times 100 \tag{12}$$

The results obtained by proposed algorithm for seventeen different test problems from TSPLIB are presented in Table 1.

In Table 1 results of STSPs and ATSPs are displayed separately. In the Table **Best** column represent the best solution obtained by proposed method and optimal solution are taken from TSPLIB. The problems whose optimal solutions (according to TSPLIB) are obtained by proposed approach are presented in bold face. It is found from the Table 1 that the algorithm produces optimal solution for most of the problems taken for the test and for others it gives solutions very near to optimal solutions. For problems like, $rat99$, $eli101$, $kroA200$, $fiv56$, the algorithm does not provide optimal solution but other parameters like **average**,

Table 1. The Result obtained by the proposed algorithm for STSP and ATSP

	Problem	Optimal	**Best**	Worst	Average	SD	Error(%)	Time(S)
STSP	gr17	2085	**2085**	2085	2085.00	0.00	00.00	1.56
	bays29	2020	**2020**	2024	2020.05	0.89	00.01	10.05
	swiss42	1273	**1273**	1273	21273.00	0.00	00.00	7.46
	eli51	426	**426**	427	426.29	0.46	0.07	19.91
	berlin52	7542	**7542**	7555	7543.29	3.90	0.01	20.28
	st70	675	**675**	681	676.00	1.73	0.14	100
	eli76	538	**538**	541	538.15	0.65	0.02	150
	rat99	1211	1212	1216	1213.90	0.99	00.07	200
	kroA100	21282	**21282**	21406	21319.00	47.79	00.17	305.01
	kroC100	20749	**20749**	20992	20862.25	45.15	00.18	350.01
	eli101	629	630	637	631.20	1.50	0.34	200.90
	lin105	14379	**14379**	14385	14379.29	1.30	00.00	320.10
	pr124	59030	**59030**	59320	59118.64	98.30	00.15	305.00
	pr152	73682	**73682**	73705	73691.64	28.26	0.12	1031.32
	kroA200	29368	29402	30016	29640.00	145.0	0.46	350.12
ATSP	br17	39	**39**	bf39	39.00	0.00	00.00	1.13
	ftv33	1286	**1286**	1286	1286.00	0.00	00.00	5.56
	ry48	14422	**14422**	14642	14452.79	64.79	0.21	15.12
	ftv56	1608	1629	1689	1642.19	18.87	0.810	25.12

Table 2. Compare results using 2-Opt and 3-Opt in test problems

	Problem	Optimal	ACO + PSO + 2-Opt	**ACO + PSO + 3-Opt**
STSP	gr17	2085	**2085**	**2085**
	bays29	2020	2028	**2020**
	swiss42	1273	1284	**1273**
	eli51	426	447	**426**
	berlin52	7542	7800	**7542**
	st70	675	699	**675**
	eli76	538	550	**538**
	rat99	1211	1270	1212
	kroA100	21282	21910	**21282**
	kroC100	20749	20892	**20749**
	eli101	629	795	630
	lin105	14379	15500	**14379**
	pr124	59030	62040	**59030**
	pr152	73682	73910	**73682**
	kroA200	29368	30290	29402
ATSP	br17	39	**39**	**39**
	ftv33	1286	1340	**1286**
	ry48	14422	14020	**14422**
	ftv56	1608	1648	1629

Sd, **Error(%)** and **Time(s)** are better compare to [30]. Small values of **SD** and **Error(%)** of the solutions of the problems ensure that obtained solutions of the problems are very close to optimal solutions.

Table 2 represent results obtained by proposed method due to different test problems using 2-Opt and 3-Opt operations in the algorithm. In the case of small size problems like $gr17$ and $br17$ both the approaches provide same solution as optimal solution. But for large size problems 2-Opt and 3-Opt produces different solutions. Problems for which optimal solutions are obtained by the algorithms are presented in bold face in the Table 2. In some problems like $rat99$, $eli101$, $kroA200$, $fiv56$, using 3-Opt, the algorithm does not provide optimal solution but it produces better solutions than that obtained by the algorithm using 2-Opt. It is also clear from Table 2 that for all the problems algorithm with 3-Opt provide better result than that using 2-Opt. So in proposed algorithm 3-Opt operation is used.

Table 3 represents the effect of swarm size in the algorithm for different test problems. Swarm sizes like 10, 20, 30 and number of city(node) of the test problems are used for the test. In the case of small size problem like $gr17$ $bays29$, $swiss42$ for STSPs and $br17$ for ASTSPs the algorithm gives same solution as optimal solution for different swarm size. For problems like $eli51$, $berlin52$, $st70$,

Table 3. The Result obtained by the proposed method for various number of ants

	Problem	Optimal	Swarm size = 10	Swarm size = 20	Swarm size = 30	Swarm size = Problem size
STSP	gr17	2085	**2085**			**2085**
	bays29	2020	**2020**	**2020**		**2020**
	swiss42	1273	**1273**	**1273**	**1273**	**1273**
	eli51	426	**426**	428	429	427
	berlin52	7542	**7542**	7590	7610	7610
	st70	675	**675**	680	689	702
	eli76	538	**538**	545	552	570
	rat99	1211	1212	1222	1230	1249
	kroA100	21282	**21282**	21492	21572	21825
	kroC100	20749	**20749**	20790	20785	20892
	eli101	629	630	680	720	790
	lin105	14379	**14379**	14420	14510	14705
	pr124	59030	**59030**	60120	60350	60480
	pr152	73682	**73682**	73710	73699	73750
	kroA200	29368	29402	29803	29901	30230
ATSP	br17	39	**39**			**39**
	ftv33	1286	**1286**	1294	1315	1301
	ry48	14422	**14424**	14460	14510	14510
	ftv56	1608	1629	1642	1672	1690

*eli*76, *rat*99, *kroA*100, *lin*105, *pr*124, *kroA*200, *ftv*33, *ry*48, *ftv*56 the algorithm provide better solution for swarm size 10. So in the proposed algorithm swarm size is taken as 10.

Table 4 represents a comparison of all computational results of the proposed algorithm with other existing algorithms in the literature. From Table 4, it is clear that proposed approach is better compared to other existing approaches in the literature both with respect to accuracy and computational time. For the test problems like *eli*51, *st*70, *eli*76, *rat*99, *eli*101, *kroA*200 the algorithm produces better values of *Avg*, *SD*, *Error* compared to other algorithm. For the test problem *kroA*100, proposed method provide optimal solution but other parameters like *Avg*, *SD*, *Error* are not better compared to WFA with 3-Opt (Othman et al., 2013). In some test problems the proposed method does not provide optimal solution but the solution are very near to optimal solution, due to minimum standard deviation (SD) compared to other algorithms in the literature.

Table 4. Comparison of results obtained by proposed approach with other method in literature

Method	Problem	eli5	berlin52	st70	eil76	rat99	kroA100	eil101	lin105	kroA200
	Optimal	**426**	**7542**	**675**	**538**	**1211**	**21282**	**629**	**14379**	**29368**
	Best	**426**	**7542**	**675**	**538**	1212	**21282**	630	**14379**	29402
RABNET-TSP (2006) [37]	Avg	438.70	8073.97	-	556.10	-	21868.47	654.83	14702.17	30257.53
	SD	3.52	270.14	-	8.03	-	245.76	6.57	328.37	342.98
	Error(%)	2.98	7.05	-	3.36	-	2.76	4.11	2.25	3.03
Modified RABNET -TSP (2009) [31]	Avg	437.47	7932.50	-	556.33	-	21522.73	648.64	14400.7	30190.27
	SD	4.20	277.25	-	5.30	-	93.34	3.85	44.03	273.38
	Error(%)	2.69	5.18	-	3.41	-	1.13	3.12	0.15	2.80
SA ACO PSO (2012) [5]	Avg	427.27	7542.00	-	540.20	-	21370.30	635.23	14406.37	29738
	SD	0.45	0.00	-	2.94	-	123.36	3.59	37.28	356.07
	Error(%)	0.30	0.00	-	0.41	-	0.41	0.99	0.19	1.26
WFA with 2-opt (2013) [35]	Avg	426.65	7542.00	-	541.22	-	21282.00	639.87	143790.00	29654.03
	SD	0.66	0.00	-	0.66	-	0.00	2.88	0.00	151.42
	Error(%)	0.15	0.00	-	0.60	-	0.00	1.73	0.00	0.97
WFA with 3-opt (2013) [35]	Avg	426.60	7542	-	539.44	-	**21282.80**	633.50	14459.40	29646.50
	SD	0.52	0.00	-	1.51	-	**0.00**	3.47	1.38	110.91
	Error(%)	0.14	0.00	-	0.27	-	**0.00**	0.72	0.56	0.95
HACO (2012) [19]	Avg	431.20	7560.54	-	-	1241.33	-	-	-	-
	SD	2.00	67.48	-	-	9.60	-	-	-	-
	Error(%)	1.22	0.23	-	-	1.42	-	-	-	-
PSO-ACO-3Opt (2015) [30]	Avg	426.45	**7543.20**	678.20	538.30	1227.40	21445.10	623.70	**14379.15**	29646.05
	SD	0.61	**2.37**	1.47	0.47	1.98	78.24	2.12	**0.48**	114.71
	Error(%)	0.11	0.02	0.47	0.06	0.28	0.77	0.59	**0.00**	0.95
Proposed Method	**Avg.**	**426.29**	7543.29	**676.00**	**538.15**	**1213.90**	21319.50	**31.20**	14379.29	**29642.00**
	SD	**0.46**	3.90	**1.73**	**0.65**	0.99	47.79	**1.50**	1.30	**165**
	Error(%)	**0.07**	**0.01**	**0.14**	**0.00**	**0.07**	0.17	**0.34**	**0.00**	**0.46**

8 Conclusion

Here for the first time combining the features of swap sequence and swap operation based PSO, ACO and K-Opt operation a new hybrid algorithm is presented to solve STSP as well as ATSP. Here ACO is used a small number of iterations to generate initial swarm of PSO. Then PSO operations are made on this swarm a sufficient number of times to find optimal path. During PSO iterations if a particle does not change its position for a predefined number of iterations then K-Opt operation (for $K = 3$) is made on it a finite number of times to improve its position. The performance of the proposed algorithm is tested using different size

standard TSPs from TSPLIB. In most of the TSPs considered for test the proposed algorithm provide optimal solution. In some test problems the proposed algorithm does not provide optimal solutions but the solutions are very close to optimal solutions. The performance of proposed method is better if and only if small numbers of ants (10 in proposed method) used in ACO. All experimental results imply that proposed approach is better compared to other existing approaches in the literature both with respect to accuracy and computational time. The algorithm can be used to solve TSPs in fuzzy environment, rough environment, rough-fuzzy environment and etc. Proposed algorithm can be used to solved solid TSP and vehicle routing problem and router (networking) related problem with minor modification.

References

1. Akhand, M.A.H., Akter, S., Rashid, M.A.: Velocity tentative particle swarm optimization to solve TSP. In: 2013 International Conference on Electrical Information and Communication Technology (EICT), pp. 1–6. IEEE Conference Publications (2014)
2. Angeline, P.J.: Evolutionary optimization versus particle swarm optimization: philosophy and performance differences. In: Porto, V.W., Saravanan, N., Waagen, D., Eiben, A.E. (eds.) EP 1998. LNCS, vol. 1447, pp. 601–610. Springer, Heidelberg (1998). doi:10.1007/BFb0040811
3. Bontoux, B., Feillet, D.: Ant colony optimization for the traveling purchaser problem. Comput. Oper. Res. **35**(2), 628–637 (2008)
4. Changdar, C., Mahapatra, G.S., Pal, R.: An efficient genetic algorithm for multi-objective solid traveling salesman problem under fuzziness. Swarm Evol. Comput. **15**, 27–37 (2014)
5. Chen, S.M., Chien, C.Y.: Solving the traveling salesman problem based on the genetic simulated annealing ant colony system with particle swarm optimization techniques. Expert Syst. Appl. **38**, 14439–14450 (2011)
6. Dantzig, G.B., Fulkerson, D.R., Johnson, S.M.: Solution of large scale traveling salesman problem. Oper. Res. **2**, 393–410 (1954)
7. Dorigo, M., Di Caro, G.: The ant colony optimization meta-heuristic. In: Corne, D., Dorigo, M., Glover, F. (eds.) New Ideas in Optimization, pp. 11–32. McGraw-Hill, London (1999)
8. Dorigo, M., Gambardella, L.M.: Ant colonies for the traveling salesman problem. Biosystems **43**, 73–81 (1997)
9. Dorigo, M., Maniezzo, V., Colorni, A.: Ant system: optimization by a colony of cooperating agents. IEEE Trans. Syst. Man Cybern. Part-B Cybern. **26**(1), 29–41 (1996)
10. Eberhart, R., Kennedy, J.: A new optimizer using particles swarm theory. In: Proceedings of the Sixth International Symposium on Micro Machine, Human Science, Nagoya, Japan, pp. 39–43. IEEE Service Center, Piscataway (1995)
11. Fan, H.: Discrete particle swarm optimization for TSP based on neighborhood. J. Comput. Inf. Syst. (JCIS) **6**, 3407–3414 (2010)
12. Focacci, F., Lodi, A., Milano, M.: A hybrid exact algorithm for the TSPTW. INFORMS J. Comput. **14**, 403–417 (2002)

13. Geng, X.T., Chen, Z.H., Yang, W., Shi, D.Q., Zhao, K.: Solving the traveling salesman problem based on an adaptive simulated annealing algorithm with greedy search. Appl. Soft Comput. **11**(4), 3680–3689 (2011)
14. Guchhait, P., Maiti, M.K., Maitia, M.: Two storage inventory model of a deteriorating item with variable demand under partial credit period. Appl. Soft Comput. **13**, 428–448 (2013)
15. Guchhait, P., Maiti, M.K., Maitia, M.: Inventory model of a deteriorating item with price and credit linked fuzzy demand: a fuzzy differential equation approach. Oper. Res. Soc. India **51**(3), 321–353 (2013)
16. Gunduz, M., Kiran, M.S., Ozceylan, E.: A hierarchic approach based on swarm intelligence to solve traveling salesman problem. Turk. J. Electr. Eng. Comput. Sci. **23**, 103–117 (2015)
17. Helsgaun, K.: General k-opt submoves for the Lin-Kernighan TSP heuristic. Math. Program. Comput. **1**, 119–163 (2009)
18. Ibaraki, T., Imahori, S., Kubo, M., Masuda, T., Uno, T., Yagiura, M.: Effective local search algorithm for routing and scheduling problems with general time window constraints. Transp. Sci. **39**(2), 206–232 (2005)
19. Junqiang, W., Aijia, O.: A hybrid algorithm of ACO and delete-cross method for TSP. In: 2012 International Conference on Industrial Control and Electronics Engineering (ICICEE), pp. 1694–1696. IEEE (2012)
20. Jolai, F., Ghanbari, A.: Integrating data transformation techniques with Hopfield neural networks for solving traveling salesman problem. Expert Syst. Appl. **37**, 5331–5335 (2010)
21. Karaboga, D., Gorkemli, B.: A combinatorial artificial bee colony algorithm for traveling salesman problem. In: 2011 International Symposium on Innovations in Intelligent Systems and Applications, Istanbul, Turkey
22. Kennedy, J., Eberhart, R.: Particle swarm optimization. IEEE Int. Conf. Neural Netw. **4**, 1942–1948 (1995)
23. Khanra, A., Maiti, M.K., Maiti, M.: Profit maximization of TSP through a hybrid algorithm. Comput. Ind. Eng. **88**, 229–236 (2015)
24. Lopez-Ibanez, M., Blum, C.: Beam-ACO for the traveling salesman problem with time windows. Comput. Oper. Res. **37**(9), 1570–1583 (2010)
25. Lawler, E.L., Lenstra, J.K., Rinnooy Kan, A.H.G., Shmoys, D.B.: The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization. Wiley, New York (1985)
26. Liang, J.J., Qin, A.K., Suganthan, P.N., Baskar, S.: Comprehensive learning particle swarm optimizer for global optimization of multimodal functions. J. Evol. Comput. **10**(3), 281–295 (2006)
27. Lin, S., Kernighan, B.W.: An effective heuristic algorithm for the traveling salesman problem. Oper. Res. **21**(2), 498–516 (1973)
28. Majumdar, J., Bhunia, A.K.: Genetic algorithm for asymmetric traveling salesman problem with imprecise travel times. J. Comput. Appl. Math. **235**(9), 3063–3078 (2011)
29. Mavrovouniotis, M., Yang, S.: Ant colony optimization with immigrants schemes for the dynamic traveling salesman problem with traffic factors. Appl. Soft Comput. **13**(10), 4023–4037 (2013)
30. Mahi, M., Baykan, O.K., Kodaz, H.: A new hybrid method based on Particle Swarm Optimization, Ant Colony Optimization and 3-Opt algorithms for Traveling Salesman Problem. Appl. Soft Comput. **30**, 484–490 (2015)

31. Masutti, T.A.S., de Castro, L.N.: A self-organizing neural network using ideas from the immune system to solve the traveling salesman problem. Inf. Sci. **179**(10), 1454–1468 (2009)
32. Miliotis, P.: Using cutting planes to solve the symmetric travelling salesman problem. Math. Program. **15**, 177–188 (1978). North-Holland Publishing Company
33. Moon, C., Kim, J., Choi, G., Seo, Y.: An efficient genetic algorithm for the traveling salesman problem with precedence constraints. Eur. J. Oper. Res. **140**, 606–617 (2002)
34. Nguyen, H.D., Yoshihara, I., Yamamori, K., Yasunaga, M.: Implementation of an effective hybrid GA for large scale traveling salesman problem. IEEE Trans. Syst. Man Cybern. Part-B Cybern. **37**(1), 92–99 (2007)
35. Othman, Z.A., Srour, A.I., Hamdan, A.R., Ling, P.Y.: Performance water flow-like algorithm for TSP by improving its local search. Int. J. Adv. Comput. Technol. **5**(14), 126 (2013)
36. Petberg, M.W., Homg, S.: On the symmetric traveling salesman problems: a computational study. Math. Program. Stud. **12**, 87–107 (1980)
37. Pasti, R., De Castro, L.N.: A Neuro-immune network for solving the traveling salesman problem. In: Proceedings of International Joint Conference on Neural Networks, IJCNN 2006, pp. 3760–3766 (2006)
38. Petersen, H.L., Madsen, O.B.G.: The double traveling salesman problem within multiple stack formulation and heuristic solution approaches. Eur. J. Oper. Res. **198**, 339–347 (2009)
39. Padberg, M., Rinaldi, G.: Optimization of a 532-city symmetric traveling salesman problem by branch and cut. Oper. Res. Lett. **6**(1), 1–7 (1987)
40. Sierksma, G.: Hamiltonicity and the 3-OPT procedure for the traveling salesman problem. Appl. Math. **22**(2), 351–358 (2014)
41. Shi, X.H., Liang, Y.C., Lee, H.P., Lu, C., Wang, Q.X.: Particle swarm optimization-based algorithms for TSP and generalized TSP. Inf. Process. Lett. **103**(5), 169–176 (2007)
42. Tsai, C.F., Tsai, C.W., Tseng, C.C.: A new hybrid heuristic approach for solving large traveling salesman problem. Inf. Sci. **166**, 67–81 (2004)
43. Wang, K.P., Huang, L., Zhou, C.G., Pang, W.: Particle swarm optimization for traveling salesman problem. Int. Conf. Mach. Learn. Cybern. **3**, 1583–1585 (2003)
44. Yan, X., Zhang, C., Luo, W., Li, W., Chen, W., Liu, H.: Solve traveling salesman problem using particle swarm optimization algorithm. Int. J. Comput. Sci. Issues **9**(6), 264–271 (2012)

FASER128: Cryptanalysis and Its Countermeasure

M.K. Dubey(✉), Navneet Gaba, and S.S. Bedi

SAG, DRDO, Metcalf House, Delhi 110054, India
kantmanish@yahoo.com, navneetgaba2000@yahoo.com, ssbedi53@hotmail.com

Abstract. Many symmetric key encryption schemes have been designed to ensure the confidentiality of data only. Data integrity plays an important role of security in various encryption scheme. Assuming this fact, many researchers have focused their research to design Authenticated Encryption (AE) schemes that provide both confidentiality and authenticity. FASER is one of them which was submitted in CAESAR competition and withdrawn in later due to an attack reported in the paper [6]. It has two parent ciphers namely FASER128 and FASER256. Cryptanalysis of FASER128 was studied by the authors in [6,7] and mentioned some serious flaws in the design of the crypto algorithm. Due to these flaws, both the parent ciphers of the FASER have been withdrawn. In this paper, we study the cryptanalysis of FASER128 by key recovery attack and discuss some weaknesses. We have also suggested some modifications of cryptoalgorithm to avoid the key recovery attack.

Keywords: Stream cipher · Key recovery attack · Authenticated encryption

1 Introduction

The CAESAR (Competition for Authenticated Encryption: Security, Applicability, and Robustness) competition was started in 2014 and its aim is to find Authenticated encryption schemes that (1) offer advantages over AES-GCM and (2) are suitable for widespread adoption. The notion of Authenticated encryption was first coined by the seminal work by Bellare and Namprempre [3] in 2000, Bellare, Kohno T. and Namprempre [4] in 2002 and then further extended by several authors. Authenticated encryption schemes are key-based cryptographic schemes comprising of both an encryption and an authentication that provides confidentiality and authenticity. Confidentiality assures that adversary cannot gain much information from ciphertext corresponding to plaintext while authenticity ensures that ciphertext has not been altered which was delivered by authentic sender to receiver. Since the security of authenticated ciphers depends on both encryption and authentication, therefore the designer's should have to take more precautions to design encryption as well as authentication schemes because due to this an attacker has more choices to execute the attack either in any one of

© Springer Nature Singapore Pte Ltd. 2017
D. Giri et al. (Eds.): ICMC 2017, CCIS 655, pp. 120–131, 2017.
DOI: 10.1007/978-981-10-4642-1_11

the encryption or the authentication or both of them simultaneously and also has more chances to get the information such as authentication tag and so on. Thus, it is more tedious job to design a good authenticated encryption scheme. FASER [5] is an Authenticated encryption schemes that consists of two parent ciphers: FASER128 and FASER256. The nomenclature represents the maximum secret key length that can be used in each cipher. The two parent ciphers of FASER, FASER128 and FASER256, had been submitted to CAESAR competition but due to key recovery attack proposed in [6], it was later withdrawn from CAESAR competition.

FASER128 and FASER256 both are comprise of two identical state registers, one for encryption E and one for authentication A followed by three components FSR, MIX, MAJ. They have also consist of three processes that is initialization, update, finalization. In this paper, we studied an attack of encryption portion only therefore the details about an authentication portion is not included. The recommended key parameter set for FASER128 includes 16 byte key(secret), 8 byte secret message number, 8 byte public message number, and 8 byte tag. The paper is organised as follows: Sect. 2 deals the structure and function of FASER128. In Sect. 3, we discuss key recovery attack on FASER128 and describe a method to recover full key of the crypto algorithm. Section 4 deals some observations about the weaknesses of FASER128 and find some suitable situations experimentally to avoid this attack and finally, we conclude with conclusion.

2 Description of FASER128

This section deals the details of structure of FASER128.

2.1 Components of FASER 128

State Register

FASER128 has two identical state registers, one is used for encryption and other is used for authentication, denoted by E and A respectively. Both the state registers are identical in size, that is, 256. It is represented as $E = (E_3, E_2, E_1, E_0)$ and $A = (A_3, A_2, A_1, A_0)$ where each E_i or A_i is 64 bits in size.

FSR E

FSR (Feedback Shift Register)is used for updation of state register of FASER128. The FSR is made up of 8 sub-FSRs, where 2 sub-FSR is comprised of one LFSR (Linear Feedback Shift Register) and one NLFSR (Nonlinear Feedback Shift Register). These two sub-FSRs are operate on 64-bits in size and coprime to each other. They are also updated independently in different region of state. The FSR X is defined as follows:

$$
\begin{aligned}
&FSR17(X) : y \leftarrow x_{16} \oplus x_{15} \oplus x_{14} \cdot x_{13} && (x_{16}, \ldots, x_1, x_0) \leftarrow (x_{15}, \ldots, x_0, y)\\
&FSR23(X) : y \leftarrow x_{22} \oplus x_{21} \oplus x_{12} \cdot x_{11} && (x_{22}, \ldots, x_1, x_0) \leftarrow (x_{21}, \ldots, x_0, y)\\
&FSR29(X) : y \leftarrow x_{28} \oplus x_{27} \oplus x_{19} \cdot x_{12} && (x_{28}, \ldots, x_1, x_0) \leftarrow (x_{27}, \ldots, x_0, y)\\
&FSR31(X) : y \leftarrow x_{30} \oplus x_{11} \oplus x_{21} \cdot x_{13} && (x_{30}, \ldots, x_1, x_0) \leftarrow (x_{29}, \ldots, x_0, y)\\
&FSR33(X) : y \leftarrow x_{32} \oplus x_{19} && (x_{32}, \ldots, x_1, x_0) \leftarrow (x_{31}, \ldots, x_0, y)\\
&FSR35(X) : y \leftarrow x_{34} \oplus x_{32} && (x_{34}, \ldots, x_1, x_0) \leftarrow (x_{33}, \ldots, x_0, y)\\
&FSR41(X) : y \leftarrow x_{40} \oplus x_{37} && (x_{40}, \ldots, x_1, x_0) \leftarrow (x_{39}, \ldots, x_0, y)\\
&FSR47(X) : y \leftarrow x_{46} \oplus x_{41} && (x_{46}, \ldots, x_1, x_0) \leftarrow (x_{45}, \ldots, x_0, y)
\end{aligned}
$$

The feedback update of the FSR E can be described as

$$FeedFSR(X) = (FSR_3(X_3), FSR_2(X_2), FSR_1(X_1), FSR_0(X_0)) \tag{1}$$

where

$$
\begin{aligned}
FSR_0(X_0) &= FSR33(H_{33}(X_0)) \parallel FSR31(L_{31}(X_0)),\\
FSR_1(X_1) &= FSR35(H_{35}(X_1)) \parallel FSR29(L_{29}(X_1)),\\
FSR_2(X_2) &= FSR41(H_{41}(X_2)) \parallel FSR23(L_{23}(X_2)),\\
FSR_3(X_3) &= FSR47(H_{47}(X_3)) \parallel FSR17(L_{17}(X_3)),
\end{aligned}
$$

$H_i(X)$ and $L_i(X)$ represent the i-th most significant (High) bits of 64-bit X and i-th least significant (Low) bits of 64-bit X respectively, and $\parallel$ denotes concatenation. In one update of the FSR E, the FSR is clocked 8 times, that is,

$$FSR(X) = (FeedFSR)^8$$

where the FSR is clocked at once then each subFSR is being clocked independently. Consequently, all 64-bits are updated in each state register of the FSR E.

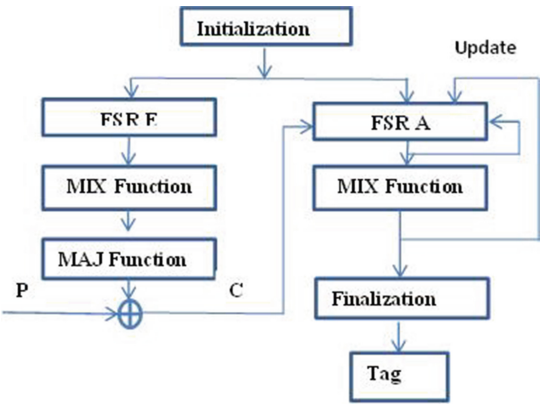


Fig. 1. FASER block diagram

Mix function

The Mix function combines the initial states from the sub-FSRs. The purpose of Mix function is to defuse the information across the state register and provide good diffusion property. The input is the entire state (X_3, X_2, X_1, X_0) and the output gives three 64-bit words such that

$$\begin{aligned} Y_0 &= (X_0 \lll 3) \oplus (X_1 \lll 12) \oplus (X_2 \lll 43) \oplus (X_3 \lll 27) \\ Y_1 &= (X_0 \lll 22) \oplus (X_1 \lll 54) \oplus (X_2 \lll 5) \oplus (X_3 \lll 30) \\ Y_2 &= (X_0 \lll 50) \oplus (X_1 \lll 35) \oplus (X_2 \lll 14) \oplus (X_3 \lll 60) \end{aligned} \tag{2}$$

where "$\lll$" denotes the bitwise rotation to the left (Fig. 1).

MAJ function

The MAJ function operates on 64-bit words which is the bitwise majority function say MAJ. The output of Mix function is used as an input of Majority function. The output of MAJ function is defined as follows:

$$Z = (Y_0 \wedge Y_1) \vee (Y_0 \wedge Y_2) \vee (Y_1 \wedge Y_2), \tag{3}$$

where $\wedge$ and $\vee$ means the bitwise multiplication and bitwise Xor respectively.

2.2 Processes of FASER128

FASER128 executes the following processes: initialization, update and finalization which are elucidated as follows:

Initialization

The purpose of initialization is to initialize the two state registers E and A using the secret key K and the public message number PMN or whole secret key. First, the inputs are directly fed into the register, least significant byte first. The remaining bytes are filled with a constant, $0x5a...5a$ to identify the key for the encryption. The register contents are then diffused so that the inputs (K and PMN) affect the entire state as follows. Here TWEAK is defined as below: TWEAK(X):$(x_{63}, x_{62}, \ldots, x_2, x_1, x_0) \leftarrow (1, x_{62}, \ldots, x_2, 0, 1)$.

Update

In each clock, the encryption of FASER128 is a synchronous stream cipher that produces a ciphertext of 64-bit word at each clock and an authentication function that accumulate the ciphertext. The FSR update function is an identical to the encryption function and the authentication function where they differ only at the initialization process. When FASER128 is clocked at once, each FSR is clocked 8 times.

Algorithm 1. Initialization (E, K, PMN)

Input: 1. $E = 0x5A \parallel \ldots \parallel 0x5A \parallel PMN \parallel K$,
Output: $E = FSR(E)$

2. For $i = 1$ **to** 8
3. $E = FSR(E)$,
4. $(Y_2, Y_1, Y_0) = MIX(E)$
5. $E = (E_3, E_2 \oplus Y_2, E_1 \oplus Y_1, E_0 \oplus Y_0)$
6. $E = (E_2, E_1, E_0, E_3)$
7. end for
8. TWEAK(E)
9. For $i = 1$ **to** 8
10. $E = FSR(E)$,

Here, we only focus on procedure for update of encryption and so we omit procedure for update of authentication. The following describes one update of FASER128 to process one 64-bit plaintext word P_i. FASER128 continues to clock until all the inputs have been processed. The pseudo-code for the procedure update is

Algorithm 2. Update(E, P_i)

$E = FSR(E)$
$(Y_2, Y_1, Y_0) = MIX(E)$
$Z = MAJ(Y_2, Y_1, Y_0)$
$C_i = P_i \oplus Z$.

Finalization

This process generates the tag based on the contents of the authenticated register. The update of the authenticated register is almost similar to the update function. We have also ignored detail of this process due to irrelevance for this attack.

3 Key Recovery Attack

In this section, we give details about the key recovery attack discussed in [7]. This attack can be divided into two phases: phase I deals the recovery the initial state of the register E of the FSR E, and phase II deals the recovery of the full secret key K by reverting the procedure of initialization of Algorithm 1.

3.1 Phase I: Find Linear Relations

This phase is devoted to find some linear relations based on the following experiments. For a 64-bit word variable X, we denote by $X^t[i]$ the i-th bit of the value of X at time $t \geq 0$, where $0 \leq i \leq 63$. The main idea behind the key recovery attack is to find experimentally the certain linear relationship between the output key bits $Z^t[i]$ and $Z^{t+1}[i + 8 \bmod 64]$ for some i to generate some linear equations among the state of the register E, and then recover the state of the FSR E by solving these linear equations. For example, set $i = 54$. We have

$$\begin{aligned} Z^t[54] &= Y_0^t[54]Y_1^t[54] \oplus Y_1^t[54]Y_2^t[54] \oplus Y_2^t[54]Y_0^t[54] \\ Y_0^t[54] &= X_0^t[51] \oplus X_1^t[42] \oplus X_2^t[11] \oplus X_3^t[27] \\ Y_1^t[54] &= X_0^t[32] \oplus X_1^t[0] \oplus X_2^t[49] \oplus X_3^t[24] \\ Y_2^t[54] &= X_0^t[4] \oplus X_1^t[19] \oplus X_2^t[40] \oplus X_3^t[58] \end{aligned} \tag{4}$$

and

$$\begin{aligned} Z^{t+1}[62] &= Y_0^{t+1}[62]Y_1^{t+1}[62] \oplus Y_1^{t+1}[62]Y_2^{t+1}[62] \oplus Y_2^{t+1}[62]Y_0^{t+1}[62] \\ Y_0^{t+1}[62] &= X_0^{t+1}[59] \oplus X_1^{t+1}[50] \oplus X_2^{t+1}[19] \oplus X_3^{t+1}[35] \\ Y_1^{t+1}[62] &= X_0^{t+1}[40] \oplus X_1^{t+1}[8] \oplus X_2^{t+1}[57] \oplus X_3^{t+1}[32] \\ Y_2^{t+1}[62] &= X_0^{t+1}[12] \oplus X_1^{t+1}[27] \oplus X_2^{t+1}[48] \oplus X_3^{t+1}[2]. \end{aligned} \tag{5}$$

To find linear equations, we observe every i^{th} bit of $Y_0^t[i], Y_1^t[i], Y_2^t[i]$ and $Y_0^{t+1}[i+8], Y_1^{t+1}[i+8], Y_2^{t+1}[i+8]$ of 64-bit registers and compare each component $X_k^t[i]$ with $X_k^{t+1}[i+8]$ for $k = 0, 1, 2, 3$ and check whether they are equal or not. The purpose that these bits belong to either in LFSR or NLFSR of the components of FSR X_k. Example, for $i = 54$, $X_0^t[p] = X_0^{t+1}[p+8]$ for $p = 51, 42, 11, 27$. Similarly, some other relations can also be obtain as follows.

$$\begin{aligned} X_0^t[p] &= X_0^{t+1}[p+8] \text{ for } p = 51, 32, 4 \\ X_1^t[p] &= X_1^{t+1}[p+8] \text{ for } p = 42, 0, 19 \\ X_2^t[p] &= X_2^{t+1}[p+8] \text{ for } p = 11, 49, 40 \\ X_3^t[p] &= X_3^{t+1}[p+8] \text{ for } p = 27, 24 \end{aligned} \tag{6}$$

Using (2), (4) and (5), we get

$$\begin{aligned} Y_0^{t+1}[62] &= Y_0^t[54] \\ Y_1^{t+1}[62] &= Y_1^t[54] \\ Y_2^{t+1}[62] &= Y_2^t[54] \oplus X_3^t[58] \oplus X_3^{t+1}[2]. \end{aligned}$$

Now, we have

$$\begin{aligned} Z^t[54] \oplus Z^{t+1}[62] = &(Y_0^{t+1)}[62]Y_1^{t+1}[62] \oplus Y_1^{t+1}[62]Y_2^{t+1}[62] \oplus Y_2^{t+1}[62]Y_0^{t+1}[62]) \\ &\oplus (Y_0^{t+1)}[62]Y_1^{t+1}[62] \oplus Y_1^{t+1}[62](Y_2^t[54] \oplus X_3^t[58] \\ &\oplus X_3^{t+1}[2]) \oplus (Y_2^t[54] \oplus X_3^t[58] \oplus X_3^{t+1}[2])Y_0^{t+1}[62]). \end{aligned}$$

which gives

$$Z^t[54] \oplus Z^{t+1}[62] = (X_3^t[58] \oplus X_3^{t+1}[2])(Y_0^t[54] \oplus Y_1^{t+1}[62]). \tag{7}$$

In particular, if $Z^t[54] \oplus Z^{t+1}[62] = 1$, then we have

$$X_3^t[58] \oplus X_3^{t+1}[2] = 1 \tag{8}$$
$$Y_0^t[54] \oplus Y_1^{t+1}[62] = 1. \tag{9}$$

The above equation also holds for $i = 55$. Indeed, we have $Z^t[55] \oplus Z^{t+1}[63] = (X_3^t[59] \oplus X_3^{t+1}[3])(Y_0^t[55] \oplus Y_1^{t+1}[63])$. Thus, when $Z^t[55] \oplus Z^{t+1}[63] = 1$, we also have

$$X_3^t[59] \oplus X_3^{t+1}[3] = 1 \tag{10}$$
$$Y_0^t[55] \oplus Y_1^{t+1}[63] = 1. \tag{11}$$

Recovering the initial states of X_3

Equations (8) and (10) involve the expression of X_3 only for the different initial states. Thus, to recover the initial states of X_3 we have to solve the expression about 64 nonlinear equations. For this purpose, we collect 64 nonlinear equations which satisfies key bits of $Z^t[54] \oplus Z^{t+1}[62] = 1$ or $Z^t[55] \oplus Z^{t+1}[63] = 1$ for each possible time $t + j (j \geq 0)$. Now, X_3 is comprise of 47-bit state of linear sub-FSR and 17-bit state of non-linear sub-FSR, therefore we get linear equations that involve 47-bit state variables $H_{47}(X_3^t)$ of the linear sub-FSR and 17-bit state variables $L_{17}(X_3^t)$ of the nonlinear sub-FSR of X_3. Assuming the 17-bit of $L_{17}(X_3^t)$ at time t are known, then 64 nonlinear equations are reduced to linear equations on 47 variables $H_{47}(X_3^t)$ of the linear sub-FSR of X_3. Further we check whether $64 * 64$ matrix is consistent or not. If not then we look for another guess. Thus we can recover 47 variable of $H_{47}(X_3^t)$ out of 64 equations. The rest of the equations are used to check the correctness of recovered X_3^t. Finally, X_3^t can be determined uniquely. In order to collect 64 linear equations in the form of Eq. (8), we require less than 400 64-bit words. We find one equation for each possible state j on average. After collecting these equations we have to guess 2^{17} possible states $L_{17}(X_3^t)$ of the nonlinear sub-FSR of X_3 to find unique solution. Finally, we solve a linear system of 47 bit variables $H_{47}(X_3^t)$ for each possible states of $L_{17}(X_3^t)$. Thus we can recover initial states of the feedback of the linear sub-FSR of X_3 in a simple manner.

Recovering the initial states of X_2

Set $i = 3$. We find from (6),

$$\begin{aligned} Y_0^{t+1}[11] &= Y_0^t[3] \\ Y_1^{t+1}[11] &= Y_1^t[3] \oplus X_2^t[62] \oplus X_2^{t+1}[6] \\ Y_2^{t+1}[11] &= Y_2^t[3]. \end{aligned} \tag{12}$$

From the computation of $Z^t[3]$ and $Z^{t+1}[11]$, we get $Z^t[3] \oplus Z^{t+1}[11] = (X_2^t[62] \oplus X_2^{t+1}[6]).(Y_0^t[3] \oplus Y_2^t[3])$. If $Z^t[3] \oplus Z^{t+1}[11] = 1$, then we have

$$X_2^t[62] \oplus X_2^{t+1}[6] = 1 \tag{13}$$

$$Y_0^t[3] \oplus Y_2^t[3] = 1. \tag{14}$$

Clearly, the Eq. (13) involves the state variables of X_2 only. Similarly, to recover X_3^t, we first collect 64 linear equations by those key bits satisfying $Z^{t+j}[3] \oplus Z^{t+j+1}[11] = 1$, for $j \geq 0$. For this purpose, we have to guess 23-bit state variable $L_{23}(X_2^t)$ of the non-linear sub-FSR of X_2 and solve 41 out of 64 linear equations of 41-bit state variables $H_{41}(X_2^t)$ of the linear sub-FSR of X_2. This process is repeated for each possible j on average until we get the solutions. The rest of the linear equations are used to check the correctness. Finally X_2^t can be determined uniquely. In this case we have to guess 2^{23} possible states $L_{23}(X_2^t)$ of the nonlinear sub-FSR of X_3.

Recovering the initial states of X_1

Set $i = 37$. Similar to previous sections, we observe that

$$\begin{aligned} Z^t[37] + Z^{t+1}[45] = (Y_1^t[37] + Y_2^{t+1}[45])(X_1^t[25] + X_1^{t+1}[33] \\ + X_2^t[58] + X_2^{t+1}[2] + X_3^t[10] + X_3^{t+1}[18]). \end{aligned}$$

If $Z^t[37] + Z^{t+1}[45] = 1$, then

$$Y_1^t[37] + Y_2^{t+1}[45] = 1 \tag{15}$$

$$X_1^t[25] + X_1^{t+1}[33] + X_2^t[58] + X_2^{t+1}[2] + X_3^t[10] + X_3^{t+1}[18] = 1. \tag{16}$$

The above relation also holds for $i = 38$, that is,

$$\begin{aligned} Z^t[38] + Z^{t+1}[46] = (Y_1^t[38] + Y_2^{t+1}[46])(X_1^t[26] + X_1^{t+1}[34] \\ + X_2^t[59] + X_2^{t+1}[3] + X_3^t[11] + X_3^{t+1}[19]) \end{aligned}$$

If $Z^t[38] + Z^{t+1}[46] = 1$, then

$$Y_1^t[38] + Y_2^{t+1}[46] = 1 \tag{17}$$

$$X_1^t[26] + X_1^{t+1}[34] + X_2^t[59] + X_2^{t+1}[3] + X_3^t[11] + X_3^{t+1}[19] = 1. \tag{18}$$

Since X_3^t and X_2^t are known by previous sections, therefore we easily find the linear equations that involves the initial state of X_1 only. Once enough linear equations are collected, we guess the state $L_{29}(X_1^t)$ of the nonlinear sub-FSR of X_1 directly and solve with 35 linear equations out of 64 linear equations of the state variables $H_{35}(X_1^t)$ of the linear sub-FSR of X_1 and the rest of the equations is used to check the correctness of X_1^t. Finally, we get the unique solution of X_1^t. In this case we have to guess 2^{29} possible states $L_{29}(X_1^t)$ of the nonlinear sub-FSR of X_1.

Recovering the initial states of X_0

Set $i = 50$. We observe that

$$Z^t[50] + Z^{t+1}[58] = (Y_0^t[50] + Y_2^{t+1}[58])(X_0^t[28] + X_0^{t+1}[36] + X_1^t[60] + X_1^{t+1}[4])$$

As similar process to previous section, we observe that if $Z^t[50] + Z^{t+1}[58] = 1$, then $Y^t[0][50] + Y_2^{t+1}[58] = 1$ and $X_0^t[28] + X_0^{t+1}[36] + X_1^t[60] + X_1^{t+1}[4] = 1$. Also, it also holds for $i = 59$, that is,

$$Z^t[51] + Z^{t+1}[59] = (Y_0^t[51] + Y_2^{t+1}[59])(X_0^t[29] + X_0^{t+1}[37] + X_1^t[61] + X_1^{t+1}[5]).$$

We also observe that

$$Z^t[56] + Z^{t+1}[0] = (Y_0^t[56] + Y_1^{t+1}[0])(X_1^t[21] + X_1^{t+1}[29] + X_3^t[60] + X_3^{t+1}[4])$$
$$Z^t[57] + Z^{t+1}[1] = (Y_0^t[57] + Y_1^{t+1}[1])(X_1^t[22] + X_1^{t+1}[30] + X_3^t[61] + X_3^{t+1}[5])$$

Now, X_3^t, X_2^t and X_1^t are known by previous sections, therefore knowing these values, we easily find the equations in terms of initial state of X_0^t only and recover the initial states of X_0^t using above equations. In this case, we guess the state $L_{31}(X_0^t)$ of the nonlinear sub-FSR of X_0 directly and solve the 33 linear equations of the state variables $H_{33}(X_0^t)$ of the linear sub-FSR of X_0 and the rest of the equations are used to check the correctness of X_0^t.

3.2 Phase II: Recovering the Key K

This section deals to recover the key K from the state register E of the FSR E. If the process of initialization is known then one can easily recover the state E, since FSRE is invertible. At initialization process, the three values of TWEAK is not known, therefore we easily get the intermediate state (that is, TWEAK) of initialization process of Algorithm 1 in 2^3 possible values of E denoted by $E_1, E_2, \ldots, E_8$ because TWEAK is not a permutation. For each possible value $E_i (1 \leq i \leq 8)$, we invert steps from 7 to 2 in turn. It is experimentally observed that the rank of the matrix of the linear transformation determined by steps 4 and 5 at initialization process is 189, therefore we have to fix three (that is, 2^3) arbitrary value of matrix to recover the state. Since steps 2–7 loop eight times, so we can get totally 2^8 possible values to reach step 2 for each E_i, denoted by $E_{i,j}$, where $1 \leq j \leq 2^8$. Finally, we verify whether the prefix of each possible $E_{i,j}$ (totally 2^{11} possible values) is equal to $0x5a5a \ldots 5a$ or not. If some $E_{i,j}$ gives the correct solution, then one candidate K is written down. Here it should be emphasized that all candidates K are valid and they are equivalent to each other.

4 Flaws in Design and Methods to Avoid the Attack

We have already seen that FASER128 is a weak cryptosystem and weaknesses are found mainly in Mix and MAJ functions. The authors reported in the paper [7]

that the rank of matrix used at initialization process of Algorithm 1 was 191 while by simulating the program, it is found that it was 189. Therefore, some complexity increases to revert the initialization process. Second observation is that this particular attack is possible if the output sequence $Z^t[i] \oplus Z^{t+1}[i + 8 \bmod 64]$ is not balanced for each i where $1 \leq i \leq 64$. The balancedness of these output sequences was not mentioned by the authors [7]. Another observation is that the mixing of key stream between LFSR and NLFSR of subFSR is not proper therefore the output sequences mentioned above is not balanced. We also observe that these equations are possible because of the poor choice of rotation parameters present in Mix function and if the output sequence $Z^t[i] \oplus Z^{t+1}[i + 8 \bmod 64]$ is not balanced then it is always possible to get linear equations in terms of linear sub-FSR and output sequences. Therefore we conclude that the attack is possible if one can easily deduce linear equations corresponding to LFSR. Based on these observations, we have done a lot of experiments to avoid this type of attack and improve key stream cipher with the same speed and the same security elucidated as follows:

4.1 Fixing the Rotation Parameters in Mix Function

We have analysed the strength of output key stream sequences by changing the various rotation parameters present in the Mix function. If we denote these rotation parameters present in Mix variable is as $[t_{ij}]$ matrix where i, $0 \leq i \leq 3$ denote the corresponding to rotation parameter of Y_j variables, where $0 \leq j \leq 2$, given in Mix function. If we set $\{t_{00} = 16, t_{10} = 30, t_{20} = 39, t_{30} = 7; t_{01} = 36, t_{11} = 54, t_{21} = 52, t_{31} = 28; t_{02} = 22, t_{12} = 37, t_{22} = 46, t_{32} = 61\}$. In this case one cannot find linear equation for any i corresponding to the condition $Z^t[i] \oplus Z^{t+1}[i + 8 \bmod 64] = 1$. Therefore one cannot mount the key recovery attack on the stream cipher in real time.

4.2 Changing the Set of the Rotation Parameters in Mix Function

By changing the several rotation parameters in Mix function it has been found that if we fix MSB (most significant bits) of first four subFSR in Eq. (2), then the following rotation parameters present in Mix function gives better results of balancedness and consequently, we get better diffusion property and so one cannot mount the key recovery attack on FASER128. The set of 16 rotation parameters in Mix function are given as follows:

4.3 Changing the Clock $Z^t[i + 16 \bmod 64]$

It is observed that by changing the clock of update function $Z^t[i + 16 \bmod 64]$ in place of $Z^t[i + 8 \bmod 64]$, the output sequences $Z^t[i] \oplus Z^{t+1}[i + 8 \bmod 64]$ for each i are almost balanced at the following rotation parameters {20 47 39 16; 37 26 52v 32; 21 7 25 0} and one cannot find linear equation in this case. The time complexity to mount the attack in this case is much high. Therefore, the key recovery attack is not possible in this scenario.

Set/t_{ij}	t_{00}	t_{10}	t_{20}	t_{30}	t_{01}	t_{11}	t_{21}	t_{31}	t_{02}	t_{12}	t_{22}	t_{32}
1	30	22	46	18	5	21	19	57	15	63	53	25
2	29	20	19	37	50	45	4	60	26	41	12	2
3	48	24	37	12	41	11	45	2	54	4	27	44
4	3	20	4	59	56	6	42	1	31	38	53	46
5	24	18	40	15	10	51	7	0	36	60	49	41
6	13	11	3	57	34	21	62	55	51	30	15	25
7	24	2	51	41	8	12	60	28	15	56	36	49
8	36	46	27	17	44	60	59	51	57	33	12	24
9	37	47	18	27	51	61	9	41	57	4	55	46
10	38	62	55	18	27	51	23	53	33	11	25	26
11	56	37	44	11	63	52	14	22	48	6	59	47
12	52	3	37	58	49	48	10	26	9	54	4	45
13	34	28	20	48	13	37	2	56	21	42	38	12
14	37	18	52	44	51	28	11	7	31	14	0	40
15	30	5	16	46	53	44	20	28	36	23	11	59
16	42	15	4	34	54	36	38	63	28	57	49	44

5 Security Analysis

Based on change of rotation parameters and apply other parameters mentioned in previous section, we analysed the following security issue in the cryptoalgorithm of FASER128.

5.1 Avalanche Effect

For a good cryptoalgorithm, output key sequences should satisfy good avalanche criteria, that is, change in single input key bits gives almost 50% change in corresponding output key stream. For this purpose, we have changed every single bit of 256 initial bits of FSR E and check the avalanche criterion of whole output key stream. It has been found that significant change exists in output key stream. This ensures that the output sequence satisfies avalanche criterion.

5.2 Algebraic Attacks

For FASER128, the number of variables in the output keystream from the linear subFSRs and nonlinear subFSRs is $v = 256 + 64n$ where n is number of rounds and the number of equations is $e = 128n$, discussed in [5]. Hence algebraic attack is not applicable in this case.

Side channel attacks and other security issues are same as discussed in [5].

6 Conclusion

In this paper, we have discussed cryptanalysis of FASER128 by method of key recovery attack [7] which require only a few key words, that is, about less than 400 words and recovered all possible keys K in real time with single PC. So, FASER128 is a very insecure cryptosystem. It is observed that some sets of rotation parameters present in Mix function mentioned in Sect. 4 gives significant improvements to avoid this particular type of attack. By changing the clock of update functions also gives better improvement to avoid this attack with the same security and the same speed mentioned in [5].

Acknowledgement. The authors greatly indebted to Ms. Neelam Verma, Scientist G and Ms. Anu Khosla, Director, SAG, DRDO for their full cooperation and financial supports. The third author wishes to express his thanks to DRDO headquarter for DRDO fellowship.

References

1. CAESAR: Competition for authenticated encryption: Security applicability and robustness. https://competitions.cr.yp.to/caeser.html
2. Argen, M., Londhahl, C., Hell, M., Johansson, T.: A survey on fast correlation attack. Crypt. Commun. **4**(3), 173–202 (2012)
3. Bellare, M., Kohno, T., Namprempre, C.: Authentication encryption in SSH: provably fixing the SSH binary packet protocol. In: ACM Conference on Computer and Communications Security (CCS-9), pp. 1–31. ACM Press (2002)
4. Bellare, M., Namprempre, C.: Authenticated encryption: relations among notions and analysis of the generic composition paradigm. In: Okamoto, T. (ed.) ASIACRYPT 2000. LNCS, vol. 1976, pp. 531–545. Springer, Heidelberg (2000). doi:10.1007/3-540-44448-3_41
5. Chaza, F., MacDonald, C., Avanzi, R.: FASER v1: Authenticated encryption in a feedback shift register, CAESER (2014)
6. Xu, C., Zhang, B., Feng, D.: Linear cryptanalysis of FASER128/256 and TriviA-ck. In: Meier, W., Mukhopadhyay, D. (eds.) INDOCRYPT 2014. LNCS, vol. 8885, pp. 237–254. Springer, Cham (2014). doi:10.1007/978-3-319-13039-2_14
7. Feng, X., Zhang, F.: A real time key recovery attack on the authenticated FASER128. Cryptology ePrint Arxhive, Report 2014/258 (2014). http://eprint.iacr.org/

Modelling of Aircraft's Dynamics Using Least Square Support Vector Machine Regression

Hari Om Verma(✉) and Naba Kumar Peyada

Aerospace Department, IIT Kharagpur, Kharagpur, West Bengal, India
homverma@gmail.com, nkpeyada@aero.iitkgp.ernet.in

Abstract. The system identification is a broad area of research in various fields of engineering. Among them, our concern is to identify the aircraft dynamics by means of the measured motion and control variables using a new approach which is based on the support vector machine (SVM) regression. Due to the computational complexity of SVM, it is suggested to adopt the advanced version of SVM i.e. least square support vector machine (LSSVM) to be used for system identification. LSSVM regression is a network-based approach which requires a user defined kernel function and a set of input-output data for its training before the prediction phase like a neural-network (NN) based procedure. In this paper, LSSVM regression has been used to identify the non-linear dynamics of aircraft using real flight data.

Keywords: System identification · LSSVM regression · Kernel function

1 Introduction

System Identification (SI) is basically concerned with the mathematical modelling which is obtained from the available measured input and output data of the system [1]. It is like solving an inverse problem from the given data implicitly [2]. There are three quantities involved in the process of identifying the system which are the inputs, mathematical functions representing the dynamical system, and the outputs. SI attempts only to find the mathematical functions [3,4].

The mathematical functions can be represented in the form of differential equations which are simply formulated based on the process of physics leading to Newtonian mechanics. This type of modelling is said to be phenomenological models which, requires a high level of information a prior, leads to a complex model [3]. So, a different type of model is required which can approximate the observed behaviour for specific input without any intention of knowing the internal dynamics of system. It is said to be a behavioural model which is easy to derive and establish an overall cause-effect relationship.

Another way of classifying the system identification process is to divide modelling based on parametric and non-parametric approaches. The parametric approach involves a well-known established structure based on physical processes just

© Springer Nature Singapore Pte Ltd. 2017
D. Giri et al. (Eds.): ICMC 2017, CCIS 655, pp. 132–140, 2017.
DOI: 10.1007/978-981-10-4642-1_12

like phenomenological model either in the linear or non-linear form. One way of structured modelling is based on state-space which can be represented in the form of linear or non-linear, continuous or discontinuous, time-invariant or time-variant, and deterministic or stochastic [5]. Another way is based on transfer function which is applicable only to represent linear system. The non-parametric approach is an alternative non-hectic strategy using some kind of a mathematical function representing an input-output relationship. Such models have been developed using artificial neural network (ANN) [6]. ANN is a multi-layer feed forward neural network with a number of neurons in each layer. The first layer of the network is input layer, then intermediate as hidden layer and finally output layer. Each neuron except the input layer gets the signals from the previous layer neurons multiplied with some weights and it processes the signal through the transfer function such as sigmoid, log sigmoid etc. For training of ANN, either supervised learning or unsupervised learning is preferred. Through supervised learning methodologies, ANN is trained for network weights so that it becomes an approximate representation of input-output relationship [7,8]. Many aerospace researchers have used ANN as a function estimator for identifying the aircraft dynamics nonlinearly [9–12]. They have used various types of neural network architecture for further investigation such as for aircraft parameters estimation.

Some of the drawbacks have been identified with neural network such as lesser generalization capability of the network and more iteration required for training. Such limitations occurred due to the concept of empirical risk minimization (ERM) principal employed by the ANN. These limitations have been overcome by using one of the statistical strategies using structural risk minimization (SRM) principle such as support vector machine (SVM) [13]. SVMs have been widely used in the field of classification, pattern recognition, and function estimation. It has been used for non-linear mapping from input space to output space which takes out the problem to a quadratic programming and hence the solution is found to be global minimum. The solution of the quadratic programming makes it computationally hard. So, a modified version of SVM, known as least square support vector machine (LSSVM), has been used for non-linear mapping which is computationally faster than SVM [14,15].

In this paper, LSSVM regression method is presented to address the problem of the identifying the aircraft's dynamics by means of using the real flight data. Section 2 represents the basic prerequisites used for system identification in the process of data gathering and its compatibility. Section 3 represents the basic mathematical formulation of LSSVM regression for non-linear mapping from input space to output space, and the input-output details for non-linear modelling. Section 4 represents the results obtained during the training and predicting phase. Section 5 represents the concluding remarks on the LSSVM regression method used for modelling of aircraft's dynamics.

2 Prerequisites for System Identification/Modelling

The process of the system identification is fully dependent on available input-output data, so the preliminary step is the real flight data gathering and the

second step is the data compatibility check to verify and improve the quality of data from biases, scale factors, and time lags etc. [3].

2.1 Data Gathering

Data gathering is a process of recording inputs and outputs while performing a certain type of experiment which is basically to excite the mode of the aircraft. As it is a data acquisition process to record the aircraft motion variables and control surface deflections, but it is fully dependent on the quality of sensors in terms of accuracy and noise, sampling rate, signal conditioners, and data recording equipment.

The first step in data gathering process is to define the type of experiment such as excitation mode of short-period, phugoid, pushover-pullup, level-turn, thrust variation, bank-to-bank roll, Dutch roll manoeuvre, and steady heading steady sideslip. To excite each of the above manoeuvres, a corresponding excitation input is given to either of the control surfaces such as elevator, aileron, rudder, and/or the throttle setting. These excitation inputs are as: (i) step, (ii) doublet, (iii) 3-2-1-1 signal, (iv) modified 3-2-1-1 signal.

The second step is to take care with the instrumentation and measurement unit for signal processing and data recording to fulfil the following criteria [3]:

- Lower sampling rate satisfying Nyquist frequency criteria.
- Anti-aliasing filter introducing the same time delay in the signals.
- Recording of raw data for further processing such as differentiation, integration, or filtering of the data.
- Highly critical measurements like translational accelerations, angular rates, and control surface deflections must be sampled at higher and uniform rates while slowly varying parameters like altitude at slower rate.
- All data channels must be synchronized with time.
- The signal-to-noise ratio of 10:1 is desirable.
- All sensors must be calibrated in laboratory with high accuracy.
- Data reduction must be avoided at the time of recording.

2.2 Data Compatibility Check

Data compatibility check is another important step after the data gathering process which checks and improves the quality of the recorded data in terms of scale factor, zero shift biases, and time lags. It uses well defined kinematic equations of aircraft motion to reconstruct the flight path with the same trim conditions as used in while doing the flight test. Thus, a mismatch in the measured and flight path reconstructed is used to determine the systematic instrument errors such as scale factors, zero shifts, and time delays using the conventional output error method [3].

A real flight data has been generated using a research aircraft “HANSA” at IIT Kanpur, India [10,11]. The short period mode of the longitudinal dynamics has been excited using the control surface – elevator from the steady state trim

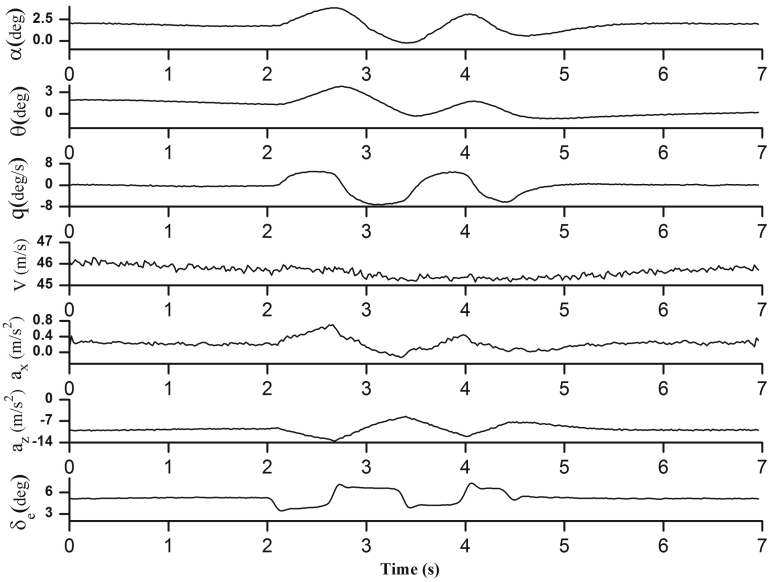


Fig. 1. Measured flight data

condition. For system identification purpose, we have chosen the longitudinal measured data of aircraft such as angle-of-attack (α), pitch angle (θ), pitch rate (q), velocity of the aircraft (V), the linear accelerations along the body axes (a_x & a_z), and the control surface deflection (δ_e). The generated flight data is shown in the Fig. 1.

3 Least Square Support Vector Machine Regression Based Modelling

This section is divided into a number of sections to ease the understanding of the modelling process step-by-step.

3.1 Mathematical Formulation of LSSVM Regression

Least Square support vector machine (LSSVM), is based on one of the statistical learning principles, and employs the structural risk minimization (SRM) principle which has been found to be superior to empirical risk minimization (ERM) principle used in fuzzy logic (FL) and neural network (NN) [14,15]. The theoretical relationship between the input space and output space is given by a function which is as follows:

$$y = f(x) = w^T \phi(x) + b; \;\; x \in R^p, \, y \in R \tag{1}$$

Where, $\phi(x)$ – a non-linear transformational matrix between the input space and the output space, x– p dimensional input vector, w – weighting vector, and b – bias.

For non-linear modelling, a finite number of sample data is obtained from measurement $\{(x_i, y_i), i = 1, 2, 3, ..., n\}$. It is desired that all of the data can be fitted by the functional relationship in the Eq. (1) with ε precision which arises two inequality conditions as follows:

$$\begin{cases} y_i - w^T\phi(x_i) - b \leq \varepsilon \\ w^T\phi(x_i) + b - y_i \leq \varepsilon \end{cases}, \quad i = 1, 2, 3, ..., n \tag{2}$$

By introducing a slack variable (ξ), the optimization goal using the SRM principle is given as follows:

$$\begin{cases} \min\limits_{w,b,\xi} J = \frac{1}{2}\|w\|^2 + c\sum_{i=1}^{n}\xi_i^2 \\ Sub.\ to\, y_i = w^T\phi(x_i) + b + \xi_i \end{cases}, \quad i = 1, 2, 3, ..., n \tag{3}$$

Where, c is a predefined constant that is to minimize the cost function J. Its value determines the training error and the regression function flatness.

One can use the Lagrange function approach to solve the above cost function subjected to the equality constraints. Thus, the Lagrange function is given as follows:

$$L = \frac{1}{2}\|w\|^2 + c\sum_{i=1}^{n}\xi_i^2 - \sum_{i=1}^{n}a_i\left(w^T\phi(x_i) + b + \xi_i - y_i\right) \tag{4}$$

The following equations are obtained from Karush-Kuhn-Tucker's condition:

$$\begin{cases} w = \sum_{i=1}^{n} a_i\phi(x_i) \\ \sum_{i=1}^{n} a_i = 0 \\ a_i = c\xi_i \\ w^T\phi(x_i) + b + \xi_i - y_i = 0,\ i = 1, 2...n \end{cases} \tag{5}$$

After eliminating w and ξ_i from the above equations, one can get the following linear system:

$$\begin{bmatrix} 0 & 1_n^T \\ 1_n & \Omega + c^{-1}I_n \end{bmatrix}\begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \tag{6}$$

Where, $y = [y_1, y_2, ..., y_n]^T$; $1_n = [1, 1, ..., 1]^T$; $a = [a_1, a_2, ..., a_n]^T$;

$$\Omega_{i,j} = K(x_i, x_j) \quad i, j = 1, 2, ..., n.$$

Now, the Eq. (6) can be easily solved by using least-square method for the parameters "a" and "b". Therefore, LSSVM regression based model is given as follows:

$$y = f(x) = \sum_{i=1}^{n} a_i K(x, x_i) + b \tag{7}$$

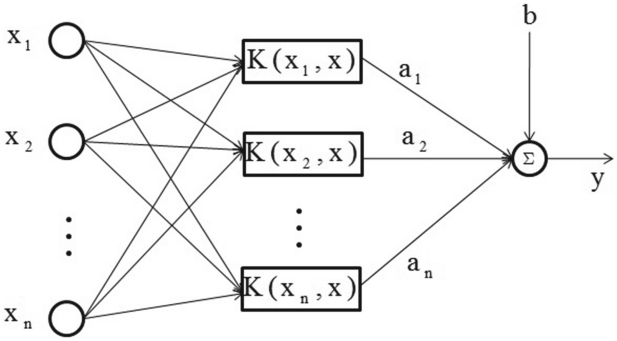


Fig. 2. Structure of LSSVM regression

In LSSVM regression methodology, a non-linear relationship requires only to solve the linear Eq. (6) with known kernel function, K and c. In Fig. 2, the network architecture of LSSVM regression is shown.

Some of the typical choices of kernel function are given below:

1. Linear Kernel Function: $K(x, x_i) = x_i^T x$
2. Multi-Layer Perceptron Kernel Function: $K(x, x_i) = \tanh(\gamma x_i^T x + r)$
3. Polynomial Kernel Function: $K(x, x_i) = (\gamma x_i^T x + r)^d, \ \gamma > 0$
4. Radial Basis Kernel Function: $K(x, x_i) = \exp\left(-\gamma \|x - x_i\|^2\right), \ \gamma > 0$

Where,γ is the kernel width. Thus, the training of LSSVM requires γ and c parameters to be well chosen so that the root-mean-square error (RMSE) can be minimized to its lowest value. In our case, radial basis kernel function has been used.

3.2 Input-Output Details for Modelling

Figure 2 shows the architecture of LSSVM regression for a multi-input single-output (MISO) system whereas our objective is to extend the concept of MISO system into multi-input multi-output (MIMO) system. The MIMO system architecture of LSSVM regression has been implemented using MATLAB in which ithsample of the input vector is given as follows:

$$x_i = [\alpha(i),\ \theta(i),\ q(i),\ V(i),\ C_D(i),\ C_L(i),\ C_m(i)]^T \tag{8}$$

Where, α - angle of attack, θ - pitch angle, q - pitch rate, V - velocity of the aircraft, and C_D, C_L and C_m are the coefficients of drag, lift and pitching moment respectively, which are represented here for longitudinal dynamics of the aircraft in a simplified form, and they are given as follows [11]:

$$C_D(i) = -C_X(i)\cos(\alpha(i)) - C_Z(i)\sin(\alpha(i)) \tag{9}$$

$$C_L(i) = C_X(i)\sin(\alpha(i)) - C_Z(i)\cos(\alpha(i)) \tag{10}$$

$$C_m(i) = [I_y\dot{q}(i) - F_{eng}Z_{enCG}]/(\bar{q}(i)S\bar{c}) \tag{11}$$

Where, the body forces coefficients (C_X and C_Z) are given as follows:

$$C_X(i) = ma_X^{CG}/\bar{q}S \tag{12}$$

$$C_Z(i) = ma_Z^{CG}/\bar{q}S \tag{13}$$

In the Eqs. (9–13), the terms used are as follows: a_X^{CG} and a_Z^{CG}- the linear body accelerations at centre of gravity (CG) of the aircraft along x and z axis, respectively, F_{eng} - total thrust, Z_{enCG} - the vertical distance between CG and the engine, I_y - the moment of inertia of the aircraft along the y-axis, $\bar{q}$ - the dynamic pressure of the ambient, S - reference area, and $\bar{c}$- aerodynamic chord length.

For the training of MIMO system based LSSVM regression, the target vector has been considered at $(i+1)^{th}$ instant which is given as follows:

$$Z(i+1) = [\alpha(i+1),\ \theta(i+1),\ q(i+1),\ V(i+1),\ a_X^{CG}(i+1),\ a_Z^{CG}(i+1)]^T \tag{14}$$

4 Results and Discussion

As the LSSVM regression mathematical formulation has been given in the Sect. 3 for multi-input single-output (MISO) case, while the nonlinear mapping has been done using MATLAB code for multi-input multi-output (MIMO) case which is the extension of the MISO case. The radial basis kernel function has been chosen for LSSVM regression model. The values of "c" and "γ" have been determined

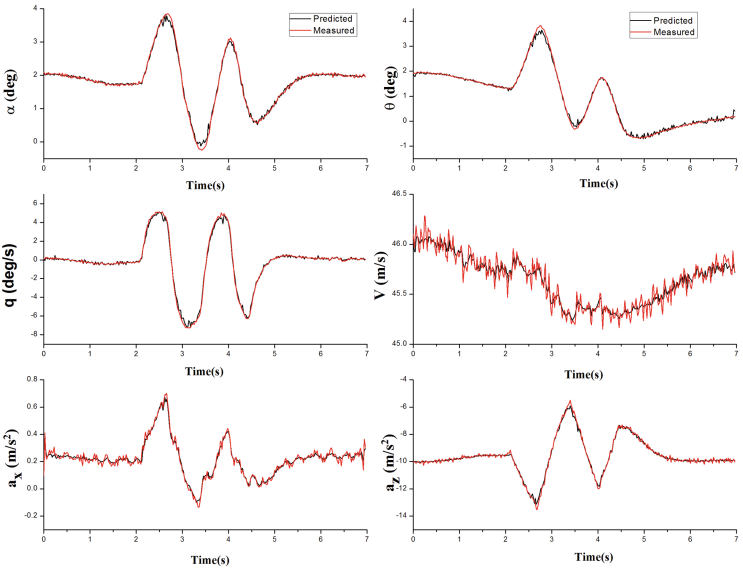


Fig. 3. Prediction case 1

Table 1. RMSE of different outputs

Output	α	θ	q	V	a_X	a_Z
RMSE	0.0010	0.0011	0.0030	0.0790	0.0254	0.1357

by trial and error method which ensures least value of the root-mean-square error (RMSE) at the output. First "γ" has been selected based on normalization of the norm of the input vectors and then "c" value has been varied from "1" to some finite value, say 10. Finally, $\gamma = 1$ and c = 2, have been chosen for our purpose. Table 1 shows the root- mean-square error of the outputs.

It is found that as the c value changes from the lower to a higher value, robustness is improved but it leads to over fitting which defines that any small change in the input value will not have any effect on the response. Once the modelling part is over, two sets of the data are used to predict from the trained model: one has the whole input data set having 349 samples while the other has a part of the whole input data set having 101 samples from 2–4 second interval of time. Figure 3 shows the first type of prediction case, in which there is a comparison between the predicted values from the trained model and the measured values of the output used at the time of training, while Fig. 4 shows the second type of prediction case. Both the results have shown a quite good matching with the measured outputs.

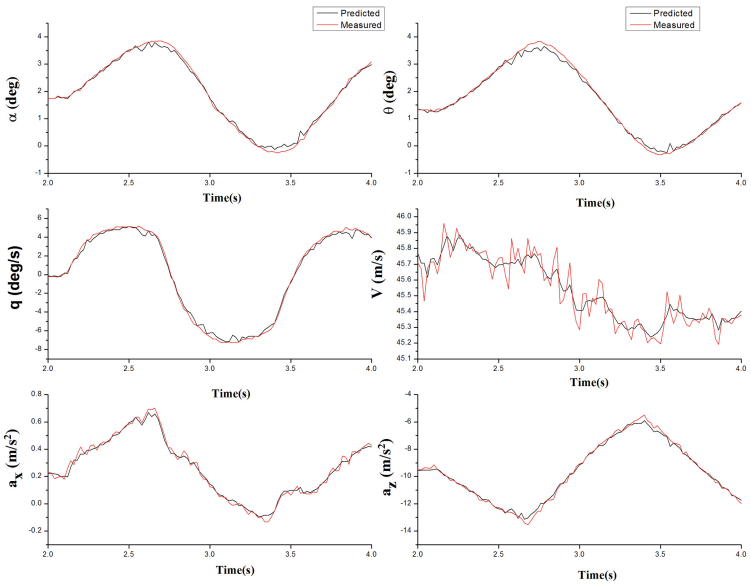


Fig. 4. Prediction case 2

5 Conclusion

In this paper, a new method has been proposed for identifying the dynamics of the aircraft using LSSVM regression method. It uses the non-linear mapping characteristics of LSSVM to establish a relationship between the chosen input and output variables which is fully dependent on the design parameters of the method. Such type of modelling concept can be used for any non-linear system where dynamic equations of motions are complex or completely unknown but input-output variables are measurable. Here, we have approached to model the short period dynamics of aircraft at a well defined operating condition which can be used in the design of control system of the aircraft. One can further extend the concept of the modelling for a global model which describes for the whole flight envelope.

References

1. Zadeh, L.A.: From circuit theory to system theory. Proc. IRE **50**, 856–865 (1962)
2. Hamel, P.G., Jategaonkar, R.V.: The evolution of flight vehicle system identification. AGARD, DLR Germany, 8–10 May 1995
3. Jategaonkar, R.V.: Flight Vehicle System Identification: A Time Domain Methodology. AIAA Progress in Aeronautics and Astronautics, vol. 216. AIAA, Weinheim (2006)
4. Klein, V., Morelli, E.: Aircraft System Identification: Theory and Practice. AIAA Education Series Inc., Reston (2006)
5. Goman, M., Khrabrov, A.: State-space representation of aerodynamic characteristics of an aircraft at high angles of attack. J. Aircr. **31**(5), 1109–1115 (1994)
6. Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural Netw. **2**, 359–366 (1989)
7. Hassoun, M.H.: Fundamental of Artificial Neural Networks. The MIT Press, Cambridge (1995)
8. Haykins, S.: Neural Networks: A Comprehensive Foundation. McMaster University, Macmillan College Publishing Company, New York (1994)
9. Raisinghani, S.C., Ghosh, A.K., Kalra, P.K.: Two new techniques for parameter estimation using neural networks. Aeronaut. J. **102**(1011), 25–29 (1998). UK
10. Singh, S., Ghosh, A.K.: Estimation of lateral-directional parameters using neural networks based modified delta method. Aeronaut. J. **111**(1124), 659–667 (2007). UK
11. Peyada, N.K., Ghosh, A.K.: Aircraft parameter estimation using new filtering technique based on neural network and Gauss-Newton method. Aeronaut. J. **113**(1142), 243–252 (2009)
12. Kumar, R., Ghosh, A.K.: Nonlinear longitudinal aerodynamic modeling using neural Gauss-Newton method. J. Aircr. **48**(5), 1809–1812 (2011). AIAA, USA
13. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, Berlin (1995)
14. Suykens, J.A.K., Gestel, T.V., Brabanter, J.D., Moor, B.D., Vandewalle, J.: Least Squares Support Vector Machines. World Scientific Publishing Co., Singapore (2002)
15. Wang, Q., Qian, W., He, K.: Unsteady aerodynamic modelling at high angles of attack using support vector machines. Chin. J. Aeronaut. **28**(3), 659–668 (2015)

Accommodative FAS-FMG Multilevel Based Meshfree Augmented RBF-FD Method for Navier-Stokes Equations in Spherical Geometry

Nikunja Bihari Barik(✉) and T.V.S. Sekhar

Indian Institute of Technology Bhubaneswar, Bhubaneswar, Odisha, India
{nbb10,sekhartvs}@iitbbs.ac.in
http://www.iitbbs.ac.in/

Abstract. The efficiency of any numerical scheme measures on the accuracy of the scheme and its computational time. An efficient meshfree augmented local radial basis function (RBF-FD) method has been developed for steady incompressible Navier-Stokes equations in spherical geometry with unbounded domain which is based on accommodative FAS-FMG multigrid method. The axi-symmetric spherical polar Navier-Stokes equations are solved without using transformation. The non-linear convective terms are handled efficiently by considering upwind type of RBF nodes. The developed scheme saves around 34% of the CPU time than the usual RBF-FD method.

Keywords: Radial basis function · Accommodative FAS-FMG multilevel method · Meshless method · Unbounded flows · Navier-Stokes equations

1 Introduction

The increasing use of computational fluid dynamics (CFD) for engineering design and analysis demands highly efficient solution methods. The discretization of numerical methods for solving elliptic Navier-Stokes(N-S) equations generally results in solving a system of algebraic equations. If the number of unknowns are large, solving by a direct method, such as Gaussian elimination, can be inefficient. Therefore, iterative methods like point Gauss-Seidel and line Gauss-Seidel are used to solve the huge linearized system of equations. For better convergence of the iterative methods, a good initial solution is essential. It was also found that Gauss-Seidel iterative method is effective for the first few iterations and then the error elimination process becomes slow. Based on this fact, a fast finite difference numerical method has been developed by Hyman [1] to solve elliptic partial differential equations with Dirichlet boundary conditions. His method is based on a local mesh refinement technique which provides a better initial guess for the iterative algorithms. The solution is achieved quickly and the CPU

© Springer Nature Singapore Pte Ltd. 2017
D. Giri et al. (Eds.): ICMC 2017, CCIS 655, pp. 141–151, 2017.
DOI: 10.1007/978-981-10-4642-1_13

time is minimized. Over the past few decades, finite difference based multigrid methods have been developed to solve the system of equations so as to improve the convergence rate of iterative methods and hence their efficiency. Ghia et al. [2] developed accommodative version of the Full Approximation Scheme-Full MultiGrid (FAS-FMG) procedure of Brandt [3] and applied this to Navier-Stokes equations. It is well known that RBF based methods suffer from high computational cost compared to conventional mesh based methods. The calculation of RBF weights corresponding to the neighboring particles of a data point, requires expansive square root and matrix inversion processes. Moreover, the calculation of derivative approximation at a given order of accuracy usually requires more number of neighboring particles (or nodes) for meshfree methods in an irregular grid than for finite difference method (FDM) on a cartesian grid. As a result, the bandwidth of matrices representing the governing algebraic equations greatly expands in case of meshfree methods [4,5]. Therefore, the iteration process gets slowed down due to the relatively dense matrix equations and the computational efficiency is reduced. At the same time, meshfree methods have the advantage of handling complex geometries efficiently. However, generation of an efficient mesh, which could ensure accurate results, is generally a tedious and time consuming task in the cartesian grid. To make the numerical scheme efficient Ding et al. [4] combined the conventional FD scheme with meshfree least square based finite differences (MLSFD). In a similar manner Javed et al. [5] used a hybrid scheme which combines RBF-FD with conventional FD schemes. The aim of the paper is to develop an efficient RBF-FD method to reduce the overall CPU time for solving Navier-Stokes equations in spherical geometry without using any transformation.

2 Augmented RBF-FD Formulation for Curvilinear Coordinates

The RBF based local method (RBF-FD) which has been proposed by Shu et al. [6], Tolstykh et al. [7], Cecil et al. [8], Wright and Fornberg [9] is spectrally accurate for a sparse matrix, better conditioned linear system and more flexibility for nonlinearities. Wright and Fornberg [9] described the derivative of a function at a given point depending on the neighborhood points like in finite difference method. That is the derivative of a function at a particular point is approximated by the linear combination of surrounding points. Chandini and Sanyasiraju (2006) applied this method for solving non-linear convection diffusion equation [10].

2.1 Augmented Radial Basis Function

Given a set of n distinct data points (r_j, θ_j) and corresponding data values $f_j, \quad j = 1, 2, \ldots, n$, the augmented RBF interpolant for axi-symmetric spherical polar coordinates is given by

$$s(r,\theta) = \sum_{j=1}^{n} \lambda_j \phi_j + \sum_{k=1}^{m} \alpha_k p_k(r,\theta) \tag{1}$$

where $\phi_j = \sqrt{1+\varepsilon^2\{(r\cos\theta - (r\cos\theta)_j)^2 + (r\sin\theta - (r\sin\theta)_j)^2\}}$, $\{p_k(r,\theta)\}_{k=1}^{m}$ is a basis for $\Pi_m(\mathbb{R}^d)$ (space of all d-variate polynomial with degree less than m) and $s(r_j,\theta_j) = f_j$. For solving the linear system m extra conditions are required. The extra conditions are chosen by taking the expansion coefficient vector $\lambda \in \mathbb{R}^n$ orthogonal to $\Pi_m(\mathbb{R}^d)$.

i.e.

$$\sum_{j=1}^{n} \lambda_j p_k(r_j,\theta_j) = 0, \quad k = 1,2,\ldots,M \tag{2}$$

To determine the expansion coefficient λ_j and α_k we solve the following symmetric linear system:

$$\left(\begin{array}{c|c} A & p \\ \hline p^T & 0 \end{array}\right)\left(\begin{array}{c} \lambda \\ \hline \alpha \end{array}\right) = \left(\begin{array}{c} \mathbf{f} \\ \hline \mathbf{0} \end{array}\right) \tag{3}$$

where A is the coefficient matrix with entries

$$a_{ij} = \sqrt{1+\varepsilon^2\{((r\cos\theta)_i - (r\cos\theta)_j)^2 + ((r\sin\theta)_i - (r\sin\theta)_j)^2\}},$$

$j = 1,2,\ldots,n$, $i = 1,2,\ldots,n$ and p is the $n \times M$ matrix with elements $p_k(r_j,\theta_j)$ for $j = 1,2,\ldots,n$ and $k = 1,2,\ldots,M$. We use Lagrange form of RBF interpolant to derive RBF-FD formulae. The interpolant is given by

$$s(r,\theta) = \sum_{j=1}^{n} \psi_j(r,\theta) u(r_j,\theta_j) \tag{4}$$

where $\psi_j(r,\theta)$ satisfies the cardinal conditions

$$\psi_j(r_k,\theta_k) = \delta_{jk} = \begin{cases} 1, & \text{if } j = k \\ 0, & \text{if } j \neq k \end{cases} \qquad k = 1,2,\ldots,n. \tag{5}$$

Closed form representation for $\psi_j(r,\theta)$ can be obtained by considering that the right hand side vector of (3) stems from each ψ_j's. Then by Cramer's rule on (3) to (4) gives

$$\psi_j(r,\theta) = \frac{det(A_j(r,\theta))}{det(A)} \tag{6}$$

where $A_j(r,\theta)$ is same as matrix A, except that the jth row is replaced by the vector

$$B(r,\theta) = [\phi_1\phi_2\ldots\phi_n | p_1(r,\theta)p_2(r,\theta)\ldots p_m(r,\theta)] \tag{7}$$

where ϕ_j are defined as above and $p_j(r,\theta) \in \Pi_m(\mathbb{R}^d)$

To approximate derivative of a function at a given point the derivation from (4) to (7) can be used. The linear differential operator of a function u at a given point (r_i,θ_i) is $l(u(r_i,\theta_i))$ and can be calculated using values of the function at

neighborhood points of (r_i, θ_i) (say n_i nodes (r_1, θ_1), (r_2, θ_2), $\ldots$, (r_{n_i}, θ_{n_i})). Then

$$l(u(r_i, \theta_i)) \approx \sum_{j=1}^{n_i} c_{ij} u(r_j, \theta_j). \tag{8}$$

By applying Lagrange RBF interpolation (4)

$$l(u(r_i, \theta_i)) \approx l(s(r_i, \theta_i)) = \sum_{j=1}^{n_i} l(\psi_j(r_i, \theta_i) u(r_j, \theta_j)). \tag{9}$$

From (8) and (9)

$$c_{ij} = l(\psi_j(r_i, \theta_i)), \quad j = 1, 2, \ldots, n_i.$$

The weights are computed by solving the linear system:

$$\left(\begin{array}{c|c} A & p \\ \hline p^T & \mathbf{0} \end{array}\right)_i \left(\frac{\mathbf{C}}{\mu}\right)_i = \left(\frac{(l(B(r,\theta)))^T}{\mathbf{0}}\right)_i$$

where A is the part of coefficient matrix of Eq. (3), $B(r, \theta)$ is the row vector in (7) and μ is a vector related to α in (1) and $C = [c_1, c_2, \ldots, c_{n_i}]'$. By using the values of C in (8) we will get an equation on $u(r_j, \theta_j)$, $j = 1, 2, \ldots, n_i$. These n_i points are some points from $u(r_i, \theta_i)$, which are nearer to the ith internal point. Clearly Eq. (2) gives $\sum_{j=1}^{n_i} c_{ij} = 0$, i is the internal points. i.e. sum of expansion coefficient is 0, like the traditional finite difference method.

2.2 Navier-Stokes Equations in Spherical Geometry

The flow of steady incompressible viscous flow past a sphere with uniform free-stream velocity U_∞ (from left to right) is considered for this study. The governing N-S equations expressed in stream function ψ and vorticity ω formulation in axi-symmetric spherical polar coordinates are

$$\frac{\partial^2 \psi}{\partial r^2} + \frac{1}{r^2} \frac{\partial^2 \psi}{\partial \theta^2} - \frac{\cot\theta}{r^2} \frac{\partial \psi}{\partial \theta} = -r\omega \sin\theta \tag{10}$$

and

$$\frac{\partial^2 \omega}{\partial r^2} + \frac{2}{r} \frac{\partial \omega}{\partial r} + \frac{1}{r^2} \frac{\partial^2 \omega}{\partial \theta^2} + \frac{\cot\theta}{r^2} \frac{\partial \omega}{\partial \theta} - \frac{\omega}{r^2 \sin^2\theta} = \frac{Re}{2} \left(q_r \frac{\partial \omega}{\partial r} + \omega \frac{\partial q_r}{\partial r} + \frac{q_r \omega}{r} + \frac{q_\theta}{r} \frac{\partial \omega}{\partial \theta} + \frac{\omega}{r} \frac{\partial q_\theta}{\partial \theta} \right). \tag{11}$$

Here Re is the Reynolds number defined as $Re = 2U_\infty a/\nu$, where a is radius of the sphere and ν is kinematic coefficient of viscosity. q_r and q_θ are the non-dimensional radial and transverse velocity components defined as

$$q_r = \frac{1}{r^2 \sin\theta} \frac{\partial \psi}{\partial \theta}, \qquad q_\theta = \frac{-1}{r \sin\theta} \frac{\partial \psi}{\partial r} \tag{12}$$

which are obtained by dividing the corresponding dimensional components by the stream velocity U_∞. They are chosen in such a way that the equation of continuity in spherical polar coordinates is satisfied.

The boundary conditions to be satisfied are:

- On the surface of the sphere ($r = 1$): $\psi = \frac{\partial \psi}{\partial r} = 0$, $\omega = -\frac{1}{\sin\theta}\frac{\partial^2 \psi}{\partial r^2}$
- At large distances from the sphere ($r \to \infty$): $\psi \to \frac{r^2}{2}\sin^2\theta$, $\omega \to 0$
- Along the axis of symmetry ($\theta = 0$ and $\theta = \pi$): $\psi = 0$, $\omega = 0$

The surface vorticity is calculated by using the procedure given in [11]

$$\omega_{1,j} = -\frac{3\psi_{2,j}}{\Delta r^2 sin\theta_j} - \frac{\omega_{2,j}}{2}$$

where Δr is the distance between the given boundary point (i.e. (1, j) points) and nearest neighborhood point (i.e. (2, j) points).

The first and second order derivatives of ψ, ω with respect to r and θ are calculated at ith point using Eqs. (4) to (8) as follows:

$$\frac{\partial \psi}{\partial r}|_{\psi=\psi_i} \approx \sum_{j=1}^{n_i} a_{ij}^r \psi(r_j, \theta_j), \qquad \frac{\partial \psi}{\partial \theta}|_{\psi=\psi_i} \approx \sum_{j=1}^{n_i} a_{ij}^\theta \psi(r_j, \theta_j),$$

$$\frac{\partial \omega}{\partial r}|_{\omega=\omega_i} \approx \sum_{j=1}^{n_i} b_{ij}^r \omega(r_j, \theta_j), \qquad \frac{\partial \omega}{\partial \theta}|_{\omega=\omega_i} \approx \sum_{j=1}^{n_i} b_{ij}^\theta \omega(r_j, \theta_j),$$

$$\frac{\partial^2 \psi}{\partial r^2}|_{\psi=\psi_i} \approx \sum_{j=1}^{n_i} a_{ij}^{rr} \psi(r_j, \theta_j), \qquad \frac{\partial^2 \psi}{\partial \theta^2}|_{\psi=\psi_i} \approx \sum_{j=1}^{n_i} a_{ij}^{\theta\theta} \psi(r_j, \theta_j),$$

$$\frac{\partial^2 \omega}{\partial r^2}|_{\omega=\omega_i} \approx \sum_{j=1}^{n_i} b_{ij}^{rr} \omega(r_j, \theta_j), \qquad \frac{\partial^2 \omega}{\partial \theta^2}|_{\omega=\omega_i} \approx \sum_{j=1}^{n_i} b_{ij}^{\theta\theta} \omega(r_j, \theta_j),$$

where a_{ij}^r, a_{ij}^θ, a_{ij}^{rr}, $a_{ij}^{\theta\theta}$, b_{ij}^r, b_{ij}^{rr}, b_{ij}^θ, $b_{ij}^{\theta\theta}$, are similar to c_{ij} in the Eq. (8).

We first solve the governing Eq. (10) for ψ by taking ω value from the previous iteration. Then we solve Eq. (11) for ω. Thus the non-linear terms like $q_r \frac{\partial \omega}{\partial r}$ in the Eq. (11) are locally linearized with known values of ψ.

The Eq. (10) is discretized at ith internal point as follows:

$$\sum_{j=1}^{n_i} (a_{ij}^{rr} + \frac{1}{r_i^2} a_{ij}^{\theta\theta} - \frac{\cot\theta_i}{r_i^2} a_{ij}^\theta) \psi(r_j, \theta_j) = r_i \omega_i \sin\theta_i.$$

As ψ_i is known now, we calculate

$$q_r = \frac{1}{r_i^2 \sin\theta_i} (\frac{\partial \psi}{\partial \theta})_{\psi=\psi_i} = \frac{1}{r_i^2 \sin\theta_i} \sum_{j=1}^{n_i} a_{ij}^\theta \psi(r_j, \theta_j) = g_i(\text{say}).$$

Similarly calculate

$$(\frac{\partial q_r}{\partial r})_{\psi=\psi_i} = d_i, (q_\theta)_{\psi=\psi_i} = \frac{-1}{r_i \sin\theta_i} (\frac{\partial \psi}{\partial r})_{\psi=\psi_i} = e_i \text{ and } (\frac{\partial q_\theta}{\partial \theta})_{\psi=\psi_i} = f_i.$$

Now Eq. (11) is discretized at *ith* internal point as follows:

$$\sum_{j=1}^{n_i}(b_{ij}^{rr}+\frac{2}{r_i}b_{ij}^{r}+\frac{1}{r_i^2}b_{ij}^{\theta\theta}+\frac{\cot\theta_i}{r_i^2}b_{ij}^{\theta}-\frac{1}{r_i^2\sin^2\theta_i})\omega(r_j,\theta_j)=$$
$$\frac{Re}{2}\sum_{j=1}^{n_i}(g_i b_{ij}^{r}+d_i+\frac{g_i}{r_i}+\frac{e_i}{r_i}b_{ij}^{\theta}+\frac{f_i}{r_i})\omega(r_j,\theta_j).$$

We finally get the following linear systems of equations for ψ and ω

$$D\psi = F_1 \tag{13}$$

and

$$E\omega = F_2 \tag{14}$$

where $D = [D_1 D_2 \ldots D_N]^{'}$ and each $D_i (i = 1, 2, \ldots, N)$ is a row vector for *ith* internal point and F_1 is the column matrix. Similarly E and F_2.

The system of linear Eqs. (13) and (14) so obtained is first solved for ψ at all internal nodes and then ω at all internal nodes using the Gauss-Seidel iterative method. This completes one iteration. The iterations are continued until the Root Mean Square(RMS) error of the dynamic residuals is less than 10^{-6}.

Upwind model supporting nodes: Upwind model supporting nodes is applied for convective terms to achieve the results at higher far fields and for high Reynolds numbers. All the other derivatives are approximated by central model supporting nodes. For convective terms, one nearest neighborhood point depending on the flow direction (radial or transverse) is chosen as supporting node. The choice of the node in the flow direction is explained below and shown in the Fig. 1 (bottom):

- $q_r < 0$, $\frac{\partial \omega}{\partial r}$ is approximated by using a forward point of reference point in radial direction.
- $q_r > 0$, $\frac{\partial \omega}{\partial r}$ is approximated by using a backward point of reference point in radial direction.
- $q_\theta < 0$, $\frac{\partial \omega}{\partial \theta}$ is approximated by using a forward point of reference point in angular direction.

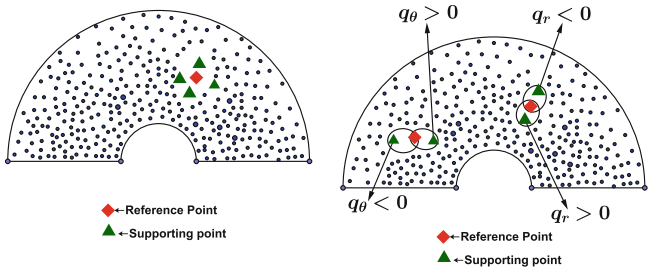


Fig. 1. Choice of local supporting nodes: (a) central model (top) and (b) upwind model (bottom)

– $q_\theta > 0$, $\frac{\partial \omega}{\partial \theta}$ is approximated by using a backward point of reference point in angular direction.

The convective terms in Eq. (11) are discretized by choosing the upwind type of nodes as described above. The modified discretization of the Eq. (11) at *ith* internal point is given below.

$$\sum_{j=1}^{n_i}(b_{ij}^{rr} + \frac{2}{r_i}b_{ij}^{r} + \frac{1}{r_i^2}b_{ij}^{\theta\theta} + \frac{\cot\theta_i}{r_i^2}b_{ij}^{\theta} - \frac{1}{r_i^2\sin^2\theta_i})\omega(r_j,\theta_j) =$$
$$\frac{Re}{2}(\sum_{j=1}^{m_i} g_i b_{ij}^{r} + \sum_{j=1}^{n_i} d_i + \sum_{j=1}^{n_i}\frac{g_i}{r_i} + \sum_{j=1}^{m_i}\frac{e_i}{r_i}b_{ij}^{\theta} + \sum_{j=1}^{n_i}\frac{f_i}{r_i})\omega(r_j,\theta_j)$$

where m_i are supporting nodes considered for the convection terms.

2.3 Accommodative FAS-FMG Multilevel Algorithm

Ghia et al. [2] developed accommodative version of the Full Approximation Scheme-Full MultiGrid (FAS-FMG) procedure of Brandt [3] and applied in finite difference method to Navier-Stokes equations. Here we consider various sets of levels for a fixed domain such that there is no pre specified connection between any two sets. This can be from a coarser level to a finer level i.e. with additional nodes than the previous set and finally the finest with some more additional nodes than the finer one, etc. i.e. $L^1, L^2, \ldots, L^m$ are the set of levels with increasing number of nodes in the same domain. Prolongation (P_{i-1}^{i}) is a operator which transfers a coarse level to a finer level solution. First we solved system of linear equations obtained in Eqs. (13) and (14) by using iterative scheme Gauss-Seidel in the coarsest level (L^1) until get convergent solutions i.e. $D^1\psi^1 = F_1^1$ and $E^1\omega^1 = F_2^1$. Then prolongate the known convergent solutions to next finer level by prolongation operator and interpolate the rest points by RBF-FD method. Repeat the procedure until get the convergent solution at finest level (L^m). The procedure as summarize as below:

1. Solve the algebraic system of linear equations $D\psi = F_1$ and $E\omega = F_2$, obtained by discretizing the governing equations using RBF-FD method, in the coarsest set of nodes (L^1) until convergence using iterative technique such as Gauss-Seidel.
2. Prolongate the coarsest set solution to the next finer set i.e. $\widehat{\psi}^2 = P(\psi^1)$ and $\widehat{\omega}^2 = P(\omega^1)$.
3. The solution at additional points can be obtained by RBF-FD interpolation. Using this as starting solution, achieve convergent solution in the finer set i.e. $D^2\psi^2 = F_1^2$ and $E^2\omega^2 = F_2^2$.
4. Repeat the above procedure for the next finer set and so on until the finest set and achieve convergent solution in the finest set.

3 Results and Discussion

The upwind model RBF-FD is used for the parameters in the range $Re = 10 - 200$ for various shape parameters and different far fields. The choice of shape

Table 1. Choice of ε for Re = 100 (upwind model) with different sets of nodes and its order of accuracy

Epsilon	65×65	76×76	97×97	113×113	129×129	141×141	151×151	Order
0.9	0.441081	0.463888	0.495871	0.509397	0.517249	0.520756	0.522337	$O(h^{2.2})$
1.0	0.467295	0.488748	0.518180	0.530440	0.537066	0.539834	0.541309	$O(h^3)$
1.1	0.491931	0.512595	0.540030	0.550902	0.557199	0.558764	0.559075	$O(h^2)$

parameter ε is also tested in comparison with finite difference model. The results are obtained from different scattered points such as $65^2, 76^2, 97^2, 113^2, 129^2, 141^2$ and 151^2 and presented in the Table 1 for different ε. To check the order of accuracy of the results in the absence of exact solution, the divided differences of the drag coefficient values $d(C_D)/dh$ for $Re = 100$ with various step sizes h of the data in Table 1 are plotted for $\varepsilon = 0.9, 1.0$ and 1.1. The decay of $d(C_D)/dh$ as function of h is presented on a log-log scale in the Fig. 2. Here, the value of 'h' in x-axis is taken as the average of step sizes of the grids corresponding to the divided differences. The slopes of the curves are parallel to the dotted lines of $O(h^{2.2}), O(h^3), O(h^2)$ respectively for $\varepsilon = 0.9, 1.0$ and 1.1. This shows that $d(C_D)/dh \to 0$ at the rate of $O(h^{2.2}), O(h^3)$ and $O(h^2)$ respectively. Hence the order of accuracy are respectively 3.2, 4 and 3. We choose $\varepsilon = 1.0$ for $Re = 100$.

The results for $Re = 100$ are tested with different far fields 30, 40 and 50 times the radius of the sphere to fix the artificial unbounded domain. For each far field, the shape parameter ε is chosen as explained above and the drag coefficient values are presented in the Table 2. From the table, we can observe that the far field of 40 times the radius of sphere is sufficient to get satisfactory results as the values are almost same with the other far fields of 30 and 50. The drag coefficient values which are obtained in a similar fashion for $Re = 10 - 200$ are tabulated in the Table 3 along with other literature values [12–18]. The drag coefficient values agree with literature values. The last column of the Table 3 shows the relative percentage error with respect to fourth order accurate based finite difference scheme [12]. The streamlines and vorticity lines are plotted for $Re = 100$ in Fig. 3 whose separation length and separation angle are found to be 3.68 and $58°.8'$ respectively. It is also found that the flow got separated initially at $Re = 20$. To the best of our knowledge, most of the numerical results available in the literature with regard to the model problem considered here are at the most second order accurate. The recent results presented for this problem in the reference [12] are fourth order accurate due to HOCS discretization. The

Table 2. Choice of far-field for Re = 100 and its order of accuracy

Far-field	Epsilon	65×65	76×76	97×97	113×113	129×129	141×141	151×151	Order
30	1.3	0.499162	0.516811	0.534433	0.540208	0.541791	0.541744	0.541015	$O(h^2)$
40	1.0	0.467295	0.488748	0.518180	0.530440	0.537066	0.539834	0.541309	$O(h^3)$
50	0.8	0.425276	0.452838	0.495113	0.514586	0.527217	0.533476	0.535483	$O(h^{1.5})$

Table 3. Comparison of drag coefficient results with other literature values for different Re

Re	Clair et al. [14] (1970)	Dennis and Walker [13] (1971)	Fornberg [15] (1988)	Juncu and Mihail [16] (1990)	Feng and Michaelides [17] (2000)	Atefi et al. [18] (2007)	Sekhar and Raju [12] (2012)	RBF-FD	Relative percentage error w.r.t. [12]
10	2.14	2.21	–	–	–	–	2.13	2.23	4.69
20	1.36	1.36	–	–	1.34	–	1.34	1.38	2.98
40	0.93	0.90	–	–	0.88	–	0.89	0.88	1.12
100	0.55	–	0.54	0.53	0.55	0.55	0.54	0.54	0.00
200	–	–	0.38	–	–	–	0.38	0.35	7.89

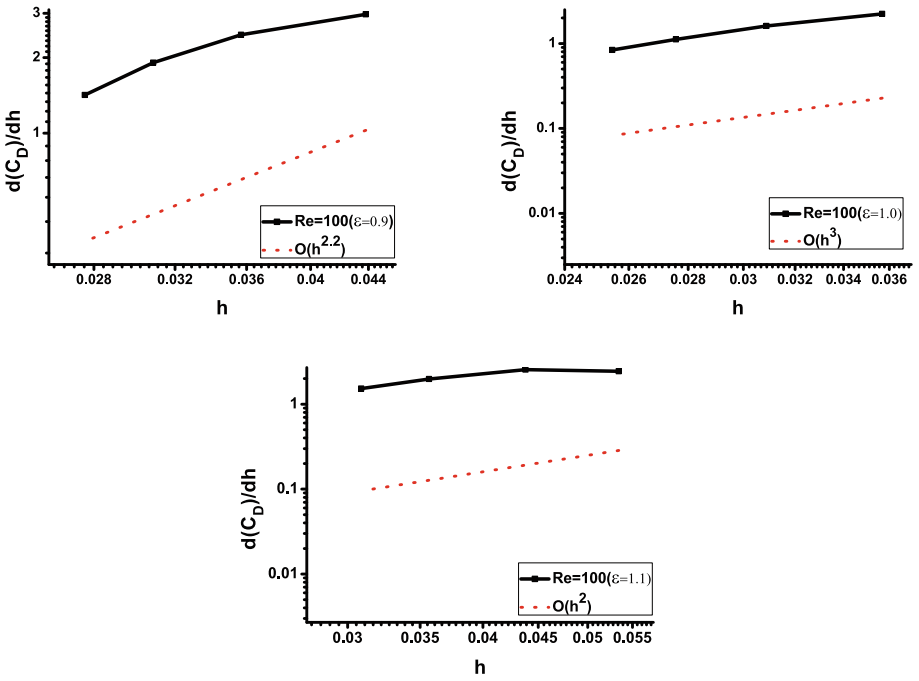


Fig. 2. Calculation of order of accuracy for $Re = 100$ with $\varepsilon = 0.9, 1.0$ and 1.1

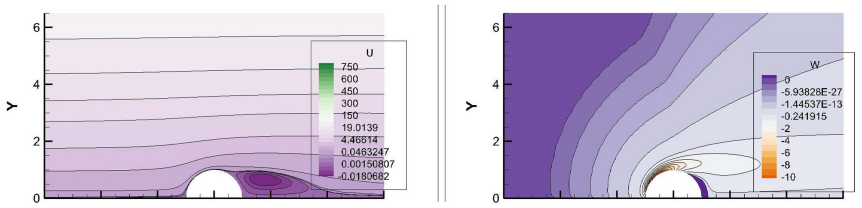


Fig. 3. Streamlines lines (top) and vorticity lines (bottom) for Re = 100

Table 4. Effect of efficient model that combined with RBF-FD in N-S equations for Re = 40

No. of steps	Finest nodes	Coarsest nodes	CPU time (mins)	%
1	129×129	129×129	13.17	–
2	129×129	65×65	8.72	33.79
3	129×129	33×33	8.64	34.4

present results with RBF-FD are also fourth order accurate and exhibits all the flow characteristics that match with experimental, theoretical and numerical results. This ensures that the RBF-FD scheme captures all flow characteristics particularly in unbounded flows, and the results are higher order accurate.

By applying the proposed efficient scheme the CPU times for $Re = 40$ with $\varepsilon = 0.55$ obtained for three sets of nodes are presented in Table 4. From the table it is clear that single set (129^2) number of nodes takes 13.17 min but if we apply two levels (i.e. 65^2 and 129^2) the same solution is coming with 8.72 min computation time. Similarly for three levels (33^2, 65^2 and 129^2) take 8.64 min thereby saving almost 34% of the CPU time when compared to the usual RBF-FD method with finest set of nodes while achieving the same level of accuracy.

4 Conclusions

An accommodative FAS-FMG multilevel augmented RBF-FD method is developed and implemented to incompressible spherical polar Navier-Stokes equations. The accommodative FAS-FMG multigrid analogy with local refinement is adopted to achieve the efficiency. The developed scheme saves almost 34% of the CPU time when compared to CPU time of the solution obtained from the finest set of nodes solely while achieving the same level of accuracy.

References

1. Hyman, J.M.: Mesh refinement and local inversion of elliptic partial differential equations. J. Comput. Phys. **23**, 124–134 (1977)
2. Ghia, U.N., Ghia, K.N., Shin, C.T.: High-Re solutions for incompressible flow using the Navier-Stokes equations and a multigrid method. J. Comput. Phys. **48**, 387–411 (1982)
3. Achi, B.: Multi-level adaptive solutions to boundary-value problems. Math. Comput. **31**, 333–390 (1977)
4. Ding, H., Shu, C., Yeo, K.S., Xu, D.: Simulation of incompressible viscous flows past a circular cylinder by hybrid FD scheme and meshless least square-based finite difference method. Comput. Methods Appl. Mech. Eng. **193**, 727–744 (2004)
5. Javed, A., Djidjeli, K., Xing, J.T., Cox, S.J.: A hybrid mesh free local RBF-Cartesian FD scheme for incompressible flow around solid bodies. Int. J. Math. Comput. Phys. Electr. Comput. Eng. **7**, 957–966 (2013)

6. Shu, C., Ding, H., Yeo, K.S.: Local radial basis function-based differential quadrature method and its application to solve two-dimensional incompressible Navier-Stokes equations. Comput. Methods Appl. Mech. Eng. **192**, 941–954 (2003)
7. Tolstykh, A.I., Lipavskii, M.V., Shirobokov, D.A.: High-accuracy discretization methods for solid mechanics. Arch. Mech. **55**, 531–553 (2003)
8. Cecil, T., Qian, J., Osher, S.: Numerical methods for high dimensional Hamilton-Jacobi equations using radial basis functions. J. Comput. Phys. **196**, 327–347 (2004)
9. Wright, G.B., Fornberg, B.: Scattered node compact finite difference-type formulas generated from radial basis functions. J. Comput. Phys. **212**, 99–123 (2006)
10. Chandhini, G., Sanyasiraju, Y.V.S.S.: Local RBF-FD solutions for steady convection-diffusion problems. Int. J. Numer. Methods Eng. **72**, 352–378 (2007)
11. Ray, R.K., Kalita, J.C.: A transformation-free HOC scheme for incompressible viscous flows on nonuniform polar grids. Int. J. Numer. Methods Fluids **62**, 683–708 (2010)
12. Sekhar, T.V.S., Raju, B.H.S.: An efficient higher order compact scheme to capture heat transfer solutions in spherical geometry. Comput. Phys. Commun. **183**, 2337–2345 (2012)
13. Dennis, S.C.R., Walker, J.D.A.: Calculation of the steady flow past a sphere at low and moderate Reynolds numbers. J. Fluid Mech. **48**, 771–789 (1971)
14. Le Clair, B.P., Hamielec, A.E., Pruppacher, H.R.: A numerical study of the drag on a sphere at low and intermediate Reynolds numbers. J. Atmos. Sci. **27**, 308–315 (1970)
15. Fornberg, B.: Steady viscous flow past a sphere at high Reynolds numbers. J. Fluid Mech. **190**, 471–489 (1988)
16. Juncu, Gh, Mihail, R.: Numerical solution of the steady incompressible Navier-Stokes equations for the flow past a sphere by a multigrid defect correction technique. Int. J. Numer. Methods Fluids **11**, 379–395 (1990)
17. Feng, Z., Michaelides, E.E.: A numerical study on the transient heat transfer from a sphere at high Reynolds and Peclet numbers. Int. J. Heat Mass Transfer **43**, 219–229 (2000)
18. Atefi, G.H., Niazmand, H., Meigounpoory, M.R.: Numerical analysis of 3-D flow past a stationary sphere with slip condition at low and moderate Reynolds numbers. J. Disper. Sci. Technol. **28**, 591–602 (2007)

Applied Mathematics

Bessel Sequences and Frames in Semi-inner Product Spaces

N.K. Sahu[1(✉)], C. Nahak[2], and Ram N. Mohapatra[3]

[1] Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar 382007, India
nabindaiict@gmail.com
[2] Department of Mathematics, Indian Institute of Technology Kharagpur, Kharagpur 721302, India
cnahak@maths.iitkgp.ernet.in
[3] Department of Mathematics, University of Central Florida, Orlando, FL 32816, USA
ramm@mail.ucf.edu

Abstract. In this paper, the concept of Bessel sequence and frame are introduced in semi-inner product spaces. Some properties of the Bessel sequences and frame are investigated in smooth uniformly convex Banach spaces. One characterization of the space of all Bessel sequences has been pointed out. Examples of frames are constructed in the real sequence spaces l^p, $1 < p < \infty$.

Keywords: Semi-inner product space · Uniformly convex smooth Banach space · Bessel sequence · Frame

1 Introduction and Preliminaries

With a view to establishing Hilbert space type arguments in Banach spaces, Lumer [19] constructed, a type of inner product called semi-inner product denoted by [.,.] with a more general axiom system. The corresponding space with this semi-inner product is called a semi-inner product space. Using the semi-inner product a norm can be defined by $\|x\| = [x, x]^{\frac{1}{2}}$. Lumer [19] showed that there are normed linear spaces, where the semi-inner product can be defined in many different ways. Subsequently, Giles [15] showed that in a fairly large class of Banach spaces it is possible to construct a semi-inner product with many of the desirable attributes of an inner product. He has shown that if X is a smooth uniformly convex Banach space, then it is possible to define a unique semi-inner product. Semi-inner product spaces have been studied by Lumer [19], Giles [15], Koehler [17] and Nanda [20].

Frame theory became popular only after 1990's. Now a days Frame theory, Wavelet analysis are rich areas of research due to their applications in signal processing, inverse-scattering problem, noise analysis and many other fields. Frame is an extension of the concept of a basis where this spanning set makes use of

© Springer Nature Singapore Pte Ltd. 2017
D. Giri et al. (Eds.): ICMC 2017, CCIS 655, pp. 155–169, 2017.
DOI: 10.1007/978-981-10-4642-1_14

its redundancy in applications. A great deal of work in frame theory has been done by Christensen [8–11]; Casazza and Christensen [5,6] and Favier and Zalik [13].

The main goal of this paper is to introduce the concept of Bessel sequence and frame in semi-inner product spaces. We first quote the following definitions:

Semi-inner product space [19]**:** Let X be a vector space over the field F of real or complex numbers. A functional $[.,.] : X \times X \to F$ is called a semi-inner product (s.i.p in short) if it satisfies the following:

1. $[x+y,z] = [x,z] + [y,z], \quad \forall x,y,z \in X;$
2. $[\lambda x, y] = \lambda[x,y], \quad \forall \lambda \in F$ and $x, y \in X;$
3. $[x,x] > 0$, for $x \neq 0;$
4. $|[x,y]|^2 \leqslant [x,x][y,y]$. The pair $(X,[.,.])$ is called a semi-inner product space.

Uniformly convex Banach space: A complete normed space X is uniformly convex if given $\epsilon > 0$, there exists $\delta(\epsilon) > 0$ such that, for $x, y \in X$ with $\|x\| = \|y\| = 1$ it holds that $\frac{\|x+y\|}{2} \leq 1 - \delta(\epsilon)$ when $\|x-y\| > \epsilon$.

Smooth Banach space: A Banach space X is said to be smooth if for any linearly independent elements x and y in X, the function $\psi(t) = \|x+ty\|$ is differentiable for all values of t.

Uniformly convex smooth Banach space: A Banach space which is uniformly convex and smooth is called a uniformly convex smooth Banach space.

Giles [15] has proved that if the underlying space is a uniformly convex smooth Banach space then it is possible to define a semi-inner product, uniquely. Also the unique semi-inner product has the following nice properties:

(i) $[x, \lambda y] = \overline{\lambda}[x,y]$ for all scalars λ.
(ii) $[x,y] = 0$ if and only if y is orthogonal to x, that is if and only if $\|y\| \leq \|y + \lambda x\|$, for all scalars λ.
(iii) Generalized Riesz representation theorem:- If f is a continuous linear functional on X then there is a unique vector $y \in X$ such that $f(x) = [x,y]$, for all $x \in X$.
(iv) The semi-inner product is continuous.

Example 1.1. The sequence space l^p, $p > 1$ and the functions space L^p, $p > 1$ are uniformly convex smooth Banach spaces. So one can define semi-inner product on these spaces, uniquely. Giles [15] has shown that the functions space L^p, $p > 1$ is a semi-inner product space with the semi-inner product defined by

$$[x,y] = \frac{1}{\|y\|_p^{p-2}} \int_X x|y|^{p-1} sgn(y) d\mu, \quad \forall\, x, y \in L^P(X,\mu).$$

Similarly the real sequence space l^p, $p > 1$ is a semi-inner product space with the semi-inner product defined by

$$[x,y] = \frac{1}{\|y\|_p^{p-2}} \sum_i |y(i)|^{p-2} y(i) x(i), \ \forall\, x = \{x(i)\}, \ y = \{y(i)\} \in l^p.$$

Bessel sequence and frame in Hilbert space are defined as follows:

Bessel sequence: A set of elements $\{f_i\}_1^\infty$ in a Hilbert space H is called a Bessel sequence if there exists a constant $B > 0$ such that
$\sum_{i=1}^{\infty} |\langle f, f_i\rangle|^2 \leq B \parallel f \parallel^2$, for all $f \in H$, where $\langle .,. \rangle$ is an inner product in H.

Frame: A family of elements $\{f_i\}_{i\in I} \subseteq H$ is called a frame for the Hilbert space H if there exist constants $A, B > 0$ such that

$$A \parallel f \parallel^2 \leq \sum_{i\in I} |\langle f, f_i\rangle|^2 \leq B \parallel f \parallel^2, \text{ for all } f \in H.$$

Frames in L^p spaces and other Banach function spaces are effective tools for modeling a variety of natural signals and images. They are also used in the numerical computation of integral and differential equations. There is plethora of literature available for frames in Banach spaces also. For classical frame theory in Banach spaces one may refer to Christensen and Heil [12], Grochenig [16]. Frames for shift invariant subspaces of L^p space are studied by Aldroubi et al. [1] in 2001. Casazza et al. [7] in 2005, characterized Banach frames in separable Banach spaces, and related them to series expansion in Banach spaces. M. Fornasier [14] studied the Banach frames and atomic decomposition characterization of α-modulation spaces in 2007. (p,Y)-Bessel operator sequences, (p,Y)-operator frames, and (p,Y)-Riesz bases for a Banach space X, are introduced and discussed by Cao et al. [3] in 2008. Liu [18] studied Schauder frames in Banach spaces in 2010. Schauder frame is a concept which is a natural generalization of frames in Hilbert spaces and Schauder bases in Banach spaces. Carando et al. [4] in 2011, discussed the reconstruction formula of Banach frames for the functions space L^p, $(1 \leq p < \infty)$ and Lorentz space $L^{p,q}$, $(1 \leq p, q < \infty)$ with respect to a solid sequence space.

The frames in Banach spaces using semi-inner product was defined by H. Zhang and J. Zhang [22] in 2011. They generalized the classical theory on frames and Riesz bases under this new perspective. They also established the Shannon sampling theorem in Banach spaces using semi-inner product structure.

In our work the concept of Bessel sequence and frame are introduced in some semi-inner product spaces, which are uniformly convex smooth Banach spaces with homogeneity property. Properties of these Bessel sequence and frame have been studied.

2 Bessel Sequence

We define Bessel sequence on a uniformly convex smooth Banach space consisting of norm $\|.\|_p$, $1 < p < \infty$. We consider our Banach space as a semi-inner product space and use the semi-inner product to define Bessel sequence in this class of Banach spaces. For the rest of the paper we assume that X is a real uniformly convex smooth Banach space with norm $\|.\|_p$ and semi-inner product $[.,.]$. We denote semi-inner product on the real sequence space l^q by $[.,.]_q$ and $\|.\|_q$.

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

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

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

We now prove some preliminary results for the existence of Bessel sequence in a uniformly convex smooth Banach space.

Lemma 2.1. Let X be a real smooth uniformly convex Banach space with $\|.\|_p$. For some sequence $y = \{y_i\}_{i=1}^{\infty} \subseteq X$ and some element $x \in X$, suppose that the series $\sum_{i=1}^{\infty} c_i y_i \frac{|[y_i, x]|^{q-2}}{\|\{[y_i, x]\}\|^{q-2}}$ is convergent for all $c = \{c_i\}_{i=1}^{\infty} \in l^q$. Also assume that $\left\{c_i \frac{|[y_i, x]|^{q-2}}{\|\{[y_i, x]\}\|^{q-2}}\right\}_{i=1}^{\infty} \in l^q$. Then the mapping $T : l^q \to X$, defined by

$$T(c) = \sum_{i=1}^{\infty} c_i y_i \frac{|[y_i, x]|^{q-2}}{\|\{[y_i, x]\}\|^{q-2}}$$

is a bounded linear operator. The generalized adjoint operator of T is $T^{\dagger} : X \to l^q$ given by $T^{\dagger}x = \{[y_i, x]\}_{i=1}^{\infty}$.

Proof. Consider the sequence of bounded linear operators $T_n : l^q \to X$ defined by

$$T_n(c) = \sum_{i=1}^{n} c_i y_i \frac{|[y_i, x]|^{q-2}}{\|\{[y_i, x]\}\|^{q-2}}.$$

One can easily see that $T_n \to T$ pointwise. Hence T is bounded. Also T is linear. For $x \in X$ and $c = \{c_i\}_{i=1}^{\infty} \in l^q$,

$$\begin{aligned}
[T(c), x] &= [\sum_{i=1}^{\infty} c_i y_i \frac{|[y_i, x]|^{q-2}}{\|\{[y_i, x]\}\|^{q-2}}, x] \\
&= \sum_{i=1}^{\infty} c_i \frac{|[y_i, x]|^{q-2}}{\|\{[y_i, x]\}\|^{q-2}} [y_i, x] \\
&= [\{c_i\}_{i=1}^{\infty}, \{[y_i, x]\}_{i=1}^{\infty}]_q \\
\Rightarrow T^{\dagger}x &= \{[y_i, x]\}_{i=1}^{\infty}.
\end{aligned}$$

Remark 2.1. Since $T : l^q \to X$ is a bounded linear operator, then $T^{\dagger} : X \to l^q$ is bounded on X and it holds that $\|T^{\dagger}(x)\|_q \leq \|T\|\|x\|_p$, for all $x \in X$ (see Pap and Pavlovic [21]). Hence

$$\begin{aligned}
(\|\{[y_i, x]\}\|_q)^q &\leq \|T\|^q (\|x\|_p)^q \\
\Rightarrow \sum_{i=1}^{\infty} |[y_i, x]|^q &\leq \|T\|^q (\|x\|_p)^q.
\end{aligned}$$

Hence the Bessel sequence on the semi-inner product space X is well defined.

We now obtain the following results for Bessel sequences.

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

Proof. Firstly, suppose that $y = \{y_i\}_{i=1}^{\infty}$ is a Bessel sequence with bound B.

Let $c = \{c_i\}_{i=1}^{\infty} \in l^q$. We have to show that $T\{c_i\}_{i=1}^{\infty}$ is well defined, that is $\sum_{i=1}^{\infty} c_i y_i \frac{|[y_i,x]|^{q-2}}{\|\{[y_i,x]\}\|^{q-2}}$ is convergent.

Let $n, m \in N$ and $n > m$. Then

$$\Big\| \sum_{i=1}^{n} c_i y_i \frac{|[y_i,x]|^{q-2}}{\|\{[y_i,x]\}\|^{q-2}} - \sum_{i=1}^{m} c_i y_i \frac{|[y_i,x]|^{q-2}}{\|\{[y_i,x]\}\|^{q-2}} \Big\|$$

$$= \Big\| \sum_{i=m+1}^{n} c_i y_i \frac{|[y_i,x]|^{q-2}}{\|\{[y_i,x]\}\|^{q-2}} \Big\|$$

$$= \sup_{\|z\|=1} \Big| \Big[\sum_{i=m+1}^{n} c_i y_i \frac{|[y_i,x]|^{q-2}}{\|\{[y_i,x]\}\|^{q-2}}, z \Big] \Big|$$

$$\leq \sup_{\|z\|=1} \sum_{i=m+1}^{n} |c_i| \frac{|[y_i,x]|^{q-2}}{\|\{[y_i,x]\}\|^{q-2}} |[y_i,z]|$$

$$\leq \Big(\sum_{i=m+1}^{n} \big(|c_i| \frac{|[y_i,x]|^{q-2}}{\|\{[y_i,x]\}\|^{q-2}}\big)^q \Big)^{\frac{1}{q}} \sup_{\|z\|=1} \Big(\sum_{i=m+1}^{n} |[y_i,z]|^p \Big)^{\frac{1}{p}}$$

$$\leq \Big(\sum_{i=m+1}^{n} \big(|c_i| \frac{|[y_i,x]|^{q-2}}{\|\{[y_i,x]\}\|^{q-2}}\big)^q \Big)^{\frac{1}{q}} \sup_{\|z\|=1} B^{\frac{1}{p}} \|z\|$$

$$= \Big(\sum_{i=m+1}^{n} \big(|c_i| \frac{|[y_i,x]|^{q-2}}{\|\{[y_i,x]\}\|^{q-2}}\big)^q \Big)^{\frac{1}{q}} B^{\frac{1}{p}}.$$

The right hand side goes to 0 as $n, m \to \infty$, since $\Big\{c_i \frac{|[y_i,x]|^{q-2}}{\|\{[y_i,x]\}\|^{q-2}}\Big\}_{i=1}^{\infty} \in l^q$ and $\sum_{i=1}^{n} |c_i| \frac{|[y_i,x]|^{q-2}}{\|\{[y_i,x]\}\|^{q-2}}$, for $n \in \mathbb{N}$, is a Cauchy sequence. Therefore $\Big\{\sum_{i=1}^{n} c_i y_i \frac{|[y_i,x]|^{q-2}}{\|\{[y_i,x]\}\|^{q-2}}\Big\}$, $n \in \mathbb{N}$, is a Cauchy sequence in X and is convergent since X is complete.

$$\sum_{i=1}^{n} c_i y_i \frac{|[y_i,x]|^{q-2}}{\|\{[y_i,x]\}\|^{q-2}} \rightarrow \sum_{i=1}^{\infty} c_i y_i \frac{|[y_i,x]|^{q-2}}{\|\{[y_i,x]\}\|^{q-2}} \text{ as } n \rightarrow \infty.$$

This implies that $T\{c_i\}_{i=1}^{\infty}$ is well defined and T is bounded.

Conversely, suppose that T is well defined and $\|T\| \leq C$, where C is any positive constant.

We have $(\|T^{\dagger}(x)\|_q)^q \leq \|T\|^q(\|x\|_p)^q$, for all $x \in X$, where $\frac{1}{p} + \frac{1}{q} = 1$ (see Pap and Pavlovic [21]).

$\Rightarrow \sum_{i=1}^{\infty} |[y_i,x]|^q \leq C^q(\|x\|_p)^q$, for all $x \in X$ and thus $\{f_i\}_{i=1}^{\infty}$ is a Bessel sequence.

We now prove a stability result for Bessel sequences.

Theorem 2. Let $y = \{y_i\}_{i=1}^{\infty}$ be a Bessel sequence in a uniformly convex smooth Banach space X. Suppose that the operator $T : l^q \rightarrow X$, defined by $T\{c_i\}_{i=1}^{\infty} = \sum_{i=1}^{\infty} c_i y_i \frac{|[y_i,x]|^{q-2}}{\|\{[y_i,x]\}\|^{q-2}}$, satisfies $\|T\| \leq M$, where M is a positive real constant. Let $g = \{g_i\}_{i=1}^{\infty}$ be another sequence in X, and assume that there exist constants $\lambda, \mu \geq 0$ such that

$$\begin{aligned} &\Big\| \sum_{i=1}^{n} c_i y_i \frac{|[y_i,x]|^{q-2}}{\|\{[y_i,x]\}\|^{q-2}} - \sum_{i=1}^{n} c_i g_i \frac{|[g_i,w]|^{q-2}}{\|\{[g_i,w]\}\|^{q-2}} \Big\| \\ &\leq \lambda \Big\| \sum_{i=1}^{n} c_i y_i \frac{|[y_i,x]|^{q-2}}{\|\{[y_i,x]\}\|^{q-2}} \Big\| + \mu \Big(\sum_{i=1}^{n} |c_i|^q\Big)^{\frac{1}{q}}, \forall \text{ scalars } \{c_n\}, n \in \mathbb{N}. \end{aligned} \tag{1}$$

Then $\{g_i\}_{i=1}^{\infty}$ is a Bessel sequence with bound $[(1+\lambda)M + \mu]^K$, where $K = K(p,q)$.

Proof. Since $y = \{y_i\}_{i=1}^{\infty}$ is a Bessel sequence, the operator $T : l^q \rightarrow X$ defined by $T\{c_i\}_{i=1}^{\infty} = \sum_{i=1}^{\infty} c_i y_i \frac{|[y_i,x]|^{q-2}}{\|\{[y_i,x]\}\|^{q-2}}$ is well defined, bounded operator and $\|T\| \leq M$. From inequality (1), we have

$$\begin{aligned} &\Big\| \sum_{i=1}^{n} c_i g_i \frac{|[g_i,w]|^{q-2}}{\|\{[g_i,w]\}\|^{q-2}} \Big\| - \Big\| \sum_{i=1}^{n} c_i y_i \frac{|[y_i,x]|^{q-2}}{\|\{[y_i,x]\}\|^{q-2}} \Big\| \\ &\leq \Big\| \sum_{i=1}^{n} c_i y_i \frac{|[y_i,x]|^{q-2}}{\|\{[y_i,x]\}\|^{q-2}} - \sum_{i=1}^{n} c_i g_i \frac{|[g_i,w]|^{q-2}}{\|\{[g_i,w]\}\|^{q-2}} \Big\| \\ &\leq \lambda \Big\| \sum_{i=1}^{n} c_i y_i \frac{|[y_i,x]|^{q-2}}{\|\{[y_i,x]\}\|^{q-2}} \Big\| + \mu \Big(\sum_{i=1}^{n} |c_i|^q\Big)^{\frac{1}{q}}. \\ \Rightarrow &\Big\| \sum_{i=1}^{n} c_i g_i \frac{|[g_i,w]|^{q-2}}{\|\{[g_i,w]\}\|^{q-2}} \Big\| \\ &\leq (1+\lambda) \Big\| \sum_{i=1}^{n} c_i y_i \frac{|[y_i,x]|^{q-2}}{\|\{[y_i,x]\}\|^{q-2}} \Big\| + \mu \Big(\sum_{i=1}^{n} |c_i|^q\Big)^{\frac{1}{q}}. \end{aligned} \tag{2}$$

The above inequality holds for all scalar sequences $c = \{c_n\}, \quad n \in \mathbb{N}$.

Now define a bounded linear operator $U : l^q \to X$ by $U(c) = \sum_{i=1}^{\infty} c_i g_i \frac{|[g_i, w]|^{q-2}}{\|\{[g_i, w]\}\|^{q-2}}$.

Clearly U is linear. Now from (2), as $n \to \infty$

$$\begin{aligned}\|U(c)\| &\leq (1+\lambda)\|T(c)\| + \mu\|c\| \\ &\leq (1+\lambda)\|T\|\|c\| + \mu\|c\| \\ &\leq [(1+\lambda)M + \mu]\|c\|, \quad \forall \ \ c = \{c_i\}_{i=1}^{\infty} \in l^q.\end{aligned}$$

Hence U is a well defined operator from l^q into X and $\|U\| \leq (1+\lambda)M + \mu$. Now from Theorem 1, it is concluded that $g = \{g_i\}_{i=1}^{\infty}$ is a Bessel sequence with Bessel bound $[(1+\lambda)M + \mu]^q$.

Next we prove that the set of all Bessel sequences in a uniformly convex smooth Banach space is a linear space.

Theorem 3. Let X be a uniformly convex smooth Banach space and B_X be the set of all Bessel sequences in X. Then B_X is a linear space.

Proof. Let $y = \{y_k\}_{k=1}^{\infty}$ and $z = \{z_k\}_{k=1}^{\infty}$ be two Bessel sequences with bounds B_1 and B_2, respectively. We show that the set $\alpha y + \beta z = \{\alpha y_k + \beta z_k\}_{k=1}^{\infty}$ where $\alpha, \beta \in C$, is also a Bessel sequence.

$$\begin{aligned}(\sum_{k=1}^{\infty} |[\alpha y_k + \beta z_k, x]|^q)^{\frac{1}{q}} &= (\sum_{k=1}^{\infty} |\alpha[y_k, x] + \beta[z_k, x]|^q)^{\frac{1}{q}} \\ &\leq (\sum_{k=1}^{\infty} |\alpha|^q |[y_k, x]^q|)^{\frac{1}{q}} + (\sum_{k=1}^{\infty} |\beta|^q |[z_k, x]^q|)^{\frac{1}{q}} \\ &= |\alpha|(\sum_{k=1}^{\infty} |[y_k, x]^q|)^{\frac{1}{q}} + |\beta|(\sum_{k=1}^{\infty} |[z_k, x]^q|)^{\frac{1}{q}} \\ &\leq |\alpha|(B_1(\|x\|_p)^q)^{\frac{1}{q}} + |\beta|(B_2(\|x\|_p)^q)^{\frac{1}{q}} \\ &= (|\alpha|B_1^{\frac{1}{q}} + |\beta|B_2^{\frac{1}{q}})(\|x\|_p).\end{aligned}$$

$\Rightarrow \sum_{k=1}^{\infty} |[\alpha y_k + \beta z_k, x]|^q \leq (|\alpha|B_1^{\frac{1}{q}} + |\beta|B_2^{\frac{1}{q}})^q (\|x\|_p)^q$.

Hence $\alpha y + \beta z$ is also a Bessel sequence with bound $(|\alpha|B_1^{\frac{1}{q}} + |\beta|B_2^{\frac{1}{q}})^q$, and consequently, B_X is a linear space.

Our next four theorems will show that the set of all Bessel sequences B_X in a uniformly convex smooth Banach space X is a Banach space and it is a BK-space as well as an AK-space (for definitions of BK-space and AK-property see Boos [2], Chap. 7).

Theorem 4. B_X is a normed linear space with the norm $\|y\|_{B_X} = \sup_{\|x\|_p \leq 1} (\sum_{k=1}^{\infty} |[y_k, x]|^q)^{\frac{1}{q}}$, for $y = \{y_k\}_{k=1}^{\infty} \in B_X$ and $x \in X$.

Proof. Clearly $\|.\|_{B_X} : B_X \to R$. Now let $y = \{y_i\}_{i=1}^{\infty}$ and $z = \{z_i\}_{i=1}^{\infty}$ be in B_X.

$$(i)\ \|y\|_{B_X} = \sup_{\|x\|_p \le 1} \Big(\sum_{k=1}^{\infty} |[y_k, x]|^q\Big)^{\frac{1}{q}} \ge 0$$

$$\begin{aligned}(ii)\ \|\alpha y\|_{B_X} &= \sup_{\|x\|_p \le 1} \Big(\sum_{k=1}^{\infty} |[\alpha y_k, x]|^q\Big)^{\frac{1}{q}} \\ &= \sup_{\|x\|_p \le 1} \Big(\sum_{k=1}^{\infty} |\alpha|^q |[y_k, x]|^q\Big)^{\frac{1}{q}} \\ &= \sup_{\|x\|_p \le 1} |\alpha| \Big(\sum_{k=1}^{\infty} |[y_k, x]|^q\Big)^{\frac{1}{q}} = |\alpha| \|y\|_{B_X}.\end{aligned}$$

$$\begin{aligned}(iii)\ \|y+z\|_{B_X} &= \sup_{\|x\|_p \le 1} \Big(\sum_{k=1}^{\infty} |[y_k + z_k, x]|^q\Big)^{\frac{1}{q}} \\ &\le \sup_{\|x\|_p \le 1} \Big(\sum_{k=1}^{\infty} \{|[y_k, x] + [z_k, x]|\}^q\Big)^{\frac{1}{q}} \\ &\le \sup_{\|x\|_p \le 1} \Big\{\Big(\sum_{k=1}^{\infty} |[y_k, x]|^q\Big)^{\frac{1}{q}} + \Big(\sum_{k=1}^{\infty} |[z_k, x]|^q\Big)^{\frac{1}{q}}\Big\} \\ &\le \sup_{\|x\|_p \le 1} \Big(\sum_{k=1}^{\infty} |[y_k, x]|^q\Big)^{\frac{1}{q}} + \sup_{\|x\|_p \le 1} \Big(\sum_{k=1}^{\infty} |[z_k, x]|^q\Big)^{\frac{1}{q}} = \|y\|_{B_X} + \|z\|_{B_X}.\end{aligned}$$

(iv) Also $\|y\|_{B_X} = \sup_{\|x\|_p \le 1} \Big(\sum_{k=1}^{\infty} |[y_k, x]|^q\Big)^{\frac{1}{q}} = 0$ if and only if $y = \{y_k\}_{k=1}^{\infty} = 0$.
Hence B_X is a normed linear space.

Theorem 5. The set of all Bessel sequences B_X in a uniformly convex smooth Banach space X is a Banach space.

Proof. Theorem 4 shows that B_X is a normed linear space with respect to the norm $\|y\|_{B_X} = \sup_{\|x\|_p \le 1} \Big(\sum_{k=1}^{\infty} |[y_k, x]|^q\Big)^{\frac{1}{q}}$, for $y = \{y_k\}_{k=1}^{\infty} \in B_X$ and $x \in X$. We prove that B_X is complete in the above norm.

Let $\{y_n\}$ be a Cauchy sequence in B_X, where $y_n = \{y_{n_k}\}$ and $x \in X$. For $n, m \in \mathbb{N},\ \ n > m$, $\|y_n - y_m\|_{B_X} \to 0$ as $n, m \to \infty$. This implies that

$$\sup_{\|x\|_p\leq 1}(\sum_{k=1}^{\infty}|[y_{n_k}-y_{m_k},x]|^q)^{\frac{1}{q}}\to 0 \text{ as } n,m\to\infty.$$
$$\Rightarrow \sup_{\|x\|_p\leq 1}\sum_{k=1}^{\infty}|[y_{n_k}-y_{m_k},x]|^q\to 0 \text{ as } n,m\to\infty.$$
$$\Rightarrow \sum_{k=1}^{\infty}|[y_{n_k},x]-[y_{m_k},x]|^q\to 0 \text{ as } n,m\to\infty.$$
$$\Rightarrow |[y_{n_k},x]-[y_{m_k},x]|\to 0 \text{ as } n,m\to\infty.$$

Hence, we see that $\{[y_{n_k},x]\}$ is a Cauchy sequence in $\mathbb{C}$. $\mathbb{C}$ is complete. Hence $\{[y_{n_k},x]\}\to[y_k,x]\in\mathbb{C}$, where $y_k=\lim_{n\to\infty}y_{n_k}$.

Now for $y=\{y_k\}_{k=1}^{\infty}$, we have

$$\begin{aligned}\|y_n-y\|_{B_X} &= \sup_{\|x\|_p\leq 1}(\sum_{k=1}^{\infty}|[y_{n_k}-y_k,x]|^q)^{\frac{1}{q}}\\ &= \sup_{\|x\|_p\leq 1}(\sum_{k=1}^{\infty}|[y_{n_k},x]-[y_k,x]|^q)^{\frac{1}{q}}.\end{aligned}$$

The right hand side of the above equation goes to 0 as $n\to\infty$ because $[y_{n_k},x]\to[y_k,x]$ as $n\to\infty$.

Next we show that $y\in B_X$. That is to show that $\sum_{k=1}^{\infty}|[y_k,x]|^q\leq B(\|x\|_p)^q, \quad \forall x\in X$. Let B_n be the corresponding Bessel bounds for the Bessel sequences y_n. Also let $B=\sup_n B_n<\infty$. Now

$$\begin{aligned}\sum_{k=1}^{\infty}|[y_k,x]|^q &= \sum_{k=1}^{\infty}|\lim_{n\to\infty}[y_{n_k},x]|^q\\ &= \sum_{k=1}^{\infty}\lim_{n\to\infty}|[y_{n_k},x]|^q\\ &= \lim_{n\to\infty}\sum_{k=1}^{\infty}|[y_{n_k},x]|^q\\ &\leq B(\|x\|_p)^q, \quad \forall x\in X.\end{aligned}$$

This shows that $y\in B_X$ and hence B_X is a Banach space.

Next, we show that B_X has two important properties as a sequence space. We require the following definitions for that purpose.

BK-space: A coordinate space Y is called a BK-space if it is a Banach space and the linear functionals defined by $f_i(y)=y_i$, for each $i\in I$ are continuous, where I is the index set and $y=\{y_i\}_{i\in I}\in Y$.

AK-space: Let Y be a BK-space and $y=\{y_i\}_{i\in I}$ be a sequence in Y. Let $y^{[n]}=(y_1,y_2,...,y_n,0,0,..)$ be the nth section of the vector y. Then Y is called an AK-space if $\lim_{n\to\infty}\|y^{[n]}-y\|_Y=0$, for all $y\in Y$.

For more on FK, BK and AK spaces, please see [2, Chap. 7]. Our next two results show that B_X is a BK-space with AK-property.

Theorem 6. B_X is a BK-space.

Proof. Let $\{y_n\}$ be a sequence in B_X and $y_n \to y \in B_X$ as $n \to \infty$.

$$\Rightarrow \|y_n - y\|_{B_X} \to 0 \quad as \quad n \to \infty$$
$$\Rightarrow \sup_{\|x\|_p \leq 1} (\sum_{k=1}^{\infty} |[y_{n_k} - y_k, x]|^q)^{\frac{1}{q}} \to 0 \quad as \quad n \to \infty$$
$$\Rightarrow (\sum_{k=1}^{\infty} |[y_{n_k} - y_k, x]|^q)^{\frac{1}{q}} \to 0 \quad as \quad n \to \infty$$
$$\Rightarrow |[y_{n_k} - y_k, x]| \to 0 \quad as \quad n \to \infty \quad and \quad \forall x \in X \quad with \quad \|x\|_p \leq 1$$
$$\Rightarrow y_{n_k} \to y_k \quad as \quad n \to \infty.$$

This implies that B_X is a BK-space.

Theorem 7. B_X is an AK-space.

Proof. Let $y = (y_1, y_2, y_3, ...) \in B_X$. Assume that $z = (y_1, y_2, ..., y_n, 0, 0, ...)$. We can see that z is also in B_X. Now for $x \in X$, we have

$$\|y - z\|_{B_X} = \sup_{\|x\|_p \leq 1} (\sum_{i=n+1}^{\infty} |[y_i, x]|^q)^{\frac{1}{q}}. \tag{3}$$

Since the series $\sum_{i=1}^{\infty} |[y_i, x]|^q$ is convergent, the remainder term $\sum_{i=n+1}^{\infty} |[y_i, x]|^q \to 0 \quad as \quad n \to \infty$. Therefore the right hand side of (3) goes to 0 as $n \to \infty$. Consequently $\|y - z\|_{B_X} \to 0$ as $n \to \infty$. This proves that B_X is an AK-space.

We have shown that the collection of all Bessel sequences form a BK-space with AK-property, it is possible to infer many of the benefits of being such a space (see [2]). It is natural to ask if we can obtain the topological and Köthe-Toeplitz duals of this sequence space. We do not have any answers at this point of time.

3 Frame

Definition 3.1. 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

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

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.

Since a frame $y = \{y_i\}_{i=1}^{\infty}$ is a Bessel sequence, the operator $T : l^q \to X$ defined by

$$T(c) = \sum_{i=1}^{\infty} c_i y_i \frac{|[y_i, x]|^{q-2}}{\|\{[y_i, x]\}\|^{q-2}}$$

is bounded and linear. Because of Lemma 2.1, the generalized adjoint operator of T is $T^{\dagger} : X \to l^q$, defined by $T^{\dagger}x = \{[y_i, x]\}_{i=1}^{\infty}$.

Remark 3.1. Taking the composition of the two operators T and $T^{\dagger}$ defined in Lemma 2.1, we get a new operator S, which is called as frame operator. The frame operator $S : X \to X$ is defined as

$$S(x) = TT^{\dagger}(x) = \sum_{i=1}^{\infty} \frac{|[y_i, x]|^{q-2}}{\|\{[y_i, x]\}\|^{q-2}} [y_i, x] y_i.$$

If X is a real semi-inner product space, then one can easily calculate that

$$[Sx, x]^{\frac{q}{2}} = \sum_{i=1}^{\infty} |[y_i, x]|^q.$$

Now we have

$$\begin{aligned} \|Sx\| = \|TT^{\dagger}x\| &\leq \|T\|\|T^{\dagger}x\| \\ &\leq \|T\|\|T\|\|x\| = \|T\|^2\|x\|. \end{aligned}$$

Hence S is bounded.

Therefore the frame operator S is a positive and bounded operator. One can easily see that S is a nonlinear operator. Hence we can not use the usual methods of Hilbert space frame theory to obtain the inverse frame operator and the reconstruction formula.

Orthogonal set: A vector x is said to be orthogonal to a vector y in a Banach space Y in the sense of semi-inner product, if $[x, y]_Y = 0$, where $[.,.]_Y$ is semi-inner product in Y. If each vector is orthogonal to all other vectors in Y in the sense of semi-inner product then Y is said to be an orthogonal set.

We now prove the following results for frames in X.

Theorem 8. Let $\{y_i\}_{i=1}^{\infty}$ be a parseval frame in a uniformly convex smooth Banach space X. Suppose that $\|y_i\|_p = 1$, for all i. Then $\{y_i\}_{i=1}^{\infty}$ is an orthonormal set in the sense of semi-inner product.

Proof. Given $\|y_i\|_p = 1$ for all i. It is to prove that $[y_i, y_j] = \delta_{ij}$, where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise.

Choose some y_k arbitrarily. Now

$$(\|y_k\|_p)^q = \sum_{i=1}^{\infty} |[y_i, y_k]|^q = (\|y_k\|_p)^{2q} + \sum_{i \neq k} |[y_i, y_k]|^q$$

$\Rightarrow 1 = 1 + \sum_{i \neq k} |[y_i, y_k]|^q$
$\Rightarrow \sum_{i \neq k} |[y_i, y_k]|^q = 0$
$\Rightarrow [y_i, y_k] = \delta_{ik}$.

Continuous semi-inner product: A semi-inner product is said to be a continuous semi-inner product if it is continuous in the second argument. Note that, a semi-inner product is automatically continuous in the first argument because of the linearity property in the first argument.

In the following theorem we assume that X is a real uniformly convex smooth Banach space with a continuous semi-inner product.

Theorem 9. Let $y = \{y_i\}_{i=1}^{\infty}$ be a sequence of elements in X. Suppose that there exist constants $A, B > 0$ such that $A(\|x\|_p)^q \leq \sum_{i=1}^{\infty} |[y_i, x]|^q \leq B(\|x\|_p)^q$ for all x in a dense subset V of X. Then $y = \{y_i\}_{i=1}^{\infty}$ is a frame for X, with B and A as upper and lower frame bounds, respectively.

Proof. To prove this theorem it is enough to show that

$$A(\|x\|_p)^q \leq \sum_{i=1}^{\infty} |[y_i, x]|^q \leq B(\|x\|_p)^q \text{ for all } x \in X. \tag{4}$$

First we prove the right hand side of the inequality (4).

Suppose to the contrary, there exists some $x_0 \in X$ such that $\sum_{i=1}^{\infty} |[y_i, x_0]|^q > B(\|x_0\|_p)^q$. Since V is dense in X, we can find a sequence $\{x_{0,j}\}_{j=1}^{\infty} \subseteq V$ such that $x_{0,j} \to x_0, \ as \ j \to \infty$. We can find a finite set $F \subseteq I$ (index set) such that $\sum_{i \in F} |[y_i, x_0]|^q > B(\|x_0\|_p)^q$.

Since $x_{0,j} \to x_0, \ as \ j \to \infty$, it follows that for very large j,

$$\sum_{i \in F} |[y_i, x_{0,j}]|^q > B(\|x_{0,j}\|_p)^q.$$

This contradicts the fact that $x_{0,j} \in V$. Hence $\sum_{i=1}^{\infty} |[y_i, x]|^q \leq B(\|x\|_p)^q$ for all $x \in X$.

Next we prove the left hand side of the inequality (4). Consider $x \in X$ and take $\{x_j\} \subseteq V$ with $x_j \to x$ as $j \to \infty$.

Since X is a continuous real semi-inner product space, we have

$$[y_i, x_j] \to [y_i, x] \text{ for } x_j \to x \text{ as } j \to \infty.$$

Hence $\sum_{i=1}^{\infty} |[y_i, x_j]|^q \to \sum_{i=1}^{\infty} |[y_i, x]|^q$ for $x_j \to x$ as $j \to \infty$.

But $A(\|x_j\|_p)^q \leq \sum_{i=1}^{\infty} |[y_i, x_j]|^q$ for all j.

This implies that $A(\|x\|_p)^q \leq \sum_{i=1}^{\infty} |[y_i, x]|^q$ as $j \to \infty$.

Thus $\{y_i\}_{i=1}^{\infty}$ is a frame for X with B and A as upper and lower frame bounds, respectively.

Example 3.1. Consider the real sequence space l^p, $1 < p < \infty$. Consider the set $\{e_i\}_{i=1}^{\infty}$, where $e_i = (0, 0, ..., 1, 0, 0..)$, where 1 is at the i^{th} coordinate and 0 at the other coordinates.

The semi-inner product of type (p) in l^p is defined as

$$[x, y] = \frac{1}{(\|y\|_p)^{p-2}} \sum_{i=1}^{\infty} |y_i|^{p-2} y_i x_i, \quad \forall x = \{x_i\}_{i=1}^{\infty} \text{ and } y = \{y_i\}_{i=1}^{\infty}.$$

We compute that $[e_i, x] = [(0, 0, .., 1, 0, 0, ..), (x_1, x_2, .., x_i, ...)] = \frac{1}{(\|x\|_p)^{p-2}} |x_i|^{p-2} x_i$ and $|[e_i, x]|^q = \frac{1}{(\|x\|_p)^{q(p-2)}} |x_i|^{q(p-1)}$. Therefore

$$\begin{aligned} \sum_{i=1}^{\infty} |[e_i, x]|^q &= \sum_{i=1}^{\infty} \frac{1}{(\|x\|_p)^{q(p-2)}} |x_i|^{q(p-1)} \\ &= \frac{1}{(\|x\|_p)^{q(p-2)}} \sum_{i=1}^{\infty} |x_i|^p, \quad as \ \frac{1}{p} + \frac{1}{q} = 1 \\ &= (\|x\|_p)^q, \quad as \ \frac{1}{p} + \frac{1}{q} = 1. \end{aligned}$$

Hence the set $\{e_i\}_{i=1}^{\infty}$ is a Parseval frame for l^p. We can also establish the reconstruction formula in this case. The set of elements $\{e_i\}_{i=1}^{\infty}$ is a Parseval frame if and only if $x = \sum_{i=1}^{\infty} \frac{|[e_i, x]|^{q-2}}{\|\{[e_i, x]\}\|^{q-2}} [e_i, x] e_i$, for all $x \in X$. We see that

$$\begin{aligned} \sum_{i=1}^{\infty} \frac{|[e_i, x]|^{q-2}}{\|\{[e_i, x]\}\|^{q-2}} [e_i, x] e_i &= \frac{1}{(\|x\|_p)^{q-2}} \sum_{i=1}^{\infty} \frac{1}{(\|x\|_p)^{(p-2)(q-2)+p-2}} |x_i|^{(p-1)(q-2)} |x_i|^{p-2} x_i e_i \\ &= \sum_{i=1}^{\infty} |x_i|^{(p-1)(q-2)+(p-2)} x_i e_i \\ &= \sum_{i=1}^{\infty} x_i e_i = x. \end{aligned}$$

Hence $\{e_i\}_{i=1}^{\infty}$ is a Parseval frame.

One can also verify that the following sets (i), (ii) and (iii) are frames for the real sequence space $l^p,\ 1 < p < \infty$.

(i) $\{e_1, 0, e_2, 0, e_3, 0, ...\}$ is a Parseval frame.
(ii) $\{e_1, e_1, e_2, e_2, ...\}$ is a tight frame with bound 2.
(iii) $\{\frac{e_1}{\sqrt{2}}, \frac{e_1}{\sqrt{2}}, \frac{e_2}{\sqrt{2}}, \frac{e_2}{\sqrt{2}}, ...\}$ is a tight frame with bound $\frac{2}{(\sqrt{2})^{\frac{p}{p-1}}}$.
(iv) $\{e_1, e_2, e_2, e_3, e_3, e_3, ...\}$ is not a frame.
(v) $\{e_1, \frac{e_2}{\sqrt{2}}, \frac{e_2}{\sqrt{2}}, \frac{e_3}{\sqrt{3}}, \frac{e_3}{\sqrt{3}}, \frac{e_3}{\sqrt{3}}, ...\}$ is not a frame unless $p = 2$.

4 Conclusion

Since the sequence spaces $l^p,\ \ p > 1$ and the function spaces $L^p,\ \ p > 1$ are uniformly convex smooth Banach spaces, the development of frame theory on these spaces using semi-inner product will lead to another new area of applied functional analysis. The frame operator which has been defined is in general nonlinear, its invertibility is an immediate open problem. The study of its invertibility and applications is a subject of future research of the authors.

Acknowledgements. The authors are thankful to the referees for their valuable suggestions which improved the presentation of the paper.

References

1. Aldroubi, A., Sun, Q., Tang, W.S.: p-frames and shift invariant spaces of L^p. J. Fourier Anal. Appl. **7**, 1–21 (2001)
2. Boos, J.: Classical and Modern Methods in Summability. Oxford University Press, Oxford (2000)
3. Cao, H.X., Li, L., Chen, Q.J., Ji, G.X.: (p, Y)-operator frames for a Banach space. J. Math. Anal. Appl. **347**, 583–591 (2008)
4. Carando, D., Lassalle, S., Schmidberg, P.: The reconstruction formula for Banach frames and duality. J. Approx. Theory **163**, 640–651 (2011)
5. Casazza, P.G., Christensen, O.: Perturbation of operators and applications to frame theory. J. Four. Anal. Appl. **3**, 543–557 (1997)
6. Casazza, P.G., Christensen, O.: Frames containing a Riesz basis and preservation of this property under perturbation. SIAM J. Math. Anal. **29**, 266–278 (1998)
7. Casazza, P.G., Christensen, O., Stoeva, D.T.: Frame expansions in separable Banach spaces. J. Math. Anal. Appl. **307**(2), 710–723 (2005)
8. Christensen, O.: Frames and Bases: An Introductory Course. Birkhauser, Boston (2008)
9. Christensen, O.: Frames and pseudo-inverse operators. J. Math. Anal. Appl. **195**, 401–414 (1995)
10. Christensen, O.: Frames perturbations. Proc. Amer. Math. Soc. **123**, 1217–1220 (1995)
11. Christensen, O.: A Paley-Wiener theorem for frames. Proc. Amer. Math. Soc. **123**, 2199–2202 (1995)

12. Christensen, O., Heil, C.: Perturbations of Banach frames and atomic decompositions. Math. Nachr. **185**, 33–47 (1997)
13. Favier, S.J., Zalik, R.A.: On the stability of frames and Riesz bases. Appl. Compu. Harm. Anal. **2**, 160–173 (1995)
14. Fornasier, M.: Banach frames for α-modulation spaces. Appl. Comput. Harmon. Anal. **22**, 157–175 (2007)
15. Giles, J.R.: Classes of semi-inner product spaces. Trans. Amer. Math. Soc. **129**, 436–446 (1967)
16. Grochenig, K.: Localization of frames, Banach frames, and the invertibility of the frame operator. J. Fourier Anal. Appl. **10**(2), 105–132 (2004)
17. Koehler, D.O.: A note on some operator theory in certain semi-inner product spaces. Proc. Amer. Math. Soc. **30**, 363–366 (1971)
18. Liu, R.: On shrinking and boundedly complete Schauder frames. J. Math. Anal. Appl. **365**, 385–398 (2010)
19. Lumer, G.: Semi-inner product spaces. Trans. Amer. Math. Soc. **100**, 29–43 (1961)
20. Nanda, S.: Numerical range for two non-linear operators in semi-inner product space. J. Nat. Acad. Math. **17**, 16–20 (2003)
21. Pap, E., Pavlovic, R.: Adjoint theorem on semi-inner product spaces of type (p). Univ. u Novom Sadu Zb. Prirod.-Mat. Fak. Ser. Mat. **25**(1), 39–46 (1995)
22. 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)

Finiteness of Criss-Cross Method in Complementarity Problem

A.K. Das[1], R. Jana[2(✉)], and Deepmala[3]

[1] SQC & OR Unit, Indian Statistical Institute, 203 B. T. Road, Kolkata 700 108, India
akdas@isical.ac.in
[2] Department of Mathematics, Jadavpur University, Kolkata 700 032, India
rwitamjanaju@gmail.com
[3] Department of Mathematics, Indian Institute of Information Technology, Design and Manufacturing, Jabalpur 482 005, India
dmrai23@gmail.com

Abstract. In this paper we consider criss-cross method for finding solution of a linear complementarity problem. The criss-cross method is a pivoting procedure. We show that the criss-cross method is able to compute solution of a linear complementarity problem in finite steps in case of some new matrix classes. We present a numerical illustration to show a comparison between criss-cross method and Lemke's algorithm with respect to number of iterations before finding a solution. Finally we raise an open problem in the context of criss-cross method.

Keywords: Criss-cross method · Complementarity problem · Lemke's algorithm · Positive subdefinite matrix · Generalized positive subdefinite matrix · Fully copositive matrix

1 Introduction

The criss-cross method is known to be finite for linear complementarity problem with positive semidefinite bisymmetric matrices and P-matrices and also for oriented matroid programming problems. We say that the criss-cross method possesses finiteness if it finds a solution or detects infeasibility in a finite number of steps. Zionts [20] proposed the criss-cross method for solving linear programming in 1969. Bland introduced smallest subscript rule for the simplex method. Using the concept of Bland [1], Chang [2], Terlaky [11] and Wang [18] independently proposed finite criss-cross method. It was observed that the proposed method works remarkably similar as the smallest subscript pivot of Bland [1] for the simplex. Recently, Fukuda, Luthi and Namiki [8] introduced a class of non-simplex pivot method which belongs to the finite criss-cross method of Chang, Terlaky and Wang. Compared to simplex method, criss-cross method is a pivoting procedure without ensuring feasibility. Hertog et al. [10] studied criss-cross method in the context of linear complementarity problem. Lemkes algorithm is a

© Springer Nature Singapore Pte Ltd. 2017
D. Giri et al. (Eds.): ICMC 2017, CCIS 655, pp. 170–180, 2017.
DOI: 10.1007/978-981-10-4642-1_15

well-known to find solution of a linear complementarity problem. It is true that both simplex and Lemke's algorithms are similar type of pivoting procedure. The limitations of Lemke's algorithm is either unable to solve several instances of linear complementarity problem or takes many iterations before to arrive at the desired solution.

The linear complementarity problem is defined as follows. Given $A \in R^{n\times n}$ and a vector $q \in R^n$, the *linear complementarity problem* LCP(q, A) is the problem of finding a solution $v \in R^n$ and $u \in R^n$ to the following system of linear equations and inequalities:

$$v - Au = q, \quad v \geq 0, \; u \geq 0 \tag{1}$$

$$v^t u = 0 \tag{2}$$

It is well studied in the literature on mathematical programming and a number of applications are reported in operations research, multiple objective programming problem, mathematical economics, geometry and engineering. Some new applications of the linear complementarity problem have been reported in the area of stochastic games. This sort of applications and the potential for future applications have motivated the study of the LCP, especially the study of the algorithms for the LCP and the study of matrix classes. In fact, much of linear complementarity theory and algorithms are based on the assumption that the matrix A belongs to a particular class of matrices. The early motivation for studying the linear complementarity problem was that the KKT optimality conditions for linear and quadratic programs reduce to an LCP. The algorithm presented by Lemke and Howson to compute an equilibrium pair of strategies to a bimatrix game, later extended by Lemke (known as Lemke's algorithm) to solve an LCP(q, A), contributed significantly to the development of the linear complementarity theory. In fact, the study of the LCP really came into prominence only when Lemke and Howson and Lemke showed that the problem of computing a Nash equilibrium point of a bimatrix game can be posed as an LCP. However, Lemke's algorithm does not solve every instance of the linear complementarity problem, and in some instances of the problem may terminate inconclusively without either computing a solution to it or showing that no solution to it exists. Extending the applicability of Lemke's algorithm to more matrix classes have been considered. For recent books on the linear complementarity problem and its applications see Cottle, Pang and Stone [5] and Murty [15].

The principal pivot transform (PPT) of LCP(q, A) with respect to α (obtained by pivoting on $A_{\alpha\alpha}$) is given by LCP(q', M) where M is the PPT of A with $q'_\alpha = -A^{-1}_{\alpha\alpha} q_\alpha$ and $q'_{\bar{\alpha}} = q_{\bar{\alpha}} - A_{\bar{\alpha}\alpha} A^{-1}_{\alpha\alpha} q_\alpha$. This problem is known as linear complementarity problem or LCP(q, A). We define $F(q, A) = \{u \in R^n_+ : q + Au \geq 0\}$ and $S(q, A) = \{u \in F(q, A) : u^t(q + Au) = 0\}$. LCP$(q, A)$ has a various application in the context of mathematical programming.

In this paper we consider finiteness of criss-cross method with respect to some new matrix classes to find solution of a linear complementarity problem. We consider the matrix classes which rely essentially on sign properties and examine the solution of linear complementarity problem. The purpose of this

paper is to characterize the new matrix classes in the context of finiteness of criss-cross method.

The paper is organized as follows. Section 2 contains some notations, definitions and a few well-known results used in the next sections. In Sect. 3 the criss-cross method and necessary properties to execute criss-cross method are discussed. Section 4 presents the characterization of the sign properties of matrices in connection with the criss-cross method. A numerical example for finding solution of an LCP(q, A) to demonstrate the effectiveness and efficiencies of criss-cross method compared with Lemke's algorithm is presented. We show that the applicability of criss-cross method can be enlarge which is illustrated with the help of an example. This issue is addressed as an open problem.

2 Preliminaries

We consider matrices and vectors with real entries. Any vector $u \in R^n$ is a column vector, u^t denotes the transpose of u. For any matrix $A \in R^{n\times n}$, A^t denotes its transpose. A vector $u \in R^n$ is said to be *unisigned* if either $u \in R^n_+$ or $-u \in R^n_+$, where R^n_+ and R^n_{++} denote the nonnegative and positive orthant in R^n respectively.

The principal pivot transform (PPT) is a fundamental concept for developing many theories and algorithms in optimization theory and plays an important role in the study of matrix classes. The *principal pivot transform* of A, a real $n \times n$ matrix, with respect to $\alpha \subseteq \{1, 2, \ldots, n\}$ is defined as the matrix given by

$$M = \begin{bmatrix} M_{\alpha\alpha} & M_{\alpha\bar{\alpha}} \\ M_{\bar{\alpha}\alpha} & M_{\bar{\alpha}\bar{\alpha}} \end{bmatrix}$$

where,
$M_{\alpha\alpha} = (A_{\alpha\alpha})^{-1}$, $M_{\alpha\bar{\alpha}} = -(A_{\alpha\alpha})^{-1}A_{\alpha\bar{\alpha}}$, $M_{\bar{\alpha}\alpha} = A_{\bar{\alpha}\alpha}(A_{\alpha\alpha})^{-1}$, $M_{\bar{\alpha}\bar{\alpha}} = A_{\bar{\alpha}\bar{\alpha}} - A_{\bar{\alpha}\alpha}(A_{\alpha\alpha})^{-1}A_{\alpha\bar{\alpha}}$.

Note that PPT is only defined with respect to those α for which $\det A_{\alpha\alpha} \neq 0$. When $\alpha = \emptyset$, by convention $\det A_{\alpha\alpha} = 1$ and $M = A$.

Lemke's algorithm is a pivotal kind of technique to compute LCP(q, A).

Step 1: Decrease u_0 so that one of the variables v_i, $1 \leq i \leq n$, say v_r is reduced to zero. We now have a basic feasible solution with u_0 in place of v_r and with exactly one pair of complementary variables (v_r, u_r) being non-basic.

Step 2: At each iteration, the complement of the variable which has been removed in the previous iteration is to be increased. In the second iteration, for instance, u_r will be increased.

Step 3: If the variable selected at step 2 to enter the basis can be arbitrarily increased, then the procedure terminates in a *secondary ray*. If a new basic feasible solution is obtained with $u_0 = 0$, we have solved (1) and (2). If in the new basic feasible solution $u_0 > 0$, we have obtained a new basic pair of complementary variables (v_s, u_s). We repeat step 2.

Lemke's algorithm consists of the repeated applications of steps 2 and 3. If non-degeneracy is assumed, the procedure terminates either in a secondary ray or in a solution to (1) and (2). Ramamurthy [17] showed that Lemke's algorithm for the linear complementarity problem can be used to check whether a given Z-matrix is a P_0-matrix and it can also be used to analyze the structure of finite Markov chains.

Several matrix classes are defined in the context of LCP(q, A). A matrix is said to be in Q if for every $q \in R^n$, $S(q, A) \neq \emptyset$. Q_0 consists the matrices for which $S(q, A) \neq \emptyset$ whenever $F(q, A) \neq \emptyset$. A matrix is said to be R_0 if LCP$(0, A)$ has unique solution. A matrix A is said to be *positive semidefinite* (PSD) if $u^t Au \geq 0$ *for all* $u \in R^n$ and A is *positive definite* (PD) if $u^t Au > 0$ *for all* $0 \neq u \in R^n$. A matrix A is said to be *column sufficient* matrix if *for all* $u \in R^n$, $u_i(Au)_i \leq 0$ *for all* i implies $u_i(Au)_i = 0$ *for all* i. A is said to be *row sufficient* if A^t is column sufficient. A is *sufficient* if A is both row and column sufficient. $A \in R^{n\times n}$ is said to be fully copositive matrix (C_0^f) if every PPT of A is a copositive matrix. $A \in R^{n\times n}$ is said to be pseudomonotone matrix if for all $u, v \geq 0$, $(v-u)^t Au \geq 0 \implies (v-u)^t Av \geq 0$.

Martos [12] proposed positive subdefinite (PSBD) matrices to address pseudo-convex functions. The nonsymmetric PSBD matrices was studied to connect generalized monotonicity and the linear complementarity problem. Later Crouzeix and Komlósi [6] enlarged PSBD class by introducing the class of GPSBD matrices. This class was studied in the context of the processability of linear complementarity problem by Lemke's algorithm. A matrix A is said to be PSBD matrix if for all $u \in R^n$, $u^t Au < 0$ implies $A^t u$ is unisigned.

A matrix $A \in R^{n\times n}$ is called GPSBD [6], [16] if $\exists\ e_i \geq 0$ and $f_i \geq 0$ with $e_i + f_i = 1,\ i = 1, 2, \ldots, n$ such that

$$\forall\ u \in R^n,\ \ u^t Au < 0 \Rightarrow \begin{cases} \text{either} & -e_i u_i + f_i (A^t u)_i \geq 0 \text{ for all } i, \\ \text{or} & -e_i u_i + f_i (A^t u)_i \leq 0 \text{ for all } i. \end{cases} \tag{3}$$

when $e_i = 0$ for all i, then A is PSBD. A is called *merely generalized positive subdefinite* (MGPSBD) *matrix* when A is GPSBD but not PSBD matrix.

We state some results which will be required in the next section.

Theorem 2.1 [7]. *Suppose $A \in R^n$ is PSBD and rank$(A) \geq 2$. Then A^t is PSBD at least one of the following conditions holds:*

(i) A is PSD,
(ii) $(A + A^t) \geq 0$,
(iii) A is $C_0^\star$.

Theorem 2.2 [7]. *A matrix $A \in R^n$ is pseudomonotone if and only A is PSBD and copositive with the additional condition that in case $A = ab^t, b_i = 0$ implies $a_i = 0$.*

Theorem 2.3 [9]. *If A is pseudomonotone, then A is a row sufficient matrix.*

Proposition 1 [3]. *Every principal submatrix of a (column, row) sufficient matrix is (column, row) sufficient.*

Proposition 2 [3]. *Both column and row sufficient matrices have nonnegative principal submatrices, and hence nonnegative diagonal elements.*

Proposition 3 [3].

(i) Let A be row sufficient with $a_{ii} = 0$ for some i. If $a_{ij} \neq 0$ for some j, then $a_{ji} \neq 0$, and in this case $a_{ij}a_{ji} < 0$.
(ii) Let A be column sufficient with $a_{ii} = 0$ for some i. If $a_{ji} \neq 0$ for some j, then $a_{ij} \neq 0$, and in this case $a_{ji}a_{ij} < 0$.
(iii) Let A be sufficient with $a_{ii} = 0$ for some i. Then $a_{ij} \neq 0$ for some j, if and only if $a_{ji} \neq 0$, and in this case $a_{ij}a_{ji} < 0$.

In case if a diagonal element of A say a_{ii} for some i is zero. Then there is a consequence of the above theorem

(i) For row sufficient matrices: If $a_{ji} \geq 0$ for all j, then $a_{ij} \leq 0$ for all j. If $a_{ji} \leq 0$ for all j, then $a_{ij} \geq 0$ for all j.
(ii) For column sufficient matrices: If $a_{ij} \geq 0$ for all j, then $a_{ji} \leq 0$ for all j. If $a_{ij} \leq 0$ for all j, then $a_{ji} \geq 0$ for all j.
(iii) For sufficient matrices: $a_{ij} \leq 0$ for all j, if and only if $a_{ji} \geq 0$ for all j. Also $a_{ij} \geq 0$ for all j, if and only if $a_{ji} \leq 0$ for all j.

Theorem 2.4 [3]. *Any principal pivotal transform of a (column, row) sufficient matrix is (column, row) sufficient.*

Theorem 2.5 [4]. *A 2×2 matrix A is sufficient if and only if for every principal pivotal transform $\bar{A}$ of A*

1. $\overline{a}_{ii} \geq 0$ and
2. for $i = 1, 2$ if $\overline{a}_{ii} = 0$, then either $\overline{a}_{ij} = \overline{a}_{ji} = 0$ or $\overline{a}_{ij}\ \ \overline{a}_{ji} < 0$ for $i \neq j$.

Theorem 2.6 [4]. *A matrix A is sufficient if and only if every principal pivotal transform $\overline{A}$ of A is sufficient of order 2.*

Theorem 2.7 [14]. *If $A \in R^{2\times2} \cap C_0^f \cap Q_0$, then A is PSD matrix.*

Theorem 2.8 [19]. *Let A be an $n \times n$ $(n \geq 2)$ pseudomonotone matrix. Then under each of the following conditions, A is column sufficient.*

(i) A is copositive plus.
(ii) $A \in R_0$.

Theorem 2.9 [10]. *Let LCP (q, A) be given, where A is a sufficient matrix, q is an arbitrary vector. Then LCP (q, A) can be processed by criss-cross method in a finite number of steps.*

3 Criss-Cross Method

Let (u, v) be the solution of a given LCP(q, A). Suppose the initial basis matrix is G, and the initial tableau is $[-A, G, q]$. A tableau is said to be complementary if u and v satisfy the complementarity condition *i.e.* $u^t v = 0$. Let $-A$ denote the non-basic part of any complementarity tableau. Non-basic part of any complementary tableau is a principal pivotal transform of the matrix $-A$. Criss-Cross method will **STOP** if it finds a solution or detects infeasibility, while **EXIT** indicates that the method fails to execute the problem. The criss-cross method is as follows:

Step 1: Let the starting basis be defined by v, and let $v = q$, $u = 0$ be the initial solution. The initial tableau is given by $[-A, G, q]$.

Step 2: Let $k := min\ \{i : v_i < 0 \ \ or \ \ u_i < 0\}$. If there is no such k, then **STOP**; a feasible complementary solution has been found. Suppose there exists a k such that $v_k < 0$, then we have to make pivot so that v_k leaves the basis.

Step 3: If $-a_{kk} < 0$, then make a diagonal pivot and repeat the procedure that is v_k leaves and u_k enters the basis. If $-a_{kk} > 0$, then **EXIT**. If $-a_{kk} = 0$, go to Step 4.

Step 4: Here $a_{kk} = 0$ is the case. Choose $r := \min\ \{j : -a_{kj} < 0\}$.

- If there is an r and $a_{rk}a_{kr} < 0$, then make an exchange pivot on (r, k) and repeat the procedure. Exchange pivot means v_k, u_r leave from the basis and u_k, v_r enter into the basis.
- Otherwise either LCP(q, A) is infeasible or criss-cross method is unable to process the solution.

Hertog et al. [10] showed that if a matrix is sufficient matrix then criss-cross method will process LCP(q, A) in a finite number of steps. We discuss the necessary and sufficient conditions which ensure not to encounter **EXIT** by criss-cross method. The method operates on diagonal and exchange pivots only, so complementarity in each step is preserved. The criss-cross method **STOP** implies either LCP(q, A) has a solution or it is infeasible.

4 Finiteness of Criss-Cross Method

Hertog et al. [10,11] define three properties so that criss cross method can process in finite number of steps. $\mathfrak{F}$ denotes the class of matrices such that for each $A \in \mathfrak{F}$ and for each vector $q \in R^n$ the problem LCP(q, A) is processed by the criss-cross method in a finite number of steps. Also suppose that $\mathfrak{F}$ is closed with respect to principal pivot transformation, and complete with respect to principal submatrices of every matrix $A \in \mathfrak{F}$. Orthogonality property ensures the finiteness of the method. Orthogonality property says that any row vector of a tableau is orthogonal to column vector of its dual tableau. Firstly we rewrite the first two properties.

Property 1. If $A \in \mathfrak{F}$, the diagonal elements of any principal pivotal transform of $-A$ are nonpositive.

Property 2. If $-a_{kk} = 0$ for some k, then $-a_{kj} < 0$ if and only if $-a_{jk} > 0$ for any j.

The first property says about the diagonal pivot whereas second property guarantees the exchange pivot. Both Properties 1 and 2 ensure the complementary and feasibility conditions. We now consider the third property which says finiteness of criss-cross method. Suppose there are two tableau defined based on sign properties and these types are exclusive for LCP(q, A) if at most one of them may exist for the problem.

Property 3. For a given LCP(q, A) we define following cases for which the pairs of cases PQ, RS, PR, and QS are exclusive for any index $1 \leq k \leq n$:
P: We have a complementary tableau with $v_i \geq 0, u_i \geq 0$ for $i < k$, and $v_k = 0, u_k < 0$.
Q: We have a complementary tableau with $v_i \geq 0, u_i \geq 0$ for $i < k$, and $v_k < 0, u_k = 0$.
R: We have a complementary tableau with $u_s < 0$ for some $s < k, and\ a_{si} \geq 0$ for $i < k$, $a_{ss} = 0$, and $a_{sk} < 0$; and symmetrically $a_{is} \leq 0$ for $i < k$, and $a_{ks} > 0$.
S: We have a complementary tableau with $v_s < 0$ for some $s < k$, and $a_{si} \geq 0$ for $i < k$, $a_{ss} = 0$, and $a_{sk} < 0$; and symmetrically $a_{is} \leq 0$ for $i < k$, and $a_{ks} > 0$.

To prove the finiteness of the method the only restrictive requirement in the property that P and Q tableau are exclusive. On the other hand remaining pairs follow from orthogonality property as shown in [11]. We now prove the following results.

Theorem 4.1. *Suppose $A \in MGPSBD \cap C_0$ with $0 < f_i < 1\ \forall\ i$. Then criss-cross method processes LCP(q, A).*

Proof. Let $I_1 = \{i\ : u_i > 0\}$ and $I_2 = \{i\ : u_i < 0\}$. We consider the following three cases (**C1, C2, C3**).

C1: $I_2 = \emptyset$. Then

$$u^t Au = u^t A^t u = \sum_i (u)_i (A^t u)_i \leq 0.$$

Since $A \in C_0$, $[(u)_i (A^t u)_i] = 0,\ \forall\ i$.

C2: $I_1 = \emptyset$. Then

$$(-u)^t A^t (-u) = u^t A^t u = \sum_i (u)_i (A^t u)_i \leq 0.$$

Since $A \in C_0$, $[(u)_i (A^t u)_i] = 0,\ \forall\ i$.

C3: Suppose $\exists\ u$ such that $(u)_i (A^t u)_i \leq 0$ for $i = 1, 2, \ldots, n$ and $(u)_k (A^t u)_k < 0$ for at least one $k \in \{1, 2, \ldots, n\}$. Let $I_1 \neq \emptyset$ and $I_2 \neq \emptyset$. Then

$$u^t A^t u = \sum_i [u_i (A^t u)_i] < 0.$$

This implies

$$-e_i u_i + f_i (A^t u)_i \geq 0, \ \forall \ i \text{ or}$$
$$-e_i u_i + f_i (A^t u)_i \leq 0, \ \forall \ i.$$

Let us consider $-e_i u_i + f_i (A^t u)_i \geq 0, \ \forall \ i$. Then for all $i \in I_1$, $-e_i u_i^2 + f_i u_i (A^t(u)_i \geq 0$. This implies $[u_i (A^t u)_i] \geq \frac{e_i}{f_i} u_i^2 > 0, \ \forall \ i \in I_1$. so, $\sum_{i \in I_1} [u_i (A^t u)_i] > 0$. Since $u_i (A^t u)_i \leq 0$ for $i = 1, \ldots, n$. Therefore, $[u_i (A^t u)_i] = 0, \ \forall \ i$.

So to show the above result it is enough to show that A satisfies the above mentioned two properties. Here Property 1 follows from the Proposition 2. Property 2 follows from Proposition 3. ■

Remark 1. *From the above result LCP(q, A) is processable by criss-cross method in general. If A, $A^t \in$ MGPSBD $\cap$ C_0 with $0 < f_i < 1$ for all i then LCP(q, A) is processable by criss-cross method in a finite number of steps.*

Our next theorem states that under some condition if A belongs to PSBD matrix class, then the criss-cross method will process LCP(q, A) in a finite number of steps.

Theorem 4.2. *Suppose A is a PSBD matrix with $rank(A) \geq 2$. Then under each of the following conditions criss-cross method processes LCP(q, A) in a finite number of steps.*

(i) A is C_0,
(ii) A is R_0.

Proof. As A is PSBD $\cap$ C_0, A^t is a PSBD $\cap$ C_0 with $rank(A^t) \geq 2$. Now A and A^t is pseudomonotone matrix by Theorem 2.1 as shown in [7]. Again any pseudomonotone matrices are row sufficient by Theorem 2.3 as shown in [9], so A and A^t are row sufficient. So A is sufficient.
To prove (ii) we proceed as follows: Here A is PSBD, so A is pseudomonotone. Hence A is row sufficient. Again as A is R_0 and by Theorem 2.8 as shown in [19] A is column sufficient, hence A is sufficient. So criss-cross method processes LCP(q, A) in a finite number of steps by Theorem 2.9 as shown in [10]. ■

Theorem 4.3. *Let $A \in C_0^f \cap Q_0$. Then criss-cross method processes LCP(q, A) in a finite number of steps.*

Proof. As $A \in C_0^f \cap Q_0$, A and all its PPTs are completely Q_0. So here all 2×2 submatrices of A or its principal pivotal transform are in $C_0^f \cap Q_0$. So all 2×2 submatrices of A are PSD matrix by Theorem 2.7 as shown in [14]. As all PSD matrices are sufficient, so here all 2×2 submatrices of A are sufficient also. So A or every matrix obtained by means of a principal pivotal transform is sufficient of order 2. By Theorem 2.5 as shown in [4] A is sufficient. So criss-cross method processes LCP(q, A) in a finite number of steps by Theorem 2.9 as shown in [10]. ■

Theorem 4.4. *Let $A, A^t \in R^{n\times n} \cap C_0^f$ with positive diagonals. Then criss-cross method processes LCP(q, A) in a finite number of steps.*

Proof. Since $A, A^t \in R^{n\times n} \cap C_0^f$ with positive diagonals, A, A^t are column sufficient [see Theorem 3.4 in [13]]. Hence A is sufficient and Hence by Theorem 2.9 as shown in [10] criss-cross method processes LCP(q, A) in a finite number of steps. ■

We make use of the following example to demonstrate the applicability of criss-cross method and a comparison with Lemke's algorithm.

Example 1. *We consider an LCP(q, A) for which* $A = \begin{bmatrix} 0 & 4 \\ -1 & 0 \end{bmatrix}$ *and* $q = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$.

Note that that $A \in$ PSBD $\cap\, C_0$ with full rank. Hence from the above theorem A is sufficient matrix. Now we apply criss-cross method to solve the above LCP(q, A) (Table 1).

In the first iteration, w_1 and w_2 are in the basis. Since diagonal pivot is not applicable, we apply exchange pivot according to the method and obtain the solutions of the given LCP(q, A). Here $u = [2,\ 1/4]^t$ and $v = [0,\ 0]^t$.

Now we apply Lemke's algorithm to solve the same LCP(q, A) (Table 2).

The Lemke's algorithm requires four iterations to solve LCP(q, A) whereas criss-cross method requires two iterations.

Table 1. Solution using criss-cross method considering Property 2

	v_1	v_2	u_1	u_2	q
v_1	1	0	0	(-4)	-1
v_2	0	1	(1)	0	2
u_2	$-1/4$	0	0	1	$1/4$
u_1	0	1	1	0	2

Table 2. Solution using Lemke's algorithm

	v_1	v_2	u_1	u_2	u_0	q
v_1	1	0	0	-4	-1	-1
v_2	0	1	1	0	-1	2
u_0	-1	0	0	4	1	1
v_2	-1	1	1	4	0	3
u_0	-1	0	0	4	1	1
u_1	-1	1	1	4	0	3
u_2	$-1/4$	0	0	1	$1/4$	$1/4$
u_1	0	1	1	0	-1	2

4.1 An Open Problem

Let us consider an LCP(q, A) for which $A = \begin{bmatrix} 0 & 2 \\ 1 & 0 \end{bmatrix}$ and $q = \begin{bmatrix} -1 \\ -2 \end{bmatrix}$. Since $a_{11} = a_{22} = 0$, we apply exchange pivot without considering Property 2 (Table 3).

Table 3. Solution using criss-cross method without considering Property 2

	v_1	v_2	u_1	u_2	q
v_1	1	0	0	(-2)	-1
v_2	0	1	(-1)	0	-2
u_2	$-1/2$	0	0	1	$1/2$
u_1	0	-1	1	0	2

The solution for LCP(q, A) is $u = [2,\ 1/2]^t$ and $v = [0,\ 0]^t$. Note that A is neither GPSBD nor sufficient matrix. However we obtain the solution of this problem by applying criss-cross method. Hence we raise the following questions. *Is it possible to apply the criss-cross method to find solutions of an LCP(q, A), where A does not belong to GPSBD or sufficient matrix?*

Acknowledgments. The second author R. Jana is thankful to the Department of Science and Technology, Govt. of India, INSPIRE Fellowship Scheme for financial support.

References

1. Bland, R.G.: New finite pivoting rules for the simplex method. Math. Oper. Res. **2**(2), 103–107 (1977)
2. Chang, Y.Y.: Least-index resolution of degeneracy in linear complementarity problems (No. TR-79-14), Department of Operations Research, Stanford University (1979)
3. Cottle, R.W.: The principal pivoting method revisited. Math. Program. **48**(1), 369–385 (1990)
4. Cottle, R.W., Guu, S.M.: Two characterizations of sufficient matrices. Linear Algebra Appl. **170**, 65–74 (1992)
5. Cottle, R.W., Pang, J.S., Stone, R.E.: The Linear Complementarity Problem. SIAM, Philadelphia (2009)
6. Crouzeix, J.P., Komlósi, S.: The linear complementarity problem and the class of generalized positive subdefinite matrices. In: Giannessi, F., Pardalos, P., Rapcsák, T. (eds.) Optimization Theory, pp. 45–63. Springer, New York (2001)
7. Crouzeix, J.P., Hassouni, A., Lahlou, A., Schaible, S.: Positive subdefinite matrices, generalized monotonicity, and linear complementarity problems. SIAM J. Matrix Anal. Appl. **22**(1), 66–85 (2000)
8. Fukuda, K., Lüthi, H.J., Namiki, M.: The existence of a short sequence of admissible pivots to an optimal basis in LP and LCP. Int. Trans. Oper. Res. **4**(4), 273–284 (1997)

9. Gowda, M.S.: Affine pseudomonotone mappings and the linear complementarity problem. SIAM J. Matrix Anal. Appl. **11**(3), 373–380 (1990)
10. Hertog, D., Roos, C., Terlaky, T.: The linear complimentarity problem, sufficient matrices, and the criss-cross method. Linear Algebra Appl. **187**, 1–14 (1993)
11. Klafszky, E., Terlaky, T.: Some generalizations of the criss-cross method for quadratic programming. Optimization **24**(1–2), 127–139 (1992)
12. Martos, B.: Subdefinite matrices and quadratic forms. SIAM J. Appl. Math. **17**(6), 1215–1223 (1969)
13. Mohan, S.R., Neogy, S.K., Das, A.K.: On the classes of fully copositive and fully semimonotone matrices. Linear Algebra Appl. **323**(1), 87–97 (2001)
14. Murthy, G.S.R., Parthasarathy, T.: Fully copositive matrices. Math. Program. **82**(3), 401–411 (1998)
15. Murty, K.G., Yu, F.T.: Linear Complementarity, Linear and Nonlinear Programming. Heldermann, Berlin (1988)
16. Neogy, S.K., Das, A.K.: Some properties of generalized positive subdefinite matrices. SIAM J. Matrix Anal. Appl. **27**(4), 988–995 (2006)
17. Ramamurthy, K.G.: A polynomial algorithm for testing the nonnegativity of principal minors of Z-matrices. Linear Algebra Appl. **83**, 39–47 (1986)
18. Wang, Z.M.: A finite conformal-elimination free algorithm over oriented matroid programming. Chin. Ann. Math. Ser. B **8**(1), 120–125 (1987)
19. Xu, S.: Notes on sufficient matrices. Linear Algebra Appl. **191**, 1–13 (1993)
20. Zionts, S.: The criss-cross method for solving linear programming problems. Manag. Sci. **15**(7), 426–445 (1969)

Imprecise Constrained Covering Solid Travelling Salesman Problem with Credibility

Anupam Mukherjee[1(✉)], Samir Maity[2], Goutam Panigrahi[1], and Manoranjan Maiti[3]

[1] Department of Mathematics, National Institute of Technology Durgapur, Durgapur 713209, West Bengal, India
mukherjee.anupam.bnk@gmail.com, panigrahi_goutam@rediffmail.com
[2] Department of Computer Science, Vidyasagar University, Midnapore 721102, West Bengal, India
maitysamir13@gmail.com
[3] Department of Applied Mathematics, Vidyasagar University, Midnapore 721102, West Bengal, India
mmaiti2005@yahoo.co.in

Abstract. In this article, we model an "Imprecise Constrained Covering Solid Travelling Salesman Problem with Credibility" (ICCSTSPC), a generalization of Covering Salesman Problem (CSP), in fuzzy environment. A salesman begins from an initial node, visits a subset of nodes exactly once using any one of appropriate vehicles available at each step, so that unvisited nodes are within a predetermined distance from the visited nodes, and returns to the initial node within a restricted time. Here the travelling costs and travelling times between any two nodes and the covering distance all are considered as fuzzy. Thus the problem reduces to find the optimal tour for a set of nodes with the proper conveyances so that total travelling cost is minimum within a restricted time. The ICCSTSPC is reduced to a set of Imprecise Constrained Covering Solid Travelling Salesman Problems by solving Unicost Set Cover Problem (USCP) using Random Insertion-Deletion (RID). These reduced Constrained Solid Travelling Salesman Problems (CSTSPs) are solved by an Improved Genetic Algorithm (IGA), which consists of probabilistic selection, order crossover, proposed generation dependent inverse mutation. A random mutation for vehicles is proposed to get a better cost at each generation of IGA by choosing an alternative vehicle for each node. Hence the ICCSTSPC is solved by a random insertion-deletion (RID) for covering set and IGA, i.e., RID-IGA. To justify the performance of the RID-IGA, some test problems are solved. The model is illustrated with some randomly generated crisp and fuzzy data.

Keywords: Solid TSP · Covering Salesman Problem · Improved GA

1 Introduction

Travelling Salesman Problem (TSP), which is one of the most well known NP-complete problems, was first formulated in 1930. Many researchers have been

© Springer Nature Singapore Pte Ltd. 2017
D. Giri et al. (Eds.): ICMC 2017, CCIS 655, pp. 181–195, 2017.
DOI: 10.1007/978-981-10-4642-1_16

developed some generalizations of TSPs, e.g., TSP with precedence constraints [13], stochastic TSP [1], symmetric TSP [12], asymmetric TSP [11] etc.

In 1989, Current and Schilling [3] first introduced the model of Covering Salesman Problem (CSP), which is a generalization of TSP. In CSP, a salesmen selects a subset of nodes from the total node set, starts from an initial node, visits all nodes of the subset exactly once, such that all other nodes be covered within a predetermined distance from the visited nodes, and at the end, comes back to the initial node. Current et al. developed a simple heuristic method to solve CSP consisting of two parts, first is the unicost set covering problem (USCP) to find minimum number of nodes to cover all other nodes, and the next step is to solve TSPs for different USCP solutions (if exist) to get the path with minimum cost. Later, Golden et al. [6] developed two local search (LS) algorithms $LS1$ and $LS2$. Salari et al. [14] developed an integer programming based LS for CSP and after that, a hybrid algorithm consisting of ant colony optimization (ACO) and dynamic programming technique for CSP was introduced by Salari et al. [15].

Genetic Algorithm (GA) is a nature inspired soft computing technique. Different types of GAs have been developed in last few decades, e.g., Adaptive GA [17], Hybrid GA [18], NSGA-II etc. [5], Fuzzy age based GA [10] etc. were developed for several research areas.

Solid travelling salesman problem (STSP) is an extension of TSP, where the salesman can avail any one kind of appropriate vehicle at each node. Considering different types of vehicles and risks, Changdar et al. [2] developed the model of STSP in crisp and fuzzy environments. Later, Maity et al. [10] extended the same problem to bi-random and random-fuzzy environments. Both Changdar et al. and Maity et al. solved their problems using their own modified GAs.

Imprecise Constrained Covering Solid Travelling Salesman Problem with Credibility (ICCSTSPC) can be defined as a generalized CSP, in which the travel costs, travel times and covering distance are taken as fuzzy, also, there are several types of conveyances at each node for travel. None has investigated this type of realistic CSP yet. Given a set of nodes N. A salesman begins from any one node and visits a subset of nodes $N' \subset N$, each node exactly once, by choosing a suitable vehicle available at each node, such that all nodes out of the tour are covered within a predetermined distance from the visited nodes, and at the end, returns to the initial node within a restricted time.

We solve the above mentioned ICCSTSPC in two steps, first we find the minimal covering sets with least nodes by solving USCP within a time bound (we take 60 s). We propose RID to solve USCP, which inserts nodes randomly (each node at most once), the insertion process stops when the feasibility of set cover is satisfied and we get a set cover. The obtained set cover is then gone through the deletion process which checks each nodes of the cover whether it can be deleted or not without violating the set cover feasibility. If such nodes exist, those are deleted to obtain a minimal set cover. This process may generates a few solutions having different number of nodes in the given runtime bound, but only the solutions with minimum number of nodes are selected for the next step, i.e., obtaining optimal paths for each of those solutions by solving CSTSPs.

For CSTSPs, we modify an improved GA (IGA), which consists of probabilistic selection, order crossover, generation dependent inverse mutation and random mutation for vehicles (at each node) are introduced. So the model ICCSTSPC is solved by a combined RID-IGA method, which is applied on randomly generated 100×100 distance matrix (crisp), and $100\times100\times3$ cost and time matrices (fuzzy) respectively are used for illustration of the model. The distance matrix is used to solve the USCP and other two for solving constrained STSP (CSTSP). To justify the performance of IGA, it is tested with some TSP benchmark problems and CSPs of Salari et al.'s [15].

2 Mathematical Preliminaries

2.1 Fuzzy Credibility Approach

Let (a, b, c) be a TFN, then the credibility measures [9] for the events $\xi \leq r$ and $\xi \geq r$ are given by:

$$Cr(\xi \leq r) = \begin{cases} 0, & \text{if } r < a; \\ \frac{r-a}{2(b-a)}, & \text{if } a \leq r \leq b; \\ \frac{1}{2}(\frac{r-b}{c-b} + 1), & \text{if } b \leq r \leq c; \\ 1, & \text{if } r > c. \end{cases} \tag{1}$$

$$Cr(\xi \geq r) = \begin{cases} 0, & \text{if } r > c; \\ \frac{c-r}{2(c-b)}, & \text{if } b \leq r \leq c; \\ \frac{1}{2}(\frac{b-r}{b-a} + 1), & \text{if } a \leq r \leq b; \\ 1, & \text{if } r < a. \end{cases} \tag{2}$$

The following lemmas can be easily proven from the above Eqs. (1) and (2):

Lemma 2.1.a: If $\xi = (a, b, c)$ be a fuzzy variable with $a < b < c$, then for a predetermined β, $0 < \beta \leq 1$, $Cr(\xi \leq r) \geq \beta$ is equivalent to
(i) $(1 - 2\beta)a + 2\beta b \leq r$, when $\beta \leq 0.5$;
(ii) $2(1 - \beta)b + (2\beta - 1)c \leq r$, when $\beta > 0.5$;

Lemma 2.1.b: If $\xi = (a, b, c)$ be a fuzzy variable with $a < b < c$, then for a predetermined β, $0 < \beta \leq 1$, $Cr(\xi \geq r) \geq \beta$ is equivalent to
(i) $2\beta b + (1 - 2\beta)c \geq r$, when $\beta \leq 0.5$;
(ii) $2(1 - \beta)a + (2\beta - 1)b \geq r$, when $\beta > 0.5$

3 Mathematical Formulations

3.1 Covering Salesman Problem

For a complete graph $G = (N, A)$, minimize the total tour cost when a salesman starts from an initial node of a subset $N' \subset N$ of nodes, visits each node exactly once and comes back to the initial node, so that the unvisited nodes be within a

predetermined distance from at least one of the visited nodes. The mathematical formulation of this problem may be stated as:

$$Minimize \quad Z = \sum_{i=1}^{|N|} \sum_{j=1}^{|N|} c_{ij} x_{ij} \tag{3}$$

Subject to:

$$\sum_{i=1}^{|N|} \sum_{j \in D_l} x_{ij} \geq 1, \forall\ l \in N \tag{4}$$

$$\sum_{i=1}^{|N|} x_{ik} = \sum_{j=1}^{|N|} x_{kj} = 0 \quad or \quad 1, \forall\ k \in N \tag{5}$$

$$x_{ij} \in \{0, 1\} \tag{6}$$

$$\sum_{i \in S} \sum_{j \in S} x_{ij} \leq |S| - 1, \quad \forall\ S \subset N' \subset N, \quad 2 \leq |S| \leq |N'| - 2 \tag{7}$$

where, N' is the set of visiting nodes, c_{ij} is the cost from the node i to the node j,

$$x_{ij} = \begin{cases} 1, \exists \text{ an edge between } i \text{ and } j, \\ 0, \text{ otherwise;} \end{cases}$$

$D_l = \{j : d_{lj} \leq \Delta_j\}$, d_{ij} = shortest distance between i and j, Δ_j = maximum covering distance at node j.

Equation (3) minimizes the total travelling cost. (4) implies that all nodes of the graph are either visited or covered by the visited nodes. Equation (5) points that each vertex has same indegree and outdegree. (6) represents the binary nature of the decision variable x_{ij} and (7) is the subtour elimination constraint.

The above Eqs. (3)–(7) can be rewritten as follows:
Let $N = \{x_1, x_2, x_3, \ldots, x_{|N|}\}$ be the set of nodes. Determine a complete tour $(x_{\alpha_1}, x_{\alpha_2}, x_{\alpha_3}, \ldots, x_{\alpha_m}, x_{\alpha_1})$, $m \leq |N|$ to

$$\text{minimize} \sum_{i=1}^{m-1} c(x_{\alpha_i} x_{\alpha_{i+1}}) + c(x_{\alpha_m} x_{\alpha_1}); \tag{8}$$

$$\text{such that, } x_j \in \bar{B}(x_{\alpha_i}, \Delta_{\alpha_i}), \quad \forall \quad x_j \in N \text{ and for some } i; \tag{9}$$

where $\alpha_i \in \{1, 2, 3, \ldots, |N|\}$ and $\alpha_i \neq \alpha_j$ for $i \neq j$, $c(i, j) = c_{ij}$, $\bar{B}(a, r)$ means closed disc with center a and radius r, Δ_j = maximum covering distance at node j.

3.2 Model-1: Constrained Covering Solid Travelling Salesman Problem (CCSTSP)

In the above mentioned CSP, let $N = \{x_1, x_2, x_3, \ldots, x_{|N|}\}$ be the set of nodes and $V = \{v_1, v_2, v_3, \ldots, v_p\}$ be the set of vehicles. Determine a complete tour $(x_{\alpha_1}, x_{\alpha_2}, x_{\alpha_3}, \ldots, x_{\alpha_m}, x_{\alpha_1})$, $m \leq |N|$ to

$$\text{minimize} \sum_{i=1}^{m-1} c(x_{\alpha_i} x_{\alpha_{i+1}}, v'_{\alpha_i}) + c(x_{\alpha_m} x_{\alpha_1}, v'_{\alpha_m}); \tag{10}$$

such that,

$$\sum_{i=1}^{m-1} t(x_{\alpha_i} x_{\alpha_{i+1}}, v'_{\alpha_i}) + t(x_{\alpha_m} x_{\alpha_1}, v'_{\alpha_m}) \leq t_{max}; \tag{11}$$

$$x_j \in \bar{B}(x_{\alpha_i}, \Delta_{\alpha_i}), \quad \forall \quad x_j \in N \text{ and for some } i. \tag{12}$$

where $\alpha_i \in \{1, 2, 3, \ldots, |N|\}$ and $\alpha_i \neq \alpha_j$ for $i \neq j$, $v'_{\alpha_i} \in V$, $\forall \quad \alpha_i \in \{1, 2, 3, \ldots, |N|\}$, $c(i, j, k) = c_{ijk}$, $t(i, j, k) = t_{ijk}$, t_{max} being the maximum allowed total time for the tour, $\bar{B}(a, r)$ means closed disc with center a and radius r, Δ_j = maximum covering distance at node j.

3.3 Model-2: Imprecise Constrained Covering Solid Travelling Salesman Problem with Credibility (ICCSTSPC)

If, in the above CCSTSP, we consider the covering distance, vehicle costs as fuzzy, also, add a time constraint, where both the time from each node to another node and the maximum total allowed time for a complete tour are also taken as fuzzy, the above model is transformed in credibility approach as: Determine a complete tour
$(x_{\alpha_1}, x_{\alpha_2}, x_{\alpha_3}, \ldots, x_{\alpha_m}, x_{\alpha_1}), \quad m \leq |N|$
to minimize F
subject to

$$Cr(\sum_{i=1}^{m-1} \mathbf{c}(x_{\alpha_i} x_{\alpha_{i+1}}, v'_{\alpha_i}) + \mathbf{c}(x_{\alpha_m} x_{\alpha_1}, v'_{\alpha_m}) \leq F) \geq \beta \tag{13}$$

$$Cr(\sum_{i=1}^{m-1} \mathbf{t}(x_{\alpha_i} x_{\alpha_{i+1}}, v'_{\alpha_i}) + \mathbf{t}(x_{\alpha_m} x_{\alpha_1}, v'_{\alpha_m}) \leq \mathbf{t_{max}}) \geq \gamma \tag{14}$$

$$Cr(\tilde{\Delta}_{\alpha_i} \geq d(x_j, x_{\alpha_i})) \geq \eta, \quad \forall \quad x_j \in N \text{ and for some } i. \tag{15}$$

where β, γ and η are the confidence levels for travelling cost, travelling time and covering distance respectively.

Using Lemmas 2.1.a and 2.1.b and subtraction formula for fuzzy numbers the above Eqs. (16), (17) and (18) can be rewritten as: Determine a complete tour $(x_{\alpha_1}, x_{\alpha_2}, x_{\alpha_3}, \ldots, x_{\alpha_m}, x_{\alpha_1}), \quad m \leq |N|$ to
minimize F
subject to

$$\begin{cases} (1-2\beta)C_1 + 2\beta C_2 \leq F, & when \quad \beta \leq 0.5; \\ 2(1-\beta)C_2 + (2\beta - 1)C_3 \leq F, & when \quad \beta > 0.5. \end{cases} \tag{16}$$

$$\begin{cases} (1-2\gamma)(T_1 - t_{max_3}) + 2\gamma(T_2 - t_{max_2}) \leq 0, & when \quad \gamma \leq 0.5; \\ 2(1-\gamma)(T_2 - t_{max_2}) + (2\gamma - 1)(T_3 - t_{max_1}) \leq 0, & when \quad \gamma > 0.5. \end{cases} \tag{17}$$

$$\begin{cases} 2\eta(\Delta_{\alpha_i})_2 + (1-2\eta)(\Delta_{\alpha_i})_3 \geq d(x_j, x_{\alpha_i}), & when \ \ \eta \leq 0.5; \\ (2\eta-1)(\Delta_{\alpha_i})_1 + 2(1-\eta)(\Delta_{\alpha_i})_2 \geq d(x_j, x_{\alpha_i}), & when \ \ \eta > 0.5, \\ & \forall \ \ x_j \in N \text{ and for some } i, \end{cases} \tag{18}$$

where, the costs, times and the maximum allowable time are taken in the form (c_1, c_2, c_3) and (t_1, t_2, t_3) and $(t_{max_1}, t_{max_2}, t_{max_3})$ respectively. $d(x_i, x_j)$ is the shortest distance between the nodes x_i and x_j. The covering distance at the node x_{α_i} is considered as TFN: $((\Delta_{\alpha_i})_1, (\Delta_{\alpha_i})_2, (\Delta_{\alpha_i})_3)$,

$$C_k = \sum_{i=1}^{m-1} (c(x_i, x_{i+1}, v_i))_k + (c(x_m, x_1, v_m))_k,$$

$$T_k = \sum_{i=1}^{m-1} (t(x_i, x_{i+1}, v_i))_k + (t(x_m, x_1, v_m))_k, \quad k = 1, 2, 3.$$

4 Solution Procedure

4.1 RID for USCP

RID for USCP

1. $\mathbf{S} \leftarrow \phi$ //**S** being the null set
2. $\mathbf{N} = \{1, 2, \ldots, n\}$ //**N** being the full set of nodes
3. $i \leftarrow 1$
4. **while** $i <$ *total no. of nodes* **do**
   - $a \in \mathbf{N} - \mathbf{S}$
   - $\mathbf{S} \leftarrow \mathbf{S} \cup a$
   - $i \leftarrow i + 1$ **if** *feasibility of SCP is satisfied* **then**
     - break;
   - **end**
   - $t \leftarrow i$

   end
5. $i \leftarrow 1$
6. **while** $i < t - 1$ **do**
   - **if** $S - \{i\}$ *is not a set cover* **then**
     - break;
   - **else**
     - $S \leftarrow S - \{i\}$;
   - **end**

   end
7. repeat the process to search for another solution
8. Mark the solutions with minimum nodes as optimal solutions

4.2 Improved Genetic Algorithm (IGA) for CSTSP

To solve the reduced CSTSPs for the marked USCP solutions, we modify an Improved GA (IGA) which includes Probabilistic Selection [10], Order crossover, generation dependent Inverse mutation and random mutation (at each node) for the vehicles.

Initialization: In GA for CSTSPs, a chromosome is formed by arranging all the nodes on the tour in any order without any repetition. Let n represent the number of nodes and m represent the number of chromosomes, and $V = \{v_1, v_2, \ldots, v_p\}$ be the total set of different conveyances. Then each chromosome X_i, $(i = 1, 2, \ldots, m)$ and corresponding vehicle set can be represented as $X_i = (x_{i1}, x_{i2}, \ldots, x_{in})$ and $V_i' = (v_{i1}', v_{i2}', \ldots, v_{in}')$ respectively.

Algorithm for initialization
Data: Number of chromosomes m, number of nodes n
Result: A set of m chromosomes each having n different nodes
while *i=1 to m* **do**
  while *j=1 to n* **do**
    label: $t = rand[1, n]$
    for *k=1 to j-1* **do**
      if $t{=}x_{ik}$ **then**
        goto label;
      end
      $x_{ij} = t$ $temp = rand[1, p]$, $v_{ij}' = v_{temp}$
    end
  end
end

The algorithms of conventional Probabilistic selection, Order crossover, proposed generation dependent inverse mutation and random mutation for vehicles are given below:

Algorithm for probabilistic selection
Data: pop-size(m), population set, probability of selection(p_s), max-gen(g)
Result: mating pool

for *i=1 to g* **do**
  for *j=1 to m* **do**
    1. a=rand[0,1]
    2. T_0=rand[10,100]
    3. b=rand[0,1]
    4. k=1+100×(i/g)
    5. T=$T_0(1-b)^k$
    6. p_b=exp $(f_{min} - f(X_j))/T$
    7. **if** $a < p_b$ **then**
      select j^th chromosome
    end
    else if $a < p_s$ **then**
      select the current chromosome
    end
    else
      select the chromosome with the value f_{min}
    end
  end
end

Algorithm for order Crossover:
Data: number of nodes N, $parent_1$, $parent_2$
Result: $offspring_1$, $offspring_2$
1. r_1=rand[1,N], r_2=rand[1,N] such that $1 \leq r_1 \leq r_2 \leq N$
2. **for** *i=r_1 to r_2* **do**
  $offspring_1[i] = parent_2[i]$
end
3. $temp = r_2 + 1$
4. **for** $i = r_2 + 1$ *to N and i=1 to* $r_1 - 1$ **do**
  if $parent_1[i] \notin offspring_1$ **then**
    $offspring_1[temp] = parent_1[i]$
    $temp \leftarrow temp + 1$
    if *temp=N+1* **then**
      $temp \leftarrow 0$
    end
  end
end
5. $s_1 = rand[1, N]$, $s_2 = rand[1, N]$ such that $1 \leq s_1 \leq s_2 \leq N$
6. **for** $i = s_1$ *to* s_2 **do**
  $offspring_2[i] = parent_1[i]$
end
7. $temp = s_2 + 1$

8. **for** $i = s_2 + 1$ *to* N *and* *i=1 to* $s_1 - 1$ **do**
   if $parent_2[i] \notin offspring_2$ **then**
      $offspring_2[temp] = parent_2[i]$
      $temp \leftarrow temp + 1$
      if $temp = N + 1$ **then**
         $temp \leftarrow 0$
      end
   end
end

Algorithm for inverse Mutation
Data: number of nodes n, chromosome
Result: mutated chromosome
1. generate r1=rand[1,n] and r2=rand[1,n] such that $r1 < r2$
2. **for** *i=r1 to r2/2* **do**
   $node[i] \leftarrow node[r2 - i + r1]$
end

Random mutation for vehicles
Data: number of nodes n, chromosome, number of vehicles p, cost matrix
Result: chromosome with mutated vehicles
for $i = 1$ *to* n **do**
   $temp = rand[1, p]$,
   if $cost(x_i, x_{i+1}, v_{temp}) \leq cost(x_i, x_{i+1}, v_i')$ **then**
      replace v_i' by v_{temp}
   end
end

Combining all the above algorithms, the whole IGA can be presented as follows:

Procedure IGA for CSTSP
Data: Maximum number of generation (maxgen), pop-size, number of nodes, cost matrix $[c_{ijk}]_{n \times n \times p}$, time matrix $[t_{ijk}]_{n \times n \times p}$, t_{max}, p_s, p_c
Result: Minimum tour cost
1. Initialization of chromosomes
2. Set $gen \leftarrow 1$, *glob-best* = *loc-best* = $MAX\text{-}INT$
3. Selection procedure
4. **for** *i=1 to pop-size* **do**
   if $rand[0, 1] < p_c$ **then**
      i^{th} is selected for crossover
   end
end
5. procedure crossover among the mating pools

6. **for** $i = 1$ *to pop-size* **do**
   p_m=$1 - \frac{0.75}{\sqrt{gen}}$
   if $rand[0,1] < p_m$ **then**
      select i^{th} chromosome for mutation and mutation for vehicles
   end
end
7. Procedure mutation and mutation for vehicles
8. **for** $i = 1$ *to pop-size* **do**
   if $cost[i] < loc\text{-}best$ **then**
      $loc\text{-}best = cost[i]$;
      $mem = i$;
   end
end
9. $gen \leftarrow gen + 1$
10. **if** $loc\text{-}best < glob\text{-}best$ *and* $time[mem] < t_{max}$ **then**
   $glob\text{-}best \leftarrow loc\text{-}best$
end
11. **if** $gen < maxgen$ **then**
   goto step 3
else
   goto step 12
end
12. end

4.3 RID-IGA Algorithm for CCSTSPs

Ultimately, the algorithm of RID-IGA used for the solution of the proposed CCSTSPs is as follows:

Algorithm of RID-IGA for CCSTSP

Data: number of nodes n, distance matrix $[d_{ij}]_{n\times n}$, cost matrix $[c_{ijk}]_{n\times n\times p}$, time matrix $[t_{ijk}]_{n\times n\times p}$, t_{max}, covering distance matrix $[\Delta_i]_{1\times n}$

Result: complete tour with minimum cost such that visited nodes cover all unvisited nodes

1. solve the **USCP** for $[\Delta_i]_{1\times n}$ using $[\Delta_i]_{1\times n}$
2. **for** *i= 1 to total no. of USCP solutions* **do**
   $mincost \leftarrow CSTSP[USCP[i]]$
   if $mincost > CSTSP[USCP[i+1]]$ **then**
      $mincost \leftarrow CSTSP[USCP[i+1]]$
   end
end

5 Numerical Experiments

5.1 Verification with Earlier Results

To test the performance of the IGA implemented in C++ code, we consider some test problems from TSBLIB and compared with best known results of those in Table 1. Salari et al. [15] used the some TSP benchmark problems and solved using a hybrid algorithm consisting of Dynamic Programming and Ant Colony Optimization (ACO). Results of some of these problems are obtained by RID-IGA algorithm and are compared in Table 2.

Table 1. Algorithm tested with benchmark problems [16]:

Problem	Best known result	IGA result	Generation
gr17	2085	2085	169
fri26	937	937	246
bayg29	1610	1610	358
bays29	2020	2020	327
dantzig42	699	699	478
eil51	426	426	542
eil76	538	538	728
kroA100	21282	21638	1041

5.2 Proposed Experiment

For computational results of the proposed CCSTSP, we generate a 100×100 distance matrix with lower bound 20 and upper bound 90 and a $100 \times 100 \times 3$ costs with lower bound 35 and upper bound 180 and time matrices with lower bound 60 and upper bound 360 such that the vehicles with higher cost assume lesser time. The third dimension of the matrices imply the number of available vehicles at each node. For ICCSTSPC, distance, cost and time matrices are formed in fuzzy environment (TFN) with the same lower and upper bounds. At first, the USCP was solved taking the covering distance as 30 distance units for CCSTSP and (26,30,33) in case of ICCSTSPC for each node within a time bound 60 s. In the second step, CSTSP was solved for all obtained USCP solution in that time bound by IGA. The paths with minimum costs among all USCP results are considered as the near optimal solutions of the CCSTSP problem, some of which (5 best solutions without time constraint and 5 solutions with time constraints) are given in Table 3. In Table 4, the experimental results of ICCSTSPC are presented for different confidence levels, where both cases—without and with time constraints are considered, η, β and γ being the confidence levels of covering distance, travelling costs and travelling times respectively.

Table 2. Comparison with Salari et al.'s [15] results:

Problem	No. of nearest nodes	Salari's method	IGA result
eil51	7	164	**158**
	9	159	**157**
	11	**147**	149
berlin52	7	**3887**	3891
	9	3430	**3362**
	11	3262	**2832**
st70	7	**288**	292
	9	259	**241**
	11	247	**233**
eil76	7	207	**184**
	9	186	**173**
	11	170	**161**
pr76	7	**50275**	51277
	9	45348	**42916**
	11	43028	**42607**
rat99	7	486	**453**
	9	455	**441**
	11	444	**423**
kroA100	7	**9674**	10558
	9	9159	**8860**
	11	**8901**	9316

Table 3. Near optimal solutions of CCSTSP:

Optimized covering path (nodes/vehicles)	Cost	t_{max}
62/0 90/1 66/1 14/2 18/0 84/2 48/2 50/0 71/1 34/2 87/0 67/1	537.6	
52/0 22/2 2/1 50/1 11/0 15/1 47/1 7/0 80/2 61/1 34/0 6/1	585.7	
47/2 4/1 32/0 31/2 93/1 49/2 53/1 35/1 44/2 9/2 66/2 65/2	629.4	
44/0 1/0 66/2 74/0 48/2 30/1 93/1 49/2 76/0 40/0 29/2 19/1	630.25	
74/2 57/2 47/1 95/0 49/0 10/0 11/2 89/2 35/2 56/2 80/2 5/2	642.4	
87/1 62/0 90/2 71/1 50/2 67/2 66/1 14/2 18/1 48/1 34/1 84/0	554.45	1200
90/1 66/2 67/2 87/2 62/2 84/1 18/1 14/2 48/1 50/0 71/1 34/0	567.4	1190
34/1 71/1 50/2 62/2 67/2 66/1 14/0 87/1 48/0 84/2 18/1 90/1	569.6	1180
67/2 66/1 14/0 18/1 90/1 62/2 50/0 84/0 34/1 71/0 48/0 87/1	575	1175
84/2 18/1 14/2 66/2 34/0 71/1 50/0 48/2 87/2 67/0 62/0 90/1	587.25	1170

Table 4. Near optimal solutions of ICCSTSPC:

η	β	γ	Path	Cost	Fuzzy Cost	t_{max}
0.7	0.5		59/2 84/0 60/0 39/2 93/1 12/1 72/1 74/0 22/0 10/2 91/1 96/2	757	738.00 757.00 775.00	
	0.6		88/0 85/2 40/0 50/2 22/2 11/1 93/0 83/0 92/0 66/2 48/1 44/2	754.4	732.20 751.20 767.20	
	0.7		44/2 88/0 47/2 91/0 31/1 93/0 51/0 48/1 50/1 34/2 66/0 35/2	755.1	730.50 747.50 766.50	
	0.8		48/2 93/0 83/0 92/0 66/1 44/2 11/0 88/1 22/1 85/1 50/2 40/2	771.15	747.75 762.75 776.75	
	0.9		60/0 39/1 72/2 93/0 74/1 12/1 10/0 22/0 96/2 91/1 59/2 84/1	777.1	745.50 763.50 780.50	
	0.95		91/2 48/2 51/1 35/0 93/2 31/1 50/1 34/1 66/1 47/2 88/2 44/2	777.55	745.35 761.35 779.35	
0.8	0.5		87/1 22/2 33/2 93/2 18/2 84/2 83/1 6/0 34/2 61/1 24/0 32/0 44/2 26/2 17/2 35/1	941.35	918.35 941.35 963.35	
	0.6		17/1 93/2 18/2 26/2 44/2 33/2 22/0 35/1 83/0 61/2 34/2 87/1 24/0 32/2 84/0 6/0	937.6	909.20 933.20 955.20	
	0.7		6/2 34/2 61/1 83/0 84/0 87/0 35/2 44/1 26/2 18/2 17/1 22/0 24/0 32/2 93/0 33/0	954.15	922.55 944.55 968.55	
	0.8		61/0 34/0 32/2 6/2 84/1 35/1 22/2 33/0 17/2 87/2 83/0 44/2 26/2 18/1 93/2 24/0	982.8	942.60 966.60 993.60	
	0.9		18/2 24/1 22/2 33/0 35/0 83/1 6/0 93/2 17/2 87/0 26/2 44/0 32/2 61/0 34/2 84/2	962.7	921.30 944.30 967.30	
	0.95		6/0 34/2 61/2 24/2 93/2 35/1 87/0 84/0 32/0 44/2 26/2 18/2 17/0 83/1 22/2 33/1	937.6	892.00 916.00 940.00	
0.95	0.5		93/2 3/0 43/1 5/0 22/2 47/2 63/1 79/1 74/1 73/1 52/0 97/0 35/1 72/2 80/0 48/1 34/1 96/0 18/2	1179.6	1151.60 1179.60 1206.60	
	0.6		97/2 43/1 96/2 34/2 35/0 5/0 63/1 47/2 73/0 22/1 80/2 3/0 79/1 74/2 93/0 72/1 52/2 18/2 48/0	1167.9	1135.30 1162.30 1190.30	
	0.7		52/0 74/0 48/0 80/1 72/2 35/1 93/2 3/0 79/1 73/1 63/1 47/2 34/2 97/1 18/0 96/1 5/2 43/2 22/1	1172.25	1135.85 1161.85 1187.85	
	0.8		3/1 74/0 48/1 34/1 35/0 18/1 93/0 79/0 52/2 43/0 5/0 63/0 97/2 72/2 80/1 96/1 22/2 73/2 47/1	1209.55	1167.55 1194.55 1219.55	
	0.9		15/2 44/1 19/1 76/2 29/2 14/1 79/0 23/0 31/0 22/1 35/2 9/1 52/2 20/1 32/2 93/1 30/1 58/2 80/0	1206.25	1156.85 1183.85 1211.85	
	0.95		22/2 47/0 3/2 93/2 48/2 35/1 73/2 74/0 63/0 97/2 34/1 96/1 18/0 52/1 72/2 80/0 5/2 79/2 43/2	1174.6	1124.50 1148.50 1177.50	
	0.6	0.6	5/1 73/0 43/1 47/2 48/1 3/2 52/0 72/2 80/1 18/1 74/0 97/2 34/1 96/1 22/1 35/0 93/0 79/0 63/1	1170.6	1139.20 1165.20 1192.20	2420 2460 2490
	0.6	0.7	34/2 79/2 80/2 72/2 18/0 52/2 74/1 93/2 35/0 5/1 73/1 22/2 48/1 47/2 63/2 3/1 97/2 43/1 96/2	1187.5	1155.10 1182.10 1209.10	2420 2460 2490
	0.7	0.7	72/2 52/1 3/0 79/0 22/0 43/0 97/2 34/2 35/1 73/1 63/1 5/1 80/0 48/1 47/1 96/1 18/1 93/0 74/0	1185.85	1144.45 1173.45 1204.45	2560 2590 2610
	0.7	0.8	47/1 35/2 18/0 80/0 5/1 63/2 97/2 72/1 48/2 52/2 22/1 74/0 79/0 3/1 93/1 96/2 43/0 34/0 73/0	1188.1	1145.90 1174.90 1207.90	2560 2590 2610

6 Discussion

From Table 1, we observe that proposed IGA algorithm gives the best known results for the first seven TSP problems. For the last problem kroA100, with order size 100, the proposed algorithm results slightly higher value than the best known one within maximum number of generation 1200.

Table 2 presents the comparisons between RID-IGA and Salari et al.'s [15] hybrid ACO-Dynamic programming algorithm for some CSP problems, which are originated from TSP benchmark problems. Here, the distance and the cost matrices are the same as there is no choice of vehicles. In most of the cases, we get better results by the proposed RID-IGA.

In Table 3, five best near optimal paths without time constraints and five best paths imposing the time constraints of the proposed CCSTSP are given. As the maximum allowable time decreases the resulting optimal cost becomes higher, which is as per our expectation.

Finally, in Table 4, where the results of the proposed ICCSTSPC are discussed briefly, it can be observed that the number of nodes increase with the increment of the η, which is the confidence level of covering distance. We also notice that for each fixed value of η, as the value β (confidence level of travelling costs) varies, we get a path with minimum value at $\beta = 0.6$. Some results with different values of γ's (confidence level of traveling time) are shown in the same table.

7 Conclusion

In the present article, a fuzzy set based Imprecise Constrained Covering Solid Travelling Salesman Problem with Credibility along with a Combined method RID-IGA has been discussed. This problem can be well applicable for the most useful real-world problems like Rural Health Care Delivery Systems, Courier Logistics, big merchant houses, government officials and other similar problems. In these types of problems, it is not always possible to attend all the cities/villages of the network in consideration, but a few places are selected for the tour so that people from the adjoining areas within an approximate range r, i.e., within a range $(r_1 - \delta_1, r_1, r_1 + \delta_2)$, can avail the facilities. This uncertainty on covering distance has not been investigated by other researchers on covering salesman problem.

Also, there may be more than one vehicles at each node, from which, any one type of suitable one can be chosen by the salesperson. The travelling costs of the vehicles depend upon several factors like availability, sudden increment of fuel price etc. and similarly the travel time also may vary due to bad condition of road or vehicle, experience of driver etc., so the imprecise travelling costs and times are taken as fuzzy numbers.

The proposed model can further be extended by imposing some restrictions like mandatory inclusion or exclusion of some particular nodes, or inducing time windows on some nodes etc. Also, the proposed algorithm can be further developed by improving selection, crossover and mutation techniques of IGA.

References

1. Chang, T., Wan, Y., Tooi, W.: A stochastic dynamic travelling salesman problem with hard time windows. Eur. J. Oper. Res. **198**(3), 748–759 (2009)
2. Changdar, C., Maiti, M.K., Maiti, M.: A Constrained solid TSP in fuzzy environment: two heuristic approaches. Iranian J. Fuzzy Syst. **10**(1), 1–28 (2013)
3. Current, J.R., Schilling, D.A.: The covering salesman problem. Transp. Sci. **23**(3), 208–213 (1989)
4. Deb, K., Agarwal, R.B.: Simulated binary crossover for continuous search space. Complex Syst. **9**, 115–148 (1995)
5. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans. Evol. Comput. **6**(2), 182–197 (2002)
6. Golden, B.L., Naji-Azimi, Z., Raghavan, S., Salari, M., Toth, P.: The generalized covering salesman problem. INFORMS J. Comput. **24**(4), 534–553 (2012)
7. Hachicha, M., Hodgson, M.J., Laporte, G., Semet, F.: Heuristics for the multivehicle covering tour problem. Comput. Oper. Res. **27**, 29–42 (2000)
8. Khanra, A., Maiti, M.K., Maiti, M.: Profit maximization of TSP through a hybrid algorithm. Comput. Ind. Eng. **88**, 229–236 (2015)
9. Liu, B.: A survey of credibility theory. Fuzzy Optim. Decis. Making **5**, 387–408 (2006)
10. Maity, S., Roy, A., Maiti, M.: A modified genetic algorithm for solving uncertain constrained solid travelling salesman problems. Comput. Ind. Eng. **83**, 273–296 (2015)
11. Majumder, A.K., Bhunia, A.K.: Genetic algorithm for asymmetric traveling salesman problem with imprecise travel times. J. Comput. Appl. Math. **235**(9), 3063–3078 (2011)
12. Mestria, M., Ochi, L.S., Martins, S.L.: GRASP with path relinking for the symmetric Euclidean clustered traveling salesman problem. Comput. Oper. Res. **40**(12), 3218–3229 (2013)
13. Moon, C., Ki, J., Choi, G., Seo, Y.: An efficient genetic algorithm for the traveling salesman problem with precedence constraints. Eur. J. Oper. Res. **140**, 606–617 (2002)
14. Salari, M., Naji-Azimi, Z.: An integer programming-based local search for the covering salesman problem. Comput. Oper. Res. **39**, 2594–2602 (2012)
15. Salari, M., Reihaneh, M., Sabbagh, M.S.: Combining ant colony optimization algorithm and dynamic programming technique for solving the covering salesman problem. Comput. Ind. Eng. **83**, 244–251 (2015)
16. TSPLIB. http://elib.zib.de/pub/mp-testdata/tsp/tsplib/tsp/
17. Xudong, S., Yunlong, X.: An improved adaptive genetic algorithm. In: International Conference on Education Technology and Management Science (ICETMS) (2013)
18. Zhao, F., Sun, J., Li, S., Liu, W.: A hybrid genetic algorithm for the traveling salesman problem with pickup and delivery. Int. J. Autom. Comput. **6**(1), 97–102 (2009)

Newton Like Line Search Method Using q-Calculus

Suvra Kanti Chakraborty(✉) and Geetanjali Panda

Department of Mathematics, Indian Institute of Technology, Kharagpur 721 302, India
{suvrakanti,geetanjali}@maths.iitkgp.ernet.in

Abstract. In this paper some Newton like methods for unconstrained optimization problem are restructured using q-calculus (quantum calculus). Two schemes are proposed, (1) q-Newton line search scheme, (2) a variant of q-Newton line search scheme. Global convergence of these schemes are discussed and numerical illustrations are provided.

Keywords: q-derivative · Newton like method · Unconstrained optimization

1 Introduction

q-calculus (quantum calculus) has been one of the research interests in the field of Mathematics and Physics for last few decades. q-analogue of ordinary derivative, first introduced by F.H. Jackson, has its wide applications in several areas like, operator Theory [2], q-Taylor formula and its remainder [10,11], mean value theorems of q-calculus [16], fractional integral and derivatives [14], integral inequalities [7]. Some recent developments using q-derivatives can be found in variational calculus [3], transform calculus [1], sampling theory [12], q-version of Bochner Theorem [9], and so on. Soterroni et al. [17] first studied the use of q-derivative in the area of unconstrained optimization, which is the q-variant of steepest descent method. However, further significant works on q-calculus in other areas of numerical optimization viz. Newton, Quasi Newton, Conjugate gradient methods and their variations are yet to study.

In this paper a new variation of Newton like method for unconstrained optimization problem is developed using q-calculus. This concept is based on q-Newton Kantorvich scheme [15]. In this paper, initially, q-derivative of gradient of the given function is used to propose a local convergent scheme and then this idea is extended by associating a line search technique to justify its global convergence property. Next, a sequence $\{q_n\}$ is introduced in the scheme instead of considering a fixed positive number q, whose limiting case is the q-version of practical line search Newton scheme. Quadratic convergence of the first scheme is proved without using the second order sufficient optimality condition. First order differentiability is sufficient to prove the global convergence of the proposed schemes. The second scheme, being a q-analogue of line search Newton method, requires weaker conditions than the classical one.

© Springer Nature Singapore Pte Ltd. 2017
D. Giri et al. (Eds.): ICMC 2017, CCIS 655, pp. 196–208, 2017.
DOI: 10.1007/978-981-10-4642-1_17

In Sect. 2, some notations and definitions from q-calculus and other prerequisites are provided, which are used in sequel throughout the paper. q-Newton line search method is introduced and its convergence analysis is provided in Sect. 3. A variant of q-Newton line search method is developed further and numerical illustrations are described in Sect. 4. Finally, concluding remarks are provided in Sect. 5.

2 Prerequisites

2.1 Notations and Definitions on Quantum Calculus

For a function $f : \mathbb{R} \to \mathbb{R}$, the q-derivative ($q \neq 1$) of f (denoted by $D_{q,x} f$), is defined as

$$D_{q,x} f(x) = \begin{cases} \frac{f(x)-f(qx)}{(1-q)x}, & x \neq 0 \\ f'(x), & x = 0 \end{cases}$$

Suppose $f : \mathbb{R}^n \to \mathbb{R}$, whose partial derivatives exist. For $x \in \mathbb{R}^n$, consider an operator $\epsilon_{q,i}$ on f as

$$(\epsilon_{q,i} f)(x) = f(x_1, x_2, \ldots, x_{i-1}, qx_i, x_{i+1}, \ldots, x_n).$$

The q-partial derivative ($q \neq 1$) of f at x with respect to x_i, denoted by $D_{q,x_i} f$, is

$$D_{q,x_i} f(x) = \begin{cases} \frac{f(x)-(\epsilon_{q,i} f)(x)}{(1-q)x_i}, & x_i \neq 0 \\ \frac{\partial f}{\partial x_i}, & x_i = 0 \end{cases}$$

Denote $g(x) = \nabla f(x) = \left[g_1(x), g_2(x), \ldots, g_n(x)\right]^T$, $g_i = \frac{\partial f}{\partial x_i}$. The Jacobi matrix of q-partial derivative of $g(x)$, denoted by $D_q g(x)$ becomes

$$D_q g(x) = \begin{pmatrix} D_{q,x_1} g_1(x) & D_{q,x_2} g_1(x) & \ldots & D_{q,x_n} g_1(x) \\ D_{q,x_1} g_2(x) & D_{q,x_2} g_2(x) & \ldots & D_{q,x_n} g_2(x) \\ \ldots & \ldots & \ldots & \ldots \\ D_{q,x_1} g_n(x) & D_{q,x_2} g_n(x) & \ldots & D_{q,x_n} g_n(x) \end{pmatrix}_{n \times n}. \quad (1)$$

In short we write,

$$D_q g(x) = [D_{q,x_j} g_i(x)]_{n \times n}, \ i, j = 1(1)n.$$

2.2 Symmetric Indefinite Factorization

A real symmetric matrix A can be expressed as $PAP^T = LBL^T$, where L is a lower triangular matrix, P is a permutation matrix and B is a block diagonal matrix which allows at most 2×2 blocks. This requires a pivot block initially. There are several pivoting strategies available in the literature (see Bunch, Kaufman and Parlett [4], Golub and Van Loan [8], and also by Duff and Reid [5], Fourer and Mehrotra [6]) to take care the sparsity of the matrix. The symmetric indefinite factorization allows to determine the inertia of B and inertia of

B remains equal to inertia of A. An indefinite factorization can be modified to ensure that the modified factors are the factors of a positive definite matrix. This idea is briefed in the following algorithm (See [13]). For this purpose MATLAB in-built command $ldl()$ is used in this paper since it is less expensive.

Algorithm 1. Modifying Symmetric Indefinite Matrix to Positive Definite [13]

Step 1: Compute the factorization $PAP^T = LBL^T$.
Step 2: Perform the spectral decomposition of B as $B = Q\Lambda Q^T$, where Q is the matrix whose columns consist of eigen vectors and Λ is the diagonal matrix whose diagonal elements are respective eigen values B.
Step 3: Construct a modification matrix F such that LBL^T is sufficiently positive definite as follows.
Suppose λ_i are the eigen values of B. Choose parameter $\delta > 0$ and define F as $F = Q\,\mathrm{diag}(\tau_i)\,Q^T$, where

$$\tau_i = \begin{cases} 0, \text{ if } \lambda_i \geq \delta \\ \delta - \lambda_i, \text{ if } \lambda_i < \delta \end{cases}$$

Step 4: A matrix E has to be added to A to make it positive definite. $P(A+E)P^T = L(B+F)L^T$ provides $E = P^T LFL^T P$. So $\lambda_{\min}(A+E) \approx \delta$.
Output: $\bar{A} \triangleq A + E$ is the positive definite matrix.

2.3 Zoutendjik Theorem

Consider k^{th} iteration of an optimization algorithm in the form $x^{(k+1)} = x^{(k)} + \alpha_k p_k$, where p_k is a descent direction and α_k satisfies Wolfe condition. Suppose f is bounded below in $\mathbb{R}^n$ and that f is continuously differentiable in an open set containing the level set $\mathscr{L} = \{x : f(x) \leq f(x_0)\}$, where x_0 is the starting point of the iteration.

Assume also that ∇f is Lipschitz continuous on $\mathscr{L}$. That is, there exists a constant $L > 0$ such that

$$\|\nabla f(x) - \nabla f(\tilde{x})\| < L\|x - \tilde{x}\| \quad \forall x,\ \tilde{x} \in \mathscr{L},$$

then

$$\sum_{k \geq 0} \cos^2 \theta_k \|\nabla f^{(k)}\|^2 < \infty,$$

where θ_k is the angle between p_k and $\nabla f^{(k)}$, $\nabla f^{(k)} = \nabla f(x^{(k)})$.

3 Proposing q-Newton Line Search Scheme for Unconstrained Optimization Problem

Consider a general unconstrained optimization problem

$$(P) \quad \min_{x\in\mathbb{R}^n} f(x),$$

where $f \in \mathscr{C}^1$ and second order partial derivatives of f exist on $x_i = 0$. q-Taylor expansion [15] of $\nabla f(x)$ at the local minimum point x^* of (P) is

$$\left[\nabla f(x^*)\right]_i \approx \left[\nabla f_i(x^{(k)})\right]_i + \left[\sum_{j=1}^{n} D_{q,x_j} \nabla f_i(x^{(k)})(x_j^* - x_j^{(k)})\right]_i \quad (i = 1, 2, \ldots, n).$$

In matrix form, this can be expressed as

$$\nabla f(x^*) \approx \nabla f(x^{(k)}) + D_q \nabla f(x^{(k)})(x^* - x^{(k)}).$$

Assuming that the optimal solution is attained at $(k+1)^{th}$ iteration, i.e. $\nabla f(x^*) = 0$ and $x^* = x^{(k+1)}$, a modified Newton scheme may be considered as

$$x^{(k+1)} = x^{(k)} - [D_q \nabla f(x^{(k)})]^{-1} \nabla f(x^{(k)}), \tag{2}$$

where $D_q \nabla f(x^{(k)})$ can be derived as in Subsect. 2.1. The matrix $D_q f(x)$ is not necessarily symmetric. For example, consider a function $f(x, y) = xy^2 + x^4$. Then $\nabla f(x, y) = [y^2 + 4x^3, 2xy]^T$ and

$$D_q \nabla f(x) = \begin{pmatrix} 4x^2(1 + q + q^2) & y(1 + q) \\ 2y & 2x \end{pmatrix},$$

which is not symmetric for all q. We may consider the symmetric counterpart $\bar{D}_q$ of D_q as $\bar{D}_q = \frac{1}{2}(D_q + D_q^T)$. In addition to this $D_q \nabla f(x)$ may not be positive definite for some q. Positive definiteness of $D_q \nabla f$ will be discussed in next section. Here we assume the symmetric counterpart $\bar{D}_q \nabla f$ and the positive definiteness of $\bar{D}_q \nabla f$ in a local neighborhood of x^*.

The modified iteration scheme (2) may be expressed as

$$x^{(k+1)} = x^{(k)} - [\bar{D}_q \nabla f(x^{(k)})]^{-1} \nabla f(x^{(k)}). \tag{3}$$

Theorem 1. *Suppose q-partial derivatives of ∇f with respect to $x_j\,(j = 1, 2, \ldots, n)$ exist in a ball $N(x^*, R)$ for some $R > 0$ and x^* be the local optimum solution to the problem (P). Moreover, $\bar{D}_q \nabla f$ is positive definite at x^* and following two assumptions hold for some $M > 0$ and $\beta > 0$.*

A1. $\|\nabla f(x) - \nabla f(y) - \bar{D}_q \nabla f(y)(x - y)\| \le M\|x - y\|^2$,

A2. $\|\bar{D}_q \nabla f(x^{(k)})^{-1}\| \le \beta$.

Then the sequence $\{x^{(k)}\}$ described in (3) converges to the solution x^* quadratically and $\|\nabla f(x)\|$ vanishes quadratically in the vicinity of x^*.

Proof. Consider a point $x^{(k+1)} = x^{(k)} + p_k$ in the vicinity of x^* along the direction $p_k \in \mathbb{R}^n$. Then

$$\begin{aligned} x^{(k+1)} - x^* &= x^{(k)} - x^* - \bar{D}_q \nabla f(x^{(k)})^{-1} \nabla f(x^{(k)}) \\ &= \bar{D}_q \nabla f(x^{(k)})^{-1} [\bar{D}_q \nabla f(x^{(k)})(x^{(k)} - x^*) - \nabla f(x^{(k)})]. \end{aligned}$$

$$\begin{aligned} \|x^{(k+1)} - x^*\| &\leq \|\bar{D}_q \nabla f(x^{(k)})^{-1}\| \, \|\bar{D}_q \nabla f(x^{(k)})(x^{(k)} - x^*) - \nabla f(x^{(k)})\| \\ &= \|\bar{D}_q \nabla f(x^{(k)})^{-1}\| \, \|\nabla f(x^*) - \nabla f(x^{(k)}) - \bar{D}_q \nabla f(x^{(k)})(x^* - x^{(k)})\| \\ &\leq \beta.M.\|x^* - x^{(k)}\|^2 \quad \text{(by assumptions)} \end{aligned}$$

Above inequality guarantees the quadratic convergence of the scheme. Since $\bar{D}_q \nabla f$ is positive definite at x^*, so there exists some $R' > 0$ such that $\bar{D}_q \nabla f$ is positive definite in the neighborhood $N_1(x^*, R')$. If the above process is repeated, we get

$$\|x^{(k+1)} - x^*\| \leq (\beta M)^{2^{k-1}} \|x^* - x^{(0)}\|^{2^k}.$$

So the initial point $x^{(0)}$ may be chosen in such a way that $x^{(0)} \in N_2(x^*, \min\big(R, R', \frac{1}{2\beta M}\big))$ to achieve quadratic order convergence.

From (3), $\bar{D}_q \nabla f(x^{(k)})(x^{(k+1)} - x^{(k)}) = -\nabla f(x^{(k)})$. Hence,

$$\begin{aligned} \|\nabla f(x^{(k+1)})\| &= \|\nabla f(x^{(k+1)}) - \nabla f(x^{(k)}) - \bar{D}_q \nabla f(x^{(k)}) p_k\| \\ &\leq M.\|x^{(k+1)} - x^{(k)}\|^2 \\ &= M.\|\bar{D}_q \nabla f(x^{(k)})^{-1} \nabla f(x^{(k)})\|^2 \\ &\leq M.\|\bar{D}_q \nabla f(x^{(k)})^{-1}\| \, \|\nabla f(x^{(k)})\|^2 \\ &= M.\beta.\|\nabla f(x^{(k)})\|^2. \end{aligned}$$

This proves that the gradient norm vanishes quadratically in the vicinity of x^*.

Note: One may note that this local scheme does not demand the existence of second order partial derivatives of f except at the points on $x_i = 0$. In the following optimization problem Newton scheme can not be applied, but the proposed scheme can be applied efficiently.

Example 1. Consider $\min_{(x,y)\in\mathbb{R}^2} f_1(x, y)$, where

$$f_1(x, y) = \begin{cases} (x-1)^3 \sin \frac{1}{x-1} + (x-1)^2 + x(y-1)^4, & \text{if } x \neq 1 \\ (y-1)^4, & \text{if } x = 1 \end{cases}$$

$f_1(x, y)$ attains the minimum at $(1, 1)$ (see Fig. 1). Since $\frac{\partial^2 f_1}{\partial x^2}$ does not exist at $(1, 1)$, f_1 is not twice differentiable. So second order sufficient conditions can not be applied to justify the existence of the minimum point as in the case of higher order numerical optimization methods. Hence, Newton method can not

be applied. But, the proposed scheme can be applied as described below. Here, for $q \neq 1$,

$$\bar{D}_q \nabla f_1(1,1) = \begin{bmatrix} 3(q-1)\sin\frac{1}{q-1} & \frac{(q-1)^3}{2} \\ \frac{(q-1)^3}{2} & 4(q-1)^2 \end{bmatrix}.$$

Let $I = \{q \in \mathbb{R} \mid 3(q-1)\sin\frac{1}{q-1} > 0,\ 12\sin\frac{1}{q-1} > \frac{(q-1)^3}{4}\}$. If we choose $q \in (0,1) \cap (1+\frac{1}{2k\pi}, 1+\frac{1}{(2k+1)\pi})$ for $k \in \mathbb{Z}$, then $\bar{D}_q f_1(1,1)$ is positive definite. On Matlab R-2013b platform, with several initial points, tolerance limit of the gradient norm as 10^{-5}, the proposed scheme (3) reaches to the solution (1,1). Results are summarized in Table 1 for several q, $(q = 0.85, 0.87, 0.89, 0.93, 0.95)$ with same set of different initial guesses and a pictorial illustration is provided in Figs. 1 and 2.

Note: Scheme (3) has following assumptions, which may be burden to the decision maker and hence the scheme should be further modified.

- The initial points are selected in the vicinity of the solution. Hence this scheme is further extended to a global convergent scheme in Subsect. 3.1, which is free from the choice of initial point.
- Selection of q is difficult. To avoid this, the global convergent scheme in Subsect. 3.1 is further modified in Sect. 4, where any sequence $\{q_k\}$ with some mild property is chosen instead of fixed q.

3.1 Global Convergence Property of the Proposed Scheme

The iterative scheme (3) has local convergence property. To achieve global convergence of the proposed scheme under some mild conditions, a line search may be associated with every iterating point, starting with any initial point. In the

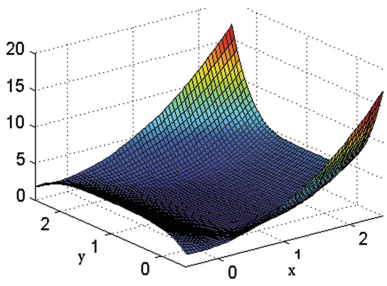


Fig. 1. Surface plot of $f_1(x, y)$

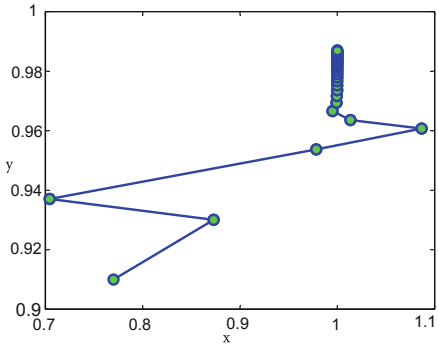


Fig. 2. Iteration points for $f_1(x, y)$ with initial guess (.77, .91)

Table 1. q-Newton iteration scheme (3) for $f_1(x, y)$

q	Initial guess	Number of iterations	Final gradient norm
.85	(.82,.85)	92	9.9104e−06
	(.72, .95)	77	9.8387e−06
	(.77, .91)	88	9.8560e−06
	(.97,.81)	94	9.8291e−06
	(1.1,.9)	90	9.9379e−06
.87	(.82,.85)	74	9.8573e−06
	(.72, .95)	24	9.9638e−06
	(.77, .91)	46	9.7381e−06
	(.97,.81)	33	9.7424e−06
	(1.1,.9)	31	9.5994e−06
.89	(.82,.85)	58	9.7938e−06
	(.72, .95)	47	9.9908e−06
	(.77, .91)	53	9.8469e−06
	(.97,.81)	59	9.9094e−06
	(1.1,.9)	56	9.9792e−06
.93	(.82,.85)	94	9.8214e−06
	(.72, .95)	24	9.9638e−06
	(.77, .91)	46	9.7381e−06
	(.97,.81)	33	9.7424e−06
	(1.1,.9)	31	9.5994e−06
.95	(.82,.85)	35	5.2470e−06
	(.72, .95)	17	9.2732e−06
	(.77, .91)	19	9.7412e−06
	(.97,.81)	23	9.9306e−06
	(1.1,.9)	21	9.7126e−06

local scheme, described in Sect. 3, $\bar{D}\nabla f$ is a symmetric matrix but not necessarily a positive definite matrix. For global convergence one needs to have a positive definite matrix. Consider the following q-Newton line search scheme as

$$x^{(k+1)} = x^{(k)} - \alpha_k (T_q^{(k)})^{-1} \nabla f(x^{(k)}), \tag{4}$$

where $T_q^{(k)}$ is a positive definite approximation of the matrix $\bar{D}_q \nabla f(x^{(k)})$ and α_k is the step length at $x^{(k)}$. $\bar{D}_q \nabla f(x^{(k)})$ is the symmetric counterpart of $D_q \nabla f(x^{(k)})$.

$D_q \nabla f$ as well as $\bar{D}_q \nabla f$ may not be positive definite at $x^{(k)}$. We need the matrix $T_q^{(k)}$ to be positive definite, which may be achieved using symmetric indefinite factorization, described in Subsect. 2.2. $\bar{D}_q \nabla f(x^{(k)})$ can be modified to $T_q^{(k)}$ as

$$T_q^{(k)} = \bar{D}_q \nabla f(x^{(k)}) + E^{(k)},$$

where $E^{(k)}$ is the matrix, which is added to $\bar{D}_q \nabla f(x^{(k)})$ to force $T_q^{(k)}$ to be positive definite.

Lemma 1. *Let $\kappa(T_q^{(k)})$ denotes the condition number of $T_q^{(k)}$. If there exists some $C > 0$ such that $\kappa(T_q^{(k)}) < C$ for every k, then under all the standard assumption of Zoutendjik Theorem, $\|\nabla f(x^k)\| \to 0$ as $k \to 0$.*

Proof. Let the eigenvalues of $T_q^{(k)}$ be $0 < \lambda_1^{(k)} \leq \lambda_2^{(k)} \leq \ldots \leq \lambda_n^{(k)}$. Since $\lambda_1^{(k)}$ is the smallest eigenvalue of $T_q^{(k)}$, for any $u \in \mathbb{R}^n$,

$$u^T T_q^{(k)} u \geq \lambda_1^{(k)} \|u\|^2.$$

Let θ_k be the angle between p_k and $\nabla f^{(k)}$, where $p_k = -(T_q^{(k)})^{-1} \nabla f^{(k)}$. Hence,

$$\cos\theta_k = -\frac{\nabla f^{(k)^T} p_k}{\|\nabla f^{(k)^T}\| \|p_k\|} = \frac{p_k^T T_q^{(k)} p_k}{\|\nabla f^{(k)^T}\| \|p_k\|} \geq \lambda_1^{(k)} \frac{\|p_k\|}{\|\nabla f^{(k)}\|}. \tag{5}$$

Again $\|\nabla f^{(k)}\| = \|T_q^{(k)} p_k\| \leq \|T_q^{(k)}\| \, \|p_k\| = \lambda_n^{(k)} \|p_k\|$. Using this in (5), we have

$$\cos\theta_k = -\frac{\nabla f^{(k)^T} p_k}{\|\nabla f^{k(T)}\| \, \|p_k\|} \geq \frac{\lambda_1^{(k)}}{\lambda_n^{(k)}} = \frac{1}{\|T_q^{(k)}\| \|T_q^{(k)^{-1}}\|} \geq \frac{1}{C}.$$

Hence, under the assumption of Zoutendjik condition (that is, $\cos^2\theta_k \|\nabla f^{(k)}\|^2 \to 0$), i.e. $\lim_{k\to\infty} \|\nabla f^{(k)}\| = 0$.

Lemma 1 justifies that q-Newton line search scheme (4) converges to a critical point. It is more likely that as the functional value reduces at every iteration, the scheme converges to a local minimum. However, the convergence rate of the scheme can be justified when $x^{(k)}$ approaches to the solution as $k \to \infty$. In the vicinity of the solution, α_k may be chosen as unit length. Moreover, in the vicinity of the solution, $\bar{D}_q \nabla f(x^{(k)})$ being positive definite, so, for sufficiently large k, $T_q^{(k)} = \bar{D}_q \nabla f(x^{(k)})$. Above discussion may be summarized as the following Algorithm.

Algorithm 2. q-Newton Scheme with Line Search for Unconstrained Optimization

Choose starting point $x^{(0)}$, tolerance limit ϵ, $k = 0$, fix $q > 0$;
for $k = 0, 1, 2 \ldots$
  Compute $\bar{D}_q \nabla f(x^{(k)})$;
  Compute $T_q^{(k)} = \bar{D}_q \nabla f(x^{(k)}) + E_k$ by Algorithm 1;
  $x^{(k+1)} = x^{(k)} - \alpha_k (T_q^{(k)})^{-1} \nabla f(x^{(k)})$, α_k is computed by Armijo-Backtracking inexact line search;
  if $\|\nabla f(x^{(k+1)})\| < \epsilon$
    Stop;
   else
      $k = k + 1$;
  end;
end;

The line search scheme (4) is an extension to global convergent version of the local convergent scheme (3). In Example 1, initial points were chosen very close to the solution. Here, for the same objective function, one may choose the initial point not necessarily in the vicinity of the solution and can apply Algorithm 2. At the initial point $(1.6, 4)$ for $q = 0.95$, $\bar{D}_q \nabla f_1 = \begin{pmatrix} 11.2535 & 835.3760 \\ 835.3760 & 161.5360 \end{pmatrix}$, which is not positive definite. For this initial point and backtracking factor 0.7 in Armijo-backtracking inexact line search with terminating condition $\|\nabla f_1\| < 10^{-5}$, 35 iterations are required to reach at the solution. The result is summarized in Table 2. One may observe that

- $\bar{D}_q \nabla f$ is not necessarily positive definite up to 23^{rd} iteration.
- after 23^{rd} iteration $\bar{D}_q \nabla f(x^{(k)})$ is positive definite which indicates that the iterating points are entering in the neighborhood of the minimum point (1,1) after 23^{rd} iteration.
- the matrix E_k corresponding to the positive definite $\bar{D}_q \nabla f(x^{(k)})$ is a null matrix.

Note : Both the iterating schemes (3) and (4) do not require the second order partial derivatives of f over the whole domain. If we further consider $f \in \mathscr{C}^2$ only in the vicinity of the solution, not necessarily in the whole domain of f, then scheme (3) behaves almost like practical Newton method. This is justified in next section for which a sequence $\{q_k\}$ is associated to the scheme instead of a fixed q at each iteration. We say this new scheme as variant of q-Newton line search scheme.

Table 2. q-Newton iteration line search scheme (Algorithm 2) for $f_1(x, y)$

k	$x^{(k)}$	$f(x^{(k)})$	$\bar{D}_q \nabla f(x^{(k)})$	E_k
0	$\begin{pmatrix} 1.6 \\ 4 \end{pmatrix}$	130.1750	$\begin{pmatrix} 11.2535 & 835.3760 \\ 835.3760 & 161.5360 \end{pmatrix}$	$\begin{pmatrix} 409.8776 & -374.6643 \\ -374.6643 & 342.4763 \end{pmatrix}$
1	$\begin{pmatrix} 5.2976 \\ 0.6200 \end{pmatrix}$	36.8833	$\begin{pmatrix} 112.2638 & -0.1573 \\ -0.1573 & 9.9466 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$
...	...	...	...	...
23	$\begin{pmatrix} 1.1828 \\ 0.9797 \end{pmatrix}$	.0290	$\begin{pmatrix} -3.6318 & -.0002 \\ -.0002 & .0314 \end{pmatrix}$	$\begin{pmatrix} 3.6318 & .0002 \\ .0002 & 0 \end{pmatrix}$
24	$\begin{pmatrix} 0.9192 \\ 0.9797 \end{pmatrix}$	.0064	$\begin{pmatrix} 4.1832 & -.0002 \\ -.0002 & 0.0244 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$
...	...	...	...	...
34	$\begin{pmatrix} 1.0000 \\ 0.9877 \end{pmatrix}$	2.2845e−08	$\begin{pmatrix} 1.7265 & -0.0001 \\ -0.0001 & .0189 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$
35	$\begin{pmatrix} 1.0000 \\ 0.9881 \end{pmatrix}$	2.0017e−08	$\begin{pmatrix} 1.7287 & -0.0001 \\ -0.0001 & .0185 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$

4 A Variant of q-Newton Line Search Method

In general the Newton direction (p_k^N) at $x^{(k)}$ satisfies $\nabla^2 f(x^{(k)})p_k^N = -\nabla f(x^{(k)})$. Since the Hessian matrix $\nabla^2 f$ may not be positive definite at the points, away from the solution of (P), so p_k^N may not be a descent direction. There are several approaches to make the Hessian positive definite. q-analogue of this practical Newton scheme is developed here.

- $f \in \mathscr{C}^1$ and ∇f is Lipschitz continuous.
- $f \in \mathscr{C}^2$ in the vicinity of the solution.

Lemma 2. *Let $\{q_k\}$ be a real sequence defined by $q_{k+1} = 1 - \frac{q_k}{(k+1)^2}$, with $0 < q_0 < 1$, a fixed number, then q_k converges to 1 as $k \to \infty$.*

The matrix $\bar{D}_{q_k}\nabla f(x^{(k)})$ is computed at the iterating point $x^{(k)}$, If

- $\bar{D}_{q_k}^{(k)}\nabla f(x(k))$ is modified as $T_{q_k}^{(k)} = \bar{D}_{q_k}^{(k)}\nabla f(x^{(k)}) + E_k$ for some matrix E_k such that $T_{q_k}^{(k)}$ becomes positive definite, and
- E_k is computed using symmetric indefinite factorization as described in Subsect. 2.2,

then the modified direction is the solution of the system $T_{q_k}^{(k)}p_k = -\nabla f(x^{(k)})$ and the corresponding scheme can be expressed as

$$x^{(k+1)} = x^{(k)} - \alpha_k (T_{q_k}^{(k)})^{-1}\nabla f(x^{(k)}). \tag{6}$$

This scheme differs from the scheme (4) in the sense that it uses the sequence $\{q_k\}$ instead of a fixed q. Following algorithm explores this concept.

Algorithm 3. Variant of q-Newton line search method

Choose starting point $x^{(0)}$, tolerance limit ϵ, $k = 0$, fix $q_0 \in (0, 1)$;
for $k = 0, 1, 2, \ldots$
  Compute $\bar{D}_q \nabla f(x^{(k)})$;
  Compute $T_{q_k}^{(k)} = \bar{D}_{q_k}^{(k)}\nabla f(x^{(k)}) + E_k$ by Algorithm 1;
  $x^{(k+1)} = x^{(k)} - \alpha_k (T_{q_k}^{(k)})^{-1}\nabla f(x^{(k)})$, α_k is computed by Armijo-Backtracking inexact line search;
  if $\|\nabla f(x^{(k+1)})\| < \epsilon$
    Stop;
    else
      $q_{k+1} = 1 - \frac{q_k}{(k+1)^2}$;
      $k = k + 1$;
  end;
end;

4.1 Convergence Analysis

Convergence proof of Algorithm 3 is similar to that of Algorithm 2 under Zoutendzik condition with the assumption of bounded condition number of the matrix $T_{q_k}^{(k)}$. Assuming the sequence $x^{(k)}$ converges to x^*, we discuss the following convergence result. The $(i,j)^{th}$ entry of $D_{q_k}\nabla f$ is $D_{q_k,x_j}\frac{\partial f}{\partial x_i}$. In the vicinity of x^*,

$$\lim_{k\to\infty}\bar{D}_{q_k}\nabla f = \lim_{q_k\to 1}\bar{D}_{q_k\to 1}\nabla f = \nabla^2 f.$$

Hence, for a local minimum point x^*, $\bar{D}_{q_k}\nabla f(x)$ is positive definite for sufficiently large k and $x \in \mathrm{Nbd}(x^*)$. So in limiting case the variant of q-Newton line search scheme reduces to Newton algorithm.

4.2 Numerical Example for Global Convergent Schemes

Consider the following optimization problem $\min_{(x,y)\in\mathbb{R}^2} f_2(x,y)$, where

$$f_2(x,y) = \begin{cases} 100(y-x^2)^2 + (1-x)^2 + c, & x \geq c, \\ \frac{x}{c}(1-x)^2 + 100(y-x^2)^2 - \frac{(1-c)^2}{c}(x-c) + c, & x < c \end{cases} \tag{7}$$

$f_2 \in \mathbb{C}^1$ and second order partial derivative of f_2 with respect to x does not exist at $x = -1.2, y \in \mathbb{R}$ and $c = -1.2$. So in general $f_2 \notin \mathbb{C}^2$. However, $f_2 \in \mathbb{C}^2$ in the vicinity of the minimum point $(1,1)$. So, for some initial points (viz. $(-1.2,1)$), the practical Newton line search can not be applied where as, q-Newton line search (Algorithm 2) and variant of q-Newton line search (Algorithm 3) can be applied. For Algorithm 2, q is fixed, say $q = 0.999$ and for Algorithm 3, a sequence q_k is considered, $q_k = 1 - \frac{q_{k-1}}{k^2}$ with $q_0 = 0.95$. Both the Algorithms are executed with same initial point $(x^{(0)}, y^{(0)}) = (-1.2, 1)$, terminating condition $\|\nabla f_2\| < 10^{-5}$. Backtracking factor and Armijo parameter are chosen to be 0.7 and 10^{-4} respectively. Solution of (7) is attained in 35 iterations in case of Algorithm 2 and 34 iterations in case of Algorithm 3. These are pictorially illustrated in Figs. 3 and 4.

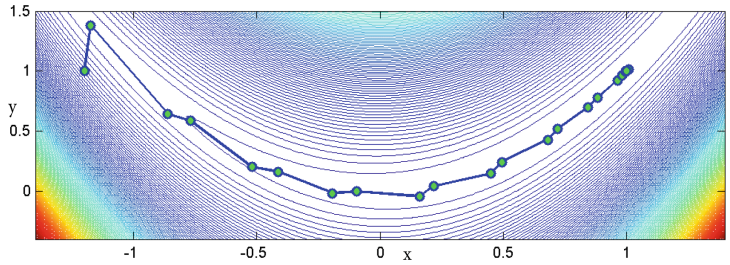


Fig. 3. Algorithm 2 for $f_2(x,y)$ with initial guess $(-1.2, 1)$

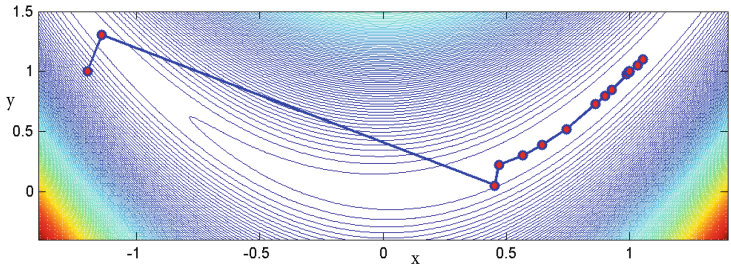


Fig. 4. Algorithm 3 for $f_2(x, y)$ with initial guess $(-1.2, 1)$

5 Conclusion

In this paper quantum calculus is used to develop Newton like schemes for unconstrained optimization problems, for which existence of second order partial derivatives at every point is not required. Further a variant of this line search scheme is proposed which behaves like practical Newton line search method in limiting case. The global convergence of both schemes have been discussed with numerical examples. The authors hope that this concept may be further modified for other numerical optimization schemes.

References

1. Abreu, L.: A q-sampling theorem related to the q-Hankel transform. Proc. Am. Math. Soc. **133**(4), 1197–1203 (2005)
2. Aral, A., Gupta, V., Agarwal, R.P.: Applications of q-Calculus in Operator Theory. Springer, New York (2013)
3. Bangerezako, G.: Variational q-calculus. J. Math. Anal. Appl. **289**(2), 650–665 (2004)
4. Bunch, J.R., Kaufman, L., Parlett, B.N.: Decomposition of a symmetric matrix. Numer. Math. **27**(1), 95–109 (1976)
5. Duff, I.S., Reid, J.K.: The multifrontal solution of indefinite sparse symmetric linear. ACM Trans. Math. Softw. (TOMS) **9**(3), 302–325 (1983)
6. Fourer, R., Mehrotra, S.: Solving symmetric indefinite systems in an interior-point method for linear programming. Math. Program. **62**(1–3), 15–39 (1993)
7. Gauchman, H.: Integral inequalities in q-calculus. Comput. Math. Appl. **47**(2), 281–300 (2004)
8. Golub, G.H., Van Loan, C.F.: Matrix Computations, vol. 3. JHU Press, Baltimore (2012)
9. Grünbaum, F.A., Haine, L.: The q-version of a theorem of bochner. J. Comput. Appl. Math. **68**(1), 103–114 (1996)
10. Ismail, M.E., Stanton, D.: Applications of q-Taylor theorems. J. Comput. Appl. Math. **153**(1), 259–272 (2003)
11. Jing, S.C., Fan, H.Y.: q-Taylor's formula with its q-remainder. Commun. Theor. Phys. **23**(1), 117 (1995)
12. Koornwinder, T.H., Swarttouw, R.F.: On q-analogues of the Fourier and Hankel transforms. Trans. Am. Math. Soc. **333**(1), 445–461 (1992)

13. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer Series in Operations Research and Financial Engineering, 2nd edn. Springer, New York (2006)
14. Rajkovic, P.M., Marinkovi, S.D., Stankovic, M.S.: Fractional integrals and derivatives in q-calculus. Appl. Anal. Discrete Math. **1**(1), 311–323 (2007)
15. Rajkovic, P.M., Marinkovic, S.D., Stankovic, M.S.: On q-Newton Kantorovich method for solving systems of equations. Appl. Math. Comput. **168**(2), 1432–1448 (2005)
16. Rajkovic, P.M., Stankovic, M.S., Marinkovic, S.D.: Mean value theorems in q-calculus. Matematicki Vesnik **54**, 171–178 (2002)
17. Soterroni, A.C., Galski, R.L., Ramos, F.M.: The q-gradient vector for unconstrained continuous optimization problems. In: Hu, B., Morasch, K., Pickl, S., Siegle, M. (eds.) Operations Research Proceedings 2010, pp. 365–370. Springer, Heidelberg (2011). doi:10.1007/978-3-642-20009-0_58

Existence Results of a Generalized Mixed Exponential Type Vector Variational-Like Inequalities

N.K. Mahato[1(✉)] and R.N. Mohapatra[2]

[1] Indian Institute of Information Technology, Design and Manufacturing (IIITDM), Jabalpur, India
nihariitkgp@gmail.com
[2] Central University Florida, Orlando, USA
Ram.Mohapatra@ucf.edu

Abstract. In this paper, we introduce a new generalized mixed exponential type vector variational-like inequality problems (GMEVVLIP) and α-relaxed exponentially (p, η)-monotone mapping. We prove the existence results of (GMEVVLIP) by utilizing the KKM technique and Nadlar's results with α-relaxed exponentially (p, η)-monotone mapping in Euclidian spaces. The present work extends some corresponding results of (GMEVVLIP) [1].

Keywords: Generalized mixed exponential type vector variational-like inequality problems · α-relaxed exponentially (p, η)-monotonicity · KKM mappings

2010 Mathematics Subject Classification: 47H04 · 47H05 · 90C33

1 Introduction

Due to the wide range of applicability of vector variational inequality problem, it has been generalized in many directions and established the existence results under various conditions (see [2–6]). Exponential type vector variational-like inequality problems was introduced by [7,8] with exponential type invexities. In this paper, we have define a very new vector variational inequality problem namely generalized mixed exponential type vector variational-like inequality problems which involved an exponential type function.

In the study of vector variational inequality problem the generalized monotonicity assumption of the operator plays a very important role. Wu and Huang [9] defined the concepts of relaxed η-α pseudomonotone mappings to study vector variational-like inequality problem in Banach spaces. Ceng and Yao [10] considered generalized variational-like inequalities with generalized α-monotone multifunctions. In 2009, Usman and Khan [1] discussed the solvability of the generalized mixed vector variational-like inequality problem with relaxed $\eta - \alpha$-P-monotone mappings. Very recently, Plubtieng and Thammathiwat [11]

© Springer Nature Singapore Pte Ltd. 2017
D. Giri et al. (Eds.): ICMC 2017, CCIS 655, pp. 209–220, 2017.
DOI: 10.1007/978-981-10-4642-1_18

considered new generalized mixed vector variational-like inequalities and studied the existence of solution of the same problem with C-monotonicity assumption. In this paper we introduced a very new generalized monotone mapping called α-relaxed exponentially (p,η)-monotone mappings.

Inspired and motivated by [1,7,8,11], we introduce a more general problem generalized mixed exponential type vector variational-like inequality problems in $\mathbb{R}^n$. We introduce a new generalized mixed exponential type vector variational-like inequality problems (GMEVVLIP) and α-relaxed exponentially (p,η)-monotone mapping. We prove the existence results of (GMEVVLIP) by utilizing the KKM technique and Nadlar's results with α-relaxed exponentially (p,η)-monotone mapping in $\mathbb{R}^n$. The results presented here, are extension and improvement of some previous results in the literature.

2 Preliminaries

Let $Y = \mathbb{R}^n$ be a Euclidian space and C be a nonempty subset of Y. C is called a cone if $\lambda C \subset C$, for any $\lambda \geq 0$. Further, the cone C is called convex cone if $C + C \subset C$. C is pointed cone if C is cone and $C \cap (-C) = \{0\}$. C is said to be proper cone, if $C \neq Y$. Now, consider $C \subseteq Y$ is a pointed closed convex cone with $intC \neq \emptyset$ with apex is at origin, where $intC$ is the set of interior points of C. Then, C induced a vector ordering in Y as follows:

(i) $x \leq_C y \Leftrightarrow y - x \in C$;
(ii) $x \nleq_C y \Leftrightarrow y - x \notin C$;
(iii) $x \leq_{intC} \Leftrightarrow y - x \in intC$;
(iv) $x \nleq_{intC} y \Leftrightarrow y - x \notin intC$.

By (Y, C), we denote an ordered space with the ordering of Y defined by set C. It is obvious that the ordering relation "$\leq_C$" defined above, is a partial order. The following properties are elementary:

(i) $x \nleq_C y \Leftrightarrow x + z \nleq_C y + z$, for any $x, y, z \in Y$;
(ii) $x \nleq_C y \Leftrightarrow \lambda x \nleq_C \lambda y$, for any $\lambda \geq 0$.

Let $K \subseteq X$ be nonempty closed convex subset of a Euclidian space $X = \mathbb{R}^m$ and (Y, C) be an ordered space induced by the closed convex pointed cone C whose apex at origin with int$C \neq \emptyset$. The following definitions and lemmas will be useful in the sequel.

Lemma 2.1 ([10])**.** Let (Y, C) be an ordered space induced by the pointed closed convex cone C with $intC \neq \emptyset$. Then for any $x, y, z \in Y$, the following relationships hold:

$$z \nleq_{intC} x \geq_C y \;\Rightarrow\; z \nleq_{intC} y;$$
$$z \ngeq_{intC} x \leq_C y \;\Rightarrow\; z \ngeq_{intC} y.$$

Definition 2.1. $f : X \to Y$ is C-convex on X if $f(tx + (1-t)y) \leq_C tf(x) + (1-t)f(y)$, for all $x,\ y \in X$, $t \in [0,1]$.

Definition 2.2. A mapping $f : K \to Y$ is said to be completely continuous if for any sequence $\{x_n\} \in K$, $x_n \rightharpoonup x_0 \in K$ weakly, then $f(x_n) \to f(x_0)$.

Definition 2.3. Let $f : K \to 2^X$ be a set-valued mapping. Then f is said to be KKM mapping if for any $\{y_1, y_2, \ldots, y_n\}$ of K we have $co\{y_1, y_2, \ldots, y_n\} \subset \bigcup_{i=1}^{n} f(y_i)$, where $co\{y_1, y_2, \ldots, y_n\}$ denotes the convex hull of $y_1, y_2, \ldots, y_n$.

Lemma 2.2 ([12])**.** Let M be a nonempty subset of a Hausdorff topological vector space X and let $f : M \to 2^X$ be a KKM mapping. If $f(y)$ is closed in X for all $y \in M$ and compact for some $y \in M$, then $\bigcap_{y \in M} f(y) \neq \emptyset$.

Lemma 2.3 ([13])**.** Let E be a normed vector space and H be a Hausdorff metric on the collection $CB(E)$ of all closed and bounded subsets of E, induced by a metric d in terms of $d(x, y) = \|x - y\|$ which is defined by

$$H(A, B) = \max\left(\sup_{x \in A} \inf_{y \in B} \|x - y\|, \sup_{y \in B} \inf_{x \in A} \|x - y\|\right),$$

for A, $B \in CB(E)$. If A and B are any two members in $CB(E)$, then for each $\epsilon > 0$ and each $x \in A$, there exists $y \in B$ such that

$$\|x - y\| \leq (1 + \epsilon)H(A, B).$$

In particular, if A and B are compact subset in E, then for each $x \in A$, there exists $y \in B$ such that

$$\|x - y\| \leq H(A, B).$$

Definition 2.4. Let $\eta : X \times X \to X$ be a bi-mapping and $A : K \to L(X, Y)$ be a single-valued mapping, where $L(X, Y)$ be space of all continuous linear mappings from X to Y. Suppose $T : K \to 2^{L(X,Y)}$ be the nonempty compact set-valued mapping, then

(i) A is said to be η-hemicontinuous if $\lim_{t \to 0^+} \langle A(x + t(y - x)), \eta(y, x)\rangle = \langle Ax, \eta(y, x)\rangle$, for each $x, y \in K$.

(ii) T is said to be H-hemicontinuous, if for any given $x, y \in K$, the mapping $t \to H(T(x + t(y - x)), Tx)$ is continuous at 0^+, where H is the Hausdorff matric defined on $CB(L(X, Y))$.

Definition 2.5. A mapping $f : X \to X$ is said to be affine if for any $x_i \in K$ and $\lambda_i \geq 0$, $(1 \leq i \leq n)$, with $\sum_{i=1}^{n} \lambda_i = 1$, we have

$$f\left(\sum_{i=1}^{n} \lambda_i x_i\right) = \sum_{i=1}^{n} \lambda_i f(x_i).$$

Definition 2.6. Let X be a Euclidian space. A function $f : X \to \mathbb{R}$ is lower semicontinuous at $x_0 \in X$ if

$$f(x_0) \leq \liminf_n f(x_n)$$

for any sequence $\{x_n\} \in X$ such that x_n converges to x_0.

Definition 2.7. Let X be a Euclidian space. A function $f : X \to \mathbb{R}$ is weakly upper semicontinuous at $x_0 \in X$ if

$$f(x_0) \geq \limsup_n f(x_n)$$

for any sequence $\{x_n\} \in X$ such that x_n converges to x_0 weakly.

Lemma 2.4 (Brouwer's fixed point theorem [14])**.** Let S be a nonempty, compact and convex subset of a finite-dimensional space and $T : S \to S$ be a continuous mapping. Then there exists a $x \in S$ such that $T(x) = x$.

3 (GMEVVLIP) with α-relaxed Exponentially (p, η)-monotone

Let $K \subseteq X$ be nonempty closed convex subset of a Euclidian space X and (Y, C) be an ordered Euclidian space induced by the closed convex pointed cone C whose apex at origin with int$C \neq \emptyset$. Let $p \in \mathbb{R}$ be a nonzero real number, $\eta : K \times K \to X$ and $f : K \times K \to Y$ be two bi-mappings, $A : L(X, Y) \to L(X, Y)$ be a mapping, where $L(X, Y)$ be space of all continuous linear mappings from X to Y, and $T : K \to 2^{L(X,Y)}$ be a vector set-valued mapping. Then The generalized mixed exponential type vector variational-like inequality problems (GMEVVLIP) is to find $u \in K$ and $x \in T(u)$, such that

$$\left\langle Ax, \frac{1}{p}(e^{p\eta(v,u)} - 1) \right\rangle + f(u, v) \not\leq_{intC} 0, \ \forall v \in K. \tag{3.1}$$

(GMEVVLIP) has wide range of application to vector optimization problems and vector variational inequalities problems.

Definition 3.1. The mapping $T : K \to L(X, Y)$ is said to be α-relaxed exponentially (p, η)-monotone if for every pair of points $u, v \in K$, we have

$$\left\langle Tu - Tv, \frac{1}{p}(e^{p\eta(u,v)} - 1) \right\rangle \geq_C \alpha(u - v), \tag{3.2}$$

where $\alpha : X \to Y$ with $\alpha(tx) = t^q \alpha(x)$ for all $t > 0$ and $x \in X$, where $q > 1$, a real number.

Definition 3.2. Let $A : L(X,Y) \to L(X,Y)$. A multivalued mapping $T : K \to 2^{L(X,Y)}$ with compact valued is said to be α-relaxed exponentially (p,η)-monotone with respect to A if for each pair of points $u, v \in K$, we have

$$\left\langle Ax - Ay, \frac{1}{p}(e^{p\eta(u,v)} - 1) \right\rangle \geq_C \alpha(u - v), \ \forall x \in T(u), y \in T(v). \tag{3.3}$$

where $\alpha : X \to Y$ with $\alpha(tx) = t^q\alpha(x)$ for all $t > 0$ and $x \in X$, where $q > 1$, a real number.

Remark 3.1

(i) If $\alpha \equiv 0$ then Definition 3.1 is called exponentially (p,η)-monotone, i.e. for each pair of points $u, v \in K$, we have

$$\left\langle Tu - Tv, \frac{1}{p}(e^{p\eta(u,v)} - 1) \right\rangle \geq_C 0.$$

(ii) If $\alpha \equiv 0$ then Definition 3.2 is called exponentially (p,η)-monotone with respect to A, i.e. for each pair of points $u, v \in K$, we have

$$\left\langle Ax - Ay, \frac{1}{p}(e^{p\eta(u,v)} - 1) \right\rangle \geq_C 0, \ \forall x \in T(u), y \in T(v).$$

So every exponentially (p,η)-monotone mapping is α-relaxed exponentially (p,η)-monotone map with $\alpha \equiv 0$.

Theorem 3.1. Let K be a nonempty bounded closed convex subset of a real Euclidian space X and (Y,C) is an ordered Euclidian space induced by the pointed closed convex cone C whose apex is at origin with $\text{int}C \neq \emptyset$. Suppose $\eta : K \times K \to X$ be affine in the first argument with $\eta(x,x) = 0, \forall x \in K$. Let $f : K \times K \to Y$ be a C-convex in the second argument with the condition $f(x,x) = 0, \forall x \in K$. Let $A : L(X,Y) \to L(X,Y)$ be a continuous mapping and $T : K \to 2^{L(X,Y)}$ be a nonempty compact valued mapping, which is H-hemicontinuous and α-relaxed exponentially (p,η)-monotone with respect to A. Then the following two statements (a) and (b) are equivalent:

(a) there exists $\overline{u} \in K$ and $\overline{x} \in T(\overline{u})$ such that

$$\left\langle A\overline{x}, \frac{1}{p}(e^{p\eta(v,\overline{u})} - 1) \right\rangle + f(\overline{u}, v) \not\leq_{intC} 0, \ \forall v \in K.$$

(b) there exists $\overline{u} \in K$ such that

$$\left\langle Ay, \frac{1}{p}(e^{p\eta(v,\overline{u})} - 1) \right\rangle + f(\overline{u}, v) \not\leq_{intC} \alpha(v - \overline{u}), \ \forall v \in K, y \in T(v).$$

Proof. Let the statement (a) is true, i.e. there exist $\overline{u} \in K$ and $\overline{x} \in T(\overline{u})$ such that

$$\left\langle A\overline{x}, \frac{1}{p}(e^{p\eta(v,\overline{u})} - 1) \right\rangle + f(\overline{u}, v) \not\leqslant_{intC} 0, \ \forall v \in K. \tag{3.4}$$

Since T is α-relaxed exponentially (p,η)-monotone with respect to A, we have

$$\begin{aligned}
&\left\langle Ay - A\overline{x}, \frac{1}{p}(e^{p\eta(v,\overline{u})} - 1) \right\rangle + f(\overline{u}, v) \geq_C \alpha(v - \overline{u}) \\
&\qquad + f(\overline{u}, v), \ \forall v \in K, y \in T(v) \\
&\Rightarrow \left\langle Ay, \frac{1}{p}(e^{p\eta(v,\overline{u})} - 1) \right\rangle + f(\overline{u}, v) \geq_C \left\langle A\overline{x}, \frac{1}{p}(e^{p\eta(v,\overline{u})} - 1) \right\rangle \\
&\qquad + \alpha(v - \overline{u}) + f(\overline{u}, v), \ \forall v \in K, y \in T(v) \\
&\Rightarrow \left\langle Ay, \frac{1}{p}(e^{p\eta(v,\overline{u})} - 1) \right\rangle + f(\overline{u}, v) - \alpha(v - \overline{u}) \geq_C \left\langle A\overline{x}, \frac{1}{p}(e^{p\eta(v,\overline{u})} - 1) \right\rangle \\
&\qquad + f(\overline{u}, v), \ \forall v \in K, y \in T(v).
\end{aligned} \tag{3.5}$$

From (3.4), (3.5) and Lemma 2.1, we get

$$\left\langle Ay, \frac{1}{p}(e^{p\eta(v,\overline{u})} - 1) \right\rangle + f(\overline{u}, v) \not\leqslant_{intC} \alpha(v - \overline{u}), \ \forall v \in K, y \in T(v).$$

Conversely, suppose that the statement (b) is true, i.e. there exists $\overline{u} \in K$ such that

$$\left\langle Ay, \frac{1}{p}(e^{p\eta(v,\overline{u})} - 1) \right\rangle + f(\overline{u}, v) \not\leqslant_{intC} \alpha(v - \overline{u}), \ \forall v \in K, y \in T(v). \tag{3.6}$$

Let $v \in K$ be any point. Letting $v_t = tv + (1-t)\overline{u}$, $t \in (0,1]$, as K is convex, $v_t \in K$. Let $y_t \in T(v_t)$, we have from (3.6),

$$\left\langle Ay_t, \frac{1}{p}(e^{p\eta(v_t,\overline{u})} - 1) \right\rangle + f(\overline{u}, v) \not\leqslant_{intC} \alpha(v_t - \overline{u}) = t^q \alpha(v - \overline{u}). \tag{3.7}$$

$$\begin{aligned}
&\left\langle Ay_t, \frac{1}{p}(e^{p\eta(v_t,\overline{u})} - 1) \right\rangle + f(\overline{u}, v_t) \\
&\quad = \left\langle Ay_t, \frac{1}{p}(e^{p\eta(tv+(1-t)\overline{u},\overline{u})} - 1) \right\rangle + f(\overline{u}, tv + (1-t)\overline{u}) \\
&\quad = \left\langle Ay_t, \frac{1}{p}(e^{pt\eta(v,\overline{u})+(1-t)p\eta(\overline{u},\overline{u})} - 1) \right\rangle + f(\overline{u}, tv + (1-t)\overline{u}) \\
&\quad \leq_C \left\langle Ay_t, \frac{1}{p}(t(e^{p\eta(v,\overline{u})} - 1) + (1-t)(e^{p\eta(\overline{u},\overline{u})} - 1)) \right\rangle + tf(\overline{u}, v) \\
&\quad = t\{\left\langle Ay_t, \frac{1}{p}(e^{p\eta(v,\overline{u})} - 1) \right\rangle + tf(\overline{u}, v)\}
\end{aligned} \tag{3.8}$$

By (3.7), (3.8) and Lemma 2.1, we get

$$\left\langle Ay_t, \frac{1}{p}(e^{p\eta(v,\overline{u})} - 1)\right\rangle + f(\overline{u}, v) \not\leqslant_{intC} t^{q-1}\alpha(v - \overline{u}). \tag{3.9}$$

Since $T(v_t)$ and $T(\overline{u})$ are compact, by lemma, for each fixed $y_t \in T(v_t)$, there exists $x_t \in T(\overline{u})$ such that

$$\|y_t - x_t\| \leq H(T(v_t), T(\overline{u})). \tag{3.10}$$

Since $T(\overline{u})$ is compact, without loss off generality, we may assume that

$$x_t \to \overline{x} \in T(\overline{u}) \text{ as } t \to 0^+.$$

Also T is H-hemicontinuous, thus it follows that

$$H(T(v_t), T(\overline{u})) \to 0 \text{ as } t \to 0^+.$$

Now by (3.10) we have

$$\begin{aligned}\|y_t - \overline{x}\| &\leq \|y_t - x_t\| + \|x_t - \overline{x}\| \\ &\leq H(T(v_t), T(\overline{u})) + \|x_t - \overline{x}\| \to 0 \text{ as } t \to 0^+.\end{aligned} \tag{3.11}$$

Since A is continuous, letting $t \to 0^+$, we have

$$\begin{aligned}\| \left\langle Ay_t, \frac{1}{p}(e^{p\eta(v,\overline{u})} - 1)\right\rangle - t^{q-1}\alpha(v - \overline{u}) - \left\langle A\overline{x}, \frac{1}{p}(e^{p\eta(v,\overline{u})} - 1)\right\rangle \| \\ \leq \| \left\langle Ay_t - A\overline{x}, \frac{1}{p}(e^{p\eta(v,\overline{u})} - 1)\right\rangle \| \|t^{q-1}\alpha(v - \overline{u})\| \\ \leq \|Ay_t - A\overline{x}\| \|\frac{1}{p}(e^{p\eta(v,\overline{u})} - 1)\| + t^{q-1}\|\alpha(v - \overline{u})\| \\ \to 0 \text{ as } t \to 0^+.\end{aligned} \tag{3.12}$$

From (3.7), we have

$$\left\langle Ay_t, \frac{1}{p}(e^{p\eta(v,\overline{u})} - 1)\right\rangle + f(\overline{u}, v) - t^{p-1}\alpha(v - \overline{u} \in V/(-intC)$$

Since $V/(-intC)$ is closed, therefore from (3.12) we have

$$\left\langle A\overline{x}, \frac{1}{p}(e^{p\eta(v,\overline{u})} - 1)\right\rangle + f(\overline{u}, v) \in V/(-intC)$$

$$\Rightarrow \left\langle A\overline{x}, \frac{1}{p}(e^{p\eta(v,\overline{u})} - 1)\right\rangle + f(\overline{u}, v) \not\leqslant_{intC} 0, \ \forall v \in K.$$

Which completes the proof. □

Theorem 3.2. Let K be a nonempty bounded closed convex subset of a Euclidian space X and (Y, C) is an ordered Euclidian space induced by the proper pointed closed convex cone C whose apex is at origin with $intC \neq \emptyset$. Suppose $\eta : K \times K \to X$ be affine in the first argument with $\eta(x, x) = 0, \forall x \in K$. Let $f : K \times K \to Y$ be completely continuous in the first argument and affine in the second argument with the condition $f(x, x) = 0, \forall x \in K$. Let $\alpha : X \to Y$ is weakly lower semicontinuous. Let $A : L(X, Y) \to L(X, Y)$ be a continuous mapping and $T : K \to 2^{L(X,Y)}$ be a nonempty compact valued mapping, which is H-hemicontinuous and α-relaxed exponentially (p, η)-monotone with respect to A. Then (GMEVVLIP) is solvable, i.e. there exist $u \in K$ and $x \in T(u)$ such that

$$\left\langle Ax, \frac{1}{p}(e^{p\eta(v,u)} - 1) \right\rangle + f(u, v) \not\leqslant_{intC} 0, \ \forall v \in K.$$

Proof. Consider the set valued mapping $F : K \to 2^X$ such that

$$F(v) = \{u \in K : \left\langle Ax, \frac{1}{p}(e^{p\eta(v,u)} - 1) \right\rangle + f(u, v) \not\leqslant_{intC} 0, \ for \ some \ x \in T(u)\}, \forall v \in K.$$

First we claim that F is a KKM mapping.

If F is not a KKM mapping, then there exists $\{u_1, u_2, \ldots, u_m\} \subset K$ such that $co\{u_1, u_2, \ldots, u_m\} \not\subseteq \bigcup_{i=1}^{m} F(u_i)$, that means there exists at least a $u \in co\{u_1, u_2, \ldots, u_m\}$, $u = \sum_{i=1}^{m} t_i u_i$, where $t_i \geq 0$, $i = 1, 2, \ldots, m$, $\sum_{i=1}^{m} t_i = 1$, but $u \notin \bigcup_{i=1}^{m} F(u_i)$.

From the construction of F, for any $x \in T(u)$ we have

$$\left\langle Ax, \frac{1}{p}(e^{p\eta(u_i,u)} - 1) \right\rangle + f(u, u_i) \leq_{intC} 0; \text{!for} i = 1, 2, \ldots, m. \tag{3.13}$$

From (3.13), and since η is affine in the first argument, it follows that

$$\begin{aligned}
0 &= \left\langle Ax, \frac{1}{p}(e^{p\eta(u,u)} - 1) \right\rangle + f(u, u) \\
&= \left\langle Ax, \frac{1}{p}(e^{p\eta(\sum_{i=1}^{m} t_i u_i, u)} - 1) \right\rangle + f(u, \sum_{i=1}^{m} t_i u_i) \\
&= \left\langle Ax, \frac{1}{p}(e^{\sum_{i=1}^{m} t_i p\eta(u_i, u)} - 1) \right\rangle + \sum_{i=1}^{m} t_i f(u, u_i) \\
&\leq_C \left\langle Ax, \frac{1}{p} \sum_{i=1}^{m} t_i (e^{p\eta(u_i,u)} - 1) \right\rangle + \sum_{i=1}^{m} t_i f(u, u_i)
\end{aligned}$$

$$= \sum_{i=1}^{m} t_i \left\{ \left\langle Ax, \frac{1}{p}(e^{p\eta(u_i,u)} - 1) \right\rangle + f(u, u_i) \right\}$$
$$\leq_{intC} 0,$$

which implies that $0 \in intC$, this contradicts the fact that C is proper. Hence F is a KKM mapping.

Define another set valued mapping $G : K \to 2^X$ such that

$$G(v) = \{u \in K : \left\langle Ay, \frac{1}{p}(e^{p\eta(v,u)} - 1) \right\rangle + f(u, v) \not\leq_{intC} \alpha(v-u),\ \forall\ y \in T(v)\}, \forall v \in K.$$

Now we will prove that $F(v) \subset G(v),\ \forall v \in K$.

Let $u \in F(v)$, there exists some $x \in T(u)$ such that

$$\left\langle Ax, \frac{1}{p}(e^{p\eta(v,u)} - 1) \right\rangle + f(u, v) \not\leq_{intC} 0. \tag{3.14}$$

Since T is α-relaxed exponentially (p, η)-monotone with respect to A, therefore $\forall v \in K, y \in T(v)$, we have

$$\left\langle Ax, \tfrac{1}{p}(e^{p\eta(v,u)} - 1) \right\rangle + f(u, v) \leq_C \left\langle Ay, \tfrac{1}{p}(e^{p\eta(v,u)} - 1) \right\rangle + f(u, v) - \alpha(v-u). \tag{3.15}$$

From (3.14), (3.15) and Lemma 2.1, we have

$$\left\langle Ay, \frac{1}{p}(e^{p\eta(v,u)} - 1) \right\rangle + f(u, v) \not\leq_{intC} \alpha(v-u),\ \forall v \in K, y \in T(v).$$

Therefore $u \in G(v)$, i.e. $F(v) \subset G(v),\ \forall v \in K$.

This implies that G is also a KKM mapping.

We claim that for each $v \in K$, $G(v) \subset K$ is closed in the weak topology of X.

Let us suppose, $\overline{u} \in \overline{G(v)}^w$ the weak closure of $G(v)$. Since X is reflexive, there is sequence $\{u_n\}$ in $G(v)$ such that $\{u_n\}$ converges weakly to $\overline{u} \in K$. Then for each $y \in T(v)$, we have

$$\left\langle Ay, \frac{1}{p}(e^{p\eta(v,u_n)} - 1) \right\rangle + f(u_n, v) \not\leq_{intC} \alpha(v-u_n)$$
$$\Rightarrow \left\langle Ay, \frac{1}{p}(e^{p\eta(v,u_n)} - 1) \right\rangle + f(u_n, v) - \alpha(v-u_n) \in Y/-intC.$$

Since, Ay and f are completely continuous and $Y/-intC$ is closed, α is weakly lower semicontinuous, therefore the sequence $\{\left\langle Ay, \frac{1}{p}(e^{p\eta(v,u_n)} - 1) \right\rangle + f(u_n, v) - \alpha(v-u_n)\}$ converges to $\left\langle Ay, \frac{1}{p}(e^{p\eta(v,\overline{u})} - 1) \right\rangle + f(\overline{u}, v) - \alpha(v-\overline{u})$ and

$\left\langle Ay, \frac{1}{p}(e^{p\eta(v,\overline{u})}-1)\right\rangle + f(\overline{u},v) - \alpha(v-\overline{u}) \in Y/-intC$ therefore

$$\left\langle Ay, \frac{1}{p}(e^{p\eta(v,\overline{u})}-1)\right\rangle + f(\overline{u},v) \not\leqslant_{intC} \alpha(v-\overline{u}).$$

Hence $\overline{u} \in G(v)$. This proves that $G(v)$ is weakly closed for all $v \in K$. Furthermore, X is reflexive and $K \subset X$ is nonempty, bounded, closed and convex; therefore, K is weakly compact subset of X and so $G(v)$ is also weakly compact. Therefore from Lemma 2.2 and Theorem 3.1, it follows that

$$\bigcap_{v\in K} G(v) \neq \emptyset.$$

So there exists $\overline{u} \in K$, such that

$$\left\langle Ay, \frac{1}{p}(e^{p\eta(v,\overline{u})}-1)\right\rangle + f(\overline{u},v) \not\leqslant_{intC} \alpha(v-\overline{u}),\ \forall\ v \in K,\ y \in T(v).$$

Hence Theorem 3.1, we can conclude that there exist $\overline{u} \in K$ and $\overline{x} \in T(\overline{u})$ such that

$$\left\langle A\overline{x}, \frac{1}{p}(e^{p\eta(v,\overline{u})}-1)\right\rangle + f(\overline{u},v) \not\leqslant_{intC} 0,\ \forall\ v \in K.$$

i.e. (GMEVVLIP) is solvable. □

Theorem 3.3. Let K be a nonempty closed convex subset of a Euclidian space X with $0 \in K$ and (Y, C) is an ordered Euclidian space induced by the proper pointed closed convex cone C whose apex is at origin with $intC \neq \emptyset$. Suppose $\eta : K \times K \to X$ be affine in the first argument with $\eta(x,x) = 0, \forall x \in K$. Let $f : K \times K \to Y$ be completely continuous in the first argument and affine in the second argument with the condition $f(x,x) = 0, \forall x \in K$. Let $\alpha : X \to Y$ is weakly lower semicontinuous. Let $A : L(X,Y) \to L_c(X,Y)$ be a continuous mapping, where $L_c(X,Y)$ is the space of all completely continuous linear map from X to Y. and $T : K \to 2^{L(X,Y)}$ be a nonempty compact valued mapping, which is H-hemicontinuous and α-relaxed exponentially (p,η)-monotone with respect to A. If there exists one $r > 0$ such that

$$\left\langle Ay, \frac{1}{p}(e^{p\eta(0,v)}-1)\right\rangle + f(v,0) \not\leqslant_{intC} 0,\ \forall\ v \in K,\ y \in T(v)\ with\ \|v\| = r \tag{3.16}$$

then (GMEVVLIP) is solvable, i.e. there exist $u \in K$ and $x \in T(u)$ such that

$$\left\langle Ax, \frac{1}{p}(e^{p\eta(v,u)}-1)\right\rangle + f(u,v) \not\leqslant_{intC} 0,\ \forall v \in K.$$

Proof. For $r > 0$, assume $K_r = \{u \in X : \|u\| \leq r\}$.
By Theorem 3.2, we know that (GMEVVLIP) is solvable over K_r, i.e. there exists $u_r \in K \cap K_r$ and $x_r \in T(u_r)$ such that

$$\left\langle Ax_r, \frac{1}{p}(e^{p\eta(v,u_r)}-1)\right\rangle + f(u_r,v) \not\leqslant_{intC} 0,\ \forall v \in K \cap K_r. \tag{3.17}$$

Putting $v = 0$ in (3.17),

$$\left\langle Ax_r, \frac{1}{p}(e^{p\eta(0,u_r)} - 1) \right\rangle + f(u_r, 0) \not\leqslant_{intC} 0. \tag{3.18}$$

If $\|u_r\| = r$ for all r, then it contradicts (3.16). Hence $r > \|u_r\|$.
For any $z \in K$, let us choose $t \in (0, 1)$ small enough such that $(1-t)u_r + tz \in K \cap K_r$. Putting $v = (1-t)u_r + tz$ in (3.17), we get

$$\left\langle Ax_r, \frac{1}{p}(e^{p\eta((1-t)u_r+tz,u_r)} - 1) \right\rangle + f(u_r, (1-t)u_r + tz) \not\leqslant_{intC} 0. \tag{3.19}$$

Since η is affine in the first variable, we have

$$\begin{aligned}
&\left\langle Ax_r, \frac{1}{p}(e^{p\eta((1-t)u_r+tz,u_r)} - 1) \right\rangle + f(u_r, (1-t)u_r + tz) \\
&\quad = \left\langle Ax_r, \frac{1}{p}(e^{(1-t)p\eta(u_r,u_r)+tp\eta(z,u_r)} - 1) \right\rangle + tf(u_r, z) \\
&\quad \leq_C \left\langle Ax_r, (1-t)\frac{1}{p}(e^{p\eta(u_r,u_r)} - 1) + t\frac{1}{p}(e^{p\eta(z,u_r)} - 1) \right\rangle + tf(u_r, z) \\
&\quad = t\{\left\langle Ax_r, \frac{1}{p}(e^{p\eta(z,u_r)} - 1) \right\rangle + f(u_r, z)\}.
\end{aligned} \tag{3.20}$$

Hence from (3.19), (3.20) and Lemma 2.1, we get

$$\left\langle Ax_r, \frac{1}{p}(e^{p\eta(z,u_r)} - 1) \right\rangle + f(u_r, z) \not\leqslant_{intC} 0, \ \forall z \in K.$$

Therefore, (GMEVVLIP) is solvable. This completes the proof. □

References

1. Usman, F., Khan, S.A.: A generalized mixed vector variational-like inequality problem. Nonlinear Anal.: Theory Methods Appl. **71**(11), 5354–5362 (2009)
2. Fang, Y.P., Huang, N.J.: Variational-like inequalities with generalized monotone mappings in Banach spaces. J. Optim. Theory Appl. **118**(2), 327–338 (2003)
3. Huang, N.-J., Gao, C.-J.: Some generalized vector variational inequalities and complementarity problems for multivalued mappings. Appl. Math. Lett. **16**(7), 1003–1010 (2003)
4. Khan, S.A., Khan, Q.H., Suhel, F.: Generalized vector mixed variational-like inequality problem without monotonicity. Thai J. Math. **10**(2), 245–258 (2012)
5. Lin, K.L., Yang, D.P., Yao, J.-C.: Generalized vector variational inequalities. J. Optim. Theory Appl. **92**(1), 117–125 (1997)
6. Zeng, L.-C., Yao, J.-C.: Existence of solutions of generalized vector variational inequalities in reflexive Banach spaces. J. Glob. Optim. **36**(4), 483–497 (2006)
7. Jayswal, A., Choudhury, S., Verma, R.U.: Exponential type vector variational-like inequalities and vector optimization problems with exponential type invexities. J. Appl. Math. Comput. **45**(1–2), 87–97 (2014)

8. Jayswal, A., Choudhury, S.: Exponential type vector variational-like inequalities and nonsmooth vector optimization problems. J. Appl. Math. Comput. **49**(1–2), 127–143 (2015)
9. Wu, K.Q., Huang, N.J.: Vector variational-like inequalities with relaxed η- α Pseudomonotone mappings in Banach spaces. J. Math. Inequal. **1**, 281–290 (2007)
10. Ceng, L.-C., Yao, J.-C.: On generalized variational-like inequalities with generalized monotone multivalued mappings. Appl. Math. Lett. **22**(3), 428–434 (2009)
11. Plubtieng, S., Thammathiwat, T.: Existence of solutions of new generalized mixed vector variational-like inequalities in reflexive Banach spaces. J. Optim. Theory Appl. **162**(2), 589–604 (2014)
12. Fan, K.: A generalization of Tychonoff's fixed point theorem. Math. Ann. **142**, 305–310 (1961)
13. Nadler, S.B.: Multi-valued contraction mappings. Pac. J. Math. **30**(2), 475–488 (1969)
14. Brouwer, L.: Zur invarianz des n-dimensional gebietes. Math. Ann. **71**(3), 305–313 (1912)

On Approximate Solutions to One Class of Nonlinear Differential Equations

Inessa Matveeva[1,2(✉)]

[1] Sobolev Institute of Mathematics,
Acad. Koptyug Avenue 4, 630090 Novosibirsk, Russia
matveeva@math.nsc.ru
[2] Novosibirsk State University, Pirogov Street 2, 630090 Novosibirsk, Russia
http://www.math.nsc.ru/LBRT/d5/english/matveeva.htm

Abstract. We consider a class of systems of nonlinear ordinary differential equations with parameters. In particular, systems of such type arise when modeling the multistage synthesis of a substance. We study properties of solutions to the systems and propose a method for approximate solving the systems in the case of very large coefficients. We establish approximation estimates and show that the convergence rate depends on the parameters characterizing the nonlinearity of the systems. Moreover, the larger the coefficients of the systems, the more exact the approximate solutions. Thereby this method allows us to avoid difficulties arising inevitably when solving systems of nonlinear differential equations with very large coefficients.

Keywords: Systems of ordinary differential equations · Cauchy problem · Large coefficients · Estimates for solutions · Limit theorems

1 Introduction

Consider the following system of ordinary differential equations

$$\begin{cases} \dfrac{dx_1}{dt} = g(t, x_n) - \dfrac{n-1}{\tau} x_1, \quad t > 0, \\ \dfrac{dx_j}{dt} = \dfrac{n-1}{\tau}(x_{j-1} - x_j), \quad j = 2, \ldots, n-1, \\ \dfrac{dx_n}{dt} = \dfrac{n-1}{\tau} x_{n-1} - \theta x_n. \end{cases} \tag{1}$$

This system arises when modeling the multistage synthesis of a substance. The dimension n of the system is defined by the number of stages, the first equation describes the initiation law, the last equation does the utilization law, $\theta \geq 0$, τ is the duration of the process, $x_j(t, \tau)$ is the substance concentration at the jth

I. Matveeva—The work is supported in part by the Russian Foundation for Basic Research (project no. 16-01-00592).

© Springer Nature Singapore Pte Ltd. 2017
D. Giri et al. (Eds.): ICMC 2017, CCIS 655, pp. 221–231, 2017.
DOI: 10.1007/978-981-10-4642-1_19

stage, $x_n(t,\tau)$ is the concentration of the final product. Therefore, $x_n(t,\tau)$ is of interest from the practical viewpoint.

It should be noted that systems of the form (1) are often termed the 'Goodwin' model [1]. Ordinary differential equations of such kind and more complicated equations arise when modeling gene networks (for example, see [2], the reviews [3,4] and the bibliography therein).

If n is very large (for instance, the process consists of a great number of the stages) then finding of the last component $x_n(t,\tau)$ of the solution to (1) is a very complicated problem. A rigorous mathematical solution to this problem was given by G.V. Demidenko (see [5, Theorems 1–4]). We formulate this result below.

Suppose that the function $g(t,v) \in C(\overline{\mathbb{R}^+_2})$ is bounded and satisfies the Lipschitz condition with respect to v. Increase the dimension of (1) unboundedly and consider the Cauchy problem for each system with the zero initial conditions

$$x_j|_{t=0} = 0, \qquad j = 1, \dots, n. \tag{2}$$

Taking only the last component of the solution to each of these Cauchy problems, we obtain the sequence of the functions $\{x_n(t,\tau)\}$.

Theorem 1 (G.V. Demidenko). *The sequence* $\{x_n(t,\tau)\}$ *converges uniformly on every segment* $[0,T]$, $T > \tau$*:*

$$x_n(t,\tau) \to y(t,\tau), \quad n \to \infty.$$

The limit function $y(t,\tau)$ *is a solution to the initial value problem for the delay equation*

$$\begin{cases} \dfrac{d}{dt}y(t,\tau) = -\theta y(t,\tau) + g(t-\tau, y(t-\tau,\tau)), & t > \tau, \\ y(t,\tau) \equiv 0, \quad 0 \le t \le \tau; \end{cases} \tag{3}$$

moreover,

$$\max_{t\in[0,T]} |x_n(t,\tau) - y(t,\tau)| \le \frac{c}{n^{1/4}}, \qquad n > n_0.$$

By Theorem 1, we need not solve the Cauchy problem (1), (2) for the system of large dimension with large coefficients in order to compute approximately $x_n(t,\tau)$ for $n \gg 1$. It is sufficient to solve only the initial value problem (3) for one delay differential equation. This result gives us an effective method for approximate finding $x_n(t,\tau)$ for $n \gg 1$ by using the delay equation; moreover, the estimate established in Theorem 1 characterizes the approximation order.

Theorem 1 has become a basis for deriving similar statements for various systems of nonlinear ordinary differential equations of large dimension (see, for example, [6–12]). In particular, a perturbation of (1) was investigated in [11]. Some examples of the Cauchy problems for (1) with nonzero initial conditions were considered in [7]. On the basis of the results, a new method for approximation of solutions to initial value problems for the mentioned delay differential

equation with arbitrary initial conditions was proposed in [10]. Three different classes of systems of large dimension were studied in [6,9], [8] and [12] respectively. In the mentioned works G.V. Demidenko proposed a series of methods for proving limit theorems which establish interconnections between solutions to classes of systems of nonlinear ordinary differential equations of large dimension and generalized solutions to delay differential equations. The readers can be familiarized with some of these methods in the papers [12,13]. Using the methods, classes of essentially nonlinear systems of large dimension (every equation in the systems is nonlinear) were studied in [14–16]. It should be noted that there is a number of works devoted to the study of approximation of solutions to delay differential equations by means of solutions to systems of ordinary differential equations of large dimension (see, for instance, [17–22]). In particular, [17,18] are the first works in this direction. A brief survey of the literature and the use of the semigroup theory for approximation are given in [19]. Approximation schemes and their development are discussed in [20–22].

If $\tau \ll 1$ (for example, the synthesis process is very rapid) then the coefficients of (1) is very large as well. In [23,24] we studied the behavior of $x_n(t,\tau)$ in dependence on τ for every fixed n. In particular, the following result was obtained.

Theorem 2. *The sequence* $\{x_n(t,\tau)\}$ *converges uniformly on every segment* $[0,T]$*:*

$$x_n(t,\tau) \to z(t), \qquad \tau \to 0.$$

The limit function $z(t)$ *is a solution to the Cauchy problem*

$$\begin{cases} \dfrac{d}{dt}z = -\theta z + g(t,z), \quad t > 0, \\ z(0) = 0; \end{cases} \tag{4}$$

moreover,

$$\max_{t\in[0,T]} |x_n(t,\tau) - z(t)| \le c\,\tau, \qquad \tau \ll 1, \tag{5}$$

where $c > 0$ *depends on* θ*,* G*,* L*,* T*,* n*.*

This result gives us an effective method for approximate calculating $x_n(t,\tau)$. Indeed, we may solve the Cauchy problem (4) for one ordinary differential equation instead of the Cauchy problem (1), (2). Then, by the obtained convergence, we have $z(t) \approx x_n(t,\tau)$ for $\tau \ll 1$. Since τ is the duration of the synthesis process, then we can find approximately the concentration $x_n(t,\tau)$ of the final product in the case of very rapid passages from one stage to the other.

More detailed modeling processes of the substance synthesis leads to systems of essentially nonlinear differential equations in comparison with (1). As a rule, so-called Hill's type functions are used (for example, see [4]). Our aim is to study one class of systems of such kind described in the next section.

2 Main Results

In the present paper we consider the Cauchy problem for the class of systems of nonlinear ordinary differential equations

$$\begin{cases} \dfrac{d\widehat{x}_1}{dt} = g(t,\widehat{x}_n) - \dfrac{n-1}{\tau}\dfrac{\widehat{x}_1}{1+\rho_1\widehat{x}_1^{\gamma_1}}, \quad t>0, \\ \dfrac{d\widehat{x}_j}{dt} = \dfrac{n-1}{\tau}\left(\dfrac{\widehat{x}_{j-1}}{1+\rho_{j-1}\widehat{x}_{j-1}^{\gamma_{j-1}}} - \dfrac{\widehat{x}_j}{1+\rho_j\widehat{x}_j^{\gamma_j}}\right), \quad j=2,\dots,n-1, \\ \dfrac{d\widehat{x}_n}{dt} = \dfrac{n-1}{\tau}\dfrac{\widehat{x}_{n-1}}{1+\rho_{n-1}\widehat{x}_{n-1}^{\gamma_{n-1}}} - \theta\widehat{x}_n, \\ \widehat{x}_j|_{t=0} = 0, \quad j=1,\dots,n, \end{cases} \tag{6}$$

where

$$\theta \geq 0, \quad \tau > 0, \quad 0 \leq \rho_k \leq \rho, \quad 0 < \gamma \leq \gamma_k, \quad k = 1,\dots,n-1.$$

This system arises when modeling the multistage synthesis of a substance as well. Obviously, this system for $\rho = 0$ coincides with (1). As was shown in [14,16], the last component $\widehat{x}_n(t,\tau)$ of the solution to (6) for $n \gg 1$ is approximated by the solution to the initial value problem (3). Analogous results for a more general class of systems of nonlinear differential equations including the systems of (6) were obtained in [15] for $\rho_k = \rho$, $\gamma_k = \gamma$.

We study properties of the components $\widehat{x}_j(t,\tau)$ of the solution to the Cauchy problem (6) as functions of t and $\tau \ll 1$, when n is fixed. Assume that the function $g(t,v) \in C(\mathbb{R}_2)$ is nonnegative and bounded $0 \leq g(t,v) \leq G$ and satisfies the Lipschitz condition

$$|g(t,v_1) - g(t,v_2)| \leq L|v_1 - v_2|, \qquad v_1,\ v_2 \in \mathbb{R}.$$

Note that the Cauchy problem (6) is uniquely solvable under these conditions; moreover, the components of the solutions are nonnegative (see the detailed proof in [14–16]).

The following result holds.

Theorem 3. *The sequence $\{\widehat{x}_n(t,\tau)\}$ consisting of the last components of the solutions to the Cauchy problems of the from (6) converges uniformly on every segment $[0,T]$:*

$$\widehat{x}_n(t,\tau) \to z(t), \qquad \tau \to 0. \tag{7}$$

The limit function $z(t)$ is the solution to the Cauchy problem (4).

Proof. Denote $u(t,\tau) = \widehat{x}(t,\tau) - x(t,\tau)$, where $x(t,\tau)$ is the solution to the Cauchy problem (1), (2) and $\widehat{x}(t,\tau)$ is the solution to the Cauchy problem (6). It is not hard to verify that the vector function $u(t,\tau)$ satisfies the following system of differential equations

$$\frac{du}{dt} = Au + G_1(t) + G_2(t),$$

where A coincides with the matrix of (1)

$$A = \begin{pmatrix} -\frac{n-1}{\tau} & 0 & \cdots & \cdots & 0 \\ \frac{n-1}{\tau} & -\frac{n-1}{\tau} & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & -\frac{n-1}{\tau} & 0 \\ 0 & \cdots & 0 & \frac{n-1}{\tau} & -\theta \end{pmatrix},$$

$$G_1(t) = \begin{pmatrix} g(t, \widehat{x}_n(t,\tau)) - g(t, x_n(t,\tau)) \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

$$G_2(t) = \frac{n-1}{\tau} \begin{pmatrix} \frac{\rho_1\left(\widehat{x}_1(t,\tau)\right)^{1+\gamma_1}}{1+\rho_1\left(\widehat{x}_1(t,\tau)\right)^{\gamma_1}} \\ \frac{\rho_2\left(\widehat{x}_2(t,\tau)\right)^{1+\gamma_2}}{1+\rho_2\left(\widehat{x}_2(t,\tau)\right)^{\gamma_2}} - \frac{\rho_1\left(\widehat{x}_1(t,\tau)\right)^{1+\gamma_1}}{1+\rho_1\left(\widehat{x}_1(t,\tau)\right)^{\gamma_1}} \\ \vdots \\ \frac{\rho_{n-1}\left(\widehat{x}_{n-1}(t,\tau)\right)^{1+\gamma_{n-1}}}{1+\rho_{n-1}\left(\widehat{x}_{n-1}(t,\tau)\right)^{\gamma_{n-1}}} - \frac{\rho_{n-2}\left(\widehat{x}_{n-2}(t,\tau)\right)^{1+\gamma_{n-2}}}{1+\rho_{n-2}\left(\widehat{x}_{n-2}(t,\tau)\right)^{\gamma_{n-2}}} \\ -\frac{\rho_{n-1}\left(\widehat{x}_{n-1}(t,\tau)\right)^{1+\gamma_{n-1}}}{1+\rho_{n-1}\left(\widehat{x}_{n-1}(t,\tau)\right)^{\gamma_{n-1}}} \end{pmatrix}.$$

Taking into account that $u(0,\tau) = 0$, we obtain

$$u(t,\tau) = \int_0^t e^{(t-s)A}(G_1(s) + G_2(s))ds.$$

Remind the representation for the matrix exponent [25]

$$e^{tA} = \varphi_1(t)I + \varphi_2(t)(A - \lambda_1 I) + \varphi_3(t)(A - \lambda_1 I)(A - \lambda_2 I) + \ldots \\ + \varphi_n(t)(A - \lambda_1 I)\ldots(A - \lambda_{n-1} I),$$

where I is the unit matrix, λ_k are the eigenvalues of A,

$$\varphi_1(t) = e^{\lambda_1 t}, \quad \varphi_k(t) = \int_0^t e^{\lambda_k (t-s)} \varphi_{k-1}(s)\, ds, \quad k = 2, \ldots, n.$$

Obviously, in our case

$$\lambda_1 = -\theta, \quad \lambda_k = -\frac{n-1}{\tau}, \qquad k = 2, \ldots, n.$$

Consequently,

$$\varphi_1(t) = e^{-\theta t}, \quad \varphi_k(t) = \frac{e^{-\theta t}}{\omega^{k-1}} \left(1 - e^{-\omega t} \sum_{j=0}^{k-2} \frac{(\omega t)^j}{j!}\right), \quad k = 2, \dots, n,$$

where $\omega = \dfrac{n-1}{\tau} - \theta$.

Hence, the last component $u_n(t,\tau)$ of $u(t,\tau)$ has the form

$$\begin{aligned} u_n(t,\tau) &= \int_0^t \psi_n(t-s)(g(s,\widehat{x}_n(s,\tau)) - g(s,x_n(s,\tau)))ds \\ &\quad + \int_0^t \sum_{j=1}^{n} \psi_{n-j+1}(t-s) G_{2j}(s) ds \\ &= J_1(t,\tau) + J_2(t,\tau), \end{aligned} \tag{8}$$

where

$$\psi_k(t) = \left(\frac{n-1}{\tau}\right)^{k-1} \varphi_k(t), \qquad k = 1, \dots, n,$$

the functions $G_{2j}(t)$ are the components of the vector function $G_2(t)$.

Consider the first function $J_1(t,\tau)$. Obviously, if $\tau < \tau_1 = \frac{n-1}{\theta}$ then $\psi_k(t)$ satisfy the estimates

$$0 \le \psi_k(t) \le \frac{e^{-\theta t}}{\left(1 - \frac{\theta\tau}{n-1}\right)^{k-1}}, \qquad k = 1, \dots, n, \quad t \ge 0.$$

Consequently, by the Lipschitz condition for $g(t,v)$, we have

$$\begin{aligned} |J_1(t,\tau)| &\le \int_0^t \psi_n(t-s) |g(s,\widehat{x}_n(s,\tau)) - g(s,x_n(s,\tau))| ds \\ &\le \frac{L}{\left(1 - \frac{\theta\tau}{n-1}\right)^{n-1}} \int_0^t |u_n(s,\tau)| ds. \end{aligned} \tag{9}$$

Consider the second function $J_2(t,\tau)$. Taking into account the explicit form of $G_{2j}(t)$, we can rewrite $J_2(t,\tau)$ as follows

$$J_2(t,\tau) = \frac{n-1}{\tau} \int_0^t \sum_{j=1}^{n-1} \big(\psi_{n-j+1}(t-s) - \psi_{n-j}(t-s)\big) \frac{\rho_j\big(\widehat{x}_j(s,\tau)\big)^{1+\gamma_j}}{1 + \rho_j\big(\widehat{x}_j(s,\tau)\big)^{\gamma_j}} \, ds.$$

To estimate $J_2(t,\tau)$ we use the next lemmas.

Lemma 1. *There exists $\tau_0 > 0$ such that the components of the solution to the Cauchy problem (6) satisfy the estimates*

$$0 \le \widehat{x}_j(t,\tau) \le \frac{\tau G(1+\rho_j)}{n-1}, \quad j = 1,\dots,n-1, \quad t \in [0,T], \quad 0 < \tau < \tau_0. \tag{10}$$

Proof. This lemma can be proved in a similar way as in [14,16].

Lemma 2. *The following estimates hold*

$$\int_0^t |\psi_k(t-s) - \psi_{k-1}(t-s)|\, ds \le \frac{\tau}{n-1} \frac{2}{\left(1 - \frac{\theta\tau}{n-1}\right)^{k-1}},$$

$$k = 2,\dots,n, \qquad \tau < \tau_1 = \frac{n-1}{\theta}, \qquad t \ge 0.$$

Proof. Let $k = 2$. Then,

$$\begin{aligned} |\psi_2(t) - \psi_1(t)| &= \left| \frac{n-1}{\tau}\varphi_2(t) - \varphi_1(t) \right| = \left| \frac{n-1}{\tau} \frac{e^{-\theta t}}{\omega} \left(1 - e^{-\omega t}\right) - e^{-\theta t} \right| \\ &= \left| \left(\frac{1}{1 - \frac{\theta\tau}{n-1}} - 1 \right) e^{-\theta t} \left(1 - e^{-\omega t}\right) - e^{-(\theta+\omega)t} \right| \\ &\le \frac{\tau}{n-1} \frac{\theta e^{-\theta t}}{\left(1 - \frac{\theta\tau}{n-1}\right)} + e^{-\frac{n-1}{\tau}t}. \end{aligned}$$

Hence,

$$\begin{aligned} \int_0^t |\psi_2(t-s) - \psi_1(t-s)|\, ds &\le \frac{\tau}{n-1} \frac{1 - e^{-\theta t}}{\left(1 - \frac{\theta\tau}{n-1}\right)} + \frac{\tau}{n-1} \left(1 - e^{-\frac{n-1}{\tau}t}\right) \\ &\le \frac{\tau}{n-1} \frac{2}{\left(1 - \frac{\theta\tau}{n-1}\right)}. \end{aligned}$$

Let $k > 2$. By definition,

$$\begin{aligned} |\psi_k(t) - \psi_{k-1}(t)| &= \left| \left(\frac{n-1}{\tau}\right)^{k-1} \varphi_k(t) - \left(\frac{n-1}{\tau}\right)^{k-2} \varphi_{k-1}(t) \right| \\ &= \left| \left(\frac{n-1}{\tau}\right)^{k-1} \frac{e^{-\theta t}}{\omega^{k-1}} \left(1 - e^{-\omega t} \sum_{j=0}^{k-2} \frac{(\omega t)^j}{j!} \right) \right. \\ &\quad \left. - \left(\frac{n-1}{\tau}\right)^{k-2} \frac{e^{-\theta t}}{\omega^{k-2}} \left(1 - e^{-\omega t} \sum_{j=0}^{k-3} \frac{(\omega t)^j}{j!} \right) \right| \end{aligned}$$

$$
\begin{aligned}
&= \left| \left(\frac{n-1}{\tau}\right)^{k-1} \frac{e^{-\theta t}}{\omega^{k-1}} \left(1 - e^{-\omega t} \sum_{j=0}^{k-2} \frac{(\omega t)^j}{j!}\right) \right. \\
&\quad - \left(\frac{n-1}{\tau}\right)^{k-2} \frac{e^{-\theta t}}{\omega^{k-2}} \left(1 - e^{-\omega t} \sum_{j=0}^{k-2} \frac{(\omega t)^j}{j!}\right) \\
&\quad \left. - \left(\frac{n-1}{\tau}\right)^{k-2} \frac{e^{-(\theta+\omega)t}}{\omega^{k-2}} \frac{(\omega t)^{k-2}}{(k-2)!} \right| \\
&= \left| \frac{\tau}{n-1} \frac{\theta e^{-\theta t}}{\left(1 - \frac{\theta\tau}{n-1}\right)^{k-1}} \left(1 - e^{-\omega t} \sum_{j=0}^{k-2} \frac{(\omega t)^j}{j!}\right) \right. \\
&\quad \left. - \left(\frac{n-1}{\tau}\right)^{k-2} \frac{t^{k-2}}{(k-2)!} e^{-\frac{n-1}{\tau}t} \right| \\
&\le \frac{\tau}{n-1} \frac{\theta e^{-\theta t}}{\left(1 - \frac{\theta\tau}{n-1}\right)^{k-1}} + \left(\frac{n-1}{\tau}\right)^{k-2} \frac{t^{k-2}}{(k-2)!} e^{-\frac{n-1}{\tau}t}.
\end{aligned}
$$

Consequently,

$$
\begin{aligned}
\int_0^t |\psi_k(t-s) - \psi_{k-1}(t-s)|\, ds &\le \frac{\tau}{n-1} \frac{1 - e^{-\theta t}}{\left(1 - \frac{\theta\tau}{n-1}\right)^{k-1}} \\
&\quad + \int_0^t \left(\frac{n-1}{\tau}\right)^{k-2} \frac{(t-s)^{k-2}}{(k-2)!} e^{-\frac{n-1}{\tau}(t-s)}\, ds \\
&= \frac{\tau}{n-1} \frac{1 - e^{-\theta t}}{\left(1 - \frac{\theta\tau}{n-1}\right)^{k-1}} + \frac{\tau}{n-1} \left(1 - e^{-\frac{n-1}{\tau}t} \sum_{j=0}^{k-2} \frac{\left(\frac{n-1}{\tau}t\right)^j}{j!}\right) \\
&\le \frac{\tau}{n-1} \frac{2}{\left(1 - \frac{\theta\tau}{n-1}\right)^{k-1}}.
\end{aligned}
$$

The lemma is proved.

Using Lemma 1, for $\tau < \tau_0$, we have

$$
\frac{\rho_j\big(\widehat{x}_j(s,\tau)\big)^{1+\gamma_j}}{1 + \rho_j\big(\widehat{x}_j(s,\tau)\big)^{\gamma_j}} \le \rho_j \left(\frac{\tau G(1+\rho_j)}{n-1}\right)^{1+\gamma_j}, \quad j = 1, \dots, n-1, \quad t \in [0, T].
$$

Taking into account the conditions

$$
0 \le \rho_k \le \rho, \quad 0 < \gamma \le \gamma_k, \quad k = 1, \dots, n-1,
$$

we obtain

$$\rho_j \left(\frac{\tau G(1+\rho_j)}{n-1} \right)^{1+\gamma_j} \le \rho \left(\frac{\tau G(1+\rho)}{n-1} \right)^{1+\gamma}, \quad j = 1, \dots, n-1,$$

for $\tau < \tau_2 = \frac{n-1}{G(1+\rho)}$.

Hence, by Lemma 2, for $\tau \le \tau_* < \min\{\tau_0, \tau_1, \tau_2\}$, we obtain

$$|J_2(t,\tau)| \le \rho \left(\frac{\tau G(1+\rho)}{n-1} \right)^{1+\gamma} \sum_{j=1}^{n-1} \frac{2}{\left(1 - \frac{\theta\tau_*}{n-1}\right)^j}$$

$$= 2\rho \left(\frac{\tau G(1+\rho)}{n-1} \right)^{1+\gamma} \left(\frac{1}{\left(1 - \frac{\theta\tau_*}{n-1}\right)^{n-1}} - 1 \right) \frac{n-1}{\theta\tau_*}. \tag{11}$$

By (9) and (11), for $\tau \le \tau_*$, from (8) we have

$$|u_n(t,\tau)| \le M \int_0^t |u_n(s,\tau)|\, ds + K\rho(1+\rho)^{1+\gamma}\tau^{1+\gamma},$$

where

$$M = \frac{L}{\left(1 - \frac{\theta\tau_*}{n-1}\right)^{n-1}}, \quad K = 2\left(\frac{G}{n-1} \right)^{1+\gamma} \left(\frac{1}{\left(1 - \frac{\theta\tau_*}{n-1}\right)^{n-1}} - 1 \right) \frac{n-1}{\theta\tau_*}.$$

Applying Gronwall's inequality (for example, see [26]), we obtain

$$|u_n(t,\tau)| \le K e^{Mt} \rho (1+\rho)^{1+\gamma} \tau^{1+\gamma}.$$

Hence, the following estimate holds

$$|\widehat{x}_n(t,\tau) - x_n(t,\tau)| \le K e^{MT} \rho (1+\rho)^{1+\gamma} \tau^{1+\gamma} \tag{12}$$

on every segment $[0,T]$.

In view of Theorem 2 the sequence $\{x_n(t,\tau)\}$ converges uniformly to the solution $z(t)$ to the Cauchy problem (4) on every segment $[0,T]$; moreover, (5) holds. Then, from (12) we have the uniform convergence

$$\widehat{x}_n(t,\tau) \to z(t), \qquad \tau \to 0, \quad t \in [0,T].$$

Theorem 3 is proved.

Corollary 1. *The following estimate holds*

$$\max_{t\in[0,T]} |\widehat{x}_n(t,\tau) - z(t)| \le c_1\tau + c_2\tau^{1+\gamma}, \qquad \tau \ll 1,$$

where $c_1 > 0$ depends on θ, G, L, T, n, and $c_2 > 0$ depends on θ, G, L, T, n, ρ, γ.

Proof. Corollary 1 follows immediately from (5) and (12).

It follows from Theorem 3 that it is sufficient to solve the Cauchy problem (4) for one ordinary differential equation in order to find approximately the last component $\widehat{x}_n(t,\tau)$ of the solution to (6) for $\tau \ll 1$. This result gives us an effective method for approximate calculating $\widehat{x}_n(t,\tau)$. Moreover, the less τ, the more exact the method.

3 Conclusion

We considered the class of the systems of nonlinear ordinary differential equations with parameters. In particular, systems of such type arise when modeling the multistage synthesis of a substance. We studied properties of the solutions to the systems and proposed a method for approximate solving the systems in the case of very large coefficients. We established the approximation estimates and showed that the convergence rate depends on the parameters characterizing the nonlinearity of the systems. Moreover, the larger the coefficients of the systems, the more exact the approximate solutions. Owing to these causes, this method allows us to avoid difficulties arising inevitably when solving systems of nonlinear differential equations with very large coefficients. As an application, the proposed method can be used for approximate finding the concentration of the final product of the multistage synthesis in the case of very rapid passages from one stage to the other.

References

1. Goodwin, B.C.: Oscillatory behavior of enzymatic control processes. Adv. Enzyme Reg. **3**, 425–439 (1965)
2. Tyson, J.J., Othmer, H.G.: The dynamics of feedback control circuits in biochemical pathways. Prog. Theor. Biol. **5**, 1–62 (1978)
3. Smolen, P., Baxter, D.A., Byrne, J.H.: Modeling transcriptional control in gene networks – methods, recent results, and future directions. Bull. Math. Biol. **62**, 247–292 (2000)
4. de Jong, H.: Modeling and simulation of genetic regulatory systems: a literature review. J. Comput. Biol. **9**, 69–103 (2002)
5. Likhoshvai, V.A., Fadeev, S.I., Demidenko, G.V., Matushkin, Y.G.: Modeling multistage synthesis without branching by a delay equation. Sib. Zh. Ind. Mat. **7**, 73–94 (2004)
6. Demidenko, G.V., Likhoshvai, V.A.: On differential equations with retarded argument. Sib. Math. J. **46**, 417–430 (2005)
7. Demidenko, G.V., Khropova, Y.E.: Matrix process modelling: on properties of solutions of one delay differential equation. In: Proceedings of the Fifth International Conference on Bioinformatics of Genome Regulation and Structure, vol. 3, pp. 38–42. Institute of Cytology and Genetics, Novosibirsk (2006)
8. Demidenko, G.V., Likhoshvai, V.A., Kotova, T.V., Khropova, Y.E.: On one class of systems of differential equations and on retarded equations. Sib. Math. J. **47**, 45–54 (2006)

9. Demidenko, G.V., Likhoshvai, V.A., Mudrov, A.V.: On the relationship between solutions of delay differential equations and infinite-dimensional systems of differential equations. Diff. Equ. **45**, 33–45 (2009)
10. Demidenko, G.V., Mel'nik, I.A.: On a method of approximation of solutions to delay differential equations. Sib. Math. J. **51**, 419–434 (2010)
11. Demidenko, G.V., Kotova, T.V.: Limit properties of solutions to one class of systems of differential equations with parameters. J. Anal. Appl. **8**, 63–74 (2010)
12. Demidenko, G.V.: Systems of differential equations of higher dimension and delay equations. Sib. Math. J. **53**, 1021–1028 (2012)
13. Demidenko, G.V.: On classes of systems of differential equations of large dimensions and delay equations. Math. Forum **5**, 45–56 (2011). (Itogi Nauki. Yug Rossii)
14. Kotova, T.V., Mel'nik, I.A.: On properties of solutions of one nonlinear system to differential equations with parameters. Preprint/Sobolev Inst. Mat. **253**, 17 (2010). Novosibirsk
15. Matveeva, I.I., Mel'nik, I.A.: On the properties of solutions to a class of nonlinear systems of differential equations of large dimension. Sib. Math. J. **53**, 248–258 (2012)
16. Uvarova, I.A.: On a system of nonlinear differential equations of high dimension. J. Appl. Ind. Math. **8**, 594–603 (2014)
17. Krasovskii, N.N.: The approximation of a problem of analytic design of controls in a system with time-lag. J. Appl. Math. Mech. **28**, 876–885 (1964)
18. Repin, Y.M.: On the approximate replacement of systems with lag by ordinary dynamical systems. J. Appl. Math. Mech. **29**, 254–264 (1965)
19. Banks, H.T., Burns, J.A.: Hereditary control problems: numerical methods based on averaging approximations. SIAM J. Control Optim. **16**, 169–208 (1978)
20. Györi, I.: Two approximation techniques for functional differential equations. Comput. Math. Appl. **16**, 195–214 (1988)
21. Kraszanai, B., Györi, I., Pituk, M.: The modified chain method for a class of delay differential equations arising in neural networks. Math. Comput. Model. **51**, 452–460 (2010)
22. Ilika, S.A., Cherevko, I.M.: Approximation of nonlinear differential functional equations. J. Math. Sci. **190**, 669–682 (2013)
23. Matveeva, I.I., Popov, A.M.: Matrix process modelling: dependence of solutions of a system of differential equations on parameter. In: Proceedings of the Fifth International Conference on Bioinformatics of Genome Regulation and Structure, vol. 3, pp. 82–85. Institute of Cytology and Genetics, Novosibirsk (2006)
24. Matveeva, I.I.: On properties of solutions to a system of differential equations with a parameter. J. Anal. Appl. **7**, 75–84 (2009)
25. Gel'fand, I.M., Shilov, G.E.: Generalized Functions, Vol. 3: Theory of Differential Equations. Academic Press, New York, London (1967)
26. Hartman, P.: Ordinary Differential Equations. SIAM, Philadelphia (2002)

A Davidon-Fletcher-Powell Type Quasi-Newton Method to Solve Fuzzy Optimization Problems

Debdas Ghosh(✉)

Department of Mathematical Sciences,
Indian Institute of Technology (BHU) Varanasi,
Varanasi 221005, Uttar Pradesh, India
debdas.mat@iitbhu.ac.in, debdas.email@gmail.com

Abstract. In this article, a *Davidon-Fletcher-Powell* type quasi-Newton method is proposed to capture *nondominated solutions* of fuzzy optimization problems. The functions that we attempt to optimize here are multivariable fuzzy-number-valued functions. The decision variables are considered to be crisp. Towards developing the *quasi-Newton method*, the notion of generalized Hukuhara difference between fuzzy numbers, and hence *generalized Hukuhara differentiability* for multi-variable fuzzy-number-valued functions are used. In order to generate the iterative points, the proposed technique produces a sequence of positive definite inverse Hessian approximations. The convergence result and an algorithm of the developed method are also included. It is found that the sequence in the proposed method has *superlinear convergence rate*. To illustrate the developed technique, a numerical example is exhibited.

Keywords: Quasi-Newton method · Generalized-Hukuhara differentiability · Fuzzy optimization · Nondominated solution

1 Introduction

In order to deal with imprecise nature of the objective and constraint functions in a decision-making problem, fuzzy optimization problems are widely studied since the seminal work by Bellman and Zadeh [6], in 1970. The research article by Cadenas and Verdegay [7], the monograph by Słowínski [29], and the references therein are rich stream of this topic. Very recently, in a survey article, Luhandjula [26] reported the milestones and perspective of the theories and applications of fuzzy optimization. The survey books by Lai and Hwang [20,21] and by Lodwick and Kacprzyk [22] explored a perceptive overview on the development of fuzzy optimization problems. Recently, Ghosh and Chakraborty published a fuzzy geometrical view [9,14,15] on fuzzy optimization problems [16–19].

In order to solve an unconstrained fuzzy optimization problem, recently, Pirzada and Pathak [23] and Chalco-Cano et al. [8] developed a Newton method. Much similar to fuzzy optimization, Ghosh [12] derived a Newton method [11] and a quasi-Newton method [13] for interval optimization problems.

© Springer Nature Singapore Pte Ltd. 2017
D. Giri et al. (Eds.): ICMC 2017, CCIS 655, pp. 232–245, 2017.
DOI: 10.1007/978-981-10-4642-1_20

The real life optimization models often need to optimize a fuzzy function over a real data set. Mathematically, the problem is the following:

$$\min_{x \in \mathbb{R}^n} \widetilde{f}(x),$$

where $\widetilde{f}$ is a fuzzy-number-valued function. There is plethora of studies and numerical algorithms [1,28] to effectively tackle unconstrained conventional optimization problems. However, the way to apply those methods to solve fuzzy optimization problems is not apparent. *In this article, we develop a Davidon-Fletcher-Powell type quasi-Newton method for an unconstrained fuzzy optimization problem.*

In order to capture the optimum candidates [4,28] for a smooth optimization problem, it is natural to use the notion of differentiability. However, to find the optimal points of a fuzzy optimization problem, in addition to the idea on *differentiability of the fuzzy function*, identification of an appropriate *ordering of fuzzy numbers* is of utmost importance. Because, unlike the real number set, the set of all fuzzy numbers is not linearly ordered [16].

Towards developing the notion of *differentiability of fuzzy functions*, Dubois and Prade [10] used the extension principle, and also exhibited that a fuzzy extension of the conventional definition of differentiation requires further enquiry. Recently, the idea of Hukuhara differentiability (H-differentiability) [3,24] received substantial attention in fuzzy optimization theory. The concept of H-fuzzy-differentiation is rigorously discussed in [2]. Stefanini [30] proposed generalizations of H-differentiability (gH-differentiability) and its application in fuzzy differential equations. Bede and Stefanini [5] gave a generalized H-differentiability of fuzzy-valued functions. Chalco-Cano et al. [8] reported that gH-derivative is the most general concept for differentiability of fuzzy functions. Thus, *in this paper, we employ the gH-derivative and its calculus* [8,33] to derive the quasi-Newton method for fuzzy optimization problems.

There have been extensive literature on *ordering of fuzzy numbers* including the research articles [25,27,31]. References of [32] reports main stream on ordering of fuzzy numbers. *In this paper, we use the fuzzy-max ordering of fuzzy numbers of* Ramík and Rimanek [27]. There are two reasons behind this choice. First, it is a partial ordering in the space of fuzzy numbers [23]. Second, it has insightful association [33] with the optimality notion on fuzzy optimization.

The rest of the article is organized in the following sequence. In Sect. 2, the notations and terminologies are given which are used throughout the paper. In Sect. 3, we derive a quasi-Newton method. Convergence analysis of the proposed method is presented in Sect. 4. Section 5 includes an illustrative numerical example. Finally, we give a brief conclusions and scopes for future research in Sect. 6.

2 Preliminaries

We use upper and lower case letters with a tildebar ($\widetilde{A}, \widetilde{B}, \widetilde{C}, \ldots$ and $\widetilde{a}, \widetilde{b}, \widetilde{c}, \ldots$) to denote fuzzy subsets of $\mathbb{R}$. The membership function of a fuzzy set $\widetilde{A}$ of $\mathbb{R}$ is represented by $\mu(x|\widetilde{A})$, for x in $\mathbb{R}$, with $\mu(\mathbb{R}) \subseteq [0, 1]$.

2.1 Fuzzy Numbers

Definition 1 (α-cut of a fuzzy set [16]). *The α-cut of a fuzzy set $\widetilde{A}$ of $\mathbb{R}$ is denoted by $\widetilde{A}(\alpha)$ and is defined by:*

$$\widetilde{A}(\alpha) = \begin{cases} \{x : \mu(x|\widetilde{A}) \geq \alpha\} & if\ 0 < \alpha \leq 1 \\ closure\{x : \mu(x|\widetilde{A}) > 0\} & if\ \alpha = 0. \end{cases}$$

Definition 2 (Fuzzy number [16]). *A fuzzy set $\widetilde{N}$ of $\mathbb{R}$ is called a fuzzy number if its membership function μ has the following properties:*

(i) $\mu(x|\widetilde{N})$ *is upper semi-continuous,*
(ii) $\mu(x|\widetilde{N}) = 0$ *outside some interval* $[a, d]$*, and*
(iii) there exist real numbers b *and* c *satisfying* $a \leq b \leq c \leq d$ *such that* $\mu(x|\widetilde{N})$ *is increasing on* $[a, b]$ *and decreasing on* $[c, d]$*, and* $\mu(x|\widetilde{N}) = 1$ *for each* x *in* $[b, c]$*.*

In particular, if $b = c$, and the parts of the membership functions $\mu(x|\widetilde{N})$ in $[a, b]$ and $[c, d]$ are linear, the fuzzy number is called a *triangular fuzzy number*, denoted by $(a/c/d)$. We denote the set of all fuzzy numbers on $\mathbb{R}$ by $\mathcal{F}(\mathbb{R})$.

Since $\mu(x|\widetilde{a})$ is upper semi-continuous for a fuzzy number $\widetilde{a}$ the α-cut of $\widetilde{a}$, $\widetilde{a}(\alpha)$ is a closed and bounded interval of $\mathbb{R}$ for all α in $[0, 1]$. We write

$$\widetilde{a}(\alpha) = \left[\widetilde{a}_\alpha^L,\ \widetilde{a}_\alpha^U\right].$$

Let $\oplus$ and $\odot$ denote the extended addition and multiplication. According to the well-known *extension principle*, the membership function of $\widetilde{a} \otimes \widetilde{b}$ ($\otimes = \oplus$ or $\odot$) is defined by

$$\mu(z|\widetilde{a} \otimes \widetilde{b}) = \sup_{x \times y = z} \min\left\{\mu(x|\widetilde{a}),\ \mu(y|\widetilde{b})\right\}.$$

For any $\widetilde{a}$ and $\widetilde{b}$ in $\mathcal{F}(\mathbb{R})$, the α-cut (for any α in $[0, 1]$) of their addition and scalar multiplication can be obtained by:

$$\left(\widetilde{a} \oplus \widetilde{b}\right)(\alpha) = \left[\widetilde{a}_\alpha^L + \widetilde{b}_\alpha^L,\ \widetilde{a}_\alpha^U + \widetilde{b}_\alpha^U\right] \text{ and}$$

$$\left(\lambda \odot \widetilde{a}\right)(\alpha) = \begin{cases} \left[\lambda\widetilde{a}_\alpha^L,\ \lambda\widetilde{a}_\alpha^U\right] & \text{if } \lambda \geq 0, \\ \left[\lambda\widetilde{a}_\alpha^U,\ \lambda\widetilde{a}_\alpha^L\right] & \text{if } \lambda < 0. \end{cases}$$

Definition 3 (Generalized Hukuhara difference [30]). *Let $\widetilde{a}$ and $\widetilde{b}$ be two fuzzy numbers. If there exists a fuzzy number $\widetilde{c}$ such that $\widetilde{c} \oplus \widetilde{b} = \widetilde{a}$ or $\widetilde{b} = \widetilde{a} \ominus \widetilde{c}$, then $\widetilde{c}$ is said to be generalized Hukuhara difference (gH-difference) between $\widetilde{a}$ and $\widetilde{b}$. Hukuhara difference between $\widetilde{a}$ and $\widetilde{b}$ is denoted by $\widetilde{a} \ominus_{gH} \widetilde{b}$.*

In terms of α-cut, for all $\alpha \in [0, 1]$, we have

$$\left(\widetilde{a} \ominus_{gH} \widetilde{b}\right)(\alpha) = \left[\min\left\{\widetilde{a}_\alpha^L - \widetilde{b}_\alpha^L,\ \widetilde{a}_\alpha^U - \widetilde{b}_\alpha^U\right\},\ \max\left\{\widetilde{a}_\alpha^L - \widetilde{b}_\alpha^L,\ \widetilde{a}_\alpha^U - \widetilde{b}_\alpha^U\right\}\right].$$

2.2 Fuzzy Functions

Let $\widetilde{f} : \mathbb{R}^n \to \mathcal{F}(\mathbb{R})$ be a fuzzy function. For each x in $\mathbb{R}^n$, we present the α-cuts of $\widetilde{f}(x)$ by

$$\widetilde{f}(x)(\alpha) = \left[\widetilde{f}_\alpha^L(x),\ \widetilde{f}_\alpha^U(x)\right] \text{ for all } \alpha \in [0,1].$$

The functions $\widetilde{f}_\alpha^L$ and $\widetilde{f}_\alpha^U$ are, evidently, two real-valued functions on $\mathbb{R}^n$ and are called the lower and upper functions, respectively.

With the help of gH-difference between two fuzzy numbers, gH-differentiability of a fuzzy function is defined as follows.

Definition 4 (gH-differentiability of fuzzy functions [8]). *Let $\widetilde{f} : \mathbb{R}^n \to \mathcal{F}(\mathbb{R})$ be a fuzzy function and $x_0 = \left(x_1^0, x_2^0, \cdots, x_n^0\right)$ be an element of $\mathbb{R}^n$. For each $i = 1, 2, \cdots, n$, we define a fuzzy function $\widetilde{h}_i : \mathbb{R} \to \mathcal{F}(\mathbb{R})$ as follows*

$$\widetilde{h}_i(x_i) = \widetilde{f}\left(x_1^0, \cdots, x_{i-1}^0, x_i, x_{i+1}^0, \cdots, x_n^0\right).$$

We say h_i is gH-differentiable if the following limit exists

$$\lim_{t_i \to 0} \frac{\widetilde{h}_i(x_i^0 + t_i) \ominus_{gH} \widetilde{h}_i(x_i^0)}{t_i}.$$

If $\widetilde{h}_i$ is gH-differentiable, then we say that $\widetilde{f}$ has the i-th partial gH-derivative at x_0 and is denoted by $\frac{\partial \widetilde{f}}{\partial x_i}(x_0)$.

The function $\widetilde{f}$ is said to be gH-differentiable at $x_0 \in \mathbb{R}^n$ if all the partial gH-derivatives $\frac{\partial \widetilde{f}}{\partial x_1}(x_0)$, $\frac{\partial \widetilde{f}}{\partial x_2}(x_0)$, $\cdots$, $\frac{\partial \widetilde{f}}{\partial x_n}(x_0)$ exist on some neighborhood of x_0 and are continuous at x_0.

Proposition 1 (See [8]). *If a fuzzy function $\widetilde{f} : \mathbb{R}^n \to \mathcal{F}(\mathbb{R})$ is gH-differentiable at $x_0 \in \mathbb{R}^n$, then for each $\alpha \in [0,1]$, the real-valued function $\widetilde{f}_\alpha^L + \widetilde{f}_\alpha^U$ is differentiable at x_0. Moreover,*

$$\frac{\partial \widetilde{f}_\alpha^L}{\partial x_i}(x_0) + \frac{\partial \widetilde{f}_\alpha^U}{\partial x_i}(x_0) = \frac{\partial(\widetilde{f}_\alpha^L + \widetilde{f}_\alpha^U)}{\partial x_i}(x_0).$$

Definition 5 (gH-gradient [8]). *The gH-gradient of a fuzzy function $\widetilde{f} : \mathbb{R}^n \to \mathcal{F}(\mathbb{R})$ at a point $x_0 \in \mathbb{R}^n$ is defined by*

$$\left(\frac{\partial \widetilde{f}(x_0)}{\partial x_1}, \frac{\partial \widetilde{f}(x_0)}{\partial x_2}, \cdots, \frac{\partial \widetilde{f}(x_0)}{\partial x_n}\right)^t.$$

We denote this gH-gradient by $\nabla \widetilde{f}(x_0)$.

We define an m-times *continuously gH-differentiable* fuzzy function $\widetilde{f}$ as a function whose all the partial gH-derivatives of order m exist and are continuous. Then, we have the following immediate result.

Proposition 2 (See [8]). *Let $\widetilde{f}$ be a fuzzy function. Let at $x_0 \in \mathbb{R}^n$, the function $\widetilde{f}$ be m-times gH-differentiable. Then, the real-valued function $\widetilde{f}_\alpha^L + \widetilde{f}_\alpha^U$ is m-times differentiable at x_0.*

Definition 6 (*gH*-Hessian [8]). *Let the fuzzy function $\widetilde{f}$ be twice gH-differentiable at x_0. Then, for each i, the function $\frac{\partial \widetilde{f}}{\partial x_i}$ is gH-differentiable at x^0. The second order partial gH-derivative can be calculated through $\frac{\partial^2 \widetilde{f}}{\partial x_i \partial x_j}$. The gH-Hessian of $\widetilde{f}$ at x_0 can be captured by the square matrix*

$$\nabla^2 \widetilde{f}(x_0) = \left[\frac{\partial^2 \widetilde{f}}{\partial x_i \partial x_j}(x_0) \right]_{n \times n}.$$

2.3 Optimality Concept

Definition 7 (Dominance relation between fuzzy numbers [23]). *Let $\widetilde{a}$ and $\widetilde{b}$ be two fuzzy numbers. For any $\alpha \in [0,1]$, let $\widetilde{a}_\alpha = \left[\widetilde{a}_\alpha^L, \widetilde{a}_\alpha^U\right]$ and $\widetilde{b}_\alpha = \left[\widetilde{b}_\alpha^L, \widetilde{b}_\alpha^U\right]$. We say $\widetilde{a}$ dominates $\widetilde{b}$ if $\widetilde{a}_\alpha^L \leq \widetilde{b}_\alpha^L$ and $\widetilde{a}_\alpha^U \leq \widetilde{b}_\alpha^U$ for all $\alpha \in [0,1]$. If $\widetilde{a}$ dominates $\widetilde{b}$, then we write $\widetilde{a} \preceq \widetilde{b}$. The fuzzy number $\widetilde{b}$ is said to be strictly dominated by $\widetilde{a}$, if $\widetilde{a} \preceq \widetilde{b}$ and there exists $\beta \in [0,1]$ such that $\widetilde{a}_\beta^L < \widetilde{b}_\beta^L$ or $\widetilde{a}_\beta^U < \widetilde{b}_\beta^U$. If $\widetilde{a}$ strictly dominates $\widetilde{b}$, then we write $\widetilde{a} \prec \widetilde{b}$.*

Definition 8 (Non-dominated solution [23]). *Let $\widetilde{f} : \mathbb{R}^n \to \mathcal{F}(\mathbb{R})$ be a fuzzy function and we intend to find a solution of '$\min_{x \in \mathbb{R}^n} \widetilde{f}(x)$'. A point $\bar{x} \in \mathbb{R}^n$ is said to be a locally non-dominated solution if for any $\epsilon > 0$, there exists no $x \in N_\epsilon(\bar{x})$ such that $\widetilde{f}(x) \preceq \widetilde{f}(\bar{x})$, where $N_\epsilon(\bar{x})$ denotes ϵ-neighborhood of $\bar{x}$. A local non-dominated solution is called a local solution of '$\min_{x \in \mathbb{R}^n} \widetilde{f}(x)$'.*

For local non-dominated solution, the following result is proved in [8].

Proposition 3 (See [8]). *Let $\widetilde{f} : \mathbb{R}^n \to \mathcal{F}(\mathbb{R})$ be a fuzzy function. If x^* is a local minimizer of the real-valued function $\widetilde{f}_\alpha^L + \widetilde{f}_\alpha^U$ for all $\alpha \in [0,1]$, then x^* is a locally non-dominated solution of '$\min_{x \in \mathbb{R}^n} \widetilde{f}(x)$'.*

3 Quasi-Newton Method

In this section, we consider to solve the following unconstrained Fuzzy Optimization Problem:

$$(FOP) \quad \min_{x \in \mathbb{R}^n} \widetilde{f}(x),$$

where $\widetilde{f} : \mathbb{R}^n \to \mathcal{F}(\mathbb{R})$ is a multi-variable fuzzy-number-valued function. On finding nondominated solution of the problem, we note that the existing Newton method (see [8,23]) requires computation of the inverse of the concerned Hessian.

However, computation of the inverse of Hessian is cost effective. Thus in this article, we intend to develop a quasi-Newton method to sidestep the computational cost of the existing Newton method. Towards this end, for the considered FOP we assume that at each of the following generated sequential points x_k, the function $\widetilde{f}$, its gH-gradient and gH-Hessian are well-defined. Therefore, according to Propositions 1 and 2, we can calculate $\widetilde{f}^L_\alpha(x_k)$, $\widetilde{f}^U_\alpha(x_k)$, $\nabla\widetilde{f}^L_\alpha(x_k)$, $\nabla\widetilde{f}^U_\alpha(x_k)$, $\nabla^2\widetilde{f}^L_\alpha(x_k)$ and $\nabla^2\widetilde{f}^U_\alpha(x_k)$ for all $\alpha\in[0,1]$, for all $k=0,1,2,\cdots$. Hence, we can have a quadratic approximations of the lower and the upper functions $\widetilde{f}^L_\alpha$ and $\widetilde{f}^U_\alpha$ at each x_k.

Let the quadratic approximation models of the functions $\widetilde{f}^L_\alpha$ and $\widetilde{f}^U_\alpha$ at x_{k+1} be

$$h^L_\alpha(x)=\widetilde{f}^L_\alpha(x_{k+1})+\nabla\widetilde{f}^L_\alpha(x_{k+1})^t(x-x_{k+1})+\tfrac{1}{2}(x-x_{k+1})^t\ \nabla^2\widetilde{f}^L_\alpha(x_{k+1})\ (x-x_{k+1})$$

and

$$h^U_\alpha(x)=\widetilde{f}^U_\alpha(x_{k+1})+\nabla\widetilde{f}^U_\alpha(x_{k+1})^t(x-x_{k+1})+\tfrac{1}{2}(x-x_{k+1})^t\ \nabla^2\widetilde{f}^U_\alpha(x_{k+1})\ (x-x_{k+1}),$$

which satisfy the interpolating conditions

$$h^L_\alpha(x_{k+1})=\widetilde{f}^L_\alpha(x_{k+1}),\ h^U_\alpha(x_{k+1})=\widetilde{f}^U_\alpha(x_{k+1}),\ \nabla h^L_\alpha(x_{k+1})=\nabla\widetilde{f}^L_\alpha(x_{k+1}),\ \text{and}$$

$$\nabla h^U_\alpha(x_{k+1})=\nabla\widetilde{f}^U_\alpha(x_{k+1}).$$

The derivatives of h^L_α and h^U_α yield

$$\begin{aligned}\nabla h^L_\alpha(x)&=\nabla\widetilde{f}^L_\alpha(x_{k+1})+\nabla^2\widetilde{f}^L_\alpha(x_{k+1})\ (x-x_{k+1})\\ \text{and }\nabla h^U_\alpha(x)&=\nabla\widetilde{f}^U_\alpha(x_{k+1})+\nabla^2\widetilde{f}^U_\alpha(x_{k+1})\ (x-x_{k+1}).\end{aligned}\tag{1}$$

In the next, the Newton method (see [8,23]) attempts to find x in terms of x_{k+1} as follows

$$x=x_{k+1}-\left[\nabla^2\phi(x_{k+1})\right]^{-1}\nabla\phi(x_{k+1}),$$

where $\phi(x)=\int_0^1\left(\widetilde{f}^L_\alpha(x)+\widetilde{f}^U_\alpha(x)\right)d\alpha$. However, due to inherent computational difficulty to find $[\nabla^2\phi(x_{k+1})]^{-1}$, it is often suggested to consider an appropriate approximation. Let A_{k+1} be an approximation of $[\nabla^2\phi(x_{k+1})]^{-1}$. Then from (1) setting $x=x_k$, $\delta_k=x_{k+1}-x_k$, $\beta^L_{\alpha k}=\nabla\widetilde{f}^L_\alpha(x_{k+1})-\nabla\widetilde{f}^L_\alpha(x_k)$ and $\beta^U_{\alpha k}=\nabla\widetilde{f}^U_\alpha(x_{k+1})-\nabla\widetilde{f}^U_\alpha(x_k)$, we obtain

$$\beta^L_{\alpha k}=\nabla^2\widetilde{f}^L_\alpha(x_{k+1})\delta_k\ \text{and}\ \beta^U_{\alpha k}=\nabla^2\widetilde{f}^U_\alpha(x_{k+1})\delta_k$$

$$\begin{aligned}\implies\beta^L_{\alpha k}+\beta^U_{\alpha k}&=\left(\nabla^2\widetilde{f}^L_\alpha(x_{k+1})+\nabla^2\widetilde{f}^U_\alpha(x_{k+1})\right)\delta_k\ \text{for all}\ \alpha\in[0,1]\\ \implies\int_0^1\left(\beta^L_{\alpha k}+\beta^U_{\alpha k}\right)d\alpha&=\left(\int_0^1\nabla^2\left(\widetilde{f}^L_\alpha+\widetilde{f}^U_\alpha\right)(x_{k+1})d\alpha\right)\delta_k\\ \implies\int_0^1\left(\nabla\phi_\alpha(x_{k+1})-\nabla\phi_\alpha(x_k)\right)d\alpha&=\left(\int_0^1\nabla^2\phi_\alpha(x_{k+1})d\alpha\right)\delta_k,\end{aligned}$$

$$\text{where } \phi_\alpha(x) = \widetilde{f}_\alpha^L(x) + \widetilde{f}_\alpha^U(x)$$
$$\implies \nabla\phi(x_{k+1}) - \nabla\phi(x_k) = \nabla^2\phi(x_{k+1})\delta_k, \text{ where } \phi(x) = \int_0^1 \phi_\alpha(x)d\alpha$$
$$\implies A_{k+1}\Phi_k = \delta_k, \text{ where } \Phi_k = \nabla\phi(x_{k+1}) - \nabla\phi(x_k).$$

According to Proposition 3, to obtain nondominated solutions of (FOP), we need to have solutions of the equation $\nabla(\widetilde{f}_\alpha^L + \widetilde{f}_\alpha^U)(x) = 0$. In order to capture solutions of this equation, much similar to the Newton method [23], the above procedure clearly suggests to consider the following generating sequence

$$\delta_k = x_{k+1} - x_k = A_{k+1}\Phi_k$$
$$\implies x_{k+1} = x_k + A_{k+1}\Phi_k.$$

If A_k is an appropriate approximation of the inverse Hessian matrix $\left[\nabla^2\phi(x_{k+1})\right]^{-1}$ and $\nabla(\widetilde{f}_\alpha^L + \widetilde{f}_\alpha^U)(x_{k+1}) \approx 0$, then the equation $x_{k+1} = x_k + A_{k+1}\Phi_k$ reduces to

$$x_{k+1} = x_k - \left[\nabla^2\phi(x_{k+1})\right]^{-1}\nabla\phi(x_k),$$

which is the generating equation of the Newton method [23] and hence obviously will converge to the minimizer of $\widetilde{f}_\alpha^L + \widetilde{f}_\alpha^U$.

As we observe, the key point of the above method is to appropriately generate A_k's. Due to the inherent computational difficulty to find the inverse of the Hessian $\nabla^2\phi(x_{k+1})$ we consider an approximation that should satisfy

$$A_{k+1}\Phi_k = \delta_k. \tag{2}$$

In this article we attempt to introduce a simple *rank-two update* of the sequence $\{A_k\}$ that satisfy the above *quasi-Newton Eq.* (2).

Let A_k be the approximation of the k-th iteration. We attempt to update A_k into A_{k+1} by adding two symmetric matrices, each of rank one as follows:

$$A_{k+1} = A_k + p_k v_k v_k^t + q_k w_k w_k^t$$

where u_k and v_k are two vectors in $\mathbb{R}^n$, and p_k and q_k are two scalars which are to be determined by the quasi-Newton Eq. (2). Therefore, we now have

$$A_k\Phi_k + p_k v_k v_k^t\Phi_k + q_k w_k w_k^t\Phi_k = \delta_k. \tag{3}$$

Evidently, v_k and w_k are not uniquely determined, but their obvious choices are $v_k = \delta_k$ and $w_k = A_k\Phi_k$. Putting this values in (3), we obtain

$$p_k = \frac{1}{v_k^t\Phi_k} = \frac{1}{\delta_k^t\Phi_k} \text{ and } q_k = -\frac{1}{w_k^t\Phi_k} = -\frac{1}{\Phi_k^t A_k\Phi_k}.$$

Therefore,

$$A_{k+1} = A_k + \frac{\delta_k\delta_k^t}{\delta_k^t\Phi_k} - \frac{A_k\Phi_k\Phi_k^t A_k}{\Phi_k^t A_k\Phi_k}. \tag{4}$$

We consider this equation to generate the sequence of above mentioned inverse Hessian approximation.

Therefore, accumulating all, we follow the following sequential way (Algorithm 1) to obtain an efficient solution of the considered (FOP) $\min_{x\in\mathbb{R}^n} \widetilde{f}(x)$.

Algorithm 1. Quasi-Newton Method with Rank-two Modification to Solve FOP

Require: Given $\widetilde{f}$, the objective function
x_0, the initial point
ϵ, a termination scalar
A_0, a symmetric positive definite matrix

1: Compute $\widetilde{f}_\alpha^L$, $\widetilde{f}_\alpha^U$ and $\phi(x) = \int_0^1 (\widetilde{f}_\alpha^L(x) + \widetilde{f}_\alpha^U(x))d\alpha$
2: Set $k = 0$
3: If $\|\nabla\phi(x_k)\| < \epsilon$, then Stop
4: Compute the search direction $d_k = -A_k\nabla\phi(x_k)$
5: Compute the step length $\alpha_k := \operatorname{argmin}_{\alpha\geq 0}\ \phi(x_k + \alpha d_k)$
6: Evaluate

$$\begin{aligned} \delta_k &= \alpha_k d_k, \\ x_{k+1} &= x_k + \delta_k, \\ \Phi_k &= \nabla\phi(x_{k+1}) - \nabla\phi(x_k), \text{ and} \\ A_{k+1} &= A_k + \frac{\delta_k\delta_k^t}{\delta_k^t\Phi_k} - \frac{A_k\Phi_k\Phi_k^t A_k}{\Phi_k^t A_k\Phi_k}. \end{aligned}$$

7: Set $k = k + 1$ and go to Step 3.

4 Convergence Analysis

In this section, the convergence analysis of the proposed quasi-Newton method is performed. In the following theorem, it is found that the proposed method has *superlinear convergence rate.*

Theorem 1 (Superlinear convergence rate). *Let $\widetilde{f} : \mathbb{R}^n \to \mathcal{F}(\mathbb{R})$ be thrice continuously gH-differentiable on $\mathbb{R}^n$ and $\bar{x}$ be a point such that*

(i) $\bar{x}$ is a local minimizer of $\widetilde{f}_\alpha^L$ and $\widetilde{f}_\alpha^U$,
(ii) $\int_0^1 \nabla^2\widetilde{f}_\alpha^L(x)d\alpha$ and $\int_0^1 \nabla^2\widetilde{f}_\alpha^U(x)d\alpha$ are symmetric positive definite, and
(iii) $\int_0^1 \nabla^2\widetilde{f}_\alpha^L(x)d\alpha$ and $\int_0^1 \nabla^2\widetilde{f}_\alpha^U(x)d\alpha$ are Lipschitzian with constants γ^L and γ^U, respectively.

Then the iteration sequence $\{x_k\}$ in Algorithm 1 converges to $\bar{x}$ superlinearly if and only if

$$\lim_{k\to\infty} \frac{\left\|\left[A_{k+1}^{-1} - \nabla^2\phi(\bar{x})\right]\delta_k\right\|}{\|\delta_k\|} = 0,$$

where $\phi(x) = \int_0^1 \left(\widetilde{f}_\alpha^L(x) + \widetilde{f}_\alpha^U(x)\right)d\alpha$.

Proof. According to the hypothesis (i), at $\bar{x}$ we have

$$\nabla\phi(\bar{x}) = \int_0^1 \left(\nabla\widetilde{f}_\alpha^L(\bar{x}) + \nabla\widetilde{f}_\alpha^U(\bar{x})\right) d\alpha = 0$$

From hypothesis (ii), the Hessian matrix

$$\nabla^2\phi(\bar{x}) = \int_0^1 \nabla^2\widetilde{f}_\alpha^L(\bar{x})d\alpha + \int_0^1 \nabla^2\widetilde{f}_\alpha^U(\bar{x})d\alpha$$

is positive definite and a symmetric matrix. With the help of hypothesis (iii), the function

$$\nabla^2\phi(x) = \int_0^1 \nabla^2\widetilde{f}_\alpha^L(x)d\alpha + \int_0^1 \nabla^2\widetilde{f}_\alpha^U(x)d\alpha$$

is found to be Lipschitzian at $\bar{x}$ with constant $\gamma^L + \gamma^U$. Mathematically, there exists a neighborhood $N_\epsilon(\bar{x})$ where

$$\left\|\nabla^2\phi(x) - \nabla^2\phi(\bar{x})\right\| \leq (\gamma^L + \gamma^U)\|x - \bar{x}\| \ \ \forall\ x \in N_\epsilon(\bar{x}).$$

Towards proving the result, the following equivalence will be proved:

$$\begin{aligned} & \lim_{k\to\infty} \frac{\left\|\left[A_{k+1}^{-1} - \nabla^2\phi(\bar{x})\right]\delta_k\right\|}{\|\delta_k\|} = 0 \\ \iff & \lim_{k\to\infty} \frac{\|\Phi_{k+1}\|}{\|\delta_k\|} = 0 \\ \iff & \lim_{k\to\infty} \frac{\|x_{k+1} - \bar{x}\|}{\|x_k - \bar{x}\|} = 0. \end{aligned}$$

With the help of quasi-Newton Eq. (1), we have

$$\begin{aligned} & \left[A_{k+1}^{-1} - \nabla^2\phi(\bar{x})\right][x_{k+1} - x_k] \\ = & -\Phi_k - \nabla^2\phi(\bar{x})(x_{k+1} - x_k) \\ = & \left(\Phi_{k+1} - \Phi_k - \nabla^2\phi(\bar{x})(x_{k+1} - x_k)\right) - \Phi_{k+1}. \end{aligned}$$

Therefore,

$$\frac{\|\Phi_{k+1}\|}{\|\delta_k\|} \leq \frac{1}{\|\delta_k\|}\left[\left\|\left(A_{k+1}^{-1} - \nabla^2\phi(\bar{x})\right)\delta_k\right\| + \left\|\Phi_{k+1} - \Phi_k - \nabla^2\phi(\bar{x})\delta_k\right\|\right].$$

It is evident to note that

$$\begin{aligned} & \left\|\Phi_{k+1} - \Phi_k - \nabla^2\phi(\bar{x})\delta_k\right\| \\ = & \left\|\left(\int_0^1 \nabla^2\left(\widetilde{f}_\alpha^L + \widetilde{f}_\alpha^U\right)(x_{k+1} + t(x_{k+1} - x_k))dt\right)\delta_k - \nabla^2\phi(\bar{x})\delta_k\right\| \\ = & \left\|\left(\int_0^1 \left(\nabla^2\phi(x_{k+1} + t\delta_k) - \nabla^2\phi(\bar{x})\right)dt\right)\delta_k\right\| \\ \leq & (\gamma^L + \gamma^U)\left(\|x_{k+1} - \bar{x}\| + \|x_k - \bar{x}\|\right). \end{aligned}$$

Since $\nabla^2\phi(\bar{x})$ is positive definite, we have $\xi > 0$ and $m \in \mathbb{N}$ such that

$$\|\Phi_{k+1}\| = \|\Phi_{k+1} - \nabla\phi(\bar{x})\| \geq \xi\|x_{k+1} - \bar{x}\| \text{ for all } k \geq m.$$

Hence, we now have

$$\frac{\|\Phi_{k+1}\|}{\|\delta_k\|} \geq \frac{\xi\|x_{k+1} - \bar{x}\|}{\|x_{k+1} - \bar{x}\| + \|x_k - \bar{x}\|} = \xi\frac{c_k}{1 + c_k},$$

where $c_k = \frac{\|x_{k+1} - \bar{x}\|}{\|x_k - \bar{x}\|}$. This inequality gives

$$\begin{aligned} &\lim_{k\to\infty} \frac{c_k}{1 + c_k} = 0 \\ \Longrightarrow\ &\lim_{k\to\infty} c_k = 0 \\ \Longrightarrow\ &\lim_{k\to\infty} \frac{\|x_{k+1} - \bar{x}\|}{\|x_k - \bar{x}\|} = 0. \end{aligned}$$

This completes the proof of superlinear convergence of the sequence $\{x_k\}$ in Algorithm 1.

Conversely, since $\{x_k\}$ converges superlinearly to $\bar{x}$ and $\nabla\phi(\bar{x}) = 0$, we must have $\beta > 0$ and $p \in \mathbb{N}$ such that

$$\|\Phi_{k+1}\| \leq \beta\|x_{k+1} - \bar{x}\| \text{ for all } k \geq p.$$

Again due to superlinear convergence of $\{x_k\}$, we have

$$0 = \lim_{k\to\infty} \frac{\|x_{k+1} - \bar{x}\|}{\|x_k - \bar{x}\|} \geq \lim_{k\to\infty} \frac{\|\Phi_{k+1}\|}{\beta\|x_k - \bar{x}\|} = \lim_{k\to\infty} \tfrac{1}{\beta} \frac{\|\Phi_{k+1}\|}{\|x_{k+1} - x_k\|} \frac{\|x_{k+1} - x_k\|}{\|x_k - \bar{x}\|}.$$

Since $\lim_{k\to\infty} \frac{\|x_{k+1} - x_k\|}{\|x_k - \bar{x}\|} = 1$, this inequality implies $\lim_{k\to\infty} \frac{\|\Phi_k\|}{\|x_{k+1} - x_k\|} = 0$. Hence, the result follows.

5 Illustrative Example

In this section, an illustrative example is presented to explore the computational procedure of Algorithm 1.

Example 1. Consider the following quadratic fuzzy optimization problem:

$$\min_{(x_1,x_2)\in\mathbb{R}^2} \left(\tfrac{1}{2}/1/\tfrac{3}{2}\right) x_1^2 \oplus \left(0/\tfrac{1}{2}/1\right) x_2^2 \oplus \left(\tfrac{1}{2}/1/\tfrac{3}{2}\right) x_1x_2 \oplus \left(0/\tfrac{1}{2}/1\right) x_1 \ominus \left(0/\tfrac{1}{2}/1\right) x_2.$$

Let us consider the initial approximation to the minimizer as $x_0 = \left(x_1^0, x_2^0\right) = (0, 0)$. With the help of fuzzy arithmetic, the lower and the upper function can be obtained as

$$\widetilde{f}_\alpha^{\mathcal{L}}(x_1, x_2) = \begin{cases} \frac{1+\alpha}{2}x_1^2 + \frac{\alpha}{2}x_2^2 + \frac{1+\alpha}{2}x_1x_2 + \frac{\alpha}{2}x_1 - (1 - \frac{\alpha}{2})x_2 & \text{if } x_1 \geq 0, x_2 \geq 0 \\ \frac{1+\alpha}{2}x_1^2 + \frac{\alpha}{2}x_2^2 + \frac{3-\alpha}{2}x_1x_2 + \frac{\alpha}{2}x_1 - \frac{\alpha}{2}x_2 & \text{if } x_1 \geq 0, x_2 \leq 0 \\ \frac{1+\alpha}{2}x_1^2 + \frac{\alpha}{2}x_2^2 + \frac{3-\alpha}{2}x_1x_2 + (1 - \frac{\alpha}{2})x_1 - (1 - \frac{\alpha}{2})x_2 & \text{if } x_1 \leq 0, x_2 \geq 0 \\ \frac{1+\alpha}{2}x_1^2 + \frac{\alpha}{2}x_2^2 + \frac{1+\alpha}{2}x_1x_2 + (1 - \frac{\alpha}{2})x_1 - \frac{\alpha}{2}x_2 & \text{if } x_1 \leq 0, x_2 \geq 0 \end{cases}$$

and

$$\widetilde{f}_{\alpha}^{U}(x_1,x_2)=\begin{cases}\frac{3-\alpha}{2}x_1^2+(1-\frac{\alpha}{2})x_2^2+\frac{3-\alpha}{2}x_1x_2+(1-\frac{\alpha}{2})x_1-\frac{\alpha}{2}x_2 & \text{if } x_1\geq 0, x_2\geq 0\\ \frac{3-\alpha}{2}x_1^2+(1-\frac{\alpha}{2})x_2^2+\frac{1+\alpha}{2}x_1x_2+(1-\frac{\alpha}{2})x_1-(1-\frac{\alpha}{2})x_2 & \text{if } x_1\geq 0, x_2\leq 0\\ \frac{3-\alpha}{2}x_1^2+(1-\frac{\alpha}{2})x_2^2+\frac{1+\alpha}{2}x_1x_2+\frac{\alpha}{2}x_1-\frac{\alpha}{2}x_2 & \text{if } x_1\leq 0, x_2\geq 0\\ \frac{3-\alpha}{2}x_1^2+(1-\frac{\alpha}{2})x_2^2+\frac{3-\alpha}{2}x_1x_2+\frac{\alpha}{2}x_1-(1-\frac{\alpha}{2})x_2 & \text{if } x_1\leq 0, x_2\geq 0.\end{cases}$$

Therefore,

$$\phi(x_1,x_2)=\int_0^1\left(\widetilde{f}_{\alpha}^{L}(x_1,x_2)+\widetilde{f}_{\alpha}^{U}(x_1,x_2)\right)d\alpha=2x_1^2+x_2^2+2x_1x_2+x_1-x_2.$$

Here

$$\nabla\phi(x_1,x_2)=\begin{bmatrix}4x_1+2x_2+1\\ 2x_1+2x_2-1\end{bmatrix}.$$

Considering the initial matrix $A_0=I_2$, we calculate the sequence $\{x_k\}$, $x_k=\left(x_1^k,x_2^k\right)$, through the following equations:

$$\begin{cases}d_k &= -A_k\begin{bmatrix}4x_1^k+2x_2^k+1\\ 2x_1^k+2x_2^k-1\end{bmatrix}\\ \alpha_k &= \operatorname{argmin}_{\alpha\geq 0}\ \phi(x_k+\alpha d_k)\\ \delta_k &= \alpha_k d_k\\ x_{k+1} &= x_k+\delta_k\\ \Phi_k &= \nabla\phi(x_{k+1})-\nabla\phi(x_k)=\begin{bmatrix}4(x_1^{k+1}-x_1^k)+2(x_2^{k+1}-x_1^k)\\ 2(x_1^{k+1}-x_1^k)+2(x_2^{k+1}-x_1^k)\end{bmatrix}, \text{ and}\\ A_{k+1} &= A_k+\frac{\delta_k\delta_k^t}{\delta_k^t\Phi_k}-\frac{A_k\Phi_k\Phi_k^tA_k}{\Phi_k^tA_k\Phi_k}.\end{cases}$$

The initial iteration $(k=0)$

$$\begin{cases}x_0 &= (0,0)\\ A_0 &= \begin{bmatrix}1&0\\0&1\end{bmatrix}\\ \|\nabla\phi(x_0)\| &= \|(1,-1)^t\|=\sqrt{2}\neq 0\\ d_0 &= -\begin{bmatrix}-1\\1\end{bmatrix}\\ \alpha_0 &= 1\\ x_1 &= (-1,1)\\ \delta_1 &= (-1,1)\\ \Phi_1 &= (-2,0)\\ A_1 &= \begin{bmatrix}\frac{1}{2}&-\frac{1}{2}\\-\frac{1}{2}&\frac{3}{2}\end{bmatrix}.\end{cases}$$

The second iteration $(k = 1)$

$$\begin{cases} x_1 & = & (-1, 1) \\ d_1 & = & (0, 1) \\ \alpha_1 & = & \frac{1}{2} \\ x_2 & = & (-1, \frac{3}{2}) \\ \nabla\phi(x_2) & = & \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \end{cases}$$

Hence $\bar{x} = (-1, \frac{3}{2})$ is a nondominated solution of the considered problem. It is important to note that the method converged at the second iteration since the objective function is a quadratic fuzzy function.

6 Conclusion

In this paper, a quasi-Newton method with rank-two modification has been derived to find a non-dominated solution of an unconstrained fuzzy optimization problem. In the optimality concept, the fuzzy-max ordering of a pair of fuzzy numbers has been used. The gH-differentiability of fuzzy functions have been employed to find the non-dominated solution point. An algorithmic implementation and the convergence analysis of the proposed technique has also been presented. The technique is found to have superlinear convergence rate. A numerical example has been explored to illustrate the proposed technique.

It is to note that unlike Newton method for fuzzy optimization problems [8,23], the proposed method is derived without using the inverse of the concerned Hessian matrix. Instead, a sequence of positive definite inverse Hessian approximation $\{A_k\}$ is used which satisfies the *quasi-Newton equation* $A_{k+1}\Phi_k = \delta_k$. Thus the derived method sidestepped the inherent computational difficulty to compute inverse of the concerned Hessian matrix. In this way the proposed method is made more efficient than the existing Newton method [8,23].

The method can be easily observed as much similar to the classical DFP method [28] for conventional optimization problem. In the analogous way of the presented technique, it can be further extended to a BFGS-like method [23] for fuzzy optimization. Instead of the rank-two modification in the approximation of the inverse of the Hessian matrix, a rank-one modification can also be done. It is also to be observed that in Algorithm 1, we make use the exact line search technique along the descent direction $d_k = -A_k\nabla\phi(x_k)$. However, an inexact line search technique [28] could have also been used. A future research on this rank-one modification and inexact line search technique for fuzzy optimization problem can be performed and implemented on Newton and quasi-Newton method.

Acknowledgement. The author is truly thankful to the anonymous reviewers and editors for their valuable comments and suggestions. The financial support through Early Career Research Award (ECR/2015/000467), Science and Engineering Research Board, Government of India is also gratefully acknowledged.

References

1. Andrei, N.: Hybrid conjugate gradient algorithm for unconstrained optimization. J. Optim. Theor. Appl. **141**, 249–264 (2009)
2. Anastassiou, G.A.: Fuzzy Mathematics: Approximation Theory, vol. 251. Springer, Heidelberg (2010)
3. Banks, H.T., Jacobs, M.Q.: A differential calculus for multifunctions. J. Math. Anal. Appl. **29**, 246–272 (1970)
4. Bazaraa, M.S., Sherali, H.D., Shetty, C.M.: Nonlinear Programming: Theory and Algorithms. Wiley-interscience, 3rd edn. Wiley, New York (2006)
5. Bede, B., Stefanini, L.: Generalized differentiability of fuzzy-valued functions. Fuzzy Sets Syst. **230**, 119–141 (2013)
6. Bellman, R.E., Zadeh, L.A.: Decision making in a fuzzy environment. Manag. Sci. **17**, B141–B164 (1970)
7. Cadenas, J.M., Verdegay, J.L.: Towards a new strategy for solving fuzzy optimization problems. Fuzzy Optim. Decis. Mak. **8**, 231–244 (2009)
8. Chalco-Cano, Y., Silva, G.N., Rufián-Lizana, A.: On the Newton method for solving fuzzy optimization problems. Fuzzy Sets Syst. **272**, 60–69 (2015)
9. Chakraborty, D., Ghosh, D.: Analytical fuzzy plane geometry II. Fuzzy Sets Syst. **243**, 84–109 (2014)
10. Dubois, D., Prade, H.: Towards fuzzy differential calculus part 3: differentiation. Fuzzy Sets Syst. **8**(3), 225–233 (1982)
11. Ghosh, D.: Newton method to obtain efficient solutions of the optimization problems with interval-valued objective functions. J. Appl. Math. Comput. **53**, 709–731 (2016). (Accepted Manusctipt)
12. Ghosh, D.: A Newton method for capturing efficient solutions of interval optimization problems. Opsearch **53**, 648–665 (2016)
13. Ghosh, D.: A quasi-newton method with rank-two update to solve interval optimization problems. Int. J. Appl. Comput. Math. (2016). doi:10.1007/s40819-016-0202-7 (Accepted Manusctipt)
14. Ghosh, D., Chakraborty, D.: Analytical fuzzy plane geometry I. Fuzzy Sets Syst. **209**, 66–83 (2012)
15. Ghosh, D., Chakraborty, D.: Analytical fuzzy plane geometry III. Fuzzy Sets Syst. **283**, 83–107 (2016)
16. Ghosh, D., Chakraborty, D.: A method to capture the entire fuzzy non-dominated set of a fuzzy multi-criteria optimization problem. Fuzzy Sets Syst. **272**, 1–29 (2015)
17. Ghosh, D., Chakraborty, D.: A new method to obtain fuzzy Pareto set of fuzzy multi-criteria optimization problems. Int. J. Intel. Fuzzy Syst. **26**, 1223–1234 (2014)
18. Ghosh, D., Chakraborty, D.: Quadratic interpolation technique to minimize univariable fuzzy functions. Int. J. Appl. Comput. Math. (2016). doi:10.1007/s40819-015-0123-x (Accepted Manuscript)
19. Ghosh, D., Chakraborty, D.: A method to obtain complete fuzzy non-dominated set of fuzzy multi-criteria optimization problems with fuzzy parameters. In: Proceedings of IEEE International Conference on Fuzzy Systems, FUZZ IEEE, IEEE Xplore, pp. 1–8 (2013)
20. Lai, Y.-J., Hwang, C.-L.: Fuzzy Mathematical Programming: Methods and Applications. Lecture Notes in Economics and Mathematical Systems, vol. 394. Springer, New York (1992)

21. Lai, Y.-J., Hwang, C.-L.: Fuzzy Multiple Objective Decision Making: Methods and Applications. Lecture Notes in Economics and Mathematical Systems, vol. 404. Springer, New York (1994)
22. Lodwick, W.A., Kacprzyk, J.: Fuzzy Optimization: Recent Advances and Applications, vol. 254. Physica-Verlag, New York (2010)
23. Pirzada, U.M., Pathak, V.D.: Newton method for solving the multi-variable fuzzy optimization problem. J. Optim. Theor. Appl. **156**, 867–881 (2013)
24. Puri, M.L., Ralescu, D.A.: Differentials of fuzzy functions. J. Math. Anal. Appl. **91**(2), 552–558 (1983)
25. Lee-Kwang, H., Lee, J.-H.: A method for ranking fuzzy numbers and its application to decision-making. IEEE Trans. Fuzzy Syst. **7**(6), 677–685 (1999)
26. Luhandjula, M.K.: Fuzzy optimization: milestones and perspectives. Fuzzy Sets Syst. **274**, 4–11 (2015)
27. Ramík, J., Rimanek, J.: Inequality relation between fuzzy numbers and its use in fuzzy optimization. Fuzzy Sets Syst. **16**(2), 123–138 (1985)
28. Nocedal, J., Wright, S.J.: Numerical Optimization, 2nd edn. Springer, New York (2006)
29. Słowínski, R.: Fuzzy Sets in Decision Analysis, Operations Research and Statistics. Kluwer, Boston (1998)
30. Stefanini, L.: A generalization of Hukuhara difference and division for interval and fuzzy arithmetic. Fuzzy Sets Syst. **161**, 1564–1584 (2010)
31. Wang, X., Kerre, E.E.: Reasonable properties for the ordering of fuzzy quantities. Fuzzy Sets Syst. **118**, 375–385 (2001)
32. Wang, Z.-X., Liu, Y.-L., Fan, Z.-P., Feng, B.: Ranking L-R fuzzy number based on deviation degree. Inform. Sci. **179**, 2070–2077 (2009)
33. Wu, H.-C.: The optimality conditions for optimization problems with fuzzy-valued objective functions. Optimization **57**, 473–489 (2008)

Bifurcation Analysis of a Delayed Modified Holling-Tanner Predator-Prey Model with Refuge

Charu Arora(✉) and Vivek Kumar

Delhi Technological University, Delhi 110042, India
charuarora099@gmail.com, vivekkumar.ag@gmail.com

Abstract. This paper deals with a delayed modified Holling-Tanner predator-prey model with refuge. The proposed model highlights the impact of delay and refuge on the dynamics of the system wherein analysis of the model in terms of local stability is performed. Both theoretical and experimental works point out that delay and refuge play an important role in the stability of the model and also it has been observed that due to delay, bifurcation occurred which results in considering delay as a bifurcation parameter. For some specific values of delay, Hopf bifurcation is investigated for the proposed model and direction of Hopf bifurcation with the stability of bifurcated periodic solutions by using normal form theory and central manifold reduction is also included in the domain of this study. At the end, few numerical simulations based on hypothetical set of parameters for the support of theoretical formulation are also carried out.

Keywords: Predator-prey model · Time delay · Hopf bifurcation · Stability · Periodic solution

1 Introduction

Prey-predator dynamics is gaining popularity among applied mathematicians and ecologists. Many mathematical models for the dynamics of prey predator relation have been proposed. To describe dynamics more appropriately the researchers introduced delay and many delayed models are also available in recent literature.

The Leslie predator prey model with Holling type II functional response takes the form;

$$\begin{cases} \frac{dx}{dt} = rx(1 - \frac{x}{k}) - \frac{mx}{A+x}y, \\ \frac{dy}{dt} = y[s(1 - h\frac{y}{x})], \end{cases} \tag{1}$$

where x and y are prey and predator densities. For more detail of model (1), we refer the study of Leslie and Gower [1]. Model (1) is further modified by Lu and Liu [2], their model takes the following form;

© Springer Nature Singapore Pte Ltd. 2017
D. Giri et al. (Eds.): ICMC 2017, CCIS 655, pp. 246–254, 2017.
DOI: 10.1007/978-981-10-4642-1_21

$$\begin{cases} \frac{dx}{dt} = rx(1-\frac{x}{k}) - \frac{\alpha xy}{a+bx+cy}y, \\ \frac{dy}{dt} = y[s(1-h\frac{y(t-\tau)}{x(t-\tau)})], \end{cases} \tag{2}$$

where x and y are again prey and predator densities. The term $\frac{\alpha xy}{a+bx+cy}$ is called Beddington-DeAngelis functional response. In non-dimensioning form, the model (2) by defining $\widetilde{t} = rt$, $\widetilde{x} = \frac{x(t)}{k}$, $\widetilde{y} = \frac{\alpha y(t)}{rk}$ and dropping the tildes, is written as,

$$\begin{cases} \frac{dx}{dt} = x(1-x) - \frac{xy}{a_1+bx+c_1y}, \\ \frac{dy}{dt} = y[(\delta - \beta\frac{y(t-\tau)}{x(t-\tau)})], \end{cases} \tag{3}$$

where $\delta = \frac{s}{r}$, $\beta = \frac{sh}{\alpha}$, $a_1 = \frac{a}{k}$, $c_1 = \frac{cr}{\alpha}$, $\widetilde{\tau} = r\tau$. Permanence of (3) has been studied in paper [2]. They [2] also studied the local and global stability of the equilibrium. Model (3) is further studied by J.F. Zhang [3] and the stability of positive equilibrium and Hopf bifurcation are done. In [3], investigation of the direction of Hopf bifurcation and stability of bifurcated periodic solutions were observed.

In this study, we have reconsidered the model (3) for further modification. The main contribution of this study is to incorporate prey refuge in the model (3). Our model becomes;

$$\begin{cases} \frac{dx}{dt} = x(1-x) - \frac{(1-m)xy}{a_1+bx+c_1y}, \\ \frac{dy}{dt} = y[(\delta - \beta\frac{y(t-\tau)}{x(t-\tau)} + (1-m)x)], \end{cases} \tag{4}$$

where x and y are prey and predator densities respectively. Here the constant m denotes the prey refuge. It means $(1-m)x$ is the amount available for predation to predator. The range for the parameter is fixed as $0 < m < 1$. To understand the feasibility of refuge, we take the case of a forest which is considered as an ecosystem where deer and lion are the prey and predator species respectively, but if deer has a habitat of particular kind where lions cannot enter and with such habitat complexity, lions cannot predate deers and eventually it gives birth to the term prey refuge. Initial data for (4) is considered as $x(0) > 0, y(0) > 0$. Rest of the parameters $a_1, b, c_1, \delta, \beta$ are positive constants and are similar to (3). Primarily, the effect of refuge (here m) on the model (3) is studied in this paper.

In recent studies, properties of periodic solutions (Hopf bifurcation) is observed [3,4]. In this paper, stability of positive equilibrium is studied and also the process of Hopf bifurcation has been focussed(in brief).

Rest of the paper is structured as follows. In Sect. 2, stability of positive equilibrium and the existence of Hopf bifurcation is studied and in Sect. 3, the direction of Hopf bifurcation with the stability of bifurcated periodic solutions is observed. Numerical simulations have been performed in Sect. 4 along with a brief discussion in Sect. 5 which concludes the paper.

2 Stability of Positive Equilibrium and Hopf Bifurcation

It is easy to calculate that the system (4) has unique positive equilibrium say $E^*(x^*, y^*)$ where x^* is a root of the equation $Ax^3 + Bx^2 + Cx + D = 0$, where

$A = \frac{-c_1(1-m)}{\beta}$, $B = \frac{c_1(1-m)-b\beta-c_1\delta-(1-m)+m(1-m)}{\beta}$, $C = \frac{b\beta-a_1\beta-\delta+m\delta+c_1\delta}{\beta}$, $D = a_1$, and

$$y^* = \frac{\delta x^* + (1-m){x^*}^2}{\beta}$$

The first attempt is to investigate the dynamics of the model without delay. The Jacobian matrix has the form,

$$J = \begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix},$$

where, $\alpha_{11} = 1 - 2x^* - \frac{(1-m)y^*(a_1+c_1y^*)}{(a_1+bx^*+c_1y^*)^2}$, $\alpha_{12} = \frac{-x^*(1-m)(a_1+bx^*)}{(a_1+bx^*+c_1y^*)^2}$, $\alpha_{21} = \frac{\beta{y^*}^2}{{x^*}^2} + (1-m)y^*$, $\alpha_{22} = \delta - \frac{2\beta y^*}{\beta} + (1-m)x^*$ and the characteristic equation is given by $\lambda^2 - (Tr.J)\lambda + (Det.J) = 0$ or $\lambda^2 - (\alpha_{11}+\alpha_{22})\lambda + (\alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21}) = 0$, Using Routh-Hurwitz criteria for determining the stability of the system under consideration. We have,

$$Det.J = \frac{((1-2x^*)(a_1+bx^*+c_1y^*)^2 - y^*(1-m)(a_1+c_1y^*))(\delta x^* + (1-m){x^*}^2 - 2\beta y^*)}{x^*(a_1+bx^*+c_1y^*)^2}$$

For $Det.J > 0$, either $((1-2x^*)(a_1+bx^*+c_1y^*)^2 - y^*(1-m)(a_1+c_1y^*)) > 0$ and $(\delta x^* + (1-m){x^*}^2 - 2\beta y^*) > 0$ or $((1-2x^*)(a_1+bx^*+c_1y^*)^2 - y^*(1-m)(a_1+c_1y^*)) < 0$ and $(\delta x^* + (1-m){x^*}^2 - 2\beta y^*) < 0$.

Also condition $\alpha_{11} + \alpha_{22} < 0$ is required for the asymptotically stability of the model without delay. Hence, we can state the following theorem.

Theorem 1. *Equilibrium $E^*(x^*, y^*)$ of system (4) is locally asymptotically stable if the following conditions are satisfied:*
H(1) $\alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21} > 0$
H(2) $\alpha_{11} + \alpha_{22} < 0$

Now we investigate condition(s) for hopf bifurcation. The procedure is quite similar to J.F. Zhang [3]. The linearised version of the model is written as,

$$\begin{cases} u_1' = \alpha_{11}u_1(t) + \alpha_{12}u_2 t, \\ u_2' = \alpha_{21}u_1(t-\tau) + \alpha_{22}u_2(t-\tau). \end{cases} \tag{5}$$

The characteristic equation can be written as

$$\lambda^2 - (\alpha_{11})\lambda + (-\alpha_{22}\lambda + \alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21})\exp^{-\lambda\tau} = 0 \tag{6}$$

Now putting $\lambda = i\omega$ in Eq. 6, we get,

$$\begin{aligned} &-\omega^2 + (\alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21})\cos\omega\tau - \alpha_{22}\omega\sin\omega\tau + i - \alpha_{22}\omega\cos\omega\tau \\ &\quad - (\alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21})\sin\omega\tau - \alpha_{11}\omega = 0 + 0i \end{aligned}$$

Separating the real and imaginary parts, we get

$$-\omega^2 + (\alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21})\cos\omega\tau - \alpha_{22}\omega\sin\omega\tau = 0 \tag{7}$$

$$-\alpha_{22}\omega\cos\omega\tau-(\alpha_{11}\alpha_{22}-\alpha_{12}\alpha_{21})\sin\omega\tau-\alpha_{11}\omega=0 \tag{8}$$

On squaring and adding Eqs. 7 and 8, we get

$$\omega^4+(\alpha_{11}^2-\alpha_{22}^2)\omega^2-(\alpha_{11}\alpha_{22}-\alpha_{12}\alpha_{21})^2=0, \tag{9}$$

If we put $z=\omega^2$, we get,

$$z^2+(\alpha_{11}^2-\alpha_{22}^2)z-(\alpha_{11}\alpha_{22}-\alpha_{12}\alpha_{21})^2=0, \tag{10}$$

which is a quadratic equation, hence the roots are $z=\frac{-(\alpha_{11}^2-\alpha_{22}^2)\pm\sqrt{(\alpha_{11}^2-\alpha_{22}^2)^2+4(\alpha_{11}\alpha_{22}-\alpha_{12}\alpha_{21})^2}}{2}$. Taking only the positive root, we get, $z=\frac{-(\alpha_{11}^2-\alpha_{22}^2)+\sqrt{(\alpha_{11}^2-\alpha_{22}^2)^2+4(\alpha_{11}\alpha_{22}-\alpha_{12}\alpha_{21})^2}}{2}$.
Therefore,

$$\omega=\pm\sqrt{\frac{-(\alpha_{11}^2-\alpha_{22}^2)\pm\sqrt{(\alpha_{11}^2-\alpha_{22}^2)^2+4(\alpha_{11}\alpha_{22}-\alpha_{12}\alpha_{21})^2}}{2}}. \tag{11}$$

Again, taking only the positive roots and denoting the positive root by $\omega_{pos.}$, we get

$$\omega_{pos}=\sqrt{\frac{-(\alpha_{11}^2-\alpha_{22}^2)\pm\sqrt{(\alpha_{11}^2-\alpha_{22}^2)^2+4(\alpha_{11}\alpha_{22}-\alpha_{12}\alpha_{21})^2}}{2}}.$$

Now solving Eqs. 7 and 8, we get the values of τ say critical value(s) of the form

$$\tau_j=\frac{1}{\omega_{pos}}\cos^{-1}\frac{\omega_{pos}^2(\alpha_{11}\alpha_{22}-\alpha_{12}\alpha_{21})-\alpha_{11}\alpha_{22}\omega_{pos}}{(\alpha_{11}\alpha_{22}-\alpha_{12}\alpha_{21})^2+\alpha_{22}^2\omega_{pos}^2}+\frac{2\pi j}{\omega_{pos}},\ \text{where j}=0,1,2....$$

We denote one of the set of critical value of τ as τ_{cr} and hence, we have $\tau_{cr}=\frac{1}{\omega_{pos}}\cos^{-1}\frac{\omega_{pos}^2(\alpha_{11}\alpha_{22}-\alpha_{12}\alpha_{21})-\alpha_{11}\alpha_{22}\omega_{pos}}{(\alpha_{11}\alpha_{22}-\alpha_{12}\alpha_{21})^2+\alpha_{22}^2\omega_{pos}^2}$. Now we can state the following lemma by above discussion and also by Rouche's theorem, similar to J.F. Zhang [3].

Lemma 1. *Assume that the positive equilibrium point of the system of system (4) without delay is locally asymptotically stable. Then at*

$$\tau_j=\frac{1}{\omega_{pos}}\cos^{-1}\frac{\omega_{pos}^2(\alpha_{11}\alpha_{22}-\alpha_{12}\alpha_{21})-\alpha_{11}\alpha_{22}\omega_{pos}}{(\alpha_{11}\alpha_{22}-\alpha_{12}\alpha_{21})^2+\alpha_{22}^2\omega_{pos}^2}+\frac{2\pi j}{\omega_{pos}},(j=0,1,2...),$$

system (9) has a pair of conjugate purely imaginary roots $\pm i\omega_{pos}$, where $\omega_{pos}=\sqrt{\frac{-(\alpha_{11}^2-\alpha_{22}^2)\pm\sqrt{(\alpha_{11}^2-\alpha_{22}^2)^2+4(\alpha_{11}\alpha_{22}-\alpha_{12}\alpha_{21})^2}}{2}}$. Furthermore, we have the following results

(i) If $\tau\in[0,\tau_{cr})$, then all roots of the system (4) have negative real parts.
(ii) If $\tau=\tau_{cr}$, system (4) has a pair of conjugate purely imaginary roots $\pm\omega_{pos}$, and all other roots have negative real parts.

Let $\lambda = v(\tau) + i\omega(\tau)$ be the root of the characteristic Eq. (6) with the condition that when $\tau = \tau_{cr}$,

$$\begin{cases} v(\tau_{cr}) = 0, \\ \omega(\tau_{cr}) = \omega_{pos}. \end{cases}$$

Now, differentiating (6) and on further simplification, we get,

$$\left(\frac{d\lambda}{d\tau}\right)^{-1} = \frac{(2\lambda - \alpha_{11} - \alpha_{22}e^{-\lambda\tau})}{(\alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21} - \alpha_{22}\lambda)} - \frac{\tau}{\lambda}.$$

Therefore,

$$[Re\left(\frac{d\lambda}{d\tau}\right)^{-1}]_{i\omega_{pos}} = \frac{\omega_{pos}^2}{(\alpha_{22}^2\omega_{pos}^4 + (\alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21})^2\omega_{pos}^2)}\sqrt{(\alpha_{11}^2 - \alpha_{22}^2)^2 + 4(\alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21})^2}$$

now $(\alpha_{22}^2\omega_{pos}^4 + (\alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21})^2\omega_{pos}^2) > 0$ thus, we have $[Re\left(\frac{d\lambda}{d\tau}^{-1}\right)]_{i\omega_{pos}} > 0$, therefore the transversality condition is proved. Now we can state the bifurcation theorem for the proposed system (4);

Theorem 2. *Suppose the condition of Theorem 1 holds.*

(i) If $\tau \in [0, \tau_{cr})$, the positive equilibrium $E^*(x^*, y^*)$ of system (4) is asymptotically stable.
(ii) If $\tau > \tau_{cr}$, the positive equilibrium $E^*(x^*, y^*)$ of system (4) is unstable.
(iii) System (4) observe Hopf bifurcation at the positive equilibrium $E^*(x^*, y^*)$ when $\tau = \tau_j$, where,

$$\tau_j = \frac{1}{\omega_{pos}} \cos^{-1} \frac{\omega_{pos}^2(\alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21}) - \alpha_{11}\alpha_{22}\omega_{pos}}{(\alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21})^2 + \alpha_{22}^2\omega_{pos}^2} + \frac{2\pi j}{\omega_{pos}}, j = 0, 1, 2....$$

3 Direction of Hopf Bifurcation

In this section, we shall study the direction of Hopf bifurcation and stability of bifurcated periodic solutions by the application of normal form theory and central manifold reduction technique introduced originally long back by Hassard et al. [5]. Such theories have been studied in recent literature. We have seen in Theorem 2 that system (4) undergoes Hopf bifurcation for some specified values of τ, these values are denoted by τ_j. As a matter of generalization, we denote any one such value by $\overline{\tau}$. Thus, at $\overline{\tau}$, the characteristic Eq. (6) will have a pair of imaginary roots $\pm i\omega_{pos}$. By the procedure explained in Hassard et al. [5], we proceed further. For the reduction of system (4) a system of functional differential equation is used. Furthermore, we denote delay τ as $\tau = \overline{\tau} + \mu$, where μ is an element of the set of real numbers. Further $\mu = 0$ is the value of Hopf bifurcation of system (4). We rescale the time by $\rightarrow \frac{t}{\tau}$. System (4) takes the following form,

$$u'(t) = L_\mu u(t) + F(u_t, \mu), \tag{12}$$

in the Banach space $\mathbb{C}([-1,0],\mathbb{R}^2)$, where $\mathrm{u}=\begin{pmatrix}u_1\\u_2\end{pmatrix}\in\mathbb{C}$ and $L_\mu:\mathbb{C}\to\mathbb{R}$, $F:\mathbb{R}\times\mathbb{C}\to\mathbb{R}$ has been obtained in J.F. Zhang [3]. We can proceed exactly in a similar manner as in [3] and can obtain the following values (with usual symbols as obtained in [3]):

$$C_1(0)=\frac{i}{2\omega_{pos}\overline{\tau}}\left(g_{20}g_{11}-2|\,g_{11}\,|^2-\frac{|\,g_{02}\,|^2}{3}\right)+\frac{g_{21}}{2},$$

$$\mu_2=\frac{Re\{C_1(0)\}}{Re\{\lambda'(\overline{\tau})\}}\text{ and }\beta_2=2Re\{C_1(0)\}.$$

Now we can state the following theorem:

Theorem 3.

(i) μ_2 *determines the directions of Hopf bifurcation. If* $\mu_2>0(<0)$*, the Hopf bifurcation is supercritical(subcritical);*

(ii) β_2 *determines the stability of bifurcated periodic solutions. If* $\beta_2>0(<0)$*, the bifurcated periodic solutions are stable (unstable).*

4 Numerical Simulations

We have studied the effect of delay and refuge on the modified Holling-Tanner predator-prey model and it is to be remarked that the real parameters are not available. So, theoretical formulations have been verified by taking hypothetical set of parameters as in [3]. However, assumed parameters can demonstrate the theoretical formulation. We consider the following numerical example:

$$\begin{cases}\frac{dx}{dt}=x(1-x)-\frac{0.4xy}{0.01+3x+y},\\ \frac{dy}{dt}=y\left[3.5-2\frac{y(t-\tau)}{x(t-\tau)}+0.4x\right].\end{cases}\tag{13}$$

Clearly $(1-m)=0.4$ and after calculating, it is observed that system (13) has a positive equilibrium $E^*(1.62,0.844)$. By Lemma 1, we calculate $\omega_{pos}=3.8246$ and $\tau_{cr}=0.38$. It is also note that $\alpha_{11}+\alpha_{22}<0$ hence $E^*(1.62,0.844)$ of the system (13) is locally stable in absence of delay term τ. By Theorem 2, it may be concluded that $E^*(1.62,0.844)$ is asymptotically stable if $\tau\in[0,0.38)$ and unstable if $\tau>0.38$. In this case, Hopf bifurcation occurs at $\tau=0.38$. Solution curves of (13) for $\tau=\tau_{cr}=0.38,0.43(>\tau_{cr})$ and $0.21(<\tau_{cr})$ respectively are given in Figs. 1, 2 and 3. It is observed that the graphs so developed are consistent with the theoretical formulation. System undergoes Hopf bifurcation for a specific value of delay at the positive equilibrium which suggests that delay plays a major role. Also, effect of refuge is studied which shows that it has an important role and system changes when its effect is more (Fig. 4).

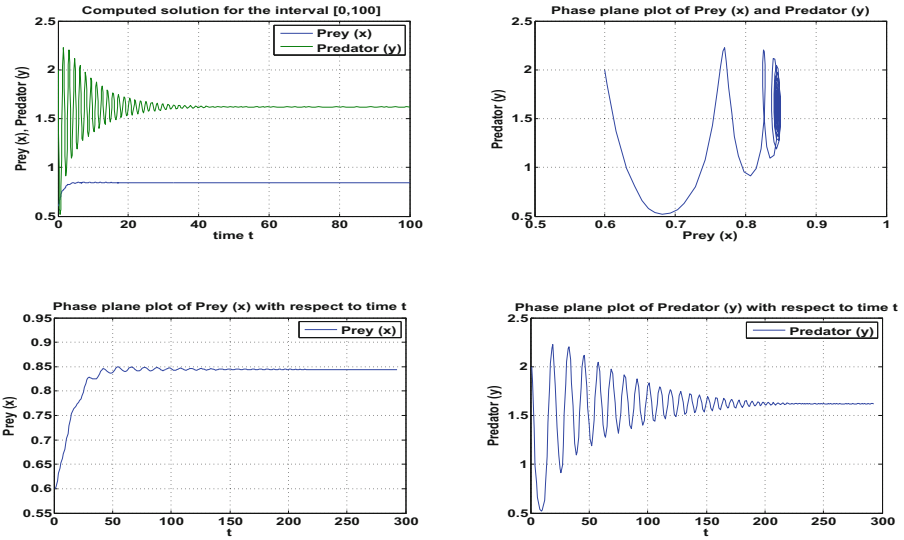


Fig. 1. Solution curves of System (4) with $(1 - m) = 0.4$ and $\tau = 0.38$ computed over the interval [0,100]

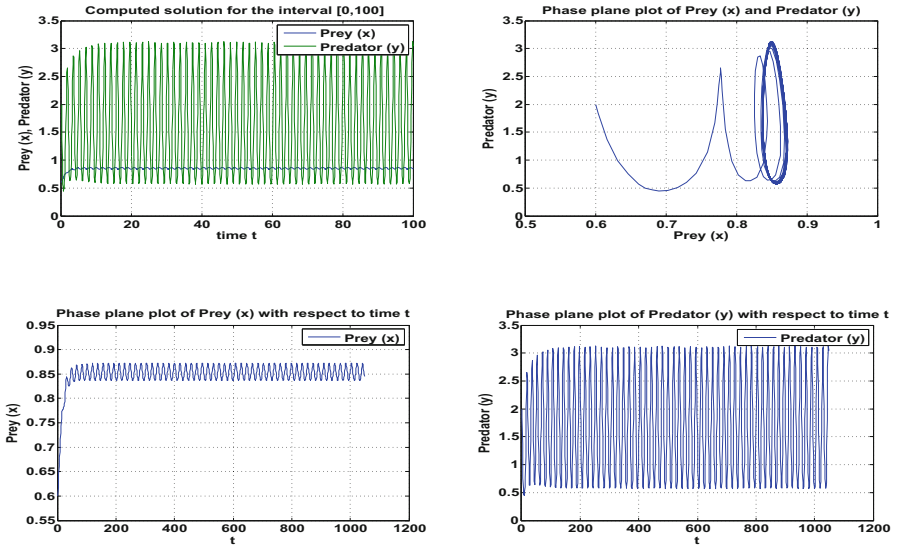


Fig. 2. Solution curves of System (4) with $(1 - m) = 0.4$ and $\tau = 0.43 > \tau_{cr} = 0.38$ computed over the interval [0,100]

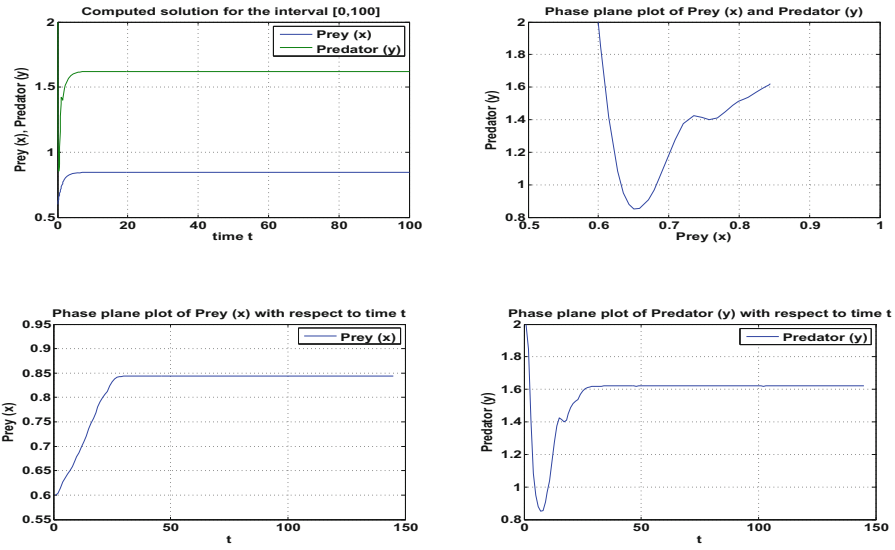


Fig. 3. Solution curves of System (4) with $(1-m) = 0.4$ and $\tau = 0.21 < \tau_{cr} = 0.38$ computed over the interval [0,100]

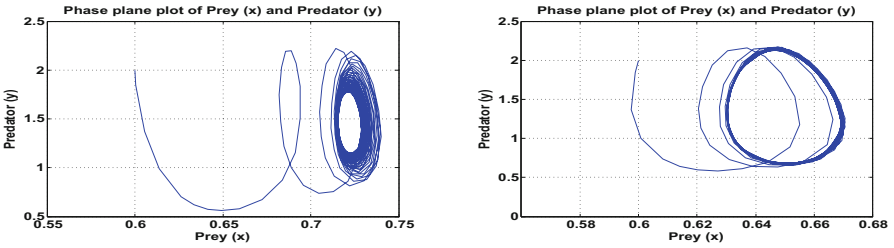


Fig. 4. Solution curves of System (4) with two different values of $(1-m)$ i.e. $(1-m) = 0.7$ and $(1-m) = 0.9$ with $\tau = 0.38$ computed over the interval [0,100]

5 Discussion

In this study, we reviewed a delayed prey-predator model with modified Holling-Tanner functional response. As a matter of fact, model (3) can be interpreted as a particular case of our proposed model and in model (4), mainly the role of refuge is studied. Limitations of this study are non availability of real parameters. Although numerical example (13), so considered is consistent with our theoretical formulation. Furthermore, for the different values of m(refuge), system has been analysed.

References

1. Leslie, P.H., Gower, J.C.: The properties of a stochastic model for the predator-prey type of interaction between two species. Biometrika **47**, 219–234 (1960)
2. Lu, Z., Liu, X.: Analysis of a predator prey model with modified Holling-Tanner functional response and time delay. Nonlinear Anal. Real World Appl. **9**, 641–650 (2008). http://www.dx.doi.org/10.1016/j.nonrwa.2006.12.016
3. Zhang, J.-F.: Bifurcation analysis of a modified Holling-Tanner predator-prey model with time delay. Appl. Math. Model. **36**, 1219–1231 (2012). http://www.dx.doi.org/10.1016/j.apm.2011.07.071
4. Kant, S., Kumar, V.: Delayed prey-predator system with habitat complexity and refuge. In: Proceedings of International Conference on Mathematical Sciences (ICMS 2014), pp. 584–591 (2014). http://www.dx.doi.org/10.13140/2.1.3493.8884
5. Hassard, B.D., Kazarinoff, N.D., Wan, Y.H.: Theory and Applications of Hopf Bifurcation. Cambridge University Press, Cambridge (1981)

A Higher Order Implicit Method for Numerical Solution of Singular Initial Value Problems

M. Kamrul Hasan[1(✉)], M. Suzan Ahamed[1], B.M. Ikramul Haque[2], M.S. Alam[1], and M. Bellal Hossain[1]

[1] Department of Mathematics, Rajshahi University of Engineering and Technology, Rajshahi 6204, Bangladesh
mkh2502@yahoo.com

[2] Department of Mathematics, Khulna University of Engineering and Technology, Khulna 9203, Bangladesh

Abstract. Recently a lower order implicit method has been presented for solving singular initial value problem. In this article a higher order implicit method has been developed to solve first or higher order problems having an initial singular point. This method is more suitable than second, third and two-stage fourth order implicit Runge-Kutta methods for first order problems. The method also provides significantly better results than the existing lower order implicit method for second order problems.

Keywords: Singular initial value problems · Implicit Runge-Kutta method · Lane–Emden type equation · Emden-Fowler type equation

Mathematical Subject Classification (2010): 35F10 · 35F25 · 35G10 · 35G25

1 Introduction

In the recent years, the studies of singular initial value problems have attracted the attention of many mathematicians and physicists. Many problems in ecology, mathematical physics and astrophysics can be modeled by second order singular initial value problems. Sometimes, first order singular initial value problems are also used *e.g.,* a leading-edge model in the computation of the run-out length of dry-flowing avalanches. Some well known problems such as Lane–Emden type and Emden-Fowler type differential equations are expressed by second order singular initial value problems. These equations have been used to explain various phenomena such as the theory of stellar structure, the thermal behavior of a spherical cloud of gas, isothermal gas spheres and the theory of thermionic currents [1–3].

Some analytical techniques were presented to solve Lane-Emden equations. Most of them were developed based on power series or perturbation techniques. Wazwaz [4–7] has presented series and exact solution to Lane-Emden

© Springer Nature Singapore Pte Ltd. 2017
D. Giri et al. (Eds.): ICMC 2017, CCIS 655, pp. 255–264, 2017.
DOI: 10.1007/978-981-10-4642-1_22

and Emden-Fowler type problems based on Adomain decomposition and modified Adomain decomposition methods. Hasan and Zhu [8,9] have solved such a singular initial value problem by the Taylor series and modified Adomain decomposition methods. Gupta and Sharma [10] have also used the Taylor series method to solve Lane–Emden and Emden–Eowler equations. Demir and Sungu [11] have presented approximate and analytic solutions of Emden-Fowler equation and Mukherjee *et al.* [12] the Lane–Emden equation by the same method *i.e.,* differential transform method. However the determination of solutions by these methods is laborious.

A few classical numerical methods have been used for numerical solution of first and second order singular initial value problem. Koch *et al.* [13,14] applied implicit Euler method (backward) to evaluate the approximate solutions of first order and second order singular initial value problem and finally used an acceleration technique known as the Iterated Defect Correction (IDeC) to improve the approximations. Benko *et al.* [15] evaluated the approximate solution of the second order singular initial value problems by implicit Euler method. The second order implicit Runge-Kutta (RK2) and third order implicit Runge-Kutta (RK3) methods are higher order solvers than the implicit Euler method for solving singular initial value problems. Lakestani and Saray [16] solved Emden-Fowler type equations numerically using Legendre scaling functions. This method consists of expanding the required approximate solution as the elements of Legendre scaling functions. Using the operational matrix of integration, the problem will be reduced to a set of algebraic equations. But utilization of this method is cumbersome.

Recently, Hasan *et al.* [17–19] derived a lower order implicit method for solving first and second order singular initial value problems, which give more accurate solution than the implicit Euler method as well as second order implicit Runge-Kutta (RK2) methods for singular initial value problems. In this article, a higher order implicit formula is presented for solving first order singular initial value problems. The method is extended for second order singular initial value problems. For second order problems the method provides significantly better results than the existing lower order method.

2 Derivation of the Present Method

First we derive the present method for solving first order singular initial value problems and then the method is extended for second order singular initial value problems.

2.1 For First Order Singular Initial Value Problems

Earlier Huq *et al.* [20] derived a formula for evaluating definite integral having an initial singular point at $x = x_0$ in the form as

$$\int_{x_0}^{x_0+3h} f(x)\,dx = \frac{3h}{4}\,[3\,f(x_0+h) + f(x_0+3h)] \tag{1}$$

Based on formula (1), Hasan *et al.* [18] derived an implicit method for solving first order singular initial value problems

$$y'(x) = f(x,\ y), \quad y(x_0) = y_0 \tag{2}$$

having the initial singular point at $x = x_0$ in the form as

$$y_{i+1} = y_i + \frac{h}{4}\ [3f(x_i + h/3,\ (y_i + (y_{i+1} - y_i)/3)) + f(x_{i+1},\ y_{i+1})] \tag{3}$$

where, $x_{i+1} = x_i + h; \quad i = 0,\ 1,\ 2, ...$

Recently, Hasan et al. [21] derived a higher order integral formula for solving singular integral having an initial singular point at $x = x_0$ in the form as

$$\int_{x_0}^{x_0+6h} f(x)\ dx = \frac{3h}{5}\ [4\ f(x_0 + h) + 5\ f(x_0 + 4h) + f(x_0 + 6h)] \tag{4}$$

Based on formula (4), a higher order implicit method has been proposed for solving first order singular initial value problems given in Eq. (2) having the initial singular point at $x = x_0$ in the form as

$$\begin{aligned} y_{i+1} = y_i + \frac{h}{10}\ &[4f(x_i + h/6,\ (y_i + (y_{i+1} - y_i)/6)) \\ &+5f(x_i + 4h/6,\ (y_i + 4(y_{i+1} - y_i)/6)) + f(x_{i+1},\ y_{i+1})] \end{aligned} \tag{5}$$

where, $x_{i+1} = x_i + h; \quad i = 0,\ 1,\ 2, ...$

It is obvious that Eq. (5) is an algebraic equation of unknown y_{i+1} and can be solve by Newton-Raphson method [22].

2.2 For Second Order Singular Initial Value Problems

Let us consider a second order singular initial value problem of the form [5]

$$y'' + \frac{2}{x}y' + f(x,\ y) = g(x), \quad 0 < x \leq 1, \quad y(0) = A,\ \ y'(0) = B \tag{6}$$

where A and B are constants, $f(x, y)$ is a continuous real valued function and $g(x) \in\ C\ [0,\ 1]$. Now Eq. (6) can be transformed into two first order initial value problems, one is non-singular and other is singular as

$$y' = z = f_1(x,\ y) \tag{7}$$

$$z' = -\frac{2}{x}z - f(\ x, y) + g(x) = f_2(x,\ y,\ z) \tag{8}$$

where $y(0) = A$, $z(0) = B$, and $y = y$, $y' = z$.

According to the present method (i.e. Eq. (5)) the approximate solutions of the Eqs. (7) and (8) are

$$y_{i+1} = y_i + \frac{h}{10}\left[4\left(z_i + (z_{i+1} - z_i)/6\right) + 5\left(z_i + 4(z_{i+1} - z_i)/6\right) + z_{i+1}\right)] \tag{9}$$

$$\begin{aligned} z_{i+1} &= z_i + \frac{h}{10}\left[4\left\{-\frac{2}{(x_0 + h/6)}(z_i + (z_{i+1} - z_i)/6)\right.\right. \\ &\qquad \left.\left. - f(x_0 + h/6,\ (y_i + (y_{i+1} - y_i)/6)) + g(x_0 + h/6)\right\}\right. \\ &+ 5\left\{-\frac{2}{(x_0 + 4h/6)}(z_i + 4(z_{i+1} - z_i)/6)\right. \\ &\qquad \left. - f(x_0 + 4h/6,\ (y_i + 4(y_{i+1} - y_i)/6)) + g(x_0 + 4h/6)\right\} \\ &\left. + \left(-\frac{2}{(x_0 + h)} z_{i+1} - f(x_0 + h, y_{i+1}) + g(x_0 + h)\right)\right]; i = 0,\ 1,\ 2, \ldots \end{aligned} \tag{9a}$$

It is obvious that Eqs. (9) and (9a) is a system of equations for two unknown y_{i+1} and z_{i+1} and can be solved by Newton-Raphson method [22].

To compare the present method to other classical methods such as the second, the third and two-stage fourth order implicit Runge-Kutta (RK2, RK3 and RK4) [23] methods are given in Eqs. (10), (11) and (12) respectively.

$$y_{i+1} = y_i + k; \qquad i = 0,\ 1,\ 2, \ldots \tag{10}$$

where, $k = h\, f(x_i + h/2,\ \ y_i + k/2)$

$$y_{i+1} = y_i + (3\ k1 + k2)/4 \tag{11}$$

where, $k1 = h\, f(x_i + h/3,\ \ y_i + k1/3)$

and $k2 = h\, f(x_i + h,\ \ y_i + k1)$

$$y_{i+1} = y_i + (k1 + k2)/2 \tag{12}$$

where, $k1 = h\, f(x_i + (1/2 - \sqrt{3}/6)h,\ \ y_i + k1/4 + (1/4 - \sqrt{3}/6)\ k2)$

and $k2 = h\, f(x_i + (1/2 + \sqrt{3}/6)h,\ \ y_i + (1/4 + \sqrt{3}/6)\ k1 + k2/4)$

3 Convergence and Stability of the Present Method

The convergence order of the present method (i.e., Eq. (5) is $O(h^4)$ *i.e.*, the truncation error is $O(h^5)$. The truncation error of the fourth order implicit Runge-Kutta (RK4) and the third order implicit Runge-Kutta (RK3) methods are $O(h^5)$and $O(h^4)$ respectively.

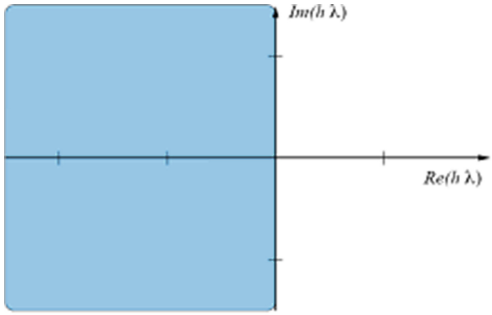


Fig. 1. Stability region of the present method.

To test the stability, consider a scalar test equation

$$y' = \lambda\, y, \quad \lambda \in C, \quad Re(\lambda) < 0 \tag{13}$$

Appling (5) to the test equation with $y' = \lambda\, y$ and $z = \lambda\, h$ yields

$$y_{i+1} = R(z)\, y_i \tag{14}$$

where, $R(z) = (1+z/2)/(1-z/2)$ is the stability function of the present method.

For $\lambda < 0$, then $|R(z)| < 1$ for any $h > 0$. Since z is imaginary, the present method is absolutely stable in the entire negative half of the complex z plane. The region of absolute stability is the set of all complex z where $|R(z)| \leq 1$. A numerical method is said to be A-stable if its stability region contains C^-, the non-positive half-plane $\{z = \lambda\, h \in C : Re(z) < 0\}$. So the present method is A-stable. The stability region of the present method is given in Fig. 1.

4 Examples

The method is illustrated by following singular initial value problems.

Example 1. Consider a first order initial value problem in the form as [18]

$$y'(x) = q\frac{y^r}{x^p}, \;\; 0 < x \leq 1, \; y(0) = 1, \quad 0 < p < 1, \;\; -1 \leq q < 0 \tag{15}$$

The exact solution of Eq. (15) is obtained as

$$\begin{aligned} y &= \left(\frac{x^{1-p}q(-1+r)+(-1+p)}{(-1+p)}\right)^{\frac{1}{1-r}}, \; r \neq 1 \\ &= e^{\frac{qx^{1-p}}{1-p}} \qquad\qquad\qquad\qquad\quad , \;\; r = 1 \end{aligned} \tag{16}$$

The absolute error of the solution of the Eq. (15) obtained by the present (i.e., Eq. (5)), Hasan (i.e., Eq. (3)), the second order implicit Runge-Kutta (RK2)

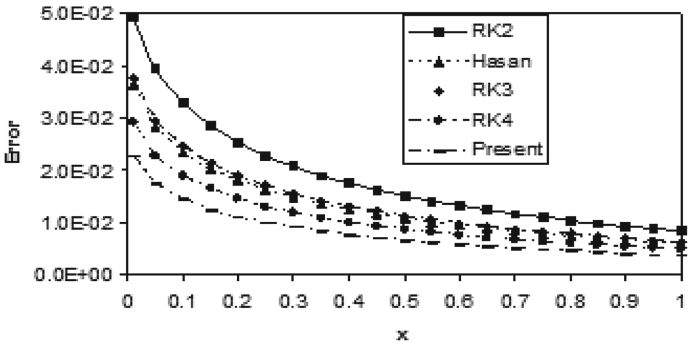


Fig. 2. The absolute error of the Eq. (15) by different method for $p = 1/2,\ q = -1,\ r = 1$ with $h = 0.01$.

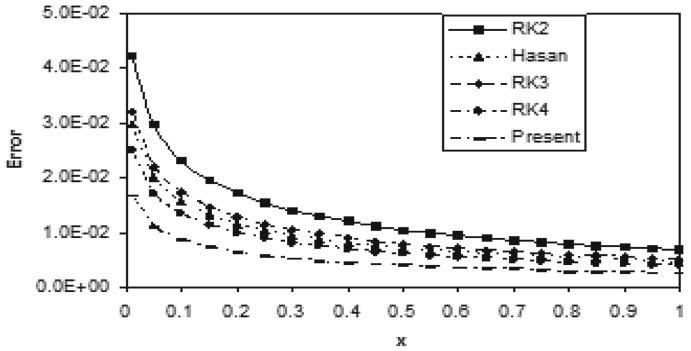


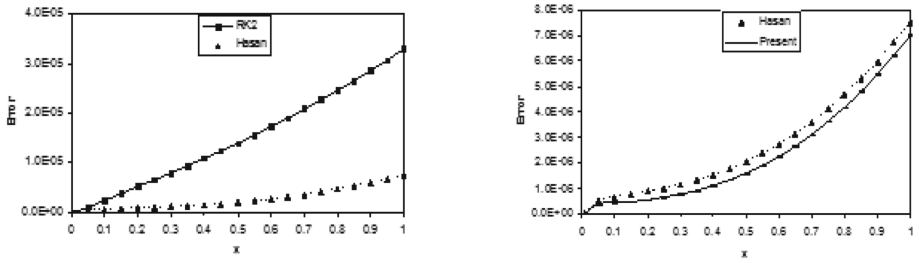
Fig. 3. The absolute error of the Eq. (15) by different method for $p = 1/2,\ q = -1,\ r = 2$ with $h = 0.01$.

(i.e., Eq. (10)), the third order implicit Runge-Kutta (RK3) (i.e., Eq. (11)) and the two-stage fourth order implicit Runge-Kutta (RK4) (i.e., Eq. (12)) methods and are plotted in Figs. 2 and 3 for $p = 1/2,\ q = -1,\ r = 1$ and $p = 1/2,\ q = -1,\ r = 2$ respectively.

Example 2. Consider a second order linear, non-homogeneous Lane-Emden equation [9]

$$y'' + \frac{2}{x}y' + y = 6 + 12x + x^2 + x^3; \quad 0 < x \leq 1, \qquad y(0) = 0,\ \ y'(0) = 0 \quad (17)$$

with the exact solution $y = x^2 + x^3$. The absolute error obtained by RK2 (i.e., Eq. (10)), Hasan (i.e., Eq. (3)) and present (i.e., Eq. (5)) methods are plotted in Fig. 4.



(a) (b)

Fig. 4. The absolute error of the Eq. (17) by RK2 and the Hasan methods in (a) the Hasan and the present methods in (b) with $h = 0.01$.

Example 3. Consider a second order linear, non-homogeneous Emden-Fowler equation [16].

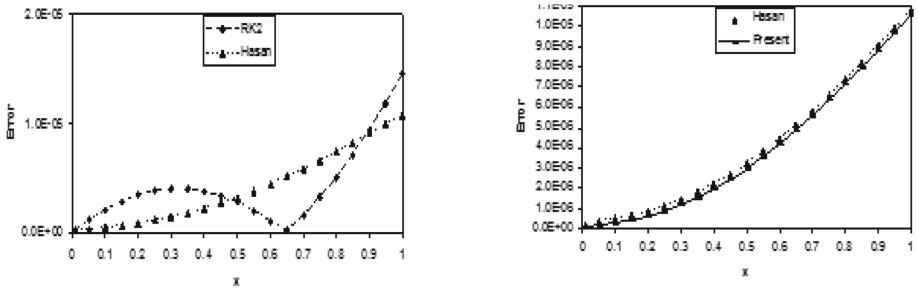
$$y'' + \frac{8}{x}y' + xy = x^5 - x^4 + 44x^2 - 30x; \quad 0 < x \leq 1, \quad y(0) = 0, \; y'(0) = 0, \tag{18}$$

with the exact solution $y = x^4 - x^3$. The absolute error obtained by RK2 (i.e., Eq. (10)), Hasan (i.e., Eq. (3)) and present (i.e., Eq. (5)) methods are plotted in Fig. 5.

Example 4. Consider a second order nonlinear, homogeneous Lane-Emden equation [15]

$$y'' + \frac{2}{x}y' + y^{1.5} = 0; \quad 0 < x \leq 1, \quad y(0) = 1, \; y'(0) = 0, \tag{19}$$

with the approximate exact solution $y = \exp(-x^2/6)$. The results of the error obtained by RK2 (i.e., Eq. (10)), Hasan (i.e., Eq. (3)) and present (i.e., Eq. (5)) methods are plotted in Fig. 6.



(a) (b)

Fig. 5. The absolute error of the Eq. (18) by RK2 and the Hasan methods in (a) the Hasan and the present methods in (b) with $h = 0.01$.

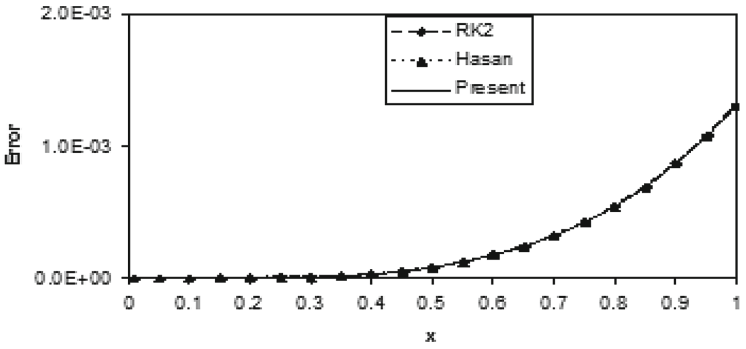


Fig. 6. The absolute error of the Eq. (19) by RK2, the Hasan and the present methods with $h = 0.01$.

5 Results and Discussion

A higher order implicit method has been presented to solve first and second order singular initial value problems. To illustrate the method, the approximate solutions of some first and second order linear and non-linear equations have been compared with their exact solutions. For first order problems, the approximate solution of Eq. (15) has been obtained by the present formula Eq. (5) and the error has been presented in Figs. 2 and 3 together with corresponding errors of RK2, RK3, RK4 and Hasan methods. Figures 2 and 3 show that the error of the present method is smaller than those obtained by Hasan method, RK2, RK3 and RK4 methods.

Then a second order linear non-homogeneous Eq. (17) has been considered. In this case the error has been presented in Fig. 4 together with corresponding errors of RK2 and Hasan methods. This figure indicates that the present method also provides better results than those obtained by RK2 and Hasan methods.

Next, a linear non-homogeneous Eq. (18) has been considered and the errors eventually found through the above three methods have been shown in Fig. 5. It is obvious from the Fig. 5 that the error by the present method is smaller than those obtained by Hasan method and RK2 method. But the errors of the second order implicit Runge-Kutta method increases rapidly after a short interval.

Finally, a non-linear homogeneous Eq. (19) has been considered and the errors of the above three methods are presented in Fig. 6. It indicates that the error of the present method is very close to that obtained by Hasan method as well as second order implicit Runge-Kutta method.

Based on these above observations, it is concluded that the present method (*i.e.*, the formula Eq. (5)) is more suitable than some existing classical methods for solving some singular initial value problems.

References

1. Chandrasekhar, S.: Introduction to the Study of Stellar Structure. Dover, NewYork (1967)
2. Davis, H.T.: Introduction to Nonlinear Differential and Integral Equations. Dover, New York (1962)
3. Richardson, O.U.: The Emission of Electricity from Hot Bodies, London (1921)
4. Wazwaz, A.M.: A new algorithm for solving differential equations of Lane-Emden type. Appl. Math. Comput. **118**, 287–310 (2001)
5. Wazwaz, A.M.: A new method for solving singular initial value problems in the second-order ordinary differential equations. Appl. Math. Comput. **28**, 45–57 (2002)
6. Wazwaz, A.M.: The modified decomposition method for analytic treatment of differential equations. Appl. Math. Comput. **173**, 165–176 (2002)
7. Wazwaz, A.M.: Adomian decomposition method for a reliable treatment of the Emden-Fowler equation. Appl. Math. Comput. **61**, 543–560 (2005)
8. Hasan, Y.Q., Zhu, L.M.: Solving singular initial value problems in the second-order ordinary differential equations. J. Appl. Sci. **7**(17), 2505–2508 (2007)
9. Hasan, Y.Q., Zhu, L.M.: Modified adomian decomposition method for singular initial value problems in the second-order ordinary differential equations. Surv. Math. Appl. **3**, 183–193 (2008)
10. Gupta, V.G., Sharma, P.: Solving singular initial value problems of Emden-Fowler and Lane-Amden type. Int. J. Appl. Math. Comput. **1**(4), 206–212 (2009)
11. Demir, H., Sungu, I.C.: Numerical solution of a class of nonlinear Emden-Fowler equations by using differential transform method. J. Arts Sci. **12**, 75–82 (2009). Sayı/Aralık
12. Mukherjee, S., Roy, B., Chaterjee, P.K.: Solution of Lane-Emden equation by differential transform method. Int. J. Nonlinear Sci. **12**(4), 478–484 (2011)
13. Koch, O., Kofler, P., Weinmuller, E.: The implicit Euler method for the numerical solution of singular initial value problems. Appl. Num. Math. **34**, 231–252 (2000)
14. Koch, O., Weinmuller, E.: Analytic and numerical treatment of a singular initial value problem in avalanche modeling. Appl. Math. Comput. **148**, 561–570 (2004)
15. Benko, D., Biles, D.C., Robinson, M.P., Spraker, J.S.: Numerical approximation for singular second order differential equations. Math. Comput. Model. **49**, 1109–1114 (2009)
16. Lakestani, M., Saray, B.N.: Numerical solution of singular IVPs of Emden-Fowler type using Legendre scaling functions. Int. J. Nonlinear Sci. **13**(2), 211–219 (2012)
17. Hasan, M.K., Huq, M.A., Rahman, M.S., Rahman, M.M., Alam, M.S.: A new implicit method for numerical solution of singular initial value problems. Int. J. Conceptions Comput. Info Technol. **2**(1), 87–91 (2014)
18. Hasan, M.K., Ahamed, M.S., Alam, M.S., Hossain, M.B.: An implicit method for numerical solution of singular and stiff initial value problems. J. Comput. Eng. **2013**, 1–5 (2013). Article ID 720812
19. Hasan, M.K., Ahamed, M.S., Huq, M.A., Alam, M.S., Hossain, M.B.: An implicit method for numerical solution of second order singular initial value problems. Open Math. J. **7**, 1–5 (2014)
20. Huq, M.A., Hasan, M.K., Rahman, M.M., Alam, M.S.: A simple and straightforward method for evaluating some singular integrals. Far East J. Math. Edu. **7**(2), 93–107 (2011)

21. Hasan, M.K., Huq, M.A., Rahaman, M.H., Haque, B.M.I.: A more accurate and straightforward method for evaluating singular integrals. Uni. J. Appl. Math. **3**(3), 53–61 (2015)
22. Balagurusamy, E.: Numerical Methods, 4th edn., pp. 145–171. Tata McGraw-Hill Publishing Company Limited (2004)
23. Jain, M.K.: Numerical Solution of Differential Equations, 2nd edn., pp. 57–59. Wiley Eastern Limited (1991)

Constrained Data Visualization Using Rational Bi-cubic Fractal Functions

S.K. Katiyar(✉), K.M. Reddy, and A.K.B. Chand

Department of Mathematics,
Indian Institute of Technology Madras, Chennai 600036, India
sbhkatiyar@gmail.com, mahipalnitw@gmail.com, chand@iitm.ac.in

Abstract. This paper addresses a method to obtain rational cubic fractal functions, which generate surfaces that lie above a plane via blending functions. In particular, the constrained bivariate interpolation discussed herein includes a method to construct fractal interpolation surfaces that preserve positivity inherent in a prescribed data set. The scaling factors and shape parameters involved in fractal boundary curves are constrained suitably such that these fractal boundary curves are above the plane whenever the given interpolation data along the grid lines are above the plane. Our rational cubic spline FIS is above the plane whenever the corresponding fractal boundary curves are above the plane. We illustrate our interpolation scheme with some numerical examples.

Keywords: Iterated function system · Fractal interpolation functions · Bicubic partially blended fractal surface · Convergence · Constrained interpolation · Positivity

MSC: 28A80 · 26C15 · 41A20 · 65D10 · 41A29 · 65D05

1 Introduction

In 1986, Barnsley [1] first put forward the concept of fractal interpolation function (FIF) by utilizing iterated function system (IFS) to handle highly irregular data in nature and scientific phenomena. A FIF is the fixed point of the Read-Bajraktarević operator defined on a suitable function space. By imposing appropriate conditions on the scaling factors, Barnsley and Harrington [2] observed that if the underlying problem is of differentiable type, then the elements of the IFS may be suitably chosen so that the corresponding FIF is smooth. Smooth FIFs can be used to generalize classical interpolation and approximation techniques. Fractal splines with general boundary conditions have studied recently [6,8,24]. Since then, many researchers have contributed to the theory of fractal functions by constructing various type of FIFs and hidden variable FIFs [11,12,20,25,35].

Shape control, shape design and shape preservation [17,26–30] are important areas for the graphical presentation of data. In computer graphics there is often

© Springer Nature Singapore Pte Ltd. 2017
D. Giri et al. (Eds.): ICMC 2017, CCIS 655, pp. 265–277, 2017.
DOI: 10.1007/978-981-10-4642-1_23

the need to construct a curve/surface from an experimental data whose form can be interactively adjusted by means of suitable parameters and which preserves salient geometric properties such positivity, monotonicity, and convexity inherent in the data. Including aforementioned FIFs, all existing polynomial FIFs are not ideal for shape preservation. Owing to this reason, our group has introduced the shape preserving rational spline FIF in the literature [9,32,33] because rational spline FIFs has an upper hand over polynomial spline FIFs as it can carry more degrees of freedom in its description. This freedom can be utilized for various purposes and objectives to be achieved in diverse real-life problems arising in different disciplines.

Fractal interpolation surface (FIS) provides a new methodology for data fitting, which not only opens up a new research field for the theory of recursive functions but also provides a powerful tool for computer graphics and widely used in modeling natural surfaces such as terrains, metals, planets, rocks and so on. A prevue of the existing theory on fractal surfaces is provided next. Massopust [22] was the first to put forward the construction of the fractal surfaces via IFS and later it is followed in earnest by researchers [4,23]. In reference [18], a construction of self-affine FIS with a triangular domain for arbitrary interpolation points is constructed and in mean time, a more general construction of hidden variable fractal interpolation surfaces, which carry additional free parameters is constructed in [19]. Xie and Sun [34] proposed a mathematical model of the bivariate FISs on the rectangular grid with arbitrary contraction factors and without any conditions on boundary points but these bivariate FIS is not a graph of a continuous function. The aforementioned result is improved and corrected in [14] by taking collinear boundary. In order to ensure the continuity of the surface, all of them assume that the interpolation nodes on the boundary are collinear. Construction of fractal interpolation surfaces for arbitrary data on a rectangular grid is given in [21] but the vertical scaling factors used in the IFS must all be equal in this case. By using function vertical scaling factors, a method of construction for the fractal interpolation surfaces on a rectangular domain with arbitrary interpolation nodes is proposed by Feng et al. [16]. Recently, Songil [31] generalized the construction of fractal surfaces in the paper [14,16]. These surfaces are self-similar, self-affine or more generally self-referential. To approximate self-affine and non-self-affine surfaces simultaneously, Chand and Kapoor [5] broached the notion of coalescence hidden variable fractal interpolation surface and extended to smooth fractal surfaces in [7]. The methods, which define surfaces via blending function schemes [15] by utilizing curves that are already available and they obtained wide attention in the literature as well as in the design environment, especially when the constructed surface is desired to preserve important geometrical properties like positivity, monotony and convexity. Note that all these FISs do not follow shape preserving aspects of prescribed surface data. Recently, the concept of positivity preserving fractal surface is introduced by Chand and coworkers in [10]. This paper has been devoted to the visualization of surface data arranged on a rectangular grid in the form of blending rational cubic spline FISs.

The remainder of this paper is organized as follows. In Sect. 2, we recall the basics of IFS theory and its connection with fractal interpolation. In Sect. 3, we construct the rational cubic spline FIFs (fractal boundary curves) in x and y

directions, and by using these fractal boundary curves and blending functions, we form a blending rational cubic spline FIS. The sufficient conditions of the scaling factors and shape parameters so that the desired surface remains above a specified plane is addressed in Sect. 4. The developed rational cubic spline FIS is illustrated through suitably chosen numerical examples in Sect. 5.

2 Basic Facts

In this section we introduce the basic objects that we will work with in this paper. We also state the intermediate proposition corresponding to the main steps of our argument. For a more extensive treatment, the reader may consult [1,2,15].

2.1 IFS for Fractal Functions

For $r \in \mathbb{N}$, let $\mathbb{N}_r$ denote the subset $\{1, 2, \ldots, r\}$ of $\mathbb{N}$. Let a set of data points $\mathscr{D} = \{(x_i, y_i) \in \mathbb{R}^2 : i \in \mathbb{N}_m\}$ satisfying $x_1 < x_2 < \cdots < x_m$, $m > 2$, be given. Set $I = [x_1, x_m]$, $I_i = [x_i, x_{i+1}]$ for $i \in \mathbb{N}_{m-1}$. Suppose contractive homeomorphisms $L_i : I \to I_i$, $i \in \mathbb{N}_{m-1}$, are given by $L_i(x) = a_i x + b_i = \frac{x_{i+1}-x_i}{x_m-x_1}x + \frac{x_m x_i - x_1 x_{i+1}}{x_m-x_1}$, and let $m-1$ continuous mappings $F_i, i \in \mathbb{N}_{m-1}$, are defined by $F_i(x, y) = \alpha_i y + r_i(x), |\alpha_i| \leq k < 1$, satisfying the join-up conditions $F_i(x_1, y_1) = y_i, \quad F_i(x_m, y_m) = y_{i+1}, i \in \mathbb{N}_{m-1}$, and $r_i : I \to \mathbb{R}$ are suitable continuous functions, generally polynomials. Define $w_i : X \to I_i \times \mathbb{R} \subseteq X, \quad w_i(x, y) = \big(L_i(x), F_i(x, y)\big) \ \forall \ i \in \mathbb{N}_{m-1}$. It is known [1] that there exists a metric on $\mathbb{R}^2$, equivalent to the Euclidean metric, with respect to which $w_i, i \in \mathbb{N}_{m-1}$, are contractions. The collection $\mathscr{I} = \{X; w_i : i \in \mathbb{N}_{m-1}\}$ is called an IFS.Associated with the IFS $\mathscr{I}$, there is a set valued Hutchinson map $W : H(X) \to H(X)$ defined by $W(B) = \bigcup_{i=1}^{m-1} w_i(B)$ for $B \in H(X)$, where $H(X)$ is the set of all nonempty compact subsets of X endowed with the Hausdorff metric h_d. The Hausdorff metric h_d completes $H(X)$. Further, W is a contraction map on the complete metric space $(H(X), h_d)$. By the Banach Fixed Point Theorem, there exists a unique set $G \in H(X)$ such that $W(G) = G$. This set G is called the attractor or deterministic fractal corresponding to the IFS $\mathscr{I}$. According to [1], the IFS $\mathscr{I}$ has a unique attractor G which is the graph of a continuous function $g : I \to \mathbb{R}, g(x_i) = y_i, i \in \mathbb{N}_m$. The function g is called a FIF or a self-referential function generated by the IFS $\mathscr{I}$, and it takes the form $g(L_i(x)) = \alpha_i g(x) + r_i(x), x \in [x_1, x_m]$. Barnsley and Harrington [2] introduced a FIF with $\mathscr{C}^r$-continuity and this result was extended to $\mathscr{C}^1$-rational spline fractal functions [9] in the following proposition:

Proposition 1. *Let $\{(x_i, y_i) : i \in \mathbb{N}_m\}$ be given interpolation data with strictly increasing abscissae and $d_i (i \in \mathbb{N}_m)$ be the derivative values at the knots. Consider the IFS $\mathscr{I}$, with $r_i(x) = \frac{p_i(x)}{q_i(x)}$, $p_i(x)$ and $q_i(x) \neq 0$ are cubic polynomials*

$\forall x \in [x_1, x_m]$, *and* $|\alpha_i| < a_i, i \in \mathbb{N}_{m-1}$. *Let* $F_{i,1}(x, y) = \frac{\alpha_i y + r_i^{(1)}(x)}{a_i}$, *where* $r_i^{(1)}(x)$ *represents the derivative of* $r_i(x)$ *with respect to* x. *If for* $i \in \mathbb{N}_{m-1}$,

$$F_i(x_1, y_1) = y_i,\ F_i(x_m, y_m) = y_{i+1},\ F_{i,1}(x_1, d_1) = d_i,\ F_{i,1}(x_m, d_m) = d_{i+1}, \quad (1)$$

then the attractor of the IFS $\mathscr{I}$ *is the graph of a Hermite rational cubic spline FIF.*

This completes our preparations for the current study, and we are now ready for our main section.

3 Bicubic Partially Blended Rational Fractal Interpolation

We wish to a construct a $\mathscr{C}^1$-continuous bivariate function $\Phi : \mathbb{R} \to \mathbb{R}$ such that $\Phi(x_i, y_j) = z_{i,j}$, $\frac{\partial \Phi}{\partial x}(x_i, y_j) = z_{i,j}^x$, and $\frac{\partial \Phi}{\partial y}(x_i, y_j) = z_{i,j}^y$ for $i \in \mathbb{N}_m, j \in \mathbb{N}_n$. This is achieved by blending the univariate rational cubic FIFs using the partially bicubic Coons technique [3]. Thus with the obvious reasons, the bivariate function Φ is termed a bicubic partially blended rational fractal interpolation surface (FIS). For the construction of the rational cubic spline FIS, first we develop the fractal boundary curves from a given set of surface data by taking $r_i(x)$ as a rational function with a cubic polynomial in numerator and preassigned quadratic polynomial with two shape parameters in denominator as follows:

3.1 Construction of Rational Cubic Spline FIFs (Fractal Boundary Curves)

Let $\Delta = \{(x_i, y_j, z_{i,j}) : i \in \mathbb{N}_m, j \in \mathbb{N}_n\}$ be a set of bivariate interpolation data, where $x_1 < x_2 < \cdots < x_m$ and $y_1 < y_2 < \cdots < y_n$ and denote $h_i = x_{i+1} - x_i$, $h_j^* = y_{j+1} - y_j$, $i \in \mathbb{N}_{m-1}$, $j \in \mathbb{N}_{n-1}$. Set $K_{i,j} = I_i \times J_j = [x_i, x_{i+1}] \times [y_j, y_{j+1}]; i \in \mathbb{N}_{m-1}$, $j \in \mathbb{N}_{n-1}$ be the generic subrectangular region and take $K = I \times J = [x_1, x_m] \times [y_1, y_n]$. Let $z_{i,j}^x$ and $z_{i,j}^y$ are the x-partial and y-partial derivatives of the original function at the point (x_i, y_j). Consider a surface data set $\{(x_i, y_j, z_{i,j}, z_{i,j}^x, z_{i,j}^y) : i \in \mathbb{N}_m, j \in \mathbb{N}_n\}$ placed on the rectangular grid K. It is plain to see that univariate data set obtained by taking sections of K with the line $y = y_j$ (along the j-th grid line parallel to x-axis), $j \in \mathbb{N}_n$, namely $R_j = \{(x_i, y_j, z_{i,j}, z_{i,j}^x) : i \in \mathbb{N}_m\}$. Let $L_i : I \to I_i$, be affine maps $L_i(x) = a_i x + b_i$ satisfying $L_i(x_1) = x_i$, $L_i(x_m) = x_{i+1}$. By considering Proposition 1 with interpolation data R_j and shape parameters $u_{i,j} > 0$, $v_{i,j} > 0$ for $i \in \mathbb{N}_{m-1}$, we construct rational cubic spline FIF (fractal boundary curve):

$$\psi(x, y_j) = \alpha_{i,j} \psi\big({L_i}^{-1}(x), y_j\big) + \frac{P_{i,j}(\theta)}{Q_{i,j}(\theta)}, \quad (2)$$

where in

$$P_{i,j}(\theta) = u_{i,j}(z_{i,j} - \alpha_{i,j}z_{1,j})(1-\theta)^3 + \{(u_{i,j}+2)z_{i,j} + u_{i,j}h_i z^x_{i,j} - \alpha_{i,j}[(u_{i,j}+2)z_{1,j} + u_{i,j}(x_m - x_1)z^x_{1,j}]\}(1-\theta)^2\theta + ((v_{i,j}+2)z_{i+1,j} - v_{i,j}h_i z^x_{i+1,j} - \alpha_{i,j}[(v_{i,j}+2)z_{m,j} - v_{i,j}(x_m - x_1)z^x_{m,j}](1-\theta)\theta^2 + v_{i,j}(z_{i+1,j} - \alpha_{i,j}z_{m,j})\theta^3,$$

$$Q_{i,j}(\theta) = u_{i,j}(1-\theta^2) + 2\theta(1-\theta) + v_{i,j}\theta^2, \theta = \frac{{L_i}^{-1}(x) - x_1}{x_m - x_1} = \frac{x - x_i}{h_i},\ x \in I_i.$$

Similarly, for each $i \in \mathbb{N}_m$, let us consider the univariate data set by taking sections of K with the line $x = x_i$ (along the i-th grid line parallel to y-axis), namely $R_i = \{(x_i, y_j, z_{i,j}, z^y_{i,j}) : j \in \mathbb{N}_n\}$. Consider the affine maps $L^*_j : [y_1, y_n] \to [y_j, y_{j+1}]$ defined by $L^*_j(y) = c_j y + d_j$ satisfying $L^*_j(y_1) = y_j$ and $L^*_j(y_n) = y_{j+1}$, $j \in \mathbb{N}_n$. For a fixed $i \in \mathbb{N}_m$, let $\alpha^*_{i,j}$ be the scaling factors along the vertical grid line $x = x_i$ such that $|\alpha^*_{i,j}| < c_j < 1$ and let the shape parameters be selected so as to satisfy $u^*_{i,j} > 0$ and $v^*_{i,j} > 0$ for all $j \in \mathbb{N}_{n-1}$. Again, following Proposition 1 with interpolation data R_i, we construct rational cubic spline FIF (fractal boundary curve):

$$\psi^*(x_i, y) = \alpha^*_{i,j}\psi^*\big(x_i, {L^*_j}^{-1}(y)\big) + \frac{P^*_{i,j}(\phi)}{Q^*_{i,j}(\phi)}, \tag{3}$$

where

$$P^*_{i,j}(\phi) = u^*_{i,j}(z_{i,j} - \alpha^*_{i,j}z_{i,1})(1-\phi)^3 + \{(u^*_{i,j}+2)z_{i,j} + u^*_{i,j}h^*_j z^y_{i,j} - \alpha^*_{i,j}[(u^*_{i,j}+2)z_{i,1} + u^*_{i,j}(y_n - y_1)z^y_{i,1}]\}(1-\phi)^2\phi + ((v^*_{i,j}+2)z_{i,j+1} - v^*_{i,j}h^*_j z^y_{i,j+1} - \alpha^*_{i,j}[(v^*_{i,j}+2)z_{i,n} - v^*_{i,j}(y_n - y_1)z^y_{i,n}])(1-\phi)\phi^2 + v^*_{i,j}(z_{i,j+1} - \alpha^*_{i,j}z_{i,n})\phi^3,$$

$$Q^*_{i,j}(\phi) = u^*_{i,j}(1-\phi^2) + 2\phi(1-\phi) + v^*_{i,j}\phi^2, \phi = \frac{{L^*_j}^{-1}(y) - y_1}{y_n - y_1} = \frac{y - y_j}{h^*_j},\ y \in J_j.$$

3.2 Formation of Blending Rational Cubic Spline FIS

In this section, we blend these univariate FIFs given in (2)–(3) using well-known bicubic partially blended Coons patch to obtain the desired surface. Consider the network of FIFs $\psi(x, y_j)$, $\psi(x, y_{j+1})$, $\psi^*(x_i, y)$, and $\psi^*(x_{i+1}, y)$ for $i \in \mathbb{N}_{m-1}, j \in \mathbb{N}_{n-1}$. Consider the cubic Hermite functions $b^i_{0,3}(x) = (1-\theta)^2(1+2\theta)$, $b^i_{3,3}(x) = \theta^2(3-2\theta)$, $b^j_{0,3}(y) = (1-\phi)^2(1+2\phi)$, and $b^j_{3,3}(y) = \phi^2(3-2\phi)$. These functions are called the blending functions, because their effect is to blend together four separate boundary curves to provide a single well-defined surface. On each individual patch $K_{i,j} = I_i \times J_j$, $i \in \mathbb{N}_{m-1}$, $j \in \mathbb{N}_{n-1}$, we define a blending rational cubic spline FIS.

$$\Phi(x, y) = -\begin{bmatrix} -1 & b^i_{0,3}(x) & b^i_{3,3}(x) \end{bmatrix} \begin{bmatrix} 0 & \psi(x, y_j) & \psi(x, y_{j+1}) \\ \psi^*(x_i, y) & z_{i,j} & z_{i,j+1} \\ \psi^*(x_{i+1}, y) & z_{i+1,j} & z_{i+1,j+1} \end{bmatrix} \begin{bmatrix} -1 \\ b^j_{0,3}(y) \\ b^j_{3,3}(y) \end{bmatrix} \tag{4}$$

The rational cubic spline FIS Φ can be written in the equivalent form to understand the geometry of Coons construction as follows:

$$\begin{aligned}\Phi(x,y) = & \left[b^i_{0,3}(x)\ b^i_{3,3}(x)\right]\begin{bmatrix}\psi^*(x_i,y)\\ \psi^*(x_{i+1},y)\end{bmatrix} + \left[b^j_{0,3}(y)\ b^j_{3,3}(y)\right]\begin{bmatrix}\psi(x,y_j)\\ \psi(x,y_{j+1})\end{bmatrix}\\ & - \left[b^i_{0,3}(x)\ b^i_{3,3}(x)\right]\begin{bmatrix}z_{i,j} & z_{i,j+1}\\ z_{i+1,j} & z_{i+1,j+1}\end{bmatrix}\begin{bmatrix}b^j_{0,3}(y)\\ b^j_{3,3}(y)\end{bmatrix},\\ := & \Phi_1(x,y) + \Phi_2(x,y) - \Phi_3(x,y).\end{aligned} \tag{5}$$

From (4), it is easy to verify that $\Phi(x_i,y_j) = z_{i,j}, \Phi(x_{i+1},y_j) = z_{i+1,j}, \Phi(x_i,y_{j+1}) = z_{i,j+1}, \Phi(x_{i+1},y_{j+1}) = z_{i+1,j+1}$. Thus Φ interpolates the given data at grid points. The following theorem is a direct consequence of the properties of the univariate FIFs forming the boundaries of Φ and the blending functions. The proof is patterned after [10].

Theorem 1. *The rational cubic spline FIS Φ (cf. (4)) satisfies the interpolation conditions $\Phi(x_i,y_j) = z_{i,j}$, $\frac{\partial\Phi}{\partial x}(x_i,y_j) = z^x_{i,j}$ and $\frac{\partial\Phi}{\partial y}(x_i,y_j) = z^y_{i,j}$, for $i \in \mathbb{N}_m$, $j \in \mathbb{N}_n$, and $\Phi \in \mathscr{C}^1(K)$.*

Remark 1. When $\alpha = [0]_{(m-1)\times n}$ and $\alpha^* = [0]_{m\times(n-1)}$, one can get the classical rational cubic surface interpolant as

$$\begin{aligned}C(x,y) = & b^j_{0,3}(y)\psi(x,y_j) + b^j_{3,3}(y)\psi(x,y_{j+1}) + b^i_{0,3}(x)\psi^*(x_i,y) + b^i_{3,3}(x)\psi^*(x_{i+1},y)\\ & - b^i_{0,3}(x)b^j_{0,3}(y)z_{i,j} - b^i_{0,3}(x)b^j_{3,3}(y)z_{i,j+1} - b^i_{3,3}(x)b^j_{0,3}(y)z_{i+1,j} - b^i_{3,3}(x)\times\\ & b^j_{3,3}(y)z_{i+1,j+1},\end{aligned}$$

where $\psi(x,y_j)$, $j \in \mathbb{N}_n$ and $\psi^*(x_i,y)$, $i \in \mathbb{N}_m$ are the classical rational cubic interpolants obtained by Sarfraz et al. [28] for the data sets $R_j, j \in \mathbb{N}_n$ and $R_i, i \in \mathbb{N}_m$ respectively.

4 Bicubic Partially Blended Rational FIS Above a Prescribe Plane

In this section we constrain the parameters so that the corresponding rational cubic spline FIS Φ (cf. (4)) would be utilized to achieve the interpolating surface when data are under consideration over an arbitrary plane.

Theorem 2. *Let $\{x_i, y_j, z_{i,j} : i \in \mathbb{N}_m, j \in \mathbb{N}_n\}$ be an interpolation data set. For each $j \in \mathbb{N}_n$, the univariate FIF $\psi(x,y_j)$ lies above the line $t = c[1 - \frac{x}{a} - \frac{y_j}{b}]$ if the scaling factors and the shape parameters are selected so as to satisfy the following conditions*

(1) The scaling factor such that $0 \le \alpha_{i,j} < min\{a_i, \frac{z_{i,j}-t_{i,j}}{z_{1,j}-t_{1,j}}, \frac{z_{i+1,j}-t_{i+1,j}}{z_{m,j}-t_{m,j}}\}$,

(2) The shape parameters $u_{i,j} > 0$ and $v_{i,j} > 0$ satisfy:

(i) $u_{i,j}[(z_{i,j} - \alpha_{i,j}z_{1,j}) + (h_i z^x_{i,j} - t_{i+1,j} - \alpha_{i,j}(x_m - x_1)z^x_{1,j} + \alpha_{i,j}t_{m,j})] + 2[z_{i,j} - t_{i,j} - \alpha_{i,j}z_{1,j} + \alpha_{i,j}t_{1,j}] \ge 0$,

(ii) $v_{i,j}[(z_{i+1,j}-\alpha_{i,j}z_{m,j})+\{h_i z^x_{i+1,j}-t_{i,j}-\alpha_{i,j}(x_m-x_1)z^x_{m,j}+\alpha_{i,j}t_{i,j}\}]+2[z_{i+1,j}-t_{i+1,j}-\alpha_{i,j}z_{m,j}+\alpha_{i,j}t_{m,j}]\geq 0.$

Proof. Let $\{x_i, y_j, z_{i,j} : i \in \mathbb{N}_m, j \in \mathbb{N}_n\}$ be an interpolation data set lying above the plane $t = c[1-\frac{x}{a}-\frac{y}{b}]$, i,e. $z_{i,j} > t_{i,j} = c[1-\frac{x}{a}-\frac{y}{b}]$ for all $i \in \mathbb{N}_m, j \in \mathbb{N}_n$. We wish to find conditions on the parameters of rational cubic spline FIS so that it lies above the aforementioned plane, that is $\Phi(x,y) > t(x,y)$ for all $(x,y) \in K$. We recall that the surface generated by the rational cubic spline FIS ψ lies above the plane if the network of boundary curves $\psi(x,y_j)\ \forall\ j \in \mathbb{N}_n$ and $\psi^*(x_i,y)\ \forall\ i \in \mathbb{N}_m$ lie above the plane [13]. Since, $\Phi(x_i,y_j) = z_{i,j} > t(x_i,y_j)$ for all $i \in \mathbb{N}_m, j \in \mathbb{N}_n$, the proof of $\psi(\tau,y_j) > t(\tau,y_j)$ for all $\tau \in I$ is equivalent to find the conditions for which $\psi(x,y_j) > t(x,y_j)$, $x \in I$ implies $\psi(L_i(x),y_j) > t(L_i(x),y_j)$ for $x \in I$. Assume $\psi(x,y_j) > t(x,y_j)$. We need to prove that

$$\alpha_{i,j}\psi(x,y_j)+\frac{P_{i,j}(\theta)}{Q_{i,j}(\theta)} > c[1-\frac{a_i x+b_i}{a}-\frac{y_j}{b}], \tag{6}$$

Since $Q_{i,j}(\theta) > 0$, in view of assumptions $\psi(x,y_j) > t(x,y_j)$ and $\alpha_{i,j} \geq 0$ for all $i \in \mathbb{N}_m, j \in \mathbb{N}_n$, we deduce that the following condition confirm (6).

$$\alpha_{i,j}c[1-\frac{x}{a}-\frac{y_j}{b}]Q_{i,j}(\theta)+P_{i,j}(\theta)-\{c[(1-\frac{y_j}{b})-\frac{a_i x+b_i}{a}]\}Q_{i,j}(\theta) > 0. \tag{7}$$

Performing some algebraic calculations by substituting $x = x_i + \theta h_i$ and $Q_{i,j}(\theta) = u_{i,j}(1-\theta)^3+(u_{i,j}+2)\theta(1-\theta)^2+(v_{i,j}+2)\theta^2(1-\theta)+v_{i,j}\theta^3$ (using the degree elevated form of $Q_{i,j}$), and using the expression from (2) for $P_{i,j}(\theta)$, we see that (7) may be reformulated as follows:

$$U_{i,1}(1-\theta)^3+U_{i,2}\theta(1-\theta)^2+U_{i,3}\theta^2(1-\theta)+U_{i,4}\theta^3 > 0, \theta \in [0,1], \tag{8}$$

where
$U_{i,1} = u_{i,j}[z_{i,j}-t_{i,j}-\alpha_{i,j}(z_{1,j}-t_{1,j})]$,
$U_{i,2} = u_{i,j}[(3z_{i,j}-2t_{i,j})-t_{i+1,j}+h_i z^x_{i,j}-\alpha_{i,j}\{(3z_{1,j}-2t_{1,j})+(x_m-x_1)z^x_{1,j}-t_{m,j}\}]$,
$U_{i,3} = v_{i,j}[(3z_{i+1,j}-2t_{i+1,j})-t_{i,j}-h_i z^x_{i+1,j}-\alpha_{i,j}\{(3z_{m,j}-2t_{m,j})-(x_m-x_1)z^x_{m,j}-t_{1,j}\}]$,
$U_{i,4} = v_{i,j}[z_{i+1,j}-t_{i+1,j}-\alpha_{i,j}(z_{m,j}-t_{m,j})]$.

With the substitution $\theta = \frac{\nu}{\nu+1}$, (8) is equivalent to

$$U_{i,4}\nu^3+U_{i,3}\nu^2+U_{i,2}\nu+U_{i,1} > 0\ \ \forall\ \ \nu > 0. \tag{9}$$

We know that [26], a cubic polynomial $\rho(\xi) = a\xi^3+b\xi^2+c\xi+d \geq 0$ for all $\xi \geq 0$, if and only if $(a,b,c,d) \in W_1 \cup W_2$, where

$W_1 = \{(a,b,c,d) : a \geq 0, b \geq 0, c \geq 0, d \geq 0\}$,
$W_2 = \{(a,b,c,d) : a \geq 0, d \geq 0, 4ac^3+4db^3+27a^2d^2-18abcd-b^2c^2 \geq 0\}$.

As the condition involved in W_2 is computationally cumbersome, to obtain a set of sufficient condition for the positivity of (9), we use comparatively efficient

and reasonably acceptable choice of parameters determined by W_1. Thus the polynomial in (9) is positive if $U_{i,1} > 0, U_{i,2} > 0, U_{i,3} > 0$ and $U_{i,4} > 0$ are satisfied. It is straight forward to see that $U_{i,1} > 0$ is satisfied if $\alpha_{i,j} < \frac{z_{i,j}-t_{i,j}}{z_{1,j}-t_{1,j}}$ and $U_{i,4} > 0$ if $\alpha_{i,j} < \frac{z_{i+1,j}-t_{i+1,j}}{z_{m,j}-t_{m,j}}$, since the shape parameters $u_{i,j} > 0$, $v_{i,j} > 0$ for $i \in \mathbb{N}_{m-1}$. It can be seen that the additional conditions on $u_{i,j}$, $v_{i,j}$ prescribed in the theorem ensure the positivity of $U_{i,2} > 0$ and $U_{i,3} > 0$. This completes the proof.

With the similar rendition of arguments as above, we can prove the following theorem.

Theorem 3. *Let $\{x_i, y_j, z_{i,j} : i \in \mathbb{N}_m, j \in \mathbb{N}_n\}$ be an interpolation data set. For each $i \in \mathbb{N}_m$, the univariate FIF $\psi^*(x_i, y)$ lies above the line $t = -\frac{c}{b}y + c(1 - \frac{x_i}{a})$ if the scaling factors and the shape parameters are selected as follows*

(1) The scaling factors such that $0 \le \alpha^*_{i,j} < min\{c_j, \frac{z_{i,j}-t_{i,j}}{z_{i,1}-t_{i,1}}, \frac{z_{i,j+1}-t_{i,j+1}}{z_{i,n}-t_{i,n}}\}$,
(2) The shape parameters $u^*_{i,j} > 0$ and $v^*_{i,j} > 0$ satisfy
(i) $u^*_{i,j}[(z_{i,j} - \alpha^*_{i,j} z_{i,1}) + (h^*_j z^y_{i,j} - t_{i,j+1} - \alpha^*_{i,j}(y_n - y_1) z^y_{i,1}) + \alpha^*_{i,j} t_{i,n}] + 2[z_{i,j} - t_{i,j} - \alpha^*_{i,j} z_{i,1} + \alpha^*_{i,j} t_{i,1}] \ge 0$,
(ii) $v^*_{i,j}[(z_{i,j+1} - \alpha^*_{i,j} z_{i,n}) + \{h^*_j z^y_{i,j+1} - t_{i,j} - \alpha^*_{i,j}(y_n - y_1) z^y_{i,n} + \alpha^*_{i,j} t_{i,1}\}] + 2[z_{i,j+1} - t_{i,j+1} - \alpha^*_{i,j} z_{i,n} + \alpha^*_{i,j} t_{i,n}] \ge 0$.

Theorem 4. *Let $\{x_i, y_j, z_{i,j} : i \in \mathbb{N}_m, j \in \mathbb{N}_n\}$ be an interpolation data set that lies above the plane $t = c[1 - \frac{x}{a} - \frac{y}{b}]$ i.e. $z_{i,j} > t_{i,j}$ for all $i \in \mathbb{N}_m, j \in \mathbb{N}_n$. Then the rational cubic spline FIS Φ (cf. (4)) lies above the plane provided the horizontal scaling parameters $\alpha_{i,j}$ for $i \in \mathbb{N}_{m-1}, j \in \mathbb{N}_n$ and the vertical scaling parameters $\alpha^*_{i,j}$ for $i \in \mathbb{N}_m, j \in \mathbb{N}_{n-1}$, the horizontal shape parameters $u_{i,j}$, $v_{i,j}$ for $i \in \mathbb{N}_{m-1}, j \in \mathbb{N}_n$ and the vertical shape parameters $u^*_{i,j}$, $v^*_{i,j}$ for $i \in \mathbb{N}_m, j \in \mathbb{N}_{n-1}$ satisfy the hypotheses of Theorems 2–3.*

Remark 2. When $\alpha = [0]_{(m-1)\times n}$ and $\alpha^* = [0]_{m\times(n-1)}$, we recover the conditions for which the traditional nonrecursive bicubic partially blended rational function C is above a prescribed plane. The foregoing theorem includes, in particular, the conditions under which rational cubic spline FIS Φ preserves the positivity property inherent in a given bivariate data set.

5 Numerical Examples

For the illustration of the developed scheme for constrained interpolation problem, consider the surface interpolation data (Table 1) with 16 points taken at random. Let us note that in Table 1, the 1^{st}, 2^{nd}, and 3^{rd} components of **(.,.,.)** represent the function value, the first order partial derivatives in x-direction and y-direction at (x_i, y_j) respectively, where $i, j \in \{1, 2, 3, 4\}$. Note that the data set reported in Table 1 lies above the plane $t = 1 - \frac{x}{8} - \frac{y}{8}$. Surface patch values are given in Table 2. Utilizing the prescription given in Theorem 4, we have calculated the restrictions on the scaling and shape parameters to obtain the

rational cubic spline FIS which lies above the plane. The details of the scaling and shape parameters used in the construction of Fig. 1a–h are given in Tables 3–4. For arbitrary choice of the matrices of the scaling and shape parameters (see Tables 3–4), a unconstrained rational cubic spline FIS is generated in Fig. 1a. It may be observed that some portion of the surface lies below the plane. This illustrates the importance of the Theorem 4. Choosing the scaling and the shape parameters according to Theorem 4 (see Tables 3–4), rational cubic spline FISs that lie above the plane are generated in Fig. 1b–h. Now we take Fig. 1b as reference fractal surface to illustrates the effects of changes in matrices of the scaling and shape parameters in the rational cubic spline FISs. We can notice the effects in the rational cubic spline FIS due to changes in the scaling factors

Table 1. Interpolation data for constrained rational cubic FISs.

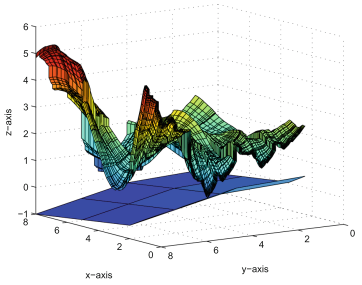
$\downarrow x/y \rightarrow$	1	2	5	8
1	(2.6301, 0.6447, −2.7380)	(1.6255, −0.5959, −2.9587)	(2.9835, 0.0216, 3.1202)	(2.0531, 2.5992, −3.2192)
4	(2.6668, −0.78, −1.2807)	(1.5811, 0.7209, −1.3839)	(3.0486, −0.0262, 1.4594)	(2.0432, −3.1445, −1.5058)
6	(1.6574, 1.1458, 0.1425)	(2.8023, −1.0589, 0.1540)	(1.2548, 0.0384, −0.1624)	(2.3150, 4.6191,0.1676)
8	(2.7101, −0.1736, −1.3830)	(1.5288, 0.1605, −1.4945)	(3.2155, −0.0058, 1.5761)	(2.0315, −0.7, −1.6261)

Table 2. Surface patch above the plane $t = 1 - \frac{x}{8} - \frac{y}{8}$.

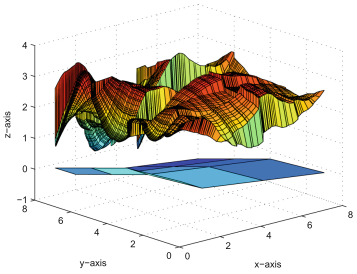
0.75	0.375	0.125	−0.125
0.625	0.25	0	−0.25
0.25	−0.125	−0.375	−0.625
−0.125	−0.5	−0.75	−1

Table 3. Scaling matrices in the construction of the blending rational cubic FISs in Fig. 1.

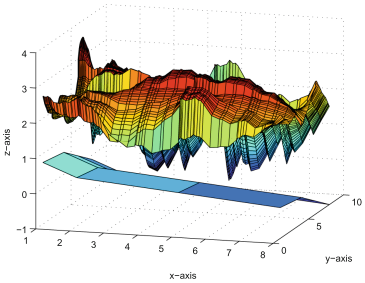
Scaling matrices in x-direction	Figs	Scaling matrices in y-direction	Figs
$\boldsymbol{\alpha} = 0.143^{*}\ [1]_{3\times 4}$	Fig. 1a	$\alpha^{*} = 0.25^{*}\ [1]_{4\times 3}$	Fig. 1a
$\alpha = \begin{bmatrix} 0.0029 & 0.0029 & 0.0029 & 0.0029 \\ 0.2886 & 0.2886 & 0.2805 & 0.2886 \\ 0.2886 & 0.2886 & 0.2886 & 0.2886 \end{bmatrix}$	Fig. 1b–f	$\alpha^{*} = \begin{bmatrix} 0.1986 & 0.0557 & 0.0557 \\ 0.1986 & 0.0557 & 0.0557 \\ 0.1986 & 0.0557 & 0.0557 \\ 0.1986 & 0.0557 & .0557 \end{bmatrix}$	Fig. 1b–c, e–f
$\alpha = \begin{bmatrix} 0.1429 & 0.1429 & 0.1429 & 0.1429 \\ 0.4286 & 0.4286 & 0.4205 & 0.4286 \\ 0.4286 & 0.4286 & 0.4286 & 0.4286 \end{bmatrix}$	Fig. 1c,e	$\alpha^{*} = \begin{bmatrix} 0.4286 & 0.2857 & 0.2857 \\ 0.4286 & 0.2857 & 0.2857 \\ 0.4286 & 0.2857 & 0.2857 \\ 0.4286 & 0.2857 & 0.2857 \end{bmatrix}$	Fig. 1d–e
$\alpha = [0]_{3\times 4}$	Fig. 1f	$\alpha^{*} = [0]_{4\times 3}$	Fig. 1f



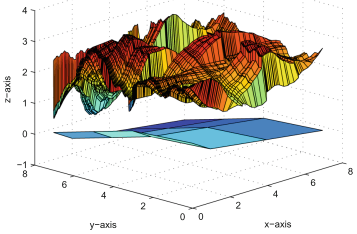
(a) Unconstrained Rational FIS



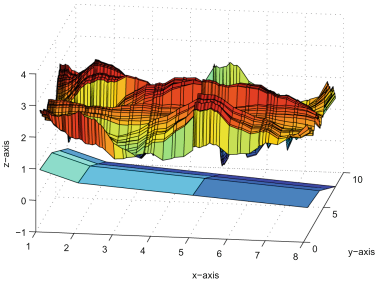
(b) Constrained rational FIS.



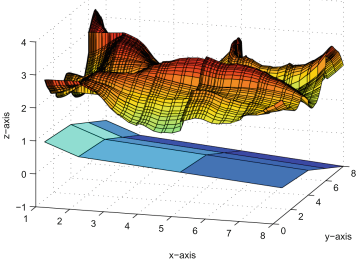
(c) Effects of α in x-direction with respect to Fig.1b.



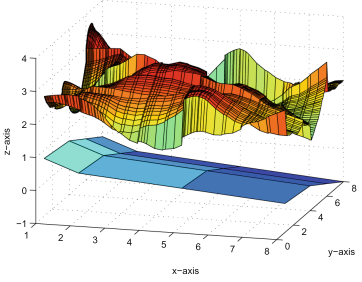
(d) Effects of α^* in y-direction with respect to Fig.1b.



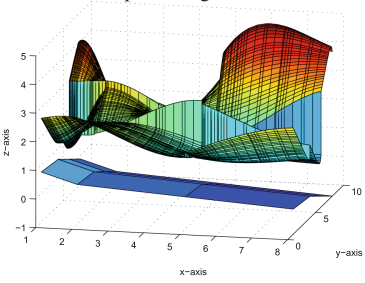
(e) Effects of α and α^* with respect to Fig.1b.



(f) Effects of change in u with respect to Fig.1b.



(g) Effects of change in v with respect to Fig.1b.



(h) Classical constrained interpolant

Fig. 1. Rational cubic FISs with constrained interpolation.

Table 4. shape parameter matrices in the construction of the blending rational cubic FISs in Fig. 1.

Matrices of shape parameters in x-direction	Figs	Matrices of shape parameters in y-direction	Figs
$u = [1]_{3\times 4}$	Fig. 1a,h	$u^* = 200^*[1]_{4\times 3}$	Fig. 1a,h
$v = [1]_{3\times 4}$	Fig. 1a,h	$v^* = [1]_{4\times 3}$	Fig. 1a,h
$u = \begin{bmatrix} 0.001 & 0.001 & 0.001 & 0.001 \\ 15.0375 & 0.001 & 0.001 & 0.001 \\ 0.001 & 0.001 & 0.001 & 0.001 \end{bmatrix}$	Fig. 1b,d,g	$u^* = \begin{bmatrix} 6.0763 & 2.1270 & 0.001 \\ 0.001 & 0.001 & 0.001 \\ 0.001 & 0.001 & 0.001 \\ 0.001 & 0.001 & 0.001 \end{bmatrix}$	Fig. 1b–c,g
$v = 0.001^*\ [1]_{3\times 4}$	Fig. 1b–c,f	$v^* = \begin{bmatrix} 1.03386 & 0.001 & 5.5899 \\ 4.9374 & 0.001 & 0.001 \\ 0.001 & 0.001 & 0.001 \\ 0.001 & 0.001 & 0.001 \end{bmatrix}$	Fig. 1b–c,f
$u = \begin{bmatrix} 0.001 & 0.001 & 0.001 & 0.001 \\ 2.7445 & 0.001 & 0.001 & 0.3758 \\ 0.001 & 0.001 & 0.001 & 0.001 \end{bmatrix}$	Fig. 1c,e	$u^* = 0.0001^*[1]_{4\times 3}$	Fig. (1)d-e
$v = 0.001^*[1]_{3\times 4}$	Fig. 1d,g	$v^* = 0.0001^*\ [1]_{4\times 3}$	Fig. 1d–e
$u = \begin{bmatrix} 0.1 & 0.1 & 0.1 & 0.1 \\ 15.1365 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 \end{bmatrix}$	Fig. 1f	$u^* = \begin{bmatrix} 7.0753 & 3.1260 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$	Fig. 1f
$v = [1]_{3\times 4}$	Fig. 1g	$v^* = \begin{bmatrix} 2.0328 & 1 & 6.5889 \\ 5.9364 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$	Fig. 1g

and/or shape parameters in Fig. 1c–h. The constrained traditional nonrecursive bicubic partially blended rational function C, which is above a prescribed plane, generated in Fig. 1h. Note that, in particular, all these fractal surfaces preserve the positivity property of prescribed data.

6 Conclusion

In the current article, we have introduced the rational cubic spline FIS for the data arranged on the rectangular grid. We have applied the well-known partially bicubic Coons technique to construct a surface whose boundaries consist of the rational cubic FIFs because it provides an elegant method of constructing shape preserving surfaces. Restrictions on the scaling and shape parameters for the rational cubic spline FIS are deduced so that it lies above a prescribed plane. In particular, we also obtain the positivity of the rational cubic spline FIS. Our scheme offers a large flexibility for simulation or modeling of objects with smooth geometric shapes because the shapes of rational cubic spline FIS can be adjusted by using different choices of the scaling factors and the shape

parameters. The proposed scheme may have wide applications in smooth surface modeling in computer graphics, in non-linear sciences, data visualization problems, and engineering design. When scaling matrices (in both the directions) are taken to be zero, the developed rational cubic spline FIS reduces to the existing traditional nonrecursive bicubic partially blended rational function. It would be interesting to study other important shape properties such as the monotonicity and convexity of the rational cubic spline FIS, which may appear elsewhere.

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. Böhm, W.: A survey of curve and surface methods in CAGD. Comput. Aided Geom. Des. **1**, 1–60 (1984)
4. Bouboulis, P., Dalla, L.: Fractal interpolation surfaces derived from fractal interpolation functions. J. Math. Anal. Appl. **336**(2), 919–936 (2007)
5. Chand, A.K.B., Kapoor, G.P.: Hidden variable bivariate fractal interpolation surfaces. Fractals **11**, 277–288 (2003)
6. Chand, A.K.B., Kapoor, G.P.: Generalized cubic spline fractal interpolation functions. SIAM J. Numer. Anal. **44**(2), 655–676 (2006)
7. Chand, A.K.B.: Natural cubic spline coalescence hidden variable fractal interpolation surfaces. Fractals **20**(2), 117–131 (2012)
8. Chand, A.K.B., Viswanathan, P.: A constructive approach to cubic Hermite fractal interpolation function and its constrained aspects. BIT Numer. Math. **53**(4), 841–865 (2013)
9. Chand, A.K.B., Vijender, N., Navascués, M.A.: Shape preservation of scientific data through rational fractal splines. Calcolo **51**, 329–362 (2013)
10. Chand, A.K.B., Vijender, N.: Positive blending Hermite rational cubic spline fractal interpolation surfaces. Calcolo **1**, 1–24 (2015)
11. Chand, A.K.B., Katiyar, S.K., Viswanathan, P.: Approximation using hidden variable fractal interpolation functions. J. Fractal Geom. **2**(1), 81–114 (2015)
12. Chand, A.K.B., Navascues, M.A., Viswanathan, P., Katiyar, S.K.: Fractal trigonometric polynomial for restricted range approximation. Fractals **24**(2), 11 (2016)
13. Casciola, G., Romani, L.: Rational interpolants with tension parameters. In: Lyche, T., Mazure, M.-L., Schumaker, L.L. (eds.) Curve and Surface Design. Modern Methods Mathematics, 41–50. Nashboro Press, Brentwood (2003).
14. Dalla, L.: Bivariate fractal interpolation functions on grids. Fractals **10**(1), 53–58 (2002)
15. Farin, G.: Curves and Surfaces for CAGD. Morgan Kaufmann, San Francisco (2002)
16. Feng, Z., Feng, Y., Yuan, Z.: Fractal interpolation surfaces with function vertical scaling factors. Appl. Math. Lett. **25**, 1896–1900 (2012)
17. Fritsch, F.N., Carlson, R.E.: Monotone piecewise cubic interpolations. SIAM J. Numer. Anal. **17**(2), 238–246 (1980)
18. Geronimo, J.S., Hardin, D.: Fractal interpolation surfaces and a related 2-D multiresolution analysis. J. Math. Anal. Appl. **176**, 561–586 (1993)

19. Geronimo, J.S., Hardin, D.: Fractal interpolation functions from R^n into R^m and their projections. Z. Anal. Anwendungen **12**, 535–548 (1993)
20. Katiyar, S.K., Chand, A.K.B., Navascués, M.A.: Hidden variable **A**-fractal functions and their monotonicity aspects. Rev. R. Acad. Cienc. Zaragoza **71**, 7–30 (2016)
21. Malysz, R.: The Minkowski dimension of the bivariate fractal interpolation surfaces. Chaos Solitons Fractals **27**, 1147–1156 (2006)
22. Massopust, P.R.: Fractal surfaces. J. Math. Anal. Appl. **151**, 275–290 (1990)
23. Metzler, W., Yun, C.H.: Construction of fractal interpolation surfaces on rectangular grids. Int. J. Bifurcat. Chaos Appl. Sci. Eng. **20**(12), 4079–4086 (2010)
24. Navascués, M.A., Sebastián, M.V.: Smooth fractal interpolation. J. Inequal. Appl. 1–20 (2006). Article ID: 78734(1)
25. Navascues, M.A., Viswanathan, P., Chand, A.K.B., Sebastian, M.V., Katiyar, S.K.: Fractal bases for Banach spaces of smooth functions. Bull. Aust. Math. Soc. **92**, 405–419 (2015)
26. Schimdt, J.W., Heß, W.: Positivity of cubic polynomial on intervals and positive spline interpolation. BIT Numer. Anal. **28**, 340–352 (1988)
27. Sarfraz, M., Hussain, M.Z.: Data visualization using rational spline interpolation. J. Comput. Appl. Math. **189**, 513–525 (2006)
28. Sarfraz, M., Hussain, M.Z., Nisar, A.: Positive data modeling using spline function. Appl. Math. Comput. **216**, 2036–2049 (2010)
29. Shaikh, T., Sarfraz, M., Hussain, M.Z.: Shape preserving constrained data visualization using rational functions. J. Prime Res. Math. **7**, 35–51 (2011)
30. Sarfraz, M., Hussain, M.Z., Hussain, M.: Shape-preserving curve interpolation. J. Comp. Math. **89**, 35–53 (2012)
31. Songil, R.: A new construction of the fractal interpolation surface. Fractals **23**, 12 (2015)
32. Viswanathan, P., Chand, A.K.B.: Fractal rational functions and their approximation properties. J. Approx. Theor. **185**, 31–50 (2014)
33. Viswanathan, P., Chand, A.K.B., Navascués, M.A.: Fractal perturbation preserving fundamental shapes: bounds on the scale factors. J. Math. Anal. Appl. **419**, 804–817 (2014)
34. Xie, H., Sun, H.: The study on bivariate fractal interpolation functions and creation of fractal interpolated surfaces. Fractals **5**, 625–634 (1997)
35. Wang, H.Y., Yu, J.S.: Fractal interpolation functions with variable parameters and their analytical properties. J. Approx. Theor. **175**, 1–18 (2013)

Electrokinetic Flow in a Surface Corrugated Microchannel

Subrata Bera[1(✉)] and S. Bhattacharyya[2]

[1] Department of Mathematics, National Institute of Technology Silchar, Silchar 788010, India
subrata.br@gmail.com
[2] Department of Mathematics, Indian Institute of Technology Khargapur, Kharagpur 721302, India

Abstract. A numerical investigation is made into the characteristics of the electrokinetic flow and its effect in the vicinity of a surface corrugated microchannel. A transformation have been used to transform the present physical domain to rectangular computational domain in order to simplify the application of boundary conditions on the channel walls. The characteristics for the electrokinetic flow are obtained by numerically solving the Laplace equation for the distribution of external electric potential; the Poisson equation for the distribution of induced electric potential; the Nernst-Planck equation for the distribution of ions and the Navier-Stokes equations for fluid flow simultaneously. These nonlinear coupled set of governing equations are solved numerically by control volume method over staggered system. Our results show that the form of the vortical flow, which develops in the vicinity of the channel wall depends on the surface roughness and thickness of the Debye layer along the homogeneous channel wall. The occurrence of electrical neutrality of fluid outside the Debye layer and recirculating vortex near channel wall suggests that the fluid flow is influenced by the induced electric field and vice-versa.

Keywords: Surface modulation · Electroosmotic flow · Electric double layer · Nernst-Planck equations

1 Introduction

The burgeoning field of microfluidics continues to expand its impact on society, finding extensive uses in areas ranging from the research laboratory to the health care industry. The micro/nano fluidic systems have a wide range of biological and chemical applications such as, drug delivery and control, rapid molecular analysis, sensing, separation and mixing, DNA manipulation and sequencing among many other applications. Most solid surfaces acquire a certain amount of electrostatic charges when they are in contact with aqueous solution. The surface charge influences the distribution of the ions within the liquid near the wall surface. As a result, positive counterions in solution have a greater affinity for the

© Springer Nature Singapore Pte Ltd. 2017
D. Giri et al. (Eds.): ICMC 2017, CCIS 655, pp. 278–289, 2017.
DOI: 10.1007/978-981-10-4642-1_24

surface, resulting in a gradient of positive ions whose concentration eventually drops to that of the bulk solution at some distance away from the wall. The negatively charged surface and the immobile positive ions adjacent to it form the Stern layer and a diffuse layer consists of a region of mobile ions. These two distinct regions form the well-known electrical double layer (EDL). The thickness of the EDL can be characterized by the Debye length, which is in the order of nanometers. Electroosmotic flow (EOF) is the bulk liquid motion that results when an externally applied electric field interacts with the net surplus of charged ions in the diffused part of an EDL (2006). When EOF is modeled in thin EDL approximation using a simple slip velocity condition known as the Helmholtz-Smoluchowski velocity (1994).

Owing to its importance, several authors investigated the various aspects of EOF in micro and nanochannels both theoretically and experimentally. In the literature, a great deal of information has been generated on EOF. Conlisk and MeFerran (2002) described a mathematical model and numerical solution for EOF due to the applied electric field in a rectangular microchannel with overlapping EDL. The EOF in micro- and nanofluidics have been studied by Wang et al. (2006) using a lattice Poisson-Boltzmann method. Erickson and Li (2003) proposed an analytical solution using Greens function for alternating current EOF through a rectangular microchannel for the case of a sinusoidal applied electric field. A mathematical model have been proposed by Qu and Li (2000) to determine electrical potential distributions and ionic concentration distributions in overlapped EDL fields between two flat plates. It may be noted that the Debye-Huckel approximation is valid only for low surface potentials (<25 mV) (Conlisk (2005)). Bera and Bhattacharyya (2013) compared the EOF between the linear model based on equilibrium Poisson-Boltzmann equations and non-linear model based on the Poisson-Nernst-Plank equations for ions. The magnitude and direction of EOF strongly depends on the magnitude and polarity of the surface charged density of the wall. This non-uniform surface potential results in difference in electrokinetic force and develops the micro-vortices which is very important to increase mixing performance of solutes, separation of ions etc. The analysis of EOF with step change in zeta-potential is studied by Fu et al. (2003). Luo (2006) investigated the two-dimensional time-dependent EOF driven by an AC electric field in micro channel with patchwise surface heterogeneities in different forms. An analytic solution for two-dimensional EOF was proposed by Horiuchi et al. (2007) in the vicinity of a step change in zeta-potential in a rectangular microchannel.

Generally, microchannel surfaces exhibit certain degrees of roughness generated by the manufacturing techniques or adhesion of biological particles from the liquids. The surface roughness of the order of few angstroms can also have a big significant factor on flow pattern. The geometric modulation of the channel wall is also created to increase the interfacial area. The surface modulation, roughness and potential heterogeneity have a great impact on flow as it disturb electro-neutrality behavior and the equilibrium EDL structure. Formation of vortices near abrupt nanochannel was numerically investigated by Ramirez and

Conlisk (2006). Chang and Yang (2004) studied the electrokinetically driven flow mixing with different patterned rectangular blocks in microchannel. A numerical model was developed by Hu et al. (2010) to simulate electroosmotic transport in microchannel with rectangular prism rough elements on the surfaces of wall. Alexander et al. (2010) studied the EOf in wavy channels by expanding the solution into a double series in terms of the dimensionless amplitudes and zeta-potential for a binary dilute electrolyte. The flow around a flow-disturbing rib located inside a rectangular microchannel was studied by Stogiannis et al. (2014) in experimentally and numerically.

One important aspect of the presence study is to consider the non-linear effects due to the present of surface modulation of the physical domain. The effect of fluid convection on ionic species distribution plays an important role in the current study. Most of the existing studies are based on the Stokes equations for the hydrodynamic flow field without considering the inertia effects. In addition the concentration distribution are based on equilibrium Boltzman distribution. The present model deals with by considering inertia effects of the full set of the Navier-Stoke equations with bodyforce terms for fluid transport. To capture the effects of convection, diffusion and electric migration of ions, the Nernst-Plank equation is considered for the ionic species distribution. There are two type of electric field present in the scenario, one is applied electric field which generated by introducing electrodes in the far upstream and downstream of the channel and another is induced electric field which developed due to the redistribution of ions near the wall. The Laplace equation for applied electric field and the Poisson equation for induced electric field will be considered in our proposed study. It may be noted that the governing equation for fluid flow, ionic concentration and potential distribution are coupled and non-linear in nature. Our aim is to solve these coupled equations using finite volume method in a staggered grid system using several upwind schemes. The influence of several important factors such as ionic concentration and surface roughness of the channel wall, on electrokinetic ion and fluid transport have been investigated thoroughly in the present study.

2 Mathematical Model

We considered a long rectangular channel of height $2h$ filled with an incompressible Newtonian electrolyte of uniform permittivity ϵ_e and viscosity μ. A schematic view of the physical domain being considered in Fig. 1(a). Because of the symmetric nature of present problem, we computed the lower half of the channel within a cycle (Fig. 1b). The external applied electric field E_0 is generated by the electrodes placed at the inlet and the outlet of the channel. The distribution of external potential ψ^* is governed by the following Laplace equation

$$\nabla^2 \psi^* = 0 \tag{1}$$

The walls are electrically insulated i.e., $\nabla \psi^* . \mathbf{n} = 0$, where $\mathbf{n}$ is the unit outward normal and far upstream and downstream, ψ^* approaches a linear function of x i.e., $\psi^* = -E_0 x^*$.

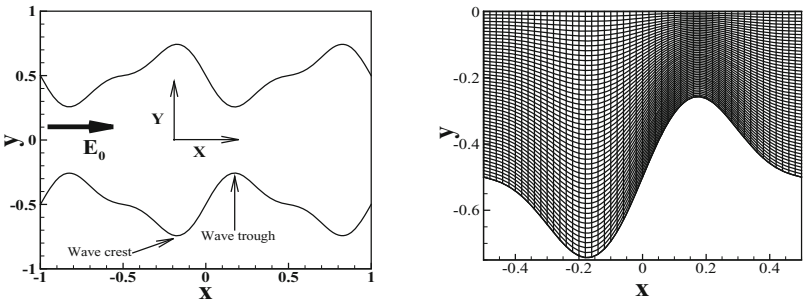


Fig. 1. (a) Schematic diagram of the surface corrugated microchannel in two cycle of wave length and (b) grid distribution of the present computational domain with wave amplitudes $\alpha_1 = 0.2$ and $\alpha_2 = 0.08$ in one cycle of wave length.

The electric field $\mathbf{E}^*$ $(= \mathbf{E}^*_x, \mathbf{E}^*_y)$ is determined by the superposition of the external electric field along with the induced electric field developed due to the redistribution of ions. The total electric potential Φ^* can be written as $\Phi^* = \psi^*(x, y) + \phi^*(x, y)$, where ϕ^* is the induced electric potential. The dimensional variables are denoted by an asterisk $(*)$. The charge density ρ^*_e is related to the electric field as

$$\nabla \cdot (\epsilon_e \mathbf{E}^*) = -\epsilon_e \nabla^2 \Phi^* = \rho^*_e = \sum_i z_i e n^*_i \tag{2}$$

Here, z_i and n^*_i are respectively, the valance and ionic concentration of the i type ion, e is the elementary electric charge and $\epsilon_e = \epsilon_0 \epsilon_r$, where ϵ_0 is the electric permittivity of vacuum and ϵ_r, the dielectric constant of the solution.

The transport equation of the ionic species i is governed by the Nernst-Planck equation as

$$\frac{\partial n^*_i}{\partial t^*} + \nabla \cdot \mathbf{N}^*_i = 0 \tag{3}$$

where $\mathbf{N}^*_i$ $(= -D_i \nabla n^*_i + n^*_i \omega_i z_i F \mathbf{E}^* + n^*_i \mathbf{q}^*)$ is the net flux of ionic species. D_i and ω_i $(= D_i/RT)$ are respectively, the diffusivity and mobility of i type species. Here, R is the gas constant and F is the Faraday's constant.

The equations for the transport of electrolyte are governed by the Navier-Stokes equations for a constant property of Newtonian fluid with an electric body force term as described by Bhattacharyya and Bera (2013)

$$\nabla \cdot \mathbf{q}^* = 0 \tag{4}$$

$$\rho \left(\frac{\partial \mathbf{q}^*}{\partial t^*} + (\mathbf{q}^* \cdot \nabla)\mathbf{q}^* \right) = -\nabla p^* + \mu \nabla^2 \mathbf{q}^* + \rho^*_e \mathbf{E} \tag{5}$$

where $\mathbf{q}^* (= u^*,\ v^*)$ is the velocity field of the fluid with u^* and v^* are the velocity components in the x and y directions, respectively. Here, ρ and μ are density and viscosity of the liquid respectively.

We scaled electric potential Φ^* by ϕ_0 $(= k_BT/e)$ and ionic concentration n_i^* by the bulk ionic concentration n_0, cartesian coordinates (x^*, y^*) by half cannel height h, the velocity field $\mathbf{q}^*$ $(= u^*,\ v^*)$ by the Helmholtz-Smoluchowski velocity $U_{HS} = \epsilon_e E_0 \phi_0/\mu$, pressure p^* is scaled by $\mu U_{HS}/h$ and time t^* by h/U_{HS}. Here, k_B is the Boltzmann constant and T is the absolute temperature of the solution. We consider a symmetric electrolyte of valance $z_i = \pm 1$. We denoted the non-dimensional concentration of cation by g and anion by f. The parameter $\kappa = [(2e^2 n_0)/(\epsilon_e k_B T)]^{1/2}$ is reciprocal of the characteristic EDL thickness (λ) and $\kappa h = h/\lambda$.

2.1 Transformation of Basic Equations

A suitable transformation is used to map the present physical domain into a rectangle domain. We have used the following coordinate transformation.

$$Y = \frac{y}{y_0(x)} \tag{6}$$

The walls of the channel are defined by a function $y_0(x)$ and is defined by

$$y_0(x) = \pm 0.5 \pm [\alpha_1 \sin(2\pi x) + \alpha_2 \sin(4\pi x)]$$

where α_1 and α_2 are the amplitudes of the two superimposed sinusoidal functions with wave crest and wave trough at $x = -0.17$ and $x = 0.17$ respectively.

The non-dimensional equations for applied electric potential (ψ), induced potential (ϕ), ionic species concentration (g, f) and flow field (u, v) in a Cartesian coordinate with origin at the midpoint of wavy centra are given by

$$\begin{aligned}
&\frac{\partial^2\psi}{\partial x^2} + Y\left[\frac{2}{y_0^2}\left(\frac{\partial y_0}{\partial x}\right)^2 - \frac{1}{y_0^2}\frac{\partial^2 y_0}{\partial x^2}\right]\frac{\partial\psi}{\partial Y} - \frac{2Y}{y_0}\frac{\partial y_0}{\partial x}\frac{\partial^2\psi}{\partial x\partial Y} \\
&\qquad + \left[\frac{Y^2}{y_0^2}\left(\frac{\partial y_0}{\partial x}\right)^2 + \frac{1}{y_0^2}\right]\frac{\partial^2\psi}{\partial Y^2} = 0 \qquad (7) \\
&\frac{\partial^2\phi}{\partial x^2} + Y\left[\frac{2}{y_0^2}\left(\frac{\partial y_0}{\partial x}\right)^2 - \frac{1}{y_0^2}\frac{\partial^2 y_0}{\partial x^2}\right]\frac{\partial\phi}{\partial Y} - \frac{2Y}{y_0}\frac{\partial y_0}{\partial x}\frac{\partial^2\phi}{\partial x\partial Y} \\
&\qquad + \left[\frac{Y^2}{y_0^2}\left(\frac{\partial y_0}{\partial x}\right)^2 + \frac{1}{y_0^2}\right]\frac{\partial^2\phi}{\partial Y^2} = -\frac{(\kappa h)^2}{2}(g - f) \qquad (8) \\
Pe\frac{\partial g}{\partial t} &+ \frac{\partial^2 g}{\partial x^2} + Y\left[\frac{2}{y_0^2}\left(\frac{\partial y_0}{\partial x}\right)^2 - \frac{1}{y_0^2}\frac{\partial^2 y_0}{\partial x^2}\right]\frac{\partial g}{\partial Y} - \frac{2Y}{y_0}\frac{\partial y_0}{\partial x}\frac{\partial^2 g}{\partial Y} \\
&\qquad + \left[\frac{Y^2}{y_0^2}\left(\frac{\partial y_0}{\partial x}\right)^2 + \frac{1}{y_0^2}\right]\frac{\partial^2 g}{\partial Y^2} = \frac{(\kappa h)^2}{2}g(g - f) \\
&\qquad + Pe\left[\frac{\partial(gu)}{\partial x} - \frac{Y}{y_0}\frac{\partial(gu)}{\partial Y} + \frac{1}{y_0}\frac{\partial(gv)}{\partial Y}\right]
\end{aligned}$$

$$+\left[\frac{\partial g}{\partial x}\frac{\partial \phi}{\partial x}+\left\{\frac{Y^2}{y_0^2}\left(\frac{\partial y_0}{\partial x}\right)^2+\frac{1}{y_0^2}\right\}\frac{\partial g}{\partial Y}\frac{\partial \phi}{\partial Y}\right]$$
$$-\frac{Y}{y_0}\frac{\partial y_0}{\partial x}\left[\frac{\partial g}{\partial x}\frac{\partial \phi}{\partial Y}+\frac{\partial g}{\partial Y}\frac{\partial \phi}{\partial x}\right]-\frac{Y}{y_0}\frac{\partial y_0}{\partial x}\left[\frac{\partial g}{\partial x}\frac{\partial \psi}{\partial Y}+\frac{\partial g}{\partial Y}\frac{\partial \psi}{\partial x}\right] \quad (9)$$

$$Pe\frac{\partial f}{\partial t}+\frac{\partial^2 f}{\partial x^2}+Y\left[\frac{2}{y_0^2}\left(\frac{\partial y_0}{\partial x}\right)^2-\frac{1}{y_0^2}\frac{\partial^2 y_0}{\partial x^2}\right]\frac{\partial f}{\partial Y}-\frac{2Y}{y_0}\frac{\partial y_0}{\partial x}\frac{\partial^2 f}{\partial Y}$$
$$+\left[\frac{Y^2}{y_0^2}\left(\frac{\partial y_0}{\partial x}\right)^2+\frac{1}{y_0^2}\right]\frac{\partial^2 f}{\partial Y^2}=-\frac{(\kappa h)^2}{2}f(g-f)$$
$$+Pe\left[\frac{\partial (fu)}{\partial x}-\frac{Y}{y_0}\frac{\partial (fu)}{\partial Y}+\frac{1}{y_0}\frac{\partial (fv)}{\partial Y}\right]$$
$$+\left[\frac{\partial f}{\partial x}\frac{\partial \phi}{\partial x}+\left\{\frac{Y^2}{y_0^2}\left(\frac{\partial y_0}{\partial x}\right)^2+\frac{1}{y_0^2}\right\}\frac{\partial f}{\partial Y}\frac{\partial \phi}{\partial Y}\right]$$
$$+\frac{Y}{y_0}\frac{\partial y_0}{\partial x}\left[\frac{\partial f}{\partial x}\frac{\partial \phi}{\partial Y}+\frac{\partial f}{\partial Y}\frac{\partial \phi}{\partial x}\right]+\frac{Y}{y_0}\frac{\partial y_0}{\partial x}\left[\frac{\partial f}{\partial x}\frac{\partial \psi}{\partial Y}+\frac{\partial f}{\partial Y}\frac{\partial \psi}{\partial x}\right] \quad (10)$$

$$\frac{\partial u}{\partial x}+\frac{Y}{y_0}\frac{\partial y_0}{\partial x}\frac{\partial u}{\partial Y}+\frac{1}{y_0}\frac{\partial v}{\partial Y}=0 \quad (11)$$

$$Re\frac{\partial u}{\partial t}+Re\left[\frac{\partial (u^2)}{\partial x}-\frac{Y}{y_0}\frac{\partial y_0}{\partial x}\frac{\partial (u^2)}{\partial Y}+\frac{1}{y_0}\frac{\partial (uv)}{\partial Y}\right]=-\left[\frac{\partial p}{\partial x}+\frac{Y}{y_0}\frac{\partial p}{\partial Y}\right]$$
$$-(g-f)\frac{(\kappa h)^2}{2}\left[\left(\frac{\partial \phi}{\partial x}-\frac{Y}{y_0}\frac{\partial y_0}{\partial x}\frac{\partial \phi}{\partial Y}\right)+\left(\frac{\partial \psi}{\partial x}-\frac{Y}{y_0}\frac{\partial y_0}{\partial x}\frac{\partial \psi}{\partial Y}\right)\right]$$
$$+\frac{\partial^2 u}{\partial X^2}+Y\left[\frac{2}{y_0^2}(\frac{\partial y_0}{\partial x})^2-\frac{1}{y_0^2}\frac{\partial^2 y_0}{\partial x^2}\right]\frac{\partial u}{\partial Y}$$
$$-\frac{2Y}{y_0}\frac{\partial y_0}{\partial x}\frac{\partial^2 u}{\partial x \partial Y}+\left[\frac{Y^2}{y_0^2}(\frac{\partial y_0}{\partial x})^2+\frac{1}{y_0^2}\right]\frac{\partial^2 u}{\partial Y^2} \quad (12)$$

$$Re\frac{\partial v}{\partial t}+Re\left[\frac{\partial (uv)}{\partial x}-\frac{Y}{y_0}\frac{\partial y_0}{\partial x}\frac{\partial (uv)}{\partial Y}+\frac{1}{y_0}\frac{\partial (v^2)}{\partial Y}\right]$$
$$=-\frac{Y}{y_0}\frac{\partial p}{\partial Y}-(g-f)\frac{(\kappa h)^2}{2}\frac{1}{y_0}\left[\frac{\partial \phi}{\partial Y}+\frac{\partial \psi}{\partial Y}\right]$$
$$+\frac{\partial^2 v}{\partial x^2}+Y\left[\frac{2}{y_0^2}(\frac{\partial y_0}{\partial x})^2-\frac{1}{y_0^2}\frac{\partial^2 y_0}{\partial x^2}\right]\frac{\partial v}{\partial Y}$$
$$-\frac{2Y}{y_0}\frac{\partial y_0}{\partial x}\frac{\partial^2 v}{\partial x \partial Y}+\left[\frac{Y^2}{y_0^2}(\frac{\partial y_0}{\partial x})^2+\frac{1}{y_0^2}\right]\frac{\partial^2 v}{\partial Y^2} \quad (13)$$

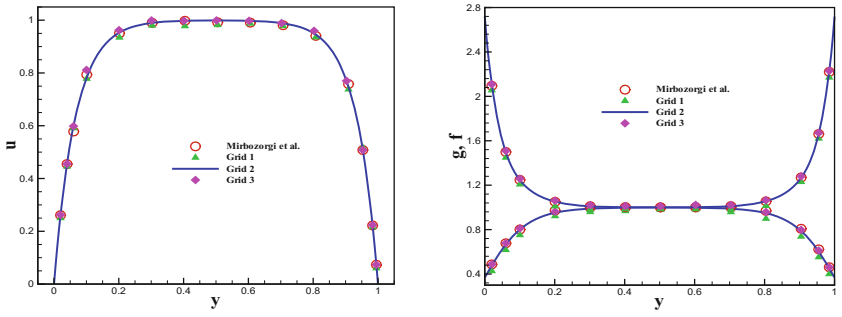
The Reynolds number based on U_{HS} is defined as $Re = U_{HS}h/\nu$, Schmidt number $Sc = \nu/D_i$, Peclet number $Pe = Re \cdot Sc$ and $\nu = \mu/\rho$, where μ is the viscosity of the electrolyte. Along the solid walls of the channel, we assumed a no-slip condition and ion impermeability i.e.,

$$u = v = 0;\ \phi = \zeta;\ \mathbf{N}_i \cdot \mathbf{n} = 0 \quad (14)$$

where **n** is the unit outward normal vector. The boundary wall bears a constant uniform surface potential ζ-potential. The flow under consideration is assumed to be axisymmetric with respect to the y-axis of symmetry. We have used the periodic boundary condition for variables at the upstream and downstream of the channel.

3 Numerical Methods

We solved the coupled set of governing non-linear equations (Eqs. 7–13) for fluid flow and ionic species concentration through a finite volume method on a staggered grid system in the transform domain. In the staggered grid system, the scalar quantities are evaluated at each cell center and the velocity components are evaluated at the midpoint of the cell sides to which they are normal. The discretized form of the governing equations are obtained by integrating the governing equations over each control volumes. Different control volumes are used to integrate different equations. We have used the second-order upwind-biased scheme, Quadratic Upwind Interpolation Convective Kinematics (QUICK) Leonard (1979), to discretize the convective and electromigration terms in both concentration and the Navier Stokes equations. The QUICK scheme uses a quadratic interpolation/ extrapolation between three nodal values to estimate the variables at the interface of the control volume. The upwind scheme imparts stability to the numerical solution in the region where a steep gradient in variables occur. An implicit first-order scheme is used for discretising the time derivative terms. The resulting discretized equations are solved iteratively through the pressure correction based iterative algorithm SIMPLE Fletcher (1991). The iteration starts by assuming the induced electric potential ϕ at every cell center. A multigrid technique may be adopted for computing elliptic type PDEs. We considered a non-uniform grid spacing along y-direction and



(a) (b)

Fig. 2. Comparison of our computed solution with Mirbozorgi et al.(2006) and the effects of grid size for fully developed EOF in a plane microchannel (*i.e.*,$y_0(x) = 1$), when channel half-height h is 10 μm, $\kappa h = 21.74$, $\zeta = -25$ mV and Re = 0.02. (a) Velocity; (b) ionic concentrations.

uniform grids along the other axis (Fig. 2 and δt was taken as 0.001. To check the effects of grid spacing, computations have been performed for three different meshes with Grid 1: 200×240, Grid 2: 400×240 and Grid 3: 400×500 for EOF in a plane microchannel for $y_0(x) = 1$, and compared with the results due to the Mirbozorgi et al. (2006). In Grid 1 and Grid 2, we considered a non-uniform grid size where δy is assumed to vary between 0.005 to 0.01 with δx is either 0.02 (for Grid 1) or $\delta x = 0.01$ (for Grid 2). In Grid 3, we considered $\delta x = 0.01$ and $0.0025 \leq \delta y \leq 0.005$. Figure 2(a) and (b) suggests that the results obtained by Grid 2 and Grid 3 agree fairly well with each other and these results are in close agreement with the result due to Mirbozorgi et al. (2006). Thus, we find that Grid 2 is optimal.

4 Results and Discussions

We have considered the half-height of the channel $h = 10\ \mu$m, viscosity $\mu = 0.001$ Kg/m s, density $\rho = 1000$ Kg/m^3, Faraday constant $F = 96,500$ C/mol and gas constant $R = 8.315$ J/mol K at temperature $T = 300$ K and the thermal voltage $\phi_0 = 0.0256$ mV. The number $\Lambda = E_0 h/\phi_0$ measures the strength of the external electric field in non-dimensional form. The external electric field is assumed to be 10^4 V/m, thus the non-dimensional parameter $\Lambda = 4.0$ when $h = 10\ \mu$m. The Reynolds number based on the Helmholtz-Smoluchowski velocity $U_{HS}(= 1.788 \times 10^{-3}$ m/s) is $Re = 1.78 \times 10^{-3}$ when $\zeta = -1$. We considered diffusion coefficient of ions are same as $D_+ = D_- = 1.3 \times 10^{-10}$ m^2/s and Schmidt number, Sc = 7692.31 and the Peclet number Pe = 13.69. The form of the applied electric fields (ψ) are presented in Fig. 3(b–c) for different value of surface roughness parameter α_1, α_2 and compared with plane channel in Fig. 3(a).

We present the streamline patterns for various values of Debye layer thickness i.e., $\kappa h = 5, 15, 60$ when $h = 10\mu$m in Fig. 4(a–c) for $\alpha_1 = 0.2$ and $\alpha_2 = 0.08$. The surface potential for upper and lower wall of the channel is considered same

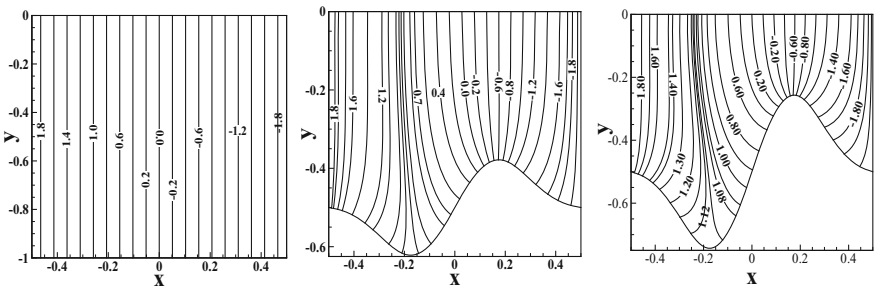


Fig. 3. Distribution of applied electric field in the surface corrugated microchannel (a) $y_0(x) = 1.0$ (plane channel), (b)$\alpha_1 = 0.1$, $\alpha_2 = 0.04$ and (c)$\alpha_1 = 0.2$, $\alpha_2 = 0.08$ when $h = 10\,\mu$m, and $E_0 = 10^4$ V/m.

as $\zeta = -1.0$. The recirculation zone appears for lower values of κh and disappear with the increase of κh. The induced pressure gradient develops due to geometric modulation and create the vortical flow in the wave crest region for low ionic concentration i.e., small κh. Vortical flow also depends on the Debye length when Debye layer thickness is in the order of the channel height. The streamlines of the liquid flow near the wall surface are distorted and a micro-vortex is generated because of the surface modulation for thin EDL.

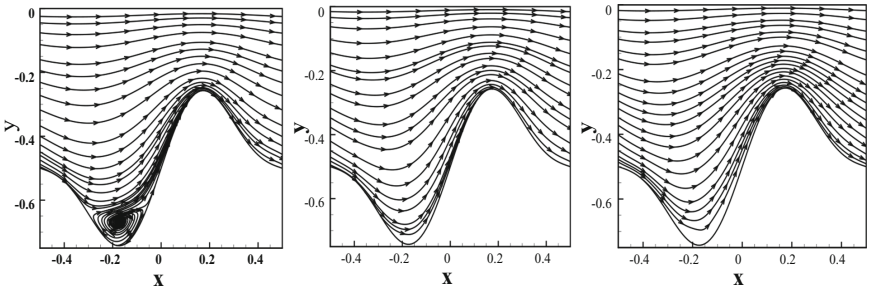
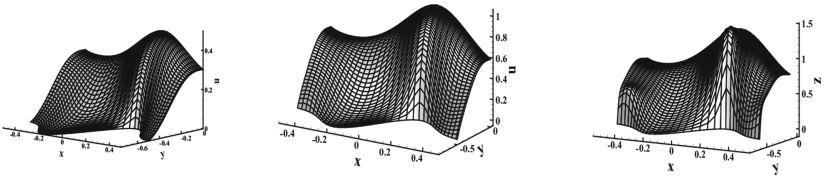


Fig. 4. Streamline profiles of EOF for different ionic concentration in the surface corrugated microchannel when $h = 10\,\mu$m, $\alpha_1 = 0.2$, $\alpha_2 = 0.08$, $\zeta = -1.0$ and $E_0 = 10^4$ V/m. (a) $\kappa h = 5$; (b) $\kappa h = 15$; and (c) $\kappa h = 60$.

The u-velocity profiles are shown in Figs. 5(a–c) for different values of ionic concentration $\kappa h = 5,\ 15,\ 60$ when the surface potential is same all over the wall and is $\zeta = -1.0$. The profiles do not resemble the classical plug-like profile for surface corrugated wall. For high ionic concentration, u -velocity is increase. An induced pressure field develops as the fluid flow rate becomes non-uniform due to the modulated surface wall. However, the flow field becomes uni-directional as it move away from the surface wall. The EOF velocity increases at a faster rate with the increase of ionic concentration for fixed channel half-height h.

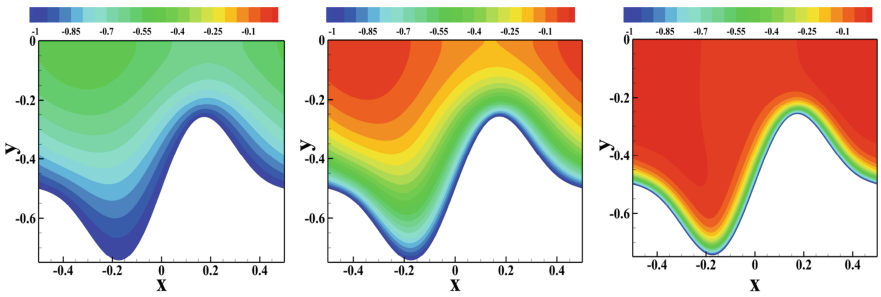
Figures 6(a–c) show the distribution of induced potential (ϕ) for different ionic concentration near the surface corrugated microchannel. The uniform surface potential ζ is applied in the wall and its value is -1 i.e., -25 mV. For low ionic concentration core region is not electro neutral. The non-uniformity of the net charge density results in nonuniform EOF velocity, which creates a pressure gradient along the primary flow direction and an induced pressure gradient develops due to the momentum loss.

The distribution of concentration profiles for cation (g) and anion (f) are shown in Fig. 7(a) for the wave crest point $x = -0.17$ and (b) the wave trough point at $x = 0.17$ for different values of κh $(= 5, 15, 60)$. The distribution of ions show that the charge density is non-zero near the corrugated wall. A non-zero charge density in the bulk region implies that the electric body force outside the Debye layer has an impact in driving the fluid motion. We find from Figs. 7(a)



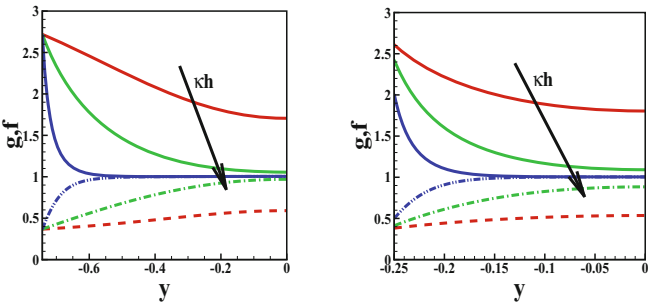
(a) (b) (c)

Fig. 5. Distribution of u-velocity (u) of EOF for different ionic concentration in the surface corrugated microchannel when $h = 10\,\mu$m, $\alpha_1 = 0.2$, $\alpha_2 = 0.08$, $\zeta = -1.0$ and $E_0 = 10^4$ V/m. (a) $\kappa h = 5$; (b) $\kappa h = 15$; and (c) $\kappa h = 60$.



(a) (b) (c)

Fig. 6. Distribution of induced potential (ϕ) of EOF for different ionic concentration in the surface corrugated microchannel when $h = 10\,\mu$m, $\alpha_1 = 0.2$, $\alpha_2 = 0.08$, $\zeta = -1.0$ and $E_0 = 10^4$ V/m. (a) $\kappa h = 5$; (b) $\kappa h = 15$; and (c) $\kappa h = 60$.



(a) (b)

Fig. 7. Distribution of cation (g) and anion (f) for different ionic concentration in the surface corrugated microchannel when $h = 10\,\mu$m, $\alpha_1 = 0.2$, $\alpha_2 = 0.08$, $\zeta = -1.0$ and $E_0 = 10^4$ V/m. (a) wave crest $x = -0.17$; and (b) wave trough $x = 0.17$. Arrow indicates the increasing direction of κh ($= 5, 15, 60$). Solid and dotted lines represent the distribution of cation (g) and anion (f) respectively.

and (b) that even for high κh, the net charge density $(g-f)$ is nonzero outside the EDL. As the electric body force on fluid flow is nonzero, it makes the governing equations for fluid flow and ion transport coupled. Besides, for lower values of κh (i.e., thick EDL), the bulk fluid is not electrically neutral and thus, the ions do not follow the Boltzmann distribution.

5 Conclusions

A numerical investigation is performed on the electroosmotic flow and its effects in the vicinity of a surface modulated microchannel. The physical modulated domain is transformed into a rectangular computational domain in order to simplify the application of boundary conditions on the channel walls. The characteristics for the electrokinetic flow are obtained by numerically solving the Laplace equation for the distribution of external electric potential; the Poisson equation for the distribution of induced electric potential; the Nernst-Planck equation for the distribution of ions and the Navier-Stokes equations for fluid flow simultaneously. These non-linear coupled set of governing equations are solved numerically by control volume method over staggered system. The recirculating vortex, which appears near the surface wall disappear and the average electroosmotic velocity increases with the increase of the electrolyte concentration. The vortical flow develops near the corrugated wall and depends on the surface roughness and Debye layer thickness. The flow field close to the wall is two dimensional. However, the streamlines shows a parallel flow faraway from the channel wall. The net charged density is not zero outside the Debye layer for surface corrugated microchannel.

References

Masiliyah, J.H., Bhattacharjee, S.: Electrokinetic and Colloid Transport Phenomena. Wiley, Hoboken (2006)

Probstein, R.F.: Physicochemical Hydrodynamics: An Introduction, 2nd edn. Wiley Interscience, New York (1994)

Conlisk, A.T., McFerran, J.: Mass transfer and flow in electrically charged micro-and nanochannels. Analytical Chemistry **74**, 2139–2150 (2002)

Wang, J., Wang, M., Li, Z.: Lattice Poisson-Boltzmann simulations of electroosmotic flows in microchannels. J. Colloid Interface Sci. **296**, 729–736 (2006)

Erickson, D., Li, D.: Analysis of alternating current electroosmotic flows in a rectangular microchannel. Langmuir **19**, 5421–5430 (2003)

Qu, W., Li, D.: A model for overlapped EDL fields. J. Colloid Interface Sci. **224**, 397–407 (2000)

Conlisk, A.T.: The Debye-Hckel approximation: its use in describing electroosmotic flow in micro- and nanochannels. Electrophoresis **26**, 1896–1912 (2005)

Bera, S., Bhattacharyya, S.: On mixed electroosmotic-pressure driven flow and mass transport in microchannels. Int. J. Eng. Sci. **62**, 165–176 (2013)

Fu, L.-M., Lin, J.-Y., Yang, R.-J.: Analysis of electroosmotic flow with step change in zeta potential. J. Colloid Interface Sci. **258**, 266–275 (2003)

Luo, W.-J.: Transient electroosmotic flow induced by AC electric field in micro-channel with patchwise surface heterogeneities. J. Colloid Interface Sci. **295**, 551–561 (2006)
Horiuchi, K., Dutta, P., Ivory, C.F.: Electroosmosis with step changes in zeta potential in microchannels. AIChE J. **53**, 2521–2533 (2007)
Ramirez, J.C., Conlisk, A.T.: Formation of vortices near abrupt nano-channel height changes in electro-osmotic flow of aqueous solutions. Biomed. Microdevices **8**, 325–330 (2006)
Chang, C.-C., Yang, R.-J.: Computational analysis of electrokinetically driven flow mixing in microchannels with patterned blocks. J. Micromech. Microeng. **14**, 550–558 (2004)
Hu, Y., Werner, C., Li, D.: Electrokinetic transport through rough microchannels. Anal. Chem. **75**, 5747–5758 (2003)
Malevich, A.E., Mityushev, V.V., Adler, P.M.: Electrokinetic phenomena in wavy channels. J. Colloid Interface Sci. **345**, 72–87 (2010)
Stogiannis, I.A., Passos, A.D., Mouza, A.A., Paras, S.V., Pênkavová, V., Tihon, J.: Flow investigation in a microchannel with a flow disturbing rib. Chem. Eng. Sci. **119**, 65–76 (2014)
Bhattacharyya, S., Bera, S.: Nonlinear electroosmosis pressure-driven flow in a wide microchannel with patchwise surface heterogeneity. J. Fluids Eng. Trans. ASME **135**, 02130 (2013)
Leonard, B.P.: Stable and accurate convective modelling procedure based on quadratic upstream interpolation. Comput. Meth. Appl. Mech. Eng. **19**, 59–98 (1979)
Fletcher, C.A.J.: Computational Techniques for Fluid Dynamics. Springer Series in Computational Physics, 2nd edn. Springer, Heidelberg (1991). vols. 1 and 2
Mirbozorgi, S.A., Niazmand, H., Renkrizbulut, M.: Electroosmotic flow in reservoir-connected flat microchannels with non-uniform zeta potential. J. Fluids Eng. Trans. ASME **128**, 1133–1143 (2006)

Pure Mathematics

Fundamental Solutions to the Laplacian in Plane Domains Bounded by Ellipses

H. Begehr(✉)

Math. Institut, FU Berlin, Arnimallee 3, 14195 Berlin, Germany
begehrh@zedat.fu-berlin.de

Abstract. Explicit harmonic Robin functions are given for the exterior of an ellipse and for a ring domain bounded by two confocal ellipses of the complex plane. The related Robin problems for the Poisson equation are explicitly solved. As the Robin functions interpolate the Green and Neumann functions the Dirichlet and Neumann problems are by the way treated.

Keywords: Robin, Green and Neumann functions · Robin boundary value problem · Plane domains bounded by ellipses · Doubly connected domain · Ring

Mathematics Subject Classifications: 31A25 · 31A30 · 35J08 · 35J25

1 Introduction

Although reflections are possible on ellipses on the plane the parqueting reflection principle does not provide harmonic Green and Neumann functions because the resulting functions fail to be meromorphic. Here the conformal invariance of the fundamental solutions to the Laplace operator, see e.g. [16], is used to construct Robin functions for the exterior of an ellipse and for a doubly connected domain bounded by two confocal ellipses.

To explain the parqueting-reflection principle the basic case of the unit disc $\mathbb{D} = \{|z| < 1\}$ is considered. A point $z \in \mathbb{D}$ is reflected at the boundary $\partial\mathbb{D} = \{|z| = 1\}$ onto $\frac{1}{\overline{z}}$. This reflection provides a covering of the complex plane $\mathbb{C} = \mathbb{D} \cup \mathbb{C} \setminus \mathbb{D}$. Choosing the elementary rational function

$$P_1(z, \zeta) = \frac{1 - \overline{z}\zeta}{\zeta - z}$$

with a simple pole at z and a simple zero at $\frac{1}{\overline{z}}$ provides the harmonic Green function for $\mathbb{D}$ in the form $G_1(z, \zeta) = \log |P_1(z, \zeta)|^2$. The harmonic Neumann function for $\mathbb{D}$ is $N_1(z, \zeta) = -\log |(\zeta - z)(1 - \overline{z}\zeta)|^2$. For convenience here twice the respective fundamental solutions are used indicated by the subscript 1, [3,4].

This principle is helpful to attain these fundamental solutions for a variety of plane domains bounded by arcs of circles and lines, see e.g. [1,5,12,17–19].

© Springer Nature Singapore Pte Ltd. 2017
D. Giri et al. (Eds.): ICMC 2017, CCIS 655, pp. 293–311, 2017.
DOI: 10.1007/978-981-10-4642-1_25

But there are plane domains for which the parqueting-reflection principle applies, not providing Green or Neumann functions. An example for such a domain is the one bounded by the ellipse [6]

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1 \quad with \quad 0<a,b,$$

in complex form given as

$$|Az-B\overline{z}|^2=1, \quad where \quad A+B=\frac{1}{b}, \quad B-A=\frac{1}{a}.$$

The reflection of a point $z, |Az-B\overline{z}|^2<1$, inside the ellipse at the ellipse is z_r, defined by $zz_r=z_1^2$, where $z_1=\frac{z}{|Az-B\overline{z}|}$ is the intersection of the ray from 0 towards z and the ellipse. Obviously, $1<|Az_r-B\overline{z_r}|^2$. This provides a parqueting of $\mathbb{C}$, see [9]. But $\widetilde{G}_1(z,\zeta)=\log\left|\frac{\zeta-z_r}{\zeta-z}\right|^2$ with $z_r=\frac{z}{|Az-B\overline{z}|^2}$ fails to be harmonic in the variable z. In order to obtain the harmonic Green and the harmonic Neumann function for the exterior domain of the ellipse the conformal map onto the unit disc is used, see [20], exercise 264(2), p. 37 and p. 297.

Using a conformal map from a ring domain bounded by two confocal ellipses onto a concentric ring domain and using the Green and Neumann functions for the latter ring [17–19], these functions can be obtained for the elliptic ring domain.

Instead, the conformal invariance is used for a (modified) harmonic Robin function to create a (modified) Robin function for the elliptic ring. The modification is required for doubly connected domains in order to achieve a Robin function interpolating the Green and Neumann functions. The modification consists in replacing the outward normal derivative on the inner boundary part by the inner normal derivative, i.e. just by changing the sign in the boundary behavior of the Robin function at one boundary part, see [17]. For particular choices of the parameters involved the Green and some Neumann functions are included. The Robin function serves to obtain a representation formula for proper functions leading to explicit solutions to related Robin boundary value problems and if required to solvabilty conditions for the Poisson equation. Particularly the Dirichlet and Neumann problems are treated. For results on the Robin problem for certain first and second order equations see e.g. [2,6–8,10,11,13–15].

2 Robin Function for the Exterior of an Ellipse

An ellipse is the zero set of the function

$$E(z)=\frac{x^2}{a^2}+\frac{y^2}{b^2}-1, a,b\in\mathbb{R},$$

of two real variables x, y. Assuming $a>b>0, a^2-b^2=1$ and introducing

$$r=a+b, \frac{1}{r}=a-b, \ \ i.e. \ \ 2a=r+\frac{1}{r}, 2b=r-\frac{1}{r}, 2(a^2+b^2)=r^2+\frac{1}{r^2}, a^2-b^2=1$$

and the complex variables

$$z = x + iy, \overline{z} = x - iy, \quad i.e. \ 2x = z + \overline{z}, 2iy = z - \overline{z}$$

this function is expressible as

$$E = \frac{(z+\overline{z})^2}{(r+\frac{1}{r})^2} - \frac{(z-\overline{z})^2}{(r-\frac{1}{r})^2} - 1.$$

According to [20], exercise 264(2), the function

$$\omega = \omega(z) = \frac{1}{r}(z \pm \sqrt{z^2-1}), \sqrt{1} = +1,$$

maps the exterior E_e of the ellipse $E = 0$, i.e. the set $0 < E$ onto a disc or the complement of a disc, dependent on the branch of the square root and the size of r. In more detail the next lemma holds.

Lemma 1. For $1 < r$ the function ω maps the outside of the ellipse E onto the disc $|\omega| < \frac{1}{r^2}$ or onto $1 < |\omega|$. If $r < 1$ then the image of $E_e = \{0 < E\}$ is either the unit disc $|\omega| < 1$ or the set $\frac{1}{r^2} < |\omega|$.

Proof. Independently of the choice of the branch of the square root the inverse mapping to ω is

$$z = \frac{1}{2}\big(r\omega + \frac{1}{r\omega}\big).$$

The choice of the square root branch and the size of the parameter r influence the chape of the image set. The function

$$\omega_1(z) = \frac{1}{r}(z - \sqrt{z^2-1}),$$

maps E_e onto a bounded domain because

$$\lim_{z\to\infty} r\omega_1(z) = \lim_{z\to\infty} z\big[1 - \sqrt{1-\frac{1}{z^2}}\big] = 0,$$

while

$$\omega_2(z) = \frac{1}{r}(z + \sqrt{z^2-1})$$

maps E_e onto an unbounded domain as obviously

$$\lim_{z\to\infty} r\omega_2(z) = \infty.$$

Inserting the inverse map the ellipse equation $E = 0$ can be rewritten as $E = 0$,

$$\begin{aligned} E &= \frac{(\omega+\overline{\omega})^2}{4}\Big(\frac{r+\frac{1}{r|\omega|^2}}{r+\frac{1}{r}}\Big)^2 - \frac{(\omega-\overline{\omega})^2}{4}\Big(\frac{r-\frac{1}{r|\omega|^2}}{r-\frac{1}{r}}\Big)^2 - 1 \\ &= u^2\big(1+\frac{\frac{1}{|\omega|^2}-1}{r^2+1}\big)^2 + v^2\big(1-\frac{\frac{1}{|\omega|^2}-1}{r^2-1}\big)^2 - 1 \\ &= \frac{(1-|\omega|^2)(1-r^4|\omega|^2)}{(a^2+b^2)|\omega|^4}\big(\frac{u^2}{4a^2}+\frac{v^2}{4b^2}\big), \end{aligned}$$

where $\omega = u + iv$.

For $0 < E$ obviously $0 < (1 - |\omega|^2)(1 - r^4|\omega|^2)$.

Case (i) $1 < r$. Then $1 - r^4|\omega|^2 < 1 - |\omega|^2$. If $0 < 1 - r^4|\omega|^2$ then $|\omega| < \frac{1}{r^2}$. But when $\frac{1}{r^2} < |\omega| < 1$ then $1 - r^4|\omega|^2 < 0 < 1 - |\omega|^2$, i.e. $E < 0$.

If $1 - |\omega|^2 < 0$ then $0 < E$, while for $\frac{1}{r^2} < |\omega| < 1$ the inequalities $1 - r^4|\omega|^2 < 0 < 1 - |\omega|^2$ and thus $E < 0$ hold. Hence, either ω maps E onto $|\omega| < \frac{1}{r^2}$ or onto $1 < |\omega|$.

Case (ii) $r < 1$. Then $1 - |\omega|^2 < 1 - r^4|\omega|^2$. If $0 < 1 - |\omega|^2$ then $0 < E$. But when $1 < |\omega| < \frac{1}{r^2}$ then $1 - |\omega|^2 < 0 < 1 - r^4|\omega|^2$, i.e. $E < 0$.

If however, $1 - r^4|\omega|^2 < 0$ this means $\frac{1}{r^2} < |\omega|$. But for $1 < |\omega| < \frac{1}{r^2}$ the relations $1 - |\omega|^2 < 0 < 1 - r^4|\omega|^2$ imply as above $E < 0$. Thus ω maps E either onto $|\omega| < 1$ or onto $\frac{1}{r^2} < |\omega|$. □

The tangent to E at the point $z = x + iy$ is determined by $y' = -\frac{x}{a^2}\frac{b^2}{y}$ with $y' = \frac{dy}{dx}$. The outward normal vector at $z_0 = x_0 + iy_0 \in E$ is described by $y - y_0 = \frac{y_0}{x_0}\frac{a^2}{b^2}(x - x_0)$. Thus the outward normal direction on the boundary of the outer domain $0 < E$ of the ellipse E at $z \in E$ is $(\nu_1, \nu_2) = -\frac{(xb^2, ya^2)}{\sqrt{x^2b^4 + y^2a^4}}$. The complex form $\nu = \nu_1 + i\nu_2$ is used to describe the outward normal unit vector on E.

Assumption. The function ω maps the ellipse E onto the unit circle $\partial\mathbb{D}$.

Because of

$$2(xb^2 + iya^2) = (a^2 + b^2)z + (b^2 - a^2)\overline{z} = r\omega - \frac{\overline{\omega}}{r}$$

then

$$\nu = -\frac{r\omega - \frac{\overline{\omega}}{r}}{|r\omega - \frac{\overline{\omega}}{r}|}.$$

Later on

$$|\big(r\omega - \frac{\overline{\omega}}{r}\big)|^2 = 2(a^2 + b^2) - \omega^2 - \overline{\omega}^2$$

will be used.

The outward normal derivative $\partial_\nu = \nu\partial_z + \overline{\nu}\partial_{\overline{z}}$ applied to real functions is just

$$\partial_\nu = 2\mathrm{Re}(\nu\partial_z) = 2\mathrm{Re}(\nu\omega'\partial_\omega).$$

Observing here $|\omega| = 1$ from

$$\sqrt{z^2 - 1} = z - r\omega = -\frac{1}{2}\big(r\omega - \frac{\overline{\omega}}{r}\big)$$

follows

$$\omega' = \omega'(z) = -\frac{\omega(z)}{\sqrt{z^2-1}} = \frac{\omega}{\frac{1}{2}\big(r\omega - \frac{\overline{\omega}}{r}\big)},$$

so that

$$\partial_\nu = -2\mathrm{Re}\Big(\nu \frac{\omega}{\frac{1}{2}\big(r\omega - \frac{\overline{\omega}}{r}\big)}\partial_\omega\Big).$$

For the ellipse E the arc length element ds is calculated as

$$ds^2 = \frac{x^2b^4 + y^2a^4}{y^2a^4}dx^2 = \frac{a^2(\omega-\overline{\omega})^2 - b^2(\omega+\overline{\omega})^2}{a^2(\omega-\overline{\omega})^2}dx^2,$$

where from $\frac{d\omega}{\omega} = -\frac{d\overline{\omega}}{\overline{\omega}}$ (for any circle with the origin as center), the relation

$$2dx = dz + d\overline{z} = a(\omega - \overline{\omega})\frac{d\omega}{\omega} = -2av\frac{d\omega}{i\omega}, \;\; \omega = u + iv,$$

and hence

$$ds^2 = [a^2(\omega-\overline{\omega})^2 - b^2(\omega+\overline{\omega})^2]\frac{1}{4}\big(\frac{d\omega}{\omega}\big)^2 = [\omega^2 + \overline{\omega}^2 - 2(a^2+b^2)]\frac{1}{4}\big(\frac{d\omega}{\omega}\big)^2$$

follow.

Taking

$$ds = \frac{1}{2}\sqrt{\omega^2 + \overline{\omega}^2 - 2(a^2+b^2)}\frac{d\omega}{\omega} = \frac{i}{2}|r\omega - \frac{\overline{\omega}}{r}|\frac{d\omega}{\omega},$$

one has for real functions

$$\partial_\nu ds = 2\mathrm{Re}(\omega\partial_\omega)\frac{d\omega}{i\omega}.$$

Remarks. If the image of the ellipse is the outside of the unit disc rather than the disc itself then because of $r\omega = z + \sqrt{z^2-1}$ there is a change of sign for ds.

For the case of mappings of E onto the circle $|\omega| = \frac{1}{r^2}$ similarly

$$\partial_\nu ds = 2\mathrm{Re}(\omega\partial_\omega)\frac{d\omega}{i\omega}.$$

As the Green function as well for the unit disc $\mathbb{D}$ as for its complement $\mathbb{C}\backslash\overline{\mathbb{D}}$ is

$$\hat{G}_1(z,\zeta) = \log|\frac{1-z\overline{\zeta}}{\zeta - z}|^2,$$

see e.g. [3,4], by the conformal invariance, see e.g. [16], the Green function for the outside of E is

$$G_1(z,\zeta) = \log|\frac{1-\omega(z)\overline{\omega(\zeta)}}{\omega(\zeta)-\omega(z)}|^2.$$

The Neumann function is

$$N_1(z,\zeta) = -\log|(\omega(\zeta)-\omega(z))(1-\omega(z)\overline{\omega(\zeta)})|^2.$$

The Poisson kernel $g_1(z,\zeta) = -\frac{1}{2}\partial_{\nu_z} G_1(z,\zeta)$ is expressed as

$$\begin{aligned} g_1(z,\zeta)ds_z &= -\Big[\frac{\omega(\zeta)}{\omega(\zeta)-\omega(z)} + \frac{\overline{\omega(\zeta)}}{\overline{\omega(\zeta)-\omega(z)}} - 1\Big]\frac{d\omega(z)}{i\omega(z)} \\ &= -\Big[\frac{\omega(\zeta)}{\omega(\zeta)-\omega(z)} + \frac{\overline{\omega(\zeta)}}{\overline{\omega(\zeta)-\omega(z)}} - 1\Big]\sigma(z)ds_z, \end{aligned}$$

while

$$\begin{aligned} \partial_{\nu_z} N_1(z,\zeta)ds_z &= 2\mathrm{Re}\Big[\frac{\omega(\zeta)}{\omega(\zeta)-\omega(z)} - \frac{\overline{\omega(\zeta)}}{\overline{\omega(\zeta)-\omega(z)}} - 1\Big]\frac{d\omega}{i\omega} \\ &= -2\frac{d\omega(z)}{i\omega(z)} = -2\sigma(z)ds_z, \quad \sigma(z) = \frac{-2}{\big|r\omega(z) - \overline{\frac{\omega(\zeta)}{r}}\big|}. \end{aligned}$$

From the Robin function for $\mathbb{D}$ or for $\mathbb{C}\setminus\overline{\mathbb{D}}$

$$\hat{R}_{1;\alpha,\beta}(z,\zeta) = \log\Big|\frac{1-z\overline{\zeta}}{\zeta-z}\Big|^2 + 2\beta\sum_{k=1}^{\infty}\frac{2\mathrm{Re}(z\overline{\zeta})^k}{\alpha+k\beta}, \quad \alpha,\beta\in\mathbb{R}, 0<\alpha^2+\beta^2, -\frac{\alpha}{\beta}\notin\mathbb{N}$$

see [11], the Robin function for the outside of the ellipse E is

$$R_{1;\alpha,\beta}(z,\zeta) = G_1(z,\zeta) + 2\beta\sum_{k=1}^{\infty}\frac{2\mathrm{Re}(\omega(z)\overline{\omega(\zeta)})^k}{\alpha+k\beta}.$$

Obviously $R_{1;\alpha,0} = G_1$ and because of

$$4\mathrm{Re}\sum_{k=1}^{\infty}\frac{(\omega(z)\overline{\omega(\zeta)})^k}{k} = -2\log|1-\omega(z)\overline{\omega(\zeta)}|^2$$

the relation $R_{1;0,\beta} = N_1$ holds. Hence the Robin function provides an interpolation of the Green and the Neumann function. For its boundary behavior

$$\begin{aligned} \partial_{\nu_z} R_{1;\alpha,\beta}(z,\zeta) &= -\frac{4}{|r\omega_1 - \overline{\frac{\omega_1}{r}}|}\mathrm{Re}\Big[\frac{\omega_1(\zeta)}{\omega_1(\zeta)-\omega_1(z)} - \frac{\omega_1(z)\overline{\omega_1(\zeta)}}{1-\omega_1(z)\overline{\omega_1(\zeta)}} - 1 \\ &\quad + 2\beta\sum_{k=1}^{\infty}\frac{k}{\alpha+k\beta}(\omega_1(z)\overline{\omega_1(\zeta)})^k\Big] \end{aligned}$$

implies

$$\begin{aligned}
&\alpha R_{1;\alpha,\beta}(z,\zeta) - \frac{\beta}{2}|r\omega_1 - \frac{\overline{\omega_1}}{r}|\partial_{\nu_z} R_{1;\alpha,\beta}(z,\zeta) \\
&= \alpha G_1(z,\zeta) + 4\beta \sum_{k=1}^{\infty} \frac{\alpha}{\alpha+k\beta} \mathrm{Re}(\omega_1(z)\overline{\omega_1(\zeta)})^k \\
&\quad + 2\mathrm{Re}\beta\big[\frac{\omega_1(\zeta)}{\omega_1(\zeta)-\omega_1(z)} - \frac{\omega_1(z)\overline{\omega_1(\zeta)}}{1-\omega_1(z)\overline{\omega_1(\zeta)}} \\
&\quad - 1 + 2\beta \sum_{k=1}^{\infty} \frac{k}{\alpha+k\beta} (\omega_1(z)\overline{\omega_1(\zeta)})^k\big].
\end{aligned}$$

Here ω_1 is used as $\mathbb{D}$ is assumed to be the target the outside of E is mapped to. Thus for $z \in E$

$$\begin{aligned}
&\alpha R_{1;\alpha,\beta}(z,\zeta) - \frac{\beta}{2}|r\omega_1 - \frac{\overline{\omega_1}}{r}|\partial_{\nu_z} R_{1;\alpha,\beta}(z,\zeta) \\
&= -2\beta\mathrm{Re}\big[2\frac{\omega_1(z)\overline{\omega_1(\zeta)}}{1-\omega_1(z)\overline{\omega_1(\zeta)}} + 1 - 2\sum_{k=1}^{\infty} (\omega_1(z)\overline{\omega_1(\zeta)})^k\big] = -2\beta.
\end{aligned}$$

Therefore with $\sigma(z) = \frac{-2}{\left|r\omega(z) - \frac{\overline{\omega(z)}}{r}\right|}$ on E

$$\begin{aligned}
&\Big[\alpha R_{1;\alpha,\beta}(z,\zeta) - \frac{\beta}{2}|r\omega_1 - \frac{\overline{\omega_1}}{r}|\partial_{\nu_z} R_{1;\alpha,\beta}(z,\zeta)\Big]\frac{d\omega}{i\omega} \\
&= \Big[\alpha\sigma(z) R_{1;\alpha,\beta}(z,\zeta) + \beta\partial_{\nu_z} R_{1;\alpha,\beta}(z,\zeta)\Big] ds_z = -2\beta\sigma(z) ds_z.
\end{aligned}$$

For $\zeta \in E$, i.e. $|\omega_1(\zeta)| = 1$, the relation

$$\begin{aligned}
&\alpha R_{1;\alpha,\beta}(z,\zeta) - \frac{\beta}{2}|r\omega_1 - \frac{\overline{\omega_1}}{r}|\partial_{\nu_z} R_{1;\alpha,\beta}(z,\zeta) \\
&= 2\mathrm{Re}\beta\big[\frac{\omega_1(\zeta)}{\omega_1(\zeta)-\omega_1(z)} + \frac{\omega_1(z)\overline{\omega_1(\zeta)}}{1-\omega_1(z)\overline{\omega_1(\zeta)}} - 1\big] \\
&= 2\beta\big[\frac{\omega_1(\zeta)}{\omega_1(\zeta)-\omega_1(z)} + \frac{\overline{\omega_1(\zeta)}}{\overline{\omega_1(\zeta)-\omega_1(z)}} - 2\big]
\end{aligned}$$

holds.

Theorem 1. The Robin function $R_{1;\alpha,\beta}$ is satisfying for any $\zeta \in E_e$

- $R_{1;\alpha,\beta}(\,\cdot\,,\zeta)$ is harmonic in $E_e \setminus \{\zeta\}$ and continuously differentiable in $\overline{E_e} \setminus \{\zeta\}$,
- $h(z,\zeta) = R_{1;\alpha,\beta}(z,\zeta) + \log|\zeta - z|^2$ is harmonic for $z \in E_e$,
- $\alpha\sigma(z) R_{1;\alpha,\beta}(z,\zeta) + \beta\partial_{\nu_z} R_{1;\alpha,\beta}(z,\zeta) = -2\beta\sigma(z)$ for $z \in E$, where the density function $\sigma(z) = \frac{-2}{\left|r\omega(z) - \frac{\overline{\omega(z)}}{r}\right|}$ on E has finite mass $\int\limits_E \sigma(z) ds_z = 2\pi$,

- $\beta \int\limits_E \sigma(z) R_{1;\alpha,\beta}(z,\zeta) ds_z = 0$ (normalization condition).

It has the further property

- $R_{1;\alpha,\beta}(z,\zeta) = R_{1;\alpha,\beta}(\zeta,z), \quad z,\zeta \in D, \quad z \neq \zeta$ (symmetry).

As a fundamental solution to the Poisson equation the Robin function provides a basic representation formula. For the unit disc $\mathbb{D}$ it can be found in [11].

Theorem. Any function $w \in C^2(\mathbb{D};\mathbb{C}) \cap C^1(\overline{\mathbb{D}};\mathbb{C})$ can be represented as

$$\begin{aligned} w(\omega_1) = &- \frac{1}{4\pi} \int\limits_{\partial\mathbb{D}} \{w(\omega)\partial_{\nu_\omega} \hat{R}_{1;\alpha,\beta}(\omega_1,\omega) - \partial_\nu w(\omega) \hat{R}_{1;\alpha,\beta}(\omega_1,\omega)\} ds_\omega \\ &- \frac{1}{\pi} \int\limits_{\mathbb{D}} \partial_\omega \partial_{\overline{\omega}} w(\omega) \hat{R}_{1;\alpha,\beta}(\omega_1,\omega) du dv, \quad \omega = u + iv, \quad \omega_1 \in \mathbb{D}, \end{aligned}$$

where $\hat{R}_{1;\alpha,\beta}$ denotes the Robin function for $\mathbb{D}$.

This representation will be transformed by introducing $r\omega_1 = r\omega_1(z) = z - \sqrt{z^2-1}, r\omega = r\omega(\zeta) = \zeta - \sqrt{\zeta^2-1}$. From

$$\begin{aligned} \partial_{\nu_\omega} &= 2\mathrm{Re}\, \omega\partial_\omega = 2\mathrm{Re}\frac{\omega}{\omega'}\partial_\zeta = \mathrm{Re}\big(r\omega - \frac{\overline{\omega}}{r}\big)\partial_\zeta, \\ ds_\omega &= \frac{d\omega}{i\omega} = -\frac{2}{|r\omega - \frac{\overline{\omega}}{r}|} ds_\zeta \end{aligned}$$

follows

$$\partial_{\nu_\omega} ds_\omega = -2\mathrm{Re}\frac{r\omega - \frac{\overline{\omega}}{r}}{|r\omega - \frac{\overline{\omega}}{r}|}\partial_\zeta ds_\zeta = \partial_{\nu_\zeta} ds_\zeta.$$

Also

$$\partial_\omega \partial_{\overline{\omega}} = \frac{1}{|\omega'|^2}\partial_\zeta \partial_{\overline{\zeta}}, \quad du dv = |\omega'|^2 d\xi d\eta.$$

Hence,

$$\begin{aligned} w(\omega_1(z)) = &- \frac{1}{4\pi} \int\limits_E \{w(\omega(\zeta))\partial_{\nu_\zeta} \hat{R}_{1;\alpha,\beta}(\omega_1(z),\omega(\zeta)) \\ &- \partial_{\nu_\zeta} w(\omega(\zeta)) \hat{R}_{1;\alpha,\beta}(\omega_1(z),\omega(\zeta))\} ds_\zeta \\ &- \frac{1}{\pi} \int\limits_{E_e} \partial_\zeta \partial_{\overline{\zeta}} w(\omega(\zeta)) \hat{R}_{1;\alpha,\beta}(\omega_1(z),\omega(\zeta)) d\xi d\eta, \quad z \in E_e. \end{aligned}$$

Therefore the following result is established.

Theorem 2. Any function $w \in C^2(E_e;\mathbb{C}) \cap C^1(E \cup E_e;\mathbb{C})$ properly decaying at ∞ can be represented as

$$w(z) = -\frac{1}{4\pi}\int\limits_E \{w(\zeta)\partial_{\nu_\zeta} R_{1;\alpha,\beta}(\zeta,z) - \partial_{\nu_\zeta} w(\zeta) R_{1;\alpha,\beta}(\zeta,z)\} ds_\zeta$$
$$-\frac{1}{\pi}\int\limits_{E_e} \partial_\zeta \partial_{\overline{\zeta}} w(\zeta) R_{1;\alpha,\beta}(z,\zeta) d\xi d\eta, \quad \zeta = \xi + i\eta.$$

Applying the boundary property of the Robin function two further representation formulas follow, proper for solving a certain Robin boundary value problem for the Poisson equation in E_e.

Corollary 1. Any function $w \in C^2(E_e;\mathbb{C}) \cap C^1(E \cup E_e;\mathbb{C})$ properly decaying at ∞ can be represented as

$$w(z) = -\frac{1}{4\pi\alpha}\int\limits_E [\alpha\sigma(\zeta)w(\zeta) + \beta\partial_\nu w(\zeta)]\frac{1}{\sigma(\zeta)}\partial_{\nu_\zeta} R_{1;\alpha,\beta}(z,\zeta) ds_\zeta$$
$$-\frac{\beta}{2\pi\alpha}\int\limits_E \partial_\nu w(\zeta) ds_\zeta$$
$$-\frac{1}{\pi}\int\limits_{E_e} \partial_\zeta \partial_{\overline{\zeta}} w(\zeta) R_{1;\alpha,\beta}(z,\zeta) d\xi d\eta, \quad \text{for } \alpha \neq 0,$$
$$w(z) = \frac{1}{4\pi\beta}\int\limits_E [\alpha\sigma(\zeta)w(\zeta) + \beta\partial_\nu w(\zeta)] R_{1;\alpha,\beta}(z,\zeta) ds_\zeta + \frac{1}{2\pi}\int\limits_E \sigma(\zeta) w(\zeta) ds_\zeta$$
$$-\frac{1}{\pi}\int\limits_{E_e} \partial_\zeta \partial_{\overline{\zeta}} w(\zeta) R_{1;\alpha,\beta}(z,\zeta) d\xi d\eta, \quad \text{for } \beta \neq 0.$$

Robin Problem. Find a solution to the Poisson equation

$$w_{z\overline{z}} = f \ \text{ in } \ E_e, \quad f \in L_{p,2}(E_e;\mathbb{C}), \quad 2 < p,$$

satisfying

$$\alpha\sigma w + \beta\partial_\nu w = \gamma \ \text{ on } \ E, \quad \alpha,\beta \in \mathbb{R}, \quad 0 < \alpha^2 + \beta^2, \quad \sigma \ \text{ as in Theorem 1}, \quad \gamma \in C(E;\mathbb{C}).$$

For the definition of the $L_{p,2}$-spaces see e.g. [3]. In analogy to the case of the unit disc, [11], the Robin problem can be solved via the next results.

Theorem 3. For $f \in L_{p,2}(E_e;\mathbb{C}), 2 < p, \ \gamma \in C(E;\mathbb{C})$, the Robin problem

$$\partial_z \partial_{\overline{z}} w = f \text{ in } E_e, \ \alpha\sigma w + \beta\partial_\nu w = \gamma \text{ on } E,$$

(i) for $\beta \neq 0$ is solvable if and only if

$$\frac{1}{2\pi}\int\limits_E \gamma(\zeta)ds_\zeta + \frac{\alpha}{2\pi}\int\limits_E \sigma(\zeta)w(\zeta)ds_\zeta = \frac{2\beta}{\pi}\int\limits_{E_e} f(\zeta)d\xi d\eta,$$

the solution being then

$$\begin{aligned} w(z) =& \frac{1}{4\pi\beta}\int\limits_E \gamma(\zeta)R_{1;\alpha,\beta}(z,\zeta)ds_\zeta + \frac{1}{2\pi}\int\limits_E \sigma(\zeta)w(\zeta)ds_\zeta \\ & - \frac{1}{\pi}\int\limits_{E_e} f(\zeta)R_{1;\alpha,\beta}(z,\zeta)d\xi d\eta, \end{aligned}$$

(ii) for $\alpha \neq 0$ is solvable if and only if

$$\frac{\beta}{2\pi}\int\limits_E \partial_\nu w(\zeta)ds_\zeta = \frac{2\beta}{\pi}\int\limits_{E_e} f(\zeta)d\xi d\eta,$$

the solution then being

$$\begin{aligned} w(z) =& -\frac{1}{4\pi\alpha}\int\limits_E \frac{\gamma(\zeta)}{\sigma(\zeta)}\partial_{\nu_\zeta} R_{1;\alpha,\beta}(\zeta,z)ds_\zeta - \frac{\beta}{2\pi\alpha}\int\limits_E \partial_\nu w(\zeta)ds_\zeta \\ & - \frac{1}{\pi}\int\limits_{E_e} f(\zeta)R_{1;\alpha,\beta}(\zeta,z)d\xi d\eta. \end{aligned}$$

Remark. If $\alpha = 0$ then $\gamma = \beta\partial_\nu w$ and the solvability condition is the known one for the Neumann problem, see e.g. [4]. In case of $\beta = 0$ there is no solvability condition!

The proof for Theorem 3 follows by direct verification on basis of the Poisson kernel for the unit disc.

3 Robin Function for a Ring Domain Between Two Confocal ellipses

The two confocal ellipses

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1, \frac{x^2}{a^2+k^2} + \frac{y^2}{b^2+k^2} = 1 \quad with \quad 0 < a, b, k \in \mathbb{R}$$

bound a doubly connected domain

$$D = \Big\{ \frac{x^2}{a^2+k^2} + \frac{y^2}{b^2+k^2} - 1 < 0 < \frac{x^2}{a^2} + \frac{y^2}{b^2} - 1 \Big\}.$$

According to [20], exercise 265, p. 37 and p. 297 the function

$$\tilde{\omega}(z) = A(z + \sqrt{z^2 - (a^2 - b^2)}), \quad A \in \mathbb{C} \text{ arbitrary},$$

maps D onto the concentric circular ring

$$\{r = a + b < |\tilde{\omega}| < \sqrt{a^2 + k^2} + \sqrt{b^2 + k^2} = \rho\}.$$

For simplicity here $A = 1$ is chosen, $a^2 - b^2 = 1$ assumed, and $\tilde{\omega}(z) = \rho\ \omega(z)$ introduced. Then ω maps D onto the ring

$$\mathcal{R} = \big\{\frac{r}{\rho} < |\omega| < 1\big\}.$$

The modulus of these two doubly connected domains is

$$\mu = \frac{\rho}{r} = \frac{a - b}{\sqrt{a^2 + k^2} - \sqrt{b^2 + k^2}}.$$

Demanding

$$r = a + b,\ \ \frac{1}{r} = a - b;\ \ \rho = \sqrt{a^2 + k^2} + \sqrt{b^2 + k^2},\ \ \frac{1}{\rho} = \sqrt{a^2 + k^2} - \sqrt{b^2 + k^2}$$

implies

$$2a = r + \frac{1}{r},\ \ 2b = r - \frac{1}{r};\ \ 2\sqrt{a^2 + k^2} = \rho + \frac{1}{\rho},\ \ 2\sqrt{b^2 + k^2} = \rho - \frac{1}{\rho};\ \ a^2 - b^2 = 1.$$

The inverse mapping of $\rho\omega(z) = z + \sqrt{z^2 - 1}$ is $z = \frac{1}{2}\big(\rho\omega + \frac{1}{\rho\omega}\big)$, mapping $|\omega| = \frac{r}{\rho}$ and $|\omega| = 1$ onto the ellipses

$$\begin{aligned} E_i &= \frac{x^2}{a^2} + \frac{y^2}{b^2} - 1 = \frac{\big(\rho + \frac{1}{\rho|\omega|^2}\big)^2}{\big(r + \frac{1}{r}\big)^2} + \frac{\big(\rho - \frac{1}{\rho|\omega|^2}\big)^2}{\big(r - \frac{1}{r}\big)^2} - 1 \\ &= \frac{(1 - r^2\rho^2|\omega|^2)(r^2 - \rho^2|\omega|^2)}{4r^2\rho^2|\omega|^4}\big(\frac{u^2}{a^2} + \frac{v^2}{b^2}\big) = 0 \end{aligned}$$

and

$$\begin{aligned} E_e &= \frac{x^2}{a^2 + k^2} + \frac{y^2}{b^2 + k^2} - 1 = \frac{\big(\rho + \frac{1}{\rho|\omega|^2}\big)^2}{\big(\rho + \frac{1}{\rho}\big)^2} + \frac{\big(\rho - \frac{1}{\rho|\omega|^2}\big)^2}{\big(\rho - \frac{1}{\rho}\big)^2} - 1 \\ &= \frac{(1 - |\omega|^2)(r^2 - |\omega|^4)}{4\rho^4|\omega|^4}\big(\frac{u^2}{a^2 + k^2} + \frac{v^2}{b^2 + k^2}\big) = 0, \end{aligned}$$

$w = u + iv$, respectively.

As in Sect. 2 the conformal invariance of the fundamental solutions to the Laplace operator is applied. The harmonic Robin function of the concentric ring $\mathcal{R}$, for $\frac{\alpha}{\beta}$ not an integer, is see [11,17,18],

$$\hat{R}_{1;\alpha,\beta}(z,\zeta) = \hat{G}_1(z,\zeta) + 2\beta \sum_{\kappa=-\infty,\kappa\neq 0}^{\infty} \frac{(z\overline{\zeta})^\kappa + (\overline{z}\zeta)^\kappa}{(\alpha + \kappa\beta)\big(1 - \big(\frac{r}{\rho}\big)^{2\kappa}\big)},$$

where

$$\hat{G}_1(z,\zeta)=\frac{\log|z|^2\log|\zeta|^2}{\log r^2-\log\rho^2}+\log\Big|\frac{1-z\overline{\zeta}}{\zeta-z}\prod_{\kappa=1}^{\infty}\frac{z\overline{\zeta}-\left(\frac{r}{\rho}\right)^{2\kappa}}{z-\left(\frac{r}{\rho}\right)^{2\kappa}\zeta}\frac{1-\left(\frac{r}{\rho}\right)^{2\kappa}z\overline{\zeta}}{\zeta-\left(\frac{r}{\rho}\right)^{2\kappa}z}\Big|^2$$

is the Green function. Hence, for D the Green function is with $\tau=\frac{r}{\rho}$

$$\begin{aligned}G_1(z,\zeta)\ &=\frac{\log|\omega(z)|^2\log|\omega(\zeta)|^2}{\log\tau^2}\\&\quad+\log\Big|\frac{1-\omega(z)\overline{\omega(\zeta)}}{\omega(\zeta)-\omega(z)}\prod_{\kappa=1}^{\infty}\frac{\omega(z)\overline{\omega(\zeta)}-\tau^{2\kappa}}{\omega(z)-\tau^{2\kappa}\omega(\zeta)}\frac{1-\tau^{2\kappa}\omega(z)\overline{\omega(\zeta)}}{\omega(\zeta)-\tau^{2\kappa}\omega(z)}\Big|^2\end{aligned}$$

and the Robin function

$$R_{1;\alpha,\beta}(z,\zeta)=G_1(z,\zeta)+2\beta\sum_{\kappa=-\infty,\kappa\neq0}^{\infty}\frac{(\omega(z)\overline{\omega(\zeta)})^\kappa+(\overline{\omega(z)}\omega(\zeta))^\kappa}{(\alpha+\kappa\beta)(1-\tau^{2\kappa})}.$$

In order to keep the interpolation property of the Robin function between Green and Neumann functions the latter has to be defined as

$$\begin{aligned}N_1(z,\zeta)&=\frac{\log|\omega(z)|^2\log|\omega(\zeta)|^2}{\log\tau^2}-\log\big|(\omega(\zeta)-\omega(z))(1-\omega(z)\overline{\omega(\zeta)})\\&\quad\times\prod_{\kappa=1}^{\infty}\frac{(\omega(z)-\tau^{2\kappa}\omega(\zeta))(\omega(\zeta)-\tau^{2\kappa}\omega(z))(\omega(z)\overline{\omega(\zeta)}-\tau^{2\kappa})(1-\tau^{2\kappa}\omega(z)\overline{\omega(\zeta)})}{|\omega(z)\overline{\omega(\zeta)}|^2}\Big|^2.\end{aligned}$$

To investigate their boundary properties on $\partial D=E_i\cup E_a$, $1<r$ is assumed and $r<\rho$ observed.

For $z\in E_i$, i.e. for $|\omega(z)|=\tau$,

$$\begin{aligned}G_1(z,\zeta)\ &=\log|\omega(\zeta)|^2+\log\Big|\frac{\overline{\omega(z)}-\tau^2\overline{\omega(\zeta)}}{(\omega(\zeta)-\omega(z))\overline{\omega(z)}}\Big|^2\\&\quad+\log\prod_{\kappa=1}^{\infty}\Big|\frac{\tau^{2\kappa}\overline{\omega(z)}-\tau^2\overline{\omega(\zeta)}}{\tau^{2\kappa}\overline{\omega(z)}-\frac{\tau^2}{\omega(\zeta)}}\frac{\tau^{2\kappa}\omega(z)-\frac{1}{\overline{\omega(\zeta)}}}{\tau^{2\kappa}\omega(z)-\omega(\zeta)}\Big|^2\\&=\log|\omega(\zeta)|^2+\log\Big|\frac{\overline{\omega(z)}-\tau^2\overline{\omega(\zeta)}}{(\omega(\zeta)-\omega(z))\overline{\omega(z)}}\Big|^2\\&\quad+\lim_{n\to\infty}\log\prod_{\kappa=1}^{n}\Big|\frac{\tau^{2(\kappa-1)}\overline{\omega(z)}-\overline{\omega(\zeta)}}{\tau^{2\kappa}\omega(z)-\omega(\zeta)}\frac{\tau^{2(\kappa+1)}\overline{\omega(\zeta)}-\overline{\omega(z)}}{\tau^{2\kappa}\omega(\zeta)-\omega(z)}\Big|^2=0.\end{aligned}$$

Similarly, for $z\in E_e$, i.e. for $|\omega(z)|=1$,

$$G_1(z,\zeta)=\log\prod_{\kappa=1}^{\infty}\Big|\frac{\tau^{2\kappa}-\omega(z)\overline{\omega(\zeta)}}{\tau^{2\kappa}-\overline{\omega(z)}\omega(\zeta)}\frac{\tau^{2\kappa}-\frac{1}{\omega(z)\overline{\omega(\zeta)}}}{\tau^{2\kappa}-\frac{1}{\overline{\omega(z)}\omega(\zeta)}}\Big|^2=0.$$

The outward normal direction is

$$\nu_z = -\frac{\rho\omega - \frac{1}{\rho\omega}}{|\rho\omega - \frac{1}{\rho\omega}|} \text{ on } E_i, \quad \nu_z = \frac{\rho\omega - \frac{1}{\rho\omega}}{|\rho\omega - \frac{1}{\rho\omega}|} \text{ on } E_e.$$

The arc length element $ds = ds_z$ satisfies

$$\begin{aligned} ds^2 &= \Big[\tau^2\Big(\frac{\omega^2}{\tau^2} + \frac{\tau^2}{\omega^2}\Big) - \tau^2\big(r^2 + \frac{1}{r^2}\big)\Big]\frac{(d\omega)^2}{4\omega^2} = -\big|\rho\omega - \frac{1}{\rho\omega}\big|^2\frac{(d\omega)^2}{4\omega^2} \text{ on } E_i, \\ ds^2 &= \big[\omega^2 + \frac{1}{\omega^2} - \big(\rho^2 + \frac{1}{\rho^2}\big)\big]\frac{(d\omega)^2}{4\omega^2} = -\big|\rho\omega - \frac{1}{\rho\omega}\big|^2\frac{(d\omega)^2}{4\omega^2} \text{ on } E_e, \end{aligned}$$

so that

$$ds = i\big|\rho\omega - \frac{1}{\rho\omega}\big|\frac{d\omega}{2\omega} \text{ on } E_i, \quad ds = -i\big|\rho\omega - \frac{1}{\rho\omega}\big|\frac{d\omega}{2\omega} \text{ on } E_e$$

are chosen. The normal derivative on ∂D applied to real-valued functions appears as

$$\partial_{\nu_z} = 2\mathrm{Re}\,\nu_z\frac{2\omega}{\rho\omega - \frac{1}{\rho\omega}}\partial_\omega.$$

Thus

$$\partial_{\nu_z} ds_z = 2\mathrm{Re}\,\omega\partial_\omega\frac{d\omega}{i\omega} \text{ on } \partial D, \quad \omega = \omega(z).$$

Introducing $\sigma = \sigma(z) = \frac{2}{\left|\rho\omega(z) - \frac{1}{\rho\omega(z)}\right|}$ thus for real functions

$$\partial_{\nu_z} = -2\sigma\mathrm{Re}\,\omega\partial_\omega \text{ on } E_i, \quad \partial_{\nu_z} = 2\sigma\mathrm{Re}\,\omega\partial_\omega \text{ on } E_e.$$

For further calculating the boundary behavior of the Robin function some relations are needed following from simple reformulations.

Lemma 2. For $|\tau| < 1$ and bounded ω

$$\sum_{\kappa=-\infty,\kappa\neq 0}^{\infty}\frac{(\omega(z)\overline{\omega(\zeta)})^\kappa}{1-\tau^{2\kappa}} = \frac{\omega(z)\overline{\omega(\zeta)}}{1-\omega(z)\overline{\omega(\zeta)}} + \sum_{\kappa=1}^{\infty}\Big[\frac{1}{1-\omega(z)\overline{\omega(\zeta)}\tau^{2\kappa}} - \frac{\omega(z)\overline{\omega(\zeta)}}{\omega(z)\overline{\omega(\zeta)} - \tau^{2\kappa}}\Big],$$

and

$$\begin{aligned} \sum_{\kappa=-\infty,\kappa\neq 0}^{\infty}\frac{\kappa(\omega(z)\overline{\omega(\zeta)})^\kappa}{1-\tau^{2\kappa}} &= \frac{\omega(z)\overline{\omega(\zeta)}}{(1-\omega(z)\overline{\omega(\zeta)})^2} \\ &+ \sum_{\kappa=1}^{\infty}\Big[\frac{\omega(z)\overline{\omega(\zeta)}\tau^{2\kappa}}{(1-\omega(z)\overline{\omega(\zeta)}\tau^{2\kappa})^2} + \frac{\omega(z)\overline{\omega(\zeta)}\tau^{2\kappa}}{(\omega(z)\overline{\omega(\zeta)} - \tau^{2\kappa})^2}\Big] \end{aligned}$$

hold.

From

$$\omega\partial_\omega G_1(z,\zeta)=\frac{\log|\omega(\zeta)|^2}{\log\tau^2}+\frac{\omega(z)}{\omega(\zeta)-\omega(z)}-\frac{\omega(z)\overline{\omega(\zeta)}}{1-\omega(z)\overline{\omega(\zeta)}}$$
$$+\sum_{\kappa=1}^{\infty}\Big[\frac{\omega(z)\overline{\omega(\zeta)}}{\omega(z)\overline{\omega(\zeta)}-\tau^{2\kappa}}-\frac{\tau^{2\kappa}\omega(z)\overline{\omega(\zeta)}}{1-\tau^{2\kappa}\omega(z)\overline{\omega(\zeta)}}+\frac{\tau^{2\kappa}\omega(z)}{\omega(\zeta)-\tau^{2\kappa}\omega(z)}-\frac{\omega(z)}{\omega(z)-\tau^{2\kappa}\omega(\zeta)}\Big]$$

and

$$\omega\partial_\omega(R_{1;\alpha,\beta}(z,\zeta)-G_1(z,\zeta))=2\beta\sum_{\kappa=-\infty,\kappa\neq0}^{\infty}\frac{\kappa}{\alpha+\kappa\beta}\frac{(\omega(z)\overline{\omega(\zeta)})^\kappa}{1-\tau^{2\kappa}},$$

using the first relation in Lemma 2, it follows for $|\omega(z)|=1,|\omega(\zeta)|<1$, i.e. on $E_e,\zeta\notin E_e$, that

$$(\alpha+2\beta\mathrm{Re}\omega(z)\partial_{\omega(z)})R_{1;\alpha,\beta}(z,\zeta)=2\beta\Big[\frac{\log|\omega(\zeta)|^2}{\log\tau^2}-1\Big]$$

and for $|\omega(z)|=\tau,\tau<|\omega(\zeta)|$, i.e. on $E_i,\zeta\notin E_i$ that

$$(\alpha+2\beta\mathrm{Re}\omega(z)\partial_{\omega(z)})R_{1;\alpha,\beta}(z,\zeta)=2\beta\frac{\log|\omega(\zeta)|^2}{\log\tau^2}.$$

For $\zeta\in E_e$, i.e. for $|\omega(\zeta)|=1$ on E_e

$$(\alpha+2\beta\mathrm{Re}\,\omega(z)\partial_{\omega(z)})R_{1;\alpha,\beta}(z,\zeta)=2\beta\Big[\frac{\omega(\zeta)}{\omega(\zeta)-\omega(z)}+\frac{\overline{\omega(\zeta)}}{\overline{\omega(\zeta)}-\overline{\omega(z)}}-2\Big],$$

while for $\zeta\in E_i$, i.e. for $|\omega(\zeta)|=\tau$, on E_i

$$(\alpha+2\beta\mathrm{Re}\,\omega(z)\partial_{\omega(z)})R_{1;\alpha,\beta}(z,\zeta)=2\beta\Big[\frac{\omega(\zeta)}{\omega(\zeta)-\omega(z)}+\frac{\overline{\omega(\zeta)}}{\overline{\omega(\zeta)}-\overline{\omega(z)}}\Big].$$

Theorem 4. The Robin function $R_{1;\alpha,\beta}$ for the elliptic ring D satisfies for any $\zeta\in D$

- $R_{1;\alpha,\beta}(\,\cdot\,,\zeta)$ is harmonic in $D\setminus\{\zeta\}$ and continuously differentiable in $\overline{D}\setminus\{\zeta\}$,
- $h(z,\zeta)=R_{1;\alpha,\beta}(z,\zeta)+\log|\zeta-z|^2$ is harmonic for $z\in D$,
- $\alpha R_{1;\alpha,\beta}(z,\zeta)+\frac{\beta}{\sigma}(1-2\frac{\log|\omega(z)|^2}{\log\tau^2})\partial_{\nu_z}R_{1;\alpha,\beta}(z,\zeta)=2\beta\Big[\frac{\log|\omega(z)|^2}{\log\tau^2}+\frac{\log|\omega(\zeta)|^2}{\log\tau^2}-1\Big]$ for $z\in E_i\cup E_e$, where the density function σ given by $\frac{2}{\sigma(z)}=\Big|\rho\omega(z)-\frac{1}{\rho\omega(z)}\Big|$ on $E_i\cup E_e$ has finite mass $\int\limits_{\partial D}\sigma(z)ds_z=4\pi$,
- $\beta\int\limits_E\sigma(z)R_{1;\alpha,\beta}(z,\zeta)ds_z=0$ (normalization condition).

It has the further property

- $R_{1;\alpha,\beta}(z,\zeta) = R_{1;\alpha,\beta}(\zeta,z), \quad z,\zeta \in D, \quad z \neq \zeta$ (symmetry).

Remarks. The operator in the boundary condition can be rewritten as

$$\alpha + \frac{\beta}{\sigma}(1 - 2\frac{\log|\omega(z)|^2}{\log\tau^2})\partial_{\nu_z} = \alpha + 2\beta\mathrm{Re}(\omega(z)\partial_{\omega(z)}).$$

The factor $1 - 2\frac{\log|\omega(z)|^2}{\log\tau^2}$ becomes 1 on E_e and -1 on E_i. The normalization condition is evident as the Robin function on the boundary is just

$$2\beta \sum_{\kappa=-\infty,\kappa\neq 0}^{\infty} \frac{(\omega(z)\overline{\omega(\zeta)})^\kappa + (\overline{\omega(z)}\omega(\zeta))^\kappa}{(\alpha+\kappa\beta)(1-\tau^{2\kappa})}.$$

Integrating this sum multiplied with $\frac{d\omega}{i\omega}$ along $|\omega| = \tau$ and $|\omega| = 1$ does not contribute nonzero terms as the sum has no $(\kappa = 0)$-term.

Theorem 5. Any function $w \in C^2(D;\mathbb{C}) \cap C^1(\overline{D};\mathbb{C})$ can be represented as

$$\begin{aligned} w(z) = &-\frac{1}{4\pi}\int_{\partial D}\{w(\zeta)\partial_{\nu_\zeta}R_{1;\alpha,\beta}(\zeta,z) - \partial_{\nu_\zeta}w(\zeta)R_{1;\alpha,\beta}(\zeta,z)\}ds_\zeta \\ &-\frac{1}{\pi}\int_D \partial_\zeta\partial_{\overline{\zeta}}w(\zeta)R_{1;\alpha,\beta}(z,\zeta)d\xi d\eta. \end{aligned}$$

A proof for regular domains is given e.g. in [11]. Applying the boundary condition of the Robin function leads to some representation formulas proper for treating a related modified Robin boundary value problem.

Corollary 2. Any function $w \in C^2(D;\mathbb{C}) \cap C^1(\overline{D};\mathbb{C})$ can be represented as

$$\begin{aligned} w(z) = &-\frac{1}{4\pi i\alpha}\int_{\partial D}[\alpha w(\zeta) + \beta(\omega(\zeta)\partial_{\omega(\zeta)} + \overline{\omega(\zeta)}\partial_{\overline{\omega(\zeta)}})w(\zeta)] \\ &\times 2\mathrm{Re}(\omega(\zeta)\partial_{\omega(\zeta)}R_{1;\alpha,\beta}(z,\zeta))\frac{d\omega(\zeta)}{\omega(\zeta)} \\ &+\frac{\beta}{2\pi i\alpha}\int_{\partial D}(\omega(\zeta)\partial_{\omega(\zeta)} + \overline{\omega(\zeta)}\partial_{\overline{\omega(\zeta)}})w(\zeta)\frac{d\omega(\zeta)}{\omega(\zeta)}\frac{\log|\omega(z)|^2}{\log\tau^2} \\ &-\frac{\beta}{2\pi i\alpha}\int_{E_e}(\omega(\zeta)\partial_{\omega(\zeta)} + \overline{\omega(\zeta)}\partial_{\overline{\omega(\zeta)}})w(\zeta)\frac{d\omega(\zeta)}{\omega(\zeta)} \\ &-\frac{1}{\pi}\int_D \partial_\zeta\partial_{\overline{\zeta}}w(\zeta)R_{1;\alpha,\beta}(z,\zeta)d\xi d\eta, \quad \text{for } \alpha \neq 0, \\ w(z) = &\frac{1}{4\pi i\beta}\int_{\partial D}[\alpha w(\zeta) + \beta(\omega(\zeta)\partial_{\omega(\zeta)} + \overline{\omega(\zeta)}\partial_{\overline{\omega(\zeta)}})w(\zeta)]R_{1;\alpha,\beta}(z,\zeta)\frac{d\omega(\zeta)}{\omega(\zeta)} \end{aligned}$$

$$-\frac{1}{2\pi i}\int\limits_{\partial D} w(\zeta)\frac{d\omega(\zeta)}{\omega(\zeta)}\frac{\log|\omega(z)|^2}{\log\tau^2}+\frac{1}{2\pi i}\int\limits_{E_e} w(\zeta)\frac{d\omega(\zeta)}{\omega(\zeta)}$$
$$-\frac{1}{\pi}\int\limits_D \partial_\zeta\partial_{\overline{\zeta}} w(\zeta)R_{1;\alpha,\beta}(z,\zeta)d\xi d\eta,\ \text{ if }\ \beta\neq 0.$$

Robin Problem. Find a solution to the Poisson equation

$$w_{z\overline{z}}=f\ \text{ in }\ D,\ \ f\in L_p(D;\mathbb{C}),\ \ 2<p,$$

satisfying for $z\in\partial D$

$$\alpha w(z)+\beta(\omega(z)\partial_{\omega(z)}+\overline{\omega(z)}\partial_{\overline{\omega(z)}})w(z)=\gamma(z),\gamma\in C(\partial D;\mathbb{C}).$$

Theorem 6. For $f\in L_p(D;\mathbb{C}),2<p,\ \gamma\in C(\partial D;\mathbb{C})$, the Robin problem

$$w_{z\overline{z}}=f \text{ in } D,\ \alpha w+\beta(\omega\partial_\omega+\overline{\omega}\partial_{\overline{\omega}})w=\gamma \text{ on } \partial D,$$

(i) for $\beta\neq 0$ is solvable if and only if

$$\frac{1}{2\pi i}\int\limits_{E_e}\gamma(\zeta)\frac{d\omega(\zeta)}{\omega(\zeta)}+\frac{\beta}{\log\tau}\frac{1}{2\pi i}\int\limits_{\partial D} w(\zeta)\frac{d\omega(\zeta)}{\omega(\zeta)}-\frac{\alpha}{2\pi i}\int\limits_{E_e} w(\zeta)\frac{d\omega(\zeta)}{\omega(\zeta)}$$
$$=-\frac{2\beta}{\pi}\int\limits_D f(\zeta)\big[\frac{\log|\omega(\zeta)|^2}{\log\tau^2}-1\big]d\xi d\eta$$

and

$$\frac{1}{2\pi i}\int\limits_{\partial D}\gamma(\zeta)\frac{d\omega(\zeta)}{\omega(\zeta)}-\frac{\alpha}{2\pi i}\int\limits_{\partial D} w(\zeta)\frac{d\omega(\zeta)}{\omega(\zeta)}=\frac{2\beta}{\pi}\int\limits_D f(\zeta)d\xi d\eta,$$

the solution being then

$$w(z)=\frac{1}{4\pi i\beta}\int\limits_{\partial D}\gamma(\zeta)R_{1;\alpha,\beta}(z,\zeta)\frac{d\omega(\zeta)}{\omega(\zeta)}-\frac{1}{2\pi i}\int\limits_{\partial D} w(\zeta)\frac{d\omega(\zeta)}{\omega(\zeta)}\frac{\log|\omega(z)|^2}{\log\tau^2}$$
$$+\frac{1}{2\pi i}\int\limits_{E_e} w(\zeta)\frac{d\omega(\zeta)}{\omega(\zeta)}-\frac{1}{\pi}\int\limits_D f(\zeta)R_{1;\alpha,\beta}(z,\zeta)d\xi d\eta,$$

(ii) for $\alpha\neq 0$ is solvable if and only if

$$\frac{\beta}{\log\tau}\big[\frac{1}{4\pi i\alpha}\int\limits_{\partial D}\gamma(\zeta)\frac{d\omega(\zeta)}{\omega(\zeta)}-\frac{\beta}{2\pi i\alpha}\int\limits_{\partial D}(\omega(\zeta)\partial_{\omega(\zeta)}+\overline{\omega(\zeta)}\partial_{\overline{\omega(\zeta)}})w(\zeta)\frac{d\omega(\zeta)}{\omega(\zeta)}\big]$$
$$+\frac{\beta}{2\pi i}\int\limits_{E_e}(\omega(\zeta)\partial_{\omega(\zeta)}+\overline{\omega(\zeta)}\partial_{\overline{\omega(\zeta)}})w(\zeta)\frac{d\omega(\zeta)}{\omega(\zeta)}=-\frac{2\beta}{\pi}\int\limits_D f(\zeta)\big[\frac{\log|\omega(\zeta)|^2}{\log\tau^2}-1\big]d\xi d\eta$$

and

$$\frac{\beta}{2\pi i}\int\limits_{\partial D}(\omega(\zeta)\partial_{\omega(\zeta)}+\overline{\omega(\zeta)}\partial_{\overline{\omega(\zeta)}})w(\zeta)\frac{d\omega(\zeta)}{\omega(\zeta)}=\frac{2\beta}{\pi}\int\limits_{D}f(\zeta)d\xi d\eta,$$

the solution then being

$$\begin{aligned}w(z)=&-\frac{1}{4\pi i\alpha}\int\limits_{\partial D}\gamma(\zeta)2\mathrm{Re}(\omega(\zeta)\partial_{\omega(\zeta)})R_{1;\alpha,\beta}(\zeta,z)\frac{d\omega(\zeta)}{\omega(\zeta)}\\&+\frac{\log|\omega(z)|^2}{\log\tau^2}\frac{\beta}{2\pi i\alpha}\int\limits_{\partial D}(\omega(\zeta)\partial_{\omega(\zeta)}+\overline{\omega(\zeta)}\partial_{\overline{\omega(\zeta)}})w(\zeta)\frac{d\omega(\zeta)}{\omega(\zeta)}\\&-\frac{\beta}{2\pi i\alpha}\int\limits_{E_e}(\omega(\zeta)\partial_{\omega(\zeta)}+\overline{\omega(\zeta)}\partial_{\overline{\omega(\zeta)}})w(\zeta)\frac{d\omega(\zeta)}{\omega(\zeta)}\\&-\frac{1}{\pi}\int\limits_{D}f(\zeta)R_{1;\alpha,\beta}(\zeta,z)d\xi d\eta.\end{aligned}$$

Remarks. The proof follows from the formulas given in Corollary 2. In particular for the case $\alpha\neq 0$ it is based on the second part of Lemma 2 and

$$\begin{aligned}&2\mathrm{Re}(\omega(\zeta)\partial_{\omega(\zeta)})R_{1;\alpha,\beta}(z,\zeta)=2\mathrm{Re}\Big[\frac{\log|\omega(z)|^2}{\log\tau^2}-\frac{\omega(\zeta)}{\omega(\zeta)-\omega(z)}-\frac{\omega(z)\overline{\omega(\zeta)}}{1-\omega(z)\overline{\omega(\zeta)}}\\&+\sum_{\kappa=1}^{\infty}\Big[\frac{\tau^{2\kappa}}{\omega(z)\overline{\omega(\zeta)}-\tau^{2\kappa}}-\frac{\tau^{2\kappa}\omega(z)\overline{\omega(\zeta)}}{1-\tau^{2\kappa}\omega(z)\overline{\omega(\zeta)}}+\frac{\tau^{2\kappa}\omega(\zeta)}{\omega(z)-\tau^{2\kappa}\omega(\zeta)}-\frac{\tau^{2\kappa}\omega(z)}{\omega(\zeta)-\tau^{2\kappa}\omega(z)}\Big]\Big]\end{aligned}$$

and also

$$\begin{aligned}&(\alpha+2\beta\mathrm{Re}(\omega(z)\partial_{\omega(z)}))2\mathrm{Re}(\omega(\zeta)\partial_{\omega(\zeta)})R_{1;\alpha,\beta}(z,\zeta)\\&=\begin{cases}-2\alpha\Big[\dfrac{\omega(\zeta)}{\omega(\zeta)-\omega(z)}+\dfrac{\overline{\omega(\zeta)}}{\overline{\omega(\zeta)-\omega(z)}}-1\Big]+\dfrac{\beta}{\log\tau}, & \text{if } z,\zeta\in E_e, \text{ or } z,\zeta\in E_i,\\ \dfrac{\beta}{\log\tau}, & \text{if } z\in E_i,\ \zeta\in E_e,\\ & \text{or } z\in E_e,\ \zeta\in E_i.\end{cases}\end{aligned}$$

For $\beta=0$ there appears no solvability condition for the Dirichlet problem in the ring domain D! The first listed solvability condition for the case $\beta\neq 0$ can be also written as

$$\begin{aligned}&\frac{1}{2\pi i}\int\limits_{E_i}\gamma(\zeta)\frac{d\omega(\zeta)}{\omega(\zeta)}+\frac{\beta}{\log\tau}\frac{1}{2\pi i}\int\limits_{\partial D}w(\zeta)\frac{d\omega(\zeta)}{\omega(\zeta)}-\frac{\alpha}{2\pi i}\int\limits_{E_i}w(\zeta)\frac{d\omega(\zeta)}{\omega(\zeta)}\\&=-\frac{2\beta}{\pi}\int\limits_{D}f(\zeta)\frac{\log|\omega(\zeta)|^2}{\log\tau^2}d\xi d\eta.\end{aligned}$$

If $\alpha = 0$ then β disappears from the conditions as β then appears to be a factor of γ, and the conditions are those for the Neumann problem.

The first solvability condition for the case $\alpha \neq 0$ can be replaced by

$$\frac{\beta}{\log \tau}\Big[\frac{1}{4\pi i\alpha}\int\limits_{\partial D}\gamma(\zeta)\frac{d\omega(\zeta)}{\omega(\zeta)} - \frac{\beta}{2\pi i\alpha}\int\limits_{\partial D}(\omega(\zeta)\partial_{\omega(\zeta)} + \overline{\omega(\zeta)}\partial_{\overline{\omega(\zeta)}})w(\zeta)\frac{d\omega(\zeta)}{\omega(\zeta)}\Big]$$
$$+\frac{\beta}{2\pi i}\int\limits_{E_i}(\omega(\zeta)\partial_{\omega(\zeta)} + \overline{\omega(\zeta)}\partial_{\overline{\omega(\zeta)}})w(\zeta)\frac{d\omega(\zeta)}{\omega(\zeta)} = -\frac{2\beta}{\pi}\int\limits_{D} f(\zeta)\frac{\log|\omega(\zeta)|^2}{\log\tau^2}d\xi d\eta.$$

4 Concluding Remarks

The formulation of proper Robin boundary conditions for the Poisson equation is not at all obvious as it seems when looking at the condition for the disc $\mathbb{D}$. A linear combination of Dirichlet and Neumann data is not adequate in general. Some density function σ has to be involved as a function of $z \in \partial D$ and the constant β has to be modified in the sense that it is a function of $z \in \partial D$ and $\zeta \in D$ but constant in $z \in \partial D$, where the constant may vary on different parts of the boundary ∂D. Having altered the Neumann condition properly then a linear combination of the Dirichlet data with this modified Neumann condition seems to be the right formulation for the Robin condition. In the presently discussed case of the elliptic ring the Neumann condition becomes

$$\big(1 - 2\frac{\log|\omega(z)|^2}{\log\tau^2}\big)\partial_{\nu_z}N_1(z,\zeta) = 2\sigma(z)\big(\frac{\log|\omega(z)|^2}{\log\tau^2} + \frac{\log|\omega(\zeta)|^2}{\log\tau^2} - 1\big).$$

In particular

$$\big(1 - 2\frac{\log|\omega(z)|^2}{\log\tau^2}\big) = \begin{cases} -1, \text{ if } z \in E_i, \\ 1, \text{ if } z \in E_e, \end{cases}$$

and

$$\big(\frac{\log|\omega(z)|^2}{\log\tau^2} + \frac{\log|\omega(\zeta)|^2}{\log\tau^2} - 1\big) = \begin{cases} \frac{\log|w(\zeta)|^2}{\log\tau^2}, & \text{if } z \in E_i, \\ \frac{\log|w(\zeta)|^2}{\log\tau^2} - 1, & \text{if } z \in E_e. \end{cases}$$

Moreover, in the general case the real constants α, β have to be replaced by real-valued functions.

References

1. Akel, M., Begehr, H.: Neumann function for a hyperbolic strip and a class of related plane domains. Math. Nachr. (to appear, 2017). doi:10.1002/mana.201500501
2. Aksoy, Ü., Çelebi, A.O.: Polyharmonic Robin problem for complex linear partial differential equations. Complex Var. Elliptic Eqs. **59**(12), 1679–1695 (2014)

3. Begehr, H.: Complex Analytic Methods for Partial Differential Equations: An Introductory Text. World Scientific, Singapore (1994)
4. Begehr, H.: Boundary value problems in complex analysis, I, II. Bol. Asoc. Mat. Venezolana **XII**, 65–85, 217–250 (2005)
5. Begehr, H.: Green function for a hyperbolic strip and a class of related plane domains. Appl. Anal. **93**, 2370–2385 (2014)
6. Begehr, H., Burgumbayeva, S., Shupeyeva, B.: Remark on Robin problem for Poisson equation. Complex Var. Elliptic Eqs. (to appear). doi:10.1080/17476933.2017.1303052
7. Begehr, H., Harutyunyan, G.: Robin boundary value problem for the Cauchy-Riemann operator. Complex Variables Theory Appl. **50**, 1125–1136 (2005)
8. Begehr, H., Harutyunyan, G.: Robin boundary value problem for the Poisson equation. J. Anal. Appl. **4**, 201–213 (2006)
9. Begehr, H., Obolashvili, E.: Some boundary value problems for a Beltrami equation. Complex Var. Theory Appl. **26**, 113–122 (1994)
10. Begehr, H., Vaitekhovich, T.: Some harmonic Robin functions in the complex plane. Adv. Pure Appl. Math. **1**, 19–34 (2010)
11. Begehr, H., Vaitekhovich, T.: Modified harmonic Robin functions. Complex Var. Elliptic Eqs. **58**, 483–496 (2013)
12. Begehr, H., Vaitekhovich, T.: Schwarz problem in lens and lune. Complex Var. Elliptic Eqs. **59**, 76–84 (2014)
13. Dittmar, B., Hantke, M.: The Robin function and its eigenvalues. Georgian Math. J. **14**, 403–417 (2007)
14. Duren, P.L., Schiffer, M.H.: Robin functions and energy functionals of multiply connected domains. Pacific J. Math. **148**, 251–273 (1991)
15. Gustafson, K., Abe, T.: The third boundary condition - was it Robin's? Math. Intelligencer **20**, 63–71 (1998)
16. Haack, W., Wendland, W.: Lectures on Partial and Pfaffian Differential Equations. Pergamon Press, Oxford (1972)
17. Vaitsiakhovich, T.: Boundary value problems for complex partial differential equations in a ring domain. Ph.D. thesis, FU Berlin (2008). www.diss.fu-berlin.de/diss/receive/FUDISS_thesis_000000003859
18. Vaitekhovich, T.S.: Boundary value problems to second order complex partial differential equations in a ring domain. Šiauliai Math. Semin. **2**(10), 117–146 (2007)
19. Vaitekhovich, T.S.: Boundary value problems to first order complex partial differential equations in a ring domain. Integr. Transf. Spec. Funct. **19**, 211–233 (2008)
20. Volkovysky, L.I., Lunts, G.L., Aranamovich, I.G.: A Collection of Problems on Complex Analysis. Pergamon Press, Oxford (1965)

A Nice Representation for a Link Between Bernstein-Durrmeyer and Kantorovich Operators

Margareta Heilmann[1(✉)] and Ioan Raşa[2]

[1] School of Mathematics and Natural Sciences, University of Wuppertal, Gaußstraße 20, 42119 Wuppertal, Germany
heilmann@math.uni-wuppertal.de
[2] Department of Mathematics, Technical University, Str. Memorandumului 28, 400114 Cluj-Napoca, Romania
Ioan.Rasa@math.utcluj.ro

Abstract. In this paper we consider a link between Bernstein-Durrmeyer operators and Kantorovich operators depending on a parameter ρ. We prove a nice representation by using the classical Bernstein polynomials and generalize the results for k-th order Kantorovich modifications.

Keywords: Linking operators · Bernstein-Durrmeyer operators · Kantorovich operators · k-th order Kantorovich modifications

1 Introduction

In [8] Păltănea introduced a class of operators $B_{n,\rho}$ depending on a nonnegative real parameter ρ which constitute a nontrivial link between the genuine Bernstein-Durrmeyer operators and the classical Bernstein operators.

For $j \in \mathbb{N}_0$, $0 \le j \le n$, the Bernstein basis functions are given by

$$p_{n,j}(x) = \binom{n}{j} x^j (1-x)^{n-j},\ 0 \le j \le n,\ x \in [0,1].$$

Moreover, for $1 \le j \le n-1$,

$$\mu_{n,j,\rho}(t) = \frac{t^{j\rho-1}(1-t)^{(n-j)\rho-1}}{B(j\rho,(n-j)\rho)}$$

with Euler's Beta function $B(x,y) = \displaystyle\int_0^1 t^{x-1}(1-t)^{y-1}dt = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$, $x,y>0$.

Let $f \in L_1[0,1]$ with finite limits at the endpoints of the interval $[0,1]$, i.e., $f(0) := \lim_{x\to 0^+} f(x)$ and $f(1) := \lim_{x\to 1^-} f(x)$. For $n \in \mathbb{N}$, $n \ge 2$, $\rho \in \mathbb{R}_+$, Păltănea [8, Definition 2.1] defined the operators $B_{n,\rho}$ by

$$B_{n,\rho}(f;x) = p_{n,0}(x)f(0) + p_{n,n}(x)f(1) + \sum_{j=1}^{n-1} p_{n,j}(x) \int_0^1 \mu_{n,j,\rho}(t) f(t)dt.$$

© Springer Nature Singapore Pte Ltd. 2017
D. Giri et al. (Eds.): ICMC 2017, CCIS 655, pp. 312–320, 2017.
DOI: 10.1007/978-981-10-4642-1_26

In [3, Theorem 2.3] Gonska and Păltănea proved the convergence of the operators $B_{n,\rho}$ to the classical Bernstein operator $B_{n,\infty}$, i.e., they proved that for every $f \in C[0,1]$

$$\lim_{\rho\to\infty} B_{n,\rho}f = B_{n,\infty}f \text{ uniformly on } [0,1].$$

In [2] Gonska and the authors of this paper investigated the k-th order Kantorovich modification of the Bernstein operators and in [5] the authors considered the kth order Kantorovich modification of the operators $B_{n,\rho}$, namely,

$$B_{n,\rho}^{(k)} = D^k \circ B_{n,\rho} \circ I_k,\ k \in \mathbb{N}_0, \rho \in \mathbb{R}_+ \cup \{\infty\},$$

where D^k denotes the k-th order ordinary differential operator and

$$I_k f = f, \text{ if } k = 0, \text{ and } I_k(f,x) = \int_0^x \frac{(x-t)^{k-1}}{(k-1)!} f(t)dt, \text{ if } k \in \mathbb{N}.$$

These operators play an important role in the investigation of simultaneous approximation.

From [5] (see the consideration of the special case $\rho \to \infty$ after Remark 2 there) we know that for each polynomial q

$$\lim_{\rho\to\infty} B_{n,\rho}^{(k)}q = B_{n,\infty}^{(k)}q \text{ uniformly on } [0,1].$$

Let $\varepsilon > 0$ be arbitrary. As the space of polynomials $\mathcal{P}$ is dense in $L_p[0,1]$, $\|\cdot\|_p$, $1 \le p < \infty$ and $C[0,1]$, $\|\cdot\|_\infty$, $p = \infty$, we can choose a polynomial q, such that $\|f - q\|_p < \varepsilon$. Then

$$\|(B_{n,\rho}^{(k)} - B_{n,\infty}^{(k)})f\|_p \le \|B_{n,\rho}^{(k)}(f-q)\|_p + \|B_{n,\infty}^{(k)}(f-q)\|_p + \|(B_{n,\rho}^{(k)} - B_{n,\infty}^{(k)})q\|_p.$$

As the operators $B_{n,\rho}^{(k)}$ and $B_{n,\infty}^{(k)}$ are bounded (see [5, Corollary 1] and [2, (3)] for the images of $e_0 = 1$) we immediately get

$$\lim_{\rho\to\infty} \|(B_{n,\rho}^{(k)} - B_{n,\infty}^{(k)})f\|_p = 0 \tag{1}$$

for each $f \in L_p[0,1]$, $\|\cdot\|_p$, $1 \le p < \infty$ and $C[0,1]$, $\|\cdot\|_\infty$, $p = \infty$.

For $\rho = 1$ and $\rho = \infty$ nice explicit representations are known, i.e.,

$$B_{n,1}^{(k)}(f;x) = \frac{n!(n-1)!}{(n-k)!(n+k-2)!} \sum_{j=0}^{n-k} p_{n-k,j}(x) \int_0^1 p_{n+k-2,j+k-1}(t) f(t)dt$$

$$B_{n,\infty}^{(k)}(f;x) = \frac{n!}{(n-k)!} \sum_{j=0}^{n-k} p_{n-k,j}(x) \Delta_{\frac{1}{n}}^k I_k\left(f; \frac{j}{n}\right),$$

where the forward difference of order k with step h for a function g is given by $\Delta_h^k g(x) = \sum_{i=0}^k \binom{k}{i}(-1)^{k-i} g(x+ih)$. By using Peano's representation theorem for divided differences (see, e.g., [9, p. 137]) this can also be written as

$$B_{n,\infty}^{(k)}(f;x) = \frac{n!}{(n-k)!} \sum_{j=0}^{n-k} p_{n-k,j}(x) \frac{1}{n^{k-1}} \int_0^1 N_{k,j}(t) f(t)dt, \tag{2}$$

where $N_{k,j}$ denotes the B-spline of order k to the equidistant knots $\frac{j}{n}, \dots \frac{j+k}{n}$, defined by

$$N_{1,j}(t) = \begin{cases} 1, \frac{j}{n} \le t < \frac{j+1}{n}, \\ 0, \text{otherwise}, \end{cases}$$
$$N_{k,j}(t) = \frac{n}{k-1}\left\{\left(t - \frac{j}{n}\right) N_{k-1,j}(t) + \left(\frac{j+k}{n} - t\right) N_{k-1,j+1}(t)\right\}.$$

Our goal is to find useful representations also for $\rho \neq 1, \infty$ for the general case $k \in \mathbb{N}$.

First we treat the case $k = 1$. In other words we prove an explicit representation for a non-trivial link between Bernstein-Durrmeyer and Kantorovich operators.

The idea is as follows. Consider

$$n \sum_{j=0}^{n-1} p_{n-1,j}(x) \int_0^1 K_{n,j,\rho}(t) f(t) dt$$

and determine a nice and easy to handle function $K_{n,j,\rho}(t)$ in such a way that

$$K_{n,j,1}(t) = p_{n-1,j}(t) \text{ and } \lim_{\rho \to \infty} K_{n,j,\rho}(t) = \begin{cases} 1, t \in \left(\frac{j}{n}, \frac{j+1}{n}\right), \\ 0, t \in [0,1] \setminus \left(\frac{j}{n}, \frac{j+1}{n}\right). \end{cases}$$

Throughout this paper we will use the following well-known formulas for the Bernstein basis functions.

$$p'_{n,j}(x) = n\left[p_{n-1,j-1}(x) - p_{n-1,j}(x)\right], \tag{3}$$
$$x p'_{n-1,j-1}(x) = (j-1) p_{n-1,j-1}(x) - j p_{n-1,j}(x). \tag{4}$$

2 A First Attempt

We start with the definition of $B^{(1)}_{n,\rho} = D \circ B_{n,\rho} \circ I$ and define

$$\omega_{n,j,\rho}(t) = \begin{cases} \int_t^1 \mu_{n,1,\rho}(u) du, & j = 0, \\ \int_0^t \left(\mu_{n,j,\rho}(u) - \mu_{n,j+1,\rho}(u)\right) du, & 1 \le j \le n-2, \\ \int_0^t \mu_{n,n-1,\rho}(u) du, & j = n-1. \end{cases} \tag{5}$$

Thus

$$\omega'_{n,j,\rho}(t) = \begin{cases} -\mu_{n,1,\rho}(t), & j = 0, \\ \mu_{n,j,\rho}(t) - \mu_{n,j+1,\rho}(t), & 1 \le j \le n-2, \\ \mu_{n,n-1,\rho}(t), & j = n-1. \end{cases}$$

Note that

$$\omega_{n,j,\rho}(0) = 0,\ 1 \le j \le n-1 \text{ and } \omega_{n,j,\rho}(1) = 0,\ 0 \le j \le n-2. \tag{6}$$

By applying (3), an appropriate index transform and the definition of $\omega_{n,j,\rho}$ we derive

$$\begin{aligned}
B^{(1)}_{n,\rho}(f;x) &= p'_{n,n}(x)I(f;1) + \sum_{j=1}^{n-1} p'_{n,j}(x) \int_0^1 \mu_{n,j,\rho}(t) I(f;t)dt \\
&= np_{n-1,n-1}(x)I(f;1) + n\sum_{j=1}^{n-1} p_{n-1,j-1}(x) \int_0^1 \mu_{n,j,\rho}(t) I(f;t)dt \\
&\quad -n\sum_{j=1}^{n-1} p_{n-1,j}(x) \int_0^1 \mu_{n,j,\rho}(t) I(f;t)dt \\
&= np_{n-1,0}(x) \int_0^1 \mu_{n,1,\rho}(t) I(f;t)dt \\
&\quad +np_{n-1,n-1}(x) \int_0^1 \mu_{n,n-1,\rho}(t) \left[I(f;1) - I(f;t)\right] dt \\
&\quad +n\sum_{j=1}^{n-2} p_{n-1,j}(x) \int_0^1 \left[\mu_{n,j+1,\rho}(t) - \mu_{n,j,\rho}(t)\right] I(f;t)dt \\
&= -np_{n-1,n-1}(x) \int_0^1 \omega'_{n,n-1,\rho}(t) \left[I(f;t) - I(f;1)\right] dt \\
&\quad -n\sum_{j=0}^{n-2} p_{n-1,j}(x) \int_0^1 \omega'_{n,j,\rho}(t) I(f;t)dt.
\end{aligned}$$

Integration by parts and observing (6) leads to

$$B^{(1)}_{n,\rho}(f;x) = n\sum_{j=0}^{n-1} p_{n-1,j}(x) \int_0^1 \omega_{n,j,\rho}(t) f(t)dt.$$

In our opinion this is not a nice representation and obviously this is not easy to handle.

3 A Second Approach

From now on we only consider $\rho \in \mathbb{N}$. We construct a function $K_{n,j,\rho}(t)$ in such a way that

$$K_{n,j,1}(t) = p_{n-1,j}(t) \text{ and } \lim_{\rho\to\infty} K_{n,j,\rho}(t) = \begin{cases} 1, t \in \left(\frac{j}{n}, \frac{j+1}{n}\right), \\ 0, t \in [0,1] \setminus \left[\frac{j}{n}, \frac{j+1}{n}\right], \end{cases}$$

and define an operator $H_{n,\rho}$ by

$$H_{n,\rho}(f;x) = n\sum_{j=0}^{n-1} p_{n-1,j}(x) \int_0^1 K_{n,j,\rho}(t) f(t)dt.$$

To do so, we consider the characteristic functions $\chi_{\left[\frac{j}{n},\frac{j+1}{n}\right]}(t)$ and for $0 \le j \le n-1$ the ρ points

$$t_i = \frac{i}{n\rho - 1},\ i = j\rho, j\rho+1, \ldots, (j+1)\rho - 1.$$

Then

$$t_i \in \begin{cases} \left(\frac{j}{n}, \frac{j+1}{n}\right), & n \ge 3,\ 1 \le j \le n-2, \\ \left[0, \frac{1}{n}\right), & n \ge 2,\ j = 0, \\ \left(\frac{n-1}{n}, 1\right], & n \ge 2,\ j = n-1. \end{cases}$$

Applying the Bernstein operator $B_{n\rho-1}$ to the functions $\chi_{\left[\frac{j}{n},\frac{j+1}{n}\right]}$ leads to

$$\begin{aligned} B_{n\rho-1,\infty}(\chi_{\left[\frac{j}{n},\frac{j+1}{n}\right]};t) &= \sum_{i=0}^{n\rho-1} p_{n\rho-1,i}(t)\chi_{\left[\frac{j}{n},\frac{j+1}{n}\right]}\left(\frac{i}{n\rho-1}\right) \\ &= \sum_{i=j\rho}^{(j+1)\rho-1} p_{n\rho-1,i}(t) \\ &= \sum_{i=0}^{\rho-1} p_{n\rho-1,i+j\rho}(t). \end{aligned}$$

For $\rho = 1$ we have $B_{n-1,\infty}(\chi_{\left[\frac{j}{n},\frac{j+1}{n}\right]};t) = p_{n-1,j}(t)$. Using a result for the application of Bernstein operators to discontinuous functions (see [6, (5.1) Theorem], [7, Theorem 1.9.1]) we derive

$$\lim_{\rho\to\infty} B_{n\rho-1,\infty}(\chi_{\left[\frac{j}{n},\frac{j+1}{n}\right]};t) = \begin{cases} 1, & \frac{j}{n} < t < \frac{j+1}{n}, \\ \frac{1}{2}, & t = \frac{j}{n}, \frac{j+1}{n}, \\ 0, & \text{otherwise.} \end{cases}$$

This is exactly what we need.
Thus we define the linking operator $H_{n,\rho}$ by

$$\begin{aligned} H_{n,\rho}(f;x) &= n\sum_{j=0}^{n-1} p_{n-1,j}(x) \int_0^1 B_{n\rho-1,\infty}(\chi_{\left[\frac{j}{n},\frac{j+1}{n}\right]};t) f(t)dt \\ &= n\sum_{j=0}^{n-1} p_{n-1,j}(x) \int_0^1 \left\{\sum_{i=0}^{\rho-1} p_{n\rho-1,i+j\rho}(t)\right\} f(t)dt. \end{aligned}$$

It is easy to see that the operators are positive linear contractions for each $f \in L_p[0,1]$, $1 \le p < \infty$, $C[0,1]$, $p = \infty$.

4 Relation Between $H_{n,\rho}$ and $B^{(1)}_{n,\rho}$

From Sects. 2 and 3 a natural question arises how the operators $H_{n,\rho}$ and $B^{(1)}_{n,\rho}$ are related. Indeed we will prove that they coincide for each $f \in L_p[0,1]$, $1 \le p < \infty$ and $C[0,1]$, respectively. First we treat the monomials $e_\nu(x) = x^\nu$, $\nu \in \mathbb{N}_0$.

Theorem 1. *For each monomial e_ν, $\nu \in \mathbb{N}_0$, we have*

$$H_{n,\rho}(e_\nu; x) = B^{(1)}_{n,\rho}(e_\nu; x).$$

Proof. By the definition of the operators $B^{(1)}_{n,\rho}$, using a representation for the images of monomials for $B_{n,\rho}$ which can be found, e.g., in [5, Proof of Theorem 1], applying (4) and an appropriate index transform we derive

$$\begin{aligned}
B^{(1)}_{n,\rho}(e_\nu; x) &= \frac{d}{dx}\left(B_{n,\rho}(I_1 e_\nu; x)\right) \\
&= \frac{1}{\nu+1}\frac{d}{dx}\left(B_{n,\rho}(e_{\nu+1}; x)\right) \\
&= \frac{n\rho}{\nu+1}\cdot\frac{(n\rho-1)!}{(n\rho+\nu)!}\cdot\frac{d}{dx}\left\{x\sum_{j=1}^{n} p_{n-1,j-1}(x)\cdot\frac{(j\rho+\nu)!}{(j\rho)!}\right\} \\
&= \frac{n\rho}{\nu+1}\cdot\frac{(n\rho-1)!}{(n\rho+\nu)!}\left\{\sum_{j=1}^{n} p_{n-1,j-1}(x)\cdot\frac{(j\rho+\nu)!}{(j\rho)!}\right. \\
&\quad \left. + \sum_{j=1}^{n} x p'_{n-1,j-1}(x)\cdot\frac{(j\rho+\nu)!}{(j\rho)!}\right\} \\
&= \frac{n\rho}{\nu+1}\cdot\frac{(n\rho-1)!}{(n\rho+\nu)!}\sum_{j=0}^{n-1} p_{n-1,j}(x) \\
&\quad \times\left\{(j+1)\cdot\frac{((j+1)\rho+\nu)!}{((j+1)\rho)!} - j\cdot\frac{(j\rho+\nu)!}{(j\rho)!}\right\}.
\end{aligned}$$

As

$$\int_0^1 p_{n\rho-1,i+j\rho}(t)t^\nu dt = \frac{(n\rho-1)!}{(n\rho+\nu)!}\cdot\frac{(i+j\rho+\nu)!}{(i+j\rho)!}$$

we have

$$H_{n,\rho}(e_\nu; x) = n\cdot\frac{(n\rho-1)!}{(n\rho+\nu)!}\sum_{j=0}^{n-1} p_{n-1,j}(x)\sum_{i=0}^{\rho-1}\frac{(i+j\rho+\nu)!}{(i+j\rho)!}.$$

Thus it remains to prove that

$$\frac{\rho}{\nu+1}\left\{(j+1)\cdot\frac{((j+1)\rho+\nu)!}{((j+1)\rho)!} - j\cdot\frac{(j\rho+\nu)!}{(j\rho)!}\right\} = \sum_{i=0}^{\rho-1}\frac{(i+j\rho+\nu)!}{(i+j\rho)!},$$

which is equivalent to

$$\binom{\rho+j\rho+\nu}{\nu+1} - \binom{j\rho+\nu}{\nu+1} = \sum_{i=0}^{\rho-1}\binom{i+j\rho+\nu}{\nu}.$$

The last equality is a combinatorial identity which can be found, e.g., in [4, (1.48)].

The equality of the two linking operators now follows as a corollary.

Corollary 1. *Let $f \in L_p[0,1]$, $1 \le p < \infty$ or $f \in C[0,1]$, $p = \infty$. Then*

$$H_{n,\rho} f = B^{(1)}_{n,\rho} f.$$

Proof. Let $\varepsilon > 0$ be arbitrary. As the space of polynomials $\mathcal{P}$ is dense in $L_p[0,1]$, $\|\cdot\|_p$, $1 \le p < \infty$ and $C[0,1]$, $\|\cdot\|_\infty$, $p = \infty$, we can choose a polynomial q, such that $\|f - q\|_p < \varepsilon$. Thus with Theorem 1 and as the operators are contractions we derive

$$\begin{aligned} \|(H_{n,\rho} - B^{(1)}_{n,\rho}) f\|_p &\le \|H_{n,\rho}(f-q)\|_p + \|B^{(1)}_{n,\rho})(q-f)\|_p \\ &\le 2\varepsilon. \end{aligned}$$

From the discussions in Sects. 2 and 3 we now get an identity for the functions $\omega_{n,j,\rho}$ defined in (5) in terms of a sum of Bernstein basis functions.

Corollary 2. *For each $0 \le j \le n-1$ we have*

$$\omega_{n,j,\rho} = \sum_{i=0}^{\rho-1} p_{n\rho-1,i+j\rho}$$

on $[0,1]$.

5 Representation for the *k*-th Order Kantorovich Modification

In this section we generalize the representation of the operators to $k \in \mathbb{N}$.

Theorem 2. *Let $n, k \in \mathbb{N}$, $n - k \ge 1$, $\rho \in \mathbb{N}$ and $f \in L_1[0,1]$. Then we have the representation*

$$\begin{aligned} B^{(k)}_{n,\rho}(f;x) = {} & \frac{n!(n\rho-1)!}{(n-k)!(n\rho+k-2)!} \sum_{j=0}^{n-k} p_{n-k,j}(x) \\ & \times \int_0^1 \sum_{i_1=0}^{\rho-1} \cdots \sum_{i_k=0}^{\rho-1} p_{n\rho+k-2,j\rho+i_1+\cdots+i_k+k-1}(t) f(t)\,dt. \end{aligned}$$

Proof. We prove the theorem by induction.
For $k = 1$ see Corollary 1.
$k \Rightarrow k+1$: From the definition of the operators $B^{(k+1)}_{n,\rho}$ we get

$$\begin{aligned} B^{(k+1)}_{n,\rho}(f;x) &= D^1 \circ B^{(k)}_{n,\rho} \circ I_1(f;x) \\ &= \frac{n!(n\rho-1)!}{(n-k)!(n\rho+k-2)!} \sum_{j=0}^{n-k} p'_{n-k,j}(x) \\ &\quad \times \int_0^1 \sum_{i_1=0}^{\rho-1} \cdots \sum_{i_k=0}^{\rho-1} p_{n\rho+k-2,j\rho+i_1+\cdots+i_k+k-1}(t) I_1(f;t)\,dt. \end{aligned}$$

By using (3) and an appropriate index transform we derive

$$B_{n,\rho}^{(k+1)}(f;x) = \frac{n!(n\rho-1)!}{(n-k-1)!(n\rho+k-2)!} \sum_{j=0}^{n-(k+1)} p_{n-(k+1),j}(x) \tag{7}$$
$$\times \int_0^1 \sum_{i_1=0}^{\rho-1} \cdots \sum_{i_k=0}^{\rho-1} \left[p_{n\rho+k-2,(j+1)\rho+i_1+\cdots+i_k+k-1}(t) \right.$$
$$\left. - p_{n\rho+k-2,j\rho+i_1+\cdots+i_k+k-1}(t) \right] I_1(f;t)dt.$$

Now we rewrite the difference of the basis functions in the integral and use again (3), i.e.,

$$\begin{aligned} & p_{n\rho+k-2,(j+1)\rho+i_1+\cdots+i_k+k-1}(t) - p_{n\rho+k-2,j\rho+i_1+\cdots+i_k+k-1}(t) \\ &= \sum_{i_{k+1}=0}^{\rho-1} \left[p_{n\rho+k-2,j\rho+i_1+\cdots+i_{k+1}+k}(t) - p_{n\rho+k-2,j\rho+i_1+\cdots+i_{k+1}+k-1}(t) \right] \\ &= -\frac{1}{n\rho+k-1} \sum_{i_{k+1}=0}^{\rho-1} p'_{n\rho+k-1,j\rho+i_1+\cdots+i_{k+1}+k}(t). \end{aligned}$$

Together with (7) and integration by parts this leads to

$$\begin{aligned} B_{n,\rho}^{(k+1)}(f;x) &= \frac{n!(n\rho-1)!}{(n-k-1)!(n\rho+k-1)!} \sum_{j=0}^{n-(k+1)} p_{n-(k+1),j}(x) \\ &\quad \times \int_0^1 \sum_{i_1=0}^{\rho-1} \cdots \sum_{i_k=0}^{\rho-1} \sum_{i_{k+1}=0}^{\rho-1} -p'_{n\rho+k-1,j\rho+i_1+\cdots+i_{k+1}+k}(t) I_1(f;t)dt \\ &= \frac{n!(n\rho-1)!}{(n-k-1)!(n\rho+k-1)!} \sum_{j=0}^{n-(k+1)} p_{n-(k+1),j}(x) \\ &\quad \times \int_0^1 \sum_{i_1=0}^{\rho-1} \cdots \sum_{i_{k+1}=0}^{\rho-1} p_{n\rho+k-1,j\rho+i_1+\cdots+i_{k+1}+k}(t) f(t)dt. \end{aligned}$$

Remark 1. From (1) we know that for $f \in C[0,1]$

$$\lim_{\rho\to\infty} B_{n,\rho}^{(k)} f = B_{n,\infty}^{(k)} f \tag{8}$$

uniformly on $[0,1]$. Now Theorem 2, (2) and (8) imply

$$\lim_{\rho\to\infty} \frac{(n\rho-1)!}{(n\rho+k-2)!} \sum_{i_1=0}^{\rho-1} \cdots \sum_{i_k=0}^{\rho-1} p_{n\rho+k-2,j\rho+i_1+\cdots+i_k+k-1}(t) = \frac{1}{n^{k-1}} N_{k,j}(t),$$

which can be written also as

$$\lim_{\rho\to\infty} \frac{1}{\rho^{k-1}} \sum_{i_1=0}^{\rho-1} \cdots \sum_{i_k=0}^{\rho-1} p_{n\rho+k-2,j\rho+i_1+\cdots+i_k+k-1}(t) = N_{k,j}(t).$$

6 Convexity of the Linking Operators

In [3, Theorem 4.1] Gonska and Păltănea considered convexity properties of the operators $B_{n,\rho}$ and proved that $(B_{n,\rho}f)^{(r)} \geq 0$ for each function $f \in C^r[0,1]$, $0 \leq r \leq n$, such that $f^{(r)} \geq 0$. By using the same method this can be generalized to $B_{n,\rho}^{(k)}$ (see [1, Theorem 4]). With the representation in Theorem 2 the convexity properties for $B_{n,\rho}^{(k)}$ now follow as a corollary.

Corollary 3. *Let $f \in C^l[0,1]$ with $f^{(l)}(x) \geq 0$, $l \in \mathbb{N}_0$, for all $x \in [0,1]$. Then*

$$D^l\left(B_{n,\rho}^{(k)}(f;x)\right) \geq 0$$

for each $k \in \mathbb{N}$, $x \in [0,1]$.

Proof.

$$\begin{aligned} D^l\left(B_{n,\rho}^{(k)}(f;x)\right) &= D^l \circ B_{n,\rho}^{(k)} \circ I_l f^{(l)} \\ &= B_{n,\rho}^{(k+l)} f^{(l)} \geq 0, \end{aligned}$$

as $f^{(l)} \geq 0$.

7 Concluding Remarks

Similar constructions can be also done for the linking operators acting on the non-compact interval $[0,\infty)$, e.g., the link between the Durrmeyer type operators and Kantorovich modifications of the Szász-Mirakjan operators. This will be treated in a forthcoming paper.

References

1. Baumann, K., Heilmann, M., Raşa, I.: Further results for kth order Kantorovich modification of linking Baskakov type operators. Results Math. **69**(3), 297–315 (2016)
2. Gonska, H., Heilmann, M., Raşa, I.: Kantorovich operators of order k. Numer. Funct. Anal. Optimiz. **32**, 717–738 (2011)
3. Gonska, H., Păltănea, R.: Simultaneous approximation by a class of Bernstein-Durrmeyer operators preserving linear functions. Czechoslovak Math. J. **60**(135), 783–799 (2010)
4. Gould, H.W.: Combinatorial identities: a standardized set of tables listing 500 binomial coefficient summations. Morgantown, W.Va. (1972)
5. Heilmann, M., Raşa, I.: k-th order Kantorovich type modification of the operators U_n^ρ. J. Appl. Funct. Anal. **9**(3–4), 320–334 (2014)
6. Herzog, F., Hill, J.D.: The Bernstein polynomials for discontinuous functions. Am. J. Math. **68**(1), 109–124 (1946)
7. Lorentz, G.G.: Bernstein Polynomials. Chelsea Publishing Company, New York (1986)
8. Păltănea, R.: A class of Durrmeyer type operators preserving linear functions. Ann. Tiberiu Popoviciu Sem. Funct. Eq. Approx. Conv. (Cluj-Napoca) **5**, 109–117 (2007)
9. Schumaker, L.L.: Spline Functions: Basic Theory. Cambridge University Press, Cambridge (2007)

Construction of Fractal Bases for Spaces of Functions

María A. Navascués[1(✉)], María V. Sebastián[2], Arya K.B. Chand[3], and Saurabh Katiyar[4]

[1] Dpto. de Matemática Aplicada, Escuela de Ingeniería y Arquitectura, Universidad de Zaragoza, C/ María de Luna, 3, 50018 Zaragoza, Spain
manavas@unizar.es

[2] Centro Universitario de la Defensa Academia General Militar, Ctra. de Huesca s/n, 50090 Zaragoza, Spain
msebasti@unizar.es

[3] Department of Mathematics, IIT Madras, Office HSB- 254H, Chennai 600036, Tamilnadu, India
chand@iitm.ac.in

[4] Department of Mathematics, IIT Madras, Room number- HSB 241 A, Chennai 600036, Tamilnadu, India
sbhkatiyar@gmail.com
http://pcmap.unizar.es/∼navascues/

Abstract. The construction of fractal versions of classical functions as polynomials, trigonometric maps, etc. by means of a particular Iterated Function System of the plane is tackled. The closeness between the classical function and its fractal analogue provides good properties of approximation and interpolation to the latter. This type of methodology opens the use of non-smooth and fractal functions in approximation. The procedure involves the definition of an operator mapping standard functions into their dual fractals. The transformation is linear and bounded and some bounds of its norm are established. Through this operator we define families of fractal functions that generalize the classical Schauder systems of Banach spaces and the orthonormal bases of Hilbert spaces. With an appropriate election of the coefficients of Iterated Function System we define sets of fractal maps that span the most important spaces of functions as $\mathcal{C}[a,b]$ or $\mathcal{L}^p[a,b]$.

Keywords: Fractal interpolation functions · Bases of functional spaces · Approximation · Interpolation · Fractals

AMS subject classifications: 28A80 · 41A10 · 58C05 · 65D05 · 26A27.

1 Introduction

The functional space $\mathcal{C}[a,b]$, composed of continuous functions defined on a compact interval, and endowed with the supremum norm, is one of the most popular

© Springer Nature Singapore Pte Ltd. 2017
D. Giri et al. (Eds.): ICMC 2017, CCIS 655, pp. 321–330, 2017.
DOI: 10.1007/978-981-10-4642-1_27

infinite dimensional Banach spaces. This set is used in almost all the fields of mathematics and broad areas of physics. Let us recall, for instance, that some of its bases facilitate the construction of random motions [2].

In many problems of applied mathematics, one needs to approximate continuous functions. A feasible way of doing this is the use of spanning systems. The natural generalization of a basis of a finite-dimensional space is a Schauder system. A sequence (x_m) of a Banach space X is a Schauder basis if every element x of X is expressed univocally as $x = \sum c_m x_m$, where c_m are scalar magnitudes. Every m-th sum of the series is an approximate value of x which may be more treatable than x.

It was early proved that $\mathcal{C}[a, b]$ owns a basis. The first system of this type was studied by G. Faber [4] and J. Schauder [12]. It comprises polygonal functions associated to a sequence that is everywhere dense in the interval. If $a = 0$, and $b = 1$, the sequence consists sometimes of dyadic rational points, first appeared in the work of Faber. Later on, the existence of bases was extended to Lebesgue spaces of p-integrable maps $\mathcal{L}^p$ where $p \geq 1$ (see for instance [5]). However this is not true for general Banach spaces, as proved by Enflo [3].

Among all the spanning systems, the interpolating families possess even more interesting properties. A basis is of this kind if the m-th approximation $S_m x$ agrees with x at some points $t_1, t_2, \ldots, t_m$. For instance, the Faber-Schauder basis is interpolatory.

Our study on fractal functions have prompted us to define mappings that are close to the classical but, at the same time, own a self-similar structure in their traces (graphs). Through these new elements, the fields of interpolation and approximation may be expanded. Our late concern is the construction of bases of functional spaces, composed of fractal functions. In the references [9,10], we defined interpolatory bases of affine fractal functions for $\mathcal{C}[a, b]$. The mappings involved are perturbations of those belonging to the Faber-Schauder family. Every map $x_m(t)$ is associated to a scale vector α_m (see next Section for its definition). The sequence α_m must tend to zero as m tends to infinity. In this instance we define a fractal basis for the same space with respect to a constant scale vector α, independent of the function. The approach can be extended to Lebesgue spaces $\mathcal{L}^p[a, b]$, for $1 \leq p < \infty$.

2 Fractal Functions Associated with Classical Continuous Maps

In this Section we describe the construction of a fractal interpolation function close to any continuous mapping $f : [a, b] \to \mathbf{R}$ (see Figs. 1 and 2).

Let $t_0 < t_1 < \ldots < t_N$ be real numbers, $N > 1$, and $I = [t_0, t_N] = [a, b]$ the smallest closed interval that contains them. Let a set of data points $(x_n)_{n=0}^N$ be given. Set $I_n = [t_{n-1}, t_n]$ and let $L_n : I \to I_n$, $n \in \{1, 2, \ldots, N\}$ be contractive homeomorphisms such that:

$$L_n(t_0) = t_{n-1}, \; L_n(t_N) = t_n \tag{1}$$

$$|L_n(c_1) - L_n(c_2)| \leq l\ |c_1 - c_2| \quad \forall\ c_1, c_2\ \in I \tag{2}$$

for some $l \in [0, 1)$.

Let $H = I \times \mathbf{R}$ and N continuous mappings, $F_n : H \to \mathbf{R}$, be given satisfying:

$$F_n(t_0, x_0) = x_{n-1}, \quad F_n(t_N, x_N) = x_n, \quad n = 1, 2, ..., N \tag{3}$$

$$|F_n(t, x) - F_n(t, y)| \leq r|x - y|, \quad t \in I, \quad x, y \in \mathbf{R}, \quad 0 \leq r < 1. \tag{4}$$

Now define functions $w_n(t, x) = (L_n(t), F_n(t, x)),\ \forall n = 1, 2, ..., N$. The next and other related results can be read in [1].

Theorem 1. *The Iterated Function System (IFS)* $\{H,\ w_n : n = 1, 2, ..., N\}$ *defined above admits a unique attractor* G. G *is the graph of a continuous function* $g : I \to \boldsymbol{R}$ *which obeys* $g(t_n) = x_n$ *for* $n = 0, 1, 2, ..., N$.

The previous function is called a Fractal Interpolation Function (FIF) corresponding to $\{(L_n(t), F_n(t, x))\}_{n=1}^{N}$. The mapping g satisfies the functional Eq. [1]:

$$g(t) = F_n(L_n^{-1}(t), g \circ L_n^{-1}(t)), \quad n = 1, 2, ..., N, \quad t \in I_n = [t_{n-1}, t_n]. \tag{5}$$

The most widely studied fractal interpolation functions so far are defined by the IFS

$$\begin{cases} L_n(t) = a_n t + b_n \\ F_n(t, x) = \alpha_n x + q_n(t) \end{cases} \tag{6}$$

where $-1 < \alpha_n < 1$, $n = 1, 2, ..., N$. α_n is called a vertical scaling factor of the transformation w_n. It follows from (1) that

$$a_n = \frac{t_n - t_{n-1}}{t_N - t_0} \qquad b_n = \frac{t_N t_{n-1} - t_0 t_n}{t_N - t_0}. \tag{7}$$

Let $f \in \mathcal{C}(I)$ be a continuous function. We consider the case

$$q_n(t) = f \circ L_n(t) - \alpha_n b(t), \tag{8}$$

where b is continuous and such that $b(t_0) = x_0$, $b(t_N) = x_N$. The set of data is here $\{(t_n, f(t_n)) : n = 0, 1, \ldots, N\}$. Using this IFS one can define fractal analogues of any continuous function [7]. In particular, we consider in this paper the case

$$b = Lf \tag{9}$$

where L is an operator of $\mathcal{C}(I)$ linear and bounded with respect to the uniform norm:

$$\|f\|_\infty = \sup\{|f(x)| : x \in I\}.$$

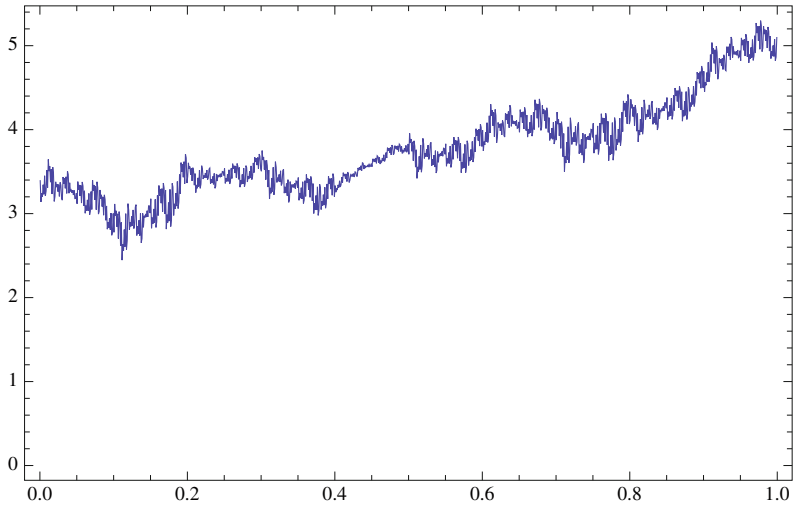


Fig. 1. Fractal function associated to a polygonal with vertices (0, 3.3), (0.1, 2.9), (0.2, 3.5), (0.3, 3.6), (0.4, 3.3), (0.5, 3.8), (0.6, 3.9), (0.7, 4), (0.8, 4.2), (0.9, 4.6), (1, 5).

Definition 1. *Let f^α be the continuous function defined by the IFS (6), (7), (8) and (9). f^α is the α-fractal function associated with f with respect to L and the partition Δ.*

f^α satisfies the fixed point Eq. (5):

$$f^\alpha(t) = f(t) + \alpha_n(f^\alpha - Lf) \circ L_n^{-1}(t), \tag{10}$$

for $t \in I_n$. From here on we will use the following notation:

$$|\alpha|_\infty = \max\{|\alpha_n| : n = 1, 2, \ldots, N\}$$

and assume that $|\alpha|_\infty < 1$.

Figure 1 represents an affine fractal function (associated with a polygonal f) with the following parameters: a uniform partition of the interval $I = [0, 1]$ of 10 subintervals $(x_n) = (3.3, 2.9, 3.5, 3.6, 3.3, 3.8, 3.9, 4, 4.2, 4.6, 5)$; the operator $Lf(t) = r(t)$, where $r(t)$ is the line passing through the extreme data $(t_0, f(t_0))$, $(t_N, f(t_N))$ and $\alpha = (-0.3, 0.4, 0.2, 0.3, -0.1, 0.3, -0.3, 0.4, 0.3, -0.3)$.

The following inequalities can be deduced easily from the fixed point Eq. (10):

$$\|f^\alpha - f\|_\infty \leq |\alpha|_\infty \|f^\alpha - Lf\|_\infty, \tag{11}$$

$$\|f^\alpha - f\|_\infty \leq \frac{|\alpha|_\infty}{1 - |\alpha|_\infty} \|f - Lf\|_\infty \tag{12}$$

for all $f \in \mathcal{C}(I)$.

Let us define the α-fractal operator with respect to Δ and L as:

$$\begin{aligned} \mathcal{F}^\alpha : \mathcal{C}(I) &\to \mathcal{C}(I) \\ f &\hookrightarrow f^\alpha \end{aligned}$$

and let us denote as $\|\mathcal{F}^\alpha\|$, $\|L\|$ the operator norms with respect to the uniform metric in $\mathcal{C}(I)$. The properties of $\mathcal{F}^\alpha$ are summarized in the next Theorem [7].

Theorem 2. *If L is linear and bounded with respect to the uniform metric:*

(a) $\mathcal{F}^\alpha$: $\mathcal{C}(I) \to \mathcal{C}(I)$ is linear and injective.
(b) If $\alpha = 0$, $\mathcal{F}^\alpha = \mathcal{I}$, where $\mathcal{I}$ is the identity operator.
(c) The operator $\mathcal{F}^\alpha$ is bounded and the following inequalities hold:

$$\|\mathcal{F}^\alpha\| \leq 1 + \frac{|\alpha|_\infty}{1 - |\alpha|_\infty} \|\mathcal{I} - L\| \tag{13}$$

$$\|\mathcal{I} - \mathcal{F}^\alpha\| \leq \frac{|\alpha|_\infty}{1 - |\alpha|_\infty} \|\mathcal{I} - L\|. \tag{14}$$

(d) If $\alpha \neq 0$, the fixed points of $\mathcal{F}^\alpha$ agree with the fixed points of L.
(e) If $\alpha \neq 0$, $\mathcal{F}^\alpha = \mathcal{I}$ if and only if $L = \mathcal{I}$.
(f) If 1 belongs to the point spectrum of L (L has non-null fixed points) then

$$1 \leq \|\mathcal{F}^\alpha\|.$$

The following result can be found in [7].

Proposition 1. *If $|\alpha|_\infty < \|L\|^{-1}$ then $\mathcal{F}^\alpha$ has closed range.*

Another important property, is the following chain of inequalities:

$$\frac{1 - |\alpha|_\infty \|L\|}{1 + |\alpha|_\infty} \|f\|_\infty \leq \|\mathcal{F}^\alpha(f)\|_\infty \leq \frac{1 + |\alpha|_\infty \|L\|}{1 - |\alpha|_\infty} \|f\|_\infty. \tag{15}$$

3 Fractal Bases of Continuous and Integrable Functions

Our objective is now the construction of fractal bases for spaces of functions. We recall the following definitions.

Definition 2. *A sequence $(x_m)_{m=0}^\infty$ of a Banach space X is a Schauder basis if $\forall x \in X$ there exists a unique representation of x as*

$$x = \sum_{m=0}^{\infty} c_m x_m,$$

where $(c_m)_{m=0}^\infty$ is a sequence of scalars.

Examples: The space $\mathcal{C}(I)$ possesses a basis of polynomials. Another important basis is composed of polygonal (triangular) functions attached to a dense sequence of points of the interval (Faber-Schauder system).

A Schauder basis is the natural generalization of a basis of a finite-dimensional vector space. It enables the identification of an element with a sequence of scalars:

$$x \approx (c_m)_{m=0}^{\infty} \Leftrightarrow x = \sum_{m=0}^{\infty} c_m x_m.$$

c_m is the m-th "coordinate" of x with respect to the basis $(x_m)_{m=0}^{\infty}$.

Definition 3. *A sequence $(x_m)_{m=0}^{\infty}$ of a Banach space is a Schauder sequence if it is a Schauder basis for $[x_m]_{m=0}^{\infty} = \overline{span}(x_m)_{m=0}^{\infty}$.*

Remark: The set $span(x_m)_{m=0}^{\infty}$ is the family of finite linear combinations of the elements x_m and $[x_m]_{m=0}^{\infty}$ is the topological closure of $span(x_m)_{m=0}^{\infty}$.

3.1 Space of Continuous Functions on a Compact Interval $\mathcal{C}(I)$

In this Subsection we define a fractal basis for the space $\mathcal{C}(I)$.

Theorem 3. *If $(f_m)_{m=0}^{\infty}$ is a Schauder basis of $\mathcal{C}(I)$ and $|\alpha|_\infty < \|L\|^{-1}$, then $(\mathcal{F}^\alpha(f_m))_{m=0}^{\infty}$ is a Schauder sequence.*

Proof. Let the range of $\mathcal{F}^\alpha$ be denoted by $rg(\mathcal{F}^\alpha)$. With the hypothesis on the scale vector $rg(\mathcal{F}^\alpha)$ is closed (Proposition 1) and it is easy to check that $[\mathcal{F}^\alpha(f_m)]_{m=0}^{\infty}$ agrees with it. $\mathcal{F}^\alpha$ is a topological isomorphism onto $rg(\mathcal{F}^\alpha)$ and this transformation preserves the bases. Consequently $(\mathcal{F}^\alpha(f_m))_{m=0}^{\infty}$ is a Schauder basis of $[\mathcal{F}^\alpha(f_m)]_{m=0}^{\infty} = rg(\mathcal{F}^\alpha)$.

Consequences: If $(f_m)_{m=0}^{\infty}$ is a Schauder basis of $\mathcal{C}(I)$ and $|\alpha|_\infty < \|L\|^{-1}$ then, the fact that $\mathcal{F}^\alpha(f_m)$ is a basis implies that

- The fractal system is finitely linearly independent.
- It is complete sequence.
- The series $\sum_{m=0}^{\infty} c_m \mathcal{F}^\alpha(f_m)$ converges if and only if $\sum_{m=0}^{\infty} c_m f_m$ is convergent (due to the isomorphism between $[f_m]_{m=0}^{\infty}$ and $[f_m^\alpha]_{m=0}^{\infty}$).

Lemma 1. [6] *If L is a bounded and linear operator from a Banach space into itself such that $\|\mathcal{I} - L\| < 1$, then L^{-1} exists and is bounded.*

Theorem 4. *If (f_m) is a Schauder basis of $\mathcal{C}(I)$ and $|\alpha|_\infty < (1 + \|\mathcal{I} - L\|)^{-1}$, then $(\mathcal{F}^\alpha(f_m))$ is a Schauder basis of $\mathcal{C}(I)$.*

Proof. The hypothesis on the scale vector implies that (14)

$$\|\mathcal{I} - \mathcal{F}^\alpha\| \leq \frac{|\alpha|_\infty}{1 - |\alpha|_\infty} \|\mathcal{I} - L\| < 1.$$

$\mathcal{F}^\alpha$ has a bounded inverse, according to the previous Lemma and the result is deduced.

Definition 4. *A basis (x_m) of a Banach space is bounded if*

$$0 < \inf \|x_m\| \leq \sup \|x_m\| < \infty.$$

The inequalities (15) imply that the fractal basis will be bounded if the original is. In particular, the affine basis associated to the classical polygonal basis of $\mathcal{C}(I)$ is bounded as

$$\frac{1 - |\alpha|_\infty \|L\|}{1 + |\alpha|_\infty} \leq \|\mathcal{F}^\alpha(f_m)\|_\infty \leq \frac{1 + |\alpha|_\infty \|L\|}{1 - |\alpha|_\infty}, \tag{16}$$

since $\|f_m\| = 1$ for any m.

These results prove that the sequence $(f_m^\alpha)_{m=0}^\infty$ of fractal functions associated to a basis is another basis if the scale vector is suitable chosen.

3.2 Space of p-integrable Functions $\mathcal{L}^p(I)$

In this Subsection we extend the concept of α-fractal function to the space $\mathcal{L}^p(I)$, for $1 \leq p < \infty$ and define fractal bases for this set.

Let us consider now the norm in $\mathcal{C}(I)$:

$$\|f\|_{\mathcal{L}^p} = \left(\int_a^b |f|^p dt \right)^{1/p} < \infty,$$

and assume that L is bounded with respect to $\|\cdot\|_p$. The properties of $\|\mathcal{F}^\alpha\|_p$ (norm associated) are very similar to those described in Theorem 2 [7]. In particular, the following bounds hold:

$$\|\mathcal{F}^\alpha\|_p \leq 1 + \frac{|\alpha|_\infty}{1 - |\alpha|_\infty} \|\mathcal{I} - L\|_p \tag{17}$$

$$\|\mathcal{I} - \mathcal{F}^\alpha\|_p \leq \frac{|\alpha|_\infty}{1 - |\alpha|_\infty} \|\mathcal{I} - L\|_p. \tag{18}$$

The following results generalize the concept of α-fractal function: any $f \in \mathcal{L}^p(I)$ will be associated with a function $\overline{f}^\alpha \in \mathcal{L}^p(I)$, for $1 \leq p < \infty$.

Figure 2 represents the fractal function associated with the Legendre polynomial of degree 3 with the following parameters: a uniform partition of the interval $I = [-1, 1]$ of 10 subintervals, the operator $Lf(t) = f(t)v(t)$, where $v(t) = \cos(2\pi t)$ and $\alpha_n = 0.3$ for all $n = 1, 2, \ldots 10$.

Lemma 2. *(Linear and Bounded Operator Theorem) [6] If an operator $S : X \to Y$ is linear and bounded, Y is Banach and X is dense in X', then S can be extended to X' preserving the norm of S.*

Remark 1. The extension $\overline{S} : X' \to Y$ is defined in the following way:

If $x' \in X'$, due to the density of X in X', there exists a sequence $(x_m) \subset X$ such that $\lim x_m = x'$. The image $\overline{S}(x')$ is then defined as $\overline{S}(x') = \lim S(x_m)$.

Since $\mathcal{C}(I)$ is dense in $\mathcal{L}^p(I)$ for $p \in [1, +\infty)$ with respect to the p-metric [11], one can use the previous Lemma to extend the operators $\mathcal{F}^\alpha$ and L to $\mathcal{L}^p(I)$ preserving the norm. Thus if $\overline{\mathcal{F}}^\alpha : \mathcal{L}^p(I) \to \mathcal{L}^p(I)$ and $\overline{L} : \mathcal{L}^p(I) \to \mathcal{L}^p(I)$ are the corresponding extensions:

$$\|\overline{\mathcal{F}}^\alpha\|_p = \|\mathcal{F}^\alpha\|_p$$

and

$$\|\overline{L}\|_p = \|L\|_p.$$

As a consequence, bearing in mind (17):

$$\|\overline{\mathcal{F}}^\alpha\|_p = \|\mathcal{F}^\alpha\|_p \leq 1 + \frac{|\alpha|_\infty}{1 - |\alpha|_\infty} \|\mathcal{I} - L\|_p \tag{19}$$

By construction $\overline{f}^\alpha = \overline{\mathcal{F}}^\alpha(f)$ is the limit of a sequence of continuous α-fractal functions $(S_m^\alpha) \subset \mathcal{C}(I)$:

$$\overline{f}^\alpha = \lim S_m^\alpha = \lim \mathcal{F}^\alpha(S_m).$$

The function $\overline{f}^\alpha$ will be the α-fractal function of $f \in \mathcal{L}^p(I)$. The properties of $\mathcal{F}^\alpha$ and L are extended to $\overline{F}^\alpha$ and $\overline{L}$ [7]. For instance, for any $f \in \mathcal{L}^p(I)$, $p \in [1, +\infty)$,

$$\|\overline{\mathcal{F}}^\alpha(f) - f\|_{\mathcal{L}^p} \leq |\alpha|_\infty \|\overline{\mathcal{F}}^\alpha(f) - \overline{L}f\|_{\mathcal{L}^p}, \tag{20}$$

$$\|\overline{\mathcal{F}}^\alpha(f) - f\|_{\mathcal{L}^p} \leq \frac{|\alpha|_\infty}{1 - |\alpha|_\infty} \|f - \overline{L}f\|_{\mathcal{L}^p}. \tag{21}$$

If $|\alpha|_\infty < \|L\|_p^{-1}$ then $\overline{\mathcal{F}}^\alpha$ is injective and its range is closed. Using arguments similar to those exposed in the previous Subsection, one has the following results.

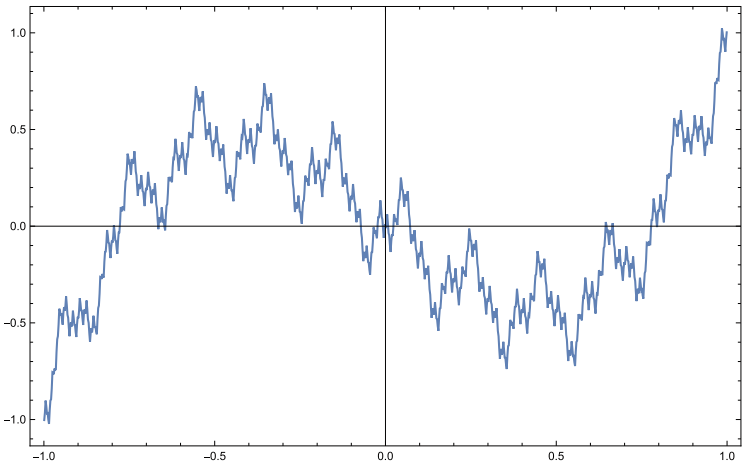


Fig. 2. Fractal function (f^α) associated with the third Legendre polynomial (f).

Theorem 5. *If $(f_m)_{m=0}^{\infty}$ is a Schauder basis of $\mathcal{L}^p(I)$ $(1 \leq p < \infty)$ and $|\alpha|_\infty < \|L\|_p^{-1}$, then $(\overline{\mathcal{F}}^{\alpha}(f_m))_{m=0}^{\infty}$ is a Schauder sequence.*

Theorem 6. *If $(f_m)_{m=0}^{\infty}$ is a Schauder basis of $\mathcal{L}^p(I)$ $(1 \leq p < \infty)$ and $|\alpha|_\infty < (1 + \|I - L\|_p)^{-1}$, then $(\overline{\mathcal{F}}^{\alpha}(f_m))_{m=0}^{\infty}$ is a Schauder basis of $\mathcal{L}^p(I)$.*

The case $p = 2$ owns some peculiarities due to the inner product operation, with the consequent structure of Hilbert space. The following results can be found in [8], Subsect. 2.3. Let $(p_m)_{m=0}^{\infty}$ be an orthonormal basis of $\mathcal{L}^2(I)$ (for instance, the system of Legendre polynomials) and $(p_m^{\alpha})_{m=0}^{\infty}$ the image sequence:

$$(p_m^{\alpha})_{m=0}^{\infty} = (\overline{\mathcal{F}}^{\alpha}(p_m))_{m=0}^{\infty}.$$

Definition 5. *A sequence $(x_m) \subseteq H$, where H is a Hilbert space, is a Bessel sequence if there exists a constant $B > 0$ such that for all $x \in H$*

$$\sum_{m=0}^{\infty} | < x, x_m > |^2 \leq B\|x\|^2.$$

Proposition 2. *For any scale vector α such that $|\alpha|_\infty < 1$, if $(p_m)_{m=0}^{\infty}$ is an orthonormal basis, $(p_m^{\alpha})_{m=0}^{\infty}$ is a Bessel sequence.*

Consequence: If $(c_m)_{m=0}^{\infty} \in l^2$ the series

$$\sum_{m=0}^{\infty} c_m p_m^{\alpha}$$

is unconditionally convergent due to the fact that $\sum c_m p_m$ is unconditional (all the orthonormal bases are unconditional).

Definition 6. *A sequence $(x_m)_{m=0}^{\infty} \subseteq H$, where H is a Hilbert space, is a Riesz basis if it is equivalent to an orthonormal basis $(y_m)_{m=0}^{\infty}$ of H, that is to say, there exists an operator T linear, bijective and bicontinuous (topological isomorphism) such that $Tx_m = y_m$.*

Definition 7. *A sequence $(x_m)_{m=0}^{\infty} \subseteq H$, where H is a Hilbert space, is a Riesz sequence if there exist $k_1, k_2 > 0$ such that for any $(c_m) \in l^2$*

$$k_1 \sum_{m=0}^{\infty} |c_m|^2 \leq \| \sum_{m=0}^{\infty} c_m x_m \|^2 \leq k_2 \sum_{m=0}^{\infty} |c_m|^2. \tag{22}$$

Proposition 3. *If $|\alpha|_\infty < \|L\|_2^{-1}$, $(p_m^{\alpha})_{m=0}^{\infty}$ is a Riesz sequence.*

And finally,

Theorem 7. *If $|\alpha|_\infty < (1+\|\mathcal{I}-L\|_2)^{-1}$ then $(p_m^{\alpha})_{m=0}^{\infty}$ is a Riesz basis of $\mathcal{L}^2(I)$.*

References

1. Barnsley, M.F.: Fractal functions and interpolation. Constr. Approx. **2**, 303–329 (1986)
2. Ciesielski, Z.: Hölder conditions for realization of Gaussian processes. Trans A.M.S. **99**, 403–413 (1961)
3. Enflo, P.: A counterexample to the approximation property in Banach spaces. Acta Math. **130**, 309–317 (1973)
4. Faber, G.: Ueber die orthogonalfunktionen des Hern Haar Jahresber. Deutsch. Math. Verein. **19**, 104–112 (1910)
5. Johnson, W.B., Rosenthal, H.P., Zippin, M.: On bases, finite dimensional decompositions and weaker structures in Banach spaces. Israel J. Math. **9**, 488–506 (1971)
6. Lebedev, L.P., Vorovich, I.I., Gladwell, G.M.L.: Functional Analysis. Applications in Mechanics and Inverse Problems, 2nd edn. Kluwer Academic Publishers, Dordrecht (2002)
7. Navascués, M.A.: Fractal approximation. Complex Anal. Oper. Th. **4**(4), 953–974 (2010)
8. Navascués, M.A.: Fractal bases of Lp spaces. Fractals **20**(2), 141–148 (2012)
9. Navascués, M.A.: Affine fractal functions as bases of continuous functions. Quaestiones Math. **37**, 1–14 (2014)
10. Navascués, M.A., Sebastián, M.V.: Construction of affine fractal functions close to classical interpolants. J. Comp. Anal. Appl. **9**(3), 271–283 (2007)
11. Reed, M., Simon, B.: Methods of Modern Mathematical Physics, I, Functional Analysis, 2nd edn. Academic Press, New York (1980)
12. Schauder, J.: Eine eigenschaft des haarschen orthogonalsystems. Math. Z. **28**, 317–320 (1928)

Infinite Matrices Bounded on Weighted c_0 Space

Riddhick Birbonshi, Arnab Patra(✉), and P.D. Srivastava

Department of Mathematics, Indian Institute of Technology Kharagpur, Kharagpur 721 302, West Bengal, India
riddhick.math@gmail.com, arnptr91@gmail.com, pds@maths.iitkgp.ernet.in

Abstract. In this paper some necessary and sufficient conditions for boundedness of an infinite matrix as a linear operator between two weighted c_0 spaces are established. Some relationship between the matrix and the weight vectors of domain and range spaces are also obtained.

Keywords: Infinite matrix · Weighted sequence space · Matrix norm · c_0 space

1 Introduction

Infinite matrices play an important role in difference equations, integral equations, infinite systems of linear algebraic or differential equations, the theory of summability of sequences and series. For detail study about infinite matrix we refer the book of Cooke [4] and for a brief review we refer to Shivakumar and Sivakumar [6].

An infinite matrix defines a linear operator on a sequence space but for a given any infinite matrix, it is not always easy to find whether it arises from a bounded operator or not. Many authors have obtained the spectrum of some operators on the c_0 space using the conditions of boundedness of an infinite matrix on c_0 [1–3,5]. So it is interesting to know the conditions under which an infinite matrix behaves as a bounded linear operator on sequence spaces such as weighted c_0 space. The conditions for boundedness of an infinite matrix on weighted l_1 space are obtained by Joseph J. Williams and Qiang Ye [7]. In this paper we find the condition under which an infinite matrix map is bounded between two weighted c_0 spaces with two different weights. Further for a given infinite matrix and a weight vector of domain, we have obtained the condition on the weight vector of range space such that the matrix is bounded. Also some necessary and sufficient conditions under which an infinite matrix is bounded for the weighted c_0 space having same weight are investigated.

2 Preliminaries and Notations

All infinite sequences and matrices throughout the paper are assumed to be indexed by $\mathbb{N}$. $c_0 = \{(x_k) : \lim_{k\to\infty} x_k = 0\}$, space of all null sequences of real

© Springer Nature Singapore Pte Ltd. 2017
D. Giri et al. (Eds.): ICMC 2017, CCIS 655, pp. 331–338, 2017.
DOI: 10.1007/978-981-10-4642-1_28

or complex numbers. We denote an infinite matrix A as $A = (a_{nk})$, $n, k = 1, 2, 3, \cdots$, an infinite sequence $x = (x_j)$, $j = 1, 2, 3, \cdots$ as a column vector and the multiplication Ax is defined as $(Ax)_i = \sum_{j=1}^{\infty} a_{ij}x_j$ for each $i \in \mathbb{N}$, where the series converges for each $i \in \mathbb{N}$. Throughout this paper, $|A|$ denotes the matrix of absolute values of A, that is $|A| = (|a_{nk}|)$. Inequalities on two real vectors or matrices are defined by component-wise, that is $x \leq y$ means $x_j \leq y_j$ for all $j \in \mathbb{N}$. If E and F are any two Banach spaces, $B(E, F)$ denotes the set of all bounded linear operators from E into F and the (operator) norm of $T \in B(E, F)$ is given by $\|T\| = \sup\{\|Tx\|_F : \|x\|_E \leq 1, x \in E\}$. For $E = F$, $B(E, F)$ is denoted by $B(E)$. Now we procure a lemma about the boundedness of an infinite matrix from c_0 to itself.

Lemma 1 [5]. *The matrix $A = (a_{nk})$ gives rise to a bounded linear operator $T \in B(c_0)$ from c_0 to itself if and only if, $\sup_{n\in\mathbb{N}} \sum_{k=1}^{\infty} |a_{nk}| < \infty$ and the columns are in c_0 which means, $\lim_{n\to\infty} a_{nk} = 0$ for each $k \in \mathbb{N}$. The operator norms of A is given by $\|A\| = \sup_{n\in\mathbb{N}} \sum_{k=1}^{\infty} |a_{nk}|$.*

3 Results

Before going to the results, first we define the weighted c_0 space.

Definition 1. *Let $r = (r_k)$ be an infinite positive real sequence, then the weighted c_0 space is defined as $c_0(r) = \{(x_k) : \lim_{k\to\infty} r_k x_k = 0\}$ and $\|x\|_r = \sup_k |x_k| r_k$.*

Remark 1. $\|x\|_r$ defines a norm on $c_0(r)$ and $(c_0(r), \|x\|_r)$ is a Banach space under this norm. It is easy to verify that if $D(r)$ is the diagonal matrix with *ith* diagonal entry r_i then, $D(r)$ is an isometric isomorphism from $c_0(r)$ to c_0 since $x = (x_k) \in c_0(r)$ then, $D(r)x = (x_k r_k) \in c_0$, and $\|x\|_r = \|D(r)x\|$.

Lemma 2. *For every infinite sequence $y = (y_k)$ there exists some weight vector $r = (r_k)$ such that $y \in c_0(r)$.*

Proof. Given $y = (y_k)$, we define $r = (r_k)$ by

$$r_k = \begin{cases} \frac{1}{k}\left(\frac{1}{|y_k|}\right), & y_k \neq 0 \\ 1, & y_k = 0 \end{cases}$$

Clearly, $(y_k) \in c_0(r)$ and $\|y\|_r = \sup_k |r_k y_k| = \sup_k \frac{1}{k} = 1$.

Theorem 1. *For two weight vectors $r = (r_n)$ and $s = (s_n)$ and an infinite matrix $A = (a_{nk})$, the matrix $A \in B(c_0(r), c_0(s))$ if and only if, $\sup_{n\in\mathbb{N}} s_n \sum_{k=1}^{\infty} |\frac{a_{nk}}{r_k}| < \infty$ and $\lim_{n\to\infty} s_k a_{nk} = 0 \ \forall\ k \in \mathbb{N}$ and $\|A\|_{r,s} = \sup_{n\in\mathbb{N}} s_n \sum_{k=1}^{\infty} |\frac{a_{nk}}{r_k}|$.*

Proof. Let $D(r)$ be the diagonal matrix with *ith* diagonal entry r_i then D_r is an isometric isomorphism from $c_0(r)$ to c_0. Now we define $T = D_s A D_r^{-1}$.

Then $A \in B(c_0(r), c_0(s))$ if and only if, $T \in B(c_0, c_0)$.

That is, if and only if, $\sup_{n\in\mathbb{N}} s_n \sum_{k=1}^{\infty} |\frac{a_{nk}}{r_k}| < \infty$ and $\lim_{n\to\infty} \frac{s_n a_{nk}}{r_k} = 0 \ \forall\ k \in \mathbb{N}$.

$\iff \sup_{n\in\mathbb{N}} s_n \sum_{k=1}^{\infty} |\frac{a_{nk}}{r_k}| < \infty$ and $\lim_{n\to\infty} s_n a_{nk} = 0 \ \forall\ k \in \mathbb{N}$.

and $\|A\|_{r,s} = \sup_{n\in\mathbb{N}} s_n \sum_{k=1}^{\infty} |\frac{a_{nk}}{r_k}|$.

In Theorem 1 the vectors r and s are given. In next lemma we consider the case where the weight vector r of domain and the matrix A are given and we obtain condition under which the existence of the weight vector s is guaranteed such that $A \in B(c_0(r), c_0(s))$.

Lemma 3. *Given an infinite matrix $A = (a_{nk})$ and a weight vector $r = (r_n)$, then the matrix $A \in B(c_0(r), c_0(s))$ for some weight vector $s = s_n$ if and only if, $\sum_{k=1}^{\infty} |\frac{a_{nk}}{r_k}| < \infty$ for all n.*

Proof. First suppose that such a $s = (s_n), s_n > 0$ exist. Then by Theorem 1, $\|A\|_{r,s} = \sup_{n\in\mathbb{N}} s_n \sum_{k=1}^{\infty} |\frac{a_{nk}}{r_k}| < \infty$ and $\lim_{n\to\infty} s_n a_{nk} = 0 \ \forall\ k \in \mathbb{N}$. Now for arbitrary $n \in \mathbb{N}$, $\|A\|_{r,s} = \sup_{n\in\mathbb{N}} s_n \sum_{k=1}^{\infty} \left|\frac{a_{nk}}{r_k}\right| \geq s_n \sum_{k=1}^{\infty} \left|\frac{a_{nk}}{r_k}\right|$. Thus $\sum_{k=1}^{\infty} \left|\frac{a_{nk}}{r_k}\right| \leq \frac{\|A\|_{r,s}}{s_n} < \infty \ \forall\ n \in \mathbb{N}$.

Conversely, suppose that $\sum_{k=1}^{\infty} |\frac{a_{nk}}{r_k}| < \infty \ \forall\ n$ is true. Now define $s = (s_n)$ as

$$s_n = \begin{cases} \frac{1}{n}\left(\frac{1}{\sum_k \left|\frac{a_{nk}}{r_k}\right|}\right), & \text{if } a_{nk} \neq 0 \text{ for at least one } k \\ 1, & \text{if } a_{nk} = 0 \text{ for all } k. \end{cases}$$

Then $\sup_{n\in\mathbb{N}} \sum_{k=1}^{\infty} \left|\frac{s_n a_{nk}}{r_k}\right| = \sup_n \frac{1}{n} = 1 < \infty$ and $\lim_{n\to\infty} \left|\frac{s_n a_{nk}}{r_k}\right| = 0$ i.e., $\lim_{n\to\infty} s_n a_{nk} = 0 \ \forall\ k \in \mathbb{N}$.

Now by Theorem 1 $\exists\ s = (s_n)$ s.t. $A \in B(c_0(r), c_0(r))$.

4 Necessary Condition for A to Be in $B(c_0(r))$

In this section we obtain some necessary conditions under the assumption $r = s$ so that $A \in B(c_0(r))$. First we give an example for which $A \notin B(c_0)$ but $A \in B(c_0(r))$ for some $r = (r_n)$. For a matrix $A \in B(c_0(r))$ we denote its norm by $\| A \|_r = \| A \|_{r,r}$.

Example 1. Define $A = (a_{nk})$ by

$$a_{nk} = \begin{cases} k & \text{if } n = 1 \\ \frac{1}{n^4} & \text{if } k = 1 \\ 0 & \text{if } n, k \geq 2 \end{cases}$$

Now $\sum_{k=1}^{\infty} |a_{1k}| = \infty$, thus $A \notin B(c_0)$ from Lemma 1. Now define $r = (r_k)$ such that $r_k = k^3$ and

$$\left| \frac{r_n a_{nk}}{r_k} \right| = \begin{cases} \frac{1}{k^2} & \text{if } n = 1 \\ \frac{1}{n} & \text{if } k = 1 \\ 0 & \text{if } n, k \geq 2 \end{cases}$$

Therefore $\| A \|_r = \sup_{n\in\mathbb{N}} \sum_{k=1}^{\infty} \left| \frac{r_n a_{nk}}{r_k} \right| = \sup \left\{ \sum_{k=1}^{\infty} \frac{1}{k^2}, \frac{1}{2}, \frac{1}{3}, \cdots \right\} < \infty$, also $\lim_{n\to\infty} a_{nk} r_n = 0, \forall k$. Hence $A \in B(c_0(r))$ by Theorem 1.

Here we discuss about the necessary condition for $A \in B(c_0(r))$.

Theorem 2. *If the infinite matrix $A = (a_{nk}) \in B(c_0(r))$ and suppose that $m = \| A \|_r = \sup_{n\in\mathbb{N}} \sum_{k=1}^{\infty} \left| \frac{r_n a_{nk}}{r_k} \right|$ then,*

1. $\sup_{k\in\mathbb{N}} \sup_{n\in\mathbb{N}} |a_{kn} a_{nk}| \leq m^2 < \infty$.
2. $\lim_{n\to\infty} a_{nk} a_{kn} = 0, \quad \forall k$.

Proof. (1) We have $m = \sup_{n\in\mathbb{N}} \sum_{k=1}^{\infty} \left| \frac{r_n a_{nk}}{r_k} \right| \geq \sup_{n\in\mathbb{N}} \left| \frac{r_n a_{nk}}{r_k} \right| \forall k$. Interchanging n and k we have, $\sup_{k\in\mathbb{N}} \left| \frac{r_k a_{kn}}{r_n} \right| \leq m \quad \forall n$. Therefore $\left| \frac{r_k a_{kn}}{r_n} \right| \leq m \quad \forall n, k$.

Now, $|a_{kn} a_{nk}| = \left| \frac{r_k a_{kn}}{r_n} \right| \left| \frac{r_n a_{nk}}{r_k} \right| \leq m \left| \frac{r_n a_{nk}}{r_k} \right|$. Therefore, $\sup_{n\in\mathbb{N}} |a_{kn} a_{nk}| \leq m \sup_{n\in\mathbb{N}} \left| \frac{r_n a_{nk}}{r_k} \right| \leq m^2 \quad \forall k$. Hence $\sup_{k\in\mathbb{N}} \sup_{n\in\mathbb{N}} |a_{kn} a_{nk}| \leq m^2$.

(2) Also $|a_{kn} a_{nk}| = \left| \frac{r_k a_{kn}}{r_n} \right| \left| \frac{r_n a_{nk}}{r_k} \right| \leq m \left| \frac{r_n a_{nk}}{r_k} \right| \quad \forall k$.

Now by Theorem 1 $\lim_{n\to\infty} r_n a_{nk} = 0 \quad \forall k$, therefore by Sandwich Theorem $\lim_{n\to\infty} |a_{nk} a_{kn}| = 0, \quad \forall k$

i.e. $\lim_{n\to\infty} a_{nk} a_{kn} = 0, \quad \forall k$.

Now we give two examples for the statements (1) and (2) of Theorem 2 respectively.

Example 2. Define $A = (a_{nk})$ by

$$a_{nk} = \begin{cases} \frac{n}{k} & \text{if } n < k \\ 0 & \text{if } n = k \\ n & \text{if } n > k \end{cases}$$

Then

$$a_{nk}a_{kn} = \begin{cases} n \text{ if } & n < k \\ 0 \text{ if } & n = k \\ k \text{ if } & n > k \end{cases}$$

Therefore $\sup_k \sup_n |a_{nk}a_{kn}| = \infty$. So by Theorem 2 there does not exist any weight vector $r = (r_n)$ such that $A \in B(c_0(r))$.

Example 3. Define $A = (a_{nk})$ by

$$a_{nk} = \begin{cases} \frac{1}{k} \text{ if } & n < k \\ 0 \text{ if } & n = k \\ n \text{ if } & n > k \end{cases}$$

Then $a_{nk}a_{kn} = 1 \quad \forall k \neq n$. So by Theorem 2 there does not exist any weight vector $r = (r_n)$ such that $A \in B(c_0(r))$.

5 Sufficient Condition for A to Be in $B(c_0(r))$

Here we discuss about the sufficient condition for A to be in $B(c_0(r))$. Before going to the main results we give a Corollary to Theorem 1.

Corollary 1. *Given a weight vector $r = (r_n)$ and an infinite matrix $A = (a_{nk})$, then $A \in B(c_0(r))$ if and only if, $|A|r' \leq \alpha r'$ for some α with $0 \leq \alpha < \infty$, and $\lim_{n\to\infty} r_n a_{nk} = 0$ where $r' = \left(\frac{1}{r_1}, \frac{1}{r_2}, \frac{1}{r_3}, \cdots\right)$.*

Proof. We know $A \in B(c_0(r))$ if and only if, $\exists \quad \alpha \in [0, \infty)$ such that

$\sup_n \sum_{k=1}^{\infty} \left|\frac{r_n a_{nk}}{r_k}\right| = \alpha < \infty$ and $\lim_{n\to\infty} r_n a_{nk} = 0$,

if and only if, $\sum_{k=1}^{\infty} \left|\frac{r_n a_{nk}}{r_k}\right| \leq \alpha \quad \forall n$, and $\lim_{n\to\infty} r_n a_{nk} = 0$,

if and only if, $\sum_{k=1}^{\infty} \left|\frac{a_{nk}}{r_k}\right| \leq \frac{\alpha}{r_n} \quad \forall n$, and $\lim_{n\to\infty} r_n a_{nk} = 0$,

if and only if, $|A|r' \leq \alpha r'$ and $\lim_{n\to\infty} r_n a_{nk} = 0$.

Now we state a basic result about the inverse of lower triangular matrices.

Lemma 4. *Let $L = (l_{ij})$ be a lower triangular matrix (i.e., $l_{ij} = 0$ if $i < j$) with $l_{ii} > 0$, $\forall i \in \mathbb{N}$ and $l_{ij} \leq 0$ for $j < i$. Then there exists an unique lower triangular matrix $X \geq 0$ such that $LX = XL = I$.*

Proof. The proof runs parallel with lines used in [7, Lemma 4].

Lemma 5. *Let $L = (l_{ij})$ where $l_{ij} \geq 0 \quad \forall i, j$ such that $l_{ij} = 0$ if $i \leq j$. Let $e_1^T = (1, 0, 0, \cdots)$ and $r^T = (r_1, r_2, \cdots)$ such that $r = (I - L)^{-1}e_1$, then $r_1 = 1$ and*

$$r_n = \sum_{i=1}^{n-1} \sum_{1<k_1<k_2<\cdots<k_{i-1}<n} l_{k_1 1} l_{k_2 k_1} \cdots l_{n k_{i-1}}, \ \forall n \geq 2. \tag{1}$$

Proof. $r = (I-L)^{-1}e_1 \Longrightarrow (I-L)r = (I-L)((I-L)^{-1}e_1)$, hence $(I-L)r = e_1$ i.e.,

$$\begin{pmatrix} 1 & 0 & 0\,0\cdots \\ -l_{21} & 1 & 0\,0\cdots \\ -l_{31} & -l_{32} & 1\,0\cdots \\ \vdots & \vdots & \vdots\,\vdots\,\ddots \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \\ r_3 \\ \vdots \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \end{pmatrix} \tag{2}$$

Equating first two entries we have $r_1 = 1$ and $-l_{21}r_1 + r_2 = 0 \Rightarrow r_2 = l_{21}$, so our result is verified for $n = 2$. Now we are going to prove the result on applying the principle of first induction.

Take any $n \in \mathbb{N}$ with $n \geq 2$ and assume result holds for $n = 2, 3, \cdots N$ then equating $(N+1)$st entries on each side of (2) we get,

$$-(l_{N+1,1}r_1 + l_{N+1,2}r_2 + l_{N+1,3}r_3 + \cdots l_{N+1,N}r_N) + r_{N+1} = 0$$
$$\Rightarrow r_{N+1} = r_1 l_{N+1,1} + \sum_{k=2}^{N} l_{N+1,k}r_k.$$

Now substitute r_j from (1) on the right side of the above equation and also using $r_1 = 1$ we have,

$$\begin{aligned} r_{N+1} =& l_{N+1,1} + \sum_{k=2}^{N} \left[\sum_{i=1}^{k-1} \sum_{1<k_1<k_2<\cdots<k_{i-1}<k} l_{k_1 1} l_{k_2 k_1} \cdots l_{n k_{i-1}} \right] l_{N+1,k} \\ =& l_{N+1,1} + \sum_{i=1}^{N-1} \left[\sum_{k=i+1}^{N} \sum_{1<k_1<k_2<\cdots<k_{i-1}<k} l_{k_1 1} l_{k_2 k_1} \cdots l_{n k_{i-1}} l_{N+1,k} \right]. \end{aligned}$$

In the first summation on replace i by $i-1$, we have

$$r_{N+1} = l_{N+1,1} + \sum_{i=2}^{N} \left[\sum_{k=i}^{N} \sum_{1<k_1<k_2<\cdots<k_{i-2}<k} l_{k_1 1} l_{k_2 k_1} \cdots l_{n k_{i-1}} l_{N+1,k} \right].$$

Now we replace k with $k_i - 1$ and we can write as,

$$\begin{aligned} r_{N+1} =& l_{N+1,1} + \sum_{i=2}^{N} \left[\sum_{1<k_1<k_2<\cdots<k_{i-1}<N+1} l_{k_1 1} l_{k_2 k_1} \cdots l_{n k_{i-1}} \right] l_{N+1,k} \\ =& \sum_{i=1}^{N-1} \left[\sum_{1<k_1<k_2<\cdots<k_{i-1}<N+1} l_{k_1 1} l_{k_2 k_1} \cdots l_{n k_{i-1}} l_{N+1,k} \right]. \end{aligned}$$

Hence by principle of induction we get the expression (1) holds for all $n \geq 2$.

Remark 2. For $i = 1$ the sum $\sum_{1<k_1<k_2<\cdots<k_{i-1}<n} l_{k_1 1} l_{k_2 k_1} \cdots l_{n k_{i-1}}$ in Eq. (1) does not make sense, but we consider the sum as l_{n1}, because $l_{k_1 1} l_{k_2 k_1} \cdots l_{n k_{i-1}}$ is a product of i factors.

Theorem 3. *Let $A = (a_{nk})$ be an infinite matrix and let L be the strictly lower triangular part of $|A|$ such that $|A| = L + U$. Let $r' = (I - L)^{-1}e_1$ and assume the following,*

(i) $\forall n \in \mathbb{N}$ *and* $n \geq 2, \quad \exists k \in \mathbb{N}, \quad n > k$ *such that* $a_{nk} \neq 0$
(ii) Ur' *exists and* $\exists \alpha \in [0, \infty)$ *such that* $Ur' \leq \alpha r'$
(iii) $\lim_{n\to\infty} r_n a_{nk} = 0 \quad \forall k$

Then $r > 0$, A is bounded on $c_0(r)$ and $\|A\|_r \leq \alpha + 1$.

Proof. From Assumption (i) it follows that $\exists$ at least one term in the expression (1) which is positive. Thus each component of the column vector $(I - L)^{-1}e_1$ is non-zero and positive. So $r' = (I-L)^{-1}e_1$ is well defined where $r' = \left(\frac{1}{r_1}, \frac{1}{r_2}, \cdots\right)$ and $r' > 0 \implies r > 0$. Also since $r' > 0$ and all terms of U are non-negative, so $Ur' \geq 0$. Now we have $r' = (I - L)^{-1}e_1$ i.e. $(I - L)r' = e_1$ thus $r' = Lr' + e_1$ which yields $Lr' \leq Lr' + e_1 = r'$.

Next $|A|r' = (L + U)r' = Lr' + Ur' \leq r' + \alpha r' = (\alpha + 1)r'$. Hence $|A|r' \leq (\alpha + 1)r'$ and from assumption (iii) $\lim_{n\to\infty} r_n a_{nk} = 0 \quad \forall k$.
Hence by Corollary 1 $A \in B(c_0(r))$ and $\|A\|_r \leq \alpha + 1$.

Example 4. Let $A = (a_{nk})$ where

$$a_{nk} = \begin{cases} \frac{1}{2^{k-2}(k-n+1)^2} & \text{if} \quad n < k \\ 1 & \text{if} \quad n \geq k \end{cases}$$

Each columns of A are not in c_0, so $A \notin B(c_0)$ by Lemma 1. Using the notations of Theorem 3 we have, $r^{'} = (I - L)^{-1}e_1 = (1, 1, 2, 4, 8, 16, \cdots)^T$. Also

$$U = \begin{pmatrix} 1 & \frac{1}{2^2} & \frac{1}{2.3^2} & \frac{1}{4.4^2} & \frac{1}{8.5^2} & \cdots \\ 0 & 1 & \frac{1}{2.2^2} & \frac{1}{4.3^2} & \frac{1}{8.4^2} & \cdots \\ 0 & 0 & 1 & \frac{1}{4.2^2} & \frac{1}{8.3^2} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Then $Ur^{'} = (1 + c, 1 + c, 2 + c, 4 + c, 8 + c, \cdots)^T$, where $c = \sum_{k=2}^{\infty} \frac{1}{k^2}, c > 0$. Now if we choose α such that $\alpha \geq (1 + c)$, then $(1 + c) \leq 1.\alpha$, $(2 + c) \leq 2.(1 + c) \leq 2\alpha$, $(4 + c) \leq 4.(1 + c) \leq 4\alpha$, and so on. So Ur' exists and $Ur' \leq \alpha r'$ for $\alpha \geq (1 + c)$. Again $r = \left(1, 1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \cdots\right)^T$ and it can be easily check that $\lim_{n\to\infty} r_n a_{nk} = 0 \quad \forall k$. So all the three conditions of Theorem 3 are satisfied. So we can say that A is bounded on $c_0(r)$ and $\|A\|_r \leq \alpha + 1$.

Theorem 4. *Let $A = (a_{nk})$ be an infinite matrix and define L, U, e_1, r, r' are defined by previous theorem and assume the following,*

(i) $\forall n \in \mathbb{N}$ *and* $n \geq 2, \quad \exists k \in \mathbb{N}, \quad n > k$ *such that* $a_{nk} \neq 0$
(ii) Ur' *exists and* $[Ur']_k \leq [LUr']_k \quad k \in \mathbb{N}, \quad k \geq 2$

(iii) $\lim_{n\to\infty} r_n a_{nk} = 0 \quad \forall k$

Then $r > 0$, A *is bounded on* $c_0(r)$ *and* $\|A\|_r \leq [Ur']_1 + 1$.

Proof. It is easy to verify that any matrix can be multiplied by a lower triangular matrix in left side, since the entries of the product are finite sum. Therefore assumption (ii) implies that LUr' exists. From Theorem 3 it follows that $r' > 0 \implies r > 0$. Let $\alpha = [Ur']_1$, then $0 \leq \alpha < \infty$. Now $[(I-L)Ur']_1 = [Ur' - LUr']_1 = [Ur']_1 = \alpha$,. Also assumption (ii) implies $[(I-L)Ur']_k \leq 0 \quad \forall k \geq 2$. Therefore we have $(I-L)Ur' \leq \alpha e_1$. Now multiplying both sides by $(I-L)^{-1}$ from left we have, $(I-L)^{-1}[(I-L)Ur'] \leq \alpha[(I-L)^{-1}e_1] = \alpha r'$, hence $Ur' \leq \alpha r'$. Now by applying Theorem 3 A is bounded on $c_0(r)$, and $\|A\|_r \leq \alpha + 1 = [Ur']_1 + 1$.

References

1. Akhmedov, A.M., Basar, F.: On the fine spectrum of the Cesaro operator in c_0. Math. J. Ibaraki Univ. **36**, 25–32 (2004)
2. Altay, B., Basar, F.: On the fine spectrum of the generalized difference operator $B(r, s)$ over the sequence spaces c_0 and c. Int. J. Math. Math. Sci. **18**, 3005–3013 (2005)
3. Altay, B., Basar, F.: On the fine spectrum of the difference operator $\triangle$ on c_0 and c. Inform. Sci. **168**, 217–224 (2004)
4. Cooke, R.G.: Infinite Matrices and Sequence Spaces. Dover Publications, New York (1955)
5. Furkan, H., Bilgiç, H., Altay, B.: On the fine spectrum of the operator $B(r, s, t)$ over c_0 and c. Comput. Math. Appl. **53**, 989–998 (2007)
6. Shivakumar, P.N., Sivakumar, K.C.: A review of infinite matrices and their applications. Linear Algebra Appl. **430**, 976–998 (2009)
7. Williams, Joseph J., Ye, Q.: Infinite matrices bounded on weighted l_1 spaces. Linear Algebra Appl. **438**, 4689–4700 (2013)

Mapping Properties of One Class of Quasielliptic Operators

Gennadii Demidenko[1,2(✉)]

[1] Sobolev Institute of Mathematics, Acad. Koptyug Avenue 4, 630090 Novosibirsk, Russia
demidenk@math.nsc.ru
[2] Novosibirsk State University, Pirogov Street 2, 630090 Novosibirsk, Russia
http://www.math.nsc.ru/LBRT/d5/english/demidenko.htm

Abstract. The paper is devoted to the theory of quasielliptic operators. We consider scalar and homogeneous quasielliptic operators $\mathcal{L}(D_x)$ with lower terms in the whole space $\mathbb{R}^n$. Our aim is to study mapping properties of these operators in weighted Sobolev spaces. We introduce a special scale of weighted Sobolev spaces $W^l_{p,q,\sigma}(\mathbb{R}^n)$. These spaces coincide with known spaces of Sobolev type for some parameters l, q, σ. We investigate mapping properties of the operators $\mathcal{L}(D_x)$ in the spaces $W^l_{p,q,\sigma}(\mathbb{R}^n)$. We indicate conditions for unique solvability of quasielliptic equations and systems in these spaces, obtain estimates for solutions and formulate an isomorphism theorem for quasielliptic operators. To prove our results we construct special regularizers for quasielliptic operators.

Keywords: Quasielliptic operators · Weighted Sobolev spaces · Isomorphism

1 Introduction

In the paper a class of quasielliptic operators $\mathcal{L}(D_x)$ is considered in the whole space $\mathbb{R}^n$. This class belongs to the classes of quasielliptic operators introduced by S.M. Nikol'skii [1] and L.R. Volevich [2]. Our aim is to study mapping properties of the operators $\mathcal{L}(D_x)$ in special weighted Sobolev spaces $W^l_{p,q,\sigma}(\mathbb{R}^n)$ and to establish isomorphism theorems.

The first isomorphism theorems for scalar elliptic operators were proved by L.A. Bagirov and V.A. Kondratiev [3], M. Cantor [4,5], R.C. McOwen [6,7]. Isomorphism theorems for matrix homogeneous elliptic operators were proved by Y. Choquet-Bruhat and D. Christodoulou [8], R.B. Lockhart and R.C. McOwen [9].

As a rule, isomorphism theorems for elliptic operators are not trivial. For example, consider the elliptic operator

$$\triangle - \varepsilon I : W^2_p(\mathbb{R}^n) \longrightarrow L_p(\mathbb{R}^n), \qquad 1 < p < \infty,$$

G. Demidenko—The work is supported in part by the Program of the Presidium of the Russian Academy of Sciences (project no. 0314-2015-0011).

© Springer Nature Singapore Pte Ltd. 2017
D. Giri et al. (Eds.): ICMC 2017, CCIS 655, pp. 339–348, 2017.
DOI: 10.1007/978-981-10-4642-1_29

where $\triangle$ is the Laplace operator. If $\varepsilon > 0$ then the mapping is an isomorphism. However, the mapping

$$\triangle : W_p^2(\mathbb{R}^n) \longrightarrow L_p(\mathbb{R}^n)$$

is not an isomorphism. Taking into account results of L.D. Kudryavtsev [10], L.A. Bagirov and V.A. Kondratyev [3], L. Nirenberg and H.F. Walker [11], M. Cantor [4,5], it is necessary to use weighted Sobolev spaces for proving isomorphism theorems for the Laplace operator. The first isomorphism theorems for the Laplace operator were proved by M. Cantor [4] and R.C. McOwen [6]. They used special weighted Sobolev spaces.

It should be noted that isomorphism theorems for matrix elliptic operators can be more complicated. For example, consider the Stokes operator

$$\begin{pmatrix} -\triangle & 0 & 0 & D_{x_1} \\ 0 & -\triangle & 0 & D_{x_2} \\ 0 & 0 & -\triangle & D_{x_3} \\ D_{x_1} & D_{x_2} & D_{x_3} & 0 \end{pmatrix}, \qquad x \in \mathbb{R}^3.$$

This operator is elliptic in the Douglis–Nirenberg sense. One can prove an isomorphism theorem for the Stokes operator (see [12]). However, it is necessary to use a product of special weighted Sobolev spaces with different components of smoothness vectors and different weights.

The first isomorphism theorems for matrix homogeneous quasielliptic operators were proved by G.V. Demidenko [13,14]. The investigations [13,14] were continued by G.N. Hile [15]. In the present paper we consider a more general class of quasielliptic operators $\mathcal{L}(D_x)$ in $\mathbb{R}^n$.

2 Quasielliptic Operators

First we consider the following scalar differential operator

$$L(D_x) = \sum_{\beta} a_\beta D_x^\beta,$$

where the coefficients a_β are constants. Suppose that its symbol $L(i\xi)$, $\xi \in \mathbb{R}^n$, satisfies the following conditions.

Condition 1. The symbol $L(i\xi)$ is homogeneous with respect to a vector $\alpha = (\alpha_1, \ldots, \alpha_n)$, $1/\alpha_j \in \mathbb{N}$, $j = 1, \ldots, n$; i.e.,

$$L(c^\alpha i\xi) = cL(i\xi), \qquad c > 0.$$

Condition 2. The equality

$$L(i\xi) = 0, \qquad \xi \in \mathbb{R}^n,$$

holds if and only if $\xi = 0$.

Definition 1. *The differential operator $L(D_x)$ is called quasielliptic, if its symbol satisfies Conditions 1, 2.*

This class of operators belongs to the class of differential operators introduced by S.M. Nikol'skii [1].

Quasielliptic operators $L(D_x)$ whose symbols $L(i\xi)$ are homogeneous with respect to a vector α are usually called *quasielliptic operators without lower terms*. Such operators can be written in the form

$$L(D_x) = \sum_{\beta\alpha=1} a_\beta D_x^\beta. \tag{1}$$

Examples of such operators are elliptic operators, $2b$-parabolic operators without lower terms, etc.

Note that the symbol of the quasielliptic operator (1) satisfies the following estimate

$$c_1\langle\xi\rangle \le |L(i\xi)| \le c_2\langle\xi\rangle, \qquad \langle\xi\rangle^2 = \sum_{j=1}^n \xi_j^{2/\alpha_j}, \qquad \xi \in \mathbb{R}^n,$$

where c_1, $c_2 > 0$ are constants.

We now consider the differential operators

$$\mathcal{L}(D_x) = L(D_x) + \sum_{\beta\alpha<1} a_\beta D_x^\beta, \tag{2}$$

where $L(D_x)$ is the quasielliptic operator (1). We will call operators of the form (2) *quasielliptic operators with lower terms*. Denote the differential operator corresponding to the lower terms by

$$L'(D_x) = \sum_{\beta\alpha<1} a_\beta D_x^\beta.$$

Condition 3. Suppose that the symbol of the differential operator (2) satisfies the estimate

$$c_3\left(\langle\xi\rangle + \langle\xi\rangle^q\right) \le |L(i\xi) + L'(i\xi)| \le c_4\left(\langle\xi\rangle + \langle\xi\rangle^q\right), \qquad \xi \in \mathbb{R}^n, \tag{3}$$

where $0 \le q < 1$, c_3, $c_4 > 0$ are constants.

Example 1. Consider the differential operator

$$\mathcal{L}(D_x) = \triangle^m + \varepsilon(-1)^{m-k}\triangle^k, \qquad m > k, \quad \varepsilon > 0. \tag{4}$$

We have

$$L(D_x) = \triangle^m, \qquad L'(D_x) = \varepsilon(-1)^{m-k}\triangle^k, \qquad \alpha_1 = \ldots = \alpha_n = 1/(2m).$$

Obviously, Conditions 1–3 are fulfilled for $q = k/m$.

We now consider the matrix differential operator

$$L(D_x) = \sum_{\beta\alpha=1} A_\beta D_x^\beta, \tag{5}$$

where the coefficients A_β are constant $(m \times m)$-matrices with real or complex entries. Suppose that its symbol $L(i\xi)$ satisfies the following condition.

Condition 4. The equality

$$\det L(i\xi) = 0, \qquad \xi \in \mathbb{R}^n,$$

holds if and only if $\xi = 0$.

Definition 2. *The matrix differential operator (5) is called homogeneous quasi-elliptic operator if its symbol satisfies Condition 4.*

This class of operators belongs to the class of differential operators introduced by L.R. Volevich [2]. Examples of such operators are homogeneous elliptic operators, 2*b*-parabolic operators without lower terms, parabolic operators with 'opposite times directions', etc.

We now consider matrix differential operators of the form

$$\mathcal{L}(D_x) = L(D_x) + \sum_{\beta\alpha<1} A'_\beta D_x^\beta, \tag{6}$$

where $L(D_x)$ is the matrix quasielliptic operator of the form (5), the coefficients A'_β are constant $(m \times m)$-matrices.

We will call operators of the form (6) *homogeneous quasielliptic operator with lower terms*. Suppose that its symbol $\mathcal{L}(i\xi)$ satisfies the following condition.

Condition 5. Suppose that the symbol of the differential operator (6) satisfies the estimate

$$c_5 \left(\langle\xi\rangle + \langle\xi\rangle^q\right)^m \le |\det \mathcal{L}(i\xi)| \le c_6 \left(\langle\xi\rangle + \langle\xi\rangle^q\right)^m, \qquad \xi \in \mathbb{R}^n, \tag{7}$$

where $0 \le q < 1$, c_5, $c_6 > 0$ are constants.

Example 2. Consider the parabolic operator with 'opposite times directions'

$$\mathcal{L}(D_x) = \begin{pmatrix} D_{x_n} - \triangle' & \alpha \\ \beta & D_{x_n} + \triangle' \end{pmatrix},$$

where $\triangle'$ is the Laplace operator in $\mathbb{R}^{n-1}$ and $\alpha\beta > 0$. Obviously,

$$\mathcal{L}(D_x) = \begin{pmatrix} D_{x_n} - \triangle' & 0 \\ 0 & D_{x_n} + \triangle' \end{pmatrix} + \begin{pmatrix} 0 & \alpha \\ \beta & 0 \end{pmatrix}.$$

Consequently, Conditions 4, 5 are fulfilled for

$$m = 2, \qquad \alpha = \left(\frac{1}{2}, \ldots, \frac{1}{2}, 1\right), \qquad q = 0.$$

Isomorphism theorems for quasielliptic operators of the forms (1), (5) without lower terms were proved in [13–15]. In our paper we study mapping properties of quasielliptic operators of the forms (2), (6). Particularly, we formulate isomorphism theorems for these operators.

3 Weighted Sobolev Spaces

We introduce the *weighted Sobolev spaces* $W^l_{p,q,\sigma}(\mathbb{R}^n)$. Using these spaces, one can solve the problem on isomorphism for quasielliptic operators $\mathcal{L}(D_x)$ of the form (2) or (6).

Definition 3. *Let* $l = (1/\alpha_1, \ldots, 1/\alpha_n)$, $1/\alpha_j \in \mathbb{N}$, $j = 1, \ldots, n$, $1 < p < \infty$, $0 \le q \le 1$, $\sigma \ge 0$. *Denote by* $W^l_{p,q,\sigma}(\mathbb{R}^n)$ *the weighted Sobolev space of functions* $u \in L_{loc}(\mathbb{R}^n)$ *having the weak derivatives* $D^\nu_x u$, $\nu\alpha \le 1$, *such that*

$$D^\nu_x u \in L_p(\mathbb{R}^n) \quad for \quad q \le \nu\alpha \le 1,$$

$$\|(1 + \langle x\rangle)^{-\sigma(q-\nu\alpha)} D^\nu_x u(x),\, L_p(\mathbb{R}^n)\| < \infty \quad for \quad 0 \le \nu\alpha < q.$$

Here $\langle x\rangle^2 = \sum\limits_{j=1}^{n} x_j^{2/\alpha_j}$.

Introduce the norm

$$\|u,\, W^l_{p,q,\sigma}(\mathbb{R}^n)\| = \sum_{q \le \nu\alpha \le 1} \|D^\nu_x u(x),\, L_p(\mathbb{R}^n)\| + \sum_{0 \le \nu\alpha < q} \|(1 + \langle x\rangle)^{-\sigma(q-\nu\alpha)} D^\nu_x u(x),\, L_p(\mathbb{R}^n)\|. \tag{8}$$

The weighted Sobolev spaces $W^l_{p,q,\sigma}(\mathbb{R}^n)$ coincide with well-known spaces for some parameters l, q, σ. We consider several examples.

Example 3. Obviously, the space $W^l_{p,q,0}(\mathbb{R}^n) = W^l_{p,0,\sigma}(\mathbb{R}^n)$ is the Sobolev space $W^l_p(\mathbb{R}^n)$.

Example 4. The space $W^l_{p,1,\sigma}(\mathbb{R}^n)$ coincides with the space $W^l_{p,\sigma}(\mathbb{R}^n)$ introduced in [16]. Indeed, by definition [16],

$$\|u,\, W^l_{p,\sigma}(\mathbb{R}^n)\| = \sum_{0 \le \nu\alpha \le 1} \|(1 + \langle x\rangle)^{-\sigma(1-\nu\alpha)} D^\nu_x u(x),\, L_p(\mathbb{R}^n)\|.$$

Example 5. In the isotropic case $1/\alpha_1 = \ldots = 1/\alpha_n = \bar{l}$ the norm (8) for $q = \sigma = 1$ is equivalent to the norm

$$\sum_{0 \le |\beta| \le \bar{l}} \|(1 + |x|)^{-(\bar{l}-|\beta|)} D^\beta_x u(x),\, L_p(\mathbb{R}^n)\|. \tag{9}$$

Then, from the work [10] of L.D. Kudryavtsev it follows that the space $W^l_{p,1,1}(\mathbb{R}^n)$ for $p > n$ coincides with the Sobolev space

$$W^{\bar{l}}_{p,\square}(\mathbb{R}^n), \quad \square = \{x \in \mathbb{R}^n : \ |x_j| < 1, \ j = 1, \dots, n\},$$

where

$$\|u, \ W^{\bar{l}}_{p,\square}(\mathbb{R}^n)\| = \int\limits_{\square} |u(x)|dx + \sum_{|\beta|=\bar{l}} \|D^\beta_x u(x), \ L_p(\mathbb{R}^n)\|.$$

Example 6. Consider the Nirenberg–Walker–Cantor space $M^p_{\ell,k}(\mathbb{R}^n)$ [4,11] whose norm is defined as

$$\|u, \ M^p_{\ell,k}(\mathbb{R}^n)\| \ = \sum_{|\beta|\le\ell} \|(1+|x|)^{k+|\beta|} D^\beta_x u(x), L_p(\mathbb{R}^n)\|.$$

Clearly, by (9) the space $W^l_{p,1,1}(\mathbb{R}^n)$ coincides with the space $M^p_{\bar{l},-\bar{l}}(\mathbb{R}^n)$ in the isotropic case $1/\alpha_1 = \ldots = 1/\alpha_n = \bar{l}$ for $q = \sigma = 1$, $p > 1$.

Definition 4. *Denote by* $\overset{\circ}{W}{}^l_{p,q,\sigma}(\mathbb{R}^n)$ *the completion of* $C^\infty_0(\mathbb{R}^n)$ *with respect to the norm (8).*

From Definitions 3 and 4 it follows that the space $\overset{\circ}{W}{}^l_{p,q,\sigma}(\mathbb{R}^n)$ is embedded in the space $W^l_{p,q,\sigma}(\mathbb{R}^n)$. One can show that the strict embedding holds

$$\overset{\circ}{W}{}^l_{p,q,\sigma}(\mathbb{R}^n) \ \subset \ W^l_{p,q,\sigma}(\mathbb{R}^n)$$

for sufficiently large $\sigma > 1$.

In the next theorem we indicate the condition when these spaces coincide. Note that theorems of such type are very important in the theory of differential operators.

Theorem 1. *If* $0 \le \sigma \le 1$ *then* $\overset{\circ}{W}{}^l_{p,q,\sigma}(\mathbb{R}^n) \ = \ W^l_{p,q,\sigma}(\mathbb{R}^n)$.

Definition 5. *Denote by*

$$L_{p,\gamma}(\mathbb{R}^n), \qquad 1 < p < \infty, \quad \gamma \in R,$$

the space of integrable functions with the norm

$$\|u, \ L_{p,\gamma}(\mathbb{R}^n)\| = \|(1+\langle x\rangle)^{-\gamma} u(x), \ L_p(\mathbb{R}^n)\|.$$

Thereafter we will say that a vector-function

$$U(x) = (u_1(x), \dots, u_m(x))^T, \qquad m \ge 1$$

belongs to the weighted Sobolev space $W^l_{p,q,\sigma}(\mathbb{R}^n)$, if every its component u_j belongs to $W^l_{p,q,\sigma}(\mathbb{R}^n)$. By definition,

$$\|U,\, W^l_{p,q,\sigma}(\mathbb{R}^n)\| = \sum_{j=1}^{m} \|u_j,\, W^l_{p,q,\sigma}(\mathbb{R}^n)\|.$$

Analogously, a vector-function

$$F(x) = (f_1(x), \dots, f_m(x))^T, \qquad m \geq 1$$

belongs to the weighted space $L_{p,\gamma}(\mathbb{R}^n)$, if every its component f_j belongs to $L_{p,\gamma}(\mathbb{R}^n)$ and

$$\|F,\, L_{p,\gamma}(\mathbb{R}^n)\| = \sum_{j=1}^{m} \|f_j,\, L_{p,\gamma}(\mathbb{R}^n)\|.$$

4 Mapping Properties of the Operators (2), (6)

Consider the quasielliptic operator $\mathcal{L}(D_x)$ defined by (2) or (6). Introduce the notation $|\alpha| = \sum\limits_{j=1}^{n} \alpha_j$.

The following theorems hold.

Theorem 2. *Let $\beta = (\beta_1, \dots, \beta_n)$, $1 \geq \beta\alpha \geq q$. Then the following estimate is satisfied for every $U \in C_0^\infty(\mathbb{R}^n)$*

$$\|D_x^\beta U(x),\, L_p(\mathbb{R}^n)\| \leq c_\beta \|\mathcal{L}(D_x)U(x),\, L_p(\mathbb{R}^n)\|,$$

where the constant $c_\beta > 0$ does not depend on U.

Theorem 3. *Let $\beta = (\beta_1, \dots, \beta_n)$, $q > \beta\alpha \geq 0$ and*

$$\frac{|\alpha|}{p} > \sigma(q - \beta\alpha) > q - \beta\alpha - \frac{|\alpha|}{p'}, \quad 1 \geq \sigma \geq 0, \quad \frac{1}{p} + \frac{1}{p'} = 1.$$

Then the following estimate is satisfied for every $U \in C_0^\infty(\mathbb{R}^n)$

$$\|\langle x\rangle^{-\sigma(q-\beta\alpha)} D_x^\beta U(x),\, L_p(\mathbb{R}^n)\| \leq c\|\langle x\rangle^{(1-\sigma)(q-\beta\alpha)} \mathcal{L}(D_x)U(x),\, L_p(\mathbb{R}^n)\|,$$

where the constant $c_\beta > 0$ does not depend on U.

Theorem 4. *Let*

$$|\alpha| > q, \quad |\alpha|/p > \sigma q > |\alpha|/p - (|\alpha| - q).$$

Then for every $F \in L_{p,(\sigma-1)q}(\mathbb{R}^n)$ there exists a unique $U \in W^l_{p,q,\sigma}(\mathbb{R}^n)$ such that

$$\mathcal{L}(D_x)U(x) = F(x), \qquad x \in \mathbb{R}^n.$$

Moreover, the estimate holds

$$\|U,\, W^l_{p,q,\sigma}(\mathbb{R}^n)\| \leq c\|F,\, L_{p,(\sigma-1)q}(\mathbb{R}^n)\|$$

with a constant $c > 0$ independent of F.

Theorem 5. *Let $|\alpha|/p > q$. Then the mapping*

$$\mathcal{L}(D_x):\ W^l_{p,q,1}(\mathbb{R}^n) \longrightarrow L_p(\mathbb{R}^n), \quad 1 < p < \infty$$

is an isomorphism.

Remark 1. Theorems 4, 5 are analogs of some theorems in [13–15] for quasielliptic operators without lower terms.

We illustrate Theorem 5 by using the differential operator (4):

$$\mathcal{L}(D_x) = \triangle^m + \varepsilon(-1)^{m-k}\triangle^k, \quad m \geq k, \quad \varepsilon > 0.$$

Taking into account Example 1, we have

$$\alpha_1 = \ldots = \alpha_n = 1/(2m), \qquad q = k/m.$$

Consequently, by Theorem 5 the mapping

$$\mathcal{L}(D_x):\ W^l_{p,q,1}(\mathbb{R}^n) \longrightarrow L_p(\mathbb{R}^n), \quad l = (2m, \ldots, 2m), \tag{10}$$

is an isomorphism for $p \in (1, \frac{n}{2k})$, $n > 2k$.

Consider the critical cases in (4): $k = 0$ and $k = m$.

In the first case $k = 0$ we have $\triangle^0 = I$, $q = 0$ and $W^l_{p,0,1}(\mathbb{R}^n) = W^l_p(\mathbb{R}^n)$. Then (10) is rewritten in the form

$$\triangle^m + \varepsilon(-1)^m I:\ W^l_p(\mathbb{R}^n) \longrightarrow L_p(\mathbb{R}^n).$$

Therefore the isomorphism theorem gives the classical result.

In the second case $k = m$, we have $q = 1$ and $W^l_{p,1,1}(\mathbb{R}^n) = W^l_{p,1}(\mathbb{R}^n)$. Then (10) is rewritten in the form

$$(1 + \varepsilon)\triangle^m:\ W^l_{p,1}(\mathbb{R}^n) \longrightarrow L_p(\mathbb{R}^n).$$

The isomorphism theorem for $p \in \left(1, \frac{n}{2m}\right)$, $n > 2m$ follows from [7].

5 Elements of Used Technique

To prove of the above results we use a technique of integral representations for regularizers of differential operators. Our technique is based on the special representation by S.V. Uspenskii [17] for integrable functions:

$$\varphi(x) = \lim_{h\to 0}(2\pi)^{-n} \int_h^{h^{-1}} v^{-|\alpha|-1} \int_{\mathbb{R}^n}\int_{\mathbb{R}^n} \exp\left(i\frac{x-y}{v^\alpha}\xi\right) G(\xi)\varphi(y)d\xi dy dv, \tag{11}$$

where

$$G(\xi) = 2M\langle\xi\rangle^{2M}\exp(-\langle\xi\rangle^{2M}), \quad \langle\xi\rangle^2 = \sum_{i=1}^n \xi_i^{2/\alpha_i}, \quad M,\ 1/\alpha_i \in \mathbb{N}.$$

Applying the integral representation (10), we construct the following integral operators

$$P_{j,h}F(x) = (2\pi)^{-n} \int_h^{h^{-1}} v^{-|\alpha|} \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} \exp\left(i\frac{x-y}{v^\alpha}\xi\right)$$

$$\times G(\xi)\left(\sum_{k=1}^{m} l^{j,k}(i\xi)F_k(y)\right) d\xi dy dv, \quad j = 1, \ldots, m, \quad h > 0,$$

where $l^{j,k}(i\xi)$ are entries of the inverse matrix $(\mathcal{L}(i\xi))^{-1}$. In the case of $m = 1$ we write $(\mathcal{L}(i\xi))^{-1}F(y)$ instead of the sum

$$\sum_{k=1}^{m} l^{j,k}(i\xi)F_k(y).$$

In the present paper we use the operators $P_{j,h}$ for $h \ll 1$ in order to construct regularizers of the quasielliptic operators (2), (6). Using these regularizers, we indicate the conditions for unique solvability of the quasielliptic equations and systems in the weighted Sobolev spaces, obtain the estimates for the solutions and formulate the isomorphism theorem for the quasielliptic operators.

References

1. Nikol'skii, S.M.: The first boundary problem for a general linear equation. Sov. Math. Dokl. **3**, 1388–1390 (1962)
2. Volevich, L.R.: Local properties of solutions to quasielliptic systems. Mat. Sb. **59**, 3–52 (1962). (in Russian)
3. Bagirov, L.A., Kondratyev, V.A.: On elliptic equations in $\mathbb{R}^n$. Differ. Uravn. **11**, 498–504 (1975). (in Russian)
4. Cantor, M.: Spaces of functions with asymptotic conditions on $\mathbb{R}^n$. Indiana Univ. Math. J. **24**, 897–902 (1975)
5. Cantor, M.: Elliptic operators and decomposition of tensor fields. Bull. AMS **5**, 235–262 (1981)
6. McOwen, R.C.: The behavior of the laplacian on weighted sobolev spaces. Comm. Pure Appl. Math. **32**, 783–795 (1979)
7. McOwen, R.C.: On elliptic operators in $\mathbb{R}^n$. Comm. Partial Diff. Eqn. **5**, 913–933 (1980)
8. Choquet-Bruhat, Y., Christodoulou, D.: Elliptic systems in $H_{s,\sigma}$ spaces on manifolds which are euclidean at infinity. Acta Math. **146**, 129–150 (1981)
9. Lockhart, R.B., McOwen, R.C.: Elliptic differential operators on noncompact manifolds. Ann. Scuola Norm. Sup. Pisa. Cl. Sci. **12**, 409–447 (1985)
10. Kudryavtsev, L.D.: Direct and reverse embedding theorems. Applications to solution of elliptic equations by the variational method. Trudy Mat. Inst. Steklov. Akad Nauk SSSR **55**, 1–182 (1959). (in Russian)

11. Nirenberg, L., Walker, H.F.: The null spaces of elliptic partial differential operators in $\mathbb{R}^n$. J. Math. Anal. Appl. **42**, 271–301 (1973)
12. Demidenko, G.V.: On one class of matrix differential operators. Sib. Math. J. **45**, 86–99 (2004)
13. Demidenko, G.V.: On quasielliptic operators in $\mathbb{R}^n$. Sib. Math. J. **39**, 884–893 (1998)
14. Demidenko, G.V.: Isomorphic properties of quasi-elliptic operators. Russ. Acad. Sci. Dokl. Math. **59**, 102–106 (1999)
15. Hile, G.N.: Fundamental solutions and mapping properties of semielliptic operators. Math. Nachr. **279**, 1538–1572 (2006)
16. Demidenko, G.V.: On weighted sobolev spaces and integral operators determined by quasi-elliptic operators. Russ. Acad. Sci. Dokl. Math. **49**, 113–118 (1994)
17. Uspenskii, S.V.: On the representation of functions defined by a class of hypoelliptic operators. Proc. Steklov Inst. Math. **117**, 343–352 (1972)

$\mathcal{I}_\lambda$-double Statistically Convergent Sequences in Topological Groups

Ekrem Savaş(✉)

İstanbul Commerce University, Department of Mathematics,
Sütlüce, İstanbul, Turkey
ekremsavas@yahoo.com, esavas@ticaret.edu.tr

Abstract. In this paper, we introduce new notion, namely, $\mathcal{I}_\lambda-$ double statistical convergence in topological groups. We mainly investigate some inclusion relations between $\mathcal{I}-$double statistical and $\mathcal{I}_\lambda-$ double statistical convergence.

Keywords: Ideal convergence · Ideal double statistical convergence · Double statistical convergence · λ-double statistical convergence · Topological groups

1 Introduction

Looking through historically to statistical convergence of sequences, we recall that the concept of statistically convergence of sequences was first introduced by Fast [10] as an extension of the usual concept of sequential limits and also independently by Buck [2]. Schoenberg [33] gave some basic properties of the statistical convergence and also studied the concept as a summability method. Over the years and under different names statistical convergence has been discussed in the theory of Fourier analysis, ergodic theory and number theory. Later on it was further investigated from the sequence space point of view and linked with summability theory by Fridy [11], Šalát [23] and many others. In recent years, generalization of statistical convergence have appeared in the study of strong integral summability. Moreover statistical convergence is closely related to the concept of convergence in probability. Most of the existing works on statistical convergence have been restricted to real or complex sequences except the works of Kolk [13], Maddox [17] and Cakalli [3]. Mursaleen [19]introduced λ-statistical convergence as a generalization of statistical convergence.

In [13], Kolk extended the statistical convergence to normed spaces and also Maddox [17] extended it to locally convex Hausdorff topological linear spaces giving a representation of the statistical convergence in terms of strongly summability by using a modulus function and Cakalli [3] extended this notion to topological Hausdorff groups. Di Maio and Kočinac [18] introduced the concept of statistical convergence in topological spaces and statistical Cauchy condition in uniform spaces and established the topological nature of this convergence. Later on Hazarika and Savaş [12] introduced $\lambda-$statistical convergence

© Springer Nature Singapore Pte Ltd. 2017
D. Giri et al. (Eds.): ICMC 2017, CCIS 655, pp. 349–357, 2017.
DOI: 10.1007/978-981-10-4642-1_30

of double sequences in n-normed spaces and also Savaş and Mohiuddine [32] introduced and studied the concepts of double λ-statistically convergent and double $\lambda-$statistically Cauchy sequences in probabilistic normed space. Cakalli and Savaş [4] studied the statistical convergence of double sequences to topological groups. Quite recently Savas [25] studied $\mathcal{I}_\lambda-$statistical convergence for sequences in topological groups where more references on this important summability method can be found. In many branches of science and engineering we often come across double sequences, i.e. sequences of matrices and certainly there are situations where either the idea of ordinary convergence does not work or the underlying space does not serve our purpose. Therefore to deal with such situations we have to introduce some new type of measures which can provide a better tool and suitable frame work.

The (relatively more general) concept of $\mathcal{I}-$convergence was introduced by Kostyrko et al. [14] in a metric space. Later on it was further studied by Dems [9] and Das et al. [6]. More investigations in this direction and more applications of ideals can be found in [1,5,7,8,15,16,22,26–30].

In [6], we used ideals to introduce the concepts of $\mathcal{I}-$statistical convergence and $\mathcal{I}-$lacunary statistical convergence which naturally extend the notions of the above mentioned convergence. The concept of statistical convergence depends on the density of subsets of the set $\mathbb{N}$ of natural numbers. If $K \subset \mathbb{N}$, then $K(m,n)$ denotes the cardinality of the set $K \cap [m,n]$. The upper and lower natural density of the subset K is defined by

$$\overline{d}(K) = \lim_{n\to\infty} \sup \frac{K(1,n)}{n} \text{ and } \underline{d}(K) = \lim_{n\to\infty} \inf \frac{K(1,n)}{n}$$

If $\overline{d}(K) = \underline{d}(K)$ then we say that the natural density of K exists and it is denoted simply by $d(K)$. Clearly $d(K) = \lim_{n\to\infty} \frac{K(1,n)}{n}$.

A sequence (x_k) of real numbers is said to be statistically convergent to L if for arbitrary $\varepsilon > 0$, the set $K(\varepsilon) = \{k \in K : |x_k - L| \geq \varepsilon\}$ has natural density zero. Throughout the paper, $\mathbb{N}$ will denote the set of all natural numbers. By X, we will denote an abelian topological Hausdorff group, written additively, which satisfies the first axiom of countability. A sequence $x = (x_k)$ in X is called to be statistically convergent to an element L of X if for each neighbourhood U of 0,

$$\lim_{n\to\infty} \frac{1}{n} |\{k \leq n : x_k - L \notin U\}| = 0$$

where the vertical bars indicate the number of elements in the enclosed set, (see, [3]). The set of all statistically convergent sequences in X is denoted by $st(X)$.

2 Preliminaries

We now recall some notations and basic definitions used in this paper.

Definition 1. *A family $\mathcal{I} \subset 2^{\mathbb{N}}$ is said to be an ideal of $\mathbb{N}$ if the following conditions hold:*

(a) $A, B \in \mathcal{I}$ *implies* $A \cup B \in \mathcal{I}$,
(b) $A \in \mathcal{I},\ B \subset A$ *implies* $B \in \mathcal{I}$,

Definition 2. *A non-empty family* $F \subset 2^{\mathbb{N}}$ *is said to be a filter of* $\mathbb{N}$ *if the following conditions hold:*

(a) $\emptyset \notin F$
(b) $A, B \in F$ *implies* $A \cap B \in F$
(c) $A \in F,\ A \subset B$ *implies* $B \in F$.

If $\mathcal{I}$ is a proper ideal of $\mathbb{N}$(*i.e.*, $\mathbb{N} \notin \mathcal{I}$), then the family of sets $F(\mathcal{I}) = \{M \subset \mathbb{N} : \exists A \in \mathcal{I} : M = \mathbb{N} \backslash A\}$ is a filter of $\mathbb{N}$. It is called the filter associated with the ideal.

Definition 3. *A proper ideal* $\mathcal{I}$ *is said to be admissible if* $\{n\} \in \mathcal{I}$ *for each* $n \in \mathbb{N}$.

Definition 4 *(See, [14]). Let* $\mathcal{I} \subset 2^{\mathbb{N}}$ *be a proper admissible ideal in* $\mathbb{N}$. *Then, the sequence* (x_k) *of elements of real numbers is said to be* $\mathcal{I}$*-convergent to* L *if for each* $\varepsilon > 0$ *the set* $A(\varepsilon) = \{k \in \mathbb{N} : |x_k - L| \geq \varepsilon\} \in \mathcal{I}$.

Let $K \subseteq \mathbb{N} \times \mathbb{N}$ be a two dimensional set of natural numbers and let $K_{m,n}$ be the numbers of (i, j) in K such that $i \leq n$ and $j \leq m$. Then the lower asymptotic density of K is defined as

$$P - \liminf_{m,n} \frac{K_{m,n}}{mn} = \delta_2(K).$$

In the case when the sequence $\left(\frac{K_{m,n}}{mn}\right)_{m,n=1,1}^{\infty,\infty}$ has a limit then we say that K has a natural density and is defined as

$$P - \lim_{m,n} \frac{K_{m,n}}{mn} = \delta_2(K).$$

For example, let $K = \{(i^2, j^2) : i, j \in \mathbb{N}\}$, where $\mathbb{N}$ is the set of natural numbers. Then

$$\delta_2(K) = P - \lim_{m,n} \frac{K_{m,n}}{mn} \leq P - \lim_{m,n} \frac{\sqrt{m}\sqrt{n}}{mn} = 0$$

(i.e. the set K has double natural density zero).

Recently the studies of double sequences has a rapid growth. The concept of double statistical convergence, for complex case, was introduced by Mursaleen and Edely [20], while the idea of statistical convergence of single sequences was first studied by Fast [10]. Savaş and Patterson (see, [24]) introduced and studied lacunary statistical convergence for double sequences and they also presented some inclusion theorems. Also recently, in [31] lacunary statistical convergence for double sequences in topological groups is studied. Mursaleen and Edely [20] has given main definition as follows:

Definition 5 *([20]). A double sequences $x = (x_{kl})$ is said to be P-statistically convergent to L provided that for each $\varepsilon > 0$*

$$P - \lim_{m,n} \frac{1}{mn} \{number\ of\ (k,l) : k < m\ and\ l < n,\ |x_{kl} - L| \geq \varepsilon\} = 0.$$

In this case we write $st^2 - \lim_{k,l} x_{k,l} = L$ and we denote the set of all statistical convergent double sequences by st^2.

It is clear that a convergent double sequence is also st^2-convergent but the converse is not true, in general. Also note that st^2- convergent need not be bounded. For example, the sequence $x = (x_{k,l})$ defined by,

$$(x_{k,l}) = \begin{cases} kl;\ \text{if} \quad k \ \text{ and } \ l \ \text{are square} \\ 1; \text{otherwise} \end{cases}$$

is st^2-convergent. Nevertheless it neither convergent nor bounded. It should be noted that in [20], the authors proved the following theorem:

Theorem 1. *The following statements are equivalent:*

(a) x is statistically convergent to L;
(b) x is statistically Cauchy;
(c) there exists a subsequence y of x such that $lim_{kl} y_{kl} = L$.

By the convergence of a double sequence we mean the convergence in Pringsheims sense ([21]). A double sequence $x = (x_{kl})$ of real numbers is said to be convergent in the Pringsheim's sense or $P-$convergent if for each $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that $|x_{kl} - L| < \varepsilon$ whenever $k, l \geq N$ and L is called Pringsheim limit (denoted by $P - limx = L$).

In a topological group X, the above definitions become as in the following: a double sequence $x = (x_{kl})$ of points in X is said to be convergent to a point to a point L in X in the Pringsheims sense if for every neighbourhood U of 0 there exists $N \in$ N such that $x_{kl} - L \in U$ whenever $k, l \geq N$. L is called the Pringsheim limit of x.

Now let $\mathcal{I}$ be a nontrivial admissible ideal in $\mathbb{N} \times \mathbb{N}$. A double sequence $x = (x_{kl})$ of real number is said to be convergent to the number L with respect to the ideal $\mathcal{I}$, if for each $\varepsilon > 0$

$$A(\varepsilon) = \{(k,l) \in \mathbb{N} \times \mathbb{N} : |x_{kl} - L| \geq \varepsilon\} \in \mathcal{I}.$$

In this case we write $\mathcal{I} - lim_{k,l} x_{kl} = L$

We now define the concept of double λ-density:

Let $\lambda = (\lambda_m)$ and $\mu = (\mu_n)$ be two non-decreasing sequences of positive real numbers each of which tends to ∞ as m and n approach ∞, respectively. Also let $\lambda_{m+1} \leq \lambda_m + 1, \lambda_1 = 0$ and $\mu_{n+1} \leq \mu_n + 1, \mu_1 = 0$. The collection of such sequence (λ, μ) will be denoted by Δ.

Let $K \subseteq \mathbb{N} \times \mathbb{N}$. The number

$$\delta_\lambda(K) = \lim_{mn} \frac{1}{\lambda_{mn}} |\{k \in I_n, l \in J_m : (k,l) \in K\}|,$$

where $I_m = [m - \lambda_m + 1, m]$ and $J_n = [n - \mu_n + 1, n]$ and $\lambda_{mn} = \lambda_m \mu_n$, is said to be the λ -density of K, provided the limit exists.

Throughout this paper we shall denote $(k \in I_m, l \in J_n)$ by $(k,l) \in I_{mn}$.

A nontrivial ideal $\mathcal{I}_2$ of $\mathbb{N} \times \mathbb{N}$ is called strongly admissible if $i \times \mathbb{N}$ and $\mathbb{N} \times i$ belong to $\mathcal{I}_2$ for each $i \in \mathbb{N}$. It is evident that a strongly admissible ideal is admissible also. In this paper, we extend a few results known in the literature from ordinary (single) sequences to double sequences in topological groups and give some important inclusion theorems.

3 Main Result

Throughout $\mathcal{I}_2$ will stand for a proper strongly admissible ideal in $\mathbb{N} \times \mathbb{N}$.

We now introduce our main definitions.

Definition 6. *A double sequence $x = (x_{kl})$ of points in a topological group X, is said to be $\mathcal{I}_2$- double statistically convergent to L or $S(\mathcal{I}_2)-$convergent to L, if for each neighbourhood U of 0 and $\delta > 0$*

$$\{(m,n) \in \mathbb{N} \times \mathbb{N} : \frac{1}{mn} |\{k \leq m \text{ and } l \leq m : x_{kl} - L \notin U\}| \geq \delta\} \in \mathcal{I}_2.$$

In this case we write $x_{kl} \to L(S(\mathcal{I}_2))$. The set of all $\mathcal{I}_2$ - double statistically convergent sequences will be denoted by simply $S(\mathcal{I}_2)(X)$.

Remark 1. For $\mathcal{I}_2 = \mathcal{I}_{2fin} = \{A \subset \mathbb{N} \times \mathbb{N}, A \text{ is a finite}\}, S(\mathcal{I}_2)-$convergence coincides with double statistical convergence in a topological group X which was studied by Cakalli and Savaş [3].

Definition 7. *A sequence $x = (x_{kl})$ of points in a topological group X, is said to be $\mathcal{I}_2^\lambda-$ double statistically convergent to L or $S(\mathcal{I}_2^\lambda)-$convergent to L if for each neighbourhood U and any $\delta > 0$*

$$\left\{(m,n) \in \mathbb{N} \times \mathbb{N} : \frac{1}{\lambda_{mn}} |\{(k,l) \in I_{mn} : x_{kl} - L \notin U\}| \geq \delta\right\} \in \mathcal{I}_2.$$

In this case, we write

$$S^\lambda(\mathcal{I}_2) - \lim_{k,l \to \infty} x_{kl} = L \text{ or } x_{kl} \to L(S^\lambda(\mathcal{I}_2))$$

and define

$$S^\lambda(\mathcal{I}_2)(X) = \{x = (x_{kl}) : \text{ for some } L, S^\lambda(\mathcal{I}_2) - \lim_{k,l \to \infty} x_{kl} = L$$

Remark 2. For $\mathcal{I}_2 = \mathcal{I}_{2fin} = \{A \subset \mathbb{N} \times \mathbb{N}, A \text{ is finite}\}, \mathcal{I}_2^{\lambda}-$ double statistical convergence becomes $\lambda-$ double statistical convergence in topological groups and for $\lambda_{mn} = mn, \mathcal{I}_2^{\lambda}-$ double statistical convergence becomes double statistical convergence in topological groups.

It is obvious that every $\mathcal{I}_2^{\lambda}-$ double statistically convergent sequence has only one limit, that is, if a double sequence is $\mathcal{I}_2^{\lambda}-$ statistically convergent to L_1 and L_2 then $L_1 = L_2$.

We now prove the following theorems.

Theorem 2.

$$S(\mathcal{I}_2)(X) \subset S^{\lambda}(\mathcal{I}_2)(X) \text{ if } \lim_{n\to\infty} \inf \frac{\lambda_{mn}}{mn} > 0.$$

Proof. Let us take any neighbourhood U of 0. Then

$$\begin{aligned}\frac{1}{mn} \left|\{k \leq m, l \leq n : x_{kl} - L \notin U\}\right| &\geq \frac{1}{mn} \left|\{(k,l) \in I_{mn} : x_{kl} - L \notin U\}\right| \\ &= \frac{\lambda_{mn}}{mn} \frac{1}{\lambda_{mn}} \left|\{(k,l) \in I_{mn} : x_{kl} - L \notin U\}\right|.\end{aligned}$$

If $\lim_{mn\to\infty} \inf \frac{\lambda_{mn}}{mn} = a$ then from definition $\left\{(m,n) \in N \times N : \frac{\lambda_{mn}}{mn} < \frac{a}{2}\right\}$ is finite. For $\delta > 0$, and any neighbourhood U of 0,

$$\begin{aligned}&\left\{(m,n) \in \mathbb{N} \times \mathbb{N} : \frac{1}{\lambda_{mn}} \left|\{(k,l) \in I_{mn} : x_{kl} - L \notin U\}\right| \geq \delta\right\} \\ \subset &\left\{(k,l) \in I_{mn} : \frac{1}{mn} \left|\{k \leq m, l \leq n : x_{kl} - L \notin U\}\right| \geq \frac{a}{2}\delta\right\} \cup \\ &\left\{(m,n) \in \mathbb{N} \times \mathbb{N} : \frac{\lambda_{mn}}{mn} < \frac{a}{2}\right\}.\end{aligned}$$

The set on the right hand side belongs to $\mathcal{I}_2$ and this completes the proof.

Theorem 3. *Let $\lambda = (\lambda_{mn})$ and $\mu = (\mu_{mn})$ be two sequences in Δ such that $\lambda_{mn} \leq \mu_{mn}$ for all $(m,n) \in \mathbb{N} \times \mathbb{N}$,*

(i) If

$$\lim_{mn\to\infty} \inf \frac{\lambda_{mn}}{\mu_{mn}} > 0 \tag{1}$$

then $S^{\mu}(\mathcal{I}_2)(X) \subseteq S^{\lambda}(\mathcal{I}_2)(X)$.

(ii) If

$$\lim_{mn\to\infty} \frac{\lambda_{mn}}{\mu_{mn}} = 1 \tag{2}$$

then $S^{\lambda}(\mathcal{I}_2)(X) \subseteq S^{\mu}(\mathcal{I}_2)(X)$.

Proof. (i) Suppose that $\lambda_{mn} \leq \mu_{mn}$ for all $(m,n) \in \mathbb{N}\times\mathbb{N}$ and let (1) be satisfied. For neighbourhood U of 0, we have

$$\{(k,l) \in J_{mn} : x_{kl} - L \notin U\} \supseteq \{(k,l) \in I_{mn} : x_{kl} - L \notin U\}.$$

Therefore we can write

$$\frac{1}{\mu_{mn}} |\{(k,l) \in J_{mn} : x_{kl} - L \notin U\}| \geq \frac{\lambda_{mn}}{\mu_{mn}} \frac{1}{\lambda_{mn}} |\{(k,l) \in I_{mn} : x_{kl} - L \notin U\}|$$

and so for all $(m,n) \in \mathbb{N} \times \mathbb{N}$ we have,

$$\left\{(m,n) \in \mathbb{N} \times \mathbb{N} : \frac{1}{\lambda_{mn}} |\{(k,l) \in I_{mn} : x_{kl} - L \notin U\}| \geq \delta\right\}$$
$$\subseteq \left\{(m,n) \in \mathbb{N} \times \mathbb{N} : \frac{1}{\mu_{mn}} |\{(k,l) \in J_{mn} : x_{kl} - L \notin U\}| \geq \delta \frac{\lambda_{mn}}{\mu_{mn}}\right\} \in \mathcal{I}_2.$$

Hence $S^{\mu}(\mathcal{I}_2)(X) \subseteq S^{\lambda}(\mathcal{I}_2)(X)$.

(ii) Let $x = (x_{kl}) \in S^{\lambda}(\mathcal{I}_2)(X)$ and let (2) be satisfied. Since $I_{mn} \subset J_{mn}$, for neighbourhood U of 0, we may write

$$\begin{aligned}
&\frac{1}{\mu_{mn}} |\{(k,l) \in J_{mn} : x_{kl} - L \notin U\}| \\
&= \frac{1}{\mu_{mn}} |\{m - \mu_m + 1 < k \leq m - \lambda_n, n - \mu_n + 1 < k \leq n - \lambda_n : x_{kl} - L \notin U\}| \\
&\quad + \frac{1}{\mu_{mn}} |\{(k,l) \in I_{mn} : x_{kl} - L \notin U\}| \\
&\leq \frac{\mu_{mn} - \lambda_{mn}}{\mu_{mn}} + \frac{1}{\lambda_{mn}} |\{(k,l) \in I_{mn} : x_{kl} - L \notin U\}| \\
&\leq \left(\frac{\mu_{mn} - \lambda_{mn}}{\lambda_{mn}}\right) + \frac{1}{\lambda_{mn}} |\{(k,l) \in I_{mn} : x_{kl} - L \notin U\}| \\
&= \left(\frac{\mu_{mn}}{\lambda_{mn}} - 1\right) + \frac{1}{\lambda_{mn}} |\{(k,l) \in I_{mn} : x_{kl} - L \notin U\}|
\end{aligned}$$

for all $(m,n) \in \mathbb{N} \times \mathbb{N}$. Hence for $\delta > 0$ we have

$$\left\{(m,n) \in \mathbb{N} \times \mathbb{N} : \frac{1}{\mu_{mn}} |\{(k,l) \in J_{mn} : x_{kl} - L \notin U\}| \geq \delta\right\}$$
$$\subseteq \left\{(m,n) \in \mathbb{N} \times \mathbb{N} : \frac{1}{\lambda_{mn}} |\{(k,l) \in I_{mn} : x_{kl} - L \notin U\}| \geq \delta\right\} \in \mathcal{I}.$$

This implies that $S^{\lambda}(\mathcal{I}_2)(X) \subseteq S^{\mu}(\mathcal{I}_2)(X)$.

Finally we prove the following theorem.

Theorem 4. *If $\lambda \in \Delta$ be such that $\lim_{m,n} \frac{\mu_{mn}}{\lambda_{mn}} = 1$, then $S^{\lambda}(\mathcal{I}_2)(X) \subset S(\mathcal{I}_2)(X)$.*

Proof. Let $\delta > 0$ be given. Since $\lim_{mn} \frac{\mu_{mn}}{\lambda_{mn}} = 1$, we can choose $(m_0, n_0) \in \mathbb{N} \times \mathbb{N}$ such that $\left|\frac{\mu_{mn}}{\lambda_{mn}} - 1\right| < \frac{\delta}{2}$, for all $m \geq m_0$ and $n \geq n_0$. Let us take any neighbourhood U of 0. Now observe that,

$$\begin{aligned}\frac{1}{mn}\left|\{k \leq m, l \leq n : x_{kl} - L \notin U\}\right| &= \frac{1}{mn}\left|\{k \leq m - \lambda_m, l \leq n - \lambda_n : x_{kl} - L \notin U\}\right| \\ &\quad + \frac{1}{mn}\left|\{(k,l) \in I_{mn} : x_{kl} - L \notin U\}\right| \\ &\leq \frac{mn - \lambda_{mn}}{mn} + \frac{1}{mn}\left|\{(k,l) \in I_{mn} : x_{kl} - L \notin U\}\right| \\ &= \frac{\delta}{2} + \frac{1}{mn}\left|\{(k,l) \in I_{mn} : x_{kl} - L \notin U\}\right|\end{aligned}$$

for all $m > m_0$ and $n > n_0$. Hence for $\delta > 0$ and any neighbourhood U of 0,

$$\begin{aligned}&\left\{(m,n) \in \mathbb{N} \times \mathbb{N} : \frac{1}{mn}\left|\{k \leq m, l \leq n : x_{kl} - L \notin U\}\right| \geq \delta\right\} \\ &\subset \left\{(m,n) \in \mathbb{N} \times \mathbb{N} : \frac{1}{\lambda_{mn}}\left|\{(k,l) \in I_{mn} : x_{kl} - L \notin U\}\right| \geq \frac{\delta}{2}\right\} \cup A\end{aligned}$$

where A is the union of the first m_0 rows and the first n_0 columns of the double sequence $\left\{\frac{\lambda_{mn}}{mn}\right\}$. If $S^{\lambda}(\mathcal{I}_2) - \lim x = L$ then the set on the right hand side belongs to $\mathcal{I}_2$ and so the set on the left hand side also belongs to $\mathcal{I}_2$. This shows that $x = (x_{kl})$ is $\mathcal{I}_2$–double statistically convergent to L.

Remark 3. We do not know whether the condition in Theorem 3 is necessary and so we leave it as an open problem.

References

1. Balcerzak, M., Dems, K., Komisarski, A.: Statistical convergence and ideal convergence for sequences of functions. J. Math. Anal. Appl. **328**, 715–729 (2007)
2. Buck, R.C.: Generalized asymptotic density. Amer J. Math. **75**, 335–346 (1953)
3. Cakalli, H.: On Statistical Convergence in topological groups. Pure Appl. Math. Sci. **43**(1–2), 27–31 (1996)
4. Cakalli, H., Savaş, E.: Statistical convergence of double seqeunces in topological groups. J. Comput. Anal. Appl. **12**(2), 421–426 (2010)
5. Das, P., Ghosal, S.: Some further results on $\mathcal{I}$-Cauchy sequences and condition (AP). Comput. Math. Appl. **59**, 2597–2600 (2010)
6. Das, P., Savaş, E., Ghosal, S.K.: On generalizations of certain summability methods using ideals. Appl. Math. Lett. **24**, 1509–1514 (2011)
7. Das, P., Kostyrko, P., Wilczynski, W., Malik, P.: $\mathcal{I}-$ and $\mathcal{I}^*-$ convergence of double sequences. Math. Slovaca **58**, 605–620 (2008)
8. Das, P., Savaş, E.: On $\mathcal{I}$-convergence of nets in locally solid Riesz spaces. Filomat **27**(1), 84–89 (2013)
9. Dems, K.: On $\mathcal{I}-$Cauchy sequences, Real Anal. Exchance **30**, 123–128 (2004/2005)
10. Fast, H.: Sur la convergence statistique. Colloq Math. **2**, 241–244 (1951)

11. Fridy, J.A.: On ststistical convergence. Analysis **5**, 301–313 (1985)
12. Hazarika, B., Savaş, E.: (λ, μ)-statistical convergence of double sequences in normed spaces. Note Mat. **32**(2), 101–114 (2012)
13. Kolk, E.: The statistical convergence in Banach spaces. Tartu Ul Toime **928**, 41–52 (1991)
14. Kostyrko, P., Šalát, T., Wilczynki, W.: $\mathcal{I}-$ convergence. Real Anal. Exch. **26**(2), 669–685 (2000/2001)
15. Komisarski, A.: Pointwise $\mathcal{I}-$convergence and $\mathcal{I}^*-$ convergence in measure of sequences of functions. J. Math. Anal. Appl. **340**, 770–779 (2008)
16. Lahiri, B.K., Das, P.: On $\mathcal{I}-$ and $\mathcal{I}^*-$ convergence of nets. Real Anal. Exch. **33**(2), 431–442 (2007/2008)
17. Maddox, I.J.: Statistical convergence in locally convex spaces. Math. Camb. Phil. Soc. **104**, 141–145 (1988)
18. Maio, G.D., Kocinac, L.D.R.: Statistical convergence in topology. Topology Appl. **156**, 28–45 (2008)
19. Mursaleen, M.: λ- statistical convergence. Math. Slovaca **50**, 111–115 (2000)
20. Mursaleen, M., Edely, O.H.: Statistical convergence of double sequences. J. Math. Anal. Appl. **288**, 223–231 (2003)
21. Pringsheim, A.: Zur theorie der zweifach unendlichen Zahlenfolgen. Math. Ann. **53**, 289–321 (1900)
22. Sahiner, A., Gurdal, M., Saltan, S., Gunawan, H.: Ideal convergence in 2-normed spaces. Taiwanese J. Math. **11**, 1477–1484 (2007)
23. Šalát, T.: On statistically convergent sequences of real numbers. Math. Slovaca **30**, 139–150 (1980)
24. Savaş, E., Patterson, R.F.: Lacunary statistical convergence of multiple sequences. Appl. Math. Lett. **19**(6), 527–534 (2006)
25. Savaş, E.: I-statistically convergent sequences in topological groups. In: International conference Kangro-100. Methods of Analysis and Algebra, dedicated to the Centennial of Professor Gunnar Kangro, Tartu, Estonia, 1–6 September (2013)
26. Savaş, E., Das, P.: A generalized statistical convergence via ideals. Appl. Math. Lett. **24**, 826–830 (2011)
27. Savaş, E., Das, P., Dutta, S.: A note on strong matrix summability via ideals. Appl. Math. Lett. **25**(4), 733–738 (2012)
28. Savaş, E.: A sequence spaces in 2-normed space defined by ideal convergence and an Orlicz function. Abstr. Appl. Anal. (2011). Art. ID 741382, 9 pages
29. Savaş, E.: On some new sequence spaces in 2-normed spaces using ideal convergence and an Orlicz function. J. Inequal. Appl. (2010). Art. ID 482392, 8 pages
30. Savaş, E.: Δ^m -strongly summable sequences spaces in 2-normed spaces defined by ideal convergence and an Orlicz function. Appl. Math. Comput. **217**(1), 271–276 (2010)
31. Savaş, E.: Lacunary statistical convergence of double sequences in topological groups, J. Inequal. Appl., Article Number: 480, 2 December 2014
32. Savaş, E., Mohiuddine, S.: λ-statistically convergent double sequences in probabilistic normed spaces. Math. Slovaca **62**(1), 99–108 (2012)
33. Schoenberg, I.J.: The integrability methods. Amer. Math. Monthly **66**, 361–375 (1959)

Superconvergence Results for Volterra-Urysohn Integral Equations of Second Kind

Moumita Mandal(✉) and Gnaneshwar Nelakanti

Department of Mathematics, Indian Institute of Technology, Kharagpur 721 302, India
abmoumita001@gmail.com, gnanesh@maths.iitkgp.ernet.in

Abstract. In this paper, we consider the Galerkin method to approximate the solution of Volterra-Urysohn integral equations of second kind with a smooth kernel, using piecewise polynomial bases. We show that the exact solution is approximated with the order of convergence $\mathcal{O}(h^r)$ for the Galerkin method, whereas the iterated Galerkin solutions converge with the order $\mathcal{O}(h^{2r})$ in uniform norm, where h is the norm of the partition and r is the smoothness of the kernel. For improving the accuracy of the approximate solution of the integral equation, the multi-Galerkin method is also discussed here and we prove that the exact solution is approximated with the order of convergence $\mathcal{O}(h^{3r})$ in uniform norm for iterated multi-Galerkin method. Numerical examples are given to illustrate the theoretical results.

Keywords: Volterra-Urysohn integral equations · Smooth kernels · Galerkin method · Multi-Galerkin method · Piecewise polynomials · Superconvergence rates

1 Introduction

We consider the second kind Volterra-Urysohn integral equation of the form

$$x(t) - \int_0^t k(t, s, x(s))\, ds = f(t), \quad 0 \le t \le 1, \tag{1.1}$$

where the kernel $k(.,.,x(.))$ and f are given smooth functions, x is the unknown function to be determined. Various projection methods such as Galerkin, collocation, Petrov-Galerkin and Nyström methods are available in literature for finding numerical solutions of nonlinear integral equations (see [2–4], [5,7,9,13,15,16,18,20–22]). In [9], a simple algorithm was given for obtaining starting value for the numerical solution of the Volterra-Urysohn integral equations. In [2], Blom and Brummer discussed the collocation and iterated collocation methods for Urysohn type second kind Volterra integral equations and proved that the order of convergence of iterated collocation method is twice that of the collocation method at the knots. In [22], an interpolation post-processing

© Springer Nature Singapore Pte Ltd. 2017
D. Giri et al. (Eds.): ICMC 2017, CCIS 655, pp. 358–379, 2017.
DOI: 10.1007/978-981-10-4642-1_31

technique was proposed in the piecewise polynomial space of degree not exceeding $(r-1)$ and obtained global superconvergence of $\mathcal{O}(h^{2r})$ for Volterra-Urysohn integral equations. In [20], Wan et al. discussed spectral Galerkin method for second-kind Volterra-Urysohn integral equations and showed that the errors of the spectral approximations decay exponentially, provided that the kernel function and the source function are sufficiently smooth. In [7], using the Picard iteration method and treating the involved integration by numerical quadrature formulas, a numerical scheme was given for the second kind Volterra-Urysohn integral equations. For enlarging the convergence region of the Picard iteration method, multistage algorithm was also proposed.

In this paper, we apply Galerkin method to solve Volterra-Urysohn integral equations (1.1) with a smooth kernel using piecewise polynomial basis functions. We will show that the exact solution is approximated with the order of convergence $\mathcal{O}(h^r)$ in Galerkin method, whereas the iterated Galerkin solutions converge with the orders $\mathcal{O}(h^{2r})$ in uniform norm, where h is the norm of the partition and r is the smoothness of the kernel.

For improving the accuracy of approximate solution of the integral equation, multi-projection method was proposed to solve the linear Fredholm integral equations of second kind (see. [8,12,14]) and in [10,11], this method was extented to solve the non-linear Fredholm integral equations. Here we also discuss the multi-projection method for Volterra-Urysohn integral equations (1.1) to improve the order of convergence. If the projection operator is an orthogonal projection operator then the corresponding multi-projection and iterated multi-projection methods are called multi-Galerkin (M-Galerkin) and iterated multi-Galerkin (iterated M-Galerkin) methods. We will prove that iterated multi-Galerkin solutions converge to the exact solution with the order of convergence $\mathcal{O}(h^{3r})$ in uniform norm. Thus the iterated multi-Galerkin method improves over the iterated Galerkin method.

We organize this paper as follows. In Sect. 2, we apply the Galerkin method to solve the Eq. (1.1) and discuss the convergence results. In Sect. 3, we consider the multi-Galerkin method and its iterated version to obtain superconvergence results. In Sect. 4, numerical results are given to illustrate the theoretical results. Throughout this paper, we assume that c is a generic constant.

2 Galerkin Method: Volterra-Urysohn Integral Equations with a Smooth Kernel

Let $\mathbb{X} = L^2[0,1]$. Consider the following Volterra integral equation of second kind

$$x(t) - \int_0^t k(t,s,x(s))\,ds = f(t),\ \ t \in [0,1], \tag{2.1}$$

where the kernel $k(.,.,x(.))$ and f are given smooth functions, x is the unknown function to be determined. Consider a transformation $s(.,.) : ([0,1] \times [0,1]) \to [0,1]$, by taking $s = t\tau$, $(t,\tau) \in ([0,1] \times [0,1])$, then Volterra integral equation (2.1) becomes

$$x(t) - \int_0^1 \ell(t, s(t,\tau), x(s(t,\tau)))\, d\tau = f(t),\ t \in [0,1], \tag{2.2}$$

where $\ell(t, s(t,\tau), x(s(t,\tau))) = tk(t, s(t,\tau), x(s(t,\tau)))$.

Define

$$\mathcal{K}(x)(t) = \int_0^1 \ell(t, s(t,\tau), x(s(t,\tau)))\, d\tau, \quad x \in \mathbb{X}.$$

Then the above Eq. (2.2) can be written as

$$x - \mathcal{K}(x) = f. \tag{2.3}$$

The Fréchet derivative $\mathcal{K}'(x)$ is defined by

$$(\mathcal{K}'(x)y)(t) = \int_0^1 \frac{\partial}{\partial x}\ell(t, s(t,\tau), x(s(t,\tau)))y(s(t,\tau))\, d\tau = \int_0^1 \ell_u(t, s(t,\tau), x(s(t,\tau)))y(s(t,\tau))\, d\tau, \quad y \in \mathbb{X},$$

where $\ell_u(t, s(t,\tau), x(s(t,\tau))) = \frac{\partial}{\partial x}\ell(t, s(t,\tau), x(s(t,\tau)))$.

Let $\mathcal{C}^r[0,1]$ denote the space of r-times continuously differentiable functions and for any $u \in \mathcal{C}^r[0,1]$, denote

$$\|u\|_{r,\infty} = \max\{\|u^{(j)}\|_\infty : 0 \le j \le r\},$$

where $u^{(j)}$ denotes the j-th derivative of u.

Throughout the paper, the following assumptions are made on f, $\ell(.,.,.)$ and $\ell_u(.,.,.)$:

(i) $f \in \mathbb{X}$.

(ii) $\ell_u(t, s(t,\tau), x(s(t,\tau))) \in \mathcal{C}([0,1]\times[0,1]\times\mathbb{R}) \subseteq L^2([0,1]\times[0,1]\times\mathbb{R})$, $M = \|\ell_u\|_{L^2} = \left[\int_0^1 |\ell_u(t, s(t,\tau), x(s(t,\tau)))|^2\right]^{\frac{1}{2}} < \infty$.

(iii) $\ell_u(t, s(t,\tau), x(s(t,\tau))) \in \mathcal{C}^r([0,1]\times[0,1]\times\mathbb{R}), r \ge 1$.

(iv) The kernel $\ell(t, s(t,\tau), x(s(t,\tau)))$ and $\ell_u(t, s(t,\tau), x(s(t,\tau)))$, satisfies Lipschitz conditions in the third variable x, i.e., for any $x_1, x_2 \in \mathbb{R}$, $\exists\ c_1, c_2 > 0$ such that

$$|\ell(t, s(t,\tau), x_1(s(t,\tau))) - \ell(t, s(t,\tau), x_2(s(t,\tau)))| \le c_1|x_1(s(t,\tau)) - x_2(s(t,\tau))|,$$

$$|\ell_u(t, s(t,\tau), x_1(s(t,\tau))) - \ell_u(t, s(t,\tau), x_2(s(t,\tau)))| \le c_2|x_1(s(t,\tau)) - x_2(s(t,\tau))|.$$

Next, we define the operator $\mathcal{T}$ on $\mathbb{X}$ by

$$\mathcal{T}u := f + \mathcal{K}(u), \quad u \in \mathbb{X},$$

then the Eq. (2.3) can be written as

$$x = \mathcal{T}x. \tag{2.4}$$

Note that if $c_1 < 1$, then using assumption (iv), it follows that the Eq. (2.4), has unique solution, say $x_0 \in \mathbb{X}$.

In the following theorem, we show that $\mathcal{K}'(x_0)$ is a compact operator on $\mathbb{X}$.

Theorem 1. *Let $x_0 \in \mathbb{X}$ and the kernel $\ell_u(.,.,.) \in \mathcal{C}([0,1]\times[0,1]\times\mathbb{R}) \subseteq L^2([0,1]\times[0,1]\times\mathbb{R})$. Then the linear operator $\mathcal{K}'(x_0) : \mathbb{X} \to \mathcal{C}[0,1]$ is a compact operator.*

Proof. Let $S = \{\mathcal{K}'(x_0)y : y \in B \subseteq L^2\}$, where B denotes the closed unit ball in L^2. In order to to prove that $\mathcal{K}'(x_0)$ is a compact operator, it is enough to prove that S is uniformly bounded and equicontinuous.
For any $x_0, y \in \mathbb{X}$, we have

$$\begin{aligned}|\mathcal{K}'(x_0)y(t)| &= \left|\int_0^1 \ell_u(t,s(t,\tau),x_0(s(t,\tau)))y(s(t,\tau))\,d\tau\right| \\ &\le \left[\int_0^1 |\ell_u(t,s(t,\tau),x_0(s(t,\tau)))|^2 d\tau\right]^{\frac{1}{2}} \left[\int_0^1 |y(s(t,\tau))|^2 d\tau\right]^{\frac{1}{2}} \\ &\le M\|y\|_{L^2} \le M.\end{aligned}$$

This implies $\|\mathcal{K}'(x_0)y\|_\infty = \sup_{t\in[0,1]} |(\mathcal{K}'(x_0))y(t)| \le M < \infty$.
Hence

$$\|\mathcal{K}'(x_0)\|_\infty \le M < \infty, \tag{2.5}$$

i.e., the set S is uniformly bounded.

Let $\epsilon > 0$ be given. Since $y \in \mathbb{X}$ and $\mathcal{C}[0,1]$ is dense in $L^2[0,1]$, it follows (see [17], p.71) that there exists $g \in \mathcal{C}[0,1] \subseteq L^2[0,1]$ such that $\|y(.) - g(.)\|_{L^2} < \frac{\epsilon}{3}$. We can also find $\delta > 0$ such that $\|g(s(t,.)) - g(s(t',.))\|_{L^2} \le \|g(s(t,.)) - g(s(t',.))\|_\infty < \frac{\epsilon}{3}$, for all $t, t' \in [0,1]$, satisfying $|t - t'| < \delta$. Since $\ell_u(.,.,.) \in \mathcal{C}([0,1]\times[0,1]\times\mathbb{R}) \subseteq L^2([0,1]\times[0,1]\times\mathbb{R})$, we have $\|\ell_u(t,s(t,\tau),x(s(t,\tau))) - \ell_u(t',s(t',\tau),x(s(t',\tau)))\|_{L^2} \le \|\ell_u(t,s(t,\tau),x(s(t,\tau))) - \ell_u(t',s(t',\tau),x(s(t',\tau)))\|_\infty \to 0$ uniformly as $t \to t'$. Using this with the boundedness of $\|\ell\|_{L^2}$, for any $t, t' \in [0,1]$, we have

$$\begin{aligned}&|\mathcal{K}'(x_0)y(t) - \mathcal{K}'(x_0)y(t')| \\ &= \left|\int_0^1 [\ell_u(t,s(t,\tau),x_0(s(t,\tau)))y(s(t,\tau)) \right. \\ &\quad \left. - \ell_u(t',s(t',\tau),x_0(s(t',\tau)))y(s(t',\tau))]\,d\tau\right| \\ &\le \left|\int_0^1 [\ell_u(t,s(t,\tau),x_0(s(t,\tau))) - \ell_u(t',s(t',\tau),x_0(s(t',\tau)))]y(s(t,\tau))\,d\tau\right| \\ &\quad + \left|\int_0^1 \ell_u(t',s(t',\tau),x_0(s(t',\tau)))[y(s(t,\tau)) - y(s(t',\tau))]\,d\tau\right| \\ &\le \|\ell_u(t,s(t,\tau),x_0(s(t,\tau))) - \ell_u(t',s(t',\tau),x_0(s(t',\tau)))\|_{L^2}\|y\|_{L^2} \\ &\quad + \|\ell_u\|_{L^2}\|y(s(t,\tau)) - y(s(t',\tau))\|_{L^2} \\ &\le \|\ell_u(t,s(t,\tau),x_0(s(t,\tau))) - \ell_u(t',s(t',\tau),x_0(s(t',\tau)))\|_{L^2} \\ &\quad + M\|y(s(t,\tau)) - y(s(t',\tau))\|_{L^2}.\end{aligned} \tag{2.6}$$

Now we consider

$$\begin{aligned}
&\|y(s(t,\tau)) - y(s(t',\tau))\|_{L^2} \\
&\quad = \|y(s(t,\tau)) - g(s(t,\tau)) + g(s(t,\tau)) - g(s(t',\tau)) + g(s(t',\tau)) - y(s(t',\tau))\|_{L^2} \\
&\quad \leq \|y(s(t,\tau)) - g(s(t,\tau))\|_{L^2} + \|g(s(t,\tau)) - g(s(t',\tau))\|_{L^2} \\
&\qquad + \|g(s(t',\tau)) - y(s(t',\tau))\|_{L^2} \\
&\quad < \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon, \qquad (2.7)
\end{aligned}$$

for all $t, t' \in [0,1]$, satisfying $|t - t'| < \delta$.
Hence combining the estimates (2.6) and (2.7), we get

$$|\mathcal{K}'(x_0)y(t) - \mathcal{K}'(x_0)y(t')| \to 0, \ as \ t \to t'.$$

Hence by Arzela-Ascoli theorem, S is a relatively compact set, i.e., $\mathcal{K}'(x)$ is a compact operator. □

Next, we apply Galerkin method to solve the Eq. (2.3). For this, we consider $\Pi_n : 0 = t_0 < t_1 < \ldots < t_n = 1$, a partition of $[0,1]$ and let $h_i = \{t_i - t_{i-1} : 1 \leq i \leq n\}$, $h = \max h_i$ denotes the norm of the partition. We assume that $h \to 0$, as $n \to \infty$. Here we let the approximating subspaces $\mathbb{X}_n = S^{\nu}_{r,n}(\Pi_n)$, the space of all piecewise polynomials of order r (i.e., of degree $\leq r-1$) with breakpoints at $t_1, \cdots, t_{n-1}$ and with ν $(-1 \leq \nu \leq r-2)$ continuous derivatives. Here $\nu = 0$ corresponds to the case of continuous piecewise polynomials. If $\nu = -1$, there is no continuity requirements at the break points, in such case $u_n \in \mathbb{X}_n$ is arbitrarily taken to be left continuous at $t_1, \ldots, t_n$ and right continuous at t_0.

Orthogonal projection operator: Let the operator $\mathcal{P}_n : L^2[0,1] \to \mathbb{X}_n$ be the orthogonal projection operator defined by

$$\langle \mathcal{P}_n u, v\rangle = \langle u, v\rangle, \ v \in \mathbb{X}_n, u \in \mathbb{X}, \qquad (2.8)$$

where $\langle u, \ v\rangle = \int_0^1 u(t)v(t)dt$.

We first quote the following lemma from Chatelin [6].

Lemma 1. *Let $\mathcal{P}_n : \mathbb{X} \to \mathbb{X}_n$ be the orthogonal projection operator defined by (2.8). Then there hold*

i) $\mathcal{P}_n$ is uniformly bounded in uniform norm, i.e., $\exists$ a constant p independent of n such that

$$\|\mathcal{P}_n\|_\infty \leq p < \infty. \qquad (2.9)$$

ii)

$$\|\mathcal{P}_n u - u\|_\infty \to 0, \ as \ n \to \infty, \ u \in \mathbb{X}. \qquad (2.10)$$

iii) In particular if $u \in \mathcal{C}^r[0,1]$, then

$$\|(\mathcal{I} - \mathcal{P}_n)u\|_\infty \leq ch^r \|u\|_{r,\infty}, \qquad (2.11)$$

where c is a constant independent of n.

The Galerkin method for solving (2.3) is seeking an approximation $x_n \in \mathbb{X}_n$ such that

$$x_n - \mathcal{P}_n\mathcal{K}(x_n) = \mathcal{P}_n f. \tag{2.12}$$

Let $\mathcal{T}_n$ be the operator defined by

$$\mathcal{T}_n(u) := \mathcal{P}_n\mathcal{K}(u) + \mathcal{P}_n f. \tag{2.13}$$

Then the Eq. (2.12) can be written as

$$x_n = \mathcal{T}_n x_n. \tag{2.14}$$

In order to obtain more accurate approximation solution for (2.3), we further consider the iterated approximate solution as

$$\tilde{x}_n = f + \mathcal{K}(x_n). \tag{2.15}$$

Using $\mathcal{P}_n\tilde{x}_n = x_n$, the Eq. (2.15) can be written as

$$\tilde{x}_n - \mathcal{K}(\mathcal{P}_n\tilde{x}_n) = f. \tag{2.16}$$

Taking $\widetilde{\mathcal{T}}_n(u) := \mathcal{K}(\mathcal{P}_n u) + f$, $u \in \mathbb{X}$, the Eq. (2.16) can be written as $\tilde{x}_n = \widetilde{\mathcal{T}}_n\tilde{x}_n$.

Now we discuss the existence and uniqueness of the approximate and iterated approximate solutions. Let $BL(\mathbb{X})$ denote the space of all bounded linear operator on $\mathbb{X}$ and we recall the definition of ν-convergence and a theorem from [1], which are useful in proving existence and convergence of approximated solutions.

Definition 1. *(ν-**convergence**) Let $\mathcal{T} \in BL(\mathbb{X})$ and $\{\mathcal{T}_n\}$ be a sequence in $BL(\mathbb{X})$, then $\{\mathcal{T}_n\}$ is said to be ν convergent to $\mathcal{T}$ if $\|\mathcal{T}_n\| \leq C, \|(\mathcal{T}_n - \mathcal{T})\mathcal{T}\| \to 0$ and $\|(\mathcal{T}_n - \mathcal{T})\mathcal{T}_n\| \to 0, as\ n \to \infty$.*

Theorem 2. *Let $\mathbb{X}$ be a Banach space and $\mathcal{T}$, $\mathcal{T}_n \in BL(\mathbb{X})$. If $\mathcal{T}_n$ is norm convergent or ν-convergent to $\mathcal{T}$ and $(\mathcal{I} - \mathcal{T})^{-1}$ exists and is bounded on $\mathbb{X}$, then for sufficiently large n, $(\mathcal{I} - \mathcal{T}_n)^{-1}$ exists and is uniformly bounded on $\mathbb{X}$.*

We quote the following theorem from [19], which gives us the condition under which the solvability of one equation leads to the solvability of another equation.

Theorem 3 [19]**.** *Let $\widehat{\mathcal{F}}$ and $\widetilde{\mathcal{F}}$ be continuous operators over an open set Ω in a Banach space $\mathbb{X}$. Let the equation $x = \widetilde{\mathcal{F}}x$ has an isolated solution $\tilde{x}_0 \in \Omega$ and let the following conditions be satisfied.*

(a) The operator $\widehat{\mathcal{F}}$ is Fréchet differentiable in some neighborhood of the point $\tilde{x}_0$, while the linear operator $\mathcal{I} - \widehat{\mathcal{F}'}(\tilde{x}_o)$ is continuously invertible.

(b) Suppose that for some $\delta > 0$ and $0 < q < 1$, the following inequalities are valid (the number δ is assumed to be sufficiently small so that the sphere $\|x - \tilde{x}_0\| \leq \delta$ is contained within Ω)

$$\sup_{\|x-\tilde{x}_0\|\le\delta} \|(\mathcal{I}-\widehat{\mathcal{F}'}(\tilde{x}_o))^{-1}(\widehat{\mathcal{F}'}(x)-\widehat{\mathcal{F}'}(\tilde{x}_o)\| \le q, \tag{2.17}$$

$$\alpha = \|(\mathcal{I}-\widehat{\mathcal{F}'}(\tilde{x}_o))^{-1}(\widehat{\mathcal{F}}(\tilde{x}_o)-\widetilde{\mathcal{F}}(\tilde{x}_o)\| \le \delta(1-q). \tag{2.18}$$

Then the equation $x = \widehat{\mathcal{F}}x$ has a unique solution $\hat{x}_0$ in the sphere $\|x-\tilde{x}_0\| \le \delta$. Moreover, the inequality

$$\frac{\alpha}{1+q} \le \|\hat{x}_0-\tilde{x}_0\| \le \frac{\alpha}{1-q} \tag{2.19}$$

is valid.

Now we discuss the existence and convergence rates of the approximate solution x_n to x_0. To do this, we first prove the following lemmas.

Lemma 2. *For any $x, y \in \mathbb{X}$, the following hold*

$$\|\mathcal{K}(x)-\mathcal{K}(y)\|_\infty \le c_1\|x-y\|_\infty,$$

$$\|\mathcal{K}'(x)-\mathcal{K}'(y)\|_\infty \le c_2\|x-y\|_\infty.$$

Proof. Using Lipschitz's continuity of $\ell(.,.,.)$, for any $x, y, z \in \mathbb{X}$, we have

$$\begin{aligned}
\|(\mathcal{K}(x)-\mathcal{K}(y))z\|_\infty &= \sup_{t\in[0,1]} |(\mathcal{K}(x)-\mathcal{K}(y))z(t)| \\
&= \sup_{t\in[0,1]} \left| \int_0^1 [\ell(t,s(t,\tau),x(s(t,\tau))) - \ell(t,s(t,\tau),y(s(t,\tau)))]z(s(t,\tau))\,d\tau \right| \\
&\le \|\ell(t,s(t,\tau),x(s(t,\tau))) - \ell(t,s(t,\tau),y(s(t,\tau)))\|_{L^2}\|z\|_{L^2} \\
&\le c_1\|x-y\|_{L^2}\|z\|_{L^2} \\
&\le c_1\|x-y\|_\infty\|z\|_\infty.
\end{aligned}$$

This implies

$$\|\mathcal{K}(x)-\mathcal{K}(y)\|_\infty \le c_1\|x-y\|_\infty.$$

On similar lines, using Lipschitz's continuity of $\ell_u(.,.,.)$, we obtain

$$\|\mathcal{K}'(x)-\mathcal{K}'(y)\|_\infty\| \le c_2\|x-y\|_\infty. \tag{2.20}$$

Hence the proof follows. □

Lemma 3. *Let $\mathcal{T}'(x_0)$ and $\widetilde{\mathcal{T}}_n'(x_0)$ be the Fréchet derivatives of $\mathcal{T}(x)$ and $\widetilde{\mathcal{T}}_n(x)$, respectively at x_0. Then*

$$\|(\mathcal{I}-\mathcal{P}_n)\widetilde{\mathcal{T}}_n'(x_0)\|_\infty \to 0, \quad n\to\infty,$$
$$\|(\mathcal{I}-\mathcal{P}_n)\mathcal{T}'(x_0)\|_\infty \to 0, \quad n\to\infty.$$

Proof. We have $\mathcal{T}'(x_0) = \mathcal{K}'(x_0)$ and since from Theorem 1, $\mathcal{K}'(x_0)$ is compact, hence we have

$$\|(\mathcal{I}-\mathcal{P}_n)\mathcal{T}'(x_0)\|_\infty = \|(\mathcal{I}-\mathcal{P}_n)\mathcal{K}'(x_0)\|_\infty \to 0, \quad as \quad n\to\infty. \tag{2.21}$$

Using estimate (2.9), we have

$$\begin{aligned}\|(\mathcal{I}-\mathcal{P}_n)\widetilde{\mathcal{T}}_n'(x_0)\|_\infty &= \|(\mathcal{I}-\mathcal{P}_n)\mathcal{K}'(\mathcal{P}_n x_0)\mathcal{P}_n\|_\infty \\ &= \|(\mathcal{I}-\mathcal{P}_n)[\mathcal{K}'(\mathcal{P}_n x_0)\mathcal{P}_n - \mathcal{K}'(x_0)\mathcal{P}_n + \mathcal{K}'(x_0)\mathcal{P}_n]\|_\infty \\ &\le \|(\mathcal{I}-\mathcal{P}_n)[\mathcal{K}'(\mathcal{P}_n x_0) - \mathcal{K}'(x_0)]\mathcal{P}_n\|_\infty + \|(\mathcal{I}-\mathcal{P}_n)\mathcal{K}'(x_0)\mathcal{P}_n\|_\infty \\ &\le (1+\|\mathcal{P}_n\|_\infty)\|\mathcal{K}'(\mathcal{P}_n x_0) - \mathcal{K}'(x_0)\|_\infty\|\mathcal{P}_n\|_\infty + \|(\mathcal{I}-\mathcal{P}_n)\mathcal{K}'(x_0)\|_\infty\|\mathcal{P}_n\|_\infty \\ &\le p\{c\|\mathcal{K}'(\mathcal{P}_n x_0) - \mathcal{K}'(x_0)\|_\infty + \|(\mathcal{I}-\mathcal{P}_n)\mathcal{K}'(x_0)\|_\infty\}. \end{aligned} \tag{2.22}$$

From estimate (2.10) and Lemma 2, the first term of estimate (2.22) becomes

$$\|\mathcal{K}'(\mathcal{P}_n x_0) - \mathcal{K}'(x_0)\|_\infty \le c_2\|(\mathcal{I}-\mathcal{P}_n)x_0\|_\infty \to 0, \ as\ n\to\infty. \tag{2.23}$$

Hence using estimates (2.21), (2.22) and (2.23), we get

$$\|(\mathcal{I}-\mathcal{P}_n)\widetilde{\mathcal{T}}_n'(x_0)\|_\infty \to 0, \quad n\to\infty.$$

This complete the proof. □

Theorem 4. *Let $x_0 \in \mathcal{C}^r[0,1], r \ge 1$, be an isolated solution of the Eq. (2.3). Assume that 1 is not an eigenvalue of the linear operator $\mathcal{K}'(x_0)$. Let $\mathcal{P}_n : \mathbb{X} \to \mathbb{X}_n$ be the orthogonal projection operator defined by (2.8). Then the Eq. (2.12) has a unique solution $x_n \in B(x_0,\delta) = \{x : \|x-x_0\|_\infty < \delta\}$ for some $\delta > 0$ and for sufficiently large n. Moreover, there exists a constant $0 < q < 1$, independent of n such that*

$$\frac{\alpha_n}{1+q} \le \|x_n - x_0\|_\infty \le \frac{\alpha_n}{1-q},$$

where $\alpha_n = \|(\mathcal{I}-\mathcal{T}_n'(x_0))^{-1}(\mathcal{T}_n(x_0)-\mathcal{T}(x_0))\|_\infty$. Further, we obtain

$$\|x_n - x_0\|_\infty = \mathcal{O}(h^r).$$

Proof. Using Lemma 3, we have

$$\|\mathcal{T}_n'(x_0) - \mathcal{T}'(x_0)\|_\infty = \|\mathcal{P}_n\mathcal{K}'(x_0) - \mathcal{K}'(x_0)\|_\infty = \|(\mathcal{I}-\mathcal{P}_n)\mathcal{K}'(x_0)\|_\infty = \|(\mathcal{I}-\mathcal{P}_n)\mathcal{T}'(x_0)\|_\infty \to 0 \ \ as \ \ n \ \ \to\infty.$$

Since 1 is not an eigenvalue of $\mathcal{T}'(x_0)$, i.e., $(\mathcal{I}-\mathcal{T}'(x_0))$ is invertible on $\mathbb{X}$, then by applying Theorem 2, we have $(\mathcal{I}-\mathcal{T}_n'(x_0))^{-1}$ exists and is uniformly bounded on $\mathbb{X}$, for some sufficiently large n, i.e., there exists a constant $A_1 > 0$ such that $\|(\mathcal{I}-\mathcal{T}_n'(x_0))^{-1}\|_\infty \le A_1 < \infty$.

Now from Lemma 2, we have for any $x \in B(x_0,\delta)$,

$$\begin{aligned}\|\mathcal{T}_n'(x_0) - \mathcal{T}_n'(x)\|_\infty &= \|\mathcal{P}_n\mathcal{K}'(x_0) - \mathcal{P}_n\mathcal{K}'(x)\|_\infty \\ &\le \|\mathcal{P}_n\|_\infty\|\mathcal{K}'(x_0) - \mathcal{K}'(x)\|_\infty \\ &\le pc_2\|x_0 - x\|_\infty \le c_2 p\delta. \end{aligned} \tag{2.24}$$

Hence, we have

$$\sup_{\|x-x_0\|\leq\delta} \|(\mathcal{I}-\mathcal{T}_n'(x_0))^{-1}(\mathcal{T}_n'(x_0)-\mathcal{T}_n'(x))\|_\infty \leq A_1 pc_2\delta \leq q(say).$$

Here we choose δ such that $0<q<1$. This proves the Eq. (2.17) of Theorem 3. Using estimate (2.10), we have

$$\begin{aligned}\alpha_n &= \|(\mathcal{I}-\mathcal{T}_n'(x_0))^{-1}(\mathcal{T}_n(x_0)-\mathcal{T}(x_0))\|_\infty\\ &\leq A_1\|\mathcal{T}_n(x_0)-\mathcal{T}(x_0)\|_\infty\\ &\leq A_1\|\mathcal{P}_n(f+\mathcal{K}x_0)-(f+\mathcal{K}x_0)\|_\infty\\ &\leq A_1\|(\mathcal{I}-\mathcal{P}_n)(f+\mathcal{K}x_0)\|_\infty\\ &\leq A_1\|(\mathcal{I}-\mathcal{P}_n)x_0\|_\infty\\ &\to 0, \quad as \quad n\to\infty. \end{aligned} \tag{2.25}$$

By choosing n large enough such that $\alpha_n \leq \delta(1-q)$, the Eq. (2.19) of Theorem 3 is satisfied, i.e.,

$$\frac{\alpha_n}{1+q} \leq \|x_n - x_0\|_\infty \leq \frac{\alpha_n}{1-q}.$$

Hence using estimate (2.11), it follows that

$$\|x_n - x_0\|_\infty \leq \frac{\alpha_n}{1-q} \leq \frac{1}{1-q}A_1\|\mathcal{T}_n(x_0)-\mathcal{T}(x_0)\|_\infty \leq cA_1\|(\mathcal{I}-\mathcal{P}_n)x_0\|_\infty = \mathcal{O}(h^r).$$

This completes the proof. □

Next we discuss the existence and convergence of the iterated approximate solutions $\tilde{x}_n$ to x_0.

Theorem 5. *$\widetilde{\mathcal{T}}_n'(x_0)$ is ν-convergent to $\mathcal{T}'(x_0)$ in uniform norm.*

Proof. We have

$$\widetilde{\mathcal{T}}_n'(x_0) = \mathcal{K}'(\mathcal{P}_n x_0)\mathcal{P}_n = \mathcal{K}'(\mathcal{P}_n x_0)\mathcal{P}_n - \mathcal{K}'(x_0)\mathcal{P}_n + \mathcal{K}'(x_0)\mathcal{P}_n. \tag{2.26}$$

This implies

$$\|\widetilde{\mathcal{T}}_n'(x_0)\|_\infty = \|\mathcal{K}'(\mathcal{P}_n x_0)\mathcal{P}_n\|_\infty \leq \|\mathcal{K}'(\mathcal{P}_n x_0)\mathcal{P}_n - \mathcal{K}'(x_0)\mathcal{P}_n\|_\infty + \|\mathcal{K}'(x_0)\|_\infty\|\mathcal{P}_n\|_\infty. \tag{2.27}$$

Hence using estimates (2.5), (2.23) and (2.27), we have

$$\|\widetilde{\mathcal{T}}_n'(x_0)\|_\infty \leq c_2 p\|(\mathcal{I}-\mathcal{P}_n)x_0\|_\infty + pM < \infty, \tag{2.28}$$

i.e., $\|\widetilde{\mathcal{T}}_n'(x_0)\|_\infty$ is uniformly bounded.

Next consider

$$\begin{aligned}\|[\widetilde{\mathcal{T}}_n'(x_0)-\mathcal{T}'(x_0)]\widetilde{\mathcal{T}}_n'(x_0)\|_\infty &= \|[\mathcal{K}'(\mathcal{P}_n x_0)\mathcal{P}_n-\mathcal{K}'(x_0)]\widetilde{\mathcal{T}}_n'(x_0)\|_\infty\\ &\le \|[\mathcal{K}'(\mathcal{P}_n x_0)-\mathcal{K}'(x_0)]\mathcal{P}_n\widetilde{\mathcal{T}}_n'(x_0)\|_\infty + \|[\mathcal{K}'(x_0)(\mathcal{I}-\mathcal{P}_n)\widetilde{\mathcal{T}}_n'(x_0)\|_\infty\\ &\le \|\mathcal{P}_n\|_\infty\|\widetilde{\mathcal{T}}_n'(x_0)\|_\infty\|\mathcal{K}'(\mathcal{P}_n x_0)-\mathcal{K}'(x_0)\|_\infty\\ &+ \|\mathcal{K}'(x_0)\|_\infty\|(\mathcal{I}-\mathcal{P}_n)\widetilde{\mathcal{T}}_n'(x_0)\|_\infty. \qquad (2.29)\end{aligned}$$

Hence using Lemma 2, Lemma 3, estimates (2.5), (2.9), (2.10) and the uniform boundedness of $\|\widetilde{\mathcal{T}}_n'(x_0)\|_\infty$, we obtain

$$\begin{aligned}\|[\widetilde{\mathcal{T}}_n'(x_0)-\mathcal{T}'(x_0)]\widetilde{\mathcal{T}}_n'(x_0)\|_\infty &\le pc_2\|\widetilde{\mathcal{T}}_n'(x_0)\|_\infty\|(\mathcal{I}-\mathcal{P}_n)x_0\|_\infty + M\|(\mathcal{I}-\mathcal{P}_n)\widetilde{\mathcal{T}}_n'(x_0)\|_\infty.\\ &\to 0, \quad as \quad n\to\infty.\end{aligned}$$

Following similar steps we can establish that

$$\|(\widetilde{\mathcal{T}}_n'(x_0)-\mathcal{T}'(x_0))\mathcal{T}'(x_0)\|_\infty \to 0, \quad as \quad n\to\infty.$$

This shows that $\widetilde{\mathcal{T}}_n'(x_0)$ is ν-convergent to $\mathcal{T}'(x_0)$ in uniform norm.

Hence the proof. □

We formulate the following result from Theorems 2 and 5.

Theorem 6. *Let $x_0 \in \mathbb{X}$ be an isolated solution of the Eq. (2.3). Assume that 1 is not an eigenvalue of the linear operator $\mathcal{K}'(x_0)$. Then for sufficiently large n, the operator $(\mathcal{I}-\widetilde{\mathcal{T}}_n'(x_0))$ is invertible on $\mathbb{X}$ and there exists a constant $L>0$ independent of n such that $\|(\mathcal{I}-\widetilde{\mathcal{T}}_n'(x_0))^{-1}\|_\infty \le L < \infty$.*

Theorem 7. *Let $x_0 \in \mathbb{X}$ be an isolated solution of the Eq. (2.3). Assume that 1 is not an eigenvalue of the linear operator $\mathcal{K}'(x_0)$. Let $\mathcal{P}_n : \mathbb{X} \to \mathbb{X}_n$ be the orthogonal projection operator defined by (2.8). Then the Eq. (2.16) has a unique solution $\tilde{x}_n \in B(x_0,\delta) = \{x : \|x-x_0\|_\infty < \delta\}$ for some $\delta > 0$ and for sufficiently large n. Moreover, there exists a constant $0<q<1$, independent of n such that*

$$\frac{\beta_n}{1+q} \le \|\tilde{x}_n - x_0\|_\infty \le \frac{\beta_n}{1-q},$$

where $\beta_n = \|(\mathcal{I}-\widetilde{\mathcal{T}}_n'(x_0))^{-1}(\widetilde{\mathcal{T}}_n(x_0)-\mathcal{T}(x_0))\|_\infty$.

Proof. By Theorem 6, we have $\|(\mathcal{I}-\widetilde{\mathcal{T}}_n'(x_0))^{-1}\|_\infty \le L < \infty$.
Now using estimate (2.9), Lemma 2, for any $x \in B(x_0,\delta)$, we have

$$\begin{aligned}\|\widetilde{\mathcal{T}}_n'(x)-\widetilde{\mathcal{T}}_n'(x_0)\|_\infty = \|\mathcal{K}'(\mathcal{P}_n x)\mathcal{P}_n-\mathcal{K}'(\mathcal{P}_n x_0)\mathcal{P}_n\|_\infty &= \|\mathcal{K}'(\mathcal{P}_n x)-\mathcal{K}'(\mathcal{P}_n x_0)\|_\infty\|\mathcal{P}_n\|_\infty\\ &\le c_2 p\|\mathcal{P}_n(x-x_0)\|_\infty\\ &\le p^2 c_2\|x-x_0\|_\infty\\ &\le p^2 c_2\delta. \qquad (2.30)\end{aligned}$$

Thus we have

$$\sup_{\|x-x_0\|\leq\delta} \|(\mathcal{I}-\widetilde{\mathcal{T}}_n'(x_0))^{-1}(\widetilde{\mathcal{T}}_n'(x)-\widetilde{\mathcal{T}}_n'(x_0))\|_\infty \leq Lp^2c_2\delta \leq q(say).$$

Here we choose δ such that $0<q<1$. This proves the Eq. (2.17) of Theorem 3. Now using the estimate (2.10) and Lemma 2, we have

$$\|\widetilde{\mathcal{T}}_n(x_0)-\mathcal{T}(x_0)\|_\infty \leq \|\mathcal{K}(\mathcal{P}_nx_0)-\mathcal{K}(x_0)\|_\infty \leq c_1\|(\mathcal{I}-\mathcal{P}_n)x_0\|_\infty \to 0, \quad as \quad n\to\infty. \tag{2.31}$$

Hence

$$\beta_n = \|(\mathcal{I}-\widetilde{\mathcal{T}}_n'(x_0))^{-1}(\widetilde{\mathcal{T}}_n(x_0)-\mathcal{T}(x_0))\|_\infty \leq L\|\widetilde{\mathcal{T}}_n(x_0)-\mathcal{T}(x_0)\|_\infty \to 0, \ as \ n\to\infty. \tag{2.32}$$

By choosing n large enough such that $\beta_n \leq \delta(1-q)$, the Eq. (2.19) of Theorem 3 is satisfied. Hence by applying Theorem 3, we obtain

$$\frac{\beta_n}{1+q} \leq \|\tilde{x}_n - x_0\|_\infty \leq \frac{\beta_n}{1-q},$$

where $\beta_n = \|(\mathcal{I}-\widetilde{\mathcal{T}}_n'(x_0))^{-1}(\widetilde{\mathcal{T}}_n(x_0)-\mathcal{T}(x_0))\|_\infty$.
This completes the proof. □

Next we discuss the convergence result for the iterated approximate solution $\tilde{x}_n$ of Galerkin method, defined by the Eq. (2.16).

Theorem 8. *Let $x_0 \in \mathcal{C}^r[0,1], r\geq 1$, be an isolated solution of the Eq. (2.3) and $\ell_u(.,.,.) \in \mathcal{C}^r([0,1]\times[0,1]\times\mathbb{R})$. Let $\tilde{x}_n$ be the iterated Galerkin approximation of x_0. Then the following holds*

$$\|\tilde{x}_n - x_0\|_\infty = \mathcal{O}(h^{2r}).$$

Proof. From Theorem 7, it follows that

$$\frac{\beta_n}{1+q} \leq \|\tilde{x}_n - x_0\|_\infty \leq \frac{\beta_n}{1-q},$$

where $\beta_n = \|(\mathcal{I}-\widetilde{\mathcal{T}}_n'(x_0))^{-1}(\widetilde{\mathcal{T}}_n(x_0)-\mathcal{T}(x_0))\|_\infty$.
Hence using Theorem 6, we get

$$\begin{aligned}\|\tilde{x}_n - x_0\|_\infty \leq \frac{\beta_n}{1-q} &\leq \frac{1}{1-q}\|(\mathcal{I}-\widetilde{\mathcal{T}}_n'(x_0))^{-1}(\widetilde{\mathcal{T}}_n(x_0)-\mathcal{T}(x_0))\|_\infty \\ &\leq c\|(\mathcal{I}-\widetilde{\mathcal{T}}_n'(x_0))^{-1}\|_\infty\|\widetilde{\mathcal{T}}_n(x_0)-\mathcal{T}(x_0)\|_\infty \\ &\leq cL\|\mathcal{K}(\mathcal{P}_nx_0)-\mathcal{K}(x_0)\|_\infty. \end{aligned} \tag{2.33}$$

Using mean-value theorem, we have

$$
\begin{aligned}
&|[\mathcal{K}(\mathcal{P}_n x_0) - \mathcal{K}(x_0)](t)| \\
&\quad = \left| \int_0^1 [\ell(t, s(t,\tau), \mathcal{P}_n x_0(s(t,\tau))) - \ell(t, s(t,\tau), x_0(s(t,\tau)))]\, d\tau \right| \\
&\quad = \left| \int_0^1 \ell_u(t, s(t,\tau), x_0(s(t,\tau)) + \theta_1(\mathcal{P}_n x_0 - x_0)(s(t,\tau)))(\mathcal{P}_n x_0 - x_0)(s(t,\tau))\, d\tau \right| \\
&\quad = \left| \int_0^1 g(t, s(t,\tau), x_0(s(t,\tau)), \mathcal{P}_n x_0(s(t,\tau)), \theta_1)(\mathcal{P}_n x_0 - x_0)(s(t,\tau))\, d\tau \right| \\
&\quad = \left| \int_0^1 [g(t, s(t,\tau), x_0(s(t,\tau)), \mathcal{P}_n x_0(s(t,\tau)), \theta_1) - g_t(s(t,\tau)) + g_t(s(t,\tau))](\mathcal{P}_n x_0 - x_0)(s(t,\tau))\, d\tau \right| \\
&\quad \le \left| \int_0^1 [g(t, s(t,\tau), x_0(s(t,\tau)), \mathcal{P}_n x_0(s(t,\tau)), \theta_1) - g_t(s(t,\tau))](\mathcal{P}_n x_0 - x_0)(s(t,\tau))\, d\tau \right| \\
&\quad + \left| \int_0^1 g_t(s(t,\tau))(\mathcal{P}_n x_0 - x_0)(s(t,\tau))\, d\tau \right|, \qquad (2.34)
\end{aligned}
$$

where $0 \le \theta_1 \le 1$ and $g(t, s(t,\tau), x_0(s(t,\tau)), x(s(t,\tau)), \theta_1) = \ell_u(t, s(t,\tau), x_0(s(t,\tau)) + \theta_1(x - x_0)(s(t,\tau)))$ and $g_t(s(t,\tau)) = \ell_u(t, s(t,\tau), x_0(s(t,\tau)))$.

For the first term of the above estimate (2.34), we have

$$
\begin{aligned}
&\left| \int_0^1 [g(t, s(t,\tau), x_0(s(t,\tau)), \mathcal{P}_n x_0(s(t,\tau)), \theta_1) - g_t(s(t,\tau))](\mathcal{P}_n x_0 - x_0)(s(t,\tau))\, d\tau \right| \\
&\quad = \left| \int_0^1 [\ell_u(t, s(t,\tau), x_0(s(t,\tau)) + \theta_1(\mathcal{P}_n x_0 - x_0)(s(t,\tau))) - \ell_u(t, s(t,\tau), x_0(s(t,\tau)))](\mathcal{P}_n x_0 - x_0)(s(t,\tau))\, d\tau \right| \\
&\quad \le c_2 \int_0^1 |(\mathcal{P}_n x_0 - x_0)(s(t,\tau))|^2\, d\tau \\
&\quad \le c_2 \|\mathcal{P}_n x_0 - x_0\|_\infty^2 = \mathcal{O}(h^{2r}). \qquad (2.35)
\end{aligned}
$$

Using the orthogonality of the projection operator $\mathcal{P}_n$ and estimate (2.11), for the second term of (2.34), we obtain

$$
\begin{aligned}
\left| \int_0^1 g_t(s(t,\tau))(\mathcal{P}_n x_0 - x_0)(s(t,\tau))\, d\tau \right| &= |\langle g_t(s(t,.)), (\mathcal{I} - \mathcal{P}_n) x_0(s(t,.))\rangle| \\
&= |\langle (\mathcal{I} - \mathcal{P}_n) g_t(s(t,.)), (\mathcal{I} - \mathcal{P}_n) x_0(s(t,.))\rangle| \\
&\le \|(\mathcal{I} - \mathcal{P}_n) g_t\|_\infty \|(\mathcal{I} - \mathcal{P}_n) x_0\|_\infty \\
&\le c h^{2r} \|g_t^{(r)}\|_\infty \|x_0^{(r)}\|_\infty. \qquad (2.36)
\end{aligned}
$$

Thus using estimates (2.34), (2.35), (2.36), we get

$$
\|\mathcal{K}(\mathcal{P}_n x_0) - \mathcal{K}(x_0)\|_\infty = \mathcal{O}(h^{2r}). \qquad (2.37)
$$

Hence from estimates (2.33) and (2.37), it follows that

$$
\|\tilde{x}_n - x_0\|_\infty = \mathcal{O}(h^{2r}). \qquad (2.38)
$$

This completes the proof. □

In the following section we discuss the multi-Galerkin method to improve the above convergence rates further.

3 Superconvergence Results by Multi-Gallerkin Method

In this section, we apply the multi-Galerkin method (M-Gallerkin method) (see [8,10,11]) for solving the Eq. (2.1) and obtain the superconvergence results. To do this, we define the multi-projection operator $\mathcal{K}_n^M$ (see [8,10,11]) by

$$\mathcal{K}_n^M(x) := \mathcal{P}_n\mathcal{K}(x) + \mathcal{K}(\mathcal{P}_n x) - \mathcal{P}_n\mathcal{K}(\mathcal{P}_n x). \tag{3.1}$$

The M-Galerkin method for Eq. (2.3) is seeking an approximate solution $x_n^M \in \mathbb{X}$ such that

$$x_n^M - \mathcal{K}_n^M(x_n^M) = f. \tag{3.2}$$

Let $\mathcal{T}_n^M(u) = \mathcal{K}_n^M(u) + f, u \in \mathbb{X}$, then the Eq. (3.2) can be written as

$$x_n^M = \mathcal{T}_n^M(x_n^M). \tag{3.3}$$

In order to obtain more accurate approximate solution, we define

$$\tilde{x}_n^M = \mathcal{K}(x_n^M) + f. \tag{3.4}$$

This is called the iterated M-Galerkin solution.
The Fréchet derivative of ${\mathcal{T}_n}^M(x)$ at x_0 is a linear operator and is given by

$$\begin{aligned}{\mathcal{T}_n}^{M'}(x_0) = {\mathcal{K}_n}^{M'}(x_0) &= \mathcal{P}_n\mathcal{K}'(x_0) + \mathcal{K}'(\mathcal{P}_n x_0)\mathcal{P}_n - \mathcal{P}_n\mathcal{K}'(\mathcal{P}_n x_0)\mathcal{P}_n \\ &= \mathcal{P}_n\mathcal{K}'(x_0) + (\mathcal{I} - \mathcal{P}_n)\mathcal{K}'(\mathcal{P}_n x_0)\mathcal{P}_n.\end{aligned}$$

Now we discuss the existence and convergence rates of the approximate solution x_n^M to x_0. To do this, we first prove the following lemma.

Lemma 4. *For any $x, y \in \mathbb{X}$, the following holds*

$$\|{\mathcal{K}_n^M}'(x) - {\mathcal{K}_n^M}'(y)\|_\infty \le (pc_2 + cc_2p^2)\|x - y\|_\infty,$$

where c is a constant independent of n.

Proof. Using the estimate (2.9) and Lemma 2, we have

$$\begin{aligned}\|{\mathcal{K}_n^M}'(x) - {\mathcal{K}_n^M}'(y)\|_\infty &= \|\mathcal{P}_n\mathcal{K}'(x) + (\mathcal{I} - \mathcal{P}_n)\mathcal{K}'(\mathcal{P}_n x)\mathcal{P}_n - \mathcal{P}_n\mathcal{K}'(y) - (\mathcal{I} - \mathcal{P}_n)\mathcal{K}'(\mathcal{P}_n y)\mathcal{P}_n\|_\infty \\ &\le \|\mathcal{P}_n\|_\infty\|\mathcal{K}'(x) - \mathcal{K}'(y)\|_\infty + (1 + \|\mathcal{P}_n\|_\infty)\|[\mathcal{K}'(\mathcal{P}_n x) - \mathcal{K}'(\mathcal{P}_n y)]\mathcal{P}_n\|_\infty \\ &\le pc_2\|x - y\|_\infty + cc_2\|\mathcal{P}_n(x - y)\|_\infty\|\mathcal{P}_n\|_\infty \\ &\le (pc_2 + cc_2p^2)\|x - y\|_\infty.\end{aligned}$$

This completes the proof. □

Theorem 9. *Let $x_0 \in \mathbb{X}$ be an isolated solution of the Eq. (2.3). Assume that 1 is not an eigenvalue of the linear operator $\mathcal{K}'(x_0)$. Then for sufficiently large n, the operator $(\mathcal{I} - {\mathcal{T}_n}^{M'}(x_0))$ is invertible on $\mathbb{X}$ and there exists a constant $L_1 > 0$ independent of n such that $\|(\mathcal{I} - {\mathcal{T}_n}^{M'}(x_0))^{-1}\|_\infty \le L_1 < \infty$.*

Proof. Consider

$$\begin{aligned}\|{\mathcal{T}_n}^{M'}(x_0) - \mathcal{T}'(x_0)\|_\infty &= \|\mathcal{P}_n\mathcal{K}'(x_0) - (\mathcal{I}-\mathcal{P}_n)\mathcal{K}'(\mathcal{P}_n x_0)\mathcal{P}_n - \mathcal{K}'(x_0)\|_\infty \\ &\leq \|(\mathcal{I}-\mathcal{P}_n)\mathcal{K}'(\mathcal{P}_n x_0)\mathcal{P}_n\|_\infty + \|(\mathcal{I}-\mathcal{P}_n)\mathcal{K}'(x_0)\|_\infty. \quad (3.5)\end{aligned}$$

Note that $\widetilde{\mathcal{T}}_n'(x_0) = \mathcal{K}'(\mathcal{P}_n x_0)\mathcal{P}_n$, hence from Lemma 3, we have

$$\|(\mathcal{I}-\mathcal{P}_n)\mathcal{K}'(\mathcal{P}_n x_0)\mathcal{P}_n\|_\infty = \|(\mathcal{I}-\mathcal{P}_n)\widetilde{\mathcal{T}}_n'(x_0)\|_\infty \to 0, \quad as \quad n \to \infty,$$

and

$$\|(\mathcal{I}-\mathcal{P}_n)\mathcal{K}'(x_0)\|_\infty = \|(\mathcal{I}-\mathcal{P}_n)\mathcal{T}'(x_0)\|_\infty \to 0, \quad as \quad n \to \infty.$$

This implies

$$\|{\mathcal{T}_n}^{M'}(x_0) - \mathcal{T}'(x_0)\|_\infty \to 0, \quad as \quad n \to \infty. \quad (3.6)$$

We assume that 1 is not an eigenvalue of $\mathcal{T}'(x_0)$, i.e., $(\mathcal{I}-\mathcal{T}'(x_0))$is invertible on $\mathbb{X}$. Then by applying Theorem 2, we have $(\mathcal{I}-{\mathcal{T}_n}^{M'}(x_0))^{-1}$ exists and is uniformly bounded on $\mathbb{X}$, for some sufficiently large n, i.e., there exists a constant $L_1 > 0$ such that $\|(\mathcal{I}-{\mathcal{T}_n}^{M'}(x_0))^{-1}\|_\infty \leq L_1 < \infty$. This completes the proof. □

Theorem 10. *Let $x_0 \in \mathcal{C}^r[0,1], r \geq 1$, be an isolated solution of the Eq. (2.3) and $\ell_u(.,.,.) \in \mathcal{C}^r([0,1]\times[0,1]\times\mathbb{R})$. Assume that 1 is not an eigenvalue of the linear operator $\mathcal{K}'(x_0)$. Let $\mathcal{P}_n : \mathbb{X} \to \mathbb{X}_n$ be the orthogonal projection operator defined by (2.8). Then the Eq. (3.2) has a unique solution $x_n^M \in B(x_0,\delta) = \{x : \|x - x_0\|_\infty < \delta\}$ for some $\delta > 0$ and for sufficiently large n. Moreover, there exists a constant $0 < q < 1$, independent of n such that*

$$\frac{\alpha_n}{1+q} \leq \|x_n^M - x_0\|_\infty \leq \frac{\alpha_n}{1-q},$$

where $\alpha_n = \|(\mathcal{I}-{\mathcal{T}_n}^{M'}(x_0))^{-1}({\mathcal{T}_n}^{M}(x_0) - \mathcal{T}(x_0))\|_\infty$. Further, we obtain

$$\|x_n^M - x_0\|_\infty = \mathcal{O}(h^{2r}).$$

Proof. Using Theorem 9, it follows that there exists some $L_1 > 0$ such that $\|(\mathcal{I}-{\mathcal{T}_n}^{M'}(x_0))^{-1}\|_\infty \leq L_1 < \infty$.
Now using Lemma 4, we have for any $x \in B(x_0,\delta)$,

$$\begin{aligned}\|{\mathcal{T}_n}^{M'}(x_0) - {\mathcal{T}_n}^{M'}(x)\|_\infty &= \|\mathcal{K}_n^{M'}(x_0) - \mathcal{K}_n^{M'}(x)\|_\infty \\ &\leq (pc_2 + cc_2p^2)\|x_0 - x\|_\infty \\ &\leq (pc_2 + cc_2p^2)\delta.\end{aligned}$$

Thus we have

$$\sup_{\|x-x_0\|\leq\delta} \|(\mathcal{I}-{\mathcal{T}_n}^{M'}(x_0))^{-1}({\mathcal{T}_n}^{M'}(x_0)-{\mathcal{T}_n}^{M'}(x))\|_\infty \leq L_1(pc_2+cc_2p^2)\delta \leq q(say).$$

Here we choose δ such that $0<q<1$. This proves the Eq. (2.17) of Theorem 3. Hence applying Lemma 2, and estimate (2.10), we have

$$\begin{aligned}
\alpha_n &= \|(\mathcal{I}-{\mathcal{T}_n}^{M'}(x_0))^{-1}({\mathcal{T}_n}^{M}(x_0)-\mathcal{T}(x_0))\|_\infty \\
&\leq L_1\|{\mathcal{T}_n}^{M}(x_0)-\mathcal{T}(x_0))\|_\infty \\
&\leq L_1\|(\mathcal{I}-\mathcal{P}_n)(\mathcal{K}(\mathcal{P}_nx_0)-\mathcal{K}(x_0))\|_\infty \\
&\leq L_1(1+\|\mathcal{P}_n\|_\infty)\|\mathcal{K}(\mathcal{P}_nx_0)-\mathcal{K}(x_0)\|_\infty \\
&\leq cc_1L_1\|(\mathcal{P}_n-\mathcal{I})x_0\|_\infty \\
&\to 0 \quad as \quad n \quad \to\infty. \qquad (3.7)
\end{aligned}$$

By choosing n large enough such that $\alpha_n \leq \delta(1-q)$, the Eq. (2.19) of Theorem 3 is satisfied, i.e.,

$$\frac{\alpha_n}{1+q} \leq \|x_n^M - x_0\|_\infty \leq \frac{\alpha_n}{1-q}.$$

Hence from estimate (2.37), it follows that

$$\|x_n^M-x_0\|_\infty \leq \frac{\alpha_n}{1-q} \leq \frac{1}{1-q}L_1\|{\mathcal{T}_n}^{M}(x_0)-\mathcal{T}(x_0))\|_\infty \leq cL_1\|\mathcal{K}(\mathcal{P}_nx_0)-\mathcal{K}(x_0)\|_\infty = \mathcal{O}(h^{2r}).$$

This completes the proof. □

Remark 1. Note that from Theorems 8 and 10, it follows that both the iterated Galerkin solution $\tilde{x}_n$ and M-Galerkin solution x_n^M have the same order of convergence $\mathcal{O}(h^{2r})$. However using Theorem 10, below we prove that the iterated M-Galerkin solution improves over both iterated Galerkin and M-Galerkin solutions.

Next we discuss the superconvergence results for the iterated approximate solution $\tilde{x}_n^M$ of M-Galerkin method, defined by Eq. (3.4). To do this, we first proof the following lemma.

Lemma 5. *Let $x_0 \in \mathbb{X}$ be an isolated solution of the Eq. (2.1) and let $\mathcal{P}_n : \mathbb{X} \to \mathbb{X}_n$ be the orthogonal projection operator defined by (2.8). Let $\tilde{x}_n^M$ be the iterated approximation of x_0. Then there holds*

$$\|\tilde{x}_n^M - x_0\|_\infty \leq C_1\|x_n^M-x_0\|_\infty^2 + (1+M_1p)\|\mathcal{K}'(x_0)(\mathcal{I}-\mathcal{P}_n)(\mathcal{K}(\mathcal{P}_nx_0)-\mathcal{K}(x_0))\|_\infty,$$

where $C_1 = (c_2 + M_1M_2)$.

Proof. Applying Theorem 9, we have

$$\|(\mathcal{I} - \mathcal{K}_n^{M'}(x_0))^{-1}\|_\infty = \|(\mathcal{I} - \mathcal{T}_n^{M'}(x_0))^{-1}\|_\infty \leq L_1 < \infty. \tag{3.8}$$

It follows from the estimate (2.5) that

$$\|\mathcal{K}'(x_0)\|_\infty \leq M.$$

Hence we obtain

$$\|\mathcal{K}'(x_0)(\mathcal{I} - \mathcal{K}_n^{M'}(x_0))^{-1}\|_\infty \leq ML_1 \leq M_1 < \infty. \tag{3.9}$$

Now from the Eqs. (2.3) and (3.4), Lemma 2 and using the mean-value theorem, we have

$$\begin{aligned}
\|\tilde{x}_n^M - x_0\|_\infty &= \|\mathcal{K}(x_n^M) - \mathcal{K}(x_0)\|_\infty \\
&= \|\mathcal{K}'(x_0 + \theta_2(x_n^M - x_0))(x_n^M - x_0)\|_\infty \\
&= \|\{[\mathcal{K}'(x_0 + \theta_2(x_n^M - x_0)) - \mathcal{K}'(x_0)] + \mathcal{K}'(x_0)\}(x_n^M - x_0)\|_\infty \\
&\leq \|[\mathcal{K}'(x_0 + \theta_2(x_n^M - x_0)) - \mathcal{K}'(x_0)](x_n^M - x_0)\|_\infty + \|\mathcal{K}'(x_0)(x_n^M - x_0)\|_\infty \\
&\leq c_2\|x_n^M - x_0\|_\infty^2 + \|\mathcal{K}'(x_0)(x_n^M - x_0)\|_\infty,
\end{aligned} \tag{3.10}$$

where $0 < \theta_2 < 1$.
For the second term of the estimate (3.10), we consider

$$\begin{aligned}
x_n^M - x_0 &= \mathcal{K}_n^M(x_n^M) - \mathcal{K}(x_0) \\
&= \mathcal{K}_n^M(x_n^M) - \mathcal{K}_n^M(x_0) - \mathcal{K}_n^{M'}(x_0)(x_n^M - x_0) + \mathcal{K}_n^{M'}(x_0)(x_n^M - x_0) + \mathcal{K}_n^M(x_0) - \mathcal{K}(x_0).
\end{aligned}$$

This implies

$$(\mathcal{I} - \mathcal{K}_n^{M'}(x_0))(x_n^M - x_0) = \mathcal{K}_n^M(x_n^M) - \mathcal{K}_n^M(x_0) - \mathcal{K}_n^{M'}(x_0)(x_n^M - x_0) + \mathcal{K}_n^M(x_0) - \mathcal{K}(x_0).$$

Using mean-value theorem, we have

$$\begin{aligned}
&(x_n^M - x_0) \\
&\quad = (\mathcal{I} - \mathcal{K}_n^{M'}(x_0))^{-1}\left[\mathcal{K}_n^M(x_n^M) - \mathcal{K}_n^M(x_0) - \mathcal{K}_n^{M'}(x_0)(x_n^M - x_0) + (\mathcal{K}_n^M)(x_0) - \mathcal{K}(x_0)\right] \\
&\quad = (\mathcal{I} - \mathcal{K}_n^{M'}(x_0))^{-1}\left[\mathcal{K}_n^{M'}(x_0 + \theta_3(x_n^M - x_0)) - \mathcal{K}_n^{M'}(x_0)\right](x_n^M - x_0) \\
&\quad + (\mathcal{I} - \mathcal{K}_n^{M'}(x_0))^{-1}\left[\mathcal{K}_n^M(x_0) - \mathcal{K}(x_0)\right],
\end{aligned} \tag{3.11}$$

where $0 < \theta_3 < 1$.

Applying $\mathcal{K}'(x_0)$ on both side and using the estimate (3.9), we obtain

$$\begin{aligned}
&\|\mathcal{K}'(x_0)(x_n^M - x_0)\|_\infty \\
&\quad \le \|\mathcal{K}'(x_0)(\mathcal{I} - \mathcal{K}_n^{M\prime}(x_0))^{-1}\|_\infty \|[\mathcal{K}_n^{M\prime}(x_0 + \theta_3(x_n^M - x_0)) - \mathcal{K}_n^{M\prime}(x_0)](x_n^M - x_0)\|_\infty \\
&\quad + \|\mathcal{K}'(x_0)(\mathcal{I} - \mathcal{K}_n^{M\prime}(x_0))^{-1}[\mathcal{K}_n^M(x_0) - \mathcal{K}(x_0)]\|_\infty \\
&\quad \le M_1 \|\mathcal{K}_n^{M\prime}(x_0 + \theta_3(x_n^M - x_0)) - \mathcal{K}_n^{M\prime}(x_0)\|_\infty \|x_n^M - x_0\|_\infty \\
&\quad + \|\mathcal{K}'(x_0)(\mathcal{I} - \mathcal{K}_n^{M\prime}(x_0))^{-1}[\mathcal{K}_n^M(x_0) - \mathcal{K}(x_0)]\|_\infty. \qquad (3.12)
\end{aligned}$$

Using the identity $(\mathcal{I} - \mathcal{K}_n^{M\prime}(x_0))^{-1} = \mathcal{I} + (\mathcal{I} - \mathcal{K}_n^{M\prime}(x_0))^{-1}\mathcal{K}_n^{M\prime}(x_0)$, and the estimate (3.9), the second term of the estimate (3.12) becomes

$$\begin{aligned}
&\|\mathcal{K}'(x_0)(\mathcal{I} - \mathcal{K}_n^{M\prime}(x_0))^{-1}[\mathcal{K}_n^M(x_0) - \mathcal{K}(x_0)]\|_\infty \\
&\quad = \|\mathcal{K}'(x_0)\{\mathcal{I} + (\mathcal{I} - \mathcal{K}_n^{M\prime}(x_0))^{-1}\mathcal{K}_n^{M\prime}(x_0)\}[\mathcal{K}_n^M(x_0) - \mathcal{K}(x_0)]\|_\infty \\
&\quad \le \|\mathcal{K}'(x_0)[\mathcal{K}_n^M(x_0) - \mathcal{K}(x_0)]\|_\infty \\
&\quad + \|\mathcal{K}'(x_0)(\mathcal{I} - \mathcal{K}_n^{M\prime}(x_0))^{-1}\mathcal{K}_n^{M\prime}(x_0)[\mathcal{K}_n^M(x_0) - \mathcal{K}(x_0)]\|_\infty \\
&\quad \le \|\mathcal{K}'(x_0)[\mathcal{K}_n^M(x_0) - \mathcal{K}(x_0)]\|_\infty + M_1 \|\mathcal{K}_n^{M\prime}(x_0)[\mathcal{K}_n^M(x_0) - \mathcal{K}(x_0)]\|_\infty. \qquad (3.13)
\end{aligned}$$

From the estimates (3.12) and (3.13), we get

$$\begin{aligned}
\|\mathcal{K}'(x_0)(x_n^M - x_0)\|_\infty \le\ & M_1 \|\mathcal{K}_n^{M\prime}(x_0 + \theta_3(x_n^M - x_0)) - \mathcal{K}_n^{M\prime}(x_0)\|_\infty \|x_n^M - x_0\|_\infty \\
& + \|\mathcal{K}'(x_0)[\mathcal{K}_n^M(x_0) - \mathcal{K}(x_0)]\|_\infty + M_1 \|\mathcal{K}_n^{M\prime}(x_0)[\mathcal{K}_n^M(x_0) - \mathcal{K}(x_0)]\|_\infty. \qquad (3.14)
\end{aligned}$$

Note that

$$\mathcal{K}_n^M(x_0) - \mathcal{K}(x_0) = (\mathcal{I} - \mathcal{P}_n)(\mathcal{K}(\mathcal{P}_n x_0) - \mathcal{K}(x_0)),$$

and

$$\mathcal{K}_n^{M\prime}(x_0)[\mathcal{K}_n^M(x_0) - \mathcal{K}(x_0)] = \mathcal{P}_n \mathcal{K}'(x_0)(\mathcal{I} - \mathcal{P}_n)(\mathcal{K}(\mathcal{P}_n x_0) - \mathcal{K}(x_0)).$$

Combining this with estimate (3.14), Lemma 4, we get

$$\begin{aligned}
&\|\mathcal{K}'(x_0)(x_n^M - x_0)\|_\infty \\
&\quad \le M_1(pc_2 + cc_2p^2)\|x_n^M - x_0\|_\infty^2 + \|\mathcal{K}'(x_0)(\mathcal{I} - \mathcal{P}_n)(\mathcal{K}(\mathcal{P}_n x_0) - \mathcal{K}(x_0))\|_\infty \\
&\quad + M_1 \|\mathcal{P}_n \mathcal{K}'(x_0)(\mathcal{I} - \mathcal{P}_n)(\mathcal{K}(\mathcal{P}_n x_0) - \mathcal{K}(x_0))\|_\infty \\
&\quad \le M_1 M_2 \|x_n^M - x_0\|_\infty^2 + \|\mathcal{K}'(x_0)(\mathcal{I} - \mathcal{P}_n)(\mathcal{K}(\mathcal{P}_n x_0) - \mathcal{K}(x_0))\|_\infty \\
&\quad + M_1 \|\mathcal{P}_n\|_\infty \|\mathcal{K}'(x_0)(\mathcal{I} - \mathcal{P}_n)(\mathcal{K}(\mathcal{P}_n x_0) - \mathcal{K}(x_0))\|_\infty, \qquad (3.15)
\end{aligned}$$

where $M_2 = (pc_2 + cc_2p^2) < \infty$.
Now combining this with the estimate (3.10), we obtain

$$\|\tilde{x}_n^M - x_0\|_\infty \le C_1 \|x_n^M - x_0\|_\infty^2 + (1 + M_1 p)\|\mathcal{K}'(x_0)(\mathcal{I} - \mathcal{P}_n)(\mathcal{K}(\mathcal{P}_n x_0) - \mathcal{K}(x_0))\|_\infty,$$

where $C_1 = (c_2 + M_1 M_2)$. This completes the proof. □

Theorem 11. *Let $\mathcal{P}_n : \mathbb{X} \to \mathbb{X}_n$ be the orthogonal projection operator defined by (2.8) and $x_0 \in \mathcal{C}^r[0,1], r \geq 1$, be an isolated solution of the Eq. (2.1). Assume $\ell_u(.,.,.) \in \mathcal{C}^r([0,1]\times[0,1]\times\mathbb{R}))$. Let $\tilde{x}_n^M$ be the iterated M-Galerkin approximation of x_0. Then the following holds*

$$\|\tilde{x}_n^M - x_0\|_\infty = \mathcal{O}(h^{3r}).$$

Proof. From Lemma 5, we obtain

$$\|\tilde{x}_n^M - x_0\|_\infty \leq C_1\|x_n^M - x_0\|_\infty^2 + (1 + M_1 p)\|\mathcal{K}'(x_0)(\mathcal{I} - \mathcal{P}_n)(\mathcal{K}(\mathcal{P}_n x_0) - \mathcal{K}(x_0))\|_\infty \quad (3.16)$$

Consider

$$\|\mathcal{K}'(x_0)(\mathcal{I} - \mathcal{P}_n)(\mathcal{K}(\mathcal{P}_n x_0) - \mathcal{K}(x_0))\|_\infty = \|\mathcal{K}'(x_0)(\mathcal{I} - \mathcal{P}_n)\|_\infty \|\mathcal{K}(\mathcal{P}_n x_0) - \mathcal{K}(x_0)\|_\infty. \quad (3.17)$$

Using orthogonality of $\mathcal{P}_n$ and the estimate (2.11), we have

$$\begin{aligned}
|\mathcal{K}'(x_0)(\mathcal{I} - \mathcal{P}_n)y(t)| &= \left| \int_0^1 \ell_u(t, s(t,\tau)), x_0(s(t,\tau)))(\mathcal{I} - \mathcal{P}_n)y(s(t,\tau))\, d\tau \right| \\
&\leq |< (\mathcal{I} - \mathcal{P}_n)\ell_u(t,.,x_0(.)), (\mathcal{I} - \mathcal{P}_n)y(.) >| \\
&\leq \|(\mathcal{I} - \mathcal{P}_n)\ell_u\|_\infty \|(\mathcal{I} - \mathcal{P}_n)y\|_\infty \\
&\leq ch^r \|\ell_u^{(r)}\|_\infty \|y\|_\infty.
\end{aligned}$$

This implies

$$\|\mathcal{K}'(x_0)(\mathcal{I} - \mathcal{P}_n)\|_\infty \leq ch^r \|\ell_u^{(r)}\|_\infty. \quad (3.18)$$

Again using estimate (2.37), we have

$$\|\mathcal{K}(\mathcal{P}_n x_0) - \mathcal{K}(x_0)\|_\infty = \mathcal{O}(h^{2r}). \quad (3.19)$$

Combining the estimates (3.16), (3.17), (3.18), (3.19) and Theorem 10, we have

$$\|\tilde{x}_n^M - x_0\|_\infty = \mathcal{O}(h^{min\{4r,3r\}}) = \mathcal{O}(h^{3r}). \quad (3.20)$$

This completes the proof. □

Remark 2. From Theorems 4, 8 and 11, we see that the order of convergence of the iterated M-Galerkin method improves over the Galerkin method and iterated-Galerkin methods.

4 Numerical Results

In this section, we present the numerical results. For that we take the piecewise polynomials as the basis functions for the subspace $\mathbb{X}_n$. We present the errors of the approximate and iterated approximate solutions of Galerkin and M-Galerkin methods in uniform norm. We denote the Galerkin, iterated Galerkin, multi-Galerkin and iterated multi-Galerkin solutions by x_n, $\tilde{x}_n$, x_n^M and $\tilde{x}_n^M$, respectively. Also we denote $\|x - x_n\|_\infty = \mathcal{O}(h^\alpha)$, $\|x - \tilde{x}_n\|_\infty = \mathcal{O}(h^a)$, $\|x - x_n^M\|_\infty = \mathcal{O}(h^\gamma)$, $\|x - \tilde{x}_n^M\|_\infty = \mathcal{O}(h^c)$. The numerical algorithm was computed on a PC with Intel Pentium 3.20GHz CPU, 4.00GB RAM by using Matlab.

Consider the uniform partition of $[0, 1]$:

$$0 = t_0 < t_1 < t_2 < ... < t_n = 1,$$

where $t_i = \frac{i}{n}$, $i = 0, 1, 2, ..., n$.
We choose the approximating subspaces as the space of piecewise constant functions ($r = 1$), which has dimension n. Then for $r = 1$, the expected orders of convergence are $\alpha = 1$, $\gamma = 2$, $a = 2$ and $c = 3$. In Tables 1 and 3, we present the errors in Galerkin and iterated Galerkin method. The errors and convergence rate in multi-Galerkin and iterated multi-Galerkin methods are given in Tables 2 and 4.

Example 1. Consider the following Volterra integral equation of second kind

$$x(t) = f(t) + \int_0^t k(t, s, x(s))\, ds, \quad t \in [0, 1],$$

with the kernel function $k(t, s, x(s)) = (t + s)[x(s)]^3$ and the function $f(t) = -(15/56)t^8 + (13/14)t^7 - (11/10)t^6 + (9/20)t^5 + t^2 - t$ and the exact solution is given by $x(t) = t^2 - t$.

Table 1. Galerkin and iterated Galerkin methods

n	$\|x - x_n\|_\infty$	α	$\|x - \tilde{x}_n\|_\infty$	a
2	$1.661527988904 \times 10^{-1}$	-	$4.645082264 \times 10^{-3}$	
4	$1.037479005790 \times 10^{-1}$	0.67	$1.014390206 \times 10^{-3}$	2.19
8	$5.687881933033 \times 10^{-2}$	0.86	$2.736000385 \times 10^{-4}$	1.89
16	$2.953533302197 \times 10^{-2}$	0.94	$6.843872691 \times 10^{-5}$	1.99
32	$1.488690530289 \times 10^{-2}$	0.98	$1.979056670 \times 10^{-5}$	1.79
64	$7.318546260251 \times 10^{-3}$	1.02	$5.566386247 \times 10^{-6}$	1.83
128	$3.473331427325 \times 10^{-3}$	1.07	$1.522815353 \times 10^{-6}$	1.87

Table 2. M-Galerkin and iterated M-Galerkin methods

n	$\|x - x_n^M\|_\infty$	γ	$\|x - \tilde{x}_n^M\|_\infty$	c
2	$2.139257640477\times10^{-3}$	-	$3.195587502679\times10^{-4}$	-
4	$5.094845289810\times10^{-4}$	2.07	$3.027253073465\times10^{-5}$	3.48
8	$1.172053429381\times10^{-4}$	2.12	$4.112276815779\times10^{-6}$	2.88
16	$2.971036713855\times10^{-5}$	1.98	$5.104839056138\times10^{-7}$	3.01
32	$7.376285660495\times10^{-6}$	2.01	$6.078829573469\times10^{-8}$	3.07
64	$1.790360597207\times10^{-6}$	2.06	$7.138998341086\times10^{-9}$	3.09

Example 2. Consider the following Volterra integral equation of second kind

$$x(t) = f(t) + \int_0^t k(t, s, x(s))\, ds, \quad t \in [0, 1],$$

with the kernel function $k(t, s, x(s)) = -s[x(s)]^3$, $f(t) = t + \frac{t^5}{5}$; and its exact solution is given by $x(t) = t$.

Table 3. Galerkin and iterated Galerkin methods

n	$\|x - x_n\|_\infty$	α	$\|x - \tilde{x}_n\|_\infty$	a
2	$2.547849617844\times10^{-1}$	-	$4.28181707196\times10^{-2}$	-
4	$1.2711877422622\times10^{-1}$	1.00	$1.15157330589\times10^{-2}$	1.89
8	$6.286526902122\times10^{-2}$	1.01	$2.76579432147\times10^{-3}$	2.05
16	$3.105266662831\times10^{-2}$	1.01	$8.28309267856\times10^{-4}$	1.73
32	$1.529607511876\times10^{-2}$	1.02	$2.42217386433\times10^{-4}$	1.77
64	$7.469603314291\times10^{-3}$	1.03	$6.62642400773\times10^{-4}$	1.87
128	$3.591926730840\times10^{-3}$	1.05	$1.60017591998\times10^{-5}$	2.05

From Tables 1 and 3, we see that the iterated approximate solutions give better convergence rates than the approximate solutions in Galerkin method.

Table 4. M-Galerkin and iterated M-Galerkin methods

n	$\|x - x_n^M\|_\infty$	γ	$\|x - \tilde{x}_n^M\|_\infty$	c
2	3.083864718×10^{-2}	-	1.287065314×10^{-2}	-
4	6.007095391×10^{-3}	2.36	2.507089260×10^{-3}	3.17
8	1.491400338×10^{-3}	2.01	3.177608653×10^{-4}	2.98
16	3.806844574×10^{-4}	1.97	4.468744470×10^{-5}	2.83
32	9.256865522×10^{-5}	2.04	5.863644532×10^{-6}	2.93
64	2.235385674×10^{-5}	2.05	7.483565210×10^{-7}	2.97

From Tables 2 and 4, we also see that the iterated multi-Galerkin method gives better convergence rates than Galerkin method and iterated Galerkin method, while the size of the system of nonlinear equations that must be solved, remains the same as in Galerkin method.

References

1. Ahues, M., Largillier, A., Limaye, B.: Spectral Computations for Bounded Operators. CRC Press, Boca Raton (2001)
2. Blom, J., Brunner, H.: The numerical solution of nonlinear Volterra integral equations of the second kind by collocation and iterated collocation methods. SIAM J. Sci. Stat. Comput. **8**(5), 806–830 (1987)
3. Brunner, H.: Iterated collocation methods and their discretizations for Volterra integral equations. SIAM J. Numer. Anal. **21**(6), 1132–1145 (1984)
4. Brunner, H., Houwen, P.: The Numerical Solution of Volterra Equation. North-Holland Publishing Co., Amsterdam (1986)
5. Brunner, H., Nørsett, S.P.: Superconvergence of collocation methods for Volterra and Abel integral equations of the second kind. Numer. Math. **36**(4), 347–358 (1981)
6. Chatelin, F.: Spectral Approximation of Linear Operators ci. SIAM (1983)
7. Chen, L., Duan, J.: Multistage numerical picard iteration methods for nonlinear Volterra integral equations of the second kind. Adv. Pure Math. **5**(11), 672 (2015)
8. Chen, Z., Long, G., Nelakanti, G.: The discrete multi-projection method for Fredholm integral equations of the second kind. J. Integr. Equ. Appl. **19**(2), 143–162 (2007)
9. Day, J.T.: A starting method for solving nonlinear Volterra integral equations. Math. Comput. **21**(98), 179–188 (1967)
10. Grammont, L., Kulkarni, R.: A superconvergent projection method for nonlinear compact operator equations. C.R. Math. **342**(3), 215–218 (2006)
11. Grammont, L., Kulkarni, R.P., Vasconcelos, P.B.: Modified projection and the iterated modified projection methods for nonlinear integral equations. J. Integr. Equ. Appl. **25**(4), 481–516 (2013)
12. Kulkarni, R.P.: A superconvergence result for solutions of compact operator equations. Bull. Aust. Math. Soc. **68**(3), 517–528 (2003)
13. Kumar, S.: Superconvergence of a collocation-type method for Hammerstein equations. IMA J. Num. Anal. **7**(3), 313–325 (1987)
14. Long, G., Sahani, M.M., Nelakanti, G.: Polynomially based multi-projection methods for Fredholm integral equations of the second kind. Appl. Math. Comput. **215**(1), 147–155 (2009)
15. Maleknejad, K., Sohrabi, S., Rostami, Y.: Numerical solution of nonlinear Volterra integral equations of the second kind by using chebyshev polynomials. Appl. Math. Comput. **188**(1), 123–128 (2007)
16. Mohsen, A., El-Gamel, M.: On the numerical solution of linear and nonlinear Volterra integral and integro-differential equations. Appl. Math. Comput. **217**(7), 3330–3337 (2010)
17. Rudin, W.: Real and Complex Analysis. Tata McGraw-Hill Education, London (1987)
18. Tang, T., Xu, X., Cheng, J.: On spectral methods for Volterra integral equations and the convergence analysis. J. Comput. Math. **26**(6), 825–837 (2008)

19. Vainikko, G.M.: Galerkin's perturbation method and the general theory of approximate methods for non-linear equations. USSR Comput. Math. Math. Phys. **7**(4), 1–41 (1967)
20. Wan, Z., Chen, Y., Huang, Y.: Legendre spectral Galerkin method for second-kind Volterra integral equations. Front. Math. China **4**(1), 181–193 (2009)
21. Xie, Z., Li, X., Tang, T.: Convergence analysis of spectral Galerkin methods for Volterra type integral equations. J. Sci. Comput. **53**(2), 414–434 (2012)
22. Zhang, S., Lin, Y., Rao, M.: Numerical solutions for second-kind Volterra integral equations by Galerkin methods. Appl. Math. Comput. **45**(1), 19–39 (2000)

Derivations on Lie Ideals of Prime Γ-Rings

Kalyan Kumar Dey[1(✉)], Akhil Chandra Paul[1], and Bijan Davvaz[2]

[1] Department of Mathematics, Rajshahi University, Rajshahi 6205, Bangladesh
kkdmath@yahoo.com, acpaulrubd_math@yahoo.com
[2] Department of Mathematics, Yazd University, Yazd, Iran
davvaz@yazd.ac.ir, bdavvaz@yahoo.com

Abstract. Let M be a 2-torsion free prime Γ-ring with $Z(M)$ as the center of M. In this paper, we prove the following: (i) If U is a Lie ideal of M and if $d \neq 0$ is a derivation of M such that $d^2(U) = 0$, then $U \subseteq Z(M)$; (ii) if $U \not\subset Z(M)$ is a Lie ideal of M and $d \neq 0$ is a derivation of M, then $Z(d(U)) \subseteq Z(M)$; (iii) If $U \not\subset Z(M)$ is a Lie ideal of M and if d is a derivation of M such that $d^3 \neq 0$, then $d(U)^*$, the subring generated by $d(U)$ contains a non-zero ideal of M. Finally, we prove that if $U \not\subset Z(M)$ is a Lie ideal of M and d_1 and d_2 are derivations of M such that $d_1 d_2(U) = 0$, then $d_1 = 0$ or $d_2 = 0$.

Keywords: Γ-ring · Prime Γ-ring · Derivation · Γ-Lie ideal

1 Introduction and Preliminaries

The notion of Γ-ring was first introduced by Nobusawa [14] as a generalization of a ring theory and afterwards, it was generalized by Barnes [2] in a broad sense. In this article, we consider M as a Γ-ring in the sense of Barnes [2] and we shall denote $Z(M)$ to be the center of M. If A is a subset of M, then we define $Z(A) = \{x \in M \ \ [x, a]_\alpha = 0, \text{ for all } a \in A, \alpha \in \Gamma\}$ which is known as the center of A with respect to M, where $[a, b]_\alpha = a\alpha b - b\alpha a$ and this is known as the commutator of a and b with respect to $\alpha \in \Gamma$. An ideal P of a Γ-ring M is said to be prime if for any ideals A and B of M, $A\Gamma B \subseteq P$ implies $A \subseteq P$ or $B \subseteq P$. A Γ-ring M is said to be prime if the zero ideal is prime.

Theorem 1 *[15]. If M is a Γ-ring, the following conditions are equivalent:*

(1) M is a prime Γ-ring.
(2) If $a, b \in M$ and $a\Gamma M\Gamma b = \langle 0\rangle$, then $a = 0$ or $b = 0$.
(3) If $\langle a\rangle$ and $\langle b\rangle$ are principal ideals of M such that $\langle a\rangle\Gamma\langle b\rangle = \langle 0\rangle$, then $a = 0$ or $b = 0$.
(4) If A and B are right ideals of M such that $A\Gamma B = \langle 0\rangle$, then $A = \langle 0\rangle$ or $B = \langle 0\rangle$.
(5) If A and B are left ideals of M such that $A\Gamma B = \langle 0\rangle$, then $A = \langle 0\rangle$ or $B = \langle 0\rangle$.

© Springer Nature Singapore Pte Ltd. 2017
D. Giri et al. (Eds.): ICMC 2017, CCIS 655, pp. 380–390, 2017.
DOI: 10.1007/978-981-10-4642-1_32

An additive subgroup U of M is said to be a Lie ideal of M if $[u, m]_\alpha \in U$, for all $u \in U$, $m \in M$ and $\alpha \in \Gamma$. It is clear that every ideal of a Γ-ring M is a Lie ideal of M but the converse is in general not true. Let M be a Γ-ring. An additive mapping $d : M \to M$ is called a derivation if $d(x\alpha y) = d(x)\alpha y + x\alpha d(y)$ holds for all $x, y \in M$ and $\alpha \in \Gamma$, and d is called a Jordan derivation if $d(x\alpha x) = d(x)\alpha x + x\alpha d(x)$ holds for all $x, y \in M$ and $\alpha \in \Gamma$. The concept of derivation and Jordan derivation of Γ-rings were first introduced by Sapanci and Nakajima in [18]. Many significant results in classical ring theory have been developed by Herstein [8–10]. Some of these results were generalized in Γ-rings by Paul and Uddin [16,17], also see [4,5]. In [17], an example of a Lie ideal of a Γ-ring is given. Many mathematicians worked on derivations on Lie ideals of classical rings theories. In [13], Kamander gave some basic commutator identities. Two of these are given as

$$[x\beta y, z]_\alpha = [x, z]_\alpha \beta y + x\beta[y, z]_\alpha + x[\beta, \alpha]_z y,$$
$$[x, y\beta z]_\alpha = y\beta[x, z]_\alpha + [x, y]_\alpha \beta z + y[\beta, \alpha]_x z,$$

for all $x, y, z \in M$ and $\alpha, \beta \in \Gamma$. Throughout our paper, we consider the following condition

$$x\alpha y\beta z = x\beta y\alpha z. \tag{1}$$

By this condition the above two identities reduce to which extensively used in our paper. Using the assumption the basic commutator identities reduce to

$$[x\beta y, z]_\alpha = [x, z]_\alpha \beta y + x\beta[y, z]_\alpha,$$
$$[x, y\beta z]_\alpha = y\beta[x, z]_\alpha + [x, y]_\alpha \beta z,$$

for all $x, y, z \in M$ and $\alpha, \beta \in \Gamma$. In [3], Bergen et al. developed a number of significant results in classical rings by means of derivations and Lie ideals. As the examples, we cited the names Herstein [11], Aydin [1], Soyturk [19], Ferrero and Haetinger [6]. In the present paper, we generalize the results of [3] in Γ-rings.

2 Derivations on Lie Ideals of Γ-Rings

First we need some lemmas to prove our results.

Lemma 1 *[7]. If $U \not\subseteq Z(M)$ be a Lie ideal of a 2-torsion free prime Γ-ring M which satisfies the condition (1) and let $a, b \in M$ and $\alpha, \beta \in \Gamma$ such that $a\alpha U\beta b = 0$. Then, $a = 0$ or $b = 0$.*

Lemma 2. *Let $U \not\subseteq Z(M)$ be a Lie ideal of a 2-torsion free prime Γ-ring M. Then, $Z(U) = Z(M)$.*

Proof. We have $Z(U)$ is both a subring and a Lie ideal of M and $Z(U)$ cannot be a non-zero ideal of M. From Lemma 6 of [7], $Z(U) \subseteq Z(M)$. Since $Z(M) \subseteq Z(U)$, it follows that $Z(U) = Z(M)$.

Lemma 3. *Let U be a Lie ideal of a 2-torsion free prime Γ-ring M which satisfies the condition (1) and let $a \in M$. If $a \in Z([U,U]_\Gamma)$, then $a \in Z(M)$. That is $Z([U,U]_\Gamma) = Z(U)$.*

Proof. If $[U,U]_\Gamma \not\subseteq Z(M)$, then by Lemma 2, $a \in Z(M)$. Hence, $Z([U,U]_\Gamma) = Z(M) = Z(U)$. That is $Z([U,U]_\Gamma) = Z(U)$. On the other hand, if $[U,U]_\Gamma \subseteq Z(M)$ and $u \in U$, $m \in M$ and $\alpha \in \Gamma$, then we have $a = [u,[u,m]_\alpha]_\alpha \in Z(M)$. Then, $a\beta u = [u,[u,u\beta m]_\alpha]_\alpha \in Z(M)$ by using the condition (1) for $\beta \in \Gamma$. If $a \neq 0$ we obtain $u \in Z(M)$, which gives $a = 0$. Thus, $[u,[u,m]_\alpha]_\alpha = 0$ for all $m \in M$ and $\alpha \in \Gamma$. But, by Sub-lemma 3.8 of [7], $u \in Z(M)$. It follows that $U \subseteq Z(M)$. Therefore, $a \in Z(U)$. It follows that $Z([U,U]_\Gamma) = Z(U)$.

Lemma 4. *Let U be a Lie ideal of a 2-torsion free prime Γ-ring M which satisfies the condition (1). Let $u \in U$ and $d : M \to M$ be a derivation such that $[u,d(x)]_\alpha = 0$, for all $x \in M$ and $\alpha \in \Gamma$. Then, $u \in Z(U)$.*

Proof. Let $d(m) = [u,x]_\alpha$, for all $x \in M$ and $\alpha \in \Gamma$. By using the condition (1), we have seen that d is a derivation on M. Also, we have $d^2(x) = 0$, for all $x \in M$. Since d is a derivation on M, we have $0 = d^2(x\alpha y) = 2d(x)\alpha d(y)$. Since M is 2-torsion free, $d(x)\alpha d(y) = 0$. Replacing y by $m\beta y$, we obtain $d(x)\alpha m\beta d(y) = 0$, for all $x, y, m \in M$ and $\alpha, \beta \in Gamma$. Since M is prime, $d(x) = 0$ or $d(y) = 0$. If $d(x) = 0$, then $[u,x]_\alpha = 0$, for all $x \in M$ and $\alpha \in \Gamma$. This shows that $u \in Z(M)$. If $d(y) = 0$, then we get $[u,y]_\alpha = 0$, for all $y \in M$ and $\alpha \in \Gamma$, and again we get $u \in Z(M)$.

Lemma 5. *Let U be a Lie ideal of a 2-torsion free prime Γ-ring M which satisfies the condition (1). If $d \neq 0$ is a derivation of M such that $d(U) = 0$, then $U \subseteq Z(M)$.*

Proof. Let $u \in U$, $m \in M$ and $\alpha \in \Gamma$. Then, $[u,m]_\alpha \in U$. Since $d(U) = 0$, we have, $d([u,m]_\alpha) = 0$. But $[u,d(m)]_\alpha = d([u,m]_\alpha) = 0$. That is $[u,d(m)]_\alpha = 0$, for all $u \in U$, $m \in M$ and $\alpha \in \Gamma$. By Lemma 4, we have $u \in Z(M)$, for all $u \in M$. Hence, $U \subseteq Z(M)$.

Lemma 6. *Let U be a Lie ideal of a 2-torsion free prime Γ-ring M which satisfies the condition (1) and let $d \neq 0$ be a derivation of M such that $d(U) \subseteq Z(M)$. Then, $U \subseteq Z(M)$.*

Proof. If $U \not\subseteq Z(M)$, by Lemma 3, $V = [U,U]_\Gamma \subseteq Z(M)$. But, if $u, v \in U$ and $\alpha \in \Gamma$, then $d(u\alpha v - v\alpha u) = (d(u)\alpha v - v\alpha d(u)) + (u\alpha d(v) - d(v)\alpha u) = 0$, since $d(u), d(v) \in Z(M)$. Thus, $d(V) = 0$, but by Lemma 5, $V \subseteq Z(M)$, a contradiction.

Some other properties of $d(U)$ are given below.

Lemma 7. *Let $U \not\subseteq Z(M)$ be a Lie ideal of a 2-torsion free prime Γ-ring M which satisfies the condition (1) and let $d \neq 0$ be a derivation of M. If $a\alpha d(U) = 0$ (or $d(U)\alpha a = 0$) for all $m \in M$ and $\alpha \in \Gamma$, then $a = 0$.*

Proof. Let $u \in U$, $x \in M$ and $\beta, \gamma \in \Gamma$, then by using the condition (1), we have $(u\beta x - x\beta u)\gamma u \in U$. Using condition (1), we get $(x\beta u - x\beta u)\gamma u = u\beta(x\gamma u) - (x\gamma u)\beta u \in U$. Thus, $a\alpha d((u\beta x - x\beta u)\gamma u) = 0$, that is $a\alpha(d(u\beta x - x\beta u))\gamma u + a\alpha(u\beta x - x\beta u)\gamma d(u) = 0$. Since $u\beta x - x\beta u \in U$, so $a\alpha d((u\beta x - x\beta u)\gamma u) = 0$, that is $a\alpha d(u\beta x - x\beta u)\gamma u = 0$. Since $u\beta x - x\beta u \in U$, so

$$a\alpha d(u\beta x - x\beta u)\gamma u = 0. \tag{2}$$

So, the above relation reduces to $a\alpha(u\beta x - x\beta u)\gamma d(u) = 0$, for all $u \in U$, $x \in M$ and $\alpha, \beta, \gamma \in \Gamma$. Let $x = d(v)\delta y$, where $v \in U$, $y \in M$ and $\delta \in \Gamma$. Then,

$$a\alpha x = a\alpha d(u)\delta y = a\alpha d(v)\delta y = 0. \tag{3}$$

Therefore, by replacing x by $d(v)\delta y$ in (2), we have

$$a\alpha(u\beta d(v)\delta y - d(v)\delta y\beta u)\gamma d(u) = 0,$$

$$a\alpha u\beta d(v)\delta y\gamma d(u) - a\alpha d(v)\delta y\beta u\gamma d(u) = 0. \tag{4}$$

Using (3), (4) reduces to $a\alpha u\beta d(v)\delta y\gamma d(u) = 0$. This gives $a\alpha u\beta d(U)\delta M \gamma d(u) = 0$. Since $d(U) \neq 0$ and M is prime, then by Lemma 1, we have $a = 0$.

Now, we prove our main results.

Theorem 2. *Let $U \not\subseteq Z(M)$ be a Lie ideal of a 2-torsion free prime Γ-ring M which satisfies the condition (1) and let $d \neq 0$ be a derivation of M such that $d^2(U) = 0$, then $U \subseteq Z(M)$.*

Proof. First we assume that $U \not\subseteq Z(M)$, by Lemma 3, $V = [U, U]_\Gamma \not\subseteq Z(M)$. So, for the proof of theorem, it is sufficient to show that $V \subseteq Z(M)$. By Lemma 1 of [16], $[I, M]_\Gamma \subseteq U$ where I is an ideal of M such that $[I, M]_\Gamma \not\subseteq Z(M)$. Let $x \in [I, M]_\Gamma \subseteq U \cap I$ and $u \in V$. Then, $w = d(u) \in d([U, U]_\Gamma) \subseteq U$, so $d(w) = d^2(u) = 0$, by hypothesis. If $y \in M$ then, since $m\beta w \in I$, $[m\beta w, y]_\alpha \in [I, M]_\Gamma \subseteq U$. Hence,

$$\begin{aligned} 0 &= d^2([m\beta w, y]_\alpha) = d^2(m\beta[w, y]_\alpha + [m, y]_\alpha\beta w) \\ &= d^2(m)\beta[w, y]_\alpha + 2d(m)\beta d([w, y]_\alpha) + 2m\beta d^2([w, y]_\alpha) \\ &\quad + d^2([m, y]_\alpha\beta w + 2d([m, y]_\alpha\beta d(w + 2[m, y]_\alpha\beta d^2(w), \end{aligned}$$

by using the condition (1). Using $d(w) = d^2([m, y]_\alpha) = d^2(m) = d^2[w, y]_\alpha. = 0$, in the above relation, we obtain $2d(m)\beta d([w, y]_\alpha) = 0$. Since M is 2-torsion free, $d(m)\beta d([w, y]_\alpha) = 0$. This gives that $d([I, M]_\Gamma)\Gamma d([d(V), M]_\Gamma) = 0$. But $[I, M]_\Gamma$ is a Lie ideal of M such that $[I, M]_\Gamma \not\subseteq Z(M)$. So, by Lemma 7, $d([d(V), M]_\Gamma) = 0$. Hence, if $u \in V$, $x \in M$ and $\alpha \in \Gamma$, $0 = d([d(u), x]_\alpha) = 0$. Therefore,

$$\begin{aligned} 0 &= d(d(u)\alpha x - x\alpha d(u)) \\ &= d^2(u)\alpha x + d(u)\alpha d(x) - d(x)\alpha d(u) - x\alpha d^2(u) \\ &= d(u)\alpha d(x) - d(x)\alpha d(u) \\ &= [d(u), d(x)]_\alpha. \end{aligned}$$

Therefore, $d(V) \in Z(d(M))$. By Lemma 4, $d(V) \subseteq Z(M)$, hence by Lemma 6, $V \subseteq Z(M)$. This completes the proof.

The special cases, when $U = M$ or U is an ideal of M, in this situation, $d = 0$ if $d^2(U) = 0$, are immediate consequences of Theorem 2. Let $a \in M$ be fixed element. If $d(x) = [a, x]_\alpha$ for all x M, then by using the condition (1), we have seen that d is the inner derivation of M. If $d^2(U) = 0$, that is, $[a, [a, U]_\Gamma] = 0$, we conclude from the theorem that if U Z(M), then $[a, U]_\Gamma = 0$ and so $a \in Z(M)$ by Lemma 4. Since the ideals (and so, the prime ideals) of M are invariant with respect to inner derivations, so we obtain the following corollary.

Corollary 1. *Let U be a Lie ideal of a 2-torsion free semiprime Γ-ring M which satisfies the condition (1). If $[a, [a, U]_\Gamma] = 0$, for some $a \in M$ and $\alpha \in \Gamma$, then $[a, U]_\Gamma = 0$.*

Theorem 3. *If $U \not\subseteq Z(M)$ be a Lie ideal of a 2-torsion free prime Γ-ring M which satisfies the condition (1) and let $d \neq 0$ be a derivation of M. Then, $Z(d(U)) = Z(M)$.*

Proof. Let $a \in Z(d(U))$, and assume that $a \notin Z(M)$. Since $U \not\subseteq Z(M)$, $V = [U, U]_\Gamma \not\subseteq Z(M))$ by Lemma 3. Moreover, $d(V) \subseteq U$. Thus, $[a, d^2(u)] = 0$, for all $u \in V$ and $\alpha \in \Gamma$. But we have $[a, d(u)]_\alpha = 0$. Applying d to this, we get $d([a, d(u)]_\alpha = 0$. That is

$$\begin{aligned} 0 &= d(a\alpha d(u) - d(u)\alpha a) \\ &= d(a)\alpha d(u) + a\alpha d^2(u) - d^2(u)\alpha a - d(u)\alpha d(a) \\ &= [d(a), d(u)]_\alpha + [a, d^2(u)]_\alpha. \end{aligned}$$

Since $[a, d^2(u)]_\alpha = 0$, we obtain $[d(a), d(u)]_\alpha = 0$. Therefore, $a \in Z(d(V))$ and $d(a) \in Z(d(V))$. But $d(a\alpha u - u\alpha a) = d(a)\alpha u + a\alpha d(u) - d(u)\alpha a u \alpha d(a) = [d(a), u]_\alpha + [a, d(u)]_\alpha$. Since $[a, d(u)]_\alpha = 0$, $d([a, u]_\alpha) = [d(a), u]_\alpha \in d(V)$. Hence, $[d(a), [d(a), V]_\alpha]_\alpha = 0$. By Theorem 2, we obtain that $[d(u), V]_\alpha = 0$, and since $V \not\subseteq Z(M)$, by Lemma 2, $d(a) \in Z(M)$. By the similar manner, since $a\alpha a \in Z(d(U))$, $2a\alpha d(u) = d(a\alpha a) \in Z(M)$, for, $a \in Z(M)$ and $d(a) \in Z(M)$, the fact that $a\alpha d(a) \in Z(M)$ forces $d(a) = 0$. Therefore, $d(a) = 0$ for all $a \in Z(d(U))$ which is not in $Z(M)$. If $d(b) \neq 0$ for some $b \in Z(d(U))$, then by the above, $b \in Z(M)$. Furthermore, if $a \in Z(d(U))$, $a \notin Z(M)$, then $d(a) = 0$, hence $d(a + b) = d(a) + d(b) = d(b) \neq 0$. Consequently, $a + b \in Z(M)$, together with $b \in Z(M)$ we have $a \in Z(M)$, a contradiction. Hence, if we assume that $Z(d(U)) \not\subseteq Z(M)$, then it is forced to $d(a) = 0$, for all $a \in Z(d(U))$.

Let $K = \{x \in M \mid d(x) = 0\}$. Then, we have $Z(d(U)) \subseteq K$. Moreover, if $a \in Z(d(U))$ and u U, then $d([a, u]_\alpha) = [a, d(u)]_\alpha = 0$ since $d(a) = 0$. Thus, $[a, U]_\Gamma \subseteq K$. Now, since $U \not\subseteq Z(M)$, by Lemma 1 of [16], $[I, M]_\Gamma \subseteq U$ where I is an ideal of M such that $[I, M]_\Gamma \not\subseteq Z(M)$. If $m \in [I, M]_\Gamma \subseteq U \cap I$, then $m\alpha a \in I$, for $\alpha \in \Gamma$, hence, for $u \in U$, and $\beta \in \Gamma$, $[m\alpha a, u]_\beta \in U$. By the condition (1), we have, is $[m, u]_\beta \alpha a + m\alpha[a, u]_\beta \in U$. Therefore,

$$a \in Z(d([m, u]_\beta \alpha a + m\alpha[a, u]_\beta)) = Z(d([m, u]_\beta)\alpha a + d(m)\alpha[a, u]_\beta),$$

since $d(a) = d([a,u]_\beta) = 0$, because $a, [a,u]_\beta \in K$. Since $a \in Z(d([m,u]_\beta))$ and $a \in Z(d(m))$, we get $d(m)\alpha[a,[a,u]_\beta]_\beta = 0$ for all $m \in [I,M]_\Gamma$, $u \in L$ and $\alpha, \beta \in \Gamma$. Thus, by the proof of the last part of the Theorem 1, we have $[a,[a,U]_\Gamma]_\Gamma = 0$. Therefore, by Theorem 2 (or Corollary 1), we get $a \in Z(M)$, since $U \not\subseteq Z(M)$. The proof of the Theorem is completed.

3 Subring Generated by $d(U)$

For the development of this section, we need a significant result due to Herstein, Theorem 1 [12]. This result can be developed in Γ-rings as follows.

Theorem 4. *Let M be any Γ-ring, d be a derivation of M such that $d^3 \neq 0$. Then, Q, the subring generated by all $d(m)$, where $m \in M$, contains a non-zero ideal of M.*

Proof. Since $d^3 \neq 0$ and $d(M) \subseteq Q$, $d^2(Q) \subseteq d^3(M) \neq 0$. Take $q \in Q$ such that $d^2(q) \neq 0$. For all $m \in M$ and $\alpha \in \Gamma$, we have $d(m\alpha q) = d(m)\alpha q + m\alpha d(q) \in Q$. Since both q and $d(m)$ are in Q, $m\alpha d(q) \in Q$. This gives that $M\Gamma d(q) \subseteq Q$. Similarly, $d(q)\Gamma M \subseteq Q$.

If $r, s \in M$ and $\alpha, \beta \in \Gamma$, then

$$d(r\alpha d(q)\beta s) = d(r)\alpha d(q)\beta s + r\alpha d^2(q)\beta s + r\alpha d(q)\beta d(s) \in Q.$$

But $d(q)\beta s \in Q$ and $r\alpha d(q) \in Q$, we have $r\alpha d^2(q)\beta s \in Q$, for all $r, s \in M$ and $\alpha, \beta \in \Gamma$. From the above, $M\Gamma d^2(q) \subseteq Q$, $d^2(q)\Gamma M \subseteq Q$, we conclude that the ideal of M generated by $d^2(q) \neq 0$ must be in Q. If $d^3 = 0$ the result need not be true. Let M be a prime Γ-ring having nilpotent elements, and let $0 \neq a \in M$ be such that $a\alpha a = 0$, for all $\alpha \in \Gamma$. Let $d : M \to M$ be defined by $d(x) = [a,x]_\alpha$. Then, d is aviation of M if M satisfies the condition (1). Since $a\alpha a = 0$, $B = a\Gamma M + M\Gamma Q$ is a subring of M and contains $d(M)$. Also, $d^3 = 0$ and $d^2 \neq 0$, if $charM \neq 2$. But B contains no nonzero ideals of M, for $a\Gamma B\Gamma a = 0$.

By the Theorem 4, we have seen that for any Γ-ring M, $d(M)^*$ contains a nonzero ideal of M when M is a prime Γ-ring and $d^3 \neq 0$. Let $U \subseteq Z(M)$ be a Lie ideal of a Γ-ring M. We denote $d(U)^*$ to be the subring generated by $d(U)$, where $d \neq 0$ is a derivation of M. For the rest of our articles, we assume that $U \subseteq Z(M)$ is a 2-torsion free prime Γ-ring M which satisfies the condition (1) and let $d \neq 0$ be a derivation of M.

We shall arrange the frequent use of the Lie ideals $K = [U,U]_\Gamma$ and $W = [K,K]_\Gamma$ which are closely related to U. Our main result in this section will follow as a consequence of several lemmas.

Lemma 8. *If $d^3 \neq 0$ and if $d(K)^*$ contains a non-zero left ideal I of M and a non-zero right ideal J of M, then $d(U)^*$ contains a non-zero ideal of M.*

Proof. Since $K = [U, U]_\Gamma$ and $d(K) \subseteq U$ we have seen that $d(d(K)^*) \subseteq d(U)^*$. Let $a \in I \subseteq d(K)^*$ and $x \in M$. Then, $d(x\alpha a) \in d(I) \subseteq d(d(K)^*) \subseteq d(U)^*$ for every $\alpha \in |Gamma$. So, $d(x)\alpha a + x\alpha d(a) \in d(U)^*$, for all $\alpha \in \Gamma$. Since $d(x)\alpha a$ is in I, and so, in $d(K)^* \subseteq d(U)^*$, we have $x\alpha d(a) \in d(U)^*$. Thus, $M\Gamma d(I) \subseteq d(U)^*$. Similarly, $d(J)\Gamma M \subseteq d(U)^*$. If $a \in I$, $u \in K$ and $\alpha \in \Gamma$, then $d([u, a]_\alpha) \in d(K)^*$, so $[d(u), a]_\alpha + [u, d(a)]_\alpha \in d(K)^*$. But $d(u)\alpha a \in I \subseteq d(K)^*$, $u\alpha d(a) \in d(U)^*$, by the above, and $a\alpha d(u) \in I\Gamma d(K) \subseteq d(K)^*$. So, $d(I)\Gamma K \subseteq d(U)^*$. Similarly, $K\Gamma d(J) \subseteq d(U)^*$. Let $P = I\Gamma K\Gamma J$. Then, P is an ideal of M and, by Lemma 1, $P \neq 0$. Also, we have $d(P) = d(I\Gamma K\Gamma J) \subseteq d(I)\Gamma K\Gamma J + I\Gamma d(K)\Gamma J + I\Gamma K\Gamma d(J) \subseteq d(U)^*$, since $d(I)\Gamma K$, $K\Gamma d(J)$, I and J are all in $d(U)^*$. Thus, $d(P)^* \subseteq d(U)^*$. But, if $d^3 \neq 0$, P is an ideal of the prime Γ-ring M, $d(I)^*$ contains a non-zero ideal of M. Therefore, $d(U)^*$ contains a non-zero ideal of M.

Lemma 9. *Let $P \neq 0$ be an ideal of M. If $d(U)^*$ does not contain both a non-zero left-ideal and a non-zero right-ideal of M and if $[c, P]_\Gamma \subseteq d(U)^*$. Then, $c \in Z(M)$.*

Proof. Let $s \in d(U)$, $p \in P$ and $\alpha, \beta \in \Gamma$. Then, $[c, s\beta k]_\alpha = [c, s]_\alpha \beta k + s\beta[c, k]_\beta$. We have $[c, s\beta p]_\alpha \in d(U)^*$, for all $\alpha, \beta \in \Gamma$. Because $s \in d(U)$, $[c, p]_\beta \in d(U)$ we get $s\beta[c, k]_\alpha \in d(U)^*$. Hence, $[c, s]_\alpha \beta p \in d(U)^*$, that is, the right ideal of M, $[c, d(U)]_\Gamma \Gamma P \subseteq d(U)^*$. Similarly $P\Gamma[c, d(U)]_{Gamma} \subseteq d(U)^*$ is a left-ideal of M lying in $d(U)^*$. By the hypothesis one of $P\Gamma[c, d(U)]_\Gamma = 0$ or $[c, d(U)]_\Gamma \Gamma P = 0$. Hence, $[c, d(U)]_\Gamma = 0$, by the primeness of M. Therefore, we conclude that $c \in Z(M)$, by Theorem 3.

Now we shall prove a highly special and somewhat messy.

Lemma 10. *If $d^2(U\Gamma U) = 0$, then $d^3(W) = 0$.*

Proof. Since $U \not\subseteq Z(M)$, by Lemma 10 none of $U, K = [U, U]_\Gamma$, $W = [K, K]_\Gamma$ is in $Z(M)$. Also, $d(K) \subseteq U$, $d(W) \subseteq K$, $d^2(W) \subseteq U$. If $u \in U$, $k \in K$, $r \in W$ and $\alpha, \beta \in \Gamma$, then for any $t \in U$, we have

$$d^2(u)\alpha d^2(d(k)\beta d^2(r)\gamma t - d^2(r)\gamma t\beta d(v)) = 0. \tag{5}$$

After calculation and making use of $d(k) \in U$, $d^2(r) \in U$ and $d^2(U\Gamma U) = 0$ and the condition (1), (2) reduces to

$$d^2(u)\alpha d(k)\beta(d^4(r)\gamma t + 2d^3(r)\gamma d(t)) = 0. \tag{6}$$

Now, we choose $t \in d(K) \subseteq U$ in the relation (6), we get from (6)

$$d^2(u)\alpha d(k)\beta d^4(r)\gamma d(k) = 0,$$

since $d^3(r)\gamma d(t) = 0$ for such t. By Lemma 4, we have seen that

$$d^2(u)\alpha d(k)\beta d^4(r) = 0.$$

Using this relation in (6), we have

$$d^2(u)\alpha d(k)\beta d^3(r)\gamma d(u) = 0, \tag{7}$$

for all $u \in U$, $k \in K$, $r \in W$ and $\alpha, \beta, \gamma \in \Gamma$. Therefore, by Lemma 4, we get

$$d^2(u)\alpha d(k)\beta d^3(r) = 0, \tag{8}$$

for all $u \in U$, $k \in K$, $r \in W$ and $\alpha, \beta, \gamma \in \Gamma$. Similarly, reversing sides, we get

$$d^3(r)\alpha d(k)\beta d^2(u) = 0, \tag{9}$$

for all $u \in U$, $k \in K$, $r \in W$ and $\alpha, \beta, \gamma \in \Gamma$. Now, consider $d^2(d(t))\alpha d^2(k\beta d(r) - d(r)\beta k) = 0$ where $t, r \in W$, $k \in K$ and $\alpha, \beta \in \Gamma$. Expanding this and making use of (8) we have that $d^3(t)\alpha k\beta d^3(r) = 0$ for all for all $k \in K$, $t, r \in W$ and $\alpha, \beta, \gamma \in \Gamma$. Thus, $d^3(W)\Gamma K\Gamma d^3(W) = 0$. By Lemma 4 we obtain $d^3(W) = 0$, as claimed.

Lemma 11. *If $d^3(U) = 0$, then $d^3 = 0$.*

Proof. Let $u \in U$, $m \in M$ and $\alpha \in Gamma$. Then,

$$0 = d^3([u, m]_\alpha) = 3[d^2(u), d(m)]_\alpha + 3[d(u), d^2(m)]_\alpha + [u, d^3(m)]_\alpha. \tag{10}$$

In this replace u by $d^2(w)$ where $w \in W$, to obtain

$$[d^2(w), d^3(m)] = 0, \tag{11}$$

for all $w \in W$, $m \in M$ and $\alpha \in Gamma$ We now replace u by $d(w)$, m by $d(m)$, where $w \in W$, in (10). By using (11), we get $[d(w), d4(m)]_\alpha = 0$, for all $w \in W$, $m \in M$ and $\alpha \in \Gamma$. Since $W \not\subseteq Z(M)$, by Theorem 3 we get that $d^4(M) \subseteq Z(M)$. Since $d^4(m) \subseteq Z(M)$ for all $m \in M$. If $u \in U$, $m \in M$ and $\alpha \in \Gamma$, then

$$0 = d^4([u, m]_\alpha) = 6[d^2(u), d2(m)]_\alpha + 4[d(u), d^3(m)]_\alpha.$$

But we also have that

$$0 = d^3([u, d(m)]_\alpha) = 3[d^2(u), d^2(m)]_\alpha + 3[d(u), d^3(m)]_\alpha.$$

Playing these last two relations off against each other leads us to $2[d(u), d^3(r)]_\alpha = 0$, and so $[d(u), d^3(r)]_\alpha = 0$, for all $u \in U$, $r \in M$ and $\alpha \in \Gamma$. By Theorem 3, $d^3(M) \subseteq Z(M)$. Thus, if $r \in M$, $u \in U$ and $\alpha \in \Gamma$, then $d^3(r\alpha d^2(u)) = d^3(r)\alpha d^2(u) \in Z(M)$. However, $d^3(M) \subseteq Z(M)$, so since $d^3(M)\Gamma d^2(U) \subseteq Z(M)$ if $d^3(M) \neq 0$, we are forced to $d_2(U) \subseteq Z(M)$. Now, suppose that $d^3(M) \neq 0$, as we have seen, we must have $d^2(U) \subseteq Z(M)$. If $r \in M$, $u \in U$ and $\alpha \in \Gamma$, then $d^4(r\alpha d(u)) = d^4(r)\alpha d(u) + 4d^3(r)\alpha d^2(u) \in Z(M)$, and since $d^3(r) \in Z(M)$, $d^2(u) \in Z(M)$, we see that $d^4(r)\alpha d(u) \in Z(M)$, that is, $d^4(M)\Gamma d(U) \subseteq Z(M)$. By Lemma 8 we know that $d(L) \not\subseteq Z(M)$, by the above we know that $d^4(M) \subseteq Z(M)$, these, combined with $d^4(M)\Gamma d(U) \subseteq Z(M)$ force $d^4(M) = 0$. Again, if $r \in M$, $u \in U$ and $\alpha \in \Gamma$, then $0 = d^4(r\alpha d(u)) = 4d^3(r)\alpha d^2(u)$, so that $d^3(M)\Gamma d^2(U) = 0$. But $d^2(U) \neq 0 \subseteq Z(M)$ (by Theorem 2) so we conclude that $d^3(M) = 0$. This completes the proof.

Now, we have come to a positive to prove our main results.

Theorem 5. *If $U \not\subseteq Z(M)$ is a Lie ideal of M and d is a derivation of M such that $d^3 \neq 0$, then $d(U)^*$ contains a non-zero ideal of M.*

Proof. If $K = [U, U]_\Gamma$ and $W = [K, K]_\Gamma$, in view of Lemma 8 it is sufficient to show that $d(K)^*$ contains a non-zero left, and a non-zero right ideal of M. We assume that this situation is not true. Then, we have to show that this leads to $d^2([W, W]_\Gamma \Gamma [W, W]_\Gamma) = 0$, by Lemmas 10 and 11 we shall arrive the contradiction $d^3 = 0$.

Let $a \in d(W)$, where $w \in [W, W]_\Gamma$. Then, for all $x \in M$ and $\alpha, \beta \in \Gamma$, $a\alpha(a\beta x - x\beta a) = a\alpha(a\beta x) - (a\alpha x)\beta a \in W$, using condition (1)

$$d(a\beta(a\alpha x - x\alpha a)) = d(a)\beta(a\alpha x - x\alpha a) + a\beta d(a\alpha x - x\alpha a) \in d(W).$$

But $a \in d(W) \subseteq d(K)$ and $d(a\alpha x - x\alpha a) \in d(V)$, whence we get

$$d(a)\beta(a\alpha x - x\alpha a) \in d(U)^*, \tag{12}$$

for all $a \in d([W, W]_\Gamma)$ and $x \in M$. On the other hand, if $u \in V$, then $d([a, u]_\alpha) = [d(a), u]_\alpha + [a, d(u)]_\alpha d(V)$, and since $a \in d(W) \subseteq d(V)$ we have that $[a, d(u)]_\alpha \in d(V)^*$. Hence,

$$[d(a), V]_\alpha \in d(U)^*, \tag{13}$$

for all $a \in d([W, W]_\Gamma)$. We also have $d(a)\beta d(a\alpha r - r\alpha a) = d(a)\beta[d(a), r]_\alpha + d(a)\beta[a, d(r)]_\alpha \in d(U)^*$, by (13), $d(a)\beta[a, d(r)]_\alpha \in d(U)^*$. The net result of the above becomes

$$d(a)\beta[d(a), r]_\alpha \in d(U)^*, \tag{14}$$

for all $a \in d([W, W]_\Gamma)$, $r \in M$ and $\alpha, \beta \in \Gamma$. We linearize (14) on a to get

$$s = d(a)\beta[d(b), r]_\alpha + d(b)\beta[d(a), r]_\alpha \in d(U)^*, \tag{15}$$

for all $a, b \in d([W, W]_\Gamma)$, $r \in M$ and $\alpha, \beta \in \Gamma$. If $t = [d(a)\beta d(b), r]_\alpha = d(a)\beta[d(b), r]_\alpha + [d(a), r]_\alpha \beta d(b)$, then

$$s - t = d(b)\beta[d(a), r]_\alpha - [d(a), r]_\alpha \beta d(b) \in d(U)^*,$$

by (13). Thus, we conclude that $t \in d(U)^*$ that is, $[d(a)\beta d(b), M]_\alpha \subseteq d(U)^*$. Because $d(U)^*$ does not contain both a non-zero left-ideal and a non-zero right-ideal of M, by Lemma 9, we have that $d(a)\beta d(b) \in Z(M)$, for all $a, b \in d([W, W]_\Gamma)$. Let $a = d(a)\beta d(b)$, by (12), $d(b)\beta(b\alpha x - x\alpha b) \in d(U)^*$ and since $d(a) \in d(V)$, we get that $a\beta(b\alpha x - x\alpha b) = d(a)\delta d(b)\beta(b\alpha x - x\alpha b) \in d(U)^*$. Because $a \in Z(M)$ this says that $[b, I]_\alpha \in d(U)^*$, where $I = a\Gamma M$ is an ideal of M. By our hypothesis on $d(U)^*$, if I 0, we would conclude by Lemma 1 that $b \in Z(M)$ for all $b \in d([W, W]_\Gamma)$, by Lemma 6 we would be led to $[W, W]_\Gamma \subseteq Z(M)$, and so $U \subseteq Z(M)$, a contradiction. Thus, $I = a\Gamma M = 0$, hence a = 0. In other words, $d(a)\Gamma d(b) = 0$, for all $a, b \in d([W, W]_\Gamma)$, that is, $d([W, W]_\Gamma)\Gamma d([W, W]_\Gamma) = 0$. By Lemmas 10 and 11 we reach the contradiction that $d^3 = 0$.

We develop a result which simultaneously implies those of Theorems 2 and 3 and come to the end of paper.

Theorem 6. *Let $U \not\subseteq Z(M)$ be a Lie ideal of M. Suppose that d_1 and d_2 are derivations of M such that $d_1 d_2(U) = 0$. Then, either $d_1 = 0$ or $d_2 = 0$.*

Proof. Suppose that $d_1 \neq 0$ and $d_2 \neq 0$. Let $K = [U, U]_\Gamma$. Then, for all $k \in K$, $d_1(k) \in U$, hence $d_1 d_2([u, d_1(k)]_\alpha) = 0$, for all $u \in U$ and $\alpha \in \Gamma$. Thus, $d_1([d_2(u), d_2(k)]_\alpha + [u, d_2^2(k)]_\alpha) = 0$, which gives us, since $d_1 d_2(U) = 0$ and d_1 is a derivation of M, that $[d_1(u), d_2^2(k)]_\alpha = 0$, for all $u \in U$, $k \in K$ and $\alpha \in \Gamma$. Thus, $d_2^2(k) \in Z(d_2(U))$, by Theorem 3. If $k \in K$, $r \in M$ and $\alpha \in \Gamma$, then $0 = d_1(d_2([d_2(k), r]_\alpha)) = d_1([d_2(k), d_1(r)]_\alpha)$ since $d_1^2(k) \in Z(M)$. Thus, expanding, we get $[d_2 d_1(k), d_1(r)]_\alpha + [d_1(k), d_2 d_1(r)]_\alpha = 0$ and so $[d_1(k), d_2 d_1(r)]_\alpha = 0$, for all $k \in K$, $r \in M$ and $\alpha \in \Gamma$, that is $[d_1(K), d_2 d_1(M)]_\alpha = 0$. Since $d \neq 0$, by Theorem 3, $d_2 d_1(M) \subseteq Z(M)$. Now, for all $k \in K$, $u \in U$ and $\alpha \in \Gamma$, we have $d_2 d_1(d_1(v)\alpha u) = d_2(d_1^2(v)\alpha u + d_1(k)\alpha d_1(u)) = d_1^2(k)\alpha d_2(u)$, since $d_2 d_1^2(k) = d_2 d_1(d_1(k)) = 0$, because $d_1(k) \in U$, and $d_2(d_1(k)d_1(u)) = 0$. Therefore, $d_1^2(K)\alpha d_2(U) \subseteq Z(M)$. But, since $U \not\subseteq Z(M)$, $d_2(U) \not\subseteq Z(M)$ by Lemma 6, in consequence, $d_1^2(K) = 0$, since we know that $d_1^2(K) \subseteq Z(M)$ and $d_1^2(K)\alpha d_2(U) \subseteq Z(M)$. Since $K \not\subseteq Z(M)$ and $d_1^2(K) = 0$, by Theorem 2 we obtain $d_1 = 0$. To see that Theorem 6 implies Theorem 2, we may choose $d_1 = d_2 = d$. As for Theorem 3, if $d_2 \neq 0$, d_1 is a derivation of M and if $a \in Z(d_2(U))$, let d_1 be defined by $d_1(x) = [a, x]_\alpha$, for $\alpha \in \Gamma$, then by the condition (1) d_1 is a derivation of M. We see that $d_1 d_2(U) = 0$. Hence, by Theorem 6 since $d_1 = 0$, $d_2 \neq 0$. Therefore, $[a, x]_\alpha = 0$, for all $x \in M$ and $\alpha \in \Gamma$, that is, $a \in Z(M)$.

References

1. Aydin, N., Soyturk, M.: $(\sigma, \tau)\Gamma$-Lie ideals in prime rings with derivation. Turk. J. Math. **19**(3), 239–244 (1995)
2. Barnes, W.E.: On the $\Gamma\Gamma$-rings of Nobusawa. Pac. J. Math. **18**, 411–422 (1966)
3. Bergen, J., Herstein, I.N., Kerr, J.W.: Lie ideals and derivations of prime rings. J. Algebra **71**(1), 259–267 (1981)
4. Ceven, Y.: Jordan left derivations on completely prime gamma rings. C.Ü. Fen-Edebiyat Fakültesi Fen Bilimleri Dergisi **23**(2), 39–43 (2002)
5. Dey, K.K., Paul, A.C., Rakhimov, I.S.: Generalized derivations in semiprime gamma-rings, Int. J. Math. Math. Sci. (2012). Art. ID 270132. 14 pages
6. Ferrero, M., Haetinger, C.: Higher derivations and a theorem by Herstein. Quaest. Math. **25**(2), 249–257 (2002)
7. Halder, A.K., Paul, A.C.: Jordan left deruvations on Lie ideals of prime Γ-rings. Punjab Univ. J. Math. **1**, 1–7 (2011)
8. Herstein, I.N.: On the Lie structure of an associative ring. J. Algebra **14**, 561–571 (1970)
9. Herstein, I.N.: A note on derivations. Canad. Math. Bull. **21**(3), 369–370 (1978)
10. Herstein, I.N.: A note on derivations. II. Canad. Math. Bull. **22**(4), 509–511 (1979)
11. Herstein, I.N.: Rings with Involution, Chicago Lectures in Mathematics. The University of Chicago Press, Chicago (1976)

12. Herstein, I.N.: Topics in Ring Theory. The University of Chicago Press, Chicago (1969)
13. Kandamar, H.: The k-derivation of a gamma Γ-ring. Turk. J. Math. **24**(3), 221–231 (2000)
14. Nabusawa, N.: On a generalization of the ring theory. Osaka J. Math. **1**, 65–75 (1964)
15. Özturk, M.A., Jun, Y.B.: On the centroid of the prime gamma rings. Comm. Korean Math. Soc. **15**(3), 469–479 (2000)
16. Paul, A.C., Sabur Uddin, M.: Lie and Jordan structure in simple gamma rings. J. Physical Sci. **14**, 77–86 (2010)
17. Paul, A.C., Sabur Uddin, M.: Lie structure in simple gamma rings. Int. J. Pure Appl. Sci. Tech. **4**(2), 63–70 (2011)
18. Sapanci, M., Nakajima, A.: Jordan derivations on completely prime gamma rings. Math. Japon. **46**(1), 47–51 (1997)
19. Soyturk, M.: On spherical convolution operators with logarithmic kernel. Hadronic J. Suppl. **14**(4), 401–408 (1999)

λ_d-Statistical Convergence, λ_d-statistical Boundedness and Strong $(V,\lambda)_d$ −summability in Metric Spaces

Emine Kayan[1] and Rifat Çolak[2(✉)]

[1] Institute of Science and Technology, Firat University, 23119 Elazığ, Turkey
eminekayan86@gmail.com
[2] Department of Mathematics, Faculty of Science Firat University, 23119 Elazığ, Turkey
rftcolak@gmail.com

Abstract. In this paper, we introduce and study λ_d-statistical convergence, λ_d-statistical boundedness and strong $(V,\lambda)_d$ −summability of sequences in metric spaces. Furthermore we establish some relations between the sets of λ_d-statistically convergent sequences, between the sets of λ_d-statistically bounded sequences, between the sets of λ_d-statistical convergent sequences and the sets of strongly $(V,\lambda)_d$ −summable sequences for various sequences $\lambda=(\lambda_n)$ in Λ. Furthermore we establish some inclusion relations between the sets of strongly $(V,\lambda)_d$ −summable sequences for various sequences $\lambda=(\lambda_n)$ in set Λ^*.

Keywords: $\lambda-$density · Statistical convergence · $\lambda-$statistical convergence · Strong summability

1 Introduction and Preliminaries

The notion of statistical convergence of a sequence (of real or complex numbers) was defined by Fast [5] and Steinhaus [15] independently in 1951. After then this subject have been studied by various mathematicians (see [2,4,6,7,11–14]).

A sequence $x=(x_k)$ of real (or complex) numbers is said to be *statistically convergent* to a number L if for every $\varepsilon>0$

$$\lim_{n\to\infty}\frac{1}{n}\left|\{k\leq n:|x_k-L|\geq\varepsilon\}\right|=0.$$

Let $\lambda=(\lambda_n)$ be a non-decreasing sequence of positive real numbers tending to ∞ such that

$$\lambda_{n+1}\leq\lambda_n+1,\ \lambda_1=1.$$

The set of all such sequences will be denoted by Λ.

Let $K\subset\mathbb{N}$, $\lambda=(\lambda_n)\in\Lambda$, and define $\lambda-$*density* of K as

$$\delta_\lambda(K)=\lim_{n\to\infty}\frac{1}{\lambda_n}\left|\{k\in I_n:\ k\in K\}\right|$$

© Springer Nature Singapore Pte Ltd. 2017
D. Giri et al. (Eds.): ICMC 2017, CCIS 655, pp. 391–403, 2017.
DOI: 10.1007/978-981-10-4642-1_33

where $I_n = [n - \lambda_n + 1, n]$ and $|.|$ denotes the number of elements of the involved set. $\lambda-$density $\delta_\lambda(K)$ reduces to the natural density $\delta(K)$ in case $\lambda_n = n$ [3].

The generalized de la Vallée-Poussin mean is defined by

$$t_n(x) = \frac{1}{\lambda_n} \sum_{k \in I_n} x_k.$$

A sequence $x = (x_k)$ is said to be (V, λ)-summable to a number L (see [10]) if

$$t_n(x) \to L \text{ as } n \to \infty.$$

If $\lambda_n = n$ for each $n \in \mathbb{N}$, then (V, λ)-summability reduces to $(C, 1)$-summability. We write

$$[C, 1] = \left\{ x = (x_k) : \lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} |x_k - L| = 0 \text{ for some } L \right\},$$

$$[V, \lambda] = \left\{ x = (x_k) : \lim_{n \to \infty} \frac{1}{\lambda_n} \sum_{k \in I_n} |x_k - L| = 0 \text{ for some } L \right\}$$

for the sets of sequences $x = (x_k)$ which are strongly Cesàro summable and strongly $(V, \lambda)-$summable, respectively.

The $\lambda-$statistical convergence was introduced by Mursaleen in [11] as follows for number sequences.

Let $\lambda = (\lambda_n) \in \Lambda$. A sequence $x = (x_k)$ is said to be $\lambda-$*statistically convergent* or S_λ-convergent to L if for every $\varepsilon > 0$

$$\lim_{n \to \infty} \frac{1}{\lambda_n} |\{k \in I_n : |x_k - L| \geq \varepsilon\}| = 0,$$

where $I_n = [n - \lambda_n + 1, n]$. In this case we write $S_\lambda - \lim x = L$ or $x_k \to L(S_\lambda)$, and $S_\lambda = \{x = (x_k) : S_\lambda - \lim x = L \text{ for some number } L\}$.

In this study, we determine the relations between the sets $S_{\lambda d}$ and $S_{\mu d}$, $S_{\mu d}$ and $BS_{\lambda d}$, $BS_{\mu d}$ and $BS_{\lambda d}$, the sets $S_{\lambda d}$ and $[V, \mu]_d$ for various sequences λ, μ in the class Λ. Furthermore we determine the relations between the sets $[V, \lambda]_d$ and $[V, \mu]_d$ for various sequences λ, μ in the class Λ^*.

Throughout the paper $c(X)$ and $l_\infty(X)$ will denote the sets of convergent and bounded sequences in metric space (X, d), respectively and by the statement "for all $n \in \mathbb{N}_{n_\circ}$" we mean "for all $n \in \mathbb{N}$ except finite numbers of positive integers" where $\mathbb{N}_{n_\circ} = \{n_\circ, n_\circ + 1, n_\circ + 2, ...\}$ for some $n_\circ \in \mathbb{N} = \{1, 2, 3, ...\}$.

2 λ_d-Statistical Convergence and λ_d-statistical Boundedness in Metric Spaces

$\lambda-$statistical convergence of number sequences was introduced and studied by Mursaleen [11]. In this section we define and study λ_d-statistical convergence

and λ_d-statistical boundedness of a sequence in a metric space and give the relations between the sets of λ_d-statistical convergent sequences and the sets of λ_d-statistical bounded sequences for various sequences $\lambda = (\lambda_n)$ in Λ.

Definition 2.1. Let (X, d) be a metric space and let $\lambda = (\lambda_n) \in \Lambda$ be given. A sequence $x = (x_k)$ in metric space (X, d) is called λ_d-statistically convergent to a point $x_o \in X$ if for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \frac{1}{\lambda_n} |\{k \in I_n : x_k \notin B_\varepsilon (x_o)\}| = 0,$$

where $I_n = [n - \lambda_n + 1, n]$ and $B_\varepsilon (x_o) = \{x \in X : d(x, x_o) < \varepsilon\}$ is the open ball of radius ε and center x_o. The class of λ_d-statistically convergent sequences in the metric space (X, d) will be denoted by $S_{\lambda d}$. If a sequence $x = (x_k)$ in metric space (X, d) is λ_d-*statistically convergent* to the point $x_o \in X$ then we write $x_k \to x_o [S_{\lambda d}]$.

In case $\lambda_n = n$, $S_{\lambda d}$ reduces to the class S_d which is the set of those sequences such that for every $\varepsilon > 0$

$$\lim_{n\to\infty} \frac{1}{n} |\{k \leq n : x_k \notin B_\varepsilon (x_o)\}| = 0$$

for some $x_o \in X$ [9].

Theorem 2.2. Let (X, d) be a metric space, $\lambda = (\lambda_n)$ and $\mu = (\mu_n)$ be two sequences in Λ such that $\lambda_n \leq \mu_n$ for all $n \in \mathbb{N}_{n_o}$. Then in metric space X

(i) Every μ_d-statistically convergent sequence is λ_d-statistically convergent, that is $S_{\mu d} \subseteq S_{\lambda d}$ if

$$\lim_{n\to\infty} \inf \frac{\lambda_n}{\mu_n} > 0. \tag{1}$$

(ii) Every λ_d-statistically convergent sequence is μ_d-statistically convergent, that is $S_{\lambda d} \subseteq S_{\mu d}$ if

$$\lim_{n\to\infty} \frac{\lambda_n}{\mu_n} = 1. \tag{2}$$

In the following proof we will use the techniques given in [2].

Proof. (i) Suppose that $\lambda_n \leq \mu_n$ for all $n \in \mathbb{N}_{n_\circ}$ and let (1) be satisfied. Then $I_n \subset J_n$ and so that for every $\varepsilon > 0$ we may write

$$|\{k \in J_n : x_k \notin B_\varepsilon (x_o)\}| \geq |\{k \in I_n : x_k \notin B_\varepsilon (x_o)\}|$$

and therefore we have

$$\frac{1}{\mu_n} |\{k \in J_n : x_k \notin B_\varepsilon (x_o)\}| \geq \frac{\lambda_n}{\mu_n} \frac{1}{\lambda_n} |\{k \in I_n : x_k \notin B_\varepsilon (x_o)\}|$$

for all $n \in \mathbb{N}_{n_o}$, where $J_n = [n - \mu_n + 1, n]$. Now taking the limit as $n \to \infty$ in the last inequality and using (1) we get $x_k \to x_o\,[S_{\mu d}] \Longrightarrow x_k \to x_o\,[S_{\lambda d}]$ so that $S_{\mu d} \subseteq S_{\lambda d}$.

(ii) Let $(x_k) \in S_{\lambda d}$ and (2) be satisfied. Since $I_n \subset J_n$, for every $\varepsilon > 0$ we may write

$$\begin{aligned}\frac{1}{\mu_n}\left|\{k \in J_n : x_k \notin B_\varepsilon(x_o)\}\right| &= \frac{1}{\mu_n}\left|\{n - \mu_n + 1 \le k \le n - \lambda_n : x_k \notin B_\varepsilon(x_o)\}\right| \\ &\quad + \frac{1}{\mu_n}\left|\{k \in I_n : x_k \notin B_\varepsilon(x_o)\}\right| \\ &\le \frac{\mu_n - \lambda_n}{\mu_n} + \frac{1}{\lambda_n}\left|\{k \in I_n : x_k \notin B_\varepsilon(x_o)\}\right| \\ &\le \left(1 - \frac{\lambda_n}{\mu_n}\right) + \frac{1}{\lambda_n}\left|\{k \in I_n : x_k \notin B_\varepsilon(x_o)\}\right|\end{aligned}$$

for all $n \in \mathbb{N}_{n_o}$.

Since $\lim_n \frac{\lambda_n}{\mu_n} = 1$ by (2) the first term and since $x = (x_k) \in S_{\lambda d}$ the second term of right hand side of above inequality tend to 0 as $n \to \infty$. This implies that $\frac{1}{\mu_n}\left|\{k \in J_n : x_k \notin B_\varepsilon(x_o)\}\right| \to 0$ as $n \to \infty$ and so that $x_k \to x_o\,[S_{\lambda d}] \Longrightarrow x_k \to x_o\,[S_{\mu d}]$. Therefore $S_{\lambda d} \subseteq S_{\mu d}$.

From Theorem 2.2 we have the following result.

Corollary 2.3. Let (X, d) be a metric space, $\lambda = (\lambda_n)$ and $\mu = (\mu_n)$ be two sequences in Λ such that $\lambda_n \le \mu_n$ for all $n \in \mathbb{N}_{n_o}$. If (2) holds then $S_{\lambda d} = S_{\mu d}$.

If we take $\mu = (\mu_n) = (n)$ in Corollary 2.3 we have the following result.

Corollary 2.4. Let (X, d) be a metric space and $\lambda = (\lambda_n) \in \Lambda$. If $\lim_{n\to\infty} \frac{\lambda_n}{n} = 1$ then $S_{\lambda d} = S_d$.

Definition 2.5. Let (X, d) be a metric space and let $\lambda = (\lambda_n) \in \Lambda$ be given. A sequence $x = (x_k)$ in metric space (X, d) is called λ_d-statistically bounded if there exist a point $x \in X$ and a real number $M > 0$ such that

$$\lim_{n\to\infty} \frac{1}{\lambda_n}\left|\{k \in I_n : d(x_k, x) \ge M\}\right| = 0.$$

The set of λ_d-statistically bounded sequences in the metric space (X, d) will be denoted by $BS_{\lambda d}$.

In case $(\lambda_n) = (n)$, λ_d-statistical boundedness reduces to statistical boundedness and the set of statistically bounded sequences will be denoted by BS_d [8].

Theorem 2.6. Any bounded sequence in a metric space (X, d) is λ_d-statistically bounded for each $\lambda \in \Lambda$.

Proof. Assume that $x = (x_k)$ is a bounded sequence in a metric space (X, d) and let $\lambda \in \Lambda$ be any element. Since the sequence (x_k) is bounded then there exist a real number $M > 0$ and a point $x \in X$ such that $d(x_k, x) < M$ for every $k \in \mathbb{N} = \{1, 2, 3, ...\}$. Then

$$\{k \leq n : d(x_k, x) \geq M\} = \emptyset \tag{3}$$

and since the inclusion

$$\{k \in I_n : d(x_k, x) \geq M\} \subset \{k \leq n : d(x_k, x) \geq M\}$$

holds we have $\{k \in I_n : d(x_k, x) \geq M\} = \emptyset$ for each $n \in \mathbb{N}$. Also we have

$$\lim_{n\to\infty} |\{k \leq n : d(x_k, x) \geq M\}| = |\{k \in \mathbb{N} : d(x_k, x) \geq M\}| = 0$$

by (3). Now easily we have

$$\lim_{n\to\infty} \frac{1}{\lambda_n} |\{k \in I_n : d(x_k, x) \geq M\}| = 0$$

since $\lambda_n \to \infty$ as $n \to \infty$. Therefore the sequence (x_k) is λ_d-statistically bounded. This completes the proof.

Remark 2.7. The invers of Theorem 2.6 may not be true. For this let us consider the following example:

Example: Let us take $X = \mathbb{R}$ with usual metric. Then the sequence (x_k) defined by

$$x_k = \begin{cases} k, & k = m^2 \\ (-1)^k, & k \neq m^2 \end{cases}$$

is not bounded. But since the inequality

$$\begin{aligned} \frac{1}{\lambda_n} |\{k \in I_n : |x_k| \geq M\}| &\leq \frac{\sqrt{n} - \sqrt{n - \lambda_n}}{\lambda_n} \\ &= \frac{\lambda_n}{\lambda_n \left(\sqrt{n} + \sqrt{n - \lambda_n}\right)} \\ &= \frac{1}{\sqrt{n} + \sqrt{n - \lambda_n}} \leq \frac{1}{\sqrt{n}} \end{aligned}$$

is satisfied and the right side of this last inequality tends to 0 as $n \to \infty$, we obtain that the sequence (x_k) is λ_d-statistically bounded.

Theorem 2.8. Let (X, d) be a metric space and $\lambda = (\lambda_n), \mu = (\mu_n) \in \Lambda$ be two sequences such that $\lambda_n \leq \mu_n$ for all $n \in \mathbb{N}_{n_o}$.

(i) Suppose that the inequality (1) is satisfied. Then if a sequence $x = (x_k)$ in X is μ_d-statistically convergent, then it is λ_d- statistically bounded that is $S_{\mu d} \subseteq BS_{\lambda d}$.

(ii) Suppose that the equality (2) is satisfied. Then if a sequence $x = (x_k)$ in X is λ_d-statistically bounded then it is μ_d- statistically bounded that is $BS_{\lambda d} \subseteq BS_{\mu d}$.

Proof. (i) Suppose that $\lambda_n \leq \mu_n$ for all $n \in \mathbb{N}_{n_o}$ and let (1) be satisfied. Assume that $x = (x_k)$ is μ_d-statistically convergent to $x_o \in X$. Then $I_n \subset J_n$ and so that for $\varepsilon > 0$ and a large $M > 0$ we may write

$$\{k \in I_n : d(x_k, x_o) \geq M\} \subset \{k \in J_n : x_k \notin B_\varepsilon(x_o)\}.$$

From this inclusion we obtain the inequality

$$|\{k \in J_n : x_k \notin B_\varepsilon(x_o)\}| \geq |\{k \in I_n : d(x_k, x_o) \geq M\}|$$

and therefore we have

$$\frac{1}{\mu_n} |\{k \in J_n : x_k \notin B_\varepsilon(x_o)\}| \geq \frac{\lambda_n}{\mu_n} \frac{1}{\lambda_n} |\{k \in I_n : d(x_k, x_o) \geq M\}|$$

for all $n \in \mathbb{N}_{n_o}$, where $J_n = [n - \mu_n + 1, n]$. Now taking the limit as $n \to \infty$ in the last inequality and using (1) we get $x = (x_k) \in BS_{\lambda d}$ so that $S_{\mu d} \subseteq BS_{\lambda d}$.

(ii) Let $(x_k) \in BS_{\lambda d}$ and (2) be satisfied. Since $I_n \subset J_n$, for a large number $M > 0$ we may write

$$\begin{aligned}
\frac{1}{\mu_n} |\{k \in J_n : d(x_k, x_o) \geq M\}| &= \frac{1}{\mu_n} |\{n - \mu_n + 1 \leq k \leq n - \lambda_n : d(x_k, x_o) \geq M\}| \\
&\quad + \frac{1}{\mu_n} |\{k \in I_n : d(x_k, x_o) \geq M\}| \\
&\leq \frac{\mu_n - \lambda_n}{\mu_n} + \frac{1}{\lambda_n} |\{k \in I_n : d(x_k, x_o) \geq M\}| \\
&\leq \left(1 - \frac{\lambda_n}{\mu_n}\right) + \frac{1}{\lambda_n} |\{k \in I_n : d(x_k, x_o) \geq M\}|
\end{aligned}$$

for all $n \in \mathbb{N}_{n_o}$. Since $\lim_n \frac{\lambda_n}{\mu_n} = 1$ by (2) the first term and since $x = (x_k) \in BS_{\lambda d}$ the second term of the right hand side of above inequality tend to 0 as $n \to \infty$. This implies that

$$\lim_{n \to \infty} \frac{1}{\mu_n} |\{k \in J_n : d(x_k, x_o) \geq M\}| = 0$$

and so that the sequence (x_k) is μ_d-statistically bounded, that is $(x_k) \in BS_{\mu d}$. Therefore since $(x_k) \in BS_{\lambda d}$ is an arbitrary element ve have $BS_{\lambda d} \subseteq BS_{\mu d}$ and this completes the proof.

Theorem 2.9. Let (X, d) be a metric space and $\lambda = (\lambda_n) \in \Lambda$ be given. Then every λ_d-statistically convergent sequence in X is λ_d-statistically bounded, that is $S_{\lambda d} \subseteq BS_{\lambda d}$.

Taking $(\mu_n) = (\lambda_n)$ the proof follows from Theorem 2.8 (i).

Corollary 2.10. Let (X, d) be a metric space and $\lambda = (\lambda_n) \in \Lambda$ be given.

(i) Every statistically convergent sequence is λ_d-statistically bounded, that is $S_d \subseteq BS_{\lambda d}$ if

$$\liminf_{n\to\infty} \frac{\lambda_n}{n} > 0. \tag{4}$$

(ii) Every λ_d-statistically bounded sequence is statistically bounded, that is $BS_{\lambda d} \subseteq BS_d$ if

$$\lim_{n\to\infty} \frac{\lambda_n}{n} = 1. \tag{5}$$

Taking $(\mu_n) = (n)$ the proof follows from Theorem 2.8 (i) and (ii) respectively.

Remark 2.11. The invers of (i) in Theorem 2.8 may not be true. For example if we take $X = \mathbb{R}$ with the usual metric, the sequence (x_k) defined by

$$x_k = \begin{cases} 0, & k = 2m+1 \\ \\ 1, & k = 2m \end{cases}$$

is not μ_d-statistically convergent but it is λ_d-statistically bounded for any $\lambda, \mu \in \Lambda$. Note that in this example we do not need the restriction $\lambda_n \leq \mu_n$ for all $n \in \mathbb{N}_{n_o}$.

3 Strong $(V, \lambda)_d$ −summability in Metric Spaces

$p-$strong summability in metric spaces was studied by Bilalov and Nazarova [1]. In this section we introduce and study strong $(V, \lambda)_d -$ summability and give the relations between the sets of strongly $(V, \lambda)_d -$ summable sequences for various sequences $\lambda = (\lambda_n)$ in Λ^* in a metric space (X, d), where

$$\Lambda^* = \{\lambda = (\lambda_n) : 0 < \lambda_n \leq \lambda_{n+1}, \text{ for every } n \text{ and } \lambda_n \to \infty \ (n \to \infty)\}.$$

Note that in order to obtain the class Λ^* we remove the conditions $\lambda_{n+1} \leq \lambda_n + 1$ and $\lambda_1 = 1$ on the class Λ. It is clear that $\Lambda \subset \Lambda^*$ and the inclusion is strict. For example $\lambda = (\lambda_n) = \left(n^2\right) \in \Lambda^* - \Lambda$.

In this chapter, we use the class Λ^* instead of Λ.

Definition 3.1. Let (X, d) be a metric space and $\lambda = (\lambda_n) \in \Lambda^*$. A sequence $x = (x_k) \subset X$ is said to be *strongly* $(V, \lambda)_d -$*summable* to $x_o \in X$ if

$$\lim_{n\to\infty} \frac{1}{\lambda_n} \sum_{k \in I_n} d(x_k, x_o) = 0.$$

We write

$$[V,\lambda]_d = \left\{ x=(x_k) : \lim_{n\to\infty} \frac{1}{\lambda_n} \sum_{k\in I_n} d\left(x_k, x_o\right) = 0 \text{ for some } x_o \in X \right\}$$

for the set of the sequences which are strongly $(V,\lambda)_d$ −summable in metric space X with the metric d. If a sequence $x=(x_k)$ in metric space (X,d) is strongly $(V,\lambda)_d$ −summable to the point $x_o \in X$ then we write $x_k \to x_o\,[V,\lambda]_d$.

Strong $(V,\lambda)_d$ −summability reduces to strong $(C,1)_d$ -summability in case $\lambda_n = n$ [1].

Theorem 3.2. Let (X,d) be a metric space and $\lambda=(\lambda_n)$, $\mu=(\mu_n) \in \Lambda^*$ and suppose that $\lambda_n \leq \mu_n$ for all $n \in \mathbb{N}_{n_\circ}$.

(i) If (1) holds then a strongly $(V,\mu)_d$ −summable sequence in the metric space X is also strongly $(V,\lambda)_d$ −summable, that is $[V,\mu]_d \subseteq [V,\lambda]_d$,

(ii) Suppose (2) holds. If $x=(x_k) \subset X$ is bounded and $x_k \to x_o\,[V,\lambda]_d$ then $x_k \to x_o\,[V,\mu]_d$.

Proof. (i) Let (X,d) be a metric space and suppose that $\lambda_n \leq \mu_n$ for all $n \in \mathbb{N}_{n_\circ}$. Then $I_n \subseteq J_n$ and so that we may write

$$\frac{1}{\mu_n} \sum_{k\in J_n} d\left(x_k, x_o\right) \geq \frac{1}{\mu_n} \sum_{k\in I_n} d\left(x_k, x_o\right)$$

for all $n \in \mathbb{N}_{n_\circ}$ and hence we may write the inequality

$$\frac{1}{\mu_n} \sum_{k\in J_n} d\left(x_k, x_o\right) \geq \frac{\lambda_n}{\mu_n} \frac{1}{\lambda_n} \sum_{k\in I_n} d\left(x_k, x_o\right).$$

Then taking limit as $n \to \infty$ in the last inequality and using (1) we obtain $x_k \to x_o\,[V,\mu]_d \Longrightarrow x_k \to x_o\,[V,\lambda]_d$. Since $x=(x_k) \in [V,\mu]_d$ is an arbitrary sequence we obtain that $[V,\mu]_d \subseteq [V,\lambda]_d$.

(ii) Let the sequence $x=(x_k) \subset X$ be bounded and $x_k \to x_o\,[V,\lambda]_d$. Suppose (2) holds. Since the sequence $x=(x_k)$ is bounded then there exists some open ball $B_r(x')$ such that $x_k \in B_r(x')$ for all $k \in \mathbb{N}$, $r>0$ and $x' \in X$. Hence we may write

$$d\left(x_k, x_\circ\right) \leq d\left(x_k, x'\right) + d\left(x', x_\circ\right) < r + d\left(x', x_\circ\right) = M.$$

Now since $\lambda_n \leq \mu_n$ and so that $\frac{1}{\mu_n} \leq \frac{1}{\lambda_n}$, and $I_n \subset J_n$ for all $n \in \mathbb{N}_{n_\circ}$, we may write

$$\frac{1}{\mu_n}\sum_{k\in J_n} d\left(x_k, x_\circ\right) = \frac{1}{\mu_n}\sum_{k\in J_n - I_n} d\left(x_k, x_\circ\right) + \frac{1}{\mu_n}\sum_{k\in I_n} d\left(x_k, x_\circ\right)$$
$$\leq \frac{\mu_n - \lambda_n}{\mu_n} M + \frac{1}{\mu_n}\sum_{k\in I_n} d\left(x_k, x_\circ\right)$$
$$\leq \left(1 - \frac{\lambda_n}{\mu_n}\right) M + \frac{1}{\lambda_n}\sum_{k\in I_n} d\left(x_k, x_\circ\right)$$

for every $n \in \mathbb{N}_{n_\circ}$. Since $\lim_n \frac{\lambda_n}{\mu_n} = 1$ by (2) the first term and since $x_k \to x_\circ$ $[V,\lambda]_d$ the second term of rigt hand side of above inequality tend to 0 as $n \to \infty$. Hence we get $x_k \to x_\circ\, [V,\lambda]_d \Longrightarrow x_k \to x_\circ\, [V,\mu]_d$.

If we take discrete metric instead of any metric in Theorem 3.2 we have the following result.

Corollary 3.3. Let (X,d) be a metric space with discrete metric, $\lambda = (\lambda_n)$ and $\mu = (\mu_n)$ be two sequences in Λ^* such that $\lambda_n \leq \mu_n$ for all $n \in \mathbb{N}_{n_\circ}$. If (2) holds then $[V,\lambda]_d = [V,\mu]_d$.

Proof. If (2) holds then $\lim_{n\to\infty} \frac{\lambda_n}{\mu_n} = 1 > 0$ so that (1) holds, too. Then from Theorem 3.2 (i) we have $[V,\mu]_d \subseteq [V,\lambda]_d$. Since any sequence in discrete metric space is bounded then any sequence in $[V,\lambda]_d$ is bounded and using (2) from Theorem 3.2 (ii) we get $[V,\lambda]_d \subseteq [V,\mu]_d$. Both inclusions give the equality $[V,\lambda]_d = [V,\mu]_d$.

4 Relations Between λ_d−statistical Convergence and Strong $(V,\lambda)_d$−summability in Metric Spaces

In this section we give the relations between the sets of λ_d−statistically convergent sequences and the sets of strongly $(V,\lambda)_d$−summable sequences for various sequences $\lambda = (\lambda_n)$ belong to Λ in metric spaces.

Theorem 4.1. Let (X,d) be a metric space and $\lambda = (\lambda_n) \in \Lambda$. Then
(i) $x_k \to x_o\, [V,\lambda]_d \Longrightarrow x_k \to x_o\, [S_{\lambda d}]$.
(ii) If (x_k) is bounded and $x_k \to x_o\, [S_{\lambda d}]$ then $x_k \to x_o\, [V,\lambda]_d$.

Proof. (i) Let $\varepsilon > 0$ and $x_k \to x_o\, [V,\lambda]_d$. We may write

$$\sum_{k\in I_n} d\left(x_k, x_o\right) \geq \sum_{\substack{k\in I_n \\ d(x_k,x_o)\geq\varepsilon}} d\left(x_k, x_o\right) \geq \varepsilon.\left|\left\{k \in I_n : x_k \notin B_\varepsilon\left(x_o\right)\right\}\right|$$

and so that

$$\frac{1}{\lambda_n}\sum_{k\in I_n} d\left(x_k, x_o\right) \geq \frac{1}{\lambda_n}\sum_{\substack{k\in I_n \\ d(x_k,x_o)\geq\varepsilon}} d\left(x_k, x_o\right) \geq \frac{1}{\lambda_n}\left|\left\{k \in I_n : x_k \notin B_\varepsilon\left(x_o\right)\right\}\right|.\varepsilon.$$

Hence we obtain that $x_k \to x_o\, [S_{\lambda d}]$ whenever $x_k \to x_o\, [V,\lambda]_d$.

(ii) Let (x_k) be a bounded sequence and $x_k \to x_o\,[S_{\lambda d}]$ in metric space (X,d). Then there is an open ball $B_r(x') \subset X$ such that $x_k \in B_r(x')$ for every $k \in \mathbb{N}$ since (x_k) is bounded, where $r>0$ and $x' \in X$.

Now we may write

$$d(x_k,x_o) \le d(x_k,x') + d(x',x_o) < r + d(x',x_o) = M$$

and since $x_k \to x_o\,[S_{\lambda d}]$ for every $\varepsilon > 0$ we have

$$\lim_{n\to\infty} \frac{1}{\lambda_n} |\{k \in I_n : x_k \notin B_\varepsilon(x_o)\}| = 0.$$

Thus we obtain

$$\begin{aligned} \frac{1}{\lambda_n}\sum_{k\in I_n} d(x_k,x_o) &= \frac{1}{\lambda_n} \sum_{\substack{k\in I_n \\ d(x_k,x_\circ)\ge\varepsilon}} d(x_k,x_o) + \frac{1}{\lambda_n} \sum_{\substack{k\in I_n \\ d(x_k,x_o)<\varepsilon}} d(x_k,x_o) \\ &< \frac{M}{\lambda_n} |\{k \in I_n : x_k \notin B_\varepsilon(x_o)\}| + \varepsilon. \end{aligned}$$

This means that $x_k \to x_o\,[V,\lambda]_d$.

Theorem 4.2. Let (X,d) be a metric space and $\lambda = (\lambda_n)$, $\mu = (\mu_n) \in \Lambda$ such that $\lambda_n \le \mu_n$ for all $n \in \mathbb{N}_{n_\circ}$.

(i) If (1) holds then

$$x_k \to x_\circ\,[V,\mu]_d \Longrightarrow x_k \to x_\circ\,[S_{\lambda d}]$$

and the inclusion $[V,\mu]_d \subset S_{\lambda d}$ is strict for some $\lambda,\mu \in \Lambda$,

(ii) If (x_k) is bounded and $x_k \to x_\circ\,[S_{\lambda d}]$ then $x_k \to x_\circ\,[V,\mu]_d$, whenever (2) holds.

Proof. (i) Let $\varepsilon > 0$ be given and let $x_k \to x_\circ\,[V,\mu]_d$. Then for every $\varepsilon > 0$ we may write

$$\sum_{k\in J_n} d(x_k,x_\circ) \ge \sum_{k\in I_n} d(x_k,x_\circ) \ge \sum_{\substack{k\in I_n \\ d(x_k,x_\circ)\ge\varepsilon}} d(x_k,x_\circ) \ge \varepsilon.\,|\{k \in I_n : d(x_k,x_\circ) \ge \varepsilon\}|$$

and so that

$$\frac{1}{\mu_n}\sum_{k\in J_n} d(x_k,x_\circ) \ge \frac{\lambda_n}{\mu_n}\frac{1}{\lambda_n} |\{k \in I_n : d(x_k,x_\circ) \ge \varepsilon\}|\,.\varepsilon$$

for all $n \in \mathbb{N}_{n_\circ}$. Then taking limit as $n \to \infty$ in the last inequality and using (1) we obtain that $x_k \to x_\circ\,[V,\mu]_d \Longrightarrow x_k \to x_\circ\,[S_{\lambda d}]$. Since $x = (x_k) \in [V,\mu]_d$ is an arbitrary sequence we obtain $[V,\mu]_d \subseteq S_{\lambda d}$.

To show that the inclusion $[V,\mu]_d \subset S_{\lambda d}$ is strict for some $\lambda, \mu \in \Lambda$ we take $X = \mathbb{R}$, $d(x,y) = |x-y|$ and $\lambda_n = \frac{n+1}{2}$, $\mu_n = n$ for all $n \in \mathbb{N}$. Then $\lim_n \frac{\lambda_n}{\mu_n} = \frac{1}{2} > 0$ and hence $[V,\mu]_d \subseteq S_{\lambda d}$.

Define $x = (x_k)$ as

$$x_k = \begin{cases} \frac{1}{k}, & k \neq m^3 \\ \\ k, & k = m^3 \end{cases}.$$

Let $\varepsilon > 0$ be given. Then there exists $k_\circ \in \mathbb{N}$ such that $|x_k| < \varepsilon$ for all $k > k_\circ$ and $k \neq m^3$. Now since

$$\begin{aligned} \frac{1}{\lambda_n} \left| \{k \in I_n : |x_k| \geq \varepsilon\} \right| &\leq \frac{1}{\lambda_n} \left(k_\circ + \sqrt[3]{n} - \sqrt[3]{\frac{n-1}{2}} \right) \\ &= \frac{2}{n+1} \left(k_\circ + \sqrt[3]{n} - \sqrt[3]{\frac{n-1}{2}} \right) \to 0 \end{aligned}$$

as $n \to \infty$ we have $x_k \to 0\,[S_\lambda]\,(\mathbb{R})$. On the other hand we know that the equality

$$1 + 2^3 + 3^3 + 4^3 + ... + n^3 = \frac{n^2 (n+1)^2}{4}$$

is satisfied for every $n \in \mathbb{N}$. Considering this equality, since $\sqrt[3]{n} < [\sqrt[3]{n}] + 1$ and so that $\frac{1}{n} > \frac{1}{([\sqrt[3]{n}]+1)^3}$ we have

$$\begin{aligned} \frac{1}{\mu_n} \sum_{k \in J_n} |x_k| &= \frac{1}{n} \sum_{k=1}^{n} x_k = \frac{1}{n} \sum_{\substack{k=1 \\ k=m^3}}^{n} x_k + \frac{1}{n} \sum_{\substack{k=1 \\ k \neq m^3}}^{n} x_k > \frac{1}{n} \sum_{\substack{k=1 \\ k=m^3}}^{n} x_k = \frac{1}{n} \sum_{\substack{k=1 \\ k=m^3}}^{n} k \\ &= \frac{1}{n} \left(1 + 2^3 + 3^3 + 4^3 + ... + \left[\sqrt[3]{n}\right]^3 \right) \\ &= \frac{\left[\sqrt[3]{n}\right]^2 \left(\left[\sqrt[3]{n}\right] + 1\right)^2}{4n} > \frac{\left[\sqrt[3]{n}\right]^2 \left(\left[\sqrt[3]{n}\right] - 1\right)^2}{4 \left(\left[\sqrt[3]{n}\right] + 1\right)^3} \to \infty \ \ (n \to \infty). \end{aligned}$$

Therefore $x = (x_k) \notin [V,\mu]\,(\mathbb{R})$. Thus the inclusion $[V,\mu]_d \subset S_{\lambda d}$ is strict.

(*ii*) Suppose that $x_k \to x_\circ\,[S_{\lambda d}]$ and $x = (x_k)$ is bounded. Then there exist a number $r > 0$ and $x' \in X$ such that $x_k \in B_r(x')$ for all $k \in \mathbb{N}$. Hence we may write

$$d(x_k, x_\circ) \leq d(x_k, x') + d(x', x_\circ) < r + d(x', x_\circ) = M.$$

Also since $\frac{1}{\mu_n} \leq \frac{1}{\lambda_n}$ then for every $\varepsilon > 0$ we may write

$$\begin{aligned}\frac{1}{\mu_n}\sum_{k\in J_n} d(x_k, x_\circ) &= \frac{1}{\mu_n}\sum_{k\in J_n - I_n} d(x_k, x_\circ) + \frac{1}{\mu_n}\sum_{k\in I_n} d(x_k, x_\circ) \\ &\leq \frac{\mu_n - \lambda_n}{\mu_n} M + \frac{1}{\mu_n}\sum_{k\in I_n} d(x_k, x_\circ) \\ &\leq \left(1 - \frac{\lambda_n}{\mu_n}\right) M + \frac{1}{\lambda_n}\sum_{k\in I_n} d(x_k, x_\circ).\end{aligned}$$

(ii) Suppose that $x_k \to x_\circ [S_{\lambda d}]$ and $x = (x_k)$ is bounded. Then there exist a number $r > 0$ and $x' \in X$ such that $x_k \in B_r(x')$ for all $k \in \mathbb{N}$. Hence we may write

$$d(x_k, x_\circ) \leq d(x_k, x') + d(x', x_\circ) < r + d(x', x_\circ) = M.$$

Also since $\frac{1}{\mu_n} \leq \frac{1}{\lambda_n}$ then for every $\varepsilon > 0$ we may write

$$\begin{aligned}\frac{1}{\mu_n}\sum_{k\in J_n} d(x_k, x_\circ) &\leq \left(1 - \frac{\lambda_n}{\mu_n}\right) M + \frac{1}{\lambda_n}\sum_{\substack{k\in I_n \\ d(x_k, x_\circ)\geq\varepsilon}} d(x_k, x_\circ) \\ &\quad + \frac{1}{\lambda_n}\sum_{\substack{k\in I_n \\ d(x_k, x_\circ)<\varepsilon}} d(x_k, x_\circ) \\ &\leq \left(1 - \frac{\lambda_n}{\mu_n}\right) M + \frac{M}{\lambda_n}\left|\{k \in I_n : d(x_k, x_\circ) \geq \varepsilon\}\right| + \varepsilon\end{aligned}$$

for all $n \in \mathbb{N}_{n_\circ}$. Using (2) we obtain that $x_k \to x_\circ [V, \mu]_d$ whenever $x_k \to x_\circ [S_{\lambda d}]$.

Corollary 4.3. If $\lim_{n\to\infty} \inf \frac{\lambda_n}{\mu_n} > 0$ then $S_{\mu d} \cap [V, \mu]_d \subset S_{\lambda d}$.

If we take $\mu_n = n$ for all n in Theorem 4.2 then we have the following results. Because $\lim_{n\to\infty} \frac{\lambda_n}{\mu_n} = 1$ implies that $\lim_{n\to\infty} \inf \frac{\lambda_n}{\mu_n} = 1 > 0$, that is (2) $\Longrightarrow$ (1).

Corollary 4.4. If $\lim_{n\to\infty} \frac{\lambda_n}{n} = 1$ then
(i) If (x_k) is bounded and $x_k \to x_\circ [S_{\lambda d}]$ then $x_k \to x_\circ [C, 1] (X_d)$,
(ii) If $x_k \to x_\circ [C, 1] (X_d)$ then $x_k \to x_\circ [S_{\lambda d}]$.

Remark 4.5. Let (X, d) be a metric space, $\lambda = (\lambda_n) \in \Lambda^*$ and $0 < p < \infty$. Define

$$[V, \lambda]_{dp} = \left\{ x = (x_k) : \lim_{n\to\infty} \frac{1}{\lambda_n}\sum_{k\in I_n} [d(x_k, x_\circ)]^p = 0 \text{ for some } x_o \in X \right\}.$$

Then Theorem 4.2 is satisfied for $[V, \lambda]_{dp}$ and $[V, \mu]_{dp}$, if we take $[V, \lambda]_{dp}$ instead of $[V, \lambda]_d$ and $[V, \mu]_{dp}$ instead of $[V, \mu]_d$.

5 Conclusion

We have introduced and studied λ_d-statistical convergence, λ_d-statistical boundedness and strong $(V,\lambda)_d-$summability for a sequence in a metric space (X,d). Furthermore we have established some inclusion relations between the sets $S_{\lambda d}$ and $S_{\mu d}$, between the sets $BS_{\lambda d}$ and $BS_{\mu d}$, between the sets $[V,\lambda]_d$ and $[V,\mu]_d$ and between the sets $S_{\lambda d}$ and $[V,\mu]_d$ under some conditions for $\lambda,\mu\in\Lambda$ in a metric space (X,d).

References

1. Bilalov, B., Nazarova, T.: On statistical convergence in metric spaces. J. Math. Res. **7**(1), 37–43 (2015)
2. Çolak, R.: On λ-statistical convergence. In: Conference on Summability and Applications, Commerce University, 12–13 May, Istanbul, Turkey (2011)
3. Çolak, R., Bektaş, Ç.A.: Lambda-statistical convergence of order alpha. Acta Math. Sci. Ser. B Engl. Ed. **31**(3), 953–959 (2011)
4. Connor, J.S.: The statistical and strong p-Cesaro convergence of sequences. Analysis **8**, 47–63 (1988)
5. Fast, H.: Sur la convergence statistique. Colloq. Math. **2**, 241–244 (1951)
6. Fridy, J.: On statistical convergence. Analysis **5**, 301–313 (1985)
7. Kolk, E.: The statistical convergence in Banach spaces. Acta Comment. Univ. Tartu **928**, 41–52 (1991)
8. Küçükarslan, M., Değer, U.: On statistical boundedness of metric valued sequences. Eur. J. Pure Appl. Math. **5**, 174–186 (2012)
9. Küçükarslan, M., Değer, U., Dovgoshey, O.: On Statistical Convergence of Metric-Valued Sequences. Ukr. Math. J. **66**, 796–805 (2014)
10. Leindler, L.: Über die de la Vallée-Pousinsche summierbarkeit allgemeiner orthogonalreihen. Acta Math. Acad. Sci. Hungar. **16**, 375–387 (1965)
11. Mursaleen, M.: $\lambda-$statistical convergence. Math. Slovaca **50**(1), 111–115 (2000)
12. Šalát, T.: On statistically convergent sequences of real numbers. Math. Slovaca **30**, 139–150 (1980)
13. Savaş, E.: Strong almost convergence and almost λ-statistical convergence. Hokkaido Math. Jour. **29**(3), 531–536 (2000)
14. Schoenberg, I.J.: The integrability of certain functions and related summability methods. Amer. Math. Monthly **66**, 361–375 (1959)
15. Steinhaus, H.: Sur la convergence ordinaire et la convergence asymptotique. Colloq. Math. **2**, 73–74 (1951)

On Γ-rings with Permuting Skew Tri-derivations

Kalyan Kumar Dey[1(✉)], Akhil Chandra Paul[1], and Bijan Davvaz[2]

[1] Department of Mathematics, Rajshahi University, Rajshahi 6205, Bangladesh
kkdmath@yahoo.com, acpaulrubd_math@yahoo.com
[2] Department of Mathematics, Yazd University, Yazd, Iran
davvaz@yazd.ac.ir, bdavvaz@yahoo.com

Abstract. The objective of this paper is to introduce the notion of a permuting skew tri-derivation on prime and semiprime Γ-rings. We prove that under certain conditions a prime Γ-ring is to be commutative by means of a nonzero permuting skew tri-derivation.

Keywords: Prime Γ-ring · Semiprime Γ-ring · Permuting skew tri-derivation · Centralizing mapping · Commuting mapping

1 Introduction

Nobuswa [10] was first introduced the concept of a Γ-ring as a generalization of a ring and then Barnes [4] has generalized the notion of a Γ-ring in the sense of Nobuswa. The study of a Γ-ring is of great interest of modern algebraists, especially for extending the significant results in classical ring theory to the topics in Γ-rings theory. In [1–3,9,11,12,14,15], permuting tri derivations in prime and semiprime Γ-rings have been studied by Ozden, Ozturk and Jun, Ozturk, Jun and Kim [13] studied orthogonal traces on semiprime Γ-rings and they obtained some conditions in order that the traces must be orthogonal. Afterwards, Dey and Paul [7] worked on the trace of a permuting tri-additive mappings in Left s-unital Γ-rings and proved the commuting conditions of a Γ-ring. In [5], Dey and Paul and Rakhimov proved the significant results related to permuting Tri-derivations of Γ-rings. Dey and Paul [6] studied and investigated some results concerning a permuting tri-derivation on a non-commutative 3-torsion free semiprime Γ-ring. They obtained some characterizations of these Γ-rings with the help of permuting Tri-derivations. The notion of symmetric skew 3-derivation of classical rings has been introduced by Fosner in [8] and he obtained commutativity conditions on prime and semiprime rings with a non-zero symmetric skew 3-derivations.

By the motivation of the works of Fosner [8], we introduce the concept of permuting skew tri-derivation on Γ-rings. In this paper, we develop the commutativity conditions on prime and semiprime Γ-rings by using the notion of a non-zero permuting skew tri-derivation.

© Springer Nature Singapore Pte Ltd. 2017
D. Giri et al. (Eds.): ICMC 2017, CCIS 655, pp. 404–411, 2017.
DOI: 10.1007/978-981-10-4642-1_34

2 Preliminaries

Let M and Γ be additive abelian groups. If there exists an additive mapping $M \times \Gamma \times M \to M$ with $(a, \alpha, y) \mapsto x\alpha y$, which satisfies the condition $(x\alpha y)\beta z = x\alpha(y\beta z)$ for all $x, y, z \in M$ and $\alpha, \beta \in \Gamma$, then M is called a Γ-ring in the sense of Barnes [4]. Throughout this paper M denotes a Γ-ring with center $Z(M)$. For any $x, y \in M$ and $\alpha \in \Gamma$, the symbol $[x, y]_\alpha$ will denote the commutator $x\alpha y - y\alpha x$. A Γ-ring M is called commutative if $[x, y]_\alpha = 0$ for all $x, y \in M$ and $\alpha \in \Gamma$. We know that

$$[x\beta y, z]_\alpha = [x, z]_\alpha \beta y + x\beta[y, z]_\alpha + x[\beta, \alpha]_z y,$$
$$[x, y\beta z]_\alpha = y\beta[x, z]_\alpha + [x, y]_\alpha \beta z + y[\beta, \alpha]_x z.$$

We take an assumption

$$x\beta z\alpha y = x\alpha z\beta y, \tag{1}$$

for all $x, y, z \in M$ and $\alpha, \beta \in \Gamma$.

Using the assumption the basic commutator identities reduce to

$$[x\beta y, z]_\alpha = [x, z]_\alpha \beta y + x\beta[y, z]_\alpha,$$
$$[x, y\beta z]_\alpha = y\beta[x, z]_\alpha + [x, y]_\alpha \beta z.$$

Recall that M is prime if $x\Gamma M\Gamma y = \{0\}$ implies that $x = 0$ or $y = 0$ and M is semiprime if $x\Gamma M\Gamma x = \{0\}$ implies that $x = 0$. Let ≥ 2 be an integer. A Γ-ring M is said to be n-torsion free if for $x \in M$, $nx = 0$ implies $x = 0$. An additive map $d : M \to M$ is called a derivation of M if $d(x\alpha y) = d(x)\alpha y + x\alpha d(y)$ for all $x, y \in M$ and $\alpha \in \Gamma$, and it is called a skew derivation of M associated with the automorphism σ if $d(x\alpha y) = d(x)\alpha y + \sigma(x)\alpha d(y)$ for all $x, y \in M$ and $\alpha \in \Gamma$. Of course, skew derivations are one of the natural generalizations of usual derivations ($\sigma = 1_M$, where 1_M denotes the identity map on M). A map $f : M \to M$ is said to be centralizing on M if $[f(x), x]_\alpha \in Z(M)$ for all $x \in M$ and $\alpha \in \Gamma$. In a special case, when $[f(x), x]_\alpha = 0$ holds for all $x \in M$ and $\alpha \in \Gamma$, a map f is said to be commuting on M. By a bi-derivation we mean a bi-additive map $D : M \times M \to M$ (i.e., D is additive in both arguments), which satisfies the relations $D(x\alpha y, z) = D(x, z)\alpha y + x\alpha D(y, z)$ and $D(x, y\beta z) = D(x, y)\beta z + y\beta D(x, z)$ for all $x, y \in M$ and $\alpha, \beta \in \Gamma$. Let D be symmetric, that is $D(x, y) = D(y, x)$ for the $x, y \in M$. The map $d : M \to M$ defined by $d(x) = D(x, x)$ for all $x \in M$ is called the trace of D. A map $D : M \times M \times M \to M$ will be said to be permuting if the equation $D(x, y, z) = D(x, z, y) = D(z, x, y) = D(y, z, x) = D(z, y, x)$ for all $x, y, z \in M$. A map $d : M \to M$ defined by $d(x) = D(x, x, x)$ for all $x \in M$, where $D : M \times M \times M \to M$ is a permuting map, is called the trace of D. It is obvious that, in case when $D : M \times M \times M \to M$ is a permuting map which is also tri-additive (i.e., additive in each argument), the trace d of D satisfies the relation $d(x + y) = d(x) + d(y) + 3D(x, x, y) + 3D(x, y, y)$ for all $x, y \in M$. Since we have $D(0, y, z) = D(0 + 0, y, z) = D(0, y, z) + D(0, y, z)$ for all $y, z \in M$, we obtain $D(0, y, z) = 0$ for all $y, z \in M$. Hence, we get $D(0, y, z) = D(xx, y, z) = D(x, y, z) + D(x, y, z) = 0$ and so we see that $D(x, y, z) = D(x, y, z)$ for all

$x, y, z \in M$. This tells us that d is an odd function. A tri-additive map $D : M \times M \times M \to M$ will be called a tri-derivation if the relations $D(x\alpha w, y, z) = D(x, y, z)\alpha w + x\alpha D(w, y, z)$, $D(x, y\alpha w, z) = D(x, y, z)\alpha w + y\alpha D(x, w, z)$ and $D(x, y, z\alpha w) = D(x, y, z)\alpha w + z\alpha D(x, y, w)$ are fulfilled for all $x, y, z, w \in M$ and $\alpha \in \Gamma$. If D is permuting, then the above three relations are equivalent to each other.

A tri-additive map $D : M \times M \times M \to M$ is a skew tri-derivation associated with the automorphism σ if

(1) for all $y, z \in M$, the map $x \mapsto D(x, y, z)$ is a skew derivation of M associated with the automorphism σ,
(2) for all $x, z \in M$, the map $y \mapsto D(x, y, z)$ is a skew derivation of M associated with the automorphism σ,
(3) for all $x, y \in M$, the map $z \mapsto D(x, y, z)$ is a skew derivation of M associated with the automorphism σ.

More precisely, for all $x, y, z, u, v, w \in M$ and $\alpha \in \Gamma$, we have

$$\begin{aligned} D(x\alpha u, y, z) &= D(x, y, z)\alpha u + \sigma(x)\alpha D(u, y, z), \\ D(x, y\alpha v, z) &= D(x, y, z)\alpha v + \sigma(y)\alpha D(x, v, z), \\ D(x, y, z\alpha w) &= D(x, y, z)\alpha w + \sigma(z)\alpha D(x, y, w). \end{aligned}$$

Of course, if D is symmetric, then the above three relations are equivalent to each other.

3 Skew Tri-derivations on Γ-rings

For proving our main results, we begin with the following lemma.

Lemma 1. *Let M be a prime Γ-ring satisfying the condition (1) and let $x, y \in M$. If $x\beta[u, y]_\alpha = 0$ for all $u \in M$ and $\alpha, \beta \in \Gamma$, then either $x = 0$ or $y \in Z(M)$.*

Proof. We have $0 = x\beta[u\gamma v, y]_\alpha = x\beta u\gamma[v, y]_\alpha + x\beta[u, y]_\alpha\gamma v = x\beta u\gamma[v, y]_\alpha$ for all $u, v \in M$ and $\alpha, \beta, \gamma \in \Gamma$. Thus, $x\Gamma M\Gamma[v, y]_\alpha = 0$ for all $v \in M$ and $\alpha \in \Gamma$. By the primeness of M, we obtain that either $x = 0$ or $[v, y]_\alpha = 0$ for all $x, y, v \in M$ and $\alpha \in \Gamma$. For the latter case, $y \in Z(M)$. Hence, we find that either $x = 0$ or $y \in Z(M)$.

Theorem 1. *Let M be a 3!-torsion free prime Γ-ring satisfying the condition (1). Let $I \neq 0$ be an ideal of M, σ be an automorphism of M, and $D : M \times M \times M \to M$ be a permuting skew tri-derivation having trace d associated with the automorphism σ. Assume that*

$$[d(x), (x)]_\alpha = 0, \tag{2}$$

for all $x \in I$ and $\alpha \in \Gamma$. Then, $D = 0$.

Proof. Putting $x+y$ instead of x in (2), we obtain $0=[d(x+y),(x+y)]_\alpha=0$ for all $x\in I$ and $\alpha\in\Gamma$. This implies that

$$\begin{aligned}0&=[d(x)+d(y)+3D(x,y,y)+3D(y,x,x),\sigma(x)+\sigma(y)]_\alpha\\&=[d(x),\sigma(x)]_\alpha+[d(y),\sigma(x)]_\alpha+3[D(x,y,y),\sigma(x)]_\alpha\\&\quad+3[D(y,x,x),\sigma(x)]_\alpha+[d(x),\sigma(y)]_\alpha+[d(y),\sigma(y)]_\alpha\\&\quad+3[D(x,y,y),\sigma(y)]_\alpha+3[D(y,x,x),\sigma(y)]_\alpha\\&=[d(x),\sigma(y)]_\alpha+[d(y),\sigma(x)]_\alpha+3[D(x,y,y),\sigma(x)]_\alpha\\&\quad+3[D(y,x,x),\sigma(x)]_\alpha+3[D(x,y,y),\sigma(y)]_\alpha+3[D(y,x,x),\sigma(y)]_\alpha,\end{aligned}\tag{3}$$

for all $x,y\in I$ and α in Γ. Putting $-x$ instead of x in (3) and we obtain

$$\begin{aligned}0&=-[d(x),\sigma(y)]_\alpha-[d(y),\sigma(x)]_\alpha+3[D(x,y,y),\sigma(x)]_\alpha\\&\quad-3[D(y,x,x),\sigma(x)]_\alpha-3[D(x,y,y),\sigma(y)]_\alpha+3[D(y,x,x),\sigma(y)]_\alpha,\end{aligned}\tag{4}$$

for all $x,y\in I$ and $\alpha\in\Gamma$. Comparing (3) and (4) and using 3!-torsion freeness, we obtain

$$[D(x,y,y),(x)]_\alpha+[D(y,x,x),\sigma(y)]_\alpha=0,\tag{5}$$

for all $x,y\in I$ and $\alpha\in\Gamma$. Replacing x by $x+y$ in (5), we find that

$$[D(x+y,y,y),(x+y)]_\alpha+[D(y,x+y,x+y),\sigma(y)]_\alpha=0,$$

for all $x,y\in I$ and $\alpha\in\Gamma$. This yields that

$$\begin{aligned}0&=[D(x,y,y),\sigma(x)]_\alpha+[D(y,y,y),\sigma(x)]_\alpha+[D(x,y,y),\sigma(y)]_\alpha\\&\quad+[D(y,x,x),\sigma(y)]_\alpha+[D(y,y,y),\sigma(x)]_\alpha+[D(x,y,y),\sigma(y)]_\alpha,\end{aligned}$$

for all $x,y\in I$ and $\alpha\in\Gamma$. This gives that

$$\begin{aligned}0&=[d(y),\sigma(x)]_\alpha+[d(y),\sigma(y)]_\alpha+([D(x,y,y),\sigma(x)]_\alpha+[D(y,x,x),\sigma(y)]_\alpha)\\&\quad+[D(x,y,y),\sigma(y)]_\alpha+[D(x,y,y),\sigma(y)]_\alpha+[D(x,y,y),\sigma(y)]_\alpha,\end{aligned}$$

for all $x,y\in I$ and $\alpha\in\Gamma$. In view (2) and (5), the above expression becomes

$$[d(y),\sigma(x)]_\alpha+3[D(x,y,y),\sigma(y)]_\alpha=0,\tag{6}$$

for all $x,y\in I$ and $\alpha\in\Gamma$. Replacing $y\beta x$ by x in (6), we get

$$\begin{aligned}0&=[d(y),(y\beta x)]_\alpha+3[D(y\beta x,y,y),\sigma(y)]_\alpha\\&=\sigma(y)[d(y),\sigma(x)]_\alpha+3[d(y)+\sigma(y)D(x,y,y),\sigma(y)]_\alpha\\&=\sigma(y)[d(y),\sigma(x)]_\alpha+3[D(x,y,y),\sigma(y)]_\alpha+3d(y)[x,\sigma(y)]_\alpha.\end{aligned}$$

By using (6) and the 3-torsion freeness of M, we obtain $d(y)[x,\sigma(y)]_\alpha=0$, for all $x,y\in I$ and $\alpha\in\Gamma$. In view of Lemma 1, it follows that $d(y)=0$ or $[x,\sigma(y)]_\alpha=0$. If $[x,\sigma(y)]_\alpha\neq 0$, then $y\in I\setminus Z(M)$. Hence, we have $d(y)=0$, for all $y\in I\setminus Z(M)$. Moreover, we will show that this is true for all $y\in I$. So, let $x\in I\cap Z(M)$ and $y\in I\setminus Z(M)$. Then, clearly we have seen that $x+y,x,-y\in I\setminus Z(M)$. Then, we obtain

$$\begin{aligned}0&=d(x+y)=d(x)+d(y)+3D(x,y,y)+3D(y,x,x)\\&=d(x)+3D(x,y,y)+3D(y,x,x).\end{aligned}\tag{7}$$

On the other hand,

$$\begin{aligned} 0 = d(x-y) &= d(x) - d(y) + 3D(x,y,y) - 3D(y,x,x) \\ &= d(x) + 3D(x,y,y) - 3D(y,x,x). \end{aligned} \tag{8}$$

Comparing (7) and (8), using 2-torsion freeness of M, we have

$$d(x) + 3D(x,y,y) = 0. \tag{9}$$

Writing $y+y$ instead of y in (9), we have

$$\begin{aligned} 0 &= d(x) + 3D(x, y+y, y+y) \\ &= d(x) + 3D(x,y,y) + 3D(x,y,y) + 3D(x,y,y) + 3D(x,y,y). \end{aligned}$$

In view of (9) and the 3!-torsion freeness of M we arrive at $D(x,y,y) = 0$. So, (9) reduces to $d(x) = 0$ for all $x \in I$. Now, suppose that $x, y \in I$. Then,

$$\begin{aligned} 0 = d(x+y) &= d(x) + d(y) + 3D(x,y,y) + 3D(y,x,x) \\ &= 3D(x,y,y) + 3D(y,x,x). \end{aligned}$$

Therefore, by using 3-torsion freeness of M, we obtain

$$D(x,y,y) + D(y,x,x) = 0. \tag{10}$$

Replacing y by $y+z$ in (10), we obtain $D(x, y+z, y+z) + D(y+z, x, x) = 0$. This implies that

$$\begin{aligned} &D(x,y,y) + D(y,x,x) + D(x,y,z) + D(x,z,y) + D(x,z,z) + D(z,x,x) = 0, \\ &D(x,y,y) + D(y,x,x) + D(x,z,z) + D(z,x,x) + D(x,y,z) + D(x,y,z) = 0. \end{aligned}$$

By using (10) and using the 2-torsion freeness of M, we arrive at

$$D(x,y,z) = 0, \tag{11}$$

for all $x, y, z \in I$. Now, we have to prove that the relation (11) holds for all $x, y, z \in M$. Let $x, y, z \in I$ and $a \in M$. Then, $a\alpha x \in I$, for all $\alpha \in \Gamma$. Now, from (11), we have

$$0 = D(a\alpha x, y, z) = D(a,y,z)\alpha x + \sigma(a)\alpha D(x,y,z) = D(a,y,z)\alpha x.$$

Thus, we have $D(a,y,z)\Gamma I = 0$ and, since M is prime and $I \neq 0$, it follows that $D(a,y,z) = 0$. Replacing y by $b\beta y$, where $b \in M$ and $\beta \in \Gamma$, we get

$$0 = D(a, b\beta y, z) = D(a,b,z)\beta y + \sigma(b)\beta D(a,y,z) = D(a,b,z)\beta y.$$

It follows that $D(a,b,z)\Gamma I = 0$ and since M is prime, we have $D(a,b,z) = 0$. Finally, writing $c\delta z$ instead of z, where $z \in M$ and $\delta \in \Gamma$, we get

$$0 = D(a,b,c\delta z) = D(a,b,c)\delta z + \sigma(c)\delta D(a,b,z) = D(a,b,c)\delta z.$$

Hence, $D(a,b,c)\Gamma I = 0$ and again, using the primeness of M, it follows that $D(a,b,c) = 0$ for all $a, b, c \in M$. This completes the proof.

Theorem 2. *Let M be a 3-torsion free semiprime Γ-ring satisfying the condition (1), $I \neq 0$ be an ideal of M, σ be an automorphism of M and $D : M \times M \times M \to M$ be a permuting skew tri-derivation having the trace d associated with the automorphism σ. Assume that the trace function d is commuting on I and $[d(x), \sigma(x)]_\alpha \in Z(M)$ for all $x \in I$ and $\alpha \in \Gamma$. Then, $[d(x), \sigma(x)]_\alpha = 0$ for all $x \in I$ and $\alpha \in \Gamma$.*

Proof. By hypothesis, we have

$$[d(x), \sigma(x)]_\alpha \in Z(M), \tag{12}$$

for all $x \in I$ and $\alpha \in \Gamma$. Linearizing (12), we find

$$\begin{aligned}&[d(x), \sigma(y)]_\alpha + [d(y), \sigma(x)]_\alpha + 3[D(x, y, y), \sigma(x)]_\alpha + 3[D(y, x, x), \sigma(x)]_\alpha \\ &+ 3[D(x, y, y), \sigma(y)]_\alpha + 3[D(y, x, x), \sigma(y)]_\alpha \in Z(M),\end{aligned}$$

for all $x, y \in I$ and $\alpha \in \Gamma$. Replacing x by $-x$ in the above relation and comparing the above relation with the obtained relation, we have

$$[D(x, y, y), \sigma(x)]_\alpha + [D(y, x, x), \sigma(y)]_\alpha \in Z(M), \tag{13}$$

for all $x, y \in I$ and $\alpha \in \Gamma$. Replacing x by $x + y$ in the above equation we obtain

$$[d(y), \sigma(x)]_\alpha + 3[D(x, y, y), \sigma(y)]_\alpha \in Z(M), \tag{14}$$

for all $x, y \in I$ and $\alpha \in \Gamma$. Replacing $y\beta x$ by x in (14), we have

$$\begin{aligned}&[d(y), \sigma(y\beta x)]_\alpha + 3[D(y\beta x, y, y), \sigma(y)]_\alpha \\ &= [d(y), \sigma(y)]_\alpha \beta\sigma(x) + \sigma(y)\beta[d(y), \sigma(x)]_\alpha + 3[d(y)\beta x + \sigma(y)\beta D(x, y, y), \sigma(y)]_\alpha \\ &= \sigma(y)\beta([d(y), \sigma(x)]_\alpha + 3[D(x, y, y), \sigma(y)]_\alpha) + (\sigma(x) + 3x)\beta[d(y), \sigma(y)]_\alpha \\ &\quad + 3d(y)\beta[x, \sigma(y)]_\alpha \in Z(M).\end{aligned}$$

Hence,

$$\begin{aligned}0 &= [\sigma(y)\beta([d(y), \sigma(x)]_\alpha + 3[D(x, y, y), \sigma(y)]_\alpha), \sigma(y)]_\alpha \\ &\quad + [(\sigma(x) + 3x)\beta[d(y), \sigma(y)]_\alpha + 3d(y)\beta[x, \sigma(y)]_\alpha, \sigma(y)]_\alpha \\ &= [(\sigma(x) + 6x), \sigma(y)]_\alpha \beta[d(y), \sigma(y)]_\alpha + 3d(y)\beta[[x, \sigma(y)]_\alpha, \sigma(y)]_\alpha,\end{aligned}$$

for all $x, y \in I$ and $\alpha \in \Gamma$. Writing $d(y)\beta[d(y), \sigma(y)]_\alpha$ for x, we obtain

$$\begin{aligned}0 &= [\sigma(d(y)\beta[d(y), \sigma(y)]_\alpha) + 6d(y)\beta[d(y), \sigma(y)]_\alpha, \sigma(y)]_\alpha \beta[d(y), \sigma(y)]_\alpha \\ &\quad + 3d(y)\beta[[d(y)\beta[d(y), \sigma(y)]_\alpha, \sigma(y)]_\alpha, \sigma(y)]_\alpha \\ &= [\sigma(d(y)\beta[d(y), \sigma(y)]_\alpha), \sigma(y)]_\alpha \beta[d(y), \sigma(y)]_\alpha \\ &\quad + 6[d(y), \sigma(y)]_\alpha \beta[d(y), \sigma(y)]_\alpha \beta[d(y), \sigma(y)]_\alpha \\ &= [\sigma(d(y)), \sigma(y)]_\alpha \beta\sigma([d(y), \sigma(y)]_\alpha)\beta[d(y), \sigma(y)]_\alpha \\ &\quad + 6[d(y), \sigma(y)]_\alpha \beta[d(y), \sigma(y)]_\alpha \beta[d(y), \sigma(y)]_\alpha.\end{aligned}$$

Since d is commuting on I, we obtain

$$2[d(y),\sigma(y)]_\alpha\beta[d(y),\sigma(y)]_\alpha\beta[d(y),\sigma(y)]_\alpha = 0,$$

for all $y \in I$ and $\alpha, \beta \in \Gamma$. It follows that

$$(M\delta[d(y),\sigma(y)]_\alpha\beta[d(y),\sigma(y)]_\alpha)\Gamma M\Gamma(2[d(y),\sigma(y)]_\alpha\beta[d(y),\sigma(y)]_\alpha) = 0$$

and since M is semiprime and by using the condition (1) and since $d(y), \sigma(y) \in Z(M)$, we have

$$2[d(y),\sigma(y)]_\alpha\beta[d(y),\sigma(y)]_\alpha = 0, \tag{15}$$

for all $y \in I$ and $\alpha, \beta \in \Gamma$. On the other hand, taking $x = y\beta y$ in (7), we get

$$\begin{aligned}
&[d(y),(y\beta y)]_\alpha + 3[D(y\beta y, y, y), \sigma(y)]_\alpha \\
&= 2\sigma(y)\beta[d(y),\sigma(y)]_\alpha + 3[d(y)\beta y + \sigma(y)\beta d(y), \sigma(y)]_\alpha \\
&= 5\sigma(y)\beta[d(y),\sigma(y)]_\alpha + 3y\beta[d(y),\sigma(y)]_\alpha + 3d(y)\beta[y,\sigma(y)]_\alpha \in Z(M).
\end{aligned}$$

Therefore,

$$\begin{aligned}
0 &= [d(y), 5\sigma(y)\beta[d(y),\sigma(y)]_\alpha + 3y\beta[d(y),\sigma(y)]_\alpha + 3d(y)\beta[y,\sigma(y)]_\alpha]_\alpha \\
&= 5[d(y),\sigma(y)]_\alpha\beta[d(y),\sigma(y)]_\alpha + 3[d(y),y]_\alpha\beta[d(y),\sigma(y)]_\alpha \\
&\quad + 3d(y)\beta[d(y),[y,\sigma(y)]_\alpha]_\alpha \\
&= 5[d(y),\sigma(y)]_\alpha\beta[d(y),\sigma(y)]_\alpha + 3[d(y),y]_\alpha\beta[d(y),\sigma(y)]_\alpha \\
&\quad + 3d(y)\beta[[d(y),y]_\alpha,\sigma(y)]_\alpha.
\end{aligned}$$

Since d is commuting on I, we get

$$5[d(y),\sigma(y)]_\alpha\beta[d(y),\sigma(y)]_\alpha = 0, \tag{16}$$

for all $y \in I$ and $\alpha, \beta \in \Gamma$. Comparing (15) and (16), we find that

$$3[d(y),\sigma(y)]_\alpha\beta[d(y),\sigma(y)]_\alpha\beta[d(y),\sigma(y)]_\alpha = 0,$$

for all $y \in I$ and $\alpha, \beta \in \Gamma$. Since M is 3-torsion free, we obtain

$$[d(y),\sigma(y)]_\alpha\beta[d(y),\sigma(y)]_\alpha = 0,$$

for all $y \in I$ and $\alpha, \beta \in \Gamma$. Since $[d(y),\sigma(y)]_\alpha \in Z(M)$ and M is semiprime, we have $[d(y),\sigma(y)]_\alpha = 0$, for all for all $y \in M$ and $\alpha \in \Gamma$. because the center of a semiprime Γ-ring contain no non-zero nilpotent element.

Corollary 1. *Let M be a 3!-torsion free prime Γ-ring satisfying the condition (1), $I \neq 0$ be an ideal of M and σ be an automorphism of M. Assume that there exists a nonzero permuting skew tri-derivation $D : M \times M \times M \to M$ having the trace d associated with the automorphism σ such that the trace function d is commuting on I and $[d(x),\sigma(x)]_\alpha \in Z(M)$ holds true for all $x \in I$ and $\alpha \in \Gamma$. Then, M is commutative.*

Proof. Assume that M is not commutative. Then, in view of Theorem 2,

$$[d(x),\sigma(x)]_\alpha = 0$$

for all $x \in I$ and $\alpha \in \Gamma$. Therefore, by Theorem 1, we obtain $D = 0$, a contradiction.

References

1. Argac, N.: On prime and semiprime rings with derivations. Algebra Colloq. **13**, 371–380 (2006)
2. Ashraf, M.: On symmetric biderivations in rings. Rend. Istit. Mat. Univ. Trieste **31**, 25–36 (1999)
3. Atteya, M.J.: Permuting 3-derivations of semiprime rings. In: Proceedings of the 7th Annual Canadian Young Researchers Conference in Mathematics and Statistics (2010)
4. Barnes, W.E.: On the Γ-rings of Nobusawa. Pac. J. Math. **18**, 411–422 (1966)
5. Dey, K.K., Paul, A.C., Rakhimov, I.S.: Tri-additiive maps and permutting tri-derivations of Gamma-rings. JP J. Algebra Number Theor. Appl. **25**(1), 29–44 (2012)
6. Dey, K.K., Paul, A.C.: Permuting tri-derivations of semiprime Gamma rings. J. Sci. Res. **5**(1), 56–65 (2013)
7. Dey, K.K., Paul, A.C.: On the trace of a permuting tri-additive mapping in left s_Γ-unital Γ-rings. J. Sci. Res. **3**(2), 331–337 (2011)
8. Fosner, A.: Prime and semiprime rings symmetric skew 3-derivations. Aequationes Mathematicae **87**(1), 191–200 (2013). doi:10.1007/s00010-013-0208-8. Springer
9. Jung, Y.-S., Park, K.-H.: On prime and semiprime rings with permuting 3-derivations. Bull. Korean Math. Soc. **44**, 789–794 (2007)
10. Nabusawa, N.: On a generalization of the ring theory. Osaka J. Math. **1**, 65–75 (1964)
11. Özden, D., Özturk, M.A., Jun, Y.B.: Permuting tri-derivations in prime and semiprime gamma rings. Kyunpook Math. J. **46**(2), 153–200 (2006)
12. Özturk, M.A., Jun, Y.B.: Tri-additive maps and permuting tri-derivations. Commun. Fac. Sci. Univ. Ank. Ser. A1, **54**(1), 1–9 (2005)
13. Özturk, M.A., Jun, Y.B., Kim, K.H.: Orthogonal traces on semiprime gamma rings. Scienticae Japonicae **4**, 423–429 (2001)
14. Park, K.-H.: On prime and semiprime rings with symmetric n-derivations. J. Chungcheong Math. Soc. **22**, 451–458 (2009)
15. Park, K.-H., Jung, Y.-S.: On permuting 3-derivations and commutativity in prime near-rings. Commun. Korean Math. Soc. **25**, 1–9 (2010)

Improvement of Analytical Solution to the Inverse Truly Nonlinear Oscillator by Extended Iterative Method

B.M. Ikramul Haque[1(✉)], Md. Asifuzzaman[1], and M. Kamrul Hasan[2]

[1] Department of Mathematics,
Khulna University of Engineering & Technology,
Khulna 9203, Bangladesh
bmihmathkuet@gmail.com
[2] Department of Mathematics,
Rajshahi University of Engineering & Technology,
Rajshahi 6204, Bangladesh

Abstract. A new approach to the Mickens extended iteration method has been presented to obtain approximate analytic solutions for nonlinear oscillatory differential equation. To illustrate the accuracy of the approximate solution of the inverse nonlinear oscillator "$\ddot{x}+x^{-1}=0$", we have used the Fourier series and utilized indispensable truncated terms in each iterative step. In this article the solution gives more accurate result significantly than other existing methods and shows a good agreement with its exact solution. The percentage of error between exact frequency and our third approximate frequency is very low. We have compared all the results to exact results and other existing results and the method is convergent as well as consistent. Finally, an example is given to show the effectiveness of the approximate solution.

Keywords: Extended iterative method · Inverse truly nonlinear oscillator · Analytical solution

AMS Subject Classification: 34A34 · 34B99

1 Introduction

Most phenomena in our world are essentially nonlinear and are described by nonlinear equations. A vast of scientific knowledge has developed over a long period of time and devoted to a description of natural phenomena. Practically, most of the differential equations involving physical phenomena are nonlinear. These equations have also demonstrated their usefulness in ecology, business cycle and biology. Therefore the solution of such problems lies essentially in solving the corresponding differential equations. In many cases it is possible to replace such a nonlinear equation by a related linear equation, which approximates the actual problem closely enough to give useful results. The method of small oscillations is a well-known example of the linearization of problems which is essentially nonlinear. However, such a "linearization" is not always feasible or possible; and when it is not, the original nonlinear equation itself

© Springer Nature Singapore Pte Ltd. 2017
D. Giri et al. (Eds.): ICMC 2017, CCIS 655, pp. 412–421, 2017.
DOI: 10.1007/978-981-10-4642-1_35

must be considered. Many methods exist for constructing analytical approximations to the solution of the oscillatory system, viz. perturbation method [1, 2], standard and modified Linstedt-Poincare [3, 4], power series approach and homotopy analysis method [5, 6], harmonic balance method [7–9], iteration method [10–21] etc.

Perturbation method is mainly used for the small nonlinear problems. The modified Lindstedt-Poincare method, power series approach and homotopy analysis method have been presented for obtaining approximate periods with large amplitude of oscillations. But they are applicable only to nonlinear equations with odd nonlinearity. The mathematical foundations of harmonic balancing have been investigated by several individuals. The harmonic balance method which is originated by Mickens [7] and farther work has been done by Wu et al. [8], Hosen [9] and so on for solving the strong nonlinear problems. It corresponds to a truncated Fourier series and allows for the systematic determination of the coefficients to the various harmonics and the angular frequency.

Now-a-day's iteration method is used widely by Mickens [10], Lim and Wu [11], Hu [12], Chen and Liu [14], Alquran [16], Turkyilmazoglu [17], Haque [19–21] etc. and it is valid for small together with large amplitude of oscillation to attain the approximate frequency and the harmonious periodic solution of such nonlinear problems. Mickens [10] provided a general basis for iteration methods as they are currently used in the calculation of approximations to the periodic solutions of nonlinear oscillatory differential equations. A generalization of this work was then given by Lim and Wu [11] and this was followed by an additional extension in Mickens. Actually iteration method is a technique for calculating approximations to the periodic solutions of TNL oscillator which is patented by R.E. Mickens in [10].

The main purpose of this article is to develop a modification of the extended iteration technique for the determination of approximate solution and angular frequency of inverse truly nonlinear oscillator. We compare the result to existing results obtained by various researchers and it is mentioned that our solution measure similar and sometimes is with better results than other existing procedures.

2 The Method

An Extended Iterative method will be used to obtain analytical solution of the inverse truly nonlinear oscillator. The procedure may be briefly described as follows.

A nonlinear oscillator will be modeled by

$$\ddot{x}+f(\ddot{x},\dot{x},x)=0,\ x(0)=A,\ \dot{x}(0)=0 \tag{1}$$

where over dots denote differentiation with respect to time, t.

We choose the natural frequency Ω of this system. Then adding $\Omega^2 x$ to both sides of Eq. (1), we obtain

$$\ddot{x}+\Omega^2 x=\Omega^2 x-f(\ddot{x},\dot{x},x)\equiv G(\ddot{x},\dot{x},x).$$

The iteration scheme of above equation is as follows

$$\ddot{x}_{k+1} + \Omega_k^2 x_{k+1} = G(\ddot{x}_k, \dot{x}_k, x_k);\ k = 0, 1, 2, \cdots. \tag{2}$$

The extended iteration scheme is

$$\begin{aligned}\ddot{x}_{k+1} + \Omega_k^2\, x_{k+1} = {} & G(x_{k-1}, \dot{x}_{k-1}, \ddot{x}_{k-1}) + G_x(x_{k-1}, \dot{x}_{k-1}, \ddot{x}_{k-1})(x_k - x_{k-1}) \\ & + G_{\dot{x}}(x_{k-1}, \dot{x}_{k-1}, \ddot{x}_{k-1})(\dot{x}_k - \dot{x}_{k-1}) + G_{\ddot{x}}(x_{k-1}, \dot{x}_{k-1}, \ddot{x}_{k-1})\,(\ddot{x}_k - \ddot{x}_{k-1})\end{aligned} \tag{3}$$

where $G_x = \frac{\partial G}{\partial x},\ G_{\dot{x}} = \frac{\partial G}{\partial \dot{x}},\ G_{\ddot{x}} = \frac{\partial G}{\partial \ddot{x}}$.

And x_{k+1} satisfies the conditions

$$x_{k+1}(0) = A,\ \dot{x}_{k+1}(0) = 0. \tag{4}$$

The starting function are taken to be [11]

$$x_{-1}(t) = x_0(t) = A\,\cos(\Omega_0 t). \tag{5}$$

The right hand side of Eq. (3) is essentially the first term in a Taylor series expansion of the function $G(x_k, \dot{x}_k, \ddot{x}_k)$ at the point $(x_{k-1}, \dot{x}_{k-1}, \ddot{x}_{k-1})$ [22].

The above procedure gives the sequence of solutions $x_1(t),\ x_2(t),\ x_3(t), \cdots$. The method can be proceed to any order of approximation; but due to growing algebraic complexity the solution is confined to a lower order usually the second [10].

Example 1. Let us consider the inverse truly nonlinear oscillator

$$\ddot{x} + x^{-1} = 0. \tag{6}$$

Adding $\Omega^2 x$ on both sides of Eq. (6), we get

$$\ddot{x} + \Omega^2 x = \Omega^2 x - x^{-1} = G(x,\, \Omega^2), \tag{7}$$

where $G(x,\, \Omega^2) = \Omega^2 x - x^{-1}$, $G_x(x,\, \Omega^2) = \Omega^2 + x^{-2}$.

According to Eq. (3), the extended iteration scheme of Eq. (7) is

$$\ddot{x}_{k+1} + \Omega_k^2 x_{k+1} = (\Omega_k^2 x_0 - x_0^{-1}) + (\Omega_k^2 + x_0^{-2})(x_k - x_0). \tag{8}$$

The first approximation $x_1(t)$ and the frequency Ω_0 will be obtained by putting $k = 0$ in Eq. (8) and using Eq. (5) we get

$$\begin{aligned}\ddot{x}_1 + \Omega_0^2 x_1 &= (\Omega_0^2 x_0 - x_0^{-1}) + (\Omega_0^2 + x_0^{-2})(x_0 - x_0) \\ &= \Omega_0^2 x_0 - x_0^{-1},\end{aligned} \tag{9}$$

where $x_0(t) = A\,\cos(\Omega_0 t) = A\,\cos\theta$.

Now substituting $x_0(t)$ and expanding the right- hand side in a Fourier cosine series, then Eq. (9) reduces to

$$\begin{aligned}\ddot{x}_1 + \Omega_0^2 x_1 &= \Omega_0^2 A\cos\theta - (2\cos\theta/A - 2\cos 3\theta/A + 2\cos 5\theta/A)\\ &= (\Omega_0^2 A - 2/A)\cos\theta + 2\cos 3\theta/A - 2\cos 5\theta/A.\end{aligned} \tag{10}$$

To avoid secular terms in the solution, we have to remove cos θ from the right hand side of Eq. (10). Thus we have

$$\Omega_0 = \frac{\sqrt{2}}{A} = \frac{1.41421}{A}. \tag{11}$$

This is the first approximate frequency of the oscillator. Note that $\Omega_{exact}(A) = \frac{1.253314}{A}$. After simplification the Eq. (10) reduces to

$$\ddot{x}_1 + \Omega_0^2 x_1 = 2\cos 3\theta/A - 2\cos 5\theta/A. \tag{12}$$

Then solving Eq. (12) and satisfying the initial condition $x_1(0) = A$, we obtain

$$x_1(t) = A(1.083333\cos\theta - 0.125\cos 3\theta). \tag{13}$$

This is the first approximate solution of the oscillator. Proceeding to the second level of iteration, $x_2(t)$ satisfies the equation

$$\begin{aligned}\ddot{x}_2 + \Omega_1^2 x_2 &= (\Omega_1^2 x_0 - x_0^{-1}) + (\Omega_1^2 + x_0^{-2})(x_1 - x_0)\\ &= \Omega_1^2 x_1 - 2x_0^{-1} + x_0^{-2} x_1,\end{aligned} \tag{14}$$

where $x_0(t) = A\cos\theta$ and $x_1(t) = A(1.083333\cos\theta - 0.125\cos 3\theta)$.

Now substituting $x_0(t)$ and $x_1(t)$ and expanding the right- hand side in a Fourier cosine series then Eq. (14) reduces to

$$\begin{aligned}\ddot{x}_2 + \Omega_1^2 x_2 &= \Omega_1^2 A(1.083333\cos\theta - 0.125\cos 3\theta) - (1.58333\cos\theta/A - 1.083333\cos 3\theta/A)\\ &= (1.083333\Omega_1^2 A - 1.58333/A)\cos\theta - (0.125\Omega_1^2 A - 1.083333/A)\cos 3\theta.\end{aligned} \tag{15}$$

To avoid secular terms in the solution, we have to remove cos θ from the right hand side of Eq. (15). Thus we have

$$\Omega_1 = \frac{1.20894}{A}. \tag{16}$$

This is the second approximate frequency of the oscillator. After simplification the Eq. (15) reduces to

$$\ddot{x}_2 + \Omega_1^2 x_2 = 0.900641 \cos 3\theta/A. \tag{17}$$

Then solving Eq. (17) and satisfying the initial condition $x_2(0) = A$, we obtain

$$x_2(t) = A(1.077029 \cos \theta - 0.077029 \cos 3\theta). \tag{18}$$

This is the second approximate solution of the oscillator. Proceeding to the third level of iteration, $x_3(t)$ satisfies the equation

$$\begin{aligned} \ddot{x}_3 + \Omega_2^2 x_3 &= (\Omega_2^2 x_2 - x_0^{-1}) + (\Omega_2^2 + x_0^{-2})(x_2 - x_0) \\ &= \Omega_2^2 x_2 - 2x_0^{-1} + x_0^{-2} x_2, \end{aligned} \tag{19}$$

where $x_0(t) = A \cos \theta$ and $x_2(t) = A(1.077029 \cos \theta - 0.077029 \cos 3\theta)$.

Now substituting $x_0(t)$ and $x_2(t)$ and expanding the right- hand side in a Fourier cosine series, then Eq. (19) reduces to

$$\begin{aligned} \ddot{x}_3 + \Omega_2^2 x_3 &= \Omega_2^2 A(1.077029 \cos \theta - 0.077029 \cos 3\theta) - (1.691886 \cos \theta/A - 1.383772 \cos 3\theta/A) \\ &= (1.077029\Omega_2^2 A - 1.691886/A) \cos \theta - (0.077029\Omega_2^2 A - 1.383772/A) \cos 3\theta. \end{aligned} \tag{20}$$

To avoid secular terms in the solution, we have to remove cos θ from the right hand side of Eq. (20). So we have

$$\Omega_2 = \frac{1.25335}{A}. \tag{21}$$

Thus Ω_0, Ω_1 and Ω_2 can be obtained by Eqs. (11), (16) and (21) respectively, which represent the approximation of frequencies of oscillator (6).

3 Results and Discussions

An iterative approach is presented to obtain approximate solution of the 'inverse truly nonlinear oscillator'. The present technique is very simple for solving algebraic equations analytically and the approach is different from the existing other approach for taking truncated Fourier series. Here we have calculated the first, second and third approximate frequencies Ω_0, Ω_1 and Ω_2 respectively. All the results are given in the following Table 1. To compare the approximate frequencies we have also given the existing results determined by Mickens iteration method [15] and Mickens Harmonic Balance method [13], Haque's iteration method [18]. To show the accuracy, we have calculated the percentage errors (denoted by Error(%)) by the definitions

$$\text{Error} = \left| \frac{\Omega_e - \Omega_k}{\Omega_e} \right| \times 100\% \text{ where } \Omega_k; k = 0, 1, 2, \cdots$$

Table 1. Comparison of the approximate frequencies with exact frequency Ω_e of $\ddot{x}+x^{-1}=0$.

Exact frequency Ω_e $\frac{1.253314}{A}$			
Amplitude A	First approximate frequencies, Ω_0	Second approximate frequencies, Ω_1	Third approximate frequencies, Ω_2
	Error (%)	Error (%)	Error (%)
Mickens direct iteration method [15]	$\frac{1.155}{A}$	$\frac{1.018}{A}$	—
	7.9	18.1	
Mickens extended iteration method [15]	$\frac{1.155}{A}$	$\frac{1.189699}{A}$	—
	7.9	5.1	
Mickens HB Method [13]	$\frac{1.414}{A}$	$\frac{1.273}{A}$	$\frac{1.2731}{A}$
	12.84	1.6	1.58
Haque's iteration Method [18]	$\frac{1.414}{A}$	$\frac{1.208}{A}$	$\frac{1.265}{A}$
	12.84	3.63	0.92
Adopted method	$\frac{1.41421}{A}$	$\frac{1.20894}{A}$	$\frac{1.25335}{A}$
	12.84	3.54	0.0029

represents the approximate frequencies obtained by the present method and Ω_e represents the corresponding exact frequency of the oscillator (Table 1).

4 Convergence and Consistency Analysis

We know that the basic idea of iteration methods is to construct a sequence of solutions x_k (as well as frequencies Ω_k) that have the property of convergence

$$x_e = \lim_{k \to \infty} x_k \text{ Or, } \Omega_e = \lim_{k \to \infty} \Omega_k$$

Here x_e is the exact solution of the given nonlinear oscillator. In the present method, it has been shown that the solution yield the less error in each iterative step compared to the previous iterative step and finally $|\Omega_2 - \Omega_e| = |0.253350 - 0.253314| < \varepsilon$, where ε is a small positive number and A is chosen to be unity. From this, it is clear that the adopted method is convergent.

An iterative method of the form represented by Eq. (3) with initial guesses given in Eq. (5) is said to be consistent if

$$\lim_{k \to \infty} |x_k - x_e| = 0 \text{ or, } \lim_{k \to \infty} |\Omega_k - \Omega_e| = 0$$

In the present analysis we see that

$$\lim_{k \to \infty} |\Omega_k - \Omega_e| = 0, \text{ as } |\Omega_2 - \Omega_e| = 0.$$

Thus the consistency of the method is achieved.

Example 2. Let us consider the nonlinear cubic oscillator

$$\ddot{x} + x^3 = 0. \tag{22}$$

Adding $\Omega^2 x$ on both sides of Eq. (22), we get

$$\ddot{x} + \Omega^2 x = \Omega^2 x - x^3 = G(x, \Omega^2), \tag{23}$$

where $G(x, \Omega^2) = \Omega^2 x - x^3, G_x(x, \Omega^2) = \Omega^2 - 3x^2$.

The extended iterative scheme of Eq. (23) is

$$\ddot{x}_{k+1} + \Omega_k^2 x_{k+1} = (\Omega_k^2 x_0 - x_0^3) + (\Omega_k^2 - 3x_0^2)(x_k - x_0). \tag{24}$$

The first approximation $x_1(t)$ and the frequency Ω_0 will be obtained by putting $k = 0$ in Eq. (24), we get

$$\begin{aligned} \ddot{x}_1 + \Omega_0^2 x_1 &= (\Omega_0^2 x_0 - x_0^3) + (\Omega_0^2 - 3x_0^2)(x_0 - x_0) \\ &= \Omega_0^2 x_0 - x_0^3, \end{aligned} \tag{25}$$

where $x_0(t) = A \cos(\Omega_0 t) = A \cos\theta$.

Now substituting $x_0(t)$ and expanding the right- hand side in a Fourier cosine series, then Eq. (25) reduces to

$$\begin{aligned} \ddot{x}_1 + \Omega_0^2 x_1 &= \Omega_0^2 A \cos\theta - (0.75A^3 \cos\theta + 0.25A^3 \cos 3\theta) \\ &= (\Omega_0^2 A - 0.75A^3) \cos\theta - 0.25A^3 \cos 3\theta. \end{aligned} \tag{26}$$

To avoid secular terms in the solution, we have to remove $\cos\theta$ from the right hand side of Eq. (26). Thus we have

$$\Omega_0 = 0.866025A. \tag{27}$$

This is the first approximate frequency of the oscillator. Note that $\Omega_{exact}(A) = 0.847213A$. After simplification the Eq. (26) reduces to

$$\ddot{x}_1 + \Omega_0^2 x_1 = -0.25A^3 \cos 3\theta. \tag{28}$$

Then solving Eq. (28) and satisfying the initial condition $x_1(0) = A$, we obtain

$$x_1(t) = A(0.958333 \cos\theta + 0.041667 \cos 3\theta). \tag{29}$$

This is the first approximate solution of the oscillator. Proceeding to the second level of iteration, $x_2(t)$ satisfies the equation

$$\begin{aligned}\ddot{x}_2 + \Omega_1^2 x_2 &= (\Omega_1^2 x_0 - x_0^3) + (\Omega_1^2 - 3x_0^2)(x_1 - x_0) \\ &= \Omega_1^2 x_1 + 2x_0^3 - 3x_0^2 x_1,\end{aligned} \tag{30}$$

where $x_0(t) = A\cos(\Omega_0 t) = A\cos\theta$ and $x_1(t) = A(0.958333\cos\theta + 0.041667\cos 3\theta)$.

Now substituting $x_0(t)$ and $x_1(t)$ and expanding the right- hand side in a Fourier cosine series, then Eq. (30) reduces to

$$\begin{aligned}\ddot{x}_2 + \Omega_1^2 x_2 &= \Omega_1^2 A\,(0.958333\cos\theta + 0.041667\cos 3\theta) \\ &\quad - (0.687500A^3\cos\theta + 0.281250A^3\cos 3\theta) \\ &= (0.958333\,\Omega_1^2 A - 0.687500\,A^3)\cos\theta \\ &\quad - (0.041667\,\Omega_1^2 A - 0.281250\,A^3)\cos 3\theta.\end{aligned} \tag{31}$$

To avoid secular terms in the solution, we have to remove $\cos\theta$ from the right hand side of Eq. (31). Thus we have

$$\Omega_1 = 0.846990A. \tag{32}$$

This is the second approximate frequency of the oscillator. After simplification the Eq. (31) reduces to

$$\ddot{x}_2 + \Omega_1^2 x_2 = -0.251359A^3\cos 3\theta. \tag{33}$$

Then solving Eq. (33) and satisfying the initial condition $x_2(0) = A$, we obtain

$$x_2(t) = A(0.956203\cos\theta + 0.0437974\cos 3\theta). \tag{34}$$

This is the second approximate solution of the oscillator.

5 Results and Discussions

An iterative approach is presented to obtain approximate solution of the 'cubic truly nonlinear oscillator'. The present technique is very simple for solving algebraic equations analytically and the approach is different from the existing other approach for taking truncated Fourier series. Here we have calculated the first and second approximate frequencies. All the results are given in the following Table 2. To compare the approximate frequencies we have also given the existing results determined by Mickens Parameter Expansion [15], Mickens Harmonic Balance Method [15] and Mickens Iterative Method [15] (Table 2).

Table 2. Comparison of the approximate frequencies with exact frequency Ω_e of $\ddot{x}+x^3=0$.

Exact frequency Ω_e 0.847213*A*		
Amplitude *A*	First approximate frequencies, Ω_0	Second approximate frequencies, Ω_1
	Error (%)	Error (%)
Mickens Parameter Expansion [15]	0.866025 *A*	—
	2.2	
Mickens HB Method [15]	0.866025 *A*	0.848875*A*
	2.2	0.20
Mickens Iterative Method [15]	0.866025 *A*	0.849326*A*
	2.2	0.25
Adopted Method	0.866025*A*	0.846990*A*
	2.22	0.026

6 Conclusion

In this article, we have established a simple but effective modification of the extended iteration method to investigate nonlinear differential equations. In most of the cases, the results are improved by the modification of the method. The modified extended iteration method has been applied to both cubic and singular oscillator. In both cases, the modified extended iteration procedure gives more accurate results than the results obtained by the existing iteration scheme and is valid for large region. The percentage error between the exact frequency and the approximate frequency obtained in this study is very small. Since, in science and engineering there are many types of oscillator and the developed scheme is applied on cubic and singular oscillators, the next research may be: whether this modification is useful to other kinds of oscillator or it needs further modification.

Acknowledgement. The authors are grateful to the honorable reviewers' for their constructive suggestions/comments to improve the quality of this article. The authors are also grateful to Md. Shahinur Alam Sarker, Assistant Professor, Department of Humanities, Khulna University of Engineering & Technology, Khulna-9203, for his assistance to prepare the revised manuscript.

References

1. Nayfeh, A.H., Mook, D.T.: Nonlinear Oscillation. Wiley, New York (1979)
2. Alam, A., Rahman, H., Haque, B.M.I., Ali Akbar, M.: Perturbation technique for analytic solutions of fourth order near critically damped nonlinear systems. Int. J. Basic Appl. Sci. **11**, 131–138 (2011)
3. He, J.H.: Modified Lindstedt-Poincare methods for some non-linear oscillations. Part III: double series expansion. Int. J. Nonlinear Sci. Numer. Simul. **2**, 317–320 (2001)
4. Ramos, J.I.: Approximate methods based on order reduction for the periodic solutions of nonlinear third-order ordinary differential equations. Appl. Math. Comput. **215**, 4304–4319 (2010)

5. Xu, H., Cang, J.: Analysis of a time fractional wave-like equation with the homotopy analysis method. Phys. Lett. A **372**, 1250–1255 (2008)
6. Belendez, A., Pascual, C., Ortuno, M., Belendez, T., Gallego, S.: Application of a modified He's homotopy perturbation method to obtain higher-order approximations to a nonlinear oscillator with discontinuities. Real World Appl. **10**, 601–610 (2009)
7. Mickens, R.E.: Comments on the method of harmonic balance. J. Sound Vib. **94**, 456–460 (1984)
8. Wu, B.S., Sun, W.P., Lim, C.W.: An analytical approximate technique for a class of strongly nonlinear oscillator. Int. J. Nonlinear Mech. **41**, 766–774 (2006)
9. Hosen, M.A.: Accurate approximate analytical solutions to an anti-symmetric quadratic nonlinear oscillator. Afr. J. Math. Comput. Sci. Res. **6**, 77–81 (2013)
10. Mickens, R.E.: Iteration Procedure for determining approximate solutions to nonlinear oscillator equation. J. Sound Vib. **116**, 185–188 (1987)
11. Lim, C.W., Wu, B.S.: A modified procedure for certain non-linear oscillators. J. Sound Vib. **257**, 202–206 (2002)
12. Hu, H., Tang, J.H.: A classical iteration procedure valid for certain strongly nonlinear oscillator. J. Sound Vib. **299**, 397–402 (2006)
13. Mickens, R.E.: Harmonic balance and iteration calculations of periodic solutions to $\ddot{y} + y^{-1} = 0$. J. Sound Vib. **306**, 968–972 (2007)
14. Chen, Y.M., Liu, J.K.: A modified Mickens iteration procedure for nonlinear oscillators. J. Sound Vib. **314**, 465–473 (2008)
15. Mickens, R.E.: Truly Nonlinear Oscillations. World Scientific, Singapore (2010)
16. Alquran, M., Al-Khaled, K.: Effective approximate methods for strongly nonlinear differential equations with oscillations. Math. Sci. **6**, 32 (2012)
17. Turkyilmazoglu, M.: An effective approach for approximate analytical solutions of the damped Duffing equation. Phys. Scr. **86**, 01530 (2012)
18. Haque, B.M.I.: Modified solutions of some oscillators by iteration procedure. J. Egypt. Math. Soc. **21**, 68–73 (2013)
19. Haque, B.M.I., Hossain, M.R.: An analytic investigation of the quadratic nonlinear oscillator by an iteration method. BJMCS **13**, 1–8 (2015)
20. Haque, B.M.I., Bostami, M.B., Hossain, M.M.A., Hossain, M.R., Rahman, M.M.: Mickens iteration like method for approximate solutions of the inverse cubic truly nonlinear oscillator. BJMCS **13**, 1–9 (2015). Article no. 22823
21. Haque, B.M.I., Hossain, M.M.A., Bostami, M.B., Hossain, M.R.: Analytical approximate solutions to the nonlinear singular oscillator: an iteration procedure. BJMCS **14**, 1–7 (2016). Article no. 23263
22. Taylor, A.E., Mann, W.R.: Advance Calculus. Wiley, New York (1983)

Author Index