University of Kashan
678

Proceedings

of the 51st Annual Iranian Mathematics Conference

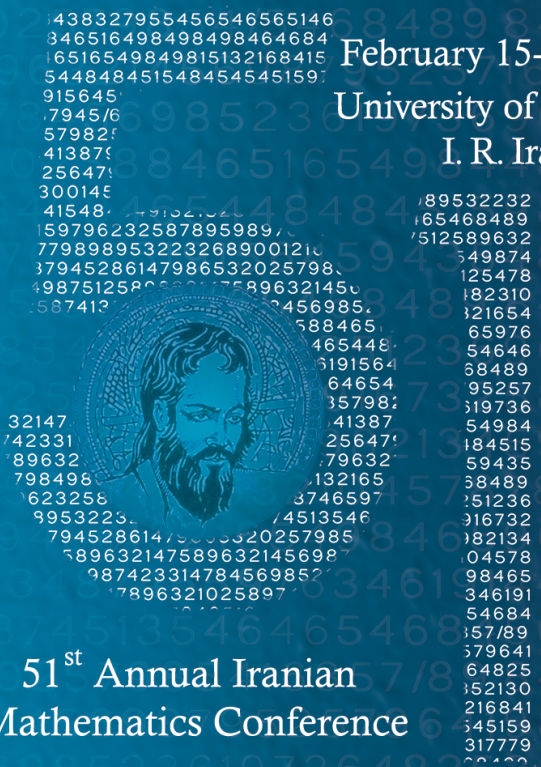
Volume 2:
Applied Mathematics, Statistics and Computer Science

Editors:

Zeinab Saeidian Tarei
(Faculty Member, University of Kashan, I. R. Iran)

Mardjan Hakimi-Nezhaad

Ali Reza Ashrafi
(Faculty Member, University of Kashan, I. R. Iran)



February 15-20, 2021
University of Kashan,
I. R. Iran

51st Annual Iranian
Mathematics Conference

51^{st} Annual Iranian Mathematics Conference

15–20 February 2021, University of Kashan

Proceedings of the 51^{st} Annual Iranian Mathematics Conference, Volume 2: Applied Mathematics, Statistics and Computer Science

Secretariat: Faculty of Mathematical Sciences, University of Kashan,
Kashan 87317-53153, I. R. Iran
Phone: (+9831) 55912918
Fax: (+9831) 55912332
Email: aimc51@kashanu.ac.ir
http://aimc51.kashanu.ac.ir

سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران

سرشناسه: کنفرانس ریاضی ایران (پنجاه و یکمین: ۱۳۹۹: کاشان)
Annual Iranian Mathematics Conference (51st: 2021: Kashan)
عنوان و نام پدیدآور: Proceedings of the 51st Annual Iranian Mathematics conference, Volume 2: Applied Mathematics, Statistics and Computer Science[Book]/ editors Zeinab Saeidian Tarei ...[et al.].
مشخصات نشر: کاشان: دانشگاه کاشان، کاشان: سوره تماشا، ۱۴۰۰=۲۰۲۱م.
مشخصات ظاهری: ۷۵۹ ص.
شابک: ۹۷۸-۶۲۲-۶۵۴۶-۲۳-۲
وضعیت فهرست‌نویسی: فیپا
یادداشت: انگلیسی.
یادداشت: editors Zeinab Saeidian Tarei, Mardjan Hakimi-Nezhaad, Ali Reza Ashrafi.
موضوع: ریاضیات – کنگره‌ها
موضوع: Mathematics – Congresses
شناسه افزوده: سعیدیان طرئی، زینب، ۱۳۶۳-، ویراستار
شناسه افزوده: Saeidian Tarei, Zeinab - ۱۹۸۴،
رده‌بندی کنگره: QA۱
رده‌بندی دیویی: ۵۱۰
شماره کتابشناسی ملی: ۷۶۶۲۱۴۲

University of Kashan
678

51st Annual Iranian Mathematics Conference

15-20 February 2021, University of Kashan

Proceedings of the 51st Annual Iranian Mathematics Conference, Volume 2: Applied Mathematics, Statistics and Computer Science

Publisher: University of Kashan Press and Sureh Tamasha Publication

Editors: Zeinab Saeidian Tarei, Mardjan Hakimi-Nezhaad, Ali Reza Ashrafi

Printing and Binding: Baran

First Printing: 2021

Print Run: 200

ISBN: 978-622-6546-23-2

Price: 1500000 IR Rials or 10 Euro

Editors:

Zeinab Saeidian Tarei
Faculty Member, University of Kashan, I. R. Iran
saeidian@kashanu.ac.ir

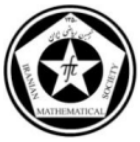
Mardjan Hakimi-Nezhaad
University of Kashan, I. R. Iran
m.hakimi20@gmail.com

Ali Reza Ashrafi
Faculty Member, University of Kashan, I. R. Iran
ashrafi@kashanu.ac.ir

Library of Congress Control Number: 7662142

ISBN: 978-622-6546-23-2

Organizers



Sponsors

Message from the Mayor of Kashan

Once again the world's top mathematicians, professors, scholars, and students of mathematics have gathered in a scientific circle in the historical city of Kashan. The Faculty of Mathematics of the University of Kashan has been honored to host the 51st Annual Iranian Mathematics Conference. Undoubtedly, the philosophy of science would be incomplete in the absence of objective examples of phenomena. Mathematics serves as the basic science for understanding the principles of existence and the basis of the order of the universe. As our grasp on mathematic theory tightens, we are humbled by the greatness of this world's creator.

It is not a secret that Kashan has long been a cradle for flourishing men and women like Ghiythal-DnJamshdKashanis who have advanced the boundaries of science.

We were also pleased to have with us an acclaimed mathematician from our city, Dr. Javad Mashreghi; president of the Canadian Mathematical Society.

As the Mayor of Kashan, I wish to welcome all scholars and mathematics enthusiasts to this conference and to thank the esteemed keynote speakers, guests, and participants. I pray that this message finds you in health and ever-increasing prosperity. I wish for a world free of pandemics and a return to normal with physical conferences.

Mayor of Kashan
Saeed Abrishami-Rad

Foreword

The 51st Annual Iranian Mathematics Conference was held at University of Kashan in cooperation with the Iranian Mathematical Society from February 15 to February 20, 2021. We were eager to host the presence of the mathematical community of Iran at University of Kashan, and by providing an intimate and academic atmosphere for opportunities for exchange and scientific participation for all in the field of mathematical sciences and their applications. University of Kashan was founded at first as an institution of higher education in 1973. It began its activities in October, 1974 by 200 students of mathematics and physics.

Being in a suitable geographical position, the cultural atmosphere of the region and the long history in science and art have provided the basis for great success for this university and now, for example, University of Kashan has been introduced as the seventh comprehensive university in Iran by ISC National University Ranking.

The Faculty of Mathematical Sciences of University of Kashan is active with nearly forty full-time faculty members in three levels of bachelor's, master's and doctoral degrees and has made a significant contribution to the development and achievements of University of Kashan.

Holding successful conferences, student competitions of the Iranian Mathematical Society and various specialized seminars have been among the activities of this faculty. The editor in chief of the "Bulletin of the Iranian Mathematical Society" and the "Journal of Mathematical Culture and Thought" by the faculty members of this faculty at various times, are some of the effective collaborations with the Iranian Mathematical Society.

Due to the outbreak of the Corona virus, the 51st Iranian Mathematical Conference is being held virtually in University of Kashan for the first time.Besides the limitations created by holding the conference virtually, new opportunities have emerged. We had the great opportunity by using the facilities of cyberspace to invite prominent national and international professors from 22 different countries.

You are all aware that due to various reasons and problems in the educational, economic and social dimensions, the number of mathematics students has decreased significantly in recent years.

The elites of the country, have emphasized on strengthening the basic sciences, especially mathematics, and have introduced them as a treasure for the development of the country. It is up to the Iranian Mathematical Society to use the opportunity and the support the authorities, to plan for the promotion and expansion of mathematics.

As a step towards taking responsibility for this, we added a new section to the conference this year called "Mathematical Promotion". This idea was welcomed by the esteemed officials of the Iranian Mathematical Society and it is hoped that it will be followed as part of the conference in the coming years. In this regard, with the help of the education department of the region, a call was made and so far we have received more than 400 articles, from interested students in different levels of elementary and high school from all over the country.

It was decided to hold the first meeting for the promotion and popularization of mathematics as part of the mathematics conference in the near future and to present the selected works.

I consider it necessary to thank the Ministry of Science, Research and Technology, esteemed officials of University of Kashan, dear colleagues in the Faculty of Mathematical Sciences of the University of Kashan, faculty members of universities and research centers across the country who helped and guided us in particular those who contributed to the accurate judging of the received papers.

I would like to thank all the participants who added value by sending valuable papers and participating in the conference. Holding a conference like Iranian Mathematics Conference virtually was a new experience for us. I hope we have been able to do this great event well and in a desirable and worthy way. Moreover, this will be an experience for the expansion of virtual activities in the future. I apologize in advance for all the shortcomings, which were mainly due to our lack of experience in holding such conferences and virtual activities.

Hoping to see you at the future conferences.

Conference Chair of AIMC51
Hassan Daghigh

Welcome to AIMC51

The Annual Iranian Mathematics Conference (AIMC) is the country's most important and oldest mathematical gathering where researchers, students, and professors at home and abroad present their latest scientific findings. The first mathematics conference of the country was held by the University of Shiraz in April 1970, the most important of which was the proposal to establish the Iranian Mathematical Society, which coincided with the second mathematical conference of the country at the Sharif University of Technology in April 1971. Since then, the conference has welcomed a large number of scholars at home and abroad each year.

The Iranian Mathematics Conference has been held for the last fifty years despite all the difficulties. The Faculty of Mathematical Sciences of the University of Kashan is now honored to hold the fifty-first gathering of this important mathematics event of the country from February 15 to February 20, 2021 in the cradle of Iranian civilization and traditional culture, the city of Kashan with seven thousand years history.

We originally planned to hold the conference in person from 7 September to 10 September 2020, but due to the corona pandemic and the laws announced to the universities by the government, we changed the time to February 2021.

AIMC 51 has 31 keynote and 7 invited speakers from 20 different countries, all of whom are among the best and most famous mathematicians in the world in their field. The scope of the conference covered various topics in mathematics, statistics and computer science. The conference was attended by more than 500 researchers from Argentina, Belarus, Brazil, Canada, Check Republic, China, Croatia, India, Iran, Iraq, Italy, Kuwait, Netherland, Nigeria, Oman, Pakistan, Romania, Russia, Saudi Arabia, Serbia, South Africa, South Korea, Thailand, Turkey and USA who held 20, 40 and 60 minutes lectures.

We have fifteen keynote speakers in pure mathematics, seven keynote speakers in applied mathematics, four keynotes in statistics and five keynotes in computer science. There are also seven young invited speakers who are famous mathematicians in their topics.

Our Keynote Speakers in Pure Mathematics are professors: Alireza Abdollahi (University of Isfahan, I. R. Iran), Javad Asadollahi (University of Isfahan, I. R. Iran), Mohammad Bagheri (Historian), Maurizio Brunetti (Universit di Napoli Federico II, Italy), Henri Darmon (McGill University, Canada), Omid Ali Shehni Karamzadeh (Shahid Chamran University of Ahvaz, I. R. Iran), Javad Mashreghi (Laval university, Canada), Mohammad Sal Moslehian (Ferdowsi University of Mashhad, I. R. Iran), Thekiso Seretlo (University of Limpopo, South Africa), Mohammad Shahryari (Sultan Qaboos University, Muscat, Oman), Andrea Solotar (University of Buenos Aires, Argentina), Teerapong Suksumran (Chiang Mai University, Thailand), Mukut Mani Tripathi (Banaras Hindu University, India), Andrei Yu. Vesnin (Russian Academy of Sciences, Russia) and Changchang Xi (Capital Normal University, China).

The AIMC51 Keynote Speakers in Applied Mathematics are professors: Tomislav Došlić (University of Zagreb, Croatia), Roberto Garrappa (University of Bari, Italy), Nezameddin Mahdavi-Amiri (Sharif University of Technology, I. R. Iran), Davoud Mirzaei (University of Isfahan, I. R. Iran), Kees Roos (Delft University of

Technology, Netherland), Majid Soleimani Damaneh (University of Tehran, I. R. Iran) and Zahra Gooya (Shahid Beheshti University, I. R. Iran).

Other main topics of AIMC 51 are Statistics and Computer Science, and the keynote speakers of these topics are professors: Masoud Asgharian (McGill university, Canada), Khalil Shafie (University of Northern Colorado, USA), Ahmad Reza Soltani (Kuwait University, Kuwait), Bijan Zohuri-Zangeneh (Sharif University of Technology, I. R. Iran), Khodakhast Bibak (Miami University, USA), Alain Bretto (University of Caen, France), Luca De Feo (University of Versailles - Saint-Quentin, France), Predrag S. Stanimirovic (University of Nis, Serbia) and Constantine Tsinakis (Vanderbilt University, USA).

Our Invited Speakers are Akbar Ali (University of Ha'il, Saudi Arabia), Mohsen Ghasemi (Urmia University, I. R. Iran), Gülistan Kaya Gök (Hakkari University. Hakkari-Turkey), Mohsen Kian (University of Bojnord, I. R. Iran), Ali Shukur (Belarusian State University, Belarus) and Ebrahim Reyhani (Shahid Rajaee Teacher Training University, I. R. Iran). The annual meeting of the Women's Committee of the Iranian Mathematical Society (WCIMS) will be started by the speech of professor Ashraf Daneshkhah, secretary of WCIMS. This meeting has professor Carolina Araujo as honorary guest. She is the Award Wiener of Ramanujan 2020, Brazil and vice president of the IMU committee for women. Professor Araujo will be presented an invited talk for AIMC 51 participants.

I am very thankful to all of my colleagues in Organizing and Scientific Committee and to all of participants. My special gratitude is going to the Keynote and Invited Speakers. I would also like to thank all the referees for the time they allocated and their help.

Chair of the Scientific Committee of AIMC51
Ali Reza Ashrafi

Conference Chair: Hassan Daghigh

Chair of Scientific Committee: Ali Reza Ashrafi

Chair of Organizing Committee: Mojtaba Bahramian

Members of Scientific Committee:

- Farshid Abdollahi — Shiraz University, I. R. Iran
- Saeid Alikhani — Yazd University, I. R. Iran
- Keyvan Amini — Razi University, I. R. Iran
- Saeid Azam — University of Isfahan, I. R. Iran
- Fariborz Azarpanah — Shahid Chamran University of Ahvaz, I. R. Iran
- Seyed Morteza Babamir — University of Kashan, I. R. Iran
- Mojtaba Bahramian — University of Kashan, I. R. Iran
- Rajab Ali Borzooei — Shahid Beheshti University , I. R. Iran
- Hassan Daghigh — University of Kashan, I. R. Iran
- Kinkar Chandra Das — Sungkyunkwan University, South Korea
- Mostafa Davtalab Olyaie — University of Kashan, I. R. Iran
- Bijan Davvaz — Yazd University, I. R. Iran
- Mohammad Ali Dehghan — Vali-e-Asr University of Rafsanjan, I. R. Iran
- Mahdi Dehghani — University of Kashan, I. R. Iran
- Sohrab Effati — Ferdowsi University Of Mashhad, I. R. Iran
- Ali Eftekhari — University of Kashan, I. R. Iran
- Taraneh Eghlidos — Sharif University of Technology, I. R. Iran
- Hossein Eshraghi — University of Kashan, I. R. Iran
- Gholamhosein Fathtabar — University of Kashan, I. R. Iran
- Majid Gazor — Isfahan University of Technology, I. R. Iran
- Faranak Goodarzi — University of Kashan, I. R. Iran
- Zahra Gooya — Shahid Beheshti University, I. R. Iran
- Massoud Hadian — Iran University of Science and Technology, I. R. Iran
- Masoud Hajarian — Shahid Beheshti University, I. R. Iran
- Ebrahim Hashemi — Shahrood University of Technology, I. R. Iran
- Ali Iranmanesh — Tarbiat Modares University, I. R. Iran
- Reza Jahani-Nezhad — University of Kashan, I. R. Iran
- Reza Kahkeshani — University of Kashan, I. R. Iran
- Vilmos Katona — University of Sopron, Hungary
- Seyed Mohammad Bagher Kashani — Tarbiat Modares University, I. R. Iran
- Rasool Kazemi Najafabadi — University of Kashan, I. R. Iran
- Stefan Kohl — University of St Andrews, Scotland
- Alireza Medghalchi — Kharazmi University, I. R. Iran
- Morteza Moniri — Shahid Beheshti University, I. R. Iran
- Ali Madanshekaf — University of Semnan, I. R. Iran
- Fereshteh Malek — KNT University of Technology, I. R. Iran
- Akbar Mohebi — University of Kashan, I. R. Iran
- Seyfollah Mosazadeh — University of Kashan, I. R. Iran
- Mark Raheb Ghamsary — Loma Linda University, USA
- Farhad Rahmati — Amirkabir University of Technology, I. R. Iran
- Abdolrahman Razani — Imam Khomeini International University, I. R. Iran
- Aliasghar Rezaei — University of Kashan, I. R. Iran
- Abbas Saadatmandi — University of Kashan, I. R. Iran
- Abbas Salemi Parizi — Shahid Bahonar University of Kerman, I. R. Iran
- Mehdi Shams — University of Kashan, I. R. Iran
- Seyyed Mansour Vaezpour — Amirkabir University of Technology, I. R. Iran

Members of Policy Conucil

- Ali Reza Ashrafi
- Mojtaba Bahramian
- Behnam Bazigaran
- Hassan Daghigh
- Ali Eftekhari
- Reza Jahani-Nezhad
- Ruhollah Jahanipur
- Akbar Mohebbi
- Amir Hossien Nokhodkar
- Ali Asghar Rezaei

Members of Organizing Committee

- Saeed Asaeedi
- Mahdi Asadi
- Jalal Askari Farsangi
- Morteza Bisheh-Niasar
- Mohammad Eghbali
- Ali Eftekhari
- Mardjan Hakimi-Nezhaad
- Elahe Khaldi
- Abolfazl Khedmati
- Mohammad Hassan Malekian
- Seyyed Ali Mohammadiyeh
- Marzieh Pourbabaee
- Mahdi Sabzevari
- Zeinab Saeedian Tarei
- Zeinab Soltani
- Fatemeh Zabihi

Other people who helped organizing the conference:

Ladies: Maryam Azizi, Narges Barzegran, Leila Goodarzi, Elham Hajirezaei, Shirin Heidari, Marzieh Sadat Hosseini, Zeinab Jafari Tadi, Nazila Jahangir, Sheyda Maddah, Elahe Mahabadian, Nasrin Malek-Mohammadi Faradonbeh, Faezeh Mohammadi, Maryam Nasr-Esfahani, Mohadeseh Nasr-Esfahani, Mahsa Rafiee, Maryam Rezaei Kashi, Mina Shafouri, Maryam Taheri-Sedeh, Ghazal Tavakoli, Armina Zare, Samaneh Zareian

Gentlemen: Jalal Abbassi, Mahdi Abedi, Ali Ghalavand, Mohammad Izadi, Bardia Jahangiri, Mostafa Karbalaei Reza, Ali Reza Khalilian, Kourosh Mavaddat-Nezhad, Sajad Raahati, Mohsen Yaghoubi

Keynote Speakers

	Name	Family	Affiliation
1	Alireza	Abdollahi	University of Isfahan, I. R. Iran
2	Javad	Asadollahi	University of Isfahan, I. R. Iran
3	Masoud	Asgharian	McGill Univesity, Canada
4	Mohammad	Bagheri	Editor in chief of the Journal of the History of Science, I. R. Iran
5	Khodakhast	Bibak	Miami University, USA
6	Alain	Bretto	Normandie University, France
7	Maurizio	Brunetti	Universita Federico II, Italy
8	Henri	Darmon	McGill University, Canada
9	Luca	De Feo	University of Versailles, Switzerland
10	Tomislav	Došlić	University of Zagreb, Croatia
11	Roberto	Garrappa	Polytechnic University of Bari, Italy
12	Zahra	Gouya	Shahid Beheshti University, I. R. Iran
13	Nezam	Mahdavi-Amiri	Sharif University of Technology, I. R. Iran
14	Javad	Mashreghi	University of Laval, Canada
15	Davoud	Mirzaei	University of Isfahan, I. R. Iran
16	Kees	Roos	Technical University Delf, Netherland
17	Mohammad	Sal Moslehian	Ferdowsi University of Mashhad, I. R. Iran
18	Thekiso Trevor	Seretlo	University of Limpopo, South Africa
19	Khalil	Shafie	University of Northern Colorado, USA
20	Omid Ali	Shehni-Karamzadeh	Shahid Chamran University of Ahvaz, I. R. Iran
21	Mohammad	Shahryari	Sultan Qaboos University, Muscat, Oman
22	Majid	Soleimani-Damaneh	University of Tehran, I. R. Iran
23	Andrea	Solotar	Universidad de Buenos Aires, Argentina
24	Ahmad Reza	Soltani	Kuwait University, Kuwait
25	Predrag	Stanimirović	University of Nis, Serbia
26	Teerapong	Suksumran	Chiang Mai University, Thailand
27	Mukut Mani	Tripathi	Banaras Hindu University, India
28	Constantine	Tsinakis	Vanderbilt University, USA
29	Andrei	Vesnin	Tomsk State University, Russia
30	Changchang	Xi	Capital Normal University, China
31	Bijan	Zohuri-Zangeneh	Sharif University of Technology, I. R. Iran

Invited Speakers

	Name	**Family**	**Affiliation**
1	Akbar	Ali	University of Hail, Saudi Arabia
2	Mohsen	Ghasemi	Urmia University, I. R. Iran
3	Gülistan	Kaya Gök	Hakkari University, Turkey
4	Mohsen	Kian	University of Bojnord, I. R. Iran
5	Ebrahim	Reihani	Shahid Rajaee Teacher Training University, I. R. Iran
6	Ali	Shukur	Belarusian State University, Belarus; The Islamic University, Iraq

Conference Participants

	First Name	Last Name	University
1	Naser	Abbasi	Lorestan University
2	Mostafa	Abbaszadeh	Amirkabir University of Technology
3	Fakhralsadat	Abdenean	Yazd University
4	Me'raj	Abdi	Bam University
5	Nasim	Abdi Kourani	Khajeh Nasir Toosi University of Technology
6	Atefeh	Abdolah Abyaneh	Kharazmi University
7	Alireza	Abdollahi	University of Isfahan
8	Farshid	Abdollahi	Shiraz University
9	Fahimeh	Abdollahi	Khajeh Nasir Toosi University of Technology
10	Alma	Abedinzadeh	University of Tehran
11	Mohammed Yahya	Abed	University of Kerbala, Iraq
12	Mahdi	Abedei	Shahid Bahonar University of Kerman
13	Marjan	Adib	Payame Noor University
14	Fatemeh Sadat	Aghaei Maybodi	Yazd University
15	Fatemeh	Ahangari	Al-Zahra University
16	Alireza	Ahmadi	Yazd University
17	Kambiz	Ahmadi	University of Shahrekord
18	Ghasem	Ahmadi	Payame Noor University
19	Razieh	Ahmadian	IPM Institute For Research In Fundamental Sciences
20	Mohammad Ali	Ahmadpoor	University of Guilan
21	Zohreh	Akbari	University of Mazandaran
22	Najmeh	Akbari	Isfahan University of Technology
23	Fahime	Akhavan Ghassabzade	University of Gonabad
24	Narges	Akhlaghinia	Shahid Beheshti University
25	Basim	Albuohimad	University of Kerbala, Iraq
26	Akbar	Ali	University of Hail, Saudi Arabia
27	Mahdi	Aliakbari	Torbat Heydarieh University
28	Ghazale	Aliasghari	Shahid Rajaee Teacher Training University
29	Saeid	Alikhani	Yazd University
30	Hajar	Alimorad	Jahrom University
31	Mohammad Reza	Alimoradi	Malayer University
32	Morteza	Alishahi	Islamic Azad University
33	Ahmed	Al-Obaidi	University of Kufa, Iraq
34	Keyvan	Amini	Razi University
35	Mostafa	Amini	Payame Noor University
36	Diba	Aminshayan Jahromi	Shiraz University
37	Letafat	Amiri	Tarbiat Modares University
38	Sadegh	Amiri	Shahid Sattari Aeronautical University
39	Hanieh	Amjadian	Amirkabir University of Technology
40	Mahdi	Anbarloei	Imam Khomeini International University
41	Hajar	Ansari	Amirkabir University of Technology
42	Ali	Ansari Ardali	University of Shahrekord
43	Fereshteh	Arad	Shahid Bahonar University of Kerman
44	Mahdi	Asadi	University of Kashan

Conference Participants

	First Name	Last Name	Affiliation
45	Mohammad Ali	Asadi	Islamic Azad University
46	Mohammad Bagher	Asadi	University of Tehran
47	Meysam	Asadipour	Yasouj University
48	Javad	Asadollahi	University of Isfahan
49	Saeed	Asaeedi	University of Kashan
50	Masoud	Asgharian	McGill Univesity, Canada
51	Ali Reza	Ashrafi	University of Kashan
52	Jalal	Askari Farsangi	University of Kashan
53	Hamed	Aslani	University of Guilan
54	Parvane	Atashpeykar	University of Bonab
55	Ahmad Reza	Attari Polsangi	Shiraz University
56	Mehrasa	Ayatollahi	Payame Noor University
57	Saeid	Azam	University of Isfahan
58	Mahdieh	Azari	Islamic Azad University
59	Fariborz	Azarpanah	Shahid Chamran University of Ahvaz
60	Seyed Morteza	Babamir	University of Kashan
61	Mohammad	Bagheri	Editor in chief of the Journal of the History of Science
62	Neda	Bagheri	University of Mazandaran
63	Karam	Bahari	Razi University
64	Shima	Baharlouei	Isfahan University of Technology
65	Erfan	Bahmani	University of Zanjan
66	Faezeh	Bahmani	University of Kashan
67	Mojtaba	Bahramian	University of Kashan
68	Fariba	Bakrani	Shahid Beheshti University
69	Seddigheh	Banihashemi	University of Mazandaran
70	Narjes Sadat	Banitaba	Yazd University
71	Ali	Barani	Lorestan University
72	Ali	Barati	Razi University
73	Hasan	Barsam	University of Jiroft
74	Ali	Barzanouni	Hakim Sabzevari University
75	Esmaeil	Bashkar	Velayat university
76	Mostafa	Bayat	Amirkabir University of Technology
77	Behnam	Bazigaran	University of Kashan
78	Fatemeh	Bazikar	University of Guilan
79	Fereshteh	Behboudi	Imam Khomeini International University
80	Reza	Beyranvand	Lorestan University
81	Khodakhast	Bibak	Miami University, USA
82	Morteza	Bisheh-niasar	University of Kashan
83	Rajab Ali	Borzooei	Shahid Beheshti University
84	Ali	Bozorgmehr	Iran University of Medical Sciences
85	Alain	Bretto	Normandie University, France
86	Maurizio	Brunetti	Universita Federico II, Italy
87	Kinkar	Chandra Das	Sungkyunkwan University, South Korea
88	Mehran	Chehlabi	Islamic Azad University
89	Abbas	Cheraghi	University of Isfahan
90	Fatemeh	Choopani	Ferdowsi University of Mashhad
91	Mohammadehsan	Dadkani	University of Sistan and Baluchestan

Conference Participants

	First Name	Last Name	Affiliation
92	Hassan	Daghigh	University of Kashan
93	Mohammadreza	Darafsheh	University of Tehran
94	Henri	Darmon	McGill University, Canada
95	Razie	Darvazeban Zade	Payame Noor University
96	Mahshid	Dashti	Malayer University
97	Zahra	Davari Shalamzari	Yazd University
98	Mostafa	Davtalab Olyaie	University of Kashan
99	Bijan	Davvaz	Yazd University
100	Luca	De Feo	University of Versailles, Switzerland
101	Mohammad Ali	Dehghan	Vali-e-Asr University of Rafsanjan
102	Sakineh	Dehghan	Shahid Beheshti University
103	Mahdi	Dehghani	University of Kashan
104	Fatemeh	Dehghani	Yazd University
105	Najmeh	Dehghani	Persian Gulf University
106	Zahra	Dehvari	Yazd University
107	Atefeh	Deris	Arak University
108	Zahra	Donyari	Shahid Chamran University of Ahvaz
109	Reza	Doostaki	Shahid Bahonar University of Kerman
110	Saeed	Doostali	University of Kashan
111	Fateme	Dorri	Ferdowsi University of Mashhad
112	Tomislav	Došlić	University of Zagreb, Croatia
113	Ghodrat	Ebadi	Tabriz University
114	Javad	Ebadpour Golanbar	Payame Noor University
115	Neda	Ebrahimi	Shahid Bahonar University of Kerman
116	Ali	Ebrahimijahan	Amirkabir University of Technology
117	Asiyeh	Ebrahimzadeh	Farhangian University
118	Sohrab	Effati	Ferdowsi University of Mashhad
119	Ali	Eftekhari	University of Kashan
120	Leila	Eftekhari	Tarbiat Modares University
121	Mohammad	Eghbali	University of Kashan
122	Taraneh	Eghlidos	Sharif University of Technology
123	Hossein	Eshraghi	University of Kashan
124	Mohammad Reza	Eslahchi	Tarbiat Modares University
125	Morteza	Essmaili	Kharazmi University
126	Masoumeh	Etebar	Shahid Chamran University of Ahvaz
127	Amin	Faghih	Sahand University of Technology
128	Farhad	Fakhar-Izadi	Amirkabir University of Technology
129	Farahnaz	Fakhraddin Arani	Nongovermental Collage Nonprofit Refah
130	Hamid	Faraji	Islamic Azad University
131	Farzaneh	Farhang Baftani	Islamic Azad University
132	Mohammad Reza	Farmani	Kharazmi University
133	Javad	Farokhi-Ostad	Birjand University of Technology
134	Fariba	Fayazi	University of Qom
135	Fateme	Fasihi	Bu-Ali Sina University of Hamedan
136	Gholam Hossien	Fathtabar Firouzjae	University of Kashan
137	Reza	Fayazi	Ferdowsi University of Mashhad
138	Mohamad Javad	Fazeli	University of Birjand
139	Fereshteh	Forouzesh	Bam University
140	Saba	Fotouhi	Amirkabir University of Technology
141	Batoul	Ganji Saffar	Al-Zahra University

Conference Participants

	First Name	Last Name	Affiliation
142	Roberto	Garrappa	Polytechnic University of Bari, Italy
143	Majid	Gazor	Isfahan University of Technology
144	Somayeh	Ghadamyari	University of Sistan and Baluchestan
145	Mansour	Ghadiri	Yazd University
146	Ali	Ghafarpanah	Salman Farsi University of Kazerun
147	Ali	Ghalavand	University of Kashan
148	Fatemeh	Ghanadian	Damghan University
149	Fatemeh	Ghandi	University of Kashan
150	Mohammad Reza	Ghanei	University of Khansar
151	Hadi	Ghasemi	Hakim Sabzevari University
152	Mohsen	Ghasemi	Urmia University
153	Mohammad Hesam	Ghasemi	Shahid Beheshti University
154	Peyman	Ghiasvand	Payame Noor University
155	Hamid	Ghorbani	University of Kashan
156	Ali Reza	Ghorchizadeh	University of Birjand
157	Azin	Golbaharan	Kharazmi University
158	Faranak	Goodarzi	University of Kashan
159	Leila	Goodarzi	University of Kashan
160	Zahra	Gooya	Shahid Beheshti University
161	Farzaneh	Gorjizadeh	University of Shahrekord
162	Punam	Gupta	Dr. Harisingh Gour University, India
163	Mahnaz	Habibi	Islamic Azad University
164	Ali	Habibi Moakher	Payame Noor University of Tehran
165	Ali	Habibirad	Shiraz University of Technology
166	Masoud	Hadian Dehkordi	Iran University of Science and Technology
167	Amir Hosein	Hadian Rasanan	Shahid Beheshti University
168	Armin	Hadjian	University of Bojnord
169	Somayeh	Hadjirezaei	Vali-e-Asr University of Rafsanjan
170	Donya	Haghighi	Imam Khomeini International University
171	Saeid	Haghjoo	Shahid Rajaee Teacher Training University
172	Narges	Haj Aboutalebi	Islamic Azad University
173	Masoud	Hajarian	Shahid Beheshti University
174	Ashraf	Haji Olov Zarnagh	University of Kashan
175	Elham	Hajirezaei	University of Kashan
176	Hamid Reza	Hajisharifi	University of Khansar
177	Nooshin	Hakamipour	Buein Zahra Technical University
178	Mardjan	Hakimi-Nezhaad	Shahid Rajaee Teacher Training University
179	Shahad	Hasan	University of Kufa, Iraq
180	Farzane	Hashemi	University of Kashan
181	Ebrahim	Hashemi	Shahrood University of Technology
182	Mehdi	Hassani	University of Zanjan
183	Mostafa	Hassanlou	Urmia University
184	Marziyeh	Hatamkhani	University of Arak
185	Sina	Hedayatian	Shahid Chamran University of Ahvaz
186	Dariush	Heidari	Mahallat Institute of Higher Education
187	Mohammad	Heidari	Kharazmi University

Conference Participants

	First Name	Last Name	Affiliation
188	Samira	Heidari	Imam Khomeini International University
189	Saghar	Heidari	Shahid Beheshti University
190	Azam	Hejazi Noghabi	Ferdowsi University of Mashhad
191	Mohammad	Hemami	Shahid Beheshti University
192	Esmail	Hesameddini	Shiraz University of Technology
193	Abdolaziz	Hesari	Shahid Chamran University of Ahvaz
194	Rasoul	Heydari Dastjerdi	Payame Noor University
195	Seyedeh Mahya	Hosseini	Payame Noor University
196	Zahra Sadat	Hosseini	Bu-Ali Sina University of Hamedan
197	Hasan	Hosseinzadeh	Islamic Azad University
198	Mohammad	Ilati	Sahand University of Technology
199	Mohammad Ali	Iranmanesh	Yazd University
200	Ali	Iranmanesh	Shahid Bahonar University of Kerman
201	Ali	Iranmanesh	Tarbiat Modares University
202	Marzieh	Izadi	Shahid Bahonar University of Kerman
203	Mehdi	Izadi	Shahid Rajaee Teacher Training University
204	Javad	Izadi	Payame Noor University
205	Mohammad Mahdi	Izadkhah	Birjand University of Technology
206	Kiyana	Izadyar	Shahid Chamran University of Ahvaz
207	Mehsin	Jabel Atteya	University of Al-Mustansiriyah, Iraq
208	Nasrin	Jafari	Yazd University
209	Habibollah	Jafari	Islamic Azad University
210	Mohammad	Jafari	Islamic Azad University
211	Hosna	Jafarmanesh	Hakim Sabzevari University
212	Nafisehsadat	Jafarzadeh	Tarbiat Modares University
213	Mehdi	Jahangiri	University of Maragheh
214	Reza	Jahani-Nezhad	University of Kashan
215	Ruhollah	Jahanipur	University of Kashan
216	Marziye	Jamali	University of Kashan
217	Sedighe	Jamshidvand	Khajeh Nasir Toosi University of Technology
218	Mohsen	Jannesari Ladani	Shahreza Higher Education Center
219	Elham	Javidmanesh	Ferdowsi University of Mashhad
220	Saeed	Johari	University of Isfahan
221	Farangis	Johari	Universidade Federal de Minas Gerais, Brazil
222	Maryam	Joulaei	Islamic Azad University
223	Alireza	Kabgani	Urmia University of Technology
224	Akram	Kabiri Samani	Payame Noor University
225	Azam	Kaheni	University of Birjand
226	Reza	Kahkeshani	University of Kashan
227	Zahra	Kamali	Islamic Azad University
228	Gholamreza	Karamali	Iran University of Science and Technology
229	Elaheh	Karimi	Islamic Azad University of Bushehr
230	Elham	Karimi	Al-Zahra University
231	Sajed	Karimy	Sharif University of Technology
232	Mohammadreza	Karimzadeh	University of Maragheh

Conference Participants

	First Name	Last Name	Affiliation
233	Seyed Mohammad Bagher	Kashani	Tarbiat Modares University
234	Roghayeh	Katani	Yasouj University
235	Vilmos	Katona	University of Sopron, Hungary
236	Gülistan	Kaya Gök	Hakkari University, Turkey
237	Mohammad Bagher	Kazemi	University of Zanjan
238	Kianoush	Kazemi	University of Birjand
239	Ramin	Kazemi	Imam Khomeini International University
240	Rasool	Kazemi Najafabadi	University of Kashan
241	Seyed Mehdi	Kazemi Torbaghan	University of Bojnord
242	Vahid	Keshavarz	Shiraz University of Technology
243	Niloufar	Keshavarz	Persian Gulf University
244	Zahra	Keshtkar	Shahid Chamran University of Ahvaz
245	Maryam	Keyvani Maraghi	University of Maragheh
246	Mahmood	Khaksar-e Oshagh	Dr. Masaheb Institute of Mathematical Research
247	Somayeh	Khalashi Ghezelahmad	Islamic Azad University
248	Elahe	Khaldi	University of Kashan
249	Ghader	Khaledi	Payame Noor University
250	Mohsen	Khaleghi Moghadam	Sari Agricultural Sciences and Natural Resources University
251	Yasser	Khalili	Sari Agricultural Sciences and Natural Resources University
252	Alireza	Khalili Asboei	Farhangian University
253	Amir	Khamseh	Kharazmi University
254	Seyed Mohammad Amin	Khatami	Birjand University of Technology
255	Abolfazl	Khedmati	University of Kashan
256	Ekhtiar	Khodadadi	Islamic Azad University
257	Hamid	Khodaei	Malayer University
258	Davod	Khojasteh Salkuyeh	University of Guilan
259	Hassan	Khosravi	Gonbad Kavous University
260	Eisa	Khosravi Dehdezi	Persian Gulf University
261	Mohsen	Kian	University of Bojnord
262	Stefan	Kohl	University of St Andrews, Scotland
263	Masoumeh	Koohestani	Khajeh Nasir Toosi University of Technology
264	Majid	Kowkabi	University of Gonabad
265	Zeinab	Kowsari	Kharazmi University
266	Behnaz	Lajmiri	Amirkabir University of Technology
267	Sanaz	Lamei	University of Guilan
268	Seyed Jalal	Langari	Farhangian University
269	Samira	Latifi	University of Mohaghegh Ardabili
270	Ehsan	Lotfali Ghasab	Shahid Chamran University of Ahvaz
271	Maryam	Lotfipour	Fasa University

Conference Participants

	First Name	Last Name	Affiliation
272	Abbas	Maarefparvar	IPM Institute For Research In Fundamental Sciences
273	Ali	Madanshekaf	University of Semnan
274	Nezam	Mahdavi-Amiri	Sharif University of Technology
275	Soheila	Mahdavi Zafarghandi	University of Kashan
276	Ali	Mahdipoor	University of Kashan
277	Zahra	Mahmoodi	Islamic Azad University
278	Mojgan	Mahmoudi	Shahid Beheshti University
279	Roya	Makrooni	University of Sistan and Baluchestan
280	Fereshteh	Malek	Khajeh Nasir Toosi University of Technology
281	Hassan	Maleki	Malayer University
282	Mohammad Hassan	Malekian	University of Kashan
283	Somayeh	Malekinejad	Payame Noor University
284	Sepideh	Maleki-Roudposhti	University of Guilan
285	Nasrin	Malek-Mohammadi	University of Kashan
286	Maryam	Malekpour	Al-Zahra University
287	Mahdiyeh	Manavi	Khajeh Nasir Toosi University of Technology
288	Mukut	Mani Tripathi	Banaras Hindu University, India
289	Hossien	Mashayekhi	University of Kashan
290	Javad	Mashreghi	University of Laval, Canada
291	Maryam	Masoudi Arani	Technical and Vocational University
292	Iman	Masoumi	Tafresh University
293	Kurosh	Mavaddat Nezhaad	University of Kashan
294	Majid	Mazrooei	University of Kashan
295	Alireza	Medghalchi	Kharazmi University
296	Hussain	Mehdi	University of Kufa, Iraq
297	Elahe	Mehraban	University of Guilan
298	Samira	Mehrangiz	Shiraz University
299	Hamid	Mehravaran	Islamic Azad University
300	Sadegh	Merati	Shiraz University
301	Ali	Mesforush	Shahrood University of Technology
302	Azar	Mirzaei	Razi University
303	Davoud	Mirzaei	University of Isfahan
304	Fatemeh	Mirzaei	Payame Noor University
305	Fatemeh	Mirzaei Gaskarei	Islamic Azad University
306	Mahsa	Mirzargar	Mahallat Institute of Higher Education
307	Mohammad Mahdi	Moayeri	Shahid Beheshti University
308	Alireza	Mofidi	Amirkabir University of Technology
309	Amir Abbas	Mofidian Naeini	Isfahan University of Technology
310	Hoda	Mohammadi	Payame Noor University
311	Maryam	Mohammadi	Isfahan University of Technology
312	Shahnaz	Mohammadi	Tabriz University
313	Reza	Mohammadiarani	Amirkabir University of Technology
314	Seyyed Ali	Mohammadiyeh	University of Kashan
315	Zahra	Mohammadzadeh	University of Birjand
316	Mahdi	Mohammadzadeh Karizaki	Torbat Heydarieh University

Conference Participants

	First Name	Last Name	Affiliation
317	Akbar	Mohebbi	University of Kashan
318	Mina	Moini	University of Malayer
319	Zahra	Mohtasham	Shahid Beheshti University
320	Reza	Mokhtari	Isfahan University of Technology
321	Tahereh	Molaee	Al-Zahra University
322	Mahdieh	Molaeiderakhtenjani	University of Birjand
323	Ehsan	Momtahan	Yasouj University
324	Morteza	Moniri	Shahid Beheshti University
325	Mansooreh	Moosapoor	Farhangian University, Bentolhoda Sadr
326	Rasoul	Moradi	Persian Gulf University
327	Sirous	Moradi	Lorestan University
328	Ali	Moradzadeh-Dehkordi	Shahreza Higher Education Center
329	Seyed Adel	Moravveji	Laval University, Canada
330	Maysam	Mosadeq	Islamic Azad University
331	Seyfollah	Mosazadeh	University of Kashan
332	Zohreh	Mostaghim	Iran University of Science and Technology
333	Marziyeh	Motahari	Tarbiat Modares University
334	Hamid	Mousavi	University of Tabriz
335	Fatemeh Sadat	Mousavinejad	Yazd University
336	Ehsan	Movahednia	Behbahan Khatam Alanbia University of Technology
337	Kamran	Musazadeh	Islamic Azad University
338	Mohammad Javad	Nadjafi-Arani	Mahallat Institute of Higher Education
339	Razieh	Naghibi	Yazd University
340	Mohammadali	Naghipoor	Jahrom University
341	Reza	Naghipour	Kharazmi University
342	Mehran	Naghizadeh Qomi	University of Mazandaran
343	Alireza	Najafizadeh	Payame Noor University
344	Maryam	Najafvand Derikvandi	University of Kashan
345	Mehran	Namjoo	Vali-e-Asr University of Rafsanjan
346	Seyed Mojtaba	Naser Sheykholislami	Semnan University
347	Nasim	Nasrabadi	University of Birjand
348	Zohreh	Nazari	Vali-e-Asr University of Rafsanjan
349	Ali Mohammad	Nazari	University of Arak
350	Tahere	Nazari	Payame Noor University
351	Ali	Naziri-Kordkandi	Payame Noor University
352	Behzad	Nemati Saray	Institute for Advanced Studies in Basic Sciences
353	Mohsen	Niazi	University of Birjand
354	Sogol	Niazian	Islamic Azad University
355	Zohre	Nikooravesh	Birjand University of Technology
356	Ashkan	Nikseresht	Shiraz University
357	Fateme	Nikzad	Payame Noor University
358	Amir Hossein	Nokhodkar	University of Kashan
359	Monireh	Nosrati	University of Bonab
360	Bahareh	Nouri	Kharazmi University

Conference Participants

	First Name	Last Name	Affiliation
361	Mohammad Reza	Oboudi	Shiraz University
362	Jafar	Ojbag	Azarbaijan Shahid Madani University
363	Fateme	Olia	Khajeh Nasir Toosi University of Technology
364	Reza	Orfi	Kharazmi University
365	Fatemeh	Parishani	University of Isfahan
366	Mehdi	Parsinia	Shahid Chamran University of Ahvaz
367	Rohollah	Parvinianzadeh	Yasouj University
368	Leila	Pedram	Imam Khomeini International University
369	Sajjad	Piradl	Payame Noor University
370	Vahid	Pirhadi	University of Kashan
371	Shima	Pirmohammadi	University of Isfahan
372	Mina	Pirzadeh	University of Guilan
373	Marzieh	Pourbabaee	University of Kashan
374	Hossein	Pourbashash	University of Garmsar
375	Alireza	Pourmoslemi	Payame Noor University
376	Maryam	Rabiee Farahani	Ferdowsi University of Mashhad
377	Farzad	Radmehr	Western Norway University of Applied Sciences, Norway
378	Majed	Raeesi	University of Sistan and Baluchestan
379	Marzieh	Raei	Malek Ashtar University of Technology
380	Ahmadreza	Raeisi Dehkordi	University of Isfahan
381	Mark	Raheb Ghamsary	Loma Linda University, USA
382	Saeid	Rahimi	Persian Gulf University
383	Morteza	Rahimi Khorzughi	University of Tehran
384	Parisa	Rahimkhani	Al-Zahra University
385	Gholamreza	Rahimlou	Technical and Vocational University
386	Mohammad Hossein	Rahmani Doust	University of Neyshabur
387	Hormoz	Rahmatan	Payame Noor University
388	Farhad	Rahmati	Amirkabir University of Technology
389	Marzieh	Rahmati	Payame Noor University
390	Ali	Rajaei	Tarbiat Modares University
391	Sayyed Mehrab	Ramezani	Yasouj University
392	Pardis	Ramezani	Payame Noor University
393	Jalil	Rashidinia	Iran University of Science and Technology
394	Abdolrahman	Razani	Imam Khomeini International University
395	Ebrahim	Reihani	Shahid Rajaee Teacher Training University
396	Parisa	Rezaei	University of Sistan and Baluchestan
397	Ali Asghar	Rezaei	University of Kashan
398	Akbar	Rezaei	Payame Noor University
399	Maryam	Rezaei Kashi	University of Kashan
400	Reza	Rezavand	University of Tehran
401	Monireh	Riahi	Damghan University
402	Mehdi	Riazi-Kermani	Wichita State University, USA
403	Mohammad Saber	Roohi	Shahid Beheshti University

Conference Participants

	First Name	Last Name	Affiliation
404	Kees	Roos	Technical University Delft, Netherland
405	Nazanin	Roshandel Tavana	Amirkabir University of Technology
406	Mehdi	Rostami	Amirkabir University of Technology
407	Salimeh	Rostami	Yazd University
408	Esmaeil	Rostami	Shahid Bahonar University of Kerman
409	Mohsen	Rostamian Delavar	University of Bojnord
410	Arta	Rouhi	Semnan University
411	Reza	Saadati	Iran University of Science and Technology
412	Maryam	Saadati	Imam Khomeini International University
413	Abbas	Saadatmandi	University of Kashan
414	Rahele	Sabagh	University of Sistan and Baluchestan
415	Samaneh	Saberali	Urmia University
416	Sedigheh	Sabermahani	Al-Zahra University
417	Mehdi	Sabzevari	University of Kashan
418	Somayeh	Sadeghi	University of Isfahan
419	Behruz	Sadeqi	Payame Noor University
420	Nasrin	Sadri	IPM Institute For Research In Fundamental Sciences
421	Farhad	Saeediyoun	Amirkabir University of Technology
422	Hojatollah	Saeidi	University of Shahrekord
423	Zeinab	Saeidian	University of Kashan
424	Jamshid	Saeidian	Kharazmi University
425	Ali	Safaie	University of Maragheh
426	Farzaneh	Safari	Imam Khomeini International University
427	Akram	Safari-Hafshejani	Payame Noor University
428	Marziyeh	Saffarian	University of Kashan
429	Maryam	Saghalorzadeh	Jundi-Shapur University of Technology
430	Zahra	Sajjadnia	Shiraz University
431	Saman	Saki	Iran University of Science and Technology
432	Mohammad	Sal Moslehian	Ferdowsi University of Mashhad
433	Worod	Salah	University of Kufa, Iraq
434	Mohammad Ali	Salahshour	University of Kashan
435	Alireza	Salehi	Petroleum University of Technology
436	Abbas	Salemi Parizi	Shahid Bahonar University of Kerman
437	Hamid	Salimi	University of Tehran
438	Mahdi	Salmanpour	University of Kashan
439	Nasrin	Samadyar	Al-Zahra University
440	Mohammad Esmael	Samei	Bu-Ali Sina University
441	Karim	Samei	Bu-Ali Sina University
442	Amir Hossein	Sanatpour	Kharazmi University
443	Samaneh	Saneifar	Yazd University
444	Behnaz	Savizi	Islamic Azad University
445	Yamin	Sayyari	Sirjan University of Technology
446	Khadijeh	Sayyari	Kharazmi University
447	Salameh	Sedaghat	Buein Zahra Technical University
448	Monireh	Sedghi	University of Tabriz

Conference Participants

	First Name	Last Name	Affiliation
449	Somayeh	Seifollahzadeh	University of Tabriz
450	Thekiso	Seretlo	University of Limpopo, South Africa
451	Zahra	Seyedi Lahrodi	Tarbiat Modares University
452	Saeed	Shaabanian	Tarbiat Modares University
453	Mohsin	Shaalan Abdulhussein Alakaashi	University of Kufa, Iraq
454	Zahra	Shabani Siahkalde	University of Sistan and Baluchestan
455	Mehrnoosh	Shadravan	Shahid Beheshti University
456	Khalil	Shafie	University of Northern Colorado, USA
457	Negur	Shehani Karamzadeh	Shahid Beheshti University
458	Amin	Shahkarami	Lorestan University
459	Hakimeh	Shahriaripour	Kerman Graduate University of Technology
460	Mohammad	Shahryari	Sultan Qaboos University, Muscat, Oman
461	Maryam	Shahsiah	University of Isfahan
462	Mina	Shamgani	University of Kashan
463	Mehdi	Shams	University of Kashan
464	Afsaneh	Shamsaki	Damghan University
465	Alireza	Shamsian	Institute for Advanced Studies in Basic Sciences
466	Reza	Sharafdini	Persian Gulf University
467	Javad	Sharafi	University of Kashan
468	Seyedeh Fatemeh	Shariati	Amirkabir University of Technology
469	Kamran	Sharifi	Shahroud University of Technology
470	Farzad	Shaveisi	Razi University
471	Omid Ali	Shehni-Karamzadeh	Shahid Chamran University of Ahvaz
472	Marjan	Sheibani Abdolyousefi	Semnan University
473	Hayder Baqer Ameen	Shelash	University of Kufa, Iraq
474	Efat	Shikhi	Payam Noor University
475	Maryam	Shirali	Shahid Chamran University of Ahvaz
476	Nasrin	Shirali	Shahid Chamran University of Ahvaz
477	Parisa	Shiri	Sahand University of Technology
478	Farrokh	Shirjian	Tarbiat Modares University
479	Hasan	Shlaka	University of Kufa, Iraq
480	Shirin	Shoaee	Shahid Beheshti University
481	Raheleh	Shokrpour	University of Tabriz
482	Ali	Shukur	Belarusian State University, Belarus; The Islamic University, Iraq
483	Amirhossein	Sobhani	Semnan University
484	Mahsa	Soheil Shamaee	University of Kashan
485	Mahdi	Sohrabi-Haghighat	University of Arak
486	Maryam	Soleimani	IPM Institute For Research In Fundamental Sciences
487	Majid	Soleimani-Damaneh	University of Tehran
488	Rasoul	Soleimani	Payame Noor University
489	Fazlollah	Soleymani	Institute for Advanced Studies in Basic Sciences
490	Mahdieh	Soleymani Baghsha	Sharif University of Technology

Conference Participants

	First Name	Last Name	Affiliation
491	Farnaz	Soliemany	Urmia University
492	Andrea	Solotar	Universidad de Buenos Aires, Argentina
493	Vali	Soltani Masih	Payame Noor University
494	Zeinab	Soltani	University of Kashan
495	Ahmad Reza	Soltani	Kuwait University, Kuwait
496	Sima	Soltani Renani	Isfahan University of Technology
497	Gholamreza	Soltaniabri	University of Tehran
498	Somayeh	Soltanpour	Petroleum University of Technology
499	Ghiyam	Soudan	Bu-Ali Sina University
500	Predrag	Stanimirović	University of Nis, Serbia
501	Teerapong	Suksumran	Chiang Mai University, Thailand
502	Hamid Reza	Tabrizidooz	University of Kashan
503	Mojgan	Taghavi	Shahid Beheshti University
504	Meysam	Taheri-Dehkordi	University of Kashan
505	Maryam	Taheri Sedeh	University of Kashan
506	Maryam	Tahmasbi	Shahid Beheshti University
507	Haleh	Tajadodi	University of Sistan and Baluchestan
508	Farkhondeh	Takhteh	Persian Gulf University
509	Ebrahim	Tamimi	Semnan University
510	Somayyeh	Tari	Azarbaijan Shahid Madani University
511	Mostafa	Tavakoli	Ferdowsi University of Mashhad
512	Mohammadreza	Tavakkoli Moghaddam	Shahid Beheshti University
513	Reza	Tayebi Khorami	Islamic Azad University
514	Atieh	Teymourzadeh	University of Mazandaran
515	Faezeh	Tiba	University of Sistan and Baluchestan
516	Abdolsaleh	Toghdori	Yazd University
517	Fateme	Torabi	Damghan University
518	Soraya	Torkaman	Yazd University
519	Vali	Torkashvand	Islamic Azad University
520	Thekiso	Trevor Seretlo	University of Limpopo, South Africa
521	Constantine	Tsinakis	Vanderbilt University, USA
522	Seyed Mansour	Vaezpour	Amirkabir University of Technology
523	Farzaneh	Vahdanipour	University of Mohaghegh Ardabili
524	Seryas	Vakili	University of Tabriz
525	Amir	Veisi	Yasouj University
526	Andrei	Vesnin	Tomsk State University, Russia
527	Changchang	Xi	Capital Normal University, China
528	Marjan	Yaghmaei	Kharazmi University
529	Mohamad	Yar Ahmadi	Shahid Chamran University of Ahvaz
530	Zahra	Yarahmadi	Islamic Azad University
531	Mohammad Reza	Yasamian	Payame Noor University
532	Azam	Yazdani	Amirkabir University of Technology
533	Mohsen	Yousefnezhad	Shiraz University
534	Omid	Zabeti	University of Sistan and Baluchestan
535	Sayyed Mahmood	Zabetzadeh	Payame Noor University
536	Fatemeh	Zabihi	University of Kashan
537	Amirhesam	Zaeim	Payame Noor University

Conference Participants

	First Name	Last Name	Affiliation
538	Hassan	Zaherifar	Yazd University
539	Elham	Zangiabadi	Vali-e-Asr University of Rafsanjan
540	Hossein	Zare	Tarbiat Modares University
541	Rohollah	Zarei	Yasouj University
542	Behnam	Zarpak	Shahed University
543	Maryam	Zeinali	Shahid Chamran University of Ahvaz
544	Ali	Zeydi Abdian	Lorestan University
545	Elham	Zeynal	Islamic Azad University
546	Mosayeb	Zohrehvand	Malayer University
547	Bijan	Zohuri-Zangeneh	Sharif University of Technology

List of Presented Papers

Keynotes and Invited Talks

On the Use of Matrix Mittag-Leffler Functions in Fractional Calculus: From Theory to Applications

Roberto Garrappa*
Department of Mathematics, Univeristy of Bari, Italy
and Marina Popolizio
Department of Electrical and Information Engineering, Polytechnic University of Bari, Italy

Abstract. The Mittag-Leffler function plays a fundamental role in fractional calculus. Its evaluation with matrix arguments has several important applications in control theory, solution of multi-term differential equations, systems of fractional differential equations and so on. After introducing the Mittag-Leffler function, its matrix extension and some of its major applications, we present here some practical methods for the computation of matrix ML functions based on the efficient numerical inversion of the Laplace transform.
Keywords: Mittag-Leffler function, Fractional derivative, Matrix function, Numerical computation.
AMS Mathematical Subject Classification [2010]: 33E12, 26A33, 65F60.

1. Introduction

Fractional calculus studies theory and applications of integrals and derivatives of non-integer order. It has been found that fractional-order differential systems are indeed more suitable to describe systems in which the action of some external source does not act instantaneously but is influenced by the whole history of the system.

Along the years, there have been proposed several ways to generalize integer-order integrals and derivatives to any real (i.e., fractional) order. One of the most attractive definitions is the one known as the fractional derivative of Dzhrbashyan-Caputo (often simply Caputo) defined for $0 < \alpha < 1$ as

$$ {}^{\mathrm{C}}\mathcal{D}_0^{\alpha} y(t) = \frac{1}{\Gamma(1-\alpha)} \int_0^t (t-\tau)^{-\alpha} y'(\tau) \mathrm{d}\tau, \quad t > 0, $$

where $\Gamma(z) = \int_0^{\infty} t^{z-1} \mathrm{e}^{-t} \mathrm{d}t$ is the Euler-Gamma function and provides a generalization of the factorial since $\Gamma(n+1) = n!$ whenever $n \in \mathbb{N}$. The importance of this definition is related to the possibility of coupling differential equations of fractional order (FDEs) by standard initial conditions of Cauchy type as in

$$ \begin{cases} {}^{\mathrm{C}}\mathcal{D}_0^{\alpha} y(t) = f(t, y(t)), \\ y(0) = y_0. \end{cases} $$

A fundamental role in the analysis and solution of FDEs is played by the Mittag-Leffler function. This function was introduced in 1904 by the Swedish

*Speaker

mathematician Magnus Gustaf Mittag-Leffler in the context of the analysis of divergent series but several years after its introduction it was recognized its great importance in fractional calculus. Nowadays it is often named as the "*Queen function of fractional calculus*" [9] since it has the same prominence of the exponential function (the king function) in integer-order calculus.

For complex parameters α and β, with $\Re(\alpha) > 0$, the ML function is defined by means of the series

$$E_{\alpha,\beta}(z) = \sum_{j=0}^{\infty} \frac{z^j}{\Gamma(\alpha j + \beta)}, \quad z \in \mathbb{C},$$

although for most of the applications it is sufficient to consider just real parameters α and β.

There are several reasons supporting the introduction, the study and the computation of the ML function in fractional calculus. One of the main reasons is that it is the eigenfunction of the Caputo fractional derivative. Indeed, it is immediate to see that ${}^{\mathrm{C}}\mathcal{D}_0^{\alpha} E_{\alpha,1}(t^{\alpha}\lambda) = \lambda E_{\alpha,1}(t^{\alpha}\lambda)$ for any real or complex λ. This property, which turns out useful in the study of stability of FDEs, can be exploited for the numerical computation of solution of FDEs and other related problems [3].

2. Mittag-Leffler with Matrix Arguments

The definition of the ML function can be extended in a straightforward way to matrix arguments as

$$E_{\alpha,\beta}(A) = \sum_{j=0}^{\infty} \frac{A^j}{\Gamma(\alpha j + \beta)},$$

where now $A \in \mathbb{C}^{n\times n}$ is any matrix with n rows and n columns and $E_{\alpha,\beta}(A)$ turns out to be a matrix of the same size. Matrix ML functions are useful, for instance, for solving semilinear systems of fractional order

$$\begin{cases} {}^{\mathrm{C}}\mathcal{D}_0^{\alpha} U(t) = AU(t) + F(t, U(t)), \\ U(0) = U_0, \end{cases} \tag{1}$$

which usually come from semi-discretization of time-fractional partial differential equations [3] or from a reformulation of multi-term fractional differential equations [10]. Moreover, they find applications in control theory where for linear time-invariant systems of fractional order

$$\begin{cases} {}^{\mathrm{C}}\mathcal{D}_0^{\alpha} x(t) = Ax(t) + Bu(t), \\ y(t) = Cx(t) + Du(t), \end{cases}$$

it is possible to study controllability and observability thanks to the corresponding Gramian matrices, defined respectively as

$$\mathcal{C}_{\alpha}(t) := \int_0^t E_{\alpha,\alpha}((t-\tau)^{\alpha}A) BB^T E_{\alpha,\alpha}((t-\tau)^{\alpha}A^T) \mathrm{d}\tau,$$

and

$$\mathcal{O}_{\alpha}(t) := \int_0^t E_{\alpha,1}(\tau^{\alpha}A^T) C^T C E_{\alpha,1}((t-\tau)^{\alpha}A) \mathrm{d}\tau.$$

Since the solution of the semi-linear system of FDEs (1) can be expressed thanks to the following variation-of-constant formula

$$U(t) = E_{\alpha,1}(t^\alpha A)U_0 + \int_0^t (t-s)^{\alpha-1} E_{\alpha,\alpha}((t-s)^\alpha A)F(s,U(s))\mathrm{d}s,$$

it is immediate to verify that suitable numerical methods can be devised by applying some specific quadrature rule whose coefficients are expressed in terms of ML matrix functions.

3. Computation of Matrix Mittag-Leffler Functions

Despite the great importance of the matrix ML function, its computation is not an easy task. Krylov subspace methods are usually used when the matrix argument is of large size [5], but these methods basically projects the matrix onto a space of smaller dimension, thus to just allow to reformulate the problem as the evaluation of ML functions of matrices of smaller size.

To this purpose, methods specifically devised for computation of matrix functions can be exploited. This is the case of the Schur–Parlett algorithm [1] based on the Schur decomposition of the matrix argument combined with the Parlett recurrence to evaluate the matrix function on the triangular factor. Since the methods requires the knowledge of derivatives of the scalar function up to a certain order (which depends on the spectrum of the matrix argument), it is essential that methods for the evaluation of derivatives of the ML function are available.

The evaluation of derivatives of the ML function, as requested by the Schur–Parlett algorithm, is a demanding task. In this talk we present some methods [6] based on the numerical inversion of the Laplace transform. Indeed, derivatives of the ML function can be expressed in terms of a three-parameter ML function (also known as the Prabhakar function) [2, 8]

$$\frac{\mathrm{d}^k}{\mathrm{d}z^k} E_{\alpha,\beta}(z) = k! E^{k+1}_{\alpha,\alpha k+\beta}(z), \quad E^\gamma_{\alpha,\beta}(z) = \frac{1}{\Gamma(\gamma)} \sum_{j=0}^{\infty} \frac{\Gamma(j+\gamma) z^j}{j!\Gamma(\alpha j+\beta)},$$

for which an analytical representation of the LT is available since

$$\mathcal{L}\left(t^{\beta-1} E^\gamma_{\alpha,\beta}(t^\alpha z)\,;\, s\right) = \frac{s^{\alpha\gamma-\beta}}{(s^\alpha - z)^\gamma}, \quad \Re(s) > 0,\ |zs^{-\alpha}| < 1.$$

Basically, a quadrature rule is applied in the formula for the inversion of the Laplace transform

$$\frac{\mathrm{d}^k}{\mathrm{d}z^k} E_{\alpha,\beta}(z) = \frac{k!}{2\pi \mathrm{i}} \int_{\sigma-\mathrm{i}\infty}^{\sigma+\mathrm{i}\infty} \mathrm{e}^s \frac{s^{\alpha\gamma-\beta}}{(s^\alpha - z)^{k+1}} \mathrm{d}s,$$

after suitably deforming the Bromwich contour $[\sigma - \mathrm{i}\infty, \sigma + \mathrm{i}\infty]$ in order to avoid numerical instability due to the presence of the exponential along the imaginary axis. An appropriate selection of the deformed contour and of the quadrature parameters allows to obtain high accuracy at a reasonable computational time, as already obtained for the scalar ML function [4, 7].

In this talk we not only discuss the main ideas underlying methods for the computation of matrix ML functions, and hence of derivatives of the scalar ML function, but we also show the efficiency of the proposed method by presenting some numerical experiments.

Some MATLAB codes for the evaluation of the ML function, with scalar and matrix arguments, are freely available on the Mathworks profile page of the speaker at the following url:
https://www.mathworks.com/matlabcentral/profile/authors/2361481

References

1. P. I. Davies and N. J. Higham, *A Schur-Parlett algorithm for computing matrix functions*, SIAM J. Matrix Anal. Appl. **25** (2) (2003) 464–485.
2. R. Garra and R. Garrappa, *The Prabhakar or three parameter Mittag-Leffler function: Theory and application*, Commun. Nonlinear Sci. Numer. Simul. **56** (2018) 314–329.
3. R. Garrappa, *Exponential integrators for time-fractional partial differential equations*, Eur. Phys. J. Spec. Top. **222** (8) (2013) 1915–1927.
4. R. Garrappa, *Numerical evaluation of two and three parameter Mittag-Leffler functions*, SIAM J. Numer. Anal. **53** (3) (2015) 1350–1369.
5. R. Garrappa, I. Moret and M. Popolizio, *On the time-fractional Schrödinger equation: Theoretical analysis and numerical solution by matrix Mittag-Leffler functions*, Comput. Math. Appl. **74** (5) (2017) 977–992.
6. R. Garrappa and M. Popolizio, *Computing the matrix Mittag-Leffler function with applications to fractional calculus*, J. Sci. Comput. **77** (1) (2018) 129–153.
7. R. Garrappa and M. Popolizio, *Fast Methods for the Computation of the Mittag-Leffler Function*, Handbook of fractional calculus with applications, Vol. 3, De Gruyter, Berlin, 2019, pp. 329–346.
8. A. Giusti, I. Colombaro, R. Garra, R. Garrappa, F. Polito, M. Popolizio and F. Mainardi, *A practical guide to Prabhakar fractional calculus*, Fract. Calc. Appl. Anal. **23** (1) (2020) 9–54.
9. R. Gorenflo, A. A. Kilbas, F. Mainardi and S. V. Rogosin, *Mittag-Leffler Functions, Related Topics and Applications*, Springer Monographs in Mathematics, Springer, Heidelberg, 2014.
10. M. Popolizio, *Numerical solution of multiterm fractional differential equations using the matrix mittag-leffler functions*, Mathematics **6** (1) (2018) 7.

E-mail: roberto.garrappa@uniba.it
E-mail: marina.popolizio@poliba.it

On the ABC Index of Graphs

Gülistan Kaya Gök*
Hakkari University, Department of Mathematics Education, Hakkari 30000, Turkey

Abstract. Atom-bond-connectivity index used to model the stability of alkanes It is an index that makes a significant contribution to chemistry, pharmacology etc. In this paper, some results for the general ABC index which has chemical applications are found using different methods. These new results for ABC index are found in terms of its edges, its vertices and its degrees.
Keywords: ABC index, Graph.
AMS Mathematical Subject Classification [2010]: 05C05, 05C12, 05C75.

1. Introduction

Graph indices are one of the topics in graph theory studies. An important part of these graph indexes are topological indices used especially in chemical graph theory.

A graph represents a molecule and expresses the topological structure of the molecule. The most well-known topological indexes consist of the relationship between vertex, edge, degree. Atom-bond-connectivity index is the best known index whose mathematical properties are reported in [2]. The atom-bond connectivity index ABC is a good example of linear and branched alkanes with tensile energy of cycloalkanes. It is an important index that correlates and calculates the strong bond between atoms with graphs. ABC index is an degree based topological index in [5] such that

$$ABC = ABC(G) = \sum_{v_iv_j \in E(G)} \sqrt{\frac{d_i + d_j - 2}{d_i d_j}}.$$

The ABC index plays a significant role in temperature studies in alkanes [1, 3, 4]. For example, ABC index of ethene (C_2H_4) is $4\sqrt{\frac{2}{3}} + \frac{2}{3}$.

The general ABC index ABC_α is described as

$$ABC_\alpha = ABC_\alpha(G) = \sum_{v_iv_j \in E(G)} (\frac{d_i + d_j - 2}{d_i d_j})^\alpha.$$

2. On the General ABC Index

In this section, G may have several connected components but G does not contain isolated vertices. Here, general ABC index for $\alpha = 1$ is found by adding an edge to G and by deleting an edge from G.

*Speaker

THEOREM 2.1. *Let i and j be nonadjacent vertices of graph G, then*

i) $ABC_1(G+ij) \leq ABC_1(G), \quad d_i \geq 2,$
ii) $ABC_1(G+ij) \geq ABC_1(G), \quad 0 < d_i \leq 2,$

where $G+ij$ is obtained by adding the ij edge to G.

THEOREM 2.2. *Let i and j be nonadjacent vertices of graph G, then*

i) $ABC_1(G) \geq ABC_1(G-ij), \quad d_i \leq 2,$
ii) $ABC_1(G) \leq ABC_1(G-ij), \quad d_i \geq 2,$

where $G-ij$ is obtained by deleting the ij edge from G.

THEOREM 2.3. *Let G be a nontrivial graph with $x, y \in \mathbb{R}$. Then,*

$$ABC_{x+y}(G)ABC_{x-y}(G) - \sigma_{x,y} \leq ABC_x(G) \leq \sqrt{ABC_{x+y}(G)ABC_{x-y}(G)},$$

with

$$\sigma_{x,y} = \begin{cases} 2^{x-2}n^2((\frac{\delta-1}{\delta^2})^x - (\frac{\Delta-1}{\Delta^2})^x), & \text{if } |x| \geq |y|, \\ 2^{x-2}n^2((\frac{\Delta-1}{\Delta^2})^{\frac{x+y}{2}}(\frac{\delta-1}{\delta^2})^{\frac{x-y}{2}} - (\frac{\delta-1}{\delta^2})^{\frac{x+y}{2}}(\frac{\Delta-1}{\Delta^2})^{\frac{x-y}{2}}), & \text{if } |x| < |y|. \end{cases}$$

THEOREM 2.4. *Let G be a nontrivial graph with m edges. G has maximum degree Δ and $2\Delta \leq m-1$. For any integer $4\alpha \geq 1$,*

$$ABC_\alpha(G) \leq (m-1)^{\alpha-1}ABC_{\frac{1}{\alpha}}(G)^\alpha.$$

3. Conclusion

ABC index is a special index in chemical graph theory. In this paper, some bounds for the general ABC index are formed by the help of degrees, edges and vertices. This paper aims to improve to the use of the ABC index.

References

1. M. B. Ahmadi, D. Dimitrov, I. Gutman and S. A. Hosseini, *Disproving a conjecture on trees with minimal atom-bond connectivity index,* MATCH Commun. Math. Comput. Chem. **72** (3) (2014) 685–698.
2. R. B. Bapat, *Graphs and Matrices*, Indian Statistical Institute, New Delhi 110016, India, 2010.
3. M. Bianchi, A. Cornaro, J. L. Palacios and A. Torriero, *New upper bounds for the ABC index*, MATCH Commun. Math. Comput. Chem. **76** (1) (2016) 117–130.
4. K. C. Das, *Atom-bond connectivity index of graphs*, Discr. Appl. Math. **158** (2010) 1181–1188.
5. I. Gutman, *Degree-based topological indices*, Croat. Chem. Acta **86** (2013) 351–361.

E-mail: gulistankayagok@hakkari.edu.tr

A Recent Progress in Localized RBF Techniques

Davoud Mirzaei*
Faculty of Mathematics and Statistics, Department of Applied Mathematics and Computer Science, University of Isfahan, 81746-73441 Isfahan, Iran

Abstract. In this talk, the new direct RBF partition of unity (D-RBF-PU) method is presented for numerical solution of boundary value problems. The D-RBF-PU method is a new localized RBF-based technique which avoids all derivatives of PU weight functions as well as all lower derivatives of local approximants. It is faster and simpler than the standard RBF-PU method, and allows the use of some discontinuous PU weight functions to develop the method in a more efficient and less expensive way. Alternatively, the new method is an RBF-generated finite difference (RBF-FD) method in a PU setting which is much faster and in some situations more accurate than the original RBF-FD. To show the generality of the idea, we will go beyond the RBFs and use other finite dimensional approximation spaces to construct the local approximants on PU cells. At the end, we will extend the method for solving surface PDEs on embedded and smooth submanifolds of the Euclidean spaces.
Keywords: Radial Basis Function (RBF), Partition of Unity (PU) Methods, RBF-FD, RBF-PU, Partial Differential Equations (PDEs).
AMS Mathematical Subject Classification [2010]: 65Nxx, 41Axx.

1. Introduction

Approximation by *radial basis functions* (RBFs) has received a lot of attention due to many attractive advantages such as ease of implementation, flexibility with respect to geometry and dimension and giving spectral accuracy in some situations [10]. However, the global RBF approximations for solving partial differential equations (PDEs) produce full and ill-conditioned matrices which make them restricted for large scale problems. So, localized approaches, such as *RBF finite difference* (RBF-FD) and *RBF partition of unity* (RBF-PU) methods, are currently being developed [1, 2, 3, 7, 8, 9]. However, RBF-FD and RBF-PU have their own disadvantages which are mostly related to their computational costs for either constructing the local approximations or solving the final linear systems.

In [4], we use the idea of *direct discretization* [5, 6] and link the RBF-PU to the RBF-FD and construct a *direct RBF-PU* (D-RBF-PU) method which is more efficient than both RBF-FD and RBF-PU methods. Our idea in not limited to the RBFs and can be easily adapted for the other approximation spaces such as multivariate polynomials. On the other hand, since the method is based on scattered point layouts instead of a predefined background mesh, the underlying domain of the PDE can be easily extended to smooth embedded submanifolds of $\mathbb{R}^d$. In this talk we will address all these issues.

*Speaker

2. Partition of Unity Methods

Let $\{\Omega_\ell\}_{\ell=1}^{N_c}$ be an open and bounded covering of Ω that means all Ω_ℓ are open and bounded and $\Omega \subset \bigcup_{\ell=1}^{N_c} \Omega_\ell$. A family of nonnegative functions $\{w_\ell\}_{\ell=1}^{N_c}$ is called a partition of unity (PU) with respect to the covering $\{\Omega_\ell\}$ if

1) $\text{supp}(w_\ell) \subseteq \Omega_\ell$,
2) $\sum_{\ell=1}^{N_c} w_\ell(x) = 1, \ \forall x \in \Omega$.

We start with an overlapping covering $\{\Omega_\ell\}_{\ell=1}^{N_c}$ of Ω. If we assume V_ℓ is an approximation space on Ω_ℓ and $s_\ell \in V_\ell$ is a local approximant of a function u on Ω_ℓ, then

$$s = \sum_{\ell=1}^{N_c} w_\ell s_\ell, \tag{1}$$

is a global approximation of u on Ω which is formed by joining the local approximants s_ℓ via PU weights w_ℓ. For example, if $X = \{x_1, \ldots, x_N\} \subset \Omega$, $X_\ell = X \cap \Omega_\ell$ and s_ℓ are u interpolants on X_ℓ then we can simply show that s is a u interpolant on X. A possible choice for w_ℓ is the Shepard's weights

$$w_\ell(x) = \frac{\psi_\ell(x)}{\sum_{j=1}^{N_c} \psi_j(x)}, \quad 1 \leqslant \ell \leqslant N_c, \tag{2}$$

where ψ_ℓ are nonnegative, nonvanishing and compactly supported functions on Ω_ℓ.

If w_ℓ and s_ℓ are smooth enough then the PU approximation (1) can be used for solving differential equations. To describe the overall approach, assume that we are looking for the approximate solution of a PDE problem of the form

$$Lu = f, \ \text{in } \Omega, \tag{3}$$

$$Bu = g, \ \text{on } \Gamma, \tag{4}$$

where $\Omega \subset \mathbb{R}^d$ is a domain, $\Gamma = \partial\Omega$ denotes its boundary and L and B are linear differential operators defined and continuous on some normed linear space in which the true solution of (3)-(4) should lie. Here, B is the boundary operator describing the Drichlet and/or Neumann boundary conditions. To obtain a numerical solution, the PDE operators L and B should operate on s (and hence on products $w_\ell s_\ell$) in (1) to get

$$Lu \approx Ls = \sum_{\ell=1}^{N_c} L(w_\ell s_\ell), \quad Bu \approx Bs = \sum_{\ell=1}^{N_c} B(w_\ell s_\ell),$$

where s_ℓ is local approximation of u in patch Ω_ℓ. The differential operators L and B should contain certain partial derivatives D^α for multi-indices $\alpha \in \mathbb{N}_0^d$. Using the Leibniz's rule we have

$$D^\alpha s = \sum_{\ell=1}^{N_c} \sum_{|\beta| \leqslant |\alpha|} \binom{\beta}{\alpha} D^\beta w_\ell D^{\alpha-\beta} s_\ell,$$

provided that both w_ℓ and s_ℓ are smooth enough. For example if $L = \Delta = D^{(2,0,\ldots,0)} + D^{(0,2,\ldots,0)} + \cdots + D^{(0,0,\ldots,2)}$, the well-known Laplacian operator in $\mathbb{R}^d$, then

$$\Delta s = \sum_{\ell=1}^{N_c} \left(s_\ell \Delta w_\ell + 2\nabla w_\ell \cdot \nabla s_\ell + w_\ell \Delta s_\ell \right),$$

where derivatives of w_ℓ are even more complicated if w_ℓ is defined as (2). This will also increase the computational costs of the method. In the next section we propose an alternative approach that avoids the above computations and reduces both computational cost and algorithmic complexity.

3. The New Method

Again consider the PDE problem (3)-(4). In this section we present an alternative approach in which Lu and Bu are *directly* approximated by the PU approximation as

$$Lu \approx \sum_{\ell=1}^{N_c} w_\ell s_\ell^L =: s^L, \quad Bu \approx \sum_{\ell=1}^{N_c} w_\ell s_\ell^B =: s^B,$$

where s_ℓ^L and s_ℓ^B are the local approximations of Lu and Bu in patch Ω_ℓ. We may assume s_ℓ^L and s_ℓ^B are identical with Ls_ℓ and Bs_ℓ on patch Ω_ℓ. While it is clear that the global approximants s^L and s^B are different from their counterparts Ls and Bs because (at least) derivatives of w_ℓ are not incorporated in the new approximation. In the new approach, we have a *direct* approximation for Lu and Bu without any detour via the local functions s_ℓ and the global approximation (1). For comparison, the Laplacian is now approximated by

$$s^\Delta = \sum_{\ell=1}^{N_c} w_\ell \Delta s_\ell.$$

Thus, not only all derivatives of w_ℓ but also many lower derivatives of the local approximants s_ℓ are not actually required. This means that the single term $w_\ell \Delta s_\ell$ will do the whole job. At the first glance, one may expect a lost in accuracy since some terms are ignored in the new approximation. But we will support the new method theoretically and show that the rates of convergence for both methods are the same.

Any finite dimensional approximation space V_ℓ can be used for constructing the local approximants s_ℓ^L and s_ℓ^B. However, we are interested in those which have good approximation properties and are easily computable.

If we combine the new approach with the RBF approximation the method is called D-RBF-PU [4] which has clear advantages over the standard and well-established RBF-PU and RBF-FD methods. Usually, compactly supported and smooth functions (on the whole Ω) are used to generate a smooth PU weight when derivatives are required. Since the derivatives of the PU weight functions are not needed, the new method can be set up on discontinuous PU weight to develop the method in a more efficient and less expensive way, and to recover the standard RBF-FD as a special case. In comparison with the RBF-FD, the new method needs to solve much fewer number of local linear systems for constructing the final differentiation matrix. This reduces the computational costs, considerably. In

our experiments, average speedups of 5x with a smooth PU weight, 10x with a constant-generated PU weight and 9x with a hybrid PU weight are observed in both 2D and 3D examples. Although for a pure Dirichlet problem both methods have approximately the same accuracy, the new method gives more accurate results for Neumann or Neumann-Dirichlet boundary value problems.

We extend the D-RBF-PU for PDE problems on smooth submanifolds embedded in $\mathbb{R}^d$. In this talk we will give some results for spherical PDEs with cost and accuracy comparison tests. Moreover, we will go beyond the RBFs and use some other approximation spaces to show the generality of the idea.

References

1. S. De Marchi, A. Martínez and E. Perracchione, *Fast and stable rational RBF-based partition of unity interpolation*, J. Comput. Appl. Math. **349** (2019) 331–343.
2. N. Flyer, E. Lehto, S. Blaise, G. B. Wright and A. St-Cyr, *A guide to RBF-generated finite differences for nonlinear transport: Shallow water simulations on a sphere*, J. Comput. Phys. **231** (2012) 4078–4095.
3. B. Fornberg and N. Flyer, *Solving PDEs with radial basis functions*, Acta Numer. **24** (2015) 215–258.
4. D. Mirzaei, *The direct radial basis function partition of unity (D-RBF-PU) method for solving PDEs*, SIAM J. Sci. Comput. **43** (2021) A54–A83.
5. D. Mirzaei, R. Schaback and M. Dehghan, *On generalized moving least squares and diffuse derivatives*, IMA J. Numer. Anal. **32** (2012) 983–1000.
6. R. Schaback, *Direct discretization with application to meshless methods for PDEs*, Dolomites Research Notes on Approximation, In: Proc. DWCAA12 **6** (2013) pp. 37–51.
7. V. Shcherbakov and E. Larsson, *Radial basis function partition of unity methods for pricing Vanilla basket options*, Comput. Math. Appl. **71** (1) (2016) 185–200.
8. A. E. Tolstykh, *On using RBF-based differencing formulas for unstructured and mixed structured-unstructured grid calculations*, in: Proc. 16th IMACS World Congress **228** (2000) pp. 4606–4624.
9. H. Wendland, *Fast evaluation of radial basis functions: methods based on partition of unity*, Approximation theory, X (St. Louis, MO, 2001), 473–483, Innov. Appl. Math., Vanderbilt Univ. Press, Nashville, TN, 2002.
10. H. Wendland, *Scattered data approximation*, Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, Cambridge, UK, 2005.

E-mail: d.mirzaei@sci.ui.ac.ir

Contributed Talks

Code and Cryptography

List Decoding of Unit Codes

Nasim Abdi Kourani*
Faculty of Mathematical, K. N. Toosi University of Technology, Tehran, Iran
Hassan Khodaiemehr
Faculty of Mathematical, K. N. Toosi University of Technology, Tehran, Iran
and Mohammad Javad Nikmehr
Faculty of Mathematical, K. N. Toosi University of Technology, Tehran, Iran

Abstract. In this paper, we propose a list decoding algorithm for the family of unit codes introduced by C. Maire and F. Oggier. Unit codes are constructed based on number fields and these codes are generalized version of number field codes for which a list decoding algorithm has already been proposed. We employ the list decoding algorithm of the number field codes presented by J. F. Biasse and G. Quintin.
Keywords: List decoding, Number field codes, Unit codes.
AMS Mathematical Subject Classification [2010]: 11T71, 68P30, 94A24.

1. Introduction

List decoding was introduced separately by Elias in [2] and Wezencraft in [7]. Biasse and Quintin proposed an algorithm for list decoding of number field (NF) codes [1]. Construction of codes using maximal order number fields was presented by Mair and Oggier in [4]. They were an extension of the codes provided by Guruswami and Lenstra [3, 6]. These codes have a high minimum distance and a high rate than the ones defined by Guruswami and Lenstra. Given that there was a list decoding algorithm for NF-codes, the question is designing a list decoding algorithm for the codes built in [1]. Hence, we propose a list decoding algorithm for the unit codes that proposed in [1].

1.1. Preliminaries. The function $E : \Sigma^{\bar{k}} \longrightarrow \Sigma^n$ is called an encoding function with parameters $\bar{k}$ and n that maps a message m with $\bar{k}$ symbols from the set of alphabets Σ to a vector of length n, which is denoted by $E(m)$. The encoded string $E(m)$ is referred to as a codeword. The list decoding problem is defined as follows: Let $r \in \Sigma^n$ be a received word, find a list of all messages $m \in \Sigma^{\bar{k}}$ such that the Hamming distance between r and $E(m)$ is at most e. e is the number of errors that the list decoding algorithm can tolerate. Let R be an integral domain; let $I_1, I_2, \ldots, I_n$ be n pairwise coprime ideals in R such that each $\frac{R}{I_j}$ is finite, and let B be an arbitrary positive real number. Further assume that there is a non-negative function Size : R $\to \mathbb{R}^+$ that associates a non-negative value to each element of the ring R. Then, the ideal-based code $C[R; I_1, I_2, \ldots, I_n; \text{Size}, B]$ is defined to be the set of codewords

*Speaker

$$\{(\tfrac{m}{I_1},\dots,\tfrac{m}{I_n}):\quad m\in R,\ \ \text{Size(m)}\leqslant B\}.$$

We denote a number field of degree d and signature (r_1,r_2) by $\mathbb{K}$ and the ring of integers of $\mathbb{K}$ by $O_{\mathbb{K}}$. In this paper we consider $\mathbb{K}$ as a totally real number field of degree r_1. If in the definition of ideal based code $R=O_{\mathbb{K}}$ and $I_1,\dots,I_n$ be n pairwise coprime ideals in $O_{\mathbb{K}}$, then this code is called NF-code. Let I be a non-zero ideal of $O_{\mathbb{K}}$, the norm I is denoted by $\aleph(I)$ and defined as $\aleph(I)=|\frac{O_{\mathbb{K}}}{I}|$. Let $(\sigma_j)_{j\leq r_1}$ be the embeddings of $\mathbb{K}$ to $\mathbb{R}$. Let $\mathbb{K}_{\mathbb{R}}:=\mathbb{K}\bigotimes_{\mathbb{Q}}\mathbb{R}\simeq\mathbb{R}^{r_1}\times\mathbb{C}^{r_2}$ and extend σ_j's to $\mathbb{K}_{\mathbb{R}}$. The Hermitian form on $\mathbb{K}_{\mathbb{R}}$, denoted by T_2, is defined as $T_2(x,x'):=\Sigma_j\sigma_j(x)\cdot\bar{\sigma_j}(x')$, where $\bar{\sigma_j}$ is the complex conjugate of σ_j. For any $x\in\mathbb{K}_{\mathbb{R}}$ we define $\|x\|:=\sqrt{T_2(x,x)}$ as the corresponding L_2-norm of element x. A valuation of the number field $\mathbb{K}$ is a function $|\cdot|\colon\mathbb{K}\longrightarrow\mathbb{R}$ such that the following properties hold:

i) $|x|\geqslant 0$ and $|x|=0$ if and only if $x=0$.
ii) $|x\cdot y|=|x|\cdot|y|$.
iii) There exist an element $c\geqslant 1$, such that for any $x,y\in\mathbb{K}$, $|x+y|\leqslant c\cdot\max\{|x|,|y|\}$.

Two valuations $|\cdot|_1$ and $|\cdot|_2$ on $\mathbb{K}$ are called equivalent if and only if there is a real number $s>0$ such that $|x|_1=|x|_2^s$ for every $x\in\mathbb{K}$. A equivalence class of the valuations of $\mathbb{K}$ is called a place of $\mathbb{K}$ and denoted by $\mathbb{P}$. The valuation $|\cdot|$ of $\mathbb{K}$ is called non-Archimedean if and only if it is satisfied $|x+y|\leq\max\{|x|,|y|\}$ for all $x,y\in\mathbb{K}$. Otherwise, it is called Archimedean. Archimedean valuation of $\mathbb{K}$ is equivalence with the infinite places of $\mathbb{K}$ and denoted by $\mathbb{P}_\infty$. Let $\{e_1,\dots,e_{\mathfrak{m}}\}$ be a linearly independent set of the vectors in $\mathbb{R}^{\mathfrak{n}}$. Let L be an additive subgroup of $(\mathbb{R}^{\mathfrak{n}},+)$ such that $L=\mathbb{Z}e_1\bigoplus\cdots\bigoplus\mathbb{Z}e_{\mathfrak{m}}$. In this case, L is called a lattice of dimension $\mathfrak{n}$ and rank $\mathfrak{m}$. L is called a full rank lattice if $\mathfrak{m}=\mathfrak{n}$.

1.2. The Structure of the Unite Codes. Here, we present the structure of the unit codes from [4]. Let $O_{\mathbb{K}}^*$ be the group of the units of $O_{\mathbb{K}}$ and let $q=|\mathcal{A}(\mathbb{F}_p)|$, where $\mathcal{A}(\mathbb{F}_p)$ is the set of alphabets over the finite field $\mathbb{F}_p$. Using Dirichlet's Unit Theorem, If $[\mathbb{K}:\mathbb{Q}]=r_1$ then $O_{\mathbb{K}}\cong\mathbb{Z}^{r_1}$ and $O_{\mathbb{K}}^*\simeq\mu_{\mathbb{K}}\times\mathbb{Z}^{r_1-1}$ in which $\mu_{\mathbb{K}}$ contains the roots of unity in $\mathbb{K}$ ($\mathbb{K}$ is a totally real number field). Also we have r_1 embeding $\sigma_1,\dots,\sigma_{r_1}$ in $\mathbb{R}$ and $O_{\mathbb{K}}^*/\mu_{\mathbb{K}}$ can be embedded as a lattice in $\mathbb{R}^{r_1}$. Let $\xi_1,\dots,\xi_{r_1-1}$ be a set of generators for the unite group modulo roots of unity. If ξ is an algebraic number, then the $(r_1-1)\times r_1$ matrix whose entries are $\log|\xi_i^j|$ for $i=1,\dots,r_1-1$, $j=1,\dots,r_1$, has the property that the sum of any row is zero. This implies that the absolute value R' of the determinate of the submatrix formed by deleting one column is independent of the column. The number R' is called the regulator of the algebraic number field. Consider the map Ψ as follows:

$$\Psi:O_{\mathbb{K}}^*\longrightarrow\mathbb{R}^{r_1},$$
$$\xi\longmapsto(\log|\sigma_1(\xi)|,\dots,\log|\sigma_{r_1}(\xi)|),$$

and $\sigma\in\mathbb{P}_\infty$. We consider a $\mathbb{Z}$-basis $\{\xi_1,\dots,\xi_{r_1-1}\}$ of $O_{\mathbb{K}}^*$ mod $\mu_{\mathbb{K}}$ and the set $\mathbb{H}_0=\Psi(O_{\mathbb{K}}^*)$. The lattice $\mathbb{H}$ in $\mathbb{R}^{r_1}$ is then as follows:

$$\mathbb{H}:=\mathbb{Z}x_0\bigoplus(\bigoplus_i\mathbb{Z}\Psi(\xi_i))=\mathbb{Z}x_0\bigoplus\mathbb{H}_0\subset\mathbb{R}^{r_1},$$

such that $x_0=(1,\dots,1)$. Restrict the map Ψ to $\Lambda=<\xi_1,\dots,\xi_{r_1-1}>$ and denoted it by Ψ_Λ. Then,

$$\Psi_\Lambda:\Lambda\longrightarrow\mathbb{H}_0,$$

$$\xi \longmapsto \Psi(\xi),$$

is an isomorphism of the groups. Let $\mathrm{Prj}_{\mathbb{H}_0} : \mathbb{H} \longrightarrow \mathbb{H}_0$ be a projection map and t be a positive real number. Define the set $\mathcal{K}(t)$ is as follows:

$$\mathcal{K}(t) = \{x \in \Pi_{\sigma\in\mathbb{P}_\infty}\mathbb{K}_\sigma \mid |\sigma(x)| \leqslant t,\ \forall\sigma\},$$

where $\mathbb{K}_\sigma$ is the completion of the number field $\mathbb{K}$ considering the valuation σ. It can be checked that $\mathcal{K}(t)$ is compact in $\mathbb{R}^{r_1}$ with volume $\mu(\mathcal{K}(t)) = (2t)^{r_1}$ where μ is a Labesgue measure. Let $\mathcal{T}$ be the fundamental domain of $\mathbb{H}$ and $z \in \mathcal{T}$ such that the following relation holds:

$$|(\mathcal{K}(t) \cap \mathbb{H})| \geqslant \frac{\mu(\mathcal{K}(t))}{\mu(\frac{\mathbb{R}^{r_1}}{\mathbb{H}})} = \frac{2^{r_1}t^{r_1}}{r_1\mathrm{Reg}_\mathbb{K}},$$

where $\mathrm{Reg}_\mathbb{K}$ is Regulator of the number field $\mathbb{K}$. Set $\mathcal{K}_z(t) = (z + \mathcal{K}(t))$ and $\mathbb{H}_{0,z}(t) = \mathrm{Prj}_{\mathbb{H}_0}(\mathcal{K}_z(t) \cap \mathbb{H})$.

LEMMA 1.1. [4, Lemma 4.3] *We have* $|\Psi_\Lambda^{-1}(\mathbb{H}_{0,z}(t))| \geq \frac{2^{r_1}t^{r_1}}{(2t+1)r_1\mathrm{Reg}_\mathbb{K}}$.

Now, consider the set of the prime ideals $\{\mathfrak{p}_1, \ldots, \mathfrak{p}_n\}$ of $O_\mathbb{K}$. Let Θ be reduction map as follows:

$$\Theta : O_\mathbb{K}^* \longrightarrow \Pi_{l=1}^n \frac{O_\mathbb{K}}{\mathfrak{p}_l},$$
$$x \longmapsto (x \mod \mathfrak{p}_1, \ldots, x \mod \mathfrak{p}_n),$$

DEFINITION 1.2. [4, Definition 4.4] The unit code of the number field $\mathbb{K}$ is defined as follows:

$$C_{z,t}(O_\mathbb{K}^*) := \Theta(\Psi_\Lambda^{-1}(\mathbb{H}_{0,z}(t))).$$

Now, we give a brief description of the list decoding algorithm of the NF-codes presented in [1]. Let $B = \Pi_{i\leqslant\mathfrak{k}}\aleph(\mathfrak{p}_\mathfrak{i})^{\frac{1}{d}}$ where $0 < \mathfrak{k} < n$. In this case, Size $= \|\cdot\|$ and the code $\mathcal{M}_C = \{m \in O_\mathbb{K} \mid \|m\| \leqslant B\}$ was Considered by Biasse and Quintin [1].

Let $m \in O_\mathbb{K}$ such that m is encoded as

$$O_\mathbb{K} \longrightarrow \frac{O_\mathbb{K}}{\mathfrak{p}_1} \times \cdots \times \frac{O_\mathbb{K}}{\mathfrak{p}_n},$$
$$m \longmapsto (m \mod \mathfrak{p}_1, \ldots, m \mod \mathfrak{p}_n),$$

where $\mathfrak{p}_l$'s are pairwise comaximal non-zero ideals of $O_\mathbb{K}$ for $l = 1, \ldots, n$. Let $(r_1, \ldots, r_n) \in \Pi_{l=1}^n \frac{O_\mathbb{K}}{\mathfrak{p}_l}$ be a received word. Let $\gamma_1, \ldots, \gamma_n$ be the positive integer numbers and Z be a parameter. All the codewords m retrieved in [1] satisfy $\Sigma_{l=1}^n \mathfrak{a}_l\gamma_l > Z$, where $\mathfrak{a}_l = 1$ if $m \mod \mathfrak{p}_l = r_l$ and 0 otherwise (it is called that m and r_l have weighted agreement Z). Let $f \in O_\mathbb{K}[y]$ be a polynomial of degree at most $\mathfrak{d}$. The codewords m with favorite weighted agreement are found by computing roots of f satisfying the following relation:

$$\|m\| \leqslant B \Longrightarrow \|f(m)\| < F, \tag{1}$$

where F is a proper bound. Let $J_l = \{\mathfrak{a}(y)(y - r_l) + \mathfrak{p} \cdot \mathfrak{b}(y) \mid \mathfrak{a}, \mathfrak{b} \in O_\mathbb{K}[y],\ \mathfrak{p} \in \mathfrak{p}_l\}$ be ideals of $O_\mathbb{K}[y]$ such that $1 \leqslant l \leqslant n$. For each ideal J_l we assign a positive integer γ_l for $l = 1, \ldots, n$. Let $f \in \Pi_{l=1}^n J_l^{\gamma_l} \subset O_\mathbb{K}[y]$, then $f(m) \in \Pi_{l=1}^n \mathfrak{p}_l^{\mathfrak{a}_l\gamma_l}$, where $\mathfrak{a}_l = 1$ if $f(m) \mod \mathfrak{p}_l = r_l$ and 0 otherwise.

Let $f(m) \neq 0$, then

$$\aleph(f(m)) \geq \Pi_{l=1}^n \aleph(\mathfrak{p}_l)^{\mathfrak{a}_l\gamma_l}.$$

The inequality between algorithmetic and geometric mean are concluded as follows:

$$\|f(m)\| \geq \sqrt{d}\aleph(f(m))^{\frac{1}{d}}.$$

If the following relation hold:

$$\Sigma_{l=1}^{n}\mathfrak{a}_l\gamma_l \log \aleph(\mathfrak{p}_l) > \tfrac{-d}{2}\log(d) + d\log F,$$

which results in

$$\sqrt{d}(\Pi_{l=1}^{n}\aleph(\mathfrak{p}_l)^{\gamma_l\mathfrak{a}_l})^{\frac{1}{d}} > F,$$

then $f(m) = 0$, otherwise it contradicts with (1). Now, let's illustrate the list decoding Algorithm 1 presented in [1] for NF-codes.

Algorithm 1 List decoding algorithm for NF-codes

Require: $O_{\mathbb{K}}$, $\gamma_1, \ldots, \gamma_n$, B, Z, $(r_1, \ldots, r_n) \in \Pi_{l=1}^{n}\frac{O_{\mathbb{K}}}{\mathfrak{p}_l}$.
Ensure: All m such that $\Sigma_{l=1}^{n}\mathfrak{a}_l\gamma_l > Z$.

(1) Calculate $\mathfrak{d}$ and F.
(2) Find $f \in \Pi_{l=1}^{n}J_l^{\gamma_l} \subset O_{\mathbb{K}}[y]$ of degree at most $\mathfrak{d}$ such that $\|m\| \leqslant B \Rightarrow \|f(m)\| < F$.
(3) Find all roots of f and annunce those roots ξ such that $\|\xi\| \leqslant B$ and $\Sigma_{l=1}^{n}\mathfrak{a}_l z_l > Z$.

2. Main Results

2.1. List Decoding of Unit Codes when $\mathbb{K}$ is Totally Real. In this section, let $\mathbb{K}$ be a totally real number field of degree r_1. Our goal is the list decoding of the unit codes. We assume that

$$\mathfrak{r} = (r_1, \ldots, r_n) = (\beta \mod \mathfrak{p}_1, \ldots, \beta \mod \mathfrak{p}_n) \in \Pi_{l=1}^{n}\frac{O_{\mathbb{K}}}{\mathfrak{p}_l},$$

is a received word such that the set $\{\mathfrak{p}_1, \ldots, \mathfrak{p}_n\}$ is a set of prime ideals of $O_{\mathbb{K}}$ and $\beta \in O_{\mathbb{K}}$. Let $\{x_1, \ldots, x_M\}$ be the output list of Algorithm 1. Then, we have $\Sigma_{l=1}^{n}\mathfrak{a}_l\gamma_l > Z$, where $\mathfrak{a}_l = 1$, if $r_l = x_i \mod \mathfrak{p}_l$, otherwise $\mathfrak{a}_l = 0$ and $\|x_i\| \leqslant B$ for $i = 1, \ldots, M$. In order to employ Algorithm 1, we note that $\{\gamma_1, \ldots, \gamma_n\} \subset \mathbb{Z}$, $x_i \in O_{\mathbb{K}}$, and $Z, B \in \mathbb{R}^+$ are parameters needed to be specified appropriately. We want to choose x_i's such that $x_i \in O_{\mathbb{K}}^*$. So, we employ the following Proposition.

Proposition 2.1. [5, Proposition 4.9] *Let $x \in O_{\mathbb{K}}$. Then, $x \in O_{\mathbb{K}}^*$ if and only if $|N(x)| = 1$.*

Therefore, we have the following conclusion.

Conclusion 1. *$x \in O_{\mathbb{K}}^*$ if and only if $\Sigma_{j=1}^{r_1}\log|\sigma_j(x)| = 0$.*

Lemma 2.2. *Let t be a given parameter such that $t \in \mathbb{R}^+$ and $\mathcal{M}_C = \{x \in O_{\mathbb{K}} \mid \|x\| \leqslant B\}$. Let $\{x_1, \ldots, x_{M'}\}$ be the elements of $\mathcal{M}_C$ satisfying in Conclusion 1, and $\mathcal{T}$ be the fundamental domain of lattice $\mathbb{H} = \mathbb{Z}x_0 \bigoplus \mathbb{H}_0$ with $x_0 = (1, \ldots, 1)$ and $z = (z'_1, \ldots, z'_{r_1})$ be a point of $\mathcal{T}$ and z_1 be an arbitrary integer such that $-t - \log|\sigma_j(x_i)| + z'_j \leqslant z_1 \leqslant t - \log|\sigma_j(x_i)| + z'_j$. To run Algorithm 1 for the unit codes, it is suffice to assume*

$$B = e^{t-z_1}\sqrt{e^{2z'_1} + \cdots + e^{2z'_{r_1}}}.$$

LEMMA 2.3. *Let* $\{x_1, \ldots, x_{M'}\}$ *be the outputs of Algorithm* 1 *satisfying Conclusion* 1. *Let* $y \in \{x_1, \ldots, x_{M'}\}$ *with* $\Psi(y) \in \mathrm{Prj}_{\mathbb{H}_0}(\mathbb{H} \cap \mathcal{K}_z(t))$, *where* $z = (z'_1, \ldots, z'_{r_1})$ *is a point of fundamental domain* $\mathbb{H}$. *Then, the outputs of the list decoding algorithm for the unit codes are exactly the outputs of Algorithm* 1 *satisfy* $\Sigma_{j=1}^{r_1} \log|\sigma_j(x_{\mathfrak{i}})| = 0$ *for* $\mathfrak{i} = 1, \ldots, M'$.

The proofs of Lemma 2.2 and Lemma 2.3 are omitted due to lack of space. Now, we present our list decoding algorithm for the unit codes when $\mathbb{K}$ is a totally real number field.

Algorithm 2 List decoding algorithm for the unit codes.

Require: $t \in \mathbb{R}^+$, Z, $O_{\mathbb{K}}$, $O_{\mathbb{K}}^*$, $\mathbb{H}_0$, $\mathbb{H}$, $\mathfrak{r} = (r_1, \ldots, r_n) = (\beta \mod \mathfrak{p}_1, \ldots, \beta \mod \mathfrak{p}_n) \in \Pi_{l=1}^n \frac{O_{\mathbb{K}}}{\mathfrak{p}_l}$, $(z'_1, \ldots, z'_{r_1}) \in \mathcal{T}$, $\{\Psi(\xi_1), \ldots, \Psi(\xi_{r_1-1})\}$ and $\{\gamma_1, \ldots, \gamma_n\} \subset \mathbb{Z}$ and $\Lambda =< \xi_1, \ldots, \xi_{r_1-1} >$.

Ensure: $\Sigma_{l=1}^n \mathfrak{a}_l\gamma_l > Z$, where $\mathfrak{a}_l = 1$ if $r_l = x_i \mod \mathfrak{p}_l$ otherwise $\mathfrak{a}_l = 0$, for $i = 1, \ldots, M$. $\Sigma_{j=1}^{r_1} \log|\sigma_j(x_{\mathfrak{i}})| = 0$ for $\mathfrak{i} = 1, \ldots, M'$ and $j = 1, \ldots, r_1$, also the following relation holds:

$$-t - \log|\sigma_j(x_{\mathfrak{i}})| + z'_j \leqslant z_1 \leqslant t - \log|\sigma_j(x_{\mathfrak{i}})| + z'_j, \quad with \quad z_1 \in \mathbb{Z}.$$

(1) Call Algorithm 1 and set $B = e^{t-z_1}\sqrt{e^{2z'_1} + \ldots + e^{2z'_{r_1}}}$ and $z_1 = 0$. Let $\{x_1, \ldots, x_M\}$ be the outputs of Algorithm 1 i.e. $\Sigma_{l=1}^n \mathfrak{a}_l\gamma_l > Z$ where $\mathfrak{a}_l = 1$ if $r_l = x_i \mod \mathfrak{p}_l$ otherwise $\mathfrak{a}_l = 0$ and $\|x_i\| \leqslant B$ for $i = 1, \ldots, M$.

(2) Find all $x_{\mathfrak{i}}$'s such that $\Sigma_{j=1}^{r_1} \log|\sigma_j(x_{\mathfrak{i}})| = 0$ for $\mathfrak{i} = 1, \ldots, M'$ and denote all such $x_{\mathfrak{i}}$'s by $\{x_1, \ldots, x_{M'}\}$.

(3) Find all z_1's such that for $j = 1, \ldots, r_1$

(2) $$-t - \log|\sigma_j(x_{\mathfrak{i}})| + z'_j \leqslant z_1 \leqslant t - \log|\sigma_j(x_{\mathfrak{i}})| + z'_j \quad with \quad z_1 \in \mathbb{Z},$$

let $l_{\mathfrak{i}}$ be the number of z_1's that satisfies in (2) and $\mathfrak{i} = 1, \ldots, M'$. We denote all such z_1's by $z_{\mathfrak{i},k}$ satisfying (2) for $k = 1, \ldots, l_{\mathfrak{i}}$ and $\mathfrak{i} = 1, \ldots, M'$. If there are no such z_1 then we remove $x_{\mathfrak{i}}$ from our list.

(4) Set $Z_{\mathfrak{i},k} = z_{\mathfrak{i},k} \cdot (1, \ldots, 1)$ and $y_{\mathfrak{i},k} = \Psi(x_{\mathfrak{i}}) + Z_{\mathfrak{i},k}$ for $\mathfrak{i} = 1, \ldots, M'$, $j = 1, \ldots, r_1$ and $k = 1, \ldots, l_{\mathfrak{i}}$ such that $y_{\mathfrak{i},k} = (y_{\mathfrak{i},k}^{(1)}, \ldots, y_{\mathfrak{i},k}^{(r_1)})$, we consider the inner product of $x_0 = (1, \ldots, 1)$ and $y_{\mathfrak{i},k}$, then we must have $z_{\mathfrak{i},k} = [\frac{\Sigma_{\alpha=1}^{r_1} y_{\mathfrak{i},k}^{(\alpha)}}{r_1}]$.

(5) Project $y_{\mathfrak{i},k}$ over $\mathbb{H}_0$, then we have the projection as follow $y_{\mathfrak{i},k} - Z_{\mathfrak{i},k} = \Psi(x_{\mathfrak{i}})$.

(6) Effect Ψ_Λ^{-1} over $y_{\mathfrak{i},k} - Z_{\mathfrak{i},k} = \Psi(x_{\mathfrak{i}})$, then we have $\Psi_\Lambda^{-1}(y_{\mathfrak{i},k} - Z_{\mathfrak{i},k}) = \Psi_\Lambda^{-1}(\Psi(x_{\mathfrak{i}})) = x_{\mathfrak{i}}$.

(7) Repeat this Algorithm until finished all $x_{\mathfrak{i}}$'s.

(8) Report all $x_{\mathfrak{i}}$'s.

References

1. J. F. Biasse and G. Quintin, *An algorithm for list decoding number field codes*, IEEE Int. Symp. Inf. Theory Proc. (2012) 91–95.

2. P. Elias, *List decoding for noisy channels*, Res. Lab. Electronics, Massachusetts Institute of Technology. (1957) 94–104.
3. H. W. Lenstra Jr., *Codes from algebraic number fields*, Mathematics and computer science, II (Amsterdam, 1986), 95–104, CWI Monogr., 4, North-Holland, Amsterdam, 1986.
4. C. Maire and F. Oggier, *Maximal order codes over number fields*, J. Pure Appl. Algebra **222** (7) (2018) 1827–1858.
5. I. Stewart and D. Tall, *Algebraic Number Theory and Fermat's Last Theorem*, 4th ed., Chapman and Hall/CRC., New York, 2015.
6. G. Venkatesan, *Constructions of codes from number fields*, IEEE Trans. Inf. Theory **49** (3) (2003) 594–603.
7. J. M. Wozencraft, *List Decoding. Quarterly Progress Report*, Res. Lab. Electronics MIT. **48** (1958) 90–95.

E-mail: n.abdi@email.kntu.ac.ir
E-mail: ha.khodaiemehr@kntu.ac.ir
E-mail: nikmehr@kntu.ac.ir

A New Approach for Decoding of Cyclic Codes Over $F_2 + uF_2$

Mohammad Reza Alimoradi*
Department of Mathematics, Faculty of Mathematical Sciences, University of Malayer, Iran

Abstract. Udaya and Bonnecaze (1999) presented a decoding algorithm for cyclic codes of odd length over the ring $F_2 + uF_2$. In this study, a simpler approach for decoding cyclic codes with odd length over this ring is proposed.
Keywords: Decoding, Cyclic codes, Torsion codes, Lee distance.
AMS Mathematical Subject Classification [2010]: 94B15.

1. Introduction

Cyclic codes over rings have created a great deal of interest, because of their new role in algebraic coding theory. Hammons et al. showed binary nonlinear codes can be viewed as linear codes through a Gray mapping over Z_4 by exploiting the isometry between (Z_4^n, Lee distance) and (F_2^{2n}, Hamming distance). Since some binary codes with a good error-correcting capability are Gray images of cyclic codes over finite rings (See [3, 5]), the study of cyclic codes over finite rings is significant. So far, a few papers have been published on the decoding of codes over finite rings (See [1, 2, 4, 5]). A decoding algorithm for cyclic codes over the ring F_2+uF_2 with respect the Lee distance proposed by Udaya and Bonnecaze, also a decoding algorithm for cyclic codes over the ring $\frac{F_p[u]}{\langle u^k\rangle}$ with respect the Hamming distance proposed by Alimoradi.

In this study, we proposed a shorter decoding procedure for cyclic codes over the ring $R = F_2 + uF_2$ with respect the Lee distance. For any ring S the quotiont ring $D(S)$ is defined as $D(S) = \frac{S[x]}{\langle x^2\rangle} \simeq S \bigoplus uS$, where $u^2 = 0$. Also $\pi_1, \pi_2 : D(S) \to S$ are defined as $\pi_1(a + bu) = a, \pi_2(a + bu) = b$. We review some basic facts about Galois rings. Let $\mu : R \to F_2$ is the natural homomorphic mapping from R to its residue field F_2, where $\alpha \to \alpha(\text{mod } u)$. The map μ is extended to $\mu : R[x] \to F_2[x]$ in the usual way. The image of a polynomial $f(x) \in R[x]$ under this projection is denoted by $\overline{f}(x)$. A monic polynomial $f(x) \in R[x]$ is said to be a basic irreducible polynomial over R if $\overline{f}(x)$ is irreducible in $F_2[x]$. Let $f(x)$ be a basic monic irreducible polynomial of degree r in $R[x]$. The Galois ring of R is denoted as $GR(R, r)$ and is defined as $\frac{R[x]}{\langle f(x)\rangle}$. So $GR(R, r) \simeq D(F_{2^r})$ by Cohen structure theorem for complete rings. So by identification of these two rings, $GR(R, r)$ is a principal ideal local ring. Also the group of units of $GR(R, r) \simeq D(F_{2^r})$ is $F^\star 2^r(1 + uF_{2^r})$. Hence $G_C = F_{2^r}^\star$ is a cyclic group and $G_A = 1 + uF_{2^r} \simeq F_{2^r}$.

*Speaker

2. Decoding of Cyclic Codes Over the Ring $F_2 + uF_2$

Udaya and Bonnecaze, introduced a decoding procedure for cyclic codes over the ring $F_2 + uF_2$ by the use of a Gray map and $\langle u, u+v\rangle$ construction codes. They showed that a cyclic code C of length n over this ring has the structure of $C = \langle fh, ufg\rangle$, where $fgh = x^n - 1$ and Gray image C is equivalent to a $\langle u, u+v\rangle$ constructed code with binary codes $C_1 = Res(C) = \langle fh\rangle$ and $C_2 = Tor_1(C) = \langle f\rangle$. It is clear that $\frac{R[x]}{\langle x^n-1\rangle} \simeq D(S)$, where $S = \frac{F_2[x]}{\langle x^n-1\rangle}$. So $C_1 = Res(C) = \pi_1(C)$ and $C_2 = Tor_1(C) = \pi_2(C)$.

We assume that C is a cyclic code over the ring $F_2 + uF_2$ with Lee weight k and the number of actual errors which have occurred is less than or equal to t. Suppose that the errors occur in the unknown coordinates $l_1, l_2, \ldots, l_t$. Then by the use of the vector representation of $F_2 + uF_2$ over F_2, the error polynomial $e(x)$ can be written as follows:

$$e(x) = (e_{l_{1,0}} + ue_{l_{1,1}})x^{l_1} + (e_{l_{2,0}} + ue_{l_{2,1}})x^{l_2} + \cdots + (e_{l_{t,0}} + ue_{l_{t,1}})x^{l_t}, \tag{1}$$

where $e_{i,j} \in F_2$, for $i = 1, 2, \ldots, t$ and $j = 0, 1$. For any error-polynomial $e(x) = \sum_{i=0}^{n-1} e_i x^i$ over $F_2 + uF_2$, the error-polynomial $e_{1,\bar{u}}(x)$ is a binary polynomial such that for all $i = 0, 1, 2, \ldots, n-1$, the coefficient of x^i in $e_{1,\bar{u}}(x)$ is equal to 1 if e_i is 1 or $\bar{u}$ and is equal to 0 otherwise. Similarly, for all $i = 0, 1, 2, \ldots, n-1$, the coefficient of x^i in $e_{u,\bar{u}}(x)$ is equal to 1 if e_i is u or $\bar{u}$ and is equal to 0 otherwise. Let χ_A be the characteristic function of $A \subseteq R$. If $f(x) = \sum_{i=0}^{n-1} a_i x^i \in R[x]$, then $W_L(f(x)) = \sum_{i=0}^{n-1} 2\chi_u(a_i) + \sum_{i=0}^{n-1} \chi_{1,\bar{u}}(a_i)$. Let $f_A(x) = \sum_{i=0}^{n-1} \chi_A(a_i) \in F_2[x]$ for any $A \subseteq R$. If $f(x) = \sum_{i=0}^{n-1} a_i x^i \in F_2[x]$, then $W_H(f(x)) = \sum_{i=0}^{n-1} \chi_1(a_i)$.
So $\chi_{1,\bar{u}}(e_i) + u\chi_{u,\bar{u}}(e_i) = e_i$, where $e_i \in R$, implies that

$$e(x) = \sum_{i=0}^{n-1} e_i x^i = e_{1,\bar{u}}(x) + ue_{u,\bar{u}}(x). \tag{2}$$

Also, $e_{u,\bar{u}}(x) = e_{1,\bar{u}}(x) + e_{1,u}(x)$. With this notation $W_L(e(x)) = \sum_{i=0}^{n-1} 2\chi_u(e_i) + \sum_{i=0}^{n-1} \chi_{1,\bar{u}}(e_i)$. Also $W_H(e_{1,\bar{u}}(x)) = \sum_{i=0}^{n-1} \chi_{1,\bar{u}}(e_i)$, $W_H(e_{1,u}(x)) = \sum_{i=0}^{n-1} \chi_{1,u}(e_i)$ and $W_H(e_{u,\bar{u}}(x)) = \sum_{i=0}^{n-1} \chi_{u,\bar{u}}(e_i)$.

THEOREM 2.1. *Let $C = \langle fh, ufg\rangle$ be a cyclic code over $R = F_2 + uF_2$, $Z_1 = \{\alpha^i, \alpha^{i+1}, \ldots, \alpha^{i+t_1-1}\}$ be t_1 consecutive roots of the polynomial f and*

$$Z_2 = \{\alpha^j, \alpha^{j+1}, \ldots, \alpha^{j+t_1+t_2-1}\},$$

be $t_1 + t_2$ consecutive roots of the polynomial fh, then it is possible to completely determine an error $e(x)$ if $W_L(e(x)) \leq \lfloor \frac{t_1+t_2}{2} \rfloor$.

PROOF. Note that unlike codes over the ring Z_4, free cyclic codes over R are not interesting because of their poor minimum Lee distance. The codes are only interesting when $t_1 + t_2$ is approximately equal to $2t_1$. Let $e(x)$ is an error polynomial over R and $e_{1,u}, e_{1,\bar{u}}, e_{u,\bar{u}}$ be its associated binary error polynomials. If $W_L(e(x)) \leq t$, by the above notations $W_H(e_{1,\bar{u}}(x)) \leq t$. Also either $W_H(e_{u,\bar{u}}(x)) \leq \lfloor t/2 \rfloor$ or $W_H(e_{1,u}(x)) \leq \lfloor t/2 \rfloor$, where $\lfloor t \rfloor$ represent the largest integer less than or equal to t. Now let $t = \frac{t_1+t_2}{2}$, as the binary code $Res(C)$ is a BCH code with $t_1 + t_2$ consecutive roots of the generator polynomial of codes, then the

binary error polynomial $e_{1,\bar{u}}(x)$ will be decoded in $Res(C)$, (See [5] the Peterson-Gorenstein-Zierler algorithm.) Also, the binary code $Tor_1(C)$ is a BCH code with t_1 consecutive roots of the generator polynomial of codes, then the error polynomial $e_{u,\bar{u}}(x)$ or $e_{1,u}(x)$ will be decoded in the binary code $Tor_1(C)$, (with this assumption that t_1+t_2 is approximately equal to $2t_1$). If $W_H(e_{u,\bar{u}}(x)) \leq \lfloor t/2 \rfloor$, then $e_{u,\bar{u}}(x)$ will be decoded in the binary code $Tor_1(C)$. So it is possible to completely decode an error $e(x)$. If $W_H(e_{1,u}(x) \leq \lfloor t/2 \rfloor$, then $e_{1,u}(x))$ will be decoded in the binary code $Tor_1(C)$. So, by the use of equation $e_{u,\bar{u}}(x) = e_{1,\bar{u}}(x) + e_{1,u}(x)$, it is possible to completely decode an error $e(x)$. □

We explain [6, Example 2] by the use of above procedure.

EXAMPLE 2.2. Let $C = \langle fh, ufg \rangle$ be a cyclic code of length 31 over $F_2 + uF_2$ and α be a primitive element of order 31 in $GR(F_2 + uF_2, 5) = \frac{(F_2+uF_2)[x]}{\langle x^5+x^2+1 \rangle}$. Let $f = f_1 f_3, h = f_5 f_7$ and $g = f_0 f_{11} f_{15}$, where

$$\begin{aligned} f_0 =& x+1, f_1 = x^5+x^2+1, f_3 = x^5+x^4+x^3+x^2+1, f_5 \\ =& x^5+x^4+x^2+x+1, f_7 = x^5+x^3+x^2+x+1, f_{11} \\ =& x^5+x^4+x^3+x+1, \end{aligned}$$

and $f_{15} = x^5 + x^3 + 1$. The sets of consecutive roots of polynomials f and fh are given as follows:

$$Z_1 = \{\alpha, \alpha^2, \alpha^3, \alpha^4\}, \ Z_2 = \{\alpha, \alpha^2, \ldots, \alpha^{10}\}.$$

Since $t_1 = 4$ and $t_1 + t_2 = 10$, from Theorem 7 in Udaya and Bonnecaze, it follows that the minimum Lee distance of this code is equal to 10. Then the number of actual errors of type $1, \bar{u}$ in e is less than or equal to 5 and either the number of actual errors of type $u, \bar{u}$ in e is less than or equal to 2 or the number of actual errors of type $1, u$ in e is less than or equal to 2. By the use of MATLAB software (See [7, Section 3.5]), the error polynomial $e(x)$ will be determined in each case. In Table 1, S_0, S_1, S_2 denote the syndrome of the error polynomials $e_{1,\bar{u}}(x), e_{u,\bar{u}}(x), e_{1,u}(x)$, respectively, and $v(x)$ denote the received polynomial. We use the symbol $\star$ when the binary error-polynomial has not solution.

TABLE 1

Ex	1	2
$v(x)$	$\bar{u}x^3 + \bar{u}x^5 + \bar{u}x^{13} + \bar{u}x^{25} + \bar{u}x^{29}$	$1 + x^2 + \bar{u}(x + x^3 + x^5)$
S_0	$\{1, 1, \alpha^8, 1, \alpha^7, \alpha^{16}, \alpha^{24}, 1, \alpha^{25}, \alpha^{14}\}$	$\{\alpha^{22}, \alpha^{13}, \alpha^{30}, \alpha^{26}, \alpha^{26}, \alpha^{29}, \alpha^{30}, \alpha^{21}, \alpha^{22}, \alpha^{21}\}$
S_1	$\{1, 1, \alpha^8, 1\}$	$\{\alpha^{23}, \alpha^{15}, \alpha^8, \alpha^{30}\}$
S_2	$\{0, 0, 0, 0\}$	$\{\alpha^5, \alpha^{10}, \alpha^{27}, \alpha^{20}\}$
$e_{1,\bar{u}}(x)$	$x^3 + x^5 + x^{13} + x^{25} + x^{29}$	$1 + x + x^2 + x^3 + x^5$
$e_{u,\bar{u}}(x)$	$\star$	$\star$
$e_{1,u}(x)$	0	$1 + x^2$
$e(x)$	$\bar{u}x^3 + \bar{u}x^5 + \bar{u}x^{13} + \bar{u}x^{25} + \bar{u}x^{29}$	$1 + x^2 + \bar{u}(x + x^3 + x^5)$

References

1. M. R. Alimoradi, *Decoding of cyclic codes over the ring* $\frac{F_p[u]}{\langle U^k \rangle}$, U. P. B. Sci. Bull. Series A **79** (3) (2017) 19–32.
2. E. Byrne, M. Greferath, J. Pernas and J. Zumbrgel, *Algebraic decoding of negacyclic codes over* $\mathbb{Z}_4$, Des. Codes. Crypt. **66** (1-3) (2007) 3–16.
3. A. R. Hammons, P. V. Kumar, A. R. Calderbank, N. J. A. Sloane and P. Solé, *The* $\mathbb{Z}_4$*-linearity of kerdock, preparata, goethals and related codes*, IEEE Trans. Inf. Theory **40** (1994) 301–319.
4. W. C. Huffman and V. Pless, *Fundamentals of Error-Correcting Codes*, Cambridge University Press, New York, 2003.
5. J. C. Interlando, R. Palazzo and M. Elia, *On the decoding of Reed-Solomon and BCH codes over integer residue rings*, IEEE Trans. Inf. Theory **43** (3) (1997) 11013–1021.
6. P. Udaya and A. Bonnecaze, *Decoding of cyclic codes over* $F_2 + uF_2$, IEEE Trans. Inf. Theory **45** (1999) 2148–2157.
7. R. H. Morelos-Zaragoza, *The art of error correcting coding*, 2nd ed., John Wiley & Sons, Hoboken, 2006.

E-mail: malimoradisharif@yahoo.com

The Weight Hierarchy of $(u, u+v)-$Construction of Codes

Farzaneh Farhang Baftani*
Department of Mathematics, Ardabil Branch, Islamic Azad University, Ardabil, Iran

Abstract. Let C_i be an $[n, k_i, d_i]$ linear code over F_q for $i = 1, 2$. Let $C = \{(u, u+v); u \in C_1, v \in C_2\}$. Motivated by finding the relationship between $d_r(C)$ and $d_r(C_1)$, $d_r(C_2)$, we investigated $d_r(C)$. Hence we found an upper bound for $d_r(C)$ according to $d_r(C_1)$ and $d_r(C_2)$. In addition, we proved that $d_2(C)$ equals to an upper bound in the binary case. Note that for a linear code D over a finite field, the r-th generalized Hamming weight (r-th GHW) is defined as the minimum of the support size of r-dimensional sub-codes of D and we denote it by $d_r(D)$.
Keywords: Generalized Hamming Weight, Linear code, $(u, u+v)-$construction, Weight Hierarchy.
AMS Mathematical Subject Classification [2010]: 05C69.

1. Introduction

Generalized Hamming weight introduced by Wei in his seminal paper [7]. Then this concept of code was investigated by several authors later, see [1, 2, 4] and [5]. Generalized Hamming Weight (GHW) investigated over rings too, see [3]. We say that C is an $[n, k, d]_q$ linear code if C is a subspace of F_q^n of dimension k, and minimum distance d in which F_q is a field of order q and $n \in N$. The smallest Hamming weight of the nonzero codewords of C is denoted by $wt(C)$. Note that $d(C) = wt(C)$ for a linear code C. For a subspace C of F_q^n, the support of C, denoted by $supp(C)$, is defined as follows

$$supp(C) = \{i : \exists (v_1, v_2, \ldots, v_n) \in C; v_i \neq 0\}.$$

Also, we define the r-th generalized Hamming weight (GHW) as follows

$$d_r = d_r(C) = \min\{\|D\| : D \subset C, \dim(D) = r\},$$

where $\|D\| = |supp(D)|$. The Weight Hierarchy (WH) for a code is defined as the sequence of GHW s of that code.

In this paper we shall study the generalized Hamming weight for $(u, u+v)$-construction of codes.

2. Main Results

Definition 2.1. [6] ($(u, u+v)$-construction) Let C_i be an $[n, k_i, d_i]$ linear code over F_q for $i = 1, 2$. Then $(u, u+v)$-construction of these codes, denoted by C, is defined as follows:

$$C = \{(u, u+v) : u \in C_1, v \in C_2\}.$$

*Speaker

THEOREM 2.2. [6] *Let C_i be an $[n, k_i, d_i]$-linear code over F_q, for $i = 1, 2$. Then the code C defined as $C = \{(u, u + v) : u \in C_1, v \in C_2\}$ is a $[2n, k_1 + k_2, \min\{2d_1, d_2\}]$-linear code over F_q.*

THEOREM 2.3. [6] *For any prime power q and $x, y \in F_q^n$, we have*

$$wt(x) + wt(y) \geq wt(x + y) \geq wt(x) - wt(y).$$

THEOREM 2.4. *Let C_i be an $[n, k_i, d_i]$ linear code over F_q for $i = 1, 2$. Then the upper bound for the weight hierarchy of code $C = \{(u, u + v) : u \in C_1, v \in C_2\}$ is as follows:*

$$d_r(C) \leq \begin{cases} \min\{2d_r(C_1), d_r(C_2)\}, & 1 \leq r \leq k_1, \\ d_r(C_2), & k_1 < r \leq k_2. \end{cases}$$

PROOF. Let $\dim(C_1) = k_1$, $\dim(C_2) = k_2$, and $k_1 < k_2$. We have following cases:

i) Let $1 \leq r \leq k_1$. Then

$\exists D_1 \subseteq C_1, \dim D_1 = r, D =< \alpha_1, \alpha_2, \ldots, \alpha_r >, \| D_1 \| = d_r(C_1)$,

$\exists D_2 \subseteq C_2, \dim D_2 = r, D =< \beta_1, \beta_2, \ldots, \beta_r >, \| D_2 \| = d_r(C_2)$,

where $\alpha_1, \alpha_2, \ldots, \alpha_r \in C_1$ and $\beta_1, \beta_2, \ldots, \beta_r \in C_2$.

Let $D =< (\alpha_1, \alpha_1), (\alpha_2, \alpha_2), \ldots, (\alpha_r, \alpha_r) >$. Note that $\dim(D) = r, D \subseteq C$. Also we have $\|D\| = 2\|\alpha_1, \alpha_2, \ldots, \alpha_r\| = 2d_r(C_1)$. Let

$$D' =< (0, \beta_1), (0, \beta_2), \ldots, (0, \beta_r) > .$$

We have $\dim(D') = r$, $\| \beta_1, \beta_2, \ldots, \beta_r \| = d_r(C_2)$. Then, D and D' are satisfying the following relation

$$\{\|D\|; D \subseteq C, \dim(D) = r\},$$

which implies $d_r(C) \leq \min\{2d_r(C_1), d_r(C_2)\}$.

ii) Let $r > k_1$. We can suppose that $r = k_1 + i, 1 \leq i \leq k_2$. We continue with fixed i. There exist $\beta_1, \beta_2, \ldots, \beta_{k_1+i} \in C_2$ such that

$$d_{k_1+i}(C_2) = \| \beta_1, \beta_2, \ldots, \beta_{k_1+i} \| .$$

Let $D =< ((0, \beta_1), (0, \beta_2), \ldots, (0, \beta_{k_1+i}) >$. Hence we have

$$\|D\| = \|\beta_1, \beta_2, \ldots, \beta_{k_1+i}\| = d_{k_1+i}(C_2) = d_r(C_2).$$

Therefore $d_r(C) \leq d_r(C_2)$ for $(k_1 < r \leq k_2)\}$. The result is obtained by using (i) and (ii). □

THEOREM 2.5. *Let C_i be an $[n, k_i, d_i]$ linear code over F_2 for $i = 1, 2$. The weight hierarchy of code $C = \{(u, u + v) : u \in C_1, v \in C_2\}$ is as follows:*

$$d_2(C) = \min\{2d_2(C_1), d_2(C_2)\}.$$

PROOF. Let $d_2(C) = \|D\|$. So $D =< (\alpha_1, \alpha_1 + v_1), (\alpha_2, \alpha_2 + v_2) >$ in which $\alpha_1, \alpha_2 \in C_1$ and $v_1, v_2 \in C_2$. Hence by using the concept of *suppD* and Theorem 2.3, we have

$$\| D \| = \| \alpha_1, \alpha_2 \| + \| \alpha_1 + v_1, \alpha_2 + v_2 \| = wt(\alpha_1 - \alpha_2) + wt(\alpha_1 - \alpha_2 + v_1 - v_2)$$

$$\begin{cases} \geq wt(v_1 - v_2) = \| v_1, v_2 \| \geq d_2(C_2), & v_1 \neq v_2, \\ = 2wt(\alpha_1 - \alpha_2) = 2 \| \alpha_1, \alpha_2 \| \geq 2d_2(C_1), & v_1 = v_2. \end{cases}$$

So, $\min\{2d_2(C_1), d_2(C_2)\} \leq d_2(C)$.

On the other hand, we have

$$\exists D_1 \subseteq C_1, \dim D_1 = 2, D =< \alpha_1, \alpha_2 >, \| D_1 \| = d_2(C_1),$$

$$\exists D_2 \subseteq C_2, \dim D_2 = 2, D =< \beta_1, \beta_2 >, \| D_2 \| = d_2(C_2).$$

Let $D =< (\alpha_1, \alpha_1), (\alpha_2, \alpha_2) >$. Then $\|D\| = \|\alpha_1, \alpha_2\| + \|\alpha_1, \alpha_2\| = 2\|\alpha_1, \alpha_2\| = 2d_2(C_1)$. Also, let $D' =< (0, \beta_1), (0, \beta_2) >$. So $\|D'\| = \|\beta_1, \beta_2\| = d_2(C_2)$. Note that D and D' are satisfying

$$\{\|D\|; D \subseteq C, \dim(D) = 2\}.$$

Therefore, we have $d_2(C) \leq \min\{2d_2(C_1), d_2(C_2)\}$. Finally we have

$$d_2(C) = \min\{2d_2(C_1), d_2(C_2)\}.$$

□

Acknowledgement

I'm thankful to the referee for his/her helpful remarks which have contributed to improve the presentation of the article.

References

1. W. S. Ching, *Linear equation over commutative rings*, Linear Algebra Appl. **18** (3) (1977) 257–266.
2. S. T. Dougherty and S. Han, *Higher weights and generalized MDS codes*, Korean Math. Soc. **47** (6) (2010) 1167–1182.
3. S. T. Dougherty, S. Han and H. Liu, *Higher weights for codes over rings*, AAECC **22** (2011) 113–135.
4. F. Farhang Baftani and H. R. Maimani, *The weight hierarchy of Hadamard codes*, Facta UniverSitatis (NIS). **34** (4) (2019) 797–803.
5. G. L. Feig, K. K. Tzeng and V. K. Wei, *On the generalized Hamming weights of several classes of cyclic codes*, IEEE Trans. Inform. Theory **38** (3) (1992) 1125–1126.
6. S .Ling and C. Xing, *Coding Theory a First Course*, Cambridge university press, Cambridge, UK, 2004.
7. V. K. Wei, *Generalized Hamming weights for linear codes*, IEEE Trans. Inform. Theory **37** (5) (1991) 1412–1418.

E-mail: far_farhang2007@yahoo.com; farhangfarzaneh1006@gmail.com

Quantum Codes From Quadratic Residue Codes over $\mathbb{F}_{q^r} + v\mathbb{F}_{q^r}$

Arezoo Soufi Karbaski
Bu Ali Sina University, University of Hamedan, Hamedan, Iran
and Karim Samei*
Bu Ali Sina University, University of Hamedan, Hamedan, Iran

Abstract. In this paper, we present a method to construct quantum codes over $\mathbb{F}_{q^r}$ from the Gray images of quadratic residue codes over the ring $R = \mathbb{F}_{q^r} + v\mathbb{F}_{q^r}$, where $v^2 = v$ and q is an odd prime number. In particular, we obtain a few quantum maximum distance separable (MDS) codes over $\mathbb{F}_{q^r}$ from quadratic residue codes and their extended over R.
Keywords: Quantum codes, Quadratic residue codes, Extended quadratic residue codes.
AMS Mathematical Subject Classification [2010]: 94B05, 94B15, 81P70.

1. Introduction

The class of quantum error-correcting codes plays a very significant role in quantum communication and their successful application in quantum computation. In 2014, Kaya et al. presented the structure of QR codes over $\mathbb{F}_p + v\mathbb{F}_p$ and obtained optimal self-dual codes and formally self-dual codes which have the best minimum distance from the Gray images of QR codes over $\mathbb{F}_p + v\mathbb{F}_p$ in [5]. Recently, Samei and Soufi studied the structure of QR codes over $\mathbb{F}_{p^r} + u_1\mathbb{F}_{p^r} + \cdots + u_t\mathbb{F}_{p^r}$ in [7], where $r, t \geq 1$. Many good quantum error-correcting codes over finite field have been constructed using the Gray images of cyclic codes over finite rings.

In this paper, let $R = \mathbb{F}_{q^r} + v\mathbb{F}_{q^r}$, where q is an odd prime number, r is a finite natural number and $v^2 = v$. The ring R is a finite principal ideal ring of order q^{2r} and characteristic q. It has two maximal ideals $< 1 - v >$ and $< v >$. By Chinese Reminder Theorem, we have $R = (1 - v)\mathbb{F}_{q^r} \oplus v\mathbb{F}_{q^r}$, which implies that for any $r \in R$ there are $a, b \in \mathbb{F}_{q^r}$ such that $r = (1 - v)a + vb$.

A linear code C over ring R of length n is a R-submodule of R^n. The Hamming weight of a codeword is the number of non-zero components.

Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ be two elements of R^n. The Euclidean inner product of vectors $\mathbf{x}, \mathbf{y}$ is $< \mathbf{x}, \mathbf{y} >_E = \Sigma_{i=1}^{n} x_i y_i$. The dual or orthogonal of C denoted $C^{\perp}$ is defined as

$$C^{\perp} = \{x \in R^n : < x, y >_E = 0, \text{ for all } y \in C\}.$$

The code C is self-orthogonal provided $C \subseteq C^{\perp}$ and self-dual provided $C = C^{\perp}$. A linear code C of length n over R is said to be cyclic if for any codeword $c \in C$, we have:

$\mathbf{c} = (c_0, c_1, \ldots, c_{n-2}, c_{n-1}) \in C$ implies that $\tau(c) = (c_{n-1}, c_0, c_1, \ldots, c_{n-2}) \in C$.

*Speaker

We let $R_n = \frac{R[X]}{<X^n-1>}$. Since C is a cyclic code of length n over R if and only if C is an ideal of R_n, we associate the vector $c = (c_0, c_1, \ldots, c_{n-1})$ in R^n with the polynomial $c(x) = c_0 + c_1x + c_2x^2 + \cdots + c_{n-1}x^{n-1}$ in R_n, where $x = X+ < X^n - 1 >$. A polynomial $e(x)$ in R_n is an idempotent if $e^2(x) = e(x)$.

A code is termed *even-like* if it has only even-like codewords; and it called *odd-like* if it is not even-like (a vector $\mathbf{x} = (x_0, x_1, \ldots, x_{n-1})$ in $\mathbb{F}_{q^r}$ is *even-like* provided that $\Sigma_{i=0}^{n-1} x_i = 0$). $X^n - 1$ has no repeated factors in $\mathbb{F}_{q^r}[X]$ if and only if $\gcd(n, q) = 1$ (See [4, Exercise 201]), an assumption we make throughout this paper.

2. Quadratic Residue Codes Over $\mathbb{F}_{q^r} + v\mathbb{F}_{q^r}$

This section is a generalization of [5]. Let p be an odd prime and we will assume that q is a prime such that q^r is a square module p, where $r \in \mathbb{N}$. Then there exist QR codes of length p over $\mathbb{F}_{q^r}$. Let $D_1 =< a_1(x) >$, $D_2 =< b_1(x) >$, $C_1 =< a_2(x) >$ and $C_2 =< b_2(x) >$ are QR codes over $\mathbb{F}_{q^r}$ such that $a_1(x)$ and $b_1(x)$ be the idempotent generators of $[p, \frac{p+1}{2}]$ QR codes and $a_2(x)$ and $b_2(x)$ be the idempotent generators of $[p, \frac{p-1}{2}]$ QR codes, see section 6 of [4, ch. 6].

In this section, without loss of generality, we can assume that Q_p and N_p are the sets of nonzero quadratic residue and quadratic nonresidue modulo p, respectively.

THEOREM 2.1. *Let $C = (1 - v)C_1 \oplus vC_2$ be a cyclic code of length n over R. Then there exists a unique idempotent generator of the form $t(x) = (1 - v)t_1(x) + vt_2(x)$, where $t_i(x)$ is a idempotent generator in $\frac{\mathbb{F}_{q^r}[X]}{<X^n-1>}$ and $C^\perp =< 1-t(x^{-1}) >$.*

DEFINITION 2.2. With the above notation, we define the four QR codes over R as follows:

$$\begin{aligned} Q_1 &= (1 - v)D_1 \oplus vD_2, \\ Q_2 &= (1 - v)D_2 \oplus vD_1, \\ Q_1' &= (1 - v)C_1 \oplus vC_2, \\ Q_2' &= (1 - v)C_2 \oplus vC_1. \end{aligned}$$

By Theorem 2.1, $p_1(x) = (1-v)a_1(x)+vb_1(x)$ and $q_1(x) = (1-v)b_1(x)+va_1(x)$ and $p_2(x) = (1-v)a_2(x)+vb_2(x)$ and $q_2(x) = (1-v)b_2(x)+va_2(x)$ are idempotent generators of QR codes in the above definition, respectively.

We investigate the properties of QR codes over R for the (distinct) cases $p \equiv -1 \pmod 4$ and $p \equiv 1 \pmod 4$. It is obvious that Q_i' is the even-like subcode of Q_i, where $i = 1, 2$. We note that $\overline{j(x)} = \frac{1}{p}(1 + x + x^2 + \cdots + x^{p-1})$ in $\frac{\mathbb{F}_{q^r}[X]}{<X^p-1>}$ is the idempotent generator of repetition code of length p. We recall that two codes are said to be *permutation equivalent* (or equivalent) if they differ only by a permutation of coordinates [4].

2.1. Case I.

If $p \equiv -1 \pmod 4$, then the codes have the following properties.

THEOREM 2.3. *Let the situation be as in definition 2.2 and $p \equiv -1 \pmod 4$, then*

(1) *Q_1 and Q_1' are equivalent to Q_2 and Q_2', respectively.*

(2) $Q_1 \cap Q_2 =< \overline{j(x)} >$ *and* $Q_1 + Q_2 = R_p$.
(3) $|Q_1| = q^{r(p+1)} = |Q_2|$.
(4) $Q_1 = Q_1' + < \overline{j(x)} >$ *and* $Q_2 = Q_2' + < \overline{j(x)} >$.
(5) $|Q_1'| = q^{r(p-1)} = |Q_2'|$.
(6) $Q_1^{\perp} = Q_1'$ *and* $Q_2^{\perp} = Q_2'$ *and* Q_1' *and* Q_2' *are self-orthogonal.*
(7) $Q_1' \cap Q_2' = \{0\}$ *and* $Q_1' + Q_2' =< 1 - \overline{j(x)} >$.

3. Extended Quadratic Residue Codes Over R

In this section, we can consider the extending odd-like quadratic residue codes over R in such a way that the extensions are self-dual or dual to each other.

REMARK 3.1. Let D_i be odd-like quadratic residue codes of length p over finite field $\mathbb{F}_{q^r}$, with $i = 1, 2$. Then there exist different QR extended codes over $\mathbb{F}_{q^r}$ as follows:

(1) $\hat{D}_i = \{(c_0, c_1, \ldots, c_{p-1}, c_p) |\ \ (c_0, c_1, \ldots, c_{p-1}) \in D_i \ \ with \ \ \Sigma_{i=0}^{p} c_i = 0\}$,
(2) $\overline{D_i} = \{(c_0, c_1, \ldots, c_{p-1}, -\gamma\Sigma_{i=0}^{p-1} c_i) |\ \ (c_0, c_1, \ldots, c_{p-1}) \in D_i \ \ and \ \ \gamma$ is a solution of equation $1 + \gamma^2 p = 0$ in $\mathbb{F}_{q^r}\}$.

We note that, in general the equation $1 + \gamma^2 p = 0$ has a solution γ in $\mathbb{F}_{q^r}$ if and only if p and -1 are both squares or both nonsquares in $\mathbb{F}_{q^r}$.

DEFINITION 3.2. The extended codes over R are formed by adding the same columns that are used to extend codes over $\mathbb{F}_{q^r}$.

THEOREM 3.3. [7, Theorem 2.15] Let D_i be odd-like quadratic residue codes of length p over finite field $\mathbb{F}_{q^r}$, with $i = 1, 2$. Then the following hold:

(1) $\hat{Q}_1 = (1 - v)\hat{D}_1 \oplus v\hat{D}_2$ and $\hat{Q}_2 = (1 - v)\hat{D}_2 \oplus v\hat{D}_1$,
(2) if the equation $1 + \gamma^2 p = 0$ has a solution γ in $\mathbb{F}_{q^r}$, then

$$\overline{Q}_1 = (1 - v)\overline{D}_1 \oplus v\overline{D}_2$$

and

$$\overline{Q}_2 = (1 - v)\overline{D}_2 \oplus v\overline{D}_1.$$

THEOREM 3.4. *Let* $p \equiv -1 \pmod 4$, *then* $\overline{Q_1}$ *and* $\overline{Q_2}$ *are self-dual.*

If $p \equiv 1 \pmod 4$, we can consider extending odd-like quadratic residue codes over R in such a way that extensions are dual to each other.

DEFINITION 3.5. Let $p \equiv 1 \pmod 4$. For $i = 1, 2$, we define

$$\tilde{D}_i = \{(c_0, c_1, \ldots, c_{p-1}, \tfrac{1}{p}\Sigma_{i=0}^{p-1} c_i) |\ \ (c_0, c_1, \ldots, c_{p-1}) \in D_i\}.$$

DEFINITION 3.6. Let $p \equiv 1 \pmod 4$. The extended codes $\tilde{Q}_1$ and $\tilde{Q}_2$ over R are formed by adding the same columns that are used to extend codes over $\mathbb{F}_{q^r}$.

THEOREM 3.7. *If* $p \equiv 1 \pmod 4$, *then*

$$\tilde{Q}_1 = (1 - v)\tilde{D}_1 \oplus v\tilde{D}_2,$$

and

$$\tilde{Q}_2 = (1 - v)\tilde{D}_2 \oplus v\tilde{D}_1.$$

THEOREM 3.8. *If* $p \equiv 1 \pmod 4$, *then the dual of* $\tilde{Q}_1$ *is* $\hat{Q}_2$ *and the dual of* $\tilde{Q}_2$ *is* $\hat{Q}_1$.

4. Stabilizer Quantum Codes From Quadratic Residue Codes Over R

Let q be a prime power. A q-ary quantum code of length n ia a subspace of $\mathbb{C}^{q^n}$. For $(\mathbf{a}|\mathbf{b})$ in R^{2n}, where $\mathbf{a} = (a_1, a_2, \ldots, a_n)$ and $\mathbf{b} = (b_1, b_2, \ldots, b_n) \in R^n$ the symplectic weight is defined as

$$swt(\mathbf{a}|\mathbf{b}) = |\{k : a_k \neq 0 \text{ or } b_k \neq 0)\}|.$$

For $(\mathbf{a}|\mathbf{b})$ and $(\mathbf{a}'|\mathbf{b}')$ in R^{2n}, define the symplectic inner product by

$$< (\mathbf{a}|\mathbf{b}), (\mathbf{a}'|\mathbf{b}') >_s = \mathbf{b}.\mathbf{a}' - \mathbf{b}'.\mathbf{a},$$

where "." denotes the standard inner product. The following theorem is well known.

THEOREM 4.1. [6, Theorem 7] Let R be a Frobenius ring. An $((n, K, d))_R$ stabilizer code exists if and only if there exists an additive code $C \leq R^{2n}$ of size $|C| = \frac{|R^n|}{K}$ such that $C \leq C^{\perp_s}$ and $swt(C^{\perp} \setminus C) = d$ if $K > 1$ and $swt(C^{\perp} - \mathbf{0}) = d$ if $K = 1$.

By Theorem 4.1, we can obtain stabilizer codes over finite Frobenius ring.

THEOREM 4.2. [3, Theorem 3.3] Let R be a finite Frobenius ring. Further, let C_1 and C_2 denote two classical linear codes over R with parameters (n, K_1, d_1) and (n, K_2, d_2) such that $C_2^{\perp} \subset C_1$. Then there exists an $((n, \frac{K_1 K_2}{|R^n|}, d))_R$ stabilizer code with minimum distance $d = \min\{wt(c)|c \in (C_1 \setminus C_2^{\perp} \cup C_2 \setminus C_1^{\perp})\}$ that is pure to $\min\{d_1, d_2\}$.

COROLLARY 4.3. *Let R be a finite Frobenius ring and C be a (n, K, d) linear code over R. If $C^{\perp} \subset C$, then there exists an $((n, \frac{K^2}{|R^n|}, d))_R$ stabilizer code with minimum distance $d = d(C \setminus C^{\perp})$.*

THEOREM 4.4. *Let $p \equiv -1 \pmod 4$. Then there exist $[[2p, 2, d_L(Q_1 \setminus Q_1^{\perp})]]_{q^r}$ and $[[2p, 0, d]]_{q^r}$ quantum codes with minimum distance $d = \min\{d_L(Q_1), d_L(Q_1')\}$.*

THEOREM 4.5. *Let $p \equiv 1 \pmod 4$. Then there exists an $[[2p, 2, d_L]]_{q^r}$ quantum code, where $d_L = \min\{wt_L(c)|c \in (Q_1 \setminus Q_2^{\perp} \cup Q_2 \setminus Q_1^{\perp})\}$. Moreover, there exists an $[[2p, 0, d]]_{q^r}$ quantum code with minimum distance $d = \min\{d_L(Q_1), d_L(Q_2')\}$.*

THEOREM 4.6. *If $p \equiv -1 \pmod 4$, then there exists an $[[2(p+1), 0, d_L(\overline{Q_1})]]_{q^r}$ stabilizer quantum code.*

THEOREM 4.7. *If $p \equiv 1 \pmod 4$, then there exists an $[[2(p+1), 0, d_L(\tilde{Q}_1)]]_{q^r}$ stabilizer quantum code.*

In the following examples, all computations are carried in *magma computational algebra system* [1].

EXAMPLE 4.8. Over $\mathbb{F}_7$, we have

$$X^3 - 1 = (X + 3)(X + 5)(X + 6),$$

where they are the factorization of $X^3 - 1$ into irreducible polynomials in $\mathbb{F}_7[X]$. For $p = 3$ the quadratic residue codes D_1 and D_2 over $\mathbb{F}_7$ is generated by the idempotent generator polynomials $4x^2 + x + 3$ and $x^2 + 4x + 3$, respectively.

So $Q_1 =< (1-v)(4x^2+x+3)+v(x^2+4x+3) >$ is the quadratic residue code of length 3 over $R = \mathbb{F}_7 + v\mathbb{F}_7$ and $Q_1^{\perp} =< (1-v)(6x^2+3x+5)+v(3x^2+6x+5) >$

is its dual code. Since $d_L(Q_1 \setminus Q_1^{\perp}) = 3$, by Theorem 4.4, there exists a quantum MDS code with parameters $[[6, 2, 3]]$ over $\mathbb{F}_7$. Moreover, there exists a quantum code with parameters $[[6, 0, 3]]$ over $\mathbb{F}_7$.

We presented an optimal formally self-dual code with parameters $[12, 6, 6]_9$ in [7, Example 5.12]. Now we obtain quantum code over $\mathbb{F}_9$ from it.

EXAMPLE 4.9. Let $R = \mathbb{F}_9 + v\mathbb{F}_9$ and let ρ be a primitive element of the finite field $\mathbb{F}_9$ and $p = 5$. The QR code Q_1 of length 5 over R is generated by the idempotent generator $p_1(x) = (1-v)(2\rho e_1(x)+2\rho^3 e_2(x))+v(2\rho^3 e_1(x)+2\rho e_2(x))$, where $e_1(x) = x + x^4$ and $e_2(x) = x^2 + x^3$. It should be noted that

$$\hat{Q_1} = \overline{Q_1} = \{(c_0, c_1, c_2, c_3, c_4, -\Sigma_{i=0}^{4} c_i) : \quad (c_0, c_1, c_2, c_3, c_4) \in Q_1 \ \}.$$

So we have $d_L(\hat{Q_1}) = 6$. Then it corresponds to a $[12, 6, 6]_9$ formally self-dual code which has the best possible minimum distance by [2]. So by Theorem 4.7, there exists a quantum code with parameters $[[12, 0, 6]]$ over $\mathbb{F}_9$ which $d = 6$ is close to the best minimum distance.

References

1. W. Bosma and J. Cannon, *Handbook of Magma Functions*, Department of Mathematics, University of Sydney, 1994.
2. M. Grassl, `http://codetables.de`, accessed on 04.11.2012.
3. K. Guenda and T. A. Gulliver, *Quantum codes over rings*, Int. J. Quantum Inform. **12** (4) (2014) 1450020.
4. W. C. Huffman and V. Pless, *Fundamentals of Error-Correcting Codes*, Cambrigde University press, New York, 2003.
5. A. Kaya, B. Yildiz and I. Siap, *Quadratic residue codes over $\mathbb{F}_p + v\mathbb{F}_p$ and their Gray images*, J. Pure App. Algebra **218** (11) (2014) 1999–2011.
6. S. Nedella and A. Klappenecker, *Stabilizer codes over Frobenius rings*, In Proc. IEEE. Int. Sympt. Information Theory, Cambridge, MA, (2012) pp. 165–169.
7. K. Samei and A. Soufi, *Quadratic residue codes over $\mathbb{F}_{p^r} + u_1\mathbb{F}_{p^r} + \cdots + u_t\mathbb{F}_{p^r}$*, Adv. Math. Commun. **11** (4) (2017) 791–804.

E-mail: arezoo.sufi@basu.ac.ir
E-mail: samei@ipm.ir

Isogeny Problems in Cryptography

Leila Goodarzi*
Faculty of Mathematical Sciences, University of Kashan, Kashan, Iran
and Hassan Daghigh
Faculty of Mathematical Sciences, University of Kashan, Kashan, Iran

Abstract. Many cryptosystems are based on the difficulty of the discrete logarithm problem in a cyclic group and the integer factorization problem. There are quantum polynomial attacks on these problems. Isogeny problems are believed to be quantum-resistant. Here we give a brief review of some problems involving isogenies on elliptic curves.
Keywords: Elliptic curve, Isogeny, Cryptography.
AMS Mathematical Subject Classification [2010]: 14H52, 94A60, 11T71.

1. Introduction

Nowadays, elliptic curves are extensively used in public-key cryptography, for example, in the key agreement protocols, encryption algorithms, and digital signatures. Based on the Shore algorithm, using quantum computers, it is possible to solve the discrete logarithm problem and the problem of decomposing large numbers in polynomial time [14]. Therefore it is crucial to look for protocols with security based on problems resistant to quantum attacks. As examples of quantum-resistant cryptosystems, we can mention lattice-based cryptography, hash-based cryptography, multivariate cryptography, code-based cryptography, and lately cryptography based on isogeny problems. In isogeny-based cryptography, the key sizes and messages exchanged are smaller than those for other post-quantum cryptosystems. Although isogeny-based cryptography is less efficient than cryptosystems, such as lattice-based cryptography, due to the low availability of efficient cryptosystems in this field, the study in these cryptosystems is valuable. Isogeny cryptosystems based on ordinary curves were suggested by Couveignes [5] for the first time. The supersingular curve case was first developed in a hash function making by Charles, Lauter, and Goren [3]. Subsequent cryptosystems based on the supersingular curve were suggested by Jao and De Feo [7].

In the following, in Section 2, we briefly review the basic concepts about elliptic curves. In Section 3, we mention isogeny problems. Then we review the SIDH protocol that is based on one of these problems. In Section 4, we review methods to construct isogenies and give a concrete example.

*Speaker

2. Preliminaries

We summarize the necessary background on elliptic curves. For more details, one can see [15].

Let F_q be the finite field of order q, where q is a power of prime p, $q = p^k$ and $p \neq 2, 3$. Assume that $\overline{F_q} = \bigcup_{n \geq 1} F_{q^n}$ is the algebraic closure of F_q. An elliptic curve over F_q is defined by a Weierstrass equation $E : y^2 = x^3 + ax + b$, where $a, b \in F_q$ and $4a^3 + 27b^2 \neq 0$. The set of F_q-rational points of E is

$$E(F_q) = \{(x, y) \in F_q^2 : y^2 = x^3 + ax + b\} \cup \{\mathcal{O}\},$$

where $\mathcal{O}$ is the point $(X : Y : Z) = (0 : 1 : 0)$ on the projective curve $Y^2Z = X^3 + aXZ^2 + bZ^3$. The set $E(F_q)$ is an abelian additive group under the chord and tangent rule with $\mathcal{O}$ as the identity element. The j-invariant of E is $j(E) = 1728.4a^3/(4a^3 + 27b^2)$. For $n > 1$ the set of n-torsion points is defined as $E[n] = \{P \in E(\overline{F_q}) : [n]P = \mathcal{O}\}$. If $\gcd(p, n) = 1$, $E[n]$ is a direct product of two cyclic groups of order n and hence $\#E[n] = n^2$. $E[p]$ has either one or p elements. E is called supersingular in the first case and ordinary in the second case.

Elliptic curves $E(F_q) : y^2 = x^3 + ax + b$ and $E'(F_q) : y^2 = x^3 + a'x + b'$ over an extension field F_{q^r} are isomorphic if there exists $u \in \overline{F_{q^r}}^*$ such that $a' = u^4a$ and $b' = u^6b$. In this case the corresponding isomorphism $f : E \to E'$ is defined by $(x, y) \mapsto (u^2x, u^3y)$. There is an isomorphism $f : E \to E'$ if and only if $j(E) = j(E')$.

Let E, E' be two elliptic curves over F_q, an isogeny is a morphism $\varphi : E \to E'$ such that $\varphi(\mathcal{O}) = \mathcal{O}'$. For every isogeny $\varphi : E \to E'$ there are rational functions $R_1(x, y)$, $R_2(x, y) \in F_q[x, y]$ such that

$$\varphi(x, y) = (R_1(x, y),\ R_2(x, y)).$$

The degree of an isogeny is the number of points in the kernel (except inseparable isogenies). As an example, for every $n \in \mathbf{N}$ the multiplication by n map $[n]_E$ on an elliptic curve E which is defined by $[n]_E P = P + P + \cdots + P$ (n times) is an isogeny. It is easy to see that the kernel of $[n]$ is $E[n]$. For every isogeny $\varphi : E \to E'$ of degree l, there exists a dual isogeny $\widehat{\varphi} : E' \to E$ such that $\varphi\widehat{\varphi} = [l]_{E'}$ and $\widehat{\varphi}\varphi = [l]_E$. Given a prime $l \neq p$, the torsion group $E[l]$ contains exactly $l + 1$ cyclic subgroups of order l; each one corresponds to a different isogeny. An endomorphism on E is an isogeny from E to itself. The set of endomorphisms of an elliptic curve, denoted by $End(E)$, has a ring structure with point-wise addition and function composition. The ring $End(E)$ is either order in an imaginary quadratic field or a maximal order in a quaternion algebra.

3. Isogeny Problems

In this section, seven isogeny problems are mentioned and we review the most important protocol based on one of them and give a simple example.

Problem 1. *Let two isogenous elliptic curves E and E' are given, find an isogeny $\varphi : E \to E'$.*

According to Tate's theorem, two elliptic curves E, E' over F_q are isogenous over F_q if and only if $\#E(F_q) = \#E'(F_q)$ [15]. So the decisional problem

of whether there is an isogeny is solvable in polynomial time since the Schoof algorithm computes the number of points in polynomial time [15]. The fastest algorithm proposed for solving problem one using classical computers is Galbraith and Stolbunov method for ordinary curves with running time $\widetilde{O}(q^{1/4})$ [9] and Delfs and Galbraith method for supersingular curves with running time $\widetilde{O}(p^{1/2})$ [6]. Using quantum computers, for ordinary curves, a quantum subexponential algorithm using the commutativity of the endomorphism ring of these curves is presented in [4]. For supersingular curves, due to the noncommutativity of the endomorphism ring, this attack is not practical. The best attack known for supersingular curves is the Jao attack with running time $\widetilde{O}(p^{1/4})$ [2]. Since the attack in the supersingular case is exponential in quantum computers, this problem has been used to design quantum-resistant protocols.

Problem two, three, and four are variants of problem one with additional conditions.

PROBLEM 2. *Let E and E' be two isogenous elliptic curves, $P \in E$ and $Q \in E'$. Find an isogeny $\varphi : E \to E'$ such that $\varphi(P) = Q$.*

Problem two is a variant of problem one with an additional condition. This problem is a generalization of the discrete logarithm problem. If $E = E'$ and φ is the scalar multiplication $[m]$, then problem two will be the discrete logarithm problem in E.

PROBLEM 3. *Given two elliptic curves E, E' over a finite field, isogenous of degree m, find an isogeny $\varphi : E \to E'$ of degree m.*

For isogeny problems to be hard the isogeny must have a large degree, so that representation as a rational map not efficient enough.

PROBLEM 4. *Given two elliptic curves E, E' over a finite field F_q, such that $\#E(F_q) = \#E'(F_q)$, find an isogeny $\varphi : E \to E'$ of smooth degree.*

Note that if E, E' be two supersingular elliptic curves over F_{p^2}, then $\#E(F_{p^2}) = \#E'(F_{p^2}) = (p+1)^2$, so based on Tate's theorem, every two supersingular elliptic curves are isogenous over F_{p^2}. The fastest algorithm known for problem four uses a meet-in-the-middle strategy with running time $O(p^{1/2})$ [8].

PROBLEM 5. *Let p be a prime number, and E be a supersingular elliptic curve over F_{p^2}. Determine the endomorphism ring of E.*

There are various possible ways to show elements of $End(E)$. One method is to show explicit isogenies $\varphi : E \to E$ as rational functions. Since the degree is usually exponential, this is not a beneficial representation. Another way is the representation as a $\mathbf{Z}$-module in quaternion algebra. In this case, the endomorphism ring has a polynomial sized representation based on the basis $\{1, i, j, k\}$. Kohel regarded the endomorphism ring computation problem in his Ph.D. thesis [13]. In the supersingular case, it is believed that problem five and problem one are equivalents [11]. For the ordinary case, problem five is much easier than problem one [1, 13], Since there exists a subexponential algorithm to compute the endomorphism ring of ordinary curves [1], while the best algorithm to compute isogenies is exponential. A subexponential quantum algorithm to compute

an isogeny between ordinary curves is proposed in [4]. The endomorphism ring computation problem is a crucial problem for isogeny-based cryptography [1, 12]. Problem five has been studied for more than twenty years but not as widely as classical problems like discrete logarithm or integer factorization.

One of the prominent quantum-resistant protocols is the Jao key exchange protocol, based on the isogeny problem between supersingular elliptic curves. In the following, after a brief overview of the Jao scheme, we present an example.

Supersingular Isogeny Diffie-Hellman (SIDH). Let l_1, l_2 be distinct small primes, and $e_1, e_2 \in \mathbf{N}$. Then choose a random small integer $f \in \mathbf{N}$ so that $p = l_1^{e_1} l_2^{e_2} f - 1$ is prime. Let E be a supersingular elliptic curve over F_{p^2} such that the group structure of $E(F_{p^2})$ be a product of two cyclic groups of order $l_1^{e_1} l_2^{e_2} f$. Let $E[l_1^{e_1}] =< R_1, S_1 >$ and $E[l_2^{e_2}] =< R_2, S_2 >$. The SIDH public parameters are (E, R_1, S_1, R_2, S_2).

Alice chooses secret random integers $0 \leq m_1, n_1 < l_1^{e_1}$ and set $T_1 = [m_1]R_1 + [n_1]S_1$. Then Alice computes an isogeny $\phi_A : E \to E_A = E/ < T_1 >$ and publishes $(E_A, \varphi_A(R_2), \varphi_A(S_2))$.

Similarly, Bob chooses $0 \leq m_2, n_2 < l_2^{e_2}$ and computes $\varphi_B : E \to E_B = E/ < T_2 >$, where $T_2 = [m_2]R_2 + [n_2]S_2$ and publishes $(E_B, \varphi_B(R_1), \varphi_B(S_1))$.

To compute the shared key, Alice computes

$$T_1' = [m_1]\varphi_B(R_1) + [n_1]\varphi_B(S_1) = \varphi_B([m_1]R_1 + [n_1]S_1) = \varphi_B(T_1),$$

and an isogeny $\varphi_{BA} : E_B \to E_{BA} = E_B/ < T_1' >$.

Similarly, Bob computes an isogeny $\varphi_{AB} : E_A \to E_{AB} = E_A/ < T_2' >$, where

$$T_2' = [m_2]\varphi_A(R_2) + [n_2]\varphi_A(S_2) = \varphi_A([m_2]R_2 + [n_2]S_2) = \varphi_A(T_2).$$

The elliptic curves E_{AB} and E_{BA} are isomorphic, so $j(E_{AB}) = j(E_{BA})$. Then the shared key is $j(E_{AB})$ [7]. The security of the mentioned protocol is based on problem six, below. The security of the SIDH protocol relies on problems, seem to be easier to solve than the endomorphism ring computation problem since extra points are revealed, and special primes are used. On the other hand, there is a strong constraint on the degree of isogeny that may make it harder. Besides, if $End(E)$ and $End(E_A)$ are known, then we can compute the specific isogeny of degree $l_1^{e_1}$ (Section 4 of [12]). So the hardness of the endomorphism ring computation of supersingular elliptic curves determines the security.

Example 3.1. Let $p = 1511 = 2^3.3^3.7 - 1$ be a prime and $E(F_{1511^2}) : y^2 = x^3 + x$. Also let

$$E[2^3] =< R_1, S_1 >, \quad E[3^3] =< R_2, S_2 >,$$

$$R_1 = (974a + 186, 388a + 314), \quad S_1 = (88a + 136, 1461a + 886),$$

$$R_2 = (890a + 242, 427a + 810), \quad S_2 = (1470a + 417, 791a + 1479).$$

Alice chooses $m_1 = 2, n_1 = 5$ and compute $T_1 = (367a + 795, 879a + 181)$. Then

$$E_A(F_{1511^2}) : y^2 = x^3 + (647a + 1289)x + (1256a + 168),$$

$$\varphi_A(R_2) = (221a + 505, 1270a + 1002), \quad \varphi_A(S_2) = (1301a + 1156, 1176a + 1235).$$

Bob selects $m_2 = 14, n_2 = 22$ and compute $T_2 = (1445a + 559, 627a + 1303)$. Then

$$E_B(F_{1511^2}) : y^2 = x^3 + (926a + 1367)x + (18a + 431),$$

$$\varphi_B(R_1) = (795a + 1063, 1356a + 1037), \quad \varphi_B(S_1) = (949a + 1035, 847a + 901).$$

Then Alice and Bob compute the values

$$T_1' = (560a + 1302, 619a + 818), \quad T_2' = (520a + 1407, 448a + 737),$$

and the corresponding isogenies.

$$E_{AB} = E_{BA} : y^2 = x^3 + (1355a + 162)x + (1152a + 1331).$$

At the end, the shared key is computed as $j = 348a + 1299$. Calculations are done using SAGE software [17].

PROBLEM 6. *SIDH isogeny problem. Let (E, R_1, S_1, R_2, S_2) be a SIDH public key. Let E_A be the elliptic curve such that there is an isogeny $\varphi_A : E \to E_A$ of degree $l_1^{e_1}$. Let $R_2' = \varphi_A(R_2)$, $S_2' = \varphi_A(S_2)$. Given $(E, R_1, S_1, R_2, S_2, E_A, R_2', S_2')$, determine an isogeny $\varphi_A : E \to E_A$ of degree $l_1^{e_1}$ such that $R_2' = \varphi_A(R_2)$ and $S_2' = \varphi_A(S_2)$.*

Problem six is a kind of problem four where the images of two points are revealed. For $0 \leq x, y < l_2^{e_2}$ set $T = [x]R_2 + [y]S_2$. Then an attacker can compute $\varphi_A(T) = [x]R_2' + [y]S_2'$ and so has many pairs $(T, \varphi_A(T))$ on the graph of φ_A . By solving an interpolation problem, the attacker can compute φ_A. The difficulty is that φ_A has degree $l_1^{e_1}$ and so is described by rational functions of exponential degree. The SIDH protocol would be insecure if Alice also reveals $R_1' = \varphi_A(R_1), S_1' = \varphi_A(S_1)$. Since then an attacker can compute $x, y \in \mathbf{Z}$ such that $[x]R_1' + [y]S_1' = \mathcal{O}$ and $(x, l_1) = 1, (y, l_1) = 1$. This case is a kind of easy discrete logarithm problem because the orders of points are smooth $(l_1^{e_1})$. Then $[x]R_1 + [y]S_1$ is in the kernel of φ_A and the attacker can determine the kernel and then φ_A. There is an $O(l_1^{e_1/2})$ classical attack to SIDH problem [7] while a quantum algorithm due to Tani [1] solves the problem in $O(l_1^{e_1/3})$.

PROBLEM 7. *Decisional SIDH isogeny problem. Let (E, R_1, S_1, R_2, S_2) be a SIDH public key. Let E_A be an elliptic curve and let $R_2', S_2' \in E_A[l_2^{e_2}]$. Let $0 < n \leq e_1$ and the parameters $(E, R_1, S_1, R_2, S_2, E_A, R_2', S_2', n)$ are given, determine whether or not there exists an isogeny $\varphi : E \to E_A$ of degree l_1^n such that $R_2' = \varphi(R_2)$ and $S_2' = \varphi(S_2)$.*

If problem seven can be solved, then we can solve the SIDH isogeny problem easily. Let $u \in \mathbf{Z}$ be such that $ul_1 \equiv 1 \pmod{l_2}$. Given the $(E, R_1, S_1, R_2, S_2, E_A, R_2', S_2')$ one selects an l_1- isogeny $\psi : E_A \to E'$ and uses the decisional algorithm on $(E, R_1, S_1, R_2, S_2, E', [u]\psi(R_2'), [u]\psi(S_2'), e_1 - 1)$. If the decisional oracle says yes, then we ensure the first $e_1 - 1$ steps in the path from E to E_A is true [10]. We can solve the isogeny problem by repeating this process.

4. Computing Isogenies

Given an elliptic curve E and a subgroup G of E, there are two primary methods to find an elliptic curve E' and an isogeny $\varphi : E \to E'$ with kernel G. The first one is based on Velu's method [18], and the other is based on Kohel's approach [13]. We can consider the kernel as points in $E(\overline{F_q})$. The kernel specifies by the kernel polynomial, the lowest degree polynomial with roots only at x-coordinates of the kernel points, which is unique and monic. Velu's method takes the kernel

as input and returns the rational maps and codomain of the curve, while Kohel's approach takes the kernel polynomial as input. Velu's method includes sums over points in the G, so this method is efficient until $\#G$ is small. If the $\#G$ is smooth and not small, Jao makes a chain of isogenies by repeatedly using Velu's method and find isogeny [7].

EXAMPLE 4.1. Consider the elliptic curve $y^2 = x^3 + 8x + 13$ over finite field F_{251} and the point $R = (136, 223)$. Let $G =< R >$ be the subgroup of E of order 5. Using Velu's formula we have $v = 217$ and $w = 39$. The 5-isogeny $\varphi : E \to E'$ is given by

$$\varphi(x,y) = \Big(\frac{x^5 - 66x^4 - 104x^3 + 119x^2 + 76x - 35}{x^4 - 66x^3 - 70x^2 + 95x + 45}, \frac{x^6 - 99x^5 - 69x^4 + 60x^3 + 107x^2 + 16x + 83}{x^6 - 99x^5 - 103x^4 - 10x^3 + 76x^2 + 63x - 99}y\Big),$$

where $E' : y^2 = x^3 + 178x + 242$. We made calculations using SAGE software [17].

References

1. G. Bisson and A. V. Sutherland, *Computing the endomorphism ring of an ordinary elliptic curve over a finite field*, J. Number Theory **131** (5) (2011) 815–831.
2. J. F. Biasse, D. Jao and A. Sankar, *A Quantum Algorithm for Computing Isogenies between Supersingular Elliptic Curves*, Lecture Notes in Computer Science, Vol. 8885. Springer, Cham, 2014.
3. D. X. Charles, K. E. Lauter and E. Z. Goren, *Cryptographic hash functions from expander graphs*, J. Cryptology **22** (1) (2009) 93–113.
4. A. Childs, D. Jao and V. Soukharev, *Constructing elliptic curve isogenies in quantum subexponential time*, J. Math. Cryptol. **8** (1) (2014) 1–29.
5. J. M. Couveignes, *Hard Homogeneous Spaces*, IACR Cryptology ePrint Archive., Report **2006** (2006). `http://eprint.iacr.org/2006/291`
6. C. Delfs and S. D. Galbraith, *Computing isogenies between supersingular elliptic curves over F_p*, Des. Codes Cryptogr. **78** (2) (2016) 425–440.
7. L. De. Feo, D. Jao and J. Plût, *Towards quantum-resistant cryptosystems from supersingular elliptic curve isogenies*, J. Math. Cryptol. **8** (3) (2014) 209–247.
8. S. D. Galbraith, *Constructing isogenies between elliptic curves over finite fields*, LMS J. Comput. Math. **2** (1999) 118–138.
9. S. Galbraith and A. Stolbunov, *Improved algorithm for the isogeny problem for ordinary elliptic curves*, Appl. Algebra Engrg. Comm. Comput. **24** (2) (2013) 107–131.
10. S. D. Galbraith and F. Vercauteren, *Computational problems in supersingular elliptic curve isogenies*, Quantum Inf. Process. **17** (10) (2018). `DOI: 10.1007/s11128-018-2023-6`
11. S. D. Galbraith, C. Petit and J. Silva, *Identification protocols and signature schemes based on supersingular isogeny problems*, J. Cryptology **33** (1) (2020) 130–175.
12. S. D. Galbraith, C. Petit, B. Shani and Y. B. Ti, *On the Security of Supersingular Isogeny Cryptosystems*, Lecture Notes in Comput. Sci., Vol. 10031, Springer, Berlin, 2016.
13. D. R. Kohel, *Endomorphism Rings of Elliptic Curves over Finite Fields*, Ph. D. Thesis, University of California, Berkeley, 1996.
14. P. W. Shor, *Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer*, SIAM Rev. **41** (2) (1999) 303–332.
15. J. H. Silverman, *The Arithmetic of Elliptic Curves*, 2nd ed., Graduate Texts in Mathematics, Vol. 106. Springer, Dordrecht, 2009.
16. S. Tani, *Claw finding algorithms using quantum walk*, Theoret. Comput. Sci. **410** (50) (2009) 5285–5297.
17. *The Sage Developers. SageMath, the Sage Mathematics Software System (Version 8.2)*, 2018. `http://www.sagemath.org.`
18. J. Vélu, *Isogénies entre courbes elliptiques*, (French) C. R. Acad. Sci. Paris Sér. A-B **273** (1971) 238–241.

E-mail: l.goodarzi96@grad.kashanu.ac.ir
E-mail: hassan@kashanu.ac.ir

Contributed Talks

Differential Equations and Dynamical Systems

Emotional Rough Extreme Learning Machines for the Identification of Nonlinear Dynamic Systems

Ghasem Ahmadi*
Department of Mathematics, Payame Noor University, P. O. Box 19395-3697, Tehran, Iran

Abstract. Rough extreme learning machine (RELM) is a rough-neural network with a single hidden layer, where the weights of connections between the inputs and hidden neurons are randomly assigned and remain unchanged during the training process. In this work, on the basis of artificial emotional learning, a stable online learning algorithm for RELM is proposed. Emotional learning facilitate the error convergence in the training of neural models with increasing their memory depth. RELM with the proposed stable emotional learning algorithm that is called emotional RELM, is used to identify the discrete dynamic nonlinear systems. The efficiency of the proposed methodology are shown in simulation results.
Keywords: Discrete dynamic nonlinear system, System identification, Extreme learning machine, Emotional learning, Rough extreme learning machine.
AMS Mathematical Subject Classification [2010]: 93B30, 68T05.

1. Introduction

System identification is a field of science that concentrates on the building mathematical models for real systems from sampled data. In recent years, the neural networks are used for the identification of nonlinear systems because of their particular properties such as universal approximation and cooperating with parallel computation.

Often, in the identification of real systems, we are confronted with the uncertain and imperfect knowledge. Rough-neural network (R-NN) is introduced by Lingras [6], on the basis of rough set theory for dealing with uncertainty and imperfect knowledge in neural networks. A rough neuron is defined as a pair of conventional neurons, one for the upper bound and the other for the lower bound, where the information exchanges between them [6]. R-NN is used in different aspects such as traffic volume prediction [6], and system identification [1, 2, 3, 4].

Extreme learning machine (ELM) is a neural network with a single hidden layer, where the weights of connections between the inputs and hidden neurons are arbitrarily chosen and never updated. In 2006, for the first time ELM is proposed by Huang et al. [5]. The training process in ELMs occurs very faster than traditional neural networks. Recently, the Rough ELM (RELM) has been proposed by the author as a combination of ELM and R-NN [1], and it has been utilized for the identification of continuous-time nonlinear systems.

*Speaker

Training a neural network is very effective on their performances. Artificial emotional learning is a training strategy that has been introduced based on the emotions [7]. It has been formulated by the usage of an emotional signal which displays the emotions about the total performance of the system. Emotional learning facilitate the error convergence in the training of neural models with increasing their memory depth.

Recently, a general description of emotional learning has been stated in [4]. In this work, we use this technique for enhancing the performances of RELMs. Here, RELM with an online Lyapunov-based emotional learning algorithm that is called emotional RELM (ERELM), is used to identify the discrete dynamic nonlinear systems (DDNSs).

The reminder of work is organized as follows. The emotional learning is explained in Section 2. Section 3 describes the structure of RELM in the identification of DDNSs. In Sections 4, on the basis of emotional learning, a stable online learning algorithm is proposed for RELM. Section 5 gives the simulation results, and the conclusion is drawn in Section 6.

2. Emotional Learning

Emotional learning in artificial intelligence is a technique for accelerating the training speed. It uses the hidden information in the previous steps of training process with increasing the memory depth of neural networks. This technique has been proposed by Lucas et al. [7], and has been utilized in some problems of system identification and control. A general description of this type of emotional learning is presented in [4]. Let $\mathbf{e}_k$ be the vector of modeling error, where k is the time index. In the emotional learning, the emotional error $\mathbf{r}_k = k_1\mathbf{e}_k + k_2\Delta\mathbf{e}_k$ is used to achieve the learning laws, where k_1 and k_2 are the tuning parameters. Then, we have

$$\mathbf{r}_k = k_1\mathbf{e}_k + k_2\Delta\mathbf{e}_k = (k_1 + k_2)\mathbf{e}_k - k_2\mathbf{e}_{k-1}.$$

The proposed learning algorithm is developed using the emotional error $\mathbf{r}_k$ instead of $\mathbf{e}_k$. Emotional learning is a training strategy for neural networks which facilitates the error convergence by making it possible to use the last information of neural parameters. It is done by increasing the memory depth of neural network.

3. RELM in the Identification of DDNSs

Consider the RELM with rough neurons in the hidden layer and the conventional neurons in the output layer (Figure 1). Suppose that $\mathbf{u}_k$ and $\mathbf{y}_k$ be the input vector and the output vector of the nonlinear system, respectively. Let $\widehat{\mathbf{y}}_k$ be the output vector of RELM and

$$\mathbf{x}_k = [u^1_{k-1}, u^2_{k-1}, \ldots, u^m_{k-1}, \overline{y}^1_k, \underline{y}^1_{k-1}, \overline{y}^2_{k-1}, \underline{y}^2_{k-1}, \ldots, \overline{y}^m_{k-1}, \underline{y}^m_{k-1}, 1]^T.$$

be the input vector of RELM, where $\underline{y}^i$ is the lower bound and $\overline{y}^i$ is the upper bound of y^i ($i \in \{1, 2, \ldots, m\}$). The last component 1 of $\mathbf{x}_k$ is the input according to the biases of hidden neurons.

Let $\underline{V}_r$ and $\overline{V}_r$ be the weights of connections between all inputs and hidden lower bound neurons and the weights of connections between all inputs and hidden upper bound neurons, respectively. According to the definition of RELM, $\underline{V}_r$ and

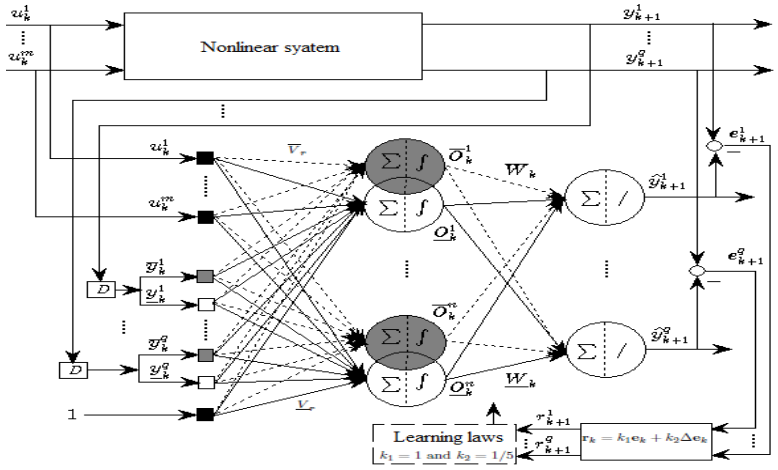


FIGURE 1. Structure of RELM model.

$\overline{V}_r$ are some randomly chosen numbers and remain unchanged during the training process. Suppose that $\underline{W}$ and $\overline{W}$ be the weights of connections between the hidden lower bound neurons and output neurons and the weights of connections between the hidden upper bound neurons and output neurons, respectively.

Further, $\underline{\phi}_k$ and $\overline{\phi}_k$ be the outputs of hidden lower bound neurons and the outputs of hidden upper bound neurons, respectively. Also, let ϕ be the activation function of hidden neurons. Let $\underline{V}_r$ and $\overline{V}_r$ contain the biases of hidden neurons. Then,

$$\underline{\phi}_k \quad = \quad \min\left(\phi(\underline{V}_r \mathbf{x}_k), \phi(\overline{V}_r \mathbf{x}_k)\right), \ \overline{\phi}_k = \max\left(\phi(\underline{V}_r \mathbf{x}_k), \phi(\overline{V}_r \mathbf{x}_k)\right),$$

and the output vector $\widehat{\mathbf{y}}_k$ of RELM is given by

$$\widehat{\mathbf{y}} \quad = \quad \underline{W}_k \underline{\phi}_k + \overline{W}_k \overline{\phi}_k. \tag{1}$$

A general form for DDNSs can be given by

$$\mathbf{z}_{k+1} = f(\mathbf{z}_k, \mathbf{u}_k), \tag{2}$$

where $\mathbf{z}_k$ and $\mathbf{u}_k$ represent the system states and inputs, respectively. By adding and subtracting $A\mathbf{z}_k$, (2) can be expressed as $\mathbf{z}_{k+1} = A\mathbf{z}_k + g(\mathbf{z}_k, \mathbf{u}_k)$, where $g(\mathbf{z}_k, \mathbf{u}_k) = f(\mathbf{z}_k, \mathbf{u}_k) - A\mathbf{z}_k$ represents the DDNS nonlinearity, and A is a matrix with eigenvalues in the unit circle. Assume that RELM can model $g(\mathbf{z}_k, \mathbf{u}_k)$ with an accuracy of ϵ_k using the parameters $\underline{W}_\star$ and $\overline{W}_\star$. Using (1), then we can write

$$\mathbf{z}_{k+1} \quad = \quad A\mathbf{z}_k + \underline{W}_\star \underline{\phi}_k + \overline{W}_\star \overline{\phi}_k + \epsilon_k. \tag{3}$$

In (3), the input vector of RELM is $\mathbf{x} = [\mathbf{u}_k, \overline{z}_k, \underline{z}_k]^T$.Parametric model of (2) can be constructed assuming the same structure as (3) by

$$\widehat{\mathbf{z}}_{k+1} = A\widehat{\mathbf{z}}_k + \widehat{\underline{W}}_k \underline{\phi}_k + \widehat{\overline{W}}_k \overline{\phi}_k,$$

where $\underline{\widehat{W}}_k$ and $\overline{\widehat{W}}_k$ represent the parameter estimations of $\underline{W}_\star$ and $\overline{W}_\star$ at the time index k, respectively. According to the structure of RELM in Fig 1, the estimated vector $\widehat{\mathbf{z}}_{k+1}$ is crisp.

4. Online Emotional Lyapunov-Based Learning Algorithm for RELM

Recently, some stable learning algorithm has been proposed for R-NN and RELM [1, 2, 4]. On the basis of these algorithms and emotional learning, we propose the following online learning algorithm for RELM:

$$\begin{aligned}
\underline{\widehat{W}}_{k+1} &= \underline{\widehat{W}}_k + \Big([k_1+k_2]P(A\mathbf{r}_k+\mathbf{r}_{k+1})\underline{\phi}_k - k_2P(A\mathbf{r}_k+\mathbf{r}_{k+1})\underline{\phi}_{k-1}\Big)\,\Gamma_1^{-1},\\
\overline{\widehat{W}}_{k+1} &= \overline{\widehat{W}}_k + \big([k_1+k_2]P(A\mathbf{r}_k+\mathbf{r}_{k+1})\overline{\phi}_k - k_2P(A\mathbf{r}_k+\mathbf{r}_{k+1})\overline{\phi}_{k-1}\big)\,\Gamma_2^{-1},
\end{aligned}$$

where Γ_1 and Γ_2 represent the gains of learning, and P is the matrix solution of Lyapunov equation $A^TPA-P=-Q$, where Q is a positive definite matrix. RELM with this learning algorithm is called ERELM.

5. Simulation Results

Consider the following DDNS with two inputs and two outputs

$$(4)\qquad \begin{cases} z^1_{k+1} &= \sin(\frac{z^1_k}{1+(z^2_k)^2}+u^1_k),\\ z^2_{k+1} &= \cos(1-\frac{z^1_kz^2_k}{1+(z^2_k)^2}-u^2_k) \quad z^1_0=z^2_0=0. \end{cases}$$

Identification of (4) is done by RELM with one hidden layer along with the external inputs of the form $u^1_k=\cos\left(\frac{2\pi k}{10}\right)$, $u^2_k=\sin\left(\frac{2\pi k}{10}\right)$. The hyperbolic tangent is the activation function of hidden neurons. The parameters $\underline{\widehat{V}}_r$ and $\overline{\widehat{V}}_r$ are some random numbers in the interval $[-1,1]$. The initial values of the parameters $\underline{\widehat{W}}_k$ and $\overline{\widehat{W}}_k$ are random numbers between -0.5 and 0.5. The input vector of RELM and ERELM is $\mathbf{x}=\left[u^1_k,u^2_k,\overline{z}^1_k,\underline{z}^1_k,\overline{z}^2_k,\underline{z}^2_k,1\right]^T$. The design parameters of learning algorithms are chosen as follow: $A=0.1I_2$, $n_h=7,\ 8$, $Q=I_2$, $\Gamma_1=\Gamma_2=10I_{n_h\times n_h}$, where n_h denotes the number of hidden rough neurons and Γ_1 and Γ_2 denote the learning rates. For ERELM, we suppose that $k_1=1$ and $k_2=1/5$. The MSEs of the identification of (4) with RELM in training and testing are listed in Table 1. Figure 2 shows the true states z^1 and z^2 of (4), the estimated states $\widehat{z}^1$ and $\widehat{z}^2$, and the errors e^1 and e^2 in testing of RELM (part (A)) and ERELM (part (B)) with seven hidden rough neurons. From the Table 1 and the Figure 2, we can conclude that in the identification of (4), the performance of ERELM is better than RELM.

6. Conclusion

In this work, on the basis of emotional learning and Lyapunov stability theory, an online learning algorithm is proposed for RELM in the identification of DDNSs. Then, the effectiveness of emotional learning in the accelerating of training process is shown in simulation results. The future works focus on the more applications of this approach in control problems.

TABLE 1. Performances comparison of RELM and ERELM in the identification of (4).

Structure	n_h	Parameters	Train MSE	Test MSE
RELM	7	28	0.0045	0.00013
RELM	8	32	0.0034	0.00003
ERELM	7	28	0.0041	0.00003
ERELM	8	32	0.0031	0.00001

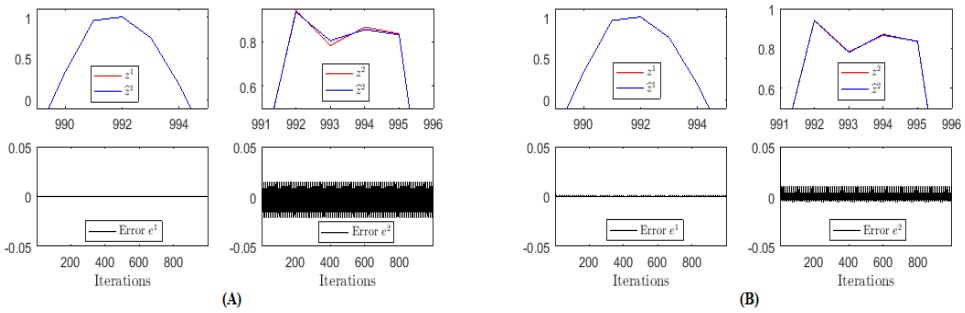


FIGURE 2. The true states z^1 and z^2 of (4), the estimated states $\widehat{z}^1$ and $\widehat{z}^2$, and the errors e^1 and e^2 in testing of RELM (part (A)) and ERELM (part (B)) with seven hidden rough neurons.

References

1. G. Ahmadi, *Stable rough extreme learning machines for the identification of uncertain continuous-time nonlinear systems*, Cont. Optim. Appl. Math. **4** (1) (2019) 83–101.
2. G. Ahmadi and M. Teshnehlab, *Designing and implementation of stable sinusoidal rough-neural identifier*, IEEE Trans. Neural Netw. Learn. Syst. **28** (8) (2017) 1774–1786.
3. G. Ahmadi and M. Teshnehlab, *Identification of multiple input-multiple output non-linear system cement rotary kiln using stochastic gradient-based rough-neural network*, J. AI Data Min. **8** (3) (2020) 417–425.
4. G. Ahmadi, M. Teshnehlab and F. Soltanian, *A higher order online Lyapunov-based emotional learning for rough-neural identifiers*, Cont. Optim. Appl. Math. **3** (1) (2018) 87–108.
5. G. B. Huang, Q. Y. Zhu and C. K. Siew, *Extreme learning machine: Theory and applications*, Neurocomputing **70** (1-3) (2006) 489–501.
6. P. Lingras, *Rough neural networks*, In Proc. of the 6th Int. Conf. on Information Processing and Management of Uncertainty in Knowledgebased Systems, Granada, (1996) pp. 1445–1450.
7. C. Lucas, A. Abbaspour, A. Gholipour, B. N. Arabi and M. Fatourechi, *Enhancing the performance of neurofuzzy predictors by emotional learning algorithm*, Informatica **27** (2) (2003) 137–146.

E-mail: g.ahmadi@pnu.ac.ir

Stability and Dynamic of the HIV Model with Logistic Growth, Treatment, Cure Rate and Cell-to-Cell Transmission

Najmeh Akbari*
Department of Mathematical Sciences, Isfahan University of Technology, Isfahan, Iran
and Rasoul Asheghi
Department of Mathematical Sciences, Isfahan University of Technology, Isfahan, Iran

Abstract. In this work, we propose a five-dimensional Ordinary Differential Equation model with logistic growth, cell-to-cell and virus-to-cell transmission rates, cellular and humoral immune responses, rate of cure, and two treatments. Then we examine the dynamic behavior of the system to investigate therapeutic effects on disease control.
Keywords: Logistic growth, Treatment rate, Cure rate, Cell-to-cell transmission, Dynamic.
AMS Mathematical Subject Classification [2010]: 34D20, 34C11, 34M10.

1. Introduction

Mathematical modeling is the attempt to present a mathematical model for a system that applies not only to the natural sciences such as physics, biology, geology, meteorology, engineering sciences, computer science, and artificial intelligence, but also to the social sciences such as economics, psychology, sociology, and medicine. Mathematical modeling helps researchers that analyze a system and predict its behavior.

The process of describing a system (e.g., disease spread) requires assumptions, access to data to estimate values of the model parameters, quantitative or qualitative predictions, and comparison of results with observational or experimental data. So the crucial role of mathematical models is to help understand a system. In recent decades, several intracellular dynamic models have been defined for the HIV-1 virus. These models describe the reaction between the virus and the host cells in diseased individuals and are valuable for understanding the dynamics of viral infections and the effectiveness of viral therapy [8].

In 2011, Sigal et al. stated that cell-to-cell expansion of HIV-1 reduces the efficacy of antiviral treatment because cell-to-cell transmission can cause many infections in target cells, which can reduce the sensitivity to antiviral drugs [6]. In 2011, Xu states the model in which the saturation collision rate is used instead of the linear collision rate for virus and cell contact [9]. In 2012, Yan and Wang described a model that included both cell-mediated and humoral immune responses and involved only the process of virus-to-cell infection [10]. In 2013, Wang et al. presented an HIV model that included CTL immune response and antiviral

*Speaker

treatment. In this model, CD_4^+T cell proliferation in the presence of the virus is expressed as a logistic function [7]. At 2015, Foutz et al. Stated the principle of HIV vaccine design based on the combined efficacy of cellular and humoral immunity [2]. In 2016, Kamboj announced a model of Reverse Transcriptase and Protease inhibitors drug therapy with the proliferation rate of logistic [3].

In 2017, Alawi et al. replaced the saturation function $\frac{\beta_1 v(t)}{1+\alpha v(t)}$ instead of the bilinear infection rate [1]. In 2018, Lin et al. proposed an HIV-1 model with virus-to-cell infection, cell-to-cell infection, cellular and humoral immune response, and saturation incidence rate of the virus are considered [5]. In 2016, Kaminski stated a method for curing infected cells by gene therapy or loss of all cccDNA from their nucleus [4].

In this paper, using many of the mathematical models presented in HIV, we develop the model of [5], regardless of the delay parameter, by replacing a logistic function for CD_4^+ healthy cells proliferation, two treatment rates to reduce infected cells and virus proliferation, and rate of cure to recover infected cells to healthy cells. Then we analyze the stability and treatment effects on it.

2. Model Formulation

Now, we extend a mathematical model for HIV infection with two treatment rates, cure rate, the transmission of infection by the virus-to-cell and cell-to-cell, logistic growth for CD_4^+ T-cell uninfected, the saturation function for the infection rate, and both types of cellular and humoral immune systems. We use five state variables in the model. Population of uninfected CD_4^+ T-cells (x), Population of infected CD_4^+ T-cells (y), Population of infectious HIV virions (v), Population of T-cells (z), Population of B-cells (w). Also, two parameters η and ε have been introduced as treatment rates of Reverse Transcriptase Inhibitors ($RTIs$) and Protease Inhibitors (PIs), respectively. Reverse Transcriptase Inhibitor prevents the transcriptase process in cells infected by the virus HIV, and Protease Inhibitor blocks the protease enzyme, thereby preventing the production of infectious and mature viruses. The proposed model is illustrated below.

$$
\begin{aligned}
\frac{dx}{dt} &= rx\left(1-\frac{x+y}{m}\right)-\frac{(1-\eta)\,\beta_1 vx}{1+\alpha v}-\beta_2 xy+\rho y-dx,\\
\frac{dy}{dt} &= \frac{(1-\eta)\,\beta_1 vx}{1+\alpha v}+\beta_2 xy-(\delta+\rho)\,y-\rho_1 yz,\\
\frac{dv}{dt} &= (1-\varepsilon)\,n\delta y-\mu v-\rho_2 vw,\\
\frac{dz}{dt} &= c_1 yz-b_1 z,\\
\frac{dw}{dt} &= c_2 vw-b_2 w.
\end{aligned}
\tag{1}
$$

All parameters in model (1) are positive and assumed to be independent of time.

3. Main Results

Proposition 3.1. *Let* $\Gamma(t)=(x(t),\,y(t),\,v(t),\,z(t),\,w(t))$, *with* $x(0)\geq 0$, $y(0)\geq 0$, $v(0)\geq 0$, $z(0)\geq 0$, $w(0)\geq 0$, *be a solution of the system* (1). *Then* $0\leq x(t)\leq$

$M, 0 \leq y(t) \leq M, 0 \leq v(t) \leq M, 0 \leq z(t) \leq M, 0 \leq w(t) \leq M$ *for all* $t \geq 0$, *for some* $M > 0$.

System (1) has two infection-free equilibrium points $E_{00} = (0, 0, 0, 0, 0)$ and $E_{01} = (\frac{m(r-d)}{r}, 0, 0, 0, 0)$.

PROPOSITION 3.2. *If* $\mathcal{R}_0 = \frac{r}{d} < 1$, *then the infection-free equilibrium point* $E_{00} = (0, 0, 0, 0, 0)$ *of system* (1) *is asympotically stable and it is unstable when* $\mathcal{R}_0 > 1$.

To simplify the calculations, we set $A = \frac{\mu(\delta+\rho)}{\delta 1}$ and $B = \frac{\mu\beta_2}{\delta_1}$.

THEOREM 3.3. *The disease-free equilibrium point* E_{01} *is asymptotically stable when* $0 < \mathcal{R}_1 = \frac{m(r-d)(\eta_1+B)}{rA} < 1$ *and it is unstable when* $\mathcal{R}_1 > 1$.

Then, we show that for $\mathcal{R}_1 > 1$, system (1) has four equilibrium points. It has equilibrium points $E_1 = (x_1, y_1, v_1, 0, 0)$, $E_2 = (x_2, y_2, v_2, z_2, 0)$, $E_3 = (x_3, y_3, v_3, 0, w_3)$ and $E_4 = (x_4, y_4, v_4, z_4, w_4)$.

THEOREM 3.4. *The following holds:*

i) *If* $\mathcal{R}_1 < 1$, *then the equilibrium point* E_1 *does not exist.*
ii) *If* $\mathcal{R}_1 = 1$, *then the equilibrium point* $E_1 = E_{01}$.
iii) *If* $\mathcal{R}_1 > 1$, *then the equilibrium point* E_1 *exists and it is locally asymptotically stable for* $v_1 < \min\left\{\frac{b_1\delta_1}{c_1\mu}, \frac{b_2}{c_2}\right\}$, *and it is unstable for* $v_1 > \frac{b_1\delta_1}{c_1\mu}$ *or* $v_1 > \frac{b_2}{c_2}$.

We set $A_2 = \frac{\delta_1}{\mu}(\eta_1 + B\alpha_2)$ and $\alpha_2 = 1 + \alpha v_2$.

THEOREM 3.5. *The following holds:*

i) *If* $\mathcal{R}_1 \leq 1$, *then the equilibrium point* E_2 *does not exist.*
ii) *If* $\mathcal{R}_1 > 1$ *and* $T_0 = \alpha_2(\delta+\rho)^2(1-\mathcal{R}_1) + \frac{A\delta_1\alpha_2 b_1}{c_1\mu}(\frac{A\alpha\delta_1^2}{\mu^2} + A_2) + \frac{\delta m b_1}{rc_1}A_2^2 < 0$, *then the equilibrium point* E_2 *exists and it is locally asymptotically stable for* $v_2 < \frac{b_2}{c_2}$.

Let $N = \frac{\rho_2}{\mu}$, $A_3 = \frac{\mu b_2}{c_2\delta_1}$ and $\alpha_3 = 1 + \alpha v_3$.

THEOREM 3.6. *The following statements are satisfied:*

i) *If* $\mathcal{R}_0 \leq 1$, *then the equilibrium point* E_3 *does not exist.*
ii) *If* $\mathcal{R}_0 > 1$ *and*

$$Q_0 = rA^2\alpha_3^2 - mA\alpha_3(r-d)(B\alpha_3 + \eta_1) + rA\alpha_3 A_1(B\alpha_3 + \eta_1) + m\delta A_3(B\alpha_3 + \eta_1)^2 < 0,$$

then the equilibrium point E_3 *exists and it is locally asymptotically stable when* $y_3 < \frac{b_1}{c_1}$.

For $\rho_2 = N\mu$, $N_1 = \frac{\delta_1 b_1 c_2}{\mu c_1 b_2}$, $N_2 = \frac{N_1\mu\rho_1}{\delta_1}$ and $\alpha_4 = 1 + \alpha v_4$.

THEOREM 3.7. *The following statements are satisfied:*

i) *If* $\mathcal{R}_0 \leq 1$, *then the equilibrium point* E_4 *does not exist.*
ii) *If* $\mathcal{R}_0 > 1$ *and*

$$\Psi_0 = rN_1^2A^2\alpha_4^2 - N_1A\alpha_4m(r-d)(BN_1\alpha_4+\eta_1) + ry_3N_1A\alpha_4(BN_1\alpha_4+\eta_1) < 0,$$

then the endemic equilibrium point E_4 *exists and it is locally asymptotically stable.*

4. Numerical Simulation

Since treatment plays an essential role in the control of AIDS. We examined its effects on disease progression. In this section, we examine the theoretical results of model (1) by numerical simulations. First, we illustrate the stability of equilibrium points for the different values of $\mathcal{R}_0$ and $\mathcal{R}_1$.

In Figure 1, by using from the Column 1 of Table 1, we have $\mathcal{R}_0 = 0.41 < 1$, then $E_0 = (0, 0, 0, 0, 0)$ or disease-free equilibrium point is stable. It shows that the disease will disappear, and only the treatment will reduce the virus load.

In Figure 2, from the Column 2 of Table 1, we obtain $\mathcal{R}_0 = 100 > 1$ and $\mathcal{R}_1 = 0.05122764706 < 1$, then $E_{01} = (990, 0, 0, 0, 0)$ is asymptotically stable. This leads to the saturation of the population of uninfected cells and eventually eliminate the disease.

In Figure 3, by replacing values the Column 3 of Table 1 in system (1), we obtain $\mathcal{R}_1 = 2.026297059 > 1$ that despite the condition $\psi_0 = -0.0002085467695 < 0$, we have $E_1 = (14841.86236, 6.496271350, 1039.403416, 0, 0)$ is asymptotically stable. In other words, by increasing the proliferation rate of the uninfected cells population and decreasing the level of treatment relative to the Column 2, the values of infected cells and the virus in Table 1 have decreased such that they have converged to y_1 and v_1. Still, the levels of cellular and humoral cells will converge to zero.

In Figure 4, from the Column 4 of Table 1, we calculate $\mathcal{R}_1 = 4.703470589 > 1$ that existence of the condition $T_0 = -1.464981642 < 0$ shows that there is

$$E_2 = (98984.77489, 9.523809524, 1047.619048, 685.1145481, 0),$$

and it is asymptotically stable. Which means that treatment reduces the level of proliferation of viral and infectious cells, also prevents excessive cellular immunity.

In Figure 5, from the Column 5 of Table 1 and put parameters in system (1), we get $\mathcal{R}_1 = 2.351735294 > 1$ and $Q_0 = -0.00005940827650 < 0$ that it can be concluded

$$E_3 = (49497.49885, 1.996874149, 100, 0, 3589.684691),$$

and it is asymptotically stable. By observing Figure 5, it can be seen that stimulation of the humoral immune system and timely treatment can significantly reduce virus replication and the number of infected cells.

In Figure 6, from the Column 6 of Table 1, we find $\mathcal{R}_1 = 20.39982353 > 1$ and $\Psi_0 = -0.001109190902 < 0$, therefore

$$E_4 = (296974, 9.523809524, 1000, 2997.258841, 1571.428572),$$

and it is asymptotically stable.

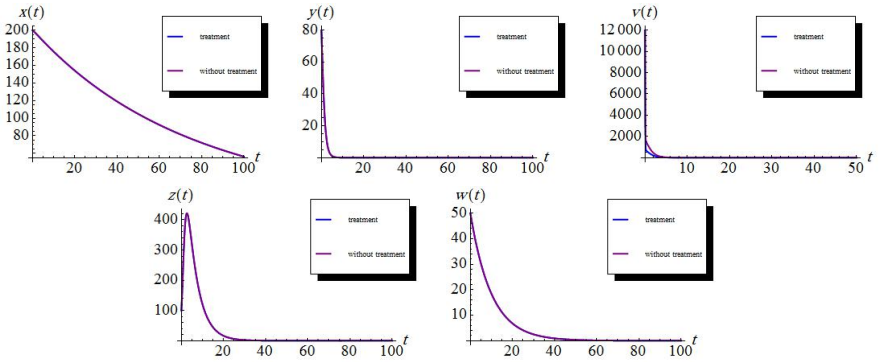


FIGURE 1. Dynamic behavior of the solutions $x(t)$, $y(t)$, $v(t)$, $z(t)$ and $w(t)$ of system (1) with treatment and without treatment for $\mathcal{R}_0 = 0.41 < 1$ at free-equilibrium point of $E_0 = (0, 0, 0, 0, 0)$.

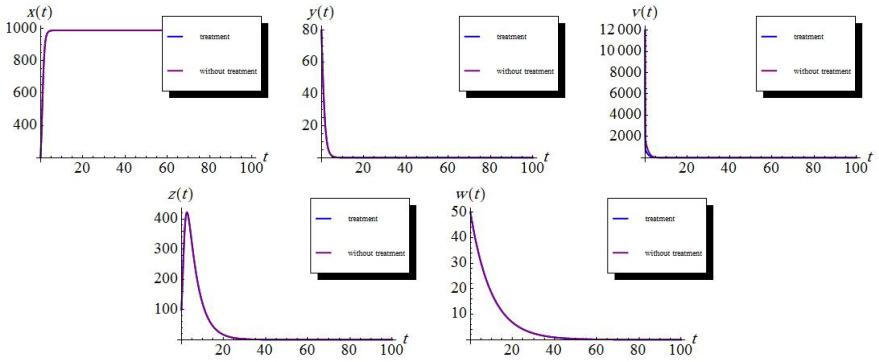


FIGURE 2. Dynamic behavior of the solutions $x(t)$, $y(t)$, $v(t)$, $z(t)$ and $w(t)$ of system (1) with treatment and without treatment at time t for $\mathcal{R}_0 = 100 > 1$ and $\mathcal{R}_1 = 0.05122764706 < 1$ at $E_{01} = (990, 0, 0, 0, 0)$.

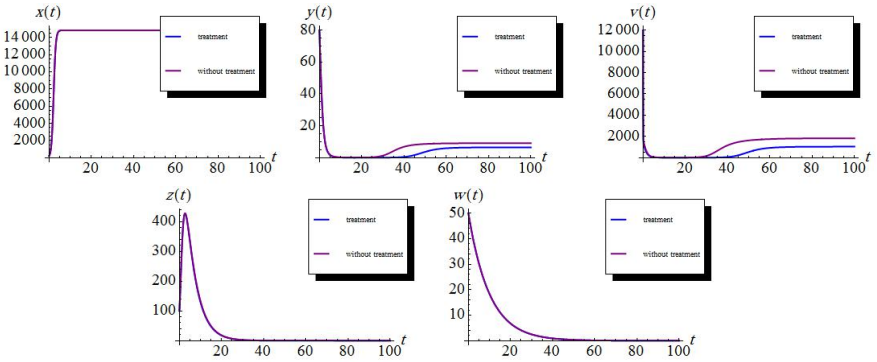


FIGURE 3. Dynamic behavior of the solutions $x(t)$, $y(t)$, $v(t)$, $z(t)$ and $w(t)$ of system (1) with treatment and without treatment at time t for $\mathcal{R}_1 = 2.026297059 > 1$ and $\psi_0 = -0.0002085467695 < 0$ at $E_1 = (14841.86236, 6.496271350, 1039.403416, 0, 0)$.

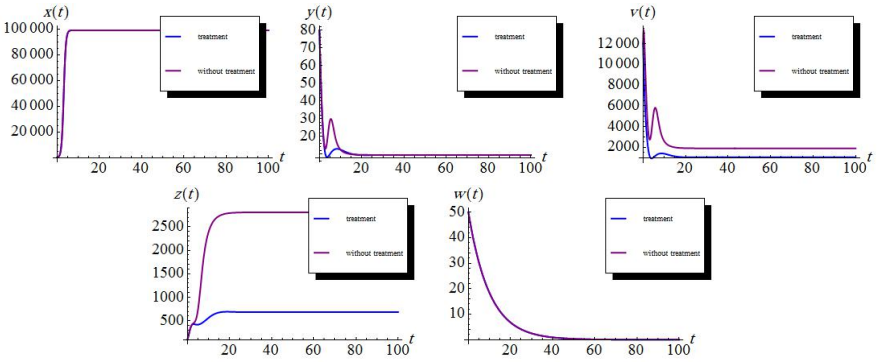


FIGURE 4. Dynamic behavior of the solutions $x(t)$, $y(t)$, $v(t)$, $z(t)$ and $w(t)$ of system (1) with treatment and without treatment at time t for $\mathcal{R}_1 = 4.703470589 > 1$ and $T_0 = -1.464981642 < 0$ at $E_2 = (98984.77489, 9.523809524, 1047.619048, 685.1145481, 0)$.

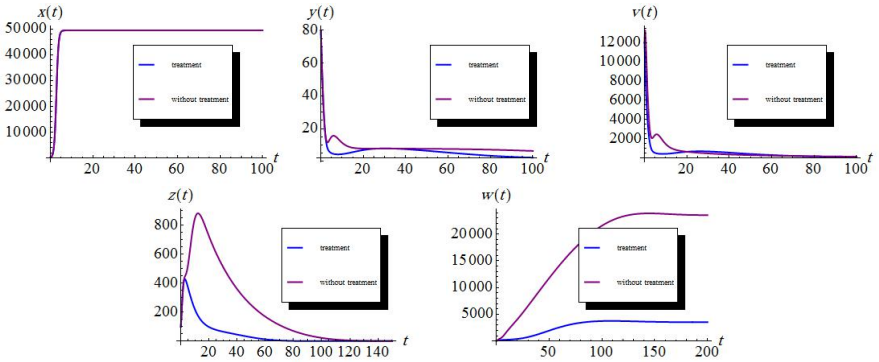


FIGURE 5. Dynamic behavior of the solutions $x(t)$, $y(t)$, $v(t)$, $z(t)$ and $w(t)$ of system (1) with treatment and without treatment at time t for $\mathcal{R}_1 = 2.351735294 > 1$ and $Q_0 = -0.00005940827650 < 0$ at $E_3 = (49497.49885, 1.996874149, 100, 0, 3589.684692)$.

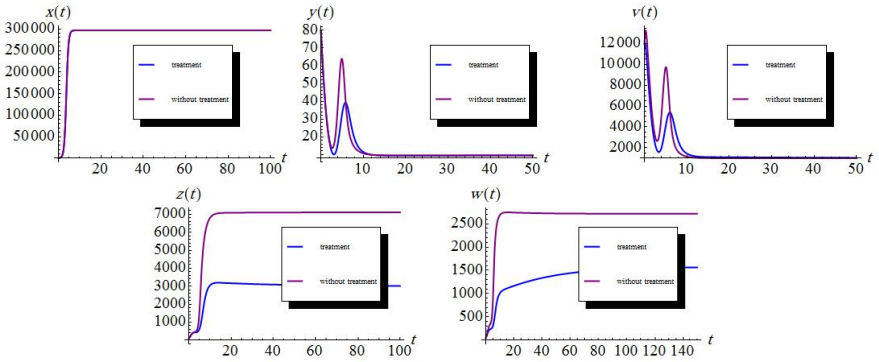


FIGURE 6. Dynamic behavior of the solutions $x(t)$, $y(t)$, $v(t)$, $z(t)$ and $w(t)$ of system (1) with treatment and without treatment at time t for $\mathcal{R}_1 = 20.39982353 > 1$ and $\Psi_0 = -0.001109190902 < 0$ at $E_4 = (296974, 9.523809524, 1000, 2997.258841, 1571.428572)$.

TABLE 1. Values of parameters in HIV mathematical model.

Parameters	unit	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6
r	day^{-1}	0.0082	2	2	2	2	2
m		1000	1000	15000	100000	5×10^4	300000
β_1	$ml\,.\,(virion\,.\,day)^{-1}$	4.8×10^{-7}	4.8×10^{-7}	4.8×10^{-7}	4.8×10^{-7}	4.8×10^{-7}	4.8×10^{-7}
β_2	$ml\,.\,(virion\,.\,day)^{-1}$	4.7×10^{-7}	4.7×10^{-7}	4.7×10^{-7}	4.7×10^{-7}	4.7×10^{-7}	4.7×10^{-7}
α	$cells^{-1}\,.\,ml$	0.001	0.001	0.001	0.001	0.001	0.001
ρ	day^{-1}	0.01	0.01	0.01	0.01	0.01	0.01
d	day^{-1}	0.02	0.02	0.02	0.02	0.02	0.02
δ	day^{-1}	0.5	0.5	0.5	0.5	0.5	0.5
ρ_1	$ml\,.\,(cells\,.\,day)^{-1}$	0.001	0.001	0.001	0.001	0.001	0.001
η		0.4	0.4	0.1	0.55	0.55	0.55
ε		0.55	0.55	0.2	0.45	0.45	0.2
n	$ml\,.\,virion$	1200	1200	1200	1200	1200	1200
μ	day^{-1}	3	3	3	3	3	3
ρ_2	$ml\,.\,(virion\,.\,day)^{-1}$	0.5	0.5	0.5	0.001	0.001	0.001
c_1	$ml\,.\,(cells\,.\,day)^{-1}$	0.021	0.021	0.021	0.021	0.021	0.021
b_1	day^{-1}	0.2	0.2	0.2	0.2	0.2	0.2
c_2	$ml\,.\,(virion\,.\,day)^{-1}$	10^{-11}	10^{-11}	10^{-11}	10^{-11}	10^{-4}	10^{-4}
b_2	day^{-1}	0.1	0.1	0.1	0.1	0.01	0.1

References

1. A. M. Elaiw, A. A. Raezah and K. Hattaf, *Stability of HIV-1 infection with saturated virus-target and infected-target incidences and CTL immune response*, Int. J. Biomath. **10** (5) (2017). DOI:10.1142/S179352451750070X
2. T. R. Fouts, K. Bagley, I. J. Prado, K. L. Bobb, J. A. Schwartz, R. Xu, R. J. Zagursky, M. A. Egan, J. H. Eldridge, C. C. LaBranche and D. C. Montefiori, *Balance of cellular and humoral immunity determines the level of protection by HIV vaccines in rhesus macaque models of HIV infection*, Proc. Natl. Acad. Sci. USA. **112** (9) (2015) pp. E992–E999.
3. D. Kamboj and M. D. Sharma, *Effects of combined drug therapy on HIV-1 infection dynamics*, Int. J. Biomath. **9** (5) (2016). DOI:10.1142/S1793524516500650
4. R. Kaminski, R. Bella, C. Yin, J. Otte, P. Ferrante, H. E. Gendelman, H. Li, R. Booze, J. Gordon, W. Hu and K. Khalili, *Excision of HIV-1 DNA by gene editing: a proof-of-concept in vivo study*, Gene Ther. **23** (8) (2016) 690–695.
5. J. Lin, R. Xu and X. Tian, *Threshold dynamics of an HIV-1 model with both viral and cellular infections, cell-mediated and humoral immune responses*, Math. Biosci. Eng. **16** (1) (2018) 292–319.
6. A. Sigal, J. T. Kim, A. B. Balazs, E. Dekel, A. Mayo, R. Milo and D. Baltimore, *Cell-to-cell spread of HIV permits ongoing replication despite antiretroviral therapy*, Nature **477** (7362) (2011) 95–98.
7. Y. Wang, Y. Zhou, F. Brauer and J. M. Heffernan, *Viral dynamics model with CTL immune response incorporating antiretroviral therapy*, J. Math. Biol. **67** (4) (2013) 901–934.
8. H. Wu and A. A. Ding, *Population HIV-1 dynamics in vivo: Applicable models and inferential tools for virological data from AIDS clinical trials*, Biometrics **55** (2) (1999) 410–418.
9. R. Xu, *Global stability of an HIV-1 infection model with saturation infection and intracellular delay*, J. Math. Anal. Appl. **375** (1) (2011) 75–81.
10. Y. Yan and W. Wang, *Global stability of a five-dimensional model with immune responses and delay*, Discrete Contin. Dyn. Syst. Ser. B **17** (1) (2012) 401–416.

E-mail: najmeh.akbari@math.iut.ac.ir
E-mail: r.asheghi@iut.ac.ir

Global Existence, Asymptotic Stability and Blow-up for Nonlinear Kirchhoff Type Equation with Damping and Coriolis Term

Hajar Ansari*
Faculty of Mathematical Sciences, Amirkabir University of Technology, Tehran, Iran
and Mahmoud Hesaaraki
Department of Mathematics, Sharif University of Technology, Tehran, Iran

Abstract. In this paper, we study the initial-boundary value problem for a nonlinear Kirchhoff type equation with Coriolis force term and damping in a bounded domain with smooth boundary. For this problem, we show that the global existence and uniqueness of solution via potential well theory and Faedo-Galerkin method. Also, we consider the asymptotic behavior of solutions. Making use of integral inequalities, multiplier technique and Lyapanov function, we establish polynomial decay and exponential decay of solution, respectively. In two different methods, we show that the energy function grows-up as exponential function when $t \to +\infty$. The first method based on a method used in Vitillaro(Arch Ration Mech Anal 149:155-182, 1999). The second method based on some energy estimates. The result of the second method seems to be much more better than the result of first one. Moreover, the blow-up of solutions are established for arbitrary initial energy by using modified concavity method.
Keywords: Kirchhoff type wave equation, Blow-up, Exponential decay, Polynomial decay.
AMS Mathematical Subject Classification [2010]: 35J60, 35J47, 35J25.

1. Introduction

In this paper, we consider the following Kirchhoff type wave equation with damping and Coriolis force term,

$$(1)\quad \begin{cases} u_{tt} - M(\|\nabla u\|^2)\Delta u - \lambda\Delta u_t + \delta u_t + \eta div(u_t) = \mu|u|^{p-1}u, & in\ \ \Omega_T, \\ u(x,0) = u_0(x),\ \ u_t(x,0) = u_1(x), & in\ \Omega, \\ u(x,t) = 0, & in\ \Gamma. \end{cases}$$

Here $\Omega \subset \mathbb{R}^N$, $N \geq 1$, is a bounded domain with smooth boundary $\partial\Omega$, T is a positive constant or $T = \infty$, $\Omega_T = \Omega \times (0,T)$, $\Gamma = \partial\Omega \times (0,T)$, $M(\|\nabla u\|^2) = m_0 + \alpha\|\nabla u(t)\|^{2\gamma}$ with $m_0 \geq 0$, $\alpha \geq 0$, $m_0 + \alpha > 0$. While $\gamma > 0$, $\lambda, \delta, \eta > 0$ and $p > 1$ are constants. Moreover, $u(x,t)$, the transversal displacement of the strip at spatial coordinate x and time t, $-\lambda\Delta u_t$ and δu_t are called a strong damping term and a weak one, respectively. Finally $\eta div(u_t)$ is Coriolis force term.

Problem (1) without damping term and Coriolis force term i.e. $\delta = \lambda = \eta = 0$ was firstly introduced by Kirchhoff [6] in 1883 as a model for small transversal

*Speaker

vibrations of an elastic string with fixed endpoints. In fact, Kirchhoff equation is an extension of the classical D'Alembert wave equation which considers the effects of changes in length of the string during vibrations. Problem (1) models several physical and biological systems, where u describes a process which depends on average of itself, as for example population density. For details, we refer the reader to [1] and the references therein for related work. Problem (1) received much attention only after Lions paper which proposed an abstract frame work to the problem.

Ono [8] considered problem (1) without Coriolis force term and weak damping. He showed that existence of a global solution by using Banach contraction mapping theorem and the decay of energy based on the method of Nakao, when the initial energy is non-negative and small. He also proved that the local solutions blow-up at a finite time with non-positive initial energy via concavity method. Moreover, he used potential well theory and concavity method to show that the blow-up of solutions with positive initial energy. Santos et al. [9] investigated

$$u_{tt} - M(\|\nabla u\|^2)\Delta u - \Delta u_t + f(u) = 0,$$

with memory condition at boundary on a bounded domain Ω. They proved global existence of solution to this problem by using Faedo-Galerkin method. They also established the energy decay exponentially and polynomially. Gazzola and Squassina [3] considered the following equation

$$u_{tt} - \Delta u - \omega\Delta u_t + \mu u_t = |u|^{p-2}u, \tag{2}$$

with Dirichlet boundary condition on a bounded domain Ω. They proved global existence of solutions with suitable initial data. They also showed blow-up of solutions with high energy initial data. Gerbi and Houari [4] investigated (2) without weak damping i.e. $\mu = 0$ under dynamic boundary conditions on a bounded domain Ω. They established local existence of solution by using Faedo-Galerkin method combined with a contraction mapping theorem. They also proved the exponential growth of the energy. Bilgin and Kalantarov [2] considered the following initial boundary value problem

$$u_{tt} - \nabla\big((a_0 + a|\nabla u|^{m-2})\nabla u\big) - b\Delta u_t = g(x,t,u,\nabla u) + |u|^{p-2}u,$$

with Dirichlet boundary condition on a bounded domain Ω. They also gave some sufficient conditions on initial data for which blow up occurs in a finite time.

Kim et al. [5] considered

$$u_{tt} - M(x,t,\|\nabla u\|^2)\Delta u + \rho(x,t,u_t,\nabla u,\nabla u_t) = 0,$$

with boundary feedback control on a bounded domain, $\Omega \subset \mathbb{R}^N$, with smooth boundary. They proved existence and uniqueness of strong solution via techniques of functional analysis, mainly a theorem of compactness for the analysis of approximation of the Faedo-Galerkin method and estimate a decay rate for the energy. Yang and Da [10] investigated longtime dynamics of the Kirchhoff wave equation with strong damping and critical nonlinearities

$$u_{tt} - (1 + \epsilon\|\nabla u\|^2)\Delta u + \Delta u_t + h(u_t) + g(u) = f(x),$$

with $\epsilon \in [0,1]$. The well-posedness and the existence of global and exponential attractors were established, and the stability of the attractors on the perturbation parameter ϵ was proved for the IBVP of the equation provided that both nonlinearities $h(s)$ and $g(s)$ are of critical growth.

In this paper, we consider problem (1) and we prove global existence and uniqueness of solution via Galerkin method and potential well theory. We also show exponential decay and polynomial decay of solutions by using Lyapanov function, integral inequalities and multiplier technique. Besides,we establish that the energy function grows-up as exponential function when $t \to +\infty$, by using two different methods. One of the methods is the method used in Vitillaro. The second method based on energy estimates. Moreover blow-up of solutions are proved under some conditions on initial data and the coefficients for all initial energy making use of modified concavity method.

2. Main Results

We study the global existence of the solution to problem (1). In order to do this, we first define some notations.

We use standard Lebesgue space $L^p(\Omega)$ with usual norm

$$\|u\|_p = \Big(\int_\Omega |u|^p dx \Big)^{\frac{1}{p}}.$$

The norm and scalar product in $L^2(\Omega)$ is denoted by $\|.\|$ and $< \cdot, \cdot >$ respectively. We use Sobolev spaces $H_0^1(\Omega)$ and $H^2(\Omega)$ with usual norms. We denote the norm of a Banach space X by $\|\cdot\|_X$. We denote by $L^p(0,T;X)$, $1 \le p \le \infty$, the Banach space of the real functions $u : (0,T) \to X$ measurable, such that

$$\|u\|_{L^p(0,T;X)} = \Big(\int_0^T \|u(t)\|_X^p dt \Big)^{\frac{1}{p}}, \quad for \ \ 1 \le p < \infty,$$

and

$$\|u\|_{L^\infty(0,T;X)} = \inf \sup_{0<t<T} \|u(t)\|_X, \quad for \ \ p = \infty.$$

We define the energy function as

$$E(t) = \frac{1}{2}\|u_t\|^2 + \frac{m_0}{2}\|\nabla u\|^2 + \frac{\alpha}{2\gamma+2}\|\nabla u\|^{2\gamma+2} - \frac{\mu}{p+1}\|u\|_{p+1}^{p+1}. \tag{3}$$

Definition 2.1. A weak solution to problem (1) is a function $u(x,t)$ such that

i) $u \in L^2(0,T;H_0^1(\Omega) \cap H^2(\Omega))$, $u_t \in L^2(0,T;L^2(\Omega) \cap H_0^1(\Omega))$ and $u_{tt} \in L^2(0,T;L^2(\Omega))$,

ii) for all $v \in C_0^\infty([0,T] \times \Omega)$ satisfies the generalized formula

$$\begin{aligned} \int_0^T \big\langle u_{tt}(\tau), v(\tau) \big\rangle d\tau + \int_0^T \big(m_0 + \alpha\|\nabla u\|^{2\gamma}(\tau)\big)\big\langle \nabla u, \nabla v \big\rangle d\tau \\ + \lambda \int_0^T \big\langle \nabla u_t(\tau), \nabla v(\tau) \big\rangle d\tau + \delta \int_0^T \big\langle u_t(\tau), v(\tau) \big\rangle d\tau \\ + \eta \big\langle div(u_t(\tau)), v(\tau) \big\rangle - \mu \int_0^T \big\langle |u(\tau)|^{p-1}u(\tau), v(\tau) \big\rangle d\tau = 0, \end{aligned}$$

iii) satisfies the initial conditions, i.e.,

$$\begin{cases} u(x,0) = u_0(x), \ u_0 \in H_0^1(\Omega) \cap H^2(\Omega), \\ u_t(x,0) = u_1(x), \quad u_1 \in H_0^1(\Omega) \cap H^2(\Omega). \end{cases}$$

Now, we state our first main results.

THEOREM 2.2. *Assume that* $\delta \geq \frac{\eta}{2}$, $\lambda \geq \frac{\eta}{2}$ *and*

$$p > 1 \quad for \ n = 1, 2 \quad or \quad 1 < p < \frac{n}{n-2} \quad for \ n \geq 3, \tag{4}$$

and

$$u_0 \in H_0^1(\Omega) \cap H^2(\Omega) \ and \ u_1 \in H_0^1(\Omega) \cap H^2(\Omega). \tag{5}$$

Then problem (1) *has a unique weak global solution,* $u(t)$ *such that*

$$\begin{cases} u \in C\big([0,T], H_0^1(\Omega) \cap H^2(\Omega)\big) \cap C^1\big([0,T], L^2(\Omega)\big), \\ u_t \in C\big([0,T], H_0^1(\Omega)\big) \cap L^2\big(\Omega \times (0,T)\big). \end{cases}$$

We show that the energy function decays exponentially via constructing of a Lyapunov function by performing a suitable modification of the energy function.

THEOREM 2.3. (Exponential decay) *Suppose the assumptions in Theorem* 2.2 *hold. Then the solution* u *to problem* (1) *satisfies the following energy decay estimates*

$$E(t) \leq (\epsilon_1 - \kappa\epsilon_2)^{-1}(\epsilon_1 + \kappa\epsilon_2)^{-1} L(0) \exp\big(-\sigma(\epsilon_1 + \kappa\epsilon_2)^{-1} t\big), \quad t \geq 0,$$

where $L(t) = \epsilon_1 E(0) + \epsilon_2 \int_\Omega u_0 u_1 dx + \frac{\epsilon_2 \lambda}{2} \|\nabla u_0\|^2$ *and* σ, κ, ϵ_1, ϵ_2 *are positive constants which will be determined in the proof.*

In the sequel, we establish polynomial decay estimates for energy function by using some integral inequalities and multiplier techniques.

THEOREM 2.4. (Polynomial decay) *Suppose the assumptions in Theorem* 2.2 *hold. Then the solution* u *to problem* (1) *satisfies the following energy decay estimates*

$$E(t) \leq E(0) \left(\frac{(S_0 + \Gamma_3)(1+\theta)}{\theta t + S_0 + \Gamma_3} \right)^{\frac{1}{\theta}},$$

where S_0, Γ_3 *and* θ *are determined later in the proof.*

Now, we established an exponential growth result for certain solutions with positive initial energy to problem (1).

THEOREM 2.5. *Assume that* $\lambda, \delta > \frac{\eta}{2}$, $p \geq 1$ *for* $n = 1, 2$ *or* $1 \leq p \leq \frac{n+2}{n-2}$ *for* $n \geq 3$ *and* $p > 2\gamma + 1$. *Let* u *be a solution to problem* (1) *with initial data satisfying* $\int_\Omega u_0 u_1 dx > 0$, $E(0) < E_1$ *and* $\|\nabla u_0\| > \beta_0$. *Then* u *grows as an exponential function, where* E_1, β_0 *are introduced in Lthe proof.*

Now, we prove the energy function, $E(t)$ grows up by using another method. In fact, we take advantage from some energy estimates. The result will be different from the one in Theorem 2.5.

THEOREM 2.6. *Assume that* $\lambda, \delta > \frac{\eta}{2}$, $p > 1$ *for* $n = 1, 2$ *or* $1 < p \leq \frac{n+2}{n-2}$ *for* $n \geq 3$ *and* $p > 2\gamma + 1$. *Let* u *be a solution to problem* (1) *with initial data satisfying* $2\int_\Omega u_0 u_1 dx + \delta\|u_0\|^2 + \lambda\|\nabla u_0\|^2 > \frac{2(p+1)}{\kappa_0} E(0)$. *Then* $E(t)$ *grows as an exponential function.*

We will prove blow-up of solutions to problem (1) with suitable initial conditions and arbitrary initial energy. We will use the concavity argument developed by Levine [7].

THEOREM 2.7. *Let* $\delta, \lambda > \frac{p+1}{2(p-1)-\sigma}\eta$, $0 < \eta < (p-1)\sqrt{m_0}$ *and* $u_0, u_1 \in H^2(\Omega) \cap H_0^1(\Omega)$, *where* σ *will be introduced in the proof. Then the solution* $u(x,t)$ *with arbitrary initial energy blows up in finite time provided that the initial conditions satisfy* $\int_\Omega u_0 u_1 dx > 0$.

Acknowledgement

Acknowledgements could be placed at the end of the text but before the references.

References

1. C. O. Alves and F. J. S. A. Correa, *On existence of solutions for a class of problem involving a nonlinear operator*, Comm. Appl. Nonlinear Anal. **8** (2) (2001) 43–56.
2. B. A. Bilgin and V. K. Kalantarov, *Blow up of solutions to the initial boundary value problem for quasilinear strongly damped wave equations*, J. Math. Anal. Appl. **403** (1) (2013) 89–94.
3. F. Gazzola and M. Squassina, *Global solutions and finite time blow up for damped semilinear wave equations*, Ann. Inst. H. Poincaré Anal. Non Linéaire **23** (2) (2006) 185–207.
4. S. Gerbi and B. Said-Houari, *Asymptotic stability and blow up for a semilinear damped wave equation with dynamic boundary conditions*, Nonlinear Anal. **74** (18) (2011) 7137–7150.
5. D. Kim, S. Kim and I. H. Jung, *Stabilization for the Kirchhoff type equation from an axially moving heterogeneous string modeling with boundary feedback control*, Nonlinear Anal. **75** (8) (2012) 3598–3617.
6. G. Kirchhoff, *Vorlesunden über Mathematiche Physik* Mechanik, Teubner, Leipzig, 1876.
7. H. A. Levine, *Instability and nonexistence of global solutions to nonlinear wave equations of the form* $Pu_{tt} = -Au + \mathfrak{F}(u)$, Trans. Amer. Math. Soc. **192** (1974) 1–21.
8. K. Ono, *On global existence, asymptotic stability and blowing up of solutions for some degenerate non-linear wave equations of Kirchhoff type with a strong dissipation*, Math. Methods Appl. Sci. **20** (2) (1997) 151–177.
9. M. L. Santos, J. Ferreira, D. C. Pereira and C. A. Raposo, *Global existence and stability for wave equation of Kirchhoff type with memory condition at the boundary*, Nonlinear Anal. **54** (5) (2003) 959–976.
10. Z. Yang and F. Da, *Stability of attractors for the Kirchhoff wave equation with strong damping and critical nonlinearities*, J. Math. Anal. Appl. **469** (1) (2019) 298–320.

E-mail: ha2578@aut.ac.ir
E-mail: hesaraki@sharif.ir

Conservation Laws by Scaling Method for the Fifth-Order Kudryashov and Sinelshchikov Equations

Mehdi Jafari
Department of Mathematics, Payame Noor University, P. O. BOX 19395-3697, Tehran, Iran
Mohammad Hadi Moslehi
Department of Mathematics, Payame Noor University, P. O. BOX 19395-3697, Tehran, Iran
and Razie Darvazeban Zade*
Department of Mathematics, Payame Noor University, P. O. BOX 19395-3697, Tehran, Iran

Abstract. In this paper, by the scaling method, we obtain new conservation laws of the fifth-order Kudryashov and Sinelshchikov equation which is generalization of the famous Kawahara equation. Scaling method applies tools from variational calculus and linear algebra and based on scaling symmetry of the PDE. We use this method to construct conservation laws of rank 3 and 5 for the fifth-order Kudryashov and Sinelshchikov equations.
Keywords: Fifth-order Kudryashov and Sinelshchikov equation, Conservation Law, Scaling symmetry.
AMS Mathematical Subject Classification [2010]: 76M60, 70S10, 35L65.

1. Introduction

Conservation laws are fundamental laws in physics. These laws state that some properties of a physical system will remain unchanged over time. Obtaining the conservation law is one of the most important applications of symmetry in physical problems. For example, scale symmetry describes a specific case of scale freeness, in which the system is completely unchanged under scaling. There are several methods for calculating the conservation laws of nonlinear partial differential equations (PDEs) (See [4]). A common method based on the relationship between the conservation laws and symmetry is expressed in Noether's theorem [1, 5, 6]. There is another method that does not use the Noether's theorem and uses scaling symmetry to construct the density and associated flux of the conservation law. This method, which is called the scaling method, is based on tools from calculus, differential geometry, calculus of variations and linear algebra. In this method, we consider a primitive density for conservation law with unspecified coefficients that are invariant under the scaling symmetry. Then, unspecified coefficients are determined by calculation and using the Euler operator [2, 7]. Finally the flux of conservation law is computed by homotopy operator to invert a divergence [8].

*Speaker

In this study, we will obtain by the scaling method, conservation laws of fifth-order Kudryashov and Sinelshchikov (K-S) equation. This equation is introduced by Kudryashov and Sinelshchikov which is generalization of the famous Kawahara equation. The fifth-order K-S equation is the following nonlinear evolution equation:

$$u_t + auu_x + bu_{3x} + cu_{4x} + du_{5x} = eu_{2x}, \tag{1}$$

where a, b, c, d and e are positive constant.

This paper is organized as follows. In Section 2, we present some definitions and result that will be used along this paper. In Section 3, by constructing the weight-balance equations for the fifth-order K-S equation the scaling symmetry of this equation is obtained. the primitive density of conservation law for this equation is constructed in Section 4. in Section 5, the actual density and the corresponding flux are constructed and some related results are obtained.

2. Notations and Definitions

In this section, we will provide the background definitions and results that will be used along this paper. Consider a system of equations in the evolutionary from,

$$u_t = P(x, u^{(M)}), \tag{2}$$

where $x = (x^1, \ldots, x^p)$ and $u = (u^1, \ldots, u^q)$ are independent space variables and dependent variables, respectively. A conservation law for (2) is in the form,

$$D_x\rho + DivJ = 0 \qquad on \qquad \Delta = 0, \tag{3}$$

where ρ is the conserved density and J is the associated flux. In (3), D_t is the total derivative whit respect to t and Div is the total Divergence. The algorithm will described in next section can be used to compute local conservation laws for systems that can be written in the evolutionary from (2).

Definition 2.1. The total derivative operator, D_x (in 1 D), acting on $f = f(x, t, u^{(M)}(x, t))$ of order n is defined as

$$D_x f = \frac{\partial f}{\partial x} + \sum_{j=1}^{N}\sum_{k=0}^{M_1^j} u^j_{(k+1)x}\frac{\partial f}{\partial u^j_{kx}},$$

where M_1^j is the order of f in component u^j and $M = \max\{M_1^1, \ldots, M_1^N\}$ [6].

Definition 2.2. Let f be a differential function of order M. In $1D$, f is called exact if f is a total derivative, i.e., there exists a differential function $F(x, u^{(M-1)}(x))$ such that $f = D_x F$ [7].

Definition 2.3. The $1D$ Euler operator for dependent variable $u^j(x)$ is defiend as

$$L_{u^j(x)} f = \sum_{k=0}^{M_1^j}(-D_x)^k \frac{\partial f}{\partial u^j_{kx}}, \qquad j = 1, \ldots, q. \tag{4}$$

Theorem 2.4. *A differential function f is exact if and only if $L_{u(x)} f = 0$* [7].

DEFINITION 2.5. Let f be an exact $1D$ differential function. The homotopy operator in $1D$ is defined as

$$H_{u(x)}f = \int_0^1 \sum_{j=1}^{N} I_{u^j(x)} f[\lambda u]\frac{d\lambda}{\lambda}, \qquad where \qquad u = (u^1, \ldots, u^q). \tag{5}$$

The integrand, $I_{u^j(x)}f$ is defined as

$$I_{u^j(x)}f = \sum_{k=1}^{M_1^j}(\sum_{i=0}^{k-1} u_{ix}^j(-D_x)^{k-(i+1)})\frac{\partial f}{\partial u_{kx}^j}, \tag{6}$$

where M_1^j is the order of f in the dependent variable u^j whit respect to x [8].

THEOREM 2.6. *Let f be exact, i.e, $D_xF = f$ for some differential function $F(x, u^{(M-1)}(x))$, then $F = D_x^{-1}f$* [8].

3. Computing the Scaling Symmetry of Fifth-Order K-S Equation

In this section, we obtain the scaling symmetry of the fifth-order K-S equation by the concept of weight. It is easy to see that fifth-order K-S equation is invariant under the scaling symmetry

$$(x, t, u, a, b, c, e) \longrightarrow (\lambda^{-1}x, \lambda^{-5}t, \lambda u, \lambda^3 a, \lambda^2 b, \lambda c, \lambda^3 e), \tag{7}$$

where λ is an arbitrary scaling parameter. We will prove this by the concept of weight [3].

DEFINITION 3.1. The weight of a variable is defined as the exponent p in the factor λ^p that multiplies the variable. For the scaling symmetry $x \longrightarrow \lambda^{-p}x$, the weight is denoted $W(x) = -p$. Total derivatives carry a weight. Indeed, if $W(x) = -p$, then $W(D_x) = p$.

DEFINITION 3.2. The rank of a monomial is the sum of the weights of the variables in the monomial. A differential function is uniform in rank if all monomials in the differential function have the same rank.
The weight-balance equations for the Kudryashov and Sinelshchikov equation are

$$\begin{aligned} W(u) + W(D_t) &= W(a) + 2W(u) + W(D_x) \\ &= W(b) + W(u) + 3W(D_x) \\ &= W(c) + W(u) + 4W(D_x) \\ &= W(d) + W(u) + 5W(D_x) \\ &= W(e) + W(u) + 2W(D_x). \end{aligned}$$

Solving the linear system gives

$$\begin{aligned} W(u) = W(D_x) = W(c) = 1\ ,\ W(d) = 0\ , \\ W(t) = 5\ ,\ W(b) = 2\ ,\ W(e) = W(a) = 3. \end{aligned}$$

Therefore, the relation (7) was proved. The conserved density and its associated flux must obey the scaling symmetry of the PDE. That is, the conservation law itself must be uniform in rank. Thus, according to the scaling symmetry of the fifth-order K-S equation, we can construct a primitive density that is a linear combination of terms of a pre-selected rank.

4. Constructing a Primitive Density

The primitive density is constructed by taking a linear combination with undetermined coefficients of terms that are invariant under the scaling symmetry of the PDE. Since the fifth-order K-S equation has t as evolution variable, we will compute the density ρ of (3) in a fixed rank, for example in rank 5. At first, consider a set P including all powers of dependent variable that have rank 5 or less,

$$\begin{aligned} P \quad &= \{u^5, b^2u, bu^3, eu^2, au^2, c^4u, cu^4, ecu, acu, u^4, bu^2, \\ &\quad c^3u, bcu, eu, , au, u^3, bu, c^2u, cu^2, u^2, cu, u\}. \end{aligned}$$

Then we utilize the total derivative operator with respect to the space variable in order to increase the terms in P up to rank 5 and put them into a new list,

$$\begin{aligned} Q \quad &= \{u^5, b^2u, bu^3, eu^2, au^2, c^4u, cu^4, ecu, acu, u^3u_x, buu_x, c^3u_x, cu^2u_x, bcu_x, \\ &\quad eu_x, au_x, uu_{x^2}, u^2u_{xx}, bu_{2x}, c^2u_{2x}, cu_x^2, cuu_x, u_xu_{xx}, uu_{3x}, u_{4x}\}. \end{aligned} \tag{8}$$

Now, we omit all terms that are divergences or divergence-equivalent to other terms in Q. Therefore, by applying the Euler operator (4) to each term in (8), we find

$$\begin{aligned} L_{u(x)}Q \quad &= \{(5u^4, 3bu^2, b^2, 3eu^2, 2au, c^4, 4cu^3, ec, ac, 0, 0, 0, 0, 0, 0, 0, \\ &\quad -u_x^2 - 2uu_{xx}, 4uu_{xx} + 2u_x^2, 0, 0, -2cu_{xx}, 0, 0, 0, 0\}. \end{aligned} \tag{9}$$

According to the Theorem 2.4, $u^3u_x, buu_x, c^3u_x, cu^2u_x, bcu_x, eu_x, au_x$, bu_{xx}, c^2u_{xx}, cuu_x, u_xu_{xx}, uu_{xxx} and u_{xxxx} are corresponding to 0 in (9). So they are divergences terms and can be removed from Q. Next, all divergence equivalent terms should be removed. For this, attach unspecified coefficients to each term in (9), then set the sum of these terms equal to zero. By gathering like terms and equating them to zero, the divergence-equivalent terms are obtained. After canceling all divergences and divergence-equivalent terms, we have new Q,

$$Q = \{u^5, bu^3, b^2u, eu^3, au^2, c^4u, cu^4, ecu, acu, uu_x^2, u^2u_{xx}, cu_x^2\}.$$

So the candidate density is,

$$\rho \quad = C_1u^5 + C_2bu^3 + C_3b^2u + C_5au^2 + C_7cu^4 + C_{17}uu_x^2 + C_{21}cu_x^2. \tag{10}$$

5. Calculating the Actual Density and Associated Flux

For determining the actual density, we have to determined the unspecified coefficient. For this, we compute the total derivative of (10) with respect to t,

$$\begin{aligned} D_t^\rho &= (5C_1u^4 + 3C_2bu^2 + C_3b^2 + 2C_5au + 4C_7cu^3 + C_{17}u_x^2)u_t \\ &\quad + (2C_{17}uu_x + 2C_{21}cu_x)u_{xt}. \end{aligned}$$

Let $E = -D_t^\rho$. Then u_t and u_{xt} have been replaced by using (1). We must have $L_{u(x)}E = 0$ by Theorem 2.4. Apply the Euler operator to E and set the result identically equal to zero, one linear system for the undetermined coefficients C_i is obtained. Solving this system we have:

$$C_1 = C_2 = C_5 = C_7 = C_{17} = C_{21} = 0, \qquad C_3 \neq 0.$$

By setting $C_3 = 1$, we have, $\rho = b^2u$.

Now, we calculate the flux of conservation law of rank 5. By the relation (3), we have $DivJ = -D_t^\rho = E$ so $J = Div^{-1}(E)$. Therefore, we must compute

$Div^{-1}(E)$. After substitution $C_3 = 1$ into E and calculating the integrand function from relation (6) and substituting it in $1D$ homotopy operator (5), J is obtained by Theorem 2.6

$$\frac{1}{2}u^2b^2a + b^2(du_{xxxx} - eu_x + bu_{xx} + cu_{xxx}).$$

So the conservation law in rank 5 for the fifth-order K-S equation is obtained.

In a similar way, additional conservation laws in rank 3 for the fifth-order K-S equation is obtained,

$$\rho = bu,$$
$$J = \tfrac{1}{2}ba + b(du_{4x} - eu_x + bu_{2x} + cu_{3x}).$$

6. Conclusion

In this paper, we consider fifth-order Kudryashov and Sinelshchikov equation that admit scaling symmetry and are uniform in rank. So, at first by the scaling method, the density of rank 5 is constructed by the Euler operator and by the homotopy operator the flux is computed. In a similar way we obtained the conservation laws for Kudryashov and Sinelshchikov equation of rank 3.

References

1. I. M. Anderson, *The Variational Bicomplex*, Dept. Math., Utah State University: Logan, Utah, 2004.
2. A. F. Cheviakov, *Computation of fluxes of conservation laws*, J. Eng. Math. **66** (2010) 153–173.
3. A. F. Cheviakov, *GeM software package for computation of symmetries and conservation laws of diffrential equations*, Comp. Phys. Comm. **176** (1) (2007) 48–61.
4. W. Hereman, M. Colagrosso, R. Sayers, R. Ringler, B. Deconinck, M. Nivala and M. S. Hickman, *Continuous and discrete homotopy operators and the computation of conservation laws,* In Differential Equations with Symbolic Computation, D. Wang and Z. Zheng, Eds., pp. 249–285, Birkhüser, 2005.
5. I. S. Krasil'shchik and A. M. Vinogradov, *Symmetries and Conservation Laws for Differential Equation of Mathematical Physics*, AMS, Providence, Rhode Island, 1998
6. P. J. Olver, *Applications of Lie Groups to Differential Equations,* 2nd ed., Graduate Texts in Mathematics, Vol. 107, Springer-Verlag, New York, 1993.
7. D. Poole and W. Hereman, *Symbolic computation of conservation laws of nonlinear partial differential equations using homotopy operators,* Ph.D. dissertation, Colorado School of Mines, Golden, Colorado, 2009.
8. D. Poole and W. Hereman, *The homotopy operator method for symbolic integration by parts and inversion of divergences with applications*, Appl. Anal. **87** (2010) 433–455.

E-mail: m.jafarii@pnu.ac.ir
E-mail: R.darvazeban@student.pnu.ac.ir
E-mail: mh.moslehi@pnu.ac.ir

A Generalization of Katok Entropy Formula to Measure-Theoretic Pressure

Sanaz Lamei*
Faculty of Mathematical Sciences, University of Guilan, Guilan, Iran
and Maryam Razi
Faculty of Mathematical Sciences, University of Guilan, Guilan, Iran

Abstract. Katok proved that for a continuous map defined on a compact metric space being invariant under an ergodic probability measure, the topological entropy defined on a subset with measure greater than or equal to $1-\zeta$ is equal to its measure-theoretic entropy for any $0 < \zeta < 1$. We generalized this entropy to pressure function when the map is measurable.
Keywords: Katok entropy formula, Measure-theoretic pressure.
AMS Mathematical Subject Classification [2010]: 58F11, 37D35.

1. Introduction

Let (Z, d) to be a compact metric space and (Z, β, ν) to be a measure space where β is the Borel σ-algebra on Z and ν is a Borel probability measure. The map $f : Z \to Z$ is measureable. Let $C(Z)$ be the set of real valued continuous functions of Z. For $n \in \mathbb{N}$, $\varepsilon \geq 0$ and $x, y \in Z$, the Bowen metric is defined as

$$d^n(x, y) = \max\{d(f^i(x), f^i(y)) : \ i = 0, 1, \ldots, n-1\}.$$

Bowen balls are defined as $B(n, \epsilon) = \{y \in Z : d^n(x, y) < \epsilon\}$ [1, 4]. A finite subset E of Z is called an (n, ϵ)-spanning set for Z if for any $x \in Z$, there is $y \in E$ with $d_n(x, y) < \epsilon$. A finite subset F of Z is called an (n, ϵ)-separated set for Z if for any $x, y \in F$, $d_n(x, y) \geq \epsilon$.

Katok proved that for a continuous map f defined on a compact metric space (Z, d) being invariant under an ergodic probability measure ν, the topological entropy defined on a subset with measure greater than or equal to $1 - \zeta$ is equal to its measure-theoretic entropy for any $0 < \zeta < 1$ [3]. This means that for any ergodic probability measure ν, and $0 < \zeta < 1$,

$$\lim_{\epsilon \to 0} \limsup_{n \to \infty} -\log M_\nu(n, \epsilon, \zeta)^n = \lim_{\epsilon \to 0} \liminf_{n \to \infty} -\log M_\nu(n, \epsilon, \zeta)^n = h_\nu(f),$$

where $M_\nu(n, \epsilon, \zeta)$ shows the minimal number of Bowen balls $B(n, \epsilon)$ covering a subset of Z with ν-measure greater than or equal to $1 - \zeta$ by (n, ϵ)-spanning sets. Let $\psi \in C(Z)$ and $(S_n\psi)(x) := \sum_{i=0}^{n-1} \psi(f^i(x))$. Then the topological pressure of f with respect to ψ is defined as

$$(1)\ \mathcal{P}(f, \psi) = \inf \left\{ \sum_{x \in F} \exp(S_n\psi)(x) \mid \ F \text{ is an } (n, \epsilon)\text{-spanning set of } Z \right\},$$

*Speaker

and the measure-theoretic pressure of f with respect to ν is given by $\mathcal{P}_\nu(f,\psi) = h_\nu(f) + \int \psi d\nu$. The Katok entropy formula was extended to a version of measure-theoretic pressure function in [2] by using the (n,ϵ)-spanning sets. For a continuous function f, $\psi \in C(\mathbb{R})$ and a probability f-invariant measure ν, they proved (1) equals to $P_\nu(f,\psi)$.

Here, we let f to be a measurable function and define the measure-theoretic pressure of f on a subset of Z with measure greater than or equal to $1-\zeta$ for $0<\zeta<1$ by using (n,ϵ)-spanning and (n,ϵ)-separated sets. Then we show theses are equal to $\mathcal{P}_\nu(f,\psi)$ when ν is an ergodic measure and $\psi \in C(\mathbb{R})$.

1.1. Generalization of Katok Entropy Formula to Pressure Function. Let Z be a compact space endowed with metric d and also (Z,β,ν) be a measure space. Take $f: Z \to Z$ be a measurable map. For any $n \in \mathbb{N}$, $\epsilon > 0$, $\zeta \in (0,1)$, and ergodic measure ν define

$$\begin{aligned}\mathcal{P}^*_\nu(f,\psi,\varepsilon,n) &= \inf\{\textstyle\sum_{x\in E}\exp(S_n\psi)(x) \mid E, \text{ an } (n,\varepsilon)\text{-spanning set for subsets of } Z \text{ with } \nu\text{-measure more than or equal to } 1-\zeta\},\\ \mathcal{Q}^*_\nu(f,\psi,\varepsilon,n) &= \sup\{\textstyle\sum_{x\in F}\exp(S_n\psi)(x) \mid F, \text{ an } (n,\varepsilon)\text{-separated set for subsets of } Z \text{ with } \nu\text{-measure more than or equal to } 1-\zeta\}.\end{aligned}$$

Set

$$\mathcal{P}^*_\nu(f,\psi,\varepsilon) = \limsup_{n\to\infty} -\log \mathcal{P}^*_\nu(f,\psi,\varepsilon,n)^n,$$

$$\mathcal{P'}^*_\nu(f,\psi,\varepsilon) = \liminf_{n\to\infty} -\log \mathcal{P}^*_\nu(f,\psi,\varepsilon,n)^n,$$

and

$$\mathcal{Q}^*_\nu(f,\psi,\varepsilon) = \limsup_{n\to\infty} \log \mathcal{Q}^*_\nu(f,\psi,\varepsilon,n)^n,$$

$$\mathcal{Q'}^*_\nu(f,\psi,\varepsilon) \liminf_{n\to\infty} \log \mathcal{Q}^*_\nu(f,\psi,\varepsilon,n)^n.$$

Moreover, define

$$\begin{aligned}\mathcal{P}^*_\nu(f,\psi) &= \lim_{\varepsilon\to 0}\mathcal{P}^*(f,\psi,\varepsilon), \quad \mathcal{P'}^*_\nu(f,\psi) = \lim_{\varepsilon\to 0}\mathcal{P'}^*(f,\psi,\varepsilon),\\ \mathcal{Q}^*_\nu(f,\psi) &= \lim_{\varepsilon\to 0}\mathcal{Q}^*(f,\psi,\varepsilon), \quad \mathcal{Q'}^*_\nu(f,\psi) = \lim_{\varepsilon\to 0}\mathcal{Q'}^*(f,\psi,\varepsilon).\end{aligned}$$

THEOREM 1.1. *Suppose (Z,d) is a compact metric space and $f: Z \to Z$ is a measurable map. For any $\nu \in \mathcal{E}(f)$ and $\psi \in C(Z)$,*

$$\mathcal{P'}^*_\nu(f,\psi) = \mathcal{Q'}^*_\nu(f,\psi) = \mathcal{P}^*_\nu(f,\psi) = \mathcal{Q}^*_\nu(f,\psi) = \mathcal{P}_\nu(f,\psi),$$

where $\mathcal{P}_\nu(f,\psi) = h_\nu(f) + \int \psi d\nu$.

For any finite measurable partition $\mathcal{C}$ of (Z,β) let

$$\mathcal{C}^n = \mathcal{C} \vee f^{-1}\mathcal{C} \vee \cdots \vee f^{-(n-1)}\mathcal{C}.$$

Denote by $C_n(x)$ the member of the partition $\mathcal{C}^n$ to which x belongs and $h_\nu(f,\mathcal{C})$ shows the measure-theoretic entropy with respect to the partition $\mathcal{C}$.

Notations. Let Y be a subset of Z with $\nu(Y) \geq 1-\zeta$. Denote by $\mathcal{R}(f,Y,n,\varepsilon,\zeta)$ and $\mathcal{S}(f,Y,n,\varepsilon,\zeta)$, the smallest cardinality of all (n,ε)-spanning sets and the largest cardinality of all (n,ε)-separated sets on Y respectively.

LEMMA 1.2. *For any $\epsilon > 0$, suppose $\mathcal{C} = \{C_1, \cdots, C_N\}$ is a finite measurable partition of (Z, β). Then for any subset Y of Z with $\nu(Y) \geq 1 - \zeta$,*

$$\mathcal{S}(f, Y, n, \epsilon, \zeta/4) \geq \mathcal{R}(f, Y, n, \epsilon, \zeta/4) \geq \mathcal{R}(f, Y, n, \epsilon, \zeta).$$

THEOREM 1.3. [4] (Shannon-Mc Millan-Breimen) *Let f be an ergodic measure preserving transformation of the probability space (Z, β, ν) and $\mathcal{C}$ be a finite partition of (Z, β, ν). Then*

$$\log \nu(C_n(x))^n \to h_\nu(f, \mathcal{C}) a.e.$$

in $L^1(Z, \beta, \nu)$.

For any n and $s > 0$, set $M_{n,\varepsilon,s} = \{x \in Z \mid C_n(x) \in \mathcal{C}^n, \nu(C_n(x)) \geq \exp[-n(h_\nu(f, A) + s)]\}$. Let $B(t, N, n) = \sum_{m=0}^{[nt]} (N-1)^m C_m^n$ and C_m^n denotes n choose m.

LEMMA 1.4. *Let ν be an ergodic measure and $0 < \zeta < 1$. For any $\epsilon > 0$, suppose $\mathcal{C} = \{C_1, \ldots, C_N\}$ is a finite measurable partition of (Z, β). Then for the subset $M_{n,\varepsilon,s}$,*

$$\begin{array}{rcl} \mathcal{R}(f, M_{n,\varepsilon,s}, n, \varepsilon, \zeta/4) & < & \exp[-n(h_\nu(f, \mathcal{C}) + s)], \\ \mathcal{S}(f, M_{n,\varepsilon,s}, n, \varepsilon, \zeta/4) & < & \exp[-n(h_\nu(f, \mathcal{C}) + s)]. \end{array}$$

LEMMA 1.5. *Let ν be an ergodic measure and $0 < \zeta < 1$. For any $\epsilon > 0$, suppose $\mathcal{C} = \{C_1, \ldots, C_N\}$ is a finite measurable partition of (Z, β) such that $diam(\mathcal{C}) := \max\{diam(C_i) \mid C_i \in \mathcal{C}\} < \frac{\varepsilon}{2}$ and $\nu(\partial\mathcal{C}) = \nu(\cup_{i=1}^N \partial C_i) = 0$, where ∂C_i denotes the boundary of C_i. Then for any subset Y of Z with $\nu(Y) \geq 1 - \zeta$,*

$$\mathcal{S}(f, Y, n, \epsilon, \zeta/4) \geq \mathcal{R}(f, Y, n, \epsilon, \zeta/4) \geq \mathcal{R}(f, Y, n, \epsilon, \zeta), \tag{2}$$

and

$$\mathcal{R}(f, Y, n, \epsilon, \zeta) \geq \exp[n(h_\nu(f, \mathcal{C}) - \varepsilon)](1 - \zeta)/(4B(\varepsilon/2, N, n)).$$

Sketch of Proof of Theorem 1.1. Suppose Z' be a subset of Z with $\nu(Z') \geq 1-\zeta$. Let E'_{n_i} and F'_{n_i} be its (n, ε)-spanning and (n, ε)-separated sets with minimal and maximal cardinality respectively.

For any n_i, consider the set B_i with its (n_i, ϵ)-spanning set E_{n_i}. Some lines of the proof are the similar for the sets E'_{n_i} and F'_{n_i}. So, let $G \in \{E, F\}$. Then for any $x \in G'_{n_i}$, there exists $y := e(x) \in E_{n_i}$ such that $d^n(x, y) < \varepsilon$. The map ψ is continuous which gives

$$(S_{n_i}\psi)(x) \leq (S_{n_i}\psi)(e(x)) + n_i\kappa. \tag{3}$$

Therefore,

$$\begin{array}{rcl} \sum_{x \in G'_{n_i}} \exp(S_{n_i}\psi)(x) & \leq & \sum\limits_{\substack{e(x) \\ x \in G'_{n_i}}} \exp[(S_{n_i}\psi)(e(x)) + n_i\kappa] \text{ by (3)} \\ & \leq & \sum\limits_{\substack{e(x) \\ x \in G'_{n_i}}} \exp[n_i(\int \psi d\nu + \frac{1}{i} + \kappa)] \text{ by the Egorov Theorem} \\ & = & \mathrm{Card}(G'_{n_i}) \exp[n_i(\int \psi d\nu + \frac{1}{i})] \\ & \leq & \exp[n_i(h_\nu(f, \mathcal{C}) + r + \int \psi d\nu + \frac{1}{i} + \kappa)]. \text{ by Lemma 1.4} \end{array}$$

Therefore,

$$-\log \mathcal{P}^*_\nu(f, \psi, \varepsilon, n_i)^{n_i} \leq h_\nu(f, \mathcal{C}) + r + \int \psi d\nu + \frac{1}{i} + \kappa,$$

and

$$-\log \mathcal{Q}_\nu^*(f,\psi,\varepsilon,n_i)^{n_i} \leq h_\nu(f,\mathcal{C}) + r + \int \psi d\nu + \frac{1}{i} + \kappa.$$

So, letting $i, n_i \to \infty$ and $r, \kappa \to 0$, proves $\mathcal{P}'^*_\nu(f,\psi,\varepsilon) \leq \mathcal{P}_\nu(f,\psi)$ and $\mathcal{Q}'^*_\nu(f,\psi,\varepsilon) \leq \mathcal{P}_\nu(f,\psi)$.

Now to prove the reverse direction. For the set Z',

$$\begin{aligned}
\textstyle\sum_{x\in E'_{n_i}} \exp(S_{n_i}\psi)(x) &\geq \textstyle\sum_{\substack{e(x)\\ x\in E'_{n_i}}} \exp[(S_{n_i}\psi)(e(x)) - n_i\kappa] \quad \text{by (3)}\\
&\geq \textstyle\sum_{\substack{e(x)\\ x\in E'_{n_i}}} \exp[n_i(\int \psi d\nu - \frac{1}{i} - \kappa)]\\
&= \mathrm{Card}(E'_{n_i}) \exp[n_i(\textstyle\int \psi d\nu - \frac{1}{i} - \kappa)]\\
&\geq \frac{1-\zeta}{4D(\frac{\varepsilon}{2},N,n)} \exp[n_i(h_\nu(f,\mathcal{C}) - \varepsilon + \textstyle\int \psi d\nu - \frac{1}{i} - \kappa)].
\end{aligned}$$

According to (2), $\mathcal{S}(f,Y,n,\epsilon,\zeta) \geq \mathcal{R}(f,Y,n,\epsilon,\zeta)$. So,

$$Q_\nu^*(f,\psi,r,n_i) \geq P_\nu^*(f,\psi,r,n_i) \text{and} Q'^*_\nu(f,\psi,r,n_i) \geq P'^*_\nu(f,\psi,r,n_i).$$

Since

$$\lim_{n\to\infty} \log B(r,N,n)^{\frac{1}{n}} = \log(N-1)^s - \log s^s - \log(1-s)^{(1-s)},$$

by letting $i \to \infty$ and $\kappa, \varepsilon, r \to 0$, we get

$$\mathcal{Q}_\nu^*(f,\psi,\varepsilon) \geq \mathcal{P}_\nu^*(f,\psi,\varepsilon) \geq \mathcal{P}_\nu(f,\psi),$$

and

$$\mathcal{Q}'^*_\nu(f,\psi,\varepsilon) \geq \mathcal{P}'^*_\nu(f,\psi,\varepsilon) \geq \mathcal{P}_\nu(f,\psi).$$

This proves the theorem. □

References

1. R. Bowen, *Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms*, Springer Lecture Notes in Mathematics 470. Springer, Berlin, 1975.
2. L. He, J. Lv and L. Zhou, *Definition of measure-theoretic pressure using spanning sets*, Acta Math. Sinica Engl. Ser. **20** (4) (2004) 709–718.
3. A. Katok, *Lyapunov exponents, entropy and periodic orbits for diffeomorphisms*, Inst. Hautes Études Sci. Publ. Math. **51** (1980) 137–173.
4. P. Walters, *An Introduction to Ergodic Theory*, Springer, Berlin, Heidelberg, New York, 1982.

E-mail: lamei@guilan.ac.ir
E-mail: maryam.razi22@gmail.com

Poincare Map on Degenerate Centers

Mahdieh Molaei Derakhtenjani*
Department of Mathematical Sciences, University of Birjand, Birjand, Iran
Omid Rabiei Motlagh
Department of Mathematical Sciences, University of Birjand, Birjand, Iran
and Hajimohammad Mohammadi Nejad
Department of Mathematical Sciences, University of Birjand, Birjand, Iran

Abstract. We consider the differential homogeneous polynomial system of order five. We provide sufficient conditions such that the origin is a degenerate center and show that with a special perturbation, this degenerate center is a limit of a hyperbolic saddle and limit of a linear center (focus).
Keywords: Poincare map, Degenerate center.
AMS Mathematical Subject Classification [2010]: 34C07, 34C25.

1. Introduction

The second part of Hilbert 16-th problem, which proposed by D. Hilbert in 1990, is to find an upper bound on the number of limit cycle which bifurcates from the planar polynomial differential system. The method of Poincare map, Abelian integrals or Melnikov integrals, inverse integrating factor and averaging theory used to study the limit cycles which bifurcate from a center [1, 2, 3, 4]. Authors in [3] provide various conditions for which the origin is a degenerate center and also use the Poincare coefficients in polar coordinate to show that a degenerate center may be the limit of a linear center(focus), a nilpotent singularity, and even a hyperbolic saddle point.

Let $P(x, y)$ and $Q(x, y)$ are n-th degree polynomials. The origin is a degenerate center for the differential polynomial system

$$\dot{x} = P(x, y), \quad \dot{y} = Q(x, y),$$

if the origin is a center and after applying a change of variables and a suitable time rescale, the system comes into the following form

$$\dot{x} = F_1(x, y), \quad \dot{y} = F_2(x, y),$$

where F_1, F_2 are nonlinear terms. In [3], authors considered $\dot{x} = P_{2m+1}(x, y)$, $\dot{y} = Q_{2m+1}(x, y)$ as the differential system with the degenerate center in the origin and used the Poincare map method to consider the perturbed degenerate center with the homogeneous polynomials as below

$$(1)\ \dot{x} = P_{2m+1}(x, y) + \epsilon \sum_{j=1}^{2m} P_j(x, y), \quad \dot{y} = Q_{2m+1}(x, y) + \epsilon \sum_{j=1}^{2m} Q_j(x, y),$$

*Speaker

where $P_j(x,y) = \sum_{k=0}^{j} \alpha_{(k,j)} x^k y^{j-k}$ and $Q_j(x,y) = \sum_{k=0}^{j} \beta_{(k,j)} x^k y^{j-k}$. The perturbed degenerate center in polar coordinate computed as

$$\frac{dr}{d\theta} = rS(\theta) + \epsilon r\, H(r,\theta,\epsilon). \tag{2}$$

Then by considering $r(\theta, r_0, \epsilon)$ as the solution of (2) with $r(0, r_0, \epsilon) = r_0$, they assumed the Poincare map $P(r_0,\epsilon) = r(2\pi, r_0, \epsilon)$ and obtained the terms of Taylor expansion of the Poincare map with respect to ϵ, as (See [3, Lemma 8])

$$\Big(\frac{\partial^j r}{\partial \epsilon^j}\Big)_{\epsilon=0}(2\pi) = j \sum_{k=0}^{j-1} \binom{j-1}{k} \int_0^{2\pi} \chi^{-1}(\psi) \frac{\partial^k r}{\partial \epsilon^k}(\psi) \frac{\partial^{j-1-k}}{\partial \epsilon^{j-1-k}} U(r_0,\psi,0)\, d\psi.$$

According to these terms, they provided conditions such that the degenerate center of (1) is a limit of a hyperbolic saddle, a linear focus(center) and a nilpotent fixed point (See [3, Lemma 11 and Remark 5]).

In this paper, we consider the results of [3] for the differential homogeneous polynomial system of order five as

$$P(x,y) = \sum_{i=0}^{5} a_i x^i y^{5-i}, \quad Q(x,y) = \sum_{i=0}^{5} b_i x^i y^{5-i}. \tag{3}$$

First we construct the symmetric and Hamiltonian degenerate center. Next we consider the perturbation of the system (3) as below

$$\begin{aligned} P_\epsilon(x,y) &= P(x,y) + \epsilon \left(\alpha_{(0,1)} y + \alpha_{(1,1)} x + \alpha_{(0,2)} y^2 + \alpha_{(1,2)} xy + \alpha_{(2,2)} x^2\right), \\ Q_\epsilon(x,y) &= Q(x,y) + \epsilon \left(\beta_{(0,1)} y + \beta_{(1,1)} x + \beta_{(0,2)} y^2 + \beta_{(1,2)} xy + \beta_{(2,2)} x^2\right), \end{aligned} \tag{4}$$

and obtain conditions under which the degenerate center is a limit of a hyperbolic saddle and limit of linear center (focus).

2. Main Results

In the following lemma, we provide sufficient conditions such that the origin in the system (3) is a symmetric degenerate center and a Hamiltonian degenerate center.

LEMMA 2.1. *Consider the system (3).*

I) *Assume that the following conditions hold.*
   1) $a_1 = a_3 = a_5 = 0$ *and* $b_0 = b_2 = b_4 = 0$.
   2) $a_0 b_1 < 0$, $a_2 b_3 < 0$ *and* $a_4 b_5 < 0$.

   Then the origin is a symmetric degenerate center for the system (3).

II) *Assume that the following conditions hold.*
   1) $a_0 > 0, a_2 > 0$ *and* $a_4 > 0$.
   2) $b_0 = -\frac{a_1}{5}, b_1 = -\frac{a_2}{2}, b_2 = -a_3,\ b_3 = -2a_4, b_4 = -5a_5, a_5^2 \le -\frac{1}{9} a_4 b_5$.
   3) $a_1^2 \le \frac{25}{18} a_0 a_2, a_3^2 \le 2 a_2 a_4$.

   Then the system (3) is a Hamiltonian system and the origin is a degenerate center for it.

PROOF. The proof is obvious by applying [3, Theorem 1(II), Theorem 2] for $m = 2$. □

In the next lemma, we consider the perturbed degenerate center (4) and impose conditions on it, such that the degenerate center is a limit of a hyperbolic saddle and limit of a linear center (focus).

LEMMA 2.2. *Consider the system* (4) *and suppose*

$$\alpha_{(1,1)} = \alpha_{(0,2)} = \alpha_{(2,2)} = 0 \quad \textit{and} \quad \beta_{(0,1)} = \beta_{(1,2)} = 0.$$

Then the following results hold.

I) *If* $\alpha_{(0,1)} = \beta_{(1,1)} = 1$*, then the degenerate center is a limit of a hyperbolic saddle.*

II) *If* $\alpha_{(0,1)} = -1$ *and* $\beta_{(1,1)} = 1$*, then the origin is a limit of a linear center (focus).*

PROOF. The proof is obvious by considering [3, Remark 5] for $m = 2$ and $j = 1, 2$. □

Now we consider the above results in a numerical example.

EXAMPLE 2.3. Consider the system

$$P_\epsilon(x,y) = \frac{x^5}{3} + x^4 y + 2x^3 y^2 + 2x^2 y^3 + \frac{5xy^4}{4} + y^5 + \epsilon\big(\alpha_{(0,1)} y + \alpha_{(1,2)} xy\big),$$
$$Q_\epsilon(x,y) = -x^5 - \frac{5x^4 y}{3} - 2x^3 y^2 - 2x^2 y^3 - xy^4 - \frac{y^5}{4} + \epsilon\big(\beta_{(1,1)} x + \beta_{(0,2)} y^2 + \beta_{(2,2)} x^2\big).$$

The unperturbed system, i.e, $\epsilon = 0$, according to Lemma 2.1(II) is the Hamiltonian degenerate center (See Figure 1(a)). Consider the perturbed system, i.e, $\epsilon \neq 0$. let

$$\epsilon = 0.1,\ \alpha_{(0,1)} = \beta_{(1,1)} = \beta_{(2,2)} = 1,\ \alpha_{(1,2)} = \beta_{(0,2)} = 0.$$

By applying Lemma 2.2(I), the origin is a limit of a hyperbolic saddle (See Figure 1(b)). Now let

$$\epsilon = 0.2,\ \alpha_{(0,1)} = -1,\ \beta_{(1,1)} = 1,\ \alpha_{(1,2)} = 0.5,\ \beta_{(0,2)} = \beta_{(2,2)} = 0.$$

By applying Lemma 2.2(II), the origin is a limit of linear center (See Figure 1(c)).

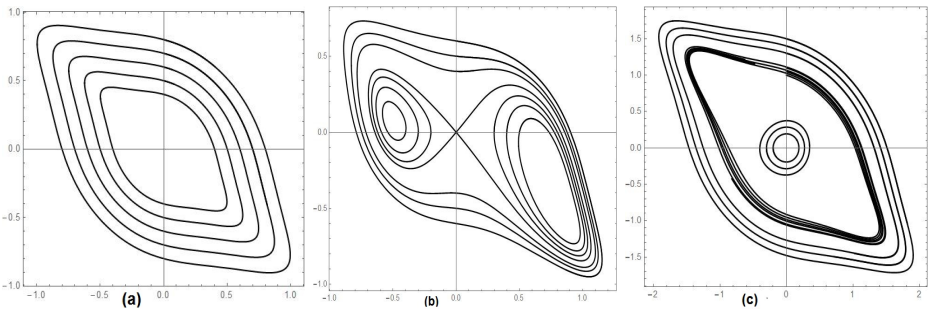


FIGURE 1. (a) Phase portrait of the unperturbed system. (b) Phase portrait of the perturbed system for $\epsilon = 0.1$. (c) Phase portrait of the perturbed system for $\epsilon = 0.2$.

References

1. J. Giné, *On the centers of planar analytic differential systems*, Internat. J. Bifur. Chaos Appl. Sci. Engrg. **17** (9) (2007) 3061–3070.
2. J. Llibre and C. Pantazi, *Limit cycles bifurcating from a degenerate center*, Math. Comput. Simulation **120** (2016) 1–11.
3. M. Molaei Derakhtenjani, O. Rabiei Motlagh and H. M. Mohammadi Nejad, *A constructive approach to degenerate center problem*, Int. J. Dyn. Syst. Differ. Equat., accepted.
4. J. Yang and P. Yu, *Nine limit cycles around a singular point by perturbing a cubic Hamiltonian system with a nilpotent center*, Appl. Math. Comput. **298** (2017) 141–152.

E-mail: m.molaei@birjand.ac.ir
E-mail: orabieimotlagh@birjand.ac.ir
E-mail: hmohammadin@birjand.ac.ir

Stability of a Stochastic Model of the Burst Neurons

Fatemeh Sadat Mousavinejad*
Department of Mathematics, Yazd University, 89195-741 Yazd, Iran
and Mehdi Fatehinia
Department of Mathematics, Yazd University, 89195-741 Yazd, Iran

Abstract. In this paper, we attempt to determine the stability of a model of the burst neurons, and resettable integrator. In order to obtain the stability of the model, we investigate, polar coordinates, Taylors expansion, and stochastic averaging method. A more comprehensive study, would include some theorems that give us some conditions which leads us to sufficient conditions on drift and diffusion coefficients for stochastic stability of the model. The most striking result to appear from the data is that the part of saccadic model in eye movements is stable under different noises.
Keywords: Noise, Saccadic model, Stability, Stochastic equation.
AMS Mathematical Subject Classification [2010]: 60H10, 34K50, 92C20.

1. Introduction

The task of human eye systems modeling, is one of the special cases for modeling and examining biomechanical systems in the nature and the body of human being. This research, which is based on the work of the saccadic model the horizontal direction, for the eye movements, indicates to introduce the stochastic model of the burst neurons [7]. In fact, the saccadic eye movements are the one which enables the humans to move eyes precisely and rapidly. Indeed, neurophysiological studies show that saccadic control signals for horizontal saccades are produced by neurons stimulated in the brain stem [2, 4].

In recent years, research regarding stochastic dynamical systems has drawn attention. As far as we know, the stochastic systems can describe natural phenomena better, such as eye movements, neural activity etc. Different theories exist in the dynamics of stochastic models [5]. There are some examples for stochastic averaging method [8], such as Hopfield neural network [3].

In this work, we decided to investigate the model of burst neurons from the saccadic system which it is based on observations the experimental findings of Van Gisbergen et al. [6]. The evidence from researches indicates that the model of burst neurons is simulated by control models [2, 4]. The achievements of these studies express that Broomhead et al. 2000 have introduced a saccadic model of slow-fast differential equations [1, 2]. More specifically, we investigate the stability of the model of the burst neurons and resettable integrator with multiplicative excitations which explained the connecting between the long and short lead burst neurons.

*Speaker

2. Main Results

In this paper, we have some theorems [5] that help us in order to find a different behaviours of the burst neurons model. And hence the stability or instability can be anticipated that.

THEOREM 2.1. [5]

i) *When* $\mu_1+\frac{1}{16}\mu_2-\frac{1}{16}\mu_4<0$, *the trivial solution of the linear Itô stochastic differential Equation is asymptotically stable with probability* 1, *thus the stochastic system is stable at the equilibrium point* O.

ii) *When* $\mu_1+\frac{1}{16}\mu_2-\frac{1}{16}\mu_4>0$, *the trivial solution of the linear Itô stochastic differential Equation is unstable with probability* 1, *which implies that the stochastic system is unstable at the equilibrium point* O.

THEOREM 2.2. [5] *When* $16\mu_1+\mu_2-\mu_4<0$ *and* $2\mu_3<\mu_4$, *the stochastic system is globally stable at the equilibrium point* O.

To begin with, we represent model of the burst neurons which is one of the case of slow-fast systems. The model is

$$\begin{aligned} \epsilon\dot{b} &= -b+\alpha sign(m)(1-e^{\frac{-|m|}{\beta}}),\\ \dot{m} &= -b, \end{aligned} \tag{1}$$

where the equilibrium point is $(0,0)$, $\alpha=800$ and $\beta=6$. Here, the net burst signal is defined by b, the motor error is m and ϵ is a small positive number.

Then, we introduce the stochastic model of the burst neurons and study its stochastic stability by some theorems. Here, we assume that $b=u$ and $m=v$, therefore the stochastic differential equations system is:

$$\begin{aligned} du &= \tfrac{1}{\epsilon}(-u+\alpha sign(v)(1-e^{\frac{-|v|}{\beta}}))dt+\sigma_1 u dW_1(t),\\ dv &= -udt+\sigma_2 v dW_2(t), \end{aligned} \tag{2}$$

where $\sigma_i(i=1,2)$. This elements are selected, based on environmental conditions. $W_i(t)$ $(i=1,2)$ are independent from standard Wiener or Brownian motion processes. According to Taylors expansion, the following equivalent system is obtained

$$\begin{aligned} du &= (\tfrac{\alpha\, v}{\beta\,\epsilon}\pm 1/2\,\tfrac{\alpha\, v^2}{\beta^2\epsilon}+1/6\,\tfrac{\alpha\, v^3}{\beta^3\epsilon}-\tfrac{u}{\epsilon}+O(4))dt+\sigma_1 u dW_1(t),\\ dv &= -udt+\sigma_2 v dW_2(t), \end{aligned}$$

In order to study the stability of system (2), we can rewritten model (2) by using the polar coordinates and the Itô stochastic differential equations. We have,

$$\begin{cases} dr = [(\mu_1+\frac{1}{16}\mu_2)r+\frac{1}{8}\mu_3 r^3]dt+(\frac{\mu_4}{8}r^2)^{\frac{1}{2}}dW_r(t),\\ d\theta = [\frac{1}{4}\mu_5+\frac{1}{8}\mu_6 r^2]dt+(\frac{\mu_2}{8})^{\frac{1}{2}}dW_\theta(t), \end{cases} \tag{3}$$

where a fix point of the model is $r=0$. The following notations are used here:

$$\begin{aligned} \mu_1 &= -\tfrac{1}{2\epsilon}, & \mu_2 &= \sigma_1^2+\sigma_2^2,\\ \mu_3 &= 0, & \mu_4 &= 3\sigma_1^2+\sigma_2^2,\\ \mu_5 &= -2-2\tfrac{\alpha}{\beta\epsilon}, & \mu_6 &= -\tfrac{\alpha}{6\beta^3\epsilon}. \end{aligned}$$

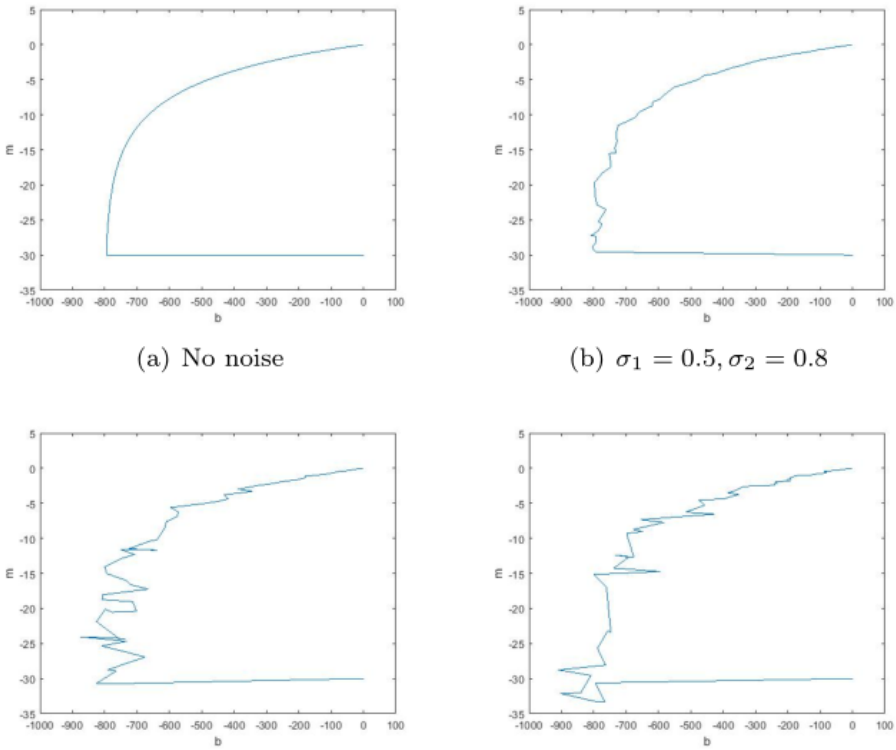
Now, according to theorems provided by Theorems 2.1 and 2.2 in Chaoliang Luo and Shangjiang Guo [5], we investigate that the system (2) is stable at the singular point O.

THEOREM 2.3. *When* $-\frac{1}{2\epsilon} - \frac{\sigma_1^2}{8} < 0$, *the trivial solution of the linear Itô stochastic differential model* (3) *is asymptotically stable with probability* 1.

Therefore, the stochastic system (2) is stable at the singularity point O by Theorem 2.1.

THEOREM 2.4. *When* $-\frac{8}{\epsilon} - 2\sigma_1^2 < 0$ *and* $0 < 3\sigma_1^2 + \sigma_2^2$, *the stochastic system* (2) *is globally stable at the singularity point* O.

Now, we illustrate that our results in the numerical simulation for the stochastic model of the burst neurons agree with our results in Theorems 2.3 and 2.4. Figure 1 shows the stability of system (2) for different noise, where we assume $\alpha = 800$ and $\beta = 6$ with the initial value $(u_0, v_0) = (-30, 0)$.



(a) No noise (b) $\sigma_1 = 0.5, \sigma_2 = 0.8$

(c) $\sigma_1 = 2, \sigma_2 = 1$ (d) $\sigma_1 = 3, \sigma_2 = 2$

FIGURE 1. Phase portrait for system (2) for the initial value $(b_0, m_0) = (-30, 0)$.

The paper presented stochastic burst model in different noise parameters, via stochastic averaging method. In fact, a stable equilibrium exists in a normal

saccadic system steady fixation. This study has identified the origin as a unique stable equilibrium point of the saccadic system. It corresponds exactly to the endpoint of the burst signal. Moreover, these results must be made available to provide for further studies checking pharmaceutical treatment for nystagmus.

References

1. O. E. Akman, D. S. Broomhead, R. V. Abadi and R. A. Clement, *Eye movement instabilities and nystagmus can be predicted by a nonlinear dynamics model of the saccadic system*, J. Math. Biol. **51** (6) (2005) 661–694.
2. O. Akman, D. S. Broomhead and R. A. Clement, *Mathematical models of eye movements*, Math. Today (Southend-on-Sea) **39** (2) (2003) 54–59.
3. Z. Huang, Q. G. Yang and J. Cao, *Stochastic stability and bifurcation analysis on Hopfield neural networks with noise*, Expert Syst. Appl. **38** (8) (2011) 10437–10445.
4. D. Laptev, O. E. Akman and R. A. Clement, *Stability of the saccadic oculomotor system*, Biol. Cybernet. **95** (3) (2006) 281–287.
5. C. Luo and S. Guo, *Stability and bifurcation of two-dimensional stochastic differential equations with multiplicative excitations*, Bull. Malays. Math. Sci. Soc. **40** (2) (2017) 795–817.
6. J. A. Van Gisbergen, D. A. Robinson and S. Gielen, *A quantitative analysis of generation of saccadic eye movements by burst neurons*, J. Neurophysiol. **45** (3) (1981) 417–442.
7. Z. Xiangyun and W. Zhiqiang, *Dynamics of a horizontal saccadic oculomotor system with colored noise*, Chin. J. Phys. **56** (5) (2018) 2052–2060.
8. W. Q. Zhu, Z. L. Huang and Y. Suzuki, *Response and stability of strongly non-linear oscillators under wider-band random excitation*, Internat. J. Non-Linear Mech. **36** (8) (2001) 1235–1250.

E-mail: f.s.mousavinejad@stu.yazd.ac.ir
E-mail: fatehiniam@yazd.ac.ir

Laplace-Adomian Decomposition Method for Solving a Model of HIV Infection on CD4$^+$ Cells

Monireh Nosrati Sahlan*
Faculty of Mathematical Sciences, University of Bonab, Bonab, Iran
and Mahin Aas
Faculty of Mathematical Sciences, University of Bonab, Bonab, Iran

Abstract. The dynamic of HIV infection of CD4$^+$ T cells is considered as a fractional order nonlinear ordinary differential equations system. In this paper using Laplace transform and Adomian decomposition method the fractional nonlinear system reduces to a linear algebraic system. By solving the algebraic system, the solutions are calculated. The numerical solution of illustrative case study shows that the purposed method is easy implement and accurate.
Keywords: Fractional model of HIV infection, Adomian decomposition method, Pade approximant.
AMS Mathematical Subject Classification [2010]: 34A34, 26A33.

1. Introduction

CD4$^+$ T cells have an important role in adjusting the immune system. Pereslon introduced the infection model of HIV into the human immune system [5] in 1980. HIV destroys the human body defense soldiers, CD4$^+$ T cells, in the blood. So the human body becomes defenseless against all other infections. In the case of treatment on the right and early time, the immune system can be protected and HIV progress is going to be controlled. Indeed the determination of the numbers of infected T cells and uninfected T cells is vital in treatment of the infection. A mathematical model for the dynamic of the HIV infection on the CD4$^+$ T cells is derived in [1] as follows:

$$\begin{cases} {}_cD^{\gamma_1} S = \rho - \mu S + \omega S \left(1 - \frac{S+I}{S_{max}}\right) - \beta V S, \\ {}_cD^{\gamma_2} I = \beta V S - \delta I, \\ {}_cD^{\gamma_3} V = N\delta I - \alpha V, \end{cases} \tag{1}$$

subject to the initial conditions

$$S(0) = \lambda_s, \qquad I(0) = \lambda_i, \qquad V(0) = \lambda_v,$$

and the descriptions of unknowns and parameters of system (1) are given in Table 1. Also S_{max} is the maximum CD4$^+$ T cell concentration in the body, N is the virus particles produced by infected CD4^+T cells and the factor $1 - \frac{S+I}{S_{max}}$ presents

*Speaker

the the logistic growth of the healthy CD4$^+$ T cells and fractional order derivative is assumed in Caputo sense.

The main aim of this paper is finding the unknowns $S(t), I(t)$ and $V(t)$ for determine the dynamic of infection of HIV. For this purpose first by using Laplace transform the system of nonlinear differential equations reduced to a system of algebraic equations and then Adomian decomposition method is applied for obtaining the solutions in the power series form. We will truncate the obtained solutions and introduce the approximated solutions of system (1).

2. Caputo Fractional Derivative

Now we briefly present the known definition of Caputo fractional derivative.

Definition 2.1. The Caputo fractional derivative operator of order nonnegative ι is defined as [2]

$$ {}_cD^{\alpha} f(x) = \frac{1}{\Gamma(n-\alpha)} \int_0^x \frac{f^{(n)}(t)}{(x-t)^{\alpha+1-n}} dt, \quad n-1 < \alpha \leq n, \quad n \in N. $$

Based on applying the Laplace transform on the system (1), we express the following property for Caputo derivative.
The Laplace transform of Caputo fractional derivative is as follows

$$ \mathcal{L}\left\{ {}_cD^{\alpha} f \right\} = s^{\alpha} \mathcal{L}\{f\} - \sum_{k=0}^{n-1} s^{\alpha-k-1} f^{(k)}(0), \qquad n-1 < \alpha \leq n. $$

Table 1. Model variables and parameters and their descriptions.

variables	description
$S(t)$	the concentration of susceptible CD4$^+$ T cells
$I(t)$	the concentration of infected CD4$^+$ T cells
$V(t)$	free HIV virus particles in the blood
parameters	description
ρ	Rate of CD4$^+$ T cells produced in body
μ	Natural turnover rates of uninfected T cells
ω	Growth rate of CD4+T cell
β	Infection rate
δ	Natural turnover rate of infected T cells
α	Natural turnover rate of virus particles

3. Numerical Implementation

In this section first applying Laplace transform to both sides of the model (1) we get

$$ (2) \quad \begin{cases} s^{\gamma_1} \mathcal{L}\{S\} = s^{\gamma_1 - 1} \lambda_s + \frac{\rho}{s} + (\omega - \mu) \mathcal{L}\{S\} - \frac{\omega}{S_{max}} \mathcal{L}\{S(S+I)\} - \beta \mathcal{L}\{VS\}, \\ s^{\gamma_2} \mathcal{L}\{I\} = s^{\gamma_2 - 1} \lambda_i + \beta \mathcal{L}\{VS\} - \delta \mathcal{L}\{I\}, \\ s^{\gamma_2} \mathcal{L}\{V\} = s^{\gamma_3 - 1} \lambda_v + N\delta \mathcal{L}\{I\} - \alpha \mathcal{L}\{V\}, \end{cases} $$

then by writing the unknowns of the system (2) in Adomian infinite series, we have

$$\begin{aligned} S(t) &= \sum_{j=0}^{\infty} s_j, \\ I(t) &= \sum_{j=0}^{\infty} i_j, \\ V(t) &= \sum_{j=0}^{\infty} v_j, \end{aligned} \tag{3}$$

also the nonlinear variable terms of the system (2) are written by Adomian polynomials as

$$\begin{aligned} S^2(t) &= \sum_{j=0}^{\infty} \mathcal{A}_j = \frac{1}{\Gamma(j+1)} \frac{d^j}{dt^j} \left(\sum_{l=0}^{j} \lambda^l s_l \right)^2_{\lambda=0}, \\ S(t)I(t) &= \sum_{j=0}^{\infty} \mathcal{B}_j = \frac{1}{\Gamma(j+1)} \frac{d^j}{dt^j} \left(\sum_{l=0}^{j} \lambda^l s_l \sum_{l=0}^{j} \lambda^l i_l \right)_{\lambda=0}, \\ S(t)V(t) &= \sum_{j=0}^{\infty} \mathcal{C}_j = \frac{1}{\Gamma(j+1)} \frac{d^j}{dt^j} \left(\sum_{l=0}^{j} \lambda^l s_l \sum_{l=0}^{j} \lambda^l v_l \right)_{\lambda=0}. \end{aligned} \tag{4}$$

Now by substituting equations (3)-(4) in (2), we get

$$\begin{cases} \mathcal{L}\{s_{j+1}\} = \frac{1}{s^{\gamma_1}} \Big((\omega - \mu)\mathcal{L}\{s_j\} - \frac{\omega}{S_{max}} \mathcal{L}\{\mathcal{A}_j + \mathcal{B}_j\} - \beta \mathcal{L}\{\mathcal{C}_j\} \Big), \\ \mathcal{L}\{i_{j+1}\} = \frac{1}{s^{\gamma_2}} \Big(\beta \mathcal{L}\{\mathcal{C}_j\} - \delta \mathcal{L}\{i_j\} \Big), \\ \mathcal{L}\{v_{j+1}\} = \frac{1}{s^{\gamma_3}} \Big(N\delta \mathcal{L}\{i_j\} - \alpha \mathcal{L}\{v_j\} \Big), \end{cases} \tag{5}$$

where

$$\mathcal{L}(s_0) = \frac{\lambda_s}{s} + \frac{\rho}{s^{\gamma_1+1}}, \qquad \mathcal{L}(i_0) = \frac{\lambda_i}{s}, \qquad \mathcal{L}(v_0) = \frac{\lambda_v}{s}.$$

By taking inverse Laplace transform of linear system (5), we obtain the analytical solutions of nonlinear system (1) as (3). The M-term approximation for analytical solutions of the Laplace Adomian decomposition method are

$$\begin{aligned} S_M(t) &= \sum_{j=0}^{M-1} s_j, \\ I_M(t) &= \sum_{j=0}^{M-1} i_j, \\ V_M(t) &= \sum_{j=0}^{M-1} v_j. \end{aligned} \tag{6}$$

Now for accelerating the convergence of the numerical solutions, Diagonal Pade Approximant (DPA[n,n]) is employed [4]. Pade approximant $[n, m]$ is a special case of rational approximation which approximate the function $f(t)$ as

$$PA[n, m](f) = \frac{a_n t^n + a_{n-1} t^{n-1} + \cdots + a_0}{b_m t^m + b_{m-1} t^{m-1} + \cdots + b_m},$$

where a_n, b_m and b_0 are unvanished. In this paper the DPA[3,3] is applied for the obtained solutions $S_M(t), I_M(t)$ and $V_M(t)$ for increasing the order of convegence.

4. Numerical Result

In this section we solve the fractional order nonlinear system (1) by purposed method for $\gamma_1 = \gamma_2 = \gamma_3 = 0.5, 0.75, 1$, $\rho = 0.1$, $\mu = 0.02$, $\omega = 3$, $\beta = 0.0027$, $\delta = 0.3$, $\alpha = 2.4$, $N = 10$ and $S_{max} = 1500$. The results $S(t), I(t)$ and $V(t)$ calculated for $M = 5$ in (6) and approximated by DPA[3,3], are given in Table 2 and compared with the results of [3], which applied variational iteration method for solving HIV infection of $CD4^+$ T cells. It is clear that the results of two approaches are in good agreement.

TABLE 2. Comparison of numerical solutions for $\gamma = 1$.
($I(t)*$: The results for $I(t)$ have been multiplied by 10^5.)

	Methods	$t = 0.2$	$t = 0.4$	$t = 0.6$	$t = 0.8$	$t = 1$
$S(t)$	Presented Method	0.208794	0.401055	0.699141	1.127030	2.237627
	DPA[3,3]	0.208807	0.406133	0.763571	1.397756	2.505684
	VIM [3]	0.208807	0.406135	0.762453	1.397881	2.506747
	Runge-Kutta	0.208808	0.406241	0.764424	1.414047	2.591594
	HDM [3]	0.061880	0.024391	0.024391	0.009968	0.003305
$I(t)*$	Presented Method	6.0156×10^{-1}	1.2953	2.0380	2.7956	3.5450
	DPA[3,3]	6.0325×10^{-1}	1.3149	2.1015	2.7950	2.4318
	VIM [3]	6.0326×10^{-1}	1.3149	2.1014	2.7951	2.4316
	Runge-Kutta	6.0327×10^{-1}	1.3158	2.1224	3.0177	4.0038
$V(t)$	Presented Method	0.061877	0.038233	0.023226	0.012647	0.005274
	DPA[3,3]	0.061880	0.023851	0.023421	1.397756	2.505684
	VIM [3]	0.061880	0.023920	0.023920	0.016218	0.01608
	Runge-Kutta	0.061880	0.023704	0.023704	0.014600	0.009101

References

1. W. Liancheng and Y. L. Michael, *Mathematical analysis of the global dynamics of a model for HIV infection of* $CD4^+$ *T cells*, Math. Biosci. **200** (1) (2006) 44–57.
2. I. Podlubny, *Fractional Differential Equations: An Introduction to Fractional Derivatives, Fractional Differential Equations to Methods of Their Solution and Some of Their Applications*, Academic Press, San Diego, CA, 1999.
3. A. Atangana and E. Alabaraoye, *Solving a system of fractional partial differential equations arising in the model of HIV infection of* $CD4^+$ *cells and attractor one-dimensional Keller-Segel equations*, Adv. diff. equat. **94** (2013) 1–14.
4. G. Adomian, *Solving Frontier Problems of Physics: The Decomposition Method*, Kluwer Academic Publishers Group, Dordrecht, 1994.
5. A. S. Perelson, *Modeling the Interaction of the Immune System with HIV*, 350–370, Lecture Notes in Biomath. 83, Springer, Berlin, 1989.

E-mail: nosrati@ubonab.ac.ir
E-mail: mahinaas98@gmail.com

Control Bifurcations for a Family of Linearly Uncontrollable Nilpotent Planner Plants

Majid Gazor
Department of Mathematical Sciences, Isfahan University of Technology, Isfahan 84156-83111, Iran
and Nasrin Sadri*
School of Mathematics, Institute for Research in Fundamental Sciences (IPM), P. O. Box 19395-5746, Isfahan, Iran

Abstract. This conference paper deals with control bifurcations of linear controllability for a generic family of planar differential plants with locally nilpotent linearly uncontrollable equilibrium. The results, of course, are readily applicable to higher dimensional systems via center manifold theory; *e.g.,* see [4]. We show how control bifurcations can help to design a compensator for controllers who start to fail their responsibilities. We illustrate the original idea from A. J. Krener, Kang, and Chang [7, 8] to show how one can move a linearly uncontrollable equilibrium to a linearly controllable equilibrium. Then, we apply input-state feedback linearization method for introducing a local compensator to tune its dynamics. We claim that our approach is a powerful and natural mathematical alternative method for many compensator design techniques in nonlinear control theory.
Keywords: Control bifurcations, Uncontrollable nilpotent system, Linear controllability.
AMS Mathematical Subject Classification [2010]: 58E25, 34H20, 37N35.

1. Motivation

Most controlled designed plants in control engineering and industry fail to perform properly after certain time. These failures may be due to many reasons such as reduction in the efficiency of its components or those of its environment. In any case, these will change certain parameters of the plant and therefore, the controller design fail to reach its desired dynamics. As soon as the controller fails to follow the desired dynamics, the controlled plant is called *singular* and the change in its qualitative dynamics is named a "*control bifurcation*". Note that this should not be confused with "*bifurcation control*". Control bifurcation refers to a controlled plant while bifurcation control refers to a parametric singular (differential) system. For bifurcation control, one needs to tune the parameters of the singular parametric differential system so that the system follows a desired dynamics. This is always considered a positive phenomenon. However, the control bifurcation is a controlled system whose qualitative dynamics change. Control bifurcations for controlled systems in industry are mainly considered as a negative phenomena. However, cognitive uses of control bifurcations in mathematics can

*Speaker

be utilized for a positive compensator design. Instances of mathematical control bifurcations are bifurcations in linear controllability, linear observability, linear accessability, etc.

We here are concerned with control bifurcations of linear controllability. Linear controllability is an important factor for many of nonlinear control methods such as back-stepping method, input state feedback linearization, gain scheduling, etc. In this paper we deal with control bifurcations of a generic family with nilpotent singularity. We see how the local bifurcations in this system may result in a control bifurcation of linear controllability. We apply the input state feedback linearization approach to illustrate how the control bifurcations can help in applying a input state feedback linearization and solve a regularization problem for a generic linearly uncontrollable nilpotent singular system.

2. Linear Controllability and Control Bifurcations

Control bifurcation was first introduced in [7]. Despite important and vast number of applications in control engineering and industry, this has been taken up with other researchers neither in nonlinear control community nor in dynamical system community. The main reason is due to the complexity of locating the control bifurcations in singular systems. This four pages conference paper aims to illustrate the subject and introduces our in progress project (but at its starting point) on this challenging subject. In this paper we supply a feedback controller design approach based on control bifurcations. This approach is an application of our tools in parametric normal forms and bifurcation control analysis [2, 3]. We claim that this is an ideal mathematics and natural alternative to the existing methods in modern nonlinear control theory.

Consider the linear control system

$$\dot{x} = Ax + Bu. \tag{1}$$

Here, u stands for the controller, $x \in \mathbb{R}^n$ is the state variable and A is a linear transformation on $\mathbb{R}^n$. The dimensions of u and B must be compatible and well-defined according to Eq. (1). The linear system (1) is called (linearly) controllable when for every initial and final states (x_i, x_f) and finite time T, there exists an unconstrained controller design such that the controlled system (1) transfers the state x_i to the final state x_f in finite time T; *e.g.,* see [1]. It is well-known that a nonlinear plant in the vicinity of an equilibrium is called *linearly controllable* when its linearization satisfies Kalman's controllability condition. More precisely, assume that the linear system 1 is the linearised system of the nonlinear plant, $u \in \mathbb{R}^{m\times 1}$ and $B \in \mathbb{R}^{n\times m}$. A necessary and sufficient condition for (A, B) to be linearly controllable is that Kalman's controllability matrix

$$\mathcal{C} = \left[B, AB, \ldots, A^{n-1}B\right],$$

has a full rank. We consider a nonlinear plat with an equilibrium whose linearization is given by (1) and Kalman's controllability matrix does not have a full rank. Then, many methods from nonlinear control theory is not applicable for this system. Control bifurcations of linear controllability is helpful in this case. In other words, control bifurcations make many methods from nonlinear control theory applicable for local controllability of linearly uncontrollable systems. For example,

the proposed approach is applicable to the cases uncontrollable by the classical input-state feedback linearization and back-stepping methods in nonlinear control theory; see [9].

3. Illustration of the Proposed Approach Using an Example

Kang et al. [7] studied a nonlinear system with an uncontrollable linearization at the origin. They proved that a generic system of this type has a nearby controllable equilibrium. Then, they suggested a regularization approach to stabilize an uncontrollable equilibrium by moving it into a nearby linearly controllable equilibrium. In this conference paper, we illustrate the original ideas of the theory using the example given by a linearly uncontrollable equilibrium (the origin) for the plant

$$\begin{aligned} f_1: \quad \dot{x} &:= -y^2 + xy + 2x^2, \\ f_2: \quad \dot{y} &:= -x + y^2 - xy + \frac{19}{28}x^2 + v + \mu_4 + \mu_6 y, \end{aligned} \tag{2}$$

where v, μ_4 and μ_6 are referred by controller inputs. The time-reversed truncated parametric normal form for this system is given by

$$\begin{aligned} \dot{x} &= -\frac{9}{4}{\mu_4}^2 + \nu_2 x + a_1 y^2 + b_1 xy + b_3 xy^3, \\ \dot{y} &= -x - \frac{1}{4}(9\mu_4 + 2\mu_6)y + b_1 y^2 + b_3 y^4, \end{aligned} \tag{3}$$

where

$$a_1 = b_1 = 1, \quad b_3 = \frac{-69}{350}.$$

There are Hopf bifurcation varieties

$$T_{H\pm} = \left\{ (\mu_4, \mu_6) \,|\, \mu_6 = -\frac{9}{2}\mu_4 \pm \frac{9}{2}\mu_4 - \frac{27}{32}{\mu_4}^2 \pm \frac{222507}{8960}{\mu_4}^3 \right\},$$

$$T_{HmC\pm} = \left\{ (\mu_4, \mu_6) \,|\, \mu_6 = -\frac{9}{2}\mu_4 \pm \frac{15\mu_4}{14}\sqrt{9 - \frac{619\sqrt{3}}{28}\sqrt{|\mu_4|} \pm \frac{2041875}{50176}\mu_4} \right\},$$

as homoclinic bifurcation varieties; see [2] for more details on Hopf and homoclinic bifurcation varieties. These bifurcation varieties are plotted in Figure 1. Now we aim to use these bifurcation varieties to move the equilibrium from the origin to nearby points so that the new equilibria would be linearly controllable.

For any equilibrium (x^*, y^*), we have the matrixes A, B and the Kalman's controllability matrix given by

$$A = \begin{bmatrix} \frac{\partial f_1}{\partial x}(x^*, y^*) & \frac{\partial f_1}{\partial y}(x^*, y^*) \\ \frac{\partial f_2}{\partial x}(x^*, y^*) & \frac{\partial f_2}{\partial x}(x^*, y^*) \end{bmatrix}, \qquad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \qquad \det[B, AB] \neq 0,$$

respectively. For any $\mu_4 \neq 0$, $\nu_1 < 0$. Thus, each nonzero input choices for (μ_4, μ_6) taken from either of the regions (a), (b), (c), and (d) in Figure 1 give rise to two linearly controllable local equilibria of a saddle and of either a source or a sink type. For an instance, we only focus on Region (c) in Figure (1). There are two equilibria: a saddle and a stable focus that both of them are linearly

controllable. Set $\mu_4 = -0.002$ and $\mu_6 = 0.005$. There exist an estimated saddle point $(x, y) = (-0.00197, 0.00197)$ and a stable focus.

Now we can design a local controller via input-state feedback linearization. We, first, transfer the saddle equilibria to origin by transformations $x = x_1 - 0.00197$ and $y = y_1 + 0.00197$, where x_1 and y_1 are new state variables. Therefore, the new system is approximately given by

$$\begin{aligned} F_1 : \quad & \dot{x_1} = 2{x_1}^2 + x_1 y_1 - {y_1}^2 - 0.0059x_1 - 0.0059y_1, \\ F_2 : \quad & \dot{y_1} = 0.67{x_1}^2 - x_1 y_1 + {y_1}^2 - 1.0046x_1 + 0.01y_1 + v. \end{aligned} \tag{4}$$

In the second step, we prove that the new system (4) has necessary conditions for applying input-state feedback linearization. By [9, Theorem 6.2, Page 238], the system $\dot{x} = f(x) + g(x)u$ with $x \in \mathbb{R}^n, u \in \mathbb{R}$ is feedback linearizable if and only if there is a domain $D_0 \subset D$ such that

1. The matrix $G(x) = \left[g(x), \mathrm{ad_f g(x)}, \ldots, \mathrm{ad_f^{n-1} g}\right]$ has rank n for all $x \in D_0$.
2. The distribution $\Delta = \left\{g(x), \mathrm{ad_f g(x)}, \ldots, \mathrm{ad_f^{n-2} g}\right\}$ is involutive in domain D_0.

Therefore, we have

$$\mathrm{ad_F g} = [\mathrm{F}, \mathrm{g}] = \begin{bmatrix} -x_1 + 2y_1 + 0.0059 \\ x_1 - 2y_1 - 0.01 \end{bmatrix} \text{ and } \mathrm{G} = [\mathrm{g} \quad \mathrm{ad_F g}] = \begin{bmatrix} 0 & -x_1 + 2y_1 + 0.0059 \\ 1 & x_1 - 2y_1 - 0.01 \end{bmatrix}.$$

The matrix $G = [g \quad \mathrm{ad_F g}]$ is in full rank when $det(G) = x_1 - 2y_1 - 0.0059 \neq 0$. Since for $g_1, g_2 \in \Delta$, $0 = [g_1, g_2] \in \Delta$, the set $\Delta = \{g\}$ is always involutive. Therefore, there exists a function $h(x)$ so that we can use it to linearize the system. The function $h(x)$ must satisfy

$$\frac{\partial L_f^i h}{\partial x} g = 0 \qquad i = 0, \ldots, n-2, \qquad \text{and} \qquad \frac{\partial L_f^{n-1} h}{\partial x} g \neq 0.$$

In our cases, we have

$$\frac{\partial h}{\partial x} g = 0, \frac{\partial h}{\partial y_1} = 0, \frac{\partial L_f h}{\partial x} g \neq 0, \frac{\partial h}{\partial x_1} (x_1 - 2y_1 - 0.0059) \neq 0, \frac{\partial h}{\partial x_1} \neq 0.$$

Then, we can choose $h(x) = x_1$ and the change of variables

$$z_1 = x_1, \quad z_2 = L_f h = 2{x_1}^2 + x_1 y_1 - {y_1}^2 - 0.0059x_1 - 0.0059y_1,$$

to transform the state equations into $\dot{z}_1 = z_2$, and

$$\begin{aligned} \dot{z}_2 \quad = \quad & (x_1 - 2y_1 - 0.0059)\, u + 8.67{x_1}^3 + 3.64{x_1}^2 y_1 - 3{y_1}^3 - 0.027{y_1}^2 - 1.04{x_1}^2 \\ & + 1.99x_1 y_1 + 0.006x_1 - 0.000029y_1. \end{aligned}$$

Then, the controller is given by

$$v := \frac{3{y_1}^3 - 8.67{x_1}^3 - 3.64{x_1}^2 y_1 + 0.027{y_1}^2 + 1.04{x_1}^2 - 1.99x_1 y_1 - 0.006x_1 + 29 \times 10^{-6} y_1 + v'}{x_1 - 2y_1 - 0.0059}, \tag{5}$$

and we have

$$\dot{z}_1 = z_2, \qquad \text{and} \qquad \dot{z}_2 = v'.$$

Using pole placement method, the linear controller $v' = kz$, and choosing appropriate constant k, we can guarantee that trajectories of $z_1(t)$ and $z_2(t)$ approach to 0. Let the desired poles be -1. Then, $k_1 = -1$ and $k_2 = -2$. These contribute into a state feedback controller design for the nonlinear plant (2). Note that we only need to compose all invertible transformations and then, feedback the controller

into the differential system (2). Here, the most challenging part refers to the use of parametric normal forms and its transformations. Due to the tedious computations, they cannot be made by hand calculations and the implementations in a computer algebra system are inevitable; *e.g.,* see [2, 3]. The numerical simulations are now given in Figures 1(b) and 1(c).

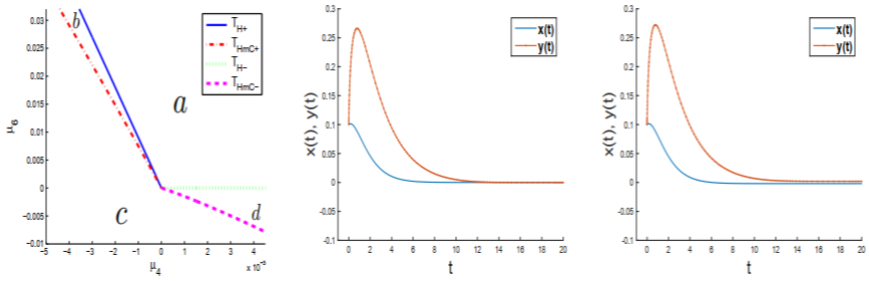


FIGURE 1. (a) Numerical transition sets for the controller inputs $(\mu_4, \mu_6$from the controlled system., (b) Tranjectories of the system (4) with controllers (5) and initial values (0.1,0.1) converge to origin., (c) Solution tranjectories (2)-(5) and initial values (0.1,0.1) converging to the origin.

References

1. R. C. Dorf and R. H. Bishop, *Modern Control System,* Prentice-Hall, Inc. Division of Simon and Schuster One Lake Street Upper Saddle River, NJ, United States, 2008.
2. M. Gazor and N. Sadri, *Bifurcation control and universal unfolding for Hopf-zero singularities with leading solenoidal terms,* SIAM J. Appl. Dyn. Syst. **15** (2016) 870–903.
3. M. Gazor and N. Sadri, *Bifurcation controller designs for the generalized cusp plants of Bogdanov–Takens singularity with an application to ship control*, SIAM J. Contr. and Opt. **57** (2019) 2122–2151.
4. B. Hamzi, W. Kang and A. J. Krener, *The controlled center dynamics,* SIAM J. Multiscale Model. Simul. **3** (2005) 838–852.
5. W. Kang, *Bifurcation and normal form of nonlinear control systems,* PART I and II, SIAM J. Contr. Opt. **36** (1998) 193–212 and 213–232.
6. W. Kang and A. J. Krener, *Extended quadratic controller normal form and dynamic state feedback linearization of nonlinear systems,* SIAM J. Contr. and Opt. **30** (1992) 1319–1337.
7. W. Kang, M. Xiao and I. A. Tall, *Controllability and local accessibility: A normal form approach,* IEEE Trans. on Automat. Contr. **48** (2003) 1724–1736.
8. A. J. Krener, W. Kang and D. E. Chang, *Control bifurcations,* IEEE Trans. Automa. Control **49** (2004) 1231–1246.
9. J. J. E. Slotine and W. Li, *Applied Nonlinear Control*, 2nd ed., Prentice-Hall, Englewood, Cliffs, 1991.

E-mail: mgazor@iut.ac.ir
E-mail: n.sadri@ipm.ir

Exitance of a Weak Solution of an Elliptic Equation

Farzaneh Safari*
Department of Pure Mathematics, Faculty Sciences, Imam Khomeini International University, Qazvin, Iran
and Abdolrahman Razani
Department of Pure Mathematics, Faculty Sciences, Imam Khomeini International University, Qazvin, Iran

Abstract. In this work we consider the elliptic problem $-\Delta_{\mathbb{H}^n} u + \lambda u = a(|\xi|_{\mathbb{H}^n})|u|^{r-2}u$, $\xi \in \Omega$, with Neumann boundary conditions on the Heisenberg group and prove the existence of at least one positive weak solution by applying a variational principle.

Keywords: Heisenberg group, Neumann problem, Variational principle.
AMS Mathematical Subject Classification [2010]: 35J20, 35R03, 46E35.

1. Introduction

Here $\mathbb{H}^n = (\mathbb{R}^{2n+1}, \circ)$ is the Heisenberg group with the following noncommutative law of product

$$(x, y, t) \circ (x', y', t') = (x + x', y + y', t + t' + 2(\langle y|x'\rangle - \langle x|y'\rangle)),$$

where $x, x', y, y' \in \mathbb{R}^n, t, t' \in \mathbb{R}$ and $\langle \ | \ \rangle$ denotes the standard inner product in $\mathbb{R}^n$. We denote by Ω the unit Korányi ball centered at the origin and $\nabla_{\mathbb{H}^n}, \Delta_{\mathbb{H}^n}$ are the Heisenberg gradient and the Kohn-Laplacian (the Heisenberg Laplacian) operators on $\mathbb{H}^n$, respectively, as defined in [5, 6, 7, 9]. Actually we consider the weighted Lebégue space

$$L^r_a(\Omega) = \{u : \int_\Omega a(|\xi|)|u|^r d\xi < \infty\},$$

endowed with the norm

$$|u|_{a,r} = (\int_\Omega a(|\xi|)|u|^r d\xi)^{\frac{1}{r}}.$$

The Heisenberg Sobolev space is defined by

$$H^1(\Omega, \mathbb{H}^n) := \{u : \Omega \to \mathbb{R} : u, |\nabla_{\mathbb{H}^n} u| \in L^2(\Omega)\},$$

endowed with the norm

$$\|u\|_{H^1} = (\int_\Omega (|\nabla_{\mathbb{H}^n} u|^2 + |u|^2) d\xi)^{\frac{1}{2}}.$$

*Speaker

We set $H_0^1(\Omega,\mathbb{H}^n) := \overline{C_0^\infty(\Omega)}^{\|u\|_{H^1}}$ where according to the Poincaré's inequality in the Heisenberg Sobolev space, the norm

$$\|u\|_* = (\int_\Omega (|\nabla_{\mathbb{H}^n} u|^2) d\xi)^{\frac{1}{2}},$$

is a norm on $H_0^1(\Omega,\mathbb{H}^n)$ which is equivalent to the standard one. Additionally, we define

$$H^2(\Omega,\mathbb{H}^n) := \{u : \Omega \to \mathbb{R} : |\Delta_H u|, |\nabla_H u| \in L^2(\Omega)\},$$

which is equipped with the norm

$$\|u\|_{H^2} = (\int_\Omega (|\Delta_{\mathbb{H}^n} u|^2 + |\nabla_{\mathbb{H}^n} u|^2 + |u|^2) d\xi)^{\frac{1}{2}}.$$

Hardy's inequality is the crucial inequality in the Sobolev spaces where this inequality in the Heisenberg Sobolev space was proved by Mokrani in 2009 [3].

LEMMA 1.1. *For $n \geq 1$ and for any $u \in H_0^1(\Omega,\mathbb{H}^n)$, we have*

$$\int_\Omega \frac{|u|^2}{(|z|^4+|t|^2)^{\frac{1}{2}}}\, d\xi \leq (\frac{n+1}{n^2})^2 \int_\Omega |\nabla_{\mathbb{H}^n} u|^2 d\xi.$$

From now on we set

$$\mathbf{X} = H_0^1(\Omega,\mathbb{H}^n) \cap H^2(\Omega,\mathbb{H}^n),$$

endowed with the norm

$$\|u\| = (\int_\Omega |\Delta_{\mathbb{H}^n} u|^2 d\xi)^{\frac{1}{2}}.$$

In the following theorem we recall compact embeddings in this space which we use to verify our main result.

THEOREM 1.2. [1] *The following embeddings are compact:*

(i) *if $Q = 4$, then $\mathbf{X} \hookrightarrow L^s(\Omega)$, $1 \leq s < \infty$,*

(ii) *if $Q > 4$, then $\mathbf{X} \hookrightarrow L^s(\Omega)$, $1 \leq s < Q^*$,*

where $Q^* = \frac{2Q}{Q-4}$ that in which $Q = 2n+2$ is homogeneous dimension of $\mathbb{H}^n$.

Now, we state the main result of this paper.

THEOREM 1.3. *Let Ω be the unit Korányi ball centered at the origin in the Heisenberg group $\mathbb{H}^n$, $2 < r < \frac{2Q}{Q-4}$ and $a \in L^\infty(0,1)$ be an increasing positive nonconstant radial function. Then the elliptic problem*

$$\begin{cases} -\Delta_{\mathbb{H}^n} u + \lambda u = a(|\xi|_{\mathbb{H}^n})|u|^{r-2}u, & \xi \in \Omega, \\ u > 0, & \xi \in \Omega, \\ \frac{\partial u}{\partial n} = 0, & \xi \in \partial\Omega, \end{cases}$$

admits at least one increasing solution in $\mathbf{X}_{rad}$ if $\lambda > 0$, where

$$\mathbf{X}_{rad} = \{u \in \mathbf{X} : \ u \text{ is a radial function}\}.$$

2. Proof of the Result

In this section, first we recall the following variational principle which is main tool for proving the result [2].

THEOREM 2.1. *Let V be a reflexive Banach space, V^* its topological dual and K be a closed convex subset of V. Let $\Phi : V \to \mathbb{R}$ be a Gâteax differentiable convex and lower semi continuous function, and let the linear operator $\Lambda : Dom(\Lambda) \subset V \to V^*$ be symmetric and positive. Assume u is a critical point of functional $I(w) = \Psi_K(w) - \Phi(w)$, where*

$$\Psi_K(w) := \begin{cases} \Phi^*(\Lambda w), & w \in K, \\ \\ +\infty, & w \notin K, \end{cases}$$

and there exists $v \in K$ satisfying the linear equation $\Lambda v = D\varphi(u)$. Then $u \in K$ is a solution of the equation

$$\Lambda u = D\Phi(u).$$

To apply Theorem 2.1, we need to consider the reflexive Banach space

$$V = \mathbf{X}_{rad} \cap L^r_a(\Omega),$$

equipped with the norm

$$\begin{aligned} |u|_V :=& \|u\|_\lambda + |u|_{a,r} \\ =& \Big(\int_\Omega |\nabla_{\mathbb{H}^n} u|^2 + \lambda |u|^2 d\xi\Big)^{\frac{1}{2}} + \Big(\int_\Omega a(|\xi|)|u|^r d\xi\Big)^{\frac{1}{r}}. \end{aligned}$$

We set $\Lambda u := -\Delta_{\mathbb{H}^n} u + \lambda u$ to be the symmetric linear operator mentioned in the Theorem 2.1 and consider $I(u) := \psi_K(u) - \varphi(u)$ as an energy functional corresponding to the problem restricted to

$$K := \{u \in V : u \text{ is positive radial function and } \frac{\partial u}{\partial n} = 0 \text{ on } \partial\Omega\},$$

where convex and lower semi continuous function φ is defined by

$$\varphi(u) = \frac{1}{r} \int_\Omega a(|\xi|_{\mathbb{H}^n})|u|^r d\xi,$$

and ψ_K is as in Theorem 2.1.

In order to verify I has a critical point, namely u, we apply Mountain Pass Geometry (MPG) Theorem established in [8]. For this reason we check the conditions of the theorem and also using embeddings mentioned above we prove that I satisfies the Palais-Smale compactness condition. Our main difficulty was passing limit through integral that we use a fact of [4, Problem 127, page 81]. Then we prove that if $h \in L^2(\Omega)$, the problem

$$\begin{cases} -\Delta_H v + \lambda v = h(\xi), & \xi \in \Omega, \\ \\ \frac{\partial v}{\partial n} = 0, & \xi \in \partial\Omega, \end{cases}$$

admits at least one solution. Using this matter, we show that for any critical point of I there exists $v \in K$ such that satisfies the linear equation $\Lambda v = D\varphi(u)$. Indeed, we check all states of the variatioal principle so u is a solution of the problem in

the weak sence. Maximal principle guarantees that u is nontrivial. Therefore, we have proved our claim.

References

1. G. Dwivedi and J. Tyagi, *Stability of positive solutions to biharmonic equations with logistic-type nonlinearities on Heisenberg group,* (2019). arXiv:1606.06413v4
2. A. Moameni, *New variational principles of symmetric boundary value problems*, J. Convex Anal. **2** (2017) 365–381.
3. H. Mokrani, *Semilinear sub-elliptic equations on the Heisenberg group with a singular potential*, Commun. Pure Appl. Anal. **8** (5) (2009) 1619–1636.
4. G. Pólya and G. Szegö, *Problems and Theorems in Analysis* I: Series. Integral Calculus. Theory of Functions. Springer-Verlag Berlin Heidelberg New York, 1978.
5. F. Safari and A. Razani, *Existence of positive radial solutions for Neumann problem on the Heisenberg group*, Bound. Value Probl. **2020** (2020) 88.
6. F. Safari and A. Razani, *Nonlinear nonhomogeneous Neumann problem on the Heisenberg group*, Appl. Anal. **2020** (2020). DOI:10.1080/00036811.2020.1807013
7. F. Safari and A. Razani, *Existence of radial solutions of the Kohn-Laplacian problem*, Complex Var. Elliptic Equ. **2020** (2020). DOI:10.1080/17476933.2020.1818733
8. A. Szulkin, *Minimax principles for lower semicontinuous functions and applications to nonlinear boundary value problems*, Ann. Inst. H. Poincaré. Anal. Non Linéaire **3** (2) (1986) 77–109.
9. J. Tyagi, *Nontrivial solutions for singular semilinear elliptic equations on the Heisenberg group*, Adv. Nonlinear Anal. **3** (2) (2014) 87–94.

E-mail: f.safari@edu.ikiu.ac.ir
E-mail: razani@sci.ikiu.ac.ir

A Frequency Domain Interpretation for the Gap Metric on the Non-Linear Operator Space: S-Gap Metric

Saman Saki*
Department of Electrical Engineering, Iran University of Science and Technology, Tehran, Iran
Hossein Bolandi
Faculty of Electrical Engineering, Iran University of Science and Technology, Tehran, Iran
and Saeed Ebadollahi
Faculty of Electrical Engineering, Iran University of Science and Technology, Tehran, Iran

Abstract. A well-known method to compare the behavior of two dynamical system is the definition of the gap metric. However, in general, there exists no explicit solution for calculation of the gap metric in the cases dealing with non-linear dynamical systems. In this paper, with a new mapping definition between the constructed graph spaces (sub-spaces of the Hilbert space) by the non-linear operators, we present a new formulation to calculate the upper bound of the gap metric. The introduced metric, called as the s-gap metric, considers the weakest topology of the most far constructed tangent spaces. The results are fruitful in the robust control theory when encountering with the stability analysis of the non-linear feedback systems.
Keywords: Robustness Analysis, Coprime Factorization, The Gap Metric, Frequency Domain Uncertainty.
AMS Mathematical Subject Classification [2010]: 93C10, 93C41.

1. Introduction

The gap metric introduced in [4] covering the linear time invariant (LTI) dynamical systems in which the weakest topology between the constructed graph of two linear dynamical systems is calculated. After a while, the application of the concept was extended to linear time varying (LTV) and non-linear dynamical systems in author's recent papers [1, 2]. Indeed, the idea behind the recent literatures in the extension of the gap metric to non-linear cases is to consider the linear systems coming from the constructed manifold topologies. This consideration is common through the control theory literatures as [3, 5] and has a drawback in which the disturbance coming from Taylor expansion is ignored in all of them.

The motivation of this paper is to formulate the gap metric in order to present a clear frequency interpretation between two dynamical systems covering the created manifold topologies and residual Taylor term.

*Speaker

1.1. Notations. For the input, state and output signals of the non-linear dynamical system, assume the signal spaces $\mathcal{U} \in \mathcal{L}_p^m$, $\mathcal{X} \in \mathcal{L}_p^n$ and $\mathcal{Y} \in \mathcal{L}_p^q$, respectively, where p is the signal norm type and also m, n and q are the dimension of the input, state and output signal. Similarly, we consider the set (in euclidean space) of the inputs, states and outputs with the symbols $\mathscr{U} \in R^{m\times 1}$, $\mathscr{X} \in R^{n\times 1}$ and $\mathscr{Y} \in R^{q\times 1}$. Now, we can consider $\mathscr{N}$ as an operator on the arbitrary graph sub-spaces $\mathcal{U}_1 \subset \mathcal{U}$ and $\mathcal{Y}_1 \subset \mathcal{Y}$ as $\mathscr{N} : \mathcal{U}_1 \to \mathcal{Y}_1$. Furthermore, to define the extended space, we need to the truncation operator as $\mathscr{T}_{a,b}u(t) = u(t)\ \forall t \in [a\ \ b]$ and zero for others. Assume the direct sum of the input and output signal spaces as $\mathcal{W} = \mathcal{U} \oplus \mathcal{Y}$. Then, the graph of the operator $\mathscr{N}$ is defined as:

$$\mathcal{G}_{\mathscr{N}} = \begin{bmatrix} I_m \\ \mathscr{N} \end{bmatrix} \mathcal{U}_1 \subset \mathcal{W}.$$

Now, consider two graph spaces constructed by operators $\mathscr{N}_1$ and $\mathscr{N}_2$, then the gap between the corresponding spaces is defined as:

$$\delta(\mathscr{N}_1, \mathscr{N}_2) = \inf_{\binom{u_2}{x_2}\in\mathcal{G}_{\mathscr{N}_2}} \sup_{\binom{u_1}{x_1}\in\mathcal{G}_{\mathscr{N}_1}} \frac{\left\| \binom{u_2}{x_2} - \binom{u_1}{x_1} \right\|}{\left\| \binom{u_1}{x_1} \right\|}. \tag{1}$$

Through the paper, we show the continues state space representation of any linear system as $\mathcal{P} = \{A, B, C, D\}$ with mapping $\mathcal{P} : u(t) \to x(t)$ and equations $\dot{x}(t) = Ax(t) + Bu(t)$ and $y(t) = Cx(t) + Du(t)$.

2. Main Results

In this section, we firstly calculate the upper bound of the non-linear gap metric and then the s-Gap metric is introduced as a meter on the non-linear operator space.

2.1. The Upper Bound of the Non-Linear Gap Metric. Consider the non-linear dynamic system of:

$$\mathscr{N} : \dot{x} = f(x(t), u(t)).$$

Then, on the operating point of $[x(t_l), u(t_l)] \in \mathscr{D}$, it can be reformulated as:

$$\mathscr{N} : \dot{x}(t) = f(x(t_l), u(t_l)) + A(x(t_l), u(t_l))(x(t) - x(t_l)) + B(x(t_l), u(t_l))(u(t) - u(t_l)),$$

where $A = \frac{\partial f}{\partial x}$ and $B = \frac{\partial f}{\partial u}$ at point of $[x(t_l), u(t_l)]$. Now, consider the following definition.

Definition 2.1. We define a mapping in which the truncated signal of $\mathscr{T}_{t_l,t_{l+1}}q(t) \in \mathcal{G}_{\mathscr{N}}$, in which $t_{l+1} \to t_l$, is mapped to the operator

$$\mathcal{P}_l^e = \{A_l, [B_l I], C_l, D_l\}.$$

Note that $\mathcal{S}$ covers all of the operators $\mathcal{P}_l^e$ constructed from linearization of any arbitrary differentiable non-linear operator. In the frequency domain, we show the operator $\mathcal{P}_l^e : [(u(t)-u(t_l))^T\ (x(t)-x(t_l))^T]^T \to x(t)$ as $\binom{N_l}{M_l} : q(t) \to \begin{pmatrix} u(t)-u(t_l) \\ x(t)-x(t_l) \end{pmatrix}$ with $M_l^* M_l + N_l^* N_l = I$. Then, we define the $\mathcal{S}_{q(t)} = \left\{ \binom{N_0}{M_0}, \binom{N_l}{M_l}, \ldots, \binom{N_l}{M_l} \right\} \in \mathcal{S}$ which is constructed from the manifold way of $q(t)$. This representation is suitable for the gap metric calculations. In the following, the norm preserving characteristic of the $S_{q(t)}$ and an explicit representation for the gap metric between two non-linear dynamic systems are represented in Theorems 2.2 and 2.3, respectively.

THEOREM 2.2. *The operator $\mathcal{S}_{q(t)} \in \mathcal{S}$ is an isometric isomorphism.*

PROOF. Let for $t \in [t_i \;\; t_f]$ the new signal illustration of

$$q(t) = \lim_{\Delta t \to 0} \sum_{l=1}^{L} \mathscr{T}_{t_l, t_l + \Delta t} q(t),$$

in which $t_{l+1} = t_l + \Delta t$ and also $L = (t_f - t_i) \diagup \Delta t$. Thus,

$$\mathcal{S}_{q(t)} q(t) = \lim_{\Delta t \to 0} \sum_{l=1}^{L} \mathscr{T}_{t_l, t_l + \Delta t} \mathcal{S}_{q(t)} q(t),$$

or equivalently, $\mathcal{S}_{q(t)} q(t) = \lim_{\Delta t \to 0} \sum_{l=1}^{L} \binom{N_l}{M_l} \mathscr{T}_{t_l, t_l + \Delta t} q(t)$. Consequently,

$$\| \mathcal{S}_{q(t)} q(t) \| = \| \lim_{\Delta t \to 0} \sum_{l=1}^{L} \binom{N_l}{M_l} \mathscr{T}_{t_l, t_l + \Delta t} q(t) \|,$$

and the fact $\| \binom{N_l}{M_l} \mathscr{T}_{t_l, t_l + \Delta t} q(t) \| = \| \mathscr{T}_{t_l, t_l + \Delta t} q(t) \|$ leads to $\| \mathcal{S}_{q(t)} q(t) \| = \| q(t) \|$. □

THEOREM 2.3. *Assume two non-linear dynamic systems as $\mathscr{N}_1$ and $\mathscr{N}_2$ with corresponding graphs $\mathcal{G}_{\mathscr{N}_1}$ and $\mathcal{G}_{\mathscr{N}_2}$. Then, the upper bound of the gap metric given by Eq. (1) can be calculated as:*

$$\begin{aligned} \delta_g(\mathscr{N}_1, \mathscr{N}_2) &\leq \inf_{Q_{mn} \in \mathcal{H}_\infty} \sup_{m,n \in \{1,2,\ldots,\infty\}} || \binom{N_{\mathscr{N}_1, n}}{M_{\mathscr{N}_1, n}} - \binom{N_{\mathscr{N}_2, m}}{M_{\mathscr{N}_2, m}} Q_{mn} ||_\infty \\ &= \delta_s \left(\mathscr{N}_1, \mathscr{N}_2 \right) \left(\mathscr{D}_1, \mathscr{D}_2 \right), \end{aligned}$$

where $\mathscr{D}_i$, $i \in \{1, 2\}$ is the equivalent domain of $\mathcal{G}_{\mathscr{N}_1}$ in the Euclidean space.

PROOF. Tacking two differential operator as $\mathcal{S}_{q_1(t)} : q_1 \to \binom{u_1}{x_1}$ and $\mathcal{S}_{q_2(t)} : q_2 \to \binom{u_2}{x_2}$. Then, the gap Eq. (1) is equivalent to:

$$\delta_g(\mathcal{G}_{\mathscr{N}_1} \mathcal{G}_{\mathscr{N}_2}) = \sup_{q_1(t) \in \mathcal{G}_{\mathscr{N}_1}} \inf_{q_2(t) \in \mathcal{G}_{\mathscr{N}_2}} \frac{|| \mathcal{S}_{q_1(t)} q_1(t) - \mathcal{S}_{q_2(t)} q_2(t) ||}{|| \mathcal{S}_{q_1(t)} q_1(t) ||}.$$

Note that $\mathcal{S}_{q_1(t)} q_1(t)$ is isometric isomorphism. Thus,

$$\begin{aligned} \frac{|| \mathcal{S}_{q_1(t)} q_1 - \mathcal{S}_{q_2(t)} q_2 ||}{|| \mathcal{S}_{q_1} q_1 ||} &\leq \frac{|| \binom{N_{\mathscr{N}_1, 1}}{M_{\mathscr{N}_1, 1}} \mathscr{T}_{t_0, t_0 + \Delta t} q_1 - \binom{N_{\mathscr{N}_2, 1}}{M_{\mathscr{N}_2, 1}} \mathscr{T}_{t_0, t_0 + \Delta t} q_2 ||}{|| q_1 ||} \\ &+ \cdots + \frac{|| \binom{N_{\mathscr{N}_1, \infty}}{M_{\mathscr{N}_1, \infty}} \mathscr{T}_{t_0, t_0 + \Delta t} q_1 - \binom{N_{\mathscr{N}_2, \infty}}{M_{\mathscr{N}_2, \infty}} \mathscr{T}_{t_0, t_0 + \Delta t} q_2 ||}{|| q_1 ||}. \end{aligned}$$

Also define, $Q_l : \mathscr{T}_{t_l, t_l + \Delta t} q_1(t) \to \mathscr{T}_{t_l, t_l + \Delta t} q_2(t)$, then,

$$\begin{aligned} \frac{|| \mathcal{S}_{q_1(t)} q_1 - \mathcal{S}_{q_2(t)} q_2 ||}{|| \mathcal{S}_{q_1} q_1 ||} &\leq \frac{\sum_{l=0}^{\infty} || \binom{N_{\mathscr{N}_1, l}}{M_{\mathscr{N}_1, l}} \mathscr{T}_{t_0, t_0 + \Delta t} q_1 - \binom{N_{\mathscr{N}_2, l}}{M_{\mathscr{N}_2, l}} \mathscr{T}_{t_0, t_0 + \Delta t} q_2 ||}{|| q_1 ||} \\ &= \frac{\sum_{l=0}^{\infty} || \binom{N_{\mathscr{N}_1, l}}{M_{\mathscr{N}_1, l}} \mathscr{T}_{t_0, t_0 + \Delta t} q_1 - \binom{N_{\mathscr{N}_2, l}}{M_{\mathscr{N}_2, l}} Q_l \mathscr{T}_{t_0, t_0 + \Delta t} q_1 ||}{|| q_1 ||} \end{aligned}$$

$$\leq \sup_{l\in\{1,\ldots,\infty\}} || \begin{pmatrix} N_{\mathscr{N}_1,l} \\ M_{\mathscr{N}_1,l} \end{pmatrix} - \begin{pmatrix} N_{\mathscr{N}_2,l} \\ M_{\mathscr{N}_2,l} \end{pmatrix} Q_l||_\infty.$$

□

2.2. The s-Gap Metric. In the previous section, we calculate the upper bound of the gap metric. Now, we are ready to extend the vinnicomb metric concepts to the non-linear cases. This helps to present a clear frequency response when comparing two non-linear dynamic systems. To that aim, we firstly need to define the following conditions for the s-gap metric.

$$\begin{cases} det\left(I + P\left(\mathscr{T}_{q_1(t)}\mathcal{G}_{\mathscr{N}_1}\right) P\left(\mathscr{T}_{q_2(t)}\mathcal{G}_{\mathscr{N}_2}\right)\right)(\omega) \neq 0,\ \forall\omega \\ wno\ det\left(I + P\left(\mathscr{T}_{q_1(t)}\mathcal{G}_{\mathscr{N}_1}\right) P\left(\mathscr{T}_{q_2(t)}\mathcal{G}_{\mathscr{N}_2}\right)\right) \\ +\mathcal{N}\left(P\left(\mathscr{T}_{q_1(t)}\mathcal{G}_{\mathscr{N}_1}\right)\right) - \mathcal{N}\left(P\left(\mathscr{T}_{q_2(t)}\mathcal{G}_{\mathscr{N}_2}\right)\right) = 0. \end{cases} \tag{2}$$

Therefore, the s-gap metric is defined as:

$$\begin{aligned} &\delta_s\left(P\left(\mathscr{T}_{q_1(t)}\mathcal{G}_{\mathscr{N}_1}\right), P\left(\mathscr{T}_{q_2(t)}\mathcal{G}_{\mathscr{N}_2}\right)\right) \\ &= \begin{cases} \sup_{m,n\in\{0,1,\ldots,\infty\}} ||\tilde{G}_{2,m}G_{1,n}||_\infty, & Eq.\ (2), \\ 1, & otherwise. \end{cases} \end{aligned} \tag{3}$$

Note that $\mathcal{N}$ shows the winding number of the argument transfer function. The idea behind definition of the Eq. (3) is the fact of

$$\delta_g(\mathscr{N}_1,\mathscr{N}_2)(\mathscr{D}_1,\mathscr{D}_2) \geq \sup_{m,n\in\{0,1,\ldots,\infty\}} ||\tilde{G}_{2,m}G_{1,n}||_\infty,$$

and the $\delta_s(\mathscr{N}_1,\mathscr{N}_2)(\mathscr{D}_1,\mathscr{D}_2)$ is the sufficient condition for the stability in the winding number point of view. We refer the reader for more details to [4]. In the following theorem, we show that $\delta_s(\mathscr{N}_1,\mathscr{N}_2)(\mathscr{D}_1,\mathscr{D}_2)$ defines a metric on the operator space.

THEOREM 2.4. *$\delta_s(\mathscr{N}_1,\mathscr{N}_2)(\mathscr{D}_1,\mathscr{D}_2)$ defines a metric on the operator space.*

PROOF. We must show that,

$$\begin{aligned} \delta_s(\mathscr{N}_1,\mathscr{N}_2)(\mathscr{D}_1,\mathscr{D}_2) &\leq \delta_s(\mathscr{N}_1, P\left(\mathscr{T}_{q(t)}\mathcal{G}_{\mathscr{N}}\right))(\mathscr{D}_1,\infty) \\ &+ \delta_s(\mathscr{N}_2, P\left(\mathscr{T}_{q(t)}\mathcal{G}_{\mathscr{N}}\right))(\mathscr{D}_2,\infty). \end{aligned}$$

Based on the Figure 1,

$$\begin{aligned} \delta_v\left(P\left(\mathscr{T}_{q_1'(t)}\mathcal{G}_{\mathscr{N}_1}\right), P\left(\mathscr{T}_{q_2'(t)}\mathcal{G}_{\mathscr{N}_2}\right)\right) &\leq \delta_v\left(P\left(\mathscr{T}_{q(t)}\mathcal{G}_{\mathscr{N}}\right), P\left(\mathscr{T}_{q_1'(t)}\mathcal{G}_{\mathscr{N}_1}\right)\right) \\ &+ \delta_v\left(P\left(\mathscr{T}_{q(t)}\mathcal{G}_{\mathscr{N}}\right), P\left(\mathscr{T}_{q_2'(t)}\mathcal{G}_{\mathscr{N}}\right)\right). \end{aligned}$$

Also clearly,

$$\delta_v\left(P\left(\mathscr{T}_{q(t)}\mathcal{G}_{\mathscr{N}}\right), P\left(\mathscr{T}_{q_1'(t)}\mathcal{G}_{\mathscr{N}_1}\right)\right) \leq \delta_v\left(P\left(\mathscr{T}_{q(t)}\mathcal{G}_{\mathscr{N}}\right), P\left(\mathscr{T}_{q_1(t)}\mathcal{G}_{\mathscr{N}_1}\right)\right),$$

for the set of $\mathscr{D}_1$, and furthermore,

$$\delta_v\left(P\left(\mathscr{T}_{q(t)}\mathcal{G}_{\mathscr{N}}\right), P\left(\mathscr{T}_{q_2'(t)}\mathcal{G}_{\mathscr{N}_2}\right)\right) \leq \delta_v\left(P\left(\mathscr{T}_{q(t)}\mathcal{G}_{\mathscr{N}}\right), P\left(\mathscr{T}_{q_2(t)}\mathcal{G}_{\mathscr{N}_2}\right)\right),$$

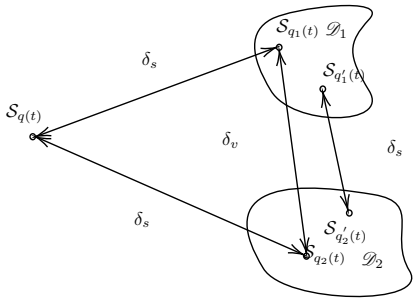


FIGURE 1. The triangular inequality concept (A domain varies).

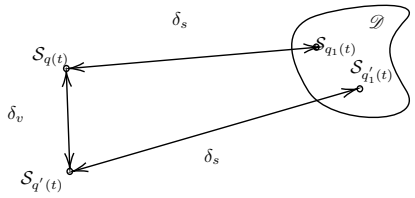


FIGURE 2. The triangular inequality concept (A single tangent space varies).

for the set of $\mathscr{D}_2$ are satisfied which conclude the first inequality. As more explanations, we point to the Figure 2. Based on this figure, we can write,

$$\begin{aligned}\delta_v\left(P\left(\mathscr{T}_{q'(t)}\mathcal{G}_{\mathscr{N}}\right),P\left(\mathscr{T}_{q'_1(t)}\mathcal{G}_{\mathscr{N}_1}\right)\right) &\leq \delta_v\left(P\left(\mathscr{T}_{q(t)}\mathcal{G}_{\mathscr{N}}\right),P\left(\mathscr{T}_{q'(t)}\mathcal{G}_{\mathscr{N}_1}\right)\right)\\ &+ \delta_v\left(P\left(\mathscr{T}_{q(t)}\mathcal{G}_{\mathscr{N}}\right),P\left(\mathscr{T}_{q'_1(t)}\mathcal{G}_{\mathscr{N}_1}\right)\right).\end{aligned}$$

Also, note that,

$$\delta_v\left(P\left(\mathscr{T}_{q(t)}\mathcal{G}_{\mathscr{N}}\right),P\left(\mathscr{T}_{q_1(t)}\mathcal{G}_{\mathscr{N}_1}\right)\right) \geq \delta_v\left(P\left(\mathscr{T}_{q(t)}\mathcal{G}_{\mathscr{N}}\right),P\left(\mathscr{T}_{q'_1(t)}\mathcal{G}_{\mathscr{N}_1}\right)\right).$$

This concludes,

$$\begin{aligned}\delta_v\left(P\left(\mathscr{T}_{q'(t)}\mathcal{G}_{\mathscr{N}}\right),P\left(\mathscr{T}_{q'_1(t)}\mathcal{G}_{\mathscr{N}_1}\right)\right) &\leq \delta_v\left(P\left(\mathscr{T}_{q(t)}\mathcal{G}_{\mathscr{N}}\right),P\left(\mathscr{T}_{q'(t)}\mathcal{G}_{\mathscr{N}_1}\right)\right)\\ &+ \delta_v\left(P\left(\mathscr{T}_{q(t)}\mathcal{G}_{\mathscr{N}}\right),P\left(\mathscr{T}_{q_1(t)}\mathcal{G}_{\mathscr{N}_1}\right)\right)\end{aligned}$$

and the proof is complete. □

Based on the given results, the $\delta_s\left(\mathscr{N}_1,\mathscr{N}_2\right)\left(\mathscr{D}_1,\mathscr{D}_2\right)$ defines a metric on the operator space. This criteria is fruitful for stability analysis of the non-linear feedback systems under uncertainty. The corresponding applications is the matter of the future publications.

References

1. H. Bolandi and S. Saki, *Design of adaptive model predictive control for a class of uncertain non-linear dynamic systems: Stability, convergence, and robustness analysis*, IET Control Theory Appl. **13** (15) (2019) 2376–2386.
2. S. Ebadollahi and S. Saki, *Wind turbine torque oscillation reduction using soft switching multiple model predictive control based on the gap metric and Kalman filter estimator*, IEEE Trans. Ind. Electron. **65** (5) (2018) 3890–3898.
3. S. Hosseini, A. Fatehi and T. A. Johansen, *Multiple model bank selection based on nonlinearity measure and H-gap metric*, J. Process Control. **22** (9) (2012) 1732–1742.
4. G. Vinnicombe, *Uncertainty and Feedback: $\mathcal{H}_\infty$ Loop-Shaping and The V-Gap Metric*, Imperial College Press, London, 2000.
5. A. Zribi, M. Chtourou and M. Djemel, *Multiple model reduction approach using gap metric and stability margin for control non-linear systems*, Int. J. Control Autom. Syst. **15** (1) (2017) 267–273.

E-mail: saman-saki@elec.iust.ac.ir
E-mail: h_bolandi@iust.ac.ir
E-mail: s_ebadollahi@iust.ac.ir

Existence of Positive Solution for Systems of Fractional q–Differential Equations via Multi-Point Boundary Value Conditions

Mohammad Esmael Samei*
Department of Mathematics, Faculty of Basic Science, Bu-Ali Sina University, Hamedan, Iran
Fateme Fasihi
Department of Mathematics, Faculty of Basic Science, Bu-Ali Sina University, Hamedan, Iran
and Hasti Zanganeh
Department of Mathematics, Faculty of Basic Science, Bu-Ali Sina University, Hamedan, Iran

Abstract. In this research, we consider the nonlinear theorems of r-concave, $(-r)$-convex and mixed monotone operators to establish the existence of positive solutions for fractional q-differential systems of operator equations on a normal cone in a real Banach space, with multipoint boundary conditions. The examples are given to confirm our results.
Keywords: Dual system, Multi-Step methods, Multi-point, q-Differential equation.
AMS Mathematical Subject Classification [2010]: 34A08, 34B16, 39A13.

1. Introduction

The quantum calculus is an old subject that was first developed by Jackson [7]. In 2011, Lv investigated the fractional differential equation ${}^c\mathbb{D}^\sigma[y](t)+\mathfrak{g}(t,y(t))=0$, for $t\in J_0=(0,1)$ under multipoint boundary conditions $y(0)=0$, ${}^c\mathbb{D}^\nu[y](1)=\sum_{i=1}^{m-2}a_i{}^c\mathbb{D}^\nu[y](e_i)$, where ${}^c\mathbb{D}^\sigma$ is the standard Riemann - Liouville fractional derivative, $n=[\sigma]+1$, here $\overline{J}_0=[0,1]$, $1<\sigma\leq 2$, $\nu\in\overline{J}_0$, with $0\leq\sigma-\nu-1$, $a_i,e_i\in J_0$, for $i=1,2,\ldots,m-2$ and $\sum_{i=1}^{m-2}a_ie_i^{\sigma-\nu-1}<1$ [8]. In [6], Henderson *et al.*, by employing the Schauder fixed point theorem investigated the system of nonlinear differential equation $y''(t)+\mu_y(t)\mathfrak{g}(z(t))=0$ and $z''(t)+\mu_v(t)\mathfrak{h}(y(t))=0$, for $0<t<T$, under the multipoint boundary conditions $\eta_1y(0)-\eta_2y'(0)=0$, $y(T)=\sum_{i=1}^{m-2}a_iy(e_i)+a_0$ and $\hat{\eta}_1z(0)-\hat{\eta}_2z'(0)=0$, $z(T)=\sum_{i=1}^{m-2}\hat{a}_iz(\hat{e}_i)+\hat{a}_0$, where $m\geq 3$.

In this study, we investigate the system of nonlinear fractional q-differential equations

$$\begin{aligned}{}^c\mathbb{D}_q^\sigma[y](t)+\eta_1\mathfrak{g}(t,z(t))+\eta_2\mathfrak{h}(t,z(t))=0,\\ {}^c\mathbb{D}_q^\sigma[z](t)+\eta_1\mathfrak{g}(t,y(t))+\eta_2\mathfrak{h}(t,y(t))=0,\end{aligned}\tag{1}$$

*Speaker

$$
\begin{aligned}
{}^{c}\mathbb{D}_q^{\sigma}[y](t) + \eta_1 \mathfrak{g}(t, y(t)) + \eta_2 \mathfrak{h}(t, z(t)) = 0,\\
{}^{c}\mathbb{D}_q^{\sigma}[z](t) + \eta_1 \mathfrak{g}(t, z(t)) + \eta_2 \mathfrak{h}(t, y(t)) = 0,
\end{aligned} \tag{2}
$$

for all $t \in J_0$, under multipoint boundary conditions

$$
{}^{c}\mathbb{D}_q^{\nu}[y](1) = \sum_{i=1}^{m-2} a_i {}^{c}\mathbb{D}_q^{\nu}[y](e_i), \quad {}^{c}\mathbb{D}_q^{\nu}[z](1) = \sum_{i=1}^{m-2} a_i {}^{c}\mathbb{D}_q^{\nu}[z](e_i), \tag{3}
$$

and $y(0) = 0$, $z(0) = 0$, where $1 < \sigma \leq 2$, $\nu \in \overline{J}_0$ with $\sigma - \nu - 1 \geq 0$, η_1, $\eta_2 \in (0, +\infty)$ with $\eta_1 \geq \eta_2$, ${}^{c}\mathbb{D}_q^{\sigma}$ is the standard Riemann-Liouville fractional q–derivative, $\mathfrak{g}$, $\mathfrak{h} : \overline{J}_0 \times [0, \infty) \to [0, \infty)$ is continuous and $a_i, e_i \in J_0$ for $i = 1, 2, \ldots, m-2$ with $\sum_{i=1}^{m-2} a_i e_i^{\sigma-\nu-1} < 1$.

2. Preliminaries, Background Materials and Some Lemmas

We recall some basic definitions, notations and results of q-fractional calculus which are used throughout this paper [2]. Let $a \in \mathbb{R}$. Define $[a]_q = (1 - q^a)/(1 - q)$ [7]. The power function $(y - z)_q^n$ with $n \in \mathbb{N}_0$ is defined by $(y - z)_q^{(n)} = \prod_{k=0}^{n-1}(y - zq^k)$, for $n \geq 1$ and $(y - z)_q^{(0)} = 1$, where y and z are real numbers and $\mathbb{N}_0 := \{0\} \cup \mathbb{N}$ [1]. The q–Gamma function is given by $\Gamma_q(\sigma) = (1 - q)^{(\sigma-1)}/(1 - q)^{\sigma-1}$, where $\sigma \in \mathbb{R}\backslash\{0, -1, -2, \ldots\}$ [7].

Note that, $\Gamma_q(\sigma + 1) = [\sigma]_q \Gamma_q(\sigma)$. The q–derivative of function $\mathfrak{g}$, is defined by $\mathbb{D}_q[\mathfrak{g}](\tau) = \frac{\mathfrak{g}(\tau) - \mathfrak{g}(q\tau)}{(1-q)\tau}$, and $\mathbb{D}_q[\mathfrak{g}](0) = \lim_{\tau \to 0} \mathbb{D}_q[\mathfrak{g}](\tau)$ [1].

Furthermore, the higher order q–derivative of a function $\mathfrak{g}$ is defined by $\mathbb{D}_q^n[\mathfrak{g}](\tau) = \mathbb{D}_q[\mathbb{D}_q^{n-1}[\mathfrak{g}]](\tau)$, for $n \geq 1$, where $\mathbb{D}_q^0[\mathfrak{g}](\tau) = \mathfrak{g}(\tau)$ [1]. The q–integral of a function $\mathfrak{g}$ is defined on $[0, b]$ by

$$
\mathbb{I}_q[\mathfrak{g}](\tau) = \int_0^{\tau} \mathfrak{g}(\xi)\, \mathrm{d}_q\xi = \tau(1 - q) \sum_{k=0}^{\infty} q^k \mathfrak{g}(\tau q^k),
$$

for $0 \leq \tau \leq b$, provided the series is absolutely converges [2].

The operator $\mathbb{I}_q^n$ is given by $\mathbb{I}_q^0[\mathfrak{g}](\tau) = \mathfrak{g}(\tau)$ and $\mathbb{I}_q^n[\mathfrak{g}](\tau) = \mathbb{I}_q[\mathbb{I}_q^{n-1}[\mathfrak{g}]](\tau)$ for $n \geq 1$ and $\mathfrak{g} \in C([0, T])$ [2]. It has been proved that $\mathbb{D}_q[\mathbb{I}_q[\mathfrak{g}]](\tau) = \mathfrak{g}(\tau)$ and $\mathbb{I}_q[\mathbb{D}_q[\mathfrak{g}]](\tau) = \mathfrak{g}(\tau) - \mathfrak{g}(0)$, whenever $\mathfrak{g}$ is continuous at $\tau = 0$ [1]. The fractional Riemann–Liouville type q–integral of the function $\mathfrak{g}$ on $J_0 = (0, 1)$ for $\sigma \geq 0$ is defined by $\mathbb{I}_q^0[\mathfrak{g}](\tau) = \mathfrak{g}(\tau)$ and

$$
\mathbb{I}_q^{\sigma}[\mathfrak{g}](\tau) = \frac{1}{\Gamma_q(\sigma)} \int_0^{\tau} (\tau - q\xi)^{(\sigma-1)} \mathfrak{g}(\xi)\, \mathrm{d}_q\xi = \tau^{\sigma}(1-q)^{\sigma} \sum_{k=0}^{\infty} q^k \frac{\prod_{i=1}^{k-1}(1 - q^{\sigma+i})}{\prod_{i=1}^{k-1}(1 - q^{i+1})} \mathfrak{g}(\tau q^k),
$$

for $t \in \overline{J}_0$ [1, 3].

The Caputo fractional q–derivative of a function $\mathfrak{g}$ is defined by

$$
{}^{c}\mathbb{D}_q^{\sigma}[\mathfrak{g}](\tau) = \mathbb{I}_q^{[\sigma]-\sigma}[\mathbb{D}_q^{[\sigma]}[\mathfrak{g}]](\tau) = \frac{1}{\Gamma_q([\sigma] - \sigma)} \int_0^{\tau} (\tau - q\xi)^{([\sigma]-\sigma-1)} \mathbb{D}_q^{[\sigma]}[\mathfrak{g}](\xi)\, \mathrm{d}_q q\xi,
$$

where $t \in \overline{J}_0$ and $\sigma > 0$ [3]. It has been proved that $\mathbb{I}_q^{\nu}[\mathbb{I}_q^{\sigma}[\mathfrak{g}]](\tau) = \mathbb{I}_q^{\sigma+\nu}[\mathfrak{g}](\tau)$, and $\mathcal{D}_q^{\sigma}[\mathbb{I}_q^{\sigma}[\mathfrak{g}]](\tau) = \mathfrak{g}(\tau)$, where $\sigma, \nu \geq 0$ [3].

LEMMA 2.1. *Let $r > \ell > 0$. Then, the formula* ${}^c\mathbb{D}_q^{\sigma}[\mathbb{I}_q^{\nu}[\mathfrak{g}]](\tau) = \mathbb{I}_q^{\nu-\sigma}[\mathfrak{g}](\tau)$ *holds almost everywhere on $\tau \in [a,b]$ for $\mathfrak{g} \in L_1[a,b]$ and it is valid at any point $\tau \in [a,b]$ if $\mathfrak{g} \in C[a,b]$.*

Let $\sigma > 0$, $\mathfrak{g} \in L^1([a,b],\mathbb{R}^+)$. Then we have $\mathbb{D}_q^{\sigma+1}[\mathfrak{g}](\tau) \leq \|\mathbb{D}_q^{\sigma}[\mathfrak{g}]\|_{L^1}$ [4]. Let $\mathcal{Y}$ be a real Banach space and P be a cone in $\mathcal{Y}$ which defined a partial ordering in $\mathcal{Y}$ by $y \leq z$ if and only if $y - z \in P$. P is said to be normal if there exists a positive constant r such that $\vartheta \leq v \leq w$ implies $\|v\| \leq N\|w\|$ [5]. P is called solid if its interior P is nonempty. An operator $\mathcal{O} : \mathcal{D} \times \mathcal{D} \to \mathcal{E}$ is said to be mixed monotone if $\mathcal{O}(v,w)$ is nondecreasing in v and non increasing in w, i.e., for all v_i, $w_i \in \mathcal{D}$ with $i = 1,2$, $v_1 \leq v_2$, and $w_2 \leq w_1$ imply $\mathcal{O}(v_1,w_1) \leq \mathcal{O}(v_2,w_2)$ [5]. For all $v, w \in \mathcal{E}$, the notation $v \sim w$ means that there exists $\ell > 0$ and $\eta > 0$ such that $\ell v \leq w \leq \eta v$ [5]. Clearly, $\sim$ is an equivalence relation. Given $h > 0$ (i.e., $h \geq 0$ and $h \neq 0$), we denote by P_h the set $P_h = \{v \in \mathcal{E} \mid v \sim h\}$. It is easy to see that $P_h \subset P$. Suppose that $0 < n-1 \leq \sigma < n$ and $y \in \overline{\mathcal{A}} \cap \overline{\mathcal{L}}$, here $\overline{\mathcal{A}} = C(\overline{J})$ and $\overline{\mathcal{L}} = L^1(\overline{J})$. Then $\mathbb{I}_q^{\sigma}[\mathbb{D}_q^{\sigma}[y]](\tau) = y(\tau) + \sum_{i=0}^{n-1} d_i\tau^i$ for some constants $d_i \in \mathbb{R}$ [9]. Let $\mathcal{D} = P$ or $\mathcal{D} = P^{\circ}$ and r be a real number with $0 \leq r < 1$. An operator $\mathcal{O} : P \to P$ is said to be r-concave ($(-r)$-convex) if it satisfies $\mathcal{O}(\ell v) \geq \ell^r \mathcal{O}(v)$, ($\mathcal{O}(\ell v) \leq \ell^{-r}\mathcal{O}(v)$), for all $\ell \in J_0$ and $v \in \mathcal{D}$ [10]. An operator $\mathcal{O} : \mathcal{E} \to \mathcal{E}$ is said to be homogeneous if it satisfies $\mathcal{O}(\ell v) = \ell\mathcal{O}(v)$, for all $\ell > 0$ and $v \in \mathcal{E}$ [11]. An operator $\mathcal{O} : P \to P$ is said to be subhomogeneous if it satisfies $\mathcal{O}(\ell v) \geq \ell\mathcal{O}(v)$, for all $\ell \in J_0$ and $v \in P$. Let P be a normal cone in a real Banach space $\mathcal{E}$ and $\mathcal{O}_1 : P \to P$ be an increasing r-concave operator and $\mathcal{O}_2 : P \to P$ be an increasing subhomogeneous operator. Assume that there is $h > 0$ such that $\mathcal{O}_1(h) \in P_h$ and $\mathcal{O}_2(h) \in P_h$ and there exists a constant $\hat{\gamma}_0 > 0$ such that $\mathcal{O}_1(v) \geq \hat{\gamma}_0\mathcal{O}_2(v)$, for all $v \in P$. Then operator equation $\mathcal{O}_1(v) + \mathcal{O}_2(v) = v$ has a unique solution v^* in P_h [11]. Moreover, constructing successively the sequence $w_n = \mathcal{O}_1(w_{n-1}) + \mathcal{O}_2(w_{n-1})$, $n \geq 1$ for any initial value $w_0 \in P_h$, we have $w_n \to v^*$ as $n \to \infty$ [11].

THEOREM 2.2. [10] *Let P be a normal cone of the real Banach space $\mathcal{E}$ and $\mathcal{O} : P \times P \to P$ be a mixed monotone operator. Suppose that for fixed w, $\mathcal{O}(\cdot,w) : P \to P$ is r_1-concave, for fixed v, $\mathcal{O}(v) : P \to P$ is $(-r_2)$-convex, where $0 \leq r_1 + r_2 < 1$, and there exist elements y_0, $z_0 \in P$ with $y_0 \leq z_0$ and a real number $\tau_0 > 0$ such that $y_0 \geq \tau_0 z_0$, $y_0 \leq \mathcal{O}(y_0,z_0)$ and $\mathcal{O}(z_0,y_0) \leq z_0$. Then $\mathcal{O}$ has exactly one fixed point $y_0 \leq v^* \leq z_0$, and constructing successively the sequence $v_n = \mathcal{O}(v_{n-1},w_{n-1})$ and $w_n = \mathcal{O}(w_{n-1},v_{n-1})$, for $n \geq 1$ and initial value $(v_0,w_0) \in [y_0,z_0] \times [y_0,z_0]$, we have $v_n \to v^*, w_n \to w^* (n \to \infty)$.*

3. Main Results

LEMMA 3.1. *Let $w \in C(\overline{J}_0)$. Then, the fractional q-differential equation*

$$ {}^c\mathbb{D}_q^{\sigma}[y](t) + w(t) = 0,$$

for $t \in J_0$, $1 < \sigma \leq 2$, under boundary conditions ${}^c\mathbb{D}_q^{\nu}[y](1) = \sum_{i=1}^{m-2} a_i {}^c\mathbb{D}_q^{\nu}[y](e_i)$, $y(0) = 0$, has a unique solution which is given by $y(t) = \int_0^1 G_q(t,\xi)w(\xi)\,\mathrm{d}_q\xi$, where $G_q(t,\xi) = {}_1G_q(t,\xi) + {}_2G_q(t,\xi)$, in which ${}_1G_q(t,\xi) = \frac{1}{\Gamma_q(\sigma)}[t^{\sigma-1}(1-\xi)^{(\sigma-\nu-1)} - (t-\xi)^{(\sigma-1)}]$, whenever $\xi \leq t$, ${}_1G_q(t,\xi) = \frac{1}{\Gamma_q(\sigma)}t^{\sigma-1}(1-\xi)^{(\sigma-\nu-1)}$, whenever $t \leq \xi$, ${}_2G_q(t,\xi) = \frac{1}{\Delta\Gamma_q(\sigma)}[\sum_{0\leq\xi\leq e_i}(a_i e_i^{\sigma-\nu-1}t^{\sigma-1}(1-\xi)^{(\sigma-\nu-1)} - a_i t^{\sigma-1}(e_i-\xi)^{(\sigma-\nu-1)})]$,

whenever $\xi \le e_i$, ${}_2G_q(t,\xi) = \frac{1}{\Delta\Gamma_q(\sigma)}[\sum_{e_i\le\xi\le 1} a_i e_i^{\sigma-\nu-1} t^{\sigma-1}(1-\xi)^{(\sigma-\nu-1)}]$, *whenever* $e_i \le \xi$, *for almost all* $t,\xi \in \overline{J}_0$, *where* $\Delta = 1 - \sum_{i=1}^{m-2} a_i e_i^{\sigma-\nu-1}$.

LEMMA 3.2. *Suppose that* $h(t) = t^{\sigma-1}$. *Then* $G_q(t,\xi)$ *is defined in Lemma 3.1 satisfies*

$$G_q(t,\xi) \le h(t)[\frac{1}{\Gamma_q(\sigma)}(1-\xi)^{(\sigma-\nu-1)} + \frac{1}{\mathcal{O}_1(\Gamma_q(\sigma))}\sum_{i=1}^{m-2} a_i e_i^{\sigma-\nu-1}(1-\xi)^{(\sigma-\nu-1)}]$$

and

$$G_q(t,\xi) \ge h(t)[\frac{1}{\Gamma_q(\sigma)}((1-\xi)^{(\sigma-\nu-1)} - (1-\xi)^{(\sigma-1)}) + \frac{1}{\mathcal{O}_1(\Gamma_q(\sigma))}\sum_{i=1}^{m-2}(a_i e_i^{\sigma-\nu-1}(1-\xi)^{(\sigma-\nu-1)} - a_i(e_i-\xi)^{(\sigma-\nu-1)})].$$

In the space $\mathcal{E} = C(\overline{J}_0,\mathbb{R})$ equipped with the norm $\|y\| = \sup_{t\in\overline{J}_0}|y(t)|$, the set $P = \{y \in \mathcal{E} : y(t) \in [0,\infty)\}$ is a cone in $\mathcal{E}$.

REMARK 3.3. From Lemma 3.1, we know that system (1)-(2) and (1)-(3) can be translated into the equations

$$\begin{aligned} y(t) &= \int_0^1 G_q(t,\xi)(\eta_1\mathfrak{g}(\xi,z(\xi)) + \eta_2\mathfrak{h}(\xi,z(\xi)))\,\mathrm{d}_q\xi, \\ z(t) &= \int_0^1 G_q(t,\xi)(\eta_1\mathfrak{g}(\xi,y(\xi)) + \eta_2\mathfrak{h}(\xi,y(\xi)))\,\mathrm{d}_q\xi, \end{aligned} \tag{4}$$

and

$$\begin{aligned} y(t) &= \int_0^1 G_q(t,\xi)(\eta_1\mathfrak{g}(\xi,y(\xi)) + \eta_2\mathfrak{h}(\xi,z(\xi)))\,\mathrm{d}_q\xi, \\ z(t) &= \int_0^1 G_q(t,\xi)(\eta_1\mathfrak{g}(\xi,z(\xi)) + \eta_2\mathfrak{h}(\xi,y(\xi)))\,\mathrm{d}_q\xi, \end{aligned} \tag{5}$$

respectively. Thus (y,z) is solution of systems (1), (2) and (3) if and only if (y,z) is a solution of system (4), and (y,z) is a solution of system (2) and (3) if and only if (y,z) is a solution of system (5). For convenience, we denote

$$P_1(\xi) = \frac{1}{\Gamma_q(\sigma)}((1-\xi)^{(\sigma-\nu-1)} - (1-\xi)^{(\sigma-1)}) + \frac{1}{\Delta\Gamma_q(\sigma)}\sum_{i=1}^{m-2}(a_i e_i^{\sigma-\nu-1}(1-\xi)^{(\sigma-\nu-1)} - a_i(e_i-q\xi)^{(\sigma-\nu-1)}),$$

$$P_2(\xi) = \frac{1}{\Gamma_q(\sigma)}(1-\xi)^{(\sigma-\nu-1)} + \frac{1}{\Delta\Gamma_q(\sigma)}\sum_{i=1}^{m-2} a_i e_i^{\sigma-\nu-1}(1-\xi)^{(\sigma-\nu-1)}.$$

THEOREM 3.4. *Suppose that the following assumptions hold.*

- (A_1) *Functions* $\mathfrak{g}$, $\mathfrak{h} \in \mathcal{C} = C[\overline{J}_0\times\mathbb{R}^+,\mathbb{R}^+]$, $\mathfrak{g}(t,w)$ *and* $\mathfrak{h}(t,w)$ *are increasing in* $w \in \mathbb{R}^+$, *and* $\mathfrak{h}(t,w) \ne 0$ *whenever* $w = 0$.
- (A_2) *There exists a constant* $r \in J_0$ *such that* $\mathfrak{g}(t,\ell w) \ge \ell^r\mathfrak{g}(t,w)$, $\mathfrak{h}(t,\ell w) \ge \ell^r\mathfrak{h}(t,w)$, *for almost all* $t \in \overline{J}_0$, $\ell \in J_0 = (0,1)$, $w \in \mathbb{R}^+$.

(A_3) *There exists a constant* $\hat{\gamma}_0 > 0$ *such that* $\mathfrak{g}(t,w) \geq \hat{\gamma}_0 \mathfrak{h}(t,w)$, $t \in \overline{J}_0$, *for* $w \in \mathbb{R}^+$.

Then, equations of system (1), (2) *and* (3) *have a unique positive solution* (y°, z°) *in* $P_h \times P_h$, *where* $h(t) = t^{\sigma-1}$, $t \in \overline{J}_0$. *Furthermore, for any initial value* y_0 *and* $z_0 \in P_h$, *constructing successively the sequence* $y_n(t) = \int_0^1 G_q(t,\xi)[\eta_1 \mathfrak{g}(\xi, z_{n-1}(\xi)) + \eta_2 \mathfrak{h}(\xi, z_{n-1}(\xi))]\,\mathrm{d}_q\xi$, $z_n(t) = \int_0^1 G_q(t,\xi)[\eta_1 \mathfrak{g}(\xi, y_{n-1}(\xi)) + \eta_2 \mathfrak{h}(\xi, y_{n-1}(\xi))]\,\mathrm{d}_q\xi$, *for* $n \in \{0\} \cup \mathbb{N}$, *we have* $(y_n(t), z_n(t)) \to (y^\circ(t), z^\circ(t))$ *as* $n \to \infty$.

Corollary 3.5. *Suppose that the following assumptions hold:*

(A_4) $\mathfrak{g} \in \mathcal{C}$, $\mathfrak{g}(t,w)$ *is increasing in* w *for* $w \in \mathbb{R}^+$, $\mathfrak{g}(t,0) \neq 0$.

(A_5) *There exists a constant* $r \in J_0$ *such that* $\mathfrak{g}(t,\lambda w) \geq \lambda^r \mathfrak{g}(t,w)$, *for almost all* $t \in \overline{J}_0$, $\lambda \in J_0$, $w \in \mathbb{R}^+$.

Then, system ${}^c\mathbb{D}_q^\sigma[y](t) + \eta_1 \mathfrak{g}(t, z(t)) = 0$, ${}^c\mathbb{D}_q^\sigma[z](t) + \eta_1 \mathfrak{g}(t, y(t)) = 0$, *for* $t \in J_0$, *with the multipoint boundary conditions* ${}^c\mathbb{D}_q^\nu[y](1) = \sum_{i=1}^{m-2} a_i {}^c\mathbb{D}_q^\nu[y](e_i)$, ${}^c\mathbb{D}_q^\nu[z](1) = \sum_{i=1}^{m-2} a_i {}^c\mathbb{D}_q^\nu[z](e_i)$ *and* $y(0) = 0$, $z(0) = 0$, *has a unique positive solution* (y°, z°) *in* $P_h \times P_h$, *where* $h(t) = t^{\sigma-1}$, $t \in \overline{J}_0$, ${}^c\mathbb{D}_q^\sigma$ *is the standard Riemann-Liouville fractional q-derivative,* $\mathfrak{g} : \overline{J}_0 \times [0,\infty) \to [0,\infty)$ *is continuous,* $1 < \sigma \leq 2$, $\nu \in \overline{J}_0$ *with* $0 \leq \sigma - \nu - 1$, $0 < a_i, e_i < 1$ *for* $i = 1, 2, \ldots, m-2$ *with* $\sum_{i=1}^{m-2} a_i e_i^{\sigma-\nu-1} < 1$ *and* η_1, $\eta_2 \in (0, +\infty)$ *with* $\eta_1 \geq \eta_2$. *Moreover, for any initial value* $y_0 \in P_h$ *and* $z_0 \in P_h$, *constructing successively the sequence* $y_n(t) = \eta_1 \int_0^1 G_q(t,\xi)\mathfrak{g}(\xi, z_{n-1}(\xi))\,\mathrm{d}_q\xi$, $z_n(t) = \eta_1 \int_0^1 G_q(t,\xi)\mathfrak{g}(\xi, y_{n-1}(\xi))\,\mathrm{d}_q\xi$, *for* $n = \{0\} \cup \mathbb{N}$, *we have* $(y_n(t), z_n(t)) \to (y^\circ(t), z^\circ(t))$ *as* $n \to \infty$.

Theorem 3.6. *Consider the following assumption:*

(A_6) $\mathfrak{g}$, $\mathfrak{h} \in \mathcal{C}$, $\mathfrak{g}(t,w)$ *and* $\mathfrak{h}(t,z)$ *are nondecreasing and nonincreasing in the second argument, respectively. In addition to,* $\mathfrak{g}(t,w)$ *and* $\mathfrak{h}(t,z)$ *are bounded in* $[\overline{J}_0 \times \mathbb{R}^+]$.

(A_7) *There exist* $0 \leq r_1 < 1$ *and* $0 \leq r_2 < 1$ *with* $0 \leq r_1 + r_2 < 1$ *such that* $\mathfrak{g}(t,kw) \geq k^{r_1}\mathfrak{g}(t,w)$ *and* $\mathfrak{h}(t,kw) \leq k^{-r_2}\mathfrak{h}(t,w)$ *respectively, for each* $k \in J_0 = (0,1)$.

Then, Eqs. (1), (2) *and* (3) *have exactly one positive solution* $(y^\circ, z^\circ) \in [y_0, z_0] \times [y_0, z_0]$, *where* $y_0, z_0 \in P$ *with* $y_0 \leq z_0$, *and constructing successively the sequence*

$$y_n(t) = \int_0^1 G_q(t,\xi)[\eta_1 \mathfrak{g}(\xi, y_{n-1}(\xi)) + \eta_2 \mathfrak{h}(\xi, z_{n-1}(\xi))]\,\mathrm{d}_q\xi,$$
$$z_n(t) = \int_0^1 G_q(t,\xi)[\eta_1 \mathfrak{g}(\xi, z_{n-1}(\xi)) + \eta_2, \mathfrak{h}(\xi, y_{n-1}(\xi))]\,\mathrm{d}_q\xi,$$

for $n \in \{0\} \cup \mathbb{N}$, *we get* $(y_n(t), z_n(t)) \to (y^\circ(t), z^\circ(t))$ *as* $n \to \infty$.

Example 3.7. Consider the system of nonlinear fractional q-differential equations

$$(6)\quad \begin{cases} {}^c\mathbb{D}_q^{\frac{17}{9}}[y](t) + \frac{125(t^3 + \sqrt[3]{z(t)}+2)}{34} + \frac{139}{71}\left[\frac{\sqrt[3]{z(t)}}{(2+t^2)(3+\sqrt[3]{z(t)})} + \frac{t^3}{4} + \sqrt{2}\right] = 0, \\ {}^c\mathbb{D}_q^{\frac{17}{9}}[z](t) + \frac{125(t^3 + \sqrt[3]{y(t)}+2)}{34} + \frac{139}{71}\left[\frac{\sqrt[3]{y(t)}}{(2+t^2)(3+\sqrt[3]{y(t)})} + \frac{t^3}{4} + \sqrt{2}\right] = 0, \end{cases}$$

for $t \in J_0$, with the multipoint boundary conditions ${}^c\mathbb{D}_q^{\frac{2}{9}}[y](1) = \frac{8}{15}{}^c\mathbb{D}_q^{\frac{2}{9}}[y](\frac{1}{5}) + \frac{3}{5}{}^c\mathbb{D}_q^{\frac{2}{9}}[y](\frac{3}{10}) + \frac{4}{7}{}^c\mathbb{D}_q^{\frac{2}{9}}[y](\frac{5}{8})$, ${}^c\mathbb{D}_q^{\frac{2}{9}}[z](1) = \frac{8}{15}{}^c\mathbb{D}_q^{\frac{2}{9}}[z](\frac{1}{5}) + \frac{3}{5}{}^c\mathbb{D}_q^{\frac{2}{9}}[z](\frac{3}{10}) + \frac{4}{7}{}^c\mathbb{D}_q^{\frac{2}{9}}[z](\frac{5}{8})$, and $y(0) = z(0) = 0$. We define $\mathfrak{g}(t, w(t))$ and $\mathfrak{h}(t, w(t))$ by $\mathfrak{g}(t, w) = t^3 + \sqrt[3]{w} + 2$, $\mathfrak{h}(t, w) = \frac{\sqrt[3]{w}}{(2+t^2)(3+\sqrt[3]{w})} + \frac{t^3}{4} + \sqrt{2}$. One can easy show that $\mathfrak{g}(\mathfrak{t}, \mathfrak{w})$, $\mathfrak{h}(t, w)$ are increasing with respect to w, $\mathfrak{h} \geq \sqrt{2} > 0$ whenever $w = 0$. Thus $\sigma - \nu - 1 = \frac{2}{3} \geq 0$ and for $m = 5$, we get $\sum_{i=1}^{m-2} a_i e_i^{\sigma-\nu-1} = \frac{1267}{1458} < 1$. Put $r = 1/3$. Then for $\gamma \in J_0$, $t \in \overline{J}_0$, $w \in \mathbb{R}^+$, we can notice that $\mathfrak{g}(t, \gamma w) \geq \gamma^r \mathfrak{g}(t, w)$, $\mathfrak{h}(t, \gamma w) \geq \gamma^r \mathfrak{h}(t, \gamma w)$. Also, we deduce that $\mathfrak{g}(t, w) \geq \hat{\gamma}_0 \mathfrak{h}(t, w)$, where $\hat{\gamma}_0 = 1 > 0$, for each $t \in \overline{J}_0$, $w \in \mathbb{R}^+$. Thus, (A_1)-(A_3) hold. Theorem 3.4 implies that system (6) has a unique positive solution in $P_h \times P_h$, where $h(t) = t^{\frac{8}{9}}$.

Example 3.8. Consider the system of nonlinear fractional q-differential equations

$$
(7) \qquad \begin{cases} {}^c\mathbb{D}_q^{\frac{15}{7}}[y](t) + \frac{21}{10}[\frac{\cos^2 t + \sqrt[5]{y(t)}}{5(1+\cos^2 t)(1+\sqrt[5]{y(t)})} + t + \sqrt[5]{125}] \\ \quad + \frac{13}{10}[\frac{|\sin t|}{\sqrt{4+|\sin t|+z(t)}} + t^2 + \sqrt{85}] = 0, \\ {}^c\mathbb{D}_q^{\frac{15}{7}}[z](t) + \frac{21}{10}[\frac{\cos^2 t + \sqrt[5]{z(t)}}{5(1+\cos^2 t)(1+\sqrt[5]{z(t)})} + t + \sqrt[5]{125}] \\ \quad + \frac{13}{10}[\frac{|\sin t|}{\sqrt{4+|\sin t|+y(t)}} + t^2 + \sqrt{85}] = 0, \end{cases}
$$

for $t \in J_0$, with the multipoint boundary conditions ${}^c\mathbb{D}_q^{\frac{1}{4}}[y](1) = \frac{1}{5}{}^c\mathbb{D}_q^{\frac{1}{4}}[y](\frac{3}{8}) + \frac{8}{13}{}^c\mathbb{D}_q^{\frac{1}{4}}[y](\frac{2}{3})$, ${}^c\mathbb{D}_q^{\frac{1}{4}}[z](1) = \frac{1}{5}{}^c\mathbb{D}_q^{\frac{1}{4}}[z](\frac{3}{8}) + \frac{8}{13}{}^c\mathbb{D}_q^{\frac{1}{4}}[z](\frac{2}{3})$ and $y(0) = z(0) = 0$. We define $\mathfrak{g}(t, w)$ and $\mathfrak{h}(t, w)$ by $\mathfrak{g}(t, w) = t + \frac{\cos^2 t + \sqrt[5]{w}}{5(1+\cos^2 t)(1+\sqrt[5]{w})} + \sqrt[5]{125}$, $\mathfrak{h}(t, w) = t^2 + \frac{|\sin t|}{(4+|\sin t|+w)^{\frac{1}{2}}} + \sqrt{85}$. Obviously, $\mathfrak{g}$, $\mathfrak{h}$ are increasing with respect to the second argument, $\mathfrak{h}(t, 0) \geq \sqrt{85} > 0$. Thus $\sigma - \nu - 1 = \frac{25}{28} \geq 0$ and for $m = 4$, we get $\sum_{i=1}^{m-2} a_i e_i^{\sigma-\nu-1} = \frac{119}{250} < 1$. Put $r = \frac{1}{5}$, $r_1 = \frac{1}{5}$, $r_2 = \frac{1}{2}$. On can easy check that $\mathfrak{g}$, $\mathfrak{h} \in C[\overline{J}_0 \times \mathbb{R}^+, \mathbb{R}^+]$ and $\mathfrak{g}(t, w)$ is nondecreasing in w and $\mathfrak{h}(t, w)$ is nondecreasing in w, $\mathfrak{g}(t, w) \leq 2 + \sqrt[5]{125}$, $\mathfrak{h}(t, w) \leq 2 + \sqrt{85}$ and $0 < r_1 + r_2 = 7/10$. Then for $\gamma \in J_0$, $t \in \overline{J}_0$, $w \in \mathbb{R}^+$, we can notice that $\mathfrak{g}(t, \gamma w) \geq \gamma^{r_1} \mathfrak{g}(t, w)$, $\mathfrak{h}(t, \gamma w) \leq \gamma^{r_2} \mathfrak{h}(t, \gamma w)$. Then all the conditions of Theorem 3.6 are fulfilled. Consequently, there exist $y_0, z_0 \in P$, and system (7) has exactly one positive solution in $[y_0, z_0] \times [y_0, z_0]$.

References

1. M. H. Annaby and Z. S. Mansour, *q-Fractional Calculus and Equations,* Springer Heidelberg, Cambridge, 2012. DOI: 10.1007/978-3-642-30898-7.
2. M. Bohner and A. Peterson, *Dynamic Equations on Time Scales,* Birkhäuser, Boston, 2001.
3. R. A. C. Ferreira, *Nontrivials solutions for fractional q–difference boundary value problems,* Elect. J. Qualit. Theory Diff. Eq. **70** (2010) 1–101.
4. C. Goodrich and A. C. Peterson, *Discrete Fractional Calculus,* Springer International Publishing, Switzerland, 2015. DOI: 10.1007/978-3-319-25562-0.
5. D. Guo and V. Lakshmikantham, *Nonlinear Problems in Abstract Cone,* Academic Press, Cambridge, MA, USA, 1988.
6. J. Henderson and R. Luca, *On a system of second-order multi-point boundary value problems,* Appl. Math. Lett. **25** (12) (2012) 2089–2094.

7. F. H. Jackson, *q-difference equations*, Amer. J. Math. **32** (1910) 305–314.
8. Z. Lv, *Positive solutions of m-point boundary value problems for fractional differential equations*, Adv. Differ. Equ. **2011** (2011) 571804. DOI:10.1155/2011/571804.
9. S. W. Vong, *Positive solutions of singular fractional differential equations with integral boundary conditions,* Math. Computer Model. **57** (2013) 1053–1059.
10. S. Xu and B. Jia, *Fixed-point theorems of ϕ concave - $(\hat{\gamma})$ - convex mixed monotone operators and applications,* J. Math. Anal. Appl. **295** (2) (2004) 645–657.
11. C. Zhang and D. R. Anderson, *A sum operator equation and applications to nonlinear elastic beam equations and Lane-Emden-Fowler equations,* J. Math. Anal. Appl. **375** (2) (2011) 388–400.

E-mail: mesamei@basu.ac.ir; mesamei@gmail.com
E-mail: f.fasihi2122@gmail.com
E-mail: zanganehhasti@gmail.com

On Weak Specification Property of Semigroup Actions

Zahra Shabani*
Department of Mathematics, Faculty of Mathematics, Statistics and Computer Science, University of Sistan and Baluchestan, Zahedan, Iran

Abstract. In this talk, we introduce the notion of weak specification property of semigroup actions on the compact metric spaces and investigate its relation with pseudo orbital specification and ergodic shadowing properties.
Keywords: Semigroup actions, Ergodic shadowing, Specification property.
AMS Mathematical Subject Classification [2010]: 37B05, 37C50.

1. Introduction

The concept of shadowing was originated from the Anosov closing lemma and because of its rich consequences, shadowing plays an important role in the general qualitative theory of dynamical systems. Another variant of shadowing is the specification property in which one can approximate distinct finite pieces of orbits by an actual orbit with a certain uniformity. It was first introduced by Bowen [3] to study the ergodic property of Axiom A diffeomorphisms. It has been shown that every mapping with the specification property is chaotic in the sense of Devaney; see [1]. The authors in [4, 5], respectively, defined some kind of specification such as weak specification and pseudo-orbital specification properties for a continuous map on a compact metric space and studied their relations with other dynamical properties.

Bahabadi [2] introduced the notions of shadowing and average shadowing properties for free semigroup actions. He obtained that a semigroup with average shadowing property is chain transitive.

Ergodic shadowing and pseudo orbital specification properties for finitely generated semigroup actions were introduced in [6], and it was proved that these properties are equivalent to the semigroup being topologically mixing and having the ordinary shadowing property.

Here, we introduce the definition of weak specification property for semigroup actions and study the connection between this property with ergodic shadowing and pseudo orbital properties. Indeed, we prove the following results.

Theorem 1.1. *Let G be a semigroup generated by the family $\{id, g_1, \dots, g_m\}$ of continuous maps on the compact metric space X, where g_i is surjective for some $i \in \{1, \dots, m\}$. Then the following properties on G are equivalent:*

1) *ergodic shadowing,*
2) *chain mixing and ordinary shadowing,*
3) *topologically mixing and ordinary shadowing,*

*Speaker

4) *weak specification and ordinary shadowing,*
5) *pseudo-orbital specification.*

2. Preliminaries

In this section, we describe the free semigroup actions and state some notations and definitions. Let $\mathbb{N} = \{1, 2, 3, \ldots\}$ and let $\mathbb{Z}^+ = \{0, 1, 2, 3, \ldots\}$. Given a finitely generated semigroup G with a finite set of generators $G_1 = \{id, g_1, g_2, \ldots, g_m\}$, where $g_i : X \to X, i \in \mathcal{P} = \{1, \ldots, m\}$, is a continuous self-map on the compact metric space (X, d). We write $G = \bigcup_{n\in\mathbb{Z}^+} G_n$, where $G_0 = id$ and

$$G_n = \{g_{i_n} \circ \cdots \circ g_{i_2} \circ g_{i_1} : g_{i_j} \in G_1\}.$$

Indeed, G_n consists of elements that are concatenations of at the most n elements of G_1.

Let $\mathbb{F}_m$ be the free semigroup with generators $\{1, \ldots, m\}$. One way to interpret this statement is to consider the itinerary map $\iota : \mathbb{F}_m \to G$ given by

$$w = w_1 w_2 \ldots w_n \to g_w^n = g_{w_1} \circ \cdots \circ g_{w_n},$$

and to regard concatenations on G as images by ι of paths on $\mathbb{F}_m$.

Let $G_1 = \{id, g_1, \ldots, g_m\}$ be a finite collection of continuous maps on the compact metric space X. Th symbolic dynamic is a way to display the elements of semigroup G associated with this family. Let Σ^m be the space of infinite sequences of m symbols $\{1, \ldots, m\}$, that is, $\Sigma^m = \{\omega = \omega_0\omega_1\omega_2 \ldots : \omega_i \in \{1, \ldots, m\}\}$. For any sequence $\omega = \omega_0\omega_1\omega_2 \ldots \in \Sigma^m$, take $g_\omega^0 := id$ and for any $n > 0$, $g_\omega^n(x) := g_{\omega_{n-1}} \circ \cdots \circ g_{\omega_0}(x)$. Let $\mathcal{A}^m$ be a set of finite words of symbols $\{1, \ldots, m\}$, that is, if $w \in \mathcal{A}^m$, then $w = w_0 \ldots w_{l-1}$, where $w_i \in \{1, \ldots, m\}$ for all $i = 0, \ldots, l-1$. Also, for $0 \le i \le l-1$, we denote $g_w^i := g_{w_{i-1}} \circ \cdots \circ g_{w_0}$.

Let (X, d) be a compact metric space and let G be a semigroup associated with the finite family $\{id, g_1, \ldots, g_m\}$ of continuous self maps on X. Given $w = w_0 \ldots w_{n-1} \in \mathcal{A}^m$ and $\epsilon > 0$. An (ϵ, w)-*chain* of semigroup G from x to y is a finite sequence $x_0 = x, x_1, \ldots, x_n = y$ such that

$$d(g_{w_i}(x_i), x_{i+1}) < \epsilon \quad \text{for any } i = 0, \ldots, n-1.$$

We say that G is *chain transitive* if for any $x, y \in X$ and any $\epsilon > 0$, there exists an ϵ-chain from x to y. Also G is called *chain mixing* if for any two points $x, y \in X$ and any $\epsilon > 0$, there is a positive integer N such that for any integer $n \ge N$, there is an ϵ-chain from x to y of length n. We say that the semigroup G is *topologically mixing*, if for any two open subsets U and V of X, there exists an integer $N \in \mathbb{N}$, such that for any $n \ge N$, $g_\omega^n(U) \cap V \neq \emptyset$, for some $\omega \in \Sigma^m$.

For a sequence $\xi = \{x_i\}_{i\ge0}$, $\delta > 0$, and $\omega = \omega_0\omega_1\omega_2 \ldots \in \Sigma^m$, put

$$Np(\xi, G, \omega, \delta) = \{i \in \mathbb{Z}^+ : d(g_{\omega_i}(x_i), x_{i+1}) \ge \delta\},$$

$$Np^c(\xi, G, \omega, \delta) = \mathbb{Z}^+ \setminus Np(\xi, G, \omega, \delta),$$

and

$$Np_n(\xi, G, \omega, \delta) = Np(\xi, G, \omega, \delta) \cap \{0, \ldots, n-1\}.$$

To simplify the notation, we denote them by $Np(\xi,\omega,\delta)$ and $Np_n(\xi,\omega,\delta)$, respectively. Given a sequence $\xi=\{x_i\}_{i\geq 0}$ and a point $z\in X$, consider

$$Ns(\xi,\omega,z,\delta)=\{i\in\mathbb{Z}^+ : d(g_\omega^i(z),x_i)\geq\epsilon\},$$

$$Ns^c(\xi,\omega,z,\delta)=\mathbb{Z}^+\setminus Ns(\xi,\omega,z,\delta),$$

and

$$Ns_n(\xi,\omega,z,\delta)=Ns(\xi,\omega,z,\delta)\cap\{0,\dots,n-1\}.$$

Definition 2.1. Let $\delta>0$ and let $\xi=\{x_i\}_{i\geq 0}\subset X$. We have the following concepts:

(1) ξ is a (δ,ω)-*pseudo orbit* of G for some $\omega=\omega_0\omega_1\ldots\in\Sigma^m$, if for any $i\in\mathbb{Z}^+$, $d(g_{\omega_i}(x_i),x_{i+1})<\delta$, see [2].

(2) ξ is a (δ,ω)-*ergodic pseudo orbit* of G for some $\omega=\omega_0\omega_1\ldots\in\Sigma^m$ provided that the set $Np(\xi,\omega,\delta)$ has zero density (See [6]), that is,

$$\lim_{n\to\infty}\frac{|Np_n(\xi,\omega,\delta)|}{n}=0.$$

Remark 2.2. Clearly, every orbit $\{g_\omega^n(x)\}_{n\geq 0}$ is a (δ,ω)-pseudo orbit, and every (δ,ω)-pseudo orbit is a (δ,ω)-ergodic pseudo orbit. Moreover a (δ,ω)-ergodic pseudo orbit may be represented as

$$\xi=\{x_0,x_1,x_2,\ldots x_{m_1};x_{m_1+1},x_{m_1+2},\ldots,x_{m_2};x_{m_2+1},x_{m_2+2},\ldots\},$$

where $\{x_{m_i+1},x_{m_i+2},\ldots,x_{m_{i+1}}\}, i\in\mathbb{Z}^+$ are finite (δ,w^i)-chains with

$$w^i=\omega_{m_i+1}\omega_{m_i+2}\ldots\omega_{m_{i+1}-1}\in\mathcal{A}^m,\ m_0=-1,$$

and $\{m_i\}_{i\in\mathbb{Z}^+}$ has zero density.

Now, we use the above notions of approximate trajectories to define the various types of shadowing properties.

Definition 2.3.

(1) [2] A semigroup G has the *shadowing property*, provided that for every $\epsilon>0$, there exists $\delta>0$ such that, for any (δ,ω)-pseudo orbit ξ of G, there is a point $z\in X$ such that for any $i\in\mathbb{Z}^+$,

$$d(g_\omega^i(z),x_i)<\epsilon.$$

(2) [6] A semigroup G has the *ergodic shadowing property* if for each $\epsilon>0$ there exists $\delta>0$ such that every (δ,ω)-ergodic pseudo orbit ξ of G can be ϵ-ergodic shadowed by some point z in X, that is, there exists $\varphi\in\Sigma^m$ with $\varphi_i=\omega_i$ for $i\in Np^c(\xi,\omega,\delta)$, such that

$$\lim_{n\to\infty}\frac{|Ns_n(\xi,\varphi,z,\epsilon)|}{n}=0.$$

Now, we introduce the notion of weak specification property for the context of semigroup actions.

Definition 2.4. We say that the semigroup G has *weak specification property* if for any $\epsilon>0$, there exists $N(\epsilon)>0$ such that for any set $\{x_1,\ldots,x_k\}$ of points of X, any sequence of nonnegative integers $a_1<b_1<a_2<b_2<\cdots<a_k<b_k$

with $a_{j+1} - b_j \geq N(\epsilon)$, and any $w^j = w_{a_j} \dots w_{b_j-1} \in \mathcal{A}^m$, $(1 \leq j \leq k)$, there exist a point $z \in X$ and $\omega \in \Sigma^m$ with $\omega_i = w_i$ for any $a_j \leq i \leq b_j - 1$ such that

$$d(g^i_\omega(z), g^{i-a_j}_{w^j}(x_j)) < \epsilon, \quad \text{for any} \quad a_j \leq i \leq b_j, \quad 1 \leq j \leq k.$$

In the following, we recall the stronger notion of specification for the semigroup G, which is called pseudo-orbital specification property and is equivalent to ergodic shadowing property (See [6]).

Definition 2.5. We say that a semigroup G has the *pseudo-orbital specification property* if for any $\epsilon > 0$, there exist $\delta = \delta(\epsilon) > 0$ and $N(\epsilon) > 0$ such that for any nonnegative integers $a_1 < b_1 < a_2 < b_2 < \dots < a_k < b_k$ with $a_{j+1} - b_j \geq N(\epsilon)$ and (δ, w^j)-pseudo orbits ξ_j with $\xi_j = \{x_{(j,i)}\}, a_j \leq i \leq b_j$, $1 \leq j \leq k$, and $w^j = w^j_{a_j} \dots w^j_{b_j-1} \in \mathcal{A}^m$, there exist a point $z \in X$ and $\omega \in \Sigma^m$ with $\omega_i = w^j_i$ for $a_j \leq i \leq b_j - 1$ and $1 \leq j \leq k$, such that

$$d(g^i_\omega(z), x_{(j,i)}) < \epsilon, \quad \text{for any} \quad a_j \leq i \leq b_j, \quad 1 \leq j \leq k.$$

It is clear from the definition that the pseudo-orbital specification property implies the weak specification property.

3. Weak Specification Property of Semigroup Action

In this section we show that if a semigroup G has the shadowing property, then weak specification and pseudo-orbital specification properties are equivalent. It is clear from the definition that any semigroup with the pseudo-orbital specification property, has the weak specification property. In the following, we present a finitely generated semigroup action with the weak specification property, that does not have the pseudo-orbital specification property.

Example 3.1. Let $X = S^1$, let g_1 be any C^1-expanding map on X and $g_2 := R_\alpha : X \to X$ be the rotation of the angle α. Let G be a semigroup with generating set $\{id, g_1, g_2\}$. The semigroup G does not have the shadowing property, as g_2 does not have not the shadowing property. Therefore it does not have the pseudo orbital specification property (since the pseudo-orbital specification property implies shadowing [6, Theorem 1.1]). We show that G has the weak specification property. Since g_1 is an expanding map by [7, Lemma 11.2.7], There exists $\epsilon_0 > 0$ such that for any $\epsilon \leq \epsilon_0$, any $x \in X$, and any $n \in \mathbb{N}$, $g^n_1(B(x, n, \epsilon)) = B(g^n_1(x), \epsilon)$, where

$$B(x, n, \epsilon) := \{y \in X, \max_{0 \leq i \leq n} d(g^i_1(x), g^i_1(y)) < \epsilon\}.$$

Also for any $\epsilon > 0$, there exists $N = N(\epsilon)$ such that $g^N_1(B(x, \epsilon)) = S^1$ for any $x \in X$. By this observation, for given $\epsilon > 0$, any set $\{x_1, \dots, x_k\}$ of points of X, any sequence of nonnegative integers $a_1 < b_1 < a_2 < b_2 < \dots < a_k < b_k$ with $a_{j+1} - b_j \geq N(\epsilon)$, and any $w^j = w_{a_j} \dots w_{b_j-1} \in \mathcal{A}^m$, $(1 \leq j \leq k)$, we can find a point $z \in X$ such that for $\omega \in \Sigma^m$ with

$$\omega_i := \begin{cases} w_i, & i \in [a_j, b_j - 1], \\ 1, & i \in \mathbb{Z}^+ \setminus [a_j, b_j - 1], \end{cases}$$

we have

$$d(g_\omega^i(z), g_{w^j}^{i-a_j}(x_j)) < \epsilon, \quad \text{for any} \quad a_j \leq i \leq b_j, \quad 1 \leq j \leq k.$$

Here, we shall show that by assuming the shadowing property for finitely generated semigroup G, weak specification property implies the pseudo-orbital specification property.

THEOREM 3.2. *If a semigroup G associated with the family of continuous self maps $\{id, g_1, \ldots, g_m\}$ satisfies that g_i is surjective for some $i \in \{1, \ldots, m\}$ and has the shadowing and weak specification properties, then it has the pseudo-orbital specification property.*

Using Theorem 3.2 and [6, Theorem 1.1], we obtain Theorem 1.1.

References

1. N. Aoki and K. Hiraide, *Topological Theory of Dynamical Systems*, North-Holland Math. Library, Vol. 52, North-Holland, Amsterdam, 1994.
2. A. Z. Bahabadi, *Shadowing and average shadowing properties for iterated function systems*, Georgian Math. J. **22** (2) (2015) 179–184.
3. R. Bowen, *Equilibrium states and ergodic theory of Anosov diffieomorphisms*, Trans. Amer. Math. Soc. **154** (1971) 377–397.
4. A. Fakhari and F. H. Ghane, *On shadowing: Ordinary and ergodic*, J. Math. Anal. Appl. **364** (1) (2010) 151–155.
5. D. Kwietniak and P. Oprocha, *A note on the average shadowing property for expansive maps*, Topology Appl. **159** (1) (2012) 19–27.
6. Z. Shabani, *Ergodic shadowing of semigroup actions*, Bull. Iranian Math. Soc. **46** (2) (2020) 303–321.
7. M. Viana and K. Oliveira, *Foundations of Ergodic Theory*, Cambridge Studies in Advanced Mathematics, 151. Cambridge University Press, Cambridge, 2016.

E-mail: zshabani@math.usb.ac.ir

$\sigma_{2,p}$-Energy Functional and Polyconvexity

Mohammad Sadegh Shahrokhi-Dehkordi
Faculty of Mathematical Sciences, University of Shahid Beheshti, Tehran, Iran
and Mojgan Taghavi*
Faculty of Mathematical Sciences, University of Shahid Beheshti, Tehran, Iran

Abstract. A class of maps reffered to as generalised twists is introduced and the system of Euler-Lagrange equations for the energy functional with polyconvex integrand over the n-dim- ensional annulus domain, based on them is presented. Further, the existence of the weak solu- tion of the Euler-Lagrange equations on the homotopy classes is investigated.
Keywords: Gnereralised twists, Euler-Lagrange equation, Polyconvex.
AMS Mathematical Subject Classification [2010]: 70S20, 58Exx.

1. Introduction

In this paper we study the energy functional of the form

$$\mathbb{E}_{\sigma_{2,p}}[u,\Omega] := \int_\Omega \left[\frac{1}{p}\sigma_2^{\frac{p}{2}}(u) + \Phi(\det \nabla u)\right] dx,$$

over the space

$$\mathcal{A}(\Omega) = \{u \in W^{1,2p}(\Omega,\mathbb{R}^n) : \det \nabla u > 0 \, for \, \mathcal{L}^n - a.e, u\big|_{\partial\Omega} = x\}.$$

Here Ω is a n-dimensional annulus, i.e., $\Omega = \{x \in \mathbb{R}^n : a < |x| < b\}$ $with$ $0 < a < b < \infty$, $2 \leq p \leq \infty$ and $\sigma_2(u) = |\wedge^2 \nabla u|^2 = \sum_{1\leq i<j\leq n} \lambda_i^2\lambda_j^2$. Notice that $0 \leq \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ are singular values of the ∇u and the convex function Φ holds in the following assumptions:

[**H1**] $\Phi : (0,\infty) \to (0,\infty)$,
[**H2**] $\Phi \in C^2(0,\infty)$,
[**H3**] $\lim_{t\downarrow 0} \Phi(t) = \lim_{t\uparrow\infty} \frac{\Phi(t)}{t} = \infty$.

By using Green's deformation tensor (See [3]), the energy functional $\mathbb{E}$ can be written as

$$\begin{aligned} \mathbb{E}_{\sigma_2,p}[u,\Omega] &= \int_\Omega \left[\frac{1}{p}\,\sigma_2^{\frac{p}{2}}(u) + \Phi(\det \nabla u)\right] dx \\ &= \int_\Omega \left[\frac{1}{p\,2^{\frac{p}{2}}}\left(\left|\nabla u\right|^4 - \left|(\nabla u)\,(\nabla u)^t\right|^2\right)^{\frac{p}{2}} + \Phi(\det \nabla u)\right] dx. \end{aligned} \tag{1}$$

*Speaker

The system of Euler-Lagrange equations relating to the above energy functional over the space $\mathcal{A}(\mathbb{X})$ as follows:

$$\begin{cases} \operatorname{div}\mathfrak{D}[\nabla u(x)] = 0, & x \in \Omega, \\ \det \nabla u(x) > 0, & x \in \Omega, \\ u(x) = x, & x \in \partial\Omega. \end{cases}$$

Noticing that the divergence operator acts row-wise and the tensor field $\mathfrak{D}$ is defined by

$$\mathfrak{D}(\xi) := \frac{1}{2^{\frac{p-2}{2}}}\left(|\xi|^4 - |\xi\,\xi^t|^2\right)^{\frac{p-2}{2}}\left[|\xi|^2\xi - \xi\,\xi^t\,\xi\right] + \Phi'(\det\xi)\mathrm{cof}\xi,$$

for all $\xi \in \mathbb{M}_{n\times n}$ satisfying $\det\xi > 0$.

A motivation for the study of such problems is nonlinear elasticity, where $\mathbb{E}_{\sigma 2,p}$ denotes the elastic energy of a homogeneous hyperelastic material and $\mathcal{A}(\Omega)$ denotes the space of orientation preserving deformations of Ω fixing the boundary pointwise (See, e.g., [1, 2]).

Now we introduce a class of maps as generalized twists, which described as

$$u(x) = G(r)\,\theta,$$

where $G(r) = f(r)\,\mathbf{Q}(r)$, $r = |x|$, $\theta = \dfrac{x}{|x|}$, $\mathbf{Q} \in C\big([a,b],\mathbf{SO}(n)\big)$ and $f \in C[a,b]$. For $\mathbf{Q}$, f to be in the $\mathcal{A}(\Omega)$, we need to consider condition L^2-summability on $\dot{f} = \frac{d}{dr}f$, $\dot{\mathbf{Q}} = \frac{d}{dr}\mathbf{Q}$ and also $\mathbf{Q}(a) = \mathbf{Q}(b) = \mathbf{I}_n$, $f(a) = a$, $f(b) = b$, $\dot{f} > 0$ on (a,b).

2. Main Results

In this section, based on the generalized twists, we introduce the system of Euler-Lagrange equations for σ_2-energy functional.

Proposition 2.1. *Let u be a generalized twist. Then the σ_2-energy functional (1) can be expressed in the form of*

$$\mathbb{G}_{\sigma_2,p}[\mathbf{Q},f] := \int_a^b \left[\mathbf{G}(r,f,\dot{f},\dot{\mathbf{Q}}) + n\omega_n\,\Phi\left(\dot{f}\left(\frac{f}{r}\right)^{n-1}\right)\right] r^{n-1}\,dr,$$

where the integrand itself is given through an integral over the unit sphere as follows:

$$\mathbf{G}(r,f,\eta,\xi) := \int_{\mathbb{S}^{n-1}} \frac{1}{p\,2^{\frac{p}{2}}}\Bigg[(n-1)(n-2)\left(\frac{f}{r}\right)^4 + 2(n-1)\left(\frac{f\eta}{r}\right)^2 + 2(n-2)\frac{f^4}{r^2}|\xi\theta|^2\Bigg]^{\frac{p}{2}} d\mathcal{H}^{n-1}(\theta),$$

over the space of admissible maps

$$\mathcal{G} = \mathcal{G}(a,b) := (\mathbf{Q},f) : \left\{\begin{array}{l} \mathbf{Q} \in W^{1,p}([a,b],\mathbf{SO}(n)), \\ \mathbf{Q}(a) = \mathbf{Q}(b) = \mathbf{I}_n, \\ f \in W^{1,p}[a,b], \\ \dot{f} > 0 \ \mathcal{L}^1\text{-a.e. on } (a,b), \\ f(a) = a, f(b) = b. \end{array}\right\},$$

PROOF. By using the fact that $u(x) = \mathbf{Q}(r)\, f(r)\theta$, we have

$$\nabla u = \big(\dot f - \frac{f}{r}\big)\, \mathbf{Q}\theta \otimes \theta + f\dot{\mathbf{Q}}\theta \otimes \theta + \frac{f}{r}\mathbf{Q}.$$

After some calculations, it is easy to obtain that

$$\begin{aligned} \big|\nabla u\big|^2 &= tr\big[(\nabla u)\,(\nabla u)^t\big] \\ &= (n-1)\bigg(\frac{f}{r}\bigg)^2 + \dot f^2 + f^2\big|\dot{\mathbf{Q}}\theta\big|^2, \end{aligned}$$

and

$$\det(\nabla u) = \det\left[\frac{f}{r}\mathbf{Q} + \left(\dot f - \frac{f}{r}\right)\mathbf{Q}\theta \otimes \theta + f\dot{\mathbf{Q}}\theta \otimes \theta\right] = \dot f\bigg(\frac{f}{r}\bigg)^{n-1}.$$

Substituting above relation into (1), the proof can be concluded. □

PROPOSITION 2.2. *Consider* $(\mathbf{Q}, f) \in \mathcal{G}$ *such that* $\mathbf{Q} \in C^2\Big((a,b), \mathbf{SO}(n)\Big)$, $f \in C^2(a,b)$, $\dot f > 0$ *on* (a,b) *and* $\mathbb{G}_{\sigma_2,p}[\mathbf{Q}, f] < \infty$. *Then the system of Euler-Lagrange equations for* $\mathbb{G}_{\sigma_2,p}[\cdot,\cdot]$ *over the* $\mathcal{G}$ *has the form as follows:*

i) *when* $n = 2$

$$\frac{d}{dr}\left[\frac{f^p\, \dot f^{p-1}}{r^{p-1}} + \Phi'\, f\right] = \frac{f^{p-1}\, \dot f^p}{r^{p-1}} + \Phi'\, \dot f,$$

ii) *when* $n \geq 3$

$$\begin{cases} \dfrac{d}{dr}\big[\mathbf{G}_\eta r^{p-1} + n\omega_n f^{n-1}\Phi'\big] = r^{n-1}\mathbf{G}_f + n(n-1)\omega_n \dot f\, f^{n-2}\Phi', \\ \dfrac{d}{dr}\big[r^{n-1}\, \mathbf{G}_\xi \mathbf{Q}^t - r^{n-1}\mathbf{Q}\mathbf{G}_\xi^t\big] = 0, \end{cases}$$

where $\Phi' = \Phi'\bigg(\dot f\, \Big(\frac{f}{r}\Big)^{n-1}\bigg)$.

PROOF. First we prove the second part. Let $f \in \mathcal{G}$. For every $\varepsilon \in \mathbb{R}$ and $\varphi \in C^\infty(a,b)$ with $\varphi(a) = \varphi(b) = 0$, set $f_\varepsilon = f + \varepsilon\varphi$. According to the assumption $f \in C^2(a,b)$, then $\dot f$ is a continuous function on (a,b). Since $supp\, \varphi \subset (a,b)$ is a compact set, there exists $c > 0$ such that $\dot f \geq c > 0$ on $supp\, \varphi$. Therefor, $\dot f_\varepsilon > 0$ when $|\varepsilon| \times \sup\limits_{(a,b)} |\dot\varphi| < c$ and the pair $(\mathbf{Q}, f_\varepsilon) \in \mathcal{G}$. Moreover by selecting ε smaller, we obtain

$$\begin{aligned} 0 &= \frac{d}{d\varepsilon}\mathbb{G}_{\sigma_2,p}[\mathbf{Q}, f_\varepsilon]\Big|_{\varepsilon=0} \\ &= \frac{d}{d\varepsilon}\left[\int_a^b \left[\mathbf{G}(r, f_\varepsilon, \dot f_\varepsilon, \dot{\mathbf{Q}}) + n\omega_n \Phi\bigg(\dot f_\varepsilon \Big(\frac{f_\varepsilon}{r}\Big)^{n-1}\bigg)\right] r\, dr\right]\Big|_{\varepsilon=0} \\ &= \int_a^b \bigg(\big[\mathbf{G}_f(r, f, \dot f, \dot{\mathbf{Q}})r^{n-1} + n(n-1)\omega_n f^{n-2}\dot f\Phi'\big]\bigg)\varphi + \\ &\qquad\qquad \bigg(\big[r^{n-1}\mathbf{G}_\eta(r, f, \dot f, \dot{\mathbf{Q}}) + n\omega_n f^{n-1}\Phi'\big]\bigg)\dot\varphi dr. \end{aligned}$$

Based on the integration by parts formula we get

$$\begin{aligned} 0 &= \int_a^b \Big(\big[\mathbf{G}_f(r,f,\dot{f},\dot{\mathbf{Q}})r^{n-1} + n(n-1)\omega_n f^{n-2}\dot{f}\Phi'\big] \\ &- \frac{d}{dr}\big[r^{n-1}\mathbf{G}_\eta(r,f,\dot{f},\dot{\mathbf{Q}}) + n\omega_n f^{n-1}\Phi'\big]\Big)\varphi\, dr. \end{aligned} \tag{2}$$

The (2) holds for all φ, then we have

$$\frac{d}{dr}\left[\mathbf{G}_\eta r^{p-1} + n\omega_n f^{n-1}\Phi'\right] = r^{n-1}\mathbf{G}_f + n(n-1)\omega_n \dot{f}\, f^{n-2}\Phi'. \tag{3}$$

□

For every fixed $\mathbf{Q} \in C^2(a,b)$, $\varepsilon \in \mathbb{R}$ and variation matrix $\mathbf{H} \in C_0^\infty\big((a,b),\mathbb{M}_{n\times n}\big)$, we set $\mathbf{Q}_\varepsilon = \mathbf{Q} + \varepsilon\,\mathbf{H}$. By using [5, Proposition 3.1], there is an arbitrary matrix $\mathbf{F} \in C_0^\infty\big((a,b),\mathbb{M}_{n\times n}\big)$ such that $\mathbf{H} := \mathbf{Q}(\mathbf{F}-\mathbf{F})$. By simple manipulation, it is derived that $\mathbf{Q}_\varepsilon \in \mathcal{G}$, hence we can write

$$\begin{aligned} 0 &= \frac{d}{d\varepsilon}\mathbb{G}_{\sigma 2,p}[\mathbf{Q}_\varepsilon, f]\Big|_{\varepsilon=0} = \int_a^b \Big(\mathbf{G}_\xi(r,f,\dot{f},\dot{\mathbf{Q}}_\varepsilon) : \frac{d}{d\varepsilon}\dot{\mathbf{Q}}_\varepsilon\Big)\Big|_{\varepsilon=0} r^{n-1}\, dr \\ &= \int_a^b \Big[\mathbf{G}_\xi(r,f,\dot{f},\dot{\mathbf{Q}}) : \big(\dot{\mathbf{F}} - \dot{\mathbf{F}}^t\big)\,\mathbf{Q} + \varepsilon\big(\mathbf{F}-\mathbf{F}^t\big)\dot{\mathbf{Q}}\Big] r^{n-1}\, dr \\ &:= \mathbf{I} + \mathbf{II}. \end{aligned}$$

In the following, we shall derive the terms **I** and **II**.

$$\begin{aligned} \mathbf{I} &= \int_a^b \big[\mathbf{G}_\xi(r,f,\dot{f},\dot{\mathbf{Q}}) : \big(\dot{\mathbf{F}} - \dot{\mathbf{F}}^t\big)\,\mathbf{Q}\big] r^{n-1} d \\ &= \int_a^b \big[r^{n-1}\mathbf{G}_\xi(r,f,\dot{f},\dot{\mathbf{Q}})\mathbf{Q}^t : \big(\dot{\mathbf{F}} - \dot{\mathbf{F}}^t\big)\big]\, dr. \end{aligned}$$

Now by utilizing integration by parts, it can be observed that

$$\mathbf{I} = -\int_a^b \left[\frac{d}{dr}\Big(r^{n-1}\mathbf{G}_\xi(r,f,\dot{f},\dot{\mathbf{Q}})\mathbf{Q}^t\Big) : \big(\dot{\mathbf{F}} - \dot{\mathbf{F}}^t\big)\right] dr.$$

For the second term, we evaluate $\mathbf{G}_\xi$ as follows

$$\mathbf{G}_\xi(r,f,\dot{f},\dot{\mathbf{Q}}) = \int_{\mathbb{S}^{n-1}} \big[(n-2)\, S\, \frac{f^4}{r^2}\dot{\mathbf{Q}}\theta\otimes\theta\big] d\mathcal{H}^{n-1}(\theta),$$

where

$$S := \frac{1}{2^{\frac{p-2}{2}}}\left[(n-1)(n-2)\left(\frac{f}{r}\right)^4 + 2(n-1)\left(\frac{f\,\dot{f}}{r}\right)^2 + 2(n-2)\frac{f^4}{r^2}\big|\dot{\mathbf{Q}}\theta\big|^2\right]^{\frac{p-2}{2}}.$$

From above relation, the second term is given by

$$\begin{aligned} \mathbf{II} &= \int_a^b \int_{\mathbb{S}^{n-1}} \Big[(n-2) f^4\, r^{n-3}\, S\langle \dot{\mathbf{Q}}\theta\otimes\theta, \big(\mathbf{F}-\mathbf{F}^t\big)\,\dot{\mathbf{Q}}\rangle\Big] d\mathcal{H}^{n-1}(\theta) \\ &= \int_a^b \int_{\mathbb{S}^{n-1}} \Big[(n-2) f^4\, r^{n-3}\, S\langle \dot{\mathbf{Q}}\theta, \big(\mathbf{F}-\mathbf{F}^t\big)\,\dot{\mathbf{Q}}\theta\rangle\Big] d\mathcal{H}^{n-1}(\theta) \\ &= 0. \end{aligned}$$

Since matrix $(\mathbf{F}-\mathbf{F}^t)$ is a skew-symmetric, so $\langle \dot{\mathbf{Q}}\theta, (\mathbf{F}-\mathbf{F}^t)\,\dot{\mathbf{Q}}\theta\rangle = 0$. Thus, summarising, we have

$$\int_a^b \left(\frac{d}{dr}\big[r^{n-1}\,\mathbf{G}_\xi(r,f,\dot{f},\dot{\mathbf{Q}})\mathbf{Q}^t : (\mathbf{F}-\mathbf{F}^t)\right) dr = 0. \tag{4}$$

The (4) holds for all skew-symmetric $(\mathbf{F}-\mathbf{F}^t) \in C_0^\infty((a,b),\mathbb{M}_{n\times n})$, then we get

$$\frac{d}{dr}\big[r^{n-1}\mathbf{G}_\xi\mathbf{Q}^t - r^{n-1}\mathbf{Q}\mathbf{G}_\xi^t\big] = 0.$$

To prove the first part, put $n=2$ in the Eq. (3).

REMARK 2.3. When $n=2$, the only solution of the Euler-Lagrange equation is $f=r$ (See [4]).

THEOREM 2.4. (Existence of σ_2-energy minimizing loops) *Let $n\geq 3$ and consider the energy functional $\mathbb{G}_{\sigma 2,p}$ over the space $\mathcal{G}$. Then, for each $\alpha\in\mathbb{Z}_2=\{0,1\}$ there exists pair $(\mathbf{Q}_\alpha, f_\alpha)\in \mathfrak{c}_\alpha[\mathcal{G}]$ such that*

$$\mathbb{G}_{\sigma_2,p}[\mathbf{Q}_\alpha, f_\alpha] = \inf_{\mathfrak{c}_\alpha[\mathcal{G}]} \mathbb{G}_{\sigma_2,p}.$$

PROOF. The proof here follows by using an adaptaion of the argument form [5] and hence will be abbreviated. □

References

1. J. M. Ball, *Convexity conditions and existence theorems in nonlinear elasticity*, Arch. Rational Mech. Anal. **63** (4) (1976) 337–403.
2. J. M. Ball, *Discontinuous equilibrium solutions and cavitation in nonlinear elasticity*, Philos. Trans. Roy. Soc. London Ser. A **306** (1496) (1982) 557–611.
3. J. Eells, *Certain Variational Principles in Riemannian Geometry*, Notes in Math., 131, Pitman, Boston, MA, 1985.
4. M. S. Shahrokhi-Dehkordi and J. Shaffaf, *$\sigma 2$-Energy as a polyconvex functional on a space of self-maps of annuli in the multi-dimensional calculus of variations*, Nonlinear Differ. Equ. Appl. **23** (2) (2016). DOI:10.1007/s00030-016-0372-3
5. M. S. Shahrokhi-Dehkordi and A. Taheri, *Polyconvexity, generalized twists and energy minimizers on a space of self-maps of annuli in the multi-dimensional calculus of variations*, Adv. Calc. Var. **2** (4) (2009) 361–396.

E-mail: M_Shahrokhi@sbu.ac.ir
E-mail: mo_taghavi@sbu.ac.ir

Contributed Talks

Interdisciplinary Mathematics

Approximate Solution of Tumor Growth Model with Cancer Stem Cells

Ghazale Aliasghari*
Department of Mathematics, Shahid Rajaee University, Tehran, Iran
and Hamid Mesgarani
Department of Mathematics, Shahid Rajaee University, Tehran, Iran

Abstract. In this paper, we investigate the reaction-diffusion system of integro-partial differential equations describing tumor growth with cancer stem cells(CSCs). We show the existence of the solution for this problem and numerical simulations confirm the evidence of tumor growth paradox, which indicates that, accelerated tumor growth with increased the death rate of cancer cells(CCs).
Keywords: Mathematical modeling of tumors, Integro-partial differential equations, Tumor growth paradox, Cancer stem cell.
AMS Mathematical Subject Classification [2010]: 93A30, 45Kxx.

1. Introduction

Mathematical modeling of tumor growth is an efficient tool to understand, predict, and improve the outcome of cancer treatment. In recent years, many papers have been devoted to analysis of these mathematical models describing the growth of tumors in presence of cancer stem cells. For instance [1, 2] and [4]. Hillen et al. proposed the following reaction-diffusion system of integro-partial differential equations [4]

$$
\begin{aligned}
(1) \qquad u_t(x,t) &= D_u\Delta u + \gamma\delta\int_\Omega K\big(x,y,p(x,t)\big)u(y,t)dy,\\
v_t(x,t) &= D_v\Delta v + (1-\delta)\gamma\int_\Omega K\big(x,y,p(x,t)\big)u(y,t)dy\\
&+ \rho\int_\Omega K\big(x,y,p(x,t)\big)v(y,t)dy - \alpha v(x,t).
\end{aligned}
$$

with initial conditions

$$u(x,0)=u_0(x)\in[0,1],\quad v(x,0)=v_0(x)\in[0,1],\quad u_0+v_0\le 1,\quad x\in\Omega.$$

Here $\Omega\subset\mathbb{R}^n$ is the domain and $x,y\in\Omega$. This model describes the evolution and distribution of $u(x,t), v(x,t)$ and $p(x,t)=u(x,t)+v(x,t)$, which indicate the density of CCs, CSCs and total tumor cell respectively. The further assumptions as follows:

(a) The function $K=K(x,y,p)$ is an integral kernel showing the probability density of the cell located at y generates a cell at x and redistributes cells

*Speaker

only within domain Ω, hence $K(x, y, p)$ is equal to 0 for all $x \notin \Omega$. It can be written as $K\big(x, y, p(x,t)\big) = F(p(x,t))K(x,y)$, where $K(x,y) \geq 0$ and $K \in C(\Omega \times \Omega)$. $F \in C1$ is a continuous, nonnegative, nonincreasing and Lipschitz function on $[0,1]$ with $F(0) = 1$, $F(1) = 0$. We can assume that $K(x,y)$ is a radial function; i.e, $K(x,y) = K(|x-y|)$, where $x, y \in \Omega$. An example of this function can be

$$K(x,y) = \frac{1}{\sqrt{\pi}\sigma} \exp(-\frac{(x-y)^2}{\sigma^2}).$$

(b) $D_u \geq 0$ and $D_v \geq 0$ are the diffusion coefficients of CCs and CSCs.
(c) $\gamma > 0$, $\rho > 0$ are positive constants representing the number of cycles per unit time for CSCs and CCs, respectively.
(d) $\delta \in [0,1]$ is the fraction of symmetric divisions of CSCs. If $\delta \to 0$ it divides into one CSC and one normal CC, whereas if $\delta \to 1$ the CSC divides into two CSCs. In most cases $\delta \ll 1$.

We intend to approximate (1) to form a system of reaction-diffusion equations. Then, we discuss the existence of the solution and finally we will give some numerical examples which confirm the evidence of tumor growth paradox.

2. The Reaction-Diffusion Model

Consider the integral appearing in (1) in the one-dimensional case and and approximate $u(y,t)$ till the second order using Taylor expansion of $u \in C^2(\Omega)$ as it was done in [2] to approximate r.h.s of (1) as follows:

$$\int_{-\infty}^{+\infty} K(|x-y|)F\big(p(x,t)\big)u(y,t)dy = F\big(p(x,t)\big)\int_{-\infty}^{+\infty} K(|x-y|)[u(x,t) + u_x(x,t)(y-x)]dy$$
$$+ F\big(p(x,t)\big)\int_{-\infty}^{+\infty} K(|x-y|)[\frac{1}{2}u_{xx}(x,t)(y-x)^2 + o(3)]dy.$$

Since K is symmetric, the first order moment cancels. A is the integral of K and B is half of its second order moment. We consider $\Omega = (-r, r)$, where $r = |y-x|$ and taking $r \to \infty$. Thus, the system (1) is rewritten as

$$\begin{aligned} u_t &= D_u u_{xx} + \gamma\delta F(p)(Au + Bu_{xx}), \\ v_t &= D_v v_{xx} + (1-\delta)\gamma F(p)(Au + Bu_{xx}) \\ &+ \rho F(p)(Av + Bv_{xx}) - \alpha v, \quad x \in \mathbb{R},\ t > 0. \end{aligned} \tag{2}$$

with initial condition of problem (1).

2.1. Existence Proof. Borsi et al. [1] proved the existence, uniqueness and boundedness of the local and global solutions. They transformed the existence problem of solutions for the nonlinear integro-differential system. Fasano et al. [2] explained the following theorem to prove the existence of the solution, too.

Also, L. Maddalena [5] analyzed the nonlinear system of integro-differential equations. She proved that an invariant limited set exists in the positive cone and this gives positivity and global existence of solutions.

Remark 2.1. It should be noted that we will demonstrate the local existence of solutions in the one-dimensional model (2). In the case of this model, if we compute the diffusion coefficients D_u and D_v we see that it depends on the concentration of cells, which is determined by p, and this goes to zero when $p \to 1$.

We can include the diffusive terms in the spirit of approximation in coefficient B. Because of this, we consider $D = 0$ in (2).

THEOREM 2.2. *Let $u_0, v_0 \in C^{2+\alpha}$ such that, for all $x \in \mathbb{R}$, $p_0(x) = u_0(x) + v_0(x) \leq 1 - M$, $M \in (0,1)$. Then the system (2) has a unique solution (u, v) in the interval $(0, T^\star)$ that satisfies $u + v < 1 - N$ for $0 < N < M$ in the region $\mathbb{R}[0, T^\star)$.*

DEFINITION 2.3. The Hölder space $C^{k,\gamma}(\bar{\Omega})$ consists of all functions $u \in C^k(\bar{\Omega})$ for which the norm

$$\|u\|_{C^{k,\gamma(\bar{\Omega})}} = \sum_{|\alpha| \leq k} \|D^\alpha\|_{C(\bar{U})} + \sum_{|\alpha|=k} [D^\alpha u]_{C^{0,\gamma(\bar{U})}},$$

is finite.

PROOF. We use the fixed point theorem to prove this claim. For some $T > 0$, we define the set $\sum = \{(u, v) \in H^{\alpha, \frac{\alpha}{2}}(\mathbb{R} \times (0, T))^2 : u(x, 0) = u_0(x), v(x, 0) = v_0(x),\ x \in \mathbb{R}\}$. Suppose $u + v < 1 - N$ for $0 < N < M$ such that $\|u\|^\alpha, \|v\|^\alpha < k$ for some $k > 0$. $\sum$ is the set of initial conditions that applies to the space $H^{\alpha, \frac{\alpha}{2}}(\mathbb{R} \times (0, T))^2$ for some $T > 0$. Now we take $(u, v) \in \sum$ fixed and solve the equations of the system (2), so we have

$$\begin{aligned} (3)\quad U_t &= \gamma\delta F(u+v)(AU + BU_{xx}), \\ V_t &= \gamma(1-\delta)F(u+v)(AU + BU_{xx}) + \rho F(u+v)(AV + BV_{xx}) - \alpha V. \end{aligned}$$

for $x \in \mathbb{R}, t > 0$, where $U(x, 0) = u_0(x), V(x, 0) = v_0(x)$. Equation (3) with it's condition can be considered as Cauchy problem that would be solved independently, provided that $U \in C^{2+\alpha}$ for all $x \in \mathbb{R}$ and $t > 0$, since $u_0(x) \in H^\alpha(\mathbb{R})$ [2]. Since, U is Hölder continuous with exponent α, the norm $\|U\|^\alpha \leq N\|u_0\|^\alpha$. Thus, we've found U in (3).

Now, we consider the function $Y(t) = \|u_0\|e^{\gamma\delta At}$ and $\omega(x, t) = Y(t) - U(x, t)$. We notice that,

$$\begin{aligned} \omega_t = Y'(t) - U(t) &= \gamma\delta A\|u_0\|e^{\gamma\delta At} - \gamma\delta F(u+v)[AU + BU_{xx}] \\ &= \gamma\delta Y(t) - \gamma\delta F(u+v)[AU + BU_{xx}]. \end{aligned}$$

As if $\omega_{xx} = -U_{xx}$, so

$$\omega_t - \gamma\delta F(u+v)[A\omega + B\omega_{xx}] = \gamma\delta A[1 - F(u+v)]Y > 0.$$

Then, we see that the operator $L = \frac{\partial}{\partial t} - \gamma\delta F(u+v)[A + B\frac{\partial}{\partial x^2}]$ is parabolic in $\mathbb{R} \times (0, T)$. Therefore, applying [3, Theorem 5] concludes that $\omega > 0$. This implies that, $Y - U > 0$ and $U \leq \|u_0\|e^{\gamma\delta At}$. Thus, we can choose $T^\star$ and $\epsilon_1 > 0$ small enough, such that $U \leq \|u_0\| + \epsilon_1$. □

3. Numerical Examples

We apply numerical simulation to represent that this model shows the tumor growth paradox i.e, a larger death rate of CC lead to a larger tumor. This effect was already found by Hillen [4] for ODE model, here we confirm that, this paradox also exist in (1) formulation.

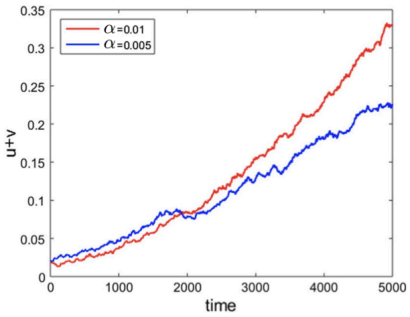


FIGURE 1. Density of CCs and CSCs as function of time for different values of mortalities.

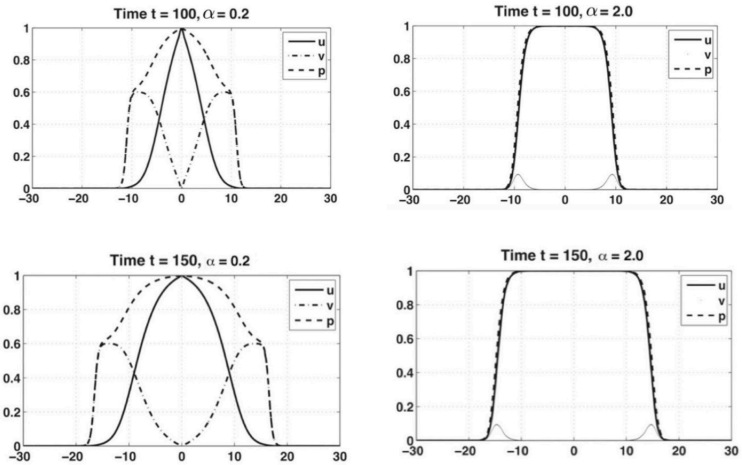


FIGURE 2. Plot of $u, v, p = u + v$ at selected times for different values of CC's death rate(α).

To solve system (1) we used finite difference scheme, with an explicit forward method in time. The r.h.s integrals approximated by the trapezoidal rule. The initial values and parameters set as follows:

$$u_0(x) = \exp(-10x^2), \quad v_0(x) = 0, \quad x \in (-30, 30),$$
$$\delta = 0.2, \quad \sigma_u = 0.5, \quad \sigma_v = 0.1, \quad \gamma = 1, \quad \rho = 0.5.$$

Figure 1 shows the results of two simulations corresponding to different mortality of CC. In particular, the two cases considered are with $\alpha = 0.005$ (blue line) and $\alpha = 0.01$ (red line).

Figure 2 shows that the distribution of u, v, p at selected times (T=100, T=150) for different value of α. For small value of α ($\alpha = 0.2$) we can see that the behavior is quite different for large α (α=2), where non-stem cancer cells play a minor role. As a result, a higher death rate for CCs leaves more space for the invasion of CSCs.

References

1. I. Borsi, A. Fasano, M. Primicerio and T. Hillen, *A non-local model for cancer stem cells and the tumour growth paradox*, Math. Med. Biol. **34** (2017) 59–75.
2. A. Fasano, A. Mancini and M. Primicerio, *Tumours with cancer stem cells: A PDE model*, Math. Biosci. **272** (2015) 76–80.
3. A. Friedman, *Partial Differential Equations of Parabolic Type*, Dover Publications, New York, 2008.
4. T. Hillen, H. Enderling and P. Hahnfeldt, *The tumor growth paradox and immune system-mediated selection for cancer stem cells*, Bull. Math. Biol. **75** (2013) 161–184.
5. L. Maddalena, *Analysis of an integro-differential system modeling tumor growth*, Appl. Math. Comput. **245** (2014) 152–157.

E-mail: gh.aliasghari@sru.ac.ir
E-mail: hmesgarani@sru.ac.ir

A Fractional-Order Model of CA3 Hippocampal Pyramidal Neurons

Leila Eftekhari*
Department of Mathematics, Tarbiat Modares University, Tehran, Iran
and Soleiman Hoseinpour
Department of Applied Mathematics, Shahrood University of Technology, Shahrood, Iran

Abstract. We study the mathematical modeling and dynamics of a two-compartment CA3 hippocampal pyramidal cell with Caputo fractional derivative. We investigate the solutions, bifurcation diagrams and chaotic behavior of the system. Chaotic regions are obtained for different values of the fractional derivative order and different injection currents. The obtained results can be considered as help to control relevant diseases caused by maximal injection currents abnormality.
Keywords: CA3 hippocampal pyramidal neurons, Caputo fractional derivative, Bifurcation analysis.
AMS Mathematical Subject Classification [2010]: 26A33, 34A08.

1. Introduction

In 1994, Pinsky and Rinzel developed a two-compartment model for CA3 hippocampal pyramidal neurons in a guinea pig. Their work was motivated by a complex 19 compartments Traub model [1]. By this reduced Pinsky-Rinzel model, it is possible to explain how interactions between the somatic and dendritic compartments occur. This recent model allows a very good computational implementation and attracted the attention of many scholars [2].

The Pinsky-Rinzel model is a non-smooth and mathematical analysis of its dynamical behavior by classic methods is hard and sometimes just impossible. In 2001, Hahn and Drand built up a mathematical analysis of the dynamical properties of the model, based on Strogatz's works [1]. They probed the changeover between resting, bursting, and spiking states affected by increasing the amount of extracellular potassium concentration.

In this paper, we survey the solutions (membrane potentials and currents) of the fractional-order CA3 hippocampal pyramidal neurons model. We study the bifurcations of the model as the fractional derivative order, α, changes and we identify regular and chaotic regimes. Next, we consider the somatic and dendritic injections, $I_{S_{app}}$ and $I_{D_{app}}$, as bifurcation parameters and explore the chaotic behavior of the system in each case. In this paper, sometimes we also name this model as the Pinsky-Rinzel model.

*Speaker

2. Mathematical Model of Fractional-Order Pinsky-Rinzel Model

In this part, we aim to extend the integer-order CA3 model to a fractional-order CA3 model. There are several definitions for fractional derivatives and integrals in fractional calculus but in this paper, we proposed the CA3 model based on the Caputo derivative.

DEFINITION 2.1. The fractional integral of order α of function $f(t)$ is defined as [3]

$$I^{\alpha} f(t) = \frac{1}{\Gamma(\alpha)} \int_0^t f(\tau)(t-\tau)^{\alpha-1} d\tau, \tag{1}$$

where $\Gamma(.)$ is Euler's Gamma function and $f(t) \in C^n$.

DEFINITION 2.2. The left fractional derivative of order α in the sense of Caputo is defined by

$${}^{C}D^{\alpha} f(t) = \begin{cases} I^{n-\alpha} D^n f(t), & n-1 < \alpha < n, \\ D^n f(t), & \alpha = n, \end{cases} \tag{2}$$

where $n \in \mathbb{N}$ and D^n is the integer derivative of order n.

The Pinsky-Rinzel model is based on two compartments, the somatic and dendritic tree, where the somatic compartment is combined with fast Sodium $\mathrm{I_{Na}}$ and delayed rectifier Potassium $\mathrm{I_{K_{DR}}}$ and leak current. The dendritic compartment has a persistent Calcium $\mathrm{I_{Ca}}$, Calcium activated Potassium $\mathrm{I_{K_{Ca}}}$ and after hyperpolarisation Potassium current $\mathrm{I_{K_{AHP}}}$ and leak current. Electronic coupling between the two compartments is modeled using two parameters; $\mathrm{g_c}$ is the strength of coupling and p is the percentage of the total area in the somatic-like compartment. $\mathrm{I_{S_{app}}}$ or $\mathrm{I_{D_{app}}}$ are coupling currents between the two compartments [2].

In this article, inspired by inter-order model [1], we propose fractional-order differential equations for the somatic ($\mathrm{V_s}$) and dendritic ($\mathrm{V_d}$) membrance potentials.

$$C_m(\alpha)\, {}^{C}D^{\alpha} V_s = -\mathrm{I_{Leak}} - \mathrm{I_{Na}} - \mathrm{I_{K_{DR}}} + \frac{\mathrm{I_{D_{app}}}}{\mathrm{p}} + \frac{\mathrm{I_{S_{app}}}}{\mathrm{p}}, \tag{3}$$

$$C_m(\alpha)\, {}^{C}D^{\alpha} V_d = -\mathrm{I_{Leak}} - \mathrm{I_{Ca}} - \mathrm{I_{K_{Ca}}} + \mathrm{I_{K_{AHP}}} + \frac{\mathrm{I_{S_{app}}}}{(1-\mathrm{p})} + \frac{\mathrm{I_{D_{app}}}}{(1-\mathrm{p})}, \tag{4}$$

where $C_m(\alpha) = \dfrac{\tau^{\alpha}}{R_m}$, R_m is the membrane resistance, and τ is the time constant. Maximal conductance parameters were taken (in $\mathrm{ms/cm^2}$) as $\mathrm{g_{Na}} = 30$, $\mathrm{g_{K_{DR}}} = 15$, $\mathrm{g_{K_{Ca}}} = 15$, $\mathrm{g_{K_{AHP}}} = 0.8$, $\mathrm{g_{Ca}} = 10$, $\mathrm{g_L} = 0.1$ and $\mathrm{g_c} = 2.1$, while reversal potentials were taken (in mV) as $\mathrm{V_{Na}} = 60$, $\mathrm{V_K} = -75$, $\mathrm{V_{Ca}} = 80$, and $\mathrm{V_L} = -60$. The size of the axosomatic compartment as a proportion of the entire cell was given by $\mathrm{p} = 0.5$ and that of the dendritic compartment as $1 - \mathrm{p}$. The capacitance is $\mathrm{C_m} = 3\ \mu\mathrm{F/cm^2}$.

The Various currents of the model are defined as follows

$$\begin{aligned}
\mathrm{I_{Na}} &= \mathrm{g_{Na}} \mathrm{m}_\infty^2(\mathrm{V_s}) \mathrm{h}(\mathrm{V_s} - \mathrm{V_{Na}}),\\
\mathrm{I_{Ca}} &= \mathrm{g_{Ca}} \mathrm{s}^2(\mathrm{V_d} - \mathrm{V_N}),\\
\mathrm{I_{K_{Ca}}} &= \mathrm{g_{k_{Ca}}} \mathrm{C}\ \chi(\mathrm{Ca})(\mathrm{V_d} - \mathrm{V_{Ca}}),\\
\mathrm{I_{SD}} &= -\mathrm{I_{DS}} = \mathrm{g_c}(\mathrm{V_d} - \mathrm{V_s}),\\
\mathrm{I_{Leak}} &= \mathrm{g_L}(\mathrm{V} - \mathrm{V_L}).
\end{aligned}$$

The activation and inactivation variables add here to these equations

$$^{C}D^{\alpha}\omega(\mathrm{V}) = \frac{\omega_\infty(\mathrm{V}) - \omega}{\tau_\omega(\mathrm{V})}, \tag{5}$$

$$\omega_\infty(\mathrm{V}) = \frac{\alpha_\omega(\mathrm{V})}{\alpha_\omega(\mathrm{V}) - \beta_\omega(\mathrm{V})}, \tag{6}$$

$$\tau_\omega(\mathrm{V}) = \frac{1}{\alpha_\omega(\mathrm{V}) + \beta_\omega(\mathrm{V})}, \tag{7}$$

where, independly, we consider $\omega = h, n, s, m, C$ and q. The rate functions are defined as follows

$$\begin{aligned}
\alpha_m(\mathrm{V_s}) &= \frac{0.32(-46.9 - \mathrm{V_s})}{\exp(\frac{-46.9+\mathrm{V_s}}{4}) - 1},\\
\beta_m(\mathrm{V_s}) &= \frac{0.28(\mathrm{V_s} + 19.9)}{\exp(\frac{\mathrm{V_s}+19.9}{5}) - 1},\\
\alpha_n(\mathrm{V_s}) &= \frac{0.016(-24.9 - \mathrm{V_s})}{\exp(\frac{-24.9+\mathrm{V_s}}{5}) - 1},\\
\beta_n(\mathrm{V_s}) &= 0.25\exp(-1 - 0.025\mathrm{V_s}),\\
\alpha_h(\mathrm{V_s}) &= 0.128\exp\Big(\frac{-43 - \mathrm{V_s}}{18}\Big),\\
\beta_h(\mathrm{V_s}) &= \frac{4}{1 + \exp(\frac{(-20-\mathrm{V_s})}{5})},\\
\alpha_s(\mathrm{V_d}) &= \frac{1.6}{1 + \exp(-0.072(\mathrm{V_d} - 5))},\\
\beta_s(\mathrm{V_d}) &= \frac{0.02(\mathrm{V_d} + 8.9)}{\exp(\frac{(\mathrm{V_d}+8.9)}{5}) - 1},\\
\alpha_C(\mathrm{V_d}) &= \frac{(1 - H(\mathrm{V_d} + 10))\exp(\frac{(\mathrm{V_d}+50)}{11} + \frac{(\mathrm{V_d}+53.5)}{27})}{18.975}\\
&\quad + H(\mathrm{V_d} + 10)(2\exp(\frac{(-53.5 + \mathrm{V_d})}{27})),\\
\beta_C(\mathrm{V_d}) &= (1 - H(\mathrm{V_d} + 10))(2\exp(\frac{(-53.5 - \mathrm{V_d})}{27}) - \alpha_c(\mathrm{V_d})),\\
\alpha_q(\mathrm{Ca}) &= \min(0.00002\mathrm{Ca}, 0.01),\\
\beta_q(\mathrm{Ca}) &= 0.001,\\
\chi(\mathrm{Ca}) &= \min\Big(\frac{\mathrm{Ca}}{250}, 1\Big),
\end{aligned}$$

where $w = h, s, n, m$ are defined simply by continuous rate functions, while the C, q and χ are formulated as discontinuous rate functions, where $H(.)$ is the Heaviside step function. To help us to get bifurcation analysis easily, we approximate these discontinuous functions. The approximated functions are as below

$$
\begin{aligned}
C_{\infty}(\mathrm{V_d}) &= \left(\frac{1}{1+\exp(\frac{10.1-V_d}{0.1016})} \right)^{0.00925}, \\
\tau_C(\mathrm{V_d}) &= 3.627\exp(0.03704V_d), \\
q_{\infty}(\mathrm{Ca}) &= 0.7894\exp(0.0002726\mathrm{Ca}) - 0.7292\ \exp(-0.01672\mathrm{Ca}), \\
\tau_q(\mathrm{Ca}) &= 657.9\exp(-0.02023\mathrm{Ca}) + 301.8\exp(-0.002381\mathrm{Ca}), \\
\chi(\mathrm{Ca}) &= 1.073\mathrm{n}(0.003453\mathrm{Ca}+0.08095) + 0.08408\mathrm{n}(0.01634\mathrm{Ca}-2.34) \\
&\quad +0.01811\mathrm{n}(0.0348\ \mathrm{Ca}-0.9918),
\end{aligned}
$$

also we have

$$\frac{\mathrm{dCa}}{\mathrm{dt}} = -0.13\mathrm{I_{Ca}} + 0.075\mathrm{Ca}.$$

The first minus sign is based on the convention that inward currents are negative. Figure 1 shows the action potential in Pinsky-Rinzel model in case $\alpha = 0.85$ and $\alpha = 1$. In fact, Figure 1 shows a comparison between fractional-order and integer-order cases. In fractional case $\alpha = 0.85$, refractory period is smaller than integer case $\alpha = 1$. We also consider $\mathrm{I_{S_{app}}}$ as the bifurcation parameter and let $\mathrm{I_{D_{app}}} = 0, \mathrm{g_c} = 2.1$, the bifurcation diagrams of the system when the parameter $\mathrm{I_{S_{app}}}$ varied with $\alpha = 1$ and $\alpha = 85$ are shown in Figure 2.

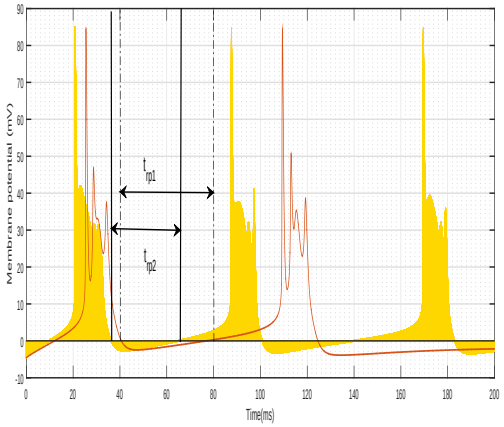


FIGURE 1. refractory period (t_{rp2}) for $\alpha = 0.85$ (yellow) and refractory period (t_{rp1}) for $\alpha = 1$ (red).

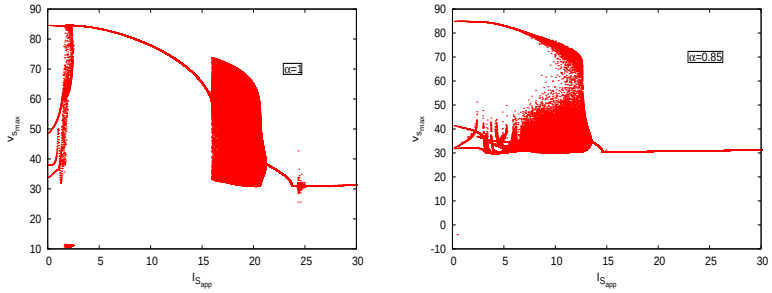


FIGURE 2. Bifurcation diagram when the parameter $I_{S_{app}}$ varied with the different values of the order α.

3. Conclusion

In this work, we investigated the bifurcation analysis of a fractional-order model of a two-compartment CA3 hippocampal pyramidal cell. Chaotic regions were achieved for different values of the fractional derivative order and different injection currents. Since the membrane capacitance has been considered to be ideal, the fractional-order model is deemed as a more accurate description of physical processes underlying a long-range memory behavior.

References

1. R. D. Traub, R. K. S. Wong, R. Miles and H. Michelson, *A model of a CA3 hippocampal pyramidal neuron incorporating voltage-clamp data on intrinsic conductances*, J. Neurophysiol. **66** (1991) 635–650.
2. E. M. Izhikevich, *Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting,* MIT Press, Cambridge, 2010.
3. I. Podlobny, *Fractional Differential Equations,* Academic press, San Diego, 1999.
4. S. H. Weinberg, *Membrane capacitive memory alters spiking in neurons described by the fractional-order Hodgkin-Huxley model*, PloS ONE **10** (5) (2015) e0126629. DOI:10.1371/journal.pone.0126629

E-mail: leila.eftekhari32@gmail.com
E-mail: soleiman.hosseinpour@gmail.com

Race Lévy Flights Model: A PDE Framework for Modeling Dynamic Decisions with Multiple Alternatives

Amir Hosein Hadian Rasanan
Institute for Cognitive and Brain Sciences, Shahid Beheshti University, Tehran, Iran
Jamal Amani Rad*
Department of Cognitive Modeling, Institute for Cognitive and Brain Sciences, Shahid Beheshti University, Tehran, Iran
and Amin Padash
Laser and Plasma Research Institute, Shahid Beheshti University, Tehran, Iran

Abstract. Lévy Flights model has attracted much attention and it performs much better than the other sequential sampling models. But there are some drawbacks with the Lévy Flights model. The first one is that it could just model decisions with only two options. Secondly, there is no exact likelihood function for this model. In this work, a new paradigm is presented for modeling the decision making that can be applied for both 2-alternative and multi-alternatives. Moreover, a space fractional partial differential equation (fPDE) is proposed for approximating the probability distribution of the first passage time of the model.
Keywords: Lévy Flights, Fractional calculus, Decision making, Sequential sampling models.
AMS Mathematical Subject Classification [2010]: 91E10, 00A06, 35R11.

1. Introduction

A wide range of psychological assessments are based on the performance (i.e. reaction time and accuracy) of the patient in a 2-alternative decision task and the balance between reaction time and accuracy (i.e. speed-accuracy tradeoff) is a very informative measure [2]. Since 1978, when Ratcliff has introduced a drift-diffusion model for modeling the process of making a decision between 2-alternatives [4], sequential sampling models have grown very much. Various sequential sampling models have been introduced by the researchers that all of them are based on accumulating a fixed amount of evidence until reaching a threshold [1]. The first model for capturing the speed-accuracy tradeoff pattern which is presented by Ratcliff can be formulated as a random walk model fluctuating between two constant boundaries and the process is stopped by overshooting or hitting one of the boundaries. The boundary that is hit or overshot declares which option should be selected. Thus, it can be formulated as follows [4]

$$\begin{cases} x(0) = z > 0, \\ x(t + \Delta t) = x(t) + v\Delta t + e\sqrt{\Delta}, \ e \sim \mathcal{N}(0, 1). \end{cases} \tag{1}$$

*Speaker

The process has stopped whenever $x(t) > a$ or $x(t) < 0$ in which a shows the upper boundary, zero stands for the lower boundary, v presents the drift rate and implies the speed of information processing, and z which is $0 < z < a$, is the starting point bias. Additional to these parameters, a non-decision time should be added to the decision time (i.e. the non-decision time stands for summation of encoding time and motor time) [7]. Figure 1 illustrates the drift-diffusion model.

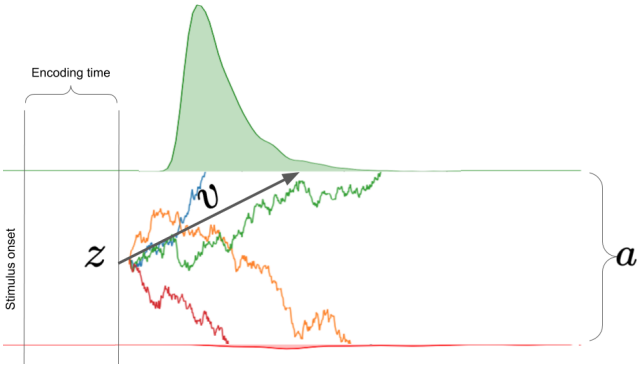


FIGURE 1. A schematic view of drift diffusion model.

As mentioned in Eq. (1), the process of information accumulation is considered as a noisy process and its noise has a normal distribution. This assumption is an add-hock assumption that the researchers have added to the model [5], but it is revised in the Lévy Flights model which is introduced by Voss and his collaborators in 2019 [8]. In this model, the noise of the accumulation process has a α-stable distribution and is formulated as follows

$$\begin{cases} x(0) = z > 0, \\ x(t + \Delta t) = x(t) + v\Delta t + e\Delta t^{\frac{1}{\alpha}}, \ e \sim stable(\alpha, \beta = 0, \gamma = \frac{1}{\sqrt{2}}, \delta = 0). \end{cases}$$

By considering the α-stable distribution as the distribution of accumulation process of noise, some jumps have occurred during the decision process. There are some psychological evidence that jumping through the accumulation process has psychological meaning [9] and should be extended to other sequential sampling models. To this end, we are going to extend the Lévy Flights model to multi-alternative decisions.

In the remaining of the paper, the Race Levy model will be introduced in Section 2. Space fractional differential equation which is utilized for obtaining the first passage time distribution of the model is presented in this section. Section 3, contains a finite difference scheme for solving the mentions space fractional equation and some simulation results of the model. Finally, a conclusion is presented in Section 4.

2. Race Lévy Flights Model

In the race accumulator framework, there is one accumulator corresponding to each option. Therefore, there is a race competition between the accumulators and

the first accumulator which reaches the threshold determines which option should be selected. Therefore, the Race Lévy model can be formulated as follows

$$\begin{cases} x_i(0) = z_i > 0, \\ x_i(t+\Delta t) = x_i(t) + v_i \Delta t + e_i \Delta t^{\frac{1}{\alpha}}, \ e_i \sim stable(\alpha, \beta = 0, \gamma = \frac{1}{\sqrt{2}}, \delta = 0), \end{cases}$$

and the process finishes when $\exists i \ x_i(t) > a$. The first passage time distribution of the model is equivalent to the response time distribution of human performance in a behavioral task. Therefore, we start with the location of the accumulator through time. If the i-th accumulator starts from z_i and the threshold locates on a, then the distance between the starting point and the threshold is $a - z_i$. So, the location of the accumulator can be obtained by [3]

$$\begin{cases} \frac{\partial}{\partial t} p_i(x,t) + v \frac{\partial}{\partial x} p_i(x,t) = D_x^{\alpha} p_i(x,t), \\ p_i(x,0) = \delta(x), p(a - z_i, t) = p(-\infty, t) = 0, \end{cases}$$

where $D_x^{\alpha} p(x,t) = \frac{-1}{2\cos\frac{\alpha\pi}{2}} \left({}_{-\infty}^{C} D_x^{\alpha} p(x,t) + {}_{x}^{C} D_{\infty}^{\alpha} p(x,t) \right)$, and ${}_{a}^{C} D_{b}^{\alpha}$ is the Caputo sense fractional derivative [3], $p_i(x,t)$ presents the distribution of the location of the accumulator trough the time and the survival probability of the i-th accumulator is $S_i(t) = \int_{-\infty}^{a-z_i} p_i(x,t) dx$. So, the first passage time of i-th accumulator calculated as $fpt_i(t) = -\frac{dS_i(t)}{dt}$. Finally, the probability of the first accumulator finishing from all accumulators is given by the following defective probability density function [6]

$$fpt_i(t) \prod_{i \neq j} S_j(t).$$

So by approximating $p_i(x,t)$, the likelihood function of the model can be obtained.

3. Model Behaviour

In this part, two examples are provided to illustrate that the model can capture the first passage time distribution of behavioral data. To this end, 3000 sample data are simulated and the $fpt(t)$ is also approximated using a finite difference scheme. In this scheme, Eq. (2) and Eq. (3) are used for discretizing the space fractional operator [3]

$$(2) \quad {}_{-\infty} D_{x_i}^{\alpha} p(x_i, t_n) = \int_{-L}^{x_i} \frac{p^{(2)}(\xi, t_n)}{(x_i - \xi)^{(\alpha-1)}} = \frac{\Delta x^{2-\alpha}}{(2-\alpha)(3-\alpha)} \sum_{k=0}^{i} \lambda_{k,i-k} p^{(2)}(x_k, t_n) + \mathcal{O}(\Delta x^2),$$

$$(3) \quad {}_{x_i} D_{d}^{\alpha} p(x_i, t_n) = \int_{x_i}^{d} \frac{p^{(2)}(\xi, t_n)}{(x_i - \xi)^{(\alpha-1)}} = \frac{\Delta x^{2-\alpha}}{(2-\alpha)(3-\alpha)} \sum_{k=i}^{N} \lambda_{k,k-i} p^{(2)}(x_k, t_n) + \mathcal{O}(\Delta x^2),$$

$$\lambda_{k,m} = \begin{cases} (m-1)^{3-\alpha} - (m-3+\alpha) m^{2-\alpha}, \ k = 0, N, \\ (m-1)^{3-\alpha} - 2m^{3-\alpha} + (m-1)^{3-\alpha}, \ 0 < k < N, k \neq i, \\ 1, \ k = i. \end{cases}$$

Moreover, by using a central scheme for first and the second-order derivatives, a linear system of algebraic equations is obtained for each time step. For more detail about the obtained linear system of algebraic equation see [3]. Figure 2 and Figure 3 show the behaviour of the model and approximation of their first passage

time distribution for two different parameters set. The first parameter set is $\{\alpha = 2, z = 0, v = 2, a = 2\}$, and the second one is $\{\alpha = 1.6, z = 0, v = 2.3, a = 3\}$.

As obvious in Figure 2 and Figure 3 the approximations and simulations are compatible with each other and the model can capture the first passage time distribution.

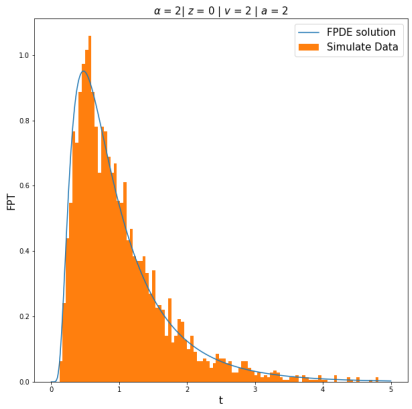


FIGURE 2. Simulation behavior of the model for $\alpha = 2$, $z = 0$, $v = 2$ and $a = 2$.

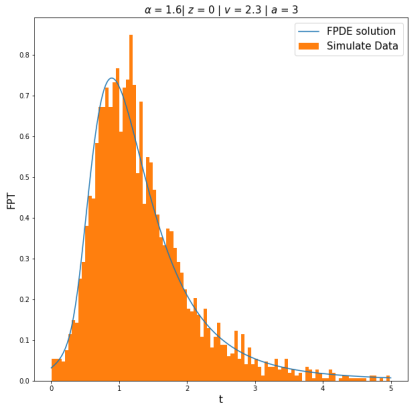


FIGURE 3. Simulation behavior of the model for $\alpha = 1.6$, $z = 0$, $v = 2.3$ and $a = 3$.

4. Conclusion

In this paper, a random walk model has been introduced for modeling the performance of patients in multi-alternative decisions. The proposed model is based on the combination of race accumulator framework with the Lévy Flights model. Additionally, a mathematically tractable procedure for approximating the likelihood function of the model is presented. This procedure is based on approximating the solution of a space partial fractional differential equation by a finite difference scheme.

References

1. R. Bogacz, E. Brown, J. Moehlis, P. Holmes and J. D. Cohen, *The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks*, Psychol. Rev. **113** (2006) 700–765.
2. R. Bogacz, E. J. Wagenmakers, B. U. Forstmann and S. Nieuwenhuis, *The neural basis of the speedaccuracy tradeoff*, Trends Neurosci. **33** (2010) 10–16.
3. A. Padash, A. V. Chechkin, B. Dybiec, I. Pavlyukevich, B. Shokri and R. Metzler, *First-passage properties of asymmetric Lévy flights*, J. Phys. A: Math. Theor. **52** (2019) 454004.
4. R. Ratcliff, *A theory of memory retrieval*, Psychol. Rev. **85** (1978) 59–108.
5. M. Shinn, N. H. Lam and J. D. Murray, *A flexible framework for simulating and fitting generalized drift-diffusion models*, eLife **9** (2020) 1–27.
6. G. Tillman, T. Van Zandtc and G. D. Loganb, *Sequential sampling models without random between-trial variability: The racing diffusion model of speeded decision making*, Psychon. Bull. Rev. **27** (2020) 911–936.
7. A. Voss, K. Rothermund and J. Voss, *Interpreting the parameters of the diffusion model: An empirical validation*, Memory & cognition **32** (2004) 1206–1220.
8. A. Voss, V. Lerche, U. Mertens and J. Voss, *Sequential sampling models with variable boundaries and non-normal noise: A comparison of six models*, Psychon Bull. Rev. **26** (2019) 813–832.
9. E. M. Wieschen, A. Voss and S. Radev, *Jumping to conclusion? A Lévy flight model of decision making*, TQMP **16** (2020) 120–132.

E-mail: amir.h.hadian@gmail.com
E-mail: j.amanirad@gmail.com; j_amanirad@sbu.ac.ir
E-mail: padash.amin@gmail.com

Analysis of Predator-Prey System with Infection

Mohammad Hossein Rahmani Doust*
Department of Mathematics, Faculty of Sciences, University of Neyshabur, Neyshabur, Iran
and Atena Ghasemabadi
Esfarayen University of Technology, Esfarayen, North Khorasan, Iran

Abstract. Mathematical modeling of diseases enables one to predict when the disease occurs, and therefore, leading to the successful control to the diseases before it gets epidemic. This paper constructs a biological model in the mathematical aspect. Solutions for a Lotka-Volterra diseased predator-prey model are analyzed. Properties such as positivity, boundedness for solutions are studied. The threshold parameters for existence of both species are determined. Based on these parameters, local and global asymptotic stability is then analyzed. Finally, a numerical simulation that verifies the obtained analytical discussion is presented.
Keywords: Prey-Predator, Lotka-Volterra Model, Threshold Parameter, Stability.
AMS Mathematical Subject Classification [2010]: 34D20, 34D23, 93D20.

1. Introduction

The main reason to use mathematical modeling for a contagious disease is "such models explain in a clear way the eco-social mechanisms that are influential in controlling the contagious disease". Some essential abbreviations are given as follows: **M**: shows the class of those who have moderate immunity to the disease. **S**: shows the class of those that are susceptible to the disease. **E**: shows those that are exposed to the disease and cannot transmit it. **I**: shows the infected class that can be transmitters. **R**: shows the class of those that have recovered. If a contact occurs between the infected individuals and those who are susceptible to the disease, the individual can be grouped in E. In E wherein the individuals are in incubation period; they are infected but cannot transmit the disease. After this period, the sick individual can be grouped among the class I who can transmit the disease. After the end of recovery, the individuals enter class R and get immune to the disease permanently or temporarily. Mathematical models for contagious diseases are in forms such as SI, SIS, SEIS, SIR, SIERS, MSIERS, SIRS and etc. For example, In SIR, the sick individual is relatively immune to the disease after recovery. Such studies and applications in biology, ecology, population dynamic, eco-epidemiological model, diseases and their transmission and prey -predator system with infection have been carried out in [3] and [5]. Linearization method has

*Speaker

been offered as a practical methodology for investigating locally asymptomatic stability [4]. A four-species model and the Lotka-Volterra models used to introducing and branching practical models applied to disease have been worked [1] and [2].

2. Main Results

We model a system of Lotka-Volterra predator-prey equations by considering assumptions: (**i**) The prey species is ill, (**ii**) Feed predator species only with ill prey, (**iii**) Disease transfers to predator. (**iv**) If prey is not ill, this species has logistic model with birth growth rate and death rate $''b_1 - \frac{a_1 r_1 N_1}{K_1}''$ and $''d_1 + (1-a_1)\frac{r_1 N_1}{K_1}''$, respectively, where $r_1 = b_1 - d_1$ and $0 \leq a_1 \leq 1$. (**v**) Parameter α is the average of contacts between preys and predators. Hence, the receive rate of diseases for susceptible prey in the duration of predation is multiple of α.

$$(1) \qquad \begin{cases} I_1^{'} = \beta_1(N_1 - I_1)I_1 - \gamma_1 I_1 - [d_1 + (1-a_1)\frac{r_1 N_1}{K_1}]I_1 - aN_2 I_1, \\ N_1^{'} = r_1(1 - \frac{N_1}{K_1})N_1 - aN_1 N_2, \\ I_2^{'} = (N_2 - I_2)(\beta_2 I_2 + \alpha I_1) - d_2 I_2 - \gamma_2 I_2, \\ N_2^{'} = kaN_1 N_2 - d_2 N_2. \end{cases}$$

Table 1. Description of parameters for system (1).

a_1	Convex combination constant of prey
b_1	Natural birth rate constant of prey
d_1	Natural death rate constant of prey
d_2	Natural death rate constant of predator
$r_1 = b_1 - d_1$	Growth rate constant of prey
K_1	Carrying capacity of the environment of prey
β_1	Daily contact rate of prey
β_2	Daily contact rate of predator
γ_1	Recovery rate constant of prey
γ_2	Recovery rate constant of predator
k	Efficiency in turning predation into new predators
α	Average number of contacts
a	Predation rate

2.1. Equilibria and Thresholds Parameters. We are going to find the equilibria. Setting the right sides of system (1) equal zero, we find equilibria as:

$$\begin{aligned} E_0 &= (0,0,0,0), \\ E_1 &= (0,K_1,0,0), \\ E_2 &= (0,N_{1E},0,N_{2E}) = (0,\frac{d_2}{ka},0,\frac{r_1}{a}(1-\frac{d_2}{kaK_1})), \\ E_3 &= (0,N_{1E},I_{2E},N_{2E}), \\ E_4 &= (I_{1E},N_{1E},I_{2E},N_{2E}). \end{aligned}$$

And so

$$N_{1E} = \frac{d_2}{ka}, N_{2E} = \frac{r_1}{a}(1 - \frac{d_2}{kaK_1}),$$
$$I_{1E} = N_{1E}(1 - \frac{\gamma_1 + d_1 + (1-a_1)r_1N_{1E}/K_1 + aN_{2E}}{\beta_1 N_{1E}}).$$

To positivity N_{2E}, we should have $d_2 < kaK_1$, and I_{2E} should be positive root of the following equation:

$$I_2^2 + I_2[\frac{d_2+\gamma_2}{\beta_2} - N_{2E} + \alpha\frac{I_{1E}}{\beta_2}] - \alpha\frac{N_{2E}I_{1E}}{\beta_2} = 0,$$
$$I_2^2 + AI_2 + B = 0, A = \frac{d_2+\gamma_2}{\beta_2} - N_{2E} + \alpha\frac{I_{1E}}{\beta_2}, B = -\alpha\frac{N_{2E}I_{1E}}{\beta_2},$$
$$I_{2E} = \frac{-A + \sqrt{A^2 - 4B}}{2}.$$

System (1) has three following threshold parameters:

$$(2)\quad R_0 = \frac{\beta_1 K_1}{\gamma_1 + b_1 - a_1\gamma_1}, \quad R_1 = \frac{\beta_1 N_{1E}}{\gamma_1 + d_1 + (1-a_1)r_1N_{1E}/K_1 + aN_{2E}},$$
$$R_2 = \frac{\beta_2 N_{2E}}{\gamma_2 + d_2}.$$

We analyze the stability of nontrivial equilibria for (1). Coefficient matrix at point E_4 is:

$$(3)\qquad J = \begin{bmatrix} a_{11} & a_{12} & 0 & a_{14} \\ 0 & a_{22} & 0 & a_{24} \\ a_{31} & 0 & a_{33} & a_{34} \\ 0 & a_{42} & 0 & 0 \end{bmatrix}.$$

$$a_{11} = -3\beta_1 I_{1E}, \quad a_{12} = I_{1E}[\beta_1 - r_1\frac{1-a_1}{K_1}], \quad a_{14} = -aI_{1E},$$
$$a_{22} = -\frac{r_1 N_{1E}}{K_1}, \quad a_{24} = -aN_{1E}, \quad a_{31} = -\alpha(I_{2E} - N_{2E}),$$
$$a_{33} = [-\beta_2(N_{2EE} - I_{2E}) - d_2 - \alpha I_{1E} - \gamma_2] - 2\beta_2 I_{2E},$$
$$a_{34} = \beta_2 I_{2E} + \alpha I_{1E}, \quad a_{42} = kaN_{2E}.$$

The characteristic equation for matrix J (3) can be obtained as follows:

$$(a_{11} - \lambda)(a_{33} - \lambda)[\lambda^2 - a_{22}\lambda - a_{24}a_{42}] = 0,$$
$$\lambda_1 = a_{11}, \lambda_2 = a_{33}, \lambda_3 = \frac{a_{22} + \sqrt{a_{22}^2 + 4a_{24}a_{42}}}{2}, \lambda_4 = \frac{a_{22} - \sqrt{a_{22}^2 + 4a_{24}a_{42}}}{2}.$$

THEOREM 2.1. *Let $R_1 > 1$ and $R_2 < 1$, where R_1 and R_2 are threshold parameters for system (1). Then E_4 is locally asymptotic stable.*

THEOREM 2.2. *Consider system (1). If $\frac{d_2}{k\,a\,K_1} > 1$, then E_1 is locally asymptotic stable. If $\frac{d_2}{k\,a\,K_1} > 1$, then E_1 is globally asymptotic stable.*

THEOREM 2.3. *If $\frac{d_2}{k\,a\,K_1} < 1$, then $\lim_{t\to+\infty} N_1(t) = N_{1E}$, and $\lim_{t\to+\infty} N_2(t) = N_{2E}$.*

THEOREM 2.4. *Consider threshold parameters (3) for (1), if $R_1 > 1$ and $R_2 < 1$. Then E_4 is globally asymptotic stable.*

3. Numerical Simulation

We simulate solutions of (1). Consider parameters of the system (1) as

$$\begin{aligned} &\beta_1 = 2, \quad K_1 = 100, \quad d_1 = 0.2, \quad \alpha = 0.7, \quad \gamma_1 = 0.6, \quad r_1 = 0.1, \\ &\beta_2 = 0.4, \quad k = 0.5, \quad d_2 = 0.2, \quad b_1 = 0.5, \quad \gamma_2 = 0.6, \quad a_1 = 0.4. \end{aligned}$$

We get the critical value $a(0) = 0.0383589737$, and, the system (1) is as follows:

$$\begin{cases} \dot{I}_1 = [2(N_1 - I_1) - 0.8 - 0.0006\, N_1 - 0.0383589737\, N_2]\, I_1, \\ \dot{N}_1 = [0.1 - 0.001 N_1 - 0.0383589737\, N_2]\ N_1, \\ \dot{I}_2 = [0.4(N_2 - I_2) - 0.8]\ I_2 + 0.7(N_2 - I_2) I_1, \\ \dot{N}_2 = [0.0191794\, N_1 - 0.2]\ N_2, \end{cases}$$

We compute $R_0 = 2.232569247 > 1$, $R_0 = 0.5 < 1$. By Theorem 2.1, If $R_1 > 1, R_2 < 1$, then $(I_{1E}, N_{1E}, I_{2E}, N_{2E}) = (22.32, 40, 4.5848, 6.5)$ is asymptotic stable verified in Figures 1, 2, 3 and 4.

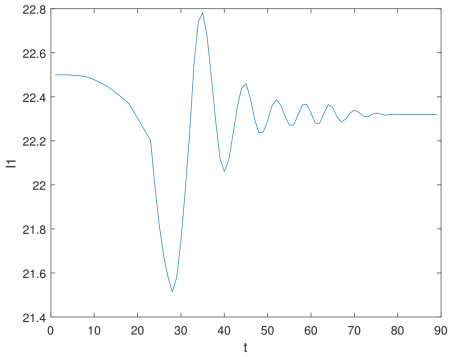


FIGURE 1. The component I_{1E} at E_4 converges to 22.32 the starting point 22.5.

4. Conclusion

One may reach main conclusion on the following points: "One way to protect species is controlling of their diseases. Most of the prevalent diseases occur in specific times and under certain conditions. Mathematical modeling of diseases enables one to predict when the disease occurs and so it leads to the successful control to the diseases before it gets epidemics. The present work has shown that gaining the threshold parameters can prevent the outbreak of a disease."

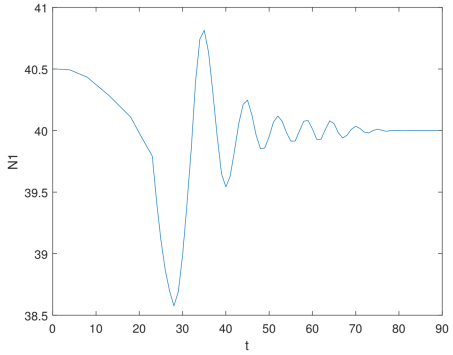


FIGURE 2. The component N_{1E} at E_4 converges to 40 the starting point 40.5.

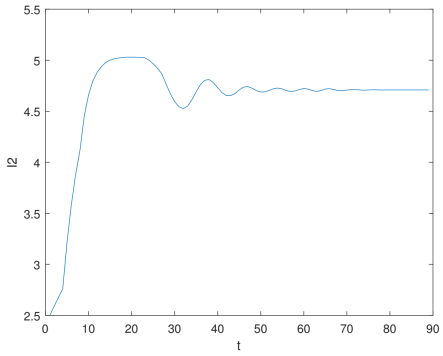


FIGURE 3. The component I_{2E} at E_4 converges to 4.5848 the starting point 2.5.

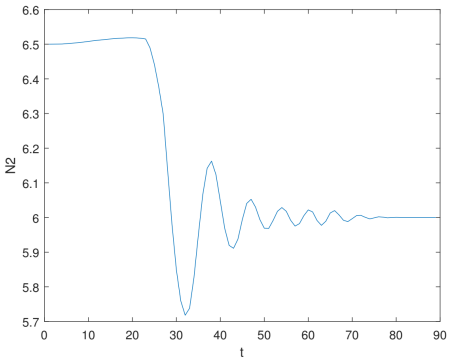


FIGURE 4. The component N_{2E} at E_4 converges to 6 the starting point 6.5.

References

1. A. Ghasemabadi, *Stability and bifurcation in a generalized delay prey-predator model,* Nonlinear Dyn. **90** (2017) 2239–2251.
2. R. Memarbashi, F. Alipour and A. Ghasemabadi, *A nonstandard finite difference scheme for a sei epidemic model,* J. Math. **49** (3) (2017) 133–147.
3. J. D. Murray, *Mathematical Biology* 1*: An Introduction*, Springer-Verlag, New York, 2003.
4. M. H. Rahmani Doust and F. Motahari Nasab, *Existence and uniqunees of asymptotic periodic solution in the cyclic four species predator-prey model,* J. Adv. Math. Model. **9** (1) (2019) 144–160.
5. Z. Xiao and Z. Zhong *Stability analysis of mutual interference predator-prey model with the fear effect*, J. Appl. Sci. Eng. **22** (2) (2019) 205–211.

E-mail: mh.rahmanidoust@neyshabur.ac.ir
E-mail: ghasemabadi.math@gmail.com

Applying Computer Algebra for Parametric Representation of the Steady States of Overlapping Generations Model

Monireh Riahi*
Department of Mathematics and Computer Sciences, Damghan University, Damghan, Iran
Abdolali Basiri
Department of Mathematics and Computer Sciences, Damghan University, Damghan, Iran
Sajjad Rahmany
Department of Mathematics and Computer Sciences, Damghan University, Damghan, Iran
and Felix Kübler
Swiss Banking Institute, University of Zurich, Zurich 8032, Switzerland

Abstract. In this paper, we address the problem of analyzing and computing the steady-states of the overlapping generation model. The computation of steady-states coincides with a geometrical representation of the algebraic variety of a polynomial ideal which tends to apply computational algebraic geometry methods to solve the problem. However, as the associated polynomial ideal to these models have parametric coefficients, it is necessary to deal with the ring of parametric polynomials. In doing so, we apply novel parametric computational tools such as comprehensive Gröbner systems to discuss the parameters space. In addition, the parameters are bounded and in fact restricted into some real intervals. This property causes to do some extra steps more than the computation a comprehensive Gröbner system. Having all the constraints on the parameters, we design a new algorithm to determine the value of each steady-state depending on the different behaviour of parameters. Doing so, the space of parameters will be divided into a finite number of algebraic sets in the way that each one determine a number of steady states, if there is any.
Keywords: Computer Algebra, Gröbner basis, Comprehensive Gröbner system, Steady-states, OLG model.
AMS Mathematical Subject Classification [2010]: 13P10, 91B52.

1. Introduction

Gröbner bases are known as effective computational tools to analyze and solve systems of polynomial equations. There are several applied problems (See [5] for instance) which will be solved by computing a suitable Gröbner basis. Before stating the main results, we review the concept of Gröbner bases in some lines. For more information, one can see [1].

*Speaker

Let $\mathbb{K}$ be a field and $\mathbf{x} = x_1, \ldots, x_n$ be n (algebraically independent) variables. Each power product $\mathbf{x}^\alpha = x_1^{\alpha_1} \ldots x_n^{\alpha_n}$ is called a monomial, where $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{Z}_{\geq 0}^n$. We can sort the set of all monomials over $\mathbb{K}$ by special types of total orderings so called monomial orderings: the total ordering $\prec$ on the set of monomials is called a monomial ordering whenever $\prec$ is well-ordering and invariant under multiplication. Among the monomial orderings, we point to the lexicographic ordering denoted by $\prec_{lex}$ as follows: assuming $x_n \prec \cdots \prec x_1$, we say that $\mathbf{x}^\alpha \prec_{lex} \mathbf{x}^\beta$ if α is smaller than β in the lexicographical sense. Each $\mathbb{K}-$linear combination of monomials is called a polynomial on $\mathbf{x}$ over $\mathbb{K}$. The set of all polynomials has the ring structure with usual polynomial addition and multiplication, and is called the polynomial ring on $\mathbf{x}$ over $\mathbb{K}$ and denoted by $\mathbb{K}[\mathbf{x}]$. Let f be a polynomial and $\prec$ be a monomial ordering. The greatest monomial w.r.t. $\prec$ contained in f is called the leading monomial of f, denoted by $\mathrm{LM}(f)$. Further, if $\mathcal{I}$ is an ideal, $\mathrm{in}(\mathcal{I})$ is the ideal generated by $\mathrm{LM}(\mathcal{I})$ and is called the initial ideal of $\mathcal{I}$. We are now going to remind the concept of Gröbner basis of a polynomial ideal: The finite set $G \subset \mathcal{I}$ is called a Gröbner basis of $\mathcal{I}$ w.r.t. the monomial ordering $\prec$ if $\mathrm{in}(\mathcal{I}) = \langle \mathrm{LM}(G) \rangle$. There are several algorithms to compute Gröbner bases which are also implemented in some software packages [2]. Let G be a Gröbner basis for $\mathcal{I}$ w.r.t. the monomial ordering $\prec$. For each polynomial f, the normal form of f w.r.t. G denoted by $\mathrm{NF}_G(f)$ is a polynomial such that none of its terms is divisible by $\mathrm{LM}(G)$. Also, by $\mathbf{V}(\mathcal{I})$ we mean the set of all common solutions of the polynomials of $\mathcal{I}$.

We turn now to the ring of parametric polynomials. Let $\mathbf{p} := p_1, \ldots, p_s$ and $\mathbf{x} := x_1, \ldots, x_n$ be the sequences of parameters and variables respectively. We call $\mathbb{K}[\mathbf{p}][\mathbf{x}]$, the parametric polynomial ring over $\mathbb{K}$, with parameters $\mathbf{p}$ and variables $\mathbf{x}$. This ring is in fact the set of all parametric polynomials as $\sum_{i=1}^m h_i \mathbf{x}^{\alpha_i}$, where $h_i \in \mathbb{K}[\mathbf{p}]$ is a polynomial on $\mathbf{p}$ with coefficients in $\mathbb{K}$, for each i.

Definition 1.1. The set of triples $\{(Z_i, W_i, G_i)\}_{i=1}^\ell$ in which for each i, $Z_i, W_i \subset \mathbb{K}[\mathbf{p}]$ and $G_i \subset \mathbb{K}[\mathbf{p}][\mathbf{x}]$ is called a comprehensive Gröbner system of the parametric ideal $\mathcal{I}$ with respect to the monomial ordering $\prec$ if for each evaluation map $\sigma : \mathbf{p} \mapsto \bar{\mathbf{p}}$ there exists an $1 \leq i \leq \ell$ in which for all $p \in Z_i$ (resp. $q \in W_i$), $p(\mathbf{p}) = \mathbf{0}$ (resp. $q(\mathbf{p}) \neq \mathbf{0}$), and $\sigma(G_i)$ is a Gröbner basis for $\sigma(\mathcal{I})$ with respect to $\prec$.

Remark that, by [6, Theorem 2.7], every parametric ideal has a comprehensive Gröbner system. Now we give an example from [3] to illustrate the definition of comprehensive Gröbner system.

Example 1.2. Consider the following parametric polynomial system in $\mathbf{Q}[a, b, c][x, y]$:

$$\Sigma : \begin{cases} ax - b &= 0, \\ by - a &= 0, \\ cx^2 - y &= 0, \\ cy^2 - x &= 0. \end{cases}$$

Choosing the graded reverse lexicographical ordering $y \prec x$, we obtain the following comprehensive Gröbner system.

TABLE 1. Comprehensive Gröbner system of the parametric polynomial ideal in Example 1.2.

i	Z_i	W_i	G_i
1	$\{\ \}$	$\{a^6-b^6, a^3c-b^3,$ $b^3c-a^3, ac^2-a, bc^2-b\}$	$\{1\}$
2	$\{a^6-b^6, a^3c-b^3, b^3c-a^3,$ $ac^2-a, bc^2-b\}$	$\{b\}$	$\{bx-acy, by-a\}$
3	$\{a, b\}$	$\{c\}$	$\{cx^2-y, cy^2-x\}$
4	$\{a, b, c\}$	$\{\ \}$	$\{x, y\}$

Regarding to the Table 1, for the specialization $\sigma_{(1,1,1)}$ for which $a \mapsto 1, b \mapsto 1$ and $c \mapsto 1$,

$$\sigma_{(1,1,1)}(\{bx-acy, by-a\}) = \{x-y, y-1\},$$

is a Gröbner basis of $\sigma_{(1,1,1)}(\langle\Sigma\rangle)$.

In this paper, we study a polynomial system which is obtained from the problem of the calculation of the equilibria of an economical model [4]. Consider the following parametric polynomial system, where A is a natural number greater than 1. Also, $l_1 \geq 0, \ldots, l_{A-1} \geq 0, l_A = 1-\sum_{a=1}^{A-1}$, $\gamma \in [0, +\infty)$, $\beta \in (0, +\infty)$, $\alpha \in (0, 1)$ and $\delta \in [0, 1]$ are the parameters of this system.

$$(1) \quad \begin{cases} c_a^{-\gamma-1} = \beta(1+r)c_{a+1}^{-\gamma-1}, \quad (a = 1, \ldots, A-1), \\ c_a = k_{a-1}(1+r) + wl_a - k_a, \quad (a = 1, \ldots, A, k_0 = k_A = 0), \\ r = \alpha K^{\alpha-1} - \delta, \\ w = (1-\alpha)K^{\alpha}, \\ K = \sum_{a=0}^{A} k_a. \end{cases}$$

Each solution of this system is called a steady-state of the assumed model. Some of the equations of this system are not in the polynomial form. To change their structure into the polynomial form, we can assume that $\gamma = 0$ and $\alpha = m/n$, where m, n are natural numbers and $m < n$. Now, we import an auxiliary variable S such that $S^n = K$ and so $K^{\alpha-1} = S^{m-n}$. Thus, we multiply the equation $r = \alpha K^{\alpha-1} - \delta$ by S^{n-m}. After these changes, we attain the following polynomial system which is equivalent to system (1).

$$(2) \quad \Sigma_{A,\alpha} := \begin{cases} c_{a+1} = \beta(1+r)c_{a+1}, \quad (a = 1, \ldots, A-1), \\ c_a = k_{a-1}(1+r) + wl_a - k_a, \quad (a = 1, \ldots, A, k_0 = k_A = 0), \\ (r+\delta)S^{n-m} = \alpha, \\ w = (1-\alpha)S^m, \\ S^n = \sum_{a=0}^{A} k_a, \\ K = S^n. \end{cases}$$

In the sequel, we begin to present the properties of the parametric polynomial ideal associated to system (2). Suppose that

$$\mathcal{I}_{A,\alpha} \subset \mathbb{Q}[\beta, \delta, l_1, \ldots, l_A][c_1, \ldots, c_A, k_0, \ldots, k_A, K, S, w, r],$$

is the parametric ideal generated by the equations of system (2).

DEFINITION 1.3. The set of triples $\{(Z_i, W_i, G_i)\}_{i=1}^{\ell}$ in which for each i, $Z_i, W_i \subset \mathbb{Q}[\beta, \delta, l_1, \ldots, l_A]$ and $G_i \subset \mathbb{Q}[\beta, \delta, l_1, \ldots, l_A][\mathbf{x}]$ is called a steady-state system of the OLG model associated to $\Sigma_{A,\alpha}$ (and is denoted by $SSS(\Sigma_{A,\alpha})$). If for each evaluation map $\sigma : (\beta, \delta, l_1, \ldots, l_A) \mapsto (\lambda_1, \lambda_2, \bar{l_1}, \ldots, \bar{l_A})$ there exists an $1 \leq i \leq \ell$ in which for all $p \in Z_i$ (resp. $q \in W_i$), $p(\lambda_1, \lambda_2, \bar{l_1}, \ldots, \bar{l_A}) = 0$ (resp. $q(\lambda_1, \lambda_2, \bar{l_1}, \ldots, \bar{l_A}) \neq 0$), and $\sigma(G_i)$ is a triangular polynomial system (Gröbner basis) for $\sigma(\mathcal{I}_{A,\alpha})$ whose solutions are the steady-states of $\sigma(\Sigma_{A,\alpha})$.

The following theorem states the existence of steady-state systems.

THEOREM 1.4. *For each value of A and α, $\Sigma_{A,\alpha}$ possesses a steady-state system.*

PROOF. Suppose that $\sigma : (\beta, \delta, l_1, \ldots, l_A) \mapsto (\lambda_1, \lambda_2, \bar{l_1}, \ldots, \bar{l_A})$ is an evaluation map on the space of parameters. It is obvious that for each value of A and α, the algebraic variety $\mathbb{V}(\sigma(\mathcal{I}_{A,\alpha}))$ contains the steady-states of $\sigma(\Sigma_{A,\alpha})$. Now, assume that $\mathcal{G}$ is a comprehensive Gröbner system of $\mathcal{I}_{A,\alpha})$ with respect to compatible monomial ordering. Note that such a system exists by [6, Theorem 2.7]. From the definition of comprehensive Gröbner system, there exists a triple $(Z, W, G) \in \mathcal{G}$ such that for all $p \in Z$ (resp. $q \in W$), $p(\lambda_1, \lambda_2, \bar{l_1}, \ldots, \bar{l_A}) = 0$ (resp. $q(\lambda_1, \lambda_2, \bar{l_1}, \ldots, \bar{l_A}) \neq 0$), and $\sigma(G)$ is a Gröbner basis such that $\mathbb{V}(\sigma(G)) = \mathbb{V}(\sigma(\mathcal{I}_{A,\alpha}))$. Therefore, $\mathbb{V}(\sigma(G))$ contains the steady-states of $\sigma(\Sigma)$ and so, $\mathcal{G}$ coincides with a steady-state system of Σ. □

REMARK 1.5. Regarding to the description of the model and Σ, the parameters can not vary all around the set of real numbers. This restriction causes to omit some triples from the comprehensive Gröbner system described in the proof of Theorem 1.4. Therefore, not only it is not sufficient just to compute a comprehensive Gröbner system, but also it is necessary to analyse the triples and verify if they contain any real solution according to the restrictions on the parameters.

In regard to the above observations, the following algorithm demonstrates the way of computing a steady-state system for $\Sigma_{A,\alpha}$.

Algorithm 1. SS-System

Require: A; the number of generations, and $\alpha = \frac{m}{n}$.
Ensure: A SS-System for $\Sigma_{A,\alpha}$
  $\mathcal{G}$:= a comprehensive Gröbner system for $\mathcal{I}_{A,\alpha}$;
  $SS := \{\}$;
  for $(Z, W, G) \in \mathcal{G}$ **do**
    if there exists $\omega \in \mathbb{R}^{A+2}$ such that $\omega_\beta > 0$, $\omega_\delta \in [0, 1]$, and $\omega_{l_a} \geq 0$ for all $a = 1, \ldots, A$ **then**
      $SS := SS \cup \{G\}$;
    end if
  end for
  Return (SS).

The following example illustrates the behaviour of the above algorithm.

EXAMPLE 1.6. Let $A = 2$ and $\alpha = 1/2$. Because of the large scale of the polynomials, we state just two triples of the output of the SS-SYSTEM algorithm (here $p = \beta(1 + r)$).

If $Z = \{\delta-1\}$ and $W = \{L_1, \beta, \beta+l_2+1, \beta+l_2+1, 4\beta^2+8\beta l_2+4l_2^2+8\beta+8l_2+4\}$ then

$G = \{(2\beta+2l_2+2)S-\beta l_1, (4\beta^2+8\beta l_2+4l_2^2+8\beta+8l_2+4)K-\beta^2 l_1^2, (4\beta+4l_2+4)w-\beta l_1, (4\beta^2+8\beta l_2+4l_2^2+8\beta+8l_2+4)c_1-\beta l_1^2 l_2-\beta l_1^2, \beta r l_1+\beta l_1-\beta-l_2-1, (4\beta^2+8\beta l_2+4l_2^2+8\beta+8l_2+4)k_1-\beta^2 l_1^2, (4\beta+4l_2+4)c_2-\beta l_1 l_2-\beta l_1, pl_1-\beta-l_2-1\}$,

and if $Z = \{\}$ and $W = \{\delta-1, L_1, \beta, \beta\delta-\beta+\delta-1\}$ then

$G = \{(4\beta\delta-4\beta+4\delta-4)S^2+(-2\beta\delta l_1+2\beta l_1-2\beta-2l_2-2)S+\beta l_1, (4\beta\delta-4\beta+4\delta-4)K+(-2\beta\delta l_1+2\beta l_1-2\beta-2l_2-2)S+\beta l_1, 2w-S, (-2\delta l_1+2\beta+2l_1+2l_2+2)S+(4\beta\delta-4\beta+4\delta-4)c_1-\beta l_1, (2\beta\delta-2\beta+2\delta-2)S+\beta l_1 r+\beta l_1-\beta-l_2-1, (-2\beta\delta l_1+2\beta l_1-2\beta-2l_2-2)S+(4\beta\delta-4\beta+4\delta-4)k_1+\beta l_1, (2\beta\delta L_1-2\beta l_1-2\beta l_2)S+(4\beta+4)c_2-\beta l_1, (2\beta\delta-2\beta+2\delta-2)S+pl_1-\beta-l_2-1\}$.

Regarding to these triples, one can substitute the values of parameters and solve the obtained polynomial system to observe the steady-states.

References

1. D. A. Cox, J. Little and D. O'shea, *Using Algebraic Geometry*, Springer, New York, 2005.
2. S. Gao, Y. Guan and F. Volny, *A new incremental algorithm for computing Gröbner bases*, Proc. of ISSAC'10, New York: ACM (2010) pp. 13–19.
3. D. Kapur, Y. Sun and D. Wang, *An efficient algorithm for computing a comprehensive Gröbner system of a parametric polynomial system*, J. Symbolic Comput. **49** (2013) 27–44.
4. D. Krueger and F. Kübler, *Computing equilibrium in OLG models with stochastic production*, J. Econ. Dyn. Control. **28** (7) (2004) 1411–1436.
5. F. Kübler and K. Schmedders, *Tackling multiplicity of equilibria with Gröbner bases*, Oper. Res. **58** (2010) 1037–1050.
6. V. Weispfenning, *Comprehensive Gröbner bases*, J. Symbolic Comput. **14** (1) (1992) 1–29.

E-mail: Monire.Riahi@gmail.com
E-mail: Basiri@du.ac.ir
E-mail: s_rahmani@du.ac.ir
E-mail: Felix.Kuebler@bf.uzh.ch

Deriving Coherent and Non-Coherent Risk Measures under the Logistic Distribution

Fazlollah Soleymani*
Department of Mathematics, Institute for Advanced Studies in Basic Sciences (IASBS), Zanjan, Iran

Abstract. Financial markets may face with high volatilities and instabilities. In such circumstances, traders and managers use some concepts such as value-at-risk (VaR) to handle the amount of risk in a financial firm. In this paper, the improved versions of VaR known as Conditional VaR (CVaR) and Entropic VaR (EVaR) are derived for the logistic distribution. Hence, closed formulations for these measures are contributed.
Keywords: Value-at-risk, Conditional value-at-risk, Entropic value-at-risk, Logistic distribution, Risk management.
AMS Mathematical Subject Classification [2010]: 91B30, 62P05, 91G70.

1. Introduction

It is famous that the value-at-risk (VaR) measure based on normal distribution, and sometimes the Conditional VaR (CVaR) measure, tend to underestimate the risk of loss, due to the heaviness of the tails of the real loss distribution, which are typically fatter than the ones modeled in a Gaussian framework. One approach for addressing this shortcoming is to substitute an alternative distribution that allows for greater weight in the tails. This can also be improved more by considering much tighter bounds for the risk measure to furnish reliable estimate for real financial data.

The VaR is defined in what follows [7]:

$$\mathrm{VaR}_\alpha(X) := \inf\{z \in \mathbb{R} | F_X(z) \geq \alpha\}, \tag{1}$$

where α is the pre-determined confidence level, X is a random variable, and $F_X(\cdot)$ is for the cumulative distribution function (shorthanded as CDF). This is straightforward that (1) is employed to obtain the loss. In fact, the measure of VaR is basically applied by banks at the portfolios to realize the occurrence and extent ratio of possible losses, [5]. To be more precise recently, the oil sector has showed un-stability in international oil prices, that have been more representative from 2004 and respond to different available factors. Accordingly, the VaR measure is not useful in several circumstances and this restricts its applicability and usefulness.

The CVaR is in fact the average loss of the given distribution in the extreme tail region and accordingly has enough superiority to be considered as an improvement over the VaR measure. It leads to higher values for the risk in comparison

*Speaker

to the VaR, [4, Chapter 15]. This measure of risk is defined by:

$$\mathrm{CVaR}_\alpha(X) := \mathbb{E}\left[X | X \geq \mathrm{VaR}_\alpha(X)\right]. \tag{2}$$

(2) is derived from (1) for a portfolio and is used in a portfolio's optimization for effective risk management. The formula (2) is also called the lower CVaR until the equality sign holds. As long as this inequality be strict, it is named as the upper CVaR, [8]. Subsequently, noting that the CVaR's application in comparison to VaR leads to a more efficient procedure based on exposure of risk.

Another important risk measure with several interesting features is the entropic VaR (EVaR). The EVaR is defined by [2]:

$$\mathrm{EVaR}_\alpha(X) := \underbrace{\inf}_{\theta} \left\{ \theta \log\left(\frac{\mathcal{M}_X(\theta)}{1-\alpha} \right) \right\}, \tag{3}$$

wherein $\mathcal{M}_X(t) = \mathbb{E}(e^{-tx})$. In this work, we denote the natural logarithm by log. The tightest higher bound that we could obtain based on the Chernoff inequality for the CVaR and VaR is the risk measure EVaR, [10].

In this work, we investigate closed forms of VaR/CVaR/EVaR under the logistic fat-tailed distribution, which can then be applied for controlling the risk of stock movements. In fact, theoretical and simulation results confirm the applicability of the logistic distribution in contrast to the Gaussian distribution for risk management.

Noting that exhausting all the non–Gaussian models in modeling stock returns is not our main goal, which is infeasible. In fact, our motivation is to adopt a fat-tailed distribution, viz., the logistic, that is rich enough to accommodate the features of financial data in terms of calculating the non-coherent VaR, coherent CVaR as well as the coherent EVaR measures. Also recalling that the concept of coherency was defined and discussed deeply at [1].

The rest of the present study is unfolded as comes next. In Section 2, the logistic distribution is defined briefly. Next, in Section 3, the risk measures of VaR/CVaR/EVaR are contributed in closed forms for this distribution. At last, a summary of the work is given in Section 4.

2. Logistic Distribution

The best-fitted distribution for a financial data set gives us a procedure to express the behavior of the underlying financial data. In fitting the economic data with a distribution, basically more than one distribution would be of interest in the matching process, [6]. In addition, to have a useful distribution supporting the fat-tail behavior of the economic and financial data sets, one remedy is to rely on fatter-tail distributions.

The logistic distribution with the parameters p and q shows a continuous statistical distribution given over the set $\mathbb{R}$ of real numbers and parameterized by a real number p (known as the "mean" of the distribution) and a positive real number q (known as "scale parameter"). Overall, the probability density function (PDF) of a logistic distribution is uni-modal with a single "peak" (i.e., a global maximum), though its overall shape (the horizontal location of its maximum, its spread, and its height) is determined by the values of p and q.

Let us consider the random variable X to be distributed as

$$X \sim \text{logistic}(p, q). \tag{4}$$

The CDF for the logistic distribution can be obtained as follows [9]:

$$F(x) = \frac{1}{e^{-\frac{x-p}{q}} + 1}, \tag{5}$$

while its PDF is given by:

$$f(x) = \frac{e^{-\frac{x-p}{q}}}{q\left(e^{-\frac{x-p}{q}} + 1\right)^2}. \tag{6}$$

The most important difference between the logistic and the normal distributions lies in the tails and in the behavior of the failure rate function. The logistic distribution has heavier tails in comparison to the normal distribution.

3. Risk Measures

The aim of this section is to contribute on closed formulations for the computation of the coherent risk measures of CVaR and EVaR. Before doing so, the non-coherent risk measure of VaR is furnished as follows. By considering (1) and using (4), one is able to write down

$$\begin{aligned} \text{VaR}_\alpha(X) &= \inf\{t \in \mathbb{R} \mid p(X \leq t) \geq \alpha\} \\ &= \inf\{t \in \mathbb{R} \mid F_X(t) \geq \alpha\} \\ &= \inf\left\{t \in \mathbb{R} \mid \frac{1}{e^{-\frac{x-p}{q}} + 1} \geq \alpha\right\} \\ &= p - q\log\left(\frac{1}{\alpha} - 1\right), \qquad 0 < \alpha < 1. \end{aligned} \tag{7}$$

The measure obtained in (7) is not convex and coherent. Now, we focus on CVaR and EVaR in the following theoretics.

THEOREM 3.1. *Assuming that $X \in L^{\mathfrak{p}}$ is a random variable showing the loss for logistic(p, q) distribution. The measure of CVaR employing is given in a closed form by (9).*

PROOF. The payoff random variable X belongs $L^{\mathfrak{p}}$ spaces, where $\mathfrak{p} \geq 1$ in order to guarantee the existence of the expectation. Having (2) in mind, we obtain

$$\begin{aligned} \text{CVaR}_\alpha(X) &= \mathbb{E}\left[X | X \geq \text{VaR}_\alpha(X)\right] \\ &= \mathbb{E}\left[X | X \geq p - q\log\left(\frac{1}{\alpha} - 1\right)\right], \quad 0 < \alpha < 1, \end{aligned} \tag{8}$$

$$= p - 19q\log(19) + 20q\log(20), \qquad \alpha = 95/100. \tag{9}$$

Here we include the final results for the most common choice of the confidence level 95%. In fact, the computation of the condition expectation in (8) for the general case is challenging and thus we restrict the final formula for this confidence level. The CVaR's application in comparison to VaR leads to an efficient procedure for the risk exposure. The choice among CVaR and VaR is not so obvious sometimes

for traders in market, but we basically employ CVaR as a double check procedure to the hypotheses of VaR when managing the risk. This ends the proof. □

As an another instance, for the confidence level $\alpha = 99/100$, we obtain

$$\mathrm{CVaR}_{99\%}(X) = p + 200q\log(10) - 99q\log(99). \tag{10}$$

THEOREM 3.2. *Under the conditions of Theorem 3.1, the EVaR measure using the logistic(p, q), is computed by (11).*

PROOF. Following the methodology of the proof Theorem 3.1 and employing (3), one obtains

$$\begin{aligned}
\mathrm{EVaR}_\alpha(X) &= \underbrace{\inf}_{z}\left\{z\log\left(\frac{\mathcal{M}_X(z)}{1-\alpha}\right)\right\},\ 0<\alpha<1,\\
&= \underbrace{\inf}_{z}\left\{z\log\left(\frac{\frac{e^{p/z}}{\mathrm{sinc}\left(\frac{\pi q}{z}\right)}}{1-\alpha}\right)\right\}\\
&= \underbrace{\inf}_{z}\left\{z\log\left(-\frac{e^{p/z}}{(\alpha-1)\mathrm{sinc}\left(\frac{\pi q}{z}\right)}\right)\right\}\\
&= \underbrace{\min}_{z}\left\{z\log\left(-\frac{e^{p/z}}{(\alpha-1)\mathrm{sinc}\left(\frac{\pi q}{z}\right)}\right)\right\}\\
&= \underbrace{\min}_{z}\left\{\varphi(z,p,q,\alpha)\right\}. \qquad (11)
\end{aligned}$$

The result of the minimization (11) for any $z, p, q > 0$ gives the value of the EVaR risk measure. The proof is complete. □

REMARK 3.3. Noting that optimization with the EVaR is tractable in terms of computation for a large class of quantitative risks, which are not effectively computable for the VaR and CVaR, [3]. To clearly put on show how the new risk measure EVaR for the logistic distribution gives upper bounds for the risk involved in the problem, Figure 1 with $\alpha = 95\%$ is furnished giving a picture of this fact. Results confirm the point that:

$$\mathrm{EVaR}_\alpha(X) \geq \mathrm{CVaR}_\alpha(X) \geq \mathrm{VaR}_\alpha(X). \tag{12}$$

4. Concluding Remarks

A coherent tight bound for risk management is the EVaR, which is based on a function minimization. Due to this importance and several shortcomings of VaR and CVaR based on normal distributions, this work has discussed closed formulas for the VaR/CVaR/EVaR measures using the fat-tailed logistic distribution. The investigation was necessary because history has showed that in a short piece of time thousands of dollars can be waisted due to failure in handling the financial risks in market. To support the theoretical discussions given in this work, an application can be pursued on time series in forecasting the prices/returns of some stocks in a period of time under time series models.

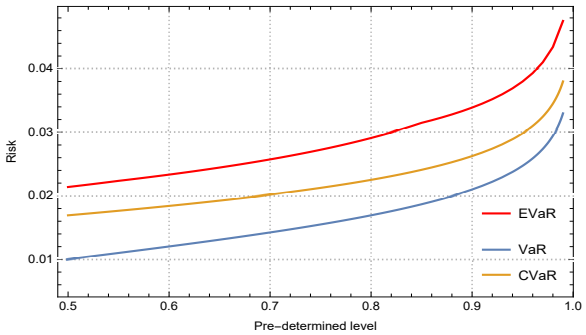


FIGURE 1. Applying the logistic distribution having $p = 0.01$, $q = 0.005$ to contrast the VaR/CVaR/EVaR measures.

References

1. C. Acerbi and T. Dirk, *On the coherence of expected shortfall*, J. Bank Finance **26** (7) (2002) 1487–1503.
2. A. Ahmadi-Javid, *An information-theoretic approach to constructing coherent risk measures*, In Info. Theory Proc. (ISIT), IEEE Int. Symp. Infor. Theory, Russia, (2011) pp. 2125–2127.
3. A. Ahmadi-Javid and M. Fallah-Tafti, *Portfolio optimization with entropic value-at-risk*, European J. Oper. Res. **279** (1) (2019) 225–241.
4. H. Albrecher, A. Binder, V. Lautscham and P. Mayer, *Introduction to Quantitative Methods for Financial Markets*, Springer, Basel, 2013.
5. M. Braione and N. K. Scholtes, *Forecasting value-at-risk under different distributional assumptions*, Econometrics **4** (1) (2016) 3.
6. R. Karim, P. Hossain, S. Begum and F. Hossain, *Rayleigh mixture distribution*, J. Appl. Math. **2011** (2011) 238290.
7. H. Markowitz, *Portfolio selection*, J. Finance **7** (1) (1952) 77–91.
8. S. Sarykalin, G. Serraino and S. Uryasev, *Value-at-risk vs. conditional value-at-risk in risk management and optimization*, In State-of-the-art decision-making tools in the information-intensive age, Informs. (2008) 270–294.
9. H. C. Yeh, *Multivariate semi-logistic distributions*, J. Multivariate Anal. **101** (4) (2010) 893–908.
10. C. Zheng and Y. Chen, *Allocation of risk capital based on iso-entropic coherent risk measure*; J. Ind. Eng. Manag. **8** (2) (2015) 530–553.

E-mail: fazlollah.soleymani@gmail.com; soleymani@iasbs.ac.ir

Fuzzy n-Fold Obstinate (Pre)Filters of EQ-Algebras

Batool Ganji Saffar*
Department of Mathematics, Faculty of Mathematical Sciences, Alzahra University, Tehran, Iran

Abstract. In this paper, we defined the concepts of fuzzy n-fold obstinate (pre)filter and fuzzy maximal (pre)filter of EQ-algebras and discussed the properties of them. We show that every fuzzy maximal (pre)filter of $\mathcal{E}$ is normalized and takes only the values $\{0,1\}$ and in good EQ-algebra, if μ is a normalized fuzzy (pre)filter of $\mathcal{E}$, then μ is a fuzzy n-fold obstinate (pre)filter of $\mathcal{E}$ if and only if every normalized fuzzy (pre)filter of quotient algebra $\mathcal{E}/\mu$ is a fuzzy n-fold obstinate (pre)filter of $\mathcal{E}/\mu$.
Keywords: EQ-algebra, Fuzzy n-fold obstinate (pre)filter, Maximal (pre)filter, Fuzzy maximal (pre)filter.
AMS Mathematical Subject Classification [2010]: 03G25, 06B10, 06B99.

1. Introduction and Preliminaries

Recently, a special algebra called EQ-algebra has been introduced by Novák. These algebras are intended to become algebras of truth values for a higher-order fuzzy logic (a fuzzy type theory, FTT). From the point of view of potential application, it seems very interesting that, we can have non-commutativity without the necessity to introduce, two kinds of implication. Filter theory plays an important role in studying logical algebras. From a logic point of view, various filters have a natural interpretation as various sets of provable formulas. Up to now, some types of n-fold filters on BCK-algebra, BL-algebra and etc., are studied. The study of fuzzy algebraic structures was started with the introduction of the concept of fuzzy sub-groups in 1971 by Rosenfeld. Since then these ideas have been applied to other algebraic structures such as semigroups, rings, ideals, modules and vector spaces.

Now, in this note, we defined the concepts of fuzzy n-fold obstinate (pre)filter and fuzzy maximal (pre)filter of EQ-algebras and discussed the properties of them.

Definition 1.1. [2] An *EQ-algebra* is an algebraic structure $\mathcal{E} = (E, \wedge, \otimes, \sim, 1)$ of type (2, 2, 2, 0) such that, for all $x, y, z, t \in E$ the following conditions hold:
$(E1)$ $\langle E, \wedge, 1\rangle$ is a commutative idempotent monoid (i.e. $\wedge$-semilattice with top element 1),
$(E2)$ $\langle E, \otimes, 1\rangle$ is a commutative monoid and $\otimes$ is isotone w.r.t. " $\leq$ ", where $x \leq y$ is defined as $x \wedge y = x$,
$(E3)$ $x \sim x = 1$, (reflexivity axiom)
$(E4)$ $((x \wedge y) \sim z) \otimes (t \sim x) \leq z \sim (t \wedge y)$, (substitution axiom)
$(E5)$ $(x \sim y) \otimes (z \sim t) \leq (x \sim z) \sim (y \sim t)$, (congruence axiom)

*Speaker

$(E6)$ $(x \wedge y \wedge z) \sim x \leq (x \wedge y) \sim x$, (monotonicity axiom)
$(E7)$ $x \otimes y \leq x \sim y$. (boundedness axiom)

Definition 1.2. [6] Let $\mathcal{E} = (E, \wedge, \otimes, \sim, 1)$ be an EQ-algebra. Then $\mathcal{E}$ is called

i) *separated* if $x \sim y = 1$, then $x = y$, for all $x, y \in E$, (in other words $x \sim y = 1$ if and only if $x = y$),
ii) *residuated* if $x \leq y \to z$ if and only if $x \otimes y \leq z$, for all $x, y, z \in E$.

Definition 1.3. [6] Let $\mathcal{E}$ be an EQ-algebra. A nonempty subset $F \subseteq E$ is called a *prefilter* of $\mathcal{E}$, if for all $x, y \in E$,
$(F1)$ $1 \in F$,
$(F2)$ If $x, x \to y \in F$, then $y \in F$.
A prefilter F is said to be a *filter*
$(F3)$ if $x \to y \in F$ implies $(x \otimes z) \to (y \otimes z) \in F$, for all $x, y, z \in E$.
A proper (pre)filter F is called a *prime (pre)filter* of $\mathcal{E}$ if $x \to y \in F$ or $y \to x \in F$, for all $x, y \in E$.

Definition 1.4. [7] A (pre)filter F of an EQ-algebra $\mathcal{E}$ is called *maximal* if and only if it is proper and no (pre)filter of $\mathcal{E}$ strictly contains F that is, for each (pre)filter G of $\mathcal{E}$, if $F \subsetneq G$, then $G = E$.

Definition 1.5. [5] A prefilter F of an EQ-algebra $\mathcal{E}$ is called an *obstinate prefilter* of $\mathcal{E}$, if $x, y \notin F$ imply $x \to y \in F$ and $y \to x \in F$.
If F is a filter of $\mathcal{E}$, then F is called an *obstinate filter* of $\mathcal{E}$.

Definition 1.6. [3] Let $\mathcal{E}$ be an EQ-algebra. A nonempty subset $F \subseteq E$ such that $1 \in F$ is called

i) an *n-fold prefilter* of $\mathcal{E}$, if for all $x, y \in E$, if $x^n, x^n \to y \in F$, then $y \in F$,
ii) an *n-fold obstinate (pre)filter* of $\mathcal{E}$, if $x, y \notin F$ imply $x^n \to y \in F$ and $y^n \to x \in F$.

Note. An n-fold prefilter F is said to be an *n-fold obstinate* filter of $\mathcal{E}$, if F satisfies in $(F3)$.

Definition 1.7. [8] Let E be a set. A fuzzy set μ in E is a function $\mu : E \to [0, 1]$.

Note. From now one, $\mathcal{E} = \langle E, \wedge, \otimes, \sim, 1\rangle$ or $\mathcal{E}$ for short, is denoted an EQ-algebra. Let μ be a fuzzy set in $\mathcal{E}$. For all $t \in [0, 1]$, the set $\mu_t = \{x \in E \mid \mu(x) \geq t\}$ is called a *level subset of* μ. Let F be a nonempty subset, we denote the characteristic function of F by χ_F. For converience, for any $a, b \in [0, 1]$, we denote $\max\{a, b\}$ and $\min\{a, b\}$ by $a \vee b$ and $a \wedge b$, respectively.

For any fuzzy sets μ and ν in $\mathcal{E}$, we define $\mu \leq \nu$ if and only if for any $x \in E$, $\mu(x) \leq \nu(x)$.

Definition 1.8. [7] Let μ be a fuzzy set in $\mathcal{E}$. Then μ is called a *fuzzy prefilter* of $\mathcal{E}$ if for all $x, y \in E$, $\mu(x) \leq \mu(1)$ and $\mu(x \to y) \wedge \mu(x) \leq \mu(y)$.

A fuzzy prefilter μ is called a *fuzzy filter* of $\mathcal{E}$ if for all $x, y, z \in E$ $\mu(x \to y) \leq \mu((x \otimes z) \to (y \otimes z))$.

Proposition 1.9. [7] *Let μ be a fuzzy filter of $\mathcal{E}$. Then for any $x, y, z \in E$, the following conditions hold:*

i) $\mu(x \otimes y) = \mu(x) \wedge \mu(y)$.
ii) $\mu(x \to y) \wedge \mu(y \to z) \leq \mu(x \to z)$.
iii) *If* $x \leq y$, *then* $\mu(x) \leq \mu(y)$, *which means* μ *is order preserving.*

DEFINITION 1.10. [4] Let μ be a fuzzy set in $\mathcal{E}$. Then μ is called a *fuzzy n-fold prefilter of* $\mathcal{E}$ if for all $x, y \in E$, $\mu(x) \leq \mu(1)$ and $\mu(x^n) \wedge \mu(x^n \to y) \leq \mu(y)$.

A fuzzy n-fold prefilter μ of $\mathcal{E}$ is called a *fuzzy n-fold filter of* $\mathcal{E}$ if for all $x, y, z \in E$, μ satisfies in

$$\mu(x \to y) \leq \mu((x \otimes z) \to (y \otimes z)).$$

Let μ be a fuzzy filter of $\mathcal{E}$. For any $x, y \in E$, define a fuzzy relation $\equiv_\mu$ on (good) $\mathcal{E}$ as follows:

$$x \equiv_\mu y \text{ if and only if } \mu(x \sim y) = \mu(1).$$

THEOREM 1.11. [4] *The relation* $\equiv_\mu$ *is a congruence relation on* $\mathcal{E}$ *and* $\mathcal{E}/\mu = (E/\mu, \otimes_\mu, \sim_\mu, \wedge_\mu, [1]_\mu)$ *is a separated (good) EQ-algebra with operations* $\otimes_\mu$, $\sim_\mu$ *and* $\wedge_\mu$ *on* E/μ *which are defined as follows:*

$$[x]_\mu \otimes_\mu [y]_\mu = [x \otimes y]_\mu,\ [x]_\mu \sim_\mu [y]_\mu = [x \sim y]_\mu \text{ and } [x]_\mu \wedge_\mu [y]_\mu = [x \wedge y]_\mu.$$

2. Fuzzy n-Fold Obstinate (Pre)Filter of EQ-Algebra

In this section, we introduce the notion of fuzzy n-fold obstinate (pre)filter and fuzzy maximal (pre)filter of EQ-algebra and some related properties of them are investigated. We show that every fuzzy maximal (pre)filter of $\mathcal{E}$ is normalized and takes only the values $\{0, 1\}$.

DEFINITION 2.1. Let μ be a fuzzy prefilter of $\mathcal{E}$. Then μ is called a *fuzzy n-fold obstinate prefilter* of $\mathcal{E}$, if for all $x, y \in E$,

$$(1 - \mu(x)) \wedge (1 - \mu(y)) \leq \mu(x^n \to y) \wedge \mu(y^n \to x).$$

A fuzzy n-fold obstinate prefilter of $\mathcal{E}$ is called a *fuzzy n-fold obstinate filter* of $\mathcal{E}$ if for all $x, y, z \in E$, μ satisfies in $\mu(x \to y) \leq \mu((x \otimes z) \to (y \otimes z))$.

EXAMPLE 2.2. Let $E = \{0, a, b, c, 1\}$ be a chain such that $0 \leq a \leq b \leq c \leq 1$. Define the operations $\wedge, \otimes$ and $\sim$ on E as follows:

$\otimes$	0	a	b	c	1
0	0	0	0	0	0
a	0	0	0	0	a
b	0	0	0	0	b
c	0	0	0	0	c
1	0	a	b	c	1

$\sim$	0	a	b	c	1
0	1	a	0	0	0
a	a	1	a	a	a
b	0	a	1	b	b
c	0	a	b	1	c
1	0	a	b	c	1

$\to$	0	a	b	c	1
0	1	1	1	1	1
a	a	1	1	1	1
b	0	a	1	1	1
c	0	a	b	1	1
1	0	a	b	c	1

where $x \wedge y = \min\{x, y\}$. Then $\mathcal{E} = (E, \wedge, \otimes, \sim, 1)$ is an EQ-algebra. Define the fuzzy set μ on E as follows:

$$\mu(0) = 0.4,\ \mu(a) = 0.4,\ \mu(b) = 0.5,\ \mu(c) = 0.6 \text{ and } \mu(1) = 0.8.$$

Then μ is a fuzzy n-fold obstinate prefilter of $\mathcal{E}$, for all $n \geq 2$. But it is not a fuzzy 1-fold obstinate prefilter of $\mathcal{E}$ because

$$0.5 = 0.5 \wedge 0.6 = (1 - \mu(a)) \wedge (1 - \mu(b)) \not\leq \mu(a \to b) \wedge \mu(b \to a)$$
$$= \mu(1) \wedge \mu(a) = \mu(a) = 0.4.$$

THEOREM 2.3. *Let F be a non-empty subset of E. Then F is an n-fold obstinate (pre)filter of $\mathcal{E}$ if and only if χ_F is a fuzzy n-fold obstinate (pre)filter of $\mathcal{E}$.*

PROPOSITION 2.4. *Every fuzzy n-fold obstinate (pre)filter of $\mathcal{E}$ is a fuzzy $n+1$-fold obstinate (pre)filter of $\mathcal{E}$.*

PROPOSITION 2.5. *Let μ and ν be two fuzzy filters of $\mathcal{E}$ such that $\mu \leq \nu$. If μ is a fuzzy n-fold obstinate filter of $\mathcal{E}$, then ν is too.*

By definition $\mu(1)$ is the largest value of μ. Sometimes we set $\mu(1) = 1$.

DEFINITION 2.6. Let $\mathcal{E}$ be an EQ-algebra. A fuzzy (pre)filter μ of $\mathcal{E}$ is called *normalized*, if $\mu(1) = 1$.

DEFINITION 2.7. Let $\mathcal{E}$ be an EQ-algebra and μ be a fuzzy (pre)filter of $\mathcal{E}$. The *normalization* μ is a fuzzy subset $\bar{\mu} : E \to [0, 1]$ given by $\bar{\mu}(x) = \mu(x) + 1 - \mu(1)$.

LEMMA 2.8. *$\bar{\mu}$ is a normalized fuzzy (pre)filter of $\mathcal{E}$.*

COROLLARY 2.9. *Let μ be a fuzzy n-fold (pre)filter of $\mathcal{E}$. Then $\bar{\mu}$ is a normalized fuzzy n-fold (pre)filter of $\mathcal{E}$.*

Note. We denote the set of all normalized fuzzy (pre)filters of $\mathcal{E}$ by $\mathcal{F}(\mathcal{E})$ and the set of all normalized fuzzy n-fold (pre)filters of $\mathcal{E}$ by $\mathcal{F}_n(\mathcal{E})$.

PROPOSITION 2.10. *Let $\mathcal{E}$ be an EQ-algebra and μ and ν be two fuzzy (pre)filters of $\mathcal{E}$. Then the following statements hold:*

i) $\mu \leq \bar{\mu}$.
ii) *If $\mu \in \mathcal{F}(\mathcal{E})$, then $\bar{\mu} = \mu$.*
iii) *$(\mathcal{F}(\mathcal{E}), \leq)$ is a $\wedge$-semilattice ($\chi_{\{[0]\}}$ is the smallest element of $\mathcal{F}(\mathcal{E})$ and $1(x) = 1$, for all $x \in E$ is the largest element of $\mathcal{F}(\mathcal{E})$).*
iv) *If $\bar{\mu}(x) = 0$, for some $x \in E$, then $\mu(x) = 0$.*
v) *If μ and ν are two fuzzy (pre)filters of $\mathcal{E}$ such that $\bar{\mu} \in \mathcal{F}(\mathcal{E})$ and $\bar{\mu} \leq \nu$, then $\nu = \bar{\nu}$.*

DEFINITION 2.11. Let $\mathcal{E}$ be an EQ-algebra and μ be a fuzzy (pre)filter of $\mathcal{E}$. We called μ is a *fuzzy maximal (pre)filter* of $\mathcal{E}$, if it is non-constant and $\bar{\mu}$ is a maximal element of $(\mathcal{F}(\mathcal{E}), \leq)$.

LEMMA 2.12. *Let μ be non-constant. If μ is a maximal element of $(\mathcal{F}(\mathcal{E}), \leq)$, then it takes only the values $\{0, 1\}$.*

THEOREM 2.13. *Let $\mathcal{E}$ be an EQ-algebra. Then every fuzzy maximal (pre)filter of $\mathcal{E}$ is normalized and takes only the values $\{0, 1\}$.*

THEOREM 2.14. *Let $\mathcal{E}$ be an EQ-algebra with bottom element "0" and $\mu \in \mathcal{F}(\mathcal{E})$. Then $E_\mu = \{x \in E \mid \mu(x) = \mu(1)\}$ is a maximal (pre)filter of $\mathcal{E}$ if and only if μ is a maximal element of $\mathcal{F}(\mathcal{E})$.*

THEOREM 2.15. *Let $\mathcal{E}$ be an EQ-algebra with bottom element "0" and μ be a fuzzy (pre)filter of $\mathcal{E}$. Then μ is a fuzzy n-fold obstinate (pre)filter of $\mathcal{E}$ if and only if for any $x \in E$, there exists $m \in \mathbb{N}$ such that $1 - \mu(x) \leq \mu((\neg x^n)^m)$.*

THEOREM 2.16. *Let $\mathcal{E}$ be an EQ-algebra with bottom element "0" and μ be a fuzzy maximal (pre)filter of $\mathcal{E}$. Then μ is a fuzzy n-fold obstinate (pre)filter of $\mathcal{E}$ if and only if for any $x \in E$, $\mu(x) = \mu(1)$ or there exists $m \in \mathbb{N}$ such that $\mu((\neg x^n)^m) = \mu(1)$, which $\neg x = x \sim 0$.*

THEOREM 2.17. *Let $\mathcal{E}$ be a residuated EQ-algebra with bottom element "0" and μ be a fuzzy (pre)filter of $\mathcal{E}$. Then the following statements are equivalent:*

i) *E_μ is a maximal (pre)filter of $\mathcal{E}$,*
ii) *for any $x \in E$, if $\mu(x) \neq \mu(1)$, then there exists $n \in \mathbb{N}$ such that $\mu(\neg x^n) = \mu(1)$, which $\neg x = x \sim 0$.*

PROPOSITION 2.18. *Let $\mathcal{E}$ be a good EQ-algebra. Then $\chi_{\{1\}}$ is a fuzzy n-fold obstinate filter of $\mathcal{E}$ if and only if every normalized fuzzy (pre)filter of $\mathcal{E}$ is a fuzzy n-fold obstinate (pre)filter of $\mathcal{E}$.*

THEOREM 2.19. *Let $\mathcal{E}$ be a good EQ-algebra and μ be a normalized fuzzy (pre)filter of $\mathcal{E}$. Then μ is a fuzzy n-fold obstinate (pre)filter of $\mathcal{E}$ if and only if every normalized fuzzy (pre)filter of quotient algebra $\mathcal{E}/\mu$ is a fuzzy n-fold obstinate (pre)filter of $\mathcal{E}/\mu$.*

References

1. M. El-Zekey, *Representable good EQ-algebra*, Soft Comput. **14** (2009) 1011–1023.
2. M. El-Zekey, V. Novák and R. Mesiar, *On good EQ-algebras*, Fuzzy Sets Syst. **178** (2011) 1–23.
3. B. Ganji Saffar, M. Aaly Kologani and R. A. Borzooei, *n-fold filters of EQ-algebras*, submitted.
4. B. Ganji Saffar, G. Muhiuddin, M. Aaly Kologani and R. A. Borzooei, *Construction of (n-fold) EQ-algebras by using fuzzy n-fold filters*, submitted.
5. L. Z. Liu and X. Y. Zhang, *Implicative and positive implicative prefilters of EQ-algebras*, J. Intell. Fuzzy Systems **26** (5) (2014) 2087–2097.
6. V. Novák and B. De Baets, *EQ-algebras*, Fuzzy Sets Syst. **160** (20) (2009) 2956–2978.
7. X. L. Xin, P. F. He and Y. W. Yang, *Characterizations of some fuzzy prefilters (filters) in EQ-algebras*, Sci. World J. (2014) 829527.
8. L. A. Zadeh, *Fuzzy sets*, Inform. Contr. **8** (1965) 338–353.

E-mail: bganji@alzahra.ac.ir

Multi-Strategy Decision-Making On Enhancing Customer Acquisition Using Neutrosophic Soft Relational Maps

Nivetha Martin*
Arul Anandar College (Autonomous), Karumathur, Madurai, Tamil Nadu, India
Florentin Smarandache
Department of Mathematics and Science, University of New Mexico, Gallup, NM 87301, USA
and Akbar Rezaei
Department of Mathematics, Payame Noor University, P. O. Box 19395-3697, Tehran, Iran

Abstract. Decision making by the business managerial on framing strategies to foster customer acquisition is a challenging task. The aim of this paper is to introduce a new method of Multi-Strategy Decision-Making (MSDM) integrated with neutrosophic soft relational maps to determine the significant and feasible strategies of customer acquisition and their inter impacts. The proposed method comprises of two-stage processes and it is validated with twenty strategies, five factors associated with customer acquisition and expert 's opinion based on multivalued neutrosophic soft sets.
Keywords: Multi-Strategy, Decision-Making, Neutrosophic soft sets, Relational maps.
AMS Mathematical Subject Classification [2010]: 94Dxx, 90B50.

1. Introduction

Decision theory is characterized by various Multi-Criteria Decision making (MCDM) (otherwise called as Multi-Objective or Multi-Attribute or Multi-Dimension Decision-Making) methods such as Analytical Hierarchy Process, ELECTRE, COPRAS, PROMTHEE, TOPSIS, SAW. MCDM methods are used in selection of alternatives subjected to criteria satisfaction. MCDM methods are extended to Fuzzy MCDM to handle uncertainty in decision making. The criterion – alternative association is represented as fuzzy values in fuzzy MCDM. Wang et al. developed Fuzzy MCDM method for sustainable supplier selection and evaluation. Peng et al. [10], Saini et al. [12] developed intuitionistic MCDM (IFMCDM) approaches with intuitionistic representation comprising of membership and non-membership values. Neutrosophic sets introduced by Smarandache [13] comprises of truth , indeterminacy and falsity values and it has been extensively used in MCDM. Athar [5], Abdel-Basset [1, 2], Nada et al. [9], Garg et al. [6] developed neutrosophic MCDM models with neutrosophic representations of criterion alternative association. Another kind of sets that also play a key role in decision making is Soft sets introduced by Molodtsov [8], which was later extended to fuzzy

*Speaker

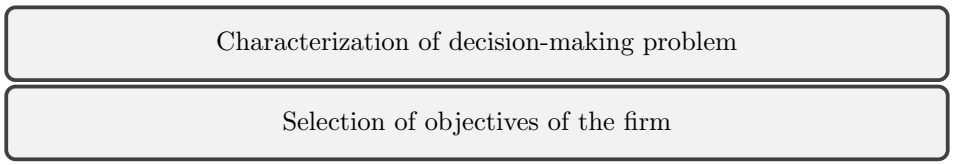
soft sets by Maji [7]. Dey et al. [3] presented the applications of multi-fuzzy soft sets in decision-making. Tripathy et al. [14] described the key role of intuitionistic fuzzy soft sets in group decision making. Faruk Karaaslan [4] elicited the implications of neutrosophic soft sets in decision making. Abu and Omar [11] extended neutrosophic soft sets to Q-neutrosophic soft sets and these sets are applied in comprehensive decision-making. In these neutrosophic soft MCDM models, the optimal ranking of the alternatives are determined. But these model do not cater to determine the impact of exercising the alternatives.

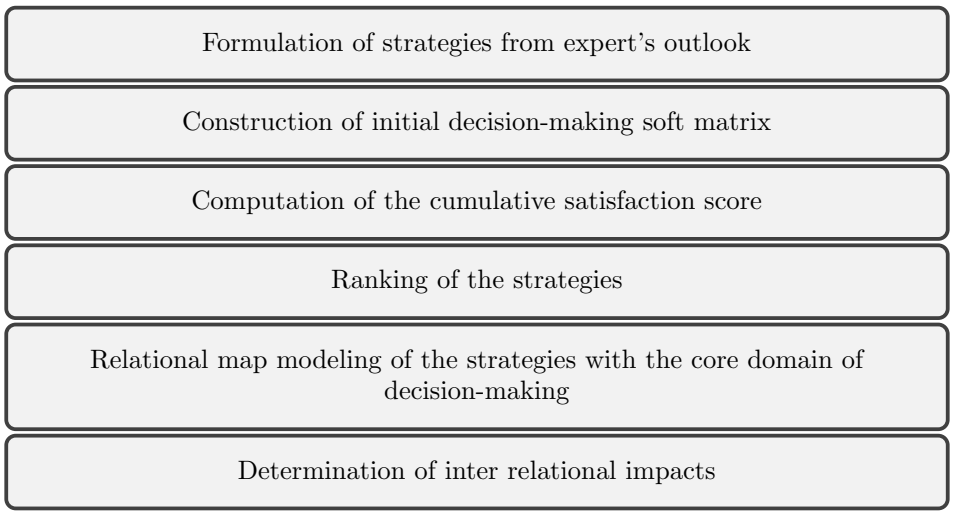
In this paper the new decision making approach based on MCDM is developed with the replacement of alternatives by strategies to make decisions and the criteria by the objectives to be fulfilled. The proposed method comprises of two-stage processes. The first stage ranks the proposed alternatives based on criteria satisfaction rate with the representation of neutrosophic soft sets and in the second stage the chosen alternatives are associated with the principles of decision making using neutrosophic soft relational maps. The integration of soft sets in relational maps is an innovative initiative of this research work. The proposed two-stage decision making process is a ground-breaking endeavor and it is validated by applying to decision making on customer acquisition strategies. Though researchers have explored strategically decision- making in various perspectives, the mathematical approach of strategy selection has not been explored so far to the best of our knowledge and this research work is an opening to it. The content of the paper is organized as follows: the methodology is presented in Section 2, the application of the proposed approach is validated in Section 3, the results are discussed in Section 4, the last section concludes the work.

2. Materials and Methods

This section presents the significance and need of MSDM and the algorithmic approach of determining optimal solution.

2.1. Multi-Strategy Decision-Making. In the approach of MSDM, the primary aim is to rank the strategies. In general, all the productions sectors construct their goals and work towards accomplishing the same. The managerial formulate strategies to achieve the goals, but the major challenge is selection and implementation of feasible strategies to yield optimum benefits. The decision-making environment does not involve only selection of alternatives with respect to criteria satisfaction, rather it involves the other dimension of choosing the right optimizing strategies. Strategical decision making is another dominating phenomenon and it has to be focused and this is how the approach of MSDM has evolved. In this new approach the method of finding the optimal strategy is a two-step process. The first step ranks the strategies and the second step associates their inter relationship with the principles of decision making. The steps are as follows:





3. Application of the Proposed MSDM Approach

This section applies the proposed two stage processes of MSDM to the decision making on customer acquisition strategies based on expert's opinion presented as below.

S_1 Selection of Advertising medium to propagate the product,
S_2 Designing user friendly products,
S_3 Customizing the product's utility to the needs of the buyers,
S_4 Attending to the diverse needs of the customers,
S_5 Developing multi-faceted products reflecting the ethos of the customers,
S_6 Scaling the cost of the product to customer's budget,
S_7 Periodic Propagation of the attributes of the product,
S_8 Product outlook modification,
S_9 Creating smart products,
S_{10} Developing innovative kind of products suiting the dynamic needs of the consumers,
S_{11} Create an ambiance to purchase product by providing offers,
S_{12} Communicating the attributes of the product to the customers,
S_{13} On line engagement with the customers,
S_{14} Establishing Trade mark of the product,
S_{15} Provision of various kinds of payment portals,
S_{16} Enrichment of the quality of the product using modern technology,
S_{17} Strengthening the consistency and reliability of the product,
S_{18} Designing products with values adding to consumer's image,
S_{19} Periodical review of product sales and marketing,
S_{20} Integrating eco-friendly characteristics with the products.

In the perspective of soft sets, let $U = \{S_1, S_2, \ldots, S_{20}\}$ and $A = \{A_1, A_2, \ldots, A_5\}$ be the set of purchasing behavior influencing factors, where A_1 = Psychological, A_2 = Personal, A_3 = Product, A_4 = Social and A_5 = Cultural.

A multivalued neutrosophic soft mapping $G : A \longrightarrow P(U)$ is represented as follows:

$$
\begin{aligned}
G(A_1) = \{ & \frac{\langle (0.9,0.1,0.2),(0.8,0.3,0.2),(0.9,0.1,0.2)\rangle}{S_1}, \frac{\langle (0.6,0.3,0.3),(0.6,0.1,0.3),(0.6,0.3,0.3)\rangle}{S_2}, \\
& \frac{\langle (0.8,0.3,0.5),(0.9,0.3,0.5),(0.8,0.3,0.5)\rangle}{S_3}, \frac{\langle (0.6,0.2,0.3),(0.6,0.2,0.3),(0.6,0.2,0.3)\rangle}{S_4}, \\
& \frac{\langle (0.7,0.5,0.2),(0.6,0.4,0.2),(.07,0.5,0.2)\rangle}{S_5}, \frac{\langle (0.9,0.1,0.1),(0.9,0.1,0.1),(0.9,0.1,0.1)\rangle}{S_6}, \\
& \frac{\langle (0.8,0.3,0.5),(0.8,0.2,0.5),(0.8,0.3,0.5)\rangle}{S7}, \frac{\langle (0.6,0.4,0.4),(0.6,0.4,0.3),(0.6,0.4,0.4)\rangle}{S_8}, \\
& \frac{\langle (0.7,0.5,0.2),(0.6,0.1,0.2),(0.7,0.5,0.2)\rangle}{S_9}, \frac{\langle (0.6,0.4,0.3),(0.6,0.4,0.3),(0.6,0.4,0.3)\rangle}{S_{10}}, \\
& \frac{\langle (0.9,0.1,0.1),(0.9,0.1,0.1),(0.9,0.1,0.1)\rangle}{S_{11}}, \frac{\langle (0.9,0.1,0.2),(0.9,0.1,0.1),(0.9,0.1,0.2)\rangle}{S_{12}}, \\
& \frac{\langle (0.8,0.3,0.5),(0.8,0.3,0.5),(0.8,0.3,0.5)\rangle}{S_{13}}, \frac{\langle (0.9,0.1,0.2),(0.9,0.1,0.2),(0.9,0.1,0.2)\rangle}{S_{14}}, \\
& \frac{\langle (0.8,0.3,0.5),(0.8,0.2,0.4),(0.8,0.3,0.5)\rangle}{S_{15}}, \frac{\langle (0.6,0.4,0.3),(0.6,0.5,0.3),(0.6,0.4,0.3)\rangle}{S_{16}}, \\
& \frac{\langle (0.8,0.3,0.5),(0.8,0.2,0.5),(0.8,0.3,0.5)\rangle}{S_{17}}, \frac{\langle (0.8,0.3,0.5),(0.8,0.2,0.5),(0.8,0.3,0.5)\rangle}{S_{18}}, \\
& \frac{\langle (0.6,0.4,0.3),(0.6,0.4,0.4),(0.6,0.4,.3)\rangle}{S_{19}}, \frac{\langle (0.7,0.5,0.2),(0.7,0.5,0.1),(0.7,0.5,0.2)\rangle}{S_{20}} \},
\end{aligned}
$$

$$
\begin{aligned}
G(A_2) = \{ & \frac{\langle (0.7,0.5,0.2),(0.6,0.4,0.2)),(0.7,0.5,0.2)\rangle}{S_1}, \frac{\langle (0.7,0.5,0.2),(0.9,0.1,0.3),(0.9,0.1,0.2)\rangle}{S_2}, \\
& \frac{\langle (0.8,0.2,0.4),(0.8,0.2,0.3),(0.8,0.2,0.4)\rangle}{S_3}, \frac{\langle (0.9,0.1,0.2),(0.9,0.3,0.2),(0.9,0.1,0.2)\rangle}{S_4}, \\
& \frac{\langle (0.8,0.2,0.4),(0.7,0.2,0.4),(0.8,0.2,0.4)\rangle}{S_5}, \frac{\langle (0.6,0.4,0.3),(0.6,0.4,0.4),(0.6,0.4,0.3)\rangle}{S_6}, \\
& \frac{\langle (0.6,0.4,0.3),(0.6,0.3,0.3),(0.6,0.4,0.3)\rangle}{S7}, \frac{\langle (0.6,0.4,0.3),(0.6,0.2,0.3),(0.6,0.4,0.3)\rangle}{S_8}, \\
& \frac{\langle (0.8,0.2,0.4),(0.7,0.2,0.4),(0.8,0.2,0.4)\rangle}{S_9}, \frac{\langle (0.9,0.2,0.3),(0.9,0.2,0.3),(0.9,0.2,0.3)\rangle}{S_{10}}, \\
& \frac{\langle (0.8,0.2,0.4),(0.8,0.2,0.4),(0.8,0.2,0.4)\rangle}{S_{11}}, \frac{\langle (0.6,0.4,0.3),(0.6,0.4,0.3),(0.6,0.4,0.3)\rangle}{S_{12}}, \\
& \frac{\langle (0.9,0.1,0.1),(0.9,0.2,0.1),(0.9,0.1,0.1)\rangle}{S_{13}}, \frac{\langle (08,0.2,0.4),(0.7,0.2,0.4),(0.8,0.2,0.4)\rangle}{S_{14}}, \\
& \frac{\langle (0.9,0.1,0.2),(0.8,0.1,0.2)(0.9,0.1,0.2)\rangle}{S_{15}}, \frac{\langle (0.9,0.1,0.1),(0.9,0.1,0.2),(0.9,0.1,0.1)\rangle}{S_{16}}, \\
& \frac{\langle (0.6,0.4,0.3),(0.6,0.4,0.2),(0.6,0.4,0.3)\rangle}{S_{17}}, \frac{\langle (0.9,0.1,0.1),(0.9,0.1,0.3),(0.9,0.1,0.1)\rangle}{S_{18}}, \\
& \frac{\langle (0.9,0.1,0.1),(0.9,0.1,0.1)\rangle,(0.9,0.1,0.1)\rangle}{S_{19}}, \frac{\langle (0.8,0.2,0.4),(0.8,0.1,0.3),(0.8,0.2,0.4)\rangle}{S_{20}} \},
\end{aligned}
$$

$$
\begin{aligned}
G(A_3) = \{ & \frac{\langle (0.8,0.3,0.5),(0.8,0.1,0.3),(0.8,0.3,0.5)\rangle}{S_1}, \frac{\langle (0.8,0.2,0.4),(0.7,0.2,0.4),(0.8,0.2,0.4)\rangle}{S_2}, \\
& \frac{\langle (0.5,0.4,0.6),(0.6,0.4,0.3),(0.5,0.4,0.6)\rangle}{S_3}, \frac{\langle (0.8,0.2,0.4),(0.7,0.5,0.3),(0.8,0.2,0.4)\rangle}{S_4}, \\
& \frac{\langle (0.9,0.2,0.3),(0.7,0.5,0.3),(0.9,0.2,0.3)\rangle}{S_5}, \frac{\langle (0.7,0.5,0.2),(0.9,0.3,0.2),(0.7,0.5,0.2)\rangle}{S_6}, \\
& \frac{\langle (0.6,0.4,0.3),(0.6,0.3,0.3),(0.6,0.4,0.3)\rangle}{S7}, \frac{\langle (0.7,0.5,0.2),(0.8,0.5,0.2),(0.7,0.5,0.2)\rangle}{S_8}, \\
& \frac{\langle (0.9,0.2,0.3),(0.8,0.1,0.4),(0.9,0.2,0.3)\rangle}{S_9}, \frac{\langle (0.8,0.2,0.4),(0.7,0.3,0.2),(0.8,0.2,0.4)\rangle}{S_{10}}, \\
& \frac{\langle (0.8,0.2,0.4),(0.8,0.5,0.2),(0.8,0.2,0.4)\rangle}{S_{11}}, \frac{\langle (0.6,0.4,0.3),(0.4,0.5,0.2),(0.7,0.5,0.2)\rangle}{S_{12}},
\end{aligned}
$$

$$\frac{\langle(0.9,0.1,0.2),(0.9,0.1,0.3),(0.7,0.5,0.2)\rangle}{S_{13}},\frac{\langle(0.9,0.1,0.2),(0.9,0.1,0.3),(0.7,0.5,0.2)\rangle}{S_{14}},$$
$$\frac{\langle(0.6,0.4,0.3),(0.6,0.4,0.3),(0.7,0.5,0.2)\rangle}{S_{15}},\frac{\langle(0.9,0.2,0.3),(0.7,0.5,0.1),(0.7,0.5,0.2)\rangle}{S_{16}},$$
$$\frac{\langle(0.9,0.1,0.2),(0.7,0.5,0.1),(0.7,0.5,0.2)\rangle}{S_{17}},\frac{\langle(0.6,0.4,0.3),(0.7,0.5,0.1),(0.7,0.5,0.2)\rangle}{S_{18}},$$
$$\frac{\langle(0.7,0.5,0.2),(0.9,0.2,0.2),(0.9,0.2,0.3)\rangle}{S_{19}},\frac{\langle(0.9,0.1,0.1),(0.6,0.4,0.4),(0.6,0.4,0.3)\rangle}{S_{20}}\},$$

$$G(A_4) = \{\frac{\langle(0.6,0.4,0.3),(0.5,0.2,0.3),(0.6,0.4,0.3)\rangle}{S_1},\frac{\langle(0.9,0.1,0.1),(0.9,0.1,0.1),(0.9,0.1,0.1)\rangle}{S_2},$$
$$\frac{\langle(0.7,0.5,0.2),(0.7,0.4,0.2),(0.7,0.5,0.2)\rangle}{S_3},\frac{\langle(0.7,0.5,0.2),(0.7,0.5,0.3),(0.7,0.5,0.2)\rangle}{S_4},$$
$$\frac{\langle(0.7,0.5,0.2),(0.7,0.5,0.3),(0.7,0.5,0.2)\rangle}{S_5},\frac{\langle(0.9,0.1,0.2),(0.9,0.3,0.2),(0.9,0.1,0.2)\rangle}{S_6},$$
$$\frac{\langle(0.5,0.4,0.6),(0.5,0.4,0.7),(0.5,0.4,0.6)\rangle}{S_7},\frac{\langle(0.7,0.5,0.2),(0.8,0.5,0.2),(0.7,0.5,0.2)\rangle}{S_8},$$
$$\frac{\langle(0.8,0.2,0.4),(0.8,0.1,0.4),(0.8,0.2,0.4)\rangle}{S_9},\frac{\langle(0.7,0.5,0.2),(0.7,0.3,0.2),(0.7,0.5,0.2)\rangle}{S_{10}},$$
$$\frac{\langle(0.7,0.5,0.2),(0.8,0.5,0.2),(0.7,0.5,0.2)\rangle}{S_{11}},\frac{\langle(0.7,0.5,0.2),(0.4,0.5,0.2),(0.7,0.5,0.2)\rangle}{S_{12}},$$
$$\frac{\langle(0.9,0.1,0.2),(0.9,0.1,0.3),(0.9,0.1,0.2)\rangle}{S_{13}},\frac{\langle(0.9,0.4,0.3),(0.7,0.2,0.3),(0.6,0.4,0.3)\rangle}{S_{14}},$$
$$\frac{\langle(0.7,0.5,0.2),(0.6,0.4,0.3),(0.7,0.5,0.2)\rangle}{S_{15}},\frac{\langle(0.7,0.5,0.2),(0.7,0.5,0.1),(0.9,0.2,0.3)\rangle}{S_{16}},$$
$$\frac{\langle(0.7,0.5,0.2),(0.7,0.5,0.1),(0.9,0.1,0.2)\rangle}{S_{17}},\frac{\langle(0.7,0.5,0.2),(0.7,0.5,0.4),(0.6,0.4,0.3)\rangle}{S_{18}},$$
$$\frac{\langle(0.9,0.2,0.3),(0.9,0.2,0.2),(0.9,0.2,0.3)\rangle}{S_{19}},\frac{\langle(0.6,0.2,0.3),(0.9,0.2,0.1),(0.9,0.2,0.3)\rangle}{S_{20}}\}$$

and

$$G(A_5) = \{\frac{\langle(0.9,0.2,0.3),(0.9,0.1,0.2),(0.9,0.2,0.3)\rangle}{S_1},\frac{\langle(0.7,0.5,0.2),(0.8,0.5,0.2),(0.7,0.5,0.2)\rangle}{S_2},$$
$$\frac{\langle(0.8,0.2,0.4),(0.7,0.2,0.4),(0.8,0.2,0.4)\rangle}{S_3},\frac{\langle(0.9,0.2,0.3),(0.9,0.2,0.4),(0.9,0.2,0.3)\rangle}{S_4},$$
$$\frac{\langle(0.8,0.2,0.4),(0.8,0.2,0.5),(0.8,0.2,0.4)\rangle}{S_5},\frac{\langle(0.7,0.5,0.2),(0.8,0.5,0.2),(0.7,0.5,0.2)\rangle}{S_6},$$
$$\frac{\langle(0.9,0.2,0.3),(0.9,0.2,0.1),(0.9,0.2,0.3)\rangle}{S_7},\frac{\langle(0.8,0.2,0.4),(0.8,0.2,0.4),(0.8,0.2,0.4)\rangle}{S_8},$$
$$\frac{\langle(0.9,0.1,0.1),(0.9,0.2,0.1),(0.9,0.1,0.1)\rangle}{S_9},\frac{\langle(0.9,0.2,0.3),(0.8,0.2,0.3),(0.9,0.2,0.3)\rangle}{S_{10}},$$
$$\frac{\langle(0.8,0.2,0.4),(0.8,02,0.4),(0.8,0.2,0.4)\rangle}{S_{11}},\frac{\langle(0.8,0.2,0.4),(0.7,0.2,0.4),(0.8,0.2,0.4)\rangle}{S_{12}},$$
$$\frac{\langle(0.7,0.5,0.2),(0.8,0.3,0.2),(0.7,0.5,0.2)\rangle}{S_{13}},\frac{\langle(0.9,0.1,0.1),(0.9,0.2,0.1),(0.9,0.1,0.1)\rangle}{S_{14}},$$
$$\frac{\langle(0.7,0.5,0.2),(0.7,0.4,0.2),(0.7,0.5,0.2)\rangle}{S_{15}},\frac{\langle(0.8,0.2,0.4),(0.8,0.2,0.3),(0.8,0.2,0.4)\rangle}{S_{16}},$$
$$\frac{\langle(0.9,0.2,0.3),(0.8,0.2,0.3),(0.9,0.2,0.3)\rangle}{S_{17}},\frac{\langle(0.9,0.2,0.3),(0.9,0.2,0.2),(0.9,0.2,0.3)\rangle}{S_{18}},$$
$$\frac{\langle(0.8,0.2,0.4),(0.8,0.2,0.3),(0.8,0.2,0.4)\rangle}{S_{19}},\frac{\langle(0.9,0.2,0.3),(0.8,0.1,0.3),(0.9,0.2,0.3)\rangle}{S_{20}}\}.$$

The score values of each of the strategies with respect to the respective association with the factors are determined by using the algorithm was discussed in [5] (See Figure 1). The following factors are considered as the core factors for the

FIGURE 1. Ranking of the Factors.

next step.

CS_1 Developing multi-faceted products reflecting the ethos of the customers,
CS_2 Scaling the cost of the product to customer's budget,
CS_3 Enrichment of the quality of the product using modern technology,
CS_4 Strengthening the consistency and reliability of the product,
CS_5 Designing products with values adding to consumer's image,
CS_6 Periodical review of product sales and marketing.

These factors are related to the various management systems of the business. The relational impacts are represented linguistic neutrosophic sets and are quantified using neutrosophic triangular fuzzy number as presented in Table 1.

TABLE 1. Quantification of Linguistic Variable.

Linguistic Variable	**Neutrosophic Triangular Number**	**Crisp Value**
Very Low (VL)	((0,0.10,0.15,0.20),0.6,0.2,0.3)	0.06
Low (L)	((0.15,0.2,0.25,0.3),0.6,0.1,0.1)	0.14
Medium (M)	((0.3,0.35,0.4,0.5),0.7,0.1,0.2)	0.23
High (H)	((0.5,0.6,0.7,0.8),0.8,0.2,0.1)	0.41
Very High (VH)	((0.8,0.9,0.95,1),0.9,0.1,0.1)	0.62

Let $U = \{CS_1, CS_2, \ldots, CS_6\}$ and $M = \{M_1, M_2, M_3, M_4\}$ be the set of management systems of business, where

$$\begin{aligned} M_1 &= \text{Product Quality Management,} \\ M_2 &= \text{Customer Loyalty Management,} \\ M_3 &= \text{Customer Relationship Management,} \\ M_4 &= \text{Marketing Management.} \end{aligned}$$

A single valued neutrosophic soft mapping $H : M \longrightarrow P(U)$ is represented as follows:

$$H(M_1) = \{\frac{VH}{CS_1}, \frac{L}{CS_2}, \frac{VH}{CS_3}, \frac{H}{CS_4}, \frac{M}{CS_5}, \frac{H}{CS_6}\},$$

TABLE 2. Fixed points of the vectors.

Initial Vector	**Fixed Point**
$X = (100000)$	$X^*M = (0.620.410.410.14)(1110) := X1$ $X1^*MT = (1.440.691.441.651.470.87) = (100110) := Y$ $Y^*M = (1.261.651.650.78)(1110) := X2$ $X2^*MT = (1.440.691.441.651.470.87) = (100110) := Y1$ $(1110)(100110)$
$X = (010000)$	$X^*M = (0.140.140.410.23)(0011) := X_1$ $X1^*MT = (0.550.640.641.030.850.85) = (010111) := Y$ $Y^*M = (1.191.611.881.49)(0110) := X_2$ $X_2^*MT = (0.820.550.821.241.240.46) = (111110) := Y_1$ $Y_1^*M = (2.022.22.471.24)(0110) := X3$ $X_3^*MT = (0.820.550.821.241.240.46) = (111110) := Y_2$ $(0110)(111110)$
$X = (001000)$	$X^*M = (0.620.410.410.23)(1110) := X_1$ $X_1^*MT = (1.440.691.441.651.470.87) = (100110) := Y$ $Y^*M = (1.261.651.650.78)(1110) := X_2$ $X_2^*MT = (1.440.691.441.651.470.87) = (100110) := Y1$ $(1110)(100110)$
$X = (000100)$	$X^*M = (0.410.620.620.41)(1111) := X_1$ $X_1^*MT = (1.580.921.672.061.71.49) = (000110) := Y$ $Y^*M = (0.641.241.240.64)(1111) := X_2$ $X_2^*MT = (1.580.921.672.061.71.49) = (000110) := Y_1$ $(1111)(000110)$
$X = (000010)$	$X^*M = (0.230.620.620.23)(1111) := X_1$ $X_1^*MT = (1.580.921.672.061.71.49) = (000110) := Y$ $Y^*M = (0.641.241.240.64)(1111) := X_2$ $X_2^*MT = (1.580.921.672.061.71.49) = (000110) := Y_1$ $(1111)(000110)$
$X = (000001)$	$X^*M = (0.410.230.230.62)(1001) := X_1$ $X_1^*MT = (0.760.370.850.820.461.03) = (001001) := Y$ $Y^*M = (1.030.640.640.85)(1001) := X_2$ $X_2^*MT = (0.760.370.850.820.461.03) = (001001) := Y_1$ $(1001)(001001)$

$$
\begin{aligned}
H(M_2) &= \{\frac{H}{CS_1}, \frac{L}{CS_2}, \frac{H}{CS_3}, \frac{VH}{CS_4}, \frac{VH}{CS_5}, \frac{M}{CS_6}\},\\
H(M_3) &= \{\frac{H}{CS_1}, \frac{H}{CS_2}, \frac{H}{CS_3}, \frac{VH}{CS_4}, \frac{VH}{CS_5}, \frac{M}{CS_6}\},\\
H(M_4) &= \{\frac{L}{CS_1}, \frac{M}{CS_2}, \frac{M}{CS_3}, \frac{H}{CS_4}, \frac{M}{CS_5}, \frac{VH}{CS_6}\}.
\end{aligned}
$$

The relational impacts are determined by using the procedure discussed in [15] (See Table 2).

4. Results and Discussions

The multivalued neutrosophic soft representation takes in the opinion of three experts into consideration. The twenty strategies taken for study are confined to six strategies based on the final scores of the association rate with the factors. The

six core factors are related with the principles of business management in various dimensions. Each of the core factors is kept in on position. The associational impacts are analyzed and the fixed points are determined. If the core factor CS_1 is kept in on position, the limit point (1110)(100110) is obtained. The factor CS_1 is highly associated with CS_4, CS_5 and M_1, M_2, M_3. By repeating the same mechanism, the associational impacts between the other core factors are determined. This approach of Multi-Strategy Decision-Making with neutrosophic soft sets representations facilitate the decision-making process and it eases the procedure of minimizing the number of strategies. The decision makers evolve many strategies, but implementing all the strategies is not possible, it is quite mandatory to explore the core strategies and to detect its relation with other decision-making principles. To make the process much comprehensive, MSDM approach is constructed in this research work.

5. Conclusion

This paper introduces the approach of Multi-Strategy Decision-Making with two stage process of decision-making. The proposed approach is validated with the decision-making environment of enhancing the customer acquisition strategies. The multivalued neutrosophic soft set representations in the first stage results in confining the number of strategies and the neutrosophic soft relational maps in the second stage is used to determine the relational impacts. This approach can be extended with other kinds of representation. This MSDM approach can be applied to any kind of decision-making environment.

References

1. M. Abdel-Baset, V. Chang and A. Gamal, *Evaluation of the green supply chain management practices: A novel neutrosophic approach*, Comput. Ind. **108** (2019) 210–220.
2. M. Abdel-Basset, N. A. Nabeeh, H. A. El-Ghareeb and A. Aboelfetouh, *Utilizing neutrosophic theory to solve transition difficulties of IoT-based enterprises*, Enterprise Inform. Syst. **14** (9-10) (2020) 1304–1324.
3. A. Dey and M. Pal, *Generalized multi fuzzy soft set and its application in decision making,* Pacific. Sci. Rev. A: Nat. Sci. Eng. **17** (1) (2015) 23–28.
4. F. Karaaslan, *Neutrosophic soft sets with applications in decision making,* Int. J. Inform. Sci. Intell. Sys. **4** (2) (2015) 1–20.
5. A. Kharal, *A neutrosophic multi Criteria decision making method,* New Math. Nat. Comput. **10** (2) (2014) 143–162.
6. H. Garg and Nancy, *Linguistic single-valued neutrosophic power aggregation operators and their applications to group decision-making problems,* IEEE/CAA J. Autom. Sinica **7** (2) (2020) 546–558.
7. P. k. Maji, R. Biswas and A. R. Roy, *Fuzzy soft sets,* J. Fuzzy Math. **9** (3) (2001) 589–602.
8. D. Molodtsov, *Soft set theory-first results,* Comput. Math. Appl. **37** (1999) 19–31.
9. N. A. Nabeeh, M. Abdel-Basset, H. A. El-Ghareeb and A. Aboelfetouh, *Neutrosophic multi-criteria decision making approach for IoT-based enterprises*, New Math. Nat. Comput. **15** (2) (2017) 307–326.
10. J. J. Peng, J. Q. Wang, J. Wang and X. H. Chen, *Multi criteria decision-making approach with hesitant interval-valued intuitionistic fuzzy sets,* Sci. World J. **2014** (2014) 868515.
11. A. Qamar and N. Hassan, *An approach toward a Q-neutrosophic soft set and its application in decision making,* Symmetry **11** (2) (2019) 139.
12. N. Saini, N. Gandotra and R. Kumar, *Multi criteria decision making under fuzzy, intuitionistic and interval-valued intuitionistic fuzzy environment: A review,* In: A. Kumar, S. Mozar

(Eds), ICCCE 2020. Lecture Notes in Electrical Engineering, Vol. 698, Springer, Singapore, 2021.
13. F. Smarandache, *Neutrosophic set, a generalization of the intuitionistic fuzzy sets,* Int. J. Pure Appl. Math. **24** (2005) 287–297.
14. B. K. Tripathy, R. K. Mohanty and T. R. Sooraj, *On intuitionistic fuzzy soft set and its application in group decision making*, Int. Conf. Emerging Trends Eng. Tech. Sci. (ICETETS), (2016) pp. 1–5.
15. W. B. Vasantha Kandaswamy and Y. Sultana, *FRM to analyses the Employee-Employer relationship model*, J. Bihar Math. Soc. **21** (2001) 25–34.

E-mail: nivetha.martin710@gmail.com
E-mail: smarand@unm.edu
E-mail: rezaei@pnu.ac.ir

Contributed Talks

Computer Science

MLIPD: A Machine Learning Approach to Identify Party and Date Hub in PPI Network

Mahnaz Habibi*
Department of Mathematics, Qazvin Branch, Islamic Azad University, Qazvin, Iran

Abstract. It has been claimed that protein interaction networks are scale free that contain a few hubs with ability to bind multiple proteins. Hubs are classified as party and date hubs. Party hubs generally bind different proteins in specific module simultaneously, while date hubs interact with multiple proteins in different modules at different times and locations. Generally, they have been divided into two classes based on the average Pearson Correlation Coefficient (avPCC) of expression over all partners or their functions. In this study, we propose a more appropriate method to identify party and date hubs based on their topological properties of network. First, we calculate some topological properties for each vertex of network. Then, using support vector machine approach, we train a model on the entire training dataset to identify party and date hubs. Finally, we evaluate our method on reference hubs based on the avPCC on network. We show that the combination of topological properties can improve the performance of each topological property approach.

Keywords: Date hub, Party hub, PPI network, Support Vector Machine.
AMS Mathematical Subject Classification [2010]: 94C15.

1. Introduction

Proteins are identified as the main agent of biological processes that can determine the phenotype of organisms. Some proteins are functional isolated form and some ones interact with other proteins or other molecules. These interactions between the proteins are often represented in the form of Protein-Protein Interaction (PPI) network [1, 7]. Since some proteins interact with multiple proteins and others interact with only a few, the PPI network has a wide range of degrees. The highly connected proteins in PPI network are referred as hubs. There are some studies that reveal the functional and structural characterization of hubs in PPI network [3, 6]. In the recent pandemic, Covid-19 (coronavirus disease), the role of hub proteins in virus-host interaction network have been highlighted to study pathogenesis of infection [8]. Prasad et al have analyzed the human PPI network and targeted hub proteins to find candidate drugs for Covid-19 [10].

In 2004, Han et al. have expressed that hub proteins which interact with most of their partners simultaneously are designated as party hubs and those that bind their different partners at different times or locations are date hubs [5]. Briefly, they have divided the highly connected proteins into two classes based on the average Pearson Correlation Coefficient (avPCC) of expression over all partners. They have studied the role of two types of hubs in molecular organization in a

*Speaker

cell. They have also shown that party and date hubs play an important role in organizing modules. Recently, some topological properties of two types of hub proteins in the PPI network are presented [2, 9].

In this study, we first present MLMLIPD algorithm (Machine Learning approach), a computational method based on a combination of topological properties of PPI network to classify party and date hubs. First, we transform the PPI network from the co-expression gene network. For each vertex, the topological properties of PPI network are calculated. We train a model on the entire training dataset based on SVM (Support Vector Machine) method and make predictions on the test dataset to classify party and date hubs. Then, we evaluate each topological property and their combination with respect to precision, recall and F-measure. Results show that the combination of these properties performs better than each existing property on PPI network.

2. Method

A PPI network $G = \langle V, E\rangle$ is a set of vertices V and a set of undirected edges E between the vertices. Generally, a vertex v of PPI network represents a given protein and each edge uv between two vertices u and v represents the connection between two proteins. A vertex u is neighbor of another vertex v, if uv is an edge of G. The set of all vertices that are the neighbors of v is the neighborhood of v and it is denoted by $N(v)$. The number of vertices of $N(v)$ is called the degree of v and denoted by $d(v)$.

A sequence $u = u_0, u_1, lcdots, u_n = v$ of distinct vertices of non-empty network G is a path between two vertices u and v, if u_iu_{i+1} for each $0 \leq i \leq n$ is an edge of G. The number of edges of a path is the length of the path. The shortest path between two vertices u and v is defined as a path with the minimum length. It is denoted by $d(u, v)$.

2.1. Representative Topological Properties. The Clustering Coefficient (CC) for each vertex v of $G =< V, E >$ is defended as following formula:

$$CC(v) = \frac{2|E(H)|}{d(v)(d(v)-1)},$$

where $H = N(v)$ and $|E(H)|$ is the number of $\{uw \in E; u, w \in N(v)\}$.

The Closeness centrality (Cl) measure for each vertex v of network $G =< V, E >$ is defined by:

$$Cl(v) = \frac{|V|-1}{\sum_{u\in V} d(u, v)}.$$

Another centrality measure of each vertex v on network $G =< V, E >$ is Sub-graph Centrality (SC). The sub-graph centrality for each vertex is defined as following formula:

$$SC(v) = \sum_{k=1}^{\infty} \delta_k(v),$$

where $\delta_k(v)$ is the number of path with length k that pass through v.

Finally, The Mean Degree Neighbor (MDN) for each vertex is calculated as following formula:

$$MDN(v) = \frac{\sum_{u \in N(v)} d(u)}{|N(v)|}.$$

2.2. MLIPD Algorithm. In this work, we propose a new algorithm named MLIPD (Machine Learning approach to Identify Party and Date hubs) from input (gene expression data). In the first step of MLIPD algorithm, using CLR algorithm [4], the the co-expression gene network is constructed from gene expression dataset. Then, we transform the PPI network from the co-expression gene network. In the second procedure, the PPI network is collectively viewed as vertex-labeled graph, where a vertex labeling is the function of vertices to the set of topological property values. In the second step, we actually combine topological properties to find a model to identify party and date hubs. We train a linear Support Vector Machine (SVM) model on training examples and make prediction on test dataset.

2.3. Performance Evaluation Measures. To evaluate the performance of our method, we use some evaluation measures. These measures are based on the relation between the number of hubs correctly predicted positive (Tp), the number of hubs correctly predicted negative (Tn), the number of hubs incorrectly predicted positive (Fp), and the number of hubs incorrectly predicted negative (Fn). The Precision ($Pre = Tp/Tp + Fp$) and recall ($Re = Tp/Tp + Fn$) are two evaluation measures. Another measure which can be used to evaluate the performance of a method is F-measure as the harmonic mean of precision and recall.

$$F - measure = \frac{2PreRe}{Pre + Re}.$$

3. Result

3.1. Dataset. In this work, we use a collection of high-throughput protein interaction data of Saccharomyces cerevisiae [4]. It contains yeast cells which grown aerobically on galactose medium. CLR algorithm identifies 45869 regular interactions between 4445 genes. We consider ten percentages of high-degree vertices as hubs. We select the avPCC cutoff at the valley between the two peaks as threshold to separate date and party hubs. Ones with relatively high avPCCs are chosen as party hubs and the other ones are defined as date hubs. This yields 465 hubs that contain 127 party and 338 date hubs. In this study, we choose randomly 200 different training and test datasets on the each of two datasets by using of randperm function in the MATLAB to evaluate methods.

3.2. Evaluation of Topological Properties. To justify using some topological properties of PPI network, we analyze the performance of each topological property with respect to predict two types of hubs. To find suitable threshold for each property, we suppose that A and B be the set of party and date hubs in PPI network respectively and S^{ψ} be the set of score values of hubs for each property, ψ . For each threshold value, α, we define:

(1) $S^{\psi}_{<\alpha}$: The set of hub vertices that their score values (S^{ψ}) is less than α.

(2) $S^{\psi}_{\geq\alpha}$: The set of hub vertices that their score values (S^{ψ}) is more than α.

Now, for each threshold value and for each property we define the F^{ψ}_{α} measure as following formula:

$$F^{\psi}_{\alpha} = \frac{\max\{|A \cap S^{\psi}_{<\alpha}|, |A \cap S^{\psi}_{\geq\alpha}|\}.\max\{|B \cap S^{\psi}_{<\alpha}|, |B \cap S^{\psi}_{\geq\alpha}|\}}{|S^{\psi}_{<\alpha}||S^{\psi}_{\geq\alpha}|},$$

where $|\cdot|$ is the number of each subset.

Then, by varying the threshold value (α) on each topological property, we calculate the best threshold value corresponding to maximum number of F^{ψ}_{α} measure. The results of the best thresholds are given in Table 1 on dataset. The relatively F-measure values justify all selected properties to identify party and date hubs.

TABLE 1. Precision and recall values for date and party hubs using the best threshold values for each topological properties.

	Threshold	Tp	Tn	Fp	Fn	Pre	Re	F-measure	Classes
SC	1.70E+32	337	107	20	1	0.94	0.99	0.97	Date
		107	337	1	20	0.99	0.84	0.91	Party
MDN	61.06	337	105	23	1	0.9	0.99	0.96	Date
		105	337	1	23	0.99	0.85	0.89	Party
CC	0.52	325	110	17	13	0.95	0.96	0.95	Date
		110	325	13	17	0.89	0.86	0.88	Party
Cl	0.27	328	113	14	10	0.95	0.97	0.96	Date
		113	328	10	14	0.91	0.8	0.90	Party

3.3. Evaluating the Combination of Topological Properties. To evaluate the performance of our method, we compare the party and date hubs predicted by MLIPD, SC, MDN and Cl with the party and date hubs which obtained from avPCC definition as real party and date hubs. In this work, we train a model on the entire training dataset and make predictions on the test dataset. We also obtain the best threshold value for each property corresponding to maximum number of F^{ψ}_{α} measure on training sets. In Table 2, we show the means and variances of F-measure values which obtained by our method and each property on 200 different testing sets on dataset. Table 2 shows that almost all methods have similar performance to identify date hubs, however MLIPD algorithm performs better compare to other methods with respect to identify party hubs.To investigate the difference in behavior of methods corresponding to different training datasets, we study the variances of F-measure values which obtained from all methods on 200 different testing sets on dataset. Our results indicate that the variance values of MLIPD are $6.90E-05$ and $8.1E-04$ on dataset related to date and party hubs respectively (as shown Table 2). The relatively small values of variance (in MLIPD algorithm) indicate that the algorithm is independent of the selected training datasets. So, MLIPD algorithm perform superior performance compare to other methods on almost all testing sets.

TABLE 2. The mean and variance of F-measure values on 200 different testing sets.

	MLIPD	SC	MDN	CC	Cl	Classes
Mean	**0.97**	0.86	0.93	0.90	0.92	Date
	0.92	0.73	0.75	0.66	0.75	Party
Variance	**6.9E-5**	0.06	0.006	0.007	0.007	Date
	8.01E-4	0.12	0.12	0.13	0.12	Party

4. Conclusion

In the first part of this work, we have presented the short preliminaries of graph theory and mentioned some topological properties of each vertex in PPI network. Then, we have proposed the MLIPD algorithm to identify party and date hubs based on combination these topological properties. We have trained a linear support vector machine model on training sets and made prediction on testing set.

In the second part of this work, we have studied the impact of each topological property to identify party and date hubs. The results on testing sets show that the MLIPD algorithm can progressively improve the performance of each property to identify party and date hubs based on common evaluating parameters (Tp, Tn, Fp, Fn and F-measure). Results indicate that MLIPD algorithm agrees well with two hub classes obtained by the average Pearson Correlation Coefficient between hub and each of respective partners for mRNA expression, and F-measure values can be increased considerably in comparison with each property.

References

1. D. Alonso-López, F. J. Campos-Laborie, M. A. Gutiérrez, L. Lambourne, M. A. Calderwood, M. Vidal and J. De Las Rivas, *APID database: Redefining protein–protein interaction experimental evidences and binary interactomes*, Database **2019** (2019). DOI:10.1093/database/baz005
2. N. Bertin, N. Simonis, D. Dupuy, M. E. Cusick, J. D. Han, H. B. Fraser, F. P. Roth and M. Vidal, *Confirmation of organized modularity in the yeast interactome*, PLoS Biol. **5** (6) (2007). DOI:10.1371/journal.pbio.0050153
3. E. Cukuroglu, E. Ozkirimli and O. Keskin, *Structural properties of hub proteins*, 5th International Symposium on Health Informatics and Bioinformatics, IEEE. (2010) pp. 194–196.
4. J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins and T. S. Gardner, *Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles*, PLoS. Biol. **5** (1) (2007). DOI:10.1186/1752-0509-4-172
5. J. D. J. Han, N. Bertin, T. Hao, D. S. Goldberg, G. F. Berriz, L. V. Zhang, D. Dupuy, A. J. Walhout, M. E. Cusick, F. P. Roth and M. Vidal, *Evidence for dynamically organized modularity in the yeast protein–protein interaction network*, Nature **430** (2004) 88–93.
6. G. Hu, Z. Wu, V. N. Uversky and L. Kurgan, *Functional analysis of human hub proteins and their interactors involved in the intrinsic disorder-enriched interactions*, Int. J. Mol. Sci. **18** (12) (2017). DOI:10.1021/pr060393m
7. K. Luck, D. K. Kim, L. Lambourne, K. Spirohn, B. E. Begg, W. Bian, R. Brignall, T. Cafarelli, F. J. Campos-Laborie, B. Charloteaux and D. Choi, *A reference map of the human binary protein interactome*, Nature **580** (2020) 402–408.
8. F. Messina, E. Giombini, C. Agrati, F. Vairo, T. A. Bartoli, S. Al Moghazi, M. Piacentini, F. Locatelli, G. Kobinger, M. Maeurer, A. Zumla, M. R. Capobianchi, F. N.

Lauria and G. Ippolito, *COVID-19: Viral–host interactome analyzed by network based-approach model to study pathogenesis of SARS-CoV-2 infectio*, J. Transl. Med. **18** (1) (2020). DOI:10.1186/s12967-020-02405-w

9. M. Mirzarezaee, B. N. Araabi and M. Sadeghi, *Features analysis for identification of date and party hubs in protein interaction network of Saccharomyces Cerevisiae*, BMC. Syst. Biol. **4** (1) (2010) 1–11.
10. K. Prasad, F. Khatoon, R. Rashid, N. Ali, A. F. Al Asmari, M. Z. Ahmed, A. S. Alqahtani, M. S. Alqahtani and V. Kumar, *Targeting hub genes and pathways of innate immune response in COVID-19: A network biology perspective*, Int. J. Biol. Macromol. **163** (2020) 1–8.

E-mail: mhabibi@ipm.ir

Face Recognition Using Ordinary and Higher-Order Singular Value Decomposition Classifier: A Comparison Study

Hamid Salimi*
School of Mathematics, Statistics and Computer Science, College of Science, University of Tehran, Tehran, Iran
Morteza Amini
School of Mathematics, Statistics and Computer Science, College of Science, University of Tehran, Tehran, Iran
and Alireza Hosseini
School of Mathematics, Statistics and Computer Science, College of Science, University of Tehran, Tehran, Iran

Abstract. The tensor based classifiers are used for classification of any data with multiple dimensions, such as images, videos, fMRI images and so on. The Higher-Order Singular Value Decomposition (HOSVD) is an essential tool for such a classifier. Although the HOSVD considers the factors of each dimension of the data separately, it needs more memory and has a higher complexity compared to the ordinary Singular Value Decomposition (SVD). In this paper, we consider the problem of face recognition and compare the performance of SVD and HOSVD classifiers in this field. It is observed that HOSVD classifier can not dominate the ordinary SVD classifier for face recognition problem.
Keywords: Multidimensional data, Sub-space classification, Tensor decomposition.
AMS Mathematical Subject Classification [2010]: 15-XX, 15A69.

1. Introduction

Classification based on SVD [1, 3] is classification method used for sparse signals, which are well characterized and classified by a few of the first singular components of the images of the same class. The ordinary SVD [8] can only be used for a matrix of samples. To use SVD for multidimensional data sets such as image, video, fMRI etc. they should be vectorized into rows of the sample matrix.

It seems that vectorizing these tensors might cause information loss about the adjacent components of the tensors. The HOSVD [4] is a generalization of the SVD method for multidimensional tensors, without vectorizing it into rows of a matrix. Using HOSVD, the SVD classification techniques are extended to the tensor based classification methods [2]. This method is used for image processing [6], face recognition [5] and handwritten classification [7].

*Speaker

Although the HOSVD decompose the data in all dimensions and considers the information of the adjacent components of the tensors, it has a higher complexity and memory consumption compared with the ordinary SVD. In this paper, we compare the performance of the SVD and HOSVD classifiers in the problem of face recognition. We observe that the HOSVD classifier can not dominate the SVD classifier for the face recognition problem using two benchmark face recognition data sets.

The rest of the paper is organized as follows: in Section 2 classification based on SVD and HOSVD algorithms is explained. Section 3 presents the experimental results and concluding remarks.

2. Classification Based On SVD and HOSVD

2.1. SVD Classifier. The main idea behind the SVD classifier is the modeling of the variation within each class using orthogonal basis vectors obtained by SVD decomposition.

THEOREM 2.1. *Every matrix $B \in \mathbb{R}^{I\times J}$ can be decomposed as product of*

$$B = \Sigma \times U \times V^T,$$

where $U \in \mathbb{R}^{I\times I}$ and $V \in \mathbb{R}^{J\times J}$ are orthogonal matrices and Σ is an $(I \times J)$ diagonal matrix with non-negative entries, ordered in the following way: $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_J \geq 0$.

The SVD can also be written as an expansion of the matrix:

$$B = \sum_{i=1}^{J} \sigma_i \times u_i \times v_i^t = \sum_{i=1}^{J} \alpha_i \times u_i,$$

where $\alpha_i = \sigma_i \times v_i^t$, u_is and v_is are column vectors of U and V respectively. This is usually called the outer product form.

To classify image date using SVD, let's vectorize the image of each class with dimension $I \times J$. Stacking all the columns of each vectorized image above each other gives a matrix. Suppose $B_i \in \mathbb{R}^{M\times N_i}$ with $M = I \times J$ be the matrix consisting of all the training images of class i $(1, \ldots, K)$.

The idea of SVD classification is to model the variation within the set of training (and test) classes of one kind using an orthogonal basis of the subspace. Therefore, it should be computed how well an unknown sample can be represented in some different bases. This can be done by computing the residual vector in least squares problems of the type:

$$\min_{(\alpha_1^i,\ldots,\alpha_N^i)} \left\| z - \sum_{j=1}^{N} \alpha_j^i \times u_j^i \right\|, \quad i = 1, \ldots, K,$$

where z represents an unknown sample and u_j^i represents the singular images of class i $(i = 1, \ldots, K)$. This problem can be written in the form $\left\{\min_{\alpha^i} \left\| z - U_N^i \alpha^i \right\|\right\}$, where $U_N^i = (u_1^i, \ldots, u_N^i)$, $i = 1, \ldots, K$. Since the columns of U_N^i are orthogonal,

the solution to this problem is given by $\alpha_i = (U_N^i)^T z$, and the new sample is classified in the class with the minimum value of

$$R_i = \left\| (I_N - U_N^i U^{i^T}_N) z \right\|, \ i = 1, \ldots, K.$$

2.2. HOSVD Classifier. An imprecise definition of a Tensor is An object with **N** indices, where **N** is the order of the Tensor. Vectors and Matrices are tensors of order 1 and 2, respectively. For example, a tensor of order 3 is denoted by $\mathcal{A} \in \mathbb{R}^{I \times J \times K}$, where I, J and K are positive integers.

One of the important definitions widely used in the tensors, is the n-mode tensor-matrix multiplication, which is defined as follows.

DEFINITION 2.2. Let $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ and $\mathrm{B} \in \mathbb{R}^{J \times I_n}$. The n-mode tensor-matrix product of $\mathcal{A}$ and B is denoted $\mathcal{A} \times_n \mathrm{B}$, and is defined by

$$(\mathcal{A} \times_n \mathrm{B})(i_1, \ldots, i_{n-1}, j, i_{n+1}, \ldots, i_N) = \sum_{i_n=1}^{I_n} \mathcal{A}(i_1, \ldots, i_N) B(j, i_n).$$

The HOSVD is defined for example for a third order tensor as follows. The definition for other orders are similar.

THEOREM 2.3. *A third order tensor $\mathcal{A} \in \mathbb{R}^{I \times J \times K}$ could be expressed as product*

$$\mathcal{A} = \mathfrak{A} \times_1 U \times_2 V \times_3 W,$$

with following properties:

1) *$U \in \mathbb{R}^{I \times I}$,$V \in \mathbb{R}^{J \times J}$ and $W \in \mathbb{R}^{K \times K}$ are orthogonal matrices.*
2) *$\mathfrak{A}$ is a real tensor of the same dimensions as $\mathcal{A}$ which satisfies:*
   (i) *The all-orthogonality property: any two different slices fixed in the same mode are orthogonal.*
   (ii) *The ordering property: norms of slices along every mode are ordered, for instance, for the first mode we have*

$$\|\mathfrak{A}(1, :, :)\| \geq \|\mathfrak{A}(2, :, :)\| \geq \cdots \geq 0.$$

Another way to reconstruct a third order tensor can be achieved:

$$\mathcal{A} = \sum_{\nu=1}^{K} \mathbf{A}_\nu \times_3 \omega_\nu, \tag{1}$$

where $\mathbf{A}_\nu = \mathfrak{A}(:, :, \nu) \times_1 U \times_2 V$, and w_ν is the νth column of W. In this form, every matrix in $\mathfrak{A}$ is a unique linear combination of the same basis matrices $\mathbf{A}_\nu$.

Below, we will deal with image classification based on HOSVD. For the image data, the train sample of size N_i the $I \times J$ pixel images in class i $(i = 1, \ldots, K)$ can be considered as a third order tensor $\mathcal{A}_i \in R^{I \times J \times N_i}$. From (1) the orthogonal basis matrices for class i $(i = 1, \ldots, K)$ are constructed. Suppose we want to classify an unknown image D with $\|D\| = 1$. For this purpose, the following minimization problem should be solved.

$$\min_{(\alpha_1^i, \ldots, \alpha_N^i)} \left\| D - \sum_{\nu=1}^{N} \alpha_\nu^i \times A_\nu^i \right\|, \qquad (i = 1, \ldots, K),$$

where α_ν^is are the unknown coefficients and the bases $A_v^i (i = 1, \ldots, K)$ are derived from HOSVD decomposition of $\mathcal{A}_i$. Because of Orthonormality of A_ν^i, the solution is given by $\hat{\alpha}_\nu^i = \langle D, A_\nu^i \rangle$. Therefore, an unknown image D belongs to class i, where has the smallest $R_i, i = 1, \ldots, K$, obtained by

$$R_i = \left\| D - \sum_{\nu=1}^{N} \hat{\alpha}_\nu^i \times A_\nu^i \right\| = \left\langle D - \sum_{\nu=1}^{N} \hat{\alpha}_\nu^i \times A_\nu^i,\ D - \sum_{\nu=1}^{N} \hat{\alpha}_\nu^i \times A_\nu^i \right\rangle$$
$$= \langle D, D \rangle - \sum_{\nu=1}^{N} \left\langle D, A_\nu^i \right\rangle^2 = 1 - \sum_{\nu=1}^{N} \left\langle D, A_\nu^i \right\rangle^2 .$$

3. Experimental Results

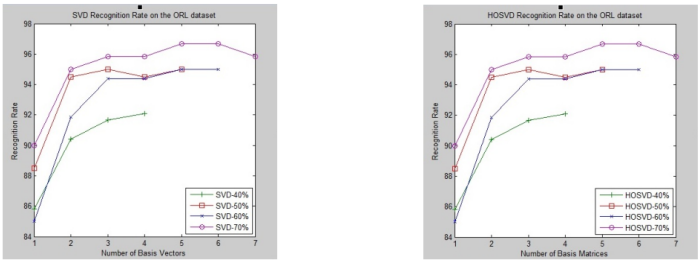
In order to compare the performance of SVD-classifier with HOSVD-classifier, 2 benchmark face recognition datasets (ORL and YALE) were used. The ORL face database contains 400 face images of 40 people (10 samples of each person). These images are different in gesture and posture such as: smiling or non-smiling, open or closed eyes and also facial details like: with and without glasses were taken with tolerance for some side movement and rotation of the face up to 20 degrees.

The YALE face dataset contains 165 grayscale images of 15 individuals. There are 11 images per subject, one per different facial expression or illumination: center-light, with glasses, happy, leftlight, without glasses, normal, right- light, sad, sleepy, surprised, and winking.

(A) ORL Dataset. (B) YALE Dataset.

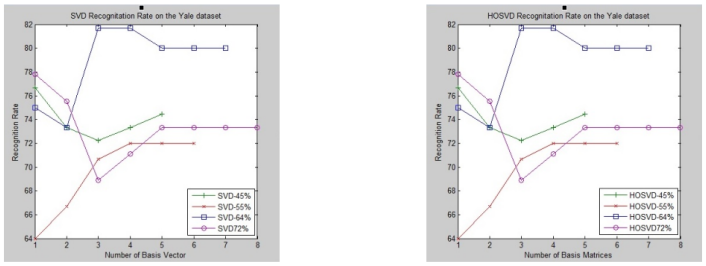
FIGURE 1. Some example images from ORL and YALE dataset.

The two algorithms were compared with 4 different schemes of train-test splinting (40%, 50%, 60% and 70% as the train set). Figures 2 and 3 show experimental results on the ORL and Yale, respectively. As it can be seen from the figures, the precision of two methods are equal for all four different schemes. We also noticed that the misclassified samples for SVD and HOSVD classifiers are the same. While the HOSVD has a higher complexity and execution time compared to the ordinary SVD, the HOSVD is not preferred for the classification of less complex data sets. Also, it is observed in the experiments that the increment of basis vectors has no effect on the precision of the methods.



(A) SVD Recognition. (B) HOSVD Recognition

FIGURE 2. Comparing SVD with HOSVD on the ORL dataset.



(A) SVD Recognition. (B) HOSVD Recognition.

FIGURE 3. Comparing SVD with HOSVD on the YALE dataset.

References

1. C. Clark and A. F. Clark, *Spectral identification by singular value decomposition*, Int. J. Remote Sensing **19** (12) (1998) 2317–2329.
2. B. Cyganek, *Embedding of the extended euclidean distance into pattern recognition with higher-order singular value decomposition of prototype tensors*, IFIP International Conference on Computer Information Systems and Industrial Management (2012) pp. 180–190.
3. S. Danaher and E. Ómongain. *Singular value decomposition in multispectral radiometry*, Int. J. Remote Sensing **13** (9) (1992) 1771–1777.
4. L. De Lathauwer, B. De Moor and J. Vandewalle, *A multilinear singular value decomposition*, SIAM J. Matrix Anal. Appl. **21** (4) (2000) 1253–1278.
5. L. Liţă, and E. Pelican, *A low-rank tensor-based algorithm for face recognition*, Appl. Math. Model. **39** (3-4) (2015) 1266–1274.
6. X. Luo, Z. Zhang, C. Zhang and X. Wu, *Multi-focus image fusion using HOSVD and edge intensity*, J. Vis. Commun. Image Represent **45** (2017) 46–61.
7. B. Savas and L. Eldén, *Handwritten digit classification using higher order singular value decomposition*, Pattern Recognition **40** (3) (2007) 993–1003.
8. G. W. Stewart, *On the early history of the singular value decomposition*, SIAM Rev. **35** (4) (1993) 551–566.

E-mail: salimi.hamid86@gmail.com
E-mail: morteza.amini@ut.ac.ir
E-mail: hosseini.alireza@ut.ac.ir

New Heuristics for Burning Graphs

Maryam Tahmasbi*
Computer Science Department, Shahid Beheshti University, Tehran, Iran
Zahra Rezai Farokh
Computer Science Department, Shahid Beheshti University, Tehran, Iran
Zahra Haj Rajab Ali Tehrani
Computer Science Department, Shahid Beheshti University, Tehran, Iran
and Yousof Buali
Department of Computer and data Sciences, Shahid Beheshti University, Tehran, Iran

Abstract. Graph burning models the spread of contagion(fire) in a graph in discrete time steps. The burning number of a graph G, $bn(G)$ is the minimum time needed to burn a graph G. Determining the burning number of a graph is NP-complete. In this paper, we develop first heuristics to solve the problem in general (connected) graphs. In order to test the performance of our algorithms, we applied them on some graph classes with known burning number and known benchmarks for NP-hard problems in graph theory. We also improved the upper bound for burning number on general graphs in terms of their distance to cluster. Then we generated a data set of 2000 random graphs with known distance to cluster and tested our heuristics on them.

Keywords: Burning number, Heuristic, Distance to cluster, Theta graphs.
AMS Mathematical Subject Classification [2010]: 05C85, 05C85, 90C06.

1. Introduction

Burning number of a graph is a new concept that measures the speed of spreading a contagion (fire) in a graph [2]. Given an undirected unweighted graph, the fire spreads in the graph in synchronous rounds as follows: in round one, a fire starts at a vertex called an activator. In each following round two events happen:

(1) The fire spreads to all neighbors of vertices that are on fire.
(2) Fire starts at a new activator that is an unburned vertex.

The process continues until all vertices of the graph are on fire. At this time we say that the burning process is complete [5]. A burning schedule specifies a burning sequence of vertices, where the ith vertex in the sequence is the activator in round i. The burning number $bn(G)$ is the minimum length of a burning sequence.

In burning number Problem the input is a simple graph G of order n and an integer $k \geq 2$. The question is whether $bn(G) \leq k$? In other words, does G contain a burning sequence $(x_1, x_2, \ldots, x_k)$?

As the first result, some of the properties of this problem including characterizations and bounds was presented in [2, 9]. Bonato et al. [2] proved that

*Speaker

the burning number of any connected graph with n vertices is at most $2\sqrt{n}-1$ and conjectured that it is always at most $\lceil\sqrt{n}\rceil$. It is proved that this problem is NP-complete even when restricted to simple graphs [1]. They developed a polynomial-time approximation algorithm with approximation factor 3 for general graphs [1]. Bonato and Lidbetter [4] developed a 3/2-approximation algorithm for path forests (disjoint union of paths). There is another approximation algorithm with an approximation ratio of 2 for trees.

In a recent study Kamali et al. [5] considered connected n-vertex graphs with minimum degree δ. They developed an algorithm that burns any such graph in at most $\sqrt{\frac{24n}{\delta+1}}$ rounds. In particular, for graphs with $\delta \in \theta(n)$, they proved that all vertices are burned in a constant number of rounds. More interestingly, even when δ is a constant that is independent of n, their algorithm answers the graph-burning conjecture in the affirmative by burning the graph in at most $\sqrt{n}$ rounds.

Šimon et al. [10] developed some heuristics for graph burning based on some centrality measures. They tested their heuristics on limited number of networks.

In this paper, we develop new heuristic algorithms for solving graph burning problem. As mentioned before, most of the studies on this problem concern limited classes of graphs. Since the problem is modeling the spread of contagion in a network, it is essential to develop algorithms for solving the problem in general graphs. We developed 6 heuristics for burning a graph. These heuristics differ in selecting the first activator and also the order of selecting the following activators.

Except for approximation algorithms [4, 6, 10], and algorithm 1 in [3] there are no official algorithms for this problem, so to test the performance of our algorithm, we used some theoretical results: we generated a random class of theta graphs and a random class of graphs with known distance to cluster and report the result of applying our algorithms on these classes. We compared our results with exact values and bounds reported in former studies. We also applied our algorithms on various graphs in known data sets: DIMACS and BHOSLIB. These data sets contain graphs with various sizes and structures.

This paper is organized as follows: in Section 2 we present some basic definitions and our heuristics. In Section 3 we present the result of our experimental study on different data sets, and in Section 4 we state conclusions and future works.

2. Algorithms

In this section, we present 6 heuristics for solving graph burning problem. The output in each algorithm is a burning sequence for the input graph G. First, we review some basic definitions and then we present our heuristics.

2.1. Basic Definitions. For a graph $G = (V, E)$, let n and m to denote the number of vertices and edges, respectively. For a vertex $v \in V$, $N(v)$ denotes the set of vertices adjacent to v and $N[v] = N(v) \cup \{v\}$ is the closed neighborhood of v. Given an integer k, $N_k[v]$ is the number of vertices with distance at most k of a vertex v. For a vertex v in a graph G, the eccentricity of v is defined as $\max\{d(u, v)|u \in V(G)\}$. The radius of G is the minimum eccentricity over the set of all vertices in G. The diameter of G is the maximum eccentricity over the set of

all vertices in G. For a subset $X \subset V(G)$, the graph $G[X]$ denotes the subgraph of G induced by X.

2.2. Heuristics. In the proposed algorithms we have two steps: the first step is to select the first candidate for burning. It seems essential since this vertex will burn vertices in distance $bn(G)$ of the graph. So, we need to select a vertex with a large set of vertices in $N_{bn(G)}[v]$. In the second step, we select the rest of the activators one by one.

Given a burning sequence $S = (x_1, x_2, \ldots, x_{bn(G)})$ of a graph G, for each vertex v, there is a vertex x_i in S such that v is burned by a fire that is started in x_i, i.e. $d(v, x_i) < d(v, x_j)$ for all $j \neq i$. we call x_i the activator of v.

We develop different heuristics based on different strategies for the first and second steps.

(1) We choose the first activator from the center of the graph. The farthest vertex to this vertex is in distance $rad(G)$ of it. So, it seems that this vertex has a big kth neighborhood. We used this in step one of heuristics *Ctr-Half dist.* and *Ctr-Far dist.*.

(2) In each time step k, for each unburned vertex v, we can calculate that in how many time steps this vertex will burn if we do not add any other activator. We call this time-to-burn of v and denote it by $t^k(v)$. Let $t^k = \max\{t^k(v) : v \in V\}$. Hence, $t^k + k$ is an upper bound for the burning number. We can select the next activator in two ways:
   (a) The next activator is a vertex v with $t^k(v) = t^k/2$. In this way the vertices with greater time-to-burn will burn in shorter time, using this new activator. (Heuristics Ctr-Half dist. and Rnd-Half dist.)
   (b) The next activator is a vertex with max -1 time-to-burn. (Heuristics Ctr-Far duist and Rnd-Far dist.)

(3) In heuristics *Rnd-Half dist.* and *Rnd-Far dist.* we select the first activator in random to see the effect of selecting the first activator in our heuristics.

Table 1 summarizes the strategies in four heuristics.

TABLE 1. Summary of first four heuristics.

Heuristics	First Activator	Next Activator
Ctr-Half dist.	Center	1/2 time-to-burn
Ctr-Far dist.	Center	max time-to-burn
Rnd-Half dist.	Random	1/2 time-to-burn
Rnd-Far dist.	Random	max time-to-burn

We developed two other heuristics with a different idea: burning a path. The main idea is finding the diameter of the graph and the path with length $diam(G)$ and then burning the vertices of this path with the same order as burning a path in $\sqrt{diam(G)}$ steps. Since computing the diameter of a graph is of large complexity, we use two approaches to find a good approximation of that. In heuristic *DFS-path* we select a random vertex and perform a DFS algorithm to find a path. In heuristic *D-BFS-path* we use the algorithm by Birmele et al. to approximate the diameter. This algorithm uses BFS twice, the first BFS starts from a random

vertex and the second one starts from one of the leaves of previous BFS. This gives a 2-approximation of the diameter of the graph. There is no guarantee that all vertices of the graph burn using only vertices of these paths. So, after burning the vertices of the path, if there is still an unburned vertex, we select them randomly as activators.

3. Experimental Study

We implemented algorithms that were introduced and explained earlier in Section 2 using Python 3. To model our graphs in a proper data structure and apply fundamental graph algorithms and measures, we used the well-known NetworkX package introduced by Hagberg et al.

3.1. Datasets. As mentioned before, there is no algorithm for burning general graphs. So, in order to evaluate our heuristics we use two types of datasets:

(1) Classic datasets that are commonly used in some NP-hard problems in graph theory such as clique number, independence set, dominating number, etc.
(2) Random graphs in some classes with the exact value or a good bound on their burning number computed.

3.2. DIMACS, BHOSLIB. We applied our 6 heuristics on all graphs in DIMACS and BHOSLIB. From 78 graphs in DIMACS, all heuristics computed a burning sequence of length 3 for 71 graphs. Roshanbin [9] proved that for a G be a graph with n vertices $bn(G) = 2$ if and only if $n \geq 2$ and G has maximum degree $n-1$ or $n-2$. The results on BHOSLIB graphs are more interesting. All heuristics compute 3 for all graphs and this is optimal.

3.3. $\theta-$Graphs. There are tight bounds on the burning number of $\theta-$graphs that are proved by Liu and et al. [8]. They showed that the burning number of order $n = q^2 + r$ with $1 \leq r \leq 2q + r$ is either q or $q + 1$. We compared our results with these bounds. In %81.7 of graphs, we computed the same value as the proved bounds. The average difference between our best results and upper bounds is 0.6 and standard deviation 1.2309. Table 2 shows the comparison of different heuristics.

Table 2. Comparison of different heuristics on theta graphs.

Heuristic	Success Rate
Ctr-Far	16.9%
Rnd-Far	7.7%
DFS-path	73.8%
D-BFS-path	1.6%

3.4. Graphs with Fixed Distance to Cluster. Kare et al. [7] computed an upper bound for graphs in terms of their distance to cluster, which is $3d + 3$. We improve this bound in the following theorem.

THEOREM 3.1. *Let G be a graph and A be a set of vertices such that $G[V(G)\backslash A]$ is a cluster graph. Then $bn(G) \leq bn(G[A]) + 2$.*

PROOF. A burning sequence of A burns all vertices except possibly vertices of complete graphs that are adjacent to ihe last vertex of the burning sequence. These complete graphs burn in at most 2 rounds. So $bn(G) \leq bn(G[A]) + 2$. □

An immediate conclusion from Theorem 3.1 is that $bn(G) \leq d + 2$ for each graph with distance to cluster d. We generated 2000 random graphs as described is Subsection 3.1 and applied our heuristics on these graphs and compared the result with $\lceil\sqrt{d}\rceil + 2$. The results show that in %98 of graphs the results meet bounds. We also compared all heuristics. Heuristics Ctr-Half dist, Rnd-Half dist and Rnd-Far dist find better solutions among our 6 heuristics. The winning heuristics are the ones that select the first activator in different ways and the following ones according to far dist. strategy.

4. Conclusion and Future Work

In this paper, we developed the first heuristics for graph burning problem. To study the performance of our heuristics, we applied them on two types of datasets: (1) Known benchmarks for NP-hard problems in graph theory. We selected DIMACS and BHOSLIB. Our heuristics computed the optimal solution in 71 graphs out of 78 graphs in DIMACS, and all the 36 graphs in BHOSLIB. (2) Randomly generated graphs in classes with a known burning number, such as 2000 $\theta-$graphs and 2000 random graphs with known distance to cluster. We computed the correct burning number in %81 of theta graphs and %98 of graphs with given distance to cluster.

Since there are very few studies (just one paper) on algorithmic approaches to solve the burning number, the results here can be used as bench marks. Finally, there is a huge body of research on the spread of influence in social networks. There are measures called centrality measures to select seeds (activators) in a social network. It is interesting to develop algorithms for burning graph using these measures.

References

1. S. Bessy, A. Bonato, J. Janssen, D. Rautenbach and E. Roshanbin, *Burning a graph is hard*, Discrete Appl. Math. **232** (2017) 73–87.
2. A. Bonato, J. Janssen and E. Roshanbin, *Burning a Graph as a Model of Social Contagion*, In: A. Bonato, F. Graham, P. Praat (Eds) Algorithms and Models for the Web Graph, WAW 2014, Lecture Notes in Computer Science, Vol. 8882, Springer, Cham. 2014.
3. A. Bonato, J. Janssen and E. Roshanbin, *How to burn a graph*, Internet Math. **12** (1-2) (2016) 85–100.
4. A. Bonato and T. Lidbetter, *Bounds on the burning numbers of spiders and path-forests*, Theoret. Comput. Sci. **794** (2019) 12–19.
5. S. Kamali, M. Avery and Zh. Kenny, *Burning Two Worlds*, Int. Conf. Current Trends in Theory & Practice Inform., Springer, Cham, (2020) 113–124.
6. S. Kamali, M. Avery and Zh. Kenny, *Burning two worlds: algorithms for burning dense and tree-like graphs*, (2016). `arXiv:1909.00530`

7. A. S. Kare and I. V. Reddy, *Parameterized algorithms for graph burning problem*, International Workshop on Combinatorial Algorithms, Springer, Cham, (2019) 304–314.
8. H. Liu, R. Zhang and X. Hu, *Burning number of theta graphs*, Appl. Math. Comput. **361** (2019) 246–257.
9. E. Roshanbin, *Burning a Graph as a Model of Social Contagion*, Ph.D. Thesis, Dalhousie University, 2016.
10. M. Šimon, L. Huraj, I. Dirgová Luptáková and J. Pospíchal, *Heuristics for spreading alarm throughout a network*, Appl. Sci. **9** (16) (2019) 3269.

E-mail: m_tahmasbi@sbu.ac.ir
E-mail: rezaifarokhz@gmail.com
E-mail: doorsatehrani@yahoo.com
E-mail: yousof.buali@gmail.com

Contributed Talks

Numerical Analysis

Simultaneous Hard Thresholding Algorithms for Multiple Measurement Vectors

Farshid Abdollahi*
Department of Mathematics, College of Sciences, Shiraz University, Shiraz, Iran

Abstract. Given $Y \in R^{m\times k}$ and a sensing matrix $A \in R^{m\times N}$ with $m \ll N$, the multiple measurement vectors (MMV) problem aims to recover row-sparse matrices $X \in R^{N\times k}$ of an underdetermined linear system $AX = Y$. In this work, we introduce two iterative algorithms, Simultaneous Null Space Tuning with Hard Thresholding with FeedBack (SNST+HT+FB) and SNST+HT with stretching for jointly sparse vectors recovery in MMV model. These algorithms are based on the null space tuning with hard thresholding techniques in single measurement vector (SMV) model of compressive sensing. Finally, some numerical results are presented to demonstrate the advantages of the algorithms.
Keywords: Compressive Sensing, Sparse recovery, Null Space Tuning, Hard Thresholding Algorithm, Multiple Measurment Vectors.
AMS Mathematical Subject Classification [2010]: 65F50, 65F10, 15A29.

1. Introduction

Compressive sensing (CS), also known as compressive sampling, has received considerable research of interest in various applications due to its superior capability to recovery a sparse signal from a much smaller number of measurements than its original dimension. In the popular, CS is referred to as a single measurement vector (SMV) model. Mathematically speaking, given a measurement matrix $A \in R^{m\times N}$ with $m \ll N$, and given a measurement vector $\mathbf{y} = A\mathbf{x} \in R^m$ associated with an s-sparse vector $\mathbf{x} \in R^N$ (a vector that has at most s nonzero entries), we want to access this vector in a numerically tractable way. A natural extension of this problem is the multiple measurement vectors (MMV) model. In the MMV model, signals are represented as matrices and are assumed to have the same sparsity structure. The main aim of MMV is to find row-sparse matrices $X \in R^{N\times k}$ of an underdetermined linear system $AX = Y$, where $A \in R^{m\times N}$ and $Y \in R^{m\times k}$ are given. The MMV model was initially motivated by a neuromagnetic inverse problem that arises in Magnetoen-cephalography (MEG), a brain imaging modality [2]. It is assumed that MEG signal is a mixture of activities at a small number of possible activation regions in the brain. MMV model has also been found in array processing, nonparametric spectrum analysis of time series, equalization of sparse communication channel, linear inverse problem, DNA microarrays and source location in sensor networks and etc. for more details see [1, 5, 7, 9, 10] and references therein.

*Speaker

For solving CS problems, there are several classes of algorithms that have been used in applications, such as l_1 minimization algorithms, greedy algorithms (for example matching pursuit (MP) and orthogonal matching pursuit (OMP)), Iterative thresholding/shrinkage algorithms and combinatorial algorithms (for more details see [3]). Greedy pursuit algorithms can be much faster than ℓ_1-based methods and are often applicable to very large sparse recovery problems. Iterative thresholding approaches offer another fast alternative, which in addition, share the near optimal recovery guarantees offered by ℓ_1-based approaches.

Li et al. in [8], proposed iterative null space tuning algorithms with hard thresholding (for example NST+HT+FB and NST+stretchedHT) to find sparse solutions, aiming at faster convergence rate and greater recovery capacity for SMV model. In this paper, we propose natural extensions of these algorithms, which we call Simultaneous null space tuning with hard thresholding with feedback (SNST+HT+FB), SNST+HT with stretching (SNST+stretchedHT) Algorithms and present the theoretical convergence studies of these algorithms. Finally, some numerical results are presented to demonstrate the advantages of the algorithms.

2. Main Results

2.1. Row-Sparse Recovery. Let us suppose that several sparse vectors $\mathbf{X}_1, \ldots, \mathbf{X}_k \in R^N$ are to be recovered from $\mathbf{Y}_1 = A\mathbf{X}_1, \ldots, \mathbf{Y}_k = A\mathbf{X}_k \in R^m$, with the additional assumption that $\mathbf{X}_1, \ldots, \mathbf{X}_k$ are jointly sparse, in other words they are all supported on a set of small cardinality. Intuitively, we foresee an achieve in computational complexity by recovering $\mathbf{X}_1, \ldots, \mathbf{X}_k$ all at the same time rather than one by one. For now, we reformulate the problem by defining the $N \times k$ matrices

$$X = \left(X_1 | \cdots | X_k\right) = \begin{pmatrix} X^1 \\ X^2 \\ \vdots \\ X^N \end{pmatrix}, Y = \left(Y_1 | \cdots | Y_k\right),$$

where X^i stands for the ith row of X and X_j stands for the jth column of $X \in R^{N \times k}$. The joint sparsity assumption just says that X is s-row-sparse, i.e., its row-support

$$supp(X) := \{i : X^i \neq 0\},$$

has cardinality at most s.

2.2. SNST+HT+FB **and** NST+stretchedHT**.** In this section we propose new algorithms that are designed to recovery of s-row-sparse matrices X from $Y = AX$. These algorithms are natural extension of the NST+HT+FB and NST+stretchedHT Algorithms which were introduced in [8]. We named these algorithms SNST+HT+FB and SNST+stretchedHT and are implemented as follows:

Algorithm 1. SNST+HT+FB Algorithm

Input: $A, Y, s, \epsilon_1, \epsilon_2$;
Output: X;

Initial: $P = I - A^*(AA^*)^{-1}A$, $\mathbf{X}^{[0]} = A^*(AA^*)^{-1}\mathbf{Y}$, $k = 0$, $U^{[-1]} = 0$, $U^{[0]} = \mathbb{T}_s(X^{[0]})$,
While $(\frac{\|AU^{[k]}-Y\|_F}{\|Y\|_F} > \epsilon_1$ **and** $\frac{\|U_S^{[k]}-U_S^{[k-1]}\|_F}{\|U_S^{[k]}\|_F} > \epsilon_2)$ $k = k + 1$,
$U^{[k]} = \mathbb{T}_s(X^{[k]}) + (A^*_{S_{[k]}}A_{S_{[k]}})^{-1}A^*_{S_{[k]}}A_{S^c_{[k]}}X^{[k]}_{S^c_{[k]}}$ **(set** $S_{[k]}$ **= indices of** s **largest** $\|(X^i)^{[k]}\|_2$**),**
$X^{[k+1]} = X^{[k]} + \boldsymbol{P}(U^{[k]} - X^{[k]})$.
End While

Algorithm 2. SNST+stretchedHT Algorithm

Input: $A, Y, s, \epsilon_1, \epsilon_2$;
Output: X;
Initial: $P = I - A^*(AA^*)^{-1}A$, $\mathbf{X}^{[0]} = A^*(AA^*)^{-1}\mathbf{Y}$, $k = 0$, $U^{[-1]} = 0$, $U^{[0]} = \mathbb{T}_s(X^{[0]})$
While $(\frac{\|AU^{[k]}-Y\|_F}{\|Y\|_F} > \epsilon_1$ **and** $\frac{\|U_S^{[k]}-U_S^{[k-1]}\|_F}{\|U_S^{[k]}\|_F} > \epsilon_2)$**,**
$k = k + 1$,
$\Theta_k = \|Y\|_F / \|A_{S_{[k]}}X^{[k]}_{S_{[k]}}\|_F$ **(set** $S_{[k]}$ **= indices of** s **largest** $\|(X^i)^{[k]}\|_2$**),**
$U^{[k]} = \Theta_k\mathbb{T}_s(X^{[k]})$,
$X^{[k+1]} = X^{[k]} + \boldsymbol{P}(U^{[k]} - X^{[k]})$.
End while

Here $P = I - A^*(AA^*)^{-1}A$ is the orthogonal projection onto $kerA$ and $\mathbf{X}^{[0]}$ is always set as the least squares solution, i.e., $\mathbf{X}^{[0]} = A^*(AA^*)^{-1}\mathbf{B}$ and $\|\cdot\|_F$ stands for Frobenius norm.

2.3. Convergence Analysis. To state the convergence results, we first recall the definition of restricted isometry property (RIP) and preconditioned restricted isometry property (P-RIP), see [6].

Definition 2.1. For an $m \times N$ measurement matrix A, the s-restricted isometry constant δ_s of A is the smallest quantity such that

$$(1 - \delta_s)\|\mathbf{x}\|_2^2 \leq \|A\mathbf{x}\|_2^2 \leq (1 + \delta_s)\|\mathbf{x}\|_2^2,$$

holds for all s-sparse signals x. Equivalently, it is given by

$$\delta_s = \max_{card(S)\leq s} \|A_S^*A_S - I\|_2.$$

If δ_s is small for reasonably large s, then matrix A is said to satisfy the s-restricted isometry property with the s-restricted isometry constant δ_s.

Definition 2.2. For an $m \times N$ measurement matrix A, the preconditioned s-restricted isometry constant γ_s of A is the smallest quantity such that

$$(1 - \gamma_s)\|\mathbf{x}\|_2^2 \leq \|(AA^*)^{-\frac{1}{2}}A\mathbf{x}\|_2^2 \leq (1 + \gamma_s)\|\mathbf{x}\|_2^2,$$

holds for all s-sparse signals x. Equivalently, it is given by

$$\gamma_s = \max_{card(S)=s} \|A_S^*(AA^*)^{-1}A_S - I\|_2.$$

By using the above definitions we can prove the following Theorem, which guarantees the convergence of the SNST-HT-FB Algorithm.

THEOREM 2.3. *Suppose that $X^{[*]}$ is a real s-row-sparse solution of $AX = Y$. If the P-RIP and RIP constants of A satisfy $\delta_{2s} + \sqrt{2}\gamma_{3s} < 1$, then $U^{[k]}$ in SNST-HT-FB satisfies*

$$\|U^{[k]} - X^{[*]}\|_F \leq \rho^k \|U^{[0]} - X^{[*]}\|_F,$$

where $\rho = \frac{\sqrt{2}\gamma_{3s}}{1-\delta_{2s}}$.

Same as the SMV case, with requirements over the restricted isometry constants of A, both procedures (SNST+HT+FB and NST+stretchedHT) reduce the error in each iteration and are guaranteed to converge to limits with error bounds depending on the tail of the real solution. Similar to the single case, although SNST+HT+FB has a pursuit spirit seen in various algorithms such as HTP [4], the feedback mechanism plays a significant role particularly for large scale problems and has led to the superiority of this method compared with others.

3. Experimental Results

In this section, we present numerical experiments of the recovering matrices at once via SNST+HT+FB and SNST+stretchedHT versus recovering their columns one by one via NST+HT+FB and NST+stretchedHT Algorithms. In Figure 1, $X \in R^{N\times k}$ is generated by randomly selecting s rows with zero mean Gaussian random entries of unit variance and letting the remaining rows to be zeros and the measurement matrix $A \in R^{m\times N}$ is a Gaussian matrix with Gaussian independent and identically distribution (i.i.d.) entries of zero mean and variance $\frac{1}{N}$. The measurement signal is given by $Y = AX$. Reconstruction performance is quantified by the relative error, which is defined by

$$\text{relative error } = \frac{\|X_{rec} - X\|_F}{\|X\|_F},$$

where X_{rec} is the reconstructed signal matrix and X is the original one. In part (a) of Figure 1 we compare NST+HT+FB and SNST+HT+FB Algorithms and in the part (b), we compare NST+stretchedHT and SNST+stretchedHT Algorithms. The experiments illustrate how the relative error of each algorithm changes along the nonzero rows s. We set $N = 500, m = 200$ and $k = 5$. Let s (sparsity level) changes from 20 to 60 and for execution-time comparison, we tested k runs of NST+HT+FB and NST+stretchedHT and 1 run of SNST+HT+FB and SNST+stretchedHT methods for some problems. For each sparsity value s, the methods are tested for 100 trials. The execution time is recorded for every trial and the averages time are then calculated. Therefore, simple numerical experiment of Figure 1, confirms that one run of SNST+HT methods is faster than k runs of NST+HT ones, moreover accuracy of SNST+HT methods is better than another ones.

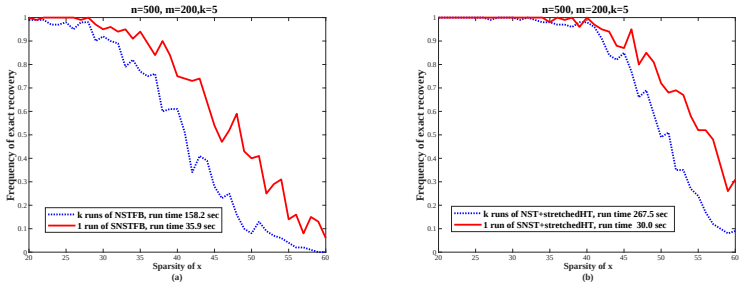


FIGURE 1. (a) Plots of $\|x-x_0\|/\|x_0\|$ as a function for NST+HT+FB and SNST+HT+FB. SNST+HT+FB does have the advantage at recovering Gaussian sparse vectors over that of NST+HT+FB. (b) Plots of $\|X - X_0\|/\|X_0\|$ as a function for NST+stretchedHT and SNST+stretchedHT. SNST+stretchedHT does have the advantage at recovering Gaussian sparse matrix with Gaussian matrix over that of NST-HT.

References

1. J. D. Blanchard, M. Cermak, D. Hanle and Y. Jing, *Greedy algorithms for joint sparse recovery*, IEEE Trans. Signal Proc. **62** (7) (2014) 1694–1704.
2. S. F. Cotter, B. D. Rao, K. Engan and K. Kreutz-Delgado, *Sparse solutions of linear inverse problems with multiple measurement vectors,* IEEE Trans. Signal Proc. **53** (7) (2005) 2477–2488.
3. Y. C. Eldar and G. Kutyniok, *Compressed Sensing: Theory and Applications*, Cambridge Univ. Press, Cambridge, UK, 2012.
4. S. Foucart, *Hard thresholding pursuit: an algorithm for compressive sensing,* SIAM J. Numer. Anal. **49** (6) (2011) 2543–2563.
5. S. Foucart, *Recovering jointly sparse vectors via hard thresholding pursuit*, In: Proc. Sampling Theory Appl. (SampTA) (2011) pp. 2–6.
6. S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing,* Birkhäuser/Springer, New York, NY, 2013.
7. J. Kim, J. Wang, L. T. Nguyen and B. Shim, *Joint sparse recovery using signal space matching pursuit,* IEEE Trans. Inf. Theory **66** (8) (2020) 5072–5096.
8. S. Li, Y. Liu and T. Mi, *Fast thresholding algorithms with feedbacks for sparse signal recovery*, Appl. Comput. Harmon. Anal. **37** (1) (2014) 69–88.
9. L. Sun, J. Liu, J. Chen and J. Ye, *Efficient recovery of jointly sparse vectors*, Adv. Neural Info. Proc. Sys. **22** (2009) 1812–1820.
10. Y. Zhu, Q. Chen, Y. Li, Y. Zhang and Y. Zhu, *Frequency-domain entropy-based blind support recovery from multiple measurement vectors*, IEEE Signal Processing Lett. **27** (2020) 980–984.

E-mail: abdollahi@shirazu.ac.ir

Analysis of the Stability of a High Order Numerical Method for Solving Unsteady Nonlinear Parabolic Differential Equations

Sadegh Amiri*
Department of Basic Sciences, Shahid Sattari Aeronautical University of Science and Technology, P. O. Box 13846-63113, Tehran, Iran

Abstract. In this study, after introducing a fourth order spacial numerical method, we demonstrate that this scheme guaranteed unconditional stability (under L_2 norm). Also, the presented method is second order in time and fourth order in space. Comparative results show that this method is accurate than the other existing methods in the literature.
Keywords: Fourth order spacial numerical method, Unconditional stability.
AMS Mathematical Subject Classification [2010]: 65Nxx, 65N06.

1. Introduction

High order numerical methods for unsteady nonlinear parabolic partial differential equations can be divided into two classes, say, wide methods and compact methods [3]. Wide fourth order numerical methods are obtained by discretizing the governing equations with fourth order central differences. In steady-state problems and the Navier-Stokes equations, many authors introduced various such schemes. For example, a two-level linearized compact ADI scheme was proposed for solving two-dimensional nonlinear reaction-diffusion equations by Wu, et al. [5]. The computational cost of their method is reduced by the use of the Newton linearized method and the ADI method. Karaa and Zhang discussed a high order ADI method for solving linear unsteady convection diffusion problems [4]. The order of their scheme is four in space and two in time.

In our analysis, we have applied both discrete perturbation stability analysis and discrete Fourier stability analysis using the discrete L_2 norm. The decay or growth of the amplification factor indicates whether or not the numerical method is stable. This paper is organized as follows. In Section 2, the formulation of the new method for initial boundary value linear parabolic differential equations is presented. In Section 3, its stability and error estimates with discrete L_2 norm is also investigated. Also, in this Section, for unsteady nonlinear parabolic PDEs, as the classical ADI method cannot be applied directly, we first suggest some suitable modifications of that method and then generalize the method of Section 2 for the nonlinear case to maintain the second order of accuracy in time and fourth in the space. The stability analysis of the method for the nonlinear case is also addressed

*Speaker

in that section. In Section 4, some numerical examples are given to illustrate the performance of the presented method.

2. Strategy of the Numerical Method

In this study, we consider the following unsteady nonlinear parabolic partial differential equations

$$U_t = \mu\ (U_{xx} + U_{yy}) + \varphi,\ 0 < x, y < 1,\ 0 < t \leq T, \tag{1}$$

in which

$$U(x, y, 0) = U_0(x, y),\ 0 < x, y < 1, \tag{2}$$

is its initial condition and

$$U(x, y, t) = g(x, y, t),\ 0 < x, y < 1,\ t \in \partial\Delta, \tag{3}$$

is its Dirichlet boundary condition with boundary $\partial\Delta$ of Δ and it is assumed that the functions U_0, g and the source term $\varphi(x, y, t, U)$ are sufficiently smooth. Also, we suppose that the diffusion parameter μ is a positive constant. The general form of the source term φ can be nonlinear in terms of the U. Applying the classical Peaceman-Rachford ADI method for Eq. (1) with $\varphi = 0$ [2] takes:

$$\frac{u^*_{i,j} - u^n_{i,j}}{k/2} = \mu\left((u_{xx})^*_{i,j} + (u_{yy})^n_{i,j}\right),\ \frac{u^{n+1}_{i,j} - u^*_{i,j}}{k/2} = \mu\left((u_{xx})^*_{i,j} + (u_{yy})^{n+1}_{i,j}\right), \tag{4}$$

$$u^n_{i,j} = u(x_i, y_j, t_n),\ u^*_{i,j} = u(x_i, y_j, t_{n+1/2}),\ x_i = ih, y_j = jh,\ t_n = nk, \tag{5}$$

where u as the approximation of U.

For obtaining higher order approximations of u_{xx} and u_{yy} we use central finite difference approximations, which are of fourth order. Furthermore, it should be mentioned that we need the values of the second order derivatives of u on the boundaries in order to solve for u_{xx} and u_{yy}. Obviously, for nonperiodic case some additional relations corresponding to the nodes near the boundary are required. For example, a second order accurate scheme can be written as follows:

$$(u_{xx})^n_{0,j} = \frac{1}{h^2}(2u_{0,j} - 5u_{1,j} + 4u_{2,j} - u_{3,j}),$$

whereas a third order accurate scheme can be written as

$$(u_{xx})^n_{0,j} + (u_{xx})^n_{1,j} = \frac{1}{h^2}(13u_{0,j} - 27u_{1,j} + 15u_{2,j} - u_{3,j}).$$

For other boundaries, one can do similarly.

3. Stability Analysis of Compact ADI Method for Linear Problem

Let $\alpha = U - u$ be the discretization error, then

$$\begin{aligned}
\frac{U^* - U^n}{k} &= \frac{\mu}{2}((U_{xx})^* + (U_{yy})^n) + e^n_1,\\
\frac{U^{n+1} - U^n}{k} &= \mu(U_{xx})^* + \frac{\mu}{2}((U_{yy})^{n+1} + (U_{yy})^n) + e^n_2,\\
e^n_1 &= O(k) + O(h^4),\\
e^n_2 &= O(k^2) + O(h^4).
\end{aligned}$$

Consequently,

$$\begin{aligned} \frac{\alpha^* - \alpha^n}{k} &= \frac{\mu}{2}((\alpha_{xx})^* + (\alpha_{yy})^n) + e_1^n, \\ \frac{\alpha^{n+1} - \alpha^n}{k} &= \mu(\alpha_{xx})^* + \frac{\mu}{2}((\alpha_{yy})^{n+1} + (\alpha_{yy})^n) + e_2^n. \end{aligned} \tag{6}$$

LEMMA 3.1. *At each point $(x_i, y_j) \in \Delta$, for ζ' and ζ_0 satisfying*

$$\frac{\zeta' - \zeta_0}{k} = \frac{\mu}{2}((\zeta')_{xx} + (\zeta_0)_{yy}) + e, \tag{7}$$

then $\zeta' = \left\{ \frac{1 - r\frac{Y_q}{1-\frac{1}{3}Y_q}}{1 + r\frac{X_p}{1-\frac{1}{3}X_p}} \right\} \zeta_0 + \frac{k}{1 + r\frac{X_p}{1-\frac{1}{3}X_p}} e$, *where* $X_p = \sin^2(\frac{\pi p h}{2})$, $Y_p = \sin^2(\frac{\pi q h}{2})$, $r = \frac{2\mu k}{h^2}$.

PROOF. Let

$$\zeta' = \sum_{p,q=1}^{N-1} c'_{p,q} \sin(\pi p x) \sin(\pi q y), \quad \zeta_0 = \sum_{p,q=1}^{N-1} c^0_{p,q} \sin(\pi p x) \sin(\pi q y),$$
$$e = \sum_{p,q=1}^{N-1} e_{p,q} \sin(\pi p x) \sin(\pi q y),$$

where $N = \frac{1}{h}$, $p, q = 1, \ldots, N$. From operating δ_x^2 and δ_y^2 on w, as a symbol for either ζ' or ζ_0, one obtains

$$\delta_x^2 w^n = -\frac{4}{h^2} \sin^2(\frac{\pi p h}{2}) w^n, \quad \delta_y^2 w^n = -\frac{4}{h^2} \sin^2(\frac{\pi q h}{2}) w^n.$$

Now by substituting the obtained relations into (7) the desired result follows. □

LEMMA 3.2. *Let ζ_0 and ζ' satisfy Lemma 3.1. At each point $(x_i, y_j) \in \Delta$ for ζ_1, ζ_0 and ζ' satisfying*

$$\frac{\zeta_1 - \zeta_0}{k} = \mu(\zeta')_{xx} + \frac{\mu}{2}((\zeta_1)_{yy} + (\zeta_0)_{yy}) + e,$$

with $\zeta_1 = \zeta_0 = \zeta' = 0$ on the boundary $\partial\Delta$, we have

$$\|\zeta_1\| \leqslant \|\zeta_0\| + \|e\| k,$$

where

$$\|\zeta\| = \left[h^2 \sum_{(x,y)\in\Delta} \zeta_{i,j}^2 \right]^{1/2}.$$

PROOF. Using Lemma 3.1 and its notations prove the result. □

THEOREM 3.3. *Let the solution U of (1), with $\varphi = 0$ and initial and boundary conditions (2), (3) are sufficiently differentiable. Then the solution u of (4), (5) converges in the L_2 norm, to U with $O(h^4 + k^2)$ discretization error.*

PROOF. Using Eq. (6) and then applying Lemma 3.2 to α^n, α^{n+1} gives the following inequality for error

$$\|\alpha^{n+1}\| \leqslant \|\alpha^n\| + \|e^n\| k, \quad n = 0, 1, 2, \ldots.$$

This verifies the stability of the compact method and also shows that the discretization error is of the fourth order in space and two in time. □

Therefore, for Eq. (1) with nonlinear term φ (with respect to U) this new method includes the following algorithmic steps:

Algorithm 1. The new compact ADI time second order scheme: "Comp-ADI"

1. With a solution at time level t_n solve the following equations for $\tilde{u}^*$
   $\left(1+\frac{h^2}{12}\delta_y^2\right)(u_{yy})_{ij}^n = \delta_y^2 u_{ij}^n,$
   $\frac{\tilde{u}_{i,j}^* - u_{i,j}^n}{k/2} = \mu((\tilde{u}_{xx})_{i,j}^* + (u_{yy})_{i,j}^n) + \varphi(x_i, y_j, t_{n+1/2}, u^n).$
2. Evaluate φ_1 and φ_2 through
   $\tilde{\psi}^n = \varphi(x_i, y_j, t_{n+1/2}, \tilde{u}_{i,j}^*),$
   $\varphi_1 = \varphi_2 := \tilde{\psi}^n.$
3. From $\tilde{\psi}^n$ obtain $u_{i,j}^{n+1}$ from the following equations
   $\frac{u_{i,j}^* - u_{i,j}^n}{k/2} = \mu((u_{xx})_{i,j}^* + (u_{yy})_{i,j}^n) + \varphi_1,$
   $\frac{u_{i,j}^{n+1} - u_{i,j}^*}{k/2} = \mu((u_{xx})_{i,j}^* + (u_{yy})_{i,j}^{n+1}) + \varphi_2.$

Remark 3.4. As we have already mentioned the before, Mitchell-Fairweather scheme has temporal order 2 and spatial order 4, only for linear PDEs and can be improved by our approach to preserve its orders for nonlinear PDEs as well. When is applied to nonlinear problems, its order of accuracy in time decreases to one, while it's joint application with our Algorithm 1 guarantees its second order of accuracy in time for nonlinear problems and actually its performance is similar to our improved method presented here.

4. Numerical Examples

In this section, the computational orders of the presented method with comparative methods that is denoted by C-order is calculated with $\frac{\log e_1 - \log e_2}{\log \Delta_1 - \log \Delta_2}$, where e_1 and e_2 are errors corresponding to grids with spatial or temporal step size Δ_1 and Δ_2, respectively. It should be mentioned again that in the following numerical examples, the Douglas scheme [1] will be denoted by "DougS", the modified Mitchell-Fairweather will be denoted by "MF-App1" and also, the presented compact method will be denoted by "Comp-ADI" which is of order two in time and four in the space.

Example 4.1. In this example the nonlinear Fitzhugh-Nagumo equation with the third degree nonlinear term is solved. This equation arises in population genetics and model the transmission of nerve impulses. Nonlinear term of this equation is $\varphi = -U(1-U)(a-U)$ and $a = 1/4$. In Fitzhugh-Nagumo equation the value of μ is 1. The exact solution of this equation is

$$U(x,y,t) = 1/(1+\exp(-\frac{1}{4}t + \frac{1}{2}(x+y))), \quad 0 \leqslant x, y \leqslant 1, t \in [0,T).$$

The initial and boundary conditions are obtained by imposing the exact solution. In this example we consider $h = \frac{1}{8}$ and simulation runs for $t \leq 0.625$.

The simulation results are illustrated in Figure 1 shows that the rate of convergence of "Comp-ADI" is better than that of "DougS". For instance with $k = 0.008$

at $t = 0.625$, the L_2 error of the solution by "Comp-ADI" is 8.434×10^{-7} whereas that of "DougS" is 6.219×10^{-4} and "MF-App1" is 4.005×10^{-5}.

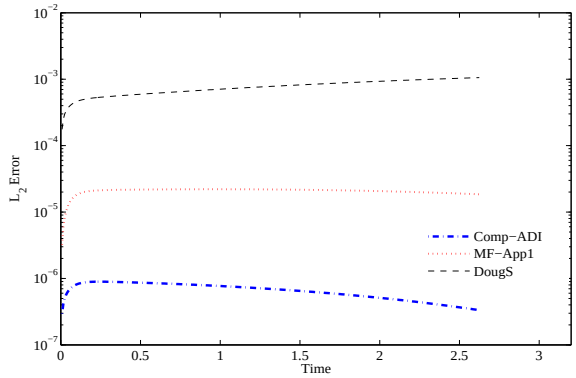


FIGURE 1. L_2 Error of DougS, MF-App1, and Comp-ADI.

5. Conclusion

A new compact scheme based on the ADI method with unconditional stability has been proposed for solving nonlinear 2D unsteady parabolic differential equations. The order of the presented method is two in time and four in the space. Another property of the presented method is that reduces the computational cost, significantly, because of have strictly diagonally dominant coefficient matrices.

References

1. J. Douglas Jr., *Alternating direction methods for three space variables,* Numer. Math. **4** (1962) 41–63.
2. D. W. Peaceman and H. H. Rachford Jr., *The numerical solution of parabolic and elliptic differential equations,* J. Soc. Ind. Appl. Math. **3** (1) (1995) 28–41.
3. P. F. A. Mancera, *A study of a numerical solution of the steady two dimensions Navier-Stokes equations in a constricted channel problem by a compact fourth order method,* Appl. Math. Comput. **146** (2-3) (2003) 771–790.
4. S. Karaa and J. Zhang, *High order ADI method for solving unsteady convection diffusion problems,* J. Comput. Phys. **198** (1) (2004) 1–9.
5. F. Wu, X. Cheng, D. Li and J. Duan, *A two-level linearized compact ADI scheme for two-dimensional nonlinear reaction-diffusion equations,* Comput. Math. Appl. **75** (8) (2018) 2835–2850.

E-mail: s.amiri@ssau.ac.ir; amirimath@yahoo.com

A Preconditioner for Three-by-Three Block Saddle Point Problems

Davod Khojasteh Salkuyeh
Faculty of Mathematical Sciences, University of Guilan, Rasht, Iran
Center of Excellence for Mathematical Modelling, Optimization and Combinational Computing (MMOCC), University of Guilan, Rasht, Iran
Hamed Aslani*
Faculty of Mathematical Sciences, University of Guilan, Rasht, Iran
and Zhao-Zheng Liang
School of Mathematics and Statistics, Lanzhou University, Lanzhou, P. R. China

Abstract. Using the idea of dimensional splitting method we present an iteration method for solving three-by-three block saddle point problems. We prove that the method is convergent unconditionally. The induced preconditioner is used to accelerate the convergence of the GMRES method for solving the problem.
Keywords: Saddle point, Block, Dimensional, Split, Preconditioner, GMRES.
AMS Mathematical Subject Classification [2010]: 65F10, 65F50, 65F08.

1. Introduction

Consider the following block three-by-three system of linear equations

$$\mathcal{A}\mathbf{x} \equiv \begin{pmatrix} A & B^T & 0 \\ -B & 0 & -C^T \\ 0 & C & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} f \\ -g \\ h \end{pmatrix} \equiv \mathbf{b}, \tag{1}$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{m\times n}$ and $C \in \mathbb{R}^{l\times m}$. Here, $f \in \mathbb{R}^n$, $g \in \mathbb{R}^m$ and $h \in \mathbb{R}^l$. In this case, the coefficient matrix of the system (1) is of order $\mathbf{n} \times \mathbf{n}$, where $\mathbf{n} = n + m + l$. Linear system of the form (1) arises from many practical scientific and engineering applications, e.g., the discrete finite element methods for solving time-dependent Maxwell equation with discontinuous coefficient [1], the least squares problems [10] and so on.

Recently Liang and Zhang in [7] established the alternating positive semi-definite splitting (APSS) method for solving the system of linear equations (1) as follows. Consider the decomposition $\mathcal{A} = \mathcal{A}_1 + \mathcal{A}_2$, where

$$\mathcal{A}_1 = \begin{pmatrix} A & B^T & 0 \\ -B & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathcal{A}_2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -C^T \\ 0 & C & 0 \end{pmatrix}. \tag{2}$$

*Speaker

Let $\alpha > 0$. Then the following splittings for the matrix $\mathcal{A}$ can be stated

$$\mathcal{A} = (\alpha\mathcal{I} + \mathcal{A}_1) - (\alpha\mathcal{I} - \mathcal{A}_2) = (\alpha\mathcal{I} + \mathcal{A}_2) - (\alpha\mathcal{I} - \mathcal{A}_1),$$

where $\mathcal{I}$ is the identity matrix of order $\mathbf{n}$. Now, using these splittings the APSS method can be written as

$$\begin{cases} (\alpha\mathcal{I} + \mathcal{A}_1)\mathbf{x}^{(k+\frac{1}{2})} = (\alpha\mathcal{I} - \mathcal{A}_2)\mathbf{x}^{(k)} + \mathbf{b}, \\ (\alpha\mathcal{I} + \mathcal{A}_2)\mathbf{x}^{(k+1)} = (\alpha\mathcal{I} - \mathcal{A}_1)\mathbf{x}^{(k+\frac{1}{2})} + \mathbf{b}, \end{cases}$$

where $\mathbf{x}^{(0)} \in \mathbb{R}^{\mathbf{n}}$ is an initial guess. By eliminating $\mathbf{x}^{(k+\frac{1}{2})}$, the iteration scheme can be rewritten as the stationary form

$$\mathbf{x}^{(k+1)} = \mathcal{T}_\alpha \mathbf{x}^{(k)} + \mathbf{f},$$

with $\mathcal{T}_\alpha = (\alpha\mathcal{I} + \mathcal{A}_2)^{-1}(\alpha\mathcal{I} - \mathcal{A}_1)(\alpha\mathcal{I} + \mathcal{A}_1)^{-1}(\alpha\mathcal{I} - \mathcal{A}_2)$, and $\mathbf{f} = 2\alpha(\alpha\mathcal{I} + \mathcal{A}_2)^{-1}(\alpha\mathcal{I}+\mathcal{A}_1)^{-1}\mathbf{b}$. It is easy to see that if we define $\mathcal{P}_\alpha = \frac{1}{2\alpha}(\alpha\mathcal{I}+\mathcal{A}_1)(\alpha\mathcal{I}+\mathcal{A}_2)$ and $\mathcal{Q}_\alpha = \frac{1}{2\alpha}(\alpha\mathcal{I} - \mathcal{A}_1)(\alpha\mathcal{I} - \mathcal{A}_2)$, then $\mathcal{A} = \mathcal{P}_\alpha - \mathcal{Q}_\alpha$ and $\mathcal{T}_\alpha = \mathcal{P}_\alpha^{-1}\mathcal{Q}_\alpha$.

2. Convergence of the APSS Iteration Method

When the (2,2)-block of $\mathcal{A}$ is symmetric positive definite (SPD), convergence of the method was presented in [7]. In this paper, we prove the convergence of the method when this block is equal to zero. To do so we first state the next lemma.

LEMMA 2.1. *Let A be SPD and the matrices B and C be of full row rank. Then $\rho(\mathcal{T}_\alpha) \leq 1$, where $\rho(.)$ denotes the spectral radius of the matrix.*

PROOF. Using the Kelloge's lemma (See [8]), it can be easily proved. □

From Lemma 2.1, for the convergence of the APSS iteration method it is enough to prove that $\rho(\mathcal{T}_\alpha) = 1$ never happens. To do so we state the following lemma.

LEMMA 2.2. *Let A be SPD and matrices B and C be of full row rank. Then the following are equivalent:*

i) *The matrix*

$$\mathcal{G}_\alpha = \begin{pmatrix} A & B^T + \frac{1}{\alpha^2}B^TC^TC & 0 \\ -B & 0 & -C^T \\ 0 & C & 0 \end{pmatrix},$$

does not have any purely imaginary eigenvalue.

ii) $\rho(\mathcal{T}_\alpha) < 1$.

PROOF. Similar to the proof of [3, Lemma 2], let λ be an eigenvalue of $\mathcal{T}_\alpha$. Then, obviousely $\lambda = 1 - \mu$, where μ is an eigenvalue of the matrix $\mathcal{P}_\alpha^{-1}\mathcal{A}$. Let $(\mu, \mathbf{x})$ be an eigenpair of $\mathcal{P}_\alpha^{-1}\mathcal{A}$. Then, we have $\mathcal{A}\mathbf{x} = \mu\mathcal{P}_\alpha\mathbf{x}$ which is equivalent to $\mathcal{A}\mathbf{x} = \frac{\mu}{2\alpha}(\mathcal{A}_1\mathcal{A}_2 + \alpha\mathcal{A} + \alpha^2\mathcal{I})\mathbf{x}$, or

$$\left(1 - \frac{1}{2}\mu\right)\mathcal{A}\mathbf{x} = \frac{\mu\alpha}{2}(\mathcal{I} + \frac{1}{\alpha^2}\mathcal{A}_1\mathcal{A}_2)\mathbf{x}. \tag{3}$$

Direct computation reveals that

$$\mathcal{H}_\alpha := \mathcal{I} + \frac{1}{\alpha^2}\mathcal{A}_1\mathcal{A}_2 = \begin{pmatrix} I & 0 & -\frac{1}{\alpha^2}B^TC^T \\ 0 & I & 0 \\ 0 & 0 & I \end{pmatrix},$$

which is obviously nonsingular. Since, both the matrices $\mathcal{A}$ and $\mathcal{H}_\alpha$ are nonsingular we deduce that $\mu \neq 0$ and $\mu \neq 2$. Then, from (3) we have $\mathcal{H}_\alpha^{-1}\mathcal{A}\mathbf{x} = \frac{\mu\alpha}{2-\mu}\mathbf{x}$. This shows that $\theta := \frac{\mu\alpha}{2-\mu}$, is an eigenvalue of

$$\mathcal{H}_\alpha^{-1}\mathcal{A} = \begin{pmatrix} A & B^T + \frac{1}{\alpha^2}B^TC^TC & 0 \\ -B & 0 & -C^T \\ 0 & C & 0 \end{pmatrix} = \mathcal{G}_\alpha. \tag{4}$$

Now, we see that $\mu = \frac{2\theta}{\alpha+\theta}$, and as a result $\lambda = 1-\mu = 1-\frac{2\theta}{\alpha+\theta} = \frac{\alpha-\theta}{\alpha+\theta}$. Hence, from Lemma 2.1 we get $|\lambda| = \left|\frac{\alpha-\theta}{\alpha+\theta}\right| \leq 1$, and $|\lambda| = 1$ if and only if $|\theta-\alpha| = |\theta+\alpha|$ which is itself equivalent to $(\Re(\theta)-\alpha)^2 + \Im(\theta)^2 = (\Re(\theta)+\alpha)^2 + \Im(\theta)^2$. The latter equation is equivalent to $\Re(\theta) = 0$. Therefore, $\rho(\mathcal{T}_\alpha) = 1$ if and only if $\mathcal{G}_\alpha$ has at least one purely imaginary eigenvalue. □

THEOREM 2.3. *Let A be SPD and matrices B and C be of full row rank. Then, the APSS iteration method unconditionally converges to the solution of (1) i.e., $\rho(\mathcal{T}_\alpha) < 1$, for all $\alpha > 0$.*

PROOF. According to Lemma 2.2, all we need is to prove that the matrix $\mathcal{G}_\alpha$ defined in (4) has no purely imaginary eigenvalue. Let $(\theta, \mathbf{x})$ be an eigenpair of the matrix $\mathcal{G}_\alpha$ with $\|\mathbf{x}\|_2 = 1$. Clearly, the matrix $\mathcal{G}_\alpha$ is nonsingular, therefore $\theta \neq 0$. Letting $\mathbf{x} = (u; v; p)$, it follows from $\mathcal{G}_\alpha\mathbf{x} = \theta\mathbf{x}$ that

$$\begin{aligned} Au + B^T(I + \frac{1}{\alpha^2}C^TC)v &= \theta u, \\ -Bu - C^Tp &= \theta v, \\ Cv &= \theta p. \end{aligned} \tag{5}$$

It is easy to see that the vectors u, v and p can not be zero. Hereafter, we assume that $u \neq 0$, $v \neq 0$ and $p \neq 0$. From $\mathcal{G}_\alpha\mathbf{x} = \theta\mathbf{x}$ and $\|\mathbf{x}\|_2 = 1$, we get $\theta = \mathbf{x}^*\mathcal{G}_\alpha\mathbf{x}$. Hence,

$$\begin{aligned} \Re(\theta) &= \frac{1}{2}\mathbf{x}^*(\mathcal{G}_\alpha + \mathcal{G}_\alpha^T)\mathbf{x} = \frac{1}{2}\begin{pmatrix} u^* & v^* & p^* \end{pmatrix}\begin{pmatrix} 2A & \frac{1}{\alpha^2}B^TC^TC & 0 \\ \frac{1}{\alpha^2}C^TCB & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} u \\ v \\ p \end{pmatrix} \\ &= u^*Au + \frac{1}{2\alpha^2}\left(u^*B^TC^TCv + v^*C^TCBu\right). \end{aligned} \tag{6}$$

On the other hand, from Eq. (5) we deduce $u^*Au + u^*B^T(I + \frac{1}{\alpha^2}C^TC)v = \theta u^*u$, and taking the conjugate of both sides, the latter equation gives $u^*Au + v^*(I + \frac{1}{\alpha^2}C^TC)Bu = \bar{\theta}u^*u$.

From the last two equations we get

$$u^*Au + \frac{1}{2}(u^*B^Tv + v^*Bu) + \frac{1}{2\alpha^2}\left(u^*B^TC^TCv + v^*C^TCBu\right) = \Re(\theta)u^*u. \tag{7}$$

From Eqs. (6) and (7) we obtain $\Re(\theta) + \frac{1}{2}(u^*B^Tv + v^*Bu) = \Re(\theta)u^*u$. Now, by contradiction we assume that $\Re(\theta) = 0$. In this case, from the above equation we deduce that

$$u^*B^Tv + v^*Bu = 0. \tag{8}$$

On the other hand, by easy manipulations we have

$$\Re(\theta) = u^*Au - \frac{1}{2\alpha^2}\left(2\Re(\theta)\|Bu\|_2^2 + \theta^2 u^*B^Tv + \bar{\theta}^2 v^*Bu\right).$$

Now, if $\Re(\theta) = 0$, then $\theta = i\xi$, where $\xi \neq 0$. Therefore, from the above equation we see that

$$\begin{aligned} 0 &= u^*Au + \frac{\xi^2}{2\alpha^2}\left(u^*B^Tv + v^*Bu\right) \\ &= u^*Au, \qquad \text{(From Eq. (8))}, \end{aligned}$$

which is a contraction, since $u \neq 0$ and A is SPD. Therefore, the proof is completed. □

3. Numerical Experiments

We present some numerical results to show the efficiency of the induced preconditioner. We first apply a symmetric diagonal scaling for the matrix $\mathcal{A}$. To do so, we replace the coefficient matrix $\mathcal{A}$ by the matrix $\mathcal{D}^{-\frac{1}{2}}\mathcal{A}\mathcal{D}^{-\frac{1}{2}}$, where $\mathcal{D} = diag(\|\mathcal{A}_1\|_2, \ldots, \|\mathcal{A}_\mathbf{n}\|_2)$ in which $\mathcal{A}_j$ is the jth column of the matrix $\mathcal{A}$. The right-hand side vector of the system is set $b = \mathcal{A}\mathbf{e}$, where $\mathbf{e}$ is a vector of all ones. We use the flexible version of the GMRES(50) [9], FGMRES(50), for solving the systems. The iteration is started from a zero vector and terminated as soon as the residual 2-norm is reduced by a factor of 10^{-6}. The maximum number of iterations is set to be 20000.

For the APSS preconditioner the subsystems are solved using the conjugate gradient (CG) method. The CG method is started from a zero vector and the iteration is stopped as soon as the residual 2-norm is reduced by a factor of 10^{-3}. The maximum number of CG iterations is set to be 200. We compare the numerical results of the APSS preconditioner, $\mathcal{M}_\alpha = (\alpha\mathcal{I} + \mathcal{A}_1)(\alpha\mathcal{I} + \mathcal{A}_2)$, in which $\mathcal{A}_1$ and $\mathcal{A}_2$ were defined in (2) with $\mathcal{P}_D$ which is presented in [6]. We report the number of iteration (denoted as "IT") and the elapsed CPU time in second (denoted as "CPU"). The value of R_k defined by $R_k = \|\mathbf{b} - \mathcal{A}\mathbf{x}^{(k)}\|_2/\|\mathbf{b}\|_2$, is also reported where $x^{(k)}$ is the computed solution at iteration k. Finally, "NA" (for Not Applicable) means that the coefficient matrix does not satisfy the assumptions of Theorem 2.3. All runs are implemented in MATLAB R2017, equipped with a Laptop with 1.80 GHz central processing unit (Intel(R) Core(TM) i7-4500), 6 GB memory and Windows 7 operating system.

We consider the problem (See [5])

$$\min_{x\in\mathbb{R}^n, y\in\mathbb{R}^l} \frac{1}{2}x^TAx + r^Tx + q^Ty \ s.t.: \ Bx + C^Ty = b, \tag{9}$$

where $r \in \mathbb{R}^n$ and $q \in \mathbb{R}^l$. To solve the above problem we define the Lagrange function

$$L(x, y, \lambda) = \frac{1}{2}x^TAx + r^Tx + q^Ty + \lambda^T(Bx + C^Ty - b),$$

TABLE 1. Numerical results.

Precond.		MOSARQP1	Liswet12
I	IT	115	77
	CPU	0.32	0.82
	RES	8.9e-07	9.9e-07
$\mathcal{M}_\alpha$	α	0.05	0.5
	IT	6	21
	CPU	0.03	0.22
	RES	5.2e-7	9.0e-7
$\mathcal{P}_D$	IT	NA	38
	CPU	-	1.28
	RES	-	9.1e-7

where the vector $\lambda \in \mathbb{R}^m$ is the Lagrange multiplier. Then the Karush-Kuhn-Tucker necessary conditions of (9) are as following (See [2])

$$\nabla_x L(x, y, \lambda) = 0, \quad \nabla_y L(x, y, \lambda) = 0 \quad \text{and} \quad \nabla_\lambda L(x, y, \lambda) = 0.$$

These equations give a three-by-three saddle point of the form (1). In this example, we have chosen the matrices A, B and C from the CUTEr collection [4]. Numerical results are presented in Table 1. As seen, $\mathcal{M}_\alpha$ is superior to the other examined preconditioners.

References

1. F. Assous, P. Degond, E. Heintze, P. A. Raviart and J. Segre, *On a finite-element method for solving the three-dimensional Maxwell equations*, J. Comput. Phys. **109** (1993) 222–237.
2. D. P. Bertsekas, *Nonlinear Programming*, 2nd ed., Athena Scientific, 1999.
3. M. Benzi and X. P. Guo, *A dimensional split preconditioner for Stokes and linearized Navier-Stokes equations*, Appl. Numer. Math. **61** (2011) 66–76.
4. N. I. M. Gould, D. Orban and P. L. Toint, *CUTEr and SifDec: A constrained and unconstrained testing environment, revisited,* ACM Trans. Math. Softw. **29** (2003) 373–394.
5. N. Huang, *Variable parameter Uzawa method for solving a class of block three-by-three saddle point problems*, Numer. Algor. **85** (2020) 1233–1254.
6. N. Huang and C. F. Ma, *Spectral analysis of the preconditioned system for the* 3×3 *block saddle point problem*, Numer. Algor. **81** (2019) 421–444.
7. Z. Z. Liang and G. F. Zhang, *Alternating positive semi-definite splitting preconditioner for double saddle point problems*, Calcolo **56** (2019) 26.
8. G. I. Marchuk, *Methods of Numerical Mathematics*, Springer-Verlag, New York, 1984.
9. Y. Saad, *Iterative Methods for Sparse Linear Systems,* PWS Press, New York, 1995.
10. J. Y. Yuan, *Numerical methods for generalized least squares problems*, J. Comput. Appl. Math. **66** (1996) 571–584.

E-mail: khojasteh@guilan.ac.ir
E-mail: hamedaslani525@gmail.com
E-mail: liangzz@lzu.edu.cn

A Stable Hybridized Discontinuous Galerkin Method for the Telegraph Equation

Shima Baharlouei*
Department of Mathematical Sciences, Isfahan University of Technology, Isfahan 84156-83111, Iran
and Reza Mokhtari
Department of Mathematical Sciences, Isfahan University of Technology, Isfahan 84156-83111, Iran

Abstract. In this paper, we present a hybridized discontinuous Galerkin (HDG) method for solving the telegraph equation. Stability of the method is perused during a theorem for periodic and Dirichlet boundary conditions. Moreover convergence of the HDG method is investigated by testing some numerical examples and we observe optimal convergence order for the approximate solution and its first temporal and spatial derivatives.
Keywords: Hybridized discontinuous Galerkin method, Telegraph equation, Stability analysis.
AMS Mathematical Subject Classification [2010]: 65M60, 65M12.

1. Introduction

One of the most important of partial differential equations is called telegraph equation which by counting the effects of finite velocity to standard heat or mass transport equation, peresents a mixed diffusion and wave propagation model [3].The general one-dimensional telegraph equation is defined as following that should be equipped by suitable initial and boundary conditions

$$\mathfrak{u}_{tt}(x,t)+\alpha\mathfrak{u}_t(x,t)+\beta\mathfrak{u}(x,t)=\mathfrak{u}_{xx}(x,t)+f(x,t),\quad x\in\Omega\subset\mathbb{R},\quad t\in[0,T], \tag{1}$$

where α and β are known constant. To date, due to the special importance of the telegraph equation in applications [5, 7], numerous numerical methods have been employed to numerically solve this equation [1, 2, 4, 6].

In this paper we present an HDG method for solving the telegraph equation numerically. The partitioning $x_L = x_{-\frac{1}{2}} < x_{\frac{1}{2}} < \cdots < x_{N-\frac{1}{2}} = x_R$ will be considered for the domain $\Omega = [x_L, x_R]$ and $\mathcal{T}_h := \{K_j\}$ displays the finite collection of disjoint elements which $K_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$ represents the j-th element for $j = 0, \ldots, N-1$. Also, we define the collection of boundaries of elements as $\partial\mathcal{T}_h = \{x^+_{-\frac{1}{2}}, x^{\pm}_{\frac{1}{2}}, \ldots, x^{\pm}_{N-\frac{3}{2}}, x^-_{N-\frac{1}{2}}\}$. The concept of jump is defined as $[[v\mathbf{n}]] = v^+\mathbf{n}^+ + v^-\mathbf{n}^-$ and $[[v\mathbf{n}]] = v\mathbf{n}$ for interior faces (ε_h^0) and boundary faces (ε_h^b), respectively. Here v^+ and v^- on face e represent respectively $v(e^+)$ and $v(e^-)$. Also we define the set of all of faces as $\varepsilon_h = \varepsilon_h^\partial \cup \varepsilon_h^0$. It is noteworthy that, $\mathbf{n}^-_{j+\frac{1}{2}} = +1$ and $\mathbf{n}^+_{j-\frac{1}{2}} = -1$ are outward unit normal vectors for any K_j. Consider

*Speaker

$\mathcal{P}^k(K)$ be the set of polynomials of degree at most k on the element $K \in \mathcal{T}_h$. W_h^k and M_h^k are called discontinuous finite element space and skeleton space (or trace space), respectively and defined by

$$\begin{aligned} W_h^k &= \Big\{w \in L^2(\Omega) : w\mid_K \in \mathcal{P}^k(K), \forall K \in \mathcal{T}_h\Big\}, \\ M_h^k &= \{\mu \in L^2(\varepsilon_h) : \mu\mid_e \in \mathcal{P}^k(e), \forall e \in \varepsilon_h\}. \end{aligned}$$

Moreover, according to boundary conditions, and displaying the L^2 projection into the skeleton space by Π, we can define the following useful subspace of M_h^k

$$M_h^{k,\mathfrak{u}}(l) := \{\mu \in M_h^k : \mu(x) = \Pi l(x), \quad x \in \Gamma_{\mathfrak{u}}\},$$

where $\Gamma_{\mathfrak{u}}$ is the set of boundary faces which boundary data are specified on $\mathfrak{u}$. By considering $(w_1, w_2)_K = \int_K w_1(x)w_2(x)dx$ and $< \mu_1, \mu_2 >_{\partial K} = \mu^-_{1,j+\frac{1}{2}}\mu^-_{2,j+\frac{1}{2}} + \mu^+_{1,j-\frac{1}{2}}\mu^+_{2,j-\frac{1}{2}}$ as an inner product and inner product on the boundaries of elements respectively, we have $(w_1, w_2)_{\mathcal{T}_h} = \sum_{K\in\mathcal{T}_h}(w_1, w_2)_K$ and $< \mu_1, \mu_2 >_{\partial\mathcal{T}_h} = \sum_{K\in\mathcal{T}_h} < \mu_1, \mu_2 >_{\partial K}$, where $w_1,\ w_2$ are defined on $\mathcal{T}_h$, and $\mu_1,\ \mu_2$ are defined on $\partial\mathcal{T}_h$.

2. Accomplishment of the HDG Method

The first step of the HDG method is obtaining the first-order system of equations of (1). By defining $\mathfrak{p} = \mathfrak{u}_x$ and $\mathfrak{v} = \mathfrak{u}_t$, the set of first-order equations corresponding to (1) are as follows

$$\mathfrak{v}_t + \alpha\mathfrak{u}_t + \beta\mathfrak{u} - \mathfrak{p}_x = f, \quad \mathfrak{p} - \mathfrak{u}_x = 0, \quad \mathfrak{v} - \mathfrak{u}_t = 0. \tag{2}$$

In corresponding weak form of (2), we intend to find $u, p, v \in W_h^k$ such that

$$\begin{aligned} (v_t, w_1)_{K_j} + \alpha(u_t, w_1)_{K_j} + \beta(u, w_1)_{K_j} + (p, w_{1x})_{K_j} + < \widehat{-p}\mathbf{n}, w_1 >_{\partial K_j} &= (f, w_1)_{K_j}, \\ (p, w_2)_{K_j} + (u, w_{2x})_{K_j} - < \hat{u}, w_2\mathbf{n} >_{\partial K_j} &= 0, \\ (v, w_3)_{K_j} - (u_t, w_3)_{K_j} &= 0, \end{aligned} \tag{3}$$

where $w_1, w_2, w_3 \in W_h^k$ are test functions and numerical flux $\widehat{-p}$ is defined as $\widehat{-p} = -p + \tau(u - \hat{u})\mathbf{n}$. Here τ is stabilization parameter and $\hat{u} \in M_h^{k,0}(b_{\mathfrak{u}})$ is called numerical trace, where $b_{\mathfrak{u}}$ denote the value of the boundary data on $\mathfrak{u}$. In detailed numerical trace $\hat{u}$ is defined as

$$\hat{u} = \begin{cases} b_{\mathfrak{u}}, & \partial K_j \cap \Gamma_{\mathfrak{u}}, \\ \lambda, & \partial K_j \setminus \Gamma_{\mathfrak{u}}, \end{cases}$$

where $\lambda \in M_h^{k,0}(0)$. Unlike the local unknowns u and p, $\hat{u}$ is a global unknown. By enforcing conservation of the numerical flux (or trace) on the element edges, one extra global equation is obtained which helps us to find the global unknown. Hence we have

$$\begin{cases} p\mathbf{n} = b_{\mathfrak{p}}, & e \in \Gamma_{\mathfrak{p}}, \\ [[\widehat{-p}\mathbf{n}]] = 0, & e \in \varepsilon_i, \end{cases} \tag{4}$$

where $\Gamma_{\mathfrak{p}}$ denotes the set of faces which boundary data are specified on $\mathfrak{p}$. Therefore, by using backward Euler method for time discretization and imposing the

definitions of numerical flux $\widehat{-p}$ and numerical trace $\hat{u}$, (3) and (4) change to

$$\begin{aligned}\frac{1}{\Delta t}(v,w_1)_{\mathcal{T}_h} &+ (\frac{\alpha}{\Delta t}+\beta)(u,w_1)_{\mathcal{T}_h}-(p_x,w_1)_{\mathcal{T}_h}\\ &+ <\tau u, w_1>_{\partial\mathcal{T}_h}-<\tau\lambda, w_1>_{\partial\mathcal{T}_h}= l_1(w_1),\end{aligned}$$

$$\begin{aligned}&(p,w_2)_{\mathcal{T}_h}+(u,w_{2x})_{\mathcal{T}_h}-<\lambda, w_2\mathbf{n}>_{\partial\mathcal{T}_h}= l_2(w_2),\\ &(v,w_3)_{\mathcal{T}_h}-\frac{1}{\Delta t}(u,w_3)_{\mathcal{T}_h}=l_3(w_3),\\ &<-p\mathbf{n},\mu>_{\partial\mathcal{T}_h}+<\tau u,\mu>_{\partial\mathcal{T}_h\setminus\varepsilon_h^\partial}-<\lambda,\mu>_{\partial\mathcal{T}_h\setminus\varepsilon_h^\partial}= l_4(\mu),\end{aligned}$$

where $\mu\in M_k^{k,0}(0)$ and

$$\begin{aligned}l_1(w_1) &= (f,w_1)_{\mathcal{T}_h}+<\tau b_{\mathrm{u}}, w_1>_{\Gamma_{\mathrm{u}}}+\frac{1}{\Delta t}(v^{n-1},w_1)_{\mathcal{T}_h}+\frac{\alpha}{\Delta t}(u^{n-1},w_1)_{\mathcal{T}_h},\\ l_2(w_2) &= <b_{\mathrm{u}}, w_2\mathbf{n}>_{\Gamma_{\mathrm{u}}},\ l_3(w_3)=\frac{1}{\Delta t}(u^{n-1},w_3)_{\mathcal{T}_h},\ l_4(\mu)=-<b_u,\mu>_{\Gamma_{\mathrm{p}}}.\end{aligned}$$

3. Main Results

In the sequal, during a theorem stability of the proposed HDG method for (1) will be checked.

THEOREM 3.1. *If $\tau>0$ then the proposed HDG method is stable provided that (1) is equipped with periodic boundary conditions.*

PROOF. Setting $w_1=u$ and $w_2=p$ in the two first equations of (3) and summing together, we have

$$\int_{K_j} v_t u\,dx+\frac{\alpha}{2}\frac{\partial}{\partial t}\|u\|_{K_j}^2+\int_{K_j}p^2\,dx+\beta\int_{K_j}u^2\,dx+\Theta_{K_j}=0, \tag{5}$$

where

$$\begin{aligned}\Theta_{K_j} &= \int_{K_j}(up)_x dx+<\widehat{-p}\mathbf{n}, u>_{\partial K_j}-<\hat{u}, p\mathbf{n}>_{\partial K_h},\\ &= \Big((\widehat{-p})^-_{j+\frac12}-(-p)^-_{j+\frac12}\Big)u^-_{j+\frac12}+\Big((-p)^+_{j-\frac12}-(\widehat{-p})^+_{j-\frac12}\Big)u^+_{j-\frac12}\\ &+ \hat{u}_{j+\frac12}(-p)^-_{j+\frac12}-\hat{u}_{j-\frac12}(-p)^+_{j-\frac12}.\end{aligned}$$

On the other hand from the last equation of (3), and substituting $w_3=-u_t$ and $w_3=v$, respectively we have

$$-\int_{K_j}vu_t\,dx+\int_{K_j}u_t^2\,dx=0,\quad \int_{K_j}v^2\,dx-\int_{K_j}u_tv\,dx=0. \tag{6}$$

Now by summing equations of (6) together, we get

$$\int_{K_j}(uv)_t\,dx=\int_{K_j}u_t^2\,dx+\int_{K_j}v^2. \tag{7}$$

Again, by summing (7) and the first equation of (6) we have

$$\int_{K_j}v_tu\,dx=\int_{K_j}(vu)_t\,dx-\int_{K_j}vu_t\,dx=\int_{K_j}v^2. \tag{8}$$

So, by using (8) in (5) we gain

$$(9)\quad \begin{aligned}\int_{K_j} v_t u\ dx \quad &+ \quad \frac{\alpha}{2}\frac{\partial}{\partial t}\|u\|^2_{K_j} + \int_{K_j} p^2\ dx + \beta\int_{K_j} u^2\ dx + \Theta_{K_j}\\ &= \quad \int_{K_j} v^2 + \frac{\alpha}{2}\frac{\partial}{\partial t}\|u\|^2_{K_j} + \int_{K_j} p^2\ dx + \beta\int_{K_j} u^2\ dx + \Theta_{K_j} = 0.\end{aligned}$$

Now by summing over all elements, (9) leads to

$$\int_{\mathcal{T}_h} v^2 + \frac{\alpha}{2}\frac{\partial}{\partial t}\|u\|^2_{\mathcal{T}_h} + \int_{\mathcal{T}_h} p^2\ dx + \beta\int_{\mathcal{T}_h} u^2\ dx + \Theta = 0,$$

where $\Theta = \sum_{K_j\in\mathcal{T}_h}\Theta_{K_j}$. As we know $\int_{\mathcal{T}_h} v^2, \int_{\mathcal{T}_h} p^2\ dx, \int_{\mathcal{T}_h} u^2\ dx \geq 0$. The aim is to prove that $\Theta \geq 0$. For this purpose, let us rewrite numerical flux $\widehat{-p}$ as below

$$(10)\quad \begin{aligned}\widehat{-p}^{-}_{j+\frac{1}{2}} &= -p^{-}_{j+\frac{1}{2}} + \tau^{-}_{j+\frac{1}{2}}\big(u^{-}_{j+\frac{1}{2}} - \hat{u}_{j+\frac{1}{2}}\big),\\ \widehat{-p}^{+}_{j-\frac{1}{2}} &= -p^{+}_{1,j-\frac{1}{2}} - \tau^{+}_{j-\frac{1}{2}}\big(u^{+}_{j-\frac{1}{2}} - \hat{u}_{j-\frac{1}{2}}\big).\end{aligned}$$

By substituting $(-p)^{-}_{j+\frac{1}{2}}$ and $(-p)^{+}_{j-\frac{1}{2}}$ from (10) into Θ_{K_j}, doing some manipulations, we get

$$\begin{aligned}\Theta_{K_j} &= \tau^{-}_{j+\frac{1}{2}}((u^{-}_{j+\frac{1}{2}})^2 - \hat{u}_{j+\frac{1}{2}}u^{-}_{j+\frac{1}{2}}) + \tau^{+}_{j-\frac{1}{2}}((u^{+}_{j-\frac{1}{2}})^2 - \hat{u}_{j-\frac{1}{2}}u^{+}_{j-\frac{1}{2}})\\ &+ \tau^{-}_{j+\frac{1}{2}}((\hat{u}_{j+\frac{1}{2}})^2 - \hat{u}_{j+\frac{1}{2}}u^{-}_{j+\frac{1}{2}}) + \tau^{+}_{j-\frac{1}{2}}((\hat{u}_{j-\frac{1}{2}})^2 - \hat{u}_{j-\frac{1}{2}}u^{+}_{j-\frac{1}{2}})\\ &- (\widehat{-p})^{-}_{j+\frac{1}{2}}\hat{u}_{j+\frac{1}{2}} - (\widehat{-p})^{+}_{j-\frac{1}{2}}\hat{u}_{j-\frac{1}{2}}.\end{aligned}$$

Now by summing over all elements, applying the conservation condition, and imposing periodic boundary conditions, we get

$$\Theta = \sum_{K_j\in\mathcal{T}_h}\Theta_{K_j} = \sum_j \tau^{-}_{j+\frac{1}{2}}\Big(u^{-}_{j+\frac{1}{2}} - \hat{u}_{j+\frac{1}{2}}\Big)^2 + \tau^{+}_{j-\frac{1}{2}}\Big(u^{+}_{j-\frac{1}{2}} - \hat{u}_{j-\frac{1}{2}}\Big)^2,$$

which is nonnegative, for all $\tau^{\mp}_{j\pm\frac{1}{2}} > 0$, or simply $\tau > 0$. Hence $\frac{1}{2}\frac{\partial}{\partial t}\|u\|^2_{\mathcal{T}_h} \leq 0$ and the method is stable. □

Corollary 3.2. *The proposed HDG method is stable if $\tau > 0$ and (1) is equipped with the boundary conditions $\mathfrak{u}(x_L, .) = \mathfrak{u}(x_R, .) = 0$.*

Example 3.3. Consider (1) with $\alpha = \beta = 1$ and $\Omega = [0, 4]$. Clearly f and the initial and boundary conditions can be extracted by the exact solution $\mathfrak{u}(x, t) = \exp(x - t)$ [2]. In Table 1 the L^2 error norm of $\mathfrak{u}$ and its derivative are given for polynomials of degree $k = 1,\ 2,\ 3$ and $\tau = 10$. As we expected u and its first temporal and spatial derivatives converge with order $k + 1$ (optimal convergence).

Example 3.4. In this example, consider (1) with periodic boundary conditions and analytical solution $\mathfrak{u}(x, t) = \exp(-t)\sin(x)$ in $\Omega = [-\pi, \pi]$ [2]. Here $f(x, t) = (2-\alpha+\beta)\exp(-t)\sin(x)$, $\mathfrak{u}(x, 0) = \sin(x)$, and $\mathfrak{u}_t(x, 0) = -\sin(x)$. In Figure 1 the numerical solution and L^2 error norms are shown for $\alpha = 6, \beta = 2$ and polynomial of degree two with $\tau = 20$. All the results are satisfactory as it turns out.

TABLE 1. L^2 error norms and corresponding numerical orders of accuracy of the associated numerical solution for Example 3.3 with $\tau = 10$ at time $T = 1$.

	Number of elements	$\|u - \mathfrak{u}\|_\Omega$	order	$\|v - \mathfrak{v}\|_\Omega$	order	$\|p - \mathfrak{p}\|_\Omega$	order
$k = 1$	10	2.8941 E-1		1.1944		7.0167 E-1	
	20	7.5230 E-2	1.94	3.1782 E-1	1.91	2.2838 E-1	1.62
	40	1.8971 E-2	1.99	8.1407 E-2	1.96	6.6768 E-2	1.77
	80	4.7456 E-3	2.00	2.0649 E-2	1.98	1.8111 E-2	1.88
$k = 2$	10	2.4691 E-1		1.1698		3.3611 E-1	
	20	3.5212 E-2	2.81	1.5886 E-1	2.88	6.3588 E-2	2.40
	40	4.5100 E-3	2.96	2.0469 E-2	2.96	9.1500 E-3	2.80
	80	5.6575 E-4	2.99	2.5892 E-3	2.98	1.2042 E-3	2.93
$k = 3$	10	2.4690 E-1		1.1696		3.3501 E-1	
	20	1.7841 E-2	3.80	8.0586 E-2	3.86	3.3870 E-2	3.31
	40	1.1309 E-3	3.98	5.1658 E-3	3.96	2.3798 E-3	3.83
	80	7.0749 E-5	4.00	3.2502 E-4	3.99	1.5298 E-4	3.96

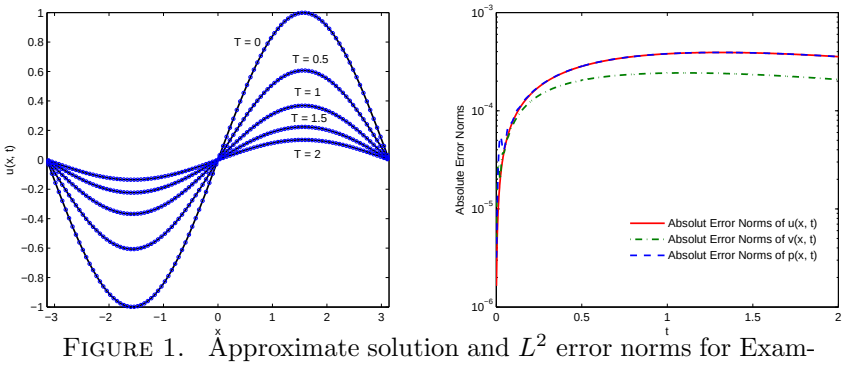


FIGURE 1. Approximate solution and L^2 error norms for Example 3.4 with $\tau = 20$ at time $T = 2$.

References

1. D. Arslan, *The numerical study of a hybrid method for solving telegraph equation*, Appl. Math. Comput. **5** (1) (2020) 293–302.
2. M. Dehghan and A. Shokri, *A numerical method for solving the hyperbolic telegraph equation*, Numer. Methods Partial Differ. Equ. **24** (4) (2008) 1080–1093.
3. M. S. El-Azab and M. El-Gamel, *A numerical algorithm for the solution of telegraph equations*, Appl. Math. Comput. **190** (1) (2007) 757–764.
4. D. Irk and E. Kirli, *Numerical solution of the homogeneous telegraph equation by using galerkin finite element method*, Int. Conf. Comput. Math. Eng., Springer, Cham. Sci. (2019) 209–217.
5. A. C. Metaxas and R. J. Meredith, *Industrial Microwave Heating*, Peter Peregrinus, London, 1993.
6. K. Nagaveni, *Haar wavelet collocation method for solving the telegraph equation with variable coefficients*, Int. J. Appl. Eng. Res. **15** (3) (2020) 235–243.
7. G. Roussy and J. A. Pearcy, *Foundations and Industrial Applications of Microwaves and Radio Frequency Fields*, Wiley and Sons, New York, 1995.

E-mail: s.baharloui@math.iut.ac.ir
E-mail: mokhtari@iut.ac.ir

A Numerical Scheme for Solving the Time-Fractional Stochastic Diffusion Equation via Orthonormal Chebyshev Polynomials

Afshin Babaei*
Department of Mathematics, University of Mazandaran, Babolsar, Iran
Hossein Jafari
Department of Mathematics, University of Mazandaran, Babolsar, Iran
and Seyedeh Seddigheh Banihashemi
Department of Mathematics, University of Mazandaran, Babolsar, Iran

Abstract. In this paper, a spectral collocation approach based on the sixth-kind Chebyshev polynomials (SKCPs) is constructed to solve a time-fractional stochastic diffusion equation (TFSDE). This method is applied to convert the solution of TFSDE to the solution of a system of nonlinear algebraic equations (NAEqs). Moreover, the convergence analysis of this suggested method is established. A numerical example is implemented to validate the efficiency of the proposed approach.
Keywords: Fractional calculus, Stochastic diffusion equation, Collocation scheme, Convergence analysis.
AMS Mathematical Subject Classification [2010]: 60H35, 26A33.

1. Introduction

In this paper, we consider the following time-fractional stochastic diffusion equation

$$\partial_t^\alpha u + \eta \Delta u = \mathbf{F}(x,t,u) + \mathbf{G}(t,u)\dot{\mathrm{B}}, \qquad \text{in} \quad \Omega \times (0,\mathtt{T}), \tag{1}$$

with the initial and boundary conditions

$$u(x,0) = u_0(x), \qquad \text{in} \quad \Omega, \tag{2}$$

$$u(x,t) = \rho(x,t), \qquad \text{in} \quad \partial\Omega \times (0,\mathtt{T}), \tag{3}$$

where $\mathtt{L}, \mathtt{T} \in \mathbb{R}^+$, $\Omega := [0,\mathtt{L}]$ and η is a positive constant. Also, $\mathbf{F} \in \mathrm{C}^1(\Omega \times (0,\mathtt{T}) \times \mathbb{R})$ and $\mathbf{G} \in \mathrm{C}^1((0,\mathtt{T}) \times \mathbb{R})$ satisfy the Lipschitz condition with respect to u and $\dot{\mathrm{B}}(t) := \frac{\mathrm{dB}(t)}{\mathrm{d}t}$ denotes a time white noise [1]. Moreover, $u_0(x)$ and $\rho(x,t)$ are the continuous functions and the operator $\partial_t^\alpha[\cdot]$ denotes the Caputo fractional derivative of order α defined as [2]:

$$\partial_t^\alpha u(x,t) = \frac{1}{\Gamma(1-\alpha)} \int_0^t (t-s)^{-\alpha} \frac{\partial u}{\partial s}(x,s)\mathrm{d}s, \quad \alpha \in (0,1),$$

where $\Gamma(\cdot)$ shows the Gamma function.

*Speaker

2. The Shifted SKCPs and their Properties

DEFINITION 2.1. The shifted SKCPs on $[0, \mathtt{T}]$ are defined by [3]:

$$\psi_j(t) = \sum_{l=0}^{j} \theta_{l,j} (t/\mathtt{T})^l,$$

where

$$\theta_{l,j} = \begin{cases} \frac{2^{2l-j}}{(2l+1)!} \sum_{i=\lfloor \frac{l+1}{2} \rfloor}^{\frac{j}{2}} \frac{(-1)^{\frac{j}{2}+1+i}(1+2i+l)!}{(2i-l)!}, & j \text{ even}, \\ \frac{2^{2l-j+1}}{(2l+1)!(j+1)} \sum_{i=\lfloor \frac{l}{2} \rfloor}^{\frac{j-1}{2}} \frac{(-1)^{\frac{j+1}{2}+l+i}(1+i)(l+2i+2)!}{(2i-l+1)!}, & j \text{ odd}. \end{cases}$$

The set of basis functions $\{\psi_j(t)\}_{j \in \mathbb{N} \cup \{0\}}$ generates a set of orthogonal functions associated with the weight function $\varpi(t) = (2\frac{t}{\mathtt{T}} - 1)^2 \sqrt{\frac{1}{\mathtt{T}}\left(t - \frac{t^2}{\mathtt{T}}\right)}$ on the interval $[0, \mathtt{T}]$.

THEOREM 2.2. [4] *Let* $\mathrm{k}(x,t) \in \mathrm{L}^2_w\Big(\Omega \times [0, \mathtt{T}]\Big)$ *with the weight function* $w(x,t) = \varpi(x)\varpi(t)$ *satisfies the expansion* $\mathrm{k}(x,t) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \mathsf{k}_{i,j} \psi_i(x) \psi_j(t)$. *Suppose* $\left\| \frac{\partial^6 \mathrm{k}(x,t)}{\partial x^3 \partial t^3} \right\|_2 \leq \bar{c}$ *for a positive constant* $\bar{c}$. *Also, if*

$$\mathrm{k}_{n,m}(x,t) = \sum_{i=0}^{n} \sum_{j=0}^{m} \mathsf{k}_{i,j} \psi_i(x) \psi_j(t),$$

be an estimation of $\mathrm{k}(x,t)$, *then we have*

$$|\mathrm{k}(x,t) - \mathrm{k}_{n,m}(x,t)| < \frac{\bar{c}}{2^{n+m}},$$

$$\left| \mathrm{k}_{xx}(x,t) - \frac{\partial^2 \mathrm{k}_{n,m}}{\partial x^2}(x,t) \right| < \kappa \frac{n^3}{2^{n+m-8}},$$

where κ *is a positive constant.*

3. Description of the Collocation Approach

To find a numerical solution of Eq. (1), assume

$$u(x,t) \simeq u_{n,m}(x,t) = \sum_{i=0}^{n} \sum_{j=0}^{m} c_{i,j} \psi_i(x) \psi_j(t) = \Psi(x)^{\mathrm{T}} \mathbf{C} \tilde{\Psi}(t), \tag{4}$$

where $\Psi(x) = [\psi_0(x), \ldots, \psi_n(x)]^{\mathrm{T}}$ and $\tilde{\Psi}(t) = [\psi_0(t), \ldots, \psi_m(t)]^{\mathrm{T}}$. Also $\mathbf{C} := \Big[c_{i,j}\Big]_{(n+1)\times(m+1)}$, $i = 0, 1, \ldots, n$, $j = 0, 1, \ldots, m$, is an $(n+1) \times (m+1)$ unknown coefficients matrix that must be determined. From Eqs. (1) and (4)

$$\begin{aligned} \mathbf{R}(x,t) &\triangleq \Psi(x)^{\mathrm{T}} \mathbf{C} \Psi^{\alpha}(t) + \eta \Psi_{xx}(x)^{\mathrm{T}} \mathbf{C} \tilde{\Psi}(t) \\ &\quad - \mathbf{F}\left(x, t, \Psi(x)^{\mathrm{T}} \mathbf{C} \tilde{\Psi}(t)\right) - \mathbf{G}\left(t, \Psi(x)^{\mathrm{T}} \mathbf{C} \tilde{\Psi}(t)\right) \dot{\mathrm{B}}(t) \simeq 0, \end{aligned} \tag{5}$$

where $\Psi^{\alpha}(t)$ is the Caputo fractional derivative of $\tilde{\Psi}(t)$ and is obtained by

$$\Psi^{\alpha}(t): \quad = \quad [0, \varrho_1^{\alpha}(t), \ldots, \varrho_m^{\alpha}(t)]^T,$$

where $\varrho_j^{\alpha}(t) = \sum_{l=1}^{j} \frac{\theta_{l,j}\Gamma(l+1)}{\mathtt{T}^l\Gamma(l+1-\alpha)} t^{l-\alpha}$ and $\Theta_{xx}(x) = [\frac{d^2}{dx^2}\psi_0(x), \ldots, \frac{d^2}{dx^2}\psi_n(x)]^{\mathrm{T}}$. According to the initial and boundary conditions (2)-(3) and Eq. (4)

$$\Phi(x) \triangleq \Psi(x)^{\mathrm{T}}\mathbf{C}\tilde{\Psi}(0) - u_0(x) \simeq 0, \tag{6}$$

$$\Pi_1(t) \triangleq \Psi(0)^{\mathrm{T}}\mathbf{C}\tilde{\Psi}(t) - \rho(0,t) \simeq 0,\ \Pi_2(t) \triangleq \Psi(\mathtt{L})^{\mathrm{T}}\mathbf{C}\tilde{\Psi}(t) - \rho(\mathtt{L},t) \simeq 0. \tag{7}$$

Let $x_0 = 0$, $x_n = \mathtt{L}$, $\{x_i;\ i = 1, \ldots, n-1\}$ are the roots of $\psi_{n-1}(x)$ and $\{t_j;\ j = 1, \ldots, m\}$ are the roots of $\psi_m(t)$. By evaluating Eqs. (5)-(7) at collocation points (x_i, t_j), a system of $(n+1) \times (m+1)$ NAEqs can be extracted as follows:

$$\begin{cases} \mathbf{R}(x_i, t_j) = 0, & i = 1, \ldots, n-1, \quad j = 1, \ldots, m, \\ \Pi_r(t_j) = 0, & r = 1, 2, \quad j = 1, \ldots, m, \\ \Lambda(x_i) = 0, & i = 0, \ldots, n. \end{cases} \tag{8}$$

Thus, the relation (8), including $(n+1)\times(m+1)$ NAEqs, can provide the unknown coefficients $c_{i,j}$, $i = 0, 1, \ldots, n$ and $j = 0, 1, \ldots, m$.

4. Convergence Analysis

In this section, error estimate of the proposed method have been discussed. Here we consider the norm

$$\|u\|_{\infty} = \mathbb{E}[\sup_{(x,t)\in\Omega\times[0,\mathtt{T}]} |u(x,t)|],$$

where $\mathbb{E}[.]$ is the mathematical expectation.

THEOREM 4.1. *Suppose $u_{n,m}(x,t)$ be the numerical solution of (1)-(3) obtained by the procedure presented in Section 3, $u(x,t)$ is the exact solution of (1)-(3) and $\mathbf{R}_{n,m}(x,t)$ is the residual error. Then, $\|\mathbf{R}_{n,m}(x,t)\|_{\infty}$ tends to zero, when $n \to \infty$ and $m \to \infty$.*

PROOF. Suppose $u_{n,m}(x,t)$, for $(x,t) \in \Omega \times [0,\mathtt{T}]$, is satisfied in the below equation

$$\partial_t^{\alpha} u_{n,m}(x,t) + \eta\Delta u_{n,m}(x,t) = \mathbf{F}(x,t,u_{n,m}(x,t)) + \mathbf{G}(t,u_{n,m}(x,t))\dot{\mathrm{B}}(t) + \mathbf{R}_{n,m}(x,t), \tag{9}$$

where $\mathbf{R}_{n,m}(x,t)$ is the residual function. From Eqs. (1) and (9)

$$\begin{aligned} \|\mathbf{R}_{n,m}(x,t)\|_{\infty} \quad \leq \quad & \Big\|\partial_t^{\alpha}\Big(u(x,t) - u_{n,m}(x,t)\Big)\Big\|_{\infty} + \eta\Big\|\Delta\Big(u(x,t) - u_{n,m}(x,t)\Big)\Big\|_{\infty} \\ + \quad & \Big\|\mathbf{F}(x,t,u(x,t)) - \mathbf{F}(x,t,u_{n,m}(x,t))\Big\|_{\infty} \\ + \quad & \|\dot{\mathrm{B}}(t)\|_{\infty}\Big\|\mathbf{G}(t,u(x,t)) - \mathbf{G}(t,u_{n,m}(x,t))\Big\|_{\infty}. \end{aligned}$$

By using Theorem 2.2 and [5, Theorem 3], we have

$$\Big\|\partial_t^{\alpha}\Big(u(x,t) - u_{n,m}(x,t)\Big)\Big\|_{\infty} < \frac{\kappa_1 \mathtt{T}^{1-\alpha} m}{\Gamma(1-\alpha)2^{n+m-2}}, \tag{10}$$

where κ_1 is a positive constant. Also, from Theorem 2.2

$$\begin{aligned}(11)\qquad \Big\|\Delta\Big(u(x,t)-u_{n,m}(x,t)\Big)\Big\|_\infty &= \sup_{(x,t)\in\Omega\times[0,\mathtt{T}]}\Big|u_{xx}(x,t)-\frac{\partial^2 u_{n,m}}{\partial x^2}(x,t)\Big| \\ &< \frac{\kappa_2 n^3}{2^{n+m-8}},\end{aligned}$$

where κ_2 is a positive constant. The functions $\mathbf{F}$ and $\mathbf{G}$ satisfy the Lipschitz condition with respect to u, hence, by applying Theorem 2.2, we obtain

$$(12)\qquad \Big\|\mathbf{F}(x,t,u(x,t))-\mathbf{F}(x,t,u_{n,m}(x,t))\Big\|_\infty \le \phi_{\mathbf{F}}\|u(x,t)-u_{n,m}(x,t)\|_\infty < \phi_{\mathbf{F}}\frac{\kappa_3}{2^{n+m}},$$

$$(13)\qquad \Big\|\mathbf{G}(t,u(x,t))-\mathbf{G}(t,u_{n,m}(x,t))\Big\|_\infty \le \phi_{\mathbf{G}}\|u(x,t)-u_{n,m}(x,t)\|_\infty < \phi_{\mathbf{G}}\frac{\kappa_4}{2^{n+m}},$$

where $\phi_{\mathbf{F}}$, $\phi_{\mathbf{G}}$, κ_3 and κ_4 are positive real constants. Let $\varphi = \|\dot{B}(\mathrm{t})\|_\infty$, then, from the relations (10)-(13), it can be concluded that

$$\|\mathbf{R}_{n,m}(x,t)\|_\infty < \frac{\kappa_1 \mathtt{T}^{1-\alpha} m}{\Gamma(1-\alpha)2^{n+m-2}} + \eta\frac{\kappa_2 n^3}{2^{n+m-8}} + \phi_{\mathbf{F}}\frac{\kappa_3}{2^{n+m}} + \varphi\phi_{\mathbf{G}}\frac{\kappa_4}{2^{n+m}}.$$

Therefore, we can see that $\|\mathbf{R}_{n,m}(x,t)\|_\infty$ tends to zero, when $n, m \to \infty$. □

5. Numerical Test Example

In this section, we investigate our proposed approach for solving TFSDE. We evaluate the numerical solution $u(x,t)$ along $\mathtt{p}$ discretized Brownian paths. The arithmetic mean over these paths is considered as the approximate solution. Consider the Eqs. (1)-(3) with $\partial_t^\alpha u + 2\Delta u = e^u + u^2\dot{\mathrm{B}} + f(x,t)$, where

$$f(x,t) = \frac{10\Gamma(5)}{\Gamma(5-\alpha)}t^{4-\alpha}\sin(\pi x) - 20\pi^2 t^4\sin(\pi x) - e^{10t^4\sin(\pi x)} - 100t^8\sin^2(\pi x)\dot{\mathrm{B}}(t).$$

With these assumptions, the exact solution is $u(x,t) = 10\ t^4\sin(\pi x)$. Figure 1 shows the exact and numerical solution of $u(x,t)$ along $\mathtt{p} = 50$ discretized Brownian paths and Figure 2 shows the numerical solution of $u(x,\mathtt{T})$ at $\mathtt{T} = 1$ along $\mathtt{p} = 50$ paths, when $\alpha = 0.75$, $n = m = 8$. Figure 3 displays absolute error and contour plot of $u(x,t)$, when $\mathtt{p} = 100$ and $n = m = 9, 12$.

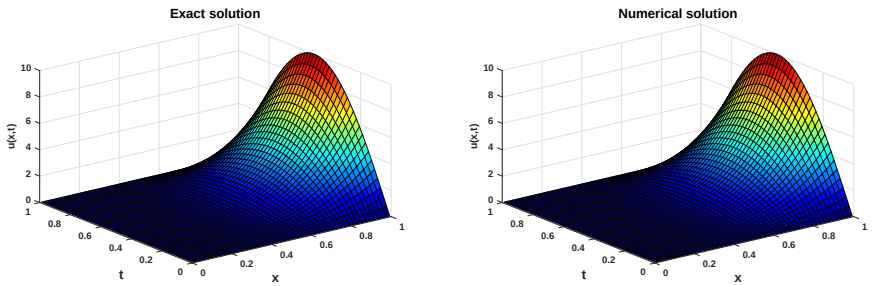


FIGURE 1. The exact and numerical solution of $u(x,t)$ with $\alpha = 0.75$ and $\mathtt{p} = 50$.

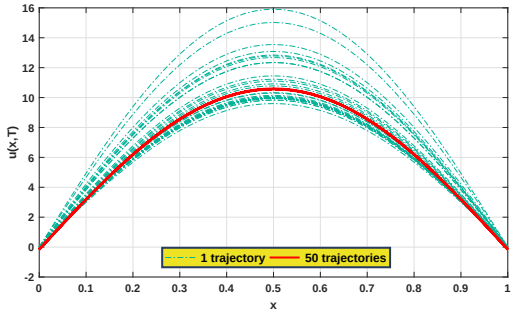


FIGURE 2. The numerical solution of $u(x, \mathtt{T})$ along $\mathtt{p} = 50$ different discretized Brownian paths (Blue) and their arithmetic mean (Red).

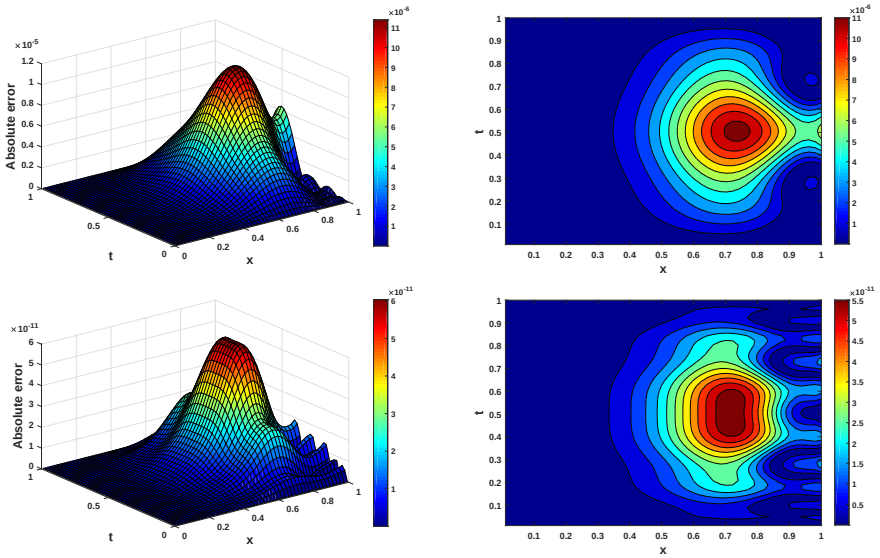


FIGURE 3. The absolute error and contour plot with $n = 9$ (up) and $n = 12$ (down).

References

1. B. Oksendal, *Stochastic Differential Equations*, Springer-Verlag Berlin Heidelberg, Berlin, 1998.
2. I. Podlubny, *Fractional Differential Equations*, Academic Press, San Diego, 1999.
3. W. M. Abd-Elhameed and Y. H. Youssri, *Sixth-kind Chebyshev spectral approach for solving fractional differential equations*, J. Nonlinear Sci. Numer. Simul. **20** (2) (2019) 191–203.
4. A. Babaei, H. Jafari and S. Banihashemi, *Numerical solution of variable order fractional nonlinear quadratic integro-differential equations based on the sixth-kind Chebyshev collocation method*, J. Comput. Appl. Math. **377** (2020) 112908.
5. A. Babaei, H. Jafari and S. Banihashemi, *A collocation approach for solving time-fractional stochastic heat equation driven by an additive noise*, *Symmetry* **12** (6) (2020) 904.

E-mail: babaei@umz.ac.ir
E-mail: jafari@umz.ac.ir
E-mail: s.banihashemi@stu.umz.ac.ir

Numerical Solutions of Time-Fractional Allen-Cahn Equation with Sinc Collocation Method

Ali Barati*
Islam abad Faculty of Engineering, Razi University, Kermanshah, Iran

Abstract. This article deals with the numerical solution of time fractional Allen-Cahn equation with Caputo derivative. The time fractional derivative is discretized by applying finite forward difference method, then the semi-discrete scheme is approximated by using the Sinc collocation method. Numerical experiments demonstrate the accuracy and efficiency of the algorithm.
Keywords: Time fractional derivative, Allen-Cahn equation, Sinc collocation method.
AMS Mathematical Subject Classification [2010]: 26A33, 65N06, 65N35.

1. Introduction

We study time-fractional Allen-Cahn equation as follows, that appears in mathematical modeling of phase separation in alloys of iron [7]:

$$\begin{aligned} &\frac{\partial^\alpha u}{\partial t^\alpha} - \frac{\partial^2 u}{\partial x^2} + u^3 - u = f(x,t), \qquad (x,t) \in \Omega = [a,b] \times [0,T], \\ &u(a,t) = g_1(t), \quad u(b,t) = g_2(t), \quad u(x,0) = u_0(x), \end{aligned} \tag{1}$$

where $f(x,t)$, $g_1(t)$, $g_2(t)$ and $u_0(x)$ are continuous functions. Also, the fractional derivative $\frac{\partial^\alpha u}{\partial t^\alpha}$ is Caputo fractional derivative defined by

$$\frac{\partial^\alpha u}{\partial t^\alpha} = \frac{1}{\Gamma(1-\alpha)} \int_0^t \frac{\partial u(x,s)}{\partial s} \frac{ds}{(t-s)^\alpha}, \qquad \alpha \in (0,1].$$

Allen-Cahn equation is considered as a model problem that is used in various fields such as quantum mechanics, the elasticity, plasma physics, mathematical biology, gas dynamics and others as well. In recent years, many researchers have investigated Allen-Cahn equation [1, 2, 6, 7, 8]. In this paper, we discrete the time fractional derivative (1) by a finite difference scheme, then we estimate solution of semi-discrete scheme by using the Sinc collocation method.

2. Description of Method

Now, we introduce time discretization and Sinc collocation in the following subsections.

*Speaker

2.1. Time Discrete Scheme. We first discrete the time fractional derivative by finite difference approximation. Suppose $t_m = m\Delta t,\ m = 0, 1, \ldots, M$, where $\Delta t = T/M$ is the time step. The time fractional derivative $\frac{\partial^\alpha u}{\partial t^\alpha}$ at t_{m+1} is estimated by

$$\begin{aligned}
\frac{\partial^\alpha u(x,t_{m+1})}{\partial t^\alpha} &= \frac{1}{\Gamma(1-\alpha)}\int_0^{t_{m+1}} \frac{\partial u(x,s)}{\partial s}\frac{ds}{(t_{m+1}-s)^\alpha} \\
&= \frac{1}{\Gamma(1-\alpha)}\sum_{k=0}^{k=m}\int_{t_k}^{t_{k+1}} \frac{\partial u(x,s)}{\partial s}\frac{ds}{(t_{m+1}-s)^\alpha} \\
&= \frac{1}{\Gamma(1-\alpha)}\sum_{k=0}^{k=m}\frac{u(x,t_{k+1})-u(x,t_k)}{\Delta t}\int_{t_k}^{t_{k+1}}\frac{ds}{(t_{m+1}-s)^\alpha} + R^{(1)}_{\Delta t}. \qquad (2)
\end{aligned}$$

Also, the truncation error $R^{(1)}_{\Delta t}$ is bounded as $R^{(1)}_{\Delta t} \le C(\Delta t)^{(2-\alpha)}$, where C is a constant [4].

By calculating the integral in (2) we have:

$$\begin{aligned}
\frac{\partial^\alpha u(x,t_{m+1})}{\partial t^\alpha} &= \frac{(\Delta t)^{-\alpha}}{\Gamma(2-\alpha)}\sum_{k=0}^{k=m}\lambda_{m-k}\Big(u(x,t_{k+1})-u(x,t_k)\Big) + R^{(1)}_{\Delta t} \\
&= \frac{(\Delta t)^{-\alpha}}{\Gamma(2-\alpha)}\sum_{k=0}^{k=m}\lambda_k(u(x,t_{m-k+1})-u(x,t_{m-k})) + R^{(1)}_{\Delta t}, \qquad (3)
\end{aligned}$$

where $\lambda_k = (k+1)^{(1-\alpha)} - k^{(1-\alpha)}$.

Now, by replacing Eq. (3) into Eq. (1), we can arrive the following relation:

$$\begin{aligned}
(4)\quad \frac{(\Delta t)^{-\alpha}}{\Gamma(2-\alpha)}\sum_{k=0}^{k=m}\lambda_k &\Big(u(x,t_{m-k+1})-u(x,t_{m-k})\Big) - \frac{\partial^2 u}{\partial x^2}(x,t_{m+1}) \\
&+ u^3(x,t_{m+1}) - u(x,t_{m+1}) = f(x,t_{m+1}) + R^{(2)}_{\Delta t}.
\end{aligned}$$

Suppose u^m be the numerical estimation to $u(x,t_m)$ and $f^{m+1} = f(x,t_{m+1})$. Ignoring the error term $R^{(2)}_{\Delta t}$ in (4), the semi-discrete scheme can be constructed as follows:

$$\begin{aligned}
&(1-\mu)u^{m+1} - \mu u^{m+1}_{xx} + \mu(u^{m+1})^3 = \tau^m,\ m = 0, 1, \ldots, M-1, \qquad (5)\\
&u^0 = u_0(x),\ u^{m+1}(a) = g_1(t_{m+1}),\ u^{m+1}(b) = g_2(t_{m+1}),
\end{aligned}$$

where $\mu = (\Delta t)^\alpha\Gamma(2-\alpha)$ and $\tau^m = \mu f^{m+1} + u^m - \sum_{k=1}^{k=m}\lambda_k\big(u^{m-k+1} - u^{m-k}\big)$.

2.2. Sinc Function. The Sinc function, which was developed by Stenger [5], is defined on $-\infty < x < \infty$ by

$$Sinc(x) = \begin{cases} \frac{\sin(\pi x)}{\pi x}, & x \neq 0, \\ 1, & x = 0. \end{cases}$$

$$S(j,h)(x) = Sinc(\frac{x-jh}{h}),\ j = 0, \pm1, \pm2, \ldots.$$

Let $f(x)$ is defined on the real line, the Whittaker cardinal expansion of f for $h > 0$ is determined as follows:

$$f(x) \approx \sum_{j=-N}^{N} f_j S(j,h)(x),\ f_j = f(x_j),\quad x_j = jh,\ h = \sqrt{\pi d/\sigma N},\ 0 < \sigma \leq 1,\ d \leq \frac{\pi}{2},$$

where N is suitably chosen and σ depends on the asymptotic behaviour of f. The approximation of $f(x)$ by Sinc function on $[a,b]$ is as follows:

$$f(x) \approx \sum_{j=-N}^{N} f_j S_j(x),\ S_j(x) = S(j,h)(x) \circ \phi(x),\ \phi(x) = \ln(\frac{x-a}{b-x}),$$

The interpolation points x_j are then given by

$$x_j = \frac{a + be^{jh}}{1 + e^{jh}}, \qquad j = 0, \pm 1, \pm 2, \ldots.$$

Also, the n-th derivative of the function f can be represented as

$$f^{(n)}(x) \approx \sum_{j=-N}^{N} f_j \frac{d^n}{dx^n}[S_j(x)].$$

Setting

$$\frac{d^i}{d\phi^i}[S_j(x)] = S_j^{(i)}(x), \qquad 0 \leq i \leq 2, \tag{6}$$

and noting that

$$\frac{d}{dx}[S_j(x)] = S_j^{(1)}(x)\phi^{'}(x),\ \frac{d^2}{dx^2}[S_j(x)] = S_j^{(2)}(x)[\phi^{'}(x)]^2 + S_j^{(1)}(x)\phi^{''}(x), \tag{7}$$

and

$$\delta_{jr}^{(l)} = h^l \frac{d^l}{d\phi^l}[S_j(x)]_{x=x_r}, \quad l = 0,1,2, \quad r = -N, -N+1, \ldots, N. \tag{8}$$

Consider (5) with the given boundary condition. Suppose $\hat{u} = u^{m+1}, \hat{\tau} = \tau^m$, we have:

$$(1-\mu)\hat{u} - \mu\hat{u}_{xx} + \mu(\hat{u})^3 = \hat{\tau}. \tag{9}$$

Now, we use the Sinc collocation method for solving (9). For this end, we take the approximate solution of (9) as:

$$\hat{u}(x) \approx \sum_{j=-N}^{N} \hat{c}_j S_j(x), \tag{10}$$

The unknown coefficients $\hat{c}_j$ in relation (10) are obtained by collocation method with Sinc functions. Also, this approximation is then used to approximate the second derivative in the points of x_r, $r = -N, -N+1, \ldots, N$. Thus, by using relations (6)-(8) we get the discrete nonlinear equations as:

$$(1-\mu)\hat{c}_r - \mu\left\{\sum_{j=-N}^{N} \hat{c}_j \left(\frac{\delta_{jr}^{(2)}}{h^2}[\phi^{'}(x_r)]^2 + \frac{\delta_{jr}^{(1)}}{h}\phi^{''}(x_r)\right)\right\} + \mu(c_r)^3 = \hat{\tau}(x_r), \tag{11}$$
$$r = -N, -N+1, \ldots, N.$$

TABLE 1. Comparison of L_∞ errors of Example 3.1 for different values α with $\Delta t = 0.001, t = 1$ and $n = 100$.

α	[6]	[3]	Our results
0.2	-	4.49e-06	1.13e-07
0.7	9.24e-04	4.98e-06	1.29e-06
0.9	7.42e-04	1.42e-05	7.54e-06

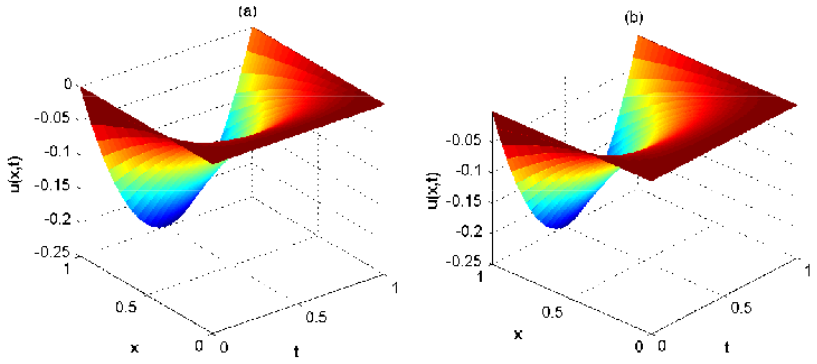


FIGURE 1. The graphs of approximate solutions of Example 3.1 for $n = 36$ with (a) $\alpha = 0.7$ (b) $\alpha = 0.9$ on $0 \le x, t \le 1$.

The nonlinear system (11) includes $2N+1$ equations and $2N+1$ unknowns which can be solved by means of Newton's method. Finally we can attain an approximate solution $\hat{u}(x)$ of (9) from (10).

3. Numerical Experiments

Now, we provide one test example to illustrate the efficiency of the method on (1). Select $\sigma = 1$ and $d = \frac{\pi}{2}$ which yield $h = \frac{\pi}{\sqrt{2N}}$, the maximum absolute errors are calculated on uniform points

$$U = \{z_0, z_1, \ldots, z_n\}, \quad z_p = (b-a)\frac{p}{n}, \quad p = 0, 1, \ldots, n.$$

EXAMPLE 3.1. Consider the time fractional Allen-Cahn equation

$$\frac{\partial^\alpha u}{\partial t^\alpha} + \frac{\partial^2 u}{\partial x^2} + u^3 - u = f(x,t), \qquad (x,t) \in \Omega = [0,1] \times [0,1],$$

where

$$f(x,t) = (\alpha+1)(x-1)xt\Gamma(1+\alpha) - (x^2-x)^3 t^{3+3\alpha} - (x^2-x+2)^3 t^{1+\alpha}.$$

The initial and boundary conditions can be achieved from the exact solution $(x^2 - x)t^{1+\alpha}$. The maximum absolute errors for this example are given in Table 1 and compare with results in [3, 6] for various values of α, the computational results are presented for $\Delta t = 0.001, t = 1$ and $n = 100$. These results state the performance and accuracy of the method. Also, for this example the approximate solutions for different values of α are drawn in Figure 1.

References

1. T. Hou, T. Tang and J. Yang, *Numerical analysis of fully discretized Crank-Nicolson scheme for fractional in space Allen-Cahn equations*, J. Sci. Comput. **72** (3) (2017) 1214–1231.
2. K. Hosseini, A. Bekir and R. Ansari, *New exact solutions of the conformable time-fractional Cahn-Allen and Cahn-Hilliard equations using the modified Kudryashov method*, Optik, Int. J. Light Electron Opt. **132** (2017) 203–209.
3. N. Khalid, M. Abbas, M. K. Iqbal and D. Baleanu, *A numerical investigation of Caputo time fractional Allen-Cahn equation using redened cubic B-spline functions*, Advances in Difference Equations **158** (2020) 1–22.
4. Y. M. Lin and C. J. Xu, *Finite difference/spectral approximations for the time-fractional diffusion equation*, J. Comput. Phys. **225** (2007) 1533–1552 .
5. F. Stenger, *Numerical Methods Based on Sinc and Analytic Functions*, Springer, New York, 1993.
6. M. G. Sakar, O. Saldir and F. Erdogan, *An iterative approximation for time-fractional Cahn-Allen equation with reproducing kernel method*, Comput. Appl. Math. **37** (5) (2018) 5951–5964.
7. H. Tariq and G. Akram, *New approach for exact solutions of time fractional Cahn-All enequation and time fractional Phi-4 equation* Phys. A, Stat. Mech. Appl. **473** (2017) 352–362.
8. B. Yin, Y. Liu, H. Li and S. He, *Fast algorithm based on TT-MFE system for space fractional Allen-Cahn equations with smooth and non-smooth solutions,* J. Comput. Phys. **379** (2019) 351–372.

E-mail: alibarati@razi.ac.ir

A Hybrid Laguerre Method for the European Exchange Option Pricing

Reza Doostaki*
Faculty of Mathematics and Computer, Shahid Bahonar University of Kerman, Kerman, Iran
Mahani Mathematical Research Center, Shahid Bahonar University of Kerman, Kerman, Iran
Mohammad Mehdi Hosseini
Faculty of Mathematics and Computer, Shahid Bahonar University of Kerman, Kerman, Iran
Mahani Mathematical Research Center, Shahid Bahonar University of Kerman, Kerman, Iran
and Abbas Salemi
Faculty of Mathematics and Computer, Shahid Bahonar University of Kerman, Kerman, Iran
Mahani Mathematical Research Center, Shahid Bahonar University of Kerman, Kerman, Iran

Abstract. In financial markets, a lot of traded options are multi-asset options. A European exchange option gives the holder the right to exchange two assets at expiration time. This paper is considered the numerical solution of two dimensional Black-Scholes partial differential equation (PDE) for evaluating the European exchange options. We use a hybrid method based on the finite difference method and Laguerre approximation method. It is shown that the two dimensional Black-Scholes PDE is reduced to a nonsingular upper triangular linear system. The numerical results demonstrate efficiency and capability of the proposed method.
Keywords: European exchange option, Two dimensional Black-Scholes PDE, Laguerre polynomials, Finite difference scheme.
AMS Mathematical Subject Classification [2010]: 65M50, 91G20.

1. Introduction

Multi-asset options are a group of options whose pay-off depends on more than one underlying asset. The multi-asset options have European and American types, as far as the time of exercising is concerned. Under the Blak-Sholes assumptions [1], each of the underlying asset prices S_1 and S_2 follow a geometric Brownian motion as

$$dS_1 = \mu_1 S_1 dt + \sigma_1 S_1 dz,$$
$$dS_2 = \mu_2 S_2 dt + \sigma_2 S_2 dw,$$

*Speaker

where μ_1 and μ_2 are the expected instantaneous rates of return of the two assets, σ_1 and σ_2 are the corresponding instantaneous volatilities, and dz and dw are two correlated Winer processes. Under the Black-Scholes environments [2, 3], the option price $u(S_1, S_2, t)$ satisfy

$$\begin{aligned}(1)\qquad \frac{\partial u}{\partial t} + \frac{1}{2}\sigma_1^2 S_1^2 \frac{\partial^2 u}{\partial S_1^2} \quad &+ \quad \sigma_1\sigma_2\rho S_1 S_2 \frac{\partial^2 u}{\partial S_1 \partial S_2} \\ &+ \quad \frac{1}{2}\sigma_2^2 S_2^2 \frac{\partial^2 u}{\partial S_2^2} + rS_1 \frac{\partial u}{\partial S_1} + rS_2 \frac{\partial u}{\partial S_2} - ru = 0,\end{aligned}$$

where constants ρ and r are the asset correlation and risk-free interest rate, respectively.

A typical example of two asset options is the exchange option that was introduced by Margrabe [4]. This option gives the holder the right to exchange asset S_2 for S_1 at expiration time $t = T$. The pay-off form an exchange option is $u(S_1, S_2, T) = \max\{S_1 - S_2, 0\}$.

This paper is considered the numerical solution of two dimensional Black-Scholes Eq. (1) for evaluating the European exchange option

2. The European Exchange Option Pricing

In orther to obtain the numerical solution of two dimensional Black-Scholes PDE (1) for evaluating the European exchange option, we use a hybrid method based on the theta finite difference method and function approximation scheme using the Laguerre polynomials. In practice, we truncate the infinite domain $[0, \infty) \times [0, \infty)$ to a finite domain $[0, S_{1,max}] \times [0, S_{2,max}]$. For evaluating the exchange options, the boundary conditions of the Black-Scholes PDE (1) are as the following:

$$\begin{aligned}&u(0, S_2, t) = 0,\\ &u(S_1, 0, t) = S_1,\\ &u(S_{1,max}, S_2, t) = S_{1,max} - S_2,\\ &u(S_1, S_{2,max}, t) = 0.\end{aligned}$$

Without loss of generality, for a given sufficiently large number E, we assume that $S_{1,max} = S_{2,max} = E$. For two-dimensional Black-Scholes Eq. (1), we can write

$$(2)\qquad \frac{\partial u(x, y, \tau)}{\partial \tau} = Du(x, y, \tau),$$

where $\tau = T - t$ and D is an operator as

$$(3)\qquad D = \frac{1}{2}\frac{\partial^2}{\partial x^2} + \sigma_1\sigma_2\rho\frac{\partial^2}{\partial x \partial y} + \frac{1}{2}\sigma_2^2\frac{\partial^2}{\partial y^2} + (\frac{1}{2}\sigma_1^2 - r)\frac{\partial}{\partial x} + (\frac{1}{2}\sigma_2^2 - r)\frac{\partial}{\partial y} - r.$$

Also the European exchange pay-off function can be considered as the initial condition of (2) as

$$u(x, y, 0) = E\ \max\{e^{-x} - e^{-y}, 0\},$$

and for the boundary conditions, we have

$$(4)\qquad \begin{aligned}&u(x, y, \tau) = 0, \quad as \quad x \to \infty,\\ &u(x, y, \tau) = Ee^{-x}, \quad as \quad y \to \infty,\\ &u(0, y, \tau) = E - Ee^{-y},\\ &u(x, 0, \tau) = 0.\end{aligned}$$

We partition the interval $[0,T]$ into J subintervals of equal length $\Delta\tau = T/J$. We set $\tau = j\Delta\tau$, and define $u^j(x,y) = u(x,y,\tau)$, for $j = 0,1,\ldots,J$. For a given arbitrary natural number M, we approximate $u^j(x,y)$ by the truncated Laguerre series as

$$u^j = L^T(x)C^jL(y),$$

where C^j is unknown coefficients matrix and $L(.)$ is an M-vector contain the Laguerre basis functions. Using the theta finite difference method and differential operator matrix of Laguerre polynomials, the Eq. (2) yield

$$A_1C^{j+1} + C^{j+1}A_2 + A_3C^{j+1}P = A_4C^j + C^jA_5 + A_6C^jP, \tag{5}$$

where

$$\begin{aligned}
A_1 &= (1+\theta\Delta\tau r)I - \frac{1}{2}\theta\Delta\tau\sigma_1^2(P^2)^T + \theta\Delta\tau(r-\frac{1}{2}\sigma_1^2)P^T,\\
A_2 &= -\frac{1}{2}\theta\Delta\tau\sigma_2^2P^2 + \theta\Delta\tau(r-\frac{1}{2}\sigma_2^2)P,\\
A_3 &= -\theta\Delta\tau\sigma_1\sigma_2\rho P^T,\\
A_4 &= (1-(1-\theta)\Delta\tau r)I + \frac{1}{2}(1-\theta)\Delta\tau\sigma_1^2(P^2)^T - (1-\theta)\Delta\tau(r-\frac{1}{2}\sigma_1^2)P^T,\\
A_5 &= \frac{1}{2}(1-\theta)\Delta\tau\sigma_2^2P^2 - (1-\theta)\Delta\tau(r-\frac{1}{2}\sigma_2^2)P,\\
A_6 &= (1-\theta)\Delta\tau\sigma_1\sigma_2\rho P^T,
\end{aligned}$$

and $I = I_{(M+1)\times(M+1)}$ is the identity matrix. The matrix Eq. (5) can be write as [5]

$$\mathbf{A}\mathbf{v}^{j+1} = \mathbf{B}\mathbf{v}^j, \quad j = 0,1,\ldots,J-1, \tag{6}$$

$$\mathbf{A} = I \bigotimes A_1 + A_2^T \bigotimes I + P^T \bigotimes A_3, \tag{7}$$

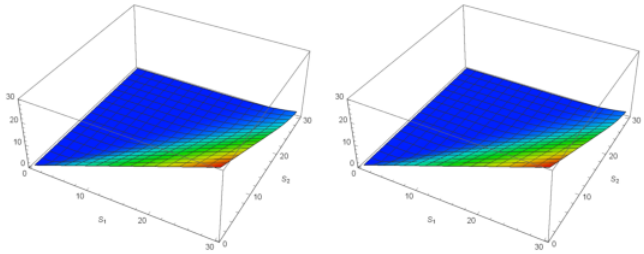
$$\mathbf{B} = I \bigotimes A_4 + A_5^T \bigotimes I + P^T \bigotimes A_6, \tag{8}$$

where $\bigotimes$ is the Kronecker product, $\mathbf{v}^j = vec(C^j)$, and $vec(C^j)$ is defined a column vector obtained by stacking the column vectors of C^j on top of one another.

LEMMA 2.1. *Let the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ is defined as (7) and (8). Then, $\boldsymbol{A}$ and $\boldsymbol{B}$ are upper triangular matrices with diagonal entries $1+\theta r\Delta\tau$ and $1-(1-\theta)r\Delta\tau$, respectively.*

PROOF. Consider the matrices A_i, $i = 1,\ldots,6$ as (5). The matrices A_1 and A_4 are upper triangular matrices with diagonal entries $1+\theta r\Delta\tau$ and $1-(1-\theta)r\Delta\tau$, respectively. Also, It can be easily seen that $A_2^T \bigotimes I$, $P^T \bigotimes A_3$, $A_5^T \bigotimes I$ and $P^T \bigotimes A_6$ are upper triangular matrices with zero diagonal entries. Hence, the matrices $\mathbf{A}$ and $\mathbf{B}$ are upper triangular matrices with diagonal entries $1+\theta r\Delta\tau$ and $1-(1-\theta)r\Delta\tau$, respectively. □

COROLLARY 2.2. *The upper triangular linear system $\boldsymbol{A}\boldsymbol{v}^{j+1} = \boldsymbol{B}\boldsymbol{v}^j$ in (6) has a unique solution.*



(A) (B)

FIGURE 1. The exact European exchange option values (a) and approximate values by the proposed method (b) for example 2.3.

PROOF. By Lemma 2.1, the coefficient matrix $\mathbf{A}$ is an upper triangular matrix with non-zeroe entries $1 + \theta r \Delta\tau$. Hence, the matrix $\mathbf{A}$ is a nonsingular matrix, and the result holds. □

Also, for the boundary conditions (4), we have

$$\mathbf{E}\mathbf{v}^{j+1} = \mathbf{G}\mathbf{v}^{j}, \quad j = 0, 1, \ldots, J-1, \tag{9}$$

where

$$\mathbf{E} = \begin{pmatrix} I \bigotimes L(x_{max})^T \\ L(y_{max}) \bigotimes I \\ I \bigotimes L(0)^T \\ L(0) \bigotimes I \end{pmatrix}_{(4M+4)\times(M+1)^2}, \quad \mathbf{G} = \begin{pmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \\ \mathbf{g}_3 \\ \mathbf{g}_4 \end{pmatrix},$$

$\mathbf{g}_i = (g_i(0), g_i(1), \ldots, g_i(M))^T$ and

$$\begin{aligned} g_1(m) &= 0, \\ g_2(m) &= E\int_0^\infty L_m(x)e^{-2x}dx, \\ g_3(m) &= E\delta_{m,0} - E\int_0^\infty L_m(y)e^{-2y}dy, \\ g_4(m) &= 0, \end{aligned}$$

for $m = 0, 1, \ldots M$, where $\delta_{m,n}$ is the Kronecker delta function. Equations (6) and (9) yield

$$\mathcal{A}\mathbf{v}^{j+1} = \mathcal{B}\mathbf{v}^{j}, \quad j = 0, 1, \ldots, J-1,$$

where

$$\mathcal{A} = \begin{pmatrix} \mathbf{A} \\ \mathbf{E} \end{pmatrix}, \quad \mathcal{B} = \begin{pmatrix} \mathbf{B} \\ \mathbf{G} \end{pmatrix}.$$

Hence, the European exchange option values is obtained from the above overdetermined linear system.

EXAMPLE 2.3. Consider the two dimensional Black-Scholes PDE (1) to evaluate the European exchange option. Let $\sigma_1 = 0.2$, $\sigma_2 = 0.2$, $\rho = 0.1$, $r = 0.1$ and $T = 1/2$. We use the proposed method for $\theta = 1/2$, $M = 100$ and $J = 100$. Figure 1 shows the exact [4] and approximate European exchange options. Also, the absolute error is shown in Figure 2. Moreover, we fix one of the underlying

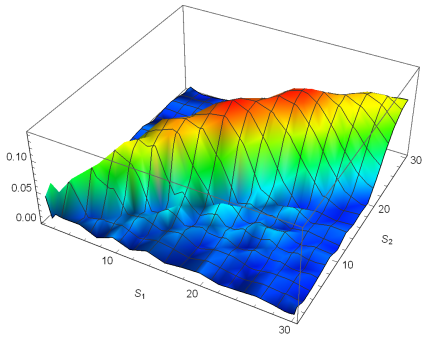
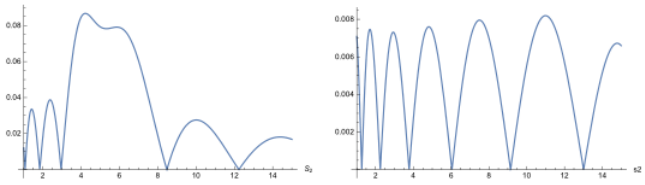
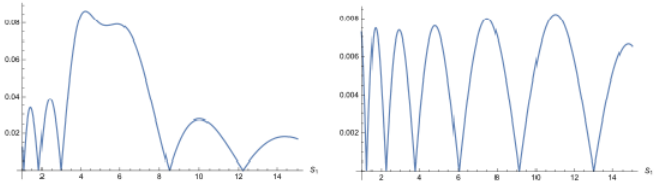


FIGURE 2. The absolute error for comparison between the exact and approximate European exchange options using the proposed method with $M = 100$ and $J = 100$.



(A) (B)

FIGURE 3. The absolute error for the European exchange option using the proposed method for $S_1 = 5$, $S_2 \in [0, 15]$ (A) and $S_1 = 26$, $S_2 \in [0, 15]$ (B) for $\theta = 1/2$, $M = 100$ and $J = 100$.



(A) (B)

FIGURE 4. The absolute error for the European exchange option using the proposed method for $S_1 = 5$, $S_2 \in [0, 15]$ (A) and $S_1 = 26$, $S_2 \in [0, 15]$ (B) for $\theta = 1/2$, $M = 100$ and $J = 100$.

asset prices equal to $S = 5, 26$. Figures 3 and 4 show the absolute errors by the proposed method for another asset price in interval $[0, 15]$.

References

1. F. Black and M. Scholes, *The pricing of options and corporate liabilities*, J. Polit. Econ. **81** (1973) 637–654.
2. E. G. Haug, *The Complete Guide to Option Pricing Formulas*, McGraw-Hill, New York, 2006.
3. P. G. Zhang, *Exotic options: A Guide to Second Generation Options*, World Scientific, Singapore, 1998.
4. W. Margrabe, *The value of an option to exchange one asset for another*, J. Finance. **33** (1978) 177–186.
5. P. Lancaster and M. Tismensky, *The Theory of Matrices: with Applications*, 2nd ed., Academic Press, San Diego, 1985.

E-mail: rdoostaki@math.uk.ac.ir; rdoostaki@yahoo.com
E-mail: mhosseini@uk.ac.ir
E-mail: salemi@uk.ac.ir

Numerical Solution of Two-Dimensional sinh-Gordon Equation via Integrated RBF-FD

Ali Ebrahimijahan*
Department of Applied Mathematics, Faculty of Mathematics and Computer Sciences, Amirkabir University of Technology, No. 424, Hafez Ave., 15914, Tehran, Iran
and Mehdi Dehghan
Department of Applied Mathematics, Faculty of Mathematics and Computer Sciences, Amirkabir University of Technology, No. 424, Hafez Ave., 15914, Tehran, Iran

Abstract. We present a method based on integrated RBF (IRBF)-finite difference (FD) for numerical solution of two-dimensional sinh-Gordon equation. An example is solved by applying IRBF-FD method to campare it with radial basis functions (RBFs) collocation based on Kansa's approach, RBF-pseudospectral (RBF-PS) technique and moving least squares (MLS) method. The aim of this paper is to show that IRBF-FD method is more accurate than other meshless methods.
Keywords: Integrated radial basis function, The sinh-Gordon equation, Integrated RBF-FD.
AMS Mathematical Subject Classification [2010]: 65L60, 34B15.

1. Introduction

Nonlinear phenomena, that appear in many areas of scientific fields such as solid state physics, plasma physics, fluid dynamics, mathematical biology and chemical kinetics, can be modeled by partial differential equations. A broad class of analytical and numerical solution methods were used to handle these problems. The search of exact solution for the nonlinear partial differential equations is very difficult. Therefore, numerical methods are useful for solving nonlinear partial differential equations.

In this paper, we investigate nonlinear sinh-Gordon equation in the following form [5]:

$$\begin{aligned}(1)\qquad \frac{\partial^2 u(x,y,t)}{\partial t^2} &- \frac{\partial^2 u(x,y,t)}{\partial x^2} - \frac{\partial^2 u(x,y,t)}{\partial y^2} \\ &+ \sinh(u(x,y,t)) = f(x,y,t),\ (x,y)\in\Omega,\ t\in(0,T),\end{aligned}$$

with initial and boundary conditions

$$\begin{aligned} &u(x,y,0) = g_1(x,y), \qquad u_t(x,y,0) = g_2(x,y), \quad (x,y)\in\Omega,\\ &u(x,y,t) = h(t), \qquad (x,y)\ on\ \partial\Omega,\ t\in(0,T).\end{aligned}$$

*Speaker

1.1. Overview of Integrated RBF. For solving partial differential equations (PDEs), there are many numerical methods such as finite difference method (FDM), finite volume method (FVM) and finite element method (FEM) usually utilize the low-order polynomial to approximate the derivatives in PDEs. Therefore, these methods are low-order methods. In order to have acceptable accuracy, for the low-order methods, they must apply a large number of grid nodes. Lately, meshless methods received a lot of attention. The author of [3] depicts meshfree methods are better than FDM for solving PDE problems especially in more spatial dimensions. One of the local meshless collocation methods is RBFs finite difference (RBFs-FD) method. About the IRBF technique, Mai Duy has done some research. In [2] May Duy and Trang Cong presented an efficient indirect RBFN-based method for solving some PDEs. The study of the performance of the RBF-FD method when the unknown function is approximated by IRBF is interesting. In this paper, to solve the governing equation the RBF-FD method with IRBF, IRBF-FD for short, is expressed and validated by solving an example.

2. Integrated RBF Based on Finite Difference (IRBF-FD) Method

Let N be the number of collocation points $\{\mathrm{x}_i\}_{i=1}^{N}$ on the region Ω and n be the number of closest neighbor nodes to form the associated local-support area Ω_i for every point $\mathrm{x}_i \in \Omega$, and $\Omega_i \bigcap \Omega_j = \emptyset$ when $i \neq j$. Similar to RBF-FD, we seek in IRBF-FD an approximation to linear operator at any node that involves a linear combination of the values of function over the stencil local domain. For example, operator $\mathcal{L}(u(\mathrm{x}))$ evaluated at x_j is approximated by a linear weighted combination of the function values of u at the points of X_j,

$$\mathcal{L}u(\mathrm{x}_j) \approx \sum_{k=1}^{n} w_k^{(j)} u(\mathrm{x}_k^{(j)}), \tag{2}$$

where $\mathrm{X}_j = \{\mathrm{x}_1^{(j)}, \mathrm{x}_2^{(j)}, \ldots, \mathrm{x}_n^{(j)}\} \subset \mathrm{X}$ is a subset containing x_j and $w_k^{(j)}$ are differentiation weights for the point x_j, which are unknown.

To calculate $w_k^{(j)}$, the local integrated radial basis function interpolation is used, assuming that, for every points $\mathrm{x}_j \in \Omega$, the function $u(\mathrm{x}_j)$ can be approximated by the IRBFs on the corresponding local domain Ω_i, (assuming that the highest-order derivative in the PDE is two).

$$\frac{\partial^2 u(\mathrm{x}_j)}{\partial x^2} \approx \sum_{k=1}^{n} \lambda_k^{(j)} \psi^{[x]}(\|\mathrm{x}_j - \mathrm{x}_k^{(j)}\|_2), \tag{3}$$

$$\begin{aligned} \frac{\partial u(\mathrm{x}_j)}{\partial x} &\approx \sum_{k=1}^{n} \lambda_k^{(j)} \int \psi^{[x]}(\|\mathrm{x}_j - \mathrm{x}_k^{(j)}\|_2) dx + c_1^{(j)}(y) \\ &= \sum_{k=1}^{n} \lambda_k^{(j)} \varphi^{[x]}(\|\mathrm{x}_j - \mathrm{x}_k^{(j)}\|_2) + c_1^{(j)}(y), \end{aligned} \tag{4}$$

$$u(\mathrm{x}_j) \approx \sum_{k=1}^{n} \lambda_k^{(j)} \int \int \psi^{[x]}(\|\mathrm{x}_j - \mathrm{x}_k^{(j)}\|_2) dx dx + x c_1^{(j)}(y) + c_2^{(j)}(y)$$

$$= \sum_{k=1}^{n} \lambda_k^{(j)} \phi^{[x]}(\|\mathrm{x}_j - \mathrm{x}_k^{(j)}\|_2) + x c_1^{(j)}(y) + c_2^{(j)}(y).$$

$$\frac{\partial^2 u(\mathrm{x}_j)}{\partial y^2} \approx \sum_{k=1}^{n} \lambda_k^{(j)} \psi^{[y]}(\|\mathrm{x}_j - \mathrm{x}_k^{(j)}\|_2), \tag{5}$$

$$\begin{aligned} \frac{\partial u(\mathrm{x}_j)}{\partial y} &\approx \sum_{k=1}^{n} \lambda_k^{(j)} \int \psi^{[y]}(\|\mathrm{x}_j - \mathrm{x}_k^{(j)}\|_2) dy + c_1^{(j)}(x) \\ &= \sum_{k=1}^{n} \lambda_k^{(j)} \varphi^{[y]}(\|\mathrm{x}_j - \mathrm{x}_k^{(j)}\|_2) + c_1^{(j)}(x), \end{aligned} \tag{6}$$

$$\begin{aligned} u(\mathrm{x}_j) &\approx \sum_{k=1}^{n} \lambda_k^{(j)} \int\int \psi^{[y]}(\|\mathrm{x}_j - \mathrm{x}_k^{(j)}\|_2) dy dy + y c_1^{(j)}(x) + c_2^{(j)}(x), \\ &= \sum_{k=1}^{n} \lambda_k^{(j)} \phi^{[y]}(\|\mathrm{x}_j - \mathrm{x}_k^{(j)}\|_2) + y c_1^{(j)}(x) + c_2^{(j)}(x), \end{aligned}$$

where $\psi(\|\mathrm{x}_j - \mathrm{x}_k^{(j)}\|_2)$ is Matern function. Acting operator $\mathcal{L}$ in Eq. (3) gives

$$\mathcal{L}u(\mathrm{x}_j) \approx \sum_{k=1}^{n} \lambda_k^{(j)} \mathcal{L}\Big[\phi(\|\mathrm{x}_j - \mathrm{x}_k^{(j)}\|_2) + x c_1^{(j)} + c_2^{(j)}\Big].$$

Combining (2)-(5), we achieved the following linear system of n algebraic equations,

$$\begin{aligned} \sum_{k=1}^{n} &\Big[\phi(\|\mathrm{x}_k^{(j)} - \mathrm{x}_{k'}^{(j)}\|_2) + x_k^{(j)} c_1^{(j)} + c_2^{(j)}\Big] w_k^{(j)} \\ &= \mathcal{L}\Big[\phi(\|\mathrm{x}_j - \mathrm{x}_k^{(j)}\|_2) + x c_1^{(j)} + c_2^{(j)}\Big],\ k' = 1, 2, \ldots, n. \end{aligned} \tag{7}$$

Solving (7) gives $w_k^{(j)}$, and substituting it into (2) leads to the approximation of $\mathcal{L}u(\mathrm{x})$ at the points of X_j. In this paper, we employ an algorithm to achieve an optimal shape parameter that has been introduced via Sarra [4]. In Algorithm (8), Φ is the interpolation matrix, $\sigma_{\max}$ and $\sigma_{\min}$ are the largest and lowest singular values of SVD decomposition, respectively, $c_{Increment} = \frac{1}{m}$ in which m is the number of points in the considered domain and $\mathcal{K}_{\min} = 110$ and $\mathcal{K}_{\min} = 1e{+}6$.

```
𝒦 = 0;
while 𝒦 < 𝒦_min, 𝒦 > 𝒦_max do
   Produce interpolation matrix Φ;
   [U, S, V] = svd(Φ);
   if 𝒦 < 𝒦_min then
      ϵ = ϵ − c_Increment;
   else
      ϵ = ϵ + c_Increment;
   end
end
```

(8)

3. Time Discrete Scheme

For discretization of time variable, we need some preliminary. We define $t_n = k\tau, k = 0, 1, \ldots, N$, where $\tau = \frac{T}{N}$ is the step size of time variable. We investigate Eq. (1) in points (x, y, t_n) then we have the following discrete form.

$$\frac{u^{n+1} - 2u^n + u^{n-1}}{\tau^2} - \frac{\partial u^{n+1}(x,y)}{\partial x^2} - \frac{\partial u^{n+1}(x,y)}{\partial y^2} + \sinh(u^n(x,y)) = f^{n+1}(x,y). \tag{9}$$

Simplifying Eq. (9) gives

$$u^{n+1} - \tau^2 u^{n+1}_{xx} - \tau^2 u^{n+1}_{yy} = 2u^n - u^{n-1} - \tau^2 \sinh(u^n) + \tau^2 f^{n+1}, \quad n \geq 1, \tag{10}$$

$$u^1 - \tau^2 u^1_{xx} - \tau^2 u^1_{yy} = 2u^0 - u^{-1} - \tau^2 \sinh(u^0) + \tau^2 f^1, \tag{11}$$

where u^{-1} is obtained by

$$\frac{u^1 - u^{-1}}{2\tau} = \frac{\partial u(x,y,t)}{\partial t}|_{t=0} = g_2(x,y), \qquad u^{-1} = u^1 - 2\tau g_2(x,y).$$

4. Implementing IRBF-FD Method for the Main Problem

According to IRBF-FD that explained in Section 2, the second- and first-order derivatives of function $u(x, y)$ at every node (i, j) of rectangular grid bring

$$\begin{aligned} \left.\frac{\partial^2 u}{\partial x^2}\right|_{(i,j)} &\approx \sum_{k\in\{i-1,i,i+1\}} w^{xx}_{(k,j)} u_{(k,j)}, \\ \left.\frac{\partial^2 u}{\partial y^2}\right|_{(i,j)} &\approx \sum_{k\in\{j-1,j,j+1\}} w^{yy}_{(k,j)} u_{(k,j)}. \end{aligned} \tag{12}$$

Assembling relations (12) at all grid point the approximation derivatives of function u arrives at

$$u_{xx} \approx U_{xx} = \mathbf{w}^{xx} U, \; u_{yy} \approx U_{yy} = \mathbf{w}^{yy} U. \tag{13}$$

Substituting Eq. (13) into Eqs. (10) and (11) will result the following algebraic system.

$$\begin{aligned} &(I - \tau^2(\mathbf{w}^{xx} + \mathbf{w}^{yy}))U^{n+1} = 2U^n - U^{n-1} - \tau^2 \sinh(U^n) + \tau^2 f^{n+1}, \quad n \geq 1, \\ &(2I - \tau^2(\mathbf{w}^{xx} + \mathbf{w}^{yy}))U^1 = 2U^0 + 2\tau g_2 - \tau^2 \sinh(U^0) + \tau^2 f^1, \end{aligned}$$

where I is identity matrix.

5. Numerical Result

Example 5.1. We consider Eq. (1) with the following exact solution form:

$$u(x, y, t) = \sin(t)(\operatorname{sech}^2(-x - y + r) + \operatorname{sech}^2(x + y + r)),$$

where the initial and boundary conditions can be obtained from the above solution. This example is solved by IRBF-FD and we compare it with results of [1]. Table 1 shows obtained maximum errors by the proposed and other meshless methods with $\tau = 0.01$ for various values of h. As can be seen via Table 1, the obtained results by IRBF-FD method are more accurate than the obtained results for methods of

[1]. Figure 1 depicts surface numerical solution false-colored by the absolute error graph with $h = 1/5, \tau = 0.01$ and $r = 2$ on $[-5, 5]^2$.

TABLE 1. Comparison between maximum error of present method and the methods of article [1] at final time $T = 1$ for Example 5.1.

		Article [1]		
h	$IRBF - FD$	$RBF - Kansa$	$RBF - PS$	MLS
$\frac{1}{5}$	1.1225×10^{-3}	2.5885×10^{-3}	2.5885×10^{-3}	2.7839×10^{-3}
$\frac{1}{10}$	1.6938×10^{-4}	6.6647×10^{-4}	6.6647×10^{-4}	2.0984×10^{-4}
$\frac{1}{15}$	5.3670×10^{-5}	2.3284×10^{-4}	2.3284×10^{-4}	5.3085×10^{-5}
$\frac{1}{20}$	4.1191×10^{-5}	8.5026×10^{-4}	8.7058×10^{-5}	4.1188×10^{-5}
$\frac{1}{25}$	3.3312×10^{-5}	4.3493×10^{-5}	4.3093×10^{-5}	3.5431×10^{-5}

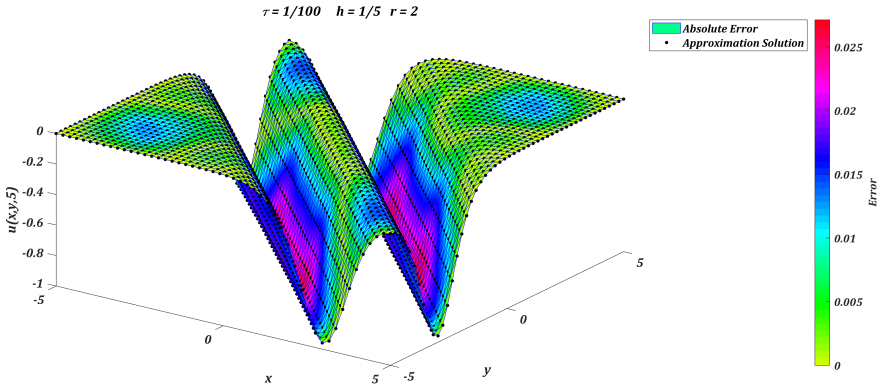


FIGURE 1. Graph of surface approximation solution false-colored at final time $T = 5$ for Example 5.1.

References

1. M. Dehghan, M. Abbaszadeh and A. Mohebbi, *The numerical solution of the two-dimensional sinh-Gordon equation via three meshless methods*, Engin. Anal. Bound. Elem. **51** (2015) 220–235.
2. N. Mai-Duy and T. Tran-Cong, *An efficient indirect RBFN-based method for numerical solution of PDEs*, Numer. Meth. PDE. **21** (4) (2005) 770–790.
3. E. T. Kansa, *Multiquadrics-A scattered data approximation scheme with applications to computational fluid dynamics: Surface approximations and partial derivative estimates*, Comput. Math. Appl. **19** (6-8) (1990) 127–145.
4. S. A. Sarra, *A local radial basis function method for advection-diffusion-reaction equations on complexly shaped domains*, Appl. Math. Comput. **218** (19) (2012) 9853–9865.
5. A. M. Wazwaz, *One and two soliton solutions for the* sinh*-Gordon equation in* $(1+1), (2+1)$ *and* $(3+1)$ *dimensions*, Appl. Math. Lett. **25** (2012) 2354–2358.

E-mail: ebrahimijahan_ali@aut.ac.ir
E-mail: mdehghan@aut.ac.ir

Robust CAS Wavelet Approach for Optimal Control of Nonlinear Volterra-Fredholm Integral Equation

Asiyeh Ebrahimzadeh*
Department of Mathematics Education, Farhangian University, Tehran, Iran

Abstract. The current paper deals with elaborating a numerical framework for estimating the optimal control and state of nonlinear Volterra-Fredholm integral equation (**VFIE**) by using the CAS wavelet bases. Wavelet bases have various resolution capability for approximating of different functions. The properties of CAS wavelet together with numerical integration and collocation method are utilized to discretize the continuous optimal control problem (**OCP**) to large-scale finite-dimensional nonlinear programming (**NLP**) problem. Also, the exact optimal control and state functions of OCP governed by VFIE can be approximated by series solutions based on CAS wavelet. The reduced problem is solved by existing well-developed algorithm in Mathematica software. Numerical experiments are reported to demonstrate the applicability and efficiency of the propounded technique.
Keywords: CAS wavelet, Volterra-Fredholm integral equation, Collocation method.
AMS Mathematical Subject Classification [2010]: 49M25, 90C30.

1. Introduction

In the current essay, we concentrate on the following OCP. Determine the real valued state-control function pair $(x^*(t), u^*(t))$, $t \in [0,1]$, that minimizes the cost functional

$$J = \int_0^1 x^2(t) + u^2(t) + f(t)\,x(t) + g(t)\,u(t)\,dt, \tag{1}$$

subject to the dynamic constraint

$$x(t) = y(t) + \lambda_1 \int_0^t k_1(t,s,x(s),u(s))\,ds + \lambda_2 \int_0^1 k_2(t,s,x(s),u(s))\,ds. \tag{2}$$

It is assumed that x, u and y are continuous real valued functions in $L^2[0,1]$. The functions k_1, k_2, f and g are continuously differentiable with respect to their arguments. λ_1 and λ_2 are real valued constants. The time interval is assumed to be $[0,1]$ for clarity of representation. Note that the time interval can be transformed from $[0,1]$ to $[t_0, t_f]$ via an affine transformation. In the literature, developing the computational techniques for solving OCPs especially for systems with integral equations is a subject of interest. It can be considered as a momentous research topic in many fields of the applied science and engineering. For instance, the proposed methods for solution of OCPs for systems governed by integral equations have been investigated in many studies such as [1, 4] and the references therein.

*Speaker

The outline of this paper is organized as follows: In Section 2, the proposed method is used to approximate the solution of the OCP, as a result, a NLP is obtained. In Section 3, we report our computational results and demonstrate the accuracy of the proposed numerical scheme by presenting numerical examples. Section 4 ends this paper with a concise conclusion.

2. Proposed Method

Recently, Yousefi and Banifatemi in [7] have introduced the CAS wavelets which are specified by

$$\psi_{nm}(t) = \begin{cases} 2^{\frac{k}{2}} CAS_m(2^k t - n), & t \in [\frac{n}{2^k}, \frac{n+1}{2^k}), \\ 0, & otherwise, \end{cases}$$

where

$$CAS_m(t) = \cos(2m\pi t) + \sin(2m\pi t),$$

in which $n = 0, 1, \ldots, 2^k - 1$, $k \in N \cup \{0\}$ and $m \in Z$. The function approximation with this basis and its integration operational matrix is given in [7]. In this section, we apply CAS wavelet for discretization of the OCP. Let $N = 2^k(2M+1)$ be the number of basis functions. The nodal point arrangement for the CAS wavelet collocation method is

$$t_i = \frac{2i-1}{2^{k+1}(2M+1)}, \quad i = 1, \ldots, N. \tag{3}$$

To construct approximation for considered OCP by using CAS wavelets, we assume that

$$\overline{x}(t) = \sum_{i=0}^{2^k-1} \sum_{j=-M}^{M} x_{ij}\psi_{ij}(t) = X^T\psi(t), \tag{4}$$

and

$$\overline{u}(t) = \sum_{i=0}^{2^k-1} \sum_{j=-M}^{M} u_{ij}\psi_{ij}(t) = U^T\psi(t), \tag{5}$$

where C and $\psi(t)$ are $(2^k(2M+1)) \times 1$ vectors given by

$$C = [C_{0(-M)}, C_{0(-M+1)}, \ldots, C_{0M}, C_{1(-M)}, \ldots, C_{1(M)}, \ldots, C_{(2^k-1)(-M)}, \ldots, C_{(2^k-1)(M)}]^T,$$

and

$$\psi = [\psi_{0(-M)}, \psi_{0(-M+1)}, \ldots, \psi_{0(M)}, \psi_{1(-M)}, \ldots, \psi_{1(M)}, \ldots, \psi_{(2^k-1)(-M)}, \ldots, \psi_{(2^k-1)(M)}]^T. \tag{6}$$

By substituting (4) and (5) in system (2), we gain

$$\begin{aligned} \overline{x}(t) &= y(t) + \lambda_1 \int_0^t k_1(t, s, \overline{x}(s), \overline{u}(s))ds \\ &+ \lambda_2 \int_0^1 k_2(t, s, \overline{x}(s), \overline{u}(s))ds, 0 \le t \le 1. \end{aligned} \tag{7}$$

We now collocate (7) at nodal points given in (3)

$$\begin{aligned}(8)\qquad \overline{x}(t_i) &= y(t_i)+\lambda_1\int_0^{t_i} k_1(t_i,s,\overline{x}(s),\overline{u}(s))ds\\ &+ \lambda_2\int_0^1 k_2(t_i,s,\overline{x}(s),\overline{u}(s))ds.\end{aligned}$$

For using Gauss-Legendre (GL) quadrature in (8), N intervals $[0,t_i]$ and the interval $[0,1]$ are transferred to the interval $[-1,1]$ by means of transformation $s=\frac{t_i}{2}(\tau+1)$ and $s=\frac{1}{2}(\tau+1)$, so we gain

$$\begin{aligned}(9)\qquad \overline{x}(t_i) &= y(t_i)+\lambda_1(\frac{t_i}{2})\int_{-1}^1 k_1\left(t_i,\frac{t_i}{2}(\tau+1),\overline{x}(\frac{t_i}{2}(\tau+1)),\overline{u}(\frac{t_i}{2}(\tau+1))\right)d\tau\\ &+ \frac{\lambda_2}{2}\int_{-1}^1 k_2\left(t_i,\frac{\tau+1}{2},\overline{x}(\frac{\tau+1}{2}),\overline{u}(\frac{\tau+1}{2})\right)d\tau.\end{aligned}$$

By applying Gauss-Legendre (GL) quadrature for approximating the integral involved in (9), we obtain

$$\begin{aligned}\overline{x}(t_i)=y(t_i) &+ \lambda_1(\frac{t_i}{2})\sum_{j=1}^{N_1} w_j k_1\left(t_i,\frac{t_i}{2}(\tau_j+1),\overline{x}(\frac{t_i}{2}(\tau_j+1)),\overline{u}(\frac{t_i}{2}(\tau_j+1))\right)\\ &+ \frac{\lambda_2}{2}\sum_{j=1}^{N_1} w_j k_2\left(t_i,\frac{\tau_j+1}{2},\overline{x}(\frac{\tau_j+1}{2}),\overline{u}(\frac{\tau_j+1}{2})\right),\end{aligned}$$

where τ_js are the GL nodes, zeros of Legendre polynomial $L_{N_1}(t)$ [3], in the interval $[-1,1]$, and w_js are the corresponding weights. While explicit formulas for quadrature nodes are not known, the weights can be expressed in closed form as $w_j=\frac{2}{(1+\tau_j^2)(L'_{N_1+1}(\tau_j))^2}$, $j=1,\ldots,N_1$. We utilize the following approximation methodology to discretize the performance index in (1). Firstly, the real valued functions $f(t)$ and $g(t)$ are approximated

$$(10)\qquad g(t)=G^T\psi(t),\quad f(t)=F^T\psi(t),$$

where $G=[g_{0(-M)},g_{0(-M+1)},\ldots,g_{(2^k-1)M}]$, $F=[f_{0(-M)},f_{0(-M+1)},\ldots,f_{(2^k-1)M}]$ and $\psi(t)$ is defined in (6). By substituting (10) in (1), we get

$$\begin{aligned}(11)\qquad J &= \int_0^1 X^T\psi(t)\psi^T(t)X+U^T\psi(t)\psi^T(t)\\ &+ F^T\psi(t)\psi^T(t)X+G^T\psi(t)\psi^T(t)Udt.\end{aligned}$$

We obtain from $\int_0^1\psi(t)\psi^T(t)dt=I$,

$$(12)\qquad \overline{J}(X,U)=X^TX+U^TU+F^TX+G^TU.$$

If the functions $f(t)=f$ and $g(t)=g$ are constant functions, then (11) is converted to

$$\overline{J}(X,U)=X^TX+U^TU+F^TPX+G^TPU,$$

where P is introduced in [7]. Therefore, the OCP is approximated to a nonlinear optimization problem with (12) as the objective function and (9) as constraints. Eventually, we can utilize many well-developed optimization algorithms to dissolve the finite-dimensional discretized optimization problem.

3. Numerical Results

In this section, we examine the accuracy of the propounded method on two examples. In order to analysis the error of the method, the following notations are introduced

$$(13)\quad \| E^x \|_\infty = \max_{1\le i\le 2^k(2M+1)} |E^x(t_i)|, \qquad \|E^u\|_\infty = \max_{1\le i\le 2^k(2M+1)} |E^u(t_i)|,$$

where $E^x(t) = \overline{x}^*(t) - x^*(t)$, $E^u(t) = \overline{u}^*(t) - u^*(t)$ and t_i, for $1 \le i \le 2^k(2M+1)$, are collocation nodes given in (3). We also define $E^J = |J^* - \overline{J}^*|$. In fact, $|E^x|$, $|E^u|$ and E^J are absolute errors.

Example 3.1. Consider the minimization of functional

$$J = \int_0^1 (x(t) - (t^2 - 2))^2 + (u(t) - t)^2 dt,$$

subject to controlled VFIE

$$x(t) = y(t) + \int_0^t (t - u(s))x^2(s)ds + \int_0^1 (t + u(s))x(s)ds,$$

where $y(t) = \frac{-1}{30}t^6 + \frac{1}{3}t^4 - t^2 + \frac{5}{3}t - \frac{5}{4}$. It can be verified that the exact optimal control and state are $x^*(t) = t^2 - 2$ and $u^*(t) = t$. Trivially, the optimal value of the cost functional is $J^* = 0$. Table 1 exhibits the results of solving this example with the proposed scheme in Section 3.

Table 1. Numerical results of Example 3.1.

k	$\| E^x \|_\infty$	$\| E^u \|_\infty$	E^J
$M = 1$			
2	4.2316E-02	2.6692E-02	1.1916E-03
3	2.5252E-02	2.0635E-02	3.0568E-04
4	8.7912E-03	7.6122E-03	4.0794E-05
5	1.3350E-03	5.6710E-03	1.6163E-05
$M = 2$			
2	4.4306E-02	2.4820E-02	5.4998E-04
3	2.6352E-02	2.0610E-02	1.1202E-04
4	1.2380E-02	7.7934E-03	2.8531E-05
5	8.1362E-03	1.4557E-03	2.7768E-06

Example 3.2. Consider the minimization of the cost functional

$$J = \int_0^1 (x(t) - 0.8182 - 2.7273t^2)^2 + (u(t) - t^2)^2 dt,$$

subject to the state dynamics

$$x(t) = y(t) + \int_0^1 x(s)(u(s) + t^2)ds,$$

where $y(t) = t^2$. The optimal solutions are $u^*(t) = t^2$ and $x^*(t) = 0.8182 + 2.7273t^2$. The optimal value of performance index is $J^* = 0$. Table 2 represents the error of proposed method obtained from (13).

TABLE 2. Numerical results of Example 3.2.

k	$\| E^x \|_\infty$	$\| E^u \|_\infty$	E^J
$M = 1$			
2	3.3435E-03	2.8499E-01	1.3515E-03
3	1.9582E-03	2.2821E-02	3.7128E-04
4	4.6086E-04	7.1998E-03	8.6605E-05
5	1.2480E-04	5.8216E-03	2.2474E-05
$M = 2$			
2	1.3679E-02	1.5848E-02	1.1103E-03
3	1.6481E-03	1.4022E-02	2.4918E-04
4	1.5095E-03	1.1112E-02	6.1521E-05
5	1.3085E-03	1.9518E-03	1.7370E-05

4. Conclusion

A collocation CAS wavelet-based method was developed to obtain the optimal control and state of systems governed by nonlinear VFIE. The main aspect of the proposed approach resides in converting the optimization problem into a mathematical programming problem which can be solved by a variety of efficient numerical approaches. This method is in the case of optimal control of IEs which plays a cardinal role in the numerous fields of science and engineering [2, 4].

References

1. A. Ebrahimzadeh and R. Khanduzi, *A directed tabu search method for solving controlled Volterra integral equations,* Math. Sci. **10** (2016) 115–122.
2. K. Maleknejad and A. Ebrahimzadeh, *The use of rationalized Haar wavelet collocation method for solving optimal control of Volterra integral equation,* J. Vib. Control **21** (2015) 1958–1967.
3. K. Maleknejad and A. Ebrahimzadeh, *Optimal control of Volterra integro-differential systems based on Legendre wavelets and collocation method,* Int. J. Math. Comput. Phys. Quantum Eng. **8** (2014) 1007–1011.
4. M. R. Peyghami, M. Hadizadeh and A. Ebrahimzadeh, *Some explicit class of hybrid methods for optimal control of Volterra integral equations,* J. Inform. Comput. Sci. **7** (2012) 253–266.
5. M. Razzaghi and S. Yousefi, *Legendre wavelets direct method for variational problems,* Math. Comput. Simulation **53** (2000) 185–192.
6. W. Shienyu, *Convergence of block pulse series approximation solution for optimal control problem,* Int. J. Syst. Sci. **21** (1990) 1355–1368.
7. S. Yousefi and A. Banifatemi, *Numerical solution of Fredholm integral equations by using CAS wavelets,* Appl. Math. Comput. **183** (2006) 458–463.

E-mail: a.ebrahimzadeh@cfu.ac.ir

A Reproducing Kernel Particle Method for 2D Time Fractional Telegraph Equation

Mohammad Reza Eslahchi*
Faculty of Mathematical Sciences, Tarbiat Modares University, Tehran, Iran
and Rezvan Salehi
Faculty of Mathematical Sciences, Tarbiat Modares University, Tehran, Iran

Abstract. This work is concerned with the numerical solution of two-dimensional time fractional telegraph equation by the reproducing kernel particle meshless method (RKPM). A meshless point collocation scheme is employed to furnish the spatial approximation. The Caputos fractional derivatives are approximated by two schemes of orders $\mathcal{O}(\tau^{3-\alpha})$ and $\mathcal{O}(\tau^{2-\alpha})$, $1/2 < \alpha < 1$. The RKPM is a meshless method that obtain desire accuracy and convergence by reproducing polynomial condition.
Keywords: Time fractional telegraph equation, Caputo's fractional derivative, Reproducing kernel particle method, Meshless method.
AMS Mathematical Subject Classification [2010]: 65M70, 35R11.

1. Introduction

The numerical methods have an important role in appropriate simulation of physical problems. These problems are usually depicted by partial differential equations. Recent developments in science and engineering have yield the new type of derivatives namely the fractional order derivatives. Due to the their memory, the fractional order differential equations may cause to obtain more valuable informations form real life problems. The fractional differential equations have many applications in fluid mechanics, physics, chemistry, viscoelasticity, finance, and etc. [1].

In this paper, the time fractional telegraph equation has been considered as follow

$$\begin{aligned} &\frac{\partial^{2\alpha}u(x,t)}{\partial t^{2\alpha}} + \frac{\partial^{\alpha}u(x,t)}{\partial t^{\alpha}} + u(x,t) = \Delta u(x,t) + f(x,t), \\ &\frac{1}{2} < \alpha < 1, \quad (x,t) \in \Omega \times [0,T], \end{aligned} \tag{1}$$

with boundary and initial conditions:

$$\begin{aligned} &u(x,0) = h_1(x), \qquad \frac{\partial u(x,t)}{\partial t}|_{t=0} = h_2(x), \quad x \in \Omega, \\ &u(x,t) = g(x,t), \quad x \in \partial\Omega, \end{aligned}$$

here $\Omega \subset \mathbb{R}^2$ is an open and bounded domain with the boundary $\partial\Omega$ and the fractional derivative $\frac{\partial^{\alpha}u(x,t)}{\partial t^{\alpha}}$ is the Caputo fractional derivative of order α which

*Speaker

is defined as [2]

$$\frac{\partial^\alpha u(x,t)}{\partial t^\alpha} = \begin{cases} \frac{1}{\Gamma(m-\alpha)} \int_0^t \frac{\partial^m u(x,s)}{\partial t^m}(t-s)^{m-\alpha-1}\mathrm{d}s, & m-1<\alpha<m, \quad m\in\mathbb{N}, \\ \frac{\partial^m u(x,t)}{\partial t^m}, & m=\alpha, \end{cases}$$

where m is the smallest integer that exceeds α.

A wide variety of numerical methods have been applied to approximate the solution of these equations. In the present works, we are going to use a meshless point collocation methods based on reproducing kernel particle (RKP) approximation for spatial approximation and a finite difference scheme of orders $\mathcal{O}(\tau^{3-\alpha})$ and $\mathcal{O}(\tau^{2-\alpha})$, $1/2 < \alpha < 1$ for Caputo's fractional derivatives. In the following section we will explain the main ideas of RKPM and time difference scheme.

2. Numerical Approach

In this section, at the first step the RKP approximation has been explained to semi-discretization of the considered equation as spatial approximation. Then, the fully discrete scheme has been yield by the aid of a finite difference approximation for fractional derivatives.

2.1. The Reproducing Kernel Particle Approximation. The reproducing kernel particle method as a correction of smoothed particle hydrodynamic (SPH) has been developed by Liu et al using Wavelets theory [3]. The main idea of RKPM is to retrieve the consistency condition using a corrected kernel. The modified kernel is actually a polynomial estimates of the kernel up to $2n$ degree. The kernel approximations are convolution integrals that are replaced by summations on the particles. In the RKPM, we set

$$u(x) = \int_\Omega u(y)\,\bar{w}(x-y)\,\mathrm{d}y,$$

where the correction kernel function $\bar{w}(x-y)$ is

$$\bar{w}(x-y) = c(x;x-y)\cdot w(x-y),$$

here $c(x;x-y)$ is the correction function that is approximated using polynomials as

$$c(x;x-y) = \sum_{i=1}^{m} p_i(x-y)\cdot b_i(x) = \mathbf{p}^T(x-y)b(x),$$

where m is the basis size, $p_i(x-y)$ are the monomial basis functions and $b_i(x)$ are the coefficients for each fixed particle x.

The reproducing property means that $u(x)$ can reproduce a set of basis function such as a polynomial basis functions of degree n with $d = dim(\Omega)$ defined as

$$\mathbf{p}(x) = \{x^\beta : \sum_{i=1}^{n} \beta_i \le n\},$$

where $\beta \in \mathbb{N}_0^d$ are multi-indexes. This property is concluded by substituting $\mathbf{p}(x)$ in (2) as

$$\mathbf{p}(x) = \int_{\Omega} \mathbf{p}(y)\, c(x; x-y) \cdot w(x-y)\, \mathrm{d}y.$$

It can be proved that for polynomial basis functions, the $u_h(x)$ is

$$u(x) = \mathbf{p}^T(\mathbf{0})\mathbf{M}^{-1}(x) \int_{\Omega} \mathbf{p}(y-x)\, \bar{w}(y-x) u(y)\, \mathrm{d}y, \tag{2}$$

where the $\mathbf{M}(x)$ is the so-called moment matrix which is defined as follow

$$\mathbf{M}(x) = \int_{\Omega} \mathbf{p}(y-x)\, \mathbf{p}^T(y-x)\, \bar{w}(y-x)\, \mathrm{d}y. \tag{3}$$

Replacing the integral of (2) by a summation over data sites $\{x_i\}_{i=1}^{NP}$, one can obtain

$$u_h(x) = \mathbf{p}^T(\mathbf{0})\mathbf{M}_h^{-1}(x) \sum_{I=1}^{NP} \mathbf{p}(x_I - x)\, \bar{w}(x_I - x) \Delta V_I U_I, \tag{4}$$

where ΔV_I is the sub-domain measure related to particle x_I and

$$\mathbf{M}_h(x) = \sum_{I=1}^{NP} \mathbf{p}(x_I - x)\, \mathbf{p}^T(x_I - x)\, \bar{w}(x_I - x)\, \Delta V_I.$$

Then, (4) can be expressed as

$$u_h(x) = \sum_{I=1}^{NP} \phi_I(x) U_I, \tag{5}$$

where $\phi_I(x)$ is the shape function for node I and given by

$$\phi_I(x) = C_I(x) w(x_I - x) \Delta V_I,$$
$$C_I(x) = \mathbf{p}^T(\mathbf{0})\mathbf{M}_h^{-1}(x)\mathbf{p}(x_I - x).$$

2.2. The Time Difference Scheme. Following the pioneer work of Sun and Wu [4], the approximation for fractional time derivatives can be concluded from following lemmas.

LEMMA 2.1. [4] *Suppose $g(t) \in C^2[0, t_n]$. Then for $0 < \beta < 1$, it holds that*

$$\left| \frac{1}{\Gamma(1-\beta)} \int_0^{t_n} \frac{g'(t)}{(t_n - t)^{\beta}}\, \mathrm{d}t - \frac{\tau^{-\beta}}{\Gamma(2-\beta)} \left[a_0 g(t_n) - \sum_{k=1}^{n-1} (a_{n-k-1} - a_{n-k}) g(t_k) - a_{n-1} g(t_0) \right] \right|$$
$$\leq \frac{1}{\Gamma(2-\beta)} \left[\frac{1-\beta}{12} + \frac{2^{2-\beta}}{2-\beta} - (1 + 2^{-\beta}) \right] \max_{0 \leq t \leq t_n} |g''(t)| \tau^{3-\beta},$$

where $a_k = \left[(k+1)^{2-\beta} - k^{2-\beta}\right]$.

LEMMA 2.2. [4] *Suppose $g(t) \in C^2[0, t_n]$. Then for $1 < \beta < 2$, it holds that*

$$\left| \int_0^{t_n} \frac{g'(t)}{(t_n - t)^{\beta-1}} \, \mathrm{d}t - \frac{1}{\tau} \left[b_0 g(t_n) - \sum_{k=1}^{n-1} (b_{n-k-1} - b_{n-k}) g(t_k) - b_{n-1} g(t_0) \right] \right|$$
$$\leq \frac{1}{2-\beta} \left[\frac{2-\beta}{12} + \frac{2^{3-\beta}}{3-\beta} - (1 + 2^{1-\beta}) \right] \max_{0 \leq t \leq t_n} |g''(t)| \tau^{3-\beta},$$

where $b_k = \frac{\tau^{2-\beta}}{2-\beta} \left[(k+1)^{2-\beta} - k^{2-\beta} \right]$.

2.3. The Fully Discrete Scheme. To get the fully discrete scheme, the time interval $[0, T]$ is partitioned into N equally spaced intervals of the length $\tau = \frac{T}{N}$. At each instance $t_k = k\,\tau$, the following notations are defined

$$u^{n-1/2} = \frac{1}{2}(u^n + u^{n-1}), \quad \delta_t u^{n-1/2} = \frac{1}{\tau}(u^n - u^{n-1}),$$

where $u^n = u(\mathbf{x}, t_n)$. Now, if we consider

$$\nu(x,t) = \frac{\partial u(x,t)}{\partial t},$$
$$\lambda(x,t) = \frac{1}{\Gamma(2-\alpha)} \int_0^t \frac{\partial \nu(x,s)}{\partial s} \frac{\mathrm{d}s}{(t-s)^{\alpha-1}},$$
$$\omega(x,t) = \frac{1}{\Gamma(1-\alpha)} \int_0^t \frac{\partial u(x,s)}{\partial s} \frac{\mathrm{d}s}{(t-s)^{\alpha}},$$

then, using these notations and Lemmas 2.1 and 2.2, the problem (1) at the instance $t_{k-1/2}$ can be written as

$$\lambda^{k-1/2} + \omega^{k-1/2} + u^{k-1/2} = \Delta u^{k-1/2} + f^{k-1/2}, \quad x \in \Omega, \ k \geq 1, \tag{6}$$

here

$$\nu^{k-1/2} = \delta_t u^{k-1/2} + \mathcal{O}(\tau^2),$$
$$\lambda^{k-1/2} = \frac{1}{\Gamma(2-2\alpha)\tau} \left[b_0 \nu^{k-1/2} - \sum_{i=1}^{k-1} (b_{k-i-1} - b_{k-i}) \nu^{i-1/2} - b_{n-1} \nu^0 \right] + \mathcal{O}(\tau^{3-2\alpha}),$$
$$\omega^{k-1/2} = \frac{\tau^{-\alpha}}{\Gamma(2-\alpha)} \left[a_0 u^{k-1/2} - \sum_{i=1}^{k-1} (a_{k-i-1} - a_{k-i}) u^{i-1/2} - a_{n-1} u^0 \right] + \mathcal{O}(\tau^{2-\alpha}).$$

Now, by eliminating the small errors in above equations and approximating $U^k = U(x, t_k)$ by (5), inserting into the semi-discrete scheme (6) and employing the collocation method at each interior node x_j, yields

$$\Lambda_j^{k-1/2} + W_j^{k-1/2} + U_j^{k-1/2} = \Delta U_j^{k-1/2} + f_j^{k-1/2}, \quad k \geq 1,$$

where

$$U_j^k = \sum_{I=1}^{NP} \phi_I(x) \hat{U}_I^k.$$

3. Numerical Results

To investigate the accuracy of the proposed scheme, the following two dimensional time fractional telegraph equations has been considered

$$\frac{\partial^{2\alpha}u}{\partial t^{2\alpha}}+\frac{\partial^{\alpha}u}{\partial t^{\alpha}}+u=\Delta u+2\sin(x)\,\sin(y)\,(\cos(t)-\sin(t)),\quad (x,y)\in\Omega,\quad t\in[0,T],$$

over unit square domain and $T=1$ with the exact solution

$$u(x,y,t)=\cos(t)\,\sin(x)\,\sin(y).$$

The initial and boundary conditions can be easily extracted from exact solution. The obtained root mean square (RMS) errors for different values of α with $\tau=0.1$ are summarized in Table 1. Also, the approximate solution and its contour plot are depicted in Figure 1 at a complex domain that is generated by criterion $r=\frac{1}{n^2}\left[1+2n+n^2-(n+1)\cos(n\theta)\right]$, where we used $n=4$.

TABLE 1. Obtained RMS errors with $\tau=0.1$ for different values of α.

	$\alpha=0.55$	$\alpha=0.75$	$\alpha=0.95$
NP	RMS	RMS	RMS
16	3.1542×10^{-2}	2.9837×10^{-2}	3.0572×10^{-2}
64	2.7641×10^{-2}	1.8614×10^{-2}	2.5862×10^{-2}
121	6.1732×10^{-3}	7.1135×10^{-3}	7.4812×10^{-3}
256	2.4534×10^{-3}	2.6271×10^{-3}	3.1492×10^{-3}

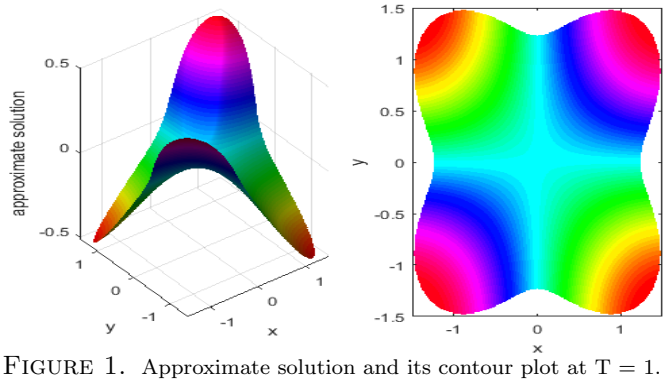


FIGURE 1. Approximate solution and its contour plot at T = 1.

References

1. R. Hilfer, *Applications of Fractional Calculus in Physics*, World Scientific, Singapore, 2000.
2. A. A. Kilbas, H. M. Srivastava and J. J. Trujillo, *Theory and Applications of Fractional Differential Equation*, Elsevier, Amsterdam, 2006.
3. W. Liu, S. Jun and Y. Zhang, *Reproducing kernel particle methods*, Int. J. Numer. Methods. Fluids. **20** (1995) 1081–1106.
4. Z .Z. Sun and X. N. Wu, *A fully discrete difference scheme for a diffusion-wave system*, Appl. Numer. Math. **56** (2006) 193–209.

E-mail: eslahchi@modares.ac.ir
E-mail: r.salehi@modares.ac.ir

Spectral Galerkin Method Using Fractional-Order Generalized Jacobi Functions for Solving Linear Systems of Fractional Differential Equations

Amin Faghih*
Department of Mathematics, Sahand University of Technology, Tabriz, Iran
and Payam Mokhtary
Department of Mathematics, Sahand University of Technology, Tabriz, Iran

Abstract. A spectral Galerkin scheme based on the newly defined fractional-order generalized Jacobi functions as basis functions are introduced to approximate the solutions of a class of systems of fractional differential equations. The numerical solvability as well as the complexity analysis of the proposed method are also investigated.
Keywords: Fractional-order generalized Jacobi functions (FGJFs), Linear systems of fractional differential equations, Galerkin method.
AMS Mathematical Subject Classification [2010]: 34A09, 65L05, 65L20.

1. Introduction

The main purpose of this paper is to develop a novel Galerkin method for the numerical solution of the following linear systems of fractional differential equations (FDEs)

$$\begin{cases} D^{\theta}V(t) = \mathcal{C}V(t) + F(t), \\ V(0) = 0,\ t \in [0,1],\ \theta = \frac{p}{q} \in (0,1), \end{cases} \tag{1}$$

where the parameters $p \geq 1$, $q \geq 2$ are two relatively prime integers, $\mathcal{C} = \{c_{ij}\}_{i,j=1}^{n}$ is the coefficient matrix, $F(t) = [f_1(t), f_2(t), \ldots, f_n(t)]^T$ and $V(t) = [v_1(t), v_2(t), \ldots, v_n(t)]^T$ are the vectors of right-hand continues functions and unknowns, respectively. D^{θ} is the Caputo type fractional derivative of order θ which is defined by [1]

$$D^{\theta}V(t) = \Big[\frac{1}{\Gamma(1-\theta)} \int_0^t (t-s)^{-\theta} v_j'(s) ds\Big]_{j=1}^{n}.$$

Our strategy produce a high accuracy and well-conditioned scheme to approximate the solutions of (1) through presenting an approach that leads to the formation of algebraic triangular systems for the unknown vector $V(t)$.

The rest of this paper is organized as follows. In the later section, FGJFs are introduced. In Section 3, We design an efficient Galerkin scheme based on the FGJFs as basis functions to approximate the solutions of (1). In Section 4, efficiency of the proposed scheme is examined by an illustrative example.

*Speaker

2. The Fractional-Order Generalized Jacobi Functions

In this section, we define new fractional-order generalized Jacobi functions (FGJFs) which are characterized by applying a suitable coordinate transformation in the generalized Jacobi polynomials/functions (GJP/Fs). This new family of orthogonal systems not only has some approximation properties for functions with singularity at boundaries, but also is a proper choice as basis function of the Galerkin or Petrov Galerkin approximation for a class of initial or boundary value systems of FDEs.

Now, let us define The GJP/Fs for each $\alpha, \beta \in \mathbb{R}$ as [4]

$$\hat{J}_n^{\alpha,\beta}(x) = (1-x)^{\hat{\alpha}}(1+x)^{\hat{\beta}} J_{\tilde{n}}^{\tilde{\alpha},\tilde{\beta}}(x), \quad x \in [-1,1], \quad \tilde{n} = n - \sigma_{\alpha,\beta} \geq 0, \ \sigma_{\alpha,\beta} = [\hat{\alpha}] + [\hat{\beta}],$$

where $J_{\tilde{n}}^{\tilde{\alpha},\tilde{\beta}}$ is the well-known classical Jacobi polynomial, and $[.]$ is the bracket function. The parameters $\hat{\alpha}, \hat{\beta}$ and $\tilde{\alpha}, \tilde{\beta}$ are defined via α, β as follows

$$\hat{\alpha} = \begin{cases} -\alpha, & \alpha \leq -1, \\ 0, & \alpha > -1, \end{cases} \qquad \tilde{\alpha} = \begin{cases} -\alpha, & \alpha \leq -1, \\ \alpha, & \alpha > -1. \end{cases}$$

Similarly for $\hat{\beta}$ and $\tilde{\beta}$ as well. These polynomials/functions are mutually $L^2_{w^{\alpha,\beta}}([-1,1])$-orthogonal, i.e.,

$$\int_{-1}^{1} \hat{J}_m^{\alpha,\beta}(x) \hat{J}_n^{\alpha,\beta}(x) w^{\alpha,\beta}(x) dx = \gamma_{\tilde{n}}^{\tilde{\alpha},\tilde{\beta}} \delta_{mn}, \quad m, n \geq \hat{n}, \tag{2}$$

in which $\gamma_{\tilde{n}}^{\tilde{\alpha},\tilde{\beta}} = \|\hat{J}_n^{\alpha,\beta}\|^2_{w^{\alpha,\beta}} = \dfrac{2^{\tilde{\alpha}+\tilde{\beta}+1}\Gamma(\tilde{n}+\tilde{\alpha}+1)\Gamma(\tilde{n}+\tilde{\beta}+1)}{(2\tilde{n}+\tilde{\alpha}+\tilde{\beta}+1)\tilde{n}!\Gamma(\tilde{n}+\tilde{\alpha}+\tilde{\beta}+1)}$, and δ_{mn} is the well-known Kronecker function.

The FGJFs $\{\hat{J}_n^{\alpha,\beta,\lambda}\}_{n \geq \sigma_{\alpha,\beta}}$ with $\lambda \in (0,1]$ and $x \in [0,1]$ are defined from the GJP/Fs through the coordinate transform $x = 2t^\lambda - 1$ as follows

$$\hat{J}_n^{\alpha,\beta,\lambda}(t) = \hat{J}_n^{\alpha,\beta}(2\ t^\lambda - 1) = 2^{\hat{\alpha}+\hat{\beta}}(1-t^\sigma)^{\hat{\alpha}} t^{\sigma\hat{\beta}} J_{\tilde{n}}^{\tilde{\alpha},\tilde{\beta}}(2\ t^\lambda - 1), \tag{3}$$

where $\alpha, \beta \in \mathbb{R}$, $n \geq \sigma_{\alpha,\beta}$.

THEOREM 2.1. *The FGJFs $\{\hat{J}_n^{\alpha,\beta,\lambda}\}_{n \geq \sigma_{\alpha,\beta}}$ form a complete mutually orthogonal system in $L^2_{w^{\alpha,\beta,\lambda}}([0,1])$ with $w^{\alpha,\beta,\lambda}(t) = \lambda(1-t^\lambda)^\alpha t^{\lambda(\beta+1)-1}$, $\alpha, \beta \in \mathbb{R}$.*

PROOF. Applying the coordinate transformation $x = 2t^\lambda - 1$ and using the relation (3), we can write

$$\int_0^1 \hat{J}_m^{\alpha,\beta,\lambda}(t) \hat{J}_n^{\alpha,\beta,\lambda}(t) w^{\alpha,\beta,\lambda}(t) dt = \frac{1}{2^{\alpha+\beta+1}} \int_{-1}^{1} \hat{J}_m^{\alpha,\beta}(x) \hat{J}_n^{\alpha,\beta}(x) w^{\alpha,\beta}(x) dx,$$

and the orthogonality relation (2) concludes the following desired result

$$\int_0^1 \hat{J}_m^{\alpha,\beta,\lambda}(t) \hat{J}_n^{\alpha,\beta,\lambda}(t) w^{\alpha,\beta,\lambda}(t) dt = \frac{1}{2^{\alpha+\beta+1}} \gamma_{\tilde{n}}^{\tilde{\alpha},\tilde{\beta}} \delta_{mn}.$$

Furthermore, suppose that the functions $u(t)$ and $U(x)$ are connected by the relation $u(t) = U(2t^\lambda - 1)$. Clearly, for any $u(t) \in L^2_{w^{\alpha,\beta,\lambda}}[0,1]$, we have $U(x) \in$

$L^2_{w^{\alpha,\beta}}[-1,1]$. Thus, the completeness of the GJP/Fs $\{\hat{J}_n^{\alpha,\beta}\}_{n\geq\sigma_{\alpha,\beta}}$ yields

$$u(t) = U(x) = \sum_{n=\sigma_{\alpha,\beta}}^{\infty} a_n \hat{J}_n^{\alpha,\beta}(x) = \sum_{n=\sigma_{\alpha,\beta}}^{\infty} a_n \hat{J}_n^{\alpha,\beta,\lambda}(t),$$

where

$$a_n = \frac{(U, \hat{J}_n^{\alpha,\beta})_{w^{\alpha,\beta}}}{\|\hat{J}_n^{\alpha,\beta}\|^2_{w^{\alpha,\beta}}} = \frac{(u, \hat{J}_n^{\alpha,\beta,\lambda})_{w^{\alpha,\beta,\lambda}}}{\|\hat{J}_n^{\alpha,\beta,\lambda}\|^2_{w^{\alpha,\beta,\lambda}}},$$

which implies the completeness of the FGJFs $\{\hat{J}_n^{\alpha,\beta,\lambda}\}_{n\geq\delta_{\alpha,\beta}}$ in $L^2_{w^{\alpha,\beta,\lambda}}[0,1]$. □

3. Numerical Approach

From existence and uniqueness theorems, it can be concluded that the $\lceil\theta\rceil$-th derivative of $v_j(t)$ often suffer from discontinuity at the initial point with the asymptotic behavior $O(t^{\lceil\theta\rceil-\theta})$, even for smooth input functions [2, 3]. This drawback affects accuracy when the GJP/Fs are applied as basis functions to obtain the Galerkin approximation of (1). To modify this weakness, we apply a new Galerkin method based on the FGJFs, which have a consistent behavior with the asymptotic behavior of the exact solutions of (1).

By substituting $\alpha = 0,\ \beta = -p$ and $\lambda = \frac{1}{q}$ into (3), the FGJFs $\{\hat{J}_r^{0,-p,\lambda}\}_{r\geq p}$ are obtained as follows

$$\hat{J}_r^{0,-p,\lambda}(t) = 2^p t^\theta J_{r-p}^{0,p}(2\ t^\lambda - 1) = \mathrm{Span}\{t^\theta, t^{\theta+\lambda}, \ldots, t^{r\lambda}\}.$$

We can write $\underline{J} = J\underline{\mathcal{T}}$, where $\underline{J} = [\hat{J}_p^{0,-p,\lambda}(t), \hat{J}_{p+1}^{0,-p,\lambda}(t), \ldots, \hat{J}_N^{0,-p,\lambda}(t), \ldots]^T$ is the vector of FGJFs with degree $(\hat{J}_r^{0,-p,\lambda}(t)) \leq r\lambda$, $\underline{\mathcal{T}} = [t^\theta, t^{\theta+\lambda}, \ldots, t^{N\lambda}, \ldots]^T$ and J is an infinite order invertible lower triangular coefficient matrix.

Now, we consider the approximate solutions as follows

$$V_N(t) = \Big[v_{j,N}(t)\Big]_{j=1}^{n}, \quad v_{j,N}(t) = \underline{d}_j\underline{J} = \underline{d}_j J\underline{\mathcal{T}}, \tag{4}$$

where $\underline{d}_j = [d_{j,p}, d_{j,p+1}, \ldots, d_{j,N}, 0, \ldots]$. Considering

$$F_N(t) = \Big[f_{j,N}(t)\Big]_{j=1}^{n}, \quad f_{j,N}(t) = \underline{f}_j\tilde{\underline{\mathcal{T}}}, \tag{5}$$

as an approximation of $F(t)$, with $\underline{f}_j = [f_{j,0}, f_{j,1}, \ldots, f_{j,N}, 0, \ldots]$,

$$\tilde{\underline{\mathcal{T}}} = [1, t^\lambda, \ldots, t^{N\lambda}, \ldots]^T,$$

and employing the relations (4) and (5) into the equivalence system of Volterra integral equations of (1), we have

$$\underline{d}\bigotimes\underline{J} = \mathcal{C}\big[\underline{d}\bigotimes I^\theta\underline{J}\big] + \underline{f}\bigotimes I^\theta\tilde{\underline{\mathcal{T}}}, \tag{6}$$

where $\underline{d} = [\underline{d}_1, \underline{d}_2, \ldots, \underline{d}_n]^T$, $\underline{f} = [\underline{f}_1, \underline{f}_2, \ldots, \underline{f}_n]^T$, and I^θ is the well-known Riemann-Liouville fractional integral operator of order θ.

Computing $I^\theta\underline{J}$ and $I^\theta\tilde{\underline{\mathcal{T}}}$, the relation (6) can be written in the following matrix formulation

$$\underline{d}\bigotimes\underline{J} = \mathcal{C}\big[\underline{d}\bigotimes(J\mathcal{M}J^{-1}\underline{J})\big] + \underline{f}\bigotimes(\mathcal{K}J^{-1}\underline{J}), \tag{7}$$

where

$$\mathcal{M}=\begin{bmatrix} \overbrace{0\dots 0}^{p} & \frac{\Gamma(\theta+1)}{\Gamma(2\theta+1)} & 0 & \dots & \\ \vdots & 0 & \frac{\Gamma(\theta+\lambda+1)}{\Gamma(2\theta+\lambda+1)} & 0 & \dots \\ \vdots & \vdots & 0 & \ddots & \ddots \end{bmatrix}, \mathcal{K}=\begin{bmatrix} \frac{1}{\Gamma(\theta+1)} & 0 & 0 & \dots \\ 0 & \frac{\Gamma(\lambda+1)}{\Gamma(\theta+\lambda+1)} & 0 & \dots \\ 0 & 0 & \frac{\Gamma(2\lambda+1)}{\Gamma(\theta+2\lambda+1)} & \dots \\ \vdots & \vdots & 0 & \ddots \end{bmatrix}.$$

Projecting (7) onto $\{\hat{J}_r^{0,-p,\lambda}\}_{r=p}^{N-p}$, and some simple manipulations the unknown coefficients satisfy the following relation

$$\underline{d}\bigotimes J=\mathcal{C}\big[\underline{d}\bigotimes(J\mathcal{M})\big]+\underline{f}\bigotimes\mathcal{K},$$

and equivalently

$$\underline{\underline{d}}=(\mathcal{C}\bigotimes\mathcal{M})\underline{\underline{d}}+\underline{\underline{f}}, \tag{8}$$

where $\underline{\underline{d}}=\underline{d}\bigotimes J$, and $\underline{\underline{f}}=\underline{f}\bigotimes\mathcal{K}$. It is noticed that, in this stage we only consider the principle sub-matrices and sub-vectors of order $\hat{N}+1$, where $\hat{N}=N-p$.

3.1. Numerical Solvability and Complexity Analysis. In this subsection, we show that the algebraic system (8) has a unique solution and propose a well-conditioned strategy to solve it. To this end, it can be easily checked that (8) can be written in the following form

$$\underline{\underline{d}}_j^{\hat{N}}=\sum_{i=1}^{n}\underline{\underline{d}}_i^{\hat{N}}(c_{ji}\mathcal{M})^{\hat{N}}+\underline{\underline{f}}_j^{\hat{N}},\quad 1\le j\le n,$$

where $\underline{\underline{f}}_j^{\hat{N}}=\underline{f}_j^{\hat{N}}\mathcal{K}^{\hat{N}}=[\underline{f}_{j,0},\underline{f}_{j,1},\dots,\underline{f}_{j,\hat{N}}]$, the corresponding index $\hat{N}$ on the top of the matrices and vectors represents the principle sub-matrices and sub-vectors of order $\hat{N}+1$ respectively, and $\underline{\underline{d}}_i^{\hat{N}}=\underline{d}_i^{\hat{N}}J^{\hat{N}}=[\underline{d}_{i,p},\underline{d}_{i,p+1},\dots,\underline{d}_{i,N}]$ is the unknown vector. Throughout, the structure of the matrix $\mathcal{M}$, it can be concluded that

$$(\underline{\underline{d}}_i^{\hat{N}}(c_{ji}\mathcal{M})^{\hat{N}})_{m=0}^{\hat{N}}=[\overbrace{0,\dots,0}^{p},G_{ji}^{0}(\underline{d}_{i,p}),G_{ji}^{1}(\underline{d}_{i,p+1}),\dots,G_{ji}^{\hat{N}-p}(\underline{d}_{i,\hat{N}})],$$

where $G_{ji}^{m-p}(\underline{d}_{i,m})=\underline{d}_{i,m}\ c_{ji}\ \frac{\Gamma(\theta+(m-p)\lambda+1)}{\Gamma(2\theta+(m-p)\lambda+1)}$, $m\ge p$, and thereby the unknown components of the unknown vectors $\{\underline{\underline{d}}_j^{\hat{N}}\}_{j=1}^{n}$ are calculated by the following recurrence relations

$$\begin{aligned}
\underline{d}_{j,p} &= \underline{f}_{j,0},\dots,\underline{d}_{j,2p-1}=\underline{f}_{j,p-1},\\
\underline{d}_{j,2p} &= \sum_{i=1}^{n}G_{ji}^{0}(\underline{d}_{i,p})+\underline{f}_{j,p},\\
&\vdots\\
\underline{d}_{j,N} &= \sum_{i=1}^{n}G_{ji}^{\hat{N}-p}(\underline{d}_{i,\hat{N}})+\underline{f}_{j,\hat{N}}.
\end{aligned}$$

Finally, obtaining $\underline{d}_i^{\hat{N}}$, $1\le i\le n$, from solving the lower triangular system $\underline{\underline{d}}_i^{\hat{N}}=\underline{d}_i^{\hat{N}}J^{\hat{N}}$, the fractional generalized Jacobi Galerkin solutions (4) can be calculated.

4. Numerical Experiment

EXAMPLE 4.1. Consider the following problem

$$\begin{cases} D^{\theta}V(t) = \mathcal{C}V(t) + F(t), \\ V(0) = 0, \ t \in [0,1], \ \theta \in (0,1), \end{cases}$$

where $V(t) = [\frac{3}{2}E_{\theta}(-t^{\theta}) - \frac{1}{2}E_{\theta}(-2t^{\theta}), \frac{3}{2}E_{\theta}(-t^{\theta}) + \frac{1}{2}E_{\theta}(-2t^{\theta})]^T$, $\mathcal{C} = \begin{bmatrix} -\frac{3}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{3}{2} \end{bmatrix}$ and $F(t) = [-\frac{1}{2}, -\frac{5}{2}]^T$.

This problem is solved via the proposed scheme for various values of θ. the numerical results are reported in Table 1 by considering 200-terms of the one parameter Mittag-Leffler functions. The listed results in Table 1 approve the reliability and high accuracy of the approximate solutions.

TABLE 1. Obtained errors for Example 4.1 with different values of θ and N.

	$\theta = \frac{1}{2}$			$\theta = \frac{2}{3}$		
N	$\|\epsilon_{1,N}\|_{L^2_{w^{0,p,\lambda}}}$	$\|\epsilon_{2,N}\|_{L^2_{w^{0,p,\lambda}}}$	*CPU-time (sec)*	$\|\epsilon_{1,N}\|_{L^2_{w^{0,p,\lambda}}}$	$\|\epsilon_{2,N}\|_{L^2_{w^{0,p,\lambda}}}$	*CPU-time*
10	5.76×10^{-1}	5.78×10^{-1}	0.34	2.50×10^{-1}	2.81×10^{-1}	0.41
20	1.18×10^{-2}	1.18×10^{-2}	0.36	1.75×10^{-2}	1.76×10^{-2}	0.44
40	1.07×10^{-8}	1.07×10^{-8}	0.42	1.68×10^{-6}	1.68×10^{-6}	0.59
80	2.52×10^{-16}	5.19×10^{-16}	0.69	5.56×10^{-16}	8.78×10^{-16}	0.84

References

1. K. Diethelm, *The Analysis of Fractional Differential Equations*, Springer, Berlin, 2010.
2. A. Faghih and P. Mokhtary, *A new fractional collocation method for a system of multi-order fractional differential equations with variable coefficients*, J. Comput. Appl. Math. **383** (2021). DOI: 10.1016/j.cam.2020.113139
3. A. Faghih and P. Mokhtary, *An efficient formulation of Chebyshev Tau method for constant coefficients systems of multi-order FDEs*, J. Sci. Comput. **82** (6) (2020). DOI: 10.1007/s10915-019-01104-z
4. B. Y. Guo, J. Shen and L. L. Wang, *Generalized Jacobi polynomials/functions and their applications*, Appl. Numer. Math. **59** (5) (2009) 1011–1028.

E-mail: a_faghih@sut.ac.ir
E-mail: mokhtary@sut.ac.ir

Numerical Solution of Nonlinear PDEs Using Modal Spectral Element Method (SEM) in Complex Geometries with Approach of Reduction of Aliasing Error

Farhad Fakhar-Izadi*
Department of Mathematics and Computer Science, Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran

Abstract. High-order SEM using orthogonal basis is proposed for solving nonlinear PDEs in complex geometries. The nonlinear terms in the weak form of equation are expanded in terms of basis by a fast Fourier transform. So, inner products of nonlinear terms can be computed using orthogonal properties of basis with reduction of aliasing error. Some examples show efficiency and accuracy of the proposed method.
Keywords: Modal spectral element, Lobatto polynomials, Aliasing error, Fast Fourier transform (FFT).
AMS Mathematical Subject Classification [2010]: 65M70, 65T50.

1. Introduction and Preliminaries

Modal SEM is used for solving nonlinear PDEs. The nonlinear terms in the Galerkin projection are expanded in terms of orthogonal modal basis using FFT. This idea can make the aliasing error so small in a lower expense than over-integration. Let $P_n^{\alpha,\beta}(x)$ denotes the nth Jacobi polynomial with real parameter $\alpha, \beta > -1$. The Lobatto polynomials are defined as $L_n(\xi) = \frac{n+2}{2} P_n^{(1,1)}(\xi)$. As mentioned in [2, 5], these polynomials maintain a high degree of orthogonality and generate sparse structure for mass and stiffness matrices. Lobatto polynomials are orthogonal in $I = [-1, 1]$ with respect to weight function $w(\xi) = (1 - \xi^2)$ as follows

$$\int_{-1}^{1} L_n(\xi) L_m(\xi) w(\xi) d\xi = \frac{2(n+1)(n+2)}{2n+3} \delta_{nm}, \tag{1}$$

where δ_{nm} is Kronecker's delta.

To ensure C_0-continuity condition of basis in the interfaces of neighbor elements, a decomposition of basis into interior and boundary modes are done [2]. So, the modal basis is defined on I as follows

$$\psi_n(\xi) = \begin{cases} \frac{1-\xi}{2}, & n = 0, \\ \frac{1-\xi^2}{4} L_{n-1}(\xi), & n = 1, \ldots, N-1, \\ \frac{1+\xi}{2}, & n = N. \end{cases} \tag{2}$$

Let the computational domain $\Omega \subseteq \mathbb{R}^2$ be partitioned by N_e non-conforming and non-overlapping quadrilateral elements $\{\Omega_e\}_{e=1}^{N_e}$ such that $\Omega = \cup_{e=1}^{N_e} \Omega_e$. If

*Speaker

$\mu : \hat{\Omega} \to \Omega_e$ is transfinite mapping which mapped $(x, y) \in \Omega_e$ on to a reference element $(\xi, \eta) \in \hat{\Omega} = I^2$, then the local mass and stiffness matrices are defined as follow

$$M^{(e)} = \int_{\hat{\Omega}} \phi_{ij}(\xi,\eta)\phi_{kl}(\xi,\eta)J^{(e)}(\xi,\eta)\ d\xi d\eta, \tag{3}$$

$$S^{(e)} = \int_{\hat{\Omega}} (D\nabla_{\xi,\eta}\phi_{ij}) \cdot (D\nabla_{\xi,\eta}\phi_{kl})J^{(e)}(\xi,\eta)\ d\xi d\eta, \tag{4}$$

where $\phi_{ij}(\xi,\eta) = \psi_i(\xi)\psi_j(\eta)$ is 2D basis on $\hat{\Omega}$ and $J^{(e)}$ is Jacobian of transformation. Also $D = \begin{pmatrix} \frac{\partial \varepsilon}{\partial x} & \frac{\partial \eta}{\partial x} \\ \frac{\partial \varepsilon}{\partial y} & \frac{\partial \eta}{\partial y} \end{pmatrix}$. For computing the entries of $M^{(e)}$, $S^{(e)}$ defined by (3)-(4), first the Jacobian term and other terms in $S^{(e)}$ are expanded in terms of Legendre polynomials. This work is done by three steps; (i) collocate the desired term at Gauss-Lobatto-Chebyshev (GLC) points; (ii) applying FFT between the physical and spectral chabyshev spaces; (iii) finding Legendre expansion coefficients of the approximation term from Chebyshev ones. This algorithm can be done in almost $O(N^d \log_2 N)$ operations in d-dimension [6]. The third step is done by a generalization of proposed algorithm in [6]. Let $\sum_{i,j} \alpha_{i,j} T_i T_j = \sum_{k,l} \beta_{k,l} P_k P_l$, where T_i and P_i are the ith Chebyshev and Legendre polynomials, respectively. Multiply both sides of above equality by $P_n P_m$ and integrate on $\hat{\Omega}$ give rises $\tilde{D} A \tilde{D}^T = CBC$, in which $A = (\alpha_{i,j})$, $B = (\beta_{k,l})$, C is diagonal matrix with $C_{ii} = \frac{2}{2i+1}$ and $\tilde{D}_{ij} = (L_i, T_j)_I$ is obtained by the following relation [6]

$$\tilde{D}_{i,j+1} = \frac{2i+2}{2i+1}\tilde{D}_{i+1,j} + \frac{2i}{2i+1}\tilde{D}_{i-1,j} - \tilde{D}_{i,j-1}. \tag{5}$$

So, $B = C^{-1}\tilde{D}A\tilde{D}^T C^{-1}$ which $\tilde{D}$ is given in (5). Finally, we have $M^{(e)} = \sum_{p,q=0}^{N_l} \beta_{p,q}(Q_p \otimes Q_q)$ and

$$S^{(e)} = \sum_{p,q=0}^{N_l} \Big[\alpha_{p,q}(G_p \otimes Q_q) + \gamma_{p,q}(Q_p \otimes G_q) + \lambda_{p,q}(F_p \otimes F_q^T + F_p^T \otimes F_q)\Big], \tag{6}$$

in which, $Q_q[n,p] = (\psi_n\psi_p, P_q)_I$, $G_q[n,p] = (\psi'_n\psi'_p, P_q)_I$, and $F_q[n,p] = (\psi'_n\psi_p, P_q)_I$ can be computed exactly using (1)-(2) and orthogonal properties of Legendre polynomials [1]. Also $\beta_{p,q}$, $\alpha_{p,q}$, $\gamma_{p,q}$, $\lambda_{p,q}$ in (6) are Legendre expansion coefficients of $J^{(e)}$, $D(:,1).D(:,1)J^{(e)}$, $D(:,2).D(:,2)J^{(e)}$, $D(:,1).D(:,2)J^{(e)}$, respectively, obtained by the aforementioned three steps algorithm.

2. Implementation and Reduction of Aliasing Error

Applying quadrature formulas with insufficient accuracy for evaluating nonlinear terms introduces aliasing error which degrades the accuracy and causes numerical instability [2, 3, 4]. In [3], super collocation or over-integration is considered for quadratic and cubic nonlinear terms on non-uniform grid for eliminating aliasing error which give rises to high computational cost and lose diagonal structure of mass matrix of nodal basis. Let $\{\upsilon_i(\xi)\}_{i=0}^P$ be an orthonormal basis of polynomial space $\mathbb{P}^P$. Define $u(\xi) = \sum_{i=0}^P \hat{u}_i \upsilon_i(\xi)$. The interesting goal is finding the expression $w(\xi) = \sum_{k=0}^P \hat{w}_k \upsilon_k(\xi)$, such that $\left\| w(\xi) - [u(\xi)]^2 \right\|_{L^2}$ is minimized. Obviously,

$\hat{w}_k$ is exactly computed as follows

$$\hat{w}_k = \sum_{i,j=0}^{P} \hat{u}_i \hat{u}_j \int_{-1}^{1} \upsilon_i(\xi)\upsilon_j(\xi)\upsilon_k(\xi)d\xi, \ \ k = 0, 1, \ldots, P. \tag{7}$$

The Gauss-Lobatto-Legendre (GLL) quadrature formula with Q nodes is exact for all integrand in $\mathbb{P}^{2Q-3}$. So, the numerical approximation of $\hat{w}_k$ in (7) (denoted by $\tilde{w}_k$) with $Q = P + 2$ nodes (which is exact in $\mathbb{P}^{2P+1}$) and the error of this approximation can be given by

$$\begin{aligned} \tilde{w}_k &= \sum_{i,j=0}^{P} \hat{u}_i \hat{u}_j [\upsilon_i(\xi)\upsilon_j(\xi), \upsilon_k(\xi)]_Q, \\ \tilde{w}_k &= \hat{w}_k - \Big[\sum_{\substack{i,j=0 \\ i+j>2P-k}}^{P} \hat{u}_i \hat{u}_j (\langle \upsilon_i\upsilon_j, \upsilon_k \rangle - [\upsilon_i\upsilon_j, \upsilon_k]_Q) \Big]. \end{aligned} \tag{8}$$

For $i + j + k \leq 2P$, there is not aliasing error because $\tilde{w}_k = \hat{w}_k$. But, in this example, $i + j + k \leq 3P$ and so aliasing error are not zero except for $k = 0$. Also, aliasing terms increase when k increases. In this paper, the nonlinear terms are computed by FFT which makes the aliasing error so small because FFT implements one of the most linearly stable interpolation scheme (Fourier interpolation on equidistant points) which has near-optimal approximation properties and is good at minimizing aliasing error in a lower expense than over-integration. Now, let $P = 35$ and $\forall i, \ \hat{u}_i = 1$. In Figure 1, the exact modes $\hat{w}_k$ are compared to $\tilde{w}_k$ which are obtained from (8) with $Q = 37$ GLL quadrature nodes. Also, in this figure, the approximation of $\hat{w}_k$ denoted by w_k^* obtained with $N = 35, 70$ Fourier collocation points are compared to the exact ones. It is evident that, in the case of FFT, $[u(\xi)]^2$ is approximated in $O(N\log_2 N)$ operations with smaller aliasing error than over-integration with $O(N^2)$ operations.

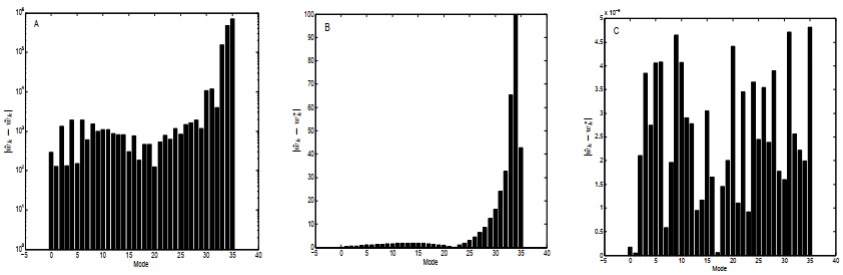


FIGURE 1. Error of approximation of $\hat{w}_k$ by A: $\tilde{w}_k$, $Q = 37$, B: w_k^*, $N = 35$, C: w_k^*, $N = 70$.

Now, the nonlinear equations in the following general form are considered

$$\frac{\partial u}{\partial t} = \mathcal{N}(u) + \mathcal{L}(u) + \mathbf{f}(\mathbf{x}, t), \quad \mathbf{x} \in \Omega \subseteq \mathbb{R}^2, \tag{9}$$

in which $\mathcal{L}(u)$, $\mathcal{N}(u)$ are linear and nonlinear terms including u and its second derivatives and integrals, respectively. Also, $\mathbf{f}(\mathbf{x}, t)$ is a smooth source function.

The boundary condition is kind of Dirichlet as $u = u_b$ on $\partial\Omega$. If $\mathbf{X} = \{u \in H^1(\Omega)|u = u_b \text{ on } \partial\Omega\}$ and $V_0 = \{v \in H^1(\Omega)|v = 0 \text{ on } \partial\Omega\}$, then the discrete variational form of (9) is finding $\hat{u}(t) \in \mathbf{X}_\delta$ such that

$$\frac{\partial}{\partial t}\langle \hat{u}, \hat{v}\rangle = \mathcal{A}(\hat{u}, \hat{v}) + \mathcal{B}(\hat{u}, \hat{v}) + \mathcal{F}(\hat{v}),\ \forall \hat{v} \in V_0^\delta, \tag{10}$$

where $\mathcal{A}(\hat{u}, \hat{v}) = (\mathcal{N}(\hat{u}), \hat{v})_\Omega$, $\mathcal{B}(\hat{u}, \hat{v}) = (\mathcal{L}(\hat{u}), \hat{v})_\Omega$ are nonlinear and bilinear forms and $\mathcal{F}(\hat{v}) = (\mathbf{f}, \hat{v})_\Omega$ is a bounded linear functional. Also, $\mathbf{X}_\delta \subseteq \mathbf{X}$, $V_0^\delta \subseteq V_0$ are finite dimensional subspaces of $\mathbf{X}$, V_0 defined by $\mathbf{X}_\delta = \{u \in \mathbf{X};\ u|_{\Omega_e} \in \mathbb{P}_N(\Omega_e),\ e = 1, \ldots, N_e\}$, $V_0^\delta = \{v \in V_0;\ v|_{\Omega_e} \in \mathbb{P}_N(\Omega_e),\ e = 1, \ldots, N_e\}$. Applying a semi-implicit time differencing scheme can linearize the discrete weak form (10). Then for computing nonlinear terms, the presented three steps algorithm in previous section can be used. For third step, a connection between coefficients of Lobatto and Chebyshev expansion of nonlinear terms on each element should be established. This connection is expressed by linear system of equations $Z^{(e)}\alpha = M^{(e)}\beta$, in which α is vector of Chebyshev expansion coefficients and β is Lobatto expansion ones. Also the entries of $Z^{(e)} = (T_i T_j, \phi_{mn} J^{(e)})_{\hat{\Omega}}$ can be computed similar to the mass and stiffness matrices [1]. In the following, the proposed method is briefly reported step by step.

i) Transfer GLC points in $\hat{\Omega}$ to local elements by transfinite mapping.
ii) Compute the nonlinear terms within each element at the GLC points.
iii) Use FFT for the nodal values to get Chebyshev expansion.
iv) Finding the modal basis expansion coefficients on each element from the Chebyshev ones using solving linear system $Z^{(e)}\alpha = M^{(e)}\beta$.
v) Assembling all coefficients vectors obtained on each element and forming the global vector.

3. Numerical Results

Let $\Omega_1 = \{(r, \theta)|\ r = R + 0.1\cos(8\theta),\ 0 \le \theta \le 2\pi\}$ and $\Omega_2 = [-0.2, 0.2]^2$. The nonlinear wave equation is considered on $\Omega = \Omega_1 - \Omega_2$ as follow

$$\frac{\partial^2 u}{\partial t^2} + \rho\frac{\partial u}{\partial t} + \beta^2 u = a\frac{\partial^2 u}{\partial x^2} + b\frac{\partial^2 u}{\partial y^2} + f(u, x, y, t),\ (x, y) \in \Omega,\ 0 < t \le T, \tag{11}$$

with $u(x, y, 0) = g_0(x, y)$, $u_t(x, y, 0) = g_1(x, y)$ as initial conditions and $u(x, y, t) = h(x, y, t)$ as boundary condition. Also g_0, g_1, h, and f are sufficiently smooth functions. If $\rho = 0$, $\beta = \sqrt{\alpha}$, $a = b = \alpha^2$ and $f(u, x, y, t) = \gamma u^3$ then Eq. (11) has an exact solution $u(x, y, t) = \sqrt{\alpha/\gamma}\tanh(\kappa(x+y-ct))$, where $\kappa = \sqrt{\alpha/(2c^2 - 4\alpha^2)}$ and c is a constant such that $c^2 > 2\alpha^2$ and $\alpha, \gamma > 0$. Also in this case Eq. (11) posses the conservation of energy

$$E = E(t) = \frac{1}{2}\int_{\mathbb{R}^2}\Big[(u_t)^2 + \alpha^2((u_x)^2 + (u_y)^2) + \alpha u^2 - \frac{\gamma}{2}u^4\Big]\mathrm{d}x\mathrm{d}y,\ \forall t > 0. \tag{12}$$

In Figure 2, numerical solution at $t = 1$ is plotted. It must be noted that BDF2-AB2 scheme is used for time differencing. L^∞ and L_2 errors are reported in Table 1. If $\tilde{E}(t)$ denotes the estimate of conserved quantity in (12), then presenting the values of $|\tilde{E}(t) - \tilde{E}(0)|$ in Table 2 shows the significance and usefulness of method.

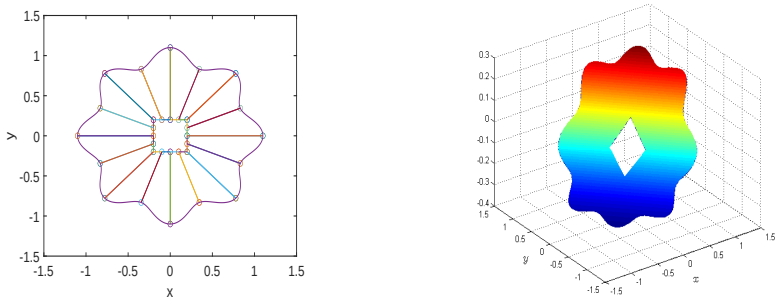


FIGURE 2. (left) dividing domain into $N_e = 16$ elements; (right) numerical solution at $t = 1$ with $N = 14$, $\Delta t = 10^{-4}$.

Table 1: Errors in the L^∞ and L_2 norms at $t = 1$ with $N_e = 16$.

δt	N	L^∞	L^2
0.001	6	1.4105858×10^{-4}	8.1562498×10^{-4}
	10	2.9807601×10^{-7}	2.3458179×10^{-6}
	14	3.3180932×10^{-7}	1.7931607×10^{-6}
0.0001	6	1.6370650×10^{-4}	8.1013011×10^{-4}
	10	2.1444105×10^{-7}	1.7333975×10^{-6}
	14	$8.9327212 \times 10^{-10}$	1.2211564×10^{-9}

Table 2: Conserved quantity of wave equation with $\Delta t = 10^{-3}$, $N = 14$, $N_e = 16$.

t	5	10	15	20	25
$\lvert\tilde{E}(t) - \tilde{E}(0)\rvert$	1.5×10^{-3}	2.5×10^{-3}	7.4×10^{-3}	7.4×10^{-3}	7.4×10^{-3}

References

1. F. Fakhar-Izadi and M. Dehghan, *Modal spectral element method in curvilinear domains*, Appl. Numer. Math. **128** (2018) 157–182.
2. G. E. Karniadakis and S. J. Sherwin, *Spectral/hp Element Methods for CFD*, 2nd ed., Oxford University Press, New York, 2004.
3. R. M. Kirby and G. E. Karniadakis, *De-aliasing on non-uniform grids: Algorithms and applications*, J. Comp. Phys. **191** (2003) 249–264.
4. R. M. Kirby and S. J. Sherwin, *Aliasing errors due to quadratic nonlinearities on triangular spectral/hp element discretisations*, J. Eng. Math. **56** (2006) 273–288.
5. C. Pozrikidis, *Introduction to Finite and Spectral Element Methods Using MATLAB*, Chapman and Hall/CRC, 2005.
6. J. Shen, *Efficient Chebyshev-Legendre Galerkin methods for elliptic problems*, In: Proc. ICOSA-HOM95, A. V. Ilin and R. Scott, eds., Houston J. Math. (1996) 233–240.

E-mail: f.fakhar@aut.ac.ir; ffizadii@gmail.com

A Polynomial Preconditioner for the LSQR Method

Somayeh Ghadamyari
Department of Mathematics, University of Sistan and Baluchestan, Zahedan, Iran
and Maryam Mojarrab*
Department of Mathematics, University of Sistan and Baluchestan, Zahedan, Iran

Abstract. LSQR is an attractive iterative method for solving the linear system $Ax = b$, and least-squares problem $\min \|Ax - b\|_2$, where A is a large and sparse matrix. Similar to other iterative methods, applying this method to ill-conditioned systems can be slow or even stagnant. To accelerate the convergence rate, we propose a polynomial type preconditioner. Some numerical examples illustrate the potency and efficiency of this preconditioned method.
Keywords: LSQR, Preconditioner, Krylov subspace methods.
AMS Mathematical Subject Classification [2010]: 15A06, 65F10, 65F20.

1. Introduction

Consider the following linear system of equations:

$$Ax = b, \tag{1}$$

where A is a large and sparse matrix in $\mathbb{R}^{m\times n}$, and $m \geq n$. b is a vector in $\mathbb{R}^m$ and x is unknown vector in $\mathbb{R}^n$. Also, it is assumed that A has full column rank. Many iterative methods have been presented for solving. Some Krylov subspace methods are based on the Arnoldi method that reduce a general square matrix to Hessenberg form.

CG-like methods with their sensible property require only a few vectors for storage and they theoretically converge at most n iterations, when A is well conditioned and many single values of A are close together and far from zero. In this condition, these methods are more beneficial. These properties occur naturally in many applications. One of the iterative methods is LSQR which has been proposed by Paige and Saunders [3]. This method does not need to store A, but in each iteration one matrix-vector product with A and one matrix-vector product with A^T are done. A sequence of approximate solutions x_k are generated such that the residual norm, $\|r_k\|_2 = \|b - Ax_k\|_2$, is reduced monotonically. For ill-conditioned problems, this method is likely to converge slowly or even stagnate. In this situation a preconditioner can improve the method. A preconditioner is a matrix or a matrix operator that, when applied to (1), can improve the condition number of A. At the same time, the solution of the new system is the same as the solution of (1).

In this paper, we propose a preconditioner called $p(A)$ for (1) as follows:

$$p(A)Ax = p(A)b, \tag{2}$$

*Speaker

where $p(A)$ is a polynomial in terms of A. The new system (2) can be solved with LSQR. This preconditioner is based on the intrinsic properties of LSQR. Then instead of (1), the linear system (2) is solved by LSQR. Numerical examples illustrate the efficiency of the preconditioned method such that it reduces the number of iterations for convergence in comparison with LSQR.

This study is organized as follows. In Section 2, a brief explanation of LSQR is recalled. The construction of the preconditioned LSQR method and some analysis discussions is presented in Section 3. Some numerical examples are implemented in Section 4. Finally, some conclusions are the subject of Section 5.

2. An Overview of LSQR

LSQR [3] is an iterative method for solving (1). This section remindes some essential properties of LSQR. In this method, Golub-Kahan procedure [1] is used to transform A to the lower bidiagonal form. The bidiagonalization process, Bidiag1, can be described as follows.

Bidiag1 (Starting vector b; reduction to lower bidiagonal form)

$$\begin{aligned} &\beta_1 u_1 = b,\ \alpha_1 v_1 = A^T u_1, \\ &\beta_{i+1} u_{i+1} = A v_i - \alpha_i u_i, \\ &\alpha_{i+1} v_{i+1} = A^T u_{i+1} - \beta_{i+1} v_i, \quad i = 1, 2, \ldots, k. \end{aligned} \tag{3}$$

The scalars $\alpha_i \geq 0$ and $\beta_i \geq 0$ are chosen so that $\|u_i\|_2 = \|v_i\|_2 = 1$. With the definitions:

$$U_k \equiv [u_1, u_2, \ldots, u_k], \quad V_k \equiv [v_1, v_2, \ldots, v_k], \qquad B_k \equiv \begin{bmatrix} \alpha_1 & & & \\ \beta_2 & \alpha_2 & & \\ & \ddots & \ddots & \\ & & \beta_k & \alpha_k \\ & & & \beta_{k+1} \end{bmatrix},$$

(3) can be rewritten as:

$$U_{k+1}(\beta_1 e_1) = b, \tag{4}$$

$$A V_k = U_{k+1} B_k, \tag{5}$$

$$A^T U_{k+1} = V_k B_k^T + \alpha_{k+1} v_{k+1} e_{k+1}^T. \tag{6}$$

Proposition 2.1. [4] *Assume that k steps of Bidiag1 have been given. Then the vectors $v_1, v_2, \ldots, v_k$ and $u_1, u_2, \ldots, u_k, u_{k+1}$ are orthonormal basis of the Krylov subspaces:*

$$\mathcal{K}_k(A^T A, v_1) = span\{v_1, A^T A v_1, \ldots, (A^T A)^{k-1} v_1\},$$
$$\mathcal{K}_{k+1}(A A^T, u_1) = span\{u_1, A A^T u_1, \ldots, (A A^T)^{k} u_1\},$$

respectively.

Proposition 2.2. [4] *Bidiag1 stops in step m if and only if $m = \min\{\mu, \lambda\}$, where μ is the grade of v_1 associated with $A^T A$ and λ is the grade of u_1 associated with $A A^T$.*

In LSQR, the approximate solution in step k is to form $x_k = V_k y_k$, where $y_k \in \mathbb{R}^k$ obtains to solve the subproblem $\min \|b - Ax\|_2$. By using (4), (5), and (6), we have:

$$\begin{aligned} r_k &= b - Ax_k = U_{k+1}(\beta_1 e_1) - AV_k y_k \\ &= U_{k+1}(\beta_1 e_1 - B_k y_k). \end{aligned} \tag{7}$$

From (7) and Proposition 2.2 the least-squares problem $\min \|b - Ax_k\|_2$ will be equivalent to

$$\min \|\beta_1 e_1 - B_k y_k\|_2. \tag{8}$$

Because U_{k+1} is a unitary matrix. (8) can be solved by using the standard QR factorization. For more details, see [3].

3. Construction of the Polynomial Preconditioner

The main idea of the polynomial preconditioned method is to construct a polynomial $p(t)$ satisfying $p(\mathcal{A}) \approx A^+$, for some matrices $\mathcal{A}$, and then solve the linear system of Eq. (2) instead of Eq. (1).

Suppose the following Krylov matrix:

$$K_k = [v_1, A^T A v_1, \ldots, (A^T A)^{k-1} v_1].$$

Since the columns of V_k span the samespace as the columns of $K_k(A^T A, v_1)$, the following relationship holds:

$$V_k = K_k R_k, \tag{9}$$

where R_k is an upper triangular matrix. It is obtained from (5)

$$A v_k = U_{k+1}(B_k)_{.,k}. \tag{10}$$

By multiplying both sides of (10) in A^T, we have

$$A^T A v_k = A^T U_{k+1}(B_k)_{.,k}. \tag{11}$$

Now, exchanging (6) in (11) gives:

$$\begin{aligned} A^T A v_k &= (V_k B_k^T + \alpha_{k+1} v_{k+1} e_{k+1}^T)(B_k)_{.,k} \\ &= V_k B_k^T (B_k)_{.,k} + \alpha_{k+1} v_{k+1} e_{k+1}^T (B_k)_{.,k} \\ &= V_k B_k^T (B_k)_{.,k} + \alpha_{k+1}\beta_{k+1} v_{k+1}, \end{aligned}$$

and

$$\alpha_{k+1}\beta_{k+1} v_{k+1} = A^T A v_k - V_k B_k^T (B_k)_{.,k}. \tag{12}$$

From (9) we have

$$v_k = K_k (R_k)_{.,k}. \tag{13}$$

Combining (12) with (13), we have

$$\begin{aligned} \alpha_{k+1}\beta_{k+1} K_{k+1}(R_{k+1})_{.,k+1} &= A^T A K_k (R_k)_{.,k} - V_k B_k^T (B_k)_{.,k} \\ &= K_{k+1} \left[\begin{pmatrix} 0 \\ (R_k)_{.,k} \end{pmatrix} - \begin{pmatrix} R_k B_k^T (B_k)_{.,k} \\ 0 \end{pmatrix} \right]. \end{aligned}$$

So a simple recursive formula for the $(k+1)$th column of R_{k+1} is obtained as followed

$$(R_{k+1})_{.,k+1} = \frac{1}{\alpha_{k+1}\beta_{k+1}} \left[\begin{pmatrix} 0 \\ (R_k)_{.,k} \end{pmatrix} - \begin{pmatrix} R_k B_k^T (B_k)_{.,k} \\ 0 \end{pmatrix} \right].$$

Hence, from (8), we have

$$\begin{aligned} x_k &= V_k y_k = K_k R_k y_k \\ &= \sum_{i=0}^{k-1} \lambda_i (A^T A)^i v_1, \end{aligned} \tag{14}$$

where $R_k y_k = [\lambda_0, \lambda_1, \cdots, \lambda_{k-1}]^T$ and y_k solves (8). Consider $q_{k-1}(t) = \sum_{i=0}^{k-1} \lambda_i t^i$ as a polynomial of degree $k-1$. Therefore, (14) can be written as

$$x_k = q_{k-1}(A^T A) v_1. \tag{15}$$

On the other hand, according to Bidiag1, $v_1 = \frac{1}{\alpha_1 \beta_1} A^T b$. So, (15) can be written as

$$x_k = \frac{1}{\alpha_1 \beta_1} q_{k-1}(A^T A) A^T b. \tag{16}$$

From the comparison (16) with (1), we conclude that $\frac{1}{\alpha_1 \beta_1} q_{k-1}(A^T A) A^T \approx A^+$, and can be used as a preconditioner for LSQR.

4. Numerical Examples

In this section some numerical experiments of using the polynomial preconditioned LSQR (PPLSQR) and the LSQR methods are provided to solve (1). In all examples, the starting guess for both methods is taken zero vector, the stopping criterion $\|r_k\|_2 \leq 10^{-10}$, and maximum number of iterations is considered 10000 iterations.

Example 4.1. In this example some matrices that are shown in Table 1 are taken from the Matrix Market [2]. The first column of Table 1 shows the names of matrices. The dimension of matrices (order) has been reported in the second column. The number of nonzero elements of each matrices is shown in the third column and the condition number of matrices that can be obtained by $cond(A) = \|A\|_2 \|A^+\|_2$ is given in the last column of Table 1. The right-hand side vector of (1) is considered such that the exact solution is a vector with entries 1. Table 2 shows the performance of PPLSQR and LSQR. For each matrix, the optimum degree (deg) shows the optimal degree of PPLSQR implementation for different degrees from $\{1, 2, \ldots, 25\}$, such that the performance with the optimal degree will result in the lowest iteration number. The number of iterations for the preconditioned method (PLSQR-It) and for LSQR (LSQR-It) have been reported in this table. As it can be seen, the preconditioned method has a better performance than the standard method. Because compared to LSQR the number of iterations for convergence has reduced, strictly.

TABLE 1. Properties of test matrices in Example 4.1.

matrix	column	row	nnz	cond
HB/well1850	1850	712	8755	$1.11e+02$
LUONG/photogrammetry2	4472	936	37056	$1.33e+08$
JGD-Taha/abtaha2	37932	331	137228	$1.22e+01$
JGD-Taha/abtaha1	14596	209	51307	$1.22e+01$

TABLE 2. Numerical results obtained for Example 4.1.

matrix	deg	PLSQR-It	PLSQR-time	LSQR-It	LSQR-time
HB/well1850	4	2	0.21	365	1.32
LUONG/photogrammetry2	3	86	048	590	2.74
JGD-Taha/abtaha2	5	109	0.4	393	1.63
JGD-Taha/abtaha1	4	123	0.11	401	0.93

5. Conclusion

To accelerate the convergence of the LSQR method, we take advantage of the LSQR method to construct a polynomial preconditioner. Numerical experiments show that the preconditioned LSQR algorithm is more efficient than the standard method without the polynomial preconditioner.

References

1. G. H. Golub and W. Kahan, *Algorithm LSQR is based on the Lanczos process and bidiagonalization procedure*, SIAM J. Numer. Anal. **2** (1965) 205–224.
2. *Matrix Market*, `http ://math.nist.gov/MatrixMarket/.`
3. C. C. Paige and M. A. Saunders, *LSQR: An algorithm for sparse linear equations and sparse least squares*, ACM Trans. Math. Softw. **8** (1) (1982) 43–71.
4. Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd ed., Society for Industrial and Applied Mathematics, Philadelphia, 2003.

E-mail: ghadamyari28@pgs.usb.ac.ir
E-mail: ma_mojarrab@math.usb.ac.ir

An Inverse Problem for the Damped BBM Equation

Fatemeh Ghanadian*
School of Mathematics and Computer Science, Damghan University, Damghan 36715-364, Iran
Reza Pourgholi
School of Mathematics and Computer Science, Damghan University, Damghan 36715-364, Iran
and Seyed Hashem Tabasi
School of Mathematics and Computer Science, Damghan University, Damghan 36715-364, Iran

Abstract. Here, we study an inverse problem related to the damped BBM equation with noisy data. By applying the quartic B-spline and Haar wavelet methods, we investigate numerically this problem. By the convergence analysis and stability, we show that our results give a fine estimation of the unknown functions of the mention inverse problem.
Keywords: BBM-type equation, Inverse problem, Quartic B-spline, Haar wavelet method.
AMS Mathematical Subject Classification [2010]: 35Q55, 65D07, 68W25, 35R30.

1. Introduction

In this paper we consider the damped generalized regularized long-wave (DGRLW) equation

$$u_t - (\varphi(x,t)u_{xt})_x - \alpha u_{xx} + u^p u_x = f(x,t), \qquad t > 0, \tag{1}$$

where x is the spatial variable, $p \geq 1$, $\varphi(x,t)$ is the variable dispersion coefficient and f is the external force. In the case $\varphi \equiv 1$, some numerical solutions are obtained based on finite difference scheme [1, 6] and finite element method [2, 3, 5, 9]. When the dispersion coefficient φ is ignored, Eq. (1) is known as the generalized Burgers equation

In this paper, we consider Eq. (1) on $(x,t) \in [0,1]\times[0,T]$ with initial condition

$$u(x,0) = p(x), \tag{2}$$

and boundary conditions, and

$$u(0,t) = g_1(t), \qquad u(1,t) = g_2(t), \tag{3}$$

where f and p are two continuous known functions and T is the final existence time. Here we study numerically an inverse problem related to (1), indeed $g_1(t)$, $g_2(t)$, and $u(x,t)$ are unknown and should be determined. There are different methods to obtain numerical solutions of DGRLW equations. Here, we present two new methods to the solution of (1). More precisely, we apply the Collocation

*Speaker

method based on a quartic B-spline basis functions and the Haar wavelet method. We will need the following assumptions: there exist two positive constants m and M such that

$$0 < m \leq \varphi(x,t) \leq M, \quad \varphi_t(x,t) \leq 0, \quad (x,t) \in [0,1] \times [0,T],$$

It is known that the use of B-splines have many different features and are effective in numerical works. One of the most important feature is that the conditions on the continuity of functions are built-in and have the smooth interpolation functions. On the other hand, as the support of each B-spline is embedded only on a few sub-intervals, the resulting matrix related to the discretized equation will be tightly banded. Moreover, if one combine with collocation, the solution procedure will be clear and shorten.

In the second part of the paper, we will use the Wavelet methods based on the Haar wavelets. Indeed, as they are made up of pairs of piecewise constant functions, they are the simplest functions in the family of the wavelet functions. On the other hand, one of the most worthy property of Haar wavelets is that one can integrate analytically these wavelets in arbitrary times, however their discontinuity is a big barrier in the general wavelet functions. One can read [4], where Çelik used the Haar wavelet method for solving magnetohydrodynamic flow equations in a rectangular duct in presence of transverse external oblique magnetic field, and also Saha in [7] applied the Haar wavelet method for the numerical solution of fractional Bagley Torvik equation. We refer the reader to study [4, 7] and references therein where the Haar wavelet techniques were used for the solutions of several differential equations. We employ this approach to see the efficiency of the Haar wavelet method for our inverse problem (1) and (2)-(3).

2. Main Results

We first use the Quartic B-spline collection method to study our inverse problem. Problem (1) and (2)-(3) will be solved with the over-specified conditions

$$u(a,t) = h_1(t), \quad u_x(a,t) = h_2(t), \quad u_{xx}(a,t) = h_3(t), \tag{4}$$

where $t \in [0,T]$ and $0 < a < 1$ is denoted as a fixed point. We consider the quartic B-spline $B_i(x)$ for $i = -2(1)N+1$ as in [8]. Let $U_m(x,t) \in \zeta$ be the B-spline approximation to the exact solution $u(x,t)$ in the form $U_m(x,t) = \sum_{i=-2}^{m+1} c_i(t) B_i(x)$, where $c_i(t)$ are time dependent parameters determined by the boundary and collocation conditions. Substituting trial functions B_j into the above equation, the nodal values of U, U', U'' and U''' are obtained in terms of the element parameters c_m by

$$\begin{aligned} U_m &= c_{m+1} + 11c_m + 11c_{m-1} + c_{m-2}, \\ U'_m &= \frac{4}{h}(c_{m+1} + 3c_m - 3c_{m-1} - c_{m-2}), \\ U''_m &= \frac{12}{h^2}(c_{m+1} - c_m - c_{m-1} + c_{m-2}), \\ U'''_m &= \frac{24}{h^3}(c_{m+1} - 3c_m + 3c_{m-1} - c_{m-2}), \end{aligned} \tag{5}$$

we use the following finite difference approximation to discretize the time variable with the uniform step size k to see $u_t^n \cong \dfrac{\delta_t}{k(1-\gamma\delta_t)}u^n$, $n \geq 0$ and $\gamma \neq 1$, where $\delta_t u^n = u^{n+1} - u^n$, $u^n = u(x,t_n)$ and $u^0 = u(x,0) = p(x)$. So

$$\frac{\delta_t}{k(1-\gamma\delta_t)}(u^n - \varphi_x(x,t_n)u_x^n - \varphi(x,t_n)u_{xx}^n) = \alpha u_{xx}^n - u^p {u_x}^n + f(x,t_n).$$

The nonlinear term is linearized by using the quasi-linearization formula as given below:

$$f(u^{n+1},u_x^{n+1}) = f(u^n,u_x^n) + (u^{n+1}-u^n)\frac{\partial f^n}{\partial u} + (u_x^{n+1}-u_x^n)\frac{\partial f^n}{\partial u_x}.$$

for $p=1$ and $p=2$ and by rearranging terms we obtain that

$$A^* c_{i-2}^{n+1} + B^* c_{i-1}^{n+1} + C^* c_i^{n+1} + D^* c_{i+1}^{n+1} = H(x_i,t_n) + H(x_i,t_{n+1}),$$

consist of $(N+1)$-linear equation with $(N+4)$ unknowns,

$$(c_{-2}, c_{-1}, c_0, \ldots, c_N, c_{N+1})^T.$$

To have a unique solution of the above system we are required the over-specified condition (4). Suppose that $a = x_s$, $1 \leq s \leq N-1$, thusly we have

$$u(x_s,t) = h_1(t), \quad u_x(x_s,t) = h_2(t), \quad u_{xx}(x_s,t) = h_3(t),$$

where $t \in [0,T]$. If we consider $m = s$ in (5), then we have $h_1(t_{n+1}) = c_{s+1}^{n+1} + 11c_{s+2}^{n+1} + 11c_{s+3}^{n+1} + c_{s+4}^{n+1}$, $h_2(t_{n+1}) = \frac{4}{h}(c_{s+4}^{n+1} + 3c_{s+3}^{n+1} - 3c_{s+2}^{n+1} - c_{s+1}^{n+1})$, $h_3(t_{n+1}) = \frac{12}{h^2}(c_{s+4}^{n+1} - c_{s+3}^{n+1} - c_{s+2}^{n+1} + c_{s+1}^{n+1})$. Consequently, $AC = B$ is a system of $(N+4)$ linear equations with $(N+4)$-unknown functions. We notice that the matrix A is ill-condition, so we obtain solution of system $AC = B$ by using the Tikhonov regularization method. We check the convergence of our algorithm. Suppose that $U(x) = \sum_{i=-2}^{N+1} c_i B_i(x)$ is the B-spline collocation approximation of $u(x)$. The following lemma and thearem will be important in our analysis that proofs of them have been done,

LEMMA 2.1. *If $\{B_{-2}, B_{-1}, B_0, \ldots, B_{N+1}\}$ be quartic B-spline, then*

$$\left|\sum_{i=-2}^{N+1} B_i(x)\right| \leq 35,$$

for $x \in [0,1]$.

THEOREM 2.2. *Let $u \in C^5[0,1]$ be an exact solution of (1) such that $\left|\frac{\partial^5 u(x,t)}{\partial x^5}\right| \leq L$ for all x,t. If $U(x,t)$ is the numerical approximation by our method of u, then $\|u(x) - U(x)\| \leq O(k^2 + h^3)$.*

We will investigate the stability by applying Von-Neuman stability analysis. Also we will obtain a numerical solution for the nonlinear inverse problem (1) which is based on the Haar wavelet method with the over-specified conditions $u(a,t) = h_1(t)$ and $u_x(a,t) = h_2(t)$, where $a \in (0,1)$ is a fixed point and $t \in [0,t_f]$. It is known that any integrable function $u \in L^2[0,1]$ can be written in terms of by the Haar coefficients with an infinite number of terms $u(x) = \sum_{i=1}^{\infty} c_i h_i(x), c_i = 2^j \int_0^1 h_i(x)u(x)\,dx$, where $i = 2^j + k + 1$ with

$j \geq 0$ and $0 \leq k \leq 2^j$. Note that $c_1 = \int_0^1 u(x)\ dx$. so $u(x) = c_1 h_1(x) + \sum_{j\geq 0} \sum_{k=0}^{2^j-1} c_{2^j+k+1} h_{2^j+k+1}(x)$. The function $u(x)$ is terminated at finite step, denoting $u_j(x)$, because $u(x)$ is piecewise constant function or can be approximated by some continuous function which are piecewisely constant in each subinterval $u_j(x) \cong c_1 h_1(x) + \sum_{j=0}^{J} \sum_{k=0}^{2^j-1} c_{2^j+k+1} h_{2^j+k+1}(x) = C_M^T H_M(x)$. Now C_M^T and the Haar function $H_M(x)$ are defined as $C_M^T = (c_1, c_2, \ldots, c_M), H_M(x) = (h_1(x), h_2(x), \ldots, h_M(x))^T$. Suppose that the interval $[0, T]$ is spitted into N subintervals of length $\Delta t = \frac{T}{N}$ and denote $t_s = (s-1)\Delta t$, $s = 1(1)N+1$. We assume that $\dot{u}''$ can be expanded in terms of Haar wavelets as,

$$\dot{u}'' \cong c_1^s h_1(x) + \sum_{j=0}^{J} \sum_{k=0}^{2^j-1} c_{2^j+k+1}^s h_{2^j+k+1}(x) = C_M^T H_M(x). \tag{6}$$

The notations $\cdot$ and $''$ denotes the differentiation with respect to t and x, respectively. By integrating Eq. (6) in t from t_s to t, and then in x from a to x, and using the over-specified conditions, we obtain that

$$\begin{aligned}
u''(x,t) &= (t-t_s) C_M^T H_M(x) + u''(x,t_s),\\
u'(x,t) &= (t-t_s) C_M^T [P_M H_M(x) - P_M H_M(a)] + u'(x,t_s) + h_2(t) - h_2(t_s),\\
\dot{u}'(x,t) &= C_M^T [P_M H_M(x) - P_M H_M(a)] + h_2'(t),
\end{aligned}$$

and

$$\begin{aligned}
u(x,t) =& (t-t_s) C_M^T [Q_M H_M(x) - Q_M H_M(a) - (x-a) P_M H_M(a)] + u(x,t_s)\\
&+ [h_1(t) - h_1(t_s)] + (x-a)[h_2(t) - h_2(t_s)].
\end{aligned} \tag{7}$$

Now, it follows by differentiating (7) in t that

$$\dot{u}(x,t) = C_M^T [Q_M H_M(x) - Q_M H_M(a) - (x-a) P_M H_M(a)] + h_1'(t) + (x-a) h_2'(t),$$

for $p = 1$ and $p = 2$, we obtained a system of M linear equations with M unknowns and can be rewritten in a form of matrix vector, as follows $AX = B$. We can calculate the approximate solution successively for $l = 1(1)M$ and $s = 1(1)N$ as follows:

$$\begin{aligned}
u(0,t_{s+1}) = g_1(t_{s+1}) &= T C_M^T [a P_M H_M(a) - Q_M H_M(a)] + u(0,t_s) + [h_1(t_{s+1}) - h_1(t_s)]\\
&- a[h_2(t_{s+1}) - h_2(t_s)],\\
u(1,t_{s+1}) = g_2(t_{s+1}) &= T C_M^T [Q_M H_M(1) - Q_M H_M(a) - (1-a) P_M H_M(a)] + u(1,t_s)\\
&+ [h_1(t_{s+1}) - h_1(t_s)] + (1-a)[h_2(t_{s+1}) - h_2(t_s)],\\
u(x_l,t_{s+1}) &= T C_M^T [Q_M H_M(x_l) - Q_M H_M(a) - (x_l - a) P_M H_M(a)] + u(x_l,t_s)\\
&+ [h_1(t_{s+1}) - h_1(t_s)] + (x_l - a)[h_2(t_{s+1}) - h_2(t_s)].
\end{aligned}$$

Example 2.3. In this example, we solve the inverse nonlinear problem (1) and (2)-(3) in $(x,t) \in [0,1] \times [0,1]$ with $\varphi(x,t) = (x^2+1)e^{-t/10}$, the initial data $u(x,0) = \sin(x)$, and the force function $f(x,t) = -\sin(x)e^{-t} - (2x\cos(x) + (x^2+1))e^{\frac{-11t}{10}}) + \frac{1}{2}\sin(2x)e^{-2t}$. The exact solution of (1) is $u(x,t) = \sin(x)e^{-t}, u(0,t) = g_1(t) = 0, u(1,t) = g_2(t) = \sin(1)e^{-t}$. Tables 1 and 2 show the numerical results of $u(0,t)$ and $u(1,t)$, respectively. Table 3 shows the numerical values of $u(x,t)$ at point $x = 0.1$.

TABLE 1. A comparison between the numerical and exact values of $g_1(t)$ in Example 2.3 by applying the quartic B-spline collocation method with $N = 100$ and Haar wavelet method with $M = 32$.

	Quartic B-spline			Haar wavelet		
t	$g_1(t)$	$g_1^*(t)$	$\lvert g_1(t) - g_1^*(t)\rvert$	$g_1(t)$	$g_1^*(t)$	$\lvert g_1(t) - g_1^*(t)\rvert$
0.1	0.000000	-0.000087	$8.680878e-05$	0.000000	0.000006	$6.003483e-06$
0.5	0.000000	0.000003	$2.585573e-06$	0.000000	0.000031	$3.131937e-05$
1	0.000000	0.000049	$4.857072e-05$	0.000000	0.000045	$4.450173e-05$
S_{g_1}	-	-	$1.5505e-06$	-	-	$1.0034e-05$

TABLE 2. A comparison between the numerical and exact values of $g_2(t)$ in Example 2.3 by applying the quartic B-spline collocation method with $N = 100$ and Haar wavelet method with $M = 32$.

	Quartic B-spline			Haar wavelet		
t	$g_2(t)$	$g_2^*(t)$	$\lvert g_2(t) - g_2^*(t)\rvert$	$g_2(t)$	$g_2^*(t)$	$\lvert g_2(t) - g_2^*(t)\rvert$
0.1	0.761394	0.762338	$9.439216e-04$	0.761394	0.770657	$9.262471e-03$
0.5	0.510378	0.512352	$1.974290e-03$	0.510378	0.520399	$1.002151e-02$
1	0.309560	0.311945	$2.384915e-03$	0.309560	0.320023	$1.046333e-02$
S_{g_2}	-	-	$5.9608e-05$	-	-	$1.0034e-05$

TABLE 3. The comparison between the exact and numerical values of function $u(0.1, t)$ in Example 2.3 extracted from the quartic B-spline collection method with $N = 100$ and Haar wavelet method with $M = 32$.

	Quartic B-spline			Haar wavelet		
t	$u(0.1,t)$	$u^*(0.1,t)$	$\lvert u(0.1,t) - u^*(0.1,t)\rvert$	$u(0.1,t)$	$u^*(0.1,t)$	$\lvert u(0.1,t) - u^*(0.1,t)\rvert$
0.1	0.090333	0.090307	$2.555846e-05$	0.090333	0.090335	$2.129506e-06$
0.5	0.060552	0.060553	$1.415859e-06$	0.060552	0.060562	$1.028834e-05$
1	0.036727	0.036742	$1.530312e-05$	0.036727	0.036741	$1.466508e-05$
S	-	-	$4.7494e-07$	-	-	$3.3021e-06$

So we have employed successfully the quartic B-spline method and the Haar wavelet to estimate unknown boundary conditions in an inverse problem related to the damped generalized regularized long-wave (DGRLW) Eq. (1) with (2)-(3). We have discussed the convergence rate of the our methods and shown the rate of the quartic B-spline method is $O(k^2 + h^3)$, while $O(\frac{1}{M})$ is the convergence rate for the Haar wavelet method.

References

1. T. Achouri, N. Khiari and K. Omrani, *On the convergence of difference schemes for the Benjamin-Bona-Mahony (BBM) equation*, Appl. Math. Comput. **182** (2006) 999–1005.

2. D. N. Arnold, J. Douglas Jr. and V. Thomée, *Superconvergence of finite element approximation to the solution of a Sobolev equation in a single space variable*, Math. Comp. **36** (1981) 53–63.
3. R. E. Ewing, *Time-stepping Galerkin methods for nonlinear Sobolev partial differential equation*, IAM J. Numer. Anal. **15** (1978) 1125–1150.
4. İ. Çelik, *Haar wavelet approximation for magnetohydrodynamic flow equations*, Appl. Math. Model. **37** (2013) 3894–3902.
5. K. Omrani, *The convergence of the fully discrete Galerkin approximations for the Benjamin-BonaMahony (BBM) equation*, Appl. Math. Comput. **180** (2006) 614–621.
6. K. Omrani and M. Ayadi, *Finite difference discretization of the BenjaminBonaMahonyBurgers (BBMB) equation*, Numer. Meth. Partial Diff. Eq. **24** (2008) 239–248.
7. S. Saha Ray, *On Haar wavelet operational matrix of general order and its application for the numerical solution of fractional Bagley Torvik equation*, Appl. Math. Comput. **218** (2012) 5239–5248.
8. B. Saka and I. Dag, *Quartic B-spline Galerkin approach to the numerical solution of the KdVB equation*, Appl. Math. Comput. **215** (2009) 746–758.
9. L. Wahlbin, *Error estimates for a Galerkin method for a class of model equations for long waves*, Numer. Math. **23** (1975) 289–303.

E-mail: ghanadian85@gmail.com
E-mail: pourgholi@du.ac.ir
E-mail: tabasi@du.ac.ir

A Numerical Meshless Method for Fractional Differential Equations

Ali Habibirad*
Department of Mathematics, Shiraz University of Technology, Shiraz, Iran
and Esmail Hesameddini
Department of Mathematics, Shiraz University of Technology, Shiraz, Iran

Abstract. This manuscript proposed an efficient meshless method for numerical solution of fractional differential equations. The main advantage of this scheme is to obtain a global approximation for this problem which reduces such problems to a system of algebraic equations. To approximate the first and derivative fractional order against the time, we use the finite difference relations. To discretization this model in space variables, we use the MK interpolation. An example is provided and the results are compared to their analytical solutions to verify the efficiency of our method.
Keywords: Fractional differential equations, Moving Kriging (MK) interpolation.
AMS Mathematical Subject Classification [2010]: 65M12, 65M60, 34A45.

1. Introduction

Fraction differential equations (FDEs) are one of the most important branches of mathematics that have been considered by many researchers in recent years. They have many applications in engineering, and sciences such as physics, chemistry, and fluid mechanics. The numerical solution of these equations is of great importance. Here we present a meshless method for numerically solving these equations. In this concept, we suppose the following time fractional model of FDEs

$$\begin{aligned} &\frac{\partial^\alpha \Psi(\mathbf{x},t)}{\partial t^\alpha} + \gamma_1 \frac{\partial^\beta \Psi(\mathbf{x},t)}{\partial t^\beta} + \gamma_2 \Psi(\mathbf{x},t) = \gamma_3 \nabla \Psi(\mathbf{x},t) + f(\mathbf{x},t), \\ &1 < \alpha \leq 2, \quad \beta = \alpha - 1, \end{aligned} \tag{1}$$

subject to the initial and boundary conditions

$$\begin{aligned} \Psi(\mathbf{x},0) &= \Psi_0(\mathbf{x}), \quad \mathbf{x} \in \Omega \subset \mathbb{R}^d, \quad \frac{\partial \Psi(\mathbf{x},0)}{\partial t} = g(\mathbf{x}), \\ \Psi(\mathbf{x},t) &= h(\mathbf{x},t), \quad \mathbf{x} \in \partial\Omega,\ t \in [0,T], \end{aligned} \tag{2}$$

where $\frac{\partial^\alpha}{\partial t^\alpha}$ and $\frac{\partial^\beta}{\partial t^\beta}$ are the Caputo sense of fractional differential operator with respect to t as follows

$$\frac{\partial^\alpha \Psi(\mathbf{x},t)}{\partial t^\alpha} = \begin{cases} \frac{1}{\Gamma(2-\alpha)} \int_0^t \frac{1}{(t-\eta)^{\alpha-1}} \frac{\partial^2 \Psi(\mathbf{x},\eta)}{\partial \eta^2} d\eta, & 1 < \alpha < 2, \\ \frac{\partial^2 \Psi(\mathbf{x},t)}{\partial t^2}, & \alpha = 2, \end{cases}$$

*Speaker

$$\frac{\partial^\beta \Psi(\mathbf{x},t)}{\partial t^\beta} = \begin{cases} \frac{1}{\Gamma(1-\beta)} \int_0^t \frac{1}{(t-\eta)^\beta} \frac{\partial \Psi(\mathbf{x},\eta)}{\partial \eta} d\eta, & 0 < \beta < 1, \\ \frac{\partial \Psi(\mathbf{x},t)}{\partial t}, & \beta = 1. \end{cases}$$

In Eq. (1) Δ is Laplacian differential operator. Also, γ_1, γ_2 and γ_3 are positive constants. Moreover, Ψ_0, g, f and h are given continuous functions. As regards the solution of FDEs is of essential importance to explain several phenomena in engineering and physics, thus solving this equation is very necessary. In this paper, we will apply an collocation meshless scheme based on Moving Kriging interpolation method to obtain the numerical solution of (1). This paper is constructed from the following sections: In Section 2, we give brief review of the MK interpolation. In Section 3 , we explain the time discretization and numerical performance of the meshless technique for the FDEs. Finally, a brief conclusion is given in Section 4.

2. The Moving Kriging Interpolation

The moving Kriging (MK) interpolation is a well-known geostatic technique for spatial interpolation in geology and mining [1, 3]. In the following, we will explain the building of meshless shape function using MK interpolation.

Assume that the problem domain $\Omega \subseteq \mathbb{R}^2$ is discretized by a set of properly scattered nodes $\mathbf{x}_i$, $i = 1, 2, \ldots, n$ and $u(\mathbf{x})$ is a function defined in Ω. Similar to the MLS approximation, assumed that only N nodes surrounding point $\mathbf{x}$ have the effect on $\Psi(\mathbf{x})$. The MK interpolation $\Psi^h(\mathbf{x})$ is defined as [4, 5, 6]

$$\Psi^h(\mathbf{x}) = \sum_{j=1}^{N} \phi_i(\mathbf{x}) \hat{\Psi}_i = \Phi(\mathbf{x}) \mathbf{\Psi}, \quad \mathbf{x} \in \Omega_x,$$

where

$$\Phi(\mathbf{x}) = \mathbf{p}^T(\mathbf{x}) A + \mathbf{r}^T(\mathbf{x}) B.$$

Matrices A and B are known by the following relations

$$A = (P^T R^{-1} P)^{-1} P^T R^{-1}, \quad B = R^{-1}(I - PA),$$

in which I is an unit matrix of sizes $N \times N$, and vector $\mathbf{p}(x)$ is

$$\mathbf{p}^T(\mathbf{x}) = [p_1(\mathbf{x}) \cdots p_m(\mathbf{x})],$$

where $p_j(\mathbf{x}), j = 1, 2, \ldots, m$ is the m polynomial basis functions, which have monomial terms. For example, for a two-dimensional problem, the linear basis is $\mathbf{p}^T(\mathbf{x}) = [1, x, y]$ and the quadratic basis is $\mathbf{p}^T(\mathbf{x}) = [1, x, y, x^2, y^2, xy]$. Note that we have used a quadratic basis in our computations in the next sections. Values of the polynomial basis functions (3) at the given set of nodes are collected in the matrix P as

$$P = \begin{bmatrix} p_1(\mathbf{x}_1) & \cdots & p_m(\mathbf{x}_1) \\ \cdots & \cdots & \cdots \\ p_1(\mathbf{x}_N) & \cdots & p_m(\mathbf{x}_N) \end{bmatrix}.$$

Also, the vector $\mathbf{r}(\mathbf{x})$ in (3) is given by

$$\mathbf{r}^T(\mathbf{x}) = \begin{bmatrix} \gamma(\mathbf{x}, \mathbf{x}_1) & \ldots & \gamma(\mathbf{x}, \mathbf{x}_N) \end{bmatrix},$$

where $\gamma(\mathbf{x}, \mathbf{x}_j)$ is the correlation function between any pair of nodes located at $\mathbf{x}$ and $\mathbf{x}_j$. Many functions can be used as a correlation function. In the current study, the following weight function is used

$$\gamma(\mathbf{x}, \mathbf{x}_j) = \begin{cases} 1 - 6d_j^2 + 8d_j^3 - 3d_j^4, & d_j \leq 1, \\ 0, & d_j > 1, \end{cases} \tag{3}$$

where $d_j = \frac{\|\mathbf{x}-\mathbf{x}_j\|}{r_j}$, in which r_j is the size of support in correlation function (3). In addition, the correlation matrix R is given in the following form

$$R = \begin{bmatrix} \gamma(\mathbf{x}_1, \mathbf{x}_1) & \cdots & \gamma(\mathbf{x}_1, \mathbf{x}_N) \\ \cdots & \cdots & \cdots \\ \gamma(\mathbf{x}_N, \mathbf{x}_1) & \cdots & \gamma(\mathbf{x}_N, \mathbf{x}_N) \end{bmatrix}.$$

The first- and second-order partial derivatives of shape function $\Phi(\mathbf{x})$ against the coordinates x and y can be easily obtained from (3) as

$$\phi_{,i}(\mathbf{x}) = P_{,i}^T(\mathbf{x})A + r_{,i}^T(\mathbf{x})B, \quad \phi_{,ii}(\mathbf{x}) = P_{,ii}^T(\mathbf{x})A + r_{,ii}^T(\mathbf{x})B.$$

The shape function $\Phi(\mathbf{x})$ obtained from the MK interpolation possesses the delta function property. For other properties of the MK based shape functions see [5].

3. Numerical Implementation

In this section, we illustrate a meshless method based on MK interpolation for solving Eq. (1). Suppose, Ω is the problem domain and $\mathbf{X}^* = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_n\}$ be an arbitrary sufficient scattered nodes in the global domain Ω. Here, instead of calculation global weak form, we construct weak form over a local subdomain like the $\Omega_{\mathbf{x}}$ which is a small region environment over each node in the general domain Ω. These subdomains can be any arbitrary geometric shapes and size [1, 3] which overlap each other and cover entire domain Ω. In our study, we take them as the circle shape in Ω. Now, for every random point $\mathbf{x}_i \in \mathbf{X}^* (1 \leq i \leq n)$ we introduce the local weak form of (1) in associated subdomain $\Omega_{\mathbf{x}}^i \subset \Omega$ to $\mathbf{x}_i$. For every point $\mathbf{x}_i$, the local weak form of Eq. (1) in $\Omega_{\mathbf{x}}^i$ is as follows

$$\frac{\partial^\alpha \Psi(\mathbf{x}_i, t)}{\partial t^\alpha} + \gamma_1 \frac{\partial^\beta \Psi(\mathbf{x}_i, t)}{\partial t^\beta} + \gamma_2 \Psi(\mathbf{x}_i, t) = \gamma_3 \nabla \Psi(\mathbf{x}_i, t) + f(\mathbf{x}_i, t), \tag{4}$$

Assuming just N points in the neighborhood node $\mathbf{x}_k$ have the effect on the numerical solution, results in

$$\Psi_h(\mathbf{x}, t) = \sum_{j=1}^{N} \Psi_j(\mathbf{x}) \hat{\Psi}_j(t). \tag{5}$$

Substituting the MK interpolation (5) in Eq. (4), the following system for all nodes will be obtained

$$C \frac{\partial^\alpha \Psi(t)}{\partial t^\alpha} + \gamma_1 C \frac{\partial^\beta \Psi(t)}{\partial t^\beta} = K\Psi(t) + F. \tag{6}$$

Note that C is an unit matrix because the shape functions obtained by the MK interpolation have the δ Kronecker property, and

$$K_{Ij} = -\gamma_2 \phi_j(\mathbf{x}_I) + \gamma_3 \Delta \phi_j, \qquad F(I) = f(\mathbf{x}_I, t).$$

To generate a fully discrete scheme of Eq. (6), suppose $\tau = \frac{T}{n}$ be the step time. So, we take $t_k = k\tau, k = 0, 1, 2, \ldots, n$, which n is a non-negative integer. Using the following relations [2] to approximate the fractional derivative

$$\begin{aligned}(7)\quad \frac{\partial^\alpha \Psi(\mathbf{x}, t_{n+1})}{\partial t^\alpha} &= c_0[\Psi_{n+1} - 2\Psi_n + \Psi_{n-1} \\ &\quad + \sum_{k=1}^{n} d_k \left(\Psi_{n-k+1} - 2\Psi_{n-k} + \Psi_{n-k-1}\right)] + O(\tau^{3-\alpha}),\end{aligned}$$

$$(8)\quad \frac{\partial^\beta \Psi(\mathbf{x}, t_{n+1})}{\partial t^\beta} = a_0[\Psi_{n+1} - \Psi_n + \sum_{k=1}^{n} b_k \left(\Psi_{n-k+1} - \Psi_{n-k}\right)] + O\left(\tau^{2-\beta}\right),$$

in which $c_0 = \frac{\tau^{-\alpha}}{\Gamma(3-\alpha)}, d_k = \left[(k+1)^{2-\alpha} - (k)^{2-\alpha}\right], a_0 = \frac{\tau^{-\beta}}{\Gamma(2-\beta)}, b_k = (k+1)^{1-\beta} - (k)^{1-\beta}$ and $\Psi_n = \Psi(\mathbf{x}, n\tau)$. Also,

$$(9)\quad \Psi = \frac{1}{2}(\Psi_{n+1} + \Psi_n).$$

Substituting Eqs. (7), (8) and (9) in (6), the following discrete scheme in time variable is resulted

$$\begin{aligned}(\mu_1 C - \frac{1}{2}K)\Psi_{n+1} &= (\mu_2 C + \frac{1}{2}K)\Psi_n - c_0 C \Psi_{n-1} \\ &- c_0 C \sum_{k=1}^{n} d_k \left(\Psi_{n-k+1} - 2\Psi_{n-k} + \Psi_{n-k-1}\right) \\ &- a_0\gamma_1 C \sum_{k=1}^{n} b_k \left(\Psi_{n-k+1} - \Psi_{n-k}\right) + F_n. \qquad (10)\end{aligned}$$

in which $\mu_1 = c_0 + a_0\gamma_1,\ \mu_2 = 2c_0 + a_0\gamma_1$ and $F_n = F(\mathbf{x}, t_n)$. For $n = 0$, from the initial condition (2) we have

$$\frac{\partial \Psi(\mathbf{x}, 0)}{\partial t} = \frac{\Psi_1 - \Psi_{-1}}{2\tau} = g(\mathbf{x}) \Rightarrow \Psi_{-1} = \Psi_1 - 2\tau G(\mathbf{x}),$$

in which $G(\mathbf{x}) = [g(\mathbf{x}_1), \ldots, g(\mathbf{x}_N)]^T$ so $(C = I)$

$$((\mu_1 + c_0)C - \frac{1}{2}K)\Psi_1 = (\mu_2 C + \frac{1}{2}K)\Psi_0 + 2c_0\tau G + F_0,$$

and for other $n (n > 0)$ we use Eq. (10).

Example 3.1. In this section, we choose an example to illustrate the validity and applicability of this scheme. To show the accuracy of proposed method, the L_∞ error is considered as follows

$$L_\infty = \max_{1 \le i \le N} |\Psi(\mathbf{x}_i) - \Psi^h(\mathbf{x}_i)|,$$

where $\Psi(\mathbf{x}_i)$ and $\Psi^h(\mathbf{x}_i)$ are the exact and numerical solutions at node $\mathbf{x}_i$.

Suppose the following fractional equation

$$\begin{aligned}\frac{\partial^\alpha u(\mathbf{x}, t)}{\partial t^\alpha} + \frac{\partial^\beta u(\mathbf{x}, t)}{\partial t^\beta} + u(\mathbf{x}, t) &= \nabla u(\mathbf{x}, t) + 2\frac{t^{2-\alpha}}{\Gamma(3-\alpha)} \\ &+ 2\frac{t^\beta}{\Gamma(2-\beta)} + \left(t^2 + x^2 + y^2\right) - 4,\end{aligned}$$

TABLE 1. The L_∞ errors between present method and analytical solutions for Example 3.1 over Ω with different final times.

T	$\alpha = 1.1, \beta = 0.9$	$\alpha = 1.5, \beta = 0.5$	$\alpha = 1.6, \beta = 0.4$	$\alpha = 1.9, \beta = 0.1$
1	$1.1740e-04$	$4.8198e-05$	$3.2552e-04$	$7.2140e-04$
2	$1.5474e-04$	$4.6447e-06$	$1.8688e-04$	$1.1251e-03$
3	$1.8845e-04$	$4.6447e-05$	$1.7790e-04$	$1.4447e-03$
5	$2.4835e-04$	$1.4892e-04$	$3.5802e-04$	$1.9942e-03$

where $\Omega = [0,1] \times [0,1]$ and $\tau = 0.01$. The exact solution is $\Psi(\mathbf{x}, t) = x^2 + y^2 + t^2$. We extract the initial and boundary conditions from the exact solution. We choose 31×31 uniform nodes in Ω. The first column of Table 1 is final time and other columns of this table are L_∞ errors of the presented method for different values of α and β for solving this example. This table reveals that our scheme is accurate and efficient for solving this problem.

4. Conclusion

In this scheme, to apply the essential boundary conditions automatically, we used moving Kriging interpolation in this method. Since the moving Kriging interpolation shape functions have Kronecker's delta properties. Also, we used finite difference relations to approximate the time fraction derivative order. We indicated the capability and accuracy of this method by an example.

References

1. A. Habibirad, E. Hesameddini and A. Taleei, *An efficient meshless method for solving multi-dimensional nonlinear Schrodinger equation*, Iran J. Sci. Technol. Trans. Sci. **44** (2020) 749–761.
2. Y. Shekari, A. Tayebi and M. H. Heydari, *A meshfree approach for solving 2D variable-order fractional nonlinear diffusion-wave equation*, Comput. Methods Appl. Mech. Eng. **350** (2019) 154–168.
3. A. Habibirad, E. Hesameddini, M. H. Heydari and R. Roohi, *An efficient meshless method based on the moving kriging interpolation for two-dimensional variable-order time fractional mobile/immobile advection-diffusion model*, Math. Methods Appl. Sci. (2020). DOI: 10.1002/mma.6759
4. L. Chen and K. M. Liew, *A local petrov-galerkin approach with moving kriging interpolation for solving transient heat conduction problems*, Comput. Mech. **47** (2011) 455–467.
5. L. Gu, *Moving kriging interpolation and element-free galerkin method*, Int. J. Numer. Methods Eng. **56** (1) (2003) 1–11.
6. Z. Baojing and D. Baodong, *A meshless local moving kriging method for two-dimensional solids*, Appl. Math. Comput. **218** (2) (2011) 563–573.

E-mail: a.habibirad@sutech.ac.ir
E-mail: hesameddini@sutech.ac.ir

The Fragile Points Method (FPM) for Solution of the Two-Dimensional Wave Equation Using Point Stiffness Matrices

Donya Haghighi*
Department of Applied Mathematics, Imam Khomeini International University, Qazvin, Iran
and Saeid Abbasbandy
Department of Applied Mathematics, Imam Khomeini International University, Qazvin, Iran

Abstract. In this paper, the Fragile Points Method (FPM) is presented for the numerical solution of Wave Equation. The generalized finite difference method has been applied to achieve the test and trial functions that these functions are discontinuous polynomials. Interior Penalty Numerical Fluxes (IPNF) has been proposed to establish the consistency of the method. Finally, numerical results are provided.
Keywords: Fragile Points Method, Interior Penalty Numerical Fluxes, Wave equation.
AMS Mathematical Subject Classification [2010]: 35L05, 65M99, 68W25.

1. Introduction

Many physical problems can be expressed as mathematical models and these models also include partial differential equations (PDEs). In many problems, numerical analysis researchers use the finite element method (FEM) [2] , Finite Volume Method (FVM) [5] and Boundary Element Method (BEM) [7] to discretize the spatial dimension. Another group of numerical methods known as meshless methods that do not require mesh for discretization such as Element Free Galerkin (EFG) [4] and Meshless Local Petrov-Galerkin (MLPG) [3]. In these methods, the test and trial functions must be continuous throughout the domain, and for this purpose, Radial Basis Function (RBF) and Moving Least Squares (MLS) approximations commonly used. Attaining trial functions by these methods cause a lot of complexity in trial functions and make it difficult to apply boundary conditions and weak form integrations. Hence by analyzing and comparing the other numerical methods, Leiting Dong and colleagues to provide a general method for solving extreme problems so that the test and trial functions are simple, local, and discontinuous polynomials introduced a new meshless method called the Fragile Points Method (FPM) [6]. In FPM, the generalized finite difference method, or the differential quadrature method or the compactly supported used to achieve test and trial functions that are discontinuous polynomials. Due to the discontinuities in these polynomials, FPM may be inconsistent. For this reason, we use

*Speaker

Numerical Flux Corrections [1]. Then we achieve a matrix of coefficients that is symmetric and sparse. This matrix is as a sum of Point Stiffness.

In this paper, wave equation

$$\frac{\partial^2 u}{\partial t^2}(\mathbf{x},t) = \alpha^2 \nabla^2 u(\mathbf{x},t) + f(\mathbf{x},t) \quad \mathbf{x} \in \Omega, \tag{1}$$

with initial conditions

$$u(\mathbf{x},0) = g_1(\mathbf{x}), \qquad \frac{\partial u}{\partial t}(\mathbf{x},0) = g_2, \tag{2}$$

and the boundary conditions

$$u(\mathbf{x},t) = h_1(\mathbf{x},t), \qquad \mathbf{x} \in \Gamma_D, \tag{3}$$

$$\nabla u.n(\mathbf{x},t) = h_2(\mathbf{x},t), \qquad \mathbf{x} \in \Gamma_N, \tag{4}$$

will be studied by Fragile Points Method (FPM) and by mentioning some example and related curves, the accuracy, and stability of the method are checked. Boundaries Γ_D (Dirichlet) and Γ_N (Neumann) satisfy that $\partial\Omega = \Gamma_D \cup \Gamma_N$ and $\Gamma_D \cap \Gamma_N = \emptyset$; n is the unit outward normal of $\partial\Omega$.

2. Main Results

2.1. Polynomial Discontinuous Trial and Test Functions. Inside the domain Ω and it's boundary $\partial\Omega$, several points are distributed sporadically. Using these points, the domain is divided into subdomains that have nothing in common, and each subdomain consists of only one point. The Voronoi Diagram method has been selected for the partition of the domain. In each subdomain of Ω, trial function can be defined according to the values u and its gradient at the internal point. For example, the trial function u_h in the subdomain E_0 which includes the point P_0 can be written as

$$u_h(\mathbf{x},t) = u_0(\mathbf{x},t) + (\mathbf{x}-\mathbf{x}_0)\nabla u(\mathbf{x},t)|_{P_0}, \qquad \mathbf{x} \in E_0. \tag{5}$$

In the above equation, u_0 is the value of u_h at P_0 and $\mathbf{x}_0$ denotes the coordinate of the point P_0.

The gradient of ∇u at P_0 is the yet unknown. We employ the Generalized Finite Difference (GFD) method to calculate ∇u at P_0 in terms of the values of u_h at several neighboring points of P_0. We name these neighboring points as $q_1, q_2, \ldots, q_m$. In the following, to calculate the amount of the gradient of ∇u at P_0, we minimize a weighted discrete L^2 norm $\mathbf{J}$ so that

$$\mathbf{J} = \sum_{i=0}^{m} \Big(\nabla u|_{P_0}.(\mathbf{x}_i - \mathbf{x}_0)^T - (u_i - u_0) \Big)^2 w_i,$$

where w_i denotes the value of weight function at q_i, $\mathbf{x}_i$ is the coordinate vector of q_i, and u_i is the value of u_h at q_i, $(i = 1, 2, \ldots, m)$. For convenience, we assume that w is constant. Due to the stationarity of $\mathbf{J}$ we have

$$\nabla u = \left(A^T A\right)^{-1} A^T \left(\mathbf{u}_m - u_0 \mathbf{I}_m\right), \tag{6}$$

where

$$A=\begin{bmatrix} x_1-x_0 & y_1-y_0 \\ x_2-x_0 & y_2-y_0 \\ \dots & \dots \\ x_m-x_0 & y_m-y_0 \end{bmatrix},\quad \mathbf{u}_m=\begin{bmatrix} u_1 \\ u_2 \\ \dots \\ u_m \end{bmatrix},\quad \mathbf{I}_m=\begin{bmatrix} 1 \\ 1 \\ \dots \\ 1 \end{bmatrix}_{m\times 1}.$$

Also Eq. (6) can be expressed at point P_0 as follows:

$$\nabla u=\mathbf{B}\mathbf{u}_E, \tag{7}$$

where

$$\mathbf{B}=\left(A^TA\right)^{-1}A^T\begin{bmatrix} -1 & 1 & 0 & \dots & 0 \\ -1 & 0 & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ -1 & 0 & \cdots & 0 & 1 \end{bmatrix}_{m\times(m+1)},\qquad \mathbf{u}_E=\begin{bmatrix} u_0 \\ u_1 \\ \vdots \\ u_m \end{bmatrix}.$$

Also by substituting (7) into (5) the relation between u_h and $\mathbf{u}_E$ will be obtained as

$$u_h=\mathbf{N}\mathbf{u}_E,\ \ \forall\mathbf{x}\in E_0,\qquad \mathbf{N}=[\mathbf{x}-\mathbf{x}_0]\mathbf{B}+[1,0,\dots,0]_{1\times(m+1)}.$$

2.2. Implementation of Numerical Flux Corrections. We can rewrite wave Eq. (1)-(4) using mixed form as following,

$$\begin{cases} \sigma=\nabla u(\mathbf{x},t), & in\ \ \Omega, \\ -\alpha^2\nabla.\sigma=-\dfrac{\partial^2u}{\partial t^2}(\mathbf{x},t)+f(\mathbf{x},t), & in\ \ \Omega, \\ u(\mathbf{x},t)=h_1(\mathbf{x},t), & in\ \ \Gamma_D, \\ \sigma.n(\mathbf{x},t)=h_2(\mathbf{x},t), & in\ \ \Gamma_N. \end{cases} \tag{8}$$

By multiplying the first and second equations in (8) by test functions τ and ν respectively and integrating it on the subdomain E, using the Green formula and by summing these equations over all subdomains we have

$$\int_\Omega \sigma_h.\tau d\Omega=-\int_\Omega u_h\nabla.\tau d\Omega+\sum_{E\in\Omega}\int_{\partial E}\hat{u}_h n.\tau d\Gamma, \tag{9}$$

$$\alpha^2\int_\Omega \sigma_h.\nabla\nu=\alpha^2\sum_{E\in\Omega}\int_{\partial E}\hat{\sigma}_h.n\nu d\Gamma-\int_\Omega\frac{\partial^2u}{\partial t^2}(\mathbf{x},t)\nu d\Omega+\int_\Omega f(\mathbf{x},t)\nu d\Omega. \tag{10}$$

In the above equations values $\hat{\sigma}_h$ and $\hat{u}_h$ represent approximations σ_h and u_h on ∂E. These values are named Numerical Fluxes. To simplify Eqs. (9) and (10), we define operators the *average* and the *jump* which by these operators, we can manage the numerical fluxes. As regards $\Gamma=\Gamma_h+\Gamma_D+\Gamma_N$, Table 3.1 in [6] and by substituting the Interior Penalty Numerical Fluxes (IPNF), we have

$$\alpha^2\sum_{E\in\Omega}\int_E\nabla u_h.\nabla\nu d\Omega-\alpha^2\sum_{e\in\Gamma_h\cup\Gamma_D}\int_e\left(\{\nabla u_h\}[\nu]+\{\nabla\nu\}[u_h]\right)d\Gamma+\alpha^2\sum_{e\in\Gamma_h\cup\Gamma_D}\frac{\eta}{h_e}\int_e[\nu][u_h]\,d\Gamma$$

$$=\int_\Omega f(\mathbf{x},t)\nu d\Omega+\alpha^2\sum_{e\in\Gamma_D}\int_e\left(\frac{\eta}{h_e}\nu-\nabla\nu.\mathbf{n}\right)h_1(\mathbf{x},t)d\Gamma+\alpha^2\sum_{e\in\Gamma_N}\int_e\nu h_2(\mathbf{x},t)d\Gamma-\int_\Omega\frac{\partial^2u}{\partial t^2}(\mathbf{x},t)\nu d\Omega.$$

The above equation is the formula of FPM, which is called FPM-Primal method. If the matrix form of this method is expressed as follows:

$$\alpha^2 \mathbf{K}\mathbf{u} + \mathbf{C}\ddot{\mathbf{u}} = \mathbf{F}. \tag{11}$$

By Substituting values $\mathbf{B}$ instead of $\nabla\nu$ and ∇u, $\mathbf{N}$ instead of u_h and ν in Eq. (11), the point stiffness matrices will be achieved as follows:

$$\begin{aligned} \mathbf{C} &= \int_E \mathbf{N}^T\mathbf{N} d\Omega, \qquad E \in \Omega, \\ \mathbf{K}_E &= \int_E \mathbf{B}^T\mathbf{B}\, d\Omega, \qquad E \in \Omega, \end{aligned}$$

$$\begin{aligned} \mathbf{K}_h &= \frac{-1}{2}\int_e (\mathbf{B}_1^T\mathbf{n}_1^T\mathbf{N}_1 + \mathbf{N}_1^T\mathbf{n}_1\mathbf{B}_1)d\Gamma + \frac{\eta}{h_e}\int_e \mathbf{N}_1^T\mathbf{N}_1 d\Gamma \\ &+\frac{-1}{2}\int_e (\mathbf{B}_2^T\mathbf{n}_2^T\mathbf{N}_2 + \mathbf{N}_2^T\mathbf{n}_2\mathbf{B}_2)d\Gamma + \frac{\eta}{h_e}\int_e \mathbf{N}_2^T\mathbf{N}_2 d\Gamma \\ &+\frac{-1}{2}\int_e (\mathbf{B}_2^T\mathbf{n}_1^T\mathbf{N}_1 + \mathbf{N}_2^T\mathbf{n}_2\mathbf{B}_1)d\Gamma + \frac{\eta}{h_e}\int_e \mathbf{N}_1^T\mathbf{N}_2 d\Gamma \\ &+\frac{-1}{2}\int_e (\mathbf{B}_1^T\mathbf{n}_2^T\mathbf{N}_2 + \mathbf{N}_1^T\mathbf{n}_1\mathbf{B}_2)d\Gamma + \frac{\eta}{h_e}\int_e \mathbf{N}_2^T\mathbf{N}_1 d\Gamma, \quad e \in \partial E_1 \cap \partial E_2, \\ \mathbf{K}_D &= -\int_e (\mathbf{B}^T\mathbf{n}^T\mathbf{N} + \mathbf{N}^T\mathbf{n}\mathbf{B})d\Gamma + \frac{\eta}{h_e}\int_e \mathbf{N}^T\mathbf{N} d\Gamma, \qquad e \in \Gamma_D, \end{aligned}$$

and we can also be written

$$\begin{aligned} F_E &= \int_E \mathbf{N}^T f(\mathbf{x},t) d\Omega, \qquad E \in \Omega, \\ F_N &= \int_e \mathbf{N}^T h_2(\mathbf{x},t) d\Gamma, \qquad e \in \Gamma_N, \\ F_D &= \int_e (\frac{\eta}{h_e}\mathbf{N}^T - \mathbf{B}^T\mathbf{n}) h_1(\mathbf{x},t) d\Gamma, \qquad e \in \Gamma_D. \end{aligned}$$

We also use finite difference schemes to deal with the time derivative in the equation. For this purpose, we can write Eq. (11) as follows:

$$\alpha^2 \mathbf{K}(\frac{\mathbf{u}^{n+2} + \mathbf{u}^{n+1}}{2}) + \mathbf{C}(\frac{\mathbf{u}^{n+2} - 2\mathbf{u}^{n+1} + \mathbf{u}^n}{\Delta t^2}) = \mathbf{F}^{(\frac{n+1}{2})}.$$

In the system of the above equations, the expression of right side means the calculation of F_E, F_N and F_D in the average time of steps n and $n-1$.

2.3. Numerical Examples. In this section, some examples to verify the accuracy and efficiency of the present method will be evaluated. The relative errors used in this section are defined as follows:

$$r_0 = \frac{\|u^h - u\|_{L^2}}{\|u\|_{L^2}}, \qquad r_1 = \frac{\|\nabla u^h - \nabla u\|_{L^2}}{\|\nabla u\|_{L^2}}.$$

EXAMPLE 2.1. Consider the two-dimensional inhomogeneous wave equation as follows

$$u_{tt}(x,y,t) = \nabla^2 u(x,y,t) + \cos(x), \quad x \in (0,\pi), \quad y \in (0,\pi), \quad t > 0.$$

Dirichlet boundary conditions $u(0,y,t) = 1 + \sin(y)\sin(t)$, $u(\pi,y,t) = -1 + \sin(y)\sin(t)$,$u(x,0,t) = \cos(x)$, $u(x,\pi,t) = \cos(x)$ and initial condition $u(x,y,0) = \cos(x)$ and $u_t(x,y,0) = \sin(y)$ and analytical solution can be expressed as $u(x,y,t) = \cos(x) + \sin(y)\sin(t)$. Relative errors r_0 and r_1 have been shown in Table 1.

TABLE 1. The relative errors of the method for Example 2.1 at $T = 1$ and $\Delta t = 0.01$.

Points	Computational Parameters	r_0	r_1	CPU Time
$N = 121$	$h_e = 1$, $\eta = 6$	1.7263×10^{-2}	9.5218×10^{-2}	$1.6s$
$N = 676$	$h_e = 0.1$, $\eta = 3$	4.4831×10^{-3}	2.7873×10^{-2}	$11s$
$N = 2601$	$h_e = 0.1$, $\eta = 3$	2.7653×10^{-3}	1.5448×10^{-2}	$91s$

EXAMPLE 2.2. In the previous example, we consider the boundary conditions as $\nabla u.n(x,0,t) = \nabla u.n(x,\pi,t) = -\sin(t)$. The Figures 1 and 2 show the accuracy of this method for this example.

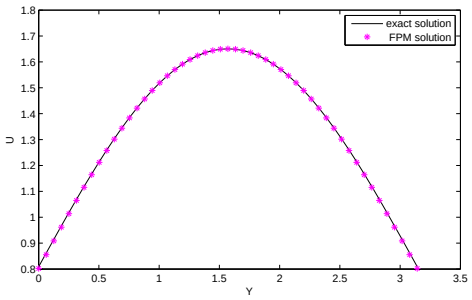


FIGURE 1. Numerical and exact solutions related to Example 2.2 for $dt = 0.01$, $Nx = Ny = 51$ ($N = 2601$), $T = 1$ and $x = 0.6283$.

According to the results of the table and comparison of the curves obtained by FPM with the exact curves, we can be seen that the method is stable and has good precision. Also, the method does not have much computational cost and depending on the number of points used, it will achieve numerical solutions with good accuracy in a short time that this is an advantage over finite element methods. Also in the finite element methods if the element is highly distorted, inaccuracy can occur.

Compared to meshless methods, we can also point out the advantage that FPM uses test and test functions that are in the form of simple and discontinuous polynomials. Therefore the computation of integrals in the weak forms will be easier. Other advantages of this method over other numerical methods are described in detail in Table 1 in [8].

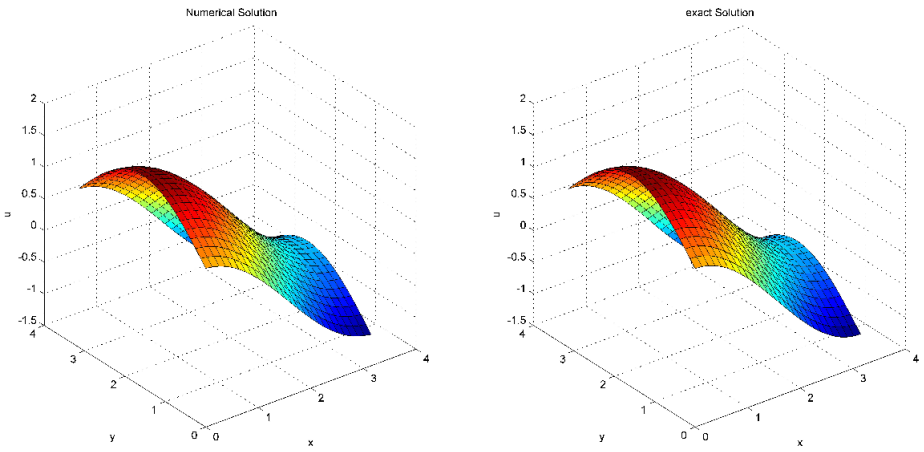


FIGURE 2. Plot of numerical and exact solutions for Example 2.2 so that $Nx = Ny = 26$ ($N = 676$), $dt = 0.01$, $T = 1$.

Therefore, with regards to the capabilities of FPM, it can be apply for types of problems. This method is also suitable for problems with discontinuous domains that these issues will be discussed in our future studies.

References

1. D. N. Arnold, F. Brezzi, B. Cockburn and L. D. Marini, *Unified analysis of discontinuous Galerkin methods for elliptic problems,* SIAM. J. Numer. Anal. **39** (5) (2002) 1749–1779.
2. M. Asadzadeh, *An Introduction to the Finite Element Method (FEM) for Differential Equations,* Chalmers: Lecture notes. 2012.
3. S. N. Atluri and T. Zhu, *A new meshless local Petrov-Galerkin (MLPG) approach in computational mechanics,* Comput. Mech. **22** (2) (1998) 117–127.
4. T. Belytschko, Y. Y. Lu and L. Gu, *Element free galerkin methods,* Int. J. Numer. Methods. Eng. **37** (2) (1994) 229–256.
5. J. C. Chai, H. S. Lee and S. V. Patankar, *Finite volume method for radiation heat transfer,* J. Thermophys. Heat. Trans. **8** (3) (1994) 419–425.
6. L. Dong, T. Yang, K. Wang and S. N. Atluri, *A new Fragile Points Method (FPM) in computational mechanics, based on the concepts of Point Stiffnesses and Numerical Flux Corrections,* Eng. Anal. Bound. Elem. **107** (2019) 124–133.
7. L. C. Wrobel, *The Boundary Element Method,* Applications in Thermo-Fluids and Acoustics, Vol. 1, John Wiley & Sons. 2002.
8. T. Yang, L. Dong and S. N. Atluri, *A simple Galerkin meshless method, the Fragile Points method using point stiffness matrices, for 2D linear elastic problems in complex domains with crack and rupture propagation,* Int. J. Numer. Meth. Eng. (2020). DOI: 10.1002/nme.6540

E-mail: haghighi.donya@edu.ikiu.ac.ir
E-mail: abbasbandy@ikiu.ac.ir

The 51st Annual Iranian Mathematics Conference University of Kashan, 15–20 February 2021

New Positive Definite RBFs via Completely Monotone Functions of Order k

Mohammad Heidari*
Faculty of Mathematical Sciences and Computer, Kharazmi University, Tehran, Iran
and Maryam Mohammadi
Faculty of Mathematical Sciences and Computer, Kharazmi University, Tehran, Iran

Abstract. In this article, we first give a recursive relation for obtaining completely monotone (CM) functions from CM functions of order k. Then the Schoenberg theorem leads to a class of new positive definite RBFs. Numerical results give accurate reconstruction of the Frank's function and original function in the well-known Runge phenomenon.
Keywords: Radial basis functions (RBFs), Interpolation, Completely monotonic functions, Positive definite.
AMS Mathematical Subject Classification [2010]: 65D05, 65D12, 65D20.

1. Introduction

Given a set of n distinct points $\{x_j\}_{j=1}^n \subset \mathbb{R}^d$ and corresponding data values $\{f_j\}_{j=1}^n$, the RBF interpolant is given by $s(x) = \sum_{j=1}^n \lambda_j \phi(\|x - x_j\|)$, where $\phi(r),\ r \geq 0$, is some radial function [4]. The expansion coefficients λ_j are determined from the interpolation conditions $s(x_j) = f_j,\ j = 1, \ldots, n$, which leads to the symmetric linear system $A\lambda = f$, where $A = [\phi(\|x_i - x_j\|)]_{1\leq i,j\leq n}$. A class of functions for which the interpolation problem is uniquely solvable for any distinct point set $\{x_j\}_{j=1}^n$ is the class of positive definite functions.

Definition 1.1. A radial basis function $\phi \in C([0,\infty))$ is called positive definite on $\mathbb{R}^d$ if and only if for any finite set of distinct points $\{x_j\}_{j=1}^n \subset \mathbb{R}^d$, the matrix $A = [\phi(\|x_i - x_j\|)]_{1\leq i,j\leq n}$, is positive definite.

Examples of such RBFs are Guassian $\phi(r) = \exp(-\frac{r^2}{2c^2})$, and Inverse multiquadrics $\phi(r) = (1 + \frac{r^2}{c^2})^{\frac{\beta}{2}},\quad \beta < 0$, where c is a positive factor called shape parameter and can be found numerically for getting accurate numerical solutions and good conditioning of the collocation matrix [1].

2. Construction

We start with a fundamental theorem by Schoenberg [3] as a relation between completely monotone functions and positive definite radial functions.

Definition 2.1. A function g is called completely monotone on $(0,\infty)$ if it satisfies $g \in C^\infty(0,\infty)$ and $(-1)^l g^{(l)}(t) \geq 0,$ for all $l \in \mathbb{N}_0,\ t > 0$. If in addition $g \in C[0,\infty)$ then g is called completely monotone on $[0,\infty)$.

*Speaker

DEFINITION 2.2. A function g is called completely monotone of order k on $(0,\infty)$, if $(-1)^k g^{(k)}(t)$ is completely monotone.

THEOREM 2.3. (Schoenberg) *A non-constant function $g:[0,\infty)\to\mathbb{R}$ is completely monotone on $[0,\infty)$ if and only if $\phi(r)=g(r^2)$ is positive definite on every $\mathbb{R}^d$.*

THEOREM 2.4. *Suppose that f is a function which is completely monotone of order k on $(0,\infty)$, and $a>b>0$. Let*

$$\begin{cases} g_1(x) = -f(x+a)+f(x+b), \\ g_i(x) = -g_{i-1}(x+a)+g_{i-1}(x+b), \quad i=2,\dots,k. \end{cases} \tag{1}$$

Then $\varphi(r)=g_k(r^2)$ is a positive definite radial basis function on every $\mathbb{R}^d$.

PROOF. Let $g_0(x)=f(x)$. We show that the function g_i, $i=1,\dots,k$, is completely monotone of order $k-i$ on $(0,\infty)$. The derivatives of $g_i(x)$ are given as $g_i^{(l)}(x)=-g_{i-1}^{(l)}(x+a)+g_{i-1}^{(l)}(x+b)$. We now deduce for $l\geqslant k-i$

$$\begin{aligned} (-1)^l g_i^{(l)}(x) &= (-1)^l\left(-g_{i-1}^{(l)}(x+a)+g_{i-1}^{(l)}(x+b)\right) \\ &= (-1)^l\left(-\int_{x+a}^{x+b} -g_{i-1}^{(l+1)}(t)dt\right) \\ &= (-1)^{l+1}\int_{x+b}^{x+a} g_{i-1}^{(l+1)}(t)dt \geqslant 0. \end{aligned}$$

The last inequality holds because g_{i-1} is completely monotone of order $k-i+1$. So g_k is completely monotone on $(0,\infty)$. Now, we prove that $g_k\in C[0,\infty)$ by induction on i. Since $f\in C(0,\infty)$, and $a>b>0$, then $g_1\in C[0,\infty)$. Let $g_{i-1}\in C[0,\infty)$, then $g_i\in C[0,\infty)$, which in turn gives $g_k\in C[0,\infty)$. So the theorem is proved according to the Schoenberg theorem. □

EXAMPLE 2.5. Since the funtion $f(x)=-\ln(x)$ is completely monotone of order 1, then the function $\phi(r)=\ln\left(\frac{r^2+a}{r^2+b}\right)$ is a positive definite RBF according to Theorem 2.4. The plots of ϕ is given for different values of a, b and shape parameter c, in Figure 1.

3. Numerical Results

In this section, we use the new positive definite RBF $\phi(\frac{r}{c})=\ln(\frac{(\frac{r}{c})^2+a}{(\frac{r}{c})^2+b})$ for interpolating two different functions. The values of all shape parameters are chosen according to algorithm 2 in [2].

Test Problem 1. (Runge function) Let $f(x)=\dfrac{1}{1+25x^2}$ be the function that we are going to interpolate on the interval $[-1,1]$. We report condition numbers of the coefficient matrices as well as L_2 and L_∞ errors for $M=125$ number of test points and different values of a, b, shape parameter c, and number of center points N, in Table 1. The numerical results show that the proposed RBF is more accurate than the Gaussian, Mutiquadric and Matern kernels. We also plot the

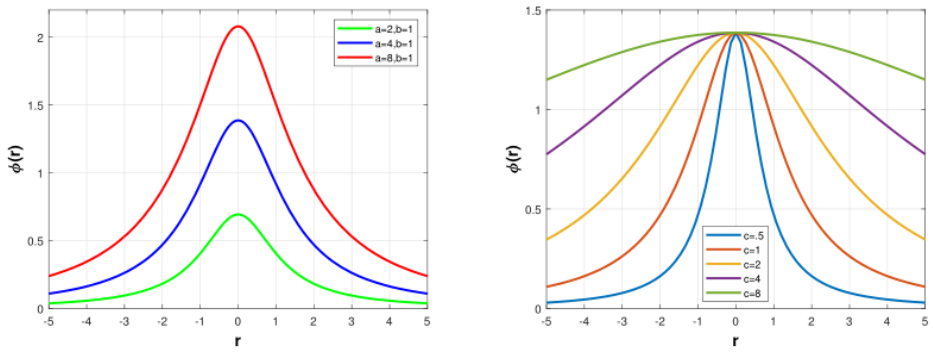


FIGURE 1. Plots of $\phi(r)$ for different values of a and b (left), and $\phi(\frac{r}{c})$ (right) for $a = 4$, $b = 1$ and different values of shape parameter c.

exact and approximate Runge function as well as point-wise error distributions for $a = 8$, $b = 1$, $N = 200$, $M = 125$, and $c = 0.0861$, in Figure 2.

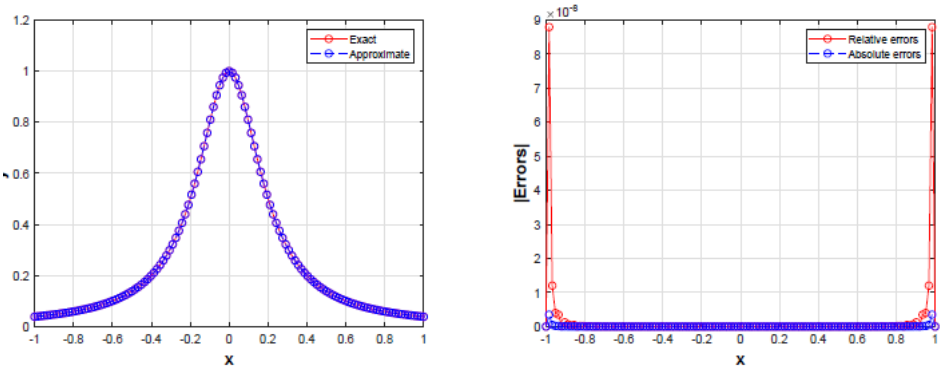


FIGURE 2. Exact and approximate solutions of the Runge function (left), and point-wise error distributions (right), with $N = 200$, $M = 125$, $a = 8$, $b = 1$ and $c = 0.0861$.

Test Problem 2. Consider the interpolation of Frankes function on $[0, 1]^2$. We report condition numbers of the coefficient matrices as well as L_2 and L_∞ errors for $M = 2601$ number of test points and different values of a, b, shape parameter c, and number of center points N, in Table 2. The numerical results show that the proposed RBF is more accurate than the Gaussian and Matern kernels and is in agreement with the Mutiquadric kernel. We also plot the exact and approximate Frank's function in Figures 3(a)-3(b) and point-wise error distributions in Figures 3(c)-3(d) for $a = 8$, $b = 1$, $N = 400$, $M = 2601$, and $c = 0.3046$, respectively.

TABLE 1. The comparison of L_2, L_∞ error, and condition number of new RBF. (Test problem 1)

Kernel	N	c	L_2 error	L_∞ error	Condition number
New kernel $a = 4$, $b = 1$	80	0.2232	5.9168e-08	3.9773e-08	9.9202e+12
	100	0.1776	1.8027e-08	1.1127e-08	9.9917e+12
	200	0.0879	1.1128e-08	7.7784e-09	1.0134e+13
	300	0.0584	8.7215e-09	6.1204e-09	1.0141e+13
	400	0.0438	3.7796e-09	2.5488e-09	1.0649e+13
New kernel $a = 8$, $b = 1$	80	0.2190	7.8833e-08	5.2942e-08	9.9476e+12
	100	0.1741	3.0312e-08	1.8654e-08	9.8673e+12
	200	0.0861	4.9953e-09	3.4897e-09	1.0031e+13
	300	0.0572	6.2254e-09	4.3668e-09	1.0120e+13
	400	0.0429	2.9899e-09	2.0113e-09	1.0670e+13
Gaussian	80	0.0636	2.2312e-06	1.5152e-06	1.0408e+13
	100	0.0506	1.2482e-06	7.9458e-07	1.0528e+13
	200	0.0251	5.7539e-07	4.0445e-07	1.0993e+13
	300	0.0167	2.2064e-07	1.5560e-07	1.1153e+13
	400	0.0125	6.6761e-08	4.6489e-08	1.0498e+13
Multiquadric	80	0.1864	1.1464e-07	7.6478e-08	9.9632e+12
	100	0.1458	7.2354e-08	4.3566e-08	9.9898e+12
	200	0.0680	3.9285e-08	2.7272e-08	9.9784e+12
	300	0.0435	1.8238e-08	1.2720e-08	1.0006e+13
	400	0.0317	7.0049e-09	4.5729e-09	1.0262e+13
Matérn/Sobolev $\nu = 3/2$	80	55.3201	2.7317e-05	1.8101e-05	9.9985e+12
	100	40.9741	1.0948e-05	7.0236e-06	1.0000e+13
	200	16.1811	6.7350e-07	4.0664e-07	1.0001e+13
	300	9.4141	1.2512e-07	7.8808e-08	1.0000e+13
	400	6.4199	3.8129e-08	2.4786e-08	1.0012e+13

TABLE 2. The comparison of L_2, L_∞ error, and condition number of new RBF. (Test problem 2)

Kernel	N	c	L_2 error	L_∞ error	Condition number
New kernel $a = 4$, $b = 1$	100	0.6829	1.8354e-01	2.5566e-02	1.0028e+13
	225	0.4261	7.2332e-03	1.4854e-03	1.0018e+13
	400	0.3108	3.2966e-04	9.6041e-05	1.0017e+13
	900	0.2020	1.6623e-04	5.6203e-05	1.0075e+13
New kernel $a = 8$, $b = 1$	100	0.6701	1.8091e-01	2.5334e-02	1.0029e+13
	225	0.4179	7.1154e-03	1.4680e-03	1.0006e+13
	400	0.3046	3.1688e-04	9.1339e-05	1.0059e+13
	900	0.1975	1.4947e-04	5.1102e-05	1.0035e+13
Gaussian	100	0.2188	1.8790e-01	2.5430e-02	1.0105e+13
	225	0.1306	9.8565e-03	1.8153e-03	1.0309e+13
	400	0.0939	8.8701e-03	2.3129e-03	1.0137e+13
	900	0.0610	7.8115e-03	2.0663e-03	1.3092e+13
Multiquadric	100	0.6067	1.7560e-01	2.4783e-02	9.9974e+12
	225	0.3718	6.8009e-03	1.4090e-03	9.9783e+12
	400	0.2661	2.5467e-04	5.7233e-05	1.0067e+13
	900	0.1669	9.0262e-05	2.6162e-05	9.9166e+12
Matérn/Sobolev $\nu = 3/2$	100	157.99	1.0134e-01	1.7967e-02	1.0002e+13
	225	81.420	1.5799e-02	3.3855e-03	1.0001e+13
	400	50.789	5.7471e-03	1.7654e-03	1.0002e+13
	900	26.058	1.4982e-03	5.1894e-04	1.0000e+13

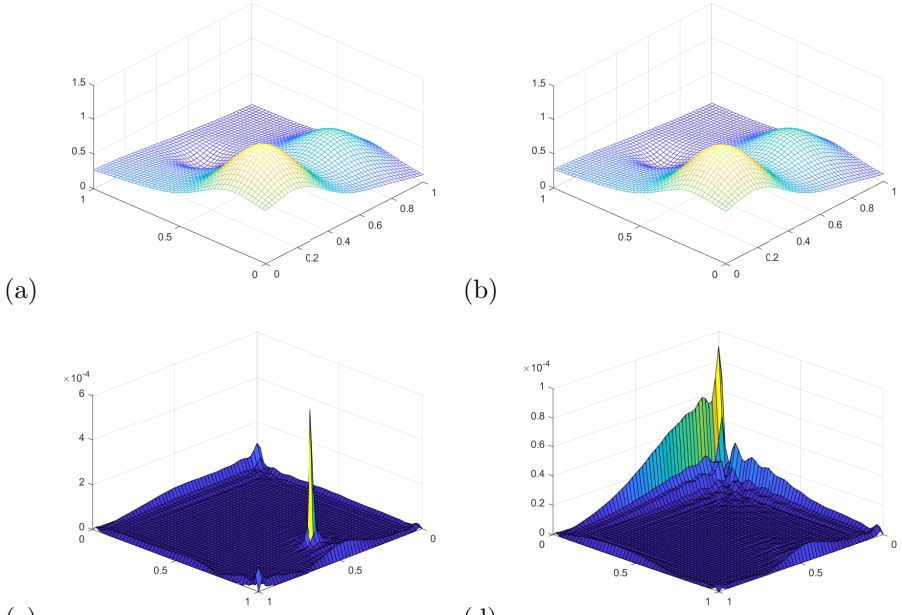


FIGURE 3. Exact (a) and approximate (b) Franke's function; Relative (c) and absolute (d) errors distributions, with $N = 400$, $M = 2601$, $a = 8$, $b = 1$ and $c = 0.3046$.

References

1. M. Bozzini, L. Lenarduzzi, M. Rossini and R. Schaback, *Interpolation by basis functions of different scales and shape*, Calcolo **41** (2) (2004) 77–87.
2. F. Saberi Zafarghandi and M. Mohammadi, *Numerical approximations for the Riesz space fractional advection-dispersion equations via radial basis functions*, Appl. Numer. Math. **144** (2019) 59–82.
3. I. J. Schoenberg, *Metric spaces and positive definite functions*, Trans. Amer. Math. Soc. **44** (1938) 522–536.
4. H. Wendland, *Scattered Data Approximation*, Cambridge University Press, Cambridge, 2005.

E-mail: Std_M.Heidari@khu.ac.ir
E-mail: m.mohammadi@khu.ac.ir

An Efficient Meshfree Machine Learning Approach to Simulate the Generalized Fitzhugh-Nagumo Equation Inspired by Neuroscience

Mohammad Hemami
Department of Computer and Data Sciences, Faculty of Mathematical Sciences, Shahid Beheshti University, Tehran, Iran
Kourosh Parand*
Department of Computer and Data Sciences, Faculty of Mathematical Sciences, Shahid Beheshti University, Tehran, Iran
Institute for Cognitive and Brain Sciences, Shahid Beheshti University, Tehran, Iran
and Jamal Amani Rad
Institute for Cognitive and Brain Sciences, Shahid Beheshti University, Tehran, Iran

Abstract. In this paper, we propose an efficient meshfree least square support vector machine regression approach (LS-SVR) to simulate the generalized Fitzhugh-Nagumo (gFHN) equation in a large spatial domain. By discretizing the problem in time, we turn it into a system of ordinary differential equations and then solve the reformed problem with LS-SVR at each time step. In addition, we have used Richardson extrapolation to increase the accuracy of the problem over time ($\Delta\tau^2$). Numerical results are tested with C^6 Wendland kernels and its comparison with the other numerical solution shows that this approach is highly accurate for solving gFHN types partial differential equations.

Keywords: Meshfree least square support vector machine, Machine learning, Fitzhugh-Nagumo, Partial differential equation, Neuroscience.
AMS Mathematical Subject Classification [2010]: 35Q92, 65M70, 68T05.

1. Introduction

Machine learning methods have been considered by many researchers during the last decade, so that the evolution and development of these methods have left admirable effects in the field of engineering and basic sciences [1]. Support Vector Machine (SVM) is one of the newest and most powerful machine learning tools used to classify and cluster data based on similarities or features, however the high flexibility of this approach and its close relationship with the concept of constrained optimization makes it possible to use this approach for curve fitting and approximating [1]. Generally, the phenomena dynamics are modeled using differential equations, which are analyzed and investigated using optimization approaches [6, 7].

*Speaker

In this paper, we intend to show the numerical simulation by LS-SVR for the generalized Fitzhugh-Nagumo equation with time-dependent coefficients given by

$$u_\tau + v(\tau)u_x - \mu(\tau)u_{xx} - \eta(\tau)u(1-u)(\rho-u) = 0,\ (x,\tau) \in [D_1,\ D_2] \times [0,T], \tag{1}$$

subject to the boundary (bc) and initial (ic) conditions

$$u(D_1,\tau) = h_1(\tau),\ u(D_2,\tau) = h_2(\tau),\ \tau \in [0,\ T], \tag{2}$$

$$u(x,0) = g(x),\ x \in [D_1,\ D_2], \tag{3}$$

where $v(\tau)$, $\mu(\tau)$ and $\eta(\tau)$ are arbitary real-valued functions of τ [5]. It should be noted that considering parameters $v(\tau) = 0$, $\mu(\tau) = 1$ and $\eta(\tau) = -1$, the popular FitzHugh-Nagumo model is obtained, which is widely used in neuroscience, and in recent years, extensive studies have been conducted on this type of model [3, 4].

2. Methodology

We first discretize the Eq. (1) in time using the Crank-Nicolson discretization in the following iterative form

$$u^{n+1} - \frac{\Delta\tau}{2}[\mu^{n+1}u_{xx}^{n+1} + \eta^{n+1}(1-u^n)(\rho-u^n)u^{n+1} - v^{n+1}u_x^{n+1}] \tag{4}$$

$$= u^n + \frac{\Delta\tau}{2}[\mu^n u_{xx}^n + \eta^n(1-u^n)(\rho-u^n)u^n - v^n u_x^n], \tag{5}$$

where n refers to the time step as $n \equiv n\Delta\tau,\ n = 0,\dots,N$ and the first time step $(n=0)$ is obtained from the initial condition of the problem and then the solutions of the problem in the next steps $(n+1)$ is calculated according to the previous step (n) which is known value.

Now, according to the LS-SVR framework, considering the solution of the problem in the primal form $u^{n+1}(x_i) \simeq \sum_{l=2}^{L} w_l^{n+1}\phi(x_i) + b^{n+1} \equiv {w^{n+1}}^T\phi_i + b^{n+1}$, we convert the time-descrete model (4) and (5) to the following Lagrangian form

$$\begin{aligned}\mathcal{L}([w,b,e_i,\alpha_i,\beta_k]^{n+1}) &= \frac{w^Tw}{2} + \frac{\gamma e^Te}{2} \\ &- \sum_{i=2}^{M-1}\alpha_i(w^T[\phi_i - \frac{\Delta\tau}{2}(\mu^{n+1}\phi_i'' + \eta^{n+1}(1-u_i^n)(\rho-u_i^n)\phi_i - v^{n+1}\phi_i')] \\ &+ b - \frac{\Delta\tau}{2}(\eta^{n+1}(1-u_i^n)(\rho-u_i^n)b) - r_i^n - e_i) - \sum_{k=1}^{2}\beta_k(w^T\phi_{*^k} + b - h_k),\end{aligned} \tag{6}$$

where ϕ is basis function, γ is a regularization parameter, $\{\alpha_i\}_{i=1}^M$, $\{\beta_k\}_{k=1}^2$ are Lagrange multipliers, r_i^n is equivalent to the value obtained to the Eq. (5) at the x_i, $\phi_{*^1} = \phi_1$ and $\phi_{*^2} = \phi_M$. Note that in this work we consider M nodal points $x \in [D_1,\ D_2]$ as $x_1, x_2, \dots, x_M$ in domain of problem. Also, to solve the problem in the dual form, we must have the basis dual matrices (kernel matrices) and its derivatives as follows (to investigate theorems of kernel, see [2, 6, 7])

$$\Phi^{(d1)}(s)^T\Phi^{(d2)}(t) = \nabla_{d1}^{d2}[K_{\hat{i}}^{\hat{j}}] = [\phi(s_{\hat{i}})^{(d1)}\phi^{(d2)}(t_{\hat{j}})]_{\hat{i},\hat{j}},\quad d1,\ d2 = 0,1,2.$$

Now, the Karush-Kuhn-Tucker (KKT) optimality conditions for Eq. (6) are as follows

$$\frac{\partial \mathcal{L}}{\partial w} = w = \sum_{i=2}^{M-1} \alpha_i([\phi_i - \frac{\Delta\tau}{2}(\mu^{n+1}\phi_i'' + \eta^{n+1}(1-u_i^n)(\rho-u_i^n)\phi_i - v^{n+1}\phi_i')]) + \sum_{k=1}^{2}\beta_k(\phi_{*^k}),$$

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial b} &= \sum_{i=2}^{M-1} \alpha_i(1 - \frac{\Delta\tau\eta^{n+1}(1-u_i^n)(\rho-u_i^n)}{2}) + \sum_{k=1}^{2}\beta_k = 0,\\
\frac{\partial \mathcal{L}}{\partial e_i} &= e_i = -\frac{\alpha_i}{\gamma},\\
\frac{\partial \mathcal{L}}{\partial \alpha_i} &= w^T[\phi_i - \frac{\Delta\tau}{2}(\mu^{n+1}\phi_i'' + \eta^{n+1}(1-u_i^n)(\rho-u_i^n)\phi_i - v^{n+1}\phi_i')]\\
&+ b - \frac{\Delta\tau}{2}(\eta^{n+1}(1-u_i^n)(\rho-u_i^n)b) - e_i = r_i^n,\\
\frac{\partial \mathcal{L}}{\partial \beta_k} &= w^T\phi_{*^k} + b = h_k.
\end{aligned}$$

After elimination of the primal variables w and $\{e_i\}_{i=2}^{M-1}$ making use of Mercers theorem, the solution is given in the dual form (See [2]), and writing dual form equations in matrix form gives the linear system as following

$$\begin{bmatrix} \nabla_0^0[K_1^1] & \Xi_1 & \nabla_0^0[K_1^M] & 1 \\ \Xi_2 & \Xi_3 & \Xi_4 & \Xi_5 \\ \nabla_0^0[K_M^1] & \Xi_6 & \nabla_0^0[K_M^M] & 1 \\ 1 & \Xi_7 & 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \boldsymbol{\alpha} \\ \beta_2 \\ b \end{bmatrix}^{n+1} = \begin{bmatrix} h_1^{n+1} \\ \mathbf{r}^n \\ h_2^{n+1} \\ 0 \end{bmatrix}, \tag{7}$$

where

$$\begin{aligned}
&\Xi_1 = \bar{A}_2^{M-1} \bullet \nabla_0^0[K_1^{2,\ldots,M-1}] + \bar{B}_2^{M-1} \bullet \nabla_1^0[K_1^{2,\ldots,M-1}]\\
&+\bar{C}_2^{M-1} \bullet \nabla_2^0[K_1^{2,\ldots,M-1}],\\
&\Xi_2 = (\bar{A}_2{}^{M-1})^T \bullet \nabla_0^0[K_{2,\ldots,M-1}^1]\\
&+(\bar{B}_2{}^{M-1})^T \bullet \nabla_0^1[K_{2,\ldots,M-1}^1] + (\bar{C}_2{}^{M-1})^T \bullet \nabla_0^2[K_{2,\ldots,M-1}^1],\\
&\Xi_3 = \hat{A}(\nabla_0^0[K_{2,\ldots,M-1}^{2,\ldots,M-1}]\hat{A} + \nabla_1^0[K_{2,\ldots,M-1}^{2,\ldots,M-1}]\hat{B} + \nabla_2^0[K_{2,\ldots,M-1}^{2,\ldots,M-1}]\hat{C})\\
&+\hat{B}(\nabla_0^1[K_{2,\ldots,M-1}^{2,\ldots,M-1}]\hat{A} + \nabla_1^1[K_{2,\ldots,M-1}^{2,\ldots,M-1}]\hat{B}\\
&+\nabla_2^1[K_{2,\ldots,M-1}^{2,\ldots,M-1}]\hat{C})\\
&+\hat{C}(\nabla_0^2[K_{2,\ldots,M-1}^{2,\ldots,M-1}]\hat{A} + \nabla_1^2[K_{2,\ldots,M-1}^{2,\ldots,M-1}]\hat{B}\\
&+\nabla_2^2[K_{2,\ldots,M-1}^{2,\ldots,M-1}]\hat{C}) + \mathbf{I}/\gamma,\\
&\Xi_4 = (\bar{A}_2{}^{M-1})^T \bullet \nabla_0^0[K_{2,\ldots,M-1}^M] + (\bar{B}_2{}^{M-1})^T \bullet \nabla_0^1[K_{2,\ldots,M-1}^M]\\
&+(\bar{C}_2{}^{M-1})^T \bullet \nabla_0^2[K_{2,\ldots,M-1}^M],\\
&\Xi_5 = (\bar{A}_2{}^{M-1})^T,\\
&\Xi_6 = \bar{A}_2^{M-1} \bullet \nabla_0^0[K_M^{2,\ldots,M-1}] + \bar{B}_2^{M-1} \bullet \nabla_1^0[K_M^{2,\ldots,M-1}]\\
&+\bar{C}_2^{M-1} \bullet \nabla_2^0[K_M^{2,\ldots,M-1}],\\
&\Xi_7 = \bar{A}_2{}^{M-1},\\
&\boldsymbol{\alpha} = [\alpha_2, \ldots, \alpha_{M-1}]^T, \qquad \mathbf{r}^n = [r_2^n, \ldots, r_{M-1}^n]^T,
\end{aligned}$$

$$\bar{A}_2^{M-1} = 1 - \eta^{n+1}\Delta\tau/2(-\rho + (1+\rho) \bullet [u^n_{2,\dots,M-1}] - [u^n_{2,\dots,M-1}] \bullet [u^n_{2,\dots,M-1}]),$$
$$\bar{B}_2^{M-1} = \Delta\tau v^{n+1}\mathbf{I}/2, \qquad \bar{C}_2^{M-1} = -\Delta\tau\mu^{n+1}\mathbf{I}/2,$$
$$\hat{A} = diag(\bar{A}_2^{M-1}), \qquad \hat{B} = diag(\bar{B}_2^{M-1}), \qquad \hat{C} = diag(\bar{C}_2^{M-1}),$$

in which $\mathbf{I}$ is an $M-2 \times M-2$ identity matrix and symbol " $\bullet$ " denoted Hadamard product. By solving the system (7) and obtaining the unknown coefficients $[\beta_1, \boldsymbol{\alpha}, \beta_2, b]$ at each time step $n+1$, we obtain an approximation of the function $u(x)$ in the time step $n+1$. To increase the accuracy of the solution, we use Richardson extrapolation to approximate the final solution, that is, we solve the problem once with the N time steps $(u_1(x)^N)$ and again with the $2N$ time steps $(u_2(x)^{2N})$, and consider the approximation as follows

$$\widehat{u}(x,\tau) \approx 2u_2(x)^{2N} - u_1(x)^N.$$

3. Results

To evaluate the efficiency of the method, we consider two different examples with specific parameters and the following conditions

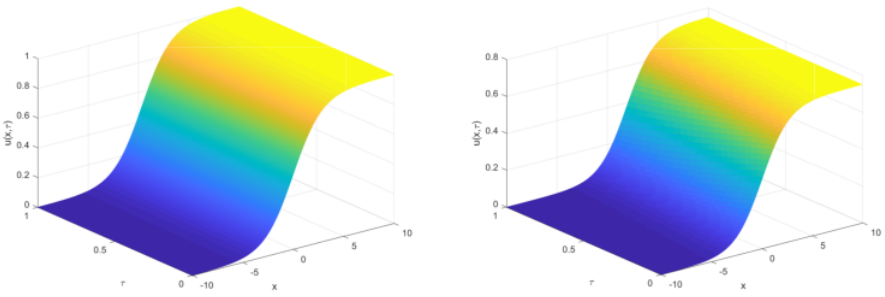
Example 3.1.

$$ic : 0.5 + 0.5\tanh(\frac{x}{2\sqrt{2}}),\ bc : \{0.5 + 0.5\tanh(\frac{1}{2\sqrt{2}})(-10 - \frac{(2\rho-1)\tau}{\sqrt{2}});$$
$$0.5 + 0.5\tanh(\frac{1}{2\sqrt{2}})(10 - \frac{(2\rho-1)\tau}{\sqrt{2}})\}, (x,\tau) \in [-10,\ 10] \times [0,1],$$
$$v = 0, \mu = 1, \eta = -1, \rho = 0.75, \gamma = 10^{10},$$

Example 3.2.

$$ic : \frac{\rho}{2} + \frac{\rho}{2}\tanh(\frac{\rho x}{2}),\ bc : \{\frac{\rho}{2} + \frac{\rho}{2}\tanh(\frac{\rho}{2}(-20 - (3-\rho)\sin(\tau)));$$
$$\frac{\rho}{2} + \frac{\rho}{2}\tanh(\frac{\rho}{2}(20 - (3-\rho)\sin(\tau)))\}, (x,\tau) \in [-10,\ 10] \times [0,1],$$
$$v = \cos(\tau), \mu = \cos(\tau), \eta = \cos(\tau), \rho = 0.75, \gamma = 10^{10}.$$

In addition, for both examples, we consider 100 nodal points M, and 1000 time steps N and 1000 evaluation points $\underline{N}$ to evaluate the error of method ; Moreover, the exact solutions of the first and second examples are $0.5 + 0.5\tanh(\frac{1}{2\sqrt{2}})(x - \frac{(2\rho-1)\tau}{\sqrt{2}})$ and $\frac{\rho}{2} + \frac{\rho}{2}\tanh(\frac{\rho}{2}(x - (3-\rho)\sin(\tau)))$, respectively. The C^6 Wendland kernel [3] $K(s,t) = (1-\sigma\|s-t\|)^8_+(1+8\sigma\|s-t\|+25\sigma\|s-t\|^2+32\sigma\|s-t\|^3)$ is also used for both examples and the optimal local parameter σ is selected as 0.0001 with try and error. The numerical simulations are carried out on a computer with Core i7 CPU 2.70 GHz 8 GB RAM, and the software programs are run under MATLAB 2017. The result of both examples is shown in Figure 1. The Table 1 shows the L_∞ error ($\max\limits_{1\le j\le M}(\max\limits_{1\le i\le \underline{N}}|u(x_i,\tau_j) - \widehat{u}(x_i,\tau_j)|)$) obtained for 1000 evaluation points and the CPU time for the proposed method. The Polynomial differential quadrature method (PDQM) [5] method is also shown in the Table 1 to compare L_∞ error and CPU time. According to this table, the L_∞ error of the proposed method is better than the PDQM method and the processor time is longer. Of course, we should note that the number of nodal points in the [5] is not specified, and so we do not know how many nodal points the CPU time is calculated for.



(A) Example 1. (B) Example 2.

FIGURE 1. Behavior of the model for both examples.

TABLE 1. Comparison of the L_∞ error and CPU time of the proposed method method with the PDQM [5].

example	presented method (L_∞)	PDQM (L_∞)	presented method (CPU time)	PDQM (CPU time)
1	1.692E-5	7.903E-4	3.8521s	0.24s
2	3.166E-4	6.328E-4	3.4325s	0.27s

4. Concluding Remarks

In the present work, we proposed a meshless method based on the LS-SVR to solve the gFHN equation over a large spatial domain. Richardson extrapolation has also been used to increase the accuracy of the problem. By evaluating two examples and comparing them with PDQM, we showed that the proposed method is efficient and accurate.

References

1. E. Alpaydin, *Introduction to Machine Learning*, 4th ed., MIT Press, Cambridge, Massachusetts, 2020.
2. J. De Brabanter and B. De moor and J. Suykens and J. Vandewalle, *Least Squares Support Vector Machines*, World Scientific Pub. Co., Singapore, 2002.
3. M. Hemami and K. Parand and J. A. Rad, *Numerical simulation of reactiondiffusion neural dynamics models and their synchronization/desynchronization: Application to epileptic seizures*, Comput. Math. Appl. **78** (2019) 3644–3677.
4. M. Hemami and J. A. Rad and K. Parand, *The use of space-splitting RBF-FD technique to simulate the controlled synchronization of neural networks arising from brain activity modeling in epileptic seizures*, J. Comput. Sci. **42** (2020) 3644–3677.
5. R. Jiwari and R. K. Gupta and V. kumar, *Polynomial differential quadrature method for numerical solutions of the generalized FitzhughNagumo equation with time-dependent coefficients*, Ain Shams Eng. J. **5** (2014) 1343–1350.
6. S. Mehrkanoon and T. Falck and J. Suykens, *Approximate solutions to ordinary differential equations using least squares support vector machines*, IEEE Trans. Neural Net. Learn. Sys. **23** (9) (2012) 1356–1367.

7. S. Mehrkanoon and J. Suykens, *Learning solutions to partial differential equations using LS-SVM*, Neurocomputing **159** (2015) 105–116.

E-mail: gaslakh@gmail.com
E-mail: K_parand@sbu.ac.ir
E-mail: J_amanirad@sbu.ac.ir

A Fast Meshless Method for Solving Coupled Nonlinear Advection-Diffusion-Reaction Systems on Irregular Domains

Mohammad Ilati*
Department of Applied Mathematics, Faculty of Basic Sciences, Sahand University of Technology, Tabriz, Iran

Abstract. In this paper, a fast meshless method is proposed for solving coupled nonlinear advection-diffusion-reaction systems on irregular domains. In this method, the Petrov-Galerkin strategy is used to build the primary local weak forms. Based on the generalized moving least squares technique, direct approximations of local weak forms are performed to construct the stiff and mass matrices. The computational efficiency is the most significant advantage of this method in comparison with the original MLPG method. This is because the numerical integrations are performed over polynomials instead of complicated MLS shape functions. The numerical results confirm the good efficiency of this method for solving coupled nonlinear advection-diffusion-reaction systems on irregular domains.

Keywords: Coupled nonlinear advection-diffusion-reaction system, Meshless method, Petrov-Galerkin formulation, Generalized moving least squares approximation.
AMS Mathematical Subject Classification [2010]: 65M99, 65N99.

1. Introduction

Meshless methods have been developed in the past decade for numerical computations of wide ranging engineering problems. These meshless methods do not require any mesh, element or lattice for discretization of problem domains, and they construct the approximate functions only via a set of nodes scattered on the computational domains. These methods can be classified in the two basic category: strong form, weak form. In the global weak form techniques, at first, the governing partial differential equations (PDEs) are transformed to a set of so-called weak-form integral equations. By a numerical integration procedure over the computational domain, these weak-form integral equations are converted to a set of algebraic equations. A set of background cells is required for this integration process and therefore these methods are not truly meshless methods. To avoid the use of global background cells, the meshless local Petrov-Galerkin method have been developed by Atluri and Shen. This method does not use global background cells to evaluate integrals and the integration process is applied on some simple shape, regular and independent sub-domains. In MLPG method, the numerical integrations are applied over moving least squares (MLS) functions and this leads to high computational costs in comparison with the finite elements method (FEM), in which integrations are done over simple polynomials. To overcome this drawback,

*Speaker

the Direct MLPG (DMLPG) method has been proposed in [2]. This technique is an improved version of MLPG and uses the generalized moving least squares (GMLS) approximation [1] instead of MLS. In DMLPG, numerical integrations on Petrov-Galerkin formulation are applied over low-degree polynomials instead of complicated MLS functions. This advantage overcomes the main drawback of meshless weak form techniques and significantly accelerate the procedure.

In this article, we propose MLPG and DMLPG methods for numerical solution of a coupled nonlinear advection-diffusion-reaction system as follows

$$(1)\quad \begin{cases} \frac{\partial u}{\partial t} + \mu(\mathbf{x}) \cdot \nabla u - \nabla \cdot (D(\mathbf{x}) \nabla u) + e_1 w_p f(u,v) = 0, \\ \\ \frac{\partial v}{\partial t} + \mu(\mathbf{x}) \cdot \nabla v - \nabla \cdot (D(\mathbf{x}) \nabla v) + e_2 w_p f(u,v) = 0, \\ \\ \frac{\partial w}{\partial t} + \mu(\mathbf{x}) \cdot \nabla w - \nabla \cdot (D(\mathbf{x}) \nabla w) + e_3 w_p f(u,v) + r(\mathbf{x}) w = 0, \end{cases}$$

where u, v and w denote the concentration of the main ground substance, aqueous solution electrolyte concentration and concentration of microorganism (e.g. bacteria), respectively [3]. The vector $\mu(\mathbf{x}) = (\mu_1(\mathbf{x}), \mu_2(\mathbf{x}))$ is the average linearized groundwater velocity, $D(x)$ is a hydrodynamic diffusion function, w_p is the total concentration of active microorganism and $w = w_p/R_M$ with a positive constant R_M. The nonlinear term is

$$f(u,v) = \frac{u}{K_u + u} . \frac{v}{K_v + v},$$

$e_i (i = 1, 2, 3)$, K_u and K_v are positive constants.

2. MLS and GMLS Methods

In MLS technique the unknown function $u(\mathbf{x})$ is approximated in terms of N scattered points $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \subset \Omega$ as

$$u(\mathbf{x}) \approx \widehat{u}(\mathbf{x}) = \sum_{j=1}^{N} \psi_j(\mathbf{x}) u_j, \qquad \mathbf{x} \in \Omega,$$

where $\psi_j(\mathbf{x})$ are MLS shape functions and are defined as follows:

$$\Psi(\mathbf{x}) := [\psi_1(\mathbf{x}), \ldots, \psi_N(\mathbf{x})] = \mathbf{p}(\mathbf{x}) \left[P^T W P\right]^{-1} P^T W,$$

where

$$\mathbf{p}(\mathbf{x}) = [p_1(\mathbf{x}), p_2(\mathbf{x}), \ldots, p_Q(\mathbf{x})],$$

$$P := \begin{bmatrix} p_1(\mathbf{x}_1) & p_2(\mathbf{x}_1) & \cdots & p_Q(\mathbf{x}_1) \\ p_1(\mathbf{x}_2) & p_2(\mathbf{x}_2) & \cdots & p_Q(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ p_1(\mathbf{x}_N) & p_2(\mathbf{x}_N) & \cdots & p_Q(\mathbf{x}_N) \end{bmatrix},$$

$$W = W(\mathbf{x}) := \operatorname{diag}\{w(\mathbf{x}, \mathbf{x}_j)\} \in \mathbb{R}^{N \times N}.$$

where w is a weight function and $\mathbb{P}_m^d = \operatorname{span}\{p_1, p_2, \ldots, p_Q\}$ is the space of d-variable polynomials of degree at most m.

For a functional λ and an unknown function u, the value of $\lambda[u]$ is approximated by $\lambda[\widehat{u}]$ as follows

$$\lambda[u(\mathbf{x})] \approx \lambda[\widehat{u}(\mathbf{x})] = \sum_{j=1}^{N} \lambda[\psi_j(\mathbf{x})] u_j. \tag{2}$$

In the above equation λ operates on shape functions ψ_j. These shape functions have no closed form and therefore computation of $\lambda[\widehat{u}]$ is a time-consuming task. If the functional λ has a complex form, this would be even more acute .
For overcoming this disadvantage, the GMLS technique [1] directly approximates the functional λ in terms of nodal values $u(\mathbf{x}_j),\ j = 1, 2, \ldots, N$ as follows

$$\lambda[u(\mathbf{x})] \approx \widehat{\lambda}[u(\mathbf{x})] = \sum_{j=1}^{N} a_j(\lambda) u(\mathbf{x}_j), \tag{3}$$

where $a_j(\lambda)$, unknown coefficients related to the functional λ, are produced as follows [2]

$$\mathbf{a}(\lambda) := [a_1(\lambda), a_2(\lambda), \ldots, a_N(\lambda)] = \lambda(\mathbf{p}) \left[P^T W P\right]^{-1} P^T W.$$

In the above equation, it can be seen that the functional λ operates only on low-order polynomials instead of complicated MLS functions. This property overcomes the main disadvantage of MLS-based methods and significantly speeds up the procedure. In the case of complicated forms of functional λ, this issue will be more important.

3. Discretization Process

For interior nodes, the PetrovGalerkin formulation of (1) against a suitable test function can be written as follows

$$\begin{cases} \int\limits_{\Omega_i} \frac{\partial u}{\partial t} \xi_i \, d\mathbf{x} + \int\limits_{\Omega_i} \mu(\mathbf{x}) \cdot \nabla u \, \xi_i \, d\mathbf{x} - \int\limits_{\Omega_i} \nabla \cdot (D(\mathbf{x}) \nabla u) \, \xi_i \, d\mathbf{x} + \int\limits_{\Omega_i} e_1 w_p f(u, v) \, \xi_i \, d\mathbf{x} = 0, \\ \int\limits_{\Omega_i} \frac{\partial v}{\partial t} \xi_i \, d\mathbf{x} + \int\limits_{\Omega_i} \mu(\mathbf{x}) \cdot \nabla v \, \xi_i \, d\mathbf{x} - \int\limits_{\Omega_i} \nabla \cdot (D(\mathbf{x}) \nabla v) \, \xi_i \, d\mathbf{x} + \int\limits_{\Omega_i} e_2 w_p f(u, v) \, \xi_i \, d\mathbf{x} = 0, \\ \int\limits_{\Omega_i} \frac{\partial w}{\partial t} \xi_i \, d\mathbf{x} + \int\limits_{\Omega_i} \mu(\mathbf{x}) \cdot \nabla w \, \xi_i \, d\mathbf{x} - \int\limits_{\Omega_i} \nabla \cdot (D(\mathbf{x}) \nabla w) \, \xi_i \, d\mathbf{x} + \int\limits_{\Omega_i} e_3 w_p f(u, v) \, \xi_i \, d\mathbf{x} \\ + \int\limits_{\Omega_i} r(\mathbf{x}) w \, \xi_i \, d\mathbf{x} = 0. \end{cases}$$

By selecting the Heaviside step function as the test function and employing the divergence theorem, the above local weak forms are transformed to the following simplified form

$$\begin{cases} \int\limits_{\Omega_i} \frac{\partial u}{\partial t} \, d\mathbf{x} + \int\limits_{\Omega_i} \mu(\mathbf{x}) \cdot \nabla u \, d\mathbf{x} - \int\limits_{\partial\Omega_i} D(\mathbf{x}) \nabla u \,.\, \mathbf{n}_i \, d\mathbf{x} + e_1 \int\limits_{\Omega_i} w_p f(u, v) \, d\mathbf{x} = 0, \\ \int\limits_{\Omega_i} \frac{\partial v}{\partial t} \, d\mathbf{x} + \int\limits_{\Omega_i} \mu(\mathbf{x}) \cdot \nabla v \, d\mathbf{x} - \int\limits_{\partial\Omega_i} D(\mathbf{x}) \nabla v \,.\, \mathbf{n}_i \, d\mathbf{x} + e_2 \int\limits_{\Omega_i} w_p f(u, v) \, d\mathbf{x} = 0, \\ \int\limits_{\Omega_i} \frac{\partial w}{\partial t} \, d\mathbf{x} + \int\limits_{\Omega_i} \mu(\mathbf{x}) \cdot \nabla w \, d\mathbf{x} - \int\limits_{\partial\Omega_i} D(\mathbf{x}) \nabla w \,.\, \mathbf{n}_i \, d\mathbf{x} + e_3 \int\limits_{\Omega_i} w_p f(u, v) \, d\mathbf{x} \\ + \int\limits_{\Omega_i} r(\mathbf{x}) w \, d\mathbf{x} = 0. \end{cases} \tag{4}$$

TABLE 1. Comparison of DMLPG and MLPG methods.

Domain	N	DMLPG				MLPG			
		$\|e_u\|_\infty$	$\|e_v\|_\infty$	$\|e_w\|_\infty$	CPU	$\|e_u\|_\infty$	$\|e_v\|_\infty$	$\|e_w\|_\infty$	CPU
Ω_1	1123	5.5473E−4	4.4534E−4	6.2538E−4	1.34	5.3206E−4	4.7612E−4	4.2532E−4	158.22
	3582	4.8756E−5	4.5443E−5	3.6567E−5	8.25	7.1947E−5	5.3408E−5	8.2294E−5	400.34
Ω_2	1324	1.5473E−4	2.6523E−4	1.2538E−4	1.76	3.0683E−4	3.2034E−4	4.2034E−4	165.45
	3867	2.4507E−5	4.0087E−5	4.9808E−5	9.45	5.6509E−5	5.0545E−5	6.9898E−5	434.65

The local sub-domain Ω_i is assumed to be

$$\Omega_i := B(\mathbf{x}_i, \rho) \cap \Omega, \quad \mathbf{x}_i \in X, \quad \rho = ch_{X,\Omega},$$

where X is a set of pairwise different scattered nodes and $h_{X,\Omega}$ is the *fill distance* of the set X, i.e.

$$h := h_{X,\Omega} = \sup_{\mathbf{x}\in\Omega} \min_{\mathbf{x}_j\in X} \|\mathbf{x} - \mathbf{x}_j\|_2 .$$

The vector $\mathbf{n_i}$ is the outward normal to the boundary $\partial\Omega_i$. Applying (2) and (3) for approximating of integrals in Eq. (4) yields MLPG and DMLPG methods, respectively. All other steps of these two methods are similar. Approximating of functionals in Eq. (4) leads to the following first-order system.

$$\begin{cases} \frac{d}{dt}U(t) = F(U,t), \\ \\ U(t^0) = U^0. \end{cases}$$

Many standard methods can be applied for discretizing of time variable. Here, the fourth-order RungeKutta method is applied for solving the above linear first-order system of ODEs.

4. Numerical Results

For an example, the advection-diffusion-reaction nonlinear system (1) is considered with the following exact solutions

$$\begin{cases} u\,(x,y,t) = \exp(-5t)\sin(\pi x)\sin(\pi y)\,, \\ \\ v\,(x,y,t) = \exp(-2t)\sin(\pi x)\sin(\pi y)\,, \\ \\ w\,(x,y,t) = \exp(-3t)\sin(\pi x)\sin(\pi y)\,. \end{cases}$$

on irregular domains demonstrated in Figure 1. The coefficients are considered to be $\mu(\mathbf{x}) = [1,1]$, $D(\mathbf{x}) = 10^{-3}$, $r(\mathbf{x}) = 2$, $R_M = 1$, $K_u = 1$, $K_v = 2$, $e_1 = 0.6$, $e_2 = 0.1$ and $e_3 = 0.8$. We solve this problem with Dirichlet boundary conditions. The quartic spline weight function $w(\mathbf{x}, \mathbf{x}_j)$ is defined as follows

$$w(\mathbf{x}, \mathbf{x}_i) = w(\delta_i = \frac{\|\mathbf{x} - \mathbf{x}_i\|}{r_s}) = \begin{cases} 1 - 6\delta_i^2 + 8\delta_i^3 - 3\delta_i^4, & \delta_i \le 1, \\ 0, & \delta_i > 1, \end{cases}$$

where r_s is the radius of the local support domain. This system is solved by DMLPG and MLPG methods and the L_∞ error norms of the components u, v and w at time $T = 1$ with $\tau = 0.001$ for various values of h are shown in Table 1. Moreover, in Table 1 the CPU times used for construction of coefficient matrix in these methods are compared.

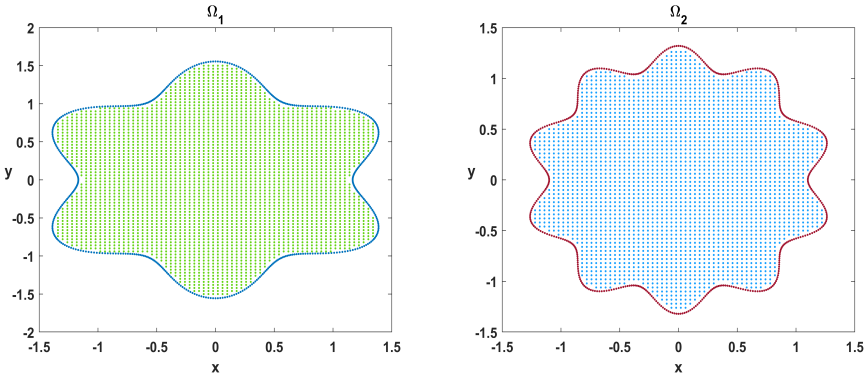


FIGURE 1. Considered irregular computational domains.

References

1. D. Mirzaei, R. Schaback and M. Dehghan, *On generalized moving least squares and diffuse derivatives*, IMA J. Numer. Anal. **32** (3) (2012) 983–1000.
2. D. Mirzaei and R. Schaback, *Direct meshless local Petrov–Galerkin (DMLPG) method: A generalized MLS approximation*, Appl. Nume. Math. **68** (2013) 73–82.
3. W. Liu, J. Huang and X. Long, *Coupled nonlinear advection-diffusion-reaction system for prevention of groundwater contamination by modified upwind finite volume element method*, Comput. Math. Appl. **69** (2015) 477–493.

E-mail: ilati@sut.ac.ir

A Hybrid of Diagonal Preconditioner and Shift-Splitting Method for Double Saddle Point Problems

Mohammad Mahdi Izadkhah*
Department of Computer Science, Faculty of Computer and Industrial Engineering, Birjand University of Technology, Birjand, Iran

Abstract. In this paper, we study a hybrid of diagonal preconditioner and shift-splitting method for numerical solution of double saddle point problems. Theoretical analysis shows that the proposed iterative method is unconditionally convergent. Some numerical results are presented to clarify the effectiveness and accuracy of the presented preconditioner for Krylov subspace method, like GMRES.
Keywords: Saddle point problem, Diagonal preconditioner, Shift-splitting, GMRES.
AMS Mathematical Subject Classification [2010]: 65F08, 65F50, 65N22.

1. Introduction

In recent years, there has been a growing interest to the saddle point problems in the field of numerical linear algebra. This kind of linear systems arise in a great deal of sciences, such as nonlinear constrained optimization, finite element approximation for solving the Navier-Stockes equation, incompressible elasticity, constrained least squares problems, and so forth [2, 3].

Double saddle point problems has been considered as the following large and sparse form

$$\mathcal{A}\mathbf{u} \equiv \begin{pmatrix} A & B & C \\ -B^T & 0 & 0 \\ -C^T & 0 & D \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} \equiv \mathbf{b}, \tag{1}$$

where $A \in \mathbb{R}^{n\times n}$ and $D \in \mathbb{R}^{p\times p}$ are symmetric positive definite(SPD), $B \in \mathbb{R}^{n\times m}$ with $\text{rank}(B) = m < n$, $C \in \mathbb{R}^{n\times p}$, $x, b_1 \in \mathbb{R}^n$, $y, b_2 \in \mathbb{R}^m$ and $z, b_3 \in \mathbb{R}^p$. The following proposition given in [2] provides a necessary and sufficient condition for the invertibility of the matrix $\mathcal{A}$ in the case that the $(1,1)$-block and $(3,3)$-block are both SPD.

Proposition 1.1. *Assume that A and D are symmetric positive definite (SPD). Then matrix $\mathcal{A}$ in (1) is invertible if and only if B has full column rank.*

Both Uzawa-type stationary methods and block preconditioned Krylov subspace methods are discussed in [2] for double saddle point problem (1).

The reminder of this paper is organized as follows. In Section 2, we preset a hybrid of diagonal preconditioner and shift-splitting iteration method. In Section

*Speaker

3, theoretical investigation will provide for the given method. Some numerical results are given in Section 4, to clarify the effectiveness and accuracy of the presented preconditioner for Krylov subspace methods.

2. Hybrid of Diagonal Preconditioner and Shift-Splitting

To establish the properties of preconditioned shift-splitting iterative method, we propose a diagonal preconditioner to system (1) as

$$P = \begin{pmatrix} A & 0 & 0 \\ 0 & Q & 0 \\ 0 & 0 & D \end{pmatrix},$$

where $Q \in \mathbb{R}^{m\times m}$ is a symmetric positive definite matrix. So, we define

$$\bar{\mathcal{A}} = P^{-\frac{1}{2}}\mathcal{A}P^{-\frac{1}{2}} = \begin{pmatrix} I & \bar{B} & \bar{C} \\ -\bar{B}^T & I & 0 \\ -\bar{C}^T & 0 & I \end{pmatrix}, \quad \begin{pmatrix} \bar{x} \\ \bar{y} \\ \bar{z} \end{pmatrix} = P^{\frac{1}{2}} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} A^{\frac{1}{2}}x \\ Q^{\frac{1}{2}}y \\ D^{\frac{1}{2}}z \end{pmatrix},$$

and

$$\begin{pmatrix} \bar{b}_1 \\ \bar{b}_2 \\ \bar{b}_3 \end{pmatrix} = P^{-\frac{1}{2}}\mathbf{b} = \begin{pmatrix} A^{-\frac{1}{2}}b_1 \\ Q^{-\frac{1}{2}}b_2 \\ D^{-\frac{1}{2}}b_3 \end{pmatrix},$$

where $\bar{B} = A^{-\frac{1}{2}}BQ^{-\frac{1}{2}} \in \mathbb{R}^{n\times m}$ also has full column rank, and $\bar{C} = A^{-\frac{1}{2}}CD^{-\frac{1}{2}} \in \mathbb{R}^{n\times p}$. Then, system (1) can be transformed into a new equivalent one

$$\begin{pmatrix} I & \bar{B} & \bar{C} \\ -\bar{B}^T & 0 & 0 \\ -\bar{C}^T & 0 & I \end{pmatrix} \begin{pmatrix} \bar{x} \\ \bar{y} \\ \bar{z} \end{pmatrix} = \begin{pmatrix} \bar{b}_1 \\ \bar{b}_2 \\ \bar{b}_3 \end{pmatrix}. \tag{2}$$

Applying shift-splitting method given in [1, 4] into (2), we obtain

$$\frac{1}{2}\begin{pmatrix} (1+\alpha)I & \bar{B} & \bar{C} \\ -\bar{B}^T & \alpha I & 0 \\ -\bar{C}^T & 0 & (1+\alpha)I \end{pmatrix} \begin{pmatrix} \bar{x}^{k+1} \\ \bar{y}^{k+1} \\ \bar{z}^{k+1} \end{pmatrix} = \frac{1}{2}\begin{pmatrix} (\alpha-1)I & -\bar{B} & -\bar{C} \\ \bar{B}^T & \alpha I & 0 \\ \bar{C}^T & 0 & (\alpha-1)I \end{pmatrix} \begin{pmatrix} \bar{x}^{k} \\ \bar{y}^{k} \\ \bar{z}^{k} \end{pmatrix} + \begin{pmatrix} \bar{b}_1 \\ \bar{b}_2 \\ \bar{b}_3 \end{pmatrix}.$$

It then follows immediately that in the original variables,

$$\frac{1}{2}\begin{pmatrix} (1+\alpha)A & B & C \\ -B^T & \alpha Q & 0 \\ -C^T & 0 & (1+\alpha)D \end{pmatrix} \begin{pmatrix} x^{k+1} \\ y^{k+1} \\ z^{k+1} \end{pmatrix} = \frac{1}{2}\begin{pmatrix} (\alpha-1)A & -B & -C \\ B^T & \alpha Q & 0 \\ C^T & 0 & (\alpha-1)D \end{pmatrix} \begin{pmatrix} x^{k} \\ y^{k} \\ z^{k} \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}.$$

It could be naturally induced a splitting preconditioner $\mathcal{P}_{DPSS}$ for Krylov subspace methods, corresponds to the diagonal preconditioned shift-splitting (DPSS) iteration as

$$\mathcal{P}_{DPSS} = \frac{1}{2}\begin{pmatrix} (1+\alpha)A & B & C \\ -B^T & \alpha Q & 0 \\ -C^T & 0 & (1+\alpha)D \end{pmatrix}.$$

We can do the following matrix factorization for splitting preconditioner $\mathcal{P}_{DPSS}$.

$$\mathcal{P}_{DPSS} = \frac{1}{2}\begin{pmatrix} I & \frac{1}{\alpha}BQ^{-1} & \frac{1}{1+\alpha}CD^{-1} \\ 0 & I & 0 \\ 0 & 0 & I \end{pmatrix} \begin{pmatrix} S & 0 & 0 \\ 0 & \alpha Q & 0 \\ 0 & 0 & (1+\alpha)D \end{pmatrix} \begin{pmatrix} I & 0 & 0 \\ -\frac{1}{\alpha}Q^{-1}B^T & I & 0 \\ -\frac{1}{1+\alpha}D^{-1}C^T & 0 & I \end{pmatrix},$$

where $S = (1+\alpha)A + \frac{1}{\alpha}BQ^{-1}B^T + \frac{1}{1+\alpha}CD^{-1}C^T \in \mathbb{R}^{n\times n}$. At each step of the DPSS iteration or applying the DPSS preconditioner $\mathcal{P}_{DPSS}$ within a Krylov subspace method, we need to solve a linear system as $\mathcal{P}_{DPSS}\mathbf{z}^{(k)} = \mathbf{r}^{(k)}$ for a

given residual vector $\mathbf{r}^{(k)}$ at each step. In the following Algorithm, Let us consider $\mathbf{r}^{(k)} = [r_1^T, r_2^T, r_3^T]^T$ and $\mathbf{z}^{(k)} = [z_1^T, z_2^T, z_3^T]^T$, where $r_1, z_1 \in \mathbb{R}^n$, $r_2, z_2 \in \mathbb{R}^m$ and $r_3, z_3 \in \mathbb{R}^p$.

Algorithm 1. Diagonal Preconditioned Shift-Splitting Iteration Method

(1) Set $k := 0$. Given inital guess $\mathbf{u}^{(0)}$ and $\alpha > 0$. Choose $\epsilon > 0$ as the precision, and k_{max} as the maximum iteration. Set $\mathbf{r}^{(0)} = \mathbf{b} - \mathcal{A}\mathbf{u}^{(0)}$, and its block entries r_1, r_2 and r_3 as defined in advance.
(2) For $\mathbf{u}^{(k)} \in \mathbb{R}^{m+n+p}$, and associated residual $\mathbf{r}^{(k)}$, if $\frac{\|\mathbf{r}^{(k)}\|_2}{\|\mathbf{r}^{(0)}\|_2} \geq \epsilon$ or $k \leq k_{max}$ continue, goto Step 3, else STOP.
(3) Solve $Dw = \frac{2}{1+\alpha} r_3$.
(4) Solve $Qy = r_2$.
(5) $w_1 = 2\left(r_1 - \frac{1}{\alpha} By\right) - Cw$.
(6) Solve $Sz_1 = w_1$.
(7) Solve $Qz_2 = \frac{1}{\alpha}(B^T z_1 + 2r_2)$.
(8) Solve $Dv = \frac{1}{1+\alpha} C^T z_1$.
(9) $z_3 = v + w$.
(10) Let $\mathbf{z}^{(k)} = [z_1^T, z_2^T, z_3^T]^T$, compute $\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} + \mathbf{z}^{(k)}$ and set $k := k+1$, goto Step 2.

Remark 2.1. From Algorithm 1, we can see that at each iteration, it is required to solve linear systems with the coefficient matrices Q, D and $(1+\alpha)A + \frac{1}{\alpha} BQ^{-1}B^T + \frac{1}{1+\alpha} CD^{-1}C^T$. Fortunately, the aforementioned matrices are symmetric positive definite, for all $\alpha > 0$.

Remark 2.2. The symmetric positive definite matrix $Q \in \mathbb{R}^{m\times m}$ should be chosen such that the linear system with coefficient matrix Q is easily solvable, and the singular values of the matrix $A^{-\frac{1}{2}} BQ^{-\frac{1}{2}} \in \mathbb{R}^{n\times m}$ are tightly clustered, or in other words, Q should be a good preconditioner to the matrix $B^T A^{-1} B \in \mathbb{R}^{m\times m}$.

3. Convergence Analysis of DPSS Method

Now, we turn to study the convergence of the diagonal preconditioned shift-splitting iteration method. Note that the iteration matrix of the proposed method is

$$\mathcal{T}_{DPSS} = \begin{pmatrix} (1+\alpha)A & B & C \\ -B^T & \alpha Q & 0 \\ -C^T & 0 & (1+\alpha)D \end{pmatrix}^{-1} \begin{pmatrix} (\alpha-1)A & -B & -C \\ B^T & \alpha Q & 0 \\ C^T & 0 & (\alpha-1)D \end{pmatrix}. \tag{3}$$

Let $\rho(\mathcal{T}_{DPSS})$ denote the spectral radius of $\mathcal{T}_{DPSS}$. To study the convergence of the diagonal preconditioned shift-splitting iteration, we first give the following lemma.

Lemma 3.1. *Let A and D be symmetric positive definite matrices, and B have full row rank. Let $\mathcal{T}_{DPSS}$ be defined as in (3). If λ is an eigenvalue of $\mathcal{T}_{DPSS}$, then $\lambda \neq \pm 1$.*

Next theorem could be proved similar in [4] for the convergence of the diagonal preconditioned shift splitting scheme proposed in this paper.

THEOREM 3.2. *Let $A \in \mathbb{R}^{n\times n}$ and $D \in \mathbb{R}^{p\times p}$ be symmetric positive definite matrices and $B \in \mathbb{R}^{n\times m}$ has full row rank, and let α be a positive number. Then, we have*

$$\rho(\mathcal{T}_{DPSS}) < 1, \quad \text{for all } \alpha > 0.$$

We propose using the Krylov subspace method like GMRES, or its restarted version GMRES(m) to accelerate the convergence of the iteration. It is easy to see that the linear system $\mathcal{A}u = b$ is equivalent to the linear system [5]

$$(I - \mathcal{T}_{DPSS})\mathbf{u} = \mathcal{P}_{DPSS}^{-1}\mathcal{A}\mathbf{u} = \mathcal{P}_{DPSS}^{-1}\mathbf{b}.$$

4. Numerical Results

In practical computations, we use left preconditioning with restarted GMRES($\sharp$) as the Krylov subspace method. Here, the integer $\sharp$ in GMRES($\sharp$) denotes that the algorithm is restarted after every $\sharp$ iterations. In this paper, we take $\sharp = 30$. All runs are started from the initial zero vector and terminated if the current iterations satisfy $ERR = \frac{\|\mathbf{r}^{(k)}\|_2}{\|\mathbf{r}^{(0)}\|_2} \leq 10^{-6}$, or if the prescribed iteration number $k_{\max} = 5000$ is exceeded. All runs are performed in MATLAB R2015a on an Intel Core i3 Laptop with 4G RAM. We consider two cases of the DPSS method as follows:

- Case I: $Q = I_m$,
- Case II: $Q = \beta B^T B$ for $\beta = 0.001$.

EXAMPLE 4.1. By the finite difference scheme of the Stokes problem, the submatrices of the coefficient matrix in the double saddle point problems have the following form

$$A = \begin{pmatrix} I \otimes T + T \otimes I & 0 \\ 0 & I \otimes T + T \otimes I \end{pmatrix} \in \mathbb{R}^{2q^2\times 2q^2}, \quad B = \begin{pmatrix} I \otimes F \\ F \otimes I \end{pmatrix} \in \mathbb{R}^{2q^2\times q^2},$$

$$D = I \otimes T + T \otimes I \in \mathbb{R}^{q^2\times q^2}, \qquad C = \begin{pmatrix} I \otimes F \\ F \otimes I \end{pmatrix} \in \mathbb{R}^{2q^2\times q^2},$$

where $T = \frac{\nu}{h^2}\text{tridiag}(-1, 2, -1) \in \mathbb{R}^{q\times q}$, $F = \frac{1}{h}\text{tridiag}(-1, 1, 0) \in \mathbb{R}^{q\times q}$, with $\otimes$ being the Kronecker product symbol and $h = \frac{1}{q+1}$ the discretization mesh size.

In Tables 1 and 2, we list the numerical results corresponding to the two ν, i.e. $\nu = 0.1, 0.01$. For each ν, three different q are used, i.e. $q = 8, 16, 24$. The parameter α in the $\mathcal{P}_{DPSS}$ is taken the same the viscosity ν. In these tables, $\mathcal{I}$, $\mathcal{P}_{DPSS}$ and $\mathcal{P}_{HSS}$ denote the GMRES(30) method without preconditioning, with the left DPSS preconditioning and with the left HSS preconditioning, respectively. IT, CPU and ERR stand for the iteration numbers, the elapsed CPU times (in seconds) and the relative error, respectively. To demonstrate efficiency of diagonal preconditioned shift-splitting method, the HSS preconditioner is considered as follows

$$\mathcal{P}_{HSS} = \begin{pmatrix} \alpha I + A & 0 & 0 \\ 0 & \alpha I & 0 \\ 0 & 0 & \alpha I + D \end{pmatrix} \begin{pmatrix} \alpha I & B & C \\ -B^T & \alpha I & 0 \\ -C^T & 0 & \alpha I \end{pmatrix}.$$

TABLE 1. Numerical results for solving Example 4.1 with $\nu = 0.1$.

Grid		$\mathcal{I}$	$\mathcal{P}_{DPSS}$		$\mathcal{P}_{HSS}$
			Case I	Case II	
8×8					
	IT	7(6)	1(4)	1(3)	3(29)
	CPU	0.347	0.031	0.023	2.047
	ERR	9.7061e-07	5.4733e-07	5.7328e-07	6.3789e-07
16×16					
	IT	12(21)	1(5)	1(4)	6(25)
	CPU	7.245	1.500	1.400	111.821
	ERR	9.7001e-07	3.6912e-08	4.0800e-08	9.97071e-07
24×24					
	IT	24(27)	1(5)	1(4)	†
	CPU	66.679	12.513	9.156	†
	ERR	9.9045e-07	6.3548e-08	1.7903e-07	†

TABLE 2. Numerical results for solving Example 4.1 with $\nu = 0.01$.

Grid		$\mathcal{I}$	$\mathcal{P}_{DPSS}$		$\mathcal{P}_{HSS}$
			Case I	Case II	
8×8					
	IT	47(26)	1(2)	1(2)	3(29)
	CPU	2.757	0.013	0.014	2.103
	ERR	9.9691e-07	4.3954e-07	3.7612e-09	8.8729e-07
16×16					
	IT	95(21)	1(2)	1(2)	7(3)
	CPU	60.123	0.477	0.461	129.335
	ERR	9.9886e-07	5.3702e-07	7.4961e-09	8.60139e-07
24×24					
	IT	124(18)	1(2)	1(2)	†
	CPU	352.326	4.067	4.205	†
	ERR	9.9951e-07	6.3462e-07	1.2512e-08	†

References

1. Z. Z. Bai, J. F. Yin and Y. F. Su, *A shift-splitting preconditioner for non-Hermitian positive definite matrices*, J. Comput. Math. **24** (2006) 539–552.
2. F. P. A. Beik and M. Benzi, *Iterative methods for double saddle point systems*, SIAM J. Matrix Anal. Appl. **39** (2) (2018) 902–921.
3. M. Benzi and G. H. Golub, *A preconditioner for generalized saddle point problems*, SIAM J. Matrix Anal. Appl. **26** (2004), 20–41.
4. M. M. Izadkhah, *Shift-splitting preconditioners for augmented systems with block* 3×3 *structure*, 48th Annual Iranian. Mathematics Conference, Bu-Ali Sina University, Hamedan, Iran, (2017) pp. 22–25.
5. Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd ed., Society for Industrial and Applied Mathematics, Philadelphia, 2003.

E-mail: izadkhah@birjandut.ac.ir

Computation of the Eigenvalues of the Sturm-Liouville Problem Using the Mittag-Leffler Function

Mohammad Jafari*
Department of Science, Payame Noor University, P. O. BOX 19395-3697, Tehran, Iran

Abstract. In this work, we have presented a method for obtaining the eigenvalues of the Sturm-Liouville fourth order problem using the Mittag-Leffler function and its the integral representation.
Keywords: Mittag-Leffler function, Sturm-Liouville problem, Asymptotic form.
AMS Mathematical Subject Classification [2010]: 26A33, 65Q10.

1. Introduction

The well-known Sturm-Liouville problems with integer derivatives have evolved over two centuries as an interesting and important field of research due to their importance in many areas of science, engineering and mathematics: see [2] and references therein. However, although a huge number of papers and books have been published in this area of research.

The main our purpose are to investigate and discuss on eigenvalues of Sturm-Liouville 4 order using the Mittag-Leffler. Here, we recall some definitions, notations and properties of fractional calculus theory used in this work.

1.1. Laplace Transform.

Definition 1.1. The Laplace transform of a function $f(t)$, defined for all real numbers $t \geq 0$, is the function $F(s)$, which is a unilateral transform defined by

$$\mathscr{L}\{f(t)\} = \int_0^\infty e^{-st} f(t)dt = F(s),$$

where s is a complex number frequency parameter.

Theorem 1.2. [1] *p-Laplace transform of $D_p^\alpha f(t)$ is defined as follows:*

$$\mathscr{L}\{f'(t)\} = sF(s) - f(0).$$

Theorem 1.3. [1] (p-convolution theorem)

$$\mathscr{L}\{f * g\} = \mathscr{L}\{f\}\mathscr{L}\{g\}.$$

Theorem 1.4. [1] *If $f(t)$ is 4-differentiable, then we have*

$$\mathscr{L}\{f^{(4)}(t)\} = s^4F(s) - s^3f(0) - s^2f'(0) - sf''(0) - f'''(0).$$

*Speaker

1.2. Mittage-Leffler Function and Theirs Properties.

DEFINITION 1.5. [3] The 2-parameter Mittage-Leffler is defined for $z, \beta \in \mathbb{C}, \Re(\alpha) > 0$,

$$E_{\alpha,\beta}(z) = \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\alpha k + \beta)}.$$

LEMMA 1.6. [3] $\Re(\alpha > 0)$, *the inverse Laplace transform of some spacial functions are as below:*

$$\mathscr{L}^{-1}\{\frac{s^\alpha}{s(s^\alpha - \lambda)}\} = E_\alpha(\lambda t^\alpha),$$
$$\mathscr{L}^{-1}\{\frac{s^{\alpha-\beta}}{s^\alpha - \lambda}\} = t^{\beta-1}E_{\alpha,\beta}(\lambda t^\alpha),$$
$$\mathscr{L}^{-1}\{\frac{k!s^{\alpha-\beta}}{(s^\alpha - \lambda)^{k+1}}\} = t^{k\alpha+\beta-1}E^{(k)}_{\alpha,\beta}(\lambda t^\alpha),$$

where $\Re(s) > |\lambda|^{\frac{1}{\alpha}}$.

2. Sturm-Liouville Problem

THEOREM 2.1. *Let us consider the SL problem as follows:*

$$y^{(4)}(t)) = \lambda y, \tag{1}$$
$$y(0) = y(1) = y''(0) = y''(1) = 0, \tag{2}$$

where $y \in AC^n[a, b]$. *The eigenvalues of SL problem* (1) *and* (2) *is*

$$\lambda = (n\pi)^4, \ n = 1, 2, \ldots. \tag{3}$$

PROOF. Apply the Laplace transform Theorem 1.4 on (1), we have

$$Y(s) = c_1\frac{s^3}{s^4 - \lambda} + c_2\frac{s^2}{s^4 - \lambda} + c_3\frac{s}{s^4 - \lambda} + c_4\frac{1}{s^4 - \lambda}. \tag{4}$$

Now by Lemma 1.6 on (4) we have

$$y(t) = c_1E_{4,1}\Big(\lambda t^4\Big) + c_2tE_{4,2}\Big(\lambda t^4\Big) + c_3t^2E_{4,3}\Big(\lambda t^4\Big) + c_4t^3E_{4,4}\Big(\lambda t^4\Big).$$

In other hand

$$y''(t) = c_1\lambda t^2E_{4,3}\Big(\lambda t^4\Big) + c_2\lambda t^3E_{4,4}\Big(\lambda t^4\Big) + c_3E_{4,1}\Big(\lambda t^4\Big) + c_4t^2E_{4,2}\Big(\lambda t^4\Big).$$

Finally by imposing the boundary conditions (2) we have

$$\begin{cases} c_2E_{4,2}\Big(\lambda\Big) + c_4E_{4,4}\Big(\lambda\Big) = 0, \\ c_2\lambda E_{4,4}\Big(\lambda\Big) + c_4E_{4,2}\Big(\lambda\Big) = 0. \end{cases}$$

We obtain

$$E^2_{4,2}\Big(\lambda\Big) - \lambda E^2_{4,4}\Big(\lambda\Big) = 0. \tag{5}$$

From the Mittage-Leffler integral representation [4], we have the following relation

$$E_{\alpha,\beta}(z) = \frac{1}{2\pi i}\int_C \frac{s^{\alpha-\beta}}{s^\alpha - z}e^s ds,$$

where C is a loop which starts and ends at $-\infty$ and encircles the circular disc $|t| \leq |z|^{\frac{1}{\alpha}}$ in the positive sense: $-\pi \leq \arg s \leq \pi$.

We see that

$$E_{4,2}\Big(\lambda\Big) = \frac{1}{2\pi i}\int_C \frac{s^2}{s^4 - \lambda}e^s ds.$$

For solving this integral, we use Cauchy's residue theorem.

$$s^4 - \lambda = 0 \Longrightarrow s_k = \Big(\lambda\Big)^{\frac{1}{4}} e^{i(\frac{k\pi}{2})}, \quad k = \ldots, -1, 0, 1, \ldots.$$

Acceptable poles are

$$s_{-1} = -i\Big(\lambda\Big)^{\frac{1}{4}}, \;\; s_0 = \Big(\lambda\Big)^{\frac{1}{4}}, \;\; s_1 = i\Big(\lambda\Big)^{\frac{1}{4}}, \;\; s_2 = -\Big(\lambda\Big)^{\frac{1}{4}}.$$

Thus

$$E_{4,2}\Big(\lambda\Big) = \frac{1}{4}\sum_{i=-1}^{2}\frac{e^{s_i}}{s_i}.$$

After calculations we have

$$E_{4,2}\Big(\lambda\Big) = \frac{1}{\Big(\lambda\Big)^{\frac{1}{4}}}\left\{\sinh\left(\lambda^{\frac{1}{4}}\right) + \sin\left(\lambda^{\frac{1}{4}}\right)\right\}. \tag{6}$$

In similarly on $E_{4,4}\Big(\lambda\Big)$, we obtain

$$E_{4,4}\Big(\lambda\Big) = \frac{1}{\Big(\lambda\Big)^{\frac{3}{4}}}\left\{\sinh\left(\lambda^{\frac{1}{4}}\right) - \sin\left(\lambda^{\frac{1}{4}}\right)\right\}. \tag{7}$$

With substitution (6) and (7) in (5) we get

$$\frac{1}{\Big(\lambda\Big)^{\frac{1}{2}}}\left\{4\sinh\left(\lambda^{\frac{1}{4}}\right)\sin\left(\lambda^{\frac{1}{4}}\right)\right\} = 0,$$

Thus (3) is obtained. □

References

1. M. R. Spiegel, *Theory and Problems of Laplace Transforms*, Schaum's Outline Series, McGraw-Hill, New York, 1965.
2. Q. M. Al-Mdallal, *An efficient method for solving fractional Sturm-Liouville problems*, Chaos Soluitions Fractals **40** (1) (2009) 138–189.
3. G. M. Mittag-Leffler, *Sur la nouvelle function E_α*, C. R. Acad. Sci. Paris (Ser. II) **137** (1903) 554–558.
4. K. B. Oldham and J. Spanier, *The Fractional Calculus*, Academic Press, New York, 1974.

E-mail: jafari536@gmail.com

A New Iterative Method for Solving a Class of Two-by-Two Block Complex Linear Systems

Davod Khojasteh Salkuyeh*
Faculty of Mathematical Sciences, University of Guilan, Rasht, Iran
Center of Excellence for Mathematical Modelling, Optimization and Combinational Computing (MMOCC), University of Guilan, Rasht, Iran

Abstract. We present an iterative method for solving the system arisen from finite element discretization of a distributed optimal control problem with time-periodic parabolic equations. We prove that the method is unconditionally convergent. Numerical results are presented to demonstrate the efficiency of the proposed method.
Keywords: Iterative, Finite element, PDE-constrained, Cptimization, Convergence.
AMS Mathematical Subject Classification [2010]: 49M25, 49K20, 65F10, 65F50.

1. Introduction

Consider the distributed control problems of the form (See [6]):

$$\begin{aligned}\min_{y,u}\ \frac{1}{2}\int_0^T\int_\Omega |y(x,t)-y_d(x,t)|^2dxdt\ +\ \frac{\nu}{2}\int_0^T\int_\Omega |u(x,t)|^2dxdt,\\ \text{s.t. } \frac{\partial}{\partial t}y(x,t)-\Delta y(x,t)\ &=\ u(x,t) \text{ in } Q_T,\\ y(x,t)\ &=\ 0 \text{ on } \Sigma_T,\\ y(x,0)\ &=\ y(x,T) \text{ on } \partial\Omega,\\ u(x,0)\ &=\ u(x,T) \text{ in } \Omega,\end{aligned}$$

where Ω is an open and bounded domain in $\mathbb{R}^d$ ($d\in\{1,2,3\}$) and its boundary $\partial\Omega$ is Lipschitz-continuous. We introduce the space-time cylinder $Q_T=\Omega\times(0,T)$ and its lateral surface $\Sigma_T=\partial Q\times(0,T)$. Here, ν is a regularization parameter, $y_d(x,t)$ is a desired state and $T>0$. We may assume that $y_d(x,t)$ is time-harmonic, i.e., $y_d(x,t)=y_d(x)e^{i\omega t}$, with $\omega=2\pi k/T$ for some $k\in\mathbb{Z}$. Substituting $y_d(x,t)$ in the problem and using finite element discretization of the problem method we get the following system of linear equations

$$\begin{pmatrix} M & 0 & K-i\omega M\\ 0 & \nu M & -M\\ K+i\omega M & -M & 0\end{pmatrix}\begin{pmatrix}\bar{y}\\ \bar{u}\\ \bar{p}\end{pmatrix}=\begin{pmatrix}M\bar{y}_d\\ 0\\ 0\end{pmatrix}, \tag{1}$$

where $M\in\mathbb{R}^{m\times m}$ and $K\in\mathbb{R}^{m\times m}$ are the mass and stiffness matrices, respectively. Both of the matrices M and K are symmetric positive definite (SPD). From

*Speaker

the second equation in (1) we have, $\bar{u} = \bar{p}/\nu$ and by substituting $\bar{u}$ in the third equation, we obtain the following system

$$\begin{cases} M\bar{y} + (K - i\omega M)\bar{p} = M\bar{y}_d, \\ (K + i\omega M)\bar{y} - \frac{1}{\nu}M\bar{p} = 0, \end{cases}$$

which is itself equivalent to

$$Ax = \begin{pmatrix} M & \sqrt{\nu}(K - i\omega M) \\ \sqrt{\nu}(K + i\omega M) & -M \end{pmatrix} \begin{pmatrix} \bar{y} \\ \bar{q} \end{pmatrix} = \begin{pmatrix} \hat{y}_d \\ 0 \end{pmatrix} = b, \tag{2}$$

where $\bar{q} = \bar{p}/\sqrt{\nu}$ and $\hat{y}_d = M\bar{y}_d$.

In [2], Krendl proposed the real block diagonal and the alternative indefinite preconditioners for the system (2). Zheng et al. in [6] proposed the block alternating splitting (BAS) iteration method which can be summarized as

$$\begin{cases} (\alpha V + H_1)x^{k+\frac{1}{2}} &= (\alpha V - S_1)x^k + \mathcal{P}_1 b, \\ (\alpha V + H_2)x^{k+1} &= (\alpha V - S_2)x^{k+\frac{1}{2}} + \mathcal{P}_2 b, \end{cases}$$

where $\alpha > 0$, $V = \text{blkdiag}(M, M)$,

$$H_1 = \text{blkdiag}(M, -M), \qquad H_2 = \text{blkdiag}(\sqrt{\nu}K, \sqrt{\nu}K),$$

$$S_1 = \frac{1}{1+\omega^2\nu}\begin{pmatrix} -i\omega\nu K & \sqrt{\nu}K \\ -\sqrt{\nu}K & i\omega\nu K \end{pmatrix} \quad \text{and} \quad S_2 = \begin{pmatrix} i\sqrt{\nu}\omega M & -M \\ M & -i\sqrt{\nu}\omega M \end{pmatrix}.$$

Numerical results presented in [6] show that the BAS iteration method outperforms the GMRES method [4]. They also showed that the parameter $\alpha = 1 + \nu\omega^2$ often gives quite suitable results. Therefore, if $\nu\omega^2 \ll 1$, then $\alpha = 1$ is a good choice.

In this paper we present a new iterative method for the system (2) and investigate its convergence properties.

2. The New Iterative Method

Using the idea of [5], the system (2) can be written in the 4-by-4 block real system

$$\mathcal{A}\mathbf{x} \equiv \begin{pmatrix} M & 0 & \sqrt{\nu}K & \omega\sqrt{\nu}M \\ 0 & M & -\omega\sqrt{\nu}M & \sqrt{\nu}K \\ \sqrt{\nu}K & -\omega\sqrt{\nu}M & -M & 0 \\ \omega\sqrt{\nu}M & \sqrt{\nu}K & 0 & -M \end{pmatrix} \begin{pmatrix} \Re(\bar{y}) \\ \Im(\bar{y}) \\ \Re(\bar{q}) \\ \Im(\bar{q}) \end{pmatrix} = \begin{pmatrix} \Re(\hat{y}_d) \\ \Im(\hat{y}_d) \\ 0 \\ 0 \end{pmatrix} \equiv \hat{\mathbf{b}}. \tag{3}$$

We define the matrices $\mathcal{G}_1$ and $\mathcal{G}_2$ as following

$$\mathcal{G}_1 = \begin{pmatrix} I & 0 & 0 & \omega\sqrt{\nu}I \\ 0 & I & -\omega\sqrt{\nu}I & 0 \\ 0 & -\omega\sqrt{\nu}I & -I & 0 \\ \omega\sqrt{\nu}I & 0 & 0 & -I \end{pmatrix}, \quad \mathcal{G}_2 = \sqrt{1+\nu\omega^2}\begin{pmatrix} 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \\ I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \end{pmatrix},$$

where $I \in \mathbb{R}^{m\times m}$ is the identity matrix. Then, the system (3) can be equivalently rewrite as

$$(\mathcal{G}_1\mathcal{M} + \frac{\sqrt{\nu}}{\sqrt{1+\nu\omega^2}}\mathcal{G}_2\hat{\mathcal{K}})\mathbf{x} = \hat{\mathbf{b}}. \tag{4}$$

where

$$\mathcal{M} = \begin{pmatrix} M & 0 & 0 & 0 \\ 0 & M & 0 & 0 \\ 0 & 0 & M & 0 \\ 0 & 0 & 0 & M \end{pmatrix} \quad \text{and} \quad \hat{\mathcal{K}} = \begin{pmatrix} K & 0 & 0 & 0 \\ 0 & K & 0 & 0 \\ 0 & 0 & K & 0 \\ 0 & 0 & 0 & K \end{pmatrix}.$$

It is easy to see that $\mathcal{G}_1$ is nonsingular and

$$\mathcal{G}_1^{-1} = \frac{1}{1+\nu\omega^2} \begin{pmatrix} I & 0 & 0 & \omega\sqrt{\nu}I \\ 0 & I & -\omega\sqrt{\nu}I & 0 \\ 0 & -\omega\sqrt{\nu}I & -I & 0 \\ \omega\sqrt{\nu}I & 0 & 0 & -I \end{pmatrix},$$

and

$$\mathcal{G} := \mathcal{G}_1^{-1}\mathcal{G}_2 = \frac{1}{\sqrt{\nu(1+\nu\omega^2)}} \begin{pmatrix} 0 & \omega\nu I & \sqrt{\nu}I & 0 \\ -\omega\nu I & 0 & 0 & \sqrt{\nu}I \\ -\sqrt{\nu}I & 0 & 0 & -\omega\nu I \\ 0 & -\sqrt{\nu}I & \omega\nu I & 0 \end{pmatrix}.$$

Moreover, we have $\mathcal{G}^2 = -I$ with I being the identity matrix of order $4m$, $\mathcal{G}^{-1} = -\mathcal{G}$ and $\mathcal{G}^T = -\mathcal{G}$. Premultiplying both sides of the system (4), gives the system

$$\mathcal{B}\mathbf{x} = (\mathcal{M} + \mathcal{G}\mathcal{K})\mathbf{x} = \mathbf{b}, \tag{5}$$

where $\mathcal{K} = \sqrt{\nu}\hat{\mathcal{K}}/\sqrt{1+\nu\omega^2}$ and $\mathbf{b} = \mathcal{G}_1^{-1}\hat{\mathbf{b}}$. Given $\alpha > 0$, we rewrite Eq. (5) as

$$(\alpha I + \mathcal{M})\mathbf{x} = (\alpha I - \mathcal{G}\mathcal{K})\mathbf{x} + \mathbf{b}. \tag{6}$$

We also rewrite Eq. (5) as

$$\mathcal{G}(\alpha I + \mathcal{K})\mathbf{x} = (\alpha\mathcal{G} - \mathcal{M})\mathbf{x} + \mathbf{b}.$$

Premultiplying both sides of this equation by $\mathcal{G}^{-1}$ and having in mind that $\mathcal{G}^{-1} = -\mathcal{G}$, gives

$$(\alpha I + \mathcal{K})\mathbf{x} = (\alpha I + \mathcal{G}\mathcal{M})\mathbf{x} - \mathcal{G}\mathbf{b}. \tag{7}$$

Now, using Eqs. (6) and (7) we establish the Alternating SPD and Scaled symmetric positive semidefinite splitting (ASSS) method for solving the system (5) as

$$\begin{cases} (\alpha I + \mathcal{M})\mathbf{x}^{(k+\frac{1}{2})} &= (\alpha I - \mathcal{G}\mathcal{K})\mathbf{x}^{(k)} + \mathbf{b}, \\ (\alpha I + \mathcal{K})\mathbf{x}^{(k+1)} &= (\alpha I + \mathcal{G}\mathcal{M})\mathbf{x}^{(k+\frac{1}{2})} - \mathcal{G}\mathbf{b}, \end{cases} \tag{8}$$

where $\mathbf{x}^{(\mathbf{0})}$ is an initial guess. Eliminating $\mathbf{x}^{(k+\frac{1}{2})}$ from Eq. (8) yields the following stationary iterative method

$$\mathbf{x}^{(k+1)} = \mathcal{T}_\alpha \mathbf{x}^{(k)} + \mathbf{f},$$

where

$$\mathcal{T}_\alpha = (\alpha I + \mathcal{K})^{-1}(\alpha I + \mathcal{G}\mathcal{M})(\alpha I + M)^{-1}(\alpha I - \mathcal{G}\mathcal{K}),$$

is the iteration matrix of the ASSS method and $\mathbf{f} = \alpha(\alpha I + \mathcal{K})^{-1}(I - \mathcal{G})(\alpha I + \mathcal{M})^{-1}\mathbf{b}$. On the other hand, if we define

$$\mathcal{P} = \frac{1}{\alpha}(I + \mathcal{G})^{-1}(\alpha I + \mathcal{M})\mathcal{G}(\alpha I + \mathcal{K}), \quad \mathcal{Q} = \frac{1}{\alpha}(I + \mathcal{G})^{-1}(\alpha\mathcal{G} - \mathcal{M})(\alpha I - \mathcal{G}\mathcal{K}),$$

then $\mathcal{B} = \mathcal{P} - \mathcal{Q}$ and $\mathcal{T}_\alpha = \mathcal{P}^{-1}\mathcal{Q}$. Hence, the matrix $\mathcal{P}$ can be used as preconditioner for the system (5). Since $\mathcal{G}^{-1} = -\mathcal{G}$, in the implementation of the preconditioner $\mathcal{P}$ in a Krylov subspace method like GMRES we only need solving two systems with the matrices $\alpha I + \mathcal{K}$ and $\alpha I + \mathcal{M}$ along with two matrix multiplications with the matrices $\mathcal{G}$ and $I + \mathcal{G}$. These systems can be solved exactly using the Cholesky factorization of the matrices $\alpha I + K$ and $\alpha I + M$, or inexactly using the conjugate gradient (CG) method. The following theorem states the convergence of the ASSS method.

THEOREM 2.1. *Assume that the matrices K and M are symmetric positive semidefinite and SPD matrices, respectively. For every $\alpha > 0$, we have*

$$\rho(\mathcal{T}_\alpha) \leq \gamma_\alpha = \max_{\mu \in \sigma(M)} \frac{\sqrt{\alpha^2 + \mu^2}}{\alpha + \mu} < 1,$$

which shows that the ASSS iteration method converges unconditionally. Here, $\sigma(.)$ and $\rho(.)$ denote the spectrum and the spectral radius of a matrix, respectively.

PROOF. Evidently the matrix $\mathcal{T}_\alpha$ is similar to the matrix $R_\alpha \mathcal{S}_\alpha$, where $\mathcal{R}_\alpha = (\alpha I + \mathcal{G}\mathcal{M})(\alpha I + \mathcal{M})^{-1}$ and $\mathcal{S}_\alpha = (\alpha I - \mathcal{G}\mathcal{K})(\alpha I + \mathcal{K})^{-1}$. So we have

$$\rho(\mathcal{T}_\alpha) = \rho(R_\alpha \mathcal{S}_\alpha) \leq \|R_\alpha \mathcal{S}_\alpha\|_2 \leq \|R_\alpha\|_2 \|\mathcal{S}_\alpha\|_2.$$

On the other hand, it follows from $\mathcal{G}^2 = -I$, $\mathcal{G}^T = -\mathcal{G}$ and $\mathcal{M}\mathcal{G} = \mathcal{G}\mathcal{M}$ that

$$\begin{aligned}
\|\mathcal{R}_\alpha\|_2^2 &= \rho\left((\alpha I + \mathcal{M})^{-1}(\alpha I - \mathcal{M}\mathcal{G})(\alpha I + \mathcal{G}\mathcal{M})(\alpha I + \mathcal{M})^{-1}\right) \\
&= \rho((\alpha I + \mathcal{M})^{-1}(\alpha^2 I + \alpha\mathcal{M}\mathcal{G} - \alpha\mathcal{G}\mathcal{M} - \mathcal{M}\mathcal{G}^2\mathcal{M})(\alpha I + \mathcal{M})^{-1}) \\
&= \rho\left((\alpha I + \mathcal{M})^{-2}(\alpha^2 I + \mathcal{M}^2)(\alpha I + \mathcal{M})^{-1}\right) \\
&= \max_{\mu \in \sigma(\mathcal{M})} \frac{\alpha^2 + \mu^2}{(\alpha + \mu)^2} = \max_{\mu \in \sigma(M)} \frac{\alpha^2 + \mu^2}{(\alpha + \mu)^2} < 1. \qquad (9)
\end{aligned}$$

In the same way, we deduce that

$$\|\mathcal{S}_\alpha\|_2^2 = \max_{\lambda \in \sigma(K)} \frac{\alpha^2 + \lambda^2}{(\alpha + \lambda)^2} \leq 1. \qquad (10)$$

Therefore, from equations (9) and (10) we see that $\rho(\mathcal{T}_\alpha) \leq \gamma_\alpha < 1$, which completes the proof. □

Similar to [1, Corollary 2.1] the minimum value of the γ_α (upper bound of the $\rho(\mathcal{T}_\alpha)$) is obtained at $\alpha^* = \sqrt{\mu_{\min}\mu_{\max}}$, where $\mu_{\min}$ and $\mu_{\max}$ are the smallest and largest eigenvalues of the matrix M, respectively.

3. Numerical Results

We consider the distributed control problem in two-dimensional case. The computational domain is the unit square $\Omega = (0,1) \times (0,1) \in \mathbb{R}^2$. The target state is chosen as

$$y_d(x,y) = \begin{cases} (2x-1)^2(2y-1)^2, & \text{if } (x,y) \in (0,\frac{1}{2}) \times (0,\frac{1}{2}), \\ 0, & \text{otherwise.} \end{cases}$$

We present the numerical results for above example. To generate the system (1) we have used the codes of the paper [3] which is available at

`www.numerical.rl.ac.uk/people/rees/`.

TABLE 1. Numerical results for $\nu = 10^{-6}$ and $\omega = 10^3$.

h	Method	α_{opt}	Iters	CPU	Err
2^{-6}	ASSS	0.0003	48	0.30	8.3e-9
	BAS	2.1	77	0.55	9.1e-9
h	Method	α_{opt}	Iters	CPU	Err
2^{-7}	ASSS	0.00003	51	1.66	7.5e-9
	BAS	1.7	76	4.21	9.3e-9

We compare the numerical results of the ASSS method with those of the BAS iterative method. The iteration is terminated as soon as the residual 2-norm is reduced by a factor of 10^8. We always use a zero vector as an initial guess. The inner systems are solved using the Cholesky factorization of the coefficient matrix in conjunction with the symmetric minimum degree reordering using the `symamd.m` command of MATLAB. In both of the methods we have used the optimal value of the parameter α (denoted by α_{opt}) which gives the minimum number of iterations. The value of α^* was computed experimentally. All runs are implemented in MATLAB R2017, equipped with a Laptop with 1.80 GHz central processing unit (Intel(R) Core(TM) i7-4500), 6 GB RAM and Windows 7 operating system.

Numerical results for $\nu = 10^{-6}$, $\omega = 10^3$ and $h = 2^6, 2^7$ have been presented in Table 1. In this table, "Its" and "CPU" stand for the number of iterations and the CPU time (in seconds), respectively. To show the accuracy of the computed solution we have also presented the relative error (denoted by "Err"). As we see, ASSS method outperforms the BAS iteration method.

References

1. Z. Z. Bai, M. Benzi and F. Chen, *Modified HSS iteration methods for a class of complex symmetric linear systems*, Computing **87** (2010) 93–111.
2. W. Krendl, V. Simoncini and W. Zulehner, *Stability estimates and structural spectral properties of saddle point problems*, Numer. Math. **124** (2013) 183–213.
3. T. Rees, H. S. Dollar and A. J. Wathen, *Optimal solvers for PDE-constrained optimization*, SIAM J. Sci. Comput. **32** (2010) 271–298.
4. Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd ed., SIAM, Philadelpha, PA, 2003.
5. M. L. Zeng, *Respectively scaled splitting iteration method for a class of block 4-by-4 linear systems from eddy current electromagnetic problems*, Japan J. Indust. Appl. Math. (2020). DOI:10.1007/s13160-020-00446-8.
6. Z. Zheng, G. F. Zhang and M. Z. Zhu, *A block alternating splitting iteration method for a class of block two-by-two complex linear systems*, J. Comput. Math. Appl. **288** (2015) 203–214.

E-mail: khojasteh@guilan.ac.ir

Higher-Order Bi-CGSTAB and Bi-CRSTAB Algorithms To Solve Some Tensor Equations

Eisa Khosravi Dehdezi*
Department of Mathematics, Persian Gulf University, Bushehr, Iran
and Saeed Karimi
Department of Mathematics, Persian Gulf University, Bushehr, Iran

Abstract. This paper investigates the tensor form of the Bi-CGSTAB and Bi-CRSTAB methods, by employing Kronecker product and vectorization, to solve the generalized coupled Sylvester tensor equations with no matricization. Some numerical examples are provided to compare the efficiency of the proposed methods.
Keywords: Tensor equations, HOBi-CGSTAB, HOBi-CRSTAB, Iterative methods, k-mode product.
AMS Mathematical Subject Classification [2010]: 15A69, 65F10, 65W05.

1. Introduction

In this paper, we are concerned with the generalized coupled Sylvester tensor equations

$$\sum_{j=1}^{n} \mathcal{X}_j \times_1 A_{ij1} \times_2 A_{ij2} \times \cdots \times_d A_{ijd} = \mathcal{E}_i, \qquad i = 1, 2, \ldots, n, \tag{1}$$

where the matrices $A_{ijl} \in \mathbb{C}^{n_{ijl} \times n_{ijl}}$ $(i, j = 1, 2, \ldots, n$ and $l = 1, 2, \ldots, d)$, tensors $\mathcal{E}_i \in \mathbb{C}^{n_{i1} \times \cdots \times n_{id}} (i = 1, 2, \ldots, n)$ are known and $\mathcal{X}_j \in \mathbb{C}^{n_{j1} \times \cdots \times n_{jd}} (j = 1, 2, \ldots, n)$ are unknown tensors and the j-mode product $\times_j$ will be defined later. The (coupled) Sylvester tensor equations often arise from the finite element, finite difference or spectral methods [3]. In [1], Khosravi Dehdezi and Karimi proposed the extended conjugate gradient squared and conjugate residual squared methods for solving (1). In this paper, tensors are written as calligraphic capital letters such as $\mathcal{A}, \mathcal{B}, \ldots$. Let N be a positive integer, an order N real tensor $\mathcal{A} \in \mathbb{R}^{\times I_1 \times \cdots \times I_N}$ is the following multidimensional array

$$\mathcal{A} = (a_{i_1 i_2 \cdots_N})(1 \leq i_j \leq n_j,\ j = 1, 2, \ldots, N),\ a_{i_1 i_2 \cdots_N} \in \mathbb{R},$$

with $H(H = n_1 n_2 \ldots n_N)$ entries [2]. Each entry of $\mathcal{A}$ is denoted by $a_{i_1 i_2 \cdots_N}$. $\mathcal{O}$ with all entries zero denote the zero tensor. With this definition of tensor, matrices are tensors of order two, where signified by capital letters, e.g., A. As usual, $\mathbb{C}$ denotes the complex number field.

Definition 1.1. [2] The operators $\times_k (k = 1, 2, \ldots, n)$ represent the k-mode product of a tensor $\mathcal{X}$ with a matrix $A \in \mathbb{C}^{m \times n_k}$ defined as follows

*Speaker

$$(\mathcal{X} \times_k A)_{i_1 i_2 \cdots k-1 j i_{k+1} \cdots d} = \sum_{i_k=1}^{n_k} x_{i_1 i_2 \cdots k-1 i_k i_{k+1} \cdots d} a_{j i_k}.$$

DEFINITION 1.2. [2] Let N, M be positive integers. The inner product of two tensors $\mathcal{X}, \mathcal{Y} \in \mathbb{C}^{I_1 \times \cdots \times I_N \times J_1 \times \cdots \times J_M}$ is defined by

$$< \mathcal{X}, \mathcal{Y} >= \sum_{j_M=1}^{J_M} \cdots \sum_{j_1=1}^{J_1} \sum_{i_N=1}^{I_N} \cdots \sum_{i_1=1}^{I_1} x_{i_1 \ldots i_N j_1 \ldots j_M} \bar{y}_{j_1 \ldots j_M i_1 \ldots i_N},$$

so the tensor norm that generated by this inner product is

$$\|\mathcal{X}\| = \sqrt{< \mathcal{X}, \mathcal{X} >} = \sqrt{\sum_{j_M=1}^{J_M} \cdots \sum_{j_1=1}^{J_1} \sum_{i_N=1}^{I_N} \cdots \sum_{i_1=1}^{I_1} |x_{i_1 \ldots i_N j_1 \ldots j_M}|^2},$$

which is the tensor Frobenius norm. We say that $\mathcal{X}, \mathcal{Y}$ are orthogonal if $< \mathcal{X}, \mathcal{Y} >= 0$.

We define a new inner product which is needed in the following.

DEFINITION 1.3. Let $H_j, j = 1, 2, \ldots, n$ be the linear space $\mathbb{C}^{n_{j1} \times \cdots \times n_{jd}}, j = 1, 2, \ldots, n$. Define

$$\begin{cases} \mathcal{L} : H_1 \times H_2 \times \cdots \times H_n \to H_1 \times H_2 \times \cdots \times H_n \\ \mathcal{L}(\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n) = \begin{pmatrix} \mathcal{L}_1(\mathcal{X}_j) \\ \mathcal{L}_2(\mathcal{X}_j) \\ \cdots \\ \mathcal{L}_n(\mathcal{X}_j) \end{pmatrix}, \end{cases}$$

where

$$\mathcal{L}_i(\mathcal{X}_j) = \sum_{j=1}^{n} \mathcal{X}_j \times_1 A_{ij1} \times_2 A_{ij2} \times \cdots \times_d A_{ijd}, \ i = 1, 2, \ldots, n.$$

According to this definition, the linear system (1) can be rewritten as

$$\mathcal{L}(\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n) = \mathcal{E}, \ \mathcal{E} = \left(\mathcal{E}_1{}^T, \mathcal{E}_2{}^T, \ldots, \mathcal{E}_n{}^T\right)^T. \tag{2}$$

The remainder of this paper is organized as follows. In Section 2,the higher-order Bi-CGSTAB and Bi-CRSTAB methods are obtained according to tensor form for solving the tensor equations (1). Finally in Section 3, we show comparative results.

2. Higher Order Bi-CGSTAB and Bi-CRSTAB Methods to Solve (1)

Two of the important iterative methods for solving large sparse non-Hermitian linear systems of equations

$$Ax = b, \qquad A \in \mathbb{R}^{n \times n}, x \in \mathbb{R}^n.$$

are the bi-conjugate gradient (Bi-CG) and bi-conjugate residual (Bi-CR) methods based on the non-symmetric Lanczos procedure. Van der Vorst in [4] introduced one of the most successful improvements, a fast and smoothly convergent variant of Bi-CG that avoids calculating the matrix A^*, known as the Bi-CGSTAB algorithm. In exact arithmetic, the Bi-CGSTAB algorithm terminates with a true solution after $j \leq n$ steps [4]. In the following, we present the higher order Bi-CGSTAB

(HOBi-CGSTAB) and higher order Bi-CRSTAB (HOBi-CRSTAB) algorithms to solve the generalized coupled Sylvester tensor Eq. (1). The mode-k matricization of a tensor $\mathcal{X} \in \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_M}$ is denoted by $\mathcal{X}(k)$ and the mode-k fibres are arranged to be the columns of the resulting matrix. The operator "vec" denotes the columns of a matrix or tensor to form a vector. For a matrix $A = (a_1, a_2, \ldots, a_n) = (a_{ij}) \in \mathbb{C}^{m \times n}$ and a matrix B, $A \otimes B = (a_{ij}B)$ is a Kronecker product and $vec(A)$ is a vector defined by $vec(A) = (a_1^T, a_2^T, ..., a_n^T)^T$, where $a_i, 1 \leq i \leq n$ is the i-th column of A and for tensor $\mathcal{X} \in \mathbb{C}^{I_1 \times I_2 \times \cdots \times I_M}$, $\mathcal{X}(1)$ is the mode-1 matricization of the tensor $\mathcal{X}$ [2]. By using the property of the Kronecker product it can be shown that tensor Eq. (1) are equivalent to the following equations

$$\sum_{j=1}^{n} G_{ij} vec(\mathcal{X}_j) = vec(\mathcal{E}_i),\ G_{ij} = A_{ijd} \otimes \cdots \otimes A_{ij2} \otimes A_{ij1},\ i, j = 1, 2, \ldots, n,$$

and $\otimes$ stands Kronecker product.

Thus, the general coupled Sylvester tensor Eq. (1) can be transformed into the following linear system

$$\underbrace{\begin{pmatrix} G_{11} & G_{12} & \ldots & G_{1n} \\ G_{21} & G_{22} & \ldots & G_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ G_{n1} & G_{n2} & \ldots & G_{nn} \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} vec(\mathcal{X}_1) \\ vec(\mathcal{X}_2) \\ \ldots \\ vec(\mathcal{X}_n) \end{pmatrix}}_{x} = \underbrace{\begin{pmatrix} vec(\mathcal{E}_1) \\ vec(\mathcal{E}_2) \\ \ldots \\ vec(\mathcal{E}_n) \end{pmatrix}}_{b}.$$

It is obvious that the size of this linear system is very large even for small values of N and thus using of the Bi-CGSTAB and Bi-CRSTAB algorithms to solve the linear system $Ax = b$ instead of corresponding tensor Eq. (1) will consume much more computer time and memory space as the dimension increases. To overcome this problem, we propose the HOBi-CGSTAB and HOBi-CRSTAB algorithms for solving the tensor Eq. (1). For this purpose, we first provide the common Bi-CGSTAB and Bi-CRSTAB algorithms for solving $Ax = b$ as follows. Let $\mathcal{X}_{i,k}$ and $\mathcal{P}_{i,k} \in \mathbb{C}^{I_1 \times \cdots \times I_N},\ i, j = 1, \ldots, n,\ k = 0, 1, 2, \ldots,$ be the k-th equation solution tensor and k-th search direction tensor, respectively. By taking

$$\mathcal{R}_{i,k} = \mathcal{E}_i - \mathcal{L}_i(\mathcal{X}_{i,k}),\ \mathcal{Q}_{i,k} = \mathcal{L}_i(\mathcal{P}_{j,k}),\ \mathcal{S}_{i,k} = \mathcal{R}_{i,k} - \alpha_k \mathcal{Q}_{i,k},$$

$$\mathcal{W}_{i,k} = \mathcal{L}_i(\mathcal{S}_{j,k}),\ \mathcal{R}_{i,k+1} = \mathcal{S}_{i,k} - \omega_k \mathcal{W}_{i,k},$$

in step k and using variables of Bi-CGSTAB and Bi-CRSTAB algorithms and choosing $\mathcal{R}^*_{i,0} \in \mathbb{C}^{I_1 \times \cdots \times I_N},\ i = 1, 2, \ldots, n,$ arbitrary, the following vectors can be rearranged with the corresponding tensors. Set

$$b = \begin{pmatrix} vec(\mathcal{E}_1) \\ vec(\mathcal{E}_2) \\ \ldots \\ vec(\mathcal{E}_n) \end{pmatrix}, x_k = \begin{pmatrix} vec(\mathcal{X}_{1,k}) \\ vec(\mathcal{X}_{2,k}) \\ \ldots \\ vec(\mathcal{X}_{n,k}) \end{pmatrix}, \tag{3}$$

$$p_k = \begin{pmatrix} vec(\mathcal{P}_{1,k}) \\ vec(\mathcal{P}_{2,k}) \\ \ldots \\ vec(\mathcal{P}_{n,k}) \end{pmatrix}, w_k = \begin{pmatrix} vec(\mathcal{W}_{1,k}) \\ vec(\mathcal{W}_{2,k}) \\ \ldots \\ vec(\mathcal{W}_{n,k}) \end{pmatrix},$$

$$(4)\qquad q_k = \begin{pmatrix} vec(\mathcal{Q}_{1,k}) \\ vec(\mathcal{Q}_{2,k}) \\ \dots \\ vec(\mathcal{Q}_{n,k}) \end{pmatrix}, r_k = \begin{pmatrix} vec(\mathcal{R}_{1,k}) \\ vec(\mathcal{R}_{2,k}) \\ \dots \\ vec(\mathcal{R}_{n,k}) \end{pmatrix}, r_0^* = \begin{pmatrix} vec(\mathcal{R}_{1,0}^*) \\ vec(\mathcal{R}_{2,0}^*) \\ \dots \\ vec(\mathcal{R}_{n,0}^*) \end{pmatrix},$$

where $\mathcal{C}_i, \mathcal{X}_{i,k}, \mathcal{P}_{i,k}, \mathcal{Q}_{i,k}, \mathcal{V}_{i,k}, \mathcal{W}_{i,k}, \mathcal{R}_{i,k}, \mathcal{R}_{i,0}^* \in \mathbb{C}^{n_{i1}\times\cdots\times n_{id}}$ for $i = 1, 2, \dots, n$ and $k = 0, 1, 2, \dots$. By using Eqs. (3) and (4), we have the following relation

$$\begin{aligned} <r_{k+1}, r_0^*> &= \left\langle \begin{pmatrix} vec(\mathcal{R}_{1,k+1}) \\ vec(\mathcal{R}_{2,k+1}) \\ \dots \\ vec(\mathcal{R}_{n,k+1}) \end{pmatrix}, \begin{pmatrix} vec(\mathcal{R}_{1,0}^*) \\ vec(\mathcal{R}_{2,0}^*) \\ \dots \\ vec(\mathcal{R}_{n,0}^*) \end{pmatrix} \right\rangle \\ &= \sum_{i=1}^{n} < vec(\mathcal{R}_{i,k+1}), vec(\mathcal{R}_{i,0}^*) > \\ &= \sum_{i=1}^{n} < \mathcal{R}_{i,k+1}, \mathcal{R}_{i,0}^* > . \end{aligned}$$

and similarly

$$\begin{aligned} < w_k, w_k > &= \sum_{i=1}^{n} < \mathcal{W}_{i,k}, \mathcal{W}_{i,k} >, < q_k, r_0^* > \\ &= \sum_{i=1}^{n} < \mathcal{Q}_{i,k}, \mathcal{R}_{i,0}^* >, < w_k, s_k > \\ &= \sum_{i=1}^{n} < \mathcal{W}_{i,k}, \mathcal{S}_{i,k} > . \end{aligned}$$

Inspired by common Bi-CGSTAB and Bi-CRSTAB, the tensors $\mathcal{P}_{i,k}, \mathcal{Q}_{i,k}$ are auxiliary tensors and $\mathcal{R}_{i,k}, i = 1, 2, \dots, n$ are k-th residual of i-th equation, i.e. $\mathcal{R}_{i,k} = \mathcal{E}_i - \mathcal{L}_i(\mathcal{X}_{j,k}),\ i = 1, 2, \dots, n,\ k = 0, 1, 2, \dots$.

In regard to (2), the k-th residual is as $\mathcal{R}_k = \mathcal{E} - \mathcal{L}(\mathcal{X}_{1,k}, \mathcal{X}_{2,k}, \dots, \mathcal{X}_{n,k})$. Therefore, the residual norm is $||\mathcal{R}_k||_* = \sqrt{\sum_{i=1}^{n} ||\mathcal{R}_{i,k}||^2}$. According to the above discussions the HOBi-CGSTAB and HOBi-CRSTAB algorithms for solving the generalized coupled Sylvester tensor Eq. (1), can be presented as follows.

Algorithm (HOBi-CGSTAB)

Input matrices A_{ijl} and tensors $\mathcal{X}_{j,0}, \mathcal{E}_i$ for $i, j = 1, 2, \dots, n$ and $l = 1, 2, \dots, d$.

(1) Set $\mathcal{P}_{i,0} = \mathcal{R}_{i,0} = \mathcal{E}_i - \mathcal{L}_i(\mathcal{X}_{j,0})$.
(2) Choose arbitrary tensors $\mathcal{R}_{i,0}^*$ such that $\sum_{i=1}^{n} < \mathcal{R}_{i,0}, \mathcal{R}_{i,0}^* > \neq 0$.
(3) For $i = 1, 2, \dots, n$ and $k = 0, 1, \dots,$ until $||\mathcal{R}_k||_*$ small enough Do
(4) $\mathcal{Q}_{i,k} = \mathcal{L}_i(\mathcal{P}_{j,k})$, $\alpha_k = (\sum_{i=1}^{n} < \mathcal{R}_{i,k}, \mathcal{R}_{i,0}^* >)/(\sum_{i=1}^{n} < \mathcal{Q}_{i,k}, \mathcal{R}_{i,0}^* >)$.
(5) $\mathcal{S}_{i,k} = \mathcal{R}_{i,k} - \alpha_k \mathcal{Q}_{i,k}$, $\mathcal{W}_{i,k} = \mathcal{L}_i(\mathcal{S}_{j,k})$,
$\omega_k = (\sum_{i=1}^{n} < \mathcal{W}_{i,k}, \mathcal{S}_{i,k} >)/\sum_{i=1}^{n} < \mathcal{W}_{i,k}, \mathcal{W}_{i,k} > .$
(6) $\mathcal{X}_{i,k+1} = \mathcal{X}_{i,k} + \alpha_k \mathcal{P}_{i,k} + \omega_k S_{i,k}$, $\mathcal{R}_{i,k+1} = \mathcal{S}_{i,k} - \omega_k \mathcal{W}_{i,k}$.
(7) $\beta_k = (\frac{\sum_{i=1}^{n} < \mathcal{R}_{i,k+1}, \mathcal{R}_{i,0}^* >}{\sum_{i=1}^{n} < \mathcal{R}_{i,k}, \mathcal{R}_{i,0}^* >})(\frac{\alpha_k}{\omega_k})$, $\mathcal{P}_{i,k+1} = \mathcal{R}_{i,k+1} + \beta_k(\mathcal{P}_{i,k} - \omega_k \mathcal{Q}_{i,k})$.
(8) End Do

Due to the similarity of HOBi-CRSTAB with the HOBi-CGSTAB algorithm and the limited number of pages of the submitted article, writing the HOBi-CRSTAB algorithm is avoided.

PROPOSITION 2.1. *Let α_k, β_k and ω_k be the parameters obtained by HOBi-CGSTAB. The iterates in HOBi-CGSTAB satisfy the following properties*

i) $\sum_{j=1}^{n} \|\mathcal{R}_{j,k+1}\|^2 \leq \sum_{j=1}^{n} \|\mathcal{S}_{j,k}\|^2$,
ii) $\sum_{j=1}^{n} < \mathcal{S}_{j,k}, \mathcal{R}^*_{j,0} >= 0$,
iii) ω_k *minimizes* $\sum_{j=1}^{n} \|\mathcal{R}_{j,k+1}\|^2$,
iv) $\mathcal{R}_{i,k} = \mathcal{E}_i - \mathcal{L}_i(\mathcal{X}_{j,k})$,

where $i = 1, 2, \ldots, n$ and $k = 0, 1, 2, \ldots$.

3. Numerical Examples

In this section, due to existing restrictions, we give only a numerical example to show the efficiency of the HOBi-CGSTAB and HOBi-CRSTAB algorithms.

EXAMPLE 3.1. Consider the generalized coupled Sylvester tensor equation

$$\begin{cases} \mathcal{X} \times_1 A_1 \times_2 A_2 + \mathcal{Y} \times_1 B_1 \times_2 B_2 = \mathcal{E}_1, \\ \mathcal{X} \times_1 D_1 \times_2 D_2 + \mathcal{Y} \times_1 E_1 \times_2 E_2 = \mathcal{E}_2, \end{cases}$$

with

$$\begin{aligned} A_1 &= ones(m,m) + diag(3.5 + diag(rand(m))), \\ A_2 &= diag(1.5 + diag(rand(n))), \\ B_1 &= ones(m,m) - diag(1.5 + diag(rand(m))), \\ B_2 &= diag(2 + diag(rand(n))), \\ D_1 &= 1.5 \times ones(m,m) + diag(1 + diag(rand(m))), \\ D_2 &= diag(1.5 + diag(rand(n))), \\ E_1 &= ones(m,m) - diag(2.5 + diag(rand(m))), \\ E_2 &= diag(1.5 + diag(rand(n))). \end{aligned}$$

For $m = 50$ and $n = 40$, we apply the mentioned algorithms to compute the approximate solution $(\mathcal{X}_k, \mathcal{Y}_k)$. The numerical results are depicted in Figure 1, where $R_k = \log_{10} \sqrt{\|\mathcal{E}_1 - \mathcal{L}_1(\mathcal{X}_k, \mathcal{Y}_k)\|^2 + \|\mathcal{E}_2 - \mathcal{L}_2(\mathcal{X}_k, \mathcal{Y}_k)\|^2}$. As shown in Figure 1, the HOBi-CGSTAB and HOBi-CRSTAB algorithms have more superiority over the other algorithms.

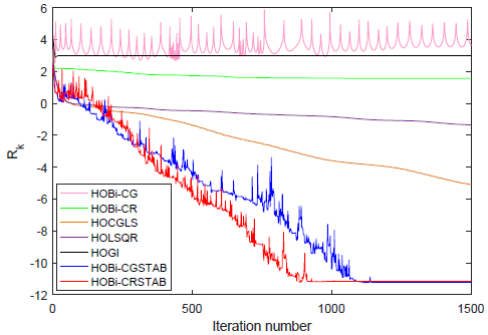


FIGURE 1. Comparison of residuals for Example 3.1.

References

1. E. Khosravi Dehdezi and S. Karimi, *Extended conjugate gradient squared and conjugate residual squared methods for solving the generalized coupled Sylvester tensor equations,* T. I. Meas. Control. (2020). DOI:10.1177/0142331220932385
2. T. G. Kolda, *Multilinear operators for higher-order decompositions,* Tech. Report SAND (2006) 2006–2081.
3. C. Lv and C. Ma, *A modified CG algorithm for solving generalized coupled Sylvester tensor equations,* Appl. Math. Comput. **365** (15) (2020) 124699.
4. H. A. Van der Vorst, *Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems,* SIAM J. Sci. Stat. Comput. **13** (2) (1992) 631–644.

E-mail: esakhosravidehdezi@gmail.com
E-mail: karimi@pgu.ac.ir

Hybrid of Finite Difference and Spectral Methods for Parabolic Time-Fractional Integro-Differential Equation

Fatemeh Mirzaei Gaskarei*
Department of Mathematics, Islamic Azad University South Tehran Branch,Tehran, Iran
and Davood Rostamy
Department of Mathematics, Islamic Azad University South Tehran Branch, Tehran, Iran & Imam Khomeini International University, Qazvin, Iran

Abstract. In the present study, a habrid of finite difference method and a Legendre-collocation spectral method are applied for solving the linear and nonlinear time-fractional parabolic integro-differential equations by the Caputo fractional derivative. In the proposed method, for space-dependent partial differential equations is used the finite difference and the time-dependent integro-differential equation is applied the spectral method. The time and space variables are on the basis of Legendre-Gauss (LG) interpolation points. We have investigated the convergence analysis of the proposed method on the L^∞-norm and L^2-norm while it is not mentioned in the paper due to high volume of calculations.
Keywords: Time-fractional parabolic integro-differential equations, Legendre-collocation spectral method, Finite difference method, Caputo derivative.
AMS Mathematical Subject Classification [2010]: 78M20, 65R20, 65M70.

1. Introduction

In this paper, we consider the following initial-boundary-value problem domains with memory term

$$D_+^\alpha u_t(x,t) + \int_0^t k(x,t-\tau)u(x,\tau)d\tau - (u^\beta(x,t))_{xx} = f(x,t), \quad (1)$$
$$(x,t) \in \Omega \times [0,T], \quad 0 < \alpha \leq 1,$$
$$u(x,0) = u_0(x), \quad x \in \Omega, \quad (2)$$
$$u(x,t) = 0, \quad (x,t) \in \partial\Omega \times [0,T], \quad (3)$$

where $\partial\Omega$ is the boundary of Ω and $\beta = 1, 2$. The functions$f$, k and u_0 are known and u is unknown function. We assumed that f , k to be sufficiently smooth and satisfies the Lipschitz condition on $\Omega \times [0,T]$. Here, D_+^α denotes the Caputo time-fractional derivative. We set $\Omega = [1,1]$. The basic idea of Legendre-collocation spectral method for solving time-fractional parabolic integro-differential equation has been proposed in [1, 2, 3, 4, 5] and these references.

*Speaker

2. Implementation Numerical Scheme

Before using collocation methods, we apply of Caputo fractional derivatives for the fractional derivative of $u(x,t)$ then, the sake of applying the theory of orthogonal polynomials, we use the variable transformations $t = \frac{T(\tau+1)}{2}, \tau \in [-1,1]$ and $s = \frac{T(y+1)}{2}, y \in [-1,\tau]$ to rewrite problems (1)-(3) as follows

$$\begin{aligned}(4)\qquad &\frac{T}{2\Gamma(1-\alpha)}\int_{-1}^{\tau}\frac{\partial v(x,y)}{\partial y}\frac{dy}{(\tau-y)^{\alpha}}+\frac{T}{2}\int_{-1}^{\tau}K(x,\tau-y)v(x,y)dy\\ &=(v^{\beta}(x,\tau))_{xx}+g(x,\tau),\quad x,\tau\in[-1,1],\quad 0<\alpha\leq 1,\end{aligned}$$

where

$$K(\tau-y):=k\left(\frac{T(1+\tau)}{2}-\frac{T(y+1)}{2}\right), v(x,\tau):=u(x,\frac{T(\tau+1)}{2}),$$
$$g(x,\tau):=f(x,\frac{T(\tau+1)}{2}),$$

and $v(x,-1):=u_0(x)$. Denote the collocation points $\{(x_j,\tau_i)\}_{j,i=0}^{N,M}$ be the set of Legendre-Gauss-Lobatto. Equation (4) at $\{(x_j,\tau_i)\}$ is

$$\begin{aligned}(5)\qquad &\frac{T}{2\Gamma(1-\alpha)}\int_{-1}^{\tau_i}\frac{\partial v(x_j,y)}{\partial y}\frac{dy}{(\tau-y)^{\alpha}}+\frac{T}{2}\int_{-1}^{\tau_i}K(x_j,\tau_i-y)v(x_j,y)dy\\ &=(v^{\beta}(x_j,\tau_i))_{xx}+g(x_j,\tau_i),\quad x,\tau\in[-1,1],\quad 0<\alpha\leq 1.\end{aligned}$$

By integrating of $v_\tau(x,\tau)$ and hold at $\{(x_j,\tau_i)\}_{j,i=0}^{N,M}$, we obtain

$$(6)\qquad v(x_j,\tau_i)=u_0(x_j)+\int_{-1}^{\tau_i}\frac{\partial v(x_j,y)}{\partial y}dy.$$

Gauss quadrature formulas will be used to compute the integral terms in Eqs. (5)-(6) so, we convert the integral interval $[-1,\tau]$ to a fixed interval $[-1,1]$

$$\begin{aligned}&\frac{T(\tau_i+1)}{4\Gamma(1-\alpha)}\int_{-1}^{1}\frac{\partial v(x_j,y(\tau_i,\theta))}{\partial\theta}\frac{d\theta}{(\tau-y(\tau_i,\theta))^{\alpha}}+\frac{T(\tau_i+1)}{4}\int_{-1}^{1}K(x_j,\tau_i-y(\tau_i,\theta))\\ &v(x_j,y(\tau_i,\theta))d\theta=(v^{\beta}(x_j,\tau_i))_{xx}+g(x_j,\tau_i),\quad x,\tau\in[-1,1],\quad 0<\alpha\leq 1,\\ &v(x_j,\tau_i)=u_0(x_j)+\frac{(\tau_i+1)}{2}\int_{-1}^{1}\frac{\partial v(x_j,y(\tau_i,\theta))}{\partial\theta}d\theta,\end{aligned}$$

by using the following variable change

$$y(\tau_i,\theta)=\frac{\tau_i+1}{2}\theta+\frac{\tau_i-1}{2},\quad \theta\in[-1,1].$$

By applying the $(p+1)$-point Legendre-Gauss type quadrature formula, using the notes and weights represented via $\{\theta_k,w_k\}_{k=0}^{p}$, we estimate the integral to obtain

$$\begin{aligned}(7)\qquad &\frac{T(\tau_i+1)}{4\Gamma(1-\alpha)}\sum_{k=0}^{p}\frac{\partial v(x_j,y(\tau_i,\theta_k))}{\partial\theta}\frac{1}{(\tau-y(\tau_i,\theta_k))^{\alpha}}w_k\\ &+\quad\frac{T(\tau_i+1)}{4}\sum_{k=0}^{p}K(x_j,\tau_i-y(\tau_i,\theta_k))v(x_j,y(\tau_i,\theta_k))w_k\end{aligned}$$

$$\begin{aligned} &\approx (v^{\beta}(x_j,\tau_i))_{xx} + g(x_j,\tau_i), \\ v(x_j,\tau_i) &\approx u_0(x_j) + \frac{(\tau_i+1)}{2}\sum_{k=0}^{p}\frac{\partial v(x_j,y(\tau_i,\theta_k))}{\partial\theta}w_k. \end{aligned} \tag{8}$$

We consider approximation solutions as follows

$$v(x,\tau) \approx I_{NM}v(x,\tau) = \sum_{n=0}^{N}\sum_{m=0}^{M} l_n(r)\rho_m(\tau)v(x_n,\tau_m), \tag{9}$$

where l_n and ρ_m are the nth and mth Lagrange interpolation polynomials based on the grid points $\{x_n\}_{n=0}^{N}$, $\{\tau_m\}_{m=0}^{M}$, respectively. Substituting these approximations Eq. (9) into Eqs. (7), (8) and appling the standard formula of numerical differentiation for estimate the spatial derivative v_{xx} in (7):

$$\begin{aligned} &\frac{T(\tau_i+1)}{4\Gamma(1-\alpha)}\sum_{k=0}^{p}v^{(1)}(x_j,\tau_i)\sum_{n=0}^{N}\sum_{m=0}^{M}\frac{1}{(\tau-y(\tau_i,\theta_k))^{\alpha}}l_n(x_j)p_m(y(\tau_i,\theta_k))w_k \\ +\ &\frac{T(\tau_i+1)}{4}\sum_{k=0}^{p}v(x_j,\tau_i)\sum_{n=0}^{N}\sum_{m=0}^{M}K(x_j,\tau_i-y(\tau_i,\theta_k))l_n(x_j)p_m(y(\tau_i,\theta_k))w_k \\ \approx\ &\frac{V_{j+1}^{\beta}(\tau_i)-2V_j^{\beta}(\tau_i)+V_{j-1}^{\beta}(\tau_i)}{h^2}+g(x_j,\tau_i), \end{aligned} \tag{10}$$

$$v(x_j,\tau_i) \approx u_0(x_j) + \frac{(\tau_i+1)}{2}\sum_{k=0}^{p}v^{(1)}(x_j,\tau_i)\sum_{n=0}^{N}\sum_{m=0}^{M}l_n(x_j)p_m(y(\tau_i,\theta_k))w_k. \tag{11}$$

To make it easier to solve the Eqs. (10) and (11) can be written in matrix form. Therefore, we denote $V_{j,i}$ and $V_{j,i}^{(1)}$ be the approximation of $v(r_j,\tau_i)$ and $v^{(1)}(r_j,\tau_i)$, respectively. So, we define matrix forms as in the following

$$\begin{aligned} &\mathbb{V}_{(N-1)(M+1)} = \mathrm{vec}[V_{j,i}], \quad \mathbb{V}^{(1)}_{(N-1)(M+1)} = \mathrm{vec}[V^{(1)}_{j,i}], \quad \mathrm{IVP} = \mathrm{vec}[u_0(x_j)], \\ &G_{(N-1)(M+1)} = \mathrm{vec}[G(x_j,\tau_i)], \quad 1\le j\le N-1, \quad 1\le i\le M+1, \end{aligned}$$

where the *vec* operator transforms a matrix to a vector via the placement of matrix rows below one other starting from the first to the last. Then, the linear systems Eqs. (10) and (11) reduce to the following matrix forms of

$$\begin{aligned} &A\mathbb{V}^{(1)}_{(N-1)(M+1)} + B\mathbb{V}_{(N-1)(M+1)} = E\mathbb{V}^{\beta}_{(N-1)(M+1)} + G_{(N-1)(M+1)}, \\ &\mathbb{V}_{(N-1)(M+1)} = \mathrm{IVP} + D\mathbb{V}^{(1)}_{(N-1)(M+1)}, \end{aligned}$$

in which for all $1\le m,n\le M+1$, $A=(A_j^i)_{j,i}$, $B=(B_j^i)_{j,i}$, and $C=(C_j^i)_{j,i}$, are block matrices. That each j,i, A_j^i, B_j^i and C_j^i are diagonal matrices with dimension of $(N-1)\times(N-1)$, where the entries of matrices A_j^i are determined

as follows:

$$(A_n^m)_{i,j} = \frac{T(\tau_i+1)}{4\Gamma(1-\alpha)}\sum_{k=0}^{p}\frac{1}{(\tau_i - y(\tau_i,\theta_k))^{\alpha}} l_n(x_j)\rho_m(y(\tau_i,\theta_k))w_k,$$

$$(B_n^m)_{i,j} = \frac{T(\tau_i+1)}{4}\sum_{k=0}^{p} k(x_j,\tau_i - y(\tau_i,\theta_k)) l_n(x_j)\rho_m(y(\tau_i,\theta_k))w_k,$$

$$(C_n^m)_{i,j} = \frac{T(\tau_i+1)}{2}\sum_{k=0}^{p} l_n(x_j)\rho_m(y(\tau_i,\theta_k))w_k,$$

we have

$$E := \frac{1}{h^2}\begin{bmatrix} -2 & 1 & \cdots & 0 \\ 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 0 & \cdots & 1 & -2 \end{bmatrix}_{(N-1)\times(N-1)}.$$

3. Numerical Results

For the following section, we use of the mentioned numerical method to solve a numerical example. We set $T = 1, p = 16$. All the calculations are supported by the software MATLAB.

EXAMPLE 3.1. Consider the problem (1) with $k(x,t) = e^{xt}$, and the exact solution is $v(x,t) = (1-x^2)\sin(t)$.

The errors in Table 1 and 2 for the linear and nonlinear problems are given with norm $\|e\|_\infty$ in the collocation node points with $N = M = 16$.

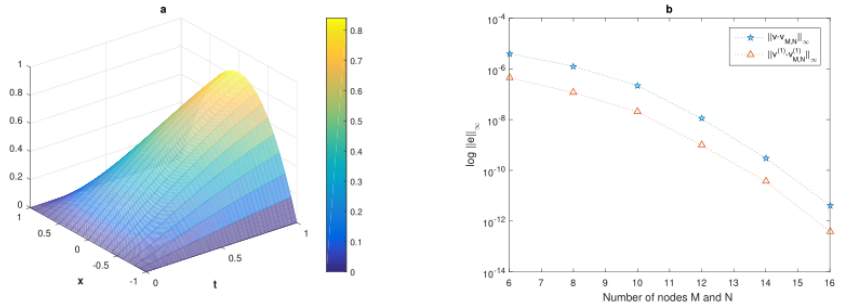


FIGURE 1. (a) Exact solution for$N = M = 16$, $\beta = 1$, $h = 0.1$ and $\alpha = 0.5$, (b) the error v and $v^{(1)}$ versus the number of collocation points.

TABLE 1. The errors $\|v - v_{M,N}\|_\infty$ for Example 3.1.

		$\alpha = 0.5$			$\alpha = 0.75$		
β	h	$M(= N)$	$\|E\|_\infty$	CPU time(s)	$M(= N)$	$\|E\|_\infty$	CPU time(s)
1	0.1	16	3.25e−12	45.32	16	6.85e−08	61.03
	0.2	16	5.91e−09	35.04	16	5.62e−07	58.67
2	0.1	16	9.32e−08	72.22	16	5.35e−06	75.29
	0.2	16	5.71e−07	65.35	16	1.21e−05	69.62

TABLE 2. The errors $\|v^{(1)} - v^{(1)}_{M,N}\|_\infty$ for Example 3.1.

		$\alpha = 0.5$			$\alpha = 0.75$		
β	h	$M(= N)$	$\|E\|_\infty$	CPU time(s)	$M(= N)$	$\|E\|_\infty$	CPU time(s)
1	0.1	16	3.25e−11	52.75	16	5.21e−07	67.52
	0.2	16	5.91e−08	41.35	16	5.62e−06	62.23
2	0.1	16	6.73e−07	81.75	16	5.35e−05	81.36
	0.2	16	7.84e−06	92.63	16	9.08e−05	92.81

References

1. Y. Jiang, *On spectral methods for volterra-type integro-differential equations*, J. Comput. Appl. Math. 230 (2009) 333–340.
2. Y. Jiang and J. Ma, *Spectral collocation methods for volterra-integro differential equations with noncompact kernels*, J. Comput. Appl. Math. **244** (2013) 115–124.
3. Y. Wei and Y. Chen, *Convergence analysis of the legendre spectral collocation methods for second order volterra integro-differential equations*, Numer. Math. **4** (2011) 419–438.
4. Y. Wei and Y. Chen, *Legendre spectral collocation method for neutral and high-order volterra integro-differential equation*, Appl. Numer. Math. **81** (2014) 15–29.
5. Y. Wei and Y. Chen, *Legendre spectral collocation method for volterra-hammerstin integral equation of the second kind*, Acta Math. Scientia. **37** (2017) 1105–1114.

E-mail: fatemeh.mirzaei64@gmail.com
E-mail: rostamy@khayam.ut.ac.ir

Desynchronization of Neural Oscillator Populations Using Least Squares Support Vector Machines

Mohammad Mahdi Moayeri*
Department of Computer and Data Sciences, Faculty of Mathematical Sciences, Shahid Beheshti University, Tehran, Iran
Kourosh Parand
Department of Computer and Data Sciences, Faculty of Mathematical Sciences, Shahid Beheshti University, Tehran, Iran
Institute for Cognitive and Brain Sciences, Shahid Beheshti University, Tehran, Iran
and Jamal Amani Rad
Institute for Cognitive and Brain Sciences, Shahid Beheshti University, Tehran, Iran

Abstract. Excessive synchronization of neurons in the brain networks can be a reason for some episodic disorders such as epilepsy. In this paper, we develop a machine learning method based on the least square support vector machine to simulate controlling synchronization in a population of noise-free and uncoupled neural oscillators. The control algorithm is based on phase reduction and uses the probability phase distribution partial differential equation to change the distribution of oscillators. We apply the proposed method on a population of Hidnmarsh-Rose neural oscillators to show the control algorithm can desynchronize the neurons efficiently.

Keywords: Phase distribution control, Neural oscillator population, Computer simulation, Support vector machine, Partial differential equations.
AMS Mathematical Subject Classification [2010]: 35Q92, 65M70, 68T05.

1. Introduction

Epilepsy is one of the most common neurological disorders which affect patients' lives quality significantly. Due to various symptoms of epilepsy and uncertainty of the cause of this disorder, there is not exist a unique definition for that [1]. However, we can define epilepsy as a neurological disorder that appears by the disproportionate synchronization, excessive excitation, or scanty inhibition in neural networks of the brain. This cognitive disorder might be closely linked with abnormal synchronization of neurons in the brain. Control strategies of the seizure are divided into three main categories: anti-epileptic drugs (AEDs), resection surgery in some acute cases and Deep Brain Stimulation (DBS). The prediction of seizure occurrence and controlling the brain functions are the main parts of the study of epilepsy disorder. In this study, we intend to prevent the seizures by controlling the synchronization in neural populations. In this paper, we consider the control algorithm proposed by Monga and Moehlis [2] and suggest an efficient numerical

*Speaker

method based on least square support vector machines (LS-SVM) to control a population of uncoupled, identical, noise-free, synchronized neural oscillator with high accuracy and low energy consumption.

2. Background

In this section, we explain the key concepts of phase reduction briefly. Phase reduction is a useful technique for describing a dynamical system that reduces the dimensionality of the system to a single-phase variable θ. Consider the following n-dimensional dynamical system [3]:

$$\frac{d\mathbf{x}}{dt} = F(\mathbf{x}) + u(t),\ \mathbf{x} \in \mathbb{R}^n, \tag{1}$$

where $u(t)$ is the control input. This system has a stable periodic orbit with periodic orbit T ,and we have $\frac{d\theta}{dt} = \frac{2\pi}{\tau} = \lambda$. By using phase reduction the aforementioned system can be written as the following one dimensional system:

$$\dot{\theta} = \omega + u(t)\mathcal{Z}(\theta), \tag{2}$$

in which $\mathcal{Z}(\theta)$ is phase response curve (PRC) which depends on neural models. We can represent the population dynamics of uncoupled, identical, noise-free by their probability distribution.

$$\frac{\partial \rho(\theta,t)}{\partial t} = -\frac{\partial}{\partial \theta}\big((\omega + u(t)\mathcal{Z}(\theta))\rho(\theta,t)\big). \tag{3}$$

Over time, by a specific control input law, we close to a desired distribution. The desired final probability distribution will be taken to be a traveling wave. For this purpose, we define L_2 norm as $\int_0^{2\pi}\Big(\rho(\theta,t) - \rho_f(\theta,t)\Big)^2 d\theta$. According to the derivative of L_2 norm and some experimental and theoretical reasons, we take the following proportional control input law:

$$u(t) = \max(\min(u_{max}, -KI(t)), u_{min}), \tag{4}$$

where K is a positive scalar and we have:

$$I(t) = 2\int_0^{2\pi}\Big(\frac{\partial \rho(\theta,t)}{\partial \theta} - \frac{\partial \rho_f(\theta,t)}{\partial \theta}\Big)\mathcal{Z}(\theta)\rho(\theta,t)d\theta. \tag{5}$$

3. Numerical Approach

In this section, we apply LS-SVM scheme to overcome the complexity of the problem and simulate the control algorithm efficiently. At first, the temporal variable in (3) is discretized using the well-known first-order Euler method. After applying the Euler algorithm on (3) we have the following equations:

$$\rho(\theta,t_k) + \Delta t\frac{\partial}{\partial \theta}\Big((\lambda + u(t_{k-1})\mathcal{Z}(\theta))\rho(\theta,t_k)\Big) = \rho(\theta,t_{k-1}),\ k = 0,...,M, \tag{6}$$

where Δt is the time step size and $t_k = k\Delta t$. By LS-SVM scheme, we approximate the solution of (6) as$\rho(\theta,t_k) \approx \rho^k(\theta) = \sum_{i=1}^N w_i\phi(\theta_i) + b = w^T\phi(theta) + b$, where

N is the number of training points. Now, the approximate solution can be obtained by solving the following optimization problem [4]:

$$\begin{aligned}(7)\quad &\min_{w,e}\ \frac{1}{2}w^Tw+\frac{\gamma}{2}e^Te\\ &s.t.\ \rho^k(\theta_i)+\Delta t\frac{\partial}{\partial\theta}\Big((\lambda+u(t_{k-1})\mathcal{Z}(\theta_i))\rho^k(\theta_i)\Big)-e_i=\rho^{k-1}(\theta_i), i=1,...,N.\end{aligned}$$

The Lagrangian function of the constrained optimization problem (7) becomes:

$$\begin{aligned}(8)\quad &\mathcal{L}(w,b,e_i,\alpha_i)=\frac{1}{2}w^Tw+\frac{\gamma}{2}e^Te-\sum_{i=1}^{N}\alpha_i\Big[w^T\Big(\phi(\theta_i)+\lambda\Delta t\phi'(\theta_i)+\\ &\Delta tu^{k-1}\mathcal{Z}'(\theta_i)\phi(\theta_i)+\Delta tu^{k-1}\mathcal{Z}(\theta_i)\phi'(\theta_i)\Big)\Big(1+\Delta tu^{k-1}\mathcal{Z}(\theta_i)\Big)b-\rho^{k-1}-e_i\Big],\end{aligned}$$

in which $\{\alpha_i\}_{i=1}^N$ are Lagrange multipliers [5]. Consider that $\phi(\theta_i)=\phi_i$; then, the KarushKuhnTucker (KKT) optimality conditions are as follow:

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial w}=0\ &\Rightarrow\ \sum_{i=1}^{N}\alpha_i\Big[\phi_i+\lambda\Delta t\phi_i'+\Delta tu^{k-1}\mathcal{Z}'(\theta_i)\phi_i+\Delta tu^{k-1}\mathcal{Z}(\theta_i)\phi_i'\Big],\\ \frac{\partial\mathcal{L}}{\partial b}=0\ &\Rightarrow\ \sum_{i=1}^{N}\alpha_i\Big[1+\Delta tu^{k-1}\mathcal{Z}(\theta_i)\Big],\\ \frac{\partial\mathcal{L}}{e_i}=0\ &\Rightarrow\ e_i=\frac{\alpha_i}{\gamma},\\ \frac{\partial\mathcal{L}}{\alpha_i}=0\ &\Rightarrow\ w^T\Big(\phi_i+\lambda\Delta t\phi_i'+\Delta tu^{k-1}\mathcal{Z}'(\theta_i)\phi_i+\Delta tu^{k-1}\mathcal{Z}(\theta_i)\phi_i'\Big)\\ &+\ \Big(1+\Delta tu^{k-1}\mathcal{Z}(\theta_i)\Big)b-e_i=\rho^{k-1}.\end{aligned}$$

In order to solve the problem in the dual form, we should construct the kernel matrix and its derivatives as:

$$(9)\qquad [\Omega_n^m]_{i,j}=\Big[\phi^{(n)}(u_i)\phi^{(m)}(v_j)\Big]_{i,j},\ m,n=0,1.$$

By elimination of the primal variables w and $\{e_i\}_{i=1}^N$, using Mercers theorem and imposing the boundary condition, the following matrix equation of the dual form is obtained:

$$(10)\qquad \Big(\mathcal{D}\left[\begin{array}{c|c}\mathcal{K}+\frac{1}{\gamma}I & \mathcal{P}\\ \hline \mathcal{F} & 0\end{array}\right]+\mathcal{Q}\Big)\begin{bmatrix}\alpha\\ b\end{bmatrix}=\mathcal{D}\begin{bmatrix}\rho^{k-1}\\ 0\end{bmatrix},$$

where

$$\begin{aligned}\mathcal{D}&=\mathsf{Diag}(0,1,1,\ldots,1),\\ \mathcal{K}&=\mathcal{A}_0^0+\mathcal{A}_0^1+\mathcal{A}_1^0+\mathcal{A}_4\Omega_1^1,\\ \mathcal{A}_0^0&=\Omega_0^0+\Delta tu^{k-1}\mathsf{Diag}(\mathcal{Z}')\Omega_0^0+\Delta tu^{k-1}\Omega_0^0\mathsf{Diag}(\mathcal{Z}')\\ &\quad+\Delta t^2(u^{k-1})^2\mathsf{Diag}(\mathcal{Z}')\Omega_0^0\mathsf{Diag}(\mathcal{Z}'),\\ \mathcal{A}_0^1&=\lambda\Delta t\Omega_0^1+\Delta tu^{k-1}\mathsf{Diag}(\mathcal{Z})\Omega_0^1+\lambda\Delta t^2u^{k-1}\Omega_0^1\mathsf{Diag}(\mathcal{Z})\\ &\quad+\Delta t^2(u^{k-1})^2\mathsf{Diag}(\mathcal{Z})\Omega_0^1\mathsf{Diag}(\mathcal{Z}'),\end{aligned}$$

$$
\begin{aligned}
\mathcal{A}_1^0 &= \lambda\Delta t\Omega_1^0 + \lambda\Delta t^2 u^{k-1}\mathsf{Diag}(\mathcal{Z}')\Omega_1^0 + \Delta t u^{k-1}\Omega_1^0\mathsf{Diag}(\mathcal{Z}) \\
&\quad + \Delta t^2(u^{k-1})^2\mathsf{Diag}(\mathcal{Z}')\Omega_1^0\mathsf{Diag}(\mathcal{Z}), \\
\mathcal{A}_1^1 &= \lambda^2\Delta t^2\Omega_1^1 + \lambda\Delta t^2 u^{k-1}\mathsf{Diag}(\mathcal{Z})\Omega_1^1 + \lambda\Delta t^2 u^{k-1}\Omega_1^1\mathsf{Diag}(\mathcal{Z}) \\
&\quad + \Delta t^2(u^{k-1})^2\mathsf{Diag}(\mathcal{Z})\Omega_1^1\mathsf{Diag}(\mathcal{Z}), \\
\mathcal{P} &= 1 + \Delta t u^{k-1}\mathcal{Z}', \\
\mathcal{F} &= 1 + \Delta t u^{k-1}\mathcal{Z}', \\
\mathcal{Q} &= \begin{bmatrix} \mathcal{B}_1^1 - \mathcal{B}_N^1 & \mathcal{B}_1^2 - \mathcal{B}_N^2 & \cdots & \mathcal{B}_1^N - \mathcal{B}_N^N \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \end{bmatrix}, \\
\mathcal{B}_p^q &= [\Omega_0^0]_{p,q} + \lambda\Delta t[\Omega_0^1]_{p,q} + \Delta t\mathcal{Z}_q'[\Omega_0^0]_{p,q} + \Delta t u^{k-1}\mathcal{Z}_q[\Omega_0^1]_{p,q}.
\end{aligned}
$$

By solving the linear algebraic system (10), the probability distribution is approximated at each time step.

4. Simulation Results

In this section, we simulate the control model on a population of uncoupled Hindmarsh-Rose oscillators. We can use the proposed method for different neural models by choosing the appropriate PRC function for the model. For HR model, PRC is defined as $\mathcal{Z}(\theta) = (1 - \cos(\theta))/2\pi$. The initial probability distribution of synchronized neurons are considered as $\rho(\theta, 0) = \frac{\exp(\zeta\cos(\theta-\pi))}{2\pi\mathcal{I}_0(b)}$, where $\mathcal{I}_0$ is the modified Bessel function of first kind of order 0 and $\zeta = 26$.The desired final distribution consider as $\rho_f(\theta, t) = \frac{1}{2\pi}$. The simulation is done with $N = 100, M = 200$. The model has a stable periodic orbit with time period $\tau = 0.2$ and $T = 5\tau$.

We use the Gaussian kernels $K(u, v) = \exp\Big(-(\frac{u-v}{\sigma})^2\Big)$ for the simulations. Moreover, suitable values for parameters are chosen by test and trial as $u_{max} = -u_{min} = 5$, $K = 100$, $\lambda = 10^4$ and $\sigma = 0.07$. The obtained results are represented in Figure 1.

5. Conclusion Remarks

Recent studies conclude that many of the cognitive disorders such as epilepsy and Parkinsons diseases might be closely linked with abnormal synchronization of neurons in the brain; so, in this paper, we developed an efficient and fast numerical algorithm to improve a control algorithm for desynchronization of uncoupled and noise-free neural oscillators all receiving the same control input. The algorithm was based on phase reduction and used a population-level partial differential equation formulation for this issue. The proposed method was applied on Hindmarsh-Rose neurons as an example to evaluate its efficacy.

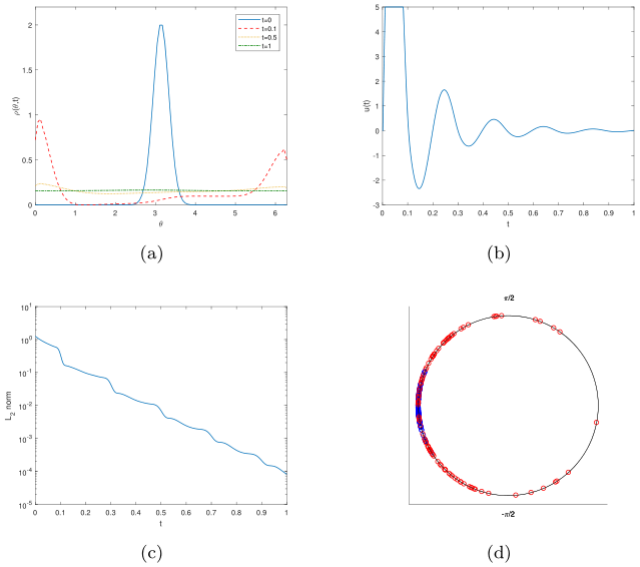


FIGURE 1. Simulation results for HR oscilators: (a) Probability distribution at various times (b) The control input (c) Logarithm of L_2 norm (d) Distribution of the firing of 100 neurons at $t = 0$ (blue circles) and $t = 5\tau$ (red circles).

References

1. S. Shorvon, *Handbook of Epilepsy Treatment*, Wiley-Blackwell, London, 2011.
2. B. Monga and J. Moehlis, *Phase distribution control of a population of oscillators*, Phys. D **398** (2019) 115 –129.
3. B. Monga and J. Moehlis, *Supervised learning algorithms for controlling underactuated dynamical systems*, Phys. D **412** (2020) 132621.
4. S. Mehrkanoon and J. Suykens, *Learning solutions to partial differential equations using LS-SVM*, Neurocomputing **159** (2015) 105–116.
5. K. Parand, A. A. Aghaei, M. Jani and A. Ghodsi, *A new approach to the numerical solution of Fredholm integral equations using least squares-support vector regression*, Math. Comput. Simul. **180** (2021) 114–128.

E-mail: m_moayeri@sbu.ac.ir
E-mail: k_parand@sbu.ac.ir
E-mail: j_amanirad@sbu.ac.ir

Solving Time-Dependent PDEs with Rational Radial Basis Function Collocation and Semi-Implicit Time Discretization

Reza Mohammadi Arani*
Department of Applied Mathematics, Amirkabir University of Technology, Tehran, Iran
and Mehdi Dehghan
Department of Applied Mathematics, Amirkabir University of Technology, Tehran, Iran

Abstract. The stability of solving time-dependent PDEs with RBF collocation method, depends on time discretization method. In many problems we use implicit methods to increase the stability range of numerical methods. Rational RBF (RRBF) is an improvement of standard RBF which has more potential to approximate discontinuous problems than standard RBF. As RRBFs are non-linear, so to avoid calculating nonlinear system of equations, we need to discretize time variable with explicit methods which they are conditionally stable and usually their stability ranges are smaller than implicit methods. In this paper we present an approach to increase the stability of solving time-dependent PDEs with RRBFs methods.

Keywords: Rational RBF, Burgers equation, Advection equation, Semi-implicit scheme.
AMS Mathematical Subject Classification [2010]: 65D05.

1. Introduction

Rational radial basis function (RRBF) is an improvement of RBF which introduced to interpolate functions with poles, which are challenging issues in applied mathematics [1, 4]. RRBFs appear more powerful and efficient than standard RBFs in interpolation problems with discontinuous solutions, but they suffer from high computational cost. So De Marchi et al. [2], introduced partition of unity RRBF (RRBF-PU), to improve the speed of method especially in higher dimensional problems. Recently Sarra [5] used RRBF and RRBF-PU to solve some time-dependent PDE problems with discontinuous solutions. He used a high order explicit method such as Runge-Kutta to discretize time direction and also used RRBF to discretize space variable. As expected, this approach was successful to solve one-dimensional pure advection equation with discontinuous initial condition and also one-dimensional inviscid Burgers equation which has a discontinuous solution [5]. These two problems are very hard to be solved with standard RBFs, we invite reader to study reference [1, 3].

Since the RRBFs are non-linear, so solving a time-dependent PDE problem with RRBFs is not simple as working with standard RBFs. To avoid solving nonlinear systems, we need to discretize time direction with explicit methods which they are conditionally stable and usually their stability ranges are smaller than stability range of implicit methods.

*Speaker

In this paper we are going to present an approach to use RRBFs for space variable discretization and also an implicit method to discretize time. The new scheme is semi-implicit approach which it is more stable than conventional RRBFs and it has their benefits in application.

The rest of this paper is organized as follows: In next section we introduce RRBF and in Section 3 we explain how to solve time-dependent PDEs with semi-implicit approach and Section 4 is devoted to numerical results.

2. Rational Radial Basis Functions

Let us consider the framework of standard RBF collocation problems. Let $\Omega \subseteq \mathbb{R}^d$ with $d = 1, 2, 3$ be area of space variable and $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N | \mathbf{x}_i \in \Omega\}$ be the sets of centers. Also let us assume that $\phi(r)$ be a radial function and define $\Phi_i(\mathbf{x}) = \phi(\|\mathbf{x}_i - \mathbf{x}\|)$ which is a radial basis function. To interpolate data by standard RBFs we have $\mathcal{R}_N f(\mathbf{x}) = \sum_{i=1}^N \lambda_i \Phi_i(\mathbf{x})$, where $\mathcal{R}_N$ is the interpolation operator and λ_i are some scalers. In RRBFs we consider the interpolation problem as follows:

$$\mathcal{R}_N f(\mathbf{x}) := \frac{p(\mathbf{x})}{q(\mathbf{x})}, \tag{1}$$

where $p, q \in \mathcal{N}_\Phi(\Omega)$ (Native space of Φ). So we have $p(\mathbf{x}) = \sum_{i=1}^N \alpha_i \Phi_i(\mathbf{x})$ and $q(\mathbf{x}) = \sum_{i=1}^N \beta_i \Phi_i(\mathbf{x})$, where α_i and β_i can be determined with collocation method if $p(\mathbf{x})$ and $q(\mathbf{x})$ are known.

Let $\mathbf{p} = (p(\mathbf{x}_1), \ldots, p(\mathbf{x}_N))^T$ and $\mathbf{q} = (q(\mathbf{x}_1), \ldots, q(\mathbf{x}_N))^T$ be the evaluation vectors of p and q, respectively. From (1) we have $\mathcal{R}_N f(\mathbf{x}_i) = f(\mathbf{x}_i) = \frac{\mathbf{p}_i}{\mathbf{q}_i}$ for $i = 1, 2, \ldots, N$, i.e.,

$$\mathbf{p} = D\mathbf{q},$$

where D is a diagonal matrix with $(f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_N))^T$ on its diagonal. Assume A is the interpolation matrix with $\Phi_i(\mathbf{x})$ basis ($A_{i,j} = \Phi_i(\mathbf{x}_j)$), then we have

$$A\alpha = \mathbf{p}, \ A\beta = \mathbf{q}.$$

To find $\mathbf{p}$ and $\mathbf{q}$ we need to solve an optimization problem as follows [4]:

$$\min_{\substack{\mathbf{q} \in \mathbb{R}^N \\ \frac{1}{\|\mathbf{f}\|_2^2}\|D\mathbf{q}\|_2^2 + \|\mathbf{q}\|_2^2 = 1}} \left(\frac{1}{\|\mathbf{f}\|_2^2} \mathbf{q}^T D^T A^{-1} D\mathbf{q} + \mathbf{q}^T A^{-1} \mathbf{q} \right). \tag{2}$$

The optimization problem (2) is equivalent to the generalized eigenvalue problem $\mathcal{M}\mathbf{q} = \lambda \mathcal{N}\mathbf{q}$, where $\mathcal{M} = \frac{1}{\|\mathbf{f}\|_2^2} D^T A^{-1} D + A^{-1}$ and $\mathcal{N} = \frac{1}{\|\mathbf{f}\|_2^2} D^T D + I_N$.

3. Implicit Method for Time Discretization

In this section we want to discretize two time-dependent PDE problems using RRBFs method for approximating the space variable and Crank-Nicolson method for time discretization.

3.1. Pure Advection Equation. We consider

$$u_t = -u_x.$$

Using Crank-Nicolson method to discretize time variable, we obtain

$$u^{n+1} + \frac{\Delta t}{2} u_x^{n+1} = u^n - \frac{\Delta t}{2} u_x^n.$$

Let $\mathcal{X} = \{\mathbf{x}_i \in \Omega | i = 1, 2, \ldots, N\}$ be the set of centers and also let

$$\mathbf{u}^n \approx (u(t_n, \mathbf{x}_1), u(t_n, \mathbf{x}_2), \ldots, u(t_n, \mathbf{x}_N))^T,$$

be the approximation solution of u at n-th time step and center points. We use the RRBFs to approximate u^n as follows

$$u^{n+1} = \frac{p^{n+1}}{q^{n+1}},$$

where $p^n, q^n \in \mathcal{N}_\Phi(\Omega)$, so we obtain

$$\frac{p^{n+1}}{q^{n+1}} + \frac{\Delta t}{2} \left(\frac{p^{n+1}}{q^{n+1}} \right)_x = u^n - \frac{\Delta t}{2} u_x^n. \tag{3}$$

Calculating the derivatives in Eq. (3), gives

$$\frac{p^{n+1}}{q^{n+1}} + \frac{\Delta t}{2} \left(\frac{p_x^{n+1} q^{n+1} - p^{n+1} q_x^{n+1}}{(q^{n+1})^2} \right) = u^n - \frac{\Delta t}{2} u_x^n, \tag{4}$$

and some easy calculations, (4) yields

$$\frac{1}{q^{n+1}} \left(p^{n+1} + \frac{\Delta t}{2} \left(p_x^{n+1} - p^{n+1} \frac{q_x^{n+1}}{q^{n+1}} \right) \right) = u^n - \frac{\Delta t}{2} u_x^n. \tag{5}$$

The non-linearity of approximating the space variable with RRBFs, is observable from Eq. (5). Now we obtain (5) for any centers in $\mathcal{X}$ and we get

$$(6) \left(\mathbf{q}^{n+1}\right)^{\cdot -1} . * \left(\mathbf{p}^{n+1} + \frac{\Delta t}{2} \left(\mathbf{p}_x^{n+1} - \mathbf{p}^{n+1} . * \mathbf{q}_x^{n+1} . * \left(\mathbf{q}^{n+1}\right)^{\cdot -1} \right) \right) = \mathbf{u}^n - \frac{\Delta t}{2} \mathbf{u}_x^n,$$

where ". " is the Hadamard product and $\mathbf{p}$ and $\mathbf{q}$ are the vectors of evaluating p and q at the center points. Using RRBF method we get

$$u^n = \frac{p^n}{q^n},$$

where $p^n, q^n \in \mathcal{N}_\Phi(\Omega)$. Assuming $\mathbf{q}^* = \mathbf{q}^n$ and

$$\mathbf{u}^{n+1} = \mathbf{p}^* . * (\mathbf{q}^*)^{\cdot -1}, \tag{7}$$

we achieve the semi-implicit RRBFs. So from (6) we assume

$$\mathbf{p}^* + \frac{\Delta t}{2} \left(\mathbf{p}_x^* - \mathbf{p}^* . * (\mathbf{q}^*)_x . * (\mathbf{q}^*)^{\cdot -1} \right) = \mathbf{q}^* . * (\mathbf{u}^n - \frac{\Delta t}{2} \mathbf{u}_x^n). \tag{8}$$

As $\mathbf{p}^*$ and $\mathbf{q}^*$ are in the Native space of Φ, so we have $\mathbf{p}_x^* = D_x \mathbf{p}^*$ and $\mathbf{q}_x^* = D_x \mathbf{q}^*$, where D_x is the derivative matrix based on kernel Φ. Let us define $\mathbf{R} = \mathbf{q}^* . * (\mathbf{u}^n - \frac{\Delta t}{2} \mathbf{u}_x^n)$, so we can rewrite (8) as follows

$$\left(I_N + \frac{\Delta t}{2} (D_x - L_x) \right) \mathbf{p}^* = \mathbf{R}, \tag{9}$$

where I_N is the identity matrix of dimension N and also L_x is a diagonal matrix where the vector $(Dx\mathbf{q}^*).*(\mathbf{q}^*)^{.-1}$ is on its diagonal. From (9) we can obtain $\mathbf{p}^*$. Also directly from Eqs. (7) and (9), $\mathbf{u}^{n+1}$ can be calculated, where $\mathbf{u}^{n+1}$ is the approximate solution of evaluating $u(x, t_{n+1})$ at center points.

3.2. Inviscid Conservative Burgers Equation. Let us consider the one dimensional inviscid conservative Burgers equation

$$u_t + uu_x = 0.$$

Using Crank-Nicolson method to discretize time variable yields

$$u^{n+1} + 0.5\Delta t\left(u^{n+1}u_x^n + u^n u_x^{n+1}\right) = u^n.$$

Let us assume $\mathbf{u}^{n+1} = \mathbf{p}^*.*(\mathbf{q}^*)^{.-1}$, then with the same calculation mentioned in Subsection 3.1, we derive

$$\left[I_N + \frac{\Delta t}{2}\left(\mathcal{D}(\mathbf{u}_x^n) + \mathcal{D}(\mathbf{u}^n)\left(D_x - L_x\right)\right)\right]\mathbf{p}^* = \mathcal{D}(\mathbf{q}^*)\mathbf{u}^n,$$

where I_N is the identity matrix of dimension N, $\mathcal{D}(v)$ is a diagonal matrix, where vector v is on its diagonal. Also D_x is the derivative matrix based on kernel Φ and $L_x = (D_x\mathbf{q}^*).*(\mathbf{q}^*)^{.-1}$. Assuming $\mathbf{q}^* = \mathbf{q}^n$, where $\mathbf{u}^n = \mathbf{p}^n.*(\mathbf{q}^n)^{.-1}$ we obtain

$$\mathbf{p}^* = \left[I_N + \frac{\Delta t}{2}\left(\mathcal{D}(\mathbf{u}_x^n) + \mathcal{D}(\mathbf{u}^n)\left(D_x - L_x\right)\right)\right]^{-1}\mathcal{D}(\mathbf{q}^*)\mathbf{u}^n,$$

so we can obtain $\mathbf{u}^{n+1}$.

4. Numerical Experiments

In this section we present some numerical experiment for solving advection and Bergurs' equations with different initial conditions. In all experiments we choose the multi-quadratic RBF (MQ)

$$\Phi_i(\mathbf{x}) = \sqrt{1 + \varepsilon^2(\mathbf{x} - \mathbf{x}_i)^2},$$

where ε is the shape parameter, for space variable and we employ the Crank-Nicolson finite difference formula for time direction discretization.

Example 4.1. We consider the one dimensional advetion equation with 3 discontinuous initial conditions, where they are as follows

$$u_1(0,x) = \begin{cases} 1, & x \le 0, \\ -1, & x > 0, \end{cases}, \quad u_2(0,x) = \begin{cases} \cos(\pi x), & x \le 0, \\ -\sin(\pi x), & x > 0, \end{cases},$$

$$u_3(0,x) = \begin{cases} e^{\pi x}, & x \le 0, \\ -e^{-\pi x}, & x > 0. \end{cases}$$

In this example we consider 100 nodes and $\Delta t = 0.001$. The results have been shown in Figure 1.

Example 4.2. In this example we consider $u_1(0,x) = -\sin(\pi x)$, $u_2(0,x) = -\sin(0.5\pi x)$ and $u_3(0,x) = -\tanh(\pi x)$ as initial condition for inviscid Burgers' equation. Also we consider 101 nodes and $\Delta t = 0.001$. The results of this example have been shown in Figure 2.

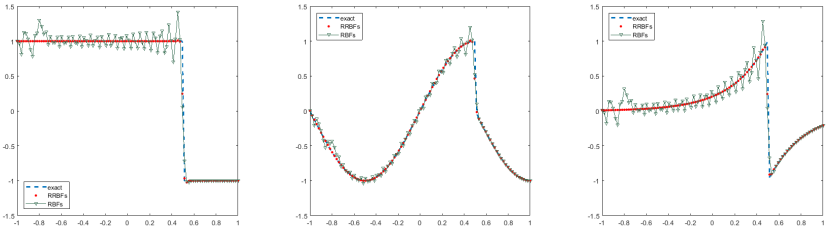


FIGURE 1. The result of advection equation at final time $\tau = 0.5$ with initial condition u_1 is Left, u_2 is center and u_3 is right figure.

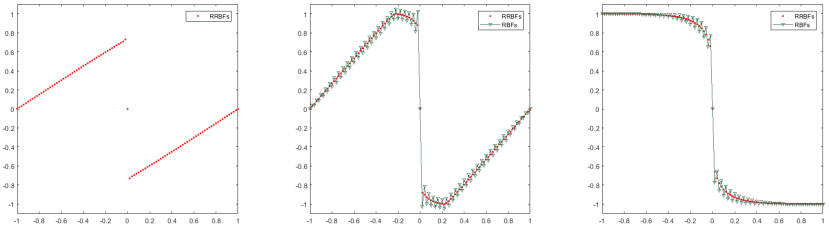


FIGURE 2. The result of Burgers equation with initial condition and final time u_1 and $\tau = 1$ is left, u_2 and $\tau = 0.75$ is center and u_3 and $\tau = 0.35$ is right figure.

5. Conclusion

Since RRBF is an appropriate tool for approximating functions with poles and discontinuity, we employed it to descretize the space variable of time-dependent PDEs. The RRBF is a non-linear approximation approach, so we needed to discretize time variable with explicit methods to avoid solving non-linear system of equations. As explicit methods usually have smaller stability range, so we presented a new method based on semi-implicit approach to discretize time direction. As shown in Figures 1 and 2, we can see the discontinuity of the solutions which have been approximated well with RRBFs approach. This ability emphasized the accuracy and efficiency of RRBFs collocations.

References

1. M. Dehghan and M. Abbaszadeh, *The space-splitting idea combined with local radial basis function meshless approach to simulate conservation laws equations*, Alexandria Eng. J. **57** (2) (2018) 1137–1156.
2. S. De Marchi, A. Martinez and E. Perracchione, *Fast and stable rational RBF-based partition of unity interpolation*, J. Comput. Appl. Math. **349** (2019) 331–343.
3. T. A. Driscoll and A. R. H. Heryudono, *Adaptive residual subsampling methods for radial basis function interpolation and collocation problems*, Comput. Math. Appl. **53** (6) (2007) 927–939.
4. S. Jakobsson, B. Andersson and F. Edelvik, *Rational radial basis function interpolation with applications to antenna design*, J. Comput. Appl. Math. **233** (4) (2009) 889–904.

5. S. A. Sarra and Y. Bai, *A rational radial basis function method for accurately resolving discontinuities and steep gradients*, Appl. Num. Math. **130** (2018) 131–142.

E-mail: r.mohammadiarani@aut.ac.ir
E-mail: mdehghan@aut.ac.ir

An Anisotropic Fractional Nonlinear Diffusion Equation for Multiplicative Noise Removal of Texture Images

Maryam Mohammadi
Department of Mathematical Sciences, Isfahan University of Technology, Isfahan 84156-83111, Iran
Reza Mokhtari*
Department of Mathematical Sciences, Isfahan University of Technology, Isfahan 84156-83111, Iran
and Nader Karimi
Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan 84156-83111, Iran

Abstract. We present here a fractional-order diffusion equation to denoise the texture images corrupted by the multiplicative noises. The fractional derivative can preserve texture image features, and the proposed gray level indicator controls anomalous diffusion and causes more details of the image to be preserved.
Keywords: Fractional-order diffusion equation, Gray level indicator, Texture images.
AMS Mathematical Subject Classification [2010]: 65M06, 35R11, 26A33.

1. Introduction

Image denoising intends to repair a noisy image to an image with higher quality. In 1990, Perona and Malik [2] proposed a nonlinear diffusion equation for additive noise removal. Multiplicative noise removal has attracted much attention in recent years. Unlike additive noise, multiplicative noise destroys almost all information of the texture images. Thus, the majority of existing methods are not so suitable for the reparation of texture images. Denote an observed image by $f = f(\boldsymbol{x})$, $\boldsymbol{x} := (x, y) \in \Omega \subset \mathbb{R}^2$, where Ω is the bounded domain of the image with two space dimensions and has a Lipschitz boundary. We assume $f = u\eta$, where u is the free-noise image and η is the gamma noise which possesses distributional function for $\eta > 0$ as $p(\eta) = \frac{L^L \eta^{L-1}}{\Gamma(L)} \exp(-L\eta)$, where $L \in \mathbb{N}$ and $\Gamma(\cdot)$ denotes the gamma function. The mean value of η is 1 and the variance of η is $\frac{1}{L}$. A classic way to solve the problem of multiplicative noise removal is to derive a variational denoising model based on the total variation. The general approach to implementing these variational denoising models derives the corresponding evolution equation and then discretizes the equation to test on the noisy image. Generally, we can obtain the following evolution equation for a variational model

$$\partial_t u = \operatorname{div}(a(|\nabla u|, u)\nabla u) - \lambda h(f, u), \qquad \boldsymbol{x} \in \Omega, \quad t \in (0, T), \tag{1}$$

*Speaker

with the boundary condition $\langle \nabla u, \vec{n} \rangle = 0$ and initial condition $u(\boldsymbol{x}, 0) = f(\boldsymbol{x})$. Aubert and Aujol [1] derived a multiplicative noise removal model to minimize energy functional

$$\min_{u \in S(\Omega)} \left\{ \int_\Omega |Du| + \lambda \int_\Omega \Big(\log u + \frac{f}{u} \Big) d\boldsymbol{x} \right\}, \tag{2}$$

where $S(\Omega) = \{u > 0, u \in \mathrm{BV}(\Omega)\}$, $\mathrm{BV}(\Omega)$ is the bounded variation on Ω, $\int_\Omega |Du|$ is the total variation regularization term, and $\int_\Omega (\log u + \frac{f}{u}) d\boldsymbol{x}$ is a fidelity term used for the preservation of details and image edges. The evolution equation associated with (2) is constructed as follows

$$\partial_t u = \mathrm{div}\Big(\frac{\nabla u}{|\nabla u|} \Big) + \lambda \frac{f - u}{u^2}, \qquad \boldsymbol{x} \in \Omega, \quad t > 0. \tag{3}$$

If we define a new fidelity term as $\int_\Omega (-\frac{u}{2} - \sqrt{f} \log u) d\boldsymbol{x}$ and place it in (2), in this case, term of the right-hand side in the evolution Eq. (3) would be changed as $\lambda(\frac{1}{2} + \frac{\sqrt{f}}{u})$. Evolution Eq. (1) is a diffusion equation in which $a(|\nabla u|, u)$ and $h(f, u)$ are the diffusion coefficient and source term, respectively.

2. Proposed Model

By inspiration of model (3) introduced by Yao et al. [4] for multiplicative noise removal for texture images and setting $\tau(u) = (1 + k|D^\alpha u|^2)^{\frac{1-\beta}{2}}$, we proposed the following equation

$$\partial_t u = -D_x^{\alpha *} \Big(b(u) \frac{D_x^\alpha u}{\tau(u)} \Big) - D_y^{\alpha *} \Big(b(u) \frac{D_y^\alpha u}{\tau(u)} \Big) + \lambda \Big(\frac{1}{2} + \frac{\sqrt{f}}{u} \Big). \tag{4}$$

with the boundary condition $u(x, y, t) = 0$ and initial condition $u(x, y, 0) = f(x, y)$. Here, $(x, y) \in \Omega \subset \mathbb{R}^2$, $t \in (0, T]$, $1 < \alpha < 2$, $0 < \beta < 1$, $k > 0$, and λ is a positive parameter that controls the fidelity of solution to the input image. In [3], an adaptive total variation model has been proposed as $\min \int_\Omega g(\boldsymbol{x}) |\nabla u| d\boldsymbol{x}$, where the weight function g controls the speed of diffusion at different points. Utilizing the idea, we proposed gray level indicator as $b(u) = (|G_\sigma * u|/M)^r$, where $M = \sup (G_\sigma * u)(x, y)$ and $r > 0$ is a parameter, $*$ is the convolution operator and G_σ for $\sigma > 0$ is the gaussian filter, which is considered as $G_\sigma(x, y) = \exp(-(x^2 + y^2)/4\sigma^2)/4\pi\sigma$. The use of gaussian convolution in the proposed model has many advantages, not only the robustness in denoising viewpoint but also the well-posedness in the theoretical perspective. The indicator $b(u)$ has these properties: $b(s)$ is monotonically increasing, $b(0) = 0$, $b(s) \geq 0$, and $b(s) \to 1$, as $s \to \sup u_{\boldsymbol{x} \in \Omega}$. Besides, D^α denotes the fractional derivative operator defined by $D^\alpha u = (D_x^\alpha u, D_y^\alpha u)$ in which D_x^α and D_y^α are respectively Grünwald-Letnikov (GL) fractional-order derivative into x and y, $|D^\alpha u|^2 = |D_x^\alpha u|^2 + |D_y^\alpha u|^2$, $D_x^{\alpha *}$ is the adjoint of D_x^α and $D_y^{\alpha *}$ is the adjoint of D_y^α. The gray level indicator $b(u)$ is a much smaller value at low gray level ($b(u) \to 0$) than at high gray levels, so that some small features at low gray levels are much less smooth and therefore are preserved. In high gray level regions, approximating $b(u)$ to 1 leads to control of diffusion coefficient by $w = 1/\big(1 + k|D^\alpha u|^2\big)^{(1-\beta)/2}$. In the regions with the high-frequency edges, $w \to 0$ (as $|D^\alpha u| \to \infty$), the image gradient is large, and this

shows that Eq. (4) can preserve edges well; in the relatively smooth regions, $w \to 1$ (as $|D^\alpha u|^2 \to 0$), the image gradient is small, and this shows that Eq. (4) is able to denoise smooth regions well. According to the theory of the diffusion equation, different choices of parameter β in w leads to different difficulties. Furthermore, in the limit case $\beta = 0$, the Eq. (4) degenerates when $|D^\alpha u| \to \infty$, while in the limit case $\beta = 1$, the Eq. (4) would never degenerate at any point with respect to $|D^\alpha u|$. The parameter β controls the speed of degeneracy, as the closer β gets to 0, the faster the Eq. (4) degenerates. Moreover, for a specific area in the noise removal process, the closer β gets to 0, the slower the regularization will be.

THEOREM 2.1. (Extremum principle) *If u is a solution of Eq. (4) in the distributional sense, which satisfies $u \in C([0,T];L^2(\Omega)) \cap L^2(0,T;V)$ and $\partial_t u \in L^2(0,T;H^1(\Omega))$, where $V = \{v | v \in W_2^\alpha(\Omega), v|_{\partial\Omega} = 0\}$, then it fulfills $\underline{u} \leq u(\cdot,\cdot,\cdot) \leq \overline{u}$, where $\underline{u} = \inf_\Omega u$ and $\overline{u} = \sup_\Omega u$.*

THEOREM 2.2. *If u is a solution of Eq. (4), then the mean value of it on the image domain Ω keeps invariant, i.e.,*

$$\frac{1}{\text{mean}(\Omega)} \int_\Omega u(x,y,t)dxdy = \frac{1}{\text{mean}(\Omega)} \int_\Omega f(x,y)dxdy.$$

Most of the existing models for multiplicative noise removal use integer-order derivatives, which may produce blur edges and not preserve some image details due to their local property. Unlike the integer derivative operator, the fractional derivative operator possesses the non-local property because the fractional derivative at a point depends on the characteristics of the entire function and not just the vicinity values of the point. A fractional-order based method applying for preserving image details achieves a good trade-off between eliminating speckle artifacts and restraining staircase effect in texture images. Because of the non-local property and long-rang dependency property of the fractional derivative, the model can perform well in texture preservation.

3. Numerical Method and Results

We consider a finite difference scheme in the spatial domain to solve the proposed problem. Let the initial discrete image has $N \times N$ pixels, $\Delta x = \Delta y = 1$ be the space size and $u(x,y) = u(x\Delta x, y\Delta y)$ for $x, y = 0, \ldots, N-1$. Left-sided GL derivative and the adjoint operator of D^α for a real function g respectively defined as follows

$$\begin{aligned} D^\alpha g(x) &= \lim_{h\to 0^+} \frac{\sum_{k\geq 0}(-1)^k C_k^\alpha g(x-kh)}{h^\alpha}, \\ D^{\alpha *} g(x) &= (-1)^m \lim_{h\to 0^+} \frac{\sum_{k\geq 0}(-1)^k C_k^\alpha g(x+kh)}{h^\alpha}, \end{aligned}$$

where $\alpha > 0$ and m is an integer satisfying $m-1 \leq \alpha < m$. C_k^α converges to zero quickly when $k \to \infty$ for fixed α. As usual, $h = 1$. We define a spatial partition (x_k, y_l) (for all $k, l = 0, 1, \ldots, N-1$) of image domain Ω. Then, we consider a discretization of the α-order fractional derivative at all points of Ω along the

x-direction by using

$$\begin{aligned} D_x^{\alpha} g(x_k, y_l) &= \sum_{i=0}^{k} \omega_i^{\alpha} g(x_{k-i}, y_l), \\ D_x^{\alpha *} g(x_k, y_l) &= (-1)^m \sum_{i=0}^{N-k-1} \omega_i^{\alpha} g(x_{k+i}, y_l), \end{aligned}$$

where $\omega_i^{\alpha} = (-1)^i C_i^{\alpha}$, $i = 0, 1, \ldots, N-1$, and $\omega_i^{\alpha} = \big(1 - \frac{1+\alpha}{i}\big)\omega_{i-1}^{\alpha}$ (for $i > 0$). For $1 < \alpha < 2$, the coefficients $\{\omega_k^{\alpha}\}_{k=0}^{\infty}$ have the following properties

$$\omega_0^{\alpha} = 1, \quad \omega_1^{\alpha} = -\alpha < 0, \quad 1 \geq \omega_2^{\alpha} \geq \cdots \geq 0, \quad \sum_{k=0}^{\infty} \omega_k^{\alpha} = 0, \quad \sum_{k=0}^{p \geq 1} \omega_k^{\alpha} \leq 0.$$

By considering fractional derivatives along the x, y-direction as N equations, fractional derivative can be presented as following matrix form

$$D_x^{\alpha} G_l := \begin{pmatrix} D_x^{\alpha} g(x_0, y_l) \\ D_x^{\alpha} g(x_1, y_l) \\ \vdots \\ D_x^{\alpha} g(x_{N-1}, y_l) \end{pmatrix} = \begin{pmatrix} \omega_0^{\alpha} & 0 & \cdots & \cdots & 0 \\ \omega_1^{\alpha} & \omega_0^{\alpha} & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \omega_{N-1}^{\alpha} & \cdots & \cdots & \cdots & \omega_0^{\alpha} \end{pmatrix} \begin{pmatrix} g(x_0, y_l) \\ g(x_1, y_l) \\ \vdots \\ g(x_{N-1}, y_l) \end{pmatrix} = B_N^{\alpha} G_l.$$

Let $U \in \mathbb{R}^{N \times N}$ be the solution matrix at all nodes (kh, lh), $k, l = 0, 1, \ldots, N-1$. Then, fractional derivative in x, y-direction is defined as $D_x^{\alpha} U = B_N^{\alpha} U$ and $D_y^{\alpha} U = U (B_N^{\alpha})^T$, respectively. Similarly, the adjoint operators of fractional derivatives are defined as $D_x^{\alpha *} U = (-1)^m (B_N^{\alpha})^T U$ and $D_y^{\alpha *} U = (-1)^m U B_N^{\alpha}$. The standard definitions of fractional derivative require a function to have zero Dirichlet boundary conditions, but these conditions are restrictive and unrealistic in practice. Since the fractional derivative is non-local, any boundary condition can influence the whole image. Besides, because of using the left-sided GL derivative, there is no need to consider boundary conditions. By such an approach, results would be satisfying to some extent. Inspired by the discrepancy principle, a stopping criterion based on the mean and variance of multiplicative noise is developed that it is independent of the information of the original image and automatically controls iterative procedure. Based on this criterion, if the denoised image approximates the original image sufficiently, the variance of $\frac{f}{u}$ would get close to $\frac{1}{L}$. The stopping criterion defined as $n_* = \min\{n \in \mathbb{N} : R(f, u^n) > \frac{1}{L}\}$, where $R(f, u^n) = \frac{1}{|\Omega|} \int_{\Omega} (\frac{f}{u^n} - \text{mean}[\frac{f}{u^2}])^2 dx$, $\text{mean}[\frac{f}{u^2}] = \frac{1}{|\Omega|} \int_{\Omega} \frac{f}{u^2} dx$. We compare the new equation's results with the results of the model presented by Yao et al. [4].

To quantify the denoising effect for a noise-free image u_0 and its denoised image u, the denoising performance is measured in terms of peak signal to noise ratio (PSNR) and mean absolute-deviation error (MAE), i.e.

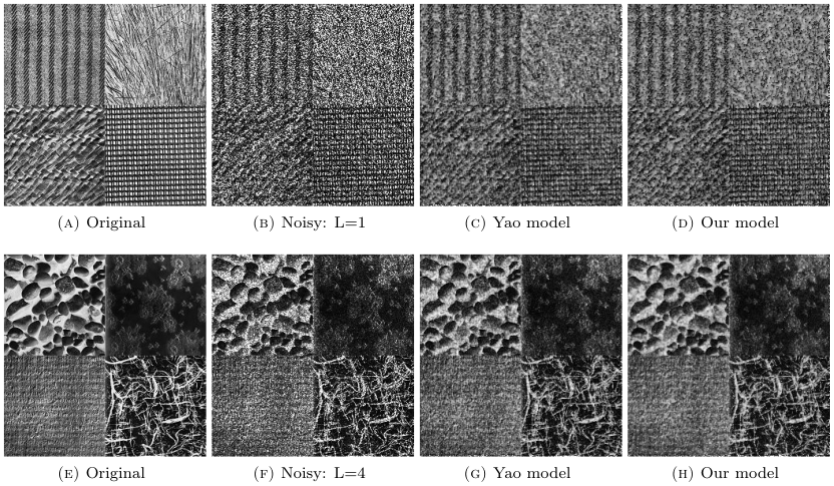
$$\begin{aligned} &\text{PSNR (dB)} = 10 \log_{10} \Big(\big((255)^2 M \times N\big) / \|u - u_0\|_{L^2}^2 \Big), \\ &\text{MAE (dB)} = \|u - u_0\|_{L^1} / \big(M \times N\big), \end{aligned}$$

where M and N are the image dimensions, and 255 is the peak signal with an 8-bit resolution. The higher value of PSNR, and the lower value of MAE, the closer

the denoised image is to the original image. For each image, a noisy observation is generated by multiplying the original image by a realization of noise according to the proposed model with the choice $L \in \{1, 4, 10\}$. We set $\alpha = 1.1$, $\Delta t = 0.02$, $r = 0.9$, $k = 0.001$, $\beta = 0.4$, $\sigma = 1$, and $\lambda = 0.6$. Experimental results demonstrate that the proposed model's PSNR and MAE values are higher and lower than Yao's results. Also, the mean value of PSNR and mean value of MAE are higher and lower, respectively.

TABLE 1. Comparison of PSNR and MAE of the different models.

	PSNR			MAE		
	the texture1 image (256×256)					
Model	L=1	L=4	L=10	L=1	L=4	L=10
Yao	12/5999	15/7955	18/2631	45/9349	31/5961	23/3952
Our	**12/7255**	**15/8659**	**18/3512**	**45/2586**	**31/1592**	**23/1081**
	the texture2 image (256×256)					
Yao	15/0453	18/2222	20/6703	32/2653	22/1883	16/4697
Our	**15/1611**	**18/3717**	**20/8467**	**31/6018**	**21/7086**	**16/1003**



(A) Original (B) Noisy: L=1 (C) Yao model (D) Our model

(E) Original (F) Noisy: L=4 (G) Yao model (H) Our model

FIGURE 1. Results to texture1 and texture2 images (256×256).

References

1. G. Aubert and J. F. Aujol, *A variational approach to removing multiplicative noise*, SIAM J. Appl. Math. **68** (4) (2008) 925–946.
2. P. Perona and J. Malik, *Scale-space and edge detection using anisotropic diffusion*, IEEE Trans. Pattern Anal. Mach. Intell. **12** (7) (1990) 629–639.
3. D. M. Strong and T. F. Chan, *Spatially and Scale Adaptive Total Variation Based Regularization and Anisotropic Diffusion in Image Processing*, Tech. Rep. CAM96-46, Univeristy of California, Los Angeles, Calif, USA, 1996.
4. W. Yao, Z. Guo, J. Sun, B. Wu and H. Gao, *Multiplicative noise removal for texture images based on adaptive anisotropic fractional diffusion equations*, SIAM J. Imaging Sci. **12** (2) (2019) 839–873.

E-mail: maryam.mohammady@math.iut.ac.ir
E-mail: mokhtari@iut.ac.ir
E-mail: nader.karimi@iut.ac.ir

A Meshless Method of Lines for the Multi-Term Time-Fractional Nonlinear Mixed Diffusion and Diffusion-Wave Equation

Tahereh Molaee*
Department of Mathematics, Faculty of Mathematical Sciences, University of Alzahra, Tehran, Iran
and Alimardan Shahrezaee
Department of Mathematics, Faculty of Mathematical Sciences, University of Alzahra, Tehran, Iran

Abstract. In this work, the multi-term time-fractional nonlinear mixed diffusion and diffusion-wave equation is considered. The time-fractional derivative is defined in Caputos sense. The spatial derivative is discretized based on finite difference and the numerical solution of nonlinear fractional ordinary differential equations system is approximated by using the radial basis functions. The numerical results demonstrate the effectiveness of the algorithm.
Keywords: Meshless method, Multi-term time-fractional equation, Mixed diffusion and diffusion-wave equation.
AMS Mathematical Subject Classification [2010]: 35R11, 65J15.

1. Introduction

Different sciences use fractional calculus in phenomena modeling to better describe unusual behaviors of the phenomena. Successful applications of fractional differential equation models are found in Physics, chemistry, geophysics, rheology, geology, robotics, engineering, bioengineering, medicine and finance [6]. Although the analytic solution of kinds of fractional differential equations can be obtained, it is often complex and difficult to evaluate. Therefore, the presentation of the numerical solution of such equations is of great importance. In recent years, the time-fractional mixed diffusion and diffusion-wave equation has been acquired attention [1, 2, 4]. However, to the best of our knowledge, in case of nonlinear with time-fractional multi-term, there is still no literature on time-fractional mixed diffusion and diffusion-wave equation.

In this work, we consider the following multi-term time-fractional nonlinear mixed diffusion and diffusion-wave equation

$$\sum_{r=1}^{s} a_r {}_0^c D_t^{\alpha_r} u(x,t) + \sum_{l=1}^{w} b_l {}_0^c D_t^{\beta_l} u(x,t) + \mathcal{N}(u(x,t)) = \triangle u(x,t) + f(x,t), \qquad (x,t) \in \Omega \times I, \tag{1}$$

*Speaker

subject to the initial and boundary conditions

$$u(x,0) = \zeta(x), \qquad \frac{\partial u(x,0)}{\partial t} = \xi(x), \qquad x \in \bar{\Omega}, \tag{2}$$

$$u(x,t) = \psi(t), \qquad x \in \partial\Omega, \tag{3}$$

where $a_r, b_l \geq 0$, $r = 1,\ldots,s$, $l = 1,\ldots,w$, $0 < \alpha_1 < \alpha_2 < \cdots < \alpha_s \leq 1$, $1 \leq \beta_1 < \beta_2 < \cdots < \beta_l \leq 2$, $\Omega = (0,L)$, $I = (0,T]$, $\triangle u = \dfrac{\partial^2 u}{\partial x^2}$. The nonlinear term $\mathcal{N}(u(x,t))$ satisfied the assumption $|\mathcal{N}(u)| \leq c|u|$, the functions $\zeta(x)$, $\xi(x)$, $\psi(t)$ and $f(x,t)$ are sufficiently smooth on a closed and bounded domain Ω, with Lipschitz boundary $\partial\Omega$. ${}_0^cD_t^\alpha$ and ${}_0^cD_t^\beta$ denote the Caputo fractional derivative defined in [1] as follows

$${}_0^cD_t^\alpha u(x,t) = \begin{cases} \frac{1}{\Gamma(1-\alpha)} \int_0^t (t-s)^{-\alpha} \frac{\partial u(x,s)}{\partial s} ds, & 0 < \alpha < 1, \\ \dfrac{du(x,t)}{dt}, & \alpha = 1, \end{cases}$$

$${}_0^cD_t^\beta u(x,t) = \begin{cases} \frac{1}{\Gamma(2-\beta)} \int_0^t (t-s)^{1-\beta} \frac{\partial^2 u(x,s)}{\partial s^2} ds, & 1 < \beta < 2, \\ \dfrac{d^2u(x,t)}{dt^2}, & \beta = 2. \end{cases}$$

The remaining part of this paper is summarized as follows. In Section 2, a numerical procedure is presented to calculate approximate solution. In Section 3, two numerical examples are considered and a conclusion is given in Section 4.

2. Numerical Procedure

In this section, the problem (2)-(3) is discretized in the space direction based on finite difference [3] and the numerical solution of nonlinear fractional ordinary differential equations system is approximated by using the radial basis functions [5] in the time direction. For this purpose, we consider spatial step size $h = \frac{L}{M}$, $x_i = ih$, $u_i = u(x_i,t)$, $i = 0,1,\ldots,M$. Evaluating the operator $\frac{\partial^2 u}{\partial x^2}$ using the second-order central difference for the space derivative at position x_i produces the nonlinear system of time-fractional equations

$$\begin{cases} \sum_{r=1}^{s} a_r {}_0^cD_t^{\alpha_r} u_i + \sum_{l=1}^{w} b_l {}_0^cD_t^{\beta_l} u_i + \mathcal{N}(u_i) \\ = \dfrac{u_{i-1} - 2u_i + u_{i+1}}{h^2} + f(x_i,t) + O(h^2), & i = 1,\ldots,M-1, \\ u(x_i,0) = \zeta(x_i), \dfrac{\partial u(x_i,0)}{\partial t} = \xi(x_i), & \\ u(0,t) = u_0 = \psi_0(t), & t \in I, \\ u(L,t) = u_M = \psi_L(t), & t \in I. \end{cases} \tag{4}$$

Now we approximate $u_i(t)$ by the radial basis functions (RBFs) method. For this purpose, we consider

$$\tilde{u}(t) \simeq \sum_{j=0}^{N} \lambda_{ij} \varphi_j(t) = \Lambda_i^T \phi(t), \tag{5}$$

where $\Lambda_i = [\lambda_{i0}, \lambda_{i1}, \ldots, \lambda_{iN}]^T$, $\phi(t) = [\varphi_0(t), \varphi_1(t), \ldots, \varphi_N(t)]^T$ and the RBF interpolant $\tilde{u}(t)$, interpolates the given function $u(t)$ at the interpolation points $\{t_\kappa\}_{\kappa=0}^N$, contained in sub-domain $\Omega_i = (x_i,t)$ by the radial basis functions $\varphi_j(t) = \varphi(\|t - t_j\|)$. These functions have different types, such as Gaussian functions,

$\varphi_j(t) = \exp(-\varepsilon d_j^2)$, inverse quadratic functions, $\varphi_j(t) = \frac{1}{\varepsilon + d_j^2}$ and multi quadratic functions, $\varphi_j(t) = \sqrt{\varepsilon + d_j^2}$, where $d_j = \|t - t_j\|$ and $\varepsilon \geq 0$ is the shape parameter. We substitute Eq. (5) in the problem (4), collocate at the interpolation points $\{t_\kappa\}_{\kappa=0}^{N}$, and express the resulting system in matrix form

$$AX = b. \tag{6}$$

The $((M+1)\times(N+1))\times((M+1)\times(N+1))$ matrix A, column vectors X and b with $(M+1)\times(N+1)$ entries in Eq. 6 are defined as follows

$$A = \begin{bmatrix} \phi(t) & 0 & 0 & \cdots & \cdots & 0 & 0 \\ 0 & \Xi & -\frac{1}{h^2}\phi(t) & \cdots & \cdots & 0 & 0 \\ 0 & -\frac{1}{h^2}\phi(t) & \Xi & -\frac{1}{h^2}\phi(t) & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 & 0 \\ 0 & 0 & 0 & -\frac{1}{h^2}\phi(t) & \Xi & -\frac{1}{h^2}\phi(t) & 0 \\ 0 & 0 & 0 & 0 & -\frac{1}{h^2}\phi(t) & \Xi & 0 \\ 0 & 0 & 0 & \cdots & \cdots & 0 & \phi(t) \end{bmatrix},$$

$$X = [\Lambda_0, \Lambda_1, \ldots, \Lambda_M]^T,$$

$$b = [\Psi_0(t), \mathbf{f}(x_1,t) + \Psi_0(t), \mathbf{f}(x_2,t), \ldots, \mathbf{f}(x_{M-1},t) + \Psi_L(t), \Psi_L(t)]^T,$$

where $\Xi = \Phi + \frac{2}{h^2}\phi(t) + \mathcal{N}(\phi(t))$, $\Phi = \sum_{r=1}^{s} a_r {}_0^c D_t^{\alpha_r}\phi(t) + \sum_{l=1}^{w} b_l {}_0^c D_t^{\beta_l}\phi(t)$, $\Psi_0(t) = [\psi_0(t_0), \psi_0(t_1), \ldots, \psi_0(t_N)]$ and $\mathbf{f}(x_i,t) = [f(x_i,t_0), f(x_i,t_1), \ldots, f(x_i,t_N)]$.

We obtain the numerical solution of problem (2)-(3) by solving the system of equations $AX = b$ in the MATLAB software.

3. Numerical Examples

For numerical purposes, we consider $L = 1$, $T = 1$ and apply the proposed procedure on two test problems by the Gaussian radial basis functions to demonstrate the efficiency and reliability of the method for problem (2)-(3). The numerical errors produced due to the performance of the scheme are measured with the root mean square (RMS) error and the maximum norm of errors (L_∞), defined as

$$RMS(u) = \sqrt{\frac{1}{N_t}\sum_{k=1}^{N_t}(u(x_k,t_k) - \tilde{u}(x_k,t_k))^2},$$

$$L_\infty = \|u(x,t) - \tilde{u}(x,t)\|_\infty = \max_{0\leq x\leq L}\max_{0\leq t\leq T}|u(x,t) - \tilde{u}(x,t)|,$$

where N_t is the total collocation points number in domain Ω, $u(x,t)$ and $\tilde{u}(x,t)$ are the exact and the numerical values at these points, respectively.

Example 3.1. We consider the following time-fractional nonlinear mixed diffusion and diffusion-wave equation

$${}_a^c D_t^{\alpha_1} u(x,t) + \frac{du(x,t)}{dt} + {}_a^c D_t^{\beta} u(x,t) + \exp(u(x,t)) = \frac{\partial^2 u}{\partial x^2} + f(x,t),$$

the corresponding forcing term, initial and boundary conditions can be obtained from the analytic solution

$$u(x,t) = t^{(3+\alpha+\beta)} \exp(-x^2).$$

Figure 1 displays the absolute error function with $M = 8$, $N = 3$, $\varepsilon = 5$, $\alpha = 0.1$ and $\beta = 1.1$ for the specific case of the Example 3.1 (without the term $\frac{du(x,t)}{dt}$).

EXAMPLE 3.2. In this test problem, we consider the multi-term time-fractional nonlinear mixed diffusion and diffusion-wave equation

$$\begin{aligned} {}_a^cD_t^{\alpha_1} u(x,t) \quad + \quad & {}_a^cD_t^{\alpha_2} u(x,t) + {}_a^cD_t^{\beta_1} u(x,t) + {}_a^cD_t^{\beta_2} u(x,t) + u^3(x,t) - u(x,t) \\ = \quad & \frac{\partial^2 u}{\partial x^2} + f(x,t), \end{aligned}$$

with the exact solution

$$u(x,t) = t^3 \sin(\pi x).$$

The absolute error function with $M = 8$, $N = 3$, $\varepsilon = 5$, $\alpha_1 = 0.2$, $\alpha_2 = 0.8$, $\beta_1 = 1.4$ and $\beta_2 = 1.8$ for Example 3.2 is given in Figure 1.

In [1], the maximum norm of errors in the final time has been obtained 3.6989×10^{-2} for the specific case of Example 3.2 (without the terms ${}_a^cD_t^{\alpha_2} u(x,t)$ and ${}_a^cD_t^{\beta_2} u(x,t)$) with $\alpha_1 = 0.05$, $\beta_1 = 1.6$, $\delta t = \frac{T}{N} = 0.2$, and in this paper, we obtain it 7.0566×10^{-3}.

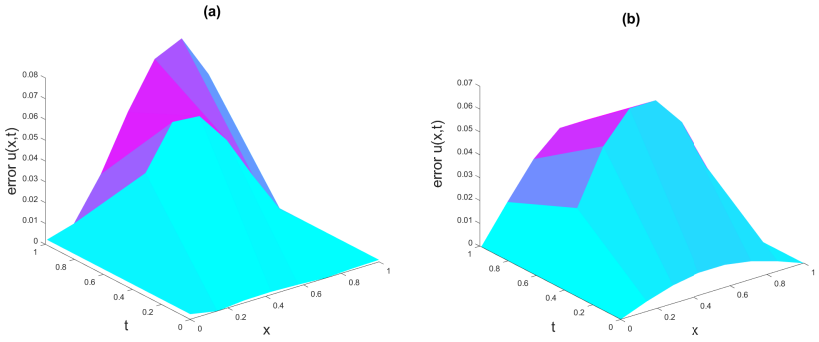


FIGURE 1. The error function u(x,t) with (a) $\alpha = 0.1$, $\beta = 1.1$ for Example 3.1 and (b) $\alpha_1 = 0.2$, $\alpha_2 = 0.8$, $\beta_1 = 1.4$ and $\beta_2 = 1.8$ for Example 3.2.

The numerical results with different values of α_r, β_l and same values for other parameters are reported in Table 1 for Example 3.1 and Example 3.2.

TABLE 1. Numerical results of Example 3.1 and Example 3.2.

	Example 3.1			Example 3.2	
α, β	RMS	L_∞	α_r,β_l	RMS	L_∞
$\alpha_1 = 0.1$			$\alpha_1 = 0.3$, $\alpha_2 = 0.8$		
$\beta = 1.08$	$5.26e-2$	$9.85588e-2$	$\beta_1 = 1.3$, $\beta_2 = 1.5$	$2.38e-2$	$8.48044e-2$
$\alpha_1 = 0.05$			$\alpha_1 = 0.1$, $\alpha_2 = 0.3$		
$\beta = 1.2$	$3.25e-2$	$9.52497e-2$	$\beta_1 = 1.1$, $\beta_2 = 1.8$	$1.68e-2$	$7.85724e-2$

4. Conclusion

In this work, we present a meshless method of lines for solving the multi-term time-fractional nonlinear mixed diffusion and diffusion-wave equation using radial basis functions on regular domain. Two test problems show the applicability and practical efficiency of this method for the problem (2)-(3) in terms of RMS error and the maximum norm of errors. The viewpoint of implementation, it is not difficult and with a few changes to the coefficient matrix, can be used for different types of problem (2)-(3) and its expanded forms, including the problem with higher dimensions, the problem with higher order derivatives with respect to x, the problem with more time-fractional terms.

References

1. A. Bhardwaj and A. Kumar, *A meshless method for time-fractional nonlinear mixed diffusion and diffusion-wave equation*, Appl. Numer. Math. **160** (2021) 146–165.
2. Z. Fu, L. Yang, H. Zhu and W. Xu, *A semi-analytical collocation Trefftz scheme for solving multi-term time fractional diffusion-wave equations*, Eng. Anal. Bound. Elem. **98** (2019) 137–146.
3. S. Kazem and M. Dehghan, *Semi-analytical solution for time-fractional diffusion equation based on finite difference method of line (MOL)*, Eng. Comput. **35** (2019) 229–241.
4. Y. Liu, H. Sun, X. Yin and L. Feng, *Fully discrete spectral method for solving a novel multi-term time-fractional mixed diffusion and diffusion-wave equation*, Z Angew Math. Phys. **71** (2020) 21.
5. J. Nazari, M. Nili Ahmadabadi and H. Almasieh, *The method of radial basis functions for the solution of nonlinear Fredholm integral equations system*, JLTA **6** (1) (2017) 11–28.
6. HG. Sun, Y. Zhang, D. Baleanu, W. Chen and YQ. Chen, *A new collection of real world applications of fractional calculus in science and engineering*, Commun. Nonlinear Sci. Numer. Simulat. **64** (2018) 213–231.

E-mail: t.molaee@alzahra.ac.ir
E-mail: ashahrezaee@alzahra.ac.ir

Realizable Interval List of Real Numbers by Interval Nonnegative Matrices via Lower Triangular Matrices

Ali Mohammad Nazari*
Department of Mathematics, Arak University, P. O. Box 38156-8-8349, Arak, Iran
Maryam Zeinali
Department of Mathematics, Shahid Rajaee Teacher Training University, Tehran, Iran
Hamid Mesgarani
Department of Mathematics, Shahid Rajaee Teacher Training University, Tehran, Iran
and Atiyeh Nezami
Department of Mathematics, Arak University, P. O. Box 38156-8-8349, Arak, Iran

Abstract. In this paper for a given set of real interval numbers σ that has one positive interval number and nonnegative summation, we find an interval nonnegative matrix C^I such that for each point set δ of given interval spectrum σ, there exists a point matrix C of C^I such that δ is its spectrum. For this purpose, we use unit lower triangular matrices and specially try to use binary unit lower triangular matrices. We also study some conditions for existence solution of the problem.
Keywords: Interval arithmetic, Interval matrix, Inverse eigenvalue problem, Nonnegative matrices.
AMS Mathematical Subject Classification [2010]: 15A18, 15A60, 15A09.

1. Introduction

A matrix L is *unit lower triangular* provided each entry on its main diagonal equals 1, and each entry above its main diagonal is zero. The inverse of a unit lower triangular matrix also is unit lower triangular and is easy to calculate. In Gaussian elimination method and LU factorization unit lower triangular matrices play a very important role. The binary unit lower triangular matrices is a unit lower triangular matrices that all entries below its main diagonal are 0 or 1.

An interval matrix is a matrix whose entries are interval numbers. The use of interval numbers began in the first half of the twentieth century and is expanding every day. In 1965, logic was fuzzy by Zadeh and interval numbers were used [1]. In 1993 J. Rohn found the inverse of interval matrices [2]. The problem of finding the eigenvalue of interval matrices is one of the most pressing issues for mathematicians, and several papers have been written in recent years, for example [5, 6, 7, 8]. In 2018 Nazari et al. started the inverse eigenvalue problem of nonnegative interval matrices, which is briefly denoted by NIIEP [9]. They solved NIIEP for matrices of order at most 3. In this paper by helping unit lower triangular matrices and use similarity of matrices we try to solve the problem for

*Speaker

order greater than 3. Nazari and Nezami solved NIEP for any order via unit lower triangular matrices [10].

When we say that the interval spectrum σ is realizable by interval matrix C^I or interval set σ is spectrum of interval matrix C^I, it means, we can find an interval nonnegative matrix C^I such that for every point set δ of interval set of eigenvalues σ (for each interval element one point), there exists a point nonnegative matrix C of C^I such that δ is its spectrum.

Now we recall some definition of interval analysis and interval matrices. The summation, subtraction, multiplication and division of two interval numbers $\mathbf{b} = [\underline{b}, \overline{b}], \mathbf{a} = [\underline{a}, \overline{a}]$ respectively, are defined as:

- $\mathbf{a} + \mathbf{b} = [\underline{a} + \underline{b}, \overline{a} + \overline{b}]$,
- $\mathbf{a} - \mathbf{b} = [\underline{a} - \overline{b}, \overline{a} - \underline{b}]$,
- $\mathbf{a} \cdot \mathbf{b} = [\min\{\underline{a} \cdot \underline{b}, \underline{a} \cdot \overline{b}, \overline{a} \cdot \underline{b}, \overline{a} \cdot \overline{b}\}, \max\{\underline{a} \cdot \underline{b}, \underline{a} \cdot \overline{b}, \overline{a} \cdot \underline{b}, \overline{a} \cdot \overline{b}\}]$,
- $\frac{\mathbf{a}}{\mathbf{b}} = \mathbf{a} \cdot \mathbf{b}^{'}$, $\mathbf{b}^{'} = [\frac{1}{\overline{b}}, \frac{1}{\underline{b}}]$, and $0 \notin \mathbf{b}$,

also the square of a interval number $\mathbf{a} = [\underline{a}, \overline{a}]$ is as following

- $\mathbf{a}^2 = \begin{cases} [\underline{a}^2, \overline{a}^2], & \text{if} \quad 0 \leq \underline{a} \leq \overline{a}, \\ \left[\overline{a}^2, \underline{a}^2\right], & \text{if} \quad \underline{a} \leq \overline{a} \leq 0, \\ \left[0, \max\{[\underline{a}^2, \overline{a}^2\}\right], & \text{if} \quad \underline{a} \leq 0 \leq \overline{a}. \end{cases}$

Definition 1.1. Let $\underline{A}$ and $\overline{A}$ be $n \times n$ real matrices, the following set

$$A^I = [\underline{A}, \overline{A}] = \{A : \underline{A} \leq A \leq \overline{A}\},$$

is called an $n \times n$ real interval matrix. The midpoint and the radius of A^I are denoted respectively by

$$A_c = \frac{\underline{A} + \overline{A}}{2}, \qquad A_\Delta = \frac{\underline{A} - \overline{A}}{2}.$$

If all interval entries of a real interval matrix ≥ 0, then A^I is called nonnegative interval matrix. The set of all real $n \times n$ interval matrices denoted by $\mathbb{IR}^{n \times n}$ and the set of all $n \times n$ nonnegative interval matrices also denoted by $\mathbb{NIR}^{n \times n}$.

Definition 1.2. Let A^I be an interval square matrix then the set of eigenvalues of A^I is defined as follows

$$\Lambda(A^I) = \{\lambda \in \mathbb{R}; Ax = \lambda x, x \neq 0, A \in A^I\}.$$

The eigenvalue of $n \times n$ nonnegative interval matrix A^I is called Perron interval eigenvalue of A^I if it is nonnegative and greater than or equal of all absolute value of eigenvalues of A^I and denoted by $\lambda_1 = [\underline{\lambda_1}, \overline{\lambda_1}]$. i.e.,

$$[\underline{\lambda_1}, \overline{\lambda_1}] \geq \left|[\underline{\lambda_i}, \overline{\lambda_i}]\right|, \qquad i = 2, 3, \ldots, n,$$

where $\left|[\underline{\lambda_i}, \overline{\lambda_i}]\right| = [\min\{\left|\underline{\lambda_i}\right|, \left|\overline{\lambda_i}\right|\}, \max\{\left|\underline{\lambda_i}\right|, \left|\overline{\lambda_i}\right|\}]$.

Some necessary conditions for NIIEP on the list of complex interval number

$$\sigma = \left\{[\underline{\lambda_1}, \overline{\lambda_1}], [\underline{\lambda_2}, \overline{\lambda_2}], \ldots, [\underline{\lambda_n}, \overline{\lambda_n}]\right\},$$

to be the spectrum of a nonnegative interval matrix are listed below.
(1) The Perron eigenvalue $\max\{|[\underline{\lambda_i}, \overline{\lambda_i}]|; [\underline{\lambda_i}, \overline{\lambda_i}] \in \sigma\}$ belongs to σ (Perron-Frobenius theorem in interval case).
(2) The list σ is closed under complex conjugation.

(3) $s_k = \sum_{i=1}^{n} \left|[\underline{\lambda_i}, \overline{\lambda_i}]\right|^k \geq 0$.
(4) $)s_k^m \leq n^{m-1} s_{km}$ for $k, m = 1, 2, \ldots$ (JLL inequality in interval case)[3, 4].

The paper is organized as follows. First we solve the NIIEP in several cases, where each element of σ is real, and σ has at least as many negative eigenvalues as positive eigenvalues. Then we solve the NIIEP in several cases where each element of σ is real and the number of negative elements of σ is less than the number of positive elements of σ.

2. Interval Real Spectrum

Let $k \leq 3$ and $\sigma^I = \left\{[\underline{\lambda_1}, \overline{\lambda_1}], [\underline{\lambda_2}, \overline{\lambda_2}], \ldots, [\underline{\lambda_n}, \overline{\lambda_n}]\right\}$ be a given spectrum such that $[\underline{\lambda_1}, \overline{\lambda_1}] \geq [\underline{\lambda_2}, \overline{\lambda_2}] \geq \cdots \geq [\underline{\lambda_k}, \overline{\lambda_k}] \geq 0 > [\underline{\lambda_n}, \overline{\lambda_n}] \geq \cdots \geq [\underline{\lambda_{k+1}}, \overline{\lambda_{k+1}}]$. We try to construct a nonnegative interval matrix C^I such that it realizes spectrum σ. At first we solve the interval spectrurm of Suleimanova. This spectrum has one positive eigenvalue and negative another eigenvalues with nonnegative summation. Suleimanova's Theorem [11] is in the interval case, which is proved below:

THEOREM 2.1. *Assume that given* $\sigma^I = \left\{[\underline{\lambda_1}, \overline{\lambda_1}], [\underline{\lambda_2}, \overline{\lambda_2}], \ldots, [\underline{\lambda_n}, \overline{\lambda_n}]\right\}$ *such that* $[\underline{\lambda_1}, \overline{\lambda_1}] \geq 0 \geq [\underline{\lambda_n}, \overline{\lambda_n}] \geq [\underline{\lambda_{n-1}}, \overline{\lambda_{n-1}}] \geq \cdots \geq [\underline{\lambda_2}, \overline{\lambda_2}]$, *and* $\sum_{i=1}^{n} [\underline{\lambda_i}, \overline{\lambda_i}] \geq 0$, *then there exists a set of nonnegative interval matrices that* σ *is its spectrum.*

PROOF. If characteristic polynomial of interval matrix will be as

$$P(\lambda) = \prod_{i=1}^{n} \left(\lambda - [\underline{\lambda_i}, \overline{\lambda_i}]\right) = \lambda^n - [\underline{a_{n-1}}, \overline{a_{n-1}}]\lambda^{n-1} - [\underline{a_{n-2}}, \overline{a_{n-2}}]\lambda^{n-2} - \cdots - [\underline{a_0}, \overline{a_0}],$$

and all $[\underline{a_i}, \overline{a_i}] \geq 0$ for $i = 0, 1, \cdots, n-1$, then it is easy to see that the following nonnegative interval companion matrix is solution of problem

$$C^I = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ & & \ddots & & \\ 0 & 0 & 0 & \cdots & 1 \\ [\underline{a_0}, \overline{a_0}] & [\underline{a_1}, \overline{a_1}] & & \cdots & [\underline{a_{n-1}}, \overline{a_{n-1}}] \end{pmatrix}.$$

On the other hand, we construct the solution via unit lower triangular matrix. Let $n = 2$, then consider the upper interval triangular matrix

$$A^I = \begin{pmatrix} [\underline{\lambda_1}, \overline{\lambda_1}] & \alpha_2 \\ 0 & [\underline{\lambda_2}, \overline{\lambda_2}] \end{pmatrix},$$

where $\alpha_2 = [\underline{\alpha_2}, \overline{\alpha_2}]$ is interval number and also we consider 2×2 unite lower triangular matrix $L = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$, therefore by similarity of matrices the following matrix

$$C^I = LA^I L^{-1} = \begin{pmatrix} [\underline{\lambda_1}, \overline{\lambda_1}] - \alpha_2 & \alpha_2 \\ [\underline{\lambda_1}, \overline{\lambda_1}] - \alpha_2 - [\underline{\lambda_2}, \overline{\lambda_2}] & \alpha_2 + [\underline{\lambda_2}, \overline{\lambda_2}] \end{pmatrix},$$

has eigenvalues $[\underline{\lambda_1}, \overline{\lambda_1}]$ and $[\underline{\lambda_2}, \overline{\lambda_2}]$ and if $-\overline{\lambda_2} \leq \underline{\alpha_2} \leq \overline{\alpha_2} \leq \underline{\lambda_1}$ then the matrix C^I is nonnegative.

For $n = 3$ we consider

$$A^I = \begin{pmatrix} [\underline{\lambda_1}, \overline{\lambda_1}] & \alpha_2 & \alpha_3 \\ 0 & [\underline{\lambda_2}, \overline{\lambda_2}] & 0 \\ 0 & 0 & [\underline{\lambda_3}, \overline{\lambda_3}] \end{pmatrix},$$

where $\alpha_2 = [\underline{\alpha_2}, \overline{\alpha_2}]$, $\alpha_3 = [\underline{\alpha_3}, \overline{\alpha_3}]$ are interval numbers and assume that

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix},$$

then the matrix

$$C^I = LA^I L^{-1} =$$

$$\begin{pmatrix} [\underline{\lambda_1}, \overline{\lambda_1}] - \alpha_2 - \alpha_3 & \alpha_2 & \alpha_3 \\ [\underline{\lambda_1}, \overline{\lambda_1}] - \alpha_2 - [\underline{\lambda_2}, \overline{\lambda_2}] - \alpha_3 & \alpha_2 + [\underline{\lambda_2}, \overline{\lambda_2}] & \alpha_3 \\ [\underline{\lambda_1}, \overline{\lambda_1}] - \alpha_2 - \alpha_3 - [\underline{\lambda_3}, \overline{\lambda_3}] & \alpha_2 & \alpha_3 + [\underline{\lambda_3}, \overline{\lambda_3}] \end{pmatrix},$$

is similar to the matrix A^I and if $-\overline{\lambda_2} \leq \underline{\alpha_2}$ and $-\overline{\lambda_3} \leq \underline{\alpha_3}$ and also $\overline{\alpha_2} + \overline{\alpha_3} \leq \underline{\lambda_1}$ then the matrix C^I is nonnegative.

To continue the proof, we follow the above process. Consider

$$A^I = \begin{pmatrix} [\underline{\lambda_1}, \overline{\lambda_1}] & \alpha_2 & \alpha_3 & \cdots & \alpha_n \\ 0 & [\underline{\lambda_2}, \overline{\lambda_2}] & 0 & \cdots & 0 \\ & & \ddots & & \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \cdots & [\underline{\lambda_n}, \overline{\lambda_n}] \end{pmatrix},$$

similarly $\alpha_2 = [\underline{\alpha_2}, \overline{\alpha_2}]$, $\alpha_3 = [\underline{\alpha_3}, \overline{\alpha_3}]$, $\ldots$, $\alpha_n = [\underline{\alpha_n}, \overline{\alpha_n}]$ are interval numbers and

$$L = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ & & \ddots & & \\ 1 & 0 & 0 & \ddots & 0 \\ 1 & 0 & 0 & \cdots & 1 \end{pmatrix}, \tag{1}$$

then the matrix

$$C^I = LA^I L^{-1} =$$

$$\begin{pmatrix} [\underline{\lambda_1}, \overline{\lambda_1}] - t & \alpha_2 & \alpha_3 & \cdots & \alpha_n \\ [\underline{\lambda_1}, \overline{\lambda_1}] - [\underline{\lambda_2}, \overline{\lambda_2}] - t & \alpha_2 + [\underline{\lambda_2}, \overline{\lambda_2}] & \alpha_3 & \cdots & \alpha_n \\ & & \ddots & & \\ [\underline{\lambda_1}, \overline{\lambda_1}] - [\underline{\lambda_{n-1}}, \overline{\lambda_{n-1}}] - t & \alpha_2 & \alpha_3 & \ddots & \alpha_n \\ [\underline{\lambda_1}, \overline{\lambda_1}] - [\underline{\lambda_n}, \overline{\lambda_n}] - t & \alpha_2 & \alpha_3 & \cdots & \alpha_n + [\underline{\lambda_n}, \overline{\lambda_n}] \end{pmatrix},$$

with $t = \sum_{i=2}^{n} \alpha_i$ is similar to the matrix A^I, and if

$$-\underline{\lambda_i} \leq \underline{\alpha_i}, \qquad i = 2, 3, \ldots, n,$$
$$\overline{\alpha_2} + \overline{\alpha_3} + \cdots + \overline{\alpha_n} \leq \underline{\lambda_1},$$

then the interval matrix C^I is nonnegative and has eigenvalues

$$\left\{ \left[\underline{\lambda_1}, \overline{\lambda_1}\right], \left[\underline{\lambda_2}, \overline{\lambda_2}\right], \cdots, \left[\underline{\lambda_n}, \overline{\lambda_n}\right] \right\}.$$

□

EXAMPLE 2.2. For the following interval set of eigenvalues find an interval matrix C^I such that realize this set.

$$\sigma^I = \{[14,17], [-4,-3], [-3,-2], [-2,-1], [-1,0]\}.$$

All of necessary conditions satisfy. At first we find the characteristic polynomial

$$P(\lambda) = \lambda^5 - [4,11]\lambda^4 - [49,159]\lambda^3 - [104,589]\lambda^2 - [60,850]\lambda - [0,408].$$

Because all coefficients of the above polynomial except λ^5 are negative, the following nonnegative interval companion matrix realizes the subset of σ

$$C^I = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ [0,408] & [60,850] & [104,589] & [49,159] & [4,11] \end{pmatrix}.$$

Also with choosing the interval matrix A^I as

$$A^I = \begin{pmatrix} [14,17] & [4,5] & [3,4] & [2,3] & [1,2] \\ 0 & [-4,-3] & 0 & 0 & 0 \\ 0 & 0 & [-3,-2] & 0 & 0 \\ 0 & 0 & 0 & [-2,-1] & 0 \\ 0 & 0 & 0 & 0 & [-1,0] \end{pmatrix},$$

and the point matrix L from (1) for $n = 6$, we can find the nonnegative interval matrix $C^I = LA^IL^{-1}$ as following

$$C^I = \begin{pmatrix} [0,7] & [4,5] & [3,4] & [2,3] & [1,2] \\ [3,11] & [0,2] & [3,4] & [2,3] & [1,2] \\ [2,10] & [4,5] & [3,4] & [2,3] & [1,2] \\ [1,9] & [4,5] & [3,4] & [2,3] & [1,2] \\ [0,8] & [4,5] & [3,4] & [2,3] & [0,2] \end{pmatrix},$$

that σ is its spectrum.

References

1. L. A. Zadeh, *Fuzzy sets*, Inform. Contr. **8** (3) (1965) 338–353.
2. J. Rohn, *Inverse interval matrix*, SIAM J. Numer. Anal. **30** (3) (1993) 864–870.
3. C. R. Johnson, *Stochastic matrices similar to doubly stochastic matrices*, Linear Multilinear Algebra **10** (2) (1981) 113–130.
4. R. Loewy and D. London, *Note on an inverse problem for nonnegative matrices*, Linear Multilinear Algebra **6** (1978/79) 83–90.
5. M. Hladik, D. Daney and E. Tsigaridas, *Bounds on real eigenvalues and singular values of interval matrices*, SIAM J. Matrix Anal. Appl. **31** (4) (2009/10) 2116–2129.
6. R. Jiri, *Perron vectors of an irreducible nonnegative interval matrix*, Linear Multilinear Algebra (54) (2006) 399–404.
7. M. Hladik, D. Daney and E. Tsigaridasl, *A filtering method for the interval eigenvalue problem*, Appl. Math. Comput. **217** (2011) 5236–5242.

8. M. Hladik, *On eigenvalues of real and complex interval matrices*, Appl. Math. Comput. **219** (2013) 5584–5591.
9. A. Nazari, M. Zeinali and H. Mesgarani, *Inverse eigenvalue problem of interval nonnegative matrices of order* ≤ 3, Appl. Math. Model. **6** (2) (2018) 187–194.
10. A. Nazari and A. Nezami, *Inverse eigenvalues problem of nonnegative matrices via unit lower triangular matrices*, (2018). `arxiv 1805.07716`
11. H. R. Suleimanova, *Stochastic matrices with real characteristic numbers*, Dokl. Akad. Nauk. **66** (1949) 343–345.

E-mail: a-nazari@araku.ac.ir
E-mail: m.zeinali64@yahoo.com
E-mail: Hmesgarani@sru.ac.ir
E-mail: atiyeh.nezami@gmail.com

A Meshless Partition of Unity Method for Electromagnetic Scattering Problem of Anisotropic Obstacle

Marzieh Raei*
Malek Ashtar University of Technology, Isfahan, Iran

Abstract. In this work, the partition of unity method based on radial basis functions as an efficient local meshless technique is examined to solve an interesting electromagnetic scattering problem. In such a problem, the scattering from infinitely long anisotropic cylinder with circular cross-section embedded in free space is investigated. The numerical results demonstrate the efficiency and accuracy of the suggested method.
Keywords: Local meshless method, Radial basis function, Partition of unity method, Electromagnetic scattering problem.
AMS Mathematical Subject Classification [2010]: 13F55, 05E40, 05C65.

1. Introduction

In recent decades, there has been a significant amount of interest in electromagnetic interaction with anisotropic materials due to its wide application in various areas such as radar cross section (RCS) control, certain types of radar absorbers, antennas and optical signal processing. Taflove et al. [1] applied the finite difference time domain method for computing the electromagnetic scattering by arbitrary-shaped anisotropic dielectric objects. Graglia et al. [2] formulated integro-differential equations for the electric and magnetic fields inside scatterers of arbitrary shape and made of anisotropic materials for two and three dimensional problems. Chen et al. [3] employed the finite-difference with the measured equation of invariance for transversally anisotropic, inhomogeneous cylinders. Esfahani et al. [4] employed a meshless method for the electromagnetic scattering problem by anisotropic cylinders.

In the current work, consider $D \subset \mathbb{R}^2$ be the cross section of an infinitely long anisotropic dielectric cylindrical scattering contained by free space as well as n is outward normal vector on boundary ∂D. The anisotropy of the scatterer is described by a symmetric positive definite matrix A, whose entries are relative magnetic permeability inside the scatterer. Furthermore, the scatterer is excited by the $e^{i\omega t}$ time-harmonic incident plane wave with TM^z polarization, where ω demonstrate the angular frequency. The direct scattering problem for an anisotropic medium with cross section D in $\mathbb{R}^2$ is formulated as a system of complex PDEs:

$$\nabla.A\nabla v + \epsilon_r k^2 v = 0 \quad \text{in} \quad D,$$
$$\nabla^2 u^s + k^2 u^s = 0 \quad \text{in} \quad \mathbb{R}^2 \backslash D,$$

*Speaker

with boundary conditions

$$\begin{aligned} &v - u^s = u^i \quad \text{on} \quad \partial D, \\ &\frac{\partial v}{\partial n_A} - \frac{\partial u^s}{\partial n} = \frac{\partial u^i}{\partial n} \quad \text{on} \quad \partial D, \\ &\frac{\partial u^s}{\partial n} + (jk + \frac{1}{2r_2})u^s = 0 \quad \text{on} \quad \partial \Omega, \end{aligned}$$

where $v \in C^2(\mathbb{C})$ and $u \in C^2(\mathbb{C})$ indicate the field inside and the field outside of the scatterer. Field u can be further decomposed into the incident u^i and the scattered field u^s, where they are the scalar scattered and incident electric fields respectively. Moreover, $k = \omega\sqrt{\mu_0 \epsilon_0} = \frac{2\pi}{\lambda}$ is the wave number of the free space which λ is the wavelength as well as ϵ_0 and μ_0 are the magnetic permeability and the constant electric permittivity of free space while ϵ_r is the relative electric permittivity of the scatterer. Also, matrix A is defined by:

$$A = \frac{1}{\mu_{xx}\mu_{yy} - \mu_{xy}^2} \begin{bmatrix} \mu_{xx} & \mu_{xy} \\ \mu_{xy} & \mu_{yy} \end{bmatrix}.$$

Let A is a symmetric and positive definite 2×2 matrix whose entries are constants and demonstrate the relative magnetic permeability inside the scatterer.

2. Spatial Discretization Scheme

In this section, the radial basis function partition of unity method (PUM) is introduced [5]. Therefore it is necessary to review the radial basis function collocation method firstly. Then the PUM can be investigated by some details. Consider the entire computational domain $\Omega \subseteq \mathbb{R}^d$ is covered by a set of scattered points $\mathcal{X}_N = \{\mathbf{x}_j\}_{j=1}^N \subset \Omega$. According to the mesless collocation method based on radial basis functions (RBFs), the approximation of function $u(x)$ on the set $\mathcal{X}_N$ leads to find an interpolant

$$S(\mathbf{x}) = \sum_{j=1}^{N} \lambda_j \phi(\|\mathbf{x} - \mathbf{x}_j\|), \tag{1}$$

where $\{\lambda_j\}_{j=1}^N$ are unknown coefficients, $\phi : \Omega \times \Omega \to \mathbb{R}$ is a smooth strictly positive definite (SPD) RBFs and $\|.\|$ denotes the Euclidean norm. Imposing the interpolation conditions

$$S(\mathbf{x}_i) = u(\mathbf{x}_i), \quad i = 1, \ldots, N,$$

to determine the coefficients $\{\lambda_j\}_{j=1}^N$, the linear system of equations is obtained

$$A\boldsymbol{\lambda} = \mathbf{u}, \tag{2}$$

where $A_{ij} = \phi(\|\mathbf{x}_i - \mathbf{x}_j\|), \quad i, j = 1, \ldots, N$, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N)^T$ and

$$\mathbf{u} = (u(\mathbf{x}_1), \ldots, u(\mathbf{x}_N))^T.$$

Since the radial kernel ϕ is SPD, the linear algebraic system (2) has a unique solution.

In the following, the interpolant (1) is expressed in Lagrange form, i.e.,

$$S(\mathbf{x}) = \sum_{j=1}^{N} l_j(\mathbf{x}) u(\mathbf{x}),$$

where $l_j(x)$ are so-called cardinal or Lagrange basis functions. Therefore, the alternative formulation for the interpolant (1) can be obtained as follows:

$$S(\mathbf{x}) = L(\mathbf{x})^T \mathbf{u}, \tag{3}$$

where $L(\mathbf{x})^T = (l_1(\mathbf{x}), \ldots, l_N(\mathbf{x}))$. Clearly, from (1), (2) and (3), the following relation between primary radial basis and cardinal basis is deduced

$$L(\mathbf{x})^T = \boldsymbol{\phi}^T(\mathbf{x}) A^{-1},$$

where $\boldsymbol{\phi}^T(\mathbf{x}) = (\phi(\|\mathbf{x} - \mathbf{x}_1\|), \ldots, \phi(\|\mathbf{x} - \mathbf{x}_N\|))$. This transformation is valid whenever the matrix A is invertible, i.e. for given distinct data points $\mathcal{X}$ and SPD radial basis functions.

In the following, the RBF-PUM will be discussed. Let Ω be an open bounded global domain and let Ω_j be open and bounded patches covering of Ω such that $\Omega \subseteq \bigcup_{j=1}^{M} \Omega_j$. Moreover, patches Ω_j satisfy some mild overlap condition. This means that each point of the global computational domain must be in the interior of at least one local subdomain. Further, the set $\mathcal{I}(\mathbf{x}) = \{j : \mathbf{x} \in \Omega_j\}$, for all $\mathbf{x} \in \Omega$, is uniformly bounded on Ω, i.e., there is the constant $\mathcal{C}$ independent of the number of patches, such that $card(\mathcal{I})(\mathbf{x}) \leq \mathcal{C}$. For each patch, the PU weight function ω_j is constructed by using the Shepard method given by

$$\omega_j(\mathbf{x}) = \frac{\varphi_j(\mathbf{x})}{\sum_{k \in \mathcal{I}(\mathbf{x})} \varphi_k(\mathbf{x})}, \quad j = 1, 2, \ldots, M,$$

where $\varphi_j(\mathbf{x})$ is the compactly supported function on Ω_j. The weight functions ω_j are non-negative, compactly supported on Ω_j, and satisfy the partition of unity property, i.e.,

$$\sum_{j \in \mathcal{I}(\mathbf{x})} \omega_j(\mathbf{x}) = 1.$$

Moreover, to ensure that weight function is non-negative and compactly support on Ω_j, the function $\varphi_j(\mathbf{x})$ is defined as follows:

$$\varphi_j(\mathbf{x}) = \varphi(\frac{\|\mathbf{x} - \xi_j\|}{R_j}), \quad j = 1, 2, \ldots, M,$$

where ξ_j and R_j are the center and radius corresponding to the $j-$th patch, respectively. In the current work, the Wendland C^2 function is used to construct the weight function.

The PUM approximation is formed a global approximation function $\mathcal{P}$ of function $u(\mathbf{x})$ in entire domain Ω as follows:

$$\mathcal{P}_u(\mathbf{x}) = \sum_{j=1}^{M} \omega_j(\mathbf{x}) S_j(\mathbf{x}) = \sum_{j \in \mathcal{I}(\mathbf{x})} \omega_j(\mathbf{x}) S_j(\mathbf{x}),$$

where $\{S_j\}_{j=1}^{N}$ are RBF based local interpolants corresponding to each patch Ω_j.

3. Numerical Results

In this section, a numerical example is considered by using uniform points to show the efficiency and accuracy of the suggested meshless method. For this purpose, the root mean square error (RMSE) and absolute error (MAXE) is applied to make comparison. In numerical implementation, inverse multiquadric (IMQ) radial basis function is used.

EXAMPLE 3.1. As an example, consider the two-dimensional electromagnetic scattering problem of circular cross section object with radius r_1 such that $kr_1 = 1$, $\epsilon_r = 1$ and anisotropic matrix $A = \frac{1}{2}\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$.

Also, The incident angle is $\theta_i = 0$ and the radius of artificial circular domain is $r2 = 2r_1$. The suggested numerical method is employed to solve the problem. The resulting error estimates, condition numbers and CPU times for various number of computational points N_v and N_us is reported in Table 1.

Moreover, all results from PUM have compared with the radial basis function-finite difference (RBF-FD) method in this table. These results show the convergency and efficiency of the suggested method for solving such problems. Also, the Figure 1 demonstrates the cross section of the circular dielectric and the computational points on domains. Furthermore, Figure 2 shows the approximated solution inside and outside of the anisotropic circular object by letting $N_v = 50$ and $N_{u^s} = 167$.

TABLE 1. Error estimates, condition numbers (CN) and CPU times for different number of computational points.

		suggested method (PUM)				RBF-FD method			
N_v / N_{u^s}	c_v / c_{u^s}	$MAXE_v$ / $MAXE_{u^s}$	$RMSE_v$ / $RMSE_{u^s}$	CN	$CPU\ time$	$MAXE_v$ / $MAXE_{u^s}$	$RMSE_v$ / $RMSE_{u^s}$	CN	$CPU\ time$
22 / 60	1.6 / 2.0	8.9442×10^{-3} / 8.9442×10^{-3}	6.4075×10^{-3} / 5.1251×10^{-3}	1.1034×10^{5}	0.1538	4.1318×10^{-2} / 2.1661×10^{-1}	3.0967×10^{-2} / 2.7510×10^{-2}	3.7512×10^{7}	0.5283
50 / 167	2.0 / 2.5	1.3286×10^{-3} / 1.3286×10^{-3}	6.2108×10^{-4} / 3.7167×10^{-4}	5.8430×10^{6}	0.6047	1.7342×10^{-2} / 1.5749×10^{-1}	1.0139×10^{-2} / 1.2451×10^{-2}	8.9136×10^{7}	1.8645
82 / 245	2.4 / 3.0	5.3038×10^{-4} / 5.3038×10^{-4}	2.6308×10^{-4} / 2.3354×10^{-4}	4.8897×10^{7}	1.5784	3.1954×10^{-3} / 3.0603×10^{-2}	1.9726×10^{-3} / 1.9795×10^{-3}	3.7150×10^{8}	4.7217
142 / 420	2.8 / 3.5	1.9472×10^{-5} / 1.9472×10^{-5}	9.9756×10^{-6} / 7.1752×10^{-6}	1.1135×10^{10}	5.5564	2.1176×10^{-3} / 3.3179×10^{-3}	1.2371×10^{-3} / 1.6737×10^{-3}	1.3332×10^{10}	14.2053
182 / 558	3.2 / 4.0	1.2937×10^{-5} / 1.2937×10^{-5}	8.3350×10^{-6} / 6.5340×10^{-6}	4.5436×10^{11}	30.9125	1.9572×10^{-3} / 2.1832×10^{-3}	1.0861×10^{-3} / 1.0916×10^{-3}	3.3612×10^{11}	43.4647

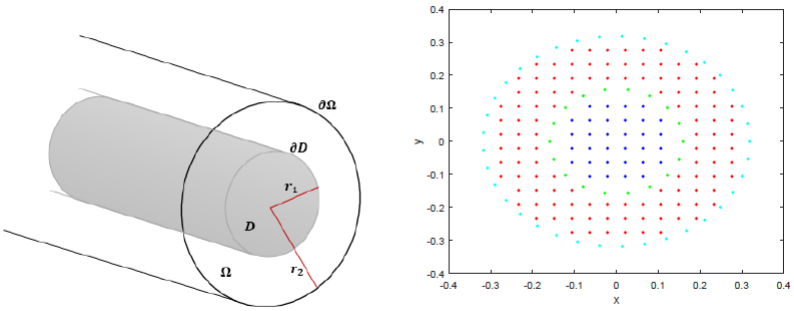


FIGURE 1. Cross section of an infinitely long dielectric cylindrical scattering (a) and computational points on domains (b).

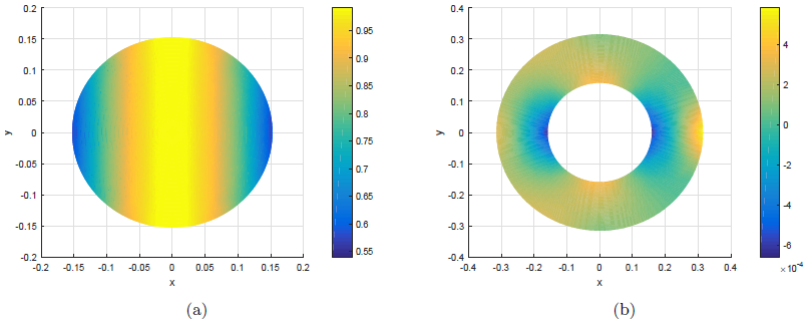


FIGURE 2. Approximated solution v inside the anisotropic dielectric circular cylinder (a) and approximated solution u^s outside the anisotropic dielectric circular cylinder (b).

References

1. A. Taflove and M. E. Brodwin, *Numerical solution of steady-state electromagnetic scattering problems using the time dependent Maxwell's equations*, IEEE Trans. Microw Theory Tech. **23** (8) (1975) 623–630.
2. R. D. Graglia and P. L. E. Vslenghi, *Electromagnetic scattering from anisotropic materials, part I: General theory*, IEEE Trans. Antennas Propag. **32** (1984) 867–869.
3. Z. N. Chen, W. Hong and W. X. Zhang, *Application of FD-MEI to electromagnetic scattering from transversally anisotropic inhomogeneous cylinders*, IEEE Trans. Electromagnetic C. **40** (2) (1998) 103–110.
4. M. H. Esfahani, H. R. Ghehsareh and S. K. Etesami, *The extended method of approximate particular solutions to simulate two-dimensional electromagnetic scattering from arbitrary shaped anisotropic objects*, Eng. Anal. Bound. Elem. **82** (2017) 91–97.
5. I. Babuška and J. M. Melenk, *The partition of unity method*, Int. J. Numer. Methods Eng. **40** (1997) 727–758.

E-mail: marzie.raei@gmail.com

Hybride of Laplace Transform and Chelyshkov Wavelets Integral Operator for Solving Fractional-Order Differential Equations with Delay

Parisa Rahimkhani*
Department of Mathematics, Faculty of Mathematical Sciences, Alzahra University, Tehran, Iran
and Yadollah Ordokhani
Department of Mathematics, Faculty of Mathematical Sciences, Alzahra University, Tehran, Iran

Abstract. In this work, we use hybride of Laplace transform and Chelyshkov wavelets integral operator for solving fractional differential equations and time-fractional partial differential equations with delay. By using Laplace transform method, fractional-order differential equations are turned into integer-order differential equations. Then, Chelyshkov wavelets integral operator and collocation method are applied for solving obtained integer-order differential equations.
Keywords: Chelyshkov wavelets, Laplace transform, Integral operator, Fractional-order delay differential equations.
AMS Mathematical Subject Classification [2010]: 34A08, 65L60, 42C40.

1. Introduction

Delay differential equations have appeared in the modeling of various problems in industrial, biological, chemical, electronic and transportation systems. Several researchers discussed the properties of the analytic solutions of these equations and also their numerical solutions such as Adomian decomposition method, shifted Chebyshev spectral Tau method, Bernstein polynomials, Legendre wavelet method, Bernoulli wavelet method [1] and so on.

2. Preliminaries and Fundamentals

2.1. Chelyshkov Wavelets. The Chelyshkov wavelets are defined on $L^2[0,h)$ as [2]

$$\psi^h_{n,m,\hat{m}}(t)=\begin{cases} 2^{\frac{k-1}{2}}\tilde{C}^h_{m,\hat{m}}(2^{k-1}t-\hat{n}), & \frac{\hat{n}}{2^{k-1}}h\le t<\frac{\hat{n}+1}{2^{k-1}}h,\\ 0, & \text{otherwise,}\end{cases}$$

with

$$\tilde{C}^h_{m,\hat{m}}(t)=\sqrt{2m+1}C^h_{m,\hat{m}}(t),$$

where $m=0,1,\ldots,\hat{m},\hat{m}=M-1$, $\hat{n}=n-1$, $n=1,2,\ldots,2^{k-1}$, $\hat{w}=2^{k-1}M$ and $C^h_{m,\hat{m}}(t)$ is Chelyshkov polynomials.

*Speaker

2.2. Integral Operator of Chelyshkov Wavelets. Integral operator of Chelyshkov wavelets $G^h(t,1)$ is obtained as [2]

$$\int_0^t \Psi^h(t)dt = G^h(t,1), \tag{1}$$

where

$$G^h(t,1) = \left[\int_0^t \psi^h_{1,0,\hat{m}}(t)dt, \int_0^t \psi^h_{1,1,\hat{m}}(t)dt, \ldots, \int_0^t \psi^h_{2^{k-1},M-1,\hat{m}}(t)dt\right],$$

that

$$\int_0^t \psi^h_{n,m,\hat{m}}(t)dt = \begin{cases} 0, & 0 \le t < \frac{\hat{n}}{2^{k-1}}h, \\ \delta_{m,k}\sum_{r=m}^{\hat{m}}\sum_{s=0}^{r}(-1)^{2r-m-s}\begin{pmatrix} \hat{m}-m \\ r-m \end{pmatrix}\begin{pmatrix} \hat{m}+r+1 \\ \hat{m}-m \end{pmatrix}\begin{pmatrix} r \\ s \end{pmatrix} & \\ \quad \frac{\hat{n}^{r-s}2^{(k-1)s}}{h^r(s+1)}[t^{s+1} - (\frac{\hat{n}}{2^{k-1}}h)^{s+1}], & \frac{\hat{n}}{2^{k-1}}h \le t < \frac{\hat{n}+1}{2^{k-1}}h, \\ \delta_{m,k}\sum_{r=m}^{\hat{m}}\sum_{s=0}^{r}(-1)^{2r-m-s}\begin{pmatrix} \hat{m}-m \\ r-m \end{pmatrix}\begin{pmatrix} \hat{m}+r+1 \\ \hat{m}-m \end{pmatrix}\begin{pmatrix} r \\ s \end{pmatrix} & \\ \quad \frac{\hat{n}^{r-s}2^{(k-1)s}}{h^r(s+1)}[(\frac{\hat{n}+1}{2^{k-1}}h)^{s+1} - (\frac{\hat{n}}{2^{k-1}}h)^{s+1}], & \frac{\hat{n}+1}{2^{k-1}}h \le t < h. \end{cases}$$

Also, we have

$$\int_0^t \int_0^t \Psi^h(t)dtdt = G^h(t,2). \tag{2}$$

2.3. Approximation of Fractional Derivative. By using properties of Laplace transform and inverse Laplace transform for partial differential equations for $0 < \alpha < 1$, we obtain [3]

$$D_t^\alpha u(x,t) \approx \alpha\frac{\partial u(x,t)}{\partial t} + (1-\alpha)[u(x,t) - u(x,0)]. \tag{3}$$

For $1 < \alpha < 2$, we have

$$D_t^\alpha u(x,t) \approx (\alpha-1)\frac{\partial^2 u(x,t)}{\partial t^2} + (2-\alpha)\frac{\partial u(x,t)}{\partial t} - (2-\alpha)u_t(x,0).$$

Also, for fractional-order ordinary differential equations with $0 < \alpha < 1$, we obtain

$$D^\alpha u(t) \approx \alpha\frac{\partial u(t)}{\partial t} + (1-\alpha)[u(t) - u(0)].$$

For $1 < \alpha < 2$, we have

$$D^\alpha u(t) \approx (\alpha-1)\frac{\partial^2 u(t)}{\partial t^2} + (2-\alpha)\frac{\partial u(t)}{\partial t} - (2-\alpha)u'(0). \tag{4}$$

3. Numerical Method

Problem a: We consider the fractional differential equations as:

$$D^{\alpha}u(t) = f(t, u(at)), \qquad u(0) = \lambda_0, u'(0) = \lambda_1, \qquad 1 < \alpha \leq 2. \tag{5}$$

Substituting Eq. (4) in Eq. (5), we obtain

$$\begin{aligned} &(\alpha - 1)\frac{\partial^2 u(t)}{\partial t^2} + (2-\alpha)\frac{\partial u(t)}{\partial t} - (2-\alpha)\lambda_1 = f(t, u(at)), \\ &u(0) = \lambda_0, u'(0) = \lambda_1. \end{aligned} \tag{6}$$

Now, we solve Eq. (6). For solving this problem, we expand $u''(t)$ as

$$u''(t) = C^T \Psi(t), \tag{7}$$

By using Eqs. (2), (6) and (7), we have

$$u(t) = C^T G(t, 2) + \lambda_0 + \lambda_1 t, \tag{8}$$

$$u(at) = C^T G(at, 2) + \lambda_0 + \lambda_1 at, \tag{9}$$

$$u'(t) = C^T G(t, 1) + \lambda_1. \tag{10}$$

Replacing Eqs. (7)-(10) in Eq. (6) and collocating this equation at the zeros of shifted Legendre polynomials, we obtain numerical solution by using (8).

Problem b: We consider the following time-fractional partial differential equations with delay as

$$\begin{aligned} &D_t^{\alpha}u(x,t) = F(x, t, u(a_0x, b_0t), u_x(a_1x, b_1x), u_{xx}(a_2x, b_2t)), \\ &u(x,0) = f_0(x), \qquad u(0,t) = g_0(t), \qquad u(1,t) = g_1(t). \end{aligned} \tag{11}$$

Substituting Eq. (3) in Eq. (11), we obtain

$$\begin{aligned} \alpha\frac{\partial u(x,t)}{\partial t} &+ (1-\alpha)[u(x,t) - f_0(x)] \\ &= F(x, t, u(a_0x, b_0t), u_x(a_1x, b_1x), u_{xx}(a_2x, b_2t)), \end{aligned} \tag{12}$$

$$u(x,0) = f_0(x), \qquad u(0,t) = g_0(t), \qquad u(1,t) = g_1(t). \tag{13}$$

Now, for numerical solution of Eqs. (12) and (13), we expand $\frac{\partial^3 u(x,t)}{\partial x^2 \partial t}$ as

$$\frac{\partial^3 u(x,t)}{\partial x^2 \partial t} \simeq \Psi^T(x) U \Psi(t).$$

By using Eqs. (1), (2), (12) and (13), we obtain

$$\frac{\partial^2 u(x,t)}{\partial x^2} \simeq \Psi^T(x) U G(t,1) + \frac{\partial^2 f_0(x)}{\partial x^2}, \tag{14}$$

$$\frac{\partial u(x,t)}{\partial t} \simeq G^T(x,2) U \Psi(t) - x G^T(1,2) U \Psi(t) + (1-x)\frac{\partial g_0(t)}{\partial t} + x\frac{\partial g_1(t)}{\partial t}, \tag{15}$$

$$u(x,t) \simeq G^T(x,2) U G(t,1) - (x) G^T(1,2) U G(t,1) + \mu(x,t), \tag{16}$$

$$\mu(x,t) = g_0(t) + f_0(x) - f_0(0) - x f_0'(0) + x(g_1(t) - g_0(t)) + x(-f_0(1) + f_0(0) + f_0'(0)).$$

$$\frac{\partial u(x,t)}{\partial x} \simeq G^T(x,1)UG(t,1) - G^T(1,2)UG(t,1) + \frac{\partial \mu(x,t)}{\partial x}. \tag{17}$$

Replacing Eqs. (14)-(17) in Eq. (11) and collocating this equation at the zeros of shifted Legendre polynomials, we obtain numerical solution by using (16).

4. Error Bound

In this part, we get error bound of best approximation in terms of Sobolev norms.

THEOREM 4.1. [4] *Suppose $u \in H^\tau(0,h)$ with $\tau \geq 0$ and $M \geq \tau$, and $\tilde{u}$ is the best approximation of u which is obtained by applying the Chelyshkov wavelets, then we get*

$$\|u - \tilde{u}\|_{L^2(0,h)} \leq c'(M-1)^{-\tau}(2^{k-1})^{-\tau}\|u^{(\tau)}\|_{L^2(0,h)}, \tag{18}$$

and for $1 \leq s \leq \tau$ we have

$$\|u - \tilde{u}\|_{H^s(0,h)} \leq c'(M-1)^{2s-\frac{1}{2}-\tau}(2^{k-1})^{s-\tau}\|u^{(\tau)}\|_{L^2(0,h)},$$

where c' is a positive constant depends on τ and h.

REMARK 4.2. This result shows that in the case u is infinitely smooth, the rate of convergence of $\tilde{u}$ to u is faster than $\frac{1}{2^{k-1}}$ to the power of $M-s$ and any power of $\frac{1}{M-1}$, which is superior to that for the classical spectral methods.

COROLLARY 4.3. *Let $u \in H^\tau(0,h)$ and $2 \leq s \leq \tau$, then we get*

$$\|u'' - \tilde{u}''\|_{L^2(0,h)} \leq c'(M-1)^{2s-\frac{1}{2}-\tau}(2^{k-1})^{s-\tau}\|u^{(\tau)}\|_{L^2(0,h)}, \tag{19}$$

and

$$\|u' - \tilde{u}'\|_{L^2(0,h)} \leq c'(M-1)^{2s-\frac{1}{2}-\tau}(2^{k-1})^{s-\tau}\|u^{(\tau)}\|_{L^2(0,h)}. \tag{20}$$

PROOF. By using definition the Sobolev norm for $2 \leq s \leq \tau$, we obtain the above results. □

THEOREM 4.4. *Suppose that $u \in H^\tau(0,h)$ and $\tilde{u} \in H^\tau(0,h)$ be the exact and the numerical solution of Eq. (5), respectively. Moreover, we consider $2 \leq s \leq \tau$ and function $f(t,u)$ satisfies the Lipschitz condition with constant η, then the error bound of the mentioned approach is obtained as*

$$\begin{aligned}\|E\|_{L^2(0,h)} &\leq |\alpha-1|c'(M-1)^{2s-\frac{1}{2}-\tau}(2^{k-1})^{s-\tau}\|u^{(\tau)}\|_{L^2(0,h)} \\ &+ |2-\alpha|c'(M-1)^{2s-\frac{1}{2}-\tau}(2^{k-1})^{s-\tau}\|u^{(\tau)}\|_{L^2(0,h)} \\ &+ \eta c'(M-1)^{-\tau}(2^{k-1})^{-\tau}\|u^{(\tau)}\|_{L^2(0,h)}.\end{aligned}$$

PROOF. By considering the Eqs. (5), (18), (19) and (20) and properties of Sobolev norms, we obtain the above result. □

5. Numerical Results

EXAMPLE 5.1. Consider the nonlinear fractional pantograph differential equation

$$D^\alpha u(t) = \frac{3}{4}u(t) + u(\frac{t}{2}) - t^2 + 2, \qquad u(0) = u'(0) = 0.$$

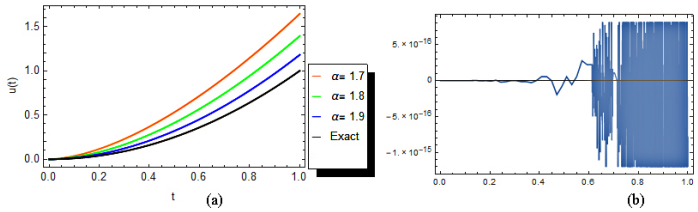


FIGURE 1. Numerical results with $k = 1, M = 6$, for different values of $\alpha = 1.7, 1.8, 1.9$ (Example 5.1).

The exact solution for $\alpha = 2$ is $u(t) = t^2$. Numerical results for $k = 1, M = 6$ and different values of α are demonstrated in Figure 1.

EXAMPLE 5.2. Consider the nonlinear fractional pantograph differential equation

$$D^{\alpha}u(t) = 1 - 2u^2(\frac{t}{2}), \qquad u(0) = 1, u'(0) = 0.$$

The exact solution for $\alpha = 2$ is $u(t) = \cos(t)$. Numerical results for $k = 1, M = 10$ and different values of α are demonstrated in Figure 2.

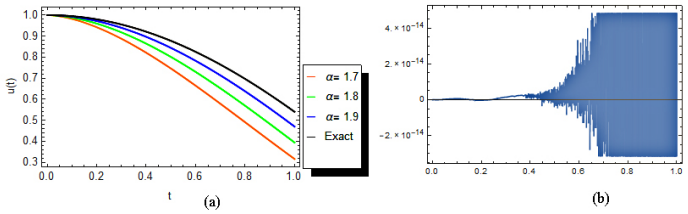


FIGURE 2. Numerical results with $k = 1, M = 10$, for different values of $\alpha = 1.7, 1.8, 1.9$ (Example 5.2).

EXAMPLE 5.3. Consider the nonlinear fractional differential equation

$$D^{\alpha}u(t) = -u^2(t) + \frac{\Gamma(6)}{\Gamma(6-\alpha)}t^{5-\alpha} + \frac{36}{\Gamma(5-\alpha)}t^{4-\alpha} - \Gamma(3+\alpha)t^2 + (t^5 + \frac{3}{2}t^4 - 2t^{2+\alpha})^2, \; u(0) = 0.$$

The exact solution is $u(t) = t^5 + \frac{3}{2}t^4 - 2t^{2+\alpha}$. Numerical results for $k = 1, M = 7$ and different values of α are demonstrated in Table 1.

EXAMPLE 5.4. Consider the time fractional Burgers equation with proportional delay as

$$D_t^{\alpha}u(x,t) = u_{xx}(x,t) + u(\frac{x}{2}, \frac{t}{2})u_x(x, \frac{t}{2}) + \frac{1}{2}u(x,t),$$
$$u(x,0) = x, \qquad u(0,t) = 0, \qquad u(1,t) = e^t.$$

The exact solution for $\alpha = 1$ is $u(x,t) = xe^t$. Table 2 displays the absolute error of suggested scheme by choosing $k = k' = 2, M = M' = 2$ and $\alpha = 1$ together with Homotopy perturbation transform method [5].

TABLE 1. The absolute error for $k = 1, M = 7$ and different values of α for (Example 5.3).

t	$\alpha = 0.8$	$\alpha = 0.9$	$\alpha = 0.99$	$\alpha = 1$
0.1	1.11×10^{-3}	5.01×10^{-4}	4.57×10^{-5}	8.46×10^{-18}
0.3	1.28×10^{-2}	6.05×10^{-3}	5.74×10^{-4}	0
0.5	2.31×10^{-2}	1.10×10^{-2}	1.04×10^{-3}	0
0.7	1.97×10^{-3}	1.73×10^{-5}	1.04×10^{-4}	2.50×10^{-16}
0.9	8.94×10^{-2}	4.65×10^{-2}	4.83×10^{-3}	2.12×10^{-15}

TABLE 2. Comparison of absolute error for $k = k' = 2, M = M' = 2, \alpha = 1$ for (Example 5.4).

x	t	*Ref.* [5]	*Our method*
0.25	0.25	2.122401×10^{-6}	0
	0.50	7.094268×10^{-5}	0
	0.75	5.634807×10^{-4}	0
	1.00	2.487124×10^{-3}	0
0.5	0.25	4.244802×10^{-6}	0
	0.50	1.418854×10^{-4}	0
	0.75	1.126961×10^{-3}	0
	1.00	4.974248×10^{-3}	0
0.75	0.25	6.369688×10^{-6}	0
	0.50	2.128250×10^{-4}	0
	0.75	1.690020×10^{-3}	0
	1.00	7.461370×10^{-3}	0
CPU times		–	0.921

References

1. P. Rahimkhani, Y. Ordokhani and E. Babolian, *Numerical solution of fractional pantograph differential equations by using generalized fractional-order Bernoulli wavelet*, J. Comput. Appl. Math. **309** (2017) 493−510.
2. P. Rahimkhani and Y. Ordokhani, *Numerical solution of Volterra-Hammerstein delay integral equations*, Iranian J. Sci. Tech. Trans. A: Sci. **44** (2020) 445−457.
3. J. Ren, Z. Z. Sun and W. Dai, *New approximations for solving the Caputo-type fractional partial differential equations,* Appl. Math. Model. **40** (4) (2016) 2625−2636.
4. P. Rahimkhani, Y. Ordokhani and E. Babolian, *Müntz-Legendre wavelet operational matrix of fractional-order integration and its applications for solving the fractional pantograph differential equations*, Numer. Algor. **77** (2018) 1283−1305.
5. B. K. Singh and P. Kumar, *Homotopy perturbation transform method for solving fractional partial differential equations with proportional delay*, SeMA **75** (1) (2017) 111−125.

E-mail: P.rahimkhani@alzahra.ac.ir
E-mail: ordokhani@alzahra.ac.ir

A New Operational Matrix of Fibonacci Polynomials for Solving a Class of Distributed Order Fractional Differential Equations

Sedigheh Sabermahani*
Department of Mathematics, Faculty of Mathematical Sciences, Alzahra University, Tehran, Iran
and Yadollah Ordokhani
Department of Mathematics, Faculty of Mathematical Sciences, Alzahra University, Tehran, Iran

Abstract. Here, we propose a numerical method for solving linear distributed-order fractional differential equations. Distributed order fractional derivative operational matrix and fractional derivative operational matrix for Fibonacci polynomials are presented. Using the operational matrices and Galerkin method, the problem is converted into a system of algebraic equations. Several examples are tests to investigate the efficiency of the technique.
Keywords: Distributed-order fractional derivative operational matrix, Distributed order fractional equation, Fibonacci polynomial, Galerkin method.
AMS Mathematical Subject Classification [2010]: 65D15, 11B39, 68M14.

1. Introduction

Firstly, distributed-order of fractional derivative is presented by Caputo [3]. Next, this concept is developed by himself and some mathematicians. This concept is used in mathematical modeling of some phenomena, then this has been considered by several researchers. Torvik and Bagley [1] investigated on the existence of the solution of distributed-order equations. The authors in [5] perused stability analysis of distributed-order of fractional differential equations. However, there is a difficult task to analytically treat these problems. Thus, it is required that a numerical scheme is established. Therefore, some numerical method for solving this class of equations have been proposed. A method based on Legendre polynomials [6] and a modified method proposed in [9] are samples of computational methods presented by some researchers.

2. Preliminaries

Definition 2.1. The Riemann-Liouville fractional integration operator of order $\gamma \geq 0$ of $f \in C_\eta, \eta > -1$ is defined as [7]

$$I^\gamma u(t) = \begin{cases} \frac{1}{\Gamma(\gamma)} \int_0^t u(s)(t-s)^{\gamma-1} ds, & \gamma > 0, t > 0, \\ u(t), & \gamma = 0. \end{cases} \tag{1}$$

*Speaker

DEFINITION 2.2. The Caputo fractional derivative of order γ of $f \in C_\eta, \eta > -1$ is defined as [7]

$$D^\gamma u(t) = \begin{cases} \frac{1}{\Gamma(n-\gamma)} \int_0^t u^{(n)}(s)(t-s)^{n-\gamma-1} ds, & n-1 < \gamma \leq n, n \in N, \\ u(t), & \gamma = 0. \end{cases} \tag{2}$$

Also, we have

1) $D^\gamma c = 0$, where c is constant.
2) $I^\gamma D^\gamma u(t) = u(t) - \sum_{i=0}^{n-1} u^{(i)}(0) \frac{t^i}{i!}$.
3) $D^\gamma t^k = \begin{cases} \frac{\Gamma(k+1)}{\Gamma(k+1-\gamma)} t^{k-\gamma}, & k \geq \gamma, \\ 0, & k < \gamma. \end{cases}$

2.1. Fibonacci Polynomials and Their Properties. Fibonacci polynomials defined as the following form [8]

$$F_m(x) = \sum_{i=0}^{\lfloor m/2 \rfloor} \begin{pmatrix} m-i \\ i \end{pmatrix} x^{m-2i}, \quad m \geq 0. \tag{3}$$

Here, the first m-terms of the Fibonacci polynomials are used to proposed an approximation. Then, an arbitrary function $u(t) \in L^2[0,1]$ may be approximate as follows

$$u(t) \simeq \sum_{j=0}^{M} u_j F_j(t) = U^T \Psi(t), \tag{4}$$

subject to

$$U = [u_0, u_1, \ldots, u_M]^T, \quad \Psi(t) = [F_0(t), F_1(t), \ldots, F_M(t)]^T. \tag{5}$$

The coefficient vector $U = (u_j)$ is given by $U^T = \Delta^T D^{-1}$, where $D = \langle \Psi, \Psi \rangle$, $\Delta = \langle u, \Psi \rangle$.

2.2. Caputo Derivative Operational Matrix of Fibonacci Polynomials. This subsection is devoted to construct the Caputo derivative operational matrix for Fibonacci polynomials. The Caputo derivative of the vector $\Psi(t)$ can be expressed by

$$D^\gamma \Psi(t) \simeq D^{(\gamma)} \Psi(t). \tag{6}$$

$D^{(\gamma)}$ is the $(M+1) \times (M+1)$ operational matrix of Caputo fractional derivative.

Now, we obtain this operational matrix. Using Eq. (3) and considering the properties of Caputo derivative, we have

$$\begin{aligned} D^\gamma F_m(t) &= \sum_{i=0}^{\lfloor m/2 \rfloor} \begin{pmatrix} m-i \\ i \end{pmatrix} D^\gamma t^{m-2i} \\ &= \sum_{i=0}^{\lfloor m/2 \rfloor} \begin{pmatrix} m-i \\ i \end{pmatrix} \eta_{m-2i-\gamma}(t), \end{aligned} \tag{7}$$

where

$$\eta_{m-2i-\gamma}(t) = \begin{cases} \frac{\Gamma(m-2i+1)}{\Gamma(m-2i+1-\gamma)} t^{m-2i-\gamma}, & m-2i \geq \gamma, \\ 0, & m-2i < \gamma. \end{cases}$$

and $m \geq 0$.

Now, by expanding the above equation regarding Fibonacci polynomials, we have

$$D^{\gamma} F_m(t) \simeq \phi_{m,\gamma}^T \Psi(t). \tag{8}$$

Then, we achieve the following relation:

$$D^{(\gamma)} = [\phi_{m,\gamma}^T], \quad m = 0, 1, \ldots, M.$$

2.3. Distributed-Order Fractional Derivative Operational Matrix of Fibonacci Polynomials. Here, the methodology of obtaining the distributed-order fractional derivative operational matrix is presented. This operational matrix is presented as follows:

$$D^{\tau(\gamma)} \Psi(t) \simeq \Theta_{\tau(\gamma)} \Psi(t), \tag{9}$$

where $D^{\tau(\gamma)} u(t) = \int_a^b \tau(\gamma) D^{\gamma} u(t) d\gamma$. Considering definition of $D^{\tau(\gamma)} u(t)$ and Eq. (8) and utilizing the Gauss-Legendre numerical integration for evaluating the existing integration, we derive

$$\begin{aligned} D^{\tau(\gamma)} F_m(t) &= \int_a^b \tau(\gamma) D^{\gamma} F_m(t) d\gamma \simeq \int_a^b \tau(\gamma) \phi_{m,\gamma}^T \Psi(t) d\gamma \\ &= (\int_a^b \tau(\gamma) \phi_{m,\gamma}^T d\gamma) \Psi(t) \\ &\simeq \frac{b-a}{2} \left(\sum_{j=1}^{N} \omega_j \tau(\frac{b-a}{2} \vartheta_j + \frac{a+b}{2}) \phi_{m,(\frac{b-a}{2}\vartheta_j + \frac{a+b}{2})}^T \right) \Psi(t) \\ &= \chi_{m,\tau(\gamma)}^T \Psi(t). \end{aligned} \tag{10}$$

Hence, we have

$$\Theta_{\tau(\gamma)} = [\chi_{m,\tau(\gamma)}^T], \quad m = 0, 1, \ldots, M.$$

3. Description of the Method

Here, we consider the following distributed-order fractional differential equation (DFDE):

$$\int_a^b \tau(\gamma) D^{\gamma} u(t) d\gamma = G(t), \quad t \in [0, 1], \tag{11}$$

subject to the following constraints

$$u^{(l)}(0) = u_{0,l}, \quad l = 0, 1, \ldots, \lceil b \rceil - 1, \tag{12}$$

and D^{γ} denotes the fractional derivative in the Caputo type of order γ. a and b are positive numbers. We approximate the function $u(t)$ by Fibonacci polynomials as follows $u(t) \simeq U^T \Psi(t)$, where U is an unknown vector and Ψ is defined in Eq. (5). Then, concerning distributed-oredr fractional derivative operational matrix, the considered problem can be written as

$$Y(t) = U^T \Theta_{\tau(\gamma)} \Psi(t) - G(t), \tag{13}$$

Now, we use the Galerkin method which is convergent [2]

$$\langle Y, F_m \rangle = 0, \quad m = 0, 1, M - l, \tag{14}$$

TABLE 1. Comparison of the L_2 errors in Example 4.1.

Methods	L_2 errors
Ref. [6]	
$m = 6$	2.73×10^{-5}
$m = 10$	1.84×10^{-6}
Present method	
$M = 3$	2.22507×10^{-7}

and for the initial conditions, we get

$$U^T D^{\gamma}\Psi(0) - u_{0,l} = 0, \quad l = 0, 1, \ldots, \lceil b \rceil - 1. \tag{15}$$

Considering the above relations, the problem translates to a system of algebraic equations. The system is solved by applying the "Find Root" package in Mathematica software to obtain the unknown coefficients vector.

4. Illustrative Examples

EXAMPLE 4.1. Consider the following DFDE [6]

$$\int_0^1 \frac{\Gamma(\frac{7}{2}-\gamma)}{\Gamma(\frac{7}{2})} D^{\gamma}u(t)d\gamma = \frac{t^{\frac{3}{2}}(t-1)}{ln(t)},$$

subject to $u(0) = 0$. The exact solution of this problem is $u(t) = t^{\frac{5}{2}}$. Table 1 reports the maximum absolute errors of the proposed technique for $M = 5$ and the results using method based on Legendre polynomials [6].

EXAMPLE 4.2. Consider the following DFDE as

$$\int_0^2 \frac{\Gamma(6-\gamma)}{120} D^{\gamma}u(t)d\gamma = \frac{t^5 - t^3}{ln(t)},$$

subject to $u(0) = u^{'}(0) = 0$, and the exact solution of this problem is $u(t) = t^5$. L_{∞} error achieved using the method for $M = 5$ is 3.75597×10^{-8} and this error for [9] for $N = 5$ is 1.06×10^{-7}.

EXAMPLE 4.3. Consider the following DFDE

$$\int_{0.2}^{1.5} \Gamma(3-\gamma) D^{\gamma}u(t)d\gamma = 2(\frac{t^{1.8} - t^{0.5}}{ln(t)}),$$

with the initial conditions $u(0) = u^{'}(0) = 0$. The exact solution is $u(t) = t^2$. by applying the described method, we solve this problem, numerically. The absolute errors of the present method $M = 2$ is displayed in Figure 1. Also, Table 2 lists the relative errors in $u(0.9)$ $(|u_{ex}(t) - u_{ap}(t)|/|u_{ex}(t)|)$ obtained by impelement the present method and method in [4].

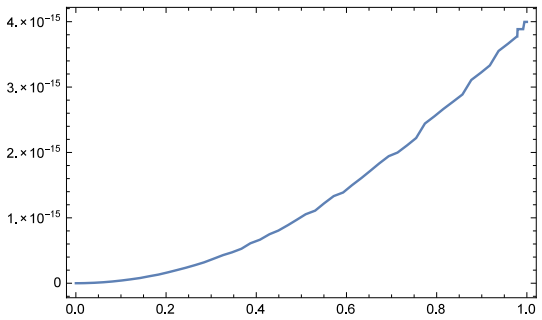


FIGURE 1. The comparison of absolute error of the present method for $M = 2$ in Example 4.3.

TABLE 2. Comparison of the relative errors in $u(0.9)$ for Example 4.3.

Methods	Relative errors
Ref. [4]	
$k = 32, h = 0.0001$	7.34×10^{-5}
$k = 64, h = 0.0001$	1.26×10^{-5}
Present method	
$M = 4$	2.96394×10^{-11}

References

1. R. L. Bagley and P. J. Torvik, *On the existence of the order domain and the solution of distributed-order equations I,* Int. J. Appl. Math. **2** (7) (2000) 865–882.
2. C. Canuto, M. Y. Hussaini, A. Quarteroni and T. A. Zang, *Spectral Methods, Fundamentals in Single Domains,* Springer-Verlag Berlin Heidelberg, New York, 2006.
3. M. Caputo, *Elasticità e Dissipazione*, Zanichelli, Bologna (in Italian), 1969.
4. J. T. Katsikadelis, *Numerical solution of distributed order fractional differential equations,* J. Comput. Phys. **259** (2014) 11–22.
5. H. S. Najafi, A. Refahi Sheikhani and A. Ansari, *Stability analysis of distributed order fractional differential equations*, Abstr. Appl. Anal. (2011) 175323. DOI: 10.1155/2011/175323
6. M. Pourbabaee and A. Saadatmandi, *A novel Legendre operational matrix for distributed order fractional differential equations*, Appl. Math. Comput. **361** (2019) 215–231.
7. S. Sabermahani, Y. Ordokhani and S. A. Yousefi, *Fractional-order general Lagrange scaling functions and their applications*, BIT Numer. Math. **60** (1) (2020) 101–128.
8. S. Sabermahani, Y. Ordokhani and S. A. Yousefi, *Fibonacci wavelets and their applications for solving two classes of time-varying delay problems*, Optim. Contr. Appl. Methods **41** (2) (2020) 395–416.
9. M. S. Semary, H. N. Hassan and A. G. Radwan, *Modified methods for solving two classes of distributed order linear fractional differential equations*, Appl. Math. Comput. **323** (2018) 106–119.

E-mail: s.saber@alzahra.ac.ir
E-mail: ordokhani@alzahra.ac.ir

On the Stability Analysis of Continuous Block Backward Differentiation Formulas up to Order 9

Hojatollah Saeidi*
Faculty of Mathematical Sciences, University of Sharekord, Sharekord, Iran
and Mohammad Shafie Dahaghin
Department of Mathematics, University of Sharekord, Shahrekord, Iran

Abstract. In this paper, we investigate the stability of continuous block backward differentiation formula (CBBDF) of orders 7, 8 and 9 and compare the stability regions of these methods with CBBDF of orders 2, 3, ..., 6. The results show that the stability regions of methods with orders 7, 8 and 9 are piecewise but larger than the methods with orders 2, 3, ..., 6 and therefore these methods are suitable for solving stiff systems.
Keywords: Continues block BDF, Collocation and interpolation, Numerical schemes, Stability region, Stiff problems.
AMS Mathematical Subject Classification [2010]: 13F55, 05E40, 05C65.

1. Introduction

Backward differentiation formulas (BDFs) are the popular implicit methods for solving ordinary differential equations. These methods were first used for the solution of stiff problems by Curtis and Hirschfelder [7]. Over the years several implicit methods have been developed and discussed extensively in literature, see [3, 4, 5]. The block methods were first introduced by Milne [10] and since then several block methods have been developed by researchers such as [6] and the references therein. Akinfenwa and Jator [1] have shown that the stability regions of continuous block backward differentiation formulas (CBBDFs) of orders 2, 3 and 4 are large. Also recently Akinfenwa, Jator and Yao [2] check out the CBBDFs of orders 4 and 6 and their stability regions. Our aim is to investigate the stability region of CBBDFs of orders up to 9.

In this paper, the concern has to do with implicit BDFs for the numerical solution of Initial Value Problems (IVPs) for first-order ODEs of the form

$$y' = f(t, y), \quad t \in (t_0, T_n), \quad y(t_0) = y_0, \tag{1}$$

which is generally written as

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h\beta_k f_{n+k},$$

where h is the step size, $\alpha_k = 1$, $\alpha_j, j = 1, 2, \ldots, k$, β_k are unknown constants which are uniquely determined such that the formula is of order k.

*Speaker

A block-by-block method is a method for computing vectors $Y_0, Y_1, \ldots$ in sequence [8]. Let the v-vector (v is the number of points within the block) for $n = mv, m = 0, 1, \ldots$ be given as $Y_\omega = (y_{n+1}, \ldots, y_{n+v})^T, F_\omega = (f_{n+1}, \ldots, f_{n+v})^T$, then the l-block v-point method for solving (1) are given by

$$Y_\omega = \sum_{i=1}^{l} A^{(i)} Y_{\omega-i} + h \sum_{i=0}^{l} B^{(i)} F_{\omega-i},$$

where $A^{(i)}, B^{(i)}, i = 0, 1, \ldots, l$ are v by v matrices [8].

Our aim is obtaining the CBBDF7, CBBDF8 and CBBDF9 and compare the stability regions of them with stability regions of CBBDF2, ..., CBBDF6, so we will have next sections.

2. Derivation of the Method

The block algorithm proposed in this paper is based on interpolation and collocation [9]. We proceed by seeking an approximate of the exact solution $y(t)$ by assuming a continuous solution $Y(t)$ of the form

$$Y(t) = \sum_{j=0}^{q+r-1} m_j \phi_j(t), \qquad t \in [t_0, T_n].$$

Such that $t \in [t_0, T_n]$, m_j are unknown coefficients and $\phi_j(t)$ are polynomial basis functions of degree $q + r - 1$, where the number of interpolation points q and the collocation point r are respectively chosen to satisfy $q = k$ and $r = 1$. The integer $k \geq 1$ denotes the step number of the method. We thus, constructed a k-step block methods with $\phi_j(t) = t^j_{n+i}$ by imposing the following conditions

$$Y(t_{n+i}) = \sum_{j=0}^{q+r-1} m_j t^j_{n+i} = y_{n+i}, \qquad i = 0, \ldots, k-1, \tag{2}$$

$$Y'(t_{n+i}) = \sum_{j=0}^{q} m_j j (t_{n+i})^{j-1} = f_{n+i}, \qquad i = k, \tag{3}$$

where y_{n+j} is the approximation for the exact solution $y(t_{n+j})$, $f_{n+j} = f(t_{n+j}, y_{n+j})$, n is the grid index and $t_{n+j} = t_n + jh$. It should be noted that equations (2) and (3) leads to a system of $q + 1$ equations of the form $AM = C$, where

$$A = \begin{pmatrix} 1 & t_n & t_n^2 & \ldots & t_n^q \\ 1 & t_{n+1} & t_{n+1}^2 & \ldots & t_{n+1}^q \\ 1 & t_{n+2} & t_{n+2}^2 & \ldots & t_{n+2}^q \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_{n+k-1} & t_{n+k-1}^2 & \ldots & t_{n+k-1}^q \\ 0 & 1 & 2t_{n+k} & \ldots & q t_{n+k}^{q-1} \end{pmatrix}, \; M = \begin{pmatrix} m_0 \\ m_1 \\ m_2 \\ \vdots \\ m_{k-1} \\ m_k \end{pmatrix}, \; C = \begin{pmatrix} y_n \\ y_{n+1} \\ y_{n+2} \\ \vdots \\ y_{n+k-1} \\ f_{n+k} \end{pmatrix},$$

which must be solved to obtain the coefficients m_j. After some algeraic computation we generated the block method as follow:

$$f_{n+i} = \frac{1}{c_i h}\left[b_i h f_{n+k} + \sum_{j=0}^{k-1} a_{ij} y_{n+j}\right],\ i = 1, 2, \ldots, k-1,$$

$$y_{n+k} = \frac{1}{c_k}\left[b_k h f_{n+k} + \sum_{j=0}^{k-1} a_{kj} y_{n+j}\right].$$

The parameters c_i, b_i and a_{ij} for $i = 1, 2, \ldots, k$ and $j = 0, 1, \ldots, k-1$ are presented in Tables 1, 2 and 3 for $k = 7, 8$ and 9 respectively.

TABLE 1. The parameters c_i, b_i and a_{ij} for $k = 7$.

i	c_i	b_i	a_{i0}	a_{i1}	a_{i2}	a_{i3}	a_{i4}	a_{i5}	a_{i6}
1	65340	-600	-9420	-94043	193500	-158100	-42705	101900	8868
2	13068	48	318	-4412	-10035	21360	3852	-10330	-753
3	5445	-15	-54	562	-3330	-1230	-1476	5270	258
4	65340	240	501	-4636	20610	-67440	36684	19135	-4854
5	65340	-600	-708	6145	-24300	59700	57483	-115900	17580
6	65340	3600	2070	-17268	64125	-140400	-233820	205350	119943
7	1089	420	60	-490	1764	-3675	-4410	4900	2940

TABLE 2. The parameters c_i, b_i and a_{ij} for $k = 8$.

i	c_i	b_i	a_{i0}	a_{i1}	a_{i2}	a_{i3}	a_{i4}	a_{i5}	a_{i6}	a_{i7}
1	45660	300	-5745	-72387	158410	-156450	-74305	127925	27762	-5210
2	958860	-2100	17385	-276360	-901117	1894200	600040	-1161825	-210315	37922
3	63924	84	-391	46627	-32354	-27825	-30394	78435	9478	-1611
4	159810	-210	597	-6328	32942	-130200	123928	3675	-29022	4408
5	958860	2100	-3687	36645	-169610	502950	470687	-1235325	450030	-51690
6	319620	-2100	2165	-20664	89705	-236600	-678440	436275	333039	74520
7	319620	14700	-7545	70070	-292334	723975	1393070	-1189475	-132440	626709
8	2283	840	-105	960	-3920	9408	15680	-14700	-11760	6720

3. Stability Analysis

In what follows, the k-step continuous block BDF rearranged and rewritten as a matrix finite difference equation of the form

$$A^{(1)} Y_{\omega+1} = A^{(0)} Y_\omega + h\beta^{(1)} F_\omega,\ \omega = 0, 1, \ldots, \tag{4}$$

where $A^{(1)}$, $A^{(0)}$ and $B^{(1)}$ are k by k matrices and

$$Y_{\omega+1} = (y_{n+1}, y_{n+2}, \ldots, y_{n+k})^T\ ,\ Y_\omega = (y_{n-k+1}, y_{n-k+2}, \ldots, y_n)^T,$$
$$F_\omega = (f_{n+1}, f_{n+2}, \ldots, f_{n+k})^T,\ n = 0, 1, \ldots, N-k.$$

TABLE 3. The parameters c_i, b_i and a_{ij} for $k = 9$.

i	c_i	b_i	a_{i0}	a_{i1}	a_{i2}	a_{i3}	a_{i4}	a_{i5}	a_{i6}	a_{i7}	a_{i8}
1	748545	-3675	-835805	-1281759	2975280	-3441760	-2504145	3400600	1294384	-432880	73860
2	855480	1200	12015	-215220	-928746	1979320	960260	-1466850	-465430	149496	-24845
3	2994180	-2100	-12115	162765	-1294020	-1817011	-2179485	44038350	939260	-283005	45261
4	1197672	672	2451	-29272	174552	-807856	1176504	-222600	-380408	101968	-15339
5	1497090	-1050	-2493	27915	-147980	513730	332493	-1523550	968660	-194970	26195
6	5988360	8400	12815	-137772	684810	-2113720	-9247140	4702950	3928022	2415240	-245205
7	2994180	-14700	-13515	141295	-674436	1952405	5702865	-3863650	-7398020	3536811	616245
8	5988360	235200	109305	-1120080	5201840	-14471072	-35354480	26886300	34531280	-28187040	12403947
9	7129	2520	280	-2835	12960	-35280	-793805	63504	70560	-45360	220680

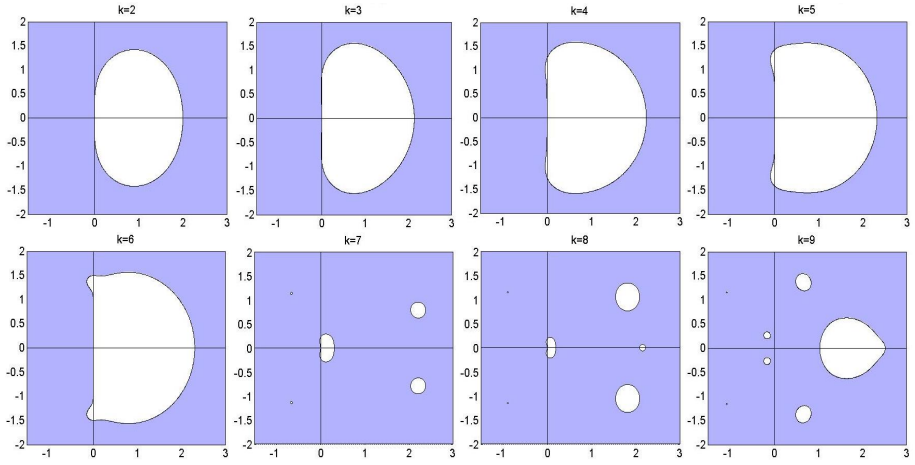


FIGURE 1. Stability regions of CBBDFs for $k = 2, 3, \ldots, 9$.

3.1. Zero Stability. As $h \to 0$, the method (4) tends to the difference system

$$A^{(1)}Y_{\omega+1} - A^{(0)}Y_{\omega} = 0,$$

with first characteristic polynomial $\rho(R) = det(RA^{(1)} - A^{(0)})$. Following Fatunla [8], the block by block method (4) is zero-stable, since $\rho(R) = 0$ satisfies $|R_j| \leq 1$ for $j = 1, 2, \ldots, k$ and for those roots with $|R_j| = 1$, the multiplicity does not exceed 1.

References

1. O. A. Akinfenwa and S. N. Jator, *On the stability of continuous block backward differentiation formula For solving stiff ordinary differential equations*, J. Mod. Meth. Numer. Math. **2** (2012) 50–58.
2. O. A. Akinfenwa, S. N. Jator and N. M. Yao, *Continuous block backward differentiation formula for solving stiff ordinary differential equations*, Comput. Appl. **65** (7) (2013) 996–1005.
3. G. Psihoyios, *A Block implicit advanced step-point (BIAS) algorithm for stiff differential systems*, Comput. Lett. **2** (12) (2006) 51–58.

4. E. Suli and D. F. Mayers, *An Introduction to Numerical Analysis,* Cambridge University Press, Cambridge, UK, 2003.
5. J. R. Cash, *Review paper, Efficient numerical methods for the solution of stiff initial-value problems and differential algebraic equations*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. **459** (2032) (2003) 797–815.
6. M. T. Chu and H. Hamilton, *Parallel solution of odes by multi-block methods*, SIAM J. Sci. Stat. Comput. **8** (1987) 342–353.
7. C. F. Curtis and J. D. Hirschfelder, *Integration of stiff equations*, Proc. Natl. Acad. Sci. **38** (1952) 235–243.
8. S. O. Fatunla, *Block methods for second order IVPs*, Int. J. Comput. Math. **41** (1991) 55–63.
9. I. Gladwell and D. K. Sayers, *Computational Techniques for Ordinary Differential Equations*, Academic Press, New York, 1976.
10. W. E. Milne, *Numerical Solution of Differential Equations*, John Wiley and Sons, New York, 1953.

E-mail: Hojat.saeidi65@gmail.com
E-mail: msh-dahaghin@sci.sku.ac.ir

Shape Preserving Interpolation by Bézier-Like Curve

Jamshid Saeidian*
Faculty of Mathematical Sciences and Computer, Kharazmi University, Tehran, Iran
and Bahareh Nouri
Faculty of Mathematical Sciences and Computer, Kharazmi University, Tehran, Iran

Abstract. In this work we study the shape preserving properties of a Bézier-like model. The model has been proposed by Yan and Liang in 2011. We prove that the proposed Bézier-like curves can preserve monotonicity and boundedness.
Keywords: Shape preserving interpolation, Monotonicity preservation, Boundedness.
AMS Mathematical Subject Classification [2010]: 65D17, 65D05.

1. Introduction

Shape-preserving interpolation is defined to be a method of constructing an interpolant curve (surface) which also preserves the shape implied by the data points. It is known as an essential curve/surface design technique in CAD/CAM and geometric design. Convexity, monotonicity, positivity and boundedness are the most important shape features which has extensively been studied in literature. When we recreate the underlying entity by interpolation from sampled values, we need to ensure that the interpolating curve adheres to these known properties.

In this work a Bézier-like curve is studied from shape-preserving point of view and it is proven that the curve preserves monotonicity and boundedness of the data.

Monotone data could be seen in many scientific phenomenon; the uric acid level in patients suffering from gout [4], erythrocyte sedimentation rate (E.S.R.) in cancer patients, rate of dissemination of drug in the blood [3] are examples of entities in medicine which only have meaning when they are monotone. To count some other cases we can refer to design of aggregation operators in multi-criteria decision making and fuzzy logic, the approximation of copulas and quasi copulas in statistics, empirical option pricing models in finance [1].

One of the hidden features in a data set may be its boundedness. This happens, for example, when the data comes from a sampling of a bounded function or they reflect the probability or efficiency of a process. Actually any quantity which is expressed as a percentage of another quantity will necessarily lie between 0 and 100. In this case, it is natural to expect the interpolant to lie between the imposed bounds.

*Speaker

2. Preliminary

Definition 2.1. [5] Let $\lambda \in [-1,1]$, for $t \in [0,1]$, the functions

$$\begin{aligned} b_{2,0}(t) &= \left(1-2\lambda t+\lambda t^2\right)(1-t)^2, \\ b_{2,1}(t) &= 2t(1-t)\left(1+\lambda-\lambda t+\lambda t^2\right), \\ b_{2,2}(t) &= \left(1-\lambda+\lambda t^2\right)t^2, \end{aligned} \tag{1}$$

are called the Bernstein-like basis functions of order 2. One can define the corresponding order n basis functions with a recursion

$$b_{n,i}(t) = (1-t)b_{n-1,i}(t) + tb_{n-1,i-1}(t),\ t \in [0,1],\ n \geq 3\ \&\ i = 0,1,2,\ldots,n. \tag{2}$$

Definition 2.2. Given control points $V_i \in \mathbb{R}^d$ $(i = 0,1,\ldots,n; n \geq 2)$, we call

$$p_n(t) = \sum_{i=0}^{n} b_{n,i}(t)V_i, \qquad t \in [0,1],$$

the Bézier-like curve of order n, where $b_{n,i}(t)(i = 0,1,\ldots,n)$ are the Bernstein-like basis functions.

The Bézier-like curve has the following properties [5]:

(a) Convex hull property: The entire curve lies inside the convex hull of its control polygon.

(b) Geometric invariance: Because $p_n(t)$ is an affine combination of the control points, the shape of the Bézier-like curve is independent of the choice of coordinate system.

(c) Symmetry:

$$p_n(t;V_0,V_1,\ldots,V_n) = \sum_{i=0}^{n} b_{n,i}(t)V_i = \sum_{j=0}^{n} b_{n,j}(1-t)V_{n-j} = p_n(1-t;V_n,V_{n-1},\ldots V_0).$$

(d) End point conditions:

$$p_n(0) = V_0,\ p_n(1) = V_n,\ p_n'(0) = (n+2\lambda)(V_1 - V_0),\ p_n'(1) = (n+2\lambda)(V_n - V_{n-1}).$$

3. Monotonicity Preservation of the Bernstein-Like Basis Functions

In this section we revisit the main results on the monotonicity preservation of curves and present a proof in the case of Bernstein-like basis functions.

Definition 3.1. A system of functions $(u_0,\ldots,u_n)$ is monotonicity preserving if for any $\alpha_0 \leq \alpha_1 \leq \cdots \leq \alpha_n$ in $\mathbb{R}$, the function $\sum_{i=0}^{n} \alpha_i u_i$ is increasing.

The following result, which characterizes monotonicity preserving systems, appears in [2, Proposition 2.3].

Proposition 3.2. *Let $(u_0,\ldots,u_n)$ be a system of functions defined on an interval $[a,b]$. Let $v_i := \sum_{j=i}^{n} u_j$ for $i \in \{0,1,\ldots,n\}$. Then $(u_0,\ldots,u_n)$ is monotonicity preserving if and only if v_0 is a constant function and the functions v_i are increasing for $i = 1,\ldots,n$.*

As a consequence of the previous proposition, we derive the following result.

THEOREM 3.3. *The Bernstein-like basis functions* $(b_{n,0}, b_{n,1}, \ldots, b_{n,n})$ *defined by* (1), (2) *is monotonicity preserving.*

PROOF. We use an induction on n, the order of basis functions. For $n = 2$, the basis, $(b_{2,0}, b_{2,1}, b_{2,2})$, is obviously monotonicity preserving according to the following calculations:

$$\begin{aligned} v_0 &= \sum_{j=0}^{2} b_{2,j}(t) = 1, \\ v_1 &= \sum_{j=1}^{2} b_{2,j}(t) = b_{2,1}(t) + b_{2,2}(t) = -\lambda t^4 + 4\lambda t^3 - (5\lambda + 1)\, t^2 + 2\,(\lambda + 1)\, t, \\ &\Rightarrow v_1'(t) = -4\lambda t^3 + 12\lambda t^2 - 2\,(5\lambda + 1)\, t + 2\,(\lambda + 1) \geq 0, \end{aligned}$$

$$v_2 = \sum_{j=2}^{2} b_{2,j}(t) = b_{2,2}(t) = \left(1 - \lambda + \lambda t^2\right) t^2 \Rightarrow v_2'(t) = 4\lambda t^3 + 2t\,(1 - \lambda) \geq 0.$$

Suppose that the Bernstein-like basis functions of order $n - 1$, i.e.

$$(b_{n-1,0}, b_{n-1,1}, \ldots, b_{n-1,n-1}),$$

is monotonicity preserving, for the basis, $(b_{n,0}, b_{n,1}, \ldots, b_{n,n})$ to be monotonicity preserving one needs v_0 to be constant, which is true according to partition of unity of basis functions. Moreover functions v_i must be increasing for $i = 1, \ldots, n$.

$$v_i = \sum_{j=i}^{n} b_{n,j}(t) \Rightarrow v_i'(t) = \sum_{j=i}^{n} b_{n,j}'(t). \tag{3}$$

Using the recursion relation, (2) we can rewrite (3) as follows:

$$\begin{aligned} v_i'(t) &= -\sum_{j=i}^{n} b_{n-1,j}(t) + (1-t)\sum_{j=i}^{n} b_{n-1,j}'(t) + \sum_{j=i}^{n} b_{n-1,j-1}(t) + t\sum_{j=i}^{n} b_{n-1,j-1}'(t) \\ &= b_{n-1,i-1}(t) + (1-t)\sum_{j=i}^{n-1} b_{n-1,j}'(t) + t\sum_{j=i-1}^{n-1} b_{n-1,j}'(t), \end{aligned}$$

which is non-negative in view of induction hypothesis and non-negativity of basis functions. □

The Bézier-like curve is a curve in $\mathbb{R}^2$ plane, so we need the following statement to cover this case. The proof is straightforward.

PROPOSITION 3.4. *Let* $(u_0, \ldots, u_n)$ *be a system of functions that is monotonicity preserving, if* $\{P_i\}_{i=0}^{n} \subseteq \mathbb{R}^2$ *be a monotone data then the function* $\sum_{i=0}^{n} P_i u_i$ *is increasing.*

4. Shape Preserving Interpolation

4.1. Monotone Interpolation. Let $\{(x_i, f_i)\}_{i=0}^n$ be a monotone data defined over the interval $[0, 1]$, we wish to find a smooth interpolant $p(x)$ on $[x_0, x_n]$ which preserves monotonicity.

We present a piece-wise Bézier-like curve as a solution to our monotone interpolation problem. For each sub-interval $[x_i, x_{i+1}]$, two auxiliary points are added and a Bézier-like curve is constructed using the control points

$$\{(x_i, f_i), (h_i, g_i), (t_i, z_i), (x_{i+1}, f_{i+1})\},$$

where h_i, t_i, g_i, z_i are unknown values. We need the piece-wise Bézier-like curve to be C^1, so in any two consecutive sub-intervals a continuity condition is imposed. Moreover we need the restrictions $x_i < h_i < t_i < x_{i+1}$ and $f_i \le g_i \le z_i \le f_{i+1}$ for having a monotone curve in each interval $[x_i, x_{i+1}]$, in this way the overall curve will be monotone. These constraints could be represented in the following system

$$\begin{cases} z_i + g_{i+1} = 2f_{i+1}, & i = 0, \dots, n-2, \\ t_i + h_{i+1} = 2x_{i+1}, & i = 0, \dots, n-2, \\ x_i < h_i < t_i < x_{i+1}, & i = 0, \dots, n-1, \\ f_i \le g_i \le z_i \le f_{i+1}, & i = 0, \dots, n-1. \end{cases}$$

This system is feasible and to present a solution, for any $g_0 \in [f_0, f_1]$ and $z_{n-1} \in [f_{n-1}, f_n]$, we assume

$$\begin{cases} z_i = g_{i+1} = f_{i+1},\ t_i = x_{i+1} - s, & i = 0, \dots, n-2, \\ h_i = x_i + s, & i = 1, \dots, n-1, \\ t_{n-1} = x_{n-1} + 2s,\ h_0 = x_1 - 2s. \end{cases}$$

Then it suffices to choose the parameter s according to

$$0 < s < \min\left\{\frac{x_1 - x_0}{2}, \frac{x_2 - x_1}{2}, \dots, \frac{x_{n-1} - x_{n-2}}{2}, \frac{x_n - x_{n-1}}{2}\right\}. \tag{4}$$

4.2. Bounded Interpolation. Consider $\{(x_i, f_i)\}_{i=0}^n$ as a data set, where $\{x_i\}$ are distinct points in real line and $\{f_i\}$ are bounded values, say $f_i \in [m, M]$. Without loss of generality, we assume that $m = 0$ and $M = 1$, so we wish to find a smooth interpolant $p(x)$ on $[x_0, x_n]$ with $0 \le p(x) \le 1$. For a given set of control points $\{P_i\} \subseteq \mathbb{R}^2$, the curve which is presented as $r_n(t) = \sum_{i=0}^n b_{n,i}(t)P_i$, is a curve which passes through the end points P_0 and P_n and also will always be completely contained inside of the convex hull of the control points. We get use of this fact and present a piece-wise Bézier-like curve as a solution to our bounded interpolation problem. For each sub-interval $[x_i, x_{i+1}]$, a Bézier-like curve is constructed using the control points $\{(x_i, f_i), (h_i, g_i), (t_i, z_i), (x_{i+1}, f_{i+1})\}$, where h_i, t_i, g_i, z_i are unknown values. We need C^1 continuity and moreover we need the restriction $z_i, g_i \in [0, 1]$, so the overall curve would be smooth and bounded into $[0, 1]$. These

constraints are illustrated as follows

$$\begin{cases} z_i + g_{i+1} = 2f_{i+1}, & i = 0, \dots, n-2, \\ t_i + h_{i+1} = 2x_{i+1}, & i = 0, \dots, n-2, \\ f_i \le z_i, g_i \le f_{i+1}, & i = 0, \dots, n-1, \\ x_i \le h_i, t_i \le x_{i+1}, & i = 0, \dots, n-1, \end{cases}$$

According to the convex hull property, this system always have a feasible solution which could be obtained similar to the monotone case, (4), so we would have a bounded piece-wise curve which interpolates the given data.

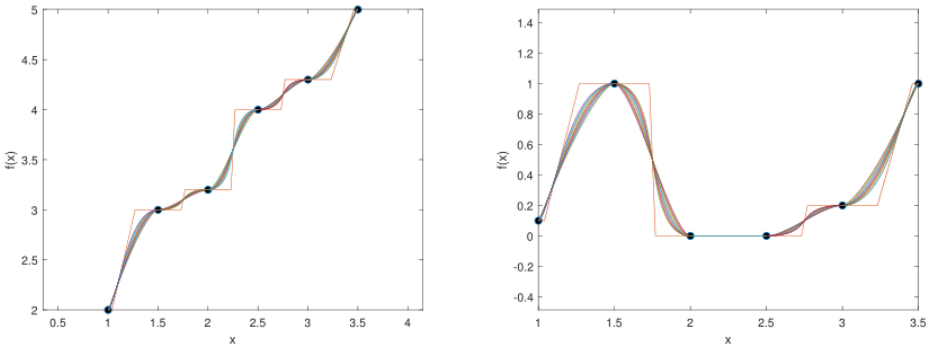
4.3. Demonstration. We illustrate the proposed methods in previous sections through numerical examples. Let us take the monotone data given in Table 1 and bounded data given in Table 2, that presents data sampled from known functions, we tried to use extreme cases to show the validity and reliability of proposed techniques. Figure 1-A demonstrates the monotone case and one can see a spectrum of curves for different values of free parameter $\lambda = -1, -0.9, \dots, 0.1, 0.2, \dots, 1$. From Figure 1-B, it is clear that the shape of the bounded data have been preserved.

TABLE 1. Monotone data.

x	1	1.5	2	2.5	3	3.5
$f(x)$	2	3	3.2	4	4.3	5

TABLE 2. Bounded data.

x	1	1.5	2	2.5	3	3.5
$f(x)$	0.1	1	0	0	0.2	1



(A) Monotonicity preserving interpolation.

(B) Bound preserving interpolation.

FIGURE 1. Shape preserving interpolations with different shape parameters.

References

1. G. Beliakov, *Monotonicity preserving approximation of multivariate scattered data,* BIT Numer. Math. **45** (2005) 653–677.
2. J. M. Carnicer, M. Garcia-Esnaola and J. M. Peña, *Convexity of rational curves and total positivity,* J. Comput. Appl. Math. **71** (2) (1996) 365–382.
3. M. Z. Hussain and M. Hussain, *Visualization of data preserving monotonicity,* Appl. Math. Comput. **190** (2) (2007) 1353–1364.

4. A. N. H. Tahat, A. R. M. Piah and Z. R. Yahya, *Shape preserving data interpolation using rational cubic ball functions,* J. Appl. Math. **2015** (2015) 1–9.
5. L. Yan and J. Liang, *An extension of the Bézier model,* Appl. Math. Comput. **218** (6) (2011) 2863–2879.

E-mail: j.saeidian@khu.ac.ir
E-mail: std_nouri411@khu.ac.ir

The Numerical Solution of Two Dimensional Variable-Order Galilei Advection Diffusion Equation

Marziyeh Saffarian*
Faculty of Mathematical Sciences, University of Kashan, Kashan, Iran
and Akbar Mohebbi
Faculty of Mathematical Sciences, University of Kashan, Kashan, Iran

Abstract. At the present work, a numerical scheme with first order temporal accuracy is developed to simulate two dimensional variable-order Galilei invariant advection diffusion equation with nonlinear source. We use the collocation meshless method to discretize this equation in spatial direction. Finally, we consider a test problem to demonstrate the accuracy and applicability of the proposed method.

Keywords: Variable-order Galilei invariant advection diffusion equation, Meshless method, Radial basis function, Thin plate spline.
AMS Mathematical Subject Classification [2010]: 65M50, 65M70, 65N35.

1. Introduction

In this work, we propose a numerical scheme for the solution of two dimensional variable-order Galilei invariant advection diffusion equation [2]

$$(1)\quad \begin{cases} u_t + \nabla u = D_{0,t}^{1-\gamma(\mathbf{x},t)}\Delta u(\mathbf{x},t) + f(u,\mathbf{x},t), & (\mathbf{x},t)\in\Omega\times(0,T],\\ u(\mathbf{x},0)=\varphi(\mathbf{x}), & \mathbf{x}\in\Omega,\\ u(\mathbf{x},t)=h(\mathbf{x},t), & (\mathbf{x},t)\in\partial\Omega\times(0,T], \end{cases}$$

where $\Omega\subset\mathbf{R}^2$, $0<\gamma_{\min}\le\gamma(\mathbf{x},t)\le\gamma_{\max}<1$ and $D_{0,t}^{1-\gamma(\mathbf{x},t)}u(\mathbf{x},t)$ is the variable-order Riemann-Liouville fractional partial derivative of order $1-\gamma(\mathbf{x},t)$ for $u(\mathbf{x},t)$ defined by [5]

$$D_{0,t}^{1-\gamma(\mathbf{x},t)}\Delta u(\mathbf{x},t)=\frac{1}{\Gamma(\gamma(\mathbf{x},t))}\frac{d}{dt}\int_0^t (t-s)^{-\gamma(\mathbf{x},t)-1}u(\mathbf{x},s)ds.$$

The radial basis function is one of common meshless method. This method was used to interpolate scatter data by Roland Hardy [3]. In 1990, Kansa developed the radial basis function method to solve the partial differential equations [4]. This method is used to diversity of the partial differential equations [6, 7, 8].

In this work, we use a WSGD scheme to approximate the variable-order Riemann-Liouville derivative in Eq. (1). In general, we approximate Eq. (1) with finite difference scheme in the temporal direction and we apply thin plate spline radial basis functions to discretize this equation in spatial direction. Then we obtain a scheme of order $\mathcal{O}(\tau)$ for defining nonlinear Eq. (1).

*Speaker

The layout of this paper is as follows: In Section 2, we explain the radial basis functions approximation method. In Section 3, we present a numerical scheme which one order of accuracy in time variable and use thin plate spline radial basis functions in space components to obtain the fully discrete scheme. The numerical results of the proposed method are given in Section 4. In Section 5, a brief conclusion is expressed. Some references that are used in this work are introduced at the end of this paper.

2. Thin Plate Spline Approximation

An unknown function $u(\mathbf{x})$ at node $\mathbf{x}$ in domain Ω can be approximated as

$$u(\mathbf{x}) \simeq \sum_{j=1}^{M} \eta_j \phi(r_j) + \psi(\mathbf{x}), \quad \mathbf{x} \in \Omega \subset R^d, \tag{2}$$

where M is the number of nodes in Ω, η_j are unknown coefficients, $\phi(r_j)$ is the radial basis function, $r_j = ||\mathbf{x} - \mathbf{x}_j||$ is Euclidean norm and $\psi(\mathbf{x})$ is the additional polynomial which can be omitted if $\phi(r_j)$ be unconditionally positive definite.

Let $p_1, p_2, \ldots, p_l$ be the basis of P_q^d (the space of $d-$variate polynomials of order not exceeding q), then ψ can be written as

$$\psi(\mathbf{x}) = \sum_{i=1}^{l} \xi_i p_i(\mathbf{x}), \tag{3}$$

where $l = \dfrac{(q-1+d)!}{d!(q-1)!}$.

To determine unknown coefficients $\eta_1, \ldots, \eta_M$ and $\xi_1, \ldots, \xi_l$, the collocation method is used. Then l equations are needed. The required equations can be achived by putting

$$\sum_{j=1}^{M} \eta_j p_i(\mathbf{x}_j) = 0. \tag{4}$$

Now let $\mathcal{L}$ be the linear partial differential operator, then we have

$$\mathcal{L}u(\mathbf{x}) \simeq \sum_{j=1}^{M} \eta_j \mathcal{L}\phi(r_j) + \mathcal{L}\psi(\mathbf{x}), \quad \mathbf{x} \in \Omega \subset R^d. \tag{5}$$

At the present work, we use thin plate spline radial basis functions which are the following form

$$\phi(r_j) = r_j^{2m} log(r_j), \qquad j = 1, 2, 3, \ldots. \tag{6}$$

In the numerical simulation we get $m = 6$.

3. The Proposed Numerical Scheme

In this section, we use WSGD scheme to approximate the variable order Riemann-Liouville derivative as [1]

$$D_{0,t}^{\alpha(\mathbf{x},t)} u(\mathbf{x}, t_n) = \frac{1}{\tau^{\alpha(\mathbf{x},t_n)}} \sum_{j=0}^{n} w_j^{\alpha(\mathbf{x},t_n)} u^{n-j} + O(\tau^2),$$

where $0 < \alpha(\mathbf{x}, t) < 1$ and

$$w_0^{\alpha(\mathbf{x},t_n)} = \frac{2+\alpha(\mathbf{x},t_n)}{2}, \qquad w_j^{\alpha(\mathbf{x},t_n)} = \frac{2+\alpha(\mathbf{x},t_n)}{2} g_j - \frac{\alpha(\mathbf{x},t_n)}{2} g_{j-1},$$

and

$$g_0^{\alpha(\mathbf{x},t_n)} = 1, \qquad g_j^{\alpha(\mathbf{x},t_n)} = (1 - \frac{\alpha(\mathbf{x},t_n)+1}{j}) g_{j-1}.$$

We use the mentioned scheme and Crank-Nicolson idea to Eq. (1), then we get

$$\begin{aligned} (7) \qquad \delta_t u^{n+1/2} + \frac{1}{2}(\nabla u^{n+1} + \nabla u^n) &= \frac{1}{2}(\frac{1}{\tau^{\beta^{n+1}}} \sum_{j=0}^{n} w_j^{\beta^{n+1}} \Delta u^{n+1-j} \\ &+ \frac{1}{\tau^{\beta^n}} \sum_{j=0}^{n} w_j^{\beta^n} \Delta u^{n-j}) + f(u^n) + O(\tau), \end{aligned}$$

where $u^n = u(\mathbf{x}, t_n)$, $\delta_t u^{n+1/2} = (u^{n+1} - u^n)/\tau$ and $\beta^n = 1 - \gamma^n$. We can approximate $u^n(\mathbf{x})$ for M interpolation nodes using relations (2)-(6). Then we have the matrix form of approximation of u^n as

$$[u]^n = B[\eta]^n.$$

We can splite the matrix B into $B = B_d + B_b + B_e$, where

$$B_d = [b_{d_{ij}}] = \begin{cases} b_{ij}, & \mathbf{x}_i \in \Omega, \\ 0, & \mathbf{x}_i \in \partial\Omega, \end{cases}, \quad B_b = [b_{b_{ij}}] = \begin{cases} b_{ij}, & \mathbf{x}_i \in \partial\Omega, \\ 0, & \mathbf{x}_i \in \Omega, \end{cases}$$

$$B_e = [b_{e_{ij}}] = \begin{cases} b_{ij}, & i = M+1 : M+3, \\ 0, & \text{elsewhere.} \end{cases}$$

Now we omit the small term in Eq. (7) and approximate $u^n(\mathbf{x})$ for M interpolation nodes using (2)-(6). Then we apply the collocation method to determine the interpolation coefficients. Finally we obtain the matrix equation as

$$C[\eta]^{n+1} = (B_d - \frac{\tau}{2}\nabla B_d)[\eta]^n + E^n + \tau[f]^n + [H]^{n+1},$$

where

$$\begin{aligned} (8) \qquad C &= B_d + \frac{\tau}{2}\nabla B_d - \frac{1}{2}[\tau^{1-\beta_{n+1}} w_0^{\alpha_{n+1}}] * \Delta B_d + B_b + B_e, \\ [H]^{n+1} &= \begin{cases} h^{n+1}(\mathbf{x}_i), & \mathbf{x}_i \in \partial\Omega, \\ 0, & \mathbf{x}_i \in \Omega, \end{cases} \\ E^n &= \frac{1}{2}[\tau^{1-\beta_{n+1}}] * \sum_{j=1}^{n+1} [w_j^{\beta_{n+1}}] * \Delta u_d^{n+1-j} + \frac{1}{2}[\tau^{1-\beta_n}] * \sum_{j=0}^{n} [w_j^{\beta_n}] * \Delta u_d^{n-j}, \end{aligned}$$

and $\nabla u_d^n = \nabla B_d[\eta]^n$. The vectors $[\tau^{1-\beta^{n+1}} w_0^{\beta_{n+1}}]$, $[w_j^{\beta_{n+1}}]$, $[w_j^{\beta_n}]$, $[\tau^{1-\beta_n}]$, $[f]^n$ are calculated using the collocation nodes $\mathbf{x}_i \in \overline{\Omega} = \Omega \cup \partial\Omega$. The symbol $*$ in Eq. (8) represents the multiplication of component by component.

4. Numerical Results

In this section, we report the numerical experiment of TPS-RBF method for a 2D test problem. Let E_1 and E_2 are error correspond to time steps τ_1 and τ_2, then computational order of the presented method can be calculate as

$$\text{C-order}= \frac{\log \frac{E_1}{E_2}}{\log \frac{\tau_1}{\tau_2}}.$$

EXAMPLE 4.1. Consider the Eq. (1) with the exact solution $u(\mathbf{x},t) = t^2 e^{x+y}$ [2]. We use TPS-RBF method on square and circle domains with $\gamma(\mathbf{x},t) = \dfrac{10-xyt}{300}$ and present the L_∞ norm of errors and computational order of the proposed method in Table 1. The graphs of numerical solution and absolute error for this problem are presented in Figure 1.

TABLE 1. Errors and computational orders for test problem 1.

	square domain with M=441		circle domain with M=476	
τ	L_∞	C-order	L_∞	C-order
1/25	2.1481×10^{-2}		8.2985×10^{-3}	
1/50	1.0603×10^{-2}	1.0186	4.0631×10^{-3}	1.0303
1/100	5.2842×10^{-3}	1.0047	2.0180×10^{-3}	1.0000
1/200	2.6380×10^{-3}	1.0022	1.0060×10^{-3}	1.0043
1/400	1.3182×10^{-3}	1.0009	5.0228×10^{-4}	1.0021

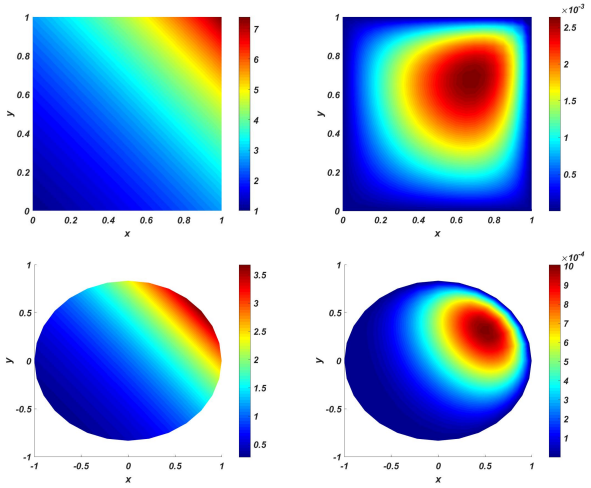


FIGURE 1. Numerical solution (left side) and absolute error (right side) with $\tau = 0.005$.

5. Conclusion

In this article, we studied the TPS-RBF method for the solution of two dimensional variable-order Galilei invariant advection diffusion equation with nonlinear source term. A finite difference scheme of order $\mathcal{O}(\tau)$ is presented for discretizing temporal direction. The accuracy of the present method is shown by an example.

References

1. J. Cao, Y. Qiu and G. Song, *A compact finite difference scheme for variable order subdiffusion equation*, Commun. Nonlinear. Sci. Numer. Simulat. **48** (2017) 140–149.
2. Ch. M. Chena, F. Liub, V. Anhb and I. Turner, *Numerical simulation for the variable-order Galilei invariant advection diffusion equation with a nonlinear source term,* Appl. Math. Comput. **217** (2011) 5729–5742.
3. R. L. Hardy, *Theory and applications of the multiquadric-biharmonic method.* 20 *years of discovery* 1968 – 1988, Comput. Math. Appl. **19** (8-9) (1990) 163–208.
4. EJ. Kansa, *Multiquadrics-a scattered data approximation scheme with applications to computational fluid dynamices-I*, Comput. Math. Appl. **19** (1990) 127–145.
5. CH. Li and F. Zeng, *Numerical Methods for Fractional Calculus*, CRC Press, New york, 1999.
6. L. Ling and E. Kansa, *Preconditioning for radial basis functions with domain decompositions methods*, Math. Comput. Model. **40** (2004) 1413–1427.
7. A. Mohebbi and M. Saffarian, *Implicit RBF meshless method for the solution of two-dimensional variable order fractional cable equation*, J. Appl. Comput. Mech. **6** (2020) 235–247.
8. F. Zabihi and M. Saffarian,*A not-a-knot meshless method with radial basis functions for numerical solutions of Gilson-Pickering equation*, Eng. Comput. **34** (2018) 37–44.

E-mail: m.Saffarian11@grad.kashanu.ac.ir
E-mail: a_mohebbi@kashanu.ac.ir

Numerical Solution of Stochastic Black-Scholes-Merton Model Occuring in Financial Market

Nasrin Samadyar*
Department of Mathematics, Faculty of Mathematical Sciences, Alzahra University, Tehran, Iran
and Yadollah Ordokhani
Department of Mathematics, Faculty of Mathematical Sciences, Alzahra University, Tehran, Iran

Abstract. Providing a suitable method for solving stochastic Black-Scholes-Merton model and investigating the efficiency of the proposed method are the most important purposes of this paper. This technique, which is based on operational matrices of hat functions, converts the mentioned model into a linear system of algebraic equations. Numerical results confirm accuracy and efficiency of suggested method.
Keywords: Stochastic differential equations, Operational matrix method, Hat functions.
AMS Mathematical Subject Classification [2010]: 60H10, 65L05.

1. Introduction

Recently, stochastic differential equations (SDEs) are employed as a widely used mathematical tool for modelling various problems in reactor dynamics, the growth of populations, financial markets, and etc. Solving SDEs exactly due to stochastic factors is very complicated or even impossible and in recent decade the attention of researchers have been attracted into numerical methods to estimate their numerical solution. For instance, Runge-Kutta method [7], collocation method [2], Wong-Zakai method [6], and meshless method [1] have been applied to provide the numerical solution of different SDEs.

The aim of this paper is approximating the solution of stochastic Black-Scholes-Merton model, which is one of the most important SDEs in financial market. The stochastic Black-Scholes-Merton model with uncertain interest rate r is formulated as follows [3]

$$\begin{cases} d\mathcal{X}(\tau) = r\mathcal{X}(\tau)d\tau + \sigma\mathcal{X}(\tau)d\mathcal{B}(\tau), & \tau \in [0,T], \\ \mathcal{X}(0) = \mathcal{X}_0, \end{cases} \tag{1}$$

where r, σ and $\mathcal{X}_0$ are known real numbers, $\mathcal{B}(\tau)$ denotes standard Brownian motion and $\mathcal{X}(\tau)$ is an unknown stochastic process which should be approximated.

*Speaker

2. Preliminaries

Definition 2.1. [5] Consider a vector of modified hat functions $\phi(\tau) = [\varphi_0(\tau), \varphi_1(\tau), \ldots, \varphi_n(\tau)]^T$, such that its components are defined as follows

$$\varphi_0(\tau) = \begin{cases} \frac{1}{2h^2}(\tau - h)(\tau - 2h), & 0 \le \tau \le 2h, \\ 0, & \text{otherwise.} \end{cases}$$

If $1 \le i \le n-1$ be an odd number, then

$$\varphi_i(\tau) = \begin{cases} \frac{-1}{h^2}(\tau - (i-1)h)(\tau - (i+1)h), & (i-1)h \le \tau \le (i+1)h, \\ 0, & \text{otherwise.} \end{cases}$$

If $2 \le i \le n-2$ be an even number, then

$$\varphi_i(\tau) = \begin{cases} \frac{1}{2h^2}(\tau - (i-1)h)(\tau - (i-2)h), & (i-2)h \le \tau \le ih, \\ \frac{1}{2h^2}(\tau - (i+1)h)(\tau - (i+2)h), & ih \le \tau \le (i+2)h, \\ 0, & \text{otherwise,} \end{cases}$$

and

$$\varphi_n(\tau) = \begin{cases} \frac{1}{2h^2}(\tau - (T-h))(\tau - (T-2h)), & T-2h \le \tau \le T, \\ 0, & \text{otherwise,} \end{cases}$$

where $n \ge 2$ is an even integer number and $h = \frac{T}{n}$.

Theorem 2.2. *The ordinary and stochastic integrals of vector $\phi(\tau)$ can be estimated as follows*

$$\int_0^\tau \phi(\varsigma)d\varsigma \simeq \mathcal{P}_o\phi(\tau), \qquad \int_0^\tau \phi(\varsigma)d\mathcal{B}(\varsigma) \simeq \mathcal{P}_s\phi(\tau), \tag{2}$$

where $\mathcal{P}_o$ and $\mathcal{P}_s$ denote ordinary and stochastic operational matrices of integration based on modified hat functions, respectively and have been calculated in paper [4].

3. Numerical Method

The given SDE in Eq. (1) can be written as the following stochastic Volterra integral equation

$$\mathcal{X}(\tau) = \mathcal{X}_0 + r\int_0^\tau \mathcal{X}(\varsigma)d\varsigma + \sigma\int_0^\tau \mathcal{X}(\varsigma)d\mathcal{B}(\varsigma). \tag{3}$$

The unknowm process $\mathcal{X}(\tau)$ and initial condition $\mathcal{X}_0$ in Eq. (3) are approximated via modified hat functions as follows

$$\mathcal{X}(\tau) \simeq \mathcal{X}_n(\tau) = \sum_{i=0}^{n} x_i\varphi_i(\tau) = X^T\phi(\tau), \qquad \mathcal{X}_0 \simeq \Upsilon^T\phi(\tau), \tag{4}$$

where

$$X = [x_0, x_1, \ldots, x_n]^T, \ \phi(\tau) = [\varphi_0(\tau), \varphi_1(\tau), \ldots, \varphi_n(\tau)]^T,$$
$$\Upsilon = [\mathcal{X}_0, \mathcal{X}_0, \ldots, \mathcal{X}_0]^T.$$

By inserting Eq. (4) into Eq. (3) and using operational matrices given in Eq. (2), we have

$$X^T\phi(\tau) \simeq \Upsilon^T\phi(\tau) + rX^T\mathcal{P}_o\phi(\tau) + \sigma X^T\mathcal{P}_s\phi(\tau). \tag{5}$$

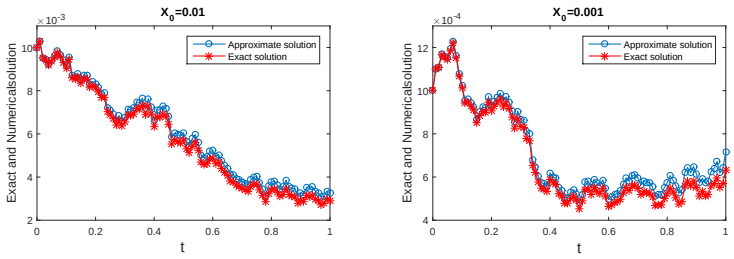


FIGURE 1. Investigating the effect of initial value $\mathcal{X}_0$ on numerical solution.

Deleting $\phi(\tau)$ from both sides of Eq. (5) and rewriting it results

$$X^T(I - r\mathcal{P}_o - \sigma\mathcal{P}_s) = \Upsilon^T. \tag{6}$$

After solving obtained linear system (6) and computing unknown vector X^T, the numerical solution of SDE (1) is calculated from Eq. (4).

4. Numerical Experiments

In this section, Black-Scholes-Merton model, which has been introduced in Eq. (1), is numerically solved via presented method to demonstrate accuracy of explained method. Using Itô formula, it is proved that the strong solution of this SDE is given by a geometric Brownian motion as follows

$$\mathcal{X}(\tau) = \mathcal{X}_0 \exp\Big((r - \frac{\sigma^2}{2})\tau + \sigma\mathcal{B}(\tau)\Big).$$

The accuracy of the proposed method are measured by the max-error and RMS-error criterions which are defined as follows

$$\text{max-error} = \max_{i=0,1,\ldots,n} \big|\mathcal{X}(\tau_i) - \mathcal{X}_n(\tau_i)\big|, \qquad \text{RMS-error} = \sqrt{\frac{1}{n+1}\sum_{i=0}^{n}\big|\mathcal{X}(\tau_i) - \mathcal{X}_n(\tau_i)\big|^2}.$$

Computation time (CPU time) and condition number (cond) of obtained coefficient matrix are also reported in Tables to confirm the efficiency of our method.

The effect of the initial value $\mathcal{X}_0$ on the numerical solution has been investigated in Table 1 and Figure 1. In this situation, we solve mentioned SDE for $T = 1, n = 100, r = 0.15, \sigma = 0.5$ and $\mathcal{X}_0 = 0.01, 0.001$, and we obtain more accurate results for smaller value of $\mathcal{X}_0$. Also, we solve this model for $T = 1, r = 0.18, \sigma = 1, \mathcal{X}_0 = 0.001$ and $n = 50, 100$ and we obtain more accurate results for $n = 100$ (See Table 2 and Figure 2). This experiment shows that a more accurate solution is obtained by considering more hat functions. In the third expriment, which its results have been reported in Table 3 and Figure 3, we consider $T = 1, r = 0.2, n = 50, \mathcal{X}_0 = 0.01$ and three different values $\sigma = 0.1, 0.3, 0.5$. These results show that better results are obtained by reducing the amount of σ.

TABLE 1. The effect of initial value $\mathcal{X}_0$ on numerical solution.

	max-error	RMS-error	CPU time (s)	cond
$\mathcal{X}_0 = 0.010$	3.9343e-04	2.9394e-04	4.044362	10.8771
$\mathcal{X}_0 = 0.001$	8.0705e-05	4.0064e-05	2.944271	9.7762

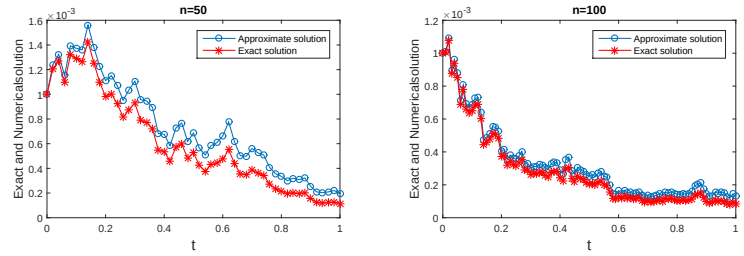


FIGURE 2. Investigating the effect of parameter n on numerical solution.

TABLE 2. The effect of parameter n on numerical solution.

	max-error	RMS-error	CPU time (s)	cond
n=50	2.2497e-04	1.3561e-04	3.582896	12.3072
n=100	6.3561e-05	4.2029e-05	3.516834	32.2248

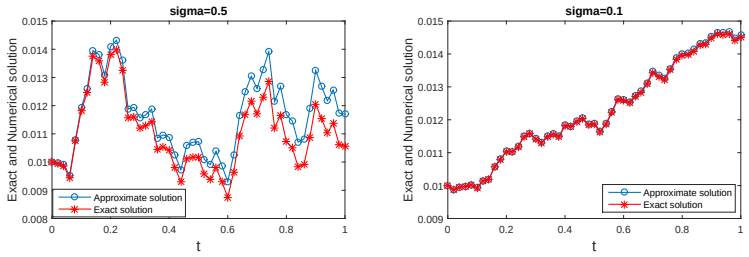


FIGURE 3. Investigating the effect of parameter σ on numerical solution.

5. Conclusion

In this paper, an efficient algorithm has been applied to solve stochastic Black-Scholes-Merton model, which is one of the most important mathematical models in financial markets. To provide more accurate numerical results, we proceed some different ways:

- increasing the number of used hat functions n,
- reduce the amount of parameter σ,
- reduce the value of initial condition $\mathcal{X}_0$.

TABLE 3. The effect of parameter σ on numerical solution.

	max-error	RMS-error	CPU time (s)	cond
$\sigma = 0.5$	1.2568e-03	6.8702e-04	2.673566	6.1187
$\sigma = 0.3$	4.9705e-04	2.5713e-04	3.365316	3.7837
$\sigma = 0.1$	6.3564e-05	3.2254e-05	3.418961	1.5294

Acknowledgement

This work is supported by the National Elites Foundation and Alzahra University. The authors would like to express our very great appreciation to reviewers for their valuable comments which have improved the quality of our paper.

References

1. M. Dehghan and M. Shirzadi, *A meshless method based on the dual reciprocity method for one-dimensional stochastic partial differential equations*, Numer. Methods. Partial Differential Eq. **32** (1) (2016) 292–306.
2. M. Kamrani and S. M. Hosseini, *Spectral collocation method for stochastic Burgers equation driven by additive noise*, Math. Comput. Simulat. **82** (9) (2012) 1630–1644.
3. F. C. Klebaner, *Introduction to Stochastic Calculus with Applications*, World Scientific Publishing Company, 2005.
4. F. Mirzaee and E. Hadadiyan, *Numerical solution of Volterra-Fredholm integral equations via modification of hat functions*, Appl. Math. Comput. **280** (2016) 110–123.
5. F. Mirzaee and N. Samadyar, *Numerical solution of nonlinear stochastic Itô-Volterra integral equations driven by fractional Brownian motion*, Math. Meth. Appl. Sci. **41** (2018) 1410–1423.
6. S. Şengül and M. Merdan, *Wong-Zakai method applications for explicitly solvable stochastic differential equations*, J. Adv. Math. Comput. Sci. (2019) 1–12.
7. A. Tocino and R. Ardanuy, *Runge–Kutta methods for numerical solution of stochastic differential equations*, J. Comput. Appl. Math. **138** (2) (2002) 219–241.

E-mail: n.samadyar@alzahra.ac.ir
E-mail: ordokhani@alzahra.ac.ir

Extrapolated Iterative Method for Solving Absolute Value Equations

Somayeh Seifollahzadeh*
Faculty of Mathematical Sciences, University of Tabriz, Tabriz, Iran
and Ghodrat Ebadi
Faculty of Mathematical Sciences, University of Tabriz, Tabriz, Iran

Abstract. In this paper, we present a generalized Newton Gauss-Seidel iteration method (NGS) to solve absolute value equations. Also we introduce extrapolated version of NGS method (ENGS) to increase the rate of convergence. Furthermore, we find upper bound for extrapolation parameter and discuss the convergence of proposed methods. Finally the efficiency of methods are illustrated by giving several examples.
Keywords: Absolute value equation, Gauss-Seidel iteration, Extrapolation Method, Convergence.
AMS Mathematical Subject Classification [2010]: 65F10, 90C05, 90C30.

1. Introduction

Consider the absolute value equation (AVE)

$$\mathcal{A}x + \mathcal{B}|x| = \mathbf{b},$$

where $\mathcal{A}, \mathcal{B} \in \mathcal{R}^{n\times n}, \mathbf{b} \in \mathcal{R}^n$ and $|x|$ indicates absolute value of each component of vector x, when $det(\mathcal{B}) \neq 0$ the above equation can be simplified to

$$Ax - |x| = b. \tag{1}$$

There exists a large body of literature in mathematics, physics, and engineering in the form of linear complementary problem (LCP) which can be written as AVE. It includes many programming problems as linear programming, convex quadratic programming, etc [7]. These programming problems are intensively used in many fields, e.g. modeling contact force, fluid in computational mechanic [1] and in finding Nash equilibrium of Bimatrix game [3]; Also it is proven NP-hard knapsack feasibility problems are AVE [6]. AVE unavoidably appears in solving interval systems of linear equations [8].

In recent decades many efforts have been devoted to showing the existence and numerical solution of AVE. We refer to some of them below. Mangasarian and Meyer discussed the existence and nonexistence of solution for (1) [7]. Prokopyev studied unique solvability of AVE and its relations with LCP and mixed it with integer programming [9]. Mangasarian generalized the Newton method (GN) to solve (1) when the singular values of A exceed 1 [10]. Salkuyeh proposed Picard-HSS (PHSS) iterative method to solve (1) when A is a nonsymmetric positive

*Speaker

matrix [11]. Also Edalatpour et al. presented generalized Gauss-Seidel (GGS) and its Preconditioned (PGGS) methods [5].

In this paper, we use the Gauss-Seidel method for solving the linear system that occurs at each iteration of Newton method for the solution of (1). Since the GN method is efficient and easy to code also it needs to calculate inverse of matrix $(A - diag(sign(x))$, that's not constant, in each iteration that takes too long. Therefore to overcome this draw back, we introduce generalized Newton Gauss-Seidel (NGS) method, where in each iteration of GN we estimate x^{k+1} using the Gauss-Seidel method in l_k iteration. Next, for improving the rate of convergence we solve (1) with an extrapolated version of NGS (ENGS), also convergence of them and finding upper bound for the extrapolation parameter is studied.

We organized the reminder of this paper as follows. In Section 2, some prerequisites are given. A generalized Newton Gauss-Seidel iterative method and it's convergence are described in Section 3, Section 4 is assigned to extrapolated generalized Newton-Gauss-Seidel iterative method, eventually numerical results are given in Section 5.

2. Preliminaries

To present the NGS method and its convergence we give some necessary lemma.

Proposition 2.1. [7]

i) *The AVE* (1) *is uniquely solvable for any* $b \in \mathbb{R}^n$ *if* $\|A^{-1}\|_2 < 1$.

ii) *The AVE* (1) *is uniquely solvable for any* $b \in \mathbb{R}^n$ *if the singular values of* A *exceed* 1.

Lemma 2.2. [4] *Let* M *be nonsingular matrix and* D *is an arbitrary matrix and* $\|M^{-1}D\|_2 < 1$. *Then,* $(M - D)$ *is invertible.*

Lemma 2.3. [4] *Let* $\|M^{-1}D\|_2 < 1$, *then* $(I - M^{-1}D)$ *is nonsingular and* $\|(I - (M^{-1}D))^{-1}\|_2 \leq \frac{1}{1-\|M^{-1}D\|_2}$.

Lemma 2.4. [2] *Let* M *be a nonsingular matrix and* $D = diag(d_i)$, *where* $d_i \in [-1, 1], i = 1, 2, \ldots, n$ *and* $\|M^{-1}\|_2 < 1$ *then* $(M - D)$ *is nonsingular and*

$$\|(M - D)^{-1}\|_2 \leq \frac{\|M^{-1}\|_2}{1 - \|M^{-1}D\|_2}.$$

3. The Generalized Newton Gauss-Seidel Method

In this paper, for solving (1) or $(A - D(x))x = b$, that $D(x) = diag(sign(x))$, we have

$$M_k x^{(k,l+1)} = N x^{(k,l)} + b, \;\; l = 0, 1, 2, \ldots, l_k - 1,$$

where $A - D(x^k) = A - D_k = M_k - N$, $M_k = tril(A) - D_k$, $N = triu(-A, 1)$ and $x^{(k+1)} = x^{(k,l_k)}$.

Let $\|M^{-1}\|_2 < 1$, according to Lemma 2.4, M_k is nonsingular. By considering $\{l_k\}$ as nonnegative integer numbers, at each iteration we have

$$x^{(k+1)} = G_k^{l_k} x^{(k)} + (G_k^{l_{k-1}} + G_k^{l_{k-2}} + \cdots + G_k + I) M_k^{-1} b,$$

where $G_k = M_k^{-1}N$. So

$$x^{(k+1)} = G_k^{l_k} x^{(k)} + (I - G_k^{l_k})(A - D_k)^{-1}b. \tag{2}$$

THEOREM 3.1. *Let matrix A in (1) be nonsingular, suppose $\|G\|_2 + \|M^{-1}\|_2 < 1$ and $\eta = \frac{2\|A^{-1}\|_2}{1-\|A^{-1}\|_2}$. If $\eta < 1$ then for nonnegative integer numbers $l_k, k = 0, 1, 2, \ldots$ and any initial value such $x^{(0)} \in \mathbb{R}^n$, the iteration sequence produced by NGS method converges to an exact solution of (1), provided that $\liminf_{k\to\infty} l_k \geq L$, where $L \in \mathbb{N}$ satisfying*

$$\|G_k^r\|_2 < \frac{1-\eta}{1+\eta} \quad \forall r \geq L.$$

PROOF. Considering $\|A^{-1}\|_2 < \frac{\|M^{-1}\|_2}{1-\|G\|_2}$, assumption $\|G\|_2 + \|M^{-1}\|_2 < 1$ and Lemma 2.2 yields (1) is uniquely solvable. Let $\tilde{x}$ be an exact solution of Eq. (1), therefore

$$(A - \tilde{D})\tilde{x} = (A - D_k - C_k)\tilde{x} = b, \tag{3}$$

where $\tilde{D} = D(\tilde{x})$, C_k is a diagonal matrix and $\tilde{D} = D_k - C_k$. Substituting (3) in (2) gets

$$x^{(k+1)} = G_k^{l_k} x^{(k)} + (I - G_k^{l_k})(I - (A - D_k)^{-1}C_k)\tilde{x}.$$

Then we have $C_k\tilde{x} = D_k(\tilde{x} - x^{(k)}) + |x^{(k)}| - |\tilde{x}|$, so

$$x^{(k+1)} - \tilde{x} = G_k^{l_k}(x^{(k)} - \tilde{x}) + (I - G_k^{l_k})(A - D_k)^{-1}(D_k(\tilde{x} - x^{(k)}) + |x^{(k)}| - |\tilde{x}|).$$

Since $\||x| - |y|\|_2 \leq \|x - y\|_2$, where $x, y \in \mathbb{R}^n$, therefore

$$\|x^{(k+1)} - \tilde{x}\|_2 \leq \|G_k^{l_k}\|_2\|x^k - \tilde{x}\|_2 + 2(1 + \|G_k^{l_k}\|)\|(A - D_k)^{-1}\|\|x^{(k)} - \tilde{x}\|_2,$$

using Lemma 2.4 we have

$$\|x^{(k+1)} - \tilde{x}\|_2 \leq \mu\|x^{(k)} - \tilde{x}\|_2, \tag{4}$$

where $\mu = (\|G_k^{l_k}\|_2 + 2\frac{\|A^{-1}\|_2(1+\|G_k^{l_k}\|_2)}{1-\|A^{-1}\|_2})$.

Since $\|G_k\|_2 \leq \frac{\|G\|_2}{1-\|M^{-1}\|_2}$, $\|G\|_2 + \|M^{-1}\|_2 < 1$ and $\eta < 1$ so

$$\lim_{r\to\infty} \|G_k\|_2^r \leq \lim_{r\to\infty} (\frac{\|G\|_2}{1 - \|M^{-1}\|_2})^r = 0 \quad k = 0, 1, \ldots,$$

therefore

$$\exists L \in \mathbb{N}, \forall k = 0, 1, \ldots \qquad s.t \qquad \|G_k^r\| < \frac{1-\eta}{1+\eta} \quad \forall r \geq L.$$

So if we set $\liminf_{k\to\infty} l_k \geq L$, using the above inequality and (4) give $\mu < 1$ and this completes the proof. □

4. Extrapolated Generalized Newton Gauss-Seidel Iterative Methods for AVE

In this part we introduce an extrapolated version of NGS method in order to increase the rate of convergence of it. In each outer iteration of NGS method, we replace $x^{(k+1)}$ by the extrapolated value

$$\alpha x^{(k+1)} + (1-\alpha)x^{(k)},$$

where $\alpha \in \mathbb{R} - \{0\}$.

Let $\{l_k\}$ be arbitrary nonnegative integer numbers, in extrapolated generalized Newton Gauss-Seidel, we calculate $x^{(k)}$ in each step using

$$x^{(k+1)} = (\alpha G_k^{l_k} + (1-\alpha)I)x^{(k)} + \alpha(I - G_k^{l_k})(A - D_k)^{-1}b.$$

If we assume $\|G\|_2 + \|M^{-1}\|_2 < 1$ and $0 < \alpha < \frac{2(1-\|M^{-1}\|_2)}{\|G\|_2 - \|M^{-1}\|_2 + 1}$, we will have

$$\rho(\alpha G_k^{l_k} + (1-\alpha)I) \leq |\alpha|\|G_k^{l_k}\|_2 + |1-\alpha| \leq \alpha\frac{\|G\|_2}{1-\|M^{-1}\|_2} + |1-\alpha| < 1.$$

THEOREM 4.1. *Let (1) be solvable,* $\|G\|_2 + \|M^{-1}\|_2 < 1$ *and also* $\eta = \frac{2\|A^{-1}\|_2}{1-\|A^{-1}\|_2}$. *If* $\eta < 1$ *then there exists an* $\alpha_0 > 1$ *such that for all* $0 < \alpha < \alpha_0$, *any initial value* $x^{(0)}$ *and arbitrary nonnegative integer numbers* $l_k, k = 0, 1, 2, \ldots$, *ENGS method converges to an exact solution, provided that* $\liminf_{k\to\infty} l_k \geq L$, *where* $L \in \mathbb{N}$ *satisfy*

$$\|G_k^r\|_2 < \frac{2 - \alpha(1+\eta)}{\alpha(1+\eta)}, \quad \forall r \geq L.$$

PROOF. Since the proof is exactly the same as previous ones, rewriting it would be redundant. □

5. Numerical Experiments

In this section, we are going to check and compare experimentally presented theories with GN, GGS and PGGS methods by several test problems. All the numerical experiments have been carried out by MATLAB R2015b (64-bit) and tested on PC with quad-core 4.2 GHz Intel Core i7 processor and 8GB Memory running by Windows 10. The process is followed by using zero vector as an initial value and the stopping criterion is

$$\frac{\|Ax^{(k)} - |x^{(k)}| - b\|}{\|b\|} \leq 10^{-7},$$

and we set a maximum number of iterations 2000. In all examples the vector $x = (x_1, x_2, \ldots, x_n)^T$ with $x_i = (-1)^i i, i = 1, 2, \ldots, n,$ is the exact solution. The optimal parameters of PGGS and ENGS methods and interval boundary of α have been found empirically. Notice that MATLAB backslash is used to solve system of equations in each iteration of generalized Newton method.

EXAMPLE 5.1. Let

$$\mathbf{A} = \begin{pmatrix} B & -I & 0 & 0 & \dots & 0 & 0 \\ -V & B & -I & 0 & \dots & 0 & 0 \\ 0 & -V & B & -I & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & B & -I \\ 0 & 0 & 0 & 0 & 0 & -V & B \end{pmatrix} \in \mathbb{R}^{n\times n},$$

where $V = tridiag(-3, 1, -2) \in \mathbb{R}^{m\times m}$, $B = tridiag(1, 4, -1) \in \mathbb{R}^{m\times m}$, $I \in \mathbb{R}^{m\times m}$ is identity matrix and $n = m^2$.

We consider (1), with above matrix A and present the numerical results in Table 1.

Notice that for $\alpha \in (0, 1.3)$, ENGS method in this example is convergent.

TABLE 1. Numerical result for Example 5.1.

Method	n	400	900	1600	2500	10000
NGS	Iter.	20	29	38	58	203
	CPU	0.008	0.016	0.030	0.063	1.014
ENGS	Iter.	14	20	27	36	89
	CPU	0.004	0.009	0.019	0.038	0.444
	α	0.76	0.76	0.76	0.76	0.76
GN	Iter.	4	7	9	12	–
	CPU	0.009	0.022	0.056	0.118	–
GGS	Iter.	112	–	–	–	–
	CPU	0.017	–	–	–	–
PGGS	Iter.	68	–	–	–	–
	CPU	0.012	–	–	–	–
	β	1.6	–	–	–	–

EXAMPLE 5.2. Let in (1), $A = M + \nu I$, where M is defined as bellow and $\nu = 0, -.5$,

$$\mathbf{M} = \begin{pmatrix} B & -I & 0 & \dots & 0 & 0 \\ -I & B & -I & \dots & 0 & 0 \\ 0 & -I & B & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \ddots & \ddots & B & -I \\ 0 & 0 & \ddots & \ddots & -I & B \end{pmatrix} \in \mathbb{R}^{m^2\times m^2},$$

where $B = trdiag(-1, 4, -1) \in \mathbb{R}^{m\times m}$. It is considerably that in example 5.2 ENGS method for $\alpha \in (0, 1.4)$ with $\nu = 0$ and $\nu = -.5$, is convergent. Numerical results is presented in Table 2.

TABLE 2. Numerical result for Example 5.2.

ν	Method	$n = m^2$	$10,000$	$40,000$	$90,000$	$160,000$
0	NGS	Iter.	13	12	12	12
		CPU	0.035	0.158	0.433	0.949
	ENGS	Iter.	10	9	9	9
		CPU	0.026	0.114	0.308	0.623
		α	1.2	1.2	1.2	1.2
	GN	Iter.	5	5	5	5
		CPU	0.051	0.235	0.594	1.232
	GGS	Iter.	122	168	–	–
		CPU	0.244	1.768	–	–
	PGGS	Iter.	18	17	17	16
		CPU	0.053	0.269	0.693	1.254
		β	1.3	1.3	1.3	1.2
-.5	NGS	Iter.	16	16	16	16
		CPU	0.045	0.241	0.632	1.230
	ENGS	Iter.	12	12	12	12
		CPU	0.031	0.158	0.431	0.849
		α	1.2	1.2	1.2	1.2
	GN	Iter.	–	–	–	–
		CPU	–	–	–	–
	GGS	Iter.	77	–	–	–
		CPU	0.155	–	–	–
	PGGS	Iter.	20	20	20	20
		CPU	0.059	0.302	0.737	1.377
		β	1.1	1.1	1.1	1.1

6. Conclusion

In this paper we presented the generalized Newton Gauss-Seidel iteration method to solve AVE and proved it's convergence, then for improving convergence rate we proposed it's extrapolated version. At the end, multiple examples have shown that the proposed iteration methods are feasible, robust and effective for AVE.

References

1. J. Bender, K. Erleben and J. Trinkle, *Interactive simulation of rigid body dynamics in computer graphics*, Comput. Graph Forum. **33** (1) (2014) 246–270.
2. L. Caccetta, B. Qu and G. Zhou, *A globally and quadratically convergent method for absolute value equations*, Comput. Optim. Appl. **48** (2011) 45–58.
3. W. Cottle, *Linear Complementarity Problem*, In: C. Floudas and P. Pardalos (eds) Encyclopedia of Optimization. Springer, Boston, MA, 2008.
4. B. N. Datta, *Numerical linear algebra and applications*, 2nd ed., SIAM Publications, Philadelphia, PA, 2010.
5. V. Edalatpour, D. Hezari and D. K. Salkuyeh, *A generalization of the Gauss-Seidel iteration method for solving absolute value equations*, Appl. Math. Comput. **293** (2017) 156–167.
6. O. L. Mangasarian, *Knapsack feasibility as an absolute value equation solvable by successive linear programming*, Optim. Lett. **3** (2) (2009) 161–170.
7. O. L. Mangasarian and R. R. Meyer, *Absolute value equations*, Linear Algebra Appl. **419** (2-3) (2006) 359–367.

8. A. X. Wang, H. J. Wang and Y. K. Deng, *Interval algorithm for absolute value equations*, Cent. Eur. J. Math. **9** (2011) 1171–1184.
9. O. Prokopyev, *On equivalent reformulations for absolute value equations*, Comput. Optim. Appl. **44** (3) (2009) 363–372.
10. O. L. Mangasarian, *A generalized Newton method for absolute value equations*, Optim. Lett. **3** (2009) 101–108.
11. D. K. Salkuyeh, *The Picard-HSS iteration method for absolute value equations*, Optim. Lett. **8** (2014) 2191–2202.

E-mail: s.sifaleh@tabrizu.ac.ir
E-mail: ghodrat_ebadi@yahoo.com; gebadi@tabrizu.ac.ir

Refinement of Diagonal and Off-Diagonal Splitting Iteration Method for Solving the Linear Systems

Ghodrat Ebadi
Faculty of Mathematical Sciences, University of Tabriz, 51666-14766, Tabriz, Iran
and Raheleh Shokrpour*
Faculty of Mathematical Sciences, University of Tabriz, 51666-14766, Tabriz, Iran

Abstract. Recently, Dehghan et al. presented the diagonal and off-diagonal splitting (DOS) iteration method for solving the linear systems $\mathcal{A}x = b$. In this paper, we present a refinement for this method (RDOS) which increases its rate of convergence up to the rate of convergence of DOS method. Few numerical examples are considered to show the efficiency of the RDOS method.
Keywords: Refinement, Splitting method, H-matrix.
AMS Mathematical Subject Classification [2010]: 65F10, 65F30.

1. Introduction

In many scientific and engineering applications, one often comes across with a problem of finding the solution of a system of linear equations are written as the following equation in matrix form:

$$\mathcal{A}u = g, \tag{1}$$

where $\mathcal{A} \in \mathbb{C}^{n\times n}$ is a nonsingular matrix with non-vanishing diagonal entries and $x, b \in \mathbb{C}^n$. In iteration method

$$u^{(p+1)} = \mathcal{T}u^{(p)} + \mathcal{C}, \qquad p = 0, 1, 2, \ldots, \tag{2}$$

$\mathcal{T}$ is called the iterative matrix depending on matrix $\mathcal{A}$. The iteration system (2) is converge if and only if the spectral radius of $\mathcal{T}$ is less than unity. For solving (1), we usually split the coefficient matrix $\mathcal{A}$, as

$$\mathcal{A} = \mathcal{D} + \mathcal{E} + \mathcal{F},$$

where $\mathcal{D}$ is a diagonal matrix, $\mathcal{E}$ is a strictly lower triangular matrix and $\mathcal{F}$ is a general matrix. In [2], authors introduced a new splitting iteration method for solving (1) based on the diagonal and off-diagonal splitting (DOS) iterative method, as follows:

The DOS Iterative Method: Given an initial guess $u^{(0)} \in \mathbb{C}^{n\times n}$ for $p = 0, 1, 2, \ldots$ until $\{u^{(p)}\}$ converges, compute the next iterate $u^{(p+1)}$ according to the following procedure:

$$\begin{cases} \mathcal{D}u^{(p+\frac{1}{2})} = [\theta_1\mathcal{D} + (\theta_1 - 1)\mathcal{E} + (\theta_1 - 1)\mathcal{F}]u^{(p)} + (1 - \theta_1)g, \\ (\mathcal{D} + \theta_2\mathcal{E})u^{(p+1)} = [(1 - \theta_2)\mathcal{D} - \theta_2\mathcal{F}]u^{(p+\frac{1}{2})} + \theta_2 g, \end{cases} \tag{3}$$

*Speaker

where θ_1 and θ_2 are given constants. We can rewrite the DOS method as

$$u^{(p+1)} = \mathcal{T}_{\theta_1,\theta_2} u^{(p)} + C_{\theta_1,\theta_2},$$

where

$$\mathcal{T}_{\theta_1,\theta_2} = (\mathcal{D} + \theta_2\mathcal{E})^{-1}[(1-\theta_2)\mathcal{D} - \theta_2\mathcal{F}]\mathcal{D}^{-1}[\theta_1\mathcal{D} + (\theta_1 - 1)\mathcal{E} + (\theta_1 - 1)\mathcal{F}],$$
$$\mathcal{C}_{\theta_1,\theta_2} = (\mathcal{D} + \theta_2\mathcal{E})^{-1}[(1-\theta_1)[(1-\theta_2)\mathcal{D} - \omega_2\mathcal{F}]\mathcal{D}^{-1} + \theta_2 I]g.$$

Since the convergence rate of the stationary iterative process depends on the spectral radius of the iterative matrix, so any reasonable refinement of the iterative matrix will reduce the spectral radius and increases the convergence rate of the method. Dafchahi [1] proposed the refinement of Jacobi (RJ) method for the solution of linear system equations. Kumar Vatti et al. [5] presented a refinement of Gauss-Seidel method and studied a refinement of AOR method [6]. Salkuyeh [4] proposed a new iterative refinement of the solution of an ill-conditioned linear system of equations.

In this work, refinement of DOS method, called RDOS iterative method. Our method accelerates the convergence of the DOS iterative method. We study some theories about the convergence of RDOS iteration method. Finally, the RDOS is compared with the DOS method. By numerical experiment and theoretic analysis, we conclude that the proposed method is superior to some existence methods.

2. The RDOS Iterative Method

In this section, we consider a refinement of DOS iterative method and obtain the following RDOS iterative method as follows.

The RDOS Iterative Method: Given an initial guess $u^{(0)}$ for $p = 0, 1, 2, \ldots$ until $\{u^{(p)}\}$ converges, compute the next iterate $u^{(p+1)}$ according to the following procedure:

$$\begin{cases} \mathcal{D}u^{(p+\frac{1}{2})} = [\theta_1\mathcal{D} + (\theta_1 - 1)\mathcal{E} + (\theta_1 - 1)\mathcal{F}]u^{(p)} + (1-\theta_1)g, \\ (\mathcal{D} + \theta_2\mathcal{E})\tilde{u}^{(p+1)} = [(1-\theta_2)\mathcal{D} - \theta_2\mathcal{F}]u^{(p+\frac{1}{2})} + \theta_2 g, \\ u^{(p+1)} = \tilde{u}^{(p+1)} + (\mathcal{D} + \theta_2\mathcal{E})^{-1}(g - \mathcal{A}\tilde{u}^{(p+1)}). \end{cases} \tag{4}$$

At each step of the RDOS iteration method, we require solutions of two systems whose coefficient matrices are $\mathcal{D}$ and $\mathcal{D} + \theta_2\mathcal{E}$. The first linear subsystem is easy to implement since $\mathcal{D}$ is a diagonal matrix, and in the second system, it is a lower triangular matrix we can use the forward substitution methods. Where $\tilde{u}^{(p+1)}$ appeared in the right hand side is as given in (3). So far (4) takes the form $u^{(p+1)} = \mathcal{T}_{r,\theta_1,\theta_2} u^{(p)} + \mathcal{C}_{r,\theta_1,\theta_2}$, where

$$\mathcal{T}_{r,\theta_1,\theta_2} = (\mathcal{D} + \theta_2\mathcal{E})^{-1}[(1-\theta_2)\mathcal{D} - \theta_2\mathcal{F}]\mathcal{D}^{-1}[\theta_1\mathcal{D} + (\theta_1 - 1)\mathcal{E} + (\theta_1 - 1)\mathcal{F}],$$
$$\mathcal{C}_{r,\theta_1,\theta_2} = (I + \mathcal{T}_{\theta_1,\theta_2})(\mathcal{D} + \theta_2\mathcal{E})^{-1}[(1-\theta_1)[(1-\theta_2)\mathcal{D} - \omega_2\mathcal{F}]\mathcal{D}^{-1} + \theta_2 I].$$

On the other hand

$$u^{(p+1)} = \mathcal{T}^2_{\theta_1,\theta_2} u^{(p)} + (I + \mathcal{T}_{\theta_1,\theta_2})\mathcal{C}_{\theta_1,\theta_2}. \tag{5}$$

Here $\mathcal{T}_{r,\theta_1,\theta_2} = \mathcal{T}^2_{\theta_1,\theta_2}$ is the iterative matrix of RDOS method. We observe that the iterative matrix of RDOS is the square of the DOS iterative matrix. The resulting algorithm is summarized as follows:

ALGORITHM 1. **RDOS Iterative Algorithm**
Let $0 \leq \theta_1 \leq 1$ *and* $0 < \theta_2 \leq 1$ *are constant.*

1. *Choose an initial guess* $u^{(0)}$.
2. *For* $p = 0, 1, 2, \ldots$ *until convergence, Do*
3. *Solve* (3) *to compute* $\tilde{u}^{(p+1)}$.
4. *Compute* $u^{(p+1)} = \tilde{u}^{(p+1)} + (\mathcal{D} + \theta_2\mathcal{E})^{-1}(g - \mathcal{A}\tilde{u}^{(p+1)})$.
5. *End for.*

3. Convergence Analysis of the RDOS Method

In this section, we indicate that the RDOS iterative method converges to the unique solution of the system (1).

THEOREM 3.1. *If the DOS method converges, then the RDOS iterative method is convergent to the exact solution of the linear system* (1) *or equivalently,* $\rho(\mathcal{T}_{r,\theta_1,\theta_2}) < 1$.

PROOF. We have $\rho(\mathcal{T}_{\theta_1,\theta_2}) < 1$. Let u be the exact solution of (1). Then the DOS iterative method can be written as $u = (I - \mathcal{T}_{\theta_1,\theta_2})^{-1}\mathcal{C}_{\theta_1,\theta_2}$. Using Eq. (5) we have

$$
\begin{aligned}
u^{(p+1)} &= \mathcal{T}^2_{\theta_1,\theta_2}u^{(p)} + [I + \mathcal{T}_{\theta_1,\theta_2}]\mathcal{C}_{\theta_1,\theta_2},\\
u^{(p+1)} &= \mathcal{T}^4_{\theta_1,\theta_2}u^{(p-1)} + [I + \mathcal{T}_{\theta_1,\theta_2} + \mathcal{T}^2_{\theta_1,\theta_2} + \mathcal{T}^3_{\theta_1,\theta_2}]\mathcal{C}_{\theta_1,\theta_2},\\
u^{(p+1)} &= \mathcal{T}^6_{\theta_1,\theta_2}u^{(p-2)} + [I + \mathcal{T}_{\theta_1,\theta_2} + \mathcal{T}^2_{\theta_1,\theta_2} + \mathcal{T}^3_{\theta_1,\theta_2} + \mathcal{T}^4_{\theta_1,\theta_2} + \mathcal{T}^5_{\theta_1,\theta_2}]\mathcal{C}_{\theta_1,\theta_2},\\
&\vdots\\
u^{(p+1)} &= \mathcal{T}^{2(p+1)}_{\theta_1,\theta_2}u^{(0)} + [I + \mathcal{T}_{\theta_1,\theta_2} + \mathcal{T}^2_{\theta_1,\theta_2} + \mathcal{T}^3_{\theta_1,\theta_2} + \mathcal{T}^4_{\theta_1,\theta_2} + \mathcal{T}^5_{\theta_1,\theta_2} + \cdots + \mathcal{T}^{2p+1}_{\theta_1,\theta_2}]\mathcal{C}_{\theta_1,\theta_2},\\
u^{(p+1)} &= \mathcal{T}^{2(p+1)}_{\theta_1,\theta_2}u^{(0)} + \Sigma_{i=0}^{\infty}\mathcal{T}^i_{\theta_1,\theta_2}\mathcal{C}_{\theta_1,\theta_2},
\end{aligned}
$$

$$
\begin{aligned}
\lim_{p\to\infty} u^{(p+1)} &= \lim_{p\to\infty} \mathcal{T}^{2(p+1)}_{\theta_1,\theta_2}u^{(0)} + (I - \mathcal{T}_{\theta_1,\theta_2})^{-1}\mathcal{C}_{\theta_1,\theta_2},\\
\lim_{p\to\infty} u^{(p+1)} &= 0 + (I - \mathcal{T}_{\theta_1,\theta_2})^{-1}\mathcal{C}_{\theta_1,\theta_2},\\
\lim_{p\to\infty} u^{(p+1)} &= u.
\end{aligned}
$$

Therefore, RDOS method converges to the solution of linear system (1). □

THEOREM 3.2. *The RDOS method converges faster than the DOS method when DOS method is convergent.*

PROOF. Let $\tilde{u}$ be the solution of (1) obtained by (5) and $\overline{u}$ be the solution of (1) obtained by (3). From (5), we have

$$\tilde{u} = \mathcal{T}_{r,\theta_1,\theta_2}\overline{u} + \mathcal{C}_{r,\theta_1,\theta_2} \Rightarrow \tilde{u} = \mathcal{T}^2_{\theta_1,\theta_2}\overline{u} + \mathcal{C}_{r,\theta_1,\theta_2}.$$

So

$$
\begin{aligned}
u^{(p+1)} - \tilde{u} &= \mathcal{T}^2_{\theta_1,\theta_2} u^{(p)} + \mathcal{C}_{r,\theta_1,\theta_2} - \tilde{u}, \\
u^{(p+1)} - \tilde{u} &= \mathcal{T}^2_{\theta_1,\theta_2} (u^{(p)} - \overline{u}) + \mathcal{C}_{r,\theta_1,\theta_2} - \tilde{u} + \mathcal{T}^2_{\theta_1,\theta_2} \overline{u}, \\
u^{(p+1)} - \tilde{u} &= \mathcal{T}^2_{\theta_1,\theta_2} (u^{(p)} - \overline{u}) - \tilde{u} + (\mathcal{T}^2_{\alpha} \overline{u} + \mathcal{C}_{r,\theta_1,\theta_2}), \\
u^{(p+1)} - \tilde{u} &= \mathcal{T}^2_{\theta_1,\theta_2} (u^{(p)} - \overline{u}) - \tilde{u} + \tilde{u}, \\
u^{(p+1)} - \tilde{u} &= \mathcal{T}^2_{\theta_1,\theta_2} (u^{(p)} - \overline{u}).
\end{aligned}
$$

Now,

$$
\begin{aligned}
\|u^{(p+1)} - \tilde{u}\|_\infty &= \|\mathcal{T}^2_{\theta_1,\theta_2} (u^{(p)} - \overline{u})\|_\infty \leq \|\mathcal{T}^2_{\theta_1,\theta_2}\|_\infty \|u^{(p)} - \overline{u}\|_\infty, \\
\|u^{(p+1)} - \tilde{u}\|_\infty &\leq \|\mathcal{T}_{\theta_1,\theta_2}\|^2_\infty \|u^{(p)} - \overline{u}\|_\infty, \\
\|u^{(p+1)} - \tilde{u}\|_\infty &\leq \|\mathcal{T}_{\theta_1,\theta_2}\|^{2n}_\infty \|u^{(1)} - \overline{u}\|_\infty.
\end{aligned}
$$

According to Theorem 3.1, $\|\mathcal{T}_{\theta_1,\theta_2}\|^2_\infty < 1$. Hence, $\|u^{(p+1)} - \tilde{u}\|_\infty \leq \|u^{(1)} - \overline{u}\|_\infty$ this is equivalent to the refinement of RDOS method converge faster than the DOS method. □

Using the singular value decomposition we can convert a nonsingular matrix $\mathcal{A}$ to a strictly diagonally dominant matrix. We can find nonsingular matrices P and Q using the $\mathcal{SVD}$ decomposition such that $P\mathcal{A}Q$ is strictly diagonally dominant [7, 8]. Also, Yuan showed that there exists a nonsingular matrix P such that $P\mathcal{A}$ is strictly diagonally dominant.

As said in [2], the DOS iteration method converges unconditionally when $\mathcal{A}$ is strictly diagonally dominant, for $0 \leq \theta_1 \leq 1$ and $0 < \theta_2 \leq 1$. It is obvious that after finding P and Q such that $P\mathcal{A}Q$ is strictly diagonally dominant, instead of solving (1) we can solve $P\mathcal{A}Qv = Pg,\ u = Qv$.

Special Cases: When $\mathcal{F}$ is strictly upper triangular matrix, we observe that for specific values of the parameters θ_1, θ_2 the RDOS iterative reduces to refinement well-known methods, for instance:

- If $\theta_1 = 0, \theta_2 = 0$, then $\mathcal{T}_{r,0,0}$ is the iteration matrix of the refinement Jacobi ($R\mathcal{J}$) method [3].
- If $\theta_1 = 1, \theta_2 = 1$, then $\mathcal{T}_{r,1,1}$ is the iteration matrix of the refinement Gauss-Seidel ($R\mathcal{GS}$) method [5]. $\mathcal{T}_{r,1-\theta_1,0}$ is the iteration matrix of the refinement Simultaneous Over-relaxation method,
- If $\theta_1 = 1, \theta_2 = free$, then $\mathcal{T}_{r,1,free}$ is the iteration matrix of the refinement Successive Over-relaxation ($R\mathcal{SOR}$) method [6].

RJ method is as fast as $\mathcal{SOR}$ method but, in compare with SOR method is easier because we don't require finding optimal parameter ω.

4. Numerical Experiments

In this section, numerical example is considered to exhibit the effectiveness of our method. We also compare the performance of the RDOS method with the DOS method from the point of view of the iteration counts (denoted as "IT"), CPU times (denoted as "CPU") and the spectral radius (denoted as "ρ"). The numerical experiment was computed in double precision in MATLAB R2016b on a PC computer with Intel(R) Core (TM) i7-7700k CPU 4.20GHz, 8.00 GB memory with

machine precision and Windows 10 operating system. In our implementations, the initial guess $u^{(0)}$ is chosen zero vector. In all examples, the stopping criterion is $\frac{\|g-\mathcal{A}u^{(p)}\|_2}{\|g\|_2} < 10^{-5}$.

In our tests, we take $h = \frac{1}{m+1}$, $n = m^2$, $\theta_1 = 0.25$, $\theta_2 = 1$ and for the tests reported in this section, $\mathcal{F}$ is strictly upper triangular matrix.

EXAMPLE 4.1. [2] Consider the linear system $(\pi\mathcal{K}_V + \mathcal{K}_H)u = g$, where $\mathcal{K}_V$ and $\mathcal{K}_H$ are the viscous and hysteretic damping matrices, respectively. Here, $\mathcal{K}_V = 10I_n$, $\mathcal{K}_H = 0.02\mathcal{W}$, $\mathcal{W} = I_m \otimes \mathcal{V}_m + \mathcal{V}_m \otimes I_m$, $\mathcal{V}_m = h^{-2}tridiag(-1, 2, -1) \in \mathbb{R}^{m\times m}$. We take $g = (-\pi^2 I_n + \mathcal{W} + \pi\mathcal{K}_V + \mathcal{K}_H)B$, where $B = (1, 1, \ldots, 1)^T$.

TABLE 1. Numerical results of Example 4.1.

	DOS			RDOS		
m	IT	CPU	$\rho(T_{\theta_1,\theta_2})$	IT	CPU	$\rho(T_{r,\theta_1,\theta_2})$
10	4	0.0052	0.03080	3	0.0048	0.0009479
20	7	0.0081	0.1935	4	0.0032	0.0374
30	12	0.0102	0.4010	7	0.0074	0.1608
40	17	0.0124	0.5661	10	0.0095	0.3204
50	24	0.0209	0.6808	14	0.0164	0.4635

Numerical result shows that the refinement of Jacobi method as fast as SOR method with optimal parameter ω, but in the refinement of Jacobi method there is no exist problem of finding ω. Also, it shows that the iteration numbers and CPU times with $RG\mathcal{S}$ and RDOS ($\theta_1 = 1, \theta_1 = 1$) are the same.

5. Conclusion

For solving the non-singular linear system, we present a refinement of DOS method and demonstrate that RDOS method converges to the unique solution of (1). We have compared the numerical result of the RDOS iterative method with the DOS iteration method. Numerical result shows that the RDOS method is superior to the DOS method in terms of the iteration counts, the CPU times and spectral radius.

References

1. F. N. Dafchahi, *A new refinement of Jacobi method for solution of linear system of equations* $AX = B$, Int. J. Contemp. Math. Sci. **3** (2008) 819–827.
2. M. Dehghan, M. D. Madiseh and M. Hajarian, *A two-step iterative method based on diagonal and off-diagonal splitting for solving linear systems*, Filomat **31** (2017) 1441–1452.
3. A. H. Laskar and S. Behera, *A refinement of iterative methods for the solution of system of linear equations* $Ax = b$, IOSR-JM **10** (3) (2014) 70–73.
4. D. K. Salkuyeh and A. Fahim, *A new iterative refinement of the solution of ill-conditioned linear system of equations*, Int. J. Comput. Math. **88** (5) (2011) 950–956.
5. V. B. K. Vatti and T. K. Eneyew, *A refinement of Gauss-Seidel method for solving linear system of equations*, Int. J. Contemp. Math. Sci. **6** (2011) 117–121.
6. V. B. K. Vatti, R. Sri and M. S. K Mylapalli, *A refinement of accelerated over relaxation method for the solution of linear systems*, IJPAM. **118** (2018) 1571–1577.

7. J. Y. Yuan, *Preconditioned diagonal dominant matrices*, Appl. Math. Comput. **114** (2–3) (2000) 255–262.
8. J. Y. Yuan and P. Y. Yalamov, *A method for constructing diagonally dominant preconditioners based on jacobi rotations*, Appl. Math. Comput. **174** (2006) 74–80.

E-mail: ghodrat_ebadi@yahoo.com
E-mail: shokrpour_raheleh@yahoo.com

A Numerical Method for Pricing Discrete Barrier Option by CAS Wavelet

Amirhossein Sobhani*
Department of Mathematics, Statistics and Computer Science, Semnan University, Semnan, Iran

Abstract. In this article, a numerical method for pricing knock-out discrete double barrier options based on CAS wavelets basis functions is proposed. According to the well-known Black-Scholes partial differential equations, the price of option could be obtained by a recursive formulas. These solutions has been approximated by CAS wavelets basis functions and expressed in operational matrix form.
Keywords: Barrier options, CAS wavelets, Option pricing.
AMS Mathematical Subject Classification [2010]: 65D15, 35E15, 46A32.

1. Introduction

A barrier option is a financial derivative that plays important role in managing risk in financial markets. A call (put) option is a financial contract that gives the holder right to buy (sell) the underlying asset at a specific price, that is called exercise price. Depending on whether we have one or two barriers, we have two types of barrier options: single and double. If the barrier option is activated (deactivated) when the stock price hits the barrier, It is called knock-in (knock-out). If the touching of barrier with stock price is checked only on fixed times, for example weakly or monthly, the barrier option is called a discrete barrier option. Various numerical and analytical methods have been proposed in recent decades for pricing barrier options. Fusai et al. obtained an analytical solution for single barrier option with the aid of z-transform [5]. In [2, 3], Fourier-cosine expansion method is used for pricing barrier options. Milev and Tagliani presented a numerical method based on quadrature method for pricing double barrier option in [6]. A numerical method for pricing barrier options based on projection methods has been presented in [4]. In [7], a numerical method with the aid of Legendre multiwavelet has been proposed. Let r, σ and S_0 are the risk-free rate, the volatility, and the initial stock price respectively. Also, assume that the stock price S_t follows geometric Brownian motion

$$dS_t = rS_t + \sigma S_t dB_t.$$

We concern in pricing knock-out discrete double barrier call option on stock, i.e. a call option that becomes worthless if the stock price hits lower barrier L or upper barrier U at the specific monitoring dates $0 = t_0 < t_1 < \cdots < t_M = T$. If the barriers are not touched in monitoring dates by underlying asset price, the pay off at maturity time T is $\max(S_T - E, 0)$, where E is exercise price. According to

*Speaker

the well-known Black-Scholes framework, the price of discretely monitored double barrier call option as a function of stock price at time $t \in (t_{m-1}, t_m)$, namely $\mathcal{P}(S, t, m-1)$, is obtained from forward solving the following partial differential Eqs. [1]:

$$-\frac{\partial \mathcal{P}}{\partial t} + rS\frac{\partial \mathcal{P}}{\partial S} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 \mathcal{P}}{\partial S^2} - r\mathcal{P} = 0, \tag{1}$$

with the following initial conditions:

$$\begin{aligned} \mathcal{P}(S, t_0, 0) &= (S - E)\,\mathbf{1}_{(\max(E,L)\le S\le U)}, \\ \mathcal{P}(S, t_m, 0) &= \mathcal{P}(S, t_m, m-1)\,\mathbf{1}_{(L\le S\le U)};\ m = 1, 2, \ldots, M-1, \end{aligned}$$

where $\mathcal{P}(S, t_m, m-1) := \lim_{t\to t_m} \mathcal{P}(S, t, m-1)$. By implementing change of variable $z = \ln\left(\frac{S}{L}\right)$ and denoting $C(z, t, m) := \mathcal{P}(S, t, m)$ PDE (1) is converted to:

$$\begin{aligned} -C_t + \mu C_z &+ \frac{\sigma^2}{2}C_{zz} = rC, \\ C(z, t_0, 0) &= L(e^z - e^{E^*})\mathbf{1}_{(\delta\le z\le\theta)}, \\ C(z, t_m, m) &= C(z, t_m, m-1)\,\mathbf{1}_{(0\le z\le\theta)};\ m = 1, 2, \ldots, M-1, \end{aligned} \tag{2}$$

where $E^* = \ln\left(\frac{E}{L}\right)$; $\mu = r - \frac{\sigma^2}{2}$; $\theta = \ln\left(\frac{U}{L}\right)$ and $\delta = \max\{E^*, 0\}$. Finally, second transformation $C(z, t_m, m) = e^{\alpha z + \beta t} h(z, t, m)$, where

$$\alpha = -\frac{\mu}{\sigma^2};\ c^2 = -\frac{\sigma^2}{2};\ \beta = \alpha\mu + \alpha^2\frac{\sigma^2}{2} - r,$$

is applied and PDE (2) is reduced to the following heat equation:

$$\begin{aligned} &-h_t + c^2 h_{zz} = 0, \\ &h(z, t_0, 0) = Le^{-\alpha z}\left(e^z - e^{E^*}\right)\mathbf{1}_{(\delta\le z\le\theta)}, \qquad m = 0, \\ &h(z, t_m, m) = h(z, t_m, m-1)\,\mathbf{1}_{(0\le\theta\le z)}, \qquad m = 1, \ldots, M-1. \end{aligned}$$

The above equation has analytical solution as below:

$$h(z, t, m) = \begin{cases} L\int_\delta^\theta k(z-\xi, t)\, e^{-\alpha\xi}\left(e^\xi - e^{E^*}\right) d\xi, & m = 0, \\ \int_0^\theta k(z-\xi, t-t_m)\, h(\xi, t_m, m-1)\, d\xi, & m = 1, 2, \ldots, M-1, \end{cases}$$

where

$$k(z, t) = \frac{1}{\sqrt{4\pi c^2 t}} e^{-\frac{z^2}{4c^2 t}}. \tag{3}$$

We consider monitoring dates of equally spaced, i.e, $t_m = m\tau$, where $\tau = \frac{T}{M}$. In this way $h(z, t_m, m-1)$ is a function of two variables z, m. Therefore from $f_m(z) := h(z, t_m, m-1)$, the following relations will be obtained:

$$f_m(z) = \int_0^\theta k(z-\xi, \tau) f_{m-1}(\xi)\, d\xi, \quad m = 2, 3, \ldots, M, \tag{4}$$

where $f_0(\xi) = Le^{-\alpha z}\left(e^z - e^{E^*}\right)\mathbf{1}_{(\delta\le\xi\le\theta)}$.

2. CAS Wavelet

Let $m \in \mathbb{Z}$ and $CAS_m(t) = \cos(2m\pi t) + \sin(2m\pi t)$, then CAS wavelets $\psi_{n,m}(t)$ for any non-negative integer k are defined on $[0,1)$ as follows:

$$\psi_{n,m}(t) = \begin{cases} 2^{\frac{k}{2}} CAS_m\left(2^k t - n\right), & \frac{n}{2^k} \le t < \frac{n+1}{2^k}, \\ 0, & otherwise, \end{cases}$$

where $n = 0, 1, 2, \ldots, 2^k - 1$. These functions construct an orthonormal basis functions for $L^2([0,1])$ [8], so we can expand any function $f(t) \in L^2([0,1))$ as

$$f(t) = \sum_{n=1}^{\infty} \sum_{m \in \mathbb{Z}} a_{m,n} \psi_{n,m}(t),$$

where $a_{m,n} = \langle f, \psi_{n,m} \rangle = \int_0^1 f(t)\psi_{n,m}(t)dt$. If we define $\tilde{\psi}_{n,m}(t) = \sqrt{\theta}^{-1} \psi_{n,m}(t/\theta)$ then $\{\tilde{\psi}_{n,m}(t); n = 1, \ldots, \infty, m \in \mathbb{Z}\}$ is an orthonormal basis for $L^2([0,\theta])$. Now, let

$$X_J = span\left\{\tilde{\psi}_{n,m}(t); n = 1, \ldots, 2^J, m = -M, \ldots, M\right\},$$

then the orthogonal projection operator $P_J : L^2([0,\theta]) \to X_J$ is defined as follow:

$$\forall f \in L^2([0,\theta])\ P_J(f) = \sum_{n=1}^{2^J} \sum_{m=-M}^{M} a_{m,n} \tilde{\psi}_{n,m}(t) = \vec{a}' \Psi_J,$$

where

$$\begin{aligned} \vec{a} = [&a_{-M,1},\ a_{-M,2}, \ldots, a_{-M,2^k}, a_{-M+1,1}, \\ &a_{-M,2}, \ldots, a_{-M+1,2^k}, \ldots, a_{M,1},\ a_{M,2}, \ldots, a_{M,2^J}], \\ \Psi = [&\tilde{\psi}_{-M,1},\ \tilde{\psi}_{-M,2}, \ldots, \tilde{\psi}_{-M,2^k}, \tilde{\psi}_{-M+1,1}, \\ &\tilde{\psi}_{-M,2}, \ldots, \tilde{\psi}_{-M+1,2^k}, \ldots, \tilde{\psi}_{M,1},\ \tilde{\psi}_{M,2}, \ldots, \tilde{\psi}_{M,2^J}]. \end{aligned}$$

3. Pricing by CAS Wavelet

Consider the compact operator $\mathcal{K} : L^2([0,\theta]) \to L^2([0,\theta])$ as follows:

$$\mathcal{K}(f)(z) := \int_0^{\theta} \kappa(z - \xi, \tau) f(\xi) d\xi,$$

where κ is defined in (3). With attention to the definition of operator $\mathcal{K}$, Eq. (4) can be rewritten as below:

$$f_m = \mathcal{K} f_{m-1} \qquad m = 1, 2, 3, \ldots, M. \tag{5}$$

Now, we define $\tilde{f}_{m,J} = P_J \mathcal{K}\left(\tilde{f}_{m-1,J}\right) = (P_J\mathcal{K})^m (f_0)$, $m \ge 2$, where $(P_J\mathcal{K})(f) = P_J(\mathcal{K}(f))$. Since the continuous projection operators P_J converge pointwise to identity operator I, then operator $P_J\mathcal{K}$ is also a compact operator and it could be shown that [4]

$$\lim_{n \to \infty} \|(P_J\mathcal{K})^m - \mathcal{K}^m\| = 0.$$

Since, $\tilde{f}_{m,J} \in V_J$ for $m \geq 1$, we can write

$$\tilde{f}_{m,J} = \sum_{i=0}^{r2^J} a_{mi}\overline{\psi}_i(z) = \Psi'_J(x)F_m,$$

where $F_m = [a_{m0}, a_{m1}, \ldots, a_{m2^J}]'$. From Eq. (5) we obtain

$$\tilde{f}_{m,J} = (P_J\mathcal{K})^{m-1}\Big(\tilde{f}_{1,J}\Big). \tag{6}$$

Because X_J is a finite dimensional linear space, so the linear operator $P_J\mathcal{K}$ on X_J could be considered as a $(2M+1)2^J \times (2M+1)2^J$ matrix K. Consequently Eq. (6) can be written as following matrix operator form

$$\tilde{f}_{m,J} = \Psi'_J K^{m-1} F_1. \tag{7}$$

For computation of the option price by (7), it is enough to calculate the matrix operator K and the vector F_1. By denoting $\overline{\psi}_i(x)$ as the i-th element of Ψ we have (See [4])

$$F_1 = [a_{11}, a_{12}, \ldots, a_{1r2^J}]', \;\; K = (k_{ij})_{(2M+1)2^J \times (2M+1)2^J},$$

where

$$a_{1i} = \int_0^\theta \int_\delta^\theta \overline{\psi}_i(\eta)\kappa(\eta-\xi,\tau)f_0(\xi)d\xi d\eta \;,\; 0 \leq i \leq (2M+1)2^J,$$
$$k_{ij} = \int_0^\theta \int_0^\theta \overline{\psi}_i(\eta)\overline{\psi}_j(\xi)\kappa(\eta-\xi,\tau)d\xi d\eta \;.$$

Therefore, the price of the knock-out discrete double barrier option can be estimated as follows:

$$\mathcal{P}\left(S_0, t_M, M-1\right) \simeq e^{\alpha z_0 + \beta t}\tilde{f}_{M,J}\left(z_0\right), \tag{8}$$

where $z_0 = \log\left(\frac{S_0}{L}\right)$ and $\tilde{f}_{M,n}$ from (7). The matrix form of relation (7) implies that the computational time of presented algorithm be nearly fixed when monitoring dates increase. Actually, if we set $N = (2M+1)2^J$ the complexity of our algorithm is $\mathcal{O}(N^2)$ that dose not depend on number of monitoring dates.

4. Numerical Result

In this section we consider a knock-out discrete double barrier option with maturity time $T = 0.5$, risk-free rate $r = 0.05$, volatility $\sigma = 0.25$, exercise price $E = 100$, different spot price S_0, upper barrier $U = 110$ and different lower barrier L. we price this option by using the presented method and the numerical results are given and compared with projection method [4] with 16 Legendre basis functions as benchmark. Table 1 shows efficiency and accuracy of presented method in comparison with the benchmark. Furthermore, we can see that CPU time of our method increases insignificantly against increases of monitoring dates. The numerical results are obtained from relation (8) with $(2M+1)2^J$ CAS wavelets basis functions. Source code of this method was written in MATLAB 2015 on a 3.2 GHz Intel Core i5 PC with 8 GB RAM.

TABLE 1. Double barrier option pricing with parameters: $T = 0.5$, $r = 0.05$, $\sigma = 0.25$, $E = 100$, $U = 110$ and $L = 95$.

M	s_0	CAS wavelets $(M = 1, J = 6)$	$\|Error\|$	CAS wavelets $(M = 1, J = 7)$	$\|Error\|$	Benchmark
5	95	0.176305	1.800e-03	0.175404	9.0600e-04	0.174498
	100	0.232373	1.3500e-04	0.232526	1.8000e-05	0.232508
	105	0.225881	1.8000e-04	0.226143	8.2000e-05	0.226061
	107	0.20748	7.5700e-04	0.206866	1.4300e-04	0.206723
	110	0.16912	1.700e-03	0.168259	8.6600e-04	0.167393
CPU		2 s		3.5 s		5.5 s
25	95	0.020119	5.9100e-04	0.019824	2.9600e-04	0.019528
	100	0.04288	-7.7000e-05	0.04296	3.0000e-06	0.042957
	105	0.040819	-9.7000e-05	0.040947	3.1000e-05	0.040916
	107	0.033305	2.9900e-04	0.033061	5.5000e-05	0.033006
	110	0.01925	5.6200e-04	0.018971	2.8300e-04	0.018688
CPU		2 s		3.5 s		5.5 s

5. Conclusion and Remarks

In this paper, we have implemented orthogonal projection method with the aid of CAS wavelets basis functions for pricing discrete double barrier options and obtained a matrix relation (7) for approximation solution of this problem. The numerical results shows the efficiency and validity of this method.

References

1. T. Björk, *Arbitrage Theory in Continuous Time*, Oxford University Press, Hawthorne, CA, USA, 2009.
2. F. Fang and C. W. Oosterlee, *A novel pricing method for european options based on fourier-cosine series expansions*, SIAM J. Sci. Comput. **31** (2) (2009) 826–848.
3. F. Fang and C. W. Oosterlee, *Pricing early-exercise and discrete barrier options by fourier-cosine series expansions*, Numer. Math. **114** (1) (2009) 27.
4. R. Farnoosh, A. Sobhani and M. H. Beheshti, *Efficient and fast numerical method for pricing discrete double barrier option by projection method*, Comput. Math. Appl. **73** (7) (2017) 1539–1545.
5. G. Fusai, I. David Abrahams and C. Sgarra, *An exact analytical solution for discrete barrier options*, Finance and Stochastics **10** (2006) 1–26.
6. M. Milev and A. Tagliani, *Numerical valuation of discrete double barrier options*, J. Comput. Appl. Math. **233** (10) (2010) 2468–2480.
7. A. Sobhani and M. Milev, *A numerical method for pricing discrete double barrier option by legendre multiwavelet*, J. Comput. Appl. Math. **328** (2018) 355–364.
8. S. Yousefi and A. Banifatemi, *Numerical solution of fredholm integral equations by using cas wavelets*, Appl. Math. Comput. **183** (1) (2006) 458–463.

E-mail: a_sobhani@semnan.ac.ir

Application of B-Spline Method for Solving Inverse Kawahara Equation

Fateme Torabi*
School of Mathematics and Computer Science, Damghan University, P. O. Box 36715-364, Damghan, Iran
Reza Pourgholi
School of Mathematics and Computer Science, Damghan University, P. O. Box 36715-364, Damghan, Iran
and Amin Esfahani
School of Mathematics and Computer Science, Damghan University, P. O. Box 36715-364, Damghan, Iran

Abstract. In this paper, a numerical method is proposed to approximate the solution of the nonlinear inverse Kawahara equation. We apply B-spline for spatial variable and derivatives which produce a system. We solve this system by using the Tikhonov regularization method. The aim of this paper is to show that the method based on B-spline is also suitable for the treatment of the nonlinear inverse parabolic partial differential equations. Numerical example also verified the efficiency and accuracy of the method that can be obtained in the MATLAB 7.10 (R2017b) and is tested on a personal computer with intel(R) core(TM)2 Duo CPU and 4GB RAM.
Keywords: B-spline method, Inverse problems, Noisy data.
AMS Mathematical Subject Classification [2010]: 65M32, 35K05.

1. Introduction

A wide range of inverse problems for the heat equation are today extensively studied, because of their clear importance in applied sciences. Inverse problems of parabolic type arise from various fields of engineering. These kind of problems have been investigated by many researchers [2, 5].

Inverse problems are often ill posed, with solutions that depend sensitively on data. In any numerical approach to the solution of such problems, regularization of some form is needed to counteract the resulting instability [6]. In this paper, we consider the nonlinear inverse Kawahara equation with forcing term

$$u_t + au^2u_x + bu_{xxx} - du_{xxxxx} = 0,\ (x,t) \in [0,1] \times [0,t_f], \tag{1}$$

where a, b and d are real constants and the following initial condition

$$u(x,0) = p(x),\ x \in [0,1], \tag{2}$$

*Speaker

with boundary conditions

$$u(0,t) = f_1(t), \qquad t \in [0, t_f], \tag{3}$$

$$u_x(0,t) = f_2(t), \qquad t \in [0, t_f], \tag{4}$$

$$u(1,t) = g_1(t), \qquad t \in [0, t_f], \tag{5}$$

$$u_x(1,t) = g_2(t), \qquad t \in [0, t_f], \tag{6}$$

$$u_{xx}(1,t) = g_3(t), \qquad t \in [0, t_f], \tag{7}$$

where $p(x)$, $g_1(t)$, $g_2(t)$, and $g_3(t)$ are continuous known functions and t_f represents the final time, while the functions $f_1(t)$, $f_2(t)$ and $u(x,t)$ are unknown which remain to be determined.

2. B-Spline Method

In this section we solve the nonlinear inverse problem (1)-(7) with the over-specified conditions

$$u(a^*,t) = h_1(t), \quad t \in [0, t_f],$$
$$u_x(a^*,t) = h_2(t), \quad t \in [0, t_f],$$

where $0 < a^* < 1$ is a fixed point. The solution domain $x \in [0,1]$ is partitioned into a mesh of uniform length $h = x_{i+1} - x_i$ by the knots x_i, where $i = 0, 1, \ldots, N-1$ such that $\Delta = 0 = x_0 < x_1 < \cdots < x_N = 1$ be the partition in $[0,1]$. B-splines are the unique nonzero splines of smallest compact support with knots at $x_0 < x_1 < \cdots < x_N$. We define the B-spline $B_i(x)$ for $i = -3, 0, \ldots, N+2$ by the following relation [4]

$$B_i(x) = \frac{1}{h^6} \begin{cases} (x - x_i + 3h)^6, & x \in [x_{i-3}, x_{i-2}), \\ (x - x_i + 3h)^6 - 7(x - x_i + 2h)^6, & x \in [x_{i-2}, x_{i-1}), \\ (x - x_i + 3h)^6 - 7(x - x_i + 2h)^6 & \\ +21(x - x_i + h)^6, & x \in [x_{i-1}, x_i), \\ (x - x_i + 3h)^6 - 7(x - x_i + 2h)^6 & \\ +21(x - x_i + h)^6 - 35(x - x_i)^6, & x \in [x_i, x_{i+1}), \\ (x - x_i - 4h)^6 - 7(x - x_i - 3h)^6 & \\ +21(x - x_i - 2h)^6, & x \in [x_{i+1}, x_{i+2}), \\ (x - x_i - 4h)^6 - 7(x - x_i - 3h)^6, & x \in [x_{i+2}, x_{i+3}), \\ (x - x_i - 4h)^6, & x \in [x_{i+3}, x_{i+4}), \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

It can be easily seen that the set of functions $\Gamma = \{B_{-3}(x), B_{-2}(x), B_{-1}(x), \ldots, B_{N+2}(x)\}$ are linearly independent on $[0,1]$, thus $\Theta = Span(\Gamma)$ is a subspace of $C^2[0,1]$ and Θ is $N+6$-dimensional. Let us consider $U_m(x,t) \in \Theta$ be the B-spline approximation to the exact solution $u(x,t)$ in the form

$$U_m(x,t) = \sum_{i=-3}^{m+2} c_i(t) B_i(x), \tag{9}$$

where $c_i(t)$ are time-dependent quantities to be determined from the boundary and over-specified conditions and collocation from of the differential equations.

Using approximate function (9) and B-spline (8), the approximate values at the knots of $U(x)$ and its derivatives up to fifth order are determined in terms of the time parameters c_m as

$$\begin{aligned}
U_m &= c_{m-3} + 57c_{m-2} + 302c_{m-1} + 302c_m + 57c_{m+1} + c_{m+2},\\
U'_m &= (6/h)(-c_{m-3} - 25c_{m-2} - 40c_{m-1} + 40c_m + 25c_{m+1} + c_{m+2}),\\
U''_m &= (30/h^2)(c_{m-3} + 9c_{m-2} - 10c_{m-1} - 10c_m + 9c_{m+1} + c_{m+2}),\\
U'''_m &= (120/h^3)(-c_{m-3} - c_{m-2} + 8c_{m-1} - 8c_m + c_{m+1} + c_{m+2}),\\
U^{(4)}_m &= (360/h^4)(-c_{m-3} + 3c_{m-2} - 2c_{m-1} + 2c_m - 3c_{m+1} + c_{m+2}),\\
U^{(5)}_m &= (720/h^5)(c_{m-3} - 5c_{m-2} + 10c_{m-1} - 10c_m + 5c_{m+1} - c_{m+2}).
\end{aligned}$$

Therefore, we have a system which solve by using Tikhonov regularization method, the coefficients c_i are obtained and using these coefficients, we can obtain the approximate solution.

3. Main Results

The purpose of this section is to illustrate the applicability of the present method described in Section 2 for solving the nonlinear inverse problem (1)-(7). We compare the exact and the approximate solutions by considering the total error S defined by [1]

$$\begin{aligned}
S_{f_1} &= \left[\frac{1}{N-1}\sum_{z=1}^{N}\Big(f_1(t_z) - f_1^*(t_z)\Big)^2\right]^{\frac{1}{2}},\\
S_{f_2} &= \left[\frac{1}{N-1}\sum_{z=1}^{N}\Big(f_2(t_z) - f_2^*(t_z)\Big)^2\right]^{\frac{1}{2}},
\end{aligned}$$

where N is the number of estimated values, f_1 and f_2 are the exact values, f_1^* and f_2^* are the estimated values.

Example 3.1. In this example we solve the nonlinear inverse Kawahara problem (1) and the comparisons are made with the exact solutions given in [3], $a = 1$, $b = -0.001$ and $d = 1$ in the Tables refex1tab1 and refex1tab2 with the noisy data (input data+0.1×rand(1)).

$$u(x,0) = \frac{3b}{\sqrt{10da}}\sec^2(\mu x),\ \mu = \frac{1}{2}\sqrt{\frac{-b}{5d}}, \frac{b}{d} < 0, \qquad 0 \le x \le 1,$$

$$u(1,t) = g_1(t) = \frac{3b}{\sqrt{10da}}\sec^2(\mu(1 - \frac{4b^2}{25d}t)), \qquad 0 \le t \le 1,$$

$$u_x(1,t) = g_2(t) = 2\mu\frac{3b}{\sqrt{10da}}\frac{\sin(\mu(1 - \frac{4b^2}{25d}t))}{\cos^3(\mu(1 - \frac{4b^2}{25d}t))}, \qquad 0 \le t \le 1,$$

$$u_{xx}(1,t) = g_3(t) = 2\mu^2\frac{3b}{\sqrt{10da}}\Big(\frac{2}{\cos^2(\mu(1 - \frac{4b^2}{25d}t))} + \frac{6\sin^2(\mu(1 - \frac{4b^2}{25d}t))}{\cos^4(\mu(1 - \frac{4b^2}{25d}t))}\Big), \quad 0 \le t \le 1.$$

The exact solutions of this problem are,

$$u(x,t) = \frac{3b}{\sqrt{10da}}\sec^2(\mu(x - \frac{4b^2}{25d}t)), \qquad 0 \le x \le 1,\ 0 \le t \le 1,$$

$$f_1(t) = u(0,t) = \frac{3b}{\sqrt{10da}}\sec^2(\mu(\frac{4b^2}{25d}t)), \qquad 0 \le t \le 1,$$

$$f_2(t) = u_x(0,t) = -2\mu\frac{3b}{\sqrt{10da}}\frac{\sin(\mu(\frac{4b^2}{25d}t))}{(\cos(\mu(\frac{4b^2}{25d}t)))^3}, \qquad 0 \le t \le 1.$$

TABLE 1. The values of $f_1(t)$ and $f_2(t)$ when $h = 0.01$, $k = 0.001$, $a^* = 0.2$.

t	$f_1(t)$		$f_2(t)$	
	Exact	$B-spline$	*Exact*	$B-spline$
0.1	$-9.486832980505137e-04$	$-2.166096298056512e-19$	$1.517893276880822e-15$	$-3.317595801594114e-17$
0.2	$-9.486832980505137e-04$	$-2.166096298231290e-19$	$3.035786553761644e-15$	$-3.317595801866043e-17$
0.3	$-9.486832980505137e-04$	$-2.166096298406067e-19$	$4.553679830642466e-15$	$-3.317595802137971e-17$
0.4	$-9.486832980505137e-04$	$-2.166096298580845e-19$	$6.071573107523287e-15$	$-3.317595802409900e-17$
0.5	$-9.486832980505137e-04$	$-2.166096298755623e-19$	$7.589466384404110e-15$	$-3.317595802681830e-17$
0.6	$-9.486832980505137e-04$	$-2.166096298930401e-19$	$9.107359661284932e-15$	$-3.317595802953759e-17$
0.7	$-9.486832980505137e-04$	$-2.166096299105178e-19$	$1.062525293816575e-14$	$-3.317595803225687e-17$
0.8	$-9.486832980505137e-04$	$-2.166096299279956e-19$	$1.214314621504657e-14$	$-3.317595803497616e-17$
0.9	$-9.486832980505137e-04$	$-2.166096299454734e-19$	$1.366103949192740e-14$	$-3.317595803769546e-17$
1	$-9.486832980505137e-04$	$-2.166096299629512e-19$	$1.517893276880822e-14$	$-3.317595804041474e-17$
S	-	$9.4923e-04$	-	$3.9511e-05$

TABLE 2. The values of $u(0.8,t)$ when $h = 0.01$, $k = 0.001$, $a^* = 0.2$.

t	$u(0.8,t)$	
	Exact	$B-spline$
0.1	$-9.487136565624832e-04$	$-1.074955673593789e-19$
0.2	$-9.487136565612688e-04$	$-1.074955673640052e-19$
0.3	$-9.487136565600545e-04$	$-1.074955673686315e-19$
0.4	$-9.487136565588401e-04$	$-1.074955673732578e-19$
0.5	$-9.487136565576257e-04$	$-1.074955673778841e-19$
0.6	$-9.487136565564114e-04$	$-1.074955673825104e-19$
0.7	$-9.487136565551969e-04$	$-1.074955673871367e-19$
0.8	$-9.487136565539826e-04$	$-1.074955673917629e-19$
0.9	$-9.487136565527682e-04$	$-1.074955673963892e-19$
1	$-9.487136565515539e-04$	$-1.074955674010155e-19$
S	-	$9.5352e-04$

References

1. J. Cabeza, G. María, J. García, M. Andrés and A. Rodríguez, *A sequential algorithm of inverse heat conduction problems using singular value decomposition*, Int. J. Therm. Sci. **44** (2005) 235–244.
2. M. Ebrahimian, R. Pourgholi, M. Emamjome and Po. Reihani, *A numerical solution of an inverse parabolic problem with unknown boundary conditions*, Appl. Math. Comput. **189** (2007) 228–234.
3. A. Jabbari and H. Kheiri, *New exact traveling wave solutions for the Kawahara and modified Kawahara equations by using modified tanh-coth method*, Acta Univ. Apulensis **23** (2010) 21–38.
4. R. Mohammadi, *Sextic B-spline collocation method for solving Euler–Bernoulli Beam Models*, Appl. Math. Comput. **241** (2014) 151–166.
5. R. Pourgholi, M. Rostamian and M. Emamjome, *A numerical method for solving a nonlinear inverse parabolic problem*, Inverse Probl. Sci. Eng. **18** (2010) 1151–1164.
6. V. Y. Tikhonov and A. N. Arsenin, *Solutions of Ill-Posed Problems*, Distributed solely by Halsted Press, New York, 1977.

E-mail: f.torabi@std.du.ac.ir
E-mail: pourgholi@du.ac.ir
E-mail: esfahani@du.ac.ir

Steffensen-Like Methods with Twelveth-Order Convergence for Solving Nonlinear Equations

Vali Torkashvand*
Young Researchers and Elite Club, Shahr-e-Qrods Branch, Islamic Azad University, Tehran, Iran
Masoud Azimi
Farhangian University, Tehran, Iran
and Manochehr Kazemi
Department of Mathematics, Ashtian Branch, Islamic Azad University, Ashtian, Iran

Abstract. In this paper, a general procedure to develop some two-parametric with-memory methods to find simple roots of nonlinear equations is proposed. The new methods are improved extensions of without memory iterative methods.We used two self-accelerating parameters to boost up the convergence order and computational efficiency of the proposed methods without using any additional function evaluations. Numerical examples are presented to support the theoretical results of the methods.
Keywords: Root finding, Two-parametric, Self-accelerated, Order of convergence, With memory method.
AMS Mathematical Subject Classification [2010]: 65H04, 65H05.

1. Introduction

According to Kung's and Traub's conjecture, an optimal iterative method without memorybased on $k+1$ evaluations can achieve an optimal convergence order of 2^k [6]. One of the best known optimal second order methods based on two evaluations for solving the equation $f(x) = 0$ is the Steffensen method, which is given as follows (SM):

$$x_{k+1} = x_k - \frac{f(x_k)}{f[x_k, w_k]},\ w_k = x_k + f(x_k),\ k = 0, 1, 2, \ldots.$$

Methods satisfying the Kung-Traub conjecture are called optimal methods. Following Traub's work (Traub, 1964), we first expose a natural classification of iterative methods relied on the required information from the current and previous iterations:

(1) Without memory methods. This type of iterative method (I.M.) is constructed by introducing the expressions $w_1(x_k)$, $w_2(x_k), \ldots, w_n(x_k)$, where x_k is the common argument. The I.M. φ, defined as

$$x_{k+1} = \varphi(x_k, w_1(x_k), \ldots, w_n(x_k)),$$

*Speaker

is called a multipoint iteration method without memory. We see from (1) that the new approximation x_{k+1} is obtained by the use of only previous approximation x_k, but through the n expressions w_i.

(2) With memory methods. Let the I.M. have arguments z_j, where each such argument represents $n+1$ quantities $x_j, w_1(x_j), \ldots, w_n(x_j)(n \geq 1)$. Then this I.M. can be represented in the general form as

$$x_{k+1} = \varphi(z_k; z_{k-1}, \ldots, z_{k-n}).$$

Such iteration function is called a multipoint iteration function with memory. Namely, in each iterative step we must preserve information of the last n approximations x_j, and for each approximation we must calculate n expressions $w_1(x_j), \ldots, w_n(x_j)$.

The aim of this paper is to develop a two parameters derivative-free optimal family of eighth-order convergent methods. With memory methods with the same number of function evaluations have a higher efficiency index than without memory methods.

2. Construction and Convergence of New Three-Point Root Solvers

2.1. Without Memory Methods. In this section, we construct a new class of third-step with memory methods. Let us consider the following iterative formula [5]:

$$(1)\quad \begin{cases} w_k = x_k + \gamma f(x_k),\ y_k = x_k - \frac{f(x_k)}{f[x_k,w_k]},\ \gamma \in \mathbb{R} - \{0\},\ k = 0,1,2,\ldots, \\ z_k = y_k - \frac{f(y_k)}{f[x_k,y_k]+f[y_k,w_k]-f[x_k,w_k]+\beta(y_k-x_k)(y_k-w_k)},\ \beta \in \mathbb{R}, \\ t_k = \frac{f(y_k)}{f(x_k)},\ u_k = \frac{f(y_k)}{f(w_k)},\ v_k = \frac{f(z_k)}{f(y_k)},\ s_k = \frac{f(z_k)}{f(w_k)},\ p_k = \frac{f(z_k)}{f(x_k)}, \\ x_{k+1} = z_k - \frac{f(z_k)(H_1(t_k)+H_2(u_k)+H_3(v_k)+H_4(s_k)+H_5(p_k))}{f[x_k,z_k]+f[z_k,y_k]-f[x_k,y_k]+\beta(z_k-y_k)(z_k-x_k)}. \end{cases}$$

The following theorem indicates underwhat conditions on the weight functions in (1) the order of convergence is eight.

THEOREM 2.1. *Let $I \subseteq \mathbb{R}$ be an open interval, $f : I \to \mathbb{R}$ be a differentiable function, and has a simple zero, say α. If x_0 is an initial guess to α, then method (1) has eight-order convergence, when the weight functions $H_1(t_k)$, $H_2(u_k)$, $H_3(v_k)$, $H_4(s_k)$ and $H_5(p_k)$ satisfy the following conditions:*

$$\begin{cases} H_1(0) = 1,\ H_1'(0) = H_1''(0) = H_1'''(0) = 0,\ |H_1^{(4)}(0)| \leq 0, \\ H_2(0) = H_2'(0) = H_2''(0) = 0,\ H_2^{(3)}(0) = -(6 + 6\gamma f[x_k, w_k]),\ |H_2^{(4)}(0)| \leq 0, \\ H_3(0) = H_3'(0) = 0,\ |H_3''(0)| \leq 0, \\ H_4(0) = 0,\ H_4'(0) = 1,\ |H_4''(0)| \leq 0, \\ H_5(0) = H_5'(0) = 0,\ |H_5''(0)| \leq 0. \end{cases}$$

Also the error equation of the method (1) is given by

$$(2)\quad \begin{aligned} e_{k+1} = {} & (1+\gamma f'(\alpha))^3 c_2^2(\beta + f'(\alpha)c_2^2 - f'(\alpha)c_3)(-f'(\alpha)(3+\gamma f'(\alpha)c_2^3 + c_2(\beta(4+3\gamma f'(\alpha)) \\ & -2f'(\alpha)(1+\gamma f'(\alpha))c_3) + f'(\alpha)(1+\gamma f'(\alpha))c_4) f'(\alpha)^{-2} e_k^8 + O(e_k^9). \end{aligned}$$

PROOF. First, we define the Taylor series of $f(x)$ as follows:

$$In\ [1] : f[e\_] = fla(e + c_2e^2 + \cdots + c_8e^8),$$

where $e = x - \alpha$, $fla = f'(\alpha)$. Note that since α is a simple zero of $f(x)$, the $f'(\alpha) \neq 0, f(\alpha) = 0$. We define

$$\begin{aligned}
&In\ [2] : f[x_, y_] = \frac{f[x] - f[y]}{x - y};\\
&In\ [3] : ew = e + \gamma f[e];\\
&In\ [4] : ey = e - Series[\frac{f[e]}{f[e, ew]}, \{e, 0, 8\}];
\end{aligned}$$

$$\begin{aligned}
In\ [5] : ez &= ey\\
&- Series[\frac{f[ey]}{f[e, ey] + f[ey, ew] - f[e, ew] + \beta(ey - e)(ey - ew)}, \{e, 0, 8\}];\\
In\ [6] : e_{k+1} &= ez\\
&- Series[f[ez](1 - (1 + \gamma f[e, ew]) * (\frac{f[e]}{f[ew]})^3 + 0 + 0 + \frac{f[ez]}{f[ew]})/(f[e, ez]\\
&+ f[ez, ey] - f[e, ey] + \beta(ez - ey)(ez - e)), \{e, 0, 8\}]//Full\ Simplify;\\
Out\ [6] : e_{k+1} &= (1 + \gamma fla)^3c_2^2(\beta + flac_2^2 - flac_3)(-fla(3 + \gamma flac_2^3 + c_2(\beta(4 + 3\gamma fla)\\
&- 2fla(1 + \gamma fla)c_3) + fla(1 + \gamma fla)c_4)fla^{-2}e^8 + O(e^9),
\end{aligned}$$

and thus proof is completed. □

2.2. New Families of Iterative Methods with Memory. It is clear from error equations (2) that the order of convergence of the family (1) is eight, when $(1 + \gamma f'(\alpha)) \neq 0$ and $(\beta + f'(\alpha)c_2^2 - f'(\alpha)c_3) \neq 0$. Therefore, it is possible to increase the convergence speed of the proposed class (1), if $(1 + \gamma f'(\alpha)) = 0$ and $(\beta + f'(\alpha)c_2^2 - f'(\alpha)c_3) = 0$ of $f'(\alpha)$, $f''(\alpha)$ and $f'''(\alpha)$ are not available in practice and such acceleration is not possible. Instead of that, we could use approximations $\tilde{f}'(\alpha) \approx f'(\alpha)$, $\tilde{f}''(\alpha) \approx f''(\alpha)$ and $\tilde{f}'''(\alpha) \approx f'''(\alpha)$, calculated by already available information. Therefore, by setting $\gamma = \frac{1}{\tilde{f}'(\alpha)}$ and $\beta = \frac{\tilde{f}'''(\alpha)}{6} - \frac{\tilde{f}''(\alpha)^2}{4\tilde{f}'(\alpha)}$ convergence order without using any new functional evaluation. Hence, the main idea in constructing methods with memory consists of the calculation of the parameters $\gamma = \gamma_k$ and $\beta = \beta_k$ as the iteration proceeds by the formula $\gamma_k = \frac{1}{\tilde{f}'(\alpha)}$ and $\beta_k = \frac{\tilde{f}'''(\alpha)}{6} - \frac{\tilde{f}''(\alpha)^2}{4\tilde{f}'(\alpha)}$ for $k = 2, 3, \ldots$. Therefore, we approximate

$$\begin{cases}
\gamma_k = -\frac{1}{\tilde{f}'(\alpha)} = -\frac{1}{N_4'(x_k)},\\
\beta_k = \frac{\tilde{f}'''(\alpha)}{6} - \frac{\tilde{f}''(\alpha)^2}{4\tilde{f}'(\alpha)} = \frac{N_6'''(y_k)}{6} - \frac{(N_6''(y_k))^2}{4N_6'(y_k)},
\end{cases}$$

where $N_4'(x_k), N_6''(y_k)$ and $N_6'''(y_k)$ are Newton's interpolation polynomials go through the nodes $\{x_k, x_{k-1}, w_{k-1}, y_{k-1}, z_{k-1}\}, \{y_k, w_k, x_k, x_{k-1}, w_{k-1}, y_{k-1}, z_{k-1}\}$,

and $\{y_k, w_k, x_k, x_{k-1}, w_{k-1}, y_{k-1}, z_{k-1}\}$, respectively. Now, we obtain the new iterative method with memory as follows:

$$
(3)\qquad \begin{cases} \gamma_k = -\frac{1}{N_4'(x_k)},\ \beta_k = \frac{N_6'''(y_k)}{6} - \frac{(N_6''(y_k))^2}{4N_6'(y_k)},\ k = 1, 2, 3, \ldots, \\ w_k = x_k + \gamma_k f(x_k),\ y_k = x_k - \frac{f(x_k)}{f[x_k, w_k]},\ \gamma_k \in \mathbb{R} - \{0\},\ k = 0, 1, 2, \ldots, \\ z_k = y_k - \frac{f(y_k)}{f[x_k,y_k]+f[y_k,w_k]-f[x_k,w_k]+\beta_k(y_k-x_k)(y_k-w_k)},\ \beta_k \in \mathbb{R}, \\ x_{k+1} = z_k - \frac{f(z_k)(1-(1+\gamma_k f[x_k,w_k])*(\frac{f(x_k)}{f(w_k)})^3+0+0+\frac{f(z_k)}{f(w_k)})}{f[x_k,z_k]+f[z_k,y_k]-f[x_k,y_k]+\beta_k(z_k-y_k)(z_k-x_k)}. \end{cases}
$$

THEOREM 2.2. *If an initial guess x_0 is sufficiently close to the zero α of $f(x)$ and the parameters γ_k and β_k in the iterative scheme (3) is recursively calculated then the R-order of convergence of with memory methods (3) is at least $\frac{1}{2}(13 + \sqrt{137}) \approx 12.3523$.*

PROOF. Firstly, we assume that the R-orders of convergence of the sequences w_k, y_k, z_k and x_k are at least r_1, r_2, r_3 and r, respectively. Hence

$$
(4)\qquad \begin{cases} e_{k+1} \sim e_k^r \sim e_k^{r^2}, \\ e_{k,z} \sim e_k^{r_3} \sim e_k^{rr_3}, \\ e_{k,y} \sim e_k^{r_2} \sim e_k^{rr_2}, \\ e_{k,w} \sim e_k^{r_1} \sim e_k^{rr_1}. \end{cases}
$$

Also, we desist from retyping the widely practiced approach in the before and put forward the self-explained Mathematica code used to supply a way that the proposed family with-memory (3) achieves R-order equal 12.3.

```
ClearAll["Global'*"];
A[t_]:=InterpolatingPolynomial[{{e, fx}, {ew, fw}, {ey, fy}, {e1, fx1}}, t];
Approximation=-1/A'[e1]//Simplify;
fx = fla*(e + c2*e^2 + c3*e^3 + c4*e^4 + c5*e^5 + c6*e^6 + c7*e^7 + c8*e^8);
fw = fla*(ew+c2*ew^2+c3*ew^3+c4*ew^4+c5*ew^5+c6*ew^6+c7*ew^7+c8*ew^8);
fy = fla*(ey+c2*ey^2+c3*ey^3+c4*ey^4+c5*ey^5+c6*ey^6+c7*ey^7+c8*ey^8);
fz = fla*(ez+c2*ez^2+c3*ez^3+c4*ez^4+c5*ez^5+c6*ez^6+c7*ez^7+c8*ez^8);
fx1 = fla*(e1+c2*e1^2+c3*e1^3+c4*e1^4+c5*e1^5+c6*e1^6+c7*e1^7+c8*e1^8);
γ = Series[Approximation, {e, 0, 2}, {ew, 0, 2}, {ey, 0, 2}, {e1, 0, 0}]//Simplify;
 Collect[Series[1 + γ * fla, {e, 0, 1}, {ew, 0, 1}, {ey, 0, 1}, {ez, 0, 1}, {e1, 0, 0}],
{e, ew, ey, ez, e1},Simplify],
```

which results in

$$c_5 eeweyez.$$

Therefore, one may obtain

$$
(5)\qquad 1 + \gamma_k f'(\alpha) \sim c_5 e_{k-1} e_{k-1,w} e_{k-1,y} e_{k-1,z}.
$$

We also have similarly

$$
(6)\qquad \beta_k + f'(\alpha)c_2^2 - f'(\alpha)c_3 \sim c_5 e_{k-1} e_{k-1,w} e_{k-1,y} e_{k-1,z}.
$$

Using relations (4), (5), and (6), we have

$$(7)\quad \begin{cases} e_{k+1} \sim (1+\gamma_k f'(\alpha))^3(\beta_k + f'(\alpha)c_2^2 - f'(\alpha)c_3)e_k^8 \sim (e_{k-1}e_{k-1,w}e_{k-1,y}e_{k-1,z})^4 e_k^8 \\ \sim e_{k-1}^{4(1+r_1+r_2+r_3)+8r}, \\ e_{k,z} \sim (1+\gamma_k f'(\alpha))^2 c_2(\beta_k + f'(\alpha)c_2^2 - f'(\alpha)c_3)e_k^4 \sim (e_{k-1}e_{k-1,w}e_{k-1,y}e_{k-1,z})^3 e_k^4 \\ \sim e_{k-1}^{3(1+r_1+r_2+r_3)+4r}, \\ e_{k,y} \sim (1+\gamma_k f'(\alpha))c_2 e_k^2 \sim e_{k-1}e_{k-1,w}e_{k-1,y}e_{k-1,z}e_k^2 \sim e_{k-1}^{1+r_1+r_2+r_3+2r}, \\ e_{k,w} \sim (1+\gamma_k f'(\alpha))e_k \sim e_{k-1}e_{k-1,w}e_{k-1,y}e_{k-1,z}e_k \sim e_{k-1}^{1+r_1+r_2+r_3+r}. \end{cases}$$

Comparing the exponents of e_{k-1} in four expressions (4) and (7) of e_{k+1}, $e_{k,z}$, $e_{k,y}$, $e_{k,w}$, we have four equations in the following system:

$$\begin{cases} rr_1 - (1+r_1+r_2+r_3) - r = 0, \\ rr_2 - (1+r_1+r_2+r_3) - 2r = 0, \\ rr_3 - 3(1+r_1+r_2+r_3) - 4r = 0, \\ r^2 - 4(1+r_1+r_2+r_3) - 8r = 0. \end{cases}$$

The positive answer to the above equations system as follows:

$$r_1 = \frac{1}{8}(5+\sqrt{137}),\ r_2 = \frac{1}{8}(13+\sqrt{137}),\ r_3 = \frac{1}{8}(23+3\sqrt{137}),\ r = \frac{1}{2}(13+\sqrt{137}),$$

which specifies the R-order of convergence of the derivative-free scheme with memory (3) is $r = \frac{1}{2}(13+\sqrt{137})$ (denoted by TAKM). □

Remark 2.3. The new three-step derivative-free methods (3) require four function evaluations and have the order of convergence 12.3523. Hence, the efficiency index of the proposed methods is $12.3523^{\frac{1}{4}} = 1.87472$ which is much better than optimal one untill four-point optimal methods without memory having efficiency indexes $EI = 2^{1/2} \simeq 1.41421, EI = 4^{1/3} \simeq 1.58740, EI = 8^{1/4} \simeq 1.68179, EI = 16^{1/5} \simeq 1.74110$, respectively.

3. Numerical Results and Discussions

Now, we further want to check the efficiency of the proposed scheme and validate the theoretical results. For this purpose, we use the following test functions [7] and display the approximate.

$$\begin{aligned} f_1(x) &= x\log(1+x\sin(x)) + e^{-1+x^2+x\cos(x)}\sin(\pi x),\ \alpha = 0,\ x_0 = 0.6, \\ f_2(x) &= 1 + \frac{1}{x^4} - \frac{1}{x} - x^2,\ \alpha = 1,\ x_0 = 1.4. \end{aligned}$$

Now, we choose our proposed scheme (3) (for $\beta = 0.1$ and $\gamma = 0.1$), called by TM for comparison with the existing robust optimal eighth-order schemes which were proposed by Lotfi et al. in [3], Kung and Traub in [2] (for $\gamma = 0.1$), Sharma and Arora (for Method 1) [4], Soleymani in [5] (for $\beta = \gamma = 0.1$), Bi et al. (for $\beta = 0$, $\gamma = -2$, $\lambda = -2.5$) [1] and respectively, called by LSSSM, KTM, SAM, BRWM and SM. For better comparisons of our proposed methods with other existing ones, we have given two types of comparison tables in each test function:

(a) Absolute error between the two consecutive iterations $|x_{n+1} - x_n|$,
(b) Absolute residual error in the corresponding function ($|f(x_n)|$).

The errors of approximations to the corresponding zeros of test functions are displayed in Table 1, where $A(h)$ denotes $A \times 10^{-h}$ and D stands for divergent. These tables include the values of the computational order of convergence COC calculated by the formula [7]

$$COC = \frac{\log |f(x_n)/f(x_{n-1})|}{\log |f(x_{n-1})/f(x_{n-2})|}.$$

TABLE 1. Comparison of the absolute error of the proposed method with other methods.

$f_1(x) = x\log(1+x\sin(x)) + e^{-1+x^2+x\cos(x)}\sin(\pi x)$, $\alpha = 0$, $x_0 = 0.6$					
Methods	$\lvert x_1 - \alpha\rvert$	$\lvert x_2 - \alpha\rvert$	$\lvert x_3 - \alpha\rvert$	COC	EI
BRWM (39) [1]	0.39745(−2)	0.23663(−18)	0.38370(−148)	8.00000	1.68179
KTM [2]	0.23230(−1)	0.33730(−13)	0.13863(−107)	8.00000	1.68179
LSSSM (14) [3]	0.42171(−2)	0.77543(−18)	0.10331(−143)	8.00000	1.68179
SAM [4]	0.94733(−1)	0.44063(−8)	0.59502(−67)	8.00000	1.68179
SM [5]	0.22337(−1)	0.15109(−12)	0.00000(0)	8.00000	1.68179
TAKM (3)	0.22337(−1)	0.80542(−18)	0.26722(−216)	12.00000	1.86121
$f_2(x) = \frac{1}{x^4} - x^2 - \frac{1}{x} + 1$, $\alpha = 1$, $x_0 = 1.4$					
Methods	$\lvert x_1 - \alpha\rvert$	$\lvert x_2 - \alpha\rvert$	$\lvert x_3 - \alpha\rvert$	COC	EI
BRWM (39) [1]	0.54274(−3)	0.12998(−23)	0.13984(−188)	8.00000	1.68179
KTM, $\gamma = 1$ [2]	0.10721(−1)	0.45584(−12)	0.50318(−95)	8.00000	1.68179
LSSSM (14) [3]	0.70023(−3)	0.18630(−22)	0.46761(−179)	8.00000	1.68179
SAM [4]	0.39080(−3)	0.38443(−26)	0.33692(−210)	8.00000	1.68179
SM [5]	0.18638(−3)	0.00000(−1)	0.00000(0)	8.00000	1.68179
TAKM (3)	0.18638(−3)	0.27205(−41)	0.76931(−497)	12.00000	1.86121

It was observed that the proposed method can be competitive to methods [1, 2, 3, 4, 5, 6] and also improve the existing methods [5]. Our approach can be continuously applied in order to improve any existing iteration formula.

Acknowledgement

The authors would like to thank the referees for their constructive comments and suggestions which substantially improved the quality of this paper.

References

1. W. Bi, Q. Wu and H. Ren, *A new family of eighth-order iterative methods for solving nonlinear equations,* Appl. Math. Comput. **214** (1) (2009) 236–245.
2. H. T. Kung and J. F. Traub, *Optimal order of one-point and multi-point iteration,* J. ACM **21** (4) (1974) 643–651.
3. T. Lotfi, S. Sharifi, M. Salimi and S. Siegmund, *A new class of three-point methods with optimal convergence order eight and its dynamics,* Numer. Algorithms **68** (2) (2015) 261–288.
4. J. R. Sharma and H. Arora, *A new family of optimal eighth order methods with dynamics for nonlinear equations,* Appl. Math. Comput. **273** (2016) 924–933.
5. F. Soleymani, *On a bi-parametric class of optimal eighth-order derivative-free methods*, Int. J. Pure Appl. Math. **72** (1) (2011) 27–37.

6. J. F. Traub, *Iterative Methods for the Solution of Equations*, Prentice-Hall, Englewood Cliffs, New Jersey, 1964.
7. V. Torkashvand, T. Lotfi and M. A. Fariborzi Araghi, *A new family of adaptive methods with memory for solving nonlinear equations*, Math. Sci. **13** (1) (2019) 1–20.

E-mail: torkashvand1978@gmail.com
E-mail: azimidr45@gmail.com
E-mail: m.kazemi@aiau.ac.ir

A New Modified Generalized Shift-Splitting Preconditioner for Saddle Point Problems

Ghodrat Ebadi
Faculty of Mathematical Sciences, University of Tabriz, Tabriz, Iran
and Seryas Vakili*
Faculty of Mathematical Sciences, University of Tabriz, Tabriz, Iran

Abstract. In this paper, a new modified generalized shift-splitting (NMGSS) method and its induced preconditioner is proposed for solving nonsymmetric saddle point problems. The convergence analysis of NMGSS iteration method is discussed. Finally the efficiency of methods are illustrated by giving one example.
Keywords: Saddle-point, Generalized shift-splitting, Preconditioner, Convergence.
AMS Mathematical Subject Classification [2010]: 65F10, 65F08.

1. Introduction

We consider a nonsymmetric saddle point problem as

$$\mathfrak{A}u = \begin{pmatrix} A & B \\ -B^T & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} f \\ -g \end{pmatrix} \equiv b, \tag{1}$$

where $A \in \mathbb{R}^{m\times m}$ is nonsymmetric positive definite, $B \in \mathbb{R}^{m\times n}$ has full column rank, $f \in \mathbb{R}^m$ and $g \in \mathbb{R}^n$, with $m \geq n$. Here, B^T is the transpose of B.

In a variety of engineering and scientific applications, such as computational fluid dynamics, optimal control, and networks computer graphics [2] solving the linear system (1) is required. When the matrices of coefficient matrix $\mathfrak{A}$, i.e. A and B are large and sparse, iterative methods are better suited for solving saddle point problems compared to direct methods [4]. If B in (1) has full column rank, then the coefficient matrix $\mathfrak{A}$ is nonsingular, in this case, the problem will be called a nonsingular saddle point problem and when B has a rank deficiency, Eq. (1) is called the singular saddle point problem and the coefficient matrix $\mathfrak{A}$ is singular. In recent years, various authors have proposed a number of useful iterative methods to solve (1). Cao et al. [5] presented the SS preconditioner as

$$\mathcal{P}_{SS} = \frac{1}{2}\begin{pmatrix} \alpha I + A & B \\ -B^T & \alpha I \end{pmatrix},$$

where $\alpha \geq 0$ and I is the unit matrix with suitable dimension. Cao et al. [6] and Chen et al. [7] introduced parameter β instead of α in last block of $\mathcal{P}_{SS}$ and provided the generalized shift-splitting (GSS) preconditioner. Huang and Su [8] utilized a modified shift-splitting ($MSSP$) preconditioner in order to increase the

*Speaker

rate of convergence of the GSS method for solving the saddle-point problem with a full ranked B matrix and symmetric positive definite (1,1) part of the form

$$\mathcal{P}_{MSSP} = \begin{pmatrix} \alpha I + 2A & 2B \\ -2B^T & \alpha I \end{pmatrix}.$$

Huang et al. [8] substituted parameter α in the final block of $\mathcal{P}_{MSSP}$ by parameter β and constructed a new preconditioner from the saddle point matrix $\mathfrak{A}$ as

$$\mathcal{P}_{MGSS} = \begin{pmatrix} \alpha I + 2A & 2B \\ -2B^T & \beta I \end{pmatrix},$$

where $\alpha \geq 0$, $\beta > 0$.

These studies led us to introduce the new modified generalized shift-splitting (NMGSS) preconditioner in order to improve the convergence rate of (1) problems. In the current study the convergence of the proposed iteration method, and the spectral properties of $NMGSS$ preconditioned matrix are investigated. We carry out a numerical example in order to show the efficiency of $NMGSS$ method and the GMRES method with the $NMGSS$ preconditioner for solving (1).

This paper is structured as follows: Section 2 will introduce the new generalized shift-splitting preconditioner and its implementation. Section 3 presents the convergence properties of the $NMGSS$ iteration method. Finally, the numerical results are provided in Section 4.

2. The New Modified Generalized Shift-Splitting Preconditioner

In this Section, using idea of [8, 9], a new splitting of matrix $\mathfrak{A}$ is presented as

$$\begin{aligned} \mathfrak{A} &= \mathcal{P}_{NMGSS} - \mathcal{Q}_{NMGSS} \\ &= \begin{pmatrix} \alpha H + 2A2B \\ -2B^T\beta I \end{pmatrix} - \begin{pmatrix} \alpha H + AB \\ -B^T\beta I \end{pmatrix}, \end{aligned} \tag{2}$$

where $\alpha \geq 0$, $\beta > 0$ and $H = \frac{A+A^T}{2}$. Therefore, using (2), we present a new method as follows:

The $NMGSS$ Iteration Method. Let $\alpha \geq 0$ and $\beta > 0$. Assume $(x^{(0)^T}, y^{(0)^T})^T$ be an initial guess for $k = 0, 1, 2, ...$, until $(x^{(0)^T}, y^{(k)^T})^T$ converges, compute

$$\mathcal{P}_{NMGSS} \begin{pmatrix} x^{(k+1)} \\ y^{(k+1)} \end{pmatrix} = \mathcal{Q}_{NMGSS} \begin{pmatrix} x^{(k)} \\ y^{(k)} \end{pmatrix} + \begin{pmatrix} f \\ -g \end{pmatrix}, \tag{3}$$

The iteration scheme (3) can be rewritten as follows

$$\begin{pmatrix} x^{(k+1)} \\ y^{(k+1)} \end{pmatrix} = \Gamma(\alpha, \beta) \begin{pmatrix} x^{(k)} \\ y^{(k)} \end{pmatrix} + \begin{pmatrix} \alpha H + 2A & 2B \\ -2B^T & \beta I \end{pmatrix}^{-1} \begin{pmatrix} f \\ -g \end{pmatrix}, \tag{4}$$

where

$$\Gamma(\alpha, \beta) = \begin{pmatrix} \alpha H + 2A & 2B \\ -B^T & \beta I \end{pmatrix}^{-1} \begin{pmatrix} \alpha H + A & B \\ -B^T & \beta I \end{pmatrix},$$

is the iteration matrix of the NMGSS method, and

$$\mathcal{P}_{NMGSS} = \begin{pmatrix} \alpha H + 2A & 2B \\ -2B^T & \beta I \end{pmatrix},$$

is called the $NMGSS$ preconditioner for $\mathfrak{A}$. At each step of (4) or applying $\mathcal{P}_{NMGSS}$ within the Krylov subspace methods, we have to solve a linear system in the following form

$$\begin{pmatrix} \alpha H + 2A & 2B \\ -2B^T & \beta I \end{pmatrix} z = r,$$

where $z = (z_1^T, z_2^T)^T$, $r = (r_1^T, r_2^T)^T$ and $z_1, r_1 \in R^m$, $z_2, r_2 \in R^n$. By decomposition of $\mathcal{P}_{NMGSS}$ a linear system with the coefficient matrix $(\alpha H + 2A + \frac{4}{\beta} BB^T)x = b$ needs to be solved. Since the matrix $(\alpha H + 2A + \frac{4}{\beta} BB^T)x = b$ for all $\alpha \geq 0$ and β is positive definite, in inexact manner, we can use the GMRES method for solving this sub-linear system by a prescribed accuracy. Also, they can be solved exactly by the LU factorization in combination with AMD or column AMD reordering.

3. The Convergence of the $NMGSS$ Iteration Method

To present the convergent properties of the $NMGSS$ iteration method, we give some necessary lemma.

LEMMA 3.1. [3] *Both roots of the complex quadratic equation $x^2 - \phi x + \psi = 0$ are less than one in modulus if and only if $|\phi - \bar{\phi}\psi| + |\psi|^2 < 1$, where $\bar{\phi}$ denotes the conjugate complex of ϕ.*

LEMMA 3.2. *Assum $A \in \mathbb{R}^{m\times m}$ be a positive definite matrix, $B \in \mathbb{R}^{m\times n}$ has full column rank, $\alpha \geq 0$ and $\beta > 0$. If λ is an eigenvalue of the $\Gamma(\alpha, \beta)$, then $\lambda \neq \pm 1$.*

LEMMA 3.3. *Assume λ be an eigenvalue of $\Gamma(\alpha, \beta)$ and $(u^*, v^*)^* \in \mathbb{C}^{m\times n}$, be the corresponding eigenvector and all the conditions in Lemma 3.2 are satisfied, then $u \neq 0$. Moreover, if $v = 0$, then $|\lambda| < 1$.*

THEOREM 3.4. *By the satisfaction conditions of the Lemma 3.2 and letting $(\lambda, (u^*, v^*)^*)$ be an eigenpair of $\Gamma(\alpha, \beta)$ of the $NMGSS$. Then the $NMGSS$ iteration method converges to the exact solution of the saddle point problem (1).*

4. Numerical Experiments

We provide an example to explain the feasibility and effectiveness of the $NMGSS$ method for solving (1). In this example, the linear system $(\alpha I + 2H + \frac{1}{\beta} BB^T)x = b$ contained in the $GMSS$ is solved inexactly by the CG method and for $MGSS$, $NMGSS$ iteration methods, we solve linear subsystems $(\alpha I + 2A + \frac{4}{\beta} BB^T)x = b$, $(\alpha H + 2A + \frac{4}{\beta} BB^T)x = b$ respectively, in an inexact manner using the GMRES methods. The inner CG and GMRES methods are terminated if the current residual of the inner iteration satisfies $\frac{\|r^{(k)}\|}{\|r^{(0)}\|} < 10^{-7}$, where $r^{(k)}$ denotes the residual of the kth CG and GMRES iteration. Moreover, the run will be terminated when $RES < 10^{-6}$ or the number of iterations exceeds $\kappa_{max} = 500$, where

$$RES = \frac{\sqrt{\|f - Ax^{(k)} - By^{(k)}\|_2^2 + \|g - B^T x^{(k)}\|_2^2}}{\sqrt{\|f\|_2^2 + \|g\|_2^2}} < 10^{-6}.$$

All the methods were solved using MATLAB (version R2015b 64-bit) and all the experiments implemented on a PC with windows system and Intel (R) Core (TM) i7-7700k CPU @ 4.20 GHz and 8.0 GB of RAM.

EXAMPLE 4.1. The problem structured as (1) was considered with the following coefficient sub-matrices [1]

$$A = \begin{pmatrix} I \otimes T + T \otimes I & 0 \\ 0 & I \otimes T + T \otimes I \end{pmatrix} \in \mathbb{R}^{2p^2 \times 2p^2},$$
$$B = \begin{pmatrix} I \otimes F \\ F \otimes I \end{pmatrix} \in \mathbb{R}^{2p^2 \times p^2},$$

and

$$T = \frac{\mu}{h^2}.tridiag(-1, 2, -1) + \frac{1}{2h}.tridiag(-1, 0, 1) \in \mathbb{R}^{p \times p},$$
$$F = \frac{1}{h}.tridiag(-1, 1, 0) \in \mathbb{R}^{p \times p},$$

where $h = \frac{1}{p+1}$ and $\otimes$ denotes the Kronecker product.

Table 1 shows the efficiency of the $NMGSS$ method by selecting small values for α and β.
Table 1 gives the numerical experiments for the different iteration methods whose

TABLE 1. Numerical results for the example with $\mu = 0.1$.

Method	p	16	32	64
GMSS	α	22	36	38
	β	16	8.3	5.9
	IT.	66	73	89
	CPU	0.038	0.16	0.97
	RES	$8.447e-07$	$9.086e-07$	$9.48e-07$
MGSS	α	0.2	0.5	0.2
	β	0.1	0.1	0.1
	IT.	21	21	21
	CPU	0.25	0.11	0.53
	RES	$9.88e-07$	$9.85e-07$	$9.57e-07$
NMGSS	α	0.01	0.01	0.01
	β	0.2	0.2	0.2
	IT.	21	21	21
	CPU	0.17	0.072	0.39
	RES	$9.622e-07$	$9.67e-07$	$9.734e-07$

optimal parameters have been found experimentally based on the minimizing of the iterations when $\mu = 0.1$ for different grids. Compared to the other two methods, the $NMGSS$ iteration method for solving Example 4.1 requires less processing time. The IT of the $GMSS$ in comparison the $MGSS$ and $NMGSS$ methods shows more sensitivity to p. We present numerical experiments of the $GMSS$, $MGSS$and $NMGSS$ preconditioned GMRES methods on different uniform grids with $\mu = 0.1$ in Tables 2. Note that I in Table 2 indicates GMRES method without preconditioning. The GMRES method with $\mathcal{P}_{NMGSS}$ preconditioning has been shown to be both feasible and efficient in Table 2. The parameters which were considered for the selected preconditions in the two Table 2 are as follows [6]:

$$\alpha_{GMSS} = \mu, \quad \beta_{GMSS} = \frac{\|B\|_2^2}{2\|H\|_2}, \qquad \alpha_{MGSS} = \mu, \quad \beta_{MGSS} = \frac{2\|B\|_2^2}{\|A\|_2},$$
$$\alpha_{NMGSS} = \mu, \quad \beta_{NMGSS} = \frac{\|B\|_2^2}{\|H\|_2}.$$

TABLE 2. Numerical of results for the three preconditioned GMRES methods.

Method	p	16	32	64
I	IT.	115	240	495
	CPU	0.1326	3.4868	81.8770
	RES	$9.50e-07$	$9.34e-07$	$9.73e-07$
$\mathcal{P}_{GMSS}$	α	0.1	0.1	0.1
	β	4.9974	4.9996	5
	IT.	23	24	25
	CPU	0.24	0.56	3.17
	RES	$6.78e-07$	$8.45e-07$	$6.72e-07$
$\mathcal{P}_{MGSS}$	α	0.1	0.1	0.1
	β	19.9861	19.9983	19.9998
	IT.	14	15	15
	CPU	0.077	0.36	1.65
	RES	$4.457e-07$	$3.123e-07$	$5.859e-07$
$\mathcal{P}_{MGSS}$	α	0.1	0.1	0.1
	β	9.9947	9.9993	9.9999
	IT.	12	13	14
	CPU	0.056	0.18	1.44
	RES	$3.9784e-07$	$5.005e-07$	$6.909e-07$

References

1. Z. Z. Bai, G. H. Golub and J. Y. Pan, *Preconditioned Hermitian and skew-Hermitian splitting methods for non-Hermitian positive semidefinite linear systems*, Numer. Math. **98** (2004) 1-32.
2. Z. Z. Bai, *Structured preconditioners for nonsingular matrices of block two-by-two structures*, Math. Comp. **75** (2006) 791–815.
3. Z. Z. Bai and Z. Q. Wang, *On parameterized inexact Uzawa methods for generalized saddle point problems*, Linear Algebra Appl. **428** (2008) 2900-2932.
4. M. Benzi, G. H. Golub and J. Liesen, *Numerical solution of saddle point problems*, Acta. Numer. **14** (2015) 1–137.
5. Y. Cao, J. Du and Q. Niu, Shift-splitting preconditioners for saddle point problems, J. Comput. Appl. Math. **272** (2014) 239–250.
6. Y. Cao, S. Li and L. Q. Yao, *A class of generalized shift-splitting preconditioners for non-symmetric saddle point problems*, Appl. Math. Lett. **49** (2015) 20–27.
7. C. R. Chen and C. F. Ma, *A generalized shift-splitting preconditioner for saddle point problems*, Appl. Math. Lett. **43** (2015) 49–55.
8. Z. G. Huang, L. G. Wang, Z. Xu and J. J. Cui, *A modified generalized shift-splitting preconditioner for nonsymmetric saddle point problems*, Numer. Algor. **78** (2018) 297–331.
9. D. K. Salkuyeh and M. Rahimian, *A modification of the generalized shift-splitting method for singular saddle point problems*, Comput. Math. Appl. **74** (12) (2017) 2940–2949.

E-mail: s.vakili@tabrizu.ac.ir

E-mail: ghodrat_ebadi@yahoo.com; gebadi@tabrizu.ac.ir

Fully Spectral Galerkin Method for the Modified Distributed-Order Anomalous Sub-Diffusion Equation

Azam Yazdani*
Department of Mathematics and Computer Science, Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran
and Farhad Fakhar-Izadi
Department of Mathematics and Computer Science, Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran

Abstract. We present a high-order spectral method for the modified time-fractional distributed-order anomalous sub-diffusion equations. First, we discretize the integral term using a Gauss-quadrature formula and convert it into a multi-term equation. The discretization leads to converting the problem to a Sylvester matrix equation. Two numerical examples represent the accuracy and efficiency of the method.
Keywords: Distributed-order anomalous sub-diffusion equation, Galerkin spectral element method, Sylvester matrix equation, Riemann-Liouville fractional derivative.
AMS Mathematical Subject Classification [2010]: 26A33, 76M22, 35R11.

1. Introduction

In this paper, we consider the following distributed-order anomalous sub-diffusion equation

$$\frac{\partial}{\partial t}u(\mathbf{x},t) = \int_0^1 \varpi(\gamma)\, {}_0D_t^{1-\gamma}\, K\triangle u(\mathbf{x},t)\, d\gamma + f(\mathbf{x},t), \quad (\mathbf{x},t)\in\Lambda\times I, \tag{1}$$

where K is positive constant and the operator ${}_0D_t^{1-\gamma}$ is the Riemann-Liouville derivative of order $1-\gamma$, with the initial and boundary conditions

$$u(\mathbf{x},0) = u_0(\mathbf{x}), \quad \mathbf{x}\in\Lambda,$$

$$\mathcal{B}u(\mathbf{x},t) = u_b(\mathbf{x},t), \quad \mathbf{x}\in\partial\Lambda\times I,$$

where $\Lambda=[0,L]$ ($[0,L]^2$ in 2D) is a bounded domain, $I=(0,T]$ and $\varpi(\gamma)$ satisfies the following conditions

$$0\le\varpi(\gamma), \quad \varpi(\gamma)\neq 0, \quad \gamma\in[0,1], \quad \int_0^1 \varpi(\gamma)d\gamma = W > 0.$$

Lang [6] considered problem (1) with the homogenous boundary condition by the backward finite difference scheme in time and Galerkin finite element method

*Speaker

in space. Fakhar-Izadi [2] proposed the space-time Petrov-Galerkin (PG) spectral method for a distributed-order time-fractional fourth-order partial differential equations (PDEs) with mixed boundary conditions. Abbaszadeh and Dehghan [1] developed the meshless Galerkin method based upon the shape functions of RKPM for solving the fractional modified distributed-order anomalous sub-diffusion equation.

In most papers, the finite difference method (FDM) is applied for discretizing time-fractional derivative because it is straightforward, but using FDM leads to algebraic convergence and limited accuracy. Also, due to the nonlocal nature of fractional operators, the cost of computing FDM is high. So in this paper, a global PG spectral method is applied for time discretization. Also, the modal spectral element method (SEM) is considered for approximation of the space variables. The proposed approach leads to obtain the approximate solution of the problem (1) through solving a Silvester matrix equation that can be solved efficiently by the QZ algorithm [4].

2. Implementation of the Proposed Method

For discretizing problem (1), first, we should approximate the distributed-order fractional derivative operator by a proper Gauss-Legendre quadrature formula. So, the following multi-term fractional sub-diffusion equation is obtained

$$\frac{\partial u(\mathbf{x},t)}{\partial t} \simeq \frac{K}{2}\sum_{s=0}^{Q} \varpi\left(\frac{\beta_s+1}{2}\right) {}_0D_t^{\frac{1-\beta_s}{2}} \triangle u(\mathbf{x},t)\,\omega_s + f(\mathbf{x},t),\ (\mathbf{x},t)\in\Lambda\times I, \tag{2}$$

$$\begin{aligned} u(\mathbf{x},0) &= u_0(\mathbf{x}),\quad \mathbf{x}\in\Lambda,\\ \mathcal{B}u(\mathbf{x},t) &= u_b(\mathbf{x},t),\quad \mathbf{x}\in\partial\Lambda\times I, \end{aligned}$$

where distinct nodes $\beta_0<\beta_1<\cdots<\beta_Q$ are roots of $(Q+1)$th Legendre polynomial and $\{\omega_s\}_{s=0}^{Q}$ are corresponding weights.

By multiplying a proper test function ν and integrating over the computational domain $\Omega=\Lambda\times I$, the variational form of (2) is given by

$$\left(\frac{\partial u}{\partial t},\nu\right)_\Omega - K\sum_{s=0}^{Q} d_s\left[{}_0D_t^{\frac{1-\beta_s}{2}}(\triangle u,\nu)_\Omega\right] \cong (f,\nu)_\Omega,$$

where $d_s=\frac{1}{2}\varpi\left(\frac{\beta_s+1}{2}\right)\omega_s$ and $(\cdot,\cdot)_\Omega$ denotes the standard L^2-inner product.

We use the eigenfunctions of the regular fractional Sturm-Liouville problem (FSLP) as trial and test basis functions in time [7]. The eigenfunctions of the first and second kind FSLP on [-1,1] are obtained respectively

$${}^{(1)}\rho_k^{\alpha,\beta,\mu}(\tau) = (1+\tau)^{-\beta+\mu-1}P_k^{\alpha-\mu+1,-\beta+\mu-1}(\tau),\ -1\le\alpha<2-\mu,\ -1\le\beta<\mu-1,$$
$${}^{(2)}\rho_k^{\alpha,\beta,\mu}(\tau) = (1-\tau)^{-\alpha+\mu-1}P_k^{-\alpha+\mu-1,\beta-\mu+1}(\tau),\ -1\le\alpha<1-\mu,\ -1\le\beta<2\mu-1.$$

We employ the fractional eigenfunctions for $\alpha=\beta=-1$

$$\phi_k^\mu(\tau) = (1+\tau)^\mu P_{k-1}^{-\mu,\mu}(\tau),\qquad \tau\in[-1,1],$$
$$\varphi_k^\mu(\tau) = (1-\tau)^\mu P_{k-1}^{\mu,-\mu}(\tau),\qquad \tau\in[-1,1],$$

as trial and test basis functions.

In the following, we examine the structure of operational temporal matrices for $1 \leq m, n \leq N$.

$$\mathbf{M}^{\mu}_{mn} = (\phi^{\mu}_{n}(t)\,,\, \varphi^{\mu}_{m}(t)\,)_{I}, \tag{3}$$

$$\mathbf{C}^{\mu}_{mn} = (\frac{d}{dt}\phi^{\mu}_{n}(t)\,,\, \varphi^{\mu}_{m}(t)\,)_{I}, \tag{4}$$

$$\mathbf{S}^{\mu}_{mn} = \sum_{s=0}^{Q} d_s \left({}^{RL}_{0}D_{t}^{\frac{1-\beta_s}{2}} \phi^{\mu}_{n}(t) \;,\; {}^{RL}_{t}D_{T}^{\frac{1-\beta_s}{2}} \varphi^{\mu}_{m}(t) \right)_{I}. \tag{5}$$

Remark 2.1. It must be noted that, the weighted inner products in (3), (4) and (5) can be computed exactly by Gauss-quadrature formulas because the integrands are polynomials.

Let the physical domain Λ be partitioned in to n_e conforming non-overlapping elements $\Lambda_e,\ e = 1, \ldots, n_e$ such that $\Lambda = \bigcup_{e=1}^{n_e} \Lambda_e$. We define the local spatial basis functions on each element such that the C_0-continuty condition of basis in the interfaces of elements is stablished. For this purpose, we use the following modal basis function on the reference element $[-1, 1]$ [5].

$$\psi_p(\xi) = \begin{cases} \frac{1-\xi}{2}, & p = 0, \\ \left(\frac{1-\xi}{2}\right)\left(\frac{1+\xi}{2}\right) P^{1,1}_{p-1}(\xi), & 0 < p < P, \\ \frac{1+\xi}{2}, & p = P. \end{cases}$$

In the following the operational matrices on the reference element $[-1, 1]$ are computed. So that local mass and stiffness matrices on each element Λ_e are obtained as follow

$$\begin{aligned} \mathbf{M}^{(e)}_{mn} &= J^{(e)} \int_{-1}^{1} \psi_n(\xi)\, \psi_m(\xi) d\xi, \\ \mathbf{S}^{(e)}_{mn} &= \frac{1}{J^{(e)}} \int_{-1}^{1} \frac{d}{d\xi}\psi_n(\xi)\, \frac{d}{d\xi}\, \psi_m(\xi) d\xi, \end{aligned}$$

in which $J^{(e)} = \frac{x_e - x_{e-1}}{2}$.

Remark 2.2. The entries of these matrices can be computed exactly using the orthogonal property of the Jacobi polynomials [3].

The elements of load vector $\mathbf{F}^{(e)}$ on each element Λ_e is obtained as follows

$$\mathbf{F}^{(e)}_{mn} = (f(x,t),\ \psi_n(x)\ \varphi^{\mu}_{m}(t))_{\Lambda_e \times I}.$$

Let u_M be the approximate solution of problem in $\Lambda_e \times I$. The fully discrete weak form in a matrix form is given by

$$\left(\mathbf{C}^{\mu} \otimes \mathbf{M}^{(g)} + \mathbf{S}^{\mu} \otimes \mathbf{S}^{(g)}\right) \alpha = \mathbf{F}^{(g)},\ in\ 1D, \tag{6}$$

$$\left(\mathbf{C}^{\mu} \otimes \left(\mathbf{M}^{(g)} \otimes \mathbf{M}^{(g)}\right) + \mathbf{S}^{\mu} \otimes \left(\mathbf{S}^{(g)} \otimes \mathbf{M}^{(g)} + \mathbf{M}^{(g)} \otimes \mathbf{S}^{(g)}\right)\right) \alpha = \mathbf{F}^{(g)},\ in\ 2D, \tag{7}$$

where α is vector of unknown expansion coefficients and $\mathbf{M}^{(g)}$ and $\mathbf{S}^{(g)}$ are global mass and stiffness matices, respictively. Also, $\mathbf{F}^{(g)}$ is global version of $\mathbf{F}^{(e)}$.

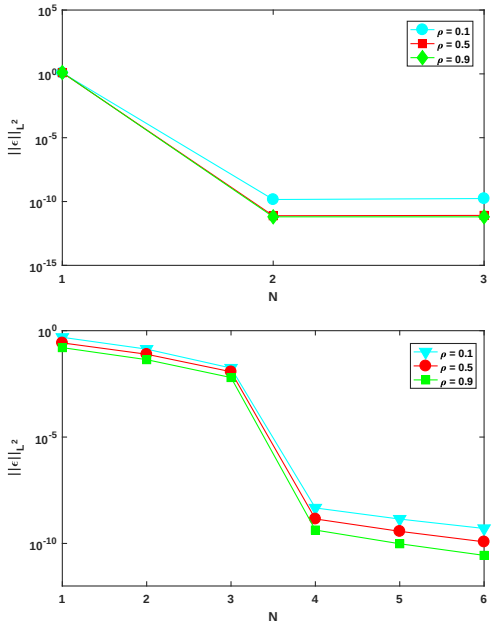


FIGURE 1. L^2 norm of error for Example 3.1 with respect to temporal refinement for (left) $p = 1$, (right) $p = 3$ with $n_e = 2$ and $M = 10$, various ρ.

It must be noted that, (6) and (7) are a kind of Sylvester matrix equation. So, we can solve them by the proposed algorithm in [4] which employs QZ factorization to structure the equation in such a way that it can be solved column-wise by a back substitution technique.

3. Numerical Results

EXAMPLE 3.1. In this example, we consider problem (2) with the exact solution

$$u(x,t) = t^{p+\rho}\cos(\pi x),$$

where ρ is singularity order of solution, and p is an integer value. Also, the distributed weight function is taken $\varpi(\gamma) = \frac{\Gamma(\gamma+p+\rho)}{\pi^2\Gamma(1+p+\rho)}$ and source function $f(x,y,t)$ is defined accordingly.

In this case, we can select the fractional parameter of temporal basis, such that the singularity order of the solution is accurately captured. When we set $\mu = \rho$, a few number of temporal basis functions are needed to achieve exponential convergence.

In the left graph of Figure 1, the L^2 norms of error are presented for different values of N and ρ with $M = 10$, $p = 1$. In the right graph of Figure 1, the L^2 norms of error are presented for different values of N with $M = 10$, $p = 3$.

EXAMPLE 3.2. We consider Eq. (2) in $2D$. The analytical solution is

$$u(x, y, t) = t^2 \sin(\pi x) \sin(\pi y).$$

Also, the distributed weight function is taken $\varpi(\gamma) = \Gamma(\gamma+2)$ and source function $f(x, y, t)$ is defined accordingly.

In Table 1, the numerical solution's errors are presented with L^∞, L^2 norms for the different values of μ, N, and M. It is observed that the choice of μ has an essential effect on the accuracy of the scheme. It can be concluded that in all examples with smooth solutions in term of the time variable (especially polynomial in time), the exponential convergence is recovered when $\mu \to 1$. In this case, the temporal basis also tends to a polynomial, and higher smoothness is achieved. Table 1 reports a comparison between the used errors obtained by the meshless Galerkin method and the present method with $n_e = 4$ for solving two-dimensional fractional modified distributed-order anomalous sub-diffusion equations.

TABLE 1. L^∞ and L^2 errors for Example 3.2 with $n_e = 4$ and different values of μ, N, and M.

(N,M)	Present Method						Method of [1]	
	$\mu = 0.2$		$\mu = 0.8$		$\mu = 1 - 10^{-15}$		$h = \tau = 0.01$	
	L^2	L^∞	L^2	L^∞	L^2	L^∞	L^2	L^∞
(2,5)	5.6e-1	5.6e-2	2.05e-1	2.05e-2	1.78e-3	2.01e-4	-	-
(2,10)	5.6e-1	5.6e-2	2.05e-1	2..05e-2	1.41e-9	1.41e-10	-	-
(3,5)	1.31e-2	1.46e-3	2.97e-2	3.11e-3	1.78e-3	2.01e-4	-	-
(3,10)	1.28e-2	1.28e-3	2.95e-2	2.94e-3	2.13e-9	2.13e-10	2.54 e-5	1.80e-6

References

1. M. Abbaszadeh and M. Dehghan, *Numerical investigation of reproducing kernel particle Galerkin method for solving fractional modified distributed-order anomalous sub-diffusion equation with error estimation*, Appl. Math. Comput. **392** (2021). DOI:10.1016/j.amc.2020.125718
2. F. Fakhar-Izadi, *Fully Petrov-Galerkin spectral method for the distributed-order time-fractional fourth-order partial differential equation*, Eng. Comput. (2020). DOI:10.1007/s00366-020-00968-2
3. F. Fakhar-Izadi and M. Dehghan, *A spectral element method using the modal basis and its application in solving second-order nonlinear partial differential equations*, Math. Methods Appl. Sci. **38** (3) (2015) 478–504.
4. J. D. Gardiner, A. J. Laub, J. J. Amato and C. B. Moler, *Solution of the Sylvester matrix equation* $AXB^T + CXD^T = E$, ACM Trans. Math. Software **18** (2) (1992) 223–231.
5. G. E. Karniadakis and S. Sherwin, *Spectral/hp Element Methods for Computational Fluid Dynamics*, 2nd ed., Numerical Mathematics and Scientific Computation, Oxford University Press, New York, 2005.
6. L. Li, F. Liu, L. Feng and I. Turner, *A Galerkin finite element method for the modified distributed-order anomalous sub-diffusion equation*, J. Comput. Appl. Math. **368** (2020). DOI:10.1016/j.cam.2019.112589
7. M. Zayernouri and G. E. Karniadakis, *Fractional Sturm-Liouville eigen-problems: Theory and numerical approximation*, J. Comput. Phys. **252** (2013), 495–517.

E-mail: azam.yazdani@aut.ac.ir
E-mail: ffizadii@gmail.com; f.fakhar@aut.ac.ir

Efficient Determination of Regularization Parameter in Tikhonov-Type Regularization of Discrete Ill-Posed Problems

Hossein Zare*
Department of Mathematical Sciences, Tarbiat Modares University, Tehran, Iran
and Masoud Hajarian
Faculty of Mathematical Sciences, Shahid Beheshti University, General Campus, Evin, Tehran 19839, Iran

Abstract. This paper presents a new approach for choosing an appropriate regularization parameter in Tikhonov-type regularization of discrete ill-posed problems. Using the basic concepts of multi-objective optimization, we derive a single-objective problem that its minimizer gives an appropriate estimation of the regularization parameter. The numerical efficiency of the presented method is compared with the L-curve and the GCV parameter choice methods.
Keywords: Multi-objective optimization, Regularization parameter, Tikhonov regularization.
AMS Mathematical Subject Classification [2010]: 65F22, 90C29.

1. Introduction

Many practical problems give rise to a linear system of equations of the form

$$Ax \approx b,\ A \in \mathbb{R}^{m\times n},\ m \geq n,\ b \in \mathbb{R}^m, \tag{1}$$

where A is an ill-conditioned matrix whose singular values "cluster" at the origin, and the vector b is contaminated by an unknown error. Such systems are commonly referred to as discrete ill-posed problems, because they usually stem from the discretization of ill-posed problems such as Fredholm integral equations of the first kind [2]. Tikhonov regularization is a popular approach to obtain a meaningful approximate solution of such problems. In this method, the linear system $Ax \approx b$ (or the linear least squares problem associated with it) is replaced by the minimization problem

$$\min_{x\in\mathbb{R}^n} \mathbf{J}_\lambda(x) = \|Ax-b\|^2 + \lambda\|Lx\|^2, \tag{2}$$

where $L \in \mathbb{R}^{k\times n}(k \leq n)$, is referred to as a regularization matrix and the scalar $\lambda > 0$ as a regularization parameter. The problem (2) is called *regularized least squares* (RLS) and its objective function is called Tikhonov functional.

We assume that $\mathcal{N}(A) \cap \mathcal{N}(L) = \{0\}$, where $\mathcal{N}$ denotes the null space of matrices. Then the problem (2) has the unique solution $x_\lambda = (A^TA + \lambda L^TL)^{-1}A^Tb$,

*Speaker

for any $\lambda > 0$. Typically, L is chosen as the identity matrix or a discrete approximation of the first or second order derivative operators. The regularization parameter λ plays an important role in computing a reliable solution x_λ. A proper choice of the regularization parameter λ is critical since if λ is too small, then x_λ is very close to the solution of original ill-posed problem. On the other hand, if λ is too large, then the connection between the problem (2) and the original problem (1) will be reduced. In this paper, we present an efficient method for choosing an appropriate regularization parameter. This method utilizes the basic concepts of multi-objective optimization. The main idea of this method is to scale the residual norm $\|Ax_\lambda - b\|^2$ and the penalty term $\|Lx_\lambda\|^2$ by a suitable method and then minimizing the sum of them.

2. RLS from Multi-Objective Optimization Point of View

In this section, we investigate the RLS problem from multi-objective optimization point of view. We refer the reader to [1, 4] for more details related to multi-objective optimization problems.

A multi-objective optimization problem can be formulated as

$$\min f(x) = [f_1(x), f_2(x), \ldots, f_k(x)]^T \quad \text{such that} \quad x \in \mathcal{S},$$

where $\mathcal{S} \subseteq \mathbb{R}^n$ is a non-empty set and $f_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, k$, are the conflicting objective functions to be minimized simultaneously. The set $\mathcal{S}$ is called the *feasible region* and the set

$$\mathcal{Z} = \{f(x) | x \in \mathcal{S}\} \subseteq \mathbb{R}^k,$$

is called the *feasible objective region*. Given two objective vectors z and z', we say that $z \leq z'$ if and only if $z_i \leq z'_i$ for all $i = 1, 2, \ldots, k$. Also, for given vectors x, x', x^* in the feasible region $\mathcal{S}$, we say that:

- $x \preceq x'$ (x weakly dominates x') if and only if $f(x) \leq f(x')$.
- $x \prec x'$ (x dominates x') if and only if $x \preceq x'$ and at least one component of $f(x)$ is strictly less than the corresponding one of $f(x')$.
- $x \sim x'$ (x is indifferent to x') if neither x dominates x' nor x' dominates x.
- x^* is a *Pareto minimizer* of f, if there is no other point in $\mathcal{S}$ that dominates it.

The set of all Pareto minimizers of f is denoted by $\mathcal{P}$. The *ideal objective vector* is defined as $z^* = (z_1^*, \ldots, z_k^*)$, where $z_i^* = \min f_i(x)$ and $x \in \mathcal{S}$. In general, because of conflicts among the objectives, $z^* \notin \mathcal{Z}$. However, it can be used as a reference point. Finally, the *nadir objective vector* is defined as $z^{\text{nad}} = (z_1^{\text{nad}}, \ldots, z_k^{\text{nad}})$ where $z_i^{\text{nad}} = \max f_i(x)$ and $x \in \mathcal{P}$.

The problem of finding an appropriate regularization parameter can be considered as a multi-objective optimization problem with two objectives. In fact, our purpose is to find a regularization parameter $\lambda^* > 0$ such that x_{λ^*} makes both $\|Ax - b\|$ and $\|Lx\|$ small, to the extent possible. This idea yields the multi-objective optimization problem

$$\min_{\lambda > 0} g(\lambda) = \begin{bmatrix} g_1(\lambda) \\ g_2(\lambda) \end{bmatrix}, \tag{3}$$

where $g_1(\lambda) = \|Ax_\lambda - b\|$, $g_2(\lambda) = \|Lx_\lambda\|$ and $x_\lambda = (A^TA + \lambda L^TL)^{-1}A^Tb$. The following theorem gives the lower and the upper bounds of the two objectives $g_1(\lambda)$ and $g_2(\lambda)$. They will be applied later to normalize the objectives $g_1(\lambda)$ and $g_2(\lambda)$ so that their values are of approximately the same magnitude.

THEOREM 2.1. [6, Theorem 2] *Given two functions $g_1(\lambda) = \|Ax_\lambda - b\|$ and $g_2(\lambda) = \|Lx_\lambda\|$, the following statements hold:*

a) *for $0 < \lambda_1 < \lambda_2$ we have $g_1(\lambda_1) < g_1(\lambda_2)$ and $g_2(\lambda_2) < g_2(\lambda_1)$. In other words, g_1 is a strictly increasing function of λ whereas g_2 is a strictly decreasing function of λ.*
b) *$g_1(\lambda) \to \|Ax_{LS} - b\|$ and $g_2(\lambda) \to \|Lx_{LS}\|$ as $\lambda \to 0$.*
a) *$g_2(\lambda) \to 0$ and $g_1(\lambda) \to \left\|(A(AP_L)^\dagger - I)b\right\|$ as $\lambda \to \infty$, where $P_L = I - L^\dagger L$ and † denotes the Moore–Penrose pseudoinverse.*

From the above theorem, we see that the vectors

$$[\|Ax_{LS} - b\|, 0]^T \quad \text{and} \quad \left[\left\|(A(AP_L)^\dagger - I)b\right\|, \|Lx_{LS}\|\right]^T$$

are the ideal objective vector and the nadir objective vector for the problem (3), respectively. It can be easily shown that composing each objective of the problem (3) with an strictly increasing function, does not change the set of its Pareto minimizers, as is described in the following theorem.

THEOREM 2.2. *Let*

$$\hat{g} = \begin{bmatrix} T_1(g_1) \\ T_2(g_2) \end{bmatrix},$$

where $T_1, T_2 : \mathbb{R} \to \mathbb{R}$ are two strictly increasing real functions. Then $\tilde{\lambda}$ is a Pareto minimizer of $\hat{g}$ if and only if it is a Pareto minimizer of g. In other words the two problems $\min_{\lambda>0} g(\lambda)$ and $\min_{\lambda>0} \hat{g}(\lambda)$ are equivalent.

One general approach for solving a multi-objective optimization problem such as (3), is converting the problem to the single-objective problem

$$\min_{\lambda>0} \|g(\lambda)\|_p^p = g_1(\lambda)^p + g_2(\lambda)^p, \qquad p > 0, \tag{4}$$

and then solving this problem by using a standard optimization method. It can be easily seen that if $\tilde{\lambda}$ is a global minimizer for the single-objective problem (4), then $\tilde{\lambda}$ is a Pareto minimizer for the multi-objective problem (3). The problem (4) is useful for our purpose, because it minimizes the distance between the feasible objective region of the problem (3) and the reference point $(0, 0)$, which is near to the ideal objective vector.

3. Description of The New Technique and Numerical Examples

We note that when the regularization parameter λ is chosen very small, the value of $\|Lx_\lambda\|$ is very large whereas the value of $\|Ax_\lambda - b\|$ is very small. Moreover, when λ is gradually increasing, the value of $g_1(\lambda) = \|Ax_\lambda - b\|$ increases slowly, while the value of $g_2(\lambda) = \|Lx_\lambda\|$ rapidly tends to zero. Also, the functions g_1 and g_2 have very different ranges. Therefore, it is not easy to minimize

$$g(\lambda) = \begin{bmatrix} g_1(\lambda) \\ g_2(\lambda) \end{bmatrix},$$

without suitable transformations. To overcome these difficulties, we firstly introduce an increasing function $T : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}$ to reduce the range of the functions g_1, g_2 at the same time. In [6], the authors suggested the use of $T(x) = \arctan(x)$. Another possible choice for T is $T(x) = 1/(1+e^{-x})$ with the range $[1/2, 1)$. Hence, we consider the following function:

$$\hat{g}(\lambda) = \begin{bmatrix} \hat{g}_1(\lambda) \\ \hat{g}_2(\lambda) \end{bmatrix} = \begin{bmatrix} T(g_1(\lambda)) \\ T(g_2(\lambda)) \end{bmatrix}.$$

According to Theorem 2.2, the problems $\min_{\lambda>0} g(\lambda)$ and $\min_{\lambda>0} \hat{g}(\lambda)$ are equivalent. In the next step, we scale $\hat{g}_1$ and $\hat{g}_2$ such that the scaled functions have a same range of values. The scaled functions are given by

$$\hat{g}_s(\lambda) = \begin{bmatrix} S(\hat{g}_1) \\ S(\hat{g}_2) \end{bmatrix},$$

where S is the following scaling function

$$S(\hat{g}_i(\lambda)) = \frac{\hat{g}_i(\lambda) - \min \hat{g}_i}{\max \hat{g}_i - \min \hat{g}_i},$$

and $\min \hat{g}_i$ and $\max \hat{g}_i$ are lower and upper bounds, respectively, for the objective function $\hat{g}_i$. Note that by using this scaling, the values of $\hat{g}_i$ lie in the interval $[0, 1]$. Again, according to Theorem 2.2, the two problems $\min_{\lambda>0} g(\lambda)$ and $\min_{\lambda>0} \hat{g}_s(\lambda)$ are equivalent. Finally, we choose the regularization parameter λ^* by finding the global minimizer of problem $\min_{\lambda>0} \|\hat{g}_s(\lambda)\|_p^p$. Let $T(x) = 1/(1 + e^{-x})$ and $p = 1$. Then we need to minimize the following function

$$E(\lambda) = \frac{T(g_1(\lambda)) - 1/2}{\max T(g_1(\lambda)) - 1/2} + \frac{T(g_2(\lambda)) - 1/2}{\max T(g_2(\lambda)) - 1/2}. \tag{5}$$

Notice that $\min T(g_1(\lambda))$ and $\min T(g_2(\lambda))$ are considered as $1/2$ in the above equation, because according to Theorem 2.1, $g_1(\lambda) \to \|Ax_{LS} - b\| \approx 0$ and $g_2(\lambda) \to 0$, as $\lambda \to 0$ and $\lambda \to \infty$, respectively. To minimize the objective function (5), we need the values of $\max T(g_1(\lambda))$ and $\max T(g_2(\lambda))$. Since g_1 is a strictly increasing function of λ, by considering Theorem 2.1, the value of $\max T(g_1(\lambda))$ can be approximated by evaluating $T(g_1(\lambda))$ at a sufficiently big number. In a similar way, since g_2 is a strictly decreasing function of λ, one may get an approximation of $\max T(g_2(\lambda))$ by evaluating $T(g_2(\lambda))$ at a sufficiently small number. One of the best characteristics of the proposed objective function (5) is its uni-modality over a sufficiently big interval containing the minimizer. Thus, the minimizer of (5) can be easily obtained by using a bracketing search procedure such as *golden section search* method [5].

The following table gives a sample comparison of the L-curve method, GCV method and presented method for three test problems taken from the Hansen's "Regularization tools" package [3]. For each problem, we generate the noise contaminated vector b as $b \leftarrow b + \sigma(e/\|e\|)$, where the elements of the noise e are created by the MATLAB "randn" function and σ is the level of the noise. We use the noise levels $\sigma = 10^{-j}, j = 1, 2, 3$, which are more compatible with real world problems. The regularization matrix L is chosen as the identity matrix I. The dimensions 40, 100, 200 and 400 are selected for each problem. Each experiment

has been executed 10 times, and the average of relative errors in the computed solutions has been demonstrated.

TABLE 1. Numerical results for $L = I$.

Problem	Size	L-curve Method			GCV Method			Presented Method		
		$\sigma = 0.1$	$\sigma = 0.01$	$\sigma = 0.001$	$\sigma = 0.1$	$\sigma = 0.01$	$\sigma = 0.001$	$\sigma = 0.1$	$\sigma = 0.01$	$\sigma = 0.001$
FOXGOOD	40	0.131	0.162	0.066	40.853	6.866	0.039	0.077	0.058	0.018
	100	0.056	0.047	0.044	0.055	0.312	0.021	0.050	0.035	0.011
	200	0.049	0.020	0.019	0.062	0.052	0.103	0.047	0.017	0.007
	400	0.027	0.014	0.014	0.101	0.022	0.115	0.027	0.006	0.003
PHILLIPS	40	0.184	0.277	0.211	5.840	0.209	0.137	0.163	0.076	0.023
	100	0.087	0.243	0.379	0.041	0.179	1.777	0.082	0.054	0.018
	200	0.097	0.111	0.188	0.208	0.050	0.014	0.092	0.031	0.011
	400	0.058	0.093	0.108	0.145	0.037	0.005	0.055	0.025	0.006
SHAW	40	0.161	0.107	0.055	0.167	0.237	7.793e+6	0.153	0.059	0.045
	100	0.112	0.091	0.054	0.129	6.770e+4	2.051	0.111	0.061	0.040
	200	0.058	0.051	0.039	0.235	0.048	0.121	0.056	0.045	0.031
	400	0.047	0.066	0.041	0.069	1.345	0.037	0.045	0.059	0.033

4. Conclusion

We presented a new method based on multi-objective optimization to find an appropriate value of the Tikhonov regularization parameter. This method does not require any information about the error in the given right-hand side and can be applied to a wide variety of discrete ill-posed problems. As numerical examples show, the presented method, in comparison with the two other methods, generally gives smaller errors on average.

References

1. K. Deb, *Multi-Objective Optimization Using Evolutionary Algorithms*, John Wiley & Sons, New York, 2001.
2. P. C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems*, Society for Industrial and Applied Mathematics, Philadelphia, 1998.
3. P. C. Hansen, *Regularization tools version* 4.0 *for MATLAB* 7.3, Numer. Algor. **46** (2007) 189–194.
4. K. Miettinen, *Nonlinear Multiobjective Optimization*, Springer Science & Business Media, New York, 2012.
5. J. A. Snyman, *Practical Mathematical Optimization*, Springer Science & Business Media, New York, 2005.
6. H. Zare and M. Hajarian, *Determination of regularization parameter via solving a multi-objective optimization problem*, Appl. Numer. Math. **156** (2020) 542–554.

E-mail: hossein.zare@modares.ac.ir
E-mail: m_hajarian@sbu.ac.ir

A Direct Method for Solving a Class of Volterra Functional Equations

Elham Zeynal*
Young Researcher and Elite Club, Yadegar-e-Imam Khomeini (RAH) Shahr-e-Rey Branch, Islamic Azad University, Tehran, Iran
and Esmail Babolian
Faculty of Mathematical Sciences and Computer, Kharazmi University, Tehran, Iran

Abstract. In this paper, we propose direct method to solve a class of Volterra delay-integro-differential equations (VDIDEs) based on vector forms of Block-Pulse Functions (BPFs). Operational matrix of integration of BPFs is applied to transform a VDIDE to a linear set of algebraic equations.
Keywords: Volterra delay-integro-differential equations, Block-Pulse Functions, Direct method.
AMS Mathematical Subject Classification [2010]: 65R20, 45D05, 34K06.

1. Introduction

Volterra delay-integro-differential equations have various applications in different sciences like biology, ecology, medicine and physics and etc. [2, 3] and [5]. The basic idea in this paper is using BPFs for solving the following VDIDE

$$(1)\qquad \begin{array}{ll} y^{'}(x) = f(x, y(x-\tau), \int_{x-\tau}^{x} k(s,x,y(s))\,ds), & x \geqslant 0, \\ y(x) = \psi_(x), & x \leqslant 0, \end{array}$$

where initial function ψ is known.

The approach simplifies the VDIDE to a set of algebraic equations. To do this, we expand the unknown function with respect to Block-Pulse functions. The operational matrix of integration and products are used to evaluate the coefficients of BPFs for the solution of Eq. (1). However, we do not use the operational matrix of delay. The paper is organized as follows: a review of BPFs and its application for resolving Eq. (1) is given in Section 2. In Section 3, an example is illustrated to show the accuracy of the method and finally, the convergence analysis is given in Section 4.

2. Block-Pulse Functions

The choice of basis functions is one of the most important steps in any numerical solution. Block-Pulse functions have been used for various problems; for instance, see [1, 6].

*Speaker

2.1. Properties of BPFs. BPFs are defined on [0,1) as

$$\varphi_i(x) = \begin{cases} 1, & x \in [ih, (i+1)h), \\ 0, & elsewhere, \end{cases}$$

for $i = 0, 1, \ldots, m-1$, a positive integer m and $h = \frac{1}{m}$. We refer to properties of the BPFs as disjointness, orthogonality, completeness and partition of unity, respectively. For disjointness

$$\varphi_i(x)\varphi_j(x) = \begin{cases} \varphi_i(x), & i = j, \\ 0, & i \neq j, \end{cases} \tag{2}$$

where $i, j = 0, 1, \ldots, m-1$. For orthogonality, we have

$$\int_0^1 \varphi_i(x)\varphi_j(x)\, dx = h\delta_{ij},$$

where δ_{ij} is the Kronecker delta.

Finally, for every function f in $\mathcal{L}^2([0,1))$, Parseval's identity holds

$$\int_0^1 |f(x)|^2\, dx = \sum_{i=0}^{\infty} f_i^2 \| \varphi_i(x) \|^2,$$

where

$$f_i = \frac{1}{h}\int_0^1 f(x)\varphi_i(x)\, dx. \tag{3}$$

Also, from the definition of BPFs, these functions form a partition of unity, i.e.

$$\sum_{i=0}^{m-1} \varphi_i(x) = 1. \tag{4}$$

2.2. Vector Forms. An m-vector of BPFs on [0,1) is presented by

$$\Phi(x) = [\varphi_0(x), \varphi_1(x), \ldots, \varphi_{m-1}(x)]^T,$$

where T stands for transpose.

Using relations (2) and (4) for all $x \in [0,1)$ we have

$$\Phi(x)\Phi^T(x) = \begin{pmatrix} \varphi_0(x) & 0 & \cdots & 0 \\ 0 & \varphi_1(x) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \varphi_{m-1}(x) \end{pmatrix}_{m\times m},$$

$$\Phi^T(x)\Phi(x) = 1,$$

$$\Phi(x)\Phi^T(x)W = \tilde{W}\Phi(x),$$

where W is an m-vector and $\tilde{W} = diag(W)$. In addition, one can conclude that for each $m \times m$ matrix V

$$\Phi^T(x)V\Phi(x) = \hat{V}^T\Phi(x),$$

where $\hat{V}$ is a column vector consisting of the diagonal components of the matrix V.

2.3. Expansion of Functions by BPFs. An arbitrary function $f \in \mathcal{L}^2([0,1))$ can be approximated as

$$f(x) \simeq \sum_{i=0}^{m-1} f_i \varphi_i(x) = F^T \Phi(x) = \Phi^T(x) F, \tag{5}$$

where $F = [f_0, f_1, \ldots, f_{m-1}]^T$ and for $i = 0, 1, \ldots, m-1$, the coefficient f_i is defined by (3). Also, for two square integrable functions f and h, using $h(x) \simeq \Phi^T(x)H$ we have

$$f(x)h(x) \simeq F^T \Phi(x) \Phi^T(x) H = F^T \tilde{H} \Phi(x),$$

where $\tilde{H} = diag(H)$.

2.4. Operational Matrix of Integration. The integration of the function f defined by (5) can be approximately obtained as

$$\int_0^x f(s)\, ds \simeq \int_0^x F^T \Phi(s)\, ds \simeq F^T P \Phi(x), \tag{6}$$

where P is operational matrix of integration of BPFs as

$$P = \frac{h}{2} \begin{pmatrix} 1 & 2 & 2 & \cdots & 2 \\ 0 & 1 & 2 & \cdots & 2 \\ 0 & 0 & 1 & \cdots & 2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}_{m \times m}.$$

3. Considering an Example

Example 3.1. Given the following VFE

$$\begin{array}{ll} y'(x) = y(x-1) + \int_{x-1}^{x} y(s)\, ds, & x \geqslant 0, \\ y(x) = e^x, & x \leqslant 0, \end{array} \tag{7}$$

with the analytical solution $y(x) = e^x$.

Assume

$$y(s) = x(s) + z(s), \tag{8}$$

for $s \in [x-1, 0]$, $x(s)$ is initial function $\psi(s)$ and for $s \in [0, x]$, $z(s)$ is the solution of Eq. (7). We substitute relation (8) into (7) and integrate Eq. (7) from 0 to x we have

$$y(x) - y(0) = \int_0^x y(x-1)\, dx + \int_0^x (1 - e^{x-1})\, dx + \int_0^x \int_0^x y(s)\, ds\, dx. \tag{9}$$

In the sequent, approximating $y(x), y(0), y(x-1)$ and $f(x)$ with respect to BPFs using (6), gives

$$\begin{cases} y(x) \simeq Y^T \Phi(x) = \Phi^T(x), \\ 5y(0) \simeq Y_0^T \Phi(x) = \Phi^T(x) Y_0, \\ y(x-1) \simeq \Psi^T \Phi(x) = \Phi^T(x) \Psi, \\ f(x) \simeq F^T \Phi(x) = \Phi^T(x) F, \end{cases} \tag{10}$$

where m-vectors Y, Y_0, Ψ and F are BPFs coefficients of $y(x), y(0), y(x-1)$ and $f(x)$, respectively and $Y_0 = [y_0, y_0, \ldots, y_0]^T$ with $y_0 = y(0)$. We substitute relations (10) to (9) and use the property of (6) to obtain

$$Y^T\Phi(x) - Y_0^T\Phi(x) \simeq \psi^T P\Phi(x) + F^T P\Phi(x) + Y^T P^2\Phi(x),$$

then, we use orthogonality of BPFs and replace $\simeq$ by $=$ to obtain

$$Y - Y_0 = P^T\psi + P^T F + (P^2)^T Y. \tag{11}$$

Relation (11) is a set of linear algebraic equations for the unknown coefficients vector Y. Therefore, $y(x) \simeq Y^T\Phi(x)$ is an approximate solution that can be calculated to Eq. (7).

Table 1 demonstrates the absolute errors computed using BPFs with $m = 10, 100, 1000$ and compared with the Adomian method [4].

TABLE 1. Obtained results of $y(x)$ for VFE.

x	Adomian	BPFs $m = 10$	BPFs $m = 100$	BPFs $m = 1000$
0	0	5.2×10^{-2}	5.0×10^{-3}	5.0×10^{-4}
0.2	2.0×10^{-2}	6.4×10^{-2}	6.1×10^{-3}	6.0×10^{-4}
0.4	8.0×10^{-2}	7.9×10^{-2}	7.5×10^{-3}	7.0×10^{-4}
0.6	1.8×10^{-1}	9.6×10^{-2}	9.2×10^{-3}	9.0×10^{-4}
0.8	3.2×10^{-1}	1.1×10^{-1}	1.1×10^{-2}	1.1×10^{-3}
1	1.3×10^{-1}	1.5×10^{-2}	2.0×10^{-4}	0
1.2	2.2×10^{-1}	4.4×10^{-2}	2.9×10^{-3}	3.0×10^{-4}
1.4	3.4×10^{-1}	7.5×10^{-2}	5.8×10^{-3}	6.0×10^{-4}
1.6	4.9×10^{-1}	1.0×10^{-1}	8.9×10^{-3}	9.0×10^{-4}
1.8	6.9×10^{-1}	1.4×10^{-1}	1.2×10^{-2}	1.2×10^{-3}
2	3.3×10^{-1}	1.5×10^{-1}	2.0×10^{-2}	2.1×10^{-3}
2.2	4.7×10^{-1}	1.3×10^{-1}	1.9×10^{-2}	2.0×10^{-3}
2.4	6.4×10^{-1}	1.1×10^{-1}	1.8×10^{-2}	1.9×10^{-3}
2.6	8.5×10^{-1}	9.9×10^{-2}	1.7×10^{-2}	1.8×10^{-3}
2.8	1.1	8.1×10^{-2}	1.6×10^{-2}	1.8×10^{-3}
3	1.2	7.2×10^{-2}	1.6×10^{-2}	1.8×10^{-3}

4. Convergence Analysis

THEOREM 4.1. [6] *Let $I = [0,1)$ and $f_m(t) = F^T\Phi(t)$ be the BPF approximation of $f \in C^1(I)$, where $F = [f_0, f_1, \ldots, f_{m-1}]^T$ and f_i is defined by (3) and suppose that there exists a positive number M such that*

$$|f^{'}(t)| \leq M, \qquad t \in [0,1).$$

Then

$$\| f - f_m \|_{\infty} \leq Mh.$$

We have $f(x) - f_m(x) = O(h)$ as a result $\int_0^s f(x)\,dx - \int_0^s f_m(x)\,dx = O(h)$. By integrating both sides of Eq. (1), we conclude $y(x) - y_m(x) = O(h)$.

References

1. E. Babolian, Z. Masouri and S. Hatamzadeh-Varmazyar, *New direct method to solve nonlinear Volterra-Fredholm integral and integro-differential equations using operational matrix with block-pulse functions*, Prog. Electromagn. Res. B. **8** (2008) 59–76.
2. G. A. Bocharov and F. A. Rihan, *Numerical modelling in biosciences with delay differential equations*, J. Comput. Appl. Math. **125** (2000) 183–199.
3. H. Brunner and P. J. van der Houwen, *The Numerical Solution of Volterra Equations*, In: CWI Monographs, Vol. 3, North-Holland, Amsterdam, 1986.
4. D. J. Evans and K. R. Raslan, *The Adomian decomposition method for solving delay differential equation*, Int. J. Comput. Math. **82** (2005) 49–54.
5. A. Jerri, *Introduction to Integral Equations with Applications*, Wiley, New York, 1999.
6. E. Zeynal, E. Babolian and T. Damercheli, *Direct methods for solving time-varying delay systems*, J. Math. Sci. **14** (2020) 159–166.

E-mail: Elzeynal@gmail.com
E-mail: babolian@khu.ac.ir

Contributed Talks

Optimization

Relaxation Method to Estimate the Nondominated Frontier of the Biobjective Quadratic Optimization Problems

Seyed Morteza Mirdehghan
Department of Mathematics, Shiraz University, Shiraz, Iran
and Diba Aminshayan Jahromi*
Department of Mathematics, Shiraz University, Shiraz, Iran

Abstract. Finding the nondominated frontier of multiobjective optimization problems is an interesting research subject for some researchers. In recent years, various researches have been conducted on finding the bounds of objective functions in quadratic optimization problems using copositive relaxation. These researches have been focused on single objective quadratic optimization problems. In this manuscript, we propose an approach to estimate a piece-wise linear nondominated frontier of the nondominated frontier of biobjective quadratic optimization problems with quadratic and linear constraints using copositive relaxation.

Keywords: Copositive optimization, Biobjective optimization, Quadratic optimization, Piece-wise linear nondominated frontier.
AMS Mathematical Subject Classification [2010]: 90C20, 90C29.

1. Introduction

The concept of copositivity was first introduced by Motzkin in 1952. After that, many articles have been published on this concept in optimization, which are focused on single-objective problems. Besides, there are many NP-hard problems that cannot be solved easily, so, the researchers try to approximate the solutions using some innovative methods. One of these methods is copositive relaxation which was introduced by Bomze in 2015 for quadratic optimization problems under quadratic and linear constraints [1]. Bomze has shown that a tight bound for these kind of problems can be identified using copositive relaxation. To extend this subject, we use the copositive relaxation for biobjective quadratic optimization problems with quadratic and linear constraints. In this regards, we propose a method to approximate the nondominated frontier of the biobjective quadratic optimization problem by a piece-wise linear map.

1.1. Preliminaries. We abbreviate the integer scalars between two integers m, n with $m \leq n$ by $[m:n] := m, m+1, \ldots, n$. The positive orthant is denoted by $\mathbb{R}^n_+ := \{x \in \mathbb{R}^n : x_i \geq 0 \ for\ all\ i \in [1:n]\}$, and the positive semidefinite matrix H by $H \succeq O$. The Frobenius duality is defined by $\langle S, X\rangle = trace(SX)$, where S and X are symmetric matrices with the same orders. The Dual of cone D is defined as follows:

$$D^* := \{S = S^T \in \mathbb{R}^{n\times n} : \langle S, X\rangle \geq 0 \ for\ all\ X \in D\},$$

*Speaker

where D is a cone of symmetric $n \times n$ matrices. For a symmetric $n \times n$ matrix Q, Q is copositive if $v^T Q v \geq 0$ for all $v \in \mathbb{R}^n_+$ and Q is strictly copositive if $v^T Q v > 0$ for all $v \in \mathbb{R}^n_+ - \{0\}$. The set of all copositive matrices is a closed and convex cone as follows:

$$C^* = \{Q = Q^T \in \mathbb{R}^{n \times n} : Q \textit{ is copositive}\}.$$

C^* is the dual cone of $C = conv\{xx^T : x \in \mathbb{R}^n_+\}$, where $conv\ S$ denotes the convex hull of set $S \subset \mathbb{R}^n$ [3]. In this manuscript, subscripts $\square_D$, $\square_P$, $\square_C$, $\square_{CD}$ and $\square_{CP}$ are used to refer to the conic dual, the primal conic problem, the co(mpletely) positive problems, the dual problem over the copositive cone, and the primal problem over the completely positive cone, respectively.

1.2. Multiobjective Optimization Problems. A multiobjective optimization problem is an optimization problem that involves multiple objective functions as follows:

$$\begin{array}{ll} \min & (q_1(x), q_2(x), \ldots, q_k(x)) \\ s.t. & x \in X, \end{array} \tag{1}$$

where $k \geq 2$ is the number of objectives and the set X is the feasible set of decision vectors. The feasible set is typically defined by some constraint functions. In addition, the vector-valued objective function is often defined as $q : X \to \mathbb{R}^k$ and $q(x) = (q_1(x), q_2(x), \ldots, q_k(x))^T$. The image of X is denoted by Y. In multi-objective optimization, there does not typically exist a feasible solution that minimizes all objective functions simultaneously. Therefore, attention is paid to Pareto optimal solutions; that is, solutions that cannot be improved in any of the objectives without degrading at least one of the other objectives. Mathematically, a feasible solution $x_1 \in X$ is dominated by another solution $x_2 \in X$, if $f_i(x_1) \leq f_i(x_2)$ for all indices $i \in \{1, 2, \ldots, k\}$ and $f_j(x_1) < f_j(x_2)$ for at least one index $j \in \{1, 2, \ldots, k\}$. A solution $x^* \in X$ (and the corresponding outcome $q(x^*)$) is called efficient, if there does not exist another solution that dominates it. The set of efficient outcomes is often called the efficient front or efficient frontier. For more information see [4].

In this article, we consider $k = 2$ and the objective functions are quadratic with quadratic and linear constraints. In the next section, we introduce these problems in details. In these problems, we don't have any efficient algorithm to find the efficient points and nondominated frontier. In [1], an approach to find a tight bound for the objective function of a single objective problem has been presented using copositive matrix and relaxation of that. In this paper, we attempt to extend this approach to biobjective quadratic optimization problems. Moreover, we try to estimate a piece-wise linear frontier of the nondominated frontier using copositive relaxation for biobjective quadratic optimization problems.

2. Copositive Relaxation in Biobjective Quadratic Optimization

Consider the following biobjective quadratic optimization problem:

$$\begin{array}{lll} \inf & (q_{01}(x), q_{02}(x)) & \\ s.t. & q_i(x) \leq 0, & i = 1, 2, \ldots, m, \\ & Ax = b, & \\ & x \geq 0, & \end{array} \tag{2}$$

where all $q_{01}(x)$, $q_{02}(x)$ and $q_i(x) = x^T Q_i x - 2b_i^T x + c_i$ for $i = 1, \ldots, m$ are quadratic functions, $b \in \mathbb{R}^p$ and A is a $p \times n$ matrix of full row rank p. We further impose the Slater condition on the linear constraints as follows:
there is a point $y \in P$ *such that* $y_j > 0$ *for all* $j \in [1:n]$, *where* $P := \{x \in \mathbb{R}^n_+ : Ax = b\}$.

The efficient frontier of (2) cannot be identified easily and clearly regarding to its nonlinear objective functions and constraints. Therefore, we try to estimate its frontier as a piece-wise linear map. For this purpose, at first we consider the following problem:

$$\begin{array}{ll} \min & z = q_{01}(x) \\ s.t. & q_i(x) \leq 0, \\ & Ax = b, \\ & x \geq 0. \end{array} \tag{3}$$

Using copositive relaxation, we consider $q_{01}(x) = x^T Q x - 2b^T x + c$ defined on $\mathbb{R}^n$ and define Shor relaxation matrix as $M(q) := \begin{bmatrix} c & -b^T \\ -b & Q \end{bmatrix}$ and also $J_0 = \begin{bmatrix} 1 & o^T \\ o & O \end{bmatrix}$.

Let $A^T = [r_1^T, r_2^T, \ldots, r_p^T]$, where r_k is the kth row of A. For all $k \in [1:p]$, we define the symmetric matrices of order $n+1$ as $A_k := \begin{bmatrix} 2b_k & -r_k \\ -r_k^T & O \end{bmatrix}$. It is sufficient to find the optimal objective of the following problem:

$$\begin{array}{ll} \min_{x \in C, \langle J_0, X \rangle = 1} & \langle M_0, X \rangle \\ s.t. & \langle M_i, X \rangle \leq 0, \ i \in [1:m], \\ & \langle A_k, X \rangle = 0, \ i \in [1:p], \end{array} \tag{4}$$

where X is a matrix of variables. We denote the optimal objective value of the above problem as z^*_{CP}. By [1, Theorem 4.1] we have $z^*_{CP} \leq z^*_+$. In the next step we should find the optimal objective value of the following problem:

$$\begin{array}{ll} \min & z = q_{02}(x) \\ s.t. & q_i(x) \leq 0, \\ & Ax = b, \\ & q_{01} \leq z^*_{CP1} + k\delta, \\ & x \geq 0, \end{array} \tag{5}$$

where δ is a step size to increase the value of z^*_{CP1} to relax the constraint corresponding to q_{01} in each iteration. To be more precise, we want to find a piece-wise linear estimation for the frontier of the nondominated space. For this estimation we use δ as a step size of it and k is started from zero up to somewhere that z^*_{CP1} remains unchanged. For problem (5), in a similar way, we can find another lower bound like z^*_{CP2} using the copositive relaxation. Using the proposed method we can find adjacent pairs $(z^*_{CP1} + k\delta, z^*_{CP2})$. By considering the convex combination of two adjacent points, we can find a piece-wise linear estimation of the nondominated frontier of (2).

EXAMPLE 2.1. Consider two quadratic functions $y_1 = x_1^2$ and $y_2 = 3x_1^2 - x_1$, where $x \in \mathbb{R}_+^n$. We have the following biobjective optimization problem:

$$\begin{array}{ll} \min & z = (x_1^2, 3x_1^2 - x_1) \\ s.t. & x_1 \geq 0. \end{array} \tag{6}$$

Using the proposed method, the following two problems are constructed:

$$\begin{array}{ll} \min & z = x_1^2 \\ s.t. & x_1 \geq 0, \end{array} \tag{7}$$

and

$$\begin{array}{ll} \min & z = 3x_1^2 - x_1 \\ s.t. & x_1 \geq 0, \\ & x_1^2 \leq z_{CP1} + k\delta. \end{array} \tag{8}$$

Considering problem (7), we have $M_0 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$, $X = \begin{bmatrix} a & b \\ b & d \end{bmatrix}$ and $J_0 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$. $\langle M_0, X \rangle = tr(\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a & b \\ b & d \end{bmatrix}) = d$ and by $\langle J_0, X \rangle = 1$, we have

$$tr(\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} a & b \\ b & d \end{bmatrix}) = a = 1.$$

Because of $X \in C$, we can let $X = xx^T$ [2], where $x = \begin{bmatrix} x_1' \\ x_2' \end{bmatrix}$, so $\begin{bmatrix} x_1'^2 & x_1'x_2' \\ x_1'x_2' & x_2'^2 \end{bmatrix} = \begin{bmatrix} a & b \\ b & d \end{bmatrix}$. Therefore, $d \geq 0$, and problem (4) according to (7) is as $\begin{array}{ll} \min & d \\ s.t. & d \geq 0. \end{array}$ and we conclude $z_{CP1}^* = 0$. In this regard, problem (8) is as follows:

$$\begin{array}{ll} \min & z = 3x_1^2 - x_1 \\ s.t. & x_1^2 \leq k\delta, \\ & x_1 \geq 0, \end{array} \tag{9}$$

choosing $\delta = \frac{1}{180}$ for $k = 0, 1, \ldots, 5$, we should solve 6 problems. For example for $k = 1$ we have

$$\begin{array}{ll} \min & -b + 3d \\ s.t. & d \leq \frac{1}{180}, \\ & d \geq 0, \end{array}$$

so, $d = \frac{1}{180}$ and $b = \frac{1}{\sqrt{180}}$. These imply that $z_{CP2}^* = \frac{-1}{\sqrt{180}} + \frac{1}{60} \cong -0.0578$ and $z_{CP1}^* \cong 0.0055$. Similarly, we can find z_{CP}^* for all 6 problems. In Table 1, the optimal objective values of these 6 problems in the structure of $(z_{CP1+k\delta}, z_{CP2})$ have been listed. The pair points represent the estimation of some nondominated points of (6).

TABLE 1. The estimation of the nondominated points of (6).

	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
$(q_1(x), q_2(x))$	(0,0)	(0.005,-0.057)	(0.011,-0.0720)	(0.0166,-0.1124)	(0.022,-0.126)	(0.0277,-0.083)

In Figure 1, the nondominated frontier of (6) have been depicted in green curve. Using pairs $(z_{CP} + k\delta, z_{CP2})$, $k = 1, \ldots, 6$, the piece-wise linear estimation of the nondominated frontier of (6) has been drawn in red color.

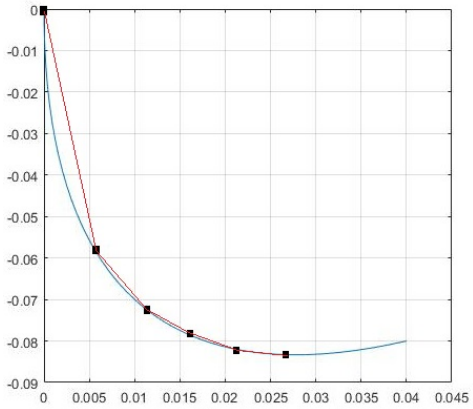


FIGURE 1.

References

1. I. M. Bomze, *Copositive relaxation beats Lagrangian dual bounds in quadratically and linearly constrained quadratic optimization problems*, SIAM J. Optim. **25** (3) (2015) 1249–1275.
2. A. Berman, *Complete positivity*, Linear Algebra Appl. **107** (1988) 57–63.
3. M. Dür, *Copositive Programming–a survey*, In: M. Diehl, F. Glineur, E. Jarlebring and W. Michiels (Eds), Recent Advances in Optimization and its Applications in Engineering, Springer, Berlin, Heidelberg, (2010) pp. 3–20.
4. C. L. Hwang and A. S. M. Masud, *Multiple Objective Decision Making–Methods and Applications: A State-of-the-Art Survey* Springer-Verlag Berlin Heidelberg, New York, 1979.

E-mail: mirdehghan@shirazu.ac.ir
E-mail: dibash1374@gmail.com

Optimality and Duality for Efficiency in Nonsmooth Multiobjective Fractional Optimization Problems

Ali Ansari Ardali*
Department of Applied Mathematics, Faculty of Mathematical Sciences, Shahrekord University, P. O. Box 115, Shahrekord, 88186-34141, Iran

Abstract. This paper is devoted to the study of optimality conditions and duality for nonsmooth multiobjective fractional optimization problems, involving inequality and equality constraints in terms of the limiting/ Mordukhovch subdifferential. Based on the concept of Mordukhovch subdifferential and using suitable generalized constraint qualification, we derive necessary and sufficient optimality conditions for these problems. In addition, we propose a type of Wolfe dual problems and examine weak/strong duality relations under generalized convexity.
Keywords: Multiobjective fractional optimization problems, Optimality conditions, Duality, Generalized convexity.
AMS Mathematical Subject Classification [2010]: 90C32, 90C46, 49J52.

1. Introduction

Optimality conditions and duality relations for multiobjective optimization problems involving locally Lipschitz functions have been investigated intensively by many researchers; see e.g., [1, 2, 3, 4] and the references therein. In this work, we employ some advanced tools of variational analysis and limiting/Mordukhovich subdifferential to establish necessary optimality conditions for Pareto/efficient solutions of a nonsmooth fractional multiobjective optimization problem with inequality and equality constraints. Since the limiting/Mordukhovich subdifferential of a real-valued function at a given point is contained in the convexified/Clarke subdifferential of such a function at the corresponding point [6], the necessary optimality conditions formulated in terms of the limiting subdifferential are sharper than the corresponding ones expressed in terms of the convexified subdifferential. Sufficient optimality conditions for such solutions to the considered problem are also provided by means of introducing generalized convex-affine functions defined in terms of the limiting subdifferential for locally Lipschitz functions. Along with optimality conditions, we state a dual problem to the primal one and explore weak and strong duality relations under assumptions of generalized convexity.

2. Preliminaries

In this section, we recall some definitions and results from nonsmooth analysis needed in what follows see, e.g., [6] for more details. For each $m \in \mathbb{N}$, we denote by $\mathbb{R}^m_+$ the nonnegative orthant of $\mathbb{R}^m$. The canonical pairing between space

*Speaker

Y and its topological dual Y^* is denoted by $\langle\cdot,\cdot\rangle$ while the symbol $\|.\|$ for the norm in the considered space. The polar cone of a set $S \subset Y$ is defined by $S^\circ := \{y^* | \langle y^*, y\rangle \leq 0,\ \forall y \in Y\}$. Given a set-valued mapping $\Psi : Y \rightrightarrows Y^*$, we denote by

$$\limsup_{y\to\hat{y}} \Psi(y) = \{y^* \in Y^* \mid \exists\ y_n \to \hat{y} \text{ and } y_n^* \rightharpoonup y^* \text{ with } y_n^* \in \Psi(y_n),\ \forall\ n \in \mathbb{N}\},$$

the sequential Painlevé-Kuratowski upper/outer limit of Ψ as $y \to \hat{y}$, where the notation $\rightharpoonup$ indicates the convergence in the weak* topology of Y^*. Let $\Theta \subset Y$ be locally closed around $\hat{y} \in \Theta$. Define the (basic/ limiting/ Mordukhovich) normal cone to Θ at $\hat{y} \in \Theta$ by

$$N_M(\hat{y};\Theta) := \limsup_{y \overset{\Theta}{\to} \hat{y}} \hat{N}(y;\Theta) = \limsup_{y \overset{\Theta}{\to} \hat{y}} \left\{ y^* \in Y^* \middle| \limsup_{y' \overset{\Theta}{\to} y} \frac{\langle y^*, y' - y\rangle}{\|y' - y\|} \leq 0 \right\}.$$

Now, let us recall the definitions of subdifferentials which will be used throughout the paper. Let $\psi : Y \to \overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$ be an extended real-valued function. The Mordukhovich subdifferential of ψ at $\hat{y} \in dom\ \psi$ are given, by

$$\partial_M \psi(\hat{y}) = \{y^* \in Y^* \mid (y^*, -1) \in N_M((\hat{y}, \psi(\hat{y}); epi\ \psi)\},$$

where $epi\ \psi = \{(y,\mu) \in Y \times \mathbb{R} |\ \mu \geq \psi(y)\}$. If $\hat{y} \notin dom\ \psi$, then one puts $\partial_M \psi(\hat{y}) = \emptyset$. It is known [6] that when ψ is a convex function, the above-defined subdifferentials coincides with the subdifferentials in the sense of convex analysis [7]. Also, the nonsmooth version of Fermats rule (See, e.g., [6, Proposition 1.114]), which is an important fact for many applications, can be formulated as follows: If $\hat{y}$ is a local minimizer for ψ, then:

$$0 \in \partial_M \psi(\hat{y}). \tag{1}$$

3. Main Results

In this section, we first establish necessary optimality conditions for efficient solutions of a multiobjective fractional optimization problem. Then by imposing assumptions of generalized convexity, we give sufficient optimality conditions for such solutions. Also, we propose a dual problem to the primal one in the sense of Mond and Weir [5] and examine weak and strong duality relations between them.

Recall Mordukhovich [6] that a set $\Theta \subset Y$ is sequentially normally compact (SNC) at $\hat{y} \in \Theta$ if for any sequences: $y_k \overset{\Theta}{\to} \hat{y}$ and $y_k^* \overset{w^*}{\to} 0$ with $y_k^* \in \widehat{N}(y_k;\Theta)$, one has $\|y_k^*\| \to 0$ as $k \longrightarrow \infty$. Also, a function $\psi : Y \longrightarrow \mathbb{R}$ is called sequentially normally compact (SNC) at $\hat{y} \in Y$ if $gph\ \psi$ is (SNC) at $(\hat{y}, \psi(\hat{y}))$.

Let us consider the following constrained multiobjective fractional optimization problem (P):

$$(P)\quad \begin{aligned} &\min_{\mathbb{R}^m_+}\ \psi(y) = \left(\frac{\tau_1(y)}{\rho_1(y)}, \ldots, \frac{\tau_m(y)}{\rho_m(y)}\right) \\ &s.t.\ \varphi_i(y) \leq 0,\ i \in I = \{1, \ldots, p\}, \\ &\quad \phi_j(y) = 0,\ j \in J = \{1, \ldots, q\},\ y \in \Theta, \end{aligned}$$

where Θ is a nonempty locally closed subset of Y and $\psi = (\frac{\tau_k(y)}{\rho_k(y)})$, $k \in K = \{1, \ldots, m\}$, $\varphi = (\varphi_i)$, $i \in I$ and $\phi = (\phi_j)$, $j \in J$ are vector functions with locally Lipschitz components defined on Y. For, the sake of convenience, we further assume that $\rho_k(y) > 0,\ k \in K$ for all $y \in \Theta$, and that $\tau_k(\hat{y}) \leq 0,\ k \in K$ for the reference point $\hat{y} \in \Theta$.

We say that feasible point $\hat{y}$ is an efficient solution of problem (P), and write $\hat{y} \in \mathcal{S}(P)$ if and only if for every feasible point y, $\psi(y) - \psi(\hat{y}) \notin -\mathbb{R}^m_+ \setminus \{0\}$. We say that constraint qualification (MFCQ) is satisfied at $\hat{y} \in \Theta$ if there do not exist $\mu_i \geq 0$, $i \in I(\hat{y}) := \{i \in I | \varphi_i(\hat{y}) = 0\}$ and $\gamma_j \geq 0$, $j \in J(\hat{y}) := \{j \in J | \phi_j(\hat{y}) = 0\}$, such that $\sum_{i \in I(\hat{y})} \mu_i + \sum_{j \in J(\hat{y})} \gamma_j \neq 0$ and

$$0 \in \sum_{i \in I(\hat{y})} \mu_i \partial_M \varphi_i(\hat{y}) + \sum_{j \in J(\hat{y})} \gamma_j \big(\partial_M \phi_j(\hat{y}) \cup \partial_M(-\phi_j)(\hat{y})\big) + N_M(\hat{y}; \Theta).$$

The following theorem gives a (KKT) type necessary optimality condition for efficient solutions of problem (P).

THEOREM 3.1. *Let $\hat{y} \in \mathcal{S}(P)$. If* (MFCQ) *holds at $\hat{y} \in \Theta$, then there exist $(\lambda_k) \in \mathbb{R}^m_+ \setminus \{0\}$, $(\mu_i) \in \mathbb{R}^p_+$ and $(\gamma_j) \in \mathbb{R}^q_+$ such that*

$$\begin{aligned} 0 \in & \sum_{k \in K} \lambda_k \big(\partial_M \tau_k(\hat{y}) - \frac{\tau_k(\hat{y})}{\rho_k(\hat{y})} \partial_M \rho_k(\hat{y})\big) + \sum_{i \in I} \mu_i \partial_M \varphi_i(\hat{y}) \\ & + \sum_{j \in J} \gamma_j \big(\partial_M \phi_j(\hat{y}) \cup \partial_M(-\phi_j)(\hat{y})\big) + N_M(\hat{y}; \Theta),\ \ \mu_i \varphi_i(\hat{y}) = 0\ i \in I. \end{aligned} \tag{2}$$

PROOF. Let $\hat{y} \in \mathcal{S}(P)$ and $\Psi(y) = \max_{k \in K} \left\{ \frac{\tau_1(y)}{\rho_1(y)} - \frac{\tau_m(y)}{\rho_m(y)} \right\}$. We are going to show that, for every feasible point y, $\Psi(\hat{y}) < \Psi(y)$. Indeed, if this is not the case, then there exists feasible point $\bar{y}$ such that $\Psi(\bar{y}) \leq \Psi(\hat{y})$. Thus, $\Psi(\bar{y}) - \Psi(\hat{y}) \in \mathbb{R}^m_+ \setminus \{0\}$, which contradicts the fact that $\hat{y} \in \mathcal{S}(P)$. Thus, $\hat{y}$ is a minimizer of the following unconstrained scalar optimization problem

$$\min_{y \in Y} \Psi(y) + \delta(y; \Omega), \tag{3}$$

where Ω is feasible set of problem (P) and $\delta(.; \Omega)$ is indicator function. Applying the nonsmooth version of Fermats rule (1), we have

$$0 \in \partial_M \big(\Psi + \delta(.; \Omega)\big)(\hat{y}). \tag{4}$$

It follows from the basic subdifferential of maximum functions [6, Theorem 3.46], the quotient rule [6, Corollary 1.111] and the sum rule [6, Theorem 3.36] we obtain

$$\partial_M \Psi(\hat{y}) \subset \left\{ \sum_{k \in K} \alpha_k \frac{\tau_k(\hat{y}) \partial_M \rho(\hat{y}) - \rho_k(\hat{y}) \partial_M \tau(\hat{y})}{\rho(\hat{y})^2} \Big| \alpha_k \geqslant 0,\ k \in K,\ \sum_{k \in K} \alpha_k = 1 \right\}. \tag{5}$$

As the (MFCQ) is satisfied at $\hat{y}$ and Ω is assumed to be (SNC) at this point, and apply [6, Corollary 3.5], it follows from (3)-(5) that for $\beta_i \geqslant 0,\ i \in I(\hat{y}), \gamma_j \geqslant$

$0,\ j \in J,$

$$0 \in \Big\{ \sum_{k\in K} \frac{\alpha_k}{\rho(\hat{y})} \Big(\partial_M \tau_k(\hat{y}) - \frac{\tau_k(\hat{y})}{\rho_k(\hat{y})} \partial_M \rho(\hat{y}) \Big) | \alpha_k \geqslant 0,\ k \in K,\ \sum_{k\in K} \alpha_k = 1 \Big\}$$
$$+ \Big\{ \sum_{i\in I(\hat{y})} \beta_i \partial_M \varphi_i(\hat{y}) + \sum_{j\in J} \gamma_j \big(\partial_M \phi_j(\hat{y}) \cup \partial_M(-\phi_j)(\hat{y}) \big) \Big\} + N_M(\hat{y}; \Omega).$$

Now, by put $\beta_i = 0,\ i \in I - I(\hat{y})$ and $\lambda_k = \frac{\alpha_k}{\rho_k(\hat{y})}$, the proof is complete. □

Definition 3.2. We say that $(f,\ g;\ h)$ is strictly generalized convex-affine on Ω at $\hat{y} \in \Omega$ if for any $y \in \Omega \setminus \{\hat{y}\}$, $\zeta_k \in \partial_M \tau_k(\hat{y})$, $\eta_k \in \partial_M \rho_k(\hat{y})\ k \in K$, $\sigma_i \in \partial_M \varphi_i(\hat{y}),\ i \in I$ and $\theta_j \in \partial h_j(\hat{y}) \cup \partial_M(-h_j)(\hat{y})$ there exists $\nu \in N_M(\hat{y}; \Omega)^-$ such that for $\omega_j \in \{1, -1\}$,

$$\tau_k(y) - \tau_k(\hat{y}) \geq \langle \zeta_k, \nu \rangle,\ k \in k,\ \rho_k(y) - \rho_k(\hat{y}) \geq \langle \eta_k, \nu \rangle,\ k \in k,$$
$$\varphi_i(y) - \varphi_i(\hat{y}) \geq \langle \sigma_i, \nu \rangle,\ i \in I,\ \phi_j(y) - \phi_j(\hat{y}) = \omega_j \langle \theta_j, \nu \rangle,\ j \in J.$$

We are now to provide sufficient conditions for a feasible point of problem (P) to be a efficient.

Theorem 3.3. *Assume that $\hat{y} \in \Omega$ satisfies condition (2). If $(\psi;\ \varphi;\ \phi)$ is strictly generalized convex-affine on Ω at $\hat{y}$, then $\hat{y} \in \mathcal{S}(P)$.*

Proof. Since $\hat{y}$ satisfies condition (2), then there exist $(\lambda_k) \in \mathbb{R}^m_+ \setminus \{0\}$, $(\mu_i) \in \mathbb{R}^p_+$ and $(\gamma_j) \in \mathbb{R}^q_+$, $\zeta^*_k \in \partial_M \tau^*_k(\hat{y}),\ \eta^*_k \in \partial_M \rho_k(\hat{y})\ k \in K$, $\sigma^*_i \in \partial_M \varphi_i(\hat{y}),\ i \in I$, with $\mu_i \varphi_i(\hat{y})$, and $\theta^*_j \in \partial h_j(\hat{y}) \cup \partial_M(-h_j)(\hat{y})$ such that

$$-\Big(\sum_{k\in K} \lambda_k \big(\zeta^*_k - \frac{\tau_k(\hat{y})}{\rho_k(\hat{y})} \eta^*_k \big) + \sum_{i\in I} \mu_i \sigma^*_i + \sum_{j\in J} \gamma_j \theta^*_j \Big) \in N_M(\hat{y}; \Omega). \tag{6}$$

Suppose to the contrary that $\hat{y} \notin \mathcal{S}(P)$. Then there is $\bar{y}$ such that $\psi(\bar{y}) - \psi(\hat{y}) \in -\mathbb{R}^m_+ \setminus \{0\}$. By the strictly generalized convex-affine property of $(\psi;\ \varphi;\ \phi)$ on Ω at $\hat{y}$, for $\bar{y}$ above, there exists $\nu \in N_M(\hat{y}; \Omega)^-$ such that

$$\Big(\sum_{k\in K} \lambda_k \big(\langle \zeta^*_k, \nu \rangle - \frac{\tau_k(\hat{y})}{\rho_k(\hat{y})} \langle \eta^*_k, \nu \rangle \big) + \sum_{i\in I} \mu_i \langle \sigma^*_i, \nu \rangle + \sum_{j\in J} \gamma_j \langle \theta^*_j, \nu \rangle \Big)$$
$$< \sum_{k\in K} \lambda_k \big[\tau_k(\bar{y}) - \tau_k(\hat{y}) - \frac{\tau_k(\hat{y})}{\rho_k(\hat{y})} \big(\rho_k(\bar{y}) - \rho_k(\hat{y}) \big] + \sum_{i\in I} \mu_i \big(\varphi_i(\bar{y}) - \varphi_i(\hat{y}) \big)$$
$$+ \sum_{j\in J} \frac{1}{\omega_j} \gamma_j \big(\phi_j(\bar{y}) - \phi_j(\hat{y}) \big),$$

where $\omega_j \in \{-1, 1\},\ j \in J$. This entails that there is $k_0 \in K$ such that $\psi_{k_0}(\bar{y}) < \psi_{k_0}(\hat{y})$. It gives a contradiction, which completes the proof. □

Now, we consider a Mond-Weir multiobjective fractional dual problem of the form:

$$(D)\ \max_{\mathbb{R}^m_+} \Big\{ \tilde{\psi}(z, \lambda, \mu, \gamma) := \big(\frac{\tau_1(y)}{\rho_1(y)}, \ldots, \frac{\tau_m(y)}{\rho_m(y)} \big) | (z, \lambda, \mu, \gamma) \in \Omega_D \Big\},$$

where, if $\mathbb{B}(0,\|\gamma\|)=\{\sigma\in\mathbb{R}^q\big|\|\sigma\|=\|\gamma\|\}$, the constraint set Ω_D is defined by

$$\begin{aligned}(7)\quad \Omega_D: \;=\; & \big\{(z,\lambda,\mu,\gamma)\big|0\in\sum_{k\in K}\lambda_k\big(\partial_M\tau_k(z)-\frac{\tau_k(z)}{\rho_k(z}\partial_M\rho_k(z)\big)\\ & +\;\sum_{i\in I}\mu_i\partial_M\varphi_i(z)+\sum_{j\in J}\gamma_j\big(\partial_M\phi_j(z)\cup\partial_M(-\phi_j)(z)\big)+N_M(\hat{z};\Theta),\\ & \langle\mu,\varphi(z)\rangle+\langle\sigma,\phi(z)\rangle\geq 0,\;\forall\sigma\in\mathbb{B}(0,\|\gamma\|)\big\},\end{aligned}$$

weak duality and strong duality relations between the primal problem (P) and the dual problem (D) read as follows.

THEOREM 3.4. (Weak Duality) *Let $y\in\Omega$ and $(z,\lambda,\mu,\gamma)\in\Omega_D$. If $(\psi;\ \varphi;\ \phi)$ is strictly generalized convex-affine on Ω at z, then $\psi(y)\not\leq\tilde{\psi}(z,\lambda,\mu,\gamma)$.*

PROOF. Assume to the contrary that $\psi(y)\leq\tilde{\psi}(z,\lambda,\mu,\gamma)$. By the strictly generalized convex-affine property of $(\psi;\ \varphi;\ \phi)$ on Ω at z, for such y, that there is $k_0\in K$ such that $\big(\tau_{k_0}(y)-\frac{\tau_{k_0}(z)}{\rho_{k_0}(z)}\rho_{k_0}(y)\big)>0$, or equivalently, $\psi_{k_0}(z)<\psi_{k_0}(y)$ which contradict and therefore completes the proof. □

THEOREM 3.5. (Strong Duality) *Let $\bar{y}\in\mathcal{S}(P)$, be such that the* (MFCQ) *is satisfied at this point. Then there exists $(\bar{\lambda},\bar{\mu},\bar{\gamma})\in(\mathbb{R}^m_+)\setminus\{0\}\times\mathbb{R}^p_+\times\mathbb{R}^q_+$ such that $(\bar{y},\bar{\lambda},\bar{\mu},\bar{\gamma})\in\Omega_D$ and $\psi(\bar{y})=\tilde{\psi}(\bar{y},\bar{\lambda},\bar{\mu},\bar{\gamma})$. If in addition $(\psi;\ \varphi;\ \phi)$ is strictly generalized convex-affine on Ω at any z, then $(\bar{y},\bar{\lambda},\bar{\mu},\bar{\gamma})\in\mathcal{S}(D)$.*

PROOF. According to Theorem 3.1 and since $(\psi;\ \varphi;\ \phi)$ is strictly generalized convex-affine on Ω at any z, by invoking of Theorem 3.4, we assert that $\tilde{\psi}(\bar{y},\bar{\lambda},\bar{\mu},\bar{\gamma})\not\leq\tilde{\psi}(y,\lambda,\mu,\gamma)$, for any $(y,\lambda,\mu,\gamma)\in\Omega_D$. Hence, $(\bar{y},\bar{\lambda},\bar{\mu},\bar{\gamma})\in\mathcal{S}(D)$. □

References

1. A. Ansari Ardali, *Boundedness of KKT multipliers in fractional programming problem using convexificators*, Iranian J. Oper. Res. **6** (1) (2015) 79–91.
2. T. D. Chuong and D. S. Kim, *Optimality conditions and duality in nonsmooth multiobjective optimization problems*, Ann. Oper. Res. **217** (2014) 117–136.
3. H. C. Lai and S. C. Ho, *Optimality and duality for nonsmooth multiobjective fractional programming problems involving exponential V-r -invexity*, Nonlinear Anal. **75** (6) (2012) 3157–3166.
4. X. J. Long, *Optimality conditions and duality for nondifferentiable multiobjective fractional programming problems with $(C,\alpha,\rho,d)-$convexity*, J. Optim. Theory Appl. **148** (1) (2011) 197–208.
5. B. Mond and T. Weir, *Generalized Concavity and Duality* Academic Press, New York, 1981.
6. B. S. Mordukhovich, *Variational Analysis and Generalized Differentiation. II*, Springer-Verlag, Berlin, 2006.
7. R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, New Jersey, 1970.

E-mail: ali.ansariardali@sku.ac.ir

Calculating Optimum Control Law for a Non-Homogeneous Linear Time-Invariant Control System via HJB Equation

Mehrasa Ayatollahi*
Department of Mathematics, Payame Noor University (PNU), Tehran, Iran

Abstract. In this paper, we consider the problem of linear quadratic continuous time optimal control. Our assumed system for this problem is a special case of non-homogeneous control systems with non-zero terms. To minimize the certain cost function assigned to this system, we will propose an optimum control strategy which is calculated by incorporating Hamilton-Jacobi-Bellman partial differential equation.
Keywords: Non-homogeneous linear control system, Optimization, Hamilton-Jacobi-Bellman equation.
AMS Mathematical Subject Classification [2010]: 49L20, 93C05.

1. Introduction

In this work, we deal with the problem of optimization of linear time invariant (LTI) continuous time control systems whose parameters do not vary with time. In practice, applications of LTI systems may be found in circuits, control theory, NMR spectroscopy, signal processing and in many other areas.

Any LTI control system is expressed as a first order ordinary differential equation of the form:

$$\dot{x}(t) = f(t, x(t), u(t)), x(0) = x_0, \tag{1}$$

where $x(t)$ and $u(t)$ are state vector and input vector, respectively. The LTI control system (1) can be classified into two categories: non-homogeneous system and homogeneous system. In a homogeneous system, no external signal is applied and we look for behavior of the states due to the presence of initial condition only. But in a non-homogeneous system, we have both the initial conditions and external input signals simultaneously [3].

The problem of optimization of LTI system is defined as determination of the best possible control strategy (usually of the optimum control vector $u(t)$), which minimizes a certain cost function or performance index. This problem is considered widely in the literature of control systems (See for example [5] and [6] and the references therein).

In what follows, we consider the problem of optimal tracking and terminal control of a non-homogeneous LTI control system in finite horizon. In general, a control optimization problem consists of minimization of a cost function as:

$$J = \int_0^T \{g(t, x(t), u(t)\}dt + q(T, x(T)),$$

*Speaker

subject to the continuous time dynamic:

$$\dot{x}(t) = f(t, x(t), u(t)), x(0) = x_0.$$

One approach for solving the mentioned problem is applying transformation on the affine system but it extends the state matrix into a time dependent form. To avoid this difficulty, we propose an approach that uses Hamilton-Jacobi-Bellman (HJB) partial differential equation. Although it is not easy to find an exact solution to HJB equation, yet there are some advantages in using this equation. HJB arises as a central aspect in optimal control theory and also provides a relatively inexpensive way to verify optimality among optimization methods if one is able to guess a solution. The HJB equation related to the above mentioned problem is expressed as:

$$-\frac{\partial V(t,x)}{\partial t} = \min_u \big[\frac{\partial V(t,x)}{\partial x} f(t, x(t), u(t)) + g(t, x(t), u(t)\big],$$

with boundary condition $V(T, x) = q(T, x)$ [2]. $V(t, x)$ is a continuously differentiable function which is called *value function* and generally it is not easy to compute it.

2. Main Results

Here, we consider a non-homogeneous linear quadratic control problem which is the minimization of the following cost function:

$$J = \frac{1}{2}x^T(T)Q_T x(T) + \frac{1}{2}\int_0^T \{x^T Q x + u^T R u\}dt, \tag{2}$$

subjec to:

$$\dot{x}(t) = Ax(t) + Bu(t) + c(t), x(0) = x_0. \tag{3}$$

In which R is a positive definite matrix and Q, R and Q_T are symmetric matrices with appropriate size. The first term in (2) refers to the cost at the end of the optimization time interval and the second term refers to the cost in the entire optimization interval. For LTI system (3) A and B are state and input matrices while the non-homogeneous term $c(t)$, is assumed to be a known vector-valued function which is such that the solution $x(t)$ of the differential equation is uniquely defined.

The above mentioned problem has a solution if and only if the original linear quadratic control problem with $c(t) = 0$ has a solution [4]. For solving this problem it is possible to rewrite it as a standard linear quadratic control problem using the transformation on the affine systems as introduced in [4], but in that case the extended state matrix becomes time dependent.

In the following theorem, by using Hamilton-Jacobi-Bellman equation we will introduce an optimal control law to solve this minimization problem while all matrices of the system remain time independent.

THEOREM 2.1. *Consider the problem of minimization of linear quadratic cost function* (2) *subject to* (3)*, in which* $c(t)$ *is an arbitrarily function such that for*

$c(t) = 0$ there exist a solution to the minimization problem. Then, the optimal control strategy that solves this problem is:

$$u^* = -R^{-1}B^T(S(t)x + m(t)),$$

where $S(t)$ is an $n \times n$ symmetric matrix with continuously differentiable entries which is the solution of the differential Riccati equation:

$$\dot{S}(t) + S(t)A + A^T S(t) - S(t)BR^{-1}B^T S(t) + Q = 0 \ , \ S(T) = Q_T,$$

and $m(t)$ is an $n \times 1$ vector with continuously differentiable entries which is the solution of the ordinary differential equation:

$$\dot{m}(t) + (A - BR^{-1}B^T S)^T m(t) + S(t)c = 0 \ , \ m(T) = 0.$$

Proof. The corresponding HJB equation to our mentioned system is:

$$-\frac{\partial V(t,x)}{\partial t} = \min_u \big[\frac{\partial V(t,x)}{\partial x}(Ax + Bu) + \frac{1}{2}(x^T Qx + u^T Ru)\big]. \tag{4}$$

Assume existence of a continuously differentiable value function of the form:

$$V(t,x) = \frac{1}{2}x^T S(t)x + m^T(t)x + n(t), \tag{5}$$

that satisfies (4). Here $S(t)$ is an $n \times n$ symmetric matrix with continuously differentiable entries, $m(t)$ is a continuously differentiable n-vector and $n(t)$ is a continuously differentiable function. To determine such $S(t)$, $m(t)$ and $n(t)$ we substitute (5) into the HJB Eq. (4) which leads to:

$$\begin{aligned} -\frac{1}{2}x^T\dot{S}(t)x - \dot{m}^T(t)x - \dot{n}(t) = \min_u \big[&\frac{1}{2}x^T Qx + \frac{1}{2}u^T Ru \\ &+ (x^T S(t) + m^T(t))(Ax + Bu + c)\big]. \end{aligned}$$

Carrying out the minimization on the right hand side yields the optimum control signal:

$$u^* = -R^{-1}B^T(S(t)x + m(t)). \tag{6}$$

For determining $S(t)$ and $m(t)$, we substitute (6) into (5) that leads to an identity relation which is readily satisfied if:

$$\begin{aligned} &\dot{S}(t) + S(t)A + A^T S(t) - S(t)BR^{-1}B^T S(t) + Q = 0, && S(T) = Q_T, \\ &\dot{m}(t) + (A - BR^{-1}B^T S(t))^T m(t) + S(t)c = 0, && m(T) = 0, \\ &\dot{n}(t) + m^T(t)c - \tfrac{1}{2}m^T(t)BR^{-1}B^T m(t) = 0, && n(T) = 0. \end{aligned} \tag{7}$$

The first equation in (7) is a matrix differential Riccati equation. According to [1], positive definiteness of Q and Q_T and controllability of (A, B) guarantee the existence of the unique S satisfying in the first equation. The two other differential equations in (7) are linear in L and G respectively. Due to the theorem of existence and uniqueness of solutions for the linear first-order differential equations [7], the existence of unique solutions to the two remaining equations in (7) is guaranteed by existence of S. This implies that the HJB Eq. (4) has a solution in the form of (5) which satisfies the boundary conditions. Now, the optimum control (6) is completely determined. □

3. Numerical Example

In this section, the proposed method is applied to a dynamical system with the following constant matrices:

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad c = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Required matrices for cost function are taken as

$$Q = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad R = I_{1\times 1}, \quad Q_T = 0_{2\times 2}.$$

Appling ODEs (7) on the controllable pair (A, B) yields

$$S = \begin{bmatrix} 1.4142 & 1 \\ 1 & 1.4142 \end{bmatrix}, \quad m = \begin{bmatrix} 1 \\ 1.4142 \end{bmatrix}.$$

So, the optimum control law will be

$$u^* = -\begin{bmatrix} 0 & 1 \end{bmatrix}(Sx + m) = -x_1 - 1.4142x_2 - 1.4142.$$

References

1. B. D. Anderson and J. B. Moore, *Optimal Control: Linear Quadratic Methods*, Prentice Hall, Englewood Cliffs, NJ, USA. 1990.
2. T. Basar and G. J. Olsder *Dynamic non-cooperative game theory*, 2nd ed., Philadelphia, PA, SIAM 1999.
3. A. Deb, S. Roychoudhury and G. Sarkar, *Analysis and Identification of Time-Invariant Systems, Time-Varying Systems, and Multi-Delay Systems using Orthogonal Hybrid Functions*, Theory and algorithms with MATLAB®. Studies in Systems, Decision and Control, 46. Springer, Cham, 2016.
4. J. Engwerda, *LQ Dynamic Optimization and Differential Games*, John Wiley and son, Tilburg University, Netherland, England, 2005.
5. W. G. dos Santos, E. M. Rocco and T. Boge, *Design of a linear time-invariant control system based on a multiobjective optimization approach*, Comput. Appl. Math. **35** (2016) 789–801.
6. S. N. Sivanandam and S. N. Deepa, *A genetic algorithm and particle swarm optimization approach for lower order modelling of linear time invariant discrete systems*, IEEE Int. Conf. Comput. Intell. Multimedia Appl., Sivakasi, Tamil Nadu, India, (2007). DOI: 10.1109/ICCIMA.2007.41
7. L. Vigano, M. Bergamasco, M. Lovera and A. Varga, *Optimal periodic output feedback control: a continuous-time approach and a case study*, Internat. J. Control. **83** (2010) 897–914.

E-mail: m_ayatollahi@pnu.ac.ir

Semidefinite Relaxation for Total Dominating Set Problem

Mehdi Djahangiri*
Department of Mathematics, Faculty of Basic Science University of Maragheh, Maragheh, Iran
and Mohsen Abdolhosseinzadeh
Department of Mathematics, Faculty of Basic Science University of Bonab, Bonab, Iran

Abstract. Finding a solution for the combinatorial optimization problems has always been important due to their applications. But most of them are NP-Complete and unsolvable in polynomial time. Therefore, the approximation algorithms have been designed for them. One of these problems is total dominating set problem. In this paper, we present a new quadratic integer programming model for total dominating set problem and design an approximation method to find a lower bound for total dominating number.
Keywords: Total dominating set, Integer programming, Semidefinite programming.
AMS Mathematical Subject Classification [2010]: 05C69, 90C10, 90C22.

1. Introduction

Consider an undirected and connected graph $G = (V, E)$, where $V = \{v_1, \ldots, v_n\}$ and E are respectively vertices and edges of G. The degree of vertex v_i is shown by $deg(v_i)$, and Δ stands for the maximum degree of the graph. A set $S \subseteq V$ is called dominating set of G if each vertex is a member of S or adjacent to a member of S. The set S is referred to as minimum dominating set if it has minimum cardinality among all dominating sets. The cardinality of minimum dominating set is called domination number and denoted by $\gamma(G)$. Domination number and its variations have been extensively studied in the literature. One of them is total domination number. A set S_t of vertices in a graph G is called a total dominating set if every vertex $v_i \in V$ is adjacent to an element of s_t. The size of total dominating set with minimum cardinality is denoted by $\gamma_t(G)$. For more details we refer the reader to [9].

The difference between these two parameters arises from this fact that the members of dominating set is not necessary be adjacent with another one in dominating set while it is necessary in total dominating set. The parts (a) and (b) in the Figure 1, show this difference. The black vertices show dominating set and total dominating set respectively in the parts (a) and (b).

Dominating set and its variants are one of the classical problems in graph theory having important applications in many fields (e.g. [3, 4] for some recent applications). In [8], more than 1200 papers on different versions of dominating set problem are listed. Despite having a lot of application and theoretical attraction,

*Speaker

Unfortunately, in [5] it has been shown the NP-completeness of dominaing set problem and subsequently the total dominating set problem. So, for any arbitrary graph, it is not expected that the total dominating set will be found in reasonable time. To overcome to this challenge, there are several methods such linear relaxation, Greedy Algorithms and metaheuristics. In this paper, the semidefinite relaxation is applied to find an approximation solution for the total dominating set problem.

The semidefinite programming is a special case of convex optimization which linear objective function is optimized over the intersection of the cone of positive semidefinite matrices with linear constraints. Let $\mathbb{S}^n$ denote the set of symmetric $n \times n$ real matrices. The cone of symmetric positive semidefinite (definite) matrices is denoted by $\mathbb{S}^n_+$ ($\mathbb{S}^n_{++}$). $B - D \succeq 0$ ($B - D \succ 0$) means that $(B - D)$ is positive semidefinite (definite). Suppose that $A_1, \ldots, A_m$ are linearly independent matrices in $\mathbb{S}^n$; $C \in \mathbb{S}^n$ and $b \in \mathbb{R}^m$. The standard form of semidefinite programming problem is written as follows:

$$\begin{array}{ll} \min & \langle C, X \rangle \\ \text{s.t.} & \langle A_i, X \rangle \geq b_i, \quad i = 1, 2, \ldots, n, \\ & X \succeq 0, \end{array}$$

where $\langle B, D \rangle = \mathrm{tr}(B^t D) = \sum_{i,j} b_{ij} d_{ij}$. The semidefinite programming model can be solved in a polynomial time with an interior point method [1]. The interested reader is referred to [2, 10] for a thorough discussion and applications of semidefinite programming. semidefinite programming relaxation is a powerful tool to approximate the optimal solution of some combinatorial problems. For example, dominating set [6] and maximum cut [7]. The good performance of semidefinite relaxation in these problems encouraged us to utilize this method to find an approximation of the k-tuple domination number.

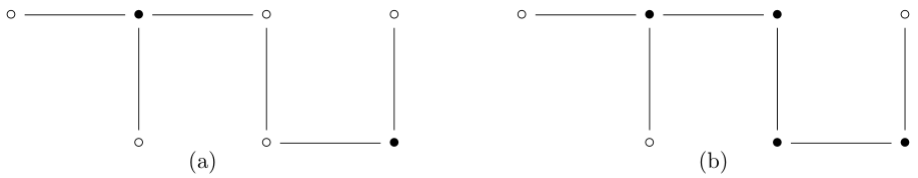


FIGURE 1.

2. Problem Description

The open neighborhood of a vertex v consists of the set of adjacent vertices to v, that is, $N(v) = \{w \in V | wv \in E\}$ and the closed neighborhood of is defined as $N[v] = N(v) \cup \{v\}$. The following labeling can be defined on V with respect to a subset $S \subseteq V$ as:

$$y(v_i) = \begin{cases} 1, & v \in S, \\ -1, & v \notin V. \end{cases}$$

For the sake of simplicity, we denote $y(v_i)$ by y_i and refer to a vertex with the label 1 as (+1)-vertex and as (-1)-vertex, otherwise. Further, $N(i)(N[i])$ stands

for the open (closed) neighborhood of the vertex v_i. It is important to mention that a vertex in a total dominating set S_t is a $(+1)$-vertex induced by S_t. From the definition of labelling, it is clear that the objective function is $\frac{1}{2}\sum_{i=1}^{n}(1+y_i)$. The next lemma gives us valid inequalities for total dominating set.

LEMMA 2.1. *$S_t \subseteq V$ is total dominating set if and only if it must satisfy in the following inequalities:*

$$\sum_{j\in N(i)}(1-y_iy_j)+\sum_{j\in N[i]}\frac{1+y_j}{2}\geq 2, \qquad i=1,2,\ldots,n. \tag{1}$$

Now, based on the (1), the quadratic integer programming model can be written as follows:

$$\begin{array}{lll} \min & \frac{1}{2}\sum_{i=1}^{n}(1+y_i) & \\ \text{s.t.} & \sum_{j\in N(i)}(1-y_iy_j)+\sum_{j\in N[i]}\frac{1+y_j}{2}\geq 2, & i=1,2,\ldots,n,\\ & y_i\in\{-1,+1\}, & i=1,2,\ldots,n. \end{array} \tag{2}$$

Observe that the objective functions of (2) and part of inequalities are linear, while analyzing of our algorithms needs a quadratic objective function. To convert these linear functions to quadratic ones, a reference variable $y_0 \in \{-1,+1\}$ is introduced and problem (2) is rephrased as follows:

$$\begin{array}{lll} \min & \frac{1}{2}\sum_{i=1}^{n}(1+y_0y_i) & \\ \text{s.t.} & \sum_{j\in N(i)}(y_0^2-y_iy_j)+\sum_{j\in N[i]}\frac{y_0^2+y_0y_j}{2}\geq 2, & i=1,2,\ldots,n,\\ & y_i\in\{-1,+1\}, & i=0,1,2,\ldots,n. \end{array} \tag{3}$$

Now suppose $\overline{y}=(y_0,y_1,\ldots,y_n)$ be the optimal solution of (3). If $y_0=+1$ then $y=(y_1,\ldots,y_n)$ is the optimal solution of (2) and if $y_0=-1$ then $y=(-y_1,\ldots,-y_n)$ is the optimal solution of (2).

3. Semidefinite Relaxation

First, for $i=0,1,\ldots,n$, the variable y_i is substituted by an $(n+1)$-dimensional vector $u_i\in\mathbb{U}$, where $\mathbb{U}=\{(+1,0,\ldots,0),(-1,0,\ldots,0)\}$. Accordingly, the restriction $y_i\in\{-1,+1\}$ is replaced by $u_i\in\mathbb{U}$ and then problem (3) is adapted as:

$$\begin{array}{lll} \min & \frac{1}{2}\sum_{i=1}^{n}(1+u_0^tu_i) & \\ \text{s.t.} & \sum_{j\in N(i)}(u_0^tu_0-u_i^tu_j)+\sum_{j\in N[i]}\frac{u_0^tu_0+u_0^tu_j}{2}\geq 2, & i=1,2,\ldots,n,\\ & u_i\in\mathbb{U}, & i=0,1,2,\ldots,n. \end{array} \tag{4}$$

Recall that $||u_i=1||$ for $u_i\in\mathbb{U}$ and this motivates to expand $\mathbb{U}$ to the standard $(n+1)$-dimensional unit sphere $\mathcal{S}^{n+1}=\{u\in\mathbb{R}^{n+1}|\ ||u||=1\}$, at the second step

of the relaxation procedure. Thus, the following problem is obtained

$$
(5)\quad \begin{aligned} \min \quad & \frac{1}{2}\sum_{i=1}^{n}(1+u_0^t u_i) \\ \text{s.t.} \quad & \sum_{j\in N(i)}(u_0^t u_0 - u_i^t u_j) + \sum_{j\in N[i]}\frac{u_0^t u_0 + u_0^t u_j}{2} \geq 2, \quad i=1,2,\ldots,n, \\ & u_i^t u_i = 1, \quad u_i \in \mathcal{S}^{n+1}, \qquad\qquad\qquad\qquad\quad i=0,1,2,\ldots,n. \end{aligned}
$$

By introducing $X_{ij} = y_i y_j, E_{ij} = e_i e_j^t$ and $A_i = \sum_{j\in N(i)} \frac{1}{2}(2E_{00} - E_{ij} - Eji) + \sum_{j\in N[i]} \frac{1}{4}(2E_{00} - E_{0j} - Ej0)$, where e_i is the i-th standard unit vector of $\mathbb{R}^{n+1}$, the model (5) is converted to the following:

$$
(6)\quad \begin{aligned} \min \quad & \frac{n}{2} + \langle C, X\rangle \\ \text{s.t.} \quad & \langle A_i, X\rangle \geq 2, \quad i=1,2,\ldots,n, \\ & X_{ii} = 1, \qquad i=0,1,2,\ldots,n, \\ & \operatorname{rank}(X) = 1, \\ & X \succeq 0, \end{aligned}
$$

where $C = (c_{ij}), c_{i0} = c_{0i} = \frac{1}{4}$ for $i = 1,\ldots,n$ and $c_{ij} = 0$ otherwise. By dropping the nonconvex constraint $\operatorname{rank}(X) = 1$ from (6), the semidefinite relaxation is formulated as:

$$
(7)\quad \begin{aligned} \min \quad & \frac{n}{2} + \langle C, X\rangle \\ \text{s.t.} \quad & \langle A_i, X\rangle \geq 2, \quad i=1,2,\ldots,n, \\ & X_{ii} = 1, \qquad i=0,1,2,\ldots,n, \\ & X \succeq 0, \end{aligned}
$$

The model (7) can be solved by interior point methods in CVX solver. Finally, the optimal solution of (7) just gives us a lower bound to total domination number. Because the total domination number has been not reported for graphs of a particular category, we solved a simple example with this method. The total domination number of graph in Figure 2 is 22 while after relaxing its model and rounding the obtained solution 100 times, geometric mean of these solutions is calculated 19.28 which is good approximation.

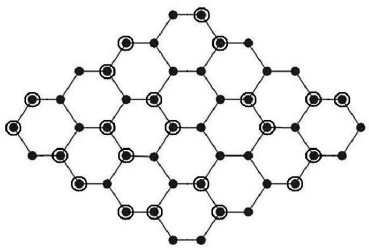


FIGURE 2.

4. Conclusion

In this paper, one of the most famous NP-complete problems in graph theory, the total dominating set problem, was investigated and a new quadratic integer programming model was presented. Finally, an SDP relaxation models are proposed. Finding the efficiency of the relaxation could be a future research direction.

References

1. F. Alizadeh, *Interior point methods in semidefinite programming with applications to combinatorial optimization*, SIAM J. Optim. **5** (1) (1995) 13–51.
2. M. F. Anjos and J. B. Lasserre (eds.), *Handbook on Semidefinite, Conic and Polynomial Optimization*, Springer US, New York, 2012.
3. J. Ceclio, J. Costa and P. Furtado, *Survey on data routing in wireless sensor networks,* in: Wireless Sensor Network Technologies for the Information Explosion Era, Springer Link, (2010) pp. 3–46.
4. P. A. Dreyer Jr., *Applications and Variations of Domination in Graphs*, Ph.D. Thesis, Rutgers University, 2000.
5. M. R. Garey and D. S. Johnson, *Computers and Intractability*, W. H. Freeman and Company, New York, 2002.
6. A. Ghaffari-Hadigheh and M. Djahangiri, *Semidefinite relaxation for the dominating set problem,* Iranian J. Oper. Res. **6** (1) (2015) 53–64.
7. M. X. Goemans and D. P. Williamson, *Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming,* J. ACM **42** (6) (1995) 1115–1145.
8. T. W. Haynes, S. Hedetniemi and P. Slater, *Fundamentals of Domination in Graphs,* CRC Press, New York, 1998.
9. M. A. Henning and A. Yeo, *Total Domination in Graphs,* Springer-Verlag, New York, 2013.
10. H. Wolkowicz, R. Saigal and L. Vandenberghe, *Handbook of Semidefinite Programming: Theory, Algorithms, and Applications,* Vol. 27, Springer, Boston, MA, 2000.

E-mail: djahangiri.mehdi@maragheh.ac.ir
E-mail: mohsen.ab@bonabu.ac.ir

A New Approach to Fuzzy Rough DEA Model

Maryam Joulaei*
Young Researchers and Elite Club, Yadegar-e-Imam Khomeini (RAH) Shahr-e-Rey Branch, Islamic Azad University, Tehran, Iran
Ali Shahabi
Young Researchers and Elite Club, Yadegar-e-Imam Khomeini (RAH) Shahr-e-Rey Branch, Islamic Azad University, Tehran, Iran
and Atefeh Armand
Young Researchers and Elite Club, Yadegar-e-Imam Khomeini (RAH) Shahr-e-Rey Branch, Islamic Azad University, Tehran, Iran

Abstract. In the real world, many data are inaccurate, and we are dealing with vague, unreliable, and inaccurate data. Measuring the performance of any creature in such uncertain conditions is inevitable. Fuzzy Rough Data Envelopment Analysis (FRDEA) provides the space to evaluate the relative performance of homogeneous organisms, known as decision units (DMUs) in the Envelopment Analysis (DEA) literature. In this paper, we used the data envelopment analysis model and assumed the performance measurements to be inaccurate. The aim of this paper is to convert the data envelopment analysis model with uncertain performance measurements into a crisp model, which is done using the principle of fuzzy expansion and the expected value of rough. Inaccurate assumption of performance measurements means fuzzy rough assumptions of inputs and outputs.
Keywords: Rough method, Data envelopment analysis, Fuzzy sets.
AMS Mathematical Subject Classification [2010]: 90C08, 03E72.

1. Introduction

Data Envelopment Analysis (DEA), as multifunctional the board and dynamic device has made significant advances in principle, technique, and applications on the planet today. This methodology is a direct nonparametric technique for estimating and assessing the general presentation of a bunch of units, first created by Charans et al. [1]. Customary information envelopment investigation models, for example, CCR and BCC don't manage ambiguous and incorrect information. In such models, it is accepted that all info and yield information is definitely indicated. In reality, although the information sources and yields are thought to be known, because of vulnerability, units containing missing data, judgment information, figure information, or consecutive inclination data may not be right. Rough set theory was first proposed by Pawlak [7] in 1982 as a conventional apparatus for displaying and preparing inadequate data in information systems. Rough set theory is generally utilized in numerous fields today. The main idea is to reduce the decision-making or classification of the rules of the problem by reducing the data

*Speaker

under the assumption of maintaining the same classification ability [8]. Rough set theory is an expansion of set hypothesis where a subset of a reference is characterized by a couple of successive sets called upper and lower approximations. A key concept in Rough Pawlak's collection model is the value relation. Fuzzy rough set was first studied by Dubois and Prade in [2, 3] and then studied by Morsi and Yakout [6], who defined the upper and lower approximations of the fuzzy set with respect to a fuzzy min-similarity relation. In this manuscript, in Section 2, the basic concepts of DEA's Rough and Rough Theory are presented. In Section 3, the answer to the RDEA model is obtained using interval programming and a method for ranking the efficiency interval is presented. Then, in Section 4, a method for solving the RFDEA model using the expected value using the fuzzy expander principle is presented. The conclusion is given in Section 5.

2. Basic Concepts

2.1. Rough Set Theory. The Rough set theory is proposed by Pawlak [7] which is an excellent mathematical tool for dealing with vague descriptions of objects. A basic assumption is that every object in the world is perceived through the information available, and this information may not be sufficient to pinpoint the exact object. One way to approximate a set is by using other sets. Thus a rough set may be defined by a pair of definite sets called the upper and lower boundaries.

Definition 2.1. [5] The collection of all sets having the same lower and upper approximations is called a Rough set, denoted by $(\underline{x}, \overline{x})$.

Note that the lower approximation is a subset containing the objects surely belonging to the set, whereas the upper approximation is a superset containing the objects possibly belonging to the set, and $\underline{x} \subset x \subset \overline{x}$. Let Λ be a nonempty set, A as a σ-algebra of subsets of Λ, and Δ an element in A, and Π a trust measure. Then is called a rough space.

Definition 2.2. [5] Let ξ be a rough variable, and $\alpha \in (0, 1]$. Then $\xi_{\sup}(\alpha) = \sup\{r | Tr\{\xi \geq r\} \geq \alpha\}$ is called the αoptimistic value to ξ, and

$$\xi_{\inf}(\alpha) = \inf\{r | Tr\{\xi \leq r\} \geq \alpha\},$$

is called the α-pessimistic value to ξ.

Theorem 2.3. [5] *Let $\xi_{\sup}(\alpha)$ and $\xi_{\inf}(\alpha)$ be the α-pessimistic and α-optimistic values of the rough variable ξ, respectively. Then we have*

- $Tr\{\xi \geq \xi_{\sup}(\alpha)\} \geq \alpha\, Tr\{\xi \leq \xi_{\inf}(\alpha)\} \geq \alpha$,
- $\xi_{\inf}(\alpha)$ *is an increasing and left-continuous function of* α,
- $\xi_{\sup}(\alpha)$ *is an decreasing and left-continuous function of* α,
- *if* $0 < \alpha \leq 1$, *then* $\xi_{\inf}(\alpha) = \xi_{\sup}(1 - \alpha)$ *and* $\xi_{\sup}(\alpha) = \xi_{\inf}(1 - \alpha)$,
- *if* $0 < \alpha \leq 0.5$, *then* $\xi_{\inf}(\alpha) \leq \xi_{\sup}(\alpha)$,
- *if* $0.5 < \alpha \leq 1$, *then* $\xi_{\inf}(\alpha) \geq \xi_{\sup}(\alpha)$.

where $\xi_{\inf}(\alpha)$ and $\xi_{\sup}(\alpha)$are an interval, the upper bound is $\xi_{\inf}(\alpha)$ and the lower bound is $\xi_{\sup}(\alpha)$denoted as $[\xi_{\sup}(\alpha), \xi_{\inf}(\alpha)]$.

LEMMA 2.4. *Let ξ be a trapezoidal fuzzy rough variable $\xi = (\bar{r}_1, \bar{r}_2, \bar{r}_3, \bar{r}_4)$, where $\bar{r}_1, \bar{r}_2, \bar{r}_3, \bar{r}_4$ are rough variables defined on $(\Lambda, \Theta, A, \pi)$, and*

$$\begin{aligned}
\bar{r}_1 &= [[P_2, P_3], [P_1, P_4]], \quad 0 < P_1 \leq P_2 \leq P_3 \leq P_4, \\
\bar{r}_2 &= [[q_2, q_3], [q_1, q_4]], \quad 0 < q_1 \leq q_2 < q_3 \leq q_4, \\
\bar{r}_3 &= [[s_2, s_3], [s_1, s_4]], \quad 0 < s_1 \leq s_2 < s_3 \leq s_4, \\
\bar{r}_4 &= [[t_2, t_3], [t_1, t_4]], \quad 0 < t_1 \leq t_2 < t_3 \leq t_4.
\end{aligned}$$

Then, the expected value of ξ is

$$E[\xi] = \frac{1}{16} \sum_{i=1}^{4} (P_i + q_i + s_i + t_i).$$

Now, the following results can be extracted.

3. General Model for Fuzzy Rough Expected Value Model (EVM)

Let a typical single objective problem with fuzzy rough parameters, as follows:

$$\begin{aligned}
&\max f(x, \xi) \\
&s.t. \quad g_i(x, \xi) \leq 0, \quad \forall i, \quad x \in X,
\end{aligned} \tag{1}$$

where $f(x, \xi)$ and $g_i(x, \xi), j = 1, \ldots, n$ are continuous functions in X and $\xi = (\xi_1, \ldots, \xi_n)$ is a fuzzy rough vector on the possibility space$(\Theta, P(\Theta), P)$. Then, it follows from the expected operator that

$$\begin{aligned}
&\max E[f(x, \xi)] \\
&s.t. \quad E[g_i(x, \xi)] \leq 0, \quad \forall i, \quad x \in X,
\end{aligned} \tag{2}$$

where E denotes the expected value operator of fuzzy rough variable. Using the expected value operator, model (2) has been converted into a certain programming and the DMs can easily obtain the optimal solution [4].

3.1. Deterministic Fuzzy Rough CCR Model with Expected Value Operator. In this subsection, we discuss the issue of evaluating the efficiencies of DMUs with fuzzy rough input and fuzzy rough output. Consider n DMUs, each of which consumes m different fuzzy rough inputs to secure s different fuzzy rough outputs. In addition, we presume that the fuzzy rough input $\tilde{\bar{x}}_{ij}, (i = 1, \ldots, m)$ and the fuzzy rough output $\tilde{\bar{y}}_{rj}, (r = 1, \ldots, s)$ are characterized, respectively, by the following two membership functions:

$$\mu_{\tilde{\bar{x}}_{ij}}(t) = \begin{cases} L(\frac{x_{ij}^{m_1} - t}{x_{ij}^{\alpha}}), & t \leq {x_{ij}}^{m_1}, \\ R(\frac{t - x_{ij}^{m_2}}{{x_{ij}}^{\beta}}), & t \geq {x_{ij}}^{m_2}. \end{cases}$$

Such $\mu_{\tilde{\bar{y}}_{rj}}(t)$ are obtained similarly, where

$$\begin{aligned}
x_{ij}^{m_1} &= ([x_{ij}^{(m_{1-a})}, x_{ij}^{(m_{1-b})}], [x_{ij}^{(m_{1-c})}, x_{ij}^{(m_{1-d})}]), \\
0 < x_{ij}^{(m_{1-c})} &\leq x_{ij}^{(m_{1-a})} \leq x_{ij}^{(m_{1-b})} \leq x_{ij}^{(m_{1-d})},
\end{aligned}$$

also $x_{ij}^{(m_2)}, y_{rj}^{(m_1)}, y_{rj}^{(m_1)}$ are created similarly. Below we consider the CCR model with fuzzy rough data.

$$
\begin{aligned}
&\max \sum_{r=1}^{s} u_r \tilde{\bar{y}}_{rp} \\
&S.t. \sum_{i=1}^{m} v_i \tilde{\bar{x}}_{ip} = 1, \\
&\sum_{r=1}^{s} u_r \tilde{\bar{y}}_{rj} - \sum_{i=1}^{m} v_i \tilde{\bar{x}}_{ij} \leq 0 \ \forall j, \quad u_r, v_i \geq 0, \ \ \forall\, (r,i).
\end{aligned} \tag{3}
$$

Model (3) involves fuzzy rough parameters and consequently, it cannot be optimized directly. Nevertheless, we employ the expected value operator to transform the fuzzy rough model into fuzzy rough EVM. Using the extension principle, the fuzzy information $\sum_{r=1}^{s} u_r \tilde{\bar{y}}_{rj}$ and $\sum_{i=1}^{m} v_i \tilde{\bar{x}}_{ij}$ can be written as follows

$$
\begin{aligned}
\sum_{i=1}^{m} v_i \tilde{\bar{x}}_{ij} &= (\sum_{i=1}^{m} v_i x_{ij}^{\alpha}, \sum_{i=1}^{m} v_i x_{ij}^{m_1}, \sum_{i=1}^{m} v_i x_{ij}^{m_2}, \sum_{i=1}^{m} v_i x_{ij}^{\beta}), \\
\sum_{r=1}^{s} u_r \tilde{\bar{y}}_{rj} &= (\sum_{r=1}^{s} u_r y_{rj}^{\alpha}, \sum_{r=1}^{s} u_r y_{rj}^{m_1}, \sum_{r=1}^{s} u_r y_{rj}^{m_2}, \sum_{r=1}^{s} u_r y_{rj}^{\beta}),
\end{aligned}
$$

where

$$
\begin{aligned}
\sum_{i=1}^{m} v_i x_{ij}^{m_1} &= ([\sum_{i=1}^{m} v_i x_{ij}^{m_1-a}, \sum_{i=1}^{m} v_i x_{ij}^{m_1-b}], [\sum_{i=1}^{m} v_i x_{ij}^{m_1-c}, \sum_{i=1}^{m} v_i x_{ij}^{m_1-d}]), \\
\sum_{i=1}^{m} v_i x_{ij}^{m_2} &= ([\sum_{i=1}^{m} v_i x_{ij}^{m_2-a}, \sum_{i=1}^{m} v_i x_{ij}^{m_2-b}], [\sum_{i=1}^{m} v_i x_{ij}^{m_2-c}, \sum_{i=1}^{m} v_i x_{ij}^{m_2-d}]), \\
\sum_{r=1}^{s} u_r y_{rj}^{m_1} &= ([\sum_{r=1}^{s} u_r y_{rj}^{m_1-a}, \sum_{r=1}^{s} u_r y_{rj}^{m_1-b}], [\sum_{r=1}^{s} u_r y_{rj}^{m_1-c}, \sum_{r=1}^{s} u_r y_{rj}^{m_1-d}]), \\
\sum_{r=1}^{s} u_r y_{rj}^{m_2} &= ([\sum_{r=1}^{s} u_r y_{rj}^{m_2-a}, \sum_{r=1}^{s} u_r y_{rj}^{m_2-b}], [\sum_{r=1}^{s} u_r y_{rj}^{m_2-c}, \sum_{r=1}^{s} u_r y_{rj}^{m_2-d}]).
\end{aligned}
$$

The deterministic CCR model based on the expected value approach is represented as follows.

$$
\begin{aligned}
&\max \sum_{r=1}^{s} u_r [(y_{rp}^{a-m_1} + y_{rp}^{b-m_1} + y_{rp}^{c-m_1} + y_{rp}^{d-m_1}) + (y_{rp}^{a-m_2} + y_{rp}^{b-m_2} + y_{rp}^{c-m_2} + y_{rp}^{d-m_2})] \\
&+4 \sum_{r=1}^{s} u_r (-y_{rp}^{\alpha} \int_0^1 L(t)dt + y_{rp}^{\beta} \int_0^1 R(t)dt)), \\
&s.t. \quad \sum_{i=1}^{m} v_i ((x_{ip}^{m_1-a} + x_{ip}^{m_1-b} + x_{ip}^{m_2-c} + x_{ip}^{m_1-d}) + (x_{ip}^{m_2-a} + x_{ip}^{m_2-b} + x_{ip}^{m_2-c} + x_{ip}^{m_2-d})) \\
&+4 \sum_{i=1}^{m} v_i (-x_{ip}^{\alpha} \int_0^1 L(t)dt + x_{ip}^{\beta} \int_0^1 R(t)dt) = 8,
\end{aligned}
$$

$$\sum_{r=1}^{s} u_r[(y_{rj}^{m_{1-a}} + y_{rj}^{m_{1-b}} + y_{rj}^{m_{1-c}} + y_{rj}^{m_{1-d}}) + (y_{rj}^{m_{2-a}} + y_{rj}^{m_{2-b}} + y_{rj}^{m_{2-c}} + y_{rj}^{m_{2-d}})]$$

$$4\sum_{r=1}^{s} u_r(-y_{rj}^{\alpha}\int_0^1 L(t)dt + y_r^{\beta}\int_0^1 R(t)dt) - \sum_{i=1}^{m} v_i((x_{ij}^{m_{1-a}} + x_{ij}^{m_{1-b}} + x_{ij}^{m_{2-c}} + x_{ij}^{m_{1-d}})$$

$$+(x_{ij}^{m_{2-a}} + x_{ij}^{m_{2-b}} + x_{ij}^{m_{2-c}} + x_{ij}^{m_{2-d}})),$$

$$(4)\quad 4\sum_{i=1}^{m} v_i(-x_{ij}^{\alpha}\int_0^1 L(t)dt + x_{ij}^{\beta}\int_0^1 R(t)dt) \le 0, \qquad \forall j, u_r, v_i \ge 0, \qquad \forall r, j.$$

4. Conclusion

In this paper, we develop a DEA model with fuzzy rough parameters to deal with uncertainty and imprecise to evaluate the relative efficiency of decision-making units under such conditions. This model is a tool to compare performance in such an uncertain and ambiguous environment. The said model under rough space has been addressed through fuzzy rough expected value operator and possibility approach to measure the relative efficiency of the DMUs. The novelty also lies in converting the fuzzy DEA model into its crisp equivalent DEA model by means of adopting concepts of fuzzy theory along with rough programming and by incorporating the α-optimistic and α-pessimistic values of rough variables to transform the rough programming model into the said crisp DEA model. The proposed approach provides insights for future research to address uncertainty in various other types of DEA models.

References

1. A. Charnes, W. W. Cooper and E. Rhodes, *Measuring the efficiency of decision making units*. European J. Oper. Res. **2** (6) (1978) 429–444.
2. D. Dubois and H. Prade, *Rough fuzzy sets and fuzzy rough sets*, Int. J. Gen. Syst. **17** (2–3) (1990) 191–209.
3. D. Dubois and H. Prade, *Twofold fuzzy sets and rough setssome issues in knowledge representation*, Fuzzy Sets and Syst. **23** (1) (1987) 3–18.
4. R. Khanjani Shiraz, V. Charles and L. Jalalzadeh, *Fuzzy rough DEA model: A possibility and expected value approaches*, Expert. Syst. Appl. **41** (2) (2014) 434–444.
5. B. Liu, *Uncertain Theory: An Introduction to its Axiomatic Foundation*, Springer, Berlin, 2004.
6. N. N. Morsi and M. M. Yakout, *Axiomatics for fuzzy rough sets*, Fuzzy Sets and Syst. **100** (1–3) (1998) 327–342.
7. Z. Pawlak, *Rough sets*, Int. J. Inf. Comput. Sci. **11** (1982) 341–356.
8. Q. M. Xiao and Z. L. Zhang, *Rough prime ideals and rough fuzzy prime ideals in semigroups*, Inf. Sci. **176** (6) (2006) 725–733.

E-mail: maryam.joulaei@yahoo.com
E-mail: shahabi@iausr.ac.ir
E-mail: atefeh.armand@ymail.com

Nonsmooth Quasiconvex Optimization Using Lower Global Subdifferential

Alireza Kabgani*
Mathematics Group, Department of Environment, Urmia University of Technology, Urmia, Iran
School of Mathematics, Institute for Research in Fundamental Sciences (IPM), P. O. Box 19395-5746, Tehran, Iran

Abstract. In this talk, some properties of the lower global subdifferential as a new notion in nonsmooth analysis are presented. Then, some KKT type optimality conditions in terms of lower global subdifferentials are derived for a quasiconvex constrained optimization problem.
Keywords: Quasiconvexity, Nonsmooth analysis, Global subdifferential, Global derivatives.
AMS Mathematical Subject Classification [2010]: 90C30, 90C26.

1. Introduction

We consider an optimization problem,

$$(P)\ \min f(x) \qquad s.t. \qquad g_i(x) \leq 0,\ i \in I := \{1, \ldots, m\},$$

where $f, g_i : \mathbb{R}^n \to \mathbb{R}$, $(i \in I)$ are quasiconvex functions. The set of feasible solutions of (P) is

$$K := \{x \in \mathbb{R}^n : g_i(x) \leq 0, i \in I\}. \tag{1}$$

For a given $\bar{x} \in K$, set $I(\bar{x}) := \{i \in I : g_i(x) = 0\}$. We recall that the function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be quasiconvex if for each $x, y \in \mathbb{R}^n$ and $\lambda \in [0, 1]$, $f(\lambda x+(1-\lambda)y) \leq \max\{f(x), f(y)\}$. Moreover, f is said to be strongly quasiconvex if for each $x, y \in \mathbb{R}^n$ with $x \neq y$ and $\lambda \in (0, 1)$, $f(\lambda x+(1-\lambda)y) < \max\{f(x), f(y)\}$. The study of quasiconvex functions and their properties has attracted great attention due to their applications in various scientific and technological areas such as mathematics, economics, image processing, machine learning, etc.; see [1, 2, 6] and the references therein. Quasiconvex functions are not necessary, differentiable, directionally differentiable or even continuous. Thus, to derive the KKT type optimality conditions for (P), we need to use some generalized derivative notions [3, 7, 8]. In this paper, we use the lower global subdifferential notion which is recently introduced by Lara and Kabgani [5]. The lower global subdifferential is defined based on the lower global directional derivative introduced in [4]. The rest of this section contains notations, preliminaries and basic definitions from generalized convexity and nonsmooth analysis. Section 2 provides necessary and sufficient conditions for ensuring optimality in (P).

*Speaker

Throughout this paper, $\mathbb{R}^n$ stands for an n-dimensional Euclidean space with Euclidean norm $\|\cdot\|$, and $\langle\cdot,\cdot\rangle$ for the standard inner product. Given a set $C \subseteq \mathbb{R}^n$, $\operatorname{co} C$, $\operatorname{cone} C$, $\operatorname{int} C$, and $\operatorname{cl} C$ denote the convex hull, convex cone hull, the interior, and the closure of C, respectively. The indicator function of C, denoted by $\iota_C(\cdot)$, is defined by $\iota_C(x) = 0$ if $x \in C$, and $\iota_C(x) = +\infty$ if $x \notin C$. The support function of C is defined by $\sigma_C(x) := \sup_{c\in C}\langle c, x\rangle$. The closed ball with center at x and radius $\delta > 0$ is denoted by $\mathbb{B}(x, \delta)$.

The cone of feasible directions and the tangent cone of C at $\overline{x} \in \operatorname{cl} C$, denoted by $D(C, \overline{x})$ and $T(C, \overline{x})$, respectively, are defined by

$$D(C,\overline{x}) := \{d \in \mathbb{R}^n :\ \exists\ \delta > 0,\ \forall\ \lambda \in (0,\delta),\ \overline{x} + \lambda d \in C\},$$
$$T(C,\overline{x}) := \{d \in \mathbb{R}^n :\ \exists\ t_n \downarrow 0,\ \exists\ \{d_n\} \subseteq \mathbb{R}^n,\ d_n \to d,\ \overline{x} + t_n d_n \in C\}.$$

If C is convex, then $\operatorname{cl} D(C, \overline{x}) = T(C, \overline{x})$. Furthermore, if C is convex, then the normal cone of C at $\overline{x} \in C$, is defined by:

$$N(C,\overline{x}) := \{d \in \mathbb{R}^n :\ \langle d, x - \overline{x}\rangle \le 0,\ \forall\ x \in C\}.$$

The polar cone of $C \subseteq \mathbb{R}^n$ is defined by

$$C^\circ := \{d \in \mathbb{R}^n : \langle d, x\rangle \le 0,\ \forall x \in C\},$$

It is seen that $N(C, \bar{x}) = T^\circ(C, \bar{x})$.

Definition 1.1. [4, Definition 3.1] Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper function and $\overline{x} \in \operatorname{dom} h$. Hence for every $\varepsilon > 0$, the lower global directional derivative of h at $\overline{x}$ in the direction $u \in \mathbb{R}^n$ is defined by:

$$h_\varepsilon(\overline{x}; u) := \inf_{0<t\le\varepsilon} \frac{h(\overline{x} + tu) - h(\overline{x})}{t}.$$

Definition 1.2. [5, Definition 3.2] Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper function and $\overline{x} \in \operatorname{dom} h$. Then for any $\varepsilon > 0$, the lower global subdifferential of h at $\overline{x} \in \operatorname{dom} h$ is defined by

$$\partial_\varepsilon h(\overline{x}) := \{\xi \in \mathbb{R}^n :\ h_\varepsilon(\overline{x}; u) \ge \langle \xi, u\rangle,\ \forall\ u \in \mathbb{B}(0,1)\},$$

The function h is called lower global subdifferentiable at $\overline{x} \in \operatorname{dom} h$ if $\partial_\varepsilon h(\overline{x}) \neq \emptyset$ for some $\varepsilon > 0$.

Clearly, $\partial_\varepsilon h(\overline{x})$ and $\partial^\varepsilon h(\overline{x})$ are closed and convex for all $\varepsilon > 0$ and all $\overline{x} \in \operatorname{dom} h$. Moreover, if h is convex, then $h_\varepsilon(\overline{x}; u) = h'(\overline{x}; u)$ by [4, Corollary 3.1 (a)], and from positively homogeneity of $h'(\overline{x}; \cdot)$, $\partial_\varepsilon h = \partial h$ for all $\varepsilon > 0$.

Definition 1.3. [5, Definition 4.1] Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper function and $\varepsilon > 0$. Then h is said to be ε-lower regular at $\overline{x} \in \operatorname{dom} h$ if for each $\varepsilon > 0$,

$$h_\varepsilon(\overline{x}; u) \ge 0,\ u\ \in \mathbb{R}^n \implies \sup_{\eta\in\partial_\varepsilon h(\overline{x})} \langle \eta, u\rangle \ge 0.$$

It is said that h is ε-lower regular on $\operatorname{dom} h$ if it is at every point on $\operatorname{dom} h$.

Theorem 1.4. [5, Theorem 3.1] *Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper function and $\overline{x} \in \operatorname{int}\operatorname{dom} h$. Then $\partial_\varepsilon h(\overline{x})$ is convex and compact for all $\varepsilon > 0$.*

2. Main Results

In this section, a sufficient optimality condition (Theorem 2.3) and a necessary optimality condition (Theorem 2.4) in terms of lower global subdifferentials of the objective function and the constraints functions are derived. In some part of the presented results, we use a generalized version of the Abadie constraint qualification.

DEFINITION 2.1. Assume that K is as presented in (1) and $\bar{x} \in K$. We say that Abadie Constraint Qualification (ACQ) holds at $\bar{x}$ if

$$\left(\bigcup_{i \in I(\bar{x})} \partial_\varepsilon g_i(\bar{x}) \right)^\circ \subseteq T(K, \bar{x}).$$

LEMMA 2.2. *Assume that K is as presented in (1) and $\bar{x} \in K$. Then, for each $\varepsilon \in (0, 1]$,*

$$\operatorname{cl}\operatorname{cone} \left(\bigcup_{i \in I(\bar{x})} \partial_\varepsilon g_i(\bar{x}) \right) \subseteq N(K, \bar{x}). \tag{2}$$

Moreover, if ACQ holds at $\bar{x}$, then the equality holds in (2) and

$$\left(\bigcup_{i \in I(\bar{x})} \partial_\varepsilon g_i(\bar{x}) \right)^\circ = T(K, \bar{x}).$$

PROOF. Let $i \in I(\bar{x})$ be arbitrary. Since K is a convex set and g_i is quasiconvex, for each $x \in K$, $(g_i)_\varepsilon(\bar{x}; x - \bar{x}) \leq 0$, for each $\varepsilon \in (0, 1]$. Thus,

$$\langle \zeta, x - \bar{x} \rangle \leq 0, \qquad \forall \left(x \in K, i \in I(\bar{x}), \zeta \in \partial_\varepsilon g_i(\bar{x}) \right).$$

Since $N(K, \bar{x})$ is a closed convex cone, (2) holds by (2).
Now, assume that ACQ holds at $\bar{x}$. We have,

$$\begin{aligned} \left(\bigcup_{i \in I(\bar{x})} \partial_\varepsilon g_i(\bar{x}) \right)^\circ \subseteq T(K, \bar{x}) \subseteq (T(K, \bar{x}))^{\circ\circ} \\ = (N(K, \bar{x}))^\circ \subseteq \left(\operatorname{cl}\operatorname{cone} \left(\bigcup_{i \in I(\bar{x})} \partial_\varepsilon g_i(\bar{x}) \right) \right)^\circ \\ = \left(\bigcup_{i \in I(\bar{x})} \partial_\varepsilon g_i(\bar{x}) \right)^\circ. \end{aligned}$$

□

In the following theorem, a sufficient optimality condition for (P) is derived.

THEOREM 2.3. *Assume that K is as presented in (1) and $\bar{x} \in K$. If f is strongly quasiconvex, $\varepsilon \in (0,1]$ and there exist $\lambda_1, \ldots, \lambda_m \geq 0$ such that*

$$0 \in \partial_\varepsilon f(\bar{x}) + \sum_{i=1}^{m} \lambda_i \partial_\varepsilon g_i(\bar{x}),$$
$$\lambda_i g_i(\bar{x}) = 0, \qquad \forall i \in I,$$

then $\bar{x}$ is an optimal solution of (P).

PROOF. By Lemma 2.2, we have

$$0 \in \partial_\varepsilon f(\bar{x}) + \sum_{i \in I(\bar{x})}^{m} \lambda_i \partial_\varepsilon g_i(\bar{x}) \subseteq \partial_\varepsilon f(\bar{x}) + N(K, \bar{x}).$$

Now, consider $\zeta \in \partial_\varepsilon f(\bar{x})$ such that $-\zeta \in N(K, \overline{x})$. Thus, $0 \leq \langle \xi, x - \overline{x} \rangle$ for each $x \in K$ and $\langle \xi, u \rangle \leq f_\varepsilon(\overline{x}; u)$ for each $u \in \mathbb{B}(0,1)$. Let $x \in K$ such that $\|x - \bar{x}\| \leq 1$. Since f is a strongly quasiconvex function and $\varepsilon \in (0,1]$,

$$\begin{aligned} 0 \leq \langle \xi, x - \overline{x} \rangle &\leq \inf_{0<t\leq\varepsilon} \frac{f(\overline{x} + t(x - \overline{x})) - f(\overline{x})}{t}, \\ \Rightarrow\ & f(\overline{x}) \leq f(\overline{x} + t(x - \overline{x})), \quad \forall\ t \in\ (0, \varepsilon], \\ \Rightarrow\ & f(\overline{x}) \leq f(x). \end{aligned}$$

Thus, $\bar{x}$ is a local minimum of (P). However, since f is strongly quasiconvex thus $\bar{x}$ is a minimizer of (P). □

The following theorem presents a necessary optimality condition for (P).

THEOREM 2.4. *Assume that K is as presented in (1) and $\bar{x} \in K \cap \operatorname{int} \operatorname{dom} f$. If ACQ holds at $\bar{x}$ and f is ε-lower regular function at $\bar{x}$, then there exist $\varepsilon > 0$ and $\lambda_1, \ldots, \lambda_m \geq 0$ such that*

$$0 \in \partial_\varepsilon f(\bar{x}) + \sum_{i=1}^{m} \lambda_i \partial_\varepsilon g_i(\bar{x}),$$
$$\lambda_i g_i(\bar{x}) = 0, \qquad \forall i \in I,$$

PROOF. First, we claim that for each $\varepsilon \in\ (0,1]$, we have

$$\sup_{\eta \in \partial_\varepsilon f(\overline{x})} \langle \eta, u \rangle \geq 0,\ \forall\ u \in D(K, \overline{x}).$$

Since $\bar{x}$ is a minimizer of (P) and f is quasiconvex, we have

$$f_1(\overline{x}; x - \overline{x}) = \inf_{0<t\leq 1} \frac{f(\overline{x} + t(x - \overline{x})) - f(\overline{x})}{t} \geq 0,\ \forall\ x \in K.$$

Then, for every $\varepsilon \in\ (0,1]$, $f_\varepsilon(\overline{x}; x - \overline{x}) \geq 0$ for all $x \in K$. Since f is ε-lower regular function at $\overline{x}$, we have

$$\begin{aligned} &\sup_{\eta \in \partial_\varepsilon f(\overline{x})} \langle \eta, x - \overline{x} \rangle \geq 0,\ \forall\ \varepsilon \in\ (0,1],\ \forall\ x \in K \\ &\Rightarrow \sup_{\eta \in \partial_\varepsilon f(\overline{x})} \langle \eta, \lambda(x - \overline{x}) \rangle \geq 0, \forall \varepsilon \in\ (0,1], \forall \lambda \geq 0, \forall x \in K. \end{aligned} \tag{3}$$

On the other hand, since K is a convex set

$$D(K,\overline{x}) = \{d \in \mathbb{R}^n : \exists\ x \in K,\ \exists\ \alpha \geq 0,\ d = \alpha(x - \overline{x})\}.$$

Hence, by Eq. (3), $\sup_{\eta \in \partial_\varepsilon f(\overline{x})} \langle \eta, u \rangle \geq 0$ for all $u \in D(K,\overline{x})$ and all $\varepsilon \in (0,1]$. Now, by [5, Theorem 4.2 (c)], we have

$$0 \in \partial_\varepsilon f(\overline{x}) + N(K,\overline{x}).$$

The proof is completed by Lemma 2.2. □

References

1. A. Agrawal and S. Boyd, *Disciplined quasiconvex programming*, Optim. Lett. **14** (2020) 1643–1657.
2. K. Hishinuma and H. Iiduka, *Fixed point quasiconvex subgradient method*, Eur. J. Oper. Res. **282** (2) (2020) 428–437.
3. N. Kanzi and M. Soleimani-Damaneh, *Characterization of the weakly efficient solutions in nonsmooth quasiconvex multiobjective optimization*, J. Global Optim. **77** (2020) 627–641.
4. F. Lara, *Optimality conditions for nonconvex nonsmooth optimization via global derivatives*, J. Optim. Theory Appl. **185** (2020) 134–150.
5. F. Lara and A. Kabgani, *On global subdifferentials with applications in nonsmooth optimization*, J. Glob. Optim. (2021). DOI:10.1007/s10898-020-00981-1
6. J. E. Martínez-Legaz, *Generalized Convex Duality and its Economic Applicatons*, In Handbook of generalized convexity and generalized monotonicity, (pp. 237–292), Springer, New York, 2005.
7. J. P. Penot, *Are Generalized Derivatives Sseful for Generalized Convex Functions?*, In: J. -P. Crouzeix, J. E. Martinez-Legaz, M. Volle, (eds.) *Generalized Convexity, Generalized Monotonicity: Recent Results*, pp. 3–59. Springer, Boston, 1998.
8. S. Suzuki, *Karush-Kuhn-Tucker type optimality condition for quasiconvex programming in terms of Greenberg-Pierskalla subdifferential*, J. Glob. Optim. **79** (2021) 191–202.

E-mail: a.kabgani@uut.ac.ir; a.kabgani@gmail.com

A Two-Step Benchmarking Approach in Value Efficiency Analysis

Nasim Nasrabadi*
Faculty of Mathematical Science and Statistics, University of Birjand, Birjand, Iran

Abstract. Basic Data Envelopment Analysis models are intrinsically preference-free. However, there exist several approaches for incorporating decision maker's preference(s) into the procedure of efficiency analysis; among them value efficiency analysis is one of the most practical approaches. In value efficiency analysis it is assumed that the decision maker has an implicit value function and he/she presents his/her preferences by means of determining the most preferred solution among all existing activities. Besides estimating the value efficiency score for each unit, value efficiency analysis is capable of setting benchmarks for value inefficient units. In this paper, we develop a two-step target setting approach in the framework of value efficiency analysis, in order to provide more realistically achievable targets.
Keywords: Benchmarking, Value efficiency, Value efficient frontier, Intermediate layer.
AMS Mathematical Subject Classification [2010]: 90B30, 90B50.

1. Introduction

Data envelopment analysis(DEA) is a non-parametric mathematical programming based approach for evaluating a set of homogeneous Decision Making Units (DMUs) with multiple inputs/multiple outputs [2, 3]. Recently, DEA has been developed and used extensively in many practical applications. For a complete history of DEA see [4].

One of the main properties of the original DEA models is that they are intrinsically preference-free. This means that they evaluate efficiency scores without taking into account any preferences among units or input and outputs. However, there are several approaches for incorporating preferences into the evaluation models [6], among them value efficiency analysis (VEA) is a well-known method. In VEA it is assumed that the decision maker (DM) has an empirical value function, which is pseudo-concave and increasing in outputs and pseudo-convex and decreasing in inputs. Moreover, he/she provides his/her preferences by means of providing his/her most preferred solution in the feasible technology. Based on these assumptions, the original VEA model in the DEA framework was formulated in [5].

On the other hand, as DEA is a powerful tool in benchmarking it is expected that the VEA model can also be used in the field of benchmarking. In this regard, a target setting approach based on VEA was developed in [7]. The main idea of this approach was to determine a target unit for each value inefficient unit which

*Speaker

is both feasible and value efficient. However, the set of all value efficient units does not necessarily form a complete envelopment for the DEA production possibility set, and may just include a small portion of the efficient frontier. Therefore, the target that that is obtained in [7] may be relatively far from the inefficient unit and so is difficult to achieve in a single step. In this regard, to develop a stepwise benchmarking approach in the framework of VEA seems necessary and useful. In this paper, we extend the model of [7] to a two-step benchmarking approach in order to obtain more realistic and achievable targets.

2. Basic Preliminaries

Suppose that we have n decision making units, where DMU_j uses input vector X_j to produce output vector Y_j, for $j = 1, \ldots, n$. As usual, we assume that X_j and Y_j are semi-positive m-vector and s-vector, respectively. The pair (X_j, Y_j) is called an observed activity, for $j = 1, \ldots, n$. The production possibility set in DEA for a constant returns to scale technology is formed as:

$$T = \{(X, Y) \mid X \geq \sum_{j=1}^{n} \lambda_j X_j, Y \leq \sum_{j=1}^{n} \lambda_j Y_j, \lambda_j \geq 0, j = 1 \ldots, n\}.$$

The formal definition of efficiency is given in the following.

Definition 2.1. A feasible activity $(\overline{X}, \overline{Y}) \in T$ is called efficient if and only if there does not exist any $(X, Y) \in T$ such that $(Y, -X) \geq (\overline{Y}, -\overline{X})$ and $(Y, -X) \neq (\overline{Y}, -\overline{X})$.

One of the well-known DEA models which can be suitably used for diagnosing efficient units is the basic additive model formulated as:

$$(1) \qquad \begin{array}{ll} \max & \sigma_o = 1^T S^- + 1^T S^+ \\ s.t. & \sum_{j=1}^{n} \lambda_j X_j + S^- = X_o, \\ & \sum_{j=1}^{n} \lambda_j Y_j - S^+ = Y_o, \\ & (S^-, S^+) \geq (0, 0), \lambda \geq 0. \end{array}$$

It can be easily shown that DMU_o is efficient if and only if the optimal value of (1) is equal to zero.

Let $E \subseteq \{1, \ldots, n\}$ denotes the set of all observed efficient units. According to [1], the set of all efficient activities is formulated as:

$$T^E = \{(X, Y) \mid \begin{array}{l} X = \sum_{j \in E} \lambda_j X_j, Y = \sum_{j \in E} \lambda_j Y_j, \\ -UY_j + VX_j - d_j = 0, \\ d_j \leq Mt_j, \lambda_j \leq M(1 - t_j), \\ t_j \in \{0, 1\}, \lambda_j, d_j \geq 0, j \in E \\ (U, V) \geq 1_{s+m}\}. \end{array}$$

Now to deal with value efficiency analysis, assume that the DM has an empirical value function which is pseudo-concave/-convex and increasing/decreasing in outputs/inputs and takes its maximum value at $(X^*, Y^*) \in T$. This means that $(X^*, Y^*) = \sum_{j=1}^{n} \lambda_j^*(X_j, Y_j)$ denotes the most preferred solution (MPS) of the DM. The additive-based value efficiency analysis model for evaluating DMU_o

is formulated as [7]:

$$\begin{array}{ll} \max & \rho_o = 1^T S^- + 1^T S^+ \\ s.t. & \sum_{j=1}^n \lambda_j X_j + S^- = X_o,\ \sum_{j=1}^n \lambda_j Y_j - S^+ = Y_o, \\ & (S^-, S^+) \geq (0,0), \lambda_j \geq 0, \quad \text{if} \quad \lambda_j^* = 0. \end{array}$$

Similarly, DMU_o is called value efficient if and only if the optimal value of the above model is equal to zero. If $VE \subseteq \{1, \ldots, n\}$ denotes the set of all observed efficient units, it is clear that $VE \subseteq E$. Moreover, the set of all value efficient activities is formulated as:

$$\begin{array}{ll} T^{VE} = \{(X,Y) \mid & X = \sum_{j\in VE} \lambda_j X_j, Y = \sum_{j\in VE} \lambda_j Y_j, \\ & -UY_j + VX_j - d_j = 0, \\ & d_j \leq Mt_j,\ \lambda_j \leq M(1-t_j),\ j \in VE, \\ & d_j, \lambda_j \geq 0,\ t_j \in \{0,1\},\ j \in VE,\ (U,V) \geq 1_{s+m}\}. \end{array}$$

It is not difficult to prove that $T^{VE} \subseteq T^E$.

3. Benchmarking

If DMU_o is value inefficient, i.e. at optimality $\rho_o > 0$ in model (3), it would be desirable to provide a corresponding target unit for it. The weighted benchmarking model which calculates the closest target for the value inefficient unit DMU_o is formulated as:

$$(2) \qquad \begin{array}{ll} \min & \sum_{i=1}^m |\frac{h_i^x}{x_{io}}| + \sum_{r=1}^s |\frac{h_r^y}{y_{ro}}| \\ s.t. & \sum_{j\in VE} \lambda_j X_j + h^x = X_o, \\ & \sum_{j\in VE} \lambda_j Y_j - h^y = Y_o, \\ & -UY_j + VX_j - d_j = 0, j \in VE, \\ & d_j \leq Mt_j,\ \lambda_j \leq M(1-t_j),\ j \in VE, \\ & d_j, \lambda_j \geq 0,\ j \in VE, \\ & (U,V) \geq 1_{s+m},\ (h^x, h^y) \quad \text{free}. \end{array}$$

Note that (2) and the model formulated in [7] both have the same feasible region which is in fact T^{VE} and the only difference between them is their objective functions. While the objective function in [7] is in fact the conventional norm-one of the slacks, in model (2) it has a weighted form. Solving (2) gives a value efficient target for DMU_o with minimum distance in terms of weighted norm-one. Moreover, it is worthwhile to note that problem (2) is in fact non-linear, due its objective function which is in the form of absolute value. However, it can be easily converted to a linear function using the following transformation for each $r \in \mathbb{R}$

$$\begin{cases} r = r^+ - r^-, \\ |r| = r^+ + r^-, \\ r^+, r^- \geq 0. \end{cases}$$

It is clear that T^{VE} does not necessarily form a complete envelope for the set T and therefore the target obtained from (2) maybe too far from DMU_o and it might be difficult to achieve in a single step. So, we need a two-step benchmarking procedure which is more practical. Here, we develop a two-step benchmarking approach which gives two targets for each value inefficient unit DMU_o, an ultimate target located on T^{VE} and an intermediate target. Towards this end, we need to formulate an intermediate layer. This layer is in fact formed by omitting the set

of all value efficient units, expect the MPS. To explain more, we assume that IVE contains all observed units which are value efficient after omitting the set of all value efficient units, except the MPS. Formally, the following algorithm calculates the set IVE:

Algorithm for IVE:

Step 1. Set $J' = (\{1, \ldots, n\} - VE) \cup \{MPS\}$.

Step 2. Solve the following problem for each $o \in \{1, \ldots, n\} - VE$:

$$
\begin{array}{rl}
\max & \rho'_o = 1^T S^- + 1^T S^+ \\
s.t. & \sum_{j \in J'} \lambda_j X_j + S^- = X_o, \\
& \sum_{j \in J'} \lambda_j Y_j - S^+ = Y_o, \\
& (S^-, S^+) \geq (0, 0), \\
& \lambda_j \geq 0, \quad \text{if} \quad \lambda^*_j = 0.
\end{array} \tag{3}
$$

Step 3. Set $IVE = \{j \mid \rho'_j = 0\}$.

Then, we form the intermediate layer T^{IVE} as:

$$
\begin{array}{ll}
T^{IVE} = \{(X, Y) \mid & X = \sum_{j \in IVE} \lambda_j X_j, Y = \sum_{j \in IVE} \lambda_j Y_j, \\
& -UY_j + VX_j - d_j = 0, \\
& d_j \leq Mt_j, \lambda_j \leq M(1 - t_j), \\
& d_j, \lambda_j \geq 0, t_j \in \{0, 1\}, j \in IVE, \\
& (U, V) \geq 1_{s+m}\}.
\end{array}
$$

Note that as the MPS belongs to both sets VE and IVE, both layers T^{VE} and T^{IVE} include the MPS. In other words, they intersects at the MPS. Now, assume that DMU_o is a value inefficient unit. If $o \in IVE$, then it has one final benchmark which is determined straightforwardly by solving (2). Otherwise, when $o \notin IVE$, we provide two sequential benchmarks for it, an intermediate benchmark on T^{IVE}, and a final benchmark on T^{VE}. Therefore, we develop the following two-step benchmarking approach for each DMU_o, where $o \notin VE \cup IVE$.

Two-Step Benchmarking Algorithm:

Step One. Intermediate Target Setting Solve model (2) by replacing VE by IVE to find the intermediate target (X^I_o, Y^I_o) as:

$$
(X^I_o, Y^I_o) = \sum_{j \in IVE} \lambda^*_j (X_j, Y_j) = (X_o - h^{x*}, Y_o + h^{y*}), \tag{4}
$$

where "*" stands for optimality of the corresponding problem.

Step Two. Final Target Setting Solve model (2) for (X^I_o, Y^I_o) instead of (X_o, Y_o) to find the final target (X^F_o, Y^F_o) as:

$$
(X^F_o, Y^F_o) = \sum_{j \in VE} \lambda^{**}_j (X_j, Y_j) = (X^I_o - h^{x**}, Y^I_o + h^{y**}), \tag{5}
$$

where "**" denotes optimality of the corresponding problem.

Note that as variables h^x and h^y are both unconstrained in model (2), none of the obtained targets, i.e. the intermediate target (4) and the final target (5), does not necessarily dominate DMU_o. However, considering the objective function of (2), one could claim that both are chosen as the closest target from the previous one w.r.t. weighted norm-one.

Example 3.1. Consider a set of 15 DMUs in a single input-single output technology operating under variable returns to scale. The corresponding data set is given in the first two rows of Table 1. Applying model (1) the set of efficient units is obtained as $E = \{1, 2, 3, 4, 5, 14, 15\}$. Now, assuming that the DM chooses unit 3 as the MPS, and by using the value efficiency model (3), the set of value efficient units is $VE = \{2, 3, 4, 15\}$. Moreover, using the given algorithm for IVE, the units on the intermediate layer is obtained as $IVE = \{3, 5, 14\}$. Note that unit 3 belongs to both VE and IVE. Finally, by implementing the proposed two-step benchmarking algorithm, we report the results in the four last rows of Table 1. Note that for each value efficient unit in VE there is no target, for units in the set IVE we have only one (final) target on T^{VE}, and for the other units we have two sequential targets, the intermediate on T^{IVE} and the final on T^{VE}. Figure (1) represents the technology along with the two sets T^{VE} and T^{IVE}.

Table 1. Data set and results.

	Units	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Units	Input	1	2	4	7	12	1.5	3.5	6.5	11	4	5.5	8.5	10	1.5	5
	Output	2	5	7.5	9	10	2.5	5.5	8	9.5	4	4.5	6.5	8	3.5	8
IVE	Input	1.5	–	–	–	–	1.5	2.75	6.5	11	1.812	2.125	8.5	10	–	–
	Output	3.5	–	–	–	–	3.5	5.5	8.281	9.688	4	4.5	8.906	9.375	–	–
VE	Input	2	–	–	–	7	2	2.75	6.5	7	2	2.125	7	7	2	–
	Output	5	–	–	–	9	5	5.937	8.75	9	5	5.156	9	9	5	–

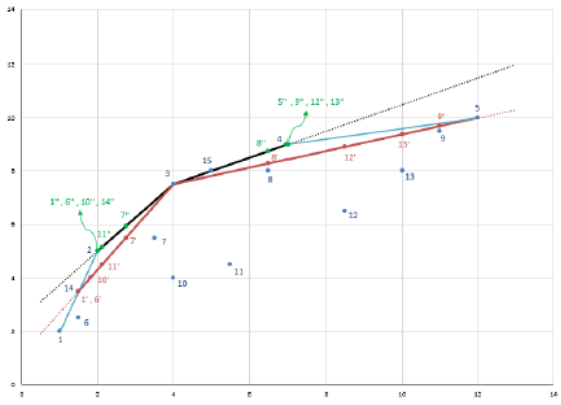


Figure 1. Illustration of value efficient frontier and intermediate layer.

References

1. J. Aparicio, J. M. Cordero and J. T. Pastor, *The determination of the least distance to the strongly efficient frontier in data envelopment analysis oriented models: Modelling and computational aspects*, Omega **71** (2017) 1–10.
2. R. D. Banker, A. Charnes and W. W. Cooper, *Some models for estimating technical and scale inefficiencies in data envelopment analysis*, Management Sci. **30** (9) (1984) 1078–1092.

3. A. Charnes, W. W. Cooper and E. Rhodes, *Measuring the efficiency of decision making units*, European J. Oper. Res. **2** (6) (1978) 429–444.
4. A. Emrouznejad and G. L. Yang, *A survey and analysis of the first* 40 *years of scholarly literature in DEA:* 1978 − 2016, Socio-Economic Planning Sci. **61** (2018) 4–8.
5. M. Halme, T. Joro, P. Korhonen, S. Salo and J. Wallenius, *A value efficiency approach to incorporating preference information in data envelopment analysis*, Management Sci. **45** (1) (1999) 103–115.
6. T. Joro and P. Korhonen, *Extension of Data Envelopment Analysis with Preference Information*, Springer, New York, 2015.
7. N. Nasrabadi, *A value efficiency-based target setting approach in data envelopment analysis*, J. New Res. Math. **5** (17) (2019) 51–72.

E-mail: nasimnasrabadi@birjand.ac.ir

An Efficient Trust Region Line Search Method for Solving the Unconstrained Optimization Problems

Zeinab Saeidian*
Faculty of Mathematical Sciences, University of Kashan, Kashan, Iran

Abstract. In this paper, we propose a new algorithm for solving unconstrained optimization problems. Using a modified definition of trust region ratio and an appropriate adaptive choice, an efficient adaptive nonmonotone scheme is provided. To avoid resolving the trust region subproblem whenever the trial step is rejected, we employ a line search strategy. Under some suitable and standard assumptions, the global convergence properties of the New Algorithm is established. Numerical experiments show the efficiency of the new proposed algorithm.
Keywords: Trust region methods, Nonmonotone adaptive technique, Line search method, Global convergence.
AMS Mathematical Subject Classification [2010]: 65K05, 90C30, 90C06.

1. Introduction

Consider the unconstrained optimization problem as follows:

$$min_{x\in\mathbb{R}^n} f(x), \tag{1}$$

in which $f : \mathbb{R}^n \to \mathbb{R}$ is a twice continuously differentiable function. For solving (1), many effective iterative procedures is provided, however, trust region and line search methods are two commonly used convergence schemes for unconstrained optimization [5]. In the procedure of line search methods, one moves along a (descent) direction as long as a sufficient reduction in the objective is achieved. On the other hand, in the standard trust region method, given x_k a trial step d_k is obtained by solving the following problem

$$\begin{aligned} \min q_k(d) &= f_k + g_k^T d + \frac{1}{2} d^T B_k d \\ s.t. &\quad \|d\| \le \Delta_k, \end{aligned} \tag{2}$$

where $f_k = f(x_k)$, $g_k = \nabla f(x_k)$, B_k is the exact Hessian $\nabla^2 f(x_k)$, or it's approximation, and $\Delta_k > 0$ is the trust region radius. Also, the agreement between the model and the objective function is measured by the trust region ratio that defined as follows:

$$r_k = \frac{f(x_k) - f(x_k + d_k)}{q_k(0) - q_k(d_k)}. \tag{3}$$

A good agreement between the model and the objective function is achieved when the ratio is closed to 1. In this case, the trust region radius is expanded for the next iteration and the trial step d_k is accepted, so we obtain the new

*Speaker

approximation point $x_{k+1} = x_k + d_k$. Otherwise, if the ratio is closed to 0 or negative, the iteration is not successful, so the trust region radius is decreased and the trial step is rejected. In such a situation, the trust region subproblem should be solved again in a smaller region.

In the classic trust region method, the objective function values were reduced monotonically. Using the monotone scheme may reduce the speed of convergence for some problems. To overcome this drawback the nonmonotone trust region method is developed [4]. The capital difference between the monotone and nonmonotone trust region method is based on the definition of the ratio r_k. One of the most efficient nonmonotone terms is proposed by Ahookhosh et al. [1] as follows

$$R_k = \epsilon_k f_{\ell(k)} + (1 - \epsilon_k) f_k, \tag{4}$$

where $f_k = f(x_k)$, $\epsilon_k \in [\epsilon_{\min}, \epsilon_{\max}] \subset [0, 1]$ and

$$f_{\ell(k)} = \max_{0 \le j \le M(k)} f_{k-j}, \tag{5}$$

that is the Grippo's nonmonotone term and $M(0) = 0$ and, for $k \geq 1$, $M(k) = \min\{k, M\}$, for given positive integer M.

The procedure of updating the trust region radius at every iteration has a crucial role in acheiving the global convergence, since the large amount of trust region radius increases the number of solving the subproblems. In addition, a small trust region radius increases the total number of iterations. Also, selecting the appropriate initial trust region radius is important. Shi and Gou in [8] defined an adaptive radius as $\Delta_k = -c^p \frac{g_k^T q_k}{q_k^T \tilde{B}_k q_k} \|q_k\|$, where $c \in (0, 1)$, p_k is a nonnegative integer and q_k satisfying $-\frac{g_k^T q_k}{\|g_k\| \|q_k\|} \geq \tau$, $\tau \in (0, 1)$ and $\tilde{B}_k = B_k + iI$ in which i is the smallest nonnegative integer that $q_k^T \tilde{B}_k q_k > 0$. In this paper, we introduce an efficient trust region method in which, when a trial step d_k is rejected, we avoid resolving the trust region subproblem.

The reminder of this paper is organized as follows: in Section 2, we present the structure of the new nonmonotone adaptive trust region method in details. The global convergence property is established in Section 3. Finally, the numerical results is reported in Section 4.

2. The New Algorithm

In this section, we deal with our modified nonmonotone trust region algorithm for solving the unconstrained optimization problems. This method combines the nonmonotone technique as proposed in [2] with an improved scalar approximation of the Hessian according to the modified secant equation as proposed in [3]. Also, if the trial step is rejected, using the line search method, we avoid to resolve the subproblem.

For given x_k, the trial step d_k is computed by (approximately) solving the following simple subproblem

$$\begin{aligned} \min q_k(d) &= g_k^T d + \frac{1}{2} d^T \gamma(x_k) d \\ s.t. \quad \|d\| &\leq \Delta_k, \end{aligned} \tag{6}$$

where $\gamma_k := \gamma(x_k)$ is a scalar approximation of the Hessian matrix. Biglari and Solimanpur in [3] proposed another simple subproblem in which the approximation of the Hessian at the current point x_k is computed by

$$\hat{\gamma}_k = \frac{4(f_{k-1} - f_k) + 3g_k^T d_{k-1} + g_{k-1}^T d_{k-1}}{d_{k-1}^T d_{k-1}}. \tag{7}$$

Whereas $\hat{\gamma}_k$, in (7), may become negative in some iterations, we modified γ_k as below [7]

$$\gamma_k = \frac{4(f_{k-1} - f_k) + (3 + \eta_k) g_k^T d_{k-1} + g_{k-1}^T d_{k-1}}{d_{k-1}^T d_{k-1}}, \tag{8}$$

where η_k in (8) is computed by

$$\eta_k = \begin{cases} \frac{4(f_k - f_{k-1}) - 3g_k^T d_{k-1} - g_{k-1}^T d_{k-1} + \delta}{{g_k}^T d_{k-1}}, & \text{if } \hat{\gamma}_k < 0, \\ 0, & \text{Otherwise,} \end{cases}$$

where δ is a small positive number and $\hat{\gamma}_k$ is defined by (7). The definition (8) implies that the scalar approximation of the Hessian became nonnegative.

Using d_k, from solving the subproblem (6), the nonmonotone trust region ratio is computed by [2]

$$r_k = \frac{R_k - f(x_k + d_k)}{Pred_k}, \tag{9}$$

where R_k is defined by (4) and $Pred_k = q_k(0) - q_k(d_k)$. For given $\mu \in (0, 1)$, the trial step is accepted whenever $r_k \geq \mu$; otherwise it is rejected.

If the trial step accepted, we set $x_{k+1} = x_k + d_k$ and we update the trust region radius appropriately. Otherwise, assuming that the function f is continuously differentiable and its gradient is Lipschitz continuous, consider [6] $\beta_k = \frac{-g_k^T d_k}{L_k \|d_k\|^2}$ in which L_k is an approximation of the Lipschitz constant L that here is updated by

$$\frac{-d_{k-1}^T y_{k-1}}{\|d_{k-1}\|^2}, \tag{10}$$

where $y_{k-1} = g_k - g_{k-1}$. From [10], the step size α_k is computed by

$$f(x_k + \alpha_k d_k) \leq R_k + \varphi \alpha_k (g_k^T d_k - \frac{1}{2} \alpha_k r L_k \|d_k\|^2), \tag{11}$$

where $\varphi \in (0, \frac{1}{2})$, and $r \in [0, +\infty)$ are real constants. Furthermore, the procedure of updating the adaptive trust region radius is as follows [7]

$$\Delta_k = \min \left\{ \nu_k \frac{\|g_k\|}{\gamma_k}, \Delta_{\max} \right\}, \tag{12}$$

where $\Delta_{\max} > 0$ is a threshold value for the radii and ν_{k+1} is updated by:

$$\nu_{k+1} = \begin{cases} \sigma_0 \nu_k, & r_k < \mu_1, \\ \nu_k, & \mu_1 \leq r_k \leq \mu_2, \\ \min\{\sigma_1 \nu_k, \nu_{\max}\}, & r_k > \mu_2, \end{cases} \tag{13}$$

where $0 < \sigma_0 < 1 < \sigma_1$, $0 < \mu_1 < \mu_2 \leq 1$ and $\nu_{\max} > 0$ are given numbers. We describe the structure of the proposed algorithm as below.

A New Algorithm

Input: $x_0 \in \mathbb{R}^n$, $0 < \mu < \mu_1 < \mu_2 \leq 1$, $0 < \sigma_0 < 1 < \sigma_1$, $0 < \epsilon_{\min} < \epsilon_{\max} < 1$, $\epsilon, \varepsilon, M, \nu_{\max}, \Delta_{\max} > 0$, $0 < \theta_1 < \theta_2$, the constant $t \in (0,1)$, $\sigma \in (0, \frac{1}{2}$, $r \in (0, \infty)$, $L_0 > 0$ and $\delta > 0$.

Step 0: Set $k = 0$, $\gamma_0 := \gamma(x_0) = 1$, $g_0 = g(x_0)$, $\nu_0 = 1$ and $\Delta_0 = \min\left\{\nu_0 \frac{\|g_0\|}{\gamma_0}, \Delta_{\max}\right\}$.

Step 1: **If** $\|g_k\| \leq \varepsilon$, **Then** Stop.

Step 2: Determine d_k by solving (6).

Step 3: Compute r_k using (9). **If** $r_k > \mu_2$, **Then** set $x_{k+1} = x_k + d_k$, Update ν_{k+1} using (13) and set $\Delta_{k+1} = \min\left\{\nu_{k+1}\frac{\|g_{k+1}\|}{\gamma_{k+1}}, \Delta_{\max}\right\}$ and goto Step 5.

Step 4: Update L_k by (10), find the step length α_k satisfying (11), and set $x_{k+1} = x_k + \alpha_k d_k$ and update Δ_k from (12).

Step 5: Compute the new Hessian approximation γ_{k+1} by (8). **If** $\gamma_{k+1} \leq \epsilon$, **Then** set $\gamma_{k+1} = \theta_1$. **If** $\gamma_{k+1} \geq \frac{1}{\epsilon}$, **Then** set $\gamma_{k+1} = \theta_2$, Set k = k + 1 and goto Step 1.

3. Convergence Analysis

For considering the global convergence, the following standard assumptions are needed:

A1: The set $\Omega = \{x \in \mathbb{R}^n | f(x) \leq f(x_0)\}$ is a closed bounded convex set, $f(x)$ is a twice continuously differentiable function in Ω and the function $\nabla f(x)$ is a Lipschitz continuous function on Ω.

A2: There exists a positive constant m such that $d^T \gamma_k d \geq m\|d\|^2, \forall d \in R^n, \forall k \in N$.

A3: The matrix γ_k is uniformly bounded, i.e. there exists a positive constant M_1 such that $\|\gamma_k\| \leq M_1, \forall k \in N$.

LEMMA 3.1. [7] *Assume that d_k is a solution of the problem (6). Then, one has*

$$Pred_k := q_k(0) - q_k(d_k) \geq \frac{1}{2}\|g_k\| \min\left\{\Delta_k, \frac{\|g_k\|}{\gamma_k}\right\}. \tag{14}$$

LEMMA 3.2. [6, 7] *Step 4 of the New Algorithm is well-defined.*

THEOREM 3.3. [6, 7] *Suppose that Assumption A1 to A3 holds and $\{x_k\}$ is the sequence generated by the New Algorithm. Then, the New Algorithm either stops at a stationary point or*

$$\liminf_{k\to\infty} \|g_k\| = 0. \tag{15}$$

4. Numerical Results

In this section, we focus on providing some computational results of applying the New Algorithm along with the following algorithms on some test problems in order to compare their performances:

NATRM: Algorithm 2.1 in [9].

FATRA: Algorithm 1 in [7].

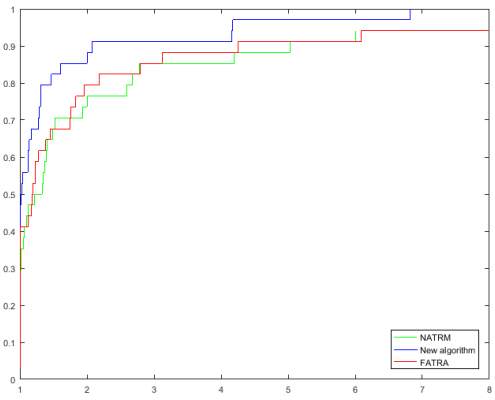


FIGURE 1. Performance based on function evaluations.

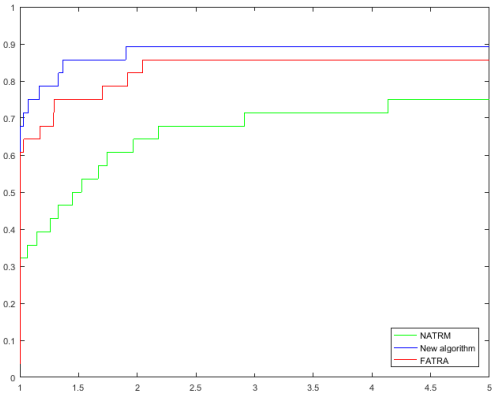


FIGURE 2. Performance based on the number of gradient.

Acknowledgement

The author would like to thank the University of Kashan for supporting this work.

References

1. M. Ahookhosh and K. Amini, *An efficient nonmonotone trust-region method for unconstrained optimization*, Numer. Algorithms **59** (2011) 523–540.
2. D. Ataee Tarzanagh, M. R. Peyghami and H. Mesgarani, *A new nonmonotone trust region method for unconstrained optimization equipped by an efficient adaptive radius*, Optim. Methods Softw. **29** (4) (2014) 819–836.
3. F. Biglari and M. Solimanpur, *Scaling on the spectral gradient method*, J. Optim. Theory Appl. **158** (2) (2013) 626–635.
4. L. Grippo, F. Lampariello and S. Lucidi, *A nonmonotone line search technique for Newton's method*, SIAM J. Numer. Anal. **23** (1986) 707–716.
5. J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, NewYork, 2006.
6. S. Rezaee and S. Babaei-Kafaki, *A modified nonmonotone trust region line search method*, J. Appl. Math. Comput. **57** (2018) 421–436.
7. Z. Saeidian and M. R. Peyghami, *An adaptive nonmonotone trust region method for unconstriained optimization problems based on a simple subproblem*, Iranian J. Nume. Anal. Optim. **5** (2) (2015) 95–117.
8. Z. Shi and J. Guo, *A new trust region method for unconstrained optimization*, J. Comput. Appl. Math. **213** (2008) 509–520.
9. Q. Zhou and C. Zhang, *A new nonmonotone adaptive trust region method based on simple quadratic models*, J. Appl. Math. Comput. **40** (2012) 111–123.
10. Z. Wan, S. Huang and X. D. Zheng, *New cautious BFGS algorithm based on modified Armijo-type linesearch*, J. Inequal. Appl. **241** (1) (2012) 1–10.

E-mail: Saeidian@Kashanu.ac.ir

Applying Game Theory in Tumor Growth Analysis

Atefeh Deris*
Faculty of Mathematical Sciences, Arak University, Arak, Iran
and Mahdi Sohrabi-Haghighat
Faculty of Mathematical Sciences, Arak University, Arak, Iran

Abstract. The behavior and growth of cancerous tumor is an interesting research subject and it has been widely analyzed from theoretical and empirical aspects. Various models have been applied to determine the growth pattern of cancerous tumor. In one of the current models, which we refer to as the competitive model, the tumor growth rate is determined based on the competition between the healthy and cancer cells. According to the effective application of this model in determining the tumor growth rate, some methods to get rid of the model restrictions are presented so that it can be used for tumor progression pattern. Finally, in order to evaluate the efficiency of the developed model, it has been implemented in some empirical examples.
Keywords: Cancerous tumor, Evolutionary game theory, Fitness, Growth rate.
AMS Mathematical Subject Classification [2010]: 13F55, 05E40, 05C65.

1. Introduction

Understanding the patterns of tumor growth is one of the important fields of study about the cancerous tumor. Different mathematical models have been introduced to explore these patterns. One of the simplest growth rules is the exponential growth model which indicates that the number of tumor cells doubles by a constant rate, meaning that the growth rate will always be constant, and thus its plot will be like a straight line in the semi-log plot [4, 5]. The exponential growth model was challenged by Gompertz in 1825, who stated that the doubling time of the tumor volume is not constant and the growth rate will decrease as the tumor volume increases. The Gompertz's growth curve is like sigmoid or S shape. The other model is the logistic growth which states that the growth rate reduces and finally reaches zero, when the population tends to the maximum carrying capacity [3]. Both models are sigmoid but Gompertz model indicates an exponentially decreasing growth rate, while in the logistic model, the growth rate decreases linearly proportional to the size of tumor.

But, these models are not sufficient to deal with the perturbed tumors (under treatment). Therefore, it is necessary to provide the biological interpretations of tumor growth mechanism with new approaches, so that it can be used in perturbed tumors regression. One of the models which has been recently introduced in this regard, and it will be referred to as the competitive model is designed based on the competition between the healthy and cancer cells [7, 8]. In this paper, we

*Speaker

intend to expand and develop the usage of the competitive model by providing some methods.

2. The Competitive Model Development for Tumor Progression Pattern

In the competitive model, the growth rate of tumor will be determined based on the competition result of the healthy and cancer cells. The healthy cells of a tissue or organism cooperate with the other cells, while the cancer cells defect the cooperation process with other cells, and each cell will receive an utility for its selected strategy. Assume that the following matrix is the game matrix of the healthy and cancer cells

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

A cancer cell will receive a utility c against a healthy cells, which is higher than the utility of the competition of a healthy cell with the other healthy cells, thus $c > a$. On the other hand, the healthy cells cooperate with each other unlike the cancer cells, therefore their utility will be more than the cancer cells in comparison, thus $a > d$. Because the healthy cells have less utility in contrast with the cancer cells, thus $d > b$. The inequalities $c > a > d > b$ recall that the matrix A is a prisoner's dilemma game matrix.

West et al. [7] introduced the amounts of $a = 3, b = 0, c = 5, d = 1$, and then simulated the tumor growth process (by considering the selection probability of each cell for the proliferation equal to the fitness of that cell). In order to develop the competitive model, two limitations are needed to be fixed. The first limitation is that the payoff matrix of each tumor is a function of that tumor's characteristics and the default values cannot be used.

To explain further, the real data of mammary tumor extracted from reference [2] were used. In Figure 1, we have illustrated the expected growth curve of cancer cells by using the constant elements and different selection intensities in the competitive model. Figure 1 clearly shows that the competitive model cannot simulate correctly the cancer cells growth by considering the mentioned constant elements and any selection intensity. In Figure 2, the previous data and two different selection intensities of 0.2 and 0.5 have been considered, but we have changed the elements of prisoner's dilemma matrix. As it can be seen in Figure 2, the growth rate of cancer cells has been approximated more accurately.

Figures 1 and 2 indicate that the matrix of prisoner's dilemma with the constant elements cannot be used for all tumors, and these elements should be selected based on the tumor behaviors.

We consider the matrix elements as parametric and let their values be determined according to the tumor characteristics that were observed in the clinical trials. Therefore, if V is the population volume of N cells (i.e. tumor carrying capacity) and v_t is the volume of tumor at time t, then the fitness of healthy and cancer cell mass with constant volume v at time t are determined by the following

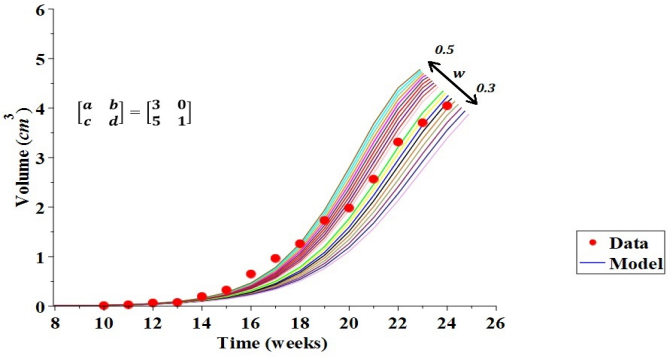
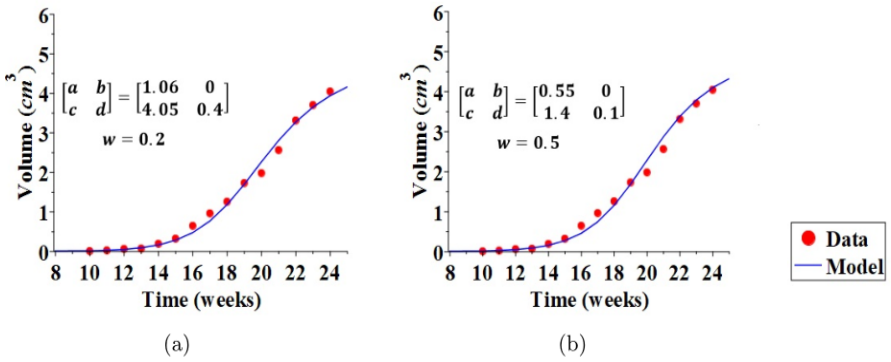


FIGURE 1. A tumor data and its simulation in the competitive model with respect to the constant elements of prisoner's dilemma matrix and different values of selection intensities.



(a) (b)

FIGURE 2. The previous tumor data (given in Figure 1) and its simulation in the competitive model with appropriate elements of prisoner's dilemma matrix and two different values of selection intensities.

formulas

$$f_t = 1 - w + wF_t,$$
$$g_t = 1 - w + wG_t,$$

where

$$F_t = \frac{a(V - v_t - 1) + b(v_t)}{V - 1},$$
$$G_t = \frac{c(V - v_t) + d(v_t - 1)}{V - 1},$$

and w is the intensity of selection which is a number in the interval $[0, 1]$, and it shows the effect of competition in the evolution process. Our empirical computations show that the numbers between 0.05 and 0.5 for the selection intensity will

lead to better results in order to calculate the growth rate and simulate the growth process in the competitive model. Considering the main idea of evolutionary game theory that considers the growth rate of species proportional to their fitness, the expected tumor volume in the next time is obtained by the following formula

$$v_{t+1} = \frac{V \cdot g_t \cdot v_t}{g_t \cdot v_t + (V - v_t) \cdot f_t}.$$

It is interesting to see that

$$\Delta v_t\big|_{t\to 0} \approx k v_t,$$

where $k = \dfrac{w(V(c-a)+(a-d))}{(V-1)(1-w+wa)}$. Based on the relationship between the matrix elements of the prisoner's dilemma problem, if $a > \dfrac{1-w}{w}$, then $k > 0$, which results an exponential growth of tumor in the initial stages of cancer cell growth process. In the following parts, we will let $b = 0$ and the condition $k > 0$ be provided.

To obtain the value of parameters a, b, c, and d, we use the clinical observations and solve a problem in order to implement the curve fitting process. Suppose that $(t_1, u_1), (t_2, u_2), \ldots, (t_k, u_k)$ are the clinical observations of tumor, where u_j is the tumor volume observed at time t_j $(j = 1, \ldots, k)$. Each time starting from (t_j, u_j), we obtain $v_{t_{(j+1)}}$ parametrically for $j = 1, \ldots, k-1$. Then, under the curve fitting process, we compute the parameters a, b, c, and d by solving the following problem

$$\min \sum_j (v_{t_j} - u_j)^2 \quad s.t. \quad c > a > d > b. \tag{1}$$

Problem (1) is a nonlinear programming problem with variables a, b, c, and d. Note that v_{t_j} is a nonlinear function on the variables a, b, c, and d, and u_j is the scalar value corresponding to the clinical observations. The optimal solution to this problem determines the prisoner's dilemma matrix elements.

If the number of clinical observations is sufficient, there is no need to constraints $c > a > d > b$ and the optimal solution of the unconstrained minimization problem will be satisfied these constrains.

As the variables of the non-linear programming problem (1) are limited only to 4 variables a, b, c and d, it can be solved by the standard mathematical softwares (especially because of the existence of appropriate initial solution such as $a = 3, b = 0, c = 5$, and $d = 1$ as well).

3. Experimental Examination of the Developed Competitive Model

In this section, we implement the developed competitive model presented in the previous section on the tumors data of 10 types of male mice with the lung cancer extracted from [1, 6]. Some experiments were conducted on ten mice, therefore, the data of each type of mouse is the average of the data belonging to the same type of mouse. The tumors data are measured in a period of 4 to 22 days on the mice with 6 to 8 weeks age. These data set display the volume range of 3 to $1449 mm^3$ (we consider 10^6 cells as $1 mm^3$ cells).

In order to show the capability of the competitive model to interpret and simulate the growth process of cancerous tumor , we have used the coefficient of determination (R^2) index to measure the accuracy of the model.

TABLE 1. The mean value (among all mice) of coefficient of determination (R^2) for different growth models.

Model	Exponential-linear	Gompertz	Generalized logistic	Power law	Exponential V_0	Von Bertalanffy	Dynamic CC	Logistic	Exponential 1	Competitive model
R^2	0.96	0.97	0.98	0.96	0.93	0.97	0.97	0.96	0.64	0.98

In [1], for the same dataset, the mean value (for all mice) of R^2 index has been calculated in different models, which have been presented in Table 1 along with the related value in the competitive model. As it can be seen, the competitive model and the generalized logistic model have the best performance in simulating the cancerous tumor growth.

These results indicate the good structure flexibility, which have provided the adaptation of competitive model with the real data.

References

1. S. Benzekry, C. Lamont, A. Beheshti, A. Tracz, J. M. Ebos, L. Hlatky and P. Hahnfeldt, *Classical mathematical models for description and prediction of experimental tumor growth,* PLoS Comput. Biol. **10** (8) (2014) e1003800.
2. S. A. Davie, J. E. Maglione, C. K. Manner, D. Young, R. D. Cardiff, C. L. MacLeod and L. G. Ellies, *Effects of FVB/NJ and C57Bl/6J strain backgrounds on mammary tumor phenotype in inducible nitric oxide synthase deficient mice*, Transgenic Res. **16** (2) (2007) 193–201.
3. L. Egghe and I. K. R. Rao, *Classification of growth models based on growth rates and its applications*, Scientometrics **25** (1) (1992) 5–46.
4. E. Frei and G. P. Canellos, *Dose: A critical factor in cancer chemotherapy*, Am. J. Med. **69** (1980) 585–594.
5. H. E. Skipper, F. M. Schabel Jr. and W. S. Wilcox, *Experimental evaluation of potential anticancer agents. xxi. on the criteria and kinetics associated with curability of experimental leukemia,* Cancer Chemother Rep. **35** (1) (1964) 1–111.
6. J. E. Talmadge, R. K. Singh, I. J. Fidler and A. Raz, *Murine models to evaluate novel and conventional therapeutic strategies for cancer,* Am. J. Pathol. **170** (3) (2007) 793–804.
7. J. West, Z. Hasnain, J. Mason and P. Newton, *The prisoner's dilemma as a cancer model*, Convergent Sci. Phys. Oncol. **2** (3) (2016) 035002.
8. J. West and P. K. Newton, *Chemotherapeutic dose scheduling based on tumor growth rates provides a case for low-dose metronomic high-entropy therapies*, Cancer Res. **77** (23) (2017) 6717–6728.
9. J. West and P. Newton, *Optimizing chemo-scheduling based on tumor growth rates,* bioRxiv (2018) 263327.

E-mail: m-haghighat@araku.ac.ir
E-mail: deris.atefeh@yahoo.com

Contributed Talks

Probability and Statistical Processes

A New Variant of the Three Towers Problem and its Simulation

Mehdi Sabzevari*
Faculty of Mathematical Sciences, University of Kashan, Kashan, Iran
Naser Noroozi
Faculty of Mathematical Sciences, University of Kashan, Kashan, Iran
and Hamid Ghorbani
Faculty of Mathematical Sciences, University of Kashan, Kashan, Iran

Abstract. In this paper, the three towers problem has been studied and a new definition of this problem has been proposed. With this new definition, an extension of the problem to n-towers is given. Finally, by means of simulation, the correctness of some derived formulas for some specific problems has been verified.

Keywords: Three towers problem, Gambler's ruin problem, Ruin time, n-Player gambler's ruin.
AMS Mathematical Subject Classification [2010]: 60G20, 60G50.

1. Introduction

The three towers problem is a rather old subject in probability theory, which due to its nature can be applied in many disciplines in science. This problem has been represented by Lennart Rade in 1972 as in the following [5]:

"We have three piles of chips, named A, B, and C, containing a, b, and c chips, respectively. We randomly select a pile (say A), then we randomly select another pile from the rest (say B) and we move one chip from A to B, in each round. We continue the rounds till one of the piles gets empty. This problem can be considered as an extension of the classical gambler's ruin problem. The objective is to find the expected duration of the rounds".

Twenty years after presentation of this problem, Arthur Engel, discovered a solution for it by turning the problem into a recursive equation. According to the procedure performed in [3], if $f(a,b,c)$ is the expected number of rounds, when the number of chips in the piles are a, b and c respectively, then

$$f(a,b,c) = 1 + \frac{1}{6} \sum_{(x,y,z)\in S} f(x,y,z), \tag{1}$$

where $S = \{(a,b+1,c-1),(a,b-1,c+1),(a+1,b,c-1),(a-1,b,c+1),(a+1,b-1,c),(a-1,b+1,c)\}$ and there exist boundary conditions of

$$f(a,b,0) = f(a,0,c) = f(0,b,c) = 0. \tag{2}$$

*Speaker

If we consider T as the duration for this process, by solving (1) and considering the boundary conditions stated in (2), the expected duration of the rounds will be obtained by the following:

$$\mathrm{E}(T) = \frac{3abc}{a+b+c}.$$

Recently, the authors of [1] have calculated the variance of the number of rounds, by using a similar method. In fact, the variance of the number of rounds when the initial amounts of the piles are a, b, and c chips respectively, is calculated as,

$$\mathrm{Var}(T) = \frac{3abc}{2(a+b+c)}\left(ab+ac+bc-1-\frac{6abc}{a+b+c}\right).$$

In this paper, first, we give a different definition of this problem and by providing a relation between this problem and the 3-player gambler's ruin problem, the expected number of rounds will be investigated. Then in the following sections, some extensions of the problem will be presented and studied.

2. A New Definition of Three Towers Problem

In this section we give a different definition of the three towers problem, which can be applied in other branches of sciences, as in the following:

There are three piles of chips called A, B, and C. In each round, one of the piles gets selected with the probability of $\frac{1}{3}$ and two chips from the other two (one from each) will be moved over to the former pile. These rounds continue till at least one of the piles gets empty. Note that in this variation, two piles can get empty simultaneously. This problem can be considered similar to the three players gamblers ruin problem, which has been studied by Sandell in [8]. Let $Z_i(n)$ be the amount of chips for pile i in round n, where $i = 1, 2, 3$ and $n = 1, 2, 3, \ldots$. By defining,

$$S_n = Z_1(n)Z_2(n)Z_3(n) + n(a+b+c-2), \tag{3}$$

and considering, $\sigma_n = \sigma\left[Z_i(k); i = 1, 2, 3, \quad k = 0, 1, 2, \ldots, n\right]$, to be the $\sigma-$algebra generated by the variables $Z_i(k)$ up to the round n, Sandell has shown that

- $\{S_n, \sigma_n\}$ is a martingale,
- The optional stopping theorem holds for $\{S_n\}$ and T, thus we have $\mathrm{E}(S_T) = \mathrm{E}(S_0)$.

On the other hand, if T denotes the first time that at least one of the piles gets empty, according to (3), $S_T = T(a+b+c-2)$, thus $\mathrm{E}(S_T) = (a+b+c-2)\mathrm{E}(T)$ and $S_0 = abc$. Consequently, for this situation, it yields that

$$\mathrm{E}(T) = \frac{abc}{a+b+c-2}. \tag{4}$$

3. The n-Towers Problem

By the definition presented in the pervious section, an extension of three towers problem, named the n-towers problem, can be proposed. In this extension, there exist n piles with $c_1, c_2, \ldots, c_n$ chips as amountes, respectively. In each round one of the piles such as pile i will be selected with the probability of $\frac{1}{n}$, and from the rest of $n-1$ piles, $n-1$ chips (1 from each) will be moved over to pile i. These rounds continue till at least one of the piles gets empty.

This problem can be considered equivalent to symmetric n-player gambler's ruin problem which is studied by Cho [2]. He derived a formula for $\mathrm{E}(T)$, but his formula was not in a simple closed-form and to be evaluated, requires the calculation of the probabilities of the various possible ruin states at time T, divided into appropriate symmetry classes. Nevertheless, the n-player gambler's ruin problem studied in [4] and [6] can be considered equal to the n-towers problem when the amount of chips are the same for all the piles. In the following, we study those.

3.1. Every Piles Has d Chips as Amount, and $1 \leq \mathbf{d} \leq \mathbf{n+1}$. Now we study the n-towers problem when the initial amount of every pile is d chips, $1 \leq d \leq n+1$. In each round one of the piles, say pile i, will be selected with the probability of p_i (not necessarily $p_i = \frac{1}{n}$), and receives one chip from every other pile, which is $n-1$ chips in total, for $i = 1, 2, \ldots, n$, $0 < p_i < 1$, $p_1 + \cdots + p_n = p$ and $0 < p \leq 1$. Also in each round, there is a chance for a tie with the probability $r = 1 - p$, $(0 \leq r < 1)$, and in this case no chip will be relocated. Obviously, when $r = 0$ there is no tie. This process continues till at least one of the piles gets empty. The n-player gambler's ruin problem corresponding to the above case has been studied in [4]. According to this paper, if E_i is the event of pile i getting empty at the end of the process, then for $m < n$:

$$P(E_{i_1} \cap E_{i_2} \cap \cdots \cap E_{i_m}) = \begin{cases} \frac{(p-p_{i_1}-p_{i_2}-\cdots-p_{i_m})^n}{p^n-\alpha}, & d = n, \\ \frac{(p-p_{i_1}-p_{i_2}-\cdots-p_{i_m})^{n+1}}{p(p^n-\beta)}, & d = n+1, \\ \frac{(p-p_{i_1}-\cdots-p_{i_m})^d}{p^d}, & 1 \leq d \leq n-1, \end{cases}$$

in which, $p = p_1 + p_2 + \cdots + p_n$, $\alpha = n! p_1 p_2 \ldots p_n$ and $\beta = \frac{n+1}{2}\alpha$.

Also the expected duration of the this process is,

$$\mathrm{E}(T) = \begin{cases} \frac{np^{n-1}}{p^n-\alpha}, & d = n, \\ \frac{np^{n-1}}{p^n-\beta} + \frac{1}{p}, & d = n+1, \\ \frac{d}{p}, & 1 \leq d \leq n-1. \end{cases} \tag{5}$$

Another point that can be obtained from this paper, is the independence of the events of "which pile gets emptied" and "the duration of the process".

Recently Sabzevari has studied the n-player gambler's ruin problem again [7]. According to his findings, variance of the duration of this process can be obtained as the following,

$$\mathrm{Var}(T) = \begin{cases} \frac{n^2\alpha p^{n-2}}{(p^n-\alpha)^2} + \frac{n(1-p)p^{n-2}}{p^n-\alpha}, & d = n, \\ \frac{n^2\beta p^{n-2}}{(p^n-\beta)^2} + \frac{n(1-p)p^{n-2}}{p^n-\beta} + \frac{1-p}{p^2}, & d = n+1, \\ \frac{d(1-p)}{p^2}, & 1 \leq d \leq n-1. \end{cases}$$

One interesting points expressed in this paper said, if there is no chance of tie, when the number of piles approaches infinity, the variance of the number of rounds for finishing the game approaches zero. In fact, if the number of piles

becomes large, the random variable T will be transformed into a degenerated random variable. In fact, when $n \to \infty$,

- $T \sim n$ if $d = n$,
- $T \sim n + 1$ if $d = n + 1$,
- $T \sim d$ if $1 \leq d \leq n - 1$.

But when there is chance of tie in the problem, the situation is totally different. If r is fixed and independent of n, then for $n \to \infty$, $\mathrm{Var}(T)$ approaches infinity too.

3.2. Every Piles Has the Amount of $n + c$ Chips, and $1 < \mathbf{c} \leq \mathbf{n}$. In this section we study the n-towers problem when the amount of each tower is $n+c$ chips, $(1 < c \leq n)$. In each round, one of the towers, say tower i, will be selected with the probability of p_i and it receives $n - 1$ chips from the other $n - 1$ towers, one from each, where $i = 1, \ldots, n$, $0 < p_i < 1$, and $p_1 + p_2 + \cdots + p_n = 1$. There is no chance of ties in this problem. In fact the corresponding n-players gamblers ruin problem has not been studied so far when there is a chance of ties.

According to findings of [6], there is no closed-form formula for calculating the expected duration for this process, and for its calculation, one should consider some factors such as the number of equivalence classes of piles amountes in different rounds. For example, when the initial amount of the towers are the same and equal to $n + 2$ chips:

$$\mathrm{E}(T) = 2 + n\left(\frac{1 + \alpha(n+2)(n+1)(3 - \alpha + p_1^2 + \cdots + p_n^2)}{4!\triangle}\right), \tag{6}$$

where $\alpha = n!p_1 \ldots p_n$ and

$$\triangle = 1 - \frac{\alpha(n^2 + 3n + 6)}{8} + \frac{\alpha^2(n+1)(n+2)}{4!}.$$

Also for the initial amount of $n+3$ chips for each tower, the expected of duration of the process has been studied as a matrix relation.

One of the interesting points which have been studied in this paper is the observation that when the initial amount of each tower is $n+2$ chips, which tower gets empty by the end of the process is not independent of the duration of the process, and this is in contrast with the previous case (amount of $n + 1$ chips for each tower).

4. Simulations

In this section, the following specific games have been considered. The aim here is to verify the correctness of the derived formulas in the pervious sections by simulation.

i) The three towers problem, see Section 2, with three piles containing five, seven, and eight chips, respectively. Using Eq. (4) yields $\mathrm{E}(T) = 15.556$.

ii) The six towers problem, see Section 3, when the initial amount of every pile is seven chips, with $r = \frac{1}{2^6}$, and

$$p_i = \frac{1}{2^i},\ i = 1, 2, \ldots, 6.$$

Using Eq. (5) yields $\mathrm{E}(T) = 7.119$.

iii) The six towers problem, see Section 3, when the initial amount of every pile is eight chips, with

$$p_i = \frac{1}{2^i},\ i = 1, 2, \ldots, 5,\ p_6 = \frac{1}{2^5}.$$

Using Eq. (6) yields $\mathrm{E}(T) = 8.032$.

Using `Maple12`, these three games were simulated N times. Table 1 shows the expected duration of these simulations. As expected, by increasing N, the number of simulated games, a good agreement between theoretical values of the expected rounds duration and their empirical counterparts are observed.

TABLE 1. The expected duration of running three mentioned games N times obtained by simulation.

N	game (i)	game (ii)	game (iii)
50	17.800	7.100	8.120
100	18.510	7.110	8.060
500	16.100	7.102	8.048
1000	16.107	7.104	8.036
5000	15.736	7.118	8.031
10000	15.634	7.125	8.033
50000	15.573	7.121	8.031
100000	15.572	7.120	8.030

References

1. J. Andel and S. Hudecova, *Variance of the game duration in gambler's ruin problem*, Statist. Probab. Lett. **82** (9) (2012) 1750–1754.
2. D. H. Cho, *A game with n players*, J. Korean Statist. Soc. **25** (2) (1996) 185–193.
3. A. Engel, *The computer solves three tower problem*, Amer. Math. Monthly **100** (1993) 62–64.
4. S. M. Hashemiparast and M. Sabzevari, *The asymmetric n-player gambler's ruin problem with ties allowed and simulation*, J. Korean Statist. Soc. **40** (3) (2011) 267–276.
5. L. Rade, *Take a Chance with your Calculator: Probability Problems for Programmable Calculators (DP series in calculators)* Dilithium Press, Forest Grove, Ore. 1977.
6. A. L. Rocha and F. Stern, *The asymmetric n-player gambler's ruin problem with equal initial fortunes*, Adv. Appl. Math. **33** (3) (2004) 512–530.
7. M. Sabzevari, *Variance of the asymmetric n-player gambler's ruin problem with ties allowed*, Comm. Statist. Simulation Comput. **47** (5) (2018) 1540–1549.
8. D. Sandell, *A game with three players*, Statist. Probab. Lett. **7** (1) (1989) 61–63.

E-mail: sabzevari@kashanu.ac.ir
E-mail: noroozi@kashanu.ac.ir
E-mail: hamidghorbani@kashanu.ac.ir

INAR(1) Model with Zero-and-One Inflated Poisson-Lindley Innovations

Zahra Sajjadnia*
Faculty of Statistics, Shiraz University, Shiraz, Iran
Zohreh Mohammadi
Department of Statistics, University of Jahrom, Jahrom, Iran
and Maryam Sharafi
Faculty of Statistics, Shiraz University, Shiraz, Iran

Abstract. In this paper, Zero and One inflated Poisson lindley distribution is introduced and some basic properties of it are obtained. The first order integer valued autoregressive model with zero and one inflated Poisson Lindley distributed innovations is presented. Some basic properties of this model are obtained and using the conditional maximum likelihood (CML) estimation method the model is fitted to the set of real data and by AIC and BIC criteria the goodness of fitting this model is demonstrated.
Keywords: INAR process, Poisson-Lindley distribution, Probability generating function, Zero and one inflated Poisson-lindley distribution.
AMS Mathematical Subject Classification [2010]: 62M10, 60E07.

1. Introduction

Time series models are one of the popular tools to analysis of time dependent data. For the data which is obtained from a random counting process, the useful time series is called an integer-valued time series. These time series are applied in different fields, such as economics, social sciences and life insurance.

At the first time, the integer valued autoregressive process of order one, which is called INAR(1) in brief, were introduced by Al-Osh and Al-Zaid (1987) [1]. In 1992, they introduced the INAR(1) model and presented some detailed discussion for the case in which the marginal distribution of the process is Geometric [2]. After that, many authors presented new integer valued processes and reviewed their properties.

Before 2010, almost all integer-valued time series which is introduced, used to model a non-negative and, consequently, non-symmetric counting observations. One of the major drawbacks of these models was that they could not be used to model both types of correlations (positive and negative). In 2010, Freeland introduced the correct symmetric stationary process, called the "true integer-valued autoregressive model of the first lag", which had a negative correlation, as well [4]. In 2019, Bourguignon et al. had extended the INAR(1) process with Poisson

*Speaker

innovations for modeling integer-valued time series with equidispersion, underdispersion, and overdispersion [3].

The Poisson-Lindley (PL) distribution, which is a compound Poisson distribution, has many properties which is useful for a good fit in some practical situations, Mohammadpour et al. (2018) considered this distribution as the marginal distribution of an INAR(1) process [6]. Along with introducing the Poisson-Lindley INAR(1) process, they established some of its properties.

In modeling count data there are some practical problems, where the number of zeros or ones exceeds or is less than a reference level and in this situations the common integer valued time series do not have an adequate fit. This shortcomming resulted in introducing zero-modified or zero and one modified distributed time series. Jazi et al. (2012a) studied the first order integer-valued AR processes with zero-inflated poisson innovations using binomial thinning operator [5]. X. Qi, et al. (2018) introduced an INAR(1) model with zero and one inflated Poisson distributed innovations [9], also M. Sharafi, et al. (2020) introduced an INAR(1) models with zero modified Poisson lindley innovations and denoted that this model is good for some real data [7]. Since the models based on PL distribution are useful for modeling real data, in some practical problems, where the number of zeros and ones exceeds, in this paper we are going to introduce a zero and one inflated INAR(1) Process with Poisson-Lindley distributed innovations.

The sections of the paper is organized as follows. In Section 1, we introduce the zero and one inflated Poisson lindley distribution and study some basic properties of it. In Section 2, we have introduced the INAR(1) model with zero and one inflated Poisson lindley innovations and explain about some basic properties of it. Finally, in Section 3, using CML estimation method we have fitted this model to the real data and have denoted the goodness of this model.

2. Preliminaries

In this section we introduce zero and one inflated Poisson lindley distribution and obtain some basic properties of it which are useful for the next sections.

Definition 2.1. The random variable Y is said to have zero and one inflated Poisson lindley distribution, denoted by $Y \sim ZOIPL(\phi_0, \phi_1, \theta)$ if its probability mass function is

$$P(Y=k) = \begin{cases} \phi_0 + (1-\phi_0-\phi_1)\frac{\theta^2(\theta+2)}{(1+\theta)^3}, & if \quad k=0, \\ \phi_1 + (1-\phi_0-\phi_1)\frac{\theta^2(\theta+3)}{(1+\theta)^4}, & if \quad k=1, \\ (1-\phi_0-\phi_1)\frac{\theta^2(\theta+2+k)}{(1+\theta)^{k+3}}, & if \quad k=2,3,\ldots, \end{cases}$$

where $0 \leq \phi_0$, $\phi_1 < 1$ and $\theta > 0$.

In the next theorem some basic properties of ZOIPL random variable are obtained.

Theorem 2.2. *Let Y be the ZOIPL random variable which is defined in the Definition 2.1. Then*

1) *the cumulative distribution function of* Y *is*

$$P(Y \leq k) = \begin{cases} 0, & if \quad k < 0, \\ \phi_0 + (1-\phi_0-\phi_1)\frac{\theta^2(\theta+2)}{(1+\theta)^3}, & if \quad 0 \leq k < 1, \\ \phi_0 + \phi_1 + (1-\phi_0-\phi_1)(1-\frac{\theta^2+([k]+3)\theta+1}{(1+\theta)^{[k]+3}}), & if \quad k \geq 1, \end{cases}$$

2) $E(Y) = \phi_1 + \phi_2 \frac{\theta+2}{\theta(1+\theta)}$, *where* $\phi_2 = 1-\phi_0-\phi_1$,
3) *the variance of* Y *is obtained as*

$$Var(Y) = \phi_1(1-\phi_1) + \phi_2 \frac{\theta^3+5\theta^2+10\theta+6}{\theta^2(1+\theta)^2} - \phi_2^2 \frac{(\theta+2)^2}{\theta^2(1+\theta)^2} - 2\phi_1\phi_2 \frac{\theta+2}{\theta(\theta+1)},$$

4) *the probability generating function of* Y *is*

$$\begin{aligned} \psi_Y(s) &= \phi_0 + \phi_1 s + (1-\phi_0-\phi_1)\frac{\theta^2(2+\theta-s)}{(1+\theta)(1+\theta-s)^2} \\ &= \phi(sp+1-p) + (1-\phi)\frac{\theta^2(2+\theta-s)}{(1+\theta)(1+\theta-s)^2}, \end{aligned}$$

where $\phi = \phi_0 + \phi_1$ *and* $p = \frac{\phi_1}{\phi_0+\phi_1}$.

We omitted the proof of this theorem because of page limitation. In the next section we introduce the INAR(1) model with ZOIPL distributed innovations.

3. The INAR(1) Model with Zero and One Inflated Poisson Lindley Distributed Innovations

In this section we introduce an INAR(1) process with zero and one inflated Poisson Lindely distributed innovations which is denoted by ZOPLINAR(1) and discuss about some basic properties of it. To do this, at first we present the definition of binomial thining operator introduced by Steutel and Van Harn (1979) [8].

Definition 3.1. (Binomial thinning operator) Let X be an arbitrary non-negative integer-valued random variable. Also, let Y_i be a sequence of independent and identically distributed Bernoulli random variables with success probability α. Then, for any $\alpha \in (0,1)$, the binomial thinning operator "$\circ$" is defined as

$$\alpha \circ X = \sum_{i=1}^{X} Y_i.$$

Moreover, for every i, Y_i is considered to be independent of X.

In this paper, we are going to introduce a new INAR(1) process based on the above thining operator.

Definition 3.2. Suppose that $\{X_t\}_{t\in N}$ is the INAR(1) process defined by

$$X_t = \alpha \circ X_{t-1} + \epsilon_t, \ t = 1, 2, \ldots, \tag{1}$$

where $\alpha \in (0,1)$, $\circ$ is the binomial thinning operator. This process is called ZOIPLINAR(1) process if the sequence $\{\epsilon_t\}_{t=0}^{\infty}$ is a sequence of iid $ZOIPL$ random variables, where ϵ_t is independent of X_{t-1} for all $t \geq 1$.

This model which is based on ZOIPL distribution are applied for modeling real data in some practical problems, where the number of zeros and ones are greater than the number which is expected from the Poisson Lindley distribution. In the next theorem we study some basic properties of the ZOIPLINAR(1) processes. In this part we are omitted the proof too.

THEOREM 3.3. *Let* $\{X_t\}$ *be the* ZOIPLINAR(1) *process defined by* (1). *Then*

1) $E(X_t|X_{t-1}) = \alpha X_{t-1} + E(\epsilon_t)$,
2) $Var(X_t|X_{t-1}) = \alpha(1-\alpha)X_{t-1} + Var(\epsilon_t)$,
3) $E(X_t) = \frac{E(\epsilon_t)}{1-\alpha}$,
4) $Var(X_t) = \frac{\alpha E(\epsilon_t)+Var(\epsilon_t)}{1-\alpha^2}$,
5) $Corr(X_t, X_{t-1}) = \alpha$,
6) *the probability generating function of* X_t *is*

$$\psi_{X_t}(s) = \psi_{X_{t-1}}(1-\alpha+\alpha s)\psi_{\epsilon_t}(s),$$

where $\psi_{\epsilon_t}(s)$ *is introduced in the Theorem* 2.2,

7) *the transition probabilities are*

$$P_{ij} = P(X_t = j|X_{t-1} = i) = \sum_{k=0}^{min(i,j)} \binom{i}{k} \alpha^k (1-\alpha)^{i-k} P(\epsilon_t = k), \tag{2}$$

where $P(\epsilon_t = k) = \phi_0 I_{\{0\}}(k) + \phi_1 I_{\{1\}}(k) + (1-\phi_0-\phi_1)\frac{\theta^2(\theta+2+k)}{(1+\theta)^{k+3}}$,

8) *the marginal and joint probability function of* X_t *is as follows:*

$$P_j = P(X_t = j) = \sum_{i=0}^{\infty} P_{ij} P_i,$$

where

$$\begin{aligned} f_{(X_1,\ldots,X_n)}(x_1,\ldots,x_n) &= P_{x_1} \prod_{i=1}^{n-1} P_{x_i x_{i+1}} \\ &= P_{x_1} \prod_{i=1}^{n-1} \left[\sum_{k=0}^{\min(x_i,x_{i+1})} \binom{x_i}{k} \alpha^k (1-\alpha)^{x_i-k} P(\epsilon_t = x_{i+1}-k) \right]. \end{aligned}$$

About the dispersion of the model we have the following remark.

REMARK 3.4. The variance of X_t is bigger than the mean of X_t if and only if the variance of ϵ_t is bigger than the mean of ϵ_t.

4. Conditional Maximum Likelihood Estimation

In the study of integer-valued time series, different estimation methods are applied. In this section, we are going to estimate the parameters of the ZOIPLINAR(1) model using conditional maximum likelihood (CML) estimation methods.

By maximizing the following conditional log-likelihood function over the parameter space, the conditional maximum likelihood estimator of $\lambda = (\alpha, \theta, \phi_0, \phi_1)$ is obtained.

$$l(\lambda) = \sum_{t=2}^{n} \log P(X_t = j \mid X_{t-1} = i),$$

where $P(X_t = j \mid X_{t-1} = i) = P_{ij}$ is the transition probabilities, presented in Eq. (2). Since there is no closed form for the CML estimates, the maximizers are achieved using numerical methods.

5. Application

In this section the data set of the monthly number of cases of polio reported by the U.S. Centers for Disease Control for the years 1970 up to 1983 is considered. There is 168 observations in this data set for which empirical mean and variance of them are 1.33 and 3.53, respectively. There are 64 zeros, which is 37.87 percent, and 55 ones, which is 32.74 percent of the observations. Using conditional maximum likelihood estimators (CMLE) we are fitted three models INAR(1), ZOINAR(1)(INAR(1) model with zero and one inflated Poisson distributed innovations) and ZOIPLINAR(1) (INAR(1) model with zero and one inflated Poisson lindley distributed innovations) to the data and as can be seen in the table below, the ZOIPLINAR(1) model has the best fitness. Because of page limitation in this abstract, we eliminated more details of this part and summaries our results in the next table.

TABLE 1. The CMLE of the parameters, AIC and BIC criteria for the Polio data.

Model	CMLE	AIC	BIC
INAR(1)	$\hat{\lambda}$=1.100 $\hat{\alpha}$=0.1848	582.1259	588.3738
ZOINAR(1)	$\hat{\lambda}$ =2.8967 $\hat{\phi_0}$=0.4011 $\hat{\phi_1}$=0.2960 $\hat{\alpha}$=0.1285	544.3156	556.8115
ZOIPLINAR(1)	$\hat{\phi_0}$=0.1322 $\hat{\phi_1}$=0.1466 $\hat{\alpha}$=0.0889 $\hat{\theta}$=1.0242	531.1109	533.3588

References

1. M. A. Al-Osh and A. A. Alzaid, *First-order integer-valued autoregressive* (INAR(1)) *process,* J. Time Ser. Anal. **8** (3) (1987) 261–275.
2. M. A. Al-Osh and A. A. Aly, *First order autoregressive time series with negative binomial and geometric marginals*, Commun. Stat. **21** (9) (1992) 2483–2492.
3. M. Bourguignon, J. Rodrigues and M. Santos-Neto, *Extended poisson inar*(1) *processes with equidispersion, underdispersion and overdispersion*, J. Appl. Stat. **46** (1) (2019) 101–118.
4. R. K. Freeland, *True integer value time series*, AStA Adv. Stat. Anal. **94** (3) (2010) 217–229.
5. M. A. Jazi, G. Jones and C. D. Lai, *First-order integer-valued AR processes with zero-inflated poisson innovations*, J. Time Series Anal. **33** (6) (2012) 954–963.
6. M. Mohammadpour, H. S. Bakouch and M. Shirozhan, *Poisson-Lindley* INAR(1) *model with applications*, Braz. J. Probab. Stat. **32** (2) (2018) 262–280.

7. M. Sharafi, Z. Sajjadnia and A. Zamani, *A first order integer-valued autoregressive process with zero modified Poisson-lindley distributed innovations*, Commun. Stat. (2020) accepted.
8. F. M. Steutel and K. Van Harn, *Discrete analogues of self-decomposability and stability*, Ann. Probab. **7** (5) (1979) 893–899.
9. X. Qi, Q. Li and F. Zhu, *Modeling time series of count with excess zeros and ones based on* INAR(1) *model with zero-and-one inflated Poisson innovations*, J. Comput. Appl. Math. **346** (2018) 572–590.

E-mail: sajjadnia@shirazu.ac.ir
E-mail: z.mohammadi@jahromu.ac.ir
E-mail: msharafi@shirazu.ac.ir

Reliability Analysis for a Class of an Exponential Distribution Based on Progressive First-Failure Censoring

Kambiz Ahmadi*
Department of Computer Sciences, Faculty of Mathematical Sciences, Shahr-e-kord University, Shahr-e-kord, Iran

Abstract. Based on progressively first-failure censored data, the problem of estimating parameters as well as reliability and hazard rate functions for a class of an exponential distribution is considered. The classic and Bayes approaches are used to estimate the parameters. The maximum likelihood estimates and exact confidence interval as well as exact confidence region for parameters are developed based on this censoring scheme. Also, when the parameters have discrete and continuous priors, several Bayes estimators with respect to squared error and linear-exponential (Linex) loss functions are derived. Finally, a real data analysis is presented to illustrate the methods of inference developed in this paper.

Keywords: Bayes estimator, Confidence region, Exponential distribution, Maximum likelihood estimator, Progressive first-failure censoring scheme.
AMS Mathematical Subject Classification [2010]: 62N01, 62N02.

1. Introduction

In many life test studies, it is common that the lifetimes of the test units may not be able to record exactly. Censoring is very common in reliability data analysis, in the past several decades. It usually applies when the exact lifetimes are known for only a portion of the products and the remainder of the lifetimes has only partial information. In some cases, the lifetime of products is quite long and so the experimental time of the progressive type-II censoring scheme can still be too long. In order to give an efficient experiment, the other test methods are proposed by statisticians, where one of them is the progressive first-failure censoring scheme (See [10]). It can be described as follows.

Suppose that n independent groups with k items within each group are put on a life test and experimenter decides beforehand the quantity m, the number of units to be failed. At the time of the first failure, $X^{\mathbf{r}}_{1;m,n,k}$, r_1 groups and the group in which the first failure is observed are randomly removed. r_2 groups and the group with observed failure are randomly removed as soon as the second failure, $X^{\mathbf{r}}_{2;m,n,k}$, has occurred. The procedure is continued until all r_m groups and the group with observed failure are removed at the time of the m-th failure, $X^{\mathbf{r}}_{m;m:k:n}$. Then $X^{\mathbf{r}}_{1;m,n,k} < X^{\mathbf{r}}_{2;m,n,k} < \cdots < X^{\mathbf{r}}_{m;m,n,k}$ are called progressively first-failure censored order statistics with the censoring scheme $\mathbf{r} = (r_1, r_2, \ldots, r_m)$. To simplify the notation, we will use X_i in place of $X^{\mathbf{r}}_{i;m,n,k}$. For a review of recent

*Speaker

developments and further discussions as well as applications of the progressive first-failure censoring scheme, we refer to [3, 6] and [8].

Suppose the lifetime random variable T has a continuous distribution with two parameters as α and λ, and with the pdf and cdf as

$$f(t;\alpha,\lambda) = \alpha\psi(t;\lambda)\exp\{-\alpha\Psi(t;\lambda)\},\ 0<t<\infty, \tag{1}$$

$$F(t;\alpha,\lambda) = 1-\exp\{-\alpha\Psi(t;\lambda)\}, \tag{2}$$

where $\psi(t;\lambda)=\frac{\partial\Psi(t;\lambda)}{\partial t}$, $\Psi(t;\lambda)$ is increasing in t with $\Psi(0;\lambda)=0$ and $\Psi(\infty;\lambda)=\infty$. The corresponding reliability and hazard rate functions become:

$$R(t) = \exp\{-\alpha\Psi(t;\lambda)\}, \qquad h(t)=\alpha\psi(t;\lambda), \tag{3}$$

respectively. This general form for lifetime model including some well-known and useful models such as Burr XII distribution with $\Psi(t;\lambda)=\ln(1+t^{\lambda})$, Pareto distribution with $\Psi(t;\lambda)=\ln t-\ln\lambda$, $t>\lambda$, Gompertz distribution with $\Psi(t;\lambda)=\frac{e^{\lambda t}-1}{\lambda}$, Weibull distribution with $\Psi(t;\lambda)=t^{\lambda}$, two parameters Rayleigh distribution with $\Psi(t;\lambda)=(t-\lambda)^2, t>\lambda$, two parameters bathtub-shaped lifetime distribution (See [4]) with $\Psi(t;\lambda)=e^{t^{\lambda}}-1$ and so on.

2. Classical Estimation

2.1. Point Estimation. Let $\mathbf{X}=(X_1,X_2,\ldots,X_m)$ be a progressive first-failure censored sample from (1), with censoring scheme $(r_1,r_2,\ldots,r_m)$. The likelihood function is given by

$$L(\alpha,\lambda;\mathbf{x}) = Ak^m\alpha^m\exp\Big\{-\alpha k\sum_{i=1}^{m}(r_i+1)\Psi(x_i;\lambda)\Big\}\prod_{i=1}^{m}\psi(x_i;\lambda), \tag{4}$$

where $A=n(n-r_1-1)(n-r_1-r_2-2)\cdots(n-r_1-r_2-\cdots-r_{m-1}-m+1)$. By setting the derivatives of the log-likelihood function with respective to α or λ to zero, the MLE of λ, say $\hat{\lambda}$, is the solution to the following likelihood equation

$$\sum_{i=1}^{m}\frac{(\partial/\partial\lambda)\psi(x_i;\lambda)}{\psi(x_i;\lambda)} = \frac{m\sum_{i=1}^{m}(r_i+1)(\partial/\partial\lambda)\Psi(x_i;\lambda)}{\sum_{i=1}^{m}(r_i+1)\Psi(x_i;\lambda)}, \tag{5}$$

and the MLE of α, say $\hat{\alpha}$, can be obtained as

$$\hat{\alpha} = \frac{m}{k\sum_{i=1}^{m}(r_i+1)\Psi(x_i;\hat{\lambda})}. \tag{6}$$

It is not easy to solve the Eq. (5) analytically in order to achieve the MLE of λ. Some numerical methods can be employed such as the Newton-Raphson method. Finally, using the invariance property, the MLEs of $R(t)$ and $h(t)$ are respectively obtained as

$$\hat{R}(t)=\exp\{-\hat{\alpha}\Psi(t;\hat{\lambda})\}, \quad \text{and} \quad \hat{h}(t)=\hat{\alpha}\psi(t;\hat{\lambda}).$$

2.2. Interval Estimation. Let $Y_i = k\alpha\Psi(X_i;\lambda)$ for $i = 1, 2, \ldots, m$. It can be seen that $Y_1 < Y_2 < \cdots < Y_m$, are the progressive first-failure censored order statistics from an exponential distribution with mean 1. Consider $Z_1 = nY_1$ and $Z_i = (n - \sum_{k=1}^{i-1} r_k - i + 1)(Y_i - Y_{i-1})$ for $i = 2, 3, \ldots, m$. The generalized spacings $Z_1, Z_2, \ldots, Z_m$ are independent and identically distributed as an exponential distribution with mean 1 (See [1, p.17-18]). Hence, for $j = 1, 2, \ldots, m-1$,

$$\tau_j = 2\sum_{i=1}^{j} Z_i = 2k\alpha[\sum_{i=1}^{j}(r_i+1)\Psi(X_i;\lambda) + \sum_{i=j+1}^{m}(r_i+1)\Psi(X_j,\lambda)], \tag{7}$$

$$\gamma_j = 2\sum_{i=j+1}^{m} Z_i = 2k\alpha \sum_{i=j+1}^{m}(r_i+1)[\Psi(X_i;\lambda) - \Psi(X_j;\lambda)], \tag{8}$$

are independently Chi-squared distributed with $2j$ and $2(m-j)$ degrees of freedom, respectively. We consider the following pivotal quantities:

$$\eta_j = \frac{j}{m-j}\cdot\frac{\sum_{i=j+1}^{m}(r_i+1)(\Psi(X_i;\lambda) - \Psi(X_j;\lambda))}{\sum_{i=1}^{j}(r_i+1)\Psi(X_i;\lambda) + \sum_{i=j+1}^{m}(r_i+1)\Psi(X_j;\lambda)},\ j = 1, 2, \ldots, m-1, \tag{9}$$

$$\xi = 2k\alpha\sum_{i=1}^{m}(r_i+1)\Psi(X_i;\lambda). \tag{10}$$

It is clearly that η_j has a F distribution with $2(m-j)$ and $2j$ degrees of freedom and ξ has a Chi-squared distribution with $2m$ degree of freedom. Meanwhile, η_j and ξ are independent. To construct an exact confidence interval for λ and the joint confidence region for the parameters α and λ, we need the following lemma.

LEMMA 2.1. *Suppose that for* $x_1 < x_2 < \cdots < x_m$,

$$w_j(\lambda) = \frac{\sum_{i=j+1}^{m}(r_i+1)(\Psi(x_i;\lambda) - \Psi(x_j;\lambda))}{\sum_{i=1}^{j}(r_i+1)\Psi(x_i;\lambda) + \sum_{i=j+1}^{m}(r_i+1)\Psi(x_j;\lambda)},\ j = 1, 2, \ldots, m-1. \tag{11}$$

Then $w_j(\lambda)$ *is strictly increasing in* λ, *if function* $\frac{\Psi'(t;\lambda)}{\Psi(t;\lambda)}$ *is strictly increasing in* t, *where* $\Psi'(t;\lambda)$ *is* $(\partial/\partial\lambda)\Psi(t;\lambda)$.

REMARK 2.2. For all of well-known lifetime distributions mentioned in Section 1, it can be shown that $\frac{\Psi'(t;\lambda)}{\Psi(t;\lambda)}$ is strictly increasing in t. For instance, when $\Psi(t;\lambda) = \ln(1+t^{\lambda})$, it turns out to be Burr XII distribution and see [9].

Let $F_{v_1,v_2}(p)$ is the percentile of F distribution with v_1 and v_2 degrees of freedom with the right-tail probability p.

THEOREM 2.3. *Suppose that* $\boldsymbol{X} = (X_1, X_2, \ldots, X_m)$ *be a progressive first-failure censored sample from* (1), *with censoring scheme* $(r_1, r_2, \ldots, r_m)$, $\frac{\Psi'(t;\lambda)}{\Psi(t;\lambda)}$ *is strictly increasing in* t, *and*

$$W_j(\lambda) = \frac{j}{m-j}w_j(\lambda),\ j = 1, 2, \ldots, m-1, \tag{12}$$

where $w_j(\lambda)$ *is defined in* (11). *Then, for any* $0 < \nu < 1$ *and* $j = 1, 2, \ldots, m-1$, *when* $F_{2(m-j),2j}(\frac{\nu}{2})$ *and* $F_{2(m-j),2j}(1-\frac{\nu}{2})$ *are in the range of the function* $W_j(\lambda)$

$$\varphi_j\big(\boldsymbol{X}, F_{2(m-j),2j}(1-\frac{\nu}{2})\big) < \lambda < \varphi_j\big(\boldsymbol{X}, F_{2(m-j),2j}(\frac{\nu}{2})\big), \tag{13}$$

is a $100(1-\nu)\%$ *confidence interval for* λ*, where* $\varphi_j(\boldsymbol{X},u)$ *is the solution of* λ *for equation* $W_j(\lambda)=u$*.*

THEOREM 2.4. *Suppose that* $\boldsymbol{X}=(X_1,X_2,\ldots,X_m)$ *be a progressive first-failure censored sample from* (1)*, with censoring scheme* $(r_1,r_2,\ldots,r_m)$*,* $\frac{\Psi'(t;\lambda)}{\Psi(t;\lambda)}$ *is strictly increasing in* t*. Then, for any* $0<\nu<1$ *and* $j=1,2,\ldots,m-1$*, when* $F_{2(m-j),2j}(\frac{1+\sqrt{1-\nu}}{2})$ *and* $F_{2(m-j),2j}(\frac{1-\sqrt{1-\nu}}{2})$*are in the range of function* $W_j(\lambda)$*, a* $100(1-\nu)\%$ *confidence region for* (α,λ) *is given by*

$$\text{(14)}\qquad \begin{cases} \varphi_j\big(\boldsymbol{X},F_{2(m-j),2j}(\frac{1+\sqrt{1-\nu}}{2})\big)<\lambda<\varphi_j\big(\boldsymbol{X},F_{2(m-j),2j}(\frac{1-\sqrt{1-\nu}}{2})\big), \\ \frac{\chi^2_{2m}(\frac{1+\sqrt{1-\nu}}{2})}{2k\sum_{i=1}^m(r_i+1)\Psi(x_i;\lambda)}<\alpha<\frac{\chi^2_{2m}(\frac{1-\sqrt{1-\nu}}{2})}{2k\sum_{i=1}^m(r_i+1)\Psi(x_i;\lambda)}, \end{cases}$$

where $\chi^2_{v_1}(p)$ *is the percentile of Chi-squared distribution with* v_1 *degree of freedom with the right-tail probability* p *and* $\varphi_j(\boldsymbol{X},u)$ *is defined in Theorem* 2.3*.*

3. Bayes Estimation

Now, we deal with the problem of estimating the parameters α and λ, as well as reliability function $R(t)$ and hazard rate function $h(t)$ against different symmetric and asymmetric loss functions. We assume that for $j=1,2,\ldots,M$, λ has a discrete prior say,

$$\text{(15)}\qquad P(\lambda=\lambda_j)=\theta_j,\ \sum_{j=1}^M\theta_j=1,$$

while the conditional distribution of α given λ_j has a conjugate prior distribution, with density

$$\text{(16)}\qquad g(\alpha|\lambda_j)=\beta_j\exp\{-\alpha\beta_j\},\ \alpha,\beta_j>0,$$

where $\beta_j,\ j=1,2,\ldots,M$, are hyper-parameters. Combining (4) and (16), the conditional posterior of the parameter α, takes the form

$$\text{(17)}\qquad \pi(\alpha|\mathbf{x},\lambda_j)=\frac{1}{\Gamma(m+1)}c_j^{m+1}\alpha^m\exp\{-\alpha c_j\},\ j=1,2,\ldots,M,$$

where $c_j=k\sum_{i=1}^m(r_i+1)\Psi(x_i;\lambda_j)+\beta_j$. Also by applying (4), (15), (16) and the discrete version of Bayes theorem, the marginal posterior distribution of λ can be expressed as

$$\text{(18)}\qquad p_j=P(\lambda=\lambda_j|\mathbf{x})=\frac{\beta_j\theta_jc_j^{-(m+1)}\prod_{i=1}^m\psi(x_i;\lambda_j)}{\sum_{j=1}^M\beta_j\theta_jc_j^{-(m+1)}\prod_{i=1}^m\psi(x_i;\lambda_j)},\ j=1,2,\ldots,M.$$

Therefore, the Bayes estimators of $\alpha,\lambda,R(t),h(t)$ under the squared error loss function are given respectively, by

$$\text{(19)}\qquad \hat{\alpha}_{SB}=(m+1)\sum_{j=1}^M\frac{p_j}{c_j},\quad \hat{\lambda}_{SB}=\sum_{j=1}^Mp_j\lambda_j,$$

$$\text{(20)}\qquad \hat{R}_{SB}(t)=\sum_{j=1}^Mp_j\Big[1+\frac{\Psi(t;\lambda_j)}{c_j}\Big]^{-(m+1)},\quad \hat{h}_{SB}(t)=(m+1)\sum_{j=1}^M\frac{p_j\psi(t;\lambda_j)}{c_j}.$$

For the loss function Linex, the Bayes estimators of α, λ, $R(t)$ and $h(t)$ are respectively obtained as

$$(21)\quad \hat{\alpha}_{LB} = -\frac{1}{c}\ln\left[\sum_{j=1}^{M} p_j\Big(1+\frac{c}{c_j}\Big)^{-(m+1)}\right],\quad \hat{\lambda}_{LB} = -\frac{1}{c}\ln\left[\sum_{j=1}^{M} p_j e^{-c\lambda_j}\right],$$

$$(22)\quad \hat{R}_{LB}(t) = -\frac{1}{c}\ln\left[\sum_{j=1}^{M}\sum_{l=0}^{\infty}\frac{(-1)^l}{\Gamma(l+1)} p_j c^l\Big(1+\frac{l\Psi(t;\lambda_j)}{c_j}\Big)^{-(m+1)}\right],$$

$$(23)\quad \hat{h}_{LB}(t) = -\frac{1}{c}\ln\left[\sum_{j=1}^{M} p_j\Big(1+\frac{c\psi(t;\lambda_j)}{c_j}\Big)^{-(m+1)}\right],$$

where $c \neq 0$ is the parameter of loss function Linex. When c is negative, underestimation is more serious than overestimation and it is opposite for positive c.

Example 3.1. (Real Data) In this example, we analyze a data set from [5], which represents the number of 1000s of cycles to failure for electrical appliances in a life test. The complete data have been used earlier by [7]. They showed that the bathtub-shaped distribution is suitable to fitting the data. The cdf of the bathtub-shaped distribution is form (2), where $\Psi(t;\lambda) = e^{t^{\lambda}} - 1, t > 0$. It can be shown that $\frac{\Psi'(t;\lambda)}{\Psi(t;\lambda)}$, is strictly increasing in t (See [4]). The data are randomly grouped

Table 1. progressively first-failure censored sample of size 8 out of 20 groups.

i	1	2	3	4	5	6	7	8
x_i	0.014	0.034	0.059	0.061	0.069	0.142	0.165	1.270
r_i	4	0	3	0	0	2	3	0

into 20 groups with $k = 3$ items within each group. The progressively first-failure censored sample is given in Table 1. For this example, 12 groups of failure times are censored, and 8 first-failures are observed. By applying (13) and (14), the 95% exact confidence intervals (CI) for λ, confidence regions (CR) for (α, λ), are obtained and the length of confidence intervals (LCI) and area for confidence regions (ACR) are presented in Table 2, where $A(\lambda) = \sum_{i=1}^{8}(r_i+1)(e^{x_i^{\lambda}} - 1)$. From Table 2, it is observed that, the 95% optimal confidence interval for λ is

Table 2. The 95% confidence intervals and regions and their some properties for λ and (α, λ).

j	CI	CR	LCI	ACR
1	$0.3933 < \lambda < 1.7034$	$0.3397 < \lambda < 1.8545$, $\frac{1.0114}{A(\lambda)} < \alpha < \frac{5.2012}{A(\lambda)}$	1.3101	1.3904
2	$0.3694 < \lambda < 1.4175$	$0.3192 < \lambda < 1.5198$, $\frac{1.0114}{A(\lambda)} < \alpha < \frac{5.2012}{A(\lambda)}$	1.0481	1.0334
3	$0.3538 < \lambda < 1.3167$	$0.3044 < \lambda < 1.4039$, $\frac{1.0114}{A(\lambda)} < \alpha < \frac{5.2012}{A(\lambda)}$	0.9629	0.9081
4	$0.2317 < \lambda < 1.0946$	$0.1920 < \lambda < 1.1696$, $\frac{1.0114}{A(\lambda)} < \alpha < \frac{5.2012}{A(\lambda)}$	0.8629	0.6805
5	$0.1391 < \lambda < 0.9320$	$0.1092 < \lambda < 1.0014$, $\frac{1.0114}{A(\lambda)} < \alpha < \frac{5.2012}{A(\lambda)}$	0.7929	0.5232
6	$0.1302 < \lambda < 0.9750$	$0.0963 < \lambda < 1.0462$, $\frac{1.0114}{A(\lambda)} < \alpha < \frac{5.2012}{A(\lambda)}$	0.8448	0.5708
7	$0.0212 < \lambda < 0.7932$	$0.0109 < \lambda < 0.8646$, $\frac{1.0114}{A(\lambda)} < \alpha < \frac{5.2012}{A(\lambda)}$	0.7720	0.4094

$(0.0212, 0.7932)$, and the optimal confidence region for (α, λ) is given by

$$0.0109 < \lambda < 0.8646 \ , \ \frac{1.0114}{A(\lambda)} < \alpha < \frac{5.2012}{A(\lambda)},$$

and $ACR = \int_{0.0109}^{0.8646} \frac{4.1898}{A(\lambda)}\, d\lambda = 0.4094$. Since there is no prior information about α, to compute the Bayes estimates, we estimate the hyper-parameters $\beta_j, j = 1, 2, \ldots, 8$, using the maximum likelihood type-II method (See [2, p. 99]). The values of β_j and p_j, for each given λ_j and θ_j, $j = 1, 2, \ldots, 8$, are summarized in Table 3. The MLEs as well as Bayes estimates of α, λ, reliability function $R(t)$, and hazard rate function $h(t)$, for $t = 0.5$, are presented in Table 4.

TABLE 3. Prior information, hyper-parameter values and the posterior probabilities.

j	1	2	3	4	5	6	7	8
λ_j	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75
θ_j	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125
β_j	3.5605	3.1398	2.7814	2.4735	2.2073	1.9756	1.7727	1.5942
p_j	0.0308	0.0549	0.0859	0.1206	0.1532	0.1778	0.1897	0.1871

TABLE 4. The ML and the Bayes estimates of $\alpha, \lambda, R(t)$ and $h(t)$, with $t = 0.5, c = 1$.

$\hat{\alpha}$	$\hat{\alpha}_{SB}$	$\hat{\alpha}_{LB}$	$\hat{\lambda}$	$\hat{\lambda}_{SB}$	$\hat{\lambda}_{LB}$
0.4800	0.4252	0.4132	0.7200	0.6268	0.6220
$\hat{R}(t)$	$\hat{R}_{SB}(t)$	$\hat{R}_{LB}(t)$	$\hat{h}(t)$	$\hat{h}_{SB}(t)$	$\hat{h}_{LB}(t)$
0.6697	0.6871	0.6833	0.7700	0.6584	0.6267

4. Conclusion

Lifetime studies are very important to assess the reliability of products. This article investigates the problem of reliability analysis for a class of an exponential distribution based on progressive first failure censoring. Both classical and Bayesian point estimations have been developed. Additionally, the exact confidence intervals and regions respectively for λ and (α, λ), have been conducted. It is noteworthy that many well-known and useful lifetime distributions which have wide application in reliability theory as well as other related fields are involved in this class.

References

1. N. Balakrishnan and R. Aggarwala, *Progressive Censoring: Theory, Methods and Applications*, Birkhäuser, Boston, 2000.
2. J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, 2nd ed., Springer-Verlag, New York, 1985.
3. Q. Bi, Y. Ma and W. Gui, *Reliability estimation for the bathtub-shaped distribution based on progressively first-failure censoring sampling*, Comm. Statist. Simulation Comput. (2020) in press. DOI:10.1080/03610918.2020.1746338
4. Z. Chen, *A new two-parameter lifetime distribution with bathtub shape or increasing failure rate function*, Statist. Probab. Lett. **49** (2) (2000) 155–161.
5. J. F. Lawless, *Statistical Models and Methods for Lifetime Data*, Wiley, New York, 2003.

6. H. S. Mohammed, S. F. Ateya and E. K. AL-Hussaini, *Estimation based on progressive first-failure censoring from exponentiated exponential distribution*, J. Appl. Stat. **44** (8) (2017) 1479–1494.
7. A. M. Sarhan, D. C. Hamilton and B. Smith, *Parameter estimation for a two-parameter bathtub-shaped lifetime distribution*, Appl. Math. Model. **36** (11) (2012) 5380–5392.
8. A. A. Soliman, A. H. Abd-Ellah, N. A. Abou-Elheggag and G. A. Abd-Elmougod, *Estimation of the parameters of life for Gompertz distribution using progressive first-failure censored data*, Comput. Statist. Data Anal. **56** (8) (2012) 2471–2485.
9. J. -W. Wu, W. -L. Hung and C. -H. Tsai, *Estimation of the parameters of the Gompertz distribution under the first-failure censored sampling plan*, Statistics **37** (6) (2003) 517–525.
10. S. J. Wu and C. Kuş, *On estimation based on progressive first-failure censored sampling*, Comput. Statist. Data Anal. **53** (10) (2009) 3659–3670.

E-mail: K.Ahmadi@sku.ac.ir

A Center-Outward Rank Test for Multivariate Paired Data

Sakineh Dehghan*
Department of Statistics, Faculty of Mathematical Sciences, Shahid Beheshti University, Tehran, Iran
and Mohammad Reza Faridrohani
Department of Statistics, Faculty of Mathematical Sciences, Shahid Beheshti University, Tehran, Iran

Abstract. In this paper, a class of test statistics is defined based on the center-outward depth ranking to test the equality of mean vectors in multivariate paired data. The tests are implemented through the idea of permutation tests that require no distributional assumption, except the symmetric paired data joint distribution assumption. Therefore, the tests have broader applicability than some of the existing tests. This class of test statistics is very easy to compute for data in any practical dimension. This distinguishes it from some of the other tests in the literature. The performance of the proposed tests is evaluated using a Monte Carlo study. The results show that the tests perform well comparing other procedures in the literature.
Keywords: Center-outward ranking, Depth function, Multivariate paired data, Permutation test.
AMS Mathematical Subject Classification [2010]: 62H15, 62G10.

1. Introduction

Many experiments in agriculture, biology, medicine etc, are performed based on the multivariate data. Therefore, analysis of the multivariate data is an important subject in statistical science. In this paper, among the various inferences for multivariate data, we'll focus on the nonparametric tests for mean vectors of the multivariate paired data. These tests are based on the center-outward ranks generated by depth functions. There is a substantial literature for this problem. Under the multivariate normality assumption, which is often difficult to justify in practice, it is common to use Hotelling's T^2 test [3]. In some cases, data may demonstrate distributions other than multivariate normal, in these situations nonparametric approaches can be suitable. Among many nonparametric tests have been introduced on this problem, we can mention [1, 2] and [6].

Most of the univariate nonparametric methods are based on rank of data points and they can be generalized to R^p if we have a ranking of data in R^p. The halfspace depth function was introduced by Tukey in 1975 [7] for imagination and ordering multivariate data. The different depth functions have been introduced and the multivariate data order as center-outward based on them. This center-outward ranking has been mostly applied in the multivariate nonparametric inference. Associated with a given distribution F on R^p, a depth function is

*Speaker

designed to provide a center-outward ordering of points $\mathbf{x}$ in R^p. Indeed, a notion of data depth is used to measure centrality/outlyingness of a point with respect to given data cloud or distribution. Let $\mathbf{X} = (X_1, \ldots, X_p)^T$ be a random vector on a probability space $(\Omega, \mathcal{F}, P)$ and F denote a distribution function corresponding to P. Zuo and Serfling [8] provided a formal definition of statistical depth function as a function $D(., F) : \mathbb{R}^p \to \mathbb{R}$ satisfying the four properties including affine invariance, maximised somewhere in the center of the distribution F, quasi-concavity and vanishing at infinity. Various depth functions have been proposed for ranking multivariate data, among which the more popular are Tukeys depth [7], Mahalanobis's depth [5], and simplicial depth [4].

In this paper, we introduce a test statistic based on center-outward depth ranking in order to test the equality of mean vectors in the multivariate paired data.

2. The Proposed Test

Let $(X_{i1}, \ldots, X_{ip})^T$ and $(Y_{i1}, \ldots, Y_{ip})^T$ are p-variate vectors of observations of random vectors $\mathbf{X}$ and $\mathbf{Y}$, respectively, from the i-th subject, $i = 1, \ldots, n$. $\mathbf{X}_i$ and $\mathbf{Y}_i$ are paired as they come from the same subject. It is commonly assumed that subjects are independent. Let the mean vector of the vector $\mathbf{W}_i = (\mathbf{X}_i^T, \mathbf{Y}_i^T)^T$ be denoted by $\boldsymbol{\mu}_W = \left(\boldsymbol{\mu}_X^T, \boldsymbol{\mu}_Y^T\right)^T$. It is assumed that the random vector $\mathbf{W}$ has a distribution centrally symmetric about $\boldsymbol{\mu}_W$. We interest to test the hypothesis equality of the mean vectors $H_0 : \boldsymbol{\mu}_X = \boldsymbol{\mu}_Y$ against $H_a : \boldsymbol{\mu}_X \neq \boldsymbol{\mu}_Y$. Let $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ip})^T$, $i = 1, \ldots, n$, be observations of p-variate vector $\mathbf{Z} = \mathbf{Y} - \mathbf{X}$. Under null hypothesis, the random vector $\mathbf{Z}$ has a distribution centrally symmetric about vector 0.

First, the random vectors $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ are ordered based on depth function $D(, F_n)$, where F_n is the sample distribution function of $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$. Let $r(\mathbf{Z}_1), \ldots, r(\mathbf{Z}_n)$ be the center-outward ranks of $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$, respectively. More precisely, for a sample point $\mathbf{Z}_i$

$$r(\mathbf{Z}_i) = \#\{\mathbf{Z}_j : \quad D(\mathbf{Z}_j, F_n) \leq D(\mathbf{Z}_i, F_n), \quad j = 1, \ldots, n\},$$

is the center-outward rank of $\mathbf{Z}_i$ with respect to the data cloud $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n$. Now, define

$$\begin{aligned} A_j &= \{\mathbf{Z}_i : \quad Z_{ij} \geq 0, \quad i = 1, \ldots, n, \quad j = 1, \ldots, p\}, \\ R_j &= \sum_{\mathbf{Z}_i \in A_j} r(\mathbf{Z}_i), \qquad j = 1, \ldots, p, \\ M_n &= \max(|2R_1 - N|, \ldots, |2R_p - N|). \end{aligned}$$

The null hypothesis would be rejected if at least one of R_j's $j = 1, \ldots, n$ is either sufficiently small or sufficiently large. It concludes that the larger values of $M_{n,D}$ is the stronger evidence against H_0 and therefore the null hypothesis H_0 is rejected for large values of $M_{n,D}$. To determine when M_n is large enough to reject H_0, we need to derive the null distribution of M_n. Define

$$P_M = p_{H_o}(M_{n,D} \geq M_o), \tag{1}$$

where M_o is the observed value of $M_{n,D}$. It seems that the asymptotic distribution of M_n will not accessible. Alternatively, we apply Fishers permutation test to determine the following p-value and complete our test procedure.

2.1. Permutation Method. We apply a permutation procedure to obtain reference distribution for $M_{n,D}$ and to estimate its finite-sample p-value. Note that an assumption behind a permutation test is that the observations are exchangeable under the null hypothesis. We first discuss how the framework is set up to satisfy this requirement and then describe the proposed permutation procedure. Define for $i = 1, \ldots, n$

$$\mathbf{W}_i = \begin{cases} \left(\mathbf{X}_i^T, \mathbf{Y}_i^T, \right)^T, & if \quad \delta_i = 1, \\ \left(\mathbf{Y}_i^T, \mathbf{X}_i^T, \right)^T, & if \quad \delta_i = -1, \end{cases}$$

where random variable δ_i takes values 1 and -1 with probability $1/2$. Under the null distribution, we have $\mathbf{Z} \stackrel{d}{=} -\mathbf{Z}$. Then, for each permuted sample $\mathbf{W}_1, \ldots, \mathbf{W}_n$, any permutation of $\mathbf{Z}$ is equal in distribution with itself. Moreover, the test statistics $M_{n,D}$ depends only on $\mathbf{Z}_i$'s, $i = 1, \ldots, n$. It concludes that, under the null distribution, the distribution of $M_{n,D}$ is invariant to any permutation.

Now, Fishers permutation test to approximate the p-value defined in (1) is applied as follows. Test statistics $M_{n,D}$ is computed for B permutation of $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ that B is sufficiently large and any prmutation randomly is selected (if n is not too large, for any permutation we compute test statistics). Under H_0, the p-value based on $M_{n,D}$ can be approximated by

$$P_{M,B} = \frac{1}{B} \sum_{i=1}^{B} I\left(M_{n,i}^* \geq M_o\right),$$

where $M_{n,i}^*$, $i = 1, \ldots, B$ and M_o are the observed values of $M_{n,D}$ based on ith permutation and the original data, respectively.

3. Simulation Study

Monte Carlo samples were generated to evaluate the performance of the proposed test procedure, including the size (i.e., type I error probability) and power of the tests. We compare the proposed tests to their counterparts T^2 Hotelling test which is derived under the assumption that all the data are normally distributed and two nonparametric tests including the test due to [2] computed with Wilcoxon score function that denoted by HP_n and the signed-rank test due to [6], MR_n. The proposed test statistic is calculated based on halfspace, simplicial and Mahalanobis depth functions as $M_{n,MD}$, $M_{n,SD}$ and $M_{n,HD}$, respectively. To investigate performance under different true distributions, we considered multivariate normal distributions and multivariate Cauchy distributions.

We simulated paired samples $\mathbf{W}_i^T = \left(\mathbf{X}_i^T, \mathbf{Y}_i^T\right)$, $i = 1, \ldots, n$, from a $(2p)$-variate distribution with mean vector $\boldsymbol{\mu}_W$ and identity covariance matrix, where $\mathbf{W}_i$, $i = 1, \ldots, n$, is either bivariate normal, or bivariate Cauchy. We let $\boldsymbol{\mu}_W = \left(\boldsymbol{\mu}_X^T, \boldsymbol{\mu}_Y^T\right)^T$ with $\boldsymbol{\mu}_X = 0$ and $\boldsymbol{\mu}_Y = \mu_Y 1_p$ for 1_p being the $p \times 1$ vector of 1s. We generated 1000 Monte Carlo samples for each setup with $n = 50$. Moreover, the

nominal level was set at 0.05 throughout. The empirical rejection probability of a test was calculated as the proportion of rejections from 1000 replicates.

The empirical rejection probabilities have been provided in Table 1. Inspection of Table 1 confirms that the performance of our test statistics is not aected by dierent depth ranking. The empirical rejection probabilities corresponding to $L = 0$ represents the proportion of rejection under the null hypothesis. These results demonstrate that all the tests would be accurate in estimating the nominal level. Table 1 clearly shows that the proposed tests perform comparably to the Hotelling T^2 test under the normal distribution, even though the former are completely nonparametric and do not utilize the normality assumption. For bivariate Cauchy distribution, $M_{n,MD}$, $M_{n,SD}$ and $M_{n,HD}$ tests outperform the Hotelling's T^2 and the nonparametric tests. Indeed, the Hotelling's T^2 is sensitive to violations of normality in data and the depth-based tests are moment-free approaches and thus more suitable for testing location parameters not derived from moments.

TABLE 1. Empirical rejection probabilities for the bivariate normal and Cauchy distribution with $\mu_Y = 0.15L$ and $\mu_Y = 0.2L$, respectively.

	Bivariate normal distribution				Bivariate Cauchy distribution			
	L				L			
Test	0	1	2	3	0	1	2	3
T^2	0.052	0.246	0.719	0.998	0.041	0.105	0.183	0.251
HD_n	0.052	0.210	0.712	0.972	0.052	0.198	0.312	0.630
MR_n	0.055	0.266	0.674	0.942	0.053	0.266	0.655	0.911
$M_{n,MD}$	0.049	0.260	0.681	0.954	0.051	0.271	0.673	0.921
$M_{n,HD}$	0.054	0.266	0.674	0.942	0.050	0.274	0.682	0.952
$M_{n,LD}$	0.051	0.252	0.666	0.942	0.052	0.271	0.691	0.953

References

1. S. Dehghan and M. R. Faridrohani, *Affine invariant depth-based tests for the multivariate one-sample location problem*, Test. **28** (3) (2019) 671–693.
2. M. Hallin and D. Paindaveine, *Optimal tests for multivariate location based on interdirections and pseudo-Mahalanobis ranks*, Ann. Statist. **30** (4) (2002) 1103–1133.
3. H. Hotelling, *The generalization of Student's ratio*, Ann. Math. Statist. **2** (3) (1931) 360–378.
4. R. Y. Liu, *On a notion of simplicial depth*, Proceedings of the National Academy of Sciences **85** (6) (1988) 1732–1734.
5. R. Y. Liu and K. Singh, *A quality index based on data depth and multivariate rank tests*, J. Amer. Statist. Assoc. **88** (421) (1993) 252–260.
6. Z. R. Mahfoud and R. H. Randles, *On multivariate signed rank tests*, J. Nonparametr. Stat. **17** (2) (2005) 201–216.
7. J. W. Tukey, *Mathematics and the picturing of data*, In Proceeding of the International Congress of Mathematicians (Vancouver, B. C., 1974) Canad. Math. Congress, Montreal, Que. (1975) pp. 523–531.
8. Y. Zuo and R. Serfling, *General notions of statistical depth function*, Ann. statist. **28** (2) (2000a) 461–482.

E-mail: sa_dehghan@sbu.ac.ir
E-mail: m_faridrohani@sbu.ac.ir

Optimal Design of Step Stress Test under Periodic Inspection for Exponential Distribution

Nooshin Hakamipour*
Department of Mathematics, Buein Zahra Technical University, Buein Zahra, Qazvin, Iran

Abstract. In this paper, we discuss the optimal step stress accelerated life test plan under periodic inspection and Type I censoring. The exponential distribution with a failure rate function that a log-quadratic function of stress and the tampered failure rate model are considered. The asymptotic variance of the maximum likelihood estimators of parameters is derived as an optimality criterion and the optimal stress change times are determined. A numerical example will be given to illustrate the proposed inferential procedures.
Keywords: Asymptotic variance, Exponential distribution, Periodic inspection, Tampered failure rate model, Three step stress test.
AMS Mathematical Subject Classification [2010]: 62N05, 90C31, 62N01.

1. Introduction

The life testing time under environment conditions may be very long and it is difficult for extremely reliable units to make life testing at use stress. The accelerated life testings (ALTs) are used to overcome this problem. ALTs are done on greater stresses than use stress and then ALTs quickly yield informations on test units. A special class of the accelerated life testing, known as step-stress testing, allows the experimenter to gradually increase the stress levels at some pre-fixed time points during the experiment for maximal flexibility and adjustability. This model has attracted great attention in the reliability literature. Since the stress-loading is non-constant for the step-stress ALT (SSALT), an additional model to explain the effect of changing stress is required. Tampered failure rate (TFR) model assumes that a change in the stress has a multiplicative effect on the failure rate function over the remaining life. In this paper TFR model has been used.

The lifetimes of test units can be examined continuously or intermittently in the SSALT. The periodic inspection of life testing time is often used due to further reduction in time and cost, on the other hand earlier studies assumed continuous inspection. The data from periodic inspection consist of only the number of failures in the inspection intervals.

Ahmad et al. [1] studied the statistical inference of model parameters and optimum test plans on the design for periodic inspection under the constant stress ALT. Moon [2] considered the estimation of model parameters and optimum plans based on Type I censored data from three step stress ALT for exponential distribution under the TFR model. Moon and Park [3] studied the optimum plan and

*Speaker

the estimation of model parameters with Type I censoring under simple SSALT. In this paper, the results of Moon and Park [3] are extended to the case of three SSALT based on periodic inspection with type I censoring under the tampered failure rate model.

In Section 2, the model and some necessary assumptions are described. In Section 3, MLEs of the parameters and the optimum plan that minimizes the asymptotic variance of the MLE of the mean lifetime at the use stress are obtained. A numerical example is presented for the proposed inferential procedures in Section 4.

2. Model and Assumptions

Suppose that there are four level stresses $y_0 < y_1 < y_2 < y_3$, where y_0 is the use stress. In the presentation of our results and without loss of generality, $x_i = \frac{y_i - y_0}{y_3 - y}$ notation is used for $i = 0, 1, 2, 3$.

All test units n are simultaneously put on stress x_1 and inspections are conducted at pre-set times $t_{11}, t_{12}, \ldots, t_{1K(1)}$, but if all units do not fail before time $t_{1K(1)}(= \tau_1)$, the surviving units are subjected to the stronger stress x_2 and observed at pre-set times $t_{21}, t_{22}, \ldots, t_{2K(2)}$, but if all units on stress x_2 do not fail before time $t_{2K(2)}(= \tau_2)$, the surviving units are subjected to the stronger stress x_3 and observed at pre-set times $t_{31}, t_{32}, \ldots, t_{3K(3)}$ and surviving units at time $t_{3K(3)}(= \tau_c)$ are censored, where $K(i)$ is the number of inspections at stress x_i, $i = 1, 2, 3$. At stress x_i, the number of failures n_{ij} are recorded corresponding p_{ij}, probability of failures in the interval $(t_{ij-1}, t_{ij}]$, $i = 1, 2, 3$, $j = 1, 2, \ldots, K(i)$ and $p_c = P(\tau_c < T)$, where $t_{10} = \tau_0 = 0$, $\tau_1 = t_{20}$, $\tau_2 = t_{30}$. Note that, $n_i = \sum_{j=1}^{K(i)} n_{ij}$ for $i = 1, 2, 3$ and n_c is the censored units at a censoring time τ_c where $n_c = n - (n_1 + n_2 + n_3)$.

Suppose that stress response relationship of each test unit has the log-quadratic function with the stress variable x_i, which is given by $\log \theta_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2$, $i = 1, 2, 3$, where β_1, β_2 and β_3 are unknown model parameters. The number of failed units n_{ij}, $i = 1, 2, 3$, $j = 1, 2, \ldots, K(i)$ are used to estimate the model parameters β_1, β_2 and β_3, and then the model is extrapolated to make statistical inferences under the use stress.

The probability distribution function $f(t)$ for a test unit lifetime T at stress x_i, $i = 1, 2, 3$ is given by

$$f(t) = \begin{cases} \frac{1}{\theta_1} \exp\left(-\frac{t}{\theta_1}\right), & 0 \le t < \tau_1, \\ \frac{1}{\theta_2} \exp\left(-\frac{t - \tau_1}{\theta_2} - \frac{\tau_1}{\theta_1}\right), & \tau_1 \le t < \tau_2, \\ \frac{1}{\theta_3} \exp\left(-\frac{t - \tau_2}{\theta_3} - \frac{\tau_2 - \tau_1}{\theta_2} - \frac{\tau_1}{\theta_1}\right), & \tau_2 \le t < \tau_c. \end{cases}$$

3. Maximum Likelihood Estimators and Optimum Plan

In this section, MLEs of the model parameters β_1, β_2 and β_3 are obtained by Newton Rapshon method and the optimum plan for searching the optimal stress change times τ_1 and τ_2, which minimize the asymptotic variance of the MLE of the logarithm of the mean lifetime at the use stress x_0. The likelihood function is

given by

$$L \propto \prod_{i=1}^{3} \prod_{j=1}^{K(i)} p_{ij}^{n_{ij}} p_c^{n_c},$$

where $p_{ij} = P(t_{ij-1} < T < t_{ij}) = \exp(-u_{ij-1}^{(0)}) - \exp(-u_{ij}^{(0)})$ and $P_c = P(T > \tau_c) = \exp(-u_{3K(3)}^{(0)})$ and

$$\begin{aligned}
u_{ij-1}^{(m)} &= (t_{ij-1} - \tau_{i-1}) x_i^m \exp(-\beta_1 - \beta_2 x_i - \beta_3 x_i^2) \\
&\quad + (\tau_{i-1} - \tau_{i-2}) x_{i-1}^m \exp(-\beta_1 - \beta_2 x_{i-1} - \beta_3 x_{i-1}^2) \\
&\quad + \tau_{i-2} x_{i-2}^m \exp(-\beta_1 - \beta_2 x_{i-2} - \beta_3 x_{i-2}^2), \\
u_{ij}^{(m)} &= (t_{ij} - \tau_{i-1}) x_i^m \exp(-\beta_1 - \beta_2 x_i - \beta_3 x_i^2) \\
&\quad + (\tau_{i-1} - \tau_{i-2}) x_{i-1}^m \exp(-\beta_1 - \beta_2 x_{i-1} - \beta_3 x_{i-1}^2) \\
&\quad + \tau_{i-2} x_{i-2}^m \exp(-\beta_1 - \beta_2 x_{i-2} - \beta_3 x_{i-2}^2), \\
u_{3K(3)}^{(m)} &= (\tau_c - \tau_2) x_3^m \exp(-\beta_1 - \beta_2 x_3 - \beta_3^2 x_3^2) \\
&\quad + (\tau_2 - \tau_1) x_2^m \exp(-\beta_1 - \beta_2 x_2 - \beta_3^2 x_2^2) \\
&\quad + \tau_1 x_1^m \exp(-\beta_1 - \beta_2 x_1 - \beta_3^2 x_2^2),
\end{aligned}$$

for $i = 1, 2, 3$, $j = 1, 2, \ldots, K(i)$ and $m = 0, 1, \ldots, 4$, where $x_0 = 0$ and $\tau_0 = 0$.

Thus, the log likelihood function which is a function of unknown parameters β_1, β_2 and β_3 is given by as follows:

$$\ell = \log L(\beta_1, \beta_2, \beta_3) \propto \sum_{i=1}^{3} \sum_{j=1}^{K(i)} n_{ij} \log p_{ij} + n_c \log p_c.$$

The MLEs for the model parameters β_1, β_2 and β_3 can be obtained by solving the following equation:

$$\frac{\partial \ell}{\partial \beta_k} = \sum_{i=1}^{3} \sum_{j=1}^{K(i)} n_{ij} \frac{1}{p_{ij}} \frac{\partial p_c}{\partial \beta_k} + n_c \frac{1}{p_c} \frac{\partial p_c}{\partial \beta_k} = 0,$$

for $k = 1, 2, 3$, where

$$\begin{aligned}
\frac{\partial p_{ij}}{\partial \beta_k} &= u_{ij-1}^{(k)} \exp(-u_{ij-1}^{(0)}) - u_{ij}^{(k)} \exp(-u_{ij}^{(0)}), \quad i = 1, 2, 3, \ j = 1, 2, \ldots, K(i), \\
\frac{\partial p_c}{\partial \beta_k} &= u_{3K(3)}^{(k)} \exp(-u_{3K(3)}^{(0)}).
\end{aligned}$$

The Fisher information matrix F is defined as $F = (f_{kl})$, $k, l = 1, 2, 3$ and can be obtained by taking the expected value of the second partial and mixed partial derivatives of ℓ with respect to β_1, β_2 and β_3 as follows:

$$\begin{aligned}
\frac{\partial^2 \ell}{\partial \beta_k^2} &= \sum_{i=1}^{3} \sum_{j=1}^{K(i)} \frac{n_{ij}}{p_{ij}} \Big[\frac{\partial^2 p_{ij}}{\partial \beta_k^2} - \frac{1}{p_{ij}} \big(\frac{\partial p_{ij}}{\partial \beta_k} \big)^2 \Big] + \frac{n_c}{p_c} \Big[\frac{\partial^2 p_c}{\partial \beta_k^2} - \frac{1}{p_c} \big(\frac{\partial p_c}{\partial \beta_k} \big)^2 \Big], \\
\frac{\partial^2 \ell}{\partial \beta_k \partial \beta_l} &= \sum_{i=1}^{3} \sum_{j=1}^{K(i)} \frac{n_{ij}}{p_{ij}} \Big[\frac{\partial^2 p_{ij}}{\partial \beta_k \partial \beta_l} - \frac{1}{p_{ij}} \big(\frac{\partial p_{ij}}{\partial \beta_k} \big) \big(\frac{\partial p_{ij}}{\partial \beta_l} \big) \Big] + \frac{n_c}{p_c} \Big[\frac{\partial^2 p_c}{\partial \beta_k \partial \beta_l} - \frac{1}{p_c} \big(\frac{\partial p_{ij}}{\partial \beta_k} \big) \big(\frac{\partial p_{ij}}{\partial \beta_l} \big) \Big].
\end{aligned}$$

for $k \neq l = 1, 2, 3$, where for $i = 1, 2, 3$,

$$\begin{aligned}
\frac{\partial^2 p_{ij}}{\partial \beta_k^2} &= \big((u_{ij-1}^{(k)})^2 - u_{ij-1}^{(2k)}\big)\exp(-u_{ij-1}^{(0)}) - \big((u_{ij}^{(k)})^2 - u_{ij}^{(2k)}\big)\exp(-u_{ij}^{(0)}),\\
\frac{\partial^2 p_{ij}}{\partial \beta_k \partial \beta_l} &= \big(u_{ij-1}^{(k)} u_{ij-1}^{(l)} - u_{ij-1}^{(k+l)}\big)\exp(-u_{ij-1}^{(0)}) - \big(u_{ij}^{(k)} u_{ij}^{(l)} - u_{ij}^{(k+1)}\big)\exp(-u_{ij}^{(0)}),\\
\frac{\partial^2 p_c}{\partial \beta_k^2} &= \big((u_{3K(3)}^{(k)})^2 - u_{3K(3)}^{(2k)}\big)\exp(-u_{3K(3)}^{(0)}),\\
\frac{\partial^2 p_c}{\partial \beta_k \partial \beta_l} &= \big(u_{3K(3)}^{(k)} u_{3K(3)}^{(l)} - u_{3K(3)}^{(k+l)}\big)\exp(-u_{3K(3)}^{(0)}).
\end{aligned}$$

The expected value of the second partial and mixed partial derivatives of ℓ with respect to β_1, β_2 and β_3 are given by

$$\begin{aligned}
f_{kk} &= -E\left(\frac{\partial^2 \ell}{\partial \beta_k^2}\right) = n\big[\sum_{i=1}^{3} Q_{ikk} + Q_{ckk}\big],\\
f_{kl} &= -E\left(\frac{\partial^2 \ell}{\partial \beta_k \partial \beta_l}\right) = n\big[\sum_{i=1}^{3} Q_{ikl} + Q_{ckl}\big],
\end{aligned}$$

where for $k, l = 1, 2, 3$ and $i = 1, 2, 3$

$$\begin{aligned}
Q_{ikl} &= \sum_{j=1}^{K(i)} \big[\frac{1}{p_{ij}}\big(\frac{\partial p_{ij}}{\partial \beta_k}\big)\big(\frac{\partial p_{ij}}{\partial \beta_l}\big) - \frac{\partial^2 p_{ij}}{\partial \beta_k \partial \beta_l}\big],\\
Q_{ckl} &= \frac{1}{p_c}\big(\frac{\partial p_c}{\partial \beta_k}\big)\big(\frac{\partial p_c}{\partial \beta_l}\big) - \frac{\partial^2 p_c}{\partial \beta_k \partial \beta_l},
\end{aligned}$$

The optimum plan for determining optimal stress change times τ_1 and τ_2 under three SSALT is presented, which minimize the asymptotic variance of $\log \hat{\theta}_0$, MLE of logarithm of mean lifetime at the use stress x_0. The asymptotic covariance matrix, V of $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$ is given by $V = F^{-1}$. And the asymptotic variance of $\log \hat{\theta}_0$ is given by

$$AV(\log \hat{\theta}_0) = (1, x_0, x_0^2)\, V\, (1, x_0, x_0^2)'. \tag{1}$$

Then the optimal change times τ_1^* and τ_2^* minimizing the $AV(\log \hat{\theta}_0)$.

4. Examples

The data from periodic inspections in three SSALT consist of only the number of failures in each inspection interval $(t_{ij-1}, t_{ij}]$, $i = 1, 2, 3$, $j = 1, 2, \ldots, K(i)$, where $K(i)$ is the number of inspection in each stress level. In practice, to find the optimal stress change times τ_1^* and τ_2^*, the parameters must be approximated by experience, similar data or preliminary test.

It is assumed that the numbers of inspections on each stress are $K(1) = 3$, $K(2) = 2$, $K(3) = 1$ and the probabilities of failure, p_{ij} in inspection intervals $(t_{ij-1}, t_{ij}]$, $i = 1, 2, 3$, $j = 1, 2, \ldots, K(i)$ are $p_{11} = 0.2$, $p_{12} = 0.15$, $p_{13} = 0.1$ on stress x_1, $p_{21} = 0.15$, $p_{22} = 0.15$ on stress x_2, $p_{31} = 0.15$ on stress x_3, $p_c = 0.1$ and three stress levels are $x_1 = 0.3$, $x_2 = 0.6$, $x_3 = 1.0$, and model parameters are $\beta_1 = 1.0$, $\beta_2 = -2.0$, $\beta_3 = -5.0$, and the stress change times are $\tau_1 = 0.56868$,

$\tau_2 = 0.67539$, $\tau_c = 0.67766$. The optimal stress change times τ_1^* and τ_2^* minimizing the $AV(\log \hat{\theta}_0)$ in (1) were obtained as $\tau_1^* = 0.44869$ and $\tau_2^* = 0.67677$. MATLAB software can be used to find the optimal points.

Now, the MLEs for model parameters β_1, β_2 and β_3 using the optimal stress change times τ_1^* and τ_2^* based on $\beta_1 = 1.0$, $\beta_2 = -2.0$ and $\beta_3 = -5.0$ are obtained for $n = 40$, $n = 30$ and $n = 25$ to examine the behavior of MLEs due to the sample size change.

All test units are simultaneously put on stress $x_1 = 0.3$ and inspections are conducted three times at pre-set times $t_{11} = 0.21226$, $t_{12} = 0.40977$ and $t_{13} = \tau_1^* = 0.44869$, but if all units do not fail before time τ_1^*, the surviving units are subjected to a stronger stress $x_2 = 0.6$ and also observed at specified times $t_{21} = 0.61178$, $t_{22} = \tau_2 = 0.67677$, but if all units do not fail before time τ_2^* , the surviving units are subjected to a stronger stress $x_3 = 1.0$ and observed until censoring time.

For $n = 40$, the number of failed test units at each inspection interval $(t_{ij-1}, t_{ij}]$, $i = 1, 2, 3$, $j = 1, 2, \ldots, K(i)$ were $n_{11} = 4$, $n_{12} = 9$, $n_{13} = 3$ on stress x_1, $n_{21} = 17$, $n_{22} = 6$ on stress x_2, $n_{31} = 1$ on stress x_3 and the number of censoring units was $n_c = 0$, where $t_{10} = 0$, $t_{20} = \tau_1^*$ and $t_{30} = \tau_2^*$. By Newton-Raphson method, the MLEs of model parameters β_1, β_2 and β_3 were obtained as $\hat{\beta}_1 = 1.12180$, $\hat{\beta}_2 = -1.96114$ and $\hat{\beta}_3 = -4.86936$.

For $n = 30$, $n_{11} = 5$, $n_{12} = 3$, $n_{13} = 4$ on stress x_1, $n_{21} = 11$, $n_{22} = 6$ on stress x_2, $n_{31} = 1$ on stress x_3, $n_c = 0$ and the MLEs of model parameters β_1, β_2 and β_3 were $\hat{\beta}_1 = 1.14961$, $\hat{\beta}_2 = -1.91964$ and $\hat{\beta}_3 = -4.86005$.

For $n = 25$, $n_{11} = 3$, $n_{12} = 8$, $n_{13} = 0$ on stress x_1, $n_{21} = 9$, $n_{22} = 3$ on stress x_2, $n_{31} = 2$ on stress x_3, $n_c = 0$ and the MLEs of model parameters β_1, β_2 and β_3 were $\hat{\beta}_1 = 1.25757$, $\hat{\beta}_2 = -2.09730$ and $\hat{\beta}_3 = -4.85060$.

As test units n changes, the MLEs of β_1, β_2 and β_3 are closer to the true values of model parameters as n increases.

The optimum plan is presented and maximum likelihood estimators of model parameters are obtained by periodic inspection and type I censored data from the step-stress accelerated life tests. This method will be very helpful in the situation that the intermittent inspection is the only practicable way of checking the status of test units under a step stress test. These results will be extended to the research associated with periodic inspection and type I censoring for multiple step stress accelerated life tests.

References

1. N. Ahmad, A. Islam and A. Salam, *Analysis of optimal accelerated life test plans for periodic inspection*, Int. J. Quality Reliability Management **23** (2006) 1019–1046.
2. G. A. Moon, *Step-stress accelerated life test for grouped and censored data*, J. Korean Data Info. Sci. Soc. **19** (2008) 697–708.
3. G. A. Moon and Y. K. Park, *Optimal step stress accelerated life tests for the exponential distribution under periodic inspection and type I censoring*, J. Korean Data Info. Sci. Soc. **20** (2009) 1169–1175.

E-mail: n.hakami@bzte.ac.ir; nooshin.hakami@aut.ac.ir

The Initial Conditions Problem in L_1 Regularization of Dynamic Random-Intercepts Models

Amir Abbas Mofidian Naeini*
Department of Mathematical Sciences, Isfahan University of Technology, Isfahan, Iran
and Reyhaneh Rikhtehgaran
Department of Mathematical Sciences, Isfahan University of Technology, Isfahan, Iran

Abstract. In this paper, we address the initial conditions problem in regularization of the random-intercepts model with the first-order lag response. This model uses random effects to cover the intra-class correlation and the first lagged response to address the serial correlation, which are the two common sources of dependency in longitudinal data. We demonstrate that ignoring the correlation between the initial response and the random effects called the initial conditions problem, can lead to biased regularized estimates.
Keywords: Penalized likelihood, Random effects, Serial correlation.
AMS Mathematical Subject Classification [2010]: 62J07.

1. Introduction

Longitudinal data are measurements or observations repeatedly collected from different subjects over time. Usually, there is dependency among observations of each subject. Hence, analysis of longitudinal data requires special models that consider this dependency. Two main sources of dependency among longitudinal observations are the intra-class correlation and the serial correlation. The intra-class correlation is due to the effects of unmeasured characteristics of each subject on its observations and this dependency can be handled by using linear mixed-effects models. The serial correlation happens due to the transferring effects of variables over time. Autoregressive models are usually applied to cover this dependency among longitudinal observations.

Usually, in longitudinal data, both sources of dependencies emerge and need to be considered in the analysis. Dynamic mixed-effects models which include both random effects and lagged responses are usually used in these situations [4]. But ignoring the correlation between initial responses and random effects in these models, called the initial conditions problem, can cause serious bias for maximum likelihood estimators (MLEs) of regression coefficients and variance components [2, 3]. Some solutions are proposed to deal with this issue in the important case of dynamic random-intercepts models with the first lagged response, e.g. [5, 6].

On the other hand, the advent of internet and the advancement of technology have made longitudinal data much easier to collect in many disciplines. In these situations, it is more likely to collect more variables and thus dimensions of data sets are increased. To achieve more interpretable models and to increase the efficiency of predictions and inferences, it is necessary to select important covariates

*Speaker

in regression analysis of longitudinal data. Regularization methods which apply penalties on the norm of regression coefficients in the structure of a penalized likelihood corresponding to a regression model are used to select important covariates and to achieve more efficient estimates.

In this paper, we discuss the initial conditions problem in regularization of dynamic random-intercepts models and show that if this issue is not handled, then, maximum penalized likelihood estimators (MPLEs) of regression coefficients with the L_1-norm penalty is inconsistent when the number of subjects is large and the number of observations of each subject is small.

The rest of this paper is organized as follows. In Section 2, we introduce dynamic random-intercepts models and the initial conditions problem in estimating parameters in these models. The MPLE is introduced in Section 3. Section 4 investigates large sample properties of the MPLE in dynamic random-intercepts models.

2. Dynamic Random-Intercepts Models

Dynamic linear mixed-effects models are specifically adopted for the analysis of longitudinal data that have both intra-class correlation and serial correlation. One of the most important models in this class is the dynamic random-intercepts model, defined as follows

$$y_{i,t} = \gamma y_{i,t-1} + \boldsymbol{x}'_{i,t}\boldsymbol{\beta} + u_{i,t},\ i = 1, \ldots, n,\ t = 1, \ldots, T, \tag{1}$$

where $y_{i,t}$ and $\boldsymbol{x}_{i,t}$ are, respectively, the response variable and the p-dimensional vector of covariates for the i-th subject at the t-th time, and $y_{i,t-1}$ is the lagged response variable. In this model, $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients with the fixed intercept β_0 as its first component and γ measures effects of the previous response on the current one.

The error terms $u_{i,t} = \alpha_i + \varepsilon_{i,t}$ in which the random intercepts α_i's represent unobserved subject effects and the residuals $\varepsilon_{i,t}$'s indicate time-varying effects that are not included in the model. It is usually assumed that $\alpha_i \overset{\text{iid}}{\sim} N(0, \sigma^2_\alpha)$, $\varepsilon_{i,t} \overset{\text{iid}}{\sim} N(0, \sigma^2_\varepsilon)$ and they are independent of each other and $\boldsymbol{x}_{i,t}$'s, for all i and t.

2.1. The Initial Conditions Problem. All the past effects of covariates and unmeasured variables on the current state of the response variable are incorporated through the lagged response, $y_{i,t-1}$, in Eq. (1). By taking backward substitution, we have

$$y_{i,t} = \gamma^t y_{i,0} + \sum_{j=0}^{t-1} \gamma^j \boldsymbol{x}'_{i,t-j}\boldsymbol{\beta} + \sum_{j=0}^{t-1} \gamma^j u_{i,t-j}.$$

Usually $y_{i,0}$ which represents the initial state of the response is correlated with the random effects α_i. If we ignore this correlation, the MLE of $\boldsymbol{\theta} = (\gamma, \boldsymbol{\beta}')'$ is inconsistent, when n is large and T is small. This problem is known as the initial conditions problem.

3. Maximum Penalized Likelihood Estimator (MPLE)

To obtain MPLE, we rewrite Eq. (1) in a vector form for each individual as

$$\boldsymbol{y}_i = \widetilde{\boldsymbol{X}}_i\boldsymbol{\theta} + \boldsymbol{u}_i,\ i = 1, \ldots, n,$$

where $\widetilde{\boldsymbol{X}}_i = (\boldsymbol{y}_{i,-1}\ \boldsymbol{X}_i)$, $\boldsymbol{y}_{i,-1} = (y_{i,0}, \ldots, y_{i,T-1})$, $\boldsymbol{X}_i = (\boldsymbol{x}_{i,1}, \ldots, \boldsymbol{x}_{i,T})^{'}$, and $\boldsymbol{\theta} = (\gamma, \boldsymbol{\beta}^{'})^{'}$. Then, the combined residual term $\mathbf{u}_i = (u_{i,1}, \ldots, u_{i,T})'$ follows $N_T(\mathbf{0}, \mathbf{V})$ with the well defined covariance matrix of the form e.g., [1],

$$\boldsymbol{V} = \sigma_c^2\overline{\boldsymbol{J}}_T + \sigma_\varepsilon^2\boldsymbol{E}_T, \tag{2}$$

where $\sigma_c^2 = \sigma_\varepsilon^2 + T\sigma_\alpha^2$. In Eq. (2), $\overline{\boldsymbol{J}}_T = (1/T)\boldsymbol{e}_T\boldsymbol{e}_T^{'}$, where $\boldsymbol{e}_T$ is a T-dimensional vector of ones and $\boldsymbol{E}_T = \boldsymbol{I}_T - \overline{\boldsymbol{J}}_T$, where $\boldsymbol{I}_T$ is a T-dimensional identity matrix.

Then, the penalized log-likelihood function is given by

$$\widetilde{l}_n(\boldsymbol{\theta}) \propto -\frac{n}{2}\log\big(det(\boldsymbol{V})\big) - \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{y}_i - \widetilde{\boldsymbol{X}}_i\boldsymbol{\theta})'\boldsymbol{V}^{-1}(\boldsymbol{y}_i - \widetilde{\boldsymbol{X}}_i\boldsymbol{\theta}) - \sum_{j=1}^{p}P_{\lambda_n}(\beta_j),$$

where $P_{\lambda_n}(\beta_j) = \lambda_n\sum_{j=1}^{p}|\beta_j|$ is the penalty function of the least absolute shrinkage and selection operator (Lasso) [7] and λ_n is the tuning parameter. By increasing λ_n, the shrinkage rate of regression coefficients toward zero increases. If variance components are known, maximizing penalized log-likelihood function for the dynamic random intercept model is equivalent to minimizing

$$Q_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i - \widetilde{\boldsymbol{X}}_i\boldsymbol{\theta})'\boldsymbol{V}^{-1}(\boldsymbol{y}_i - \widetilde{\boldsymbol{X}}_i\boldsymbol{\theta}) + \frac{\lambda_n}{n}\sum_{j=1}^{p}|\beta_j|.$$

4. Inconsistency of MPLE

Consider the following regularity conditions for the design matrix

$$\boldsymbol{C}_n = \frac{1}{n}\sum_{i=1}^{n}\widetilde{\boldsymbol{X}}_i\boldsymbol{V}^{-1}\widetilde{\boldsymbol{X}}_i^{'} \longrightarrow \boldsymbol{C}, \tag{3}$$

where $\boldsymbol{C}$ is a nonnegative-definite matrix and define the (random) function $Z_n(\boldsymbol{\phi})$ as follows

$$Z_n(\boldsymbol{\phi}) = \frac{1}{n}\sum\nolimits_{i=1}^{n}(\boldsymbol{Y}_i - \widetilde{\boldsymbol{X}}_i\boldsymbol{\phi})'\boldsymbol{V}^{-1}(\boldsymbol{Y}_i - \widetilde{\boldsymbol{X}}_i\boldsymbol{\phi}) + \frac{\lambda_n}{n}\sum_{j=2}^{p+1}\mid\phi_j\mid,$$

which is minimized at $\widehat{\boldsymbol{\theta}}_n = arg\min_{\boldsymbol{\theta}}Q_n(\boldsymbol{\theta})$.

THEOREM 4.1. *Suppose that $\boldsymbol{C}$ in (3) is nonsingular and $\lambda_n/n \to \lambda_0 \geq 0$. Also, assume that*

$$Z(\boldsymbol{\phi}) = T + (\boldsymbol{\phi} - \boldsymbol{\theta})^{'}C(\boldsymbol{\phi} - \boldsymbol{\theta}) - \frac{2}{\sigma_\varepsilon^2}P(\psi)(\gamma - \phi_1) + \lambda_0\sum_{j=2}^{p+1}\mid\phi_j\mid,$$

where $\psi = \dfrac{\sigma_\varepsilon^2}{\sigma_c^2}$, *and* $P(\psi) = \psi\varphi_T(\gamma)\sigma_{0\alpha}$, *that* $\sigma_{0\alpha} = Cov(Y_{i,0}, \alpha_i)$, *and* $\varphi_T(\gamma) = \dfrac{1-\gamma^T}{1-\gamma}$. *Then*

$$Z_n(\boldsymbol{\phi}) - Z(\boldsymbol{\phi}) \to 0,$$

in probability.

PROOF. For $Z_n(\boldsymbol{\phi})$, we have

$$\begin{aligned} Z_n(\boldsymbol{\phi}) &= (\boldsymbol{\phi}-\boldsymbol{\theta})'\frac{1}{n}\sum_{i=1}^{n}\widetilde{\boldsymbol{X}}_i'\boldsymbol{V}^{-1}\widetilde{\boldsymbol{X}}_i(\boldsymbol{\phi}-\boldsymbol{\theta}) - \frac{2(\boldsymbol{\phi}-\boldsymbol{\theta})'}{n}\sum_{i=1}^{n}\widetilde{\boldsymbol{X}}_i'\boldsymbol{V}^{-1}(\boldsymbol{Y}_i - \widetilde{\boldsymbol{X}}_i\boldsymbol{\theta}) \\ &+ \frac{1}{n}(\boldsymbol{Y}_i - \widetilde{\boldsymbol{X}}_i\boldsymbol{\theta})'\boldsymbol{V}^{-1}(\boldsymbol{Y}_i - \widetilde{\boldsymbol{X}}_i\boldsymbol{\theta}) + \frac{\lambda_n}{n}\sum_{j=2}^{p+1} \mid \phi_j \mid . \end{aligned}$$

Then, it is easy to show that

$$E\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{\boldsymbol{X}}_i'\boldsymbol{V}^{-1}(\boldsymbol{Y}_i - \widetilde{\boldsymbol{X}}_i\boldsymbol{\theta})\right) = \left(\frac{P(\psi)}{\sigma_\varepsilon^2}, \mathbf{0}'\right)', \tag{4}$$

and

$$E\left(\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{Y}_i - \widetilde{\boldsymbol{X}}_i\boldsymbol{\theta})'\boldsymbol{V}^{-1}(\boldsymbol{Y}_i - \widetilde{\boldsymbol{X}}_i\boldsymbol{\theta})\right) = T. \tag{5}$$

Therefore, with assumptions of Eq. (3), and Eqs. (4) and (5), we have

$$Z_n(\boldsymbol{\phi}) \to Z(\boldsymbol{\phi}),$$

in probability. □

COROLLARY 4.2. $\widehat{\boldsymbol{\theta}}_n \to$ *argmin* $Z(\boldsymbol{\phi})$ *in probability. Since* $Z(\boldsymbol{\phi})$ *is minimized at* $\boldsymbol{\theta}^\star \neq \boldsymbol{\theta}$, *then* $\widehat{\boldsymbol{\theta}}_n$ *is inconsistent.*

Based on the Theorem 4.1 and Corollary 4.2, we can see that ignoring the initial conditions problem in dynamic random-intercepts models can cause serious bias for the regularized maximum likelihood estimates of regression coefficients.

References

1. B. H. Baltagi, *Econometric Analysis of Panel Data*, 2nd ed., John Wiley and Sons, Chichester, 2001.
2. R. Crouchley and R. B. Davies, *A comparison of GEE and random effects models for distinguishing heterogeneity, non stationarity and state dependence in a collection of short binary event series*, Statist. Model. Int. J. **1** (2001) 271–285.
3. P. Diggle, P. Heagerty, K. Y. Liang and S. Zeger, *Analysis of Longitudinal Data*, 2nd ed., New York: University Press, Oxford, 2002.
4. C. Hsiao, *Analysis of Panel Data*, 2nd ed., University Press, Cambridge, 2002.
5. I. Kazemi and R. Crouchley, *Modelling the Initial Conditions in Dynamic Regression Models of Panel Data with Random Effects*, Ch 4. In Panel Data Econometrics, Theoretical Contributions and Empirical Applications, Baltagi BH (ed.) Elsevier: Amsterdam, Netherlands, (2006) 91–117.
6. M. Nerlove and P. Balestra, *Formulation and estimation of econometric models for panel data*, Introductory essay in Mátyás and Sevestre (1996) 3–22.

7. R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. Roy. Statist. Soc. Ser. B **58** (1) (1996) 267–288.

E-mail: r_rikhtehgaran@iut.ac.ir
E-mail: a.mofidian@math.iut.ac.ir

Numerical Evaluation of Sample Sizes in Two Stage Pretest Estimation from a Rayleigh Distribution

Mehran Naghizadeh Qomi*
Department of Statistics, University of Mazandaran, Babolsar, Iran
and Zohre Mahdizadeh
Department of Statistics, University of Mazandaran, Babolsar, Iran

Abstract. In this paper, we consider the problem of expected sample size in a two stage pretest estimation for the scale parameter σ of a Rayleigh distribution. In the presence of prior information for σ, i.e. σ_0, the probability of avoiding the second sample and the expected sample size are derived and plotted for different cases.
Keywords: Rayleigh distribution, Sample size, Two stage estimation.
AMS Mathematical Subject Classification [2010]: 62F15.

1. Introduction

Let $X_1, \ldots, X_n$ be a random sample of size n taken from a Rayleigh distribution with probability density function (p.d.f for short)

$$f(x|\sigma) = \frac{x}{\sigma} \exp\left\{-\frac{x^2}{2\sigma}\right\}, \quad x > 0. \tag{1}$$

The failure rate of the Rayleigh distribution is an increasing function of time which is suitable distribution for components that have no manufacturing defects but age rapidly with time (See [2, 3] for more information).

The maximum likelihood estimator of σ is $\hat{\sigma} = \frac{1}{2n}\sum_{i=1}^{n} X_i^2$. Suppose that we have a priori about the parameter σ in form of a point guess σ_0, i.e. the sample data come from a distribution that is close to a Rayleigh distribution with parameter σ_0. This information may be regarded as a nuisance parameter in the statistical estimation of the model. Such information about the parameter is called nonsample information or uncertain prior information.

In some situations, the researcher can consider a two stage estimator using prior information for achieving a minimum cost of experimentation: he/she consider a small first stage sample and an additional second stage sample for estimation (See [1, 4] and [5]).

In this paper, we propose a two stage pretest estimator in Rayleigh distribution. The probability of avoiding the second sample and the expected sample size are computed and evaluated numerically and graphically for different cases.

*Speaker

2. Two Stage Pretest Estimation

Let $X_{11}, X_{12}, \ldots, X_{1n_1}$ be the first sample of size n_1 taken from the Rayleigh distribution with p.d.f given in (1). The MLE of σ is then given by $\hat{\sigma}_1 = \frac{1}{2n_1}\sum_{i=1}^{n_1} X_{1i}^2$. Now, it is suspected a priori that $\sigma = \sigma_0$ may hold. This information can be tested in the form of the following hypothesis.

$$\begin{cases} H_0, & \sigma = \sigma_0, \\ H_1, & \sigma \neq \sigma_0, \end{cases}$$

at the level of significance α. A likelihood ratio test (LRT) statistic is $\frac{2n_1\hat{\sigma}_1}{\sigma_0} \sim \chi^2_{2n_1}$ under H_0 which has an acceptance region

$$A = \left\{ \hat{\sigma}_1 : \frac{q_1\sigma_0}{2n_1} \leq \hat{\sigma}_1 \leq \frac{q_2\sigma_0}{2n_1} \right\}, \tag{2}$$

where q_1 and q_2 are the values of the lower and upper $100\alpha/2\%$ points of the chi-square distribution with $2n_1$ degrees of freedom, i.e.

$$q_1 = \chi^2_{2n_1}(\frac{\alpha}{2}),\ q_2 = \chi^2_{2n_1}(1 - \frac{\alpha}{2}).$$

If H_0 is accepted, we stop sampling and take the estimator $k\hat{\sigma}_1 + (1-k)\sigma_0$, where $0 \leq k \leq 1$. If not so, we take additional observations $X_{21}, X_{22}, \ldots, X_{2n_2}$ of size n_2 and compute the pooled estimator of σ as

$$\hat{\sigma}_p = \frac{n_1\hat{\sigma}_1 + n_2\hat{\sigma}_2}{n_1 + n_2},$$

where $\hat{\sigma}_2 = \frac{1}{2n_2}\sum_{i=1}^{n_2} X_{2i}^2$ is the MLE of σ based on data in stage two.

3. Expected Sample Size

The probability of avoiding the second sample is

$$\begin{aligned} Pr(A) &= Pr\left(\frac{q_1\sigma_0}{2n_1} \leq \hat{\sigma}_1 \leq \frac{q_2\sigma_0}{2n_1}\right) = Pr\left(q_1\sigma^\star \leq \frac{2n_1\hat{\sigma}_1}{\sigma} \leq q_2\sigma^\star\right) \\ &= \int_{q_1\sigma^\star}^{q_2\sigma^\star} g(t)dt = J(\sigma^\star, \alpha, n_1), \end{aligned} \tag{3}$$

where $\sigma^\star = \sigma_0/\sigma$, A is defined in (2) and $g(t)$ is the density of $T = 2n_1\hat{\sigma}/\sigma \sim \chi^2_{2n_1}$. If $\sigma = \sigma_0$, then, by (3), we have $J(\sigma^\star, \alpha, n_1) = J(1, \alpha, n_1) = 1 - \alpha$. Figure 1 shows the shape of the probability of avoiding the second sample and $J(\sigma^\star, \alpha, n_1)$ given in (3) for selected values of α and n_1 with respect to $\sigma^\star$. We observe that $J(\sigma^\star, \alpha, n_1)$ increases in $(0, 1)$, is clearly $1 - \alpha$ at $\sigma^\star = 1$ and then decreases in $(1, \infty)$.

The expected sample size is given by

$$n^\star = E(n|\sigma) = n_1 + n_2[1 - J(\sigma^\star, \alpha, n_1)].$$

When $\sigma = \sigma_0$, the expected sample size is given by

$$n_0^\star = E(n|\sigma_0) = n_1 + n_2[1 - (1 - \alpha)] = n_1 + n_2\alpha.$$

The plot of the expected sample size, $n^\star$, is given in Figure 2 for selected values α, n_1 and n_2 with respect to $\sigma^\star$. It is observed that the expected sample size is decreasing in $(0, 1)$, is close to n_1 when $\sigma^\star = 1$ and then increasing in $(1, \infty)$. Moreover, the expected sample size increases as n_2 increases for fixed n_1, α and

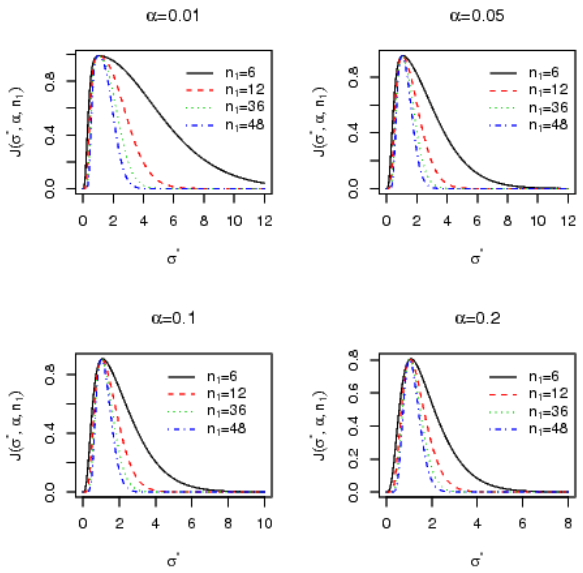


FIGURE 1. Plot of the probability of avoiding the second sample for selected values of α and n_1 with respect to $\sigma^\star$.

$\sigma^\star$. Also, the expected sample size in two stage sampling themes are small with little α.

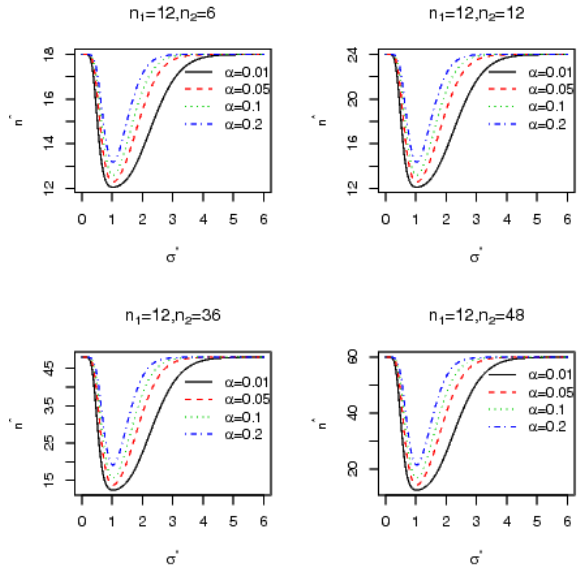


FIGURE 2. Plot of the expected sample size for selected values of n_1 and n_2 highlighting α with respect to $\sigma^\star$.

Acknowledgement

The authors are grateful to the reviewers for making helpful comments and suggestions.

References

1. N. S. Kambo, B. R. Handa and Z. A. Al-Hemyari, *Double stage shrunken estimator of the mean of a normal distribution*, J. Inform. Optim. Sci. **12** (1) (1991) 1–11.
2. M. Naghizadeh Qomi, *Bayesian shrinkage estimation based on Rayleigh type-II censored data*, Comm. Statist. Theory Methods **46** (19) (2017) 9859–9868.
3. M. Naghizadeh Qomi, *Improved estimation in Rayleigh type-II censored data under a bounded loss utilizing a point guess value*, J. Iran. Stat. Soc. **16** (2) (2017) 51–66.
4. R. Srivastava and V. Tanna, *Double stage shrinkage testimator of the scale parameter of an exponential life model under general entropy loss function*, Comm. Statist. Theory Methods **36** (2) (2007) 283–295.
5. V. B. Waikar, F. J. Schuurmann and T. E. Raghunathan, *On a two stage shrunken testimator of the mean of a Normal distribution*, Comm. Statist. Theory Methods **13** (15) (1984) 1901–1913.

E-mail: m.naghizadeh@umz.ac.ir
E-mail: mahdizade_zohre@yahoo.com

Bayesian Inference of Mortality Models in Joint Life Insurance Products

Shirin Shoaee*
Department of Actuarial Science, Faculty of Mathematical Sciences, Shahid Beheshti University, Tehran, Iran
and Akram Kohansal
Department of Statistics, Imam Khomeini International University, Qazvin, Iran

Abstract. In this paper, the Bayesian inference of mortality model is considered in joint life models. We compute the Bayesian estimations using the squared error loss function and a priori distributions that create a dependency between the hyper-parameters for this model of dependent lives. Also, we use the importance sampling method to calculate the Bayes estimations and also to create the corresponding HPD credible intervals. Finally, we analyze one real data set for illustrative purposes.
Keywords: Bayesian analysis, HPD credible interval, Importance sampling method, Joint life insurance.
AMS Mathematical Subject Classification [2010]: 62H10, 62H12, 62E15.

1. Introduction

In many fields of science, including statistics and life insurance, it is assumed that the remainder of the lives of two persons or two components is independent. But applying this assumption is not always correct. Because sometimes there may be identical risk factors for a pair of people and people are exposed to the same risk. For example, in twins, these common risk factors may be genetic or for couples, these common risk factors may be from the environment. Readers can refer to [1, 2, 3] and [4]. One classical model of dependent lives that captured our attention is called the "common shock" model. This model assumes that the lifetimes of two persons, say T_1 and T_2, are independent unless a common shock causes the death of both. For example, a contagious deadly disease, a natural catastrophe or a car accident may affect the lives of the two spouses. See [6, 7], for details.

Recently, the parameters of these models have been estimated using the maximum likelihood estimation and EM algorithm. But the estimation of the parameters by the Bayesian method has not yet been investigated. As we know, the maximum likelihood estimates do not always exist. Another important issue is the convergence of the EM algorithm, which is highly dependent on the initial value selection. Finally, it should be noted that calculating the exact confidence interval for MLEs is not easy. The constructed confidence interval based on the maximum likelihood method is determined using the asymptotic property of MLEs.

*Speaker

In this paper, we use the absolutely continuous bivariate Gopmertz (ACBGP) distribution for the modeling of dependent lives. Also, we assume that the scale parameters have a Dirichlet-Gamma prior distribution. No specific prior distributions are considered for the shape parameter. It is only assumed that this prior distribution is independent of the intended prior distribution for the scale parameters and also the probability density function is log-concave on $(0, \infty)$. It can be seen that explicit expressions cannot be obtained for the Bayesian estimation of parameters. Therefore, numerical methods should be used to calculate Bayesian estimates. So, we propose the importance sampling procedure to generate samples from the posterior distribution of the parameters and to calculate the Bayes estimations and also to construct the HPD credible intervals of the unknown parameters.

This paper is organized as follows: the ACBGP distribution is provided in Section 2. The required assumptions for prior distributions and bivariate data structures are explained in Section 3. The importance sampling structure, and their corresponding HPD credible intervals are described in Section 4. A real data set is analyzed to evaluate the proposed algorithm in Section 5. Finally, the conclusions of this article are presented in Section 6.

2. ACBGP Distribution

In this section, we introduce a classical model of dependent lives based on Gompertz distribution. Suppose T_i follows $(\sim)$ $GP(\alpha, \lambda_i)$ with probability density function $f_{GP}(t, \alpha, \lambda_i) = \alpha\lambda e^{\alpha t} e^{-\lambda_i(e^{\alpha t}-1)}$ for $i = 0, 1, 2$ and also they are independent. Define $X_i = \min\{T_0, T_i\}$, for $i = 1, 2$. Then, the random vector $\mathbf{X} = (X_1, X_2)$ is a bivariate Gompertz distribution and is denoted by $BGP(\alpha, \lambda_0, \lambda_1, \lambda_2)$. The BGP distribution has both an absolutely continuous part and a singular part. Note that the absolutely continuous bivariate Gopmertz (ACBGP) distribution can be obtained from the BGP distribution by removing the singular part and keeping only the continuous part. The joint PDF of ACBGP can be written as

$$f_{ACBGP}(x_1, x_2) = \begin{cases} f_1(x_1, x_2) = c f_{GP}(x_1, \alpha, \lambda_1 + \lambda_0) f_{GP}(x_2, \alpha, \lambda_2), \text{ if } x_2 < x_1, \\ f_2(x_1, x_2) = c f_{GP}(x_1, \alpha, \lambda_1) f_{GP}(x_2, \alpha, \lambda_2 + \lambda_0), \text{ if } x_1 < x_2, \end{cases}$$

in which c is the normalizing constantand and $c = \frac{\lambda_0+\lambda_1+\lambda_2}{\lambda_1+\lambda_2}$.

2.1. Bivariate Data Set. We also assume that

$$\mathcal{D}_1 = \{(x_{11}, x_{21}), \ldots, (x_{1n}, x_{2n})\},$$

is a random sample from the absolutely continuous bivariate Gopmertz distribution. So, the notation $I_1 = \{i : x_{2i} < x_{1i}\}$, $I_2 = \{i : x_{2i} > x_{1i}\}$, $|I_1| = n_1$, $|I_2| = n_2$, and $n = n_1 + n_2$ will be used. Consequently, according to the observations, in this case, the joint likelihood function is as follows:

$$\begin{aligned} &\ell(\mathcal{D}_1|\alpha, \lambda_0, \lambda_1, \lambda_2) = c^n \prod_{i\in I_1} f_1(x_{1i}, x_{2i}) \prod_{i\in I_2} f_2(x_{1i}, x_{2i}) \\ &= c^n \alpha^{2n} \lambda_1^{n_1} \lambda_2^{n_2} (\lambda_0 + \lambda_2)^{n_1} (\lambda_0 + \lambda_1)^{n_2} \\ &\times e^{\alpha\{\sum_{i\in I_1} x_{1i}+x_{2i}+\sum_{i\in I_2} x_{1i}+x_{2i}\}} e^{-\lambda_0\{\sum_{i\in I_1}(e^{\alpha x_{2i}}-1)+\sum_{i\in I_2}(e^{\alpha x_{1i}}-1)\}} \\ &\times e^{-\lambda_1\{\sum_{i\in I_1}(e^{\alpha x_{1i}}-1)+\sum_{i\in I_2}(e^{\alpha x_{1i}}-1)\}} e^{-\lambda_2\{\sum_{i\in I_1}(e^{\alpha x_{2i}}-1)+\sum_{i\in I_2}(e^{\alpha x_{2i}}-1)\}}. \end{aligned}$$

3. Assumptions for Prior Distributions

In this section, we will describe some of the required prior assumptions.

(I): In the first step, we assume that $\lambda = \lambda_0 + \lambda_1 + \lambda_2$ has a prior $\Gamma(a, b)$ distribution,

$$\pi_0(\lambda|a,b) = \frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}, \qquad a>0,\ b>0.$$

Also, given λ, $(\frac{\lambda_1}{\lambda}, \frac{\lambda_2}{\lambda})$ has a Dirichlet prior. We denote that it by $\pi_1(.|a_0, a_1, a_2)$ and the probability density fuction for $\lambda_0 > 0$, $\lambda_1 > 0$ and $\lambda_2 > 0$ is

$$\pi_1(\frac{\lambda_1}{\lambda}, \frac{\lambda_2}{\lambda}|\lambda, a_0, a_1, a_2) = \frac{\Gamma(a_0+a_1+a_2)}{\Gamma(a_0)\Gamma(a_1)\Gamma(a_2)}(\frac{\lambda_0}{\lambda})^{a_0-1}(\frac{\lambda_1}{\lambda})^{a_1-1}(\frac{\lambda_2}{\lambda})^{a_2-1}.$$

Therefore, the joint prior of λ_0, λ_1 and λ_2 is

$$\pi_1(\lambda_0,\lambda_1,\lambda_2|a,b,a_0,a_1,a_2) = \frac{\Gamma(\bar{a})}{\Gamma(a)}(b\lambda)^{a-\bar{a}} \times \prod_{i=0}^{2}\frac{b^{a_i}}{\Gamma(a_i)}\lambda_i^{a_i-1}e^{-b\lambda_i}, \tag{1}$$

where $\bar{a} = a_0 + a_1 + a_2$. The Eq. (1) is a Gamma-Dirichlet distribution and we denote this as $GD(a, b, a_0, a_1, a_2)$.

(II): In the second step, we explain the required assumptions for the shape parameter. We denote the prior distribution on α by $\pi_2(\alpha)$. For this prior distribution, it is only assumed that the support is non-negative on $(0, \infty)$ and that its probability distribution is log-concave. Therefore,

$$\pi(\alpha,\lambda_0,\lambda_1,\lambda_2) = \pi_1(\lambda_0,\lambda_1,\lambda_2)\pi_2(\alpha). \tag{2}$$

4. Bayesian Inference

In this section, the Bayesian estimation of parameters of ACBGP distribution and their corresponding HPD credible intervals are obtained. So, the joint posterior density of $\lambda_i, i = 0, 1, 2$ and α must be obtained using the prior distribution presented in Eq. (2) as follows:

$$\begin{aligned}\ell(\lambda_0,\lambda_1,\lambda_2,\alpha|\mathcal{D}_1) \quad &\propto \quad \lambda^{n+a-\bar{a}}(\lambda_0+\lambda_2)^{n_1}(\lambda_0+\lambda_1)^{n_2}\\ &\times \quad \frac{1}{(\lambda_1+\lambda_2)^n} \times \Gamma(a_0, T_0(\alpha)+b)\\ &\times \quad \Gamma(a_1+n_1, T_1(\alpha)+b) \times \Gamma(a_2+n_2, T_2(\alpha)+b)\\ &\times \quad \frac{\pi_2(\alpha)\alpha^{2n}\exp\{\alpha[\sum_{i\in I_1}x_{1i}+x_{2i}+\sum_{i\in I_2}x_{1i}+x_{2i}]\}}{[T_0(\alpha)+b]^{a_0}[T_1(\alpha)+b]^{n_1+a_1}[T_2(\alpha)+b]^{n_2+a_2}}.\end{aligned}$$

The Bayesian estimation of any function of λ_0, λ_1 and λ_2 as $\theta(\lambda_0, \lambda_1, \lambda_2, \alpha)$ under the squared error loss function is obtained as follows:

$$\hat{\theta}_{Bayes} = \frac{\int_0^\infty\int_0^\infty\int_0^\infty\int_0^\infty \theta(\lambda_0,\lambda_1,\lambda_2,\alpha)\ell(\lambda_0,\lambda_1,\lambda_2,\alpha|\mathcal{D}_1)d\lambda_0 d\lambda_1 d\lambda_2 d\alpha}{\int_0^\infty\int_0^\infty\int_0^\infty\int_0^\infty \ell(\lambda_0,\lambda_1,\lambda_2,\alpha|\mathcal{D}_1)d\lambda_0 d\lambda_1 d\lambda_2 d\alpha}. \tag{3}$$

It can be seen that phrase Expression (3) is not explicitly specified. The importance sampling method can be used to calculate the Bayesian estimates and the corresponding HPD credible intervals.

Algorithm.

Step 1: Use the method proposed by [5] to generate α_i from the log-concave density $\ell(\alpha|\mathcal{D}_1)$.

Step 2: Generate $(\lambda_{ji}|\alpha_i, \mathcal{D}_1) \sim \Gamma(a_j + n_j, T_j(\alpha_i) + b)$, for $j = 0, 1, 2$, $i = 1, 2, \ldots, N, n_0 = 0$.

Step 3: Therefore, the Bayes estimation is obtained as $\hat{\theta}_{Bayes} = \frac{\sum_{i=1}^{N} \theta_i h(\lambda_{0i}, \lambda_{1i}, \lambda_{2i})}{\sum_{i=1}^{N} h(\lambda_{0i}, \lambda_{1i}, \lambda_{0i})}$, where $h(\lambda_{0i}, \lambda_{1i}, \lambda_{2i}) = \lambda_i^{n+a-\bar{a}}(\lambda_{0i} + \lambda_{2i})^{n_1}(\lambda_{0i} + \lambda_{1i})^{n_2} \frac{1}{(\lambda_{1i}+\lambda_{2i})^n}$ and $\theta_i = \theta(\alpha_i, \lambda_{0i}, \lambda_{1i}, \lambda_{2i})$.

4.1. Credible Intervals. The HPD credible interval of $\theta = \theta(\lambda_0, \lambda_1, \lambda_2)$ is constructed to the same method. For this purpose, we follow the below algorithm.

Algorithm.

Step 1: Calculate $w_i = \frac{h(\lambda_{0i}, \lambda_{1i}, \lambda_{2i})}{\sum_{j=1}^{N} h(\lambda_{0j}, \lambda_{1j}, \lambda_{2j})}$.

Step 2: Rearrange $\{(\theta_1, w_1), \ldots, (\theta_N, w_N)\}$ as $\{(\theta_{(1)}, w_{(1)}), \ldots, (\theta_{(N)}, w_{(N)})\}$, where $\theta_{(1)} < \ldots < \theta_{(N)}$ but $w_{(i)}$ are not ordered and are associated with $\theta_{(i)}$.

Step 3: Compute the consistent Bayes estimate of θ_p as $\hat{\theta}_p = \theta_{(N_p)}$, where N_p is the integer satisfying $\sum_{i=1}^{N_p} w_{(i)} \leq p < \sum_{i=1}^{N_p+1} w_{(i)}$.

Step 4: Construct a $100(1-\gamma)\%$ of θ as $(\hat{\theta}_\delta, \hat{\theta}_{\delta+1-\gamma})$, for $\delta = w_{(1)}, w_{(1)}+w_{(2)}, \ldots, \sum_{i=1}^{N_\gamma} w_{(i)}$. Therefore, a $100(1-\gamma)\%$ HPD credible interval of θ becomes $(\hat{\theta}_{\delta^*}, \hat{\theta}_{\delta^*+1-\gamma})$, where δ^* satisfies $\hat{\theta}_{\delta^*+1-\gamma} - \hat{\theta}_{\delta^*} \leq \hat{\theta}_{\delta^*+1-\gamma} - \hat{\theta}_\delta$ for all δ.

5. Data Analysis

This data set include the remaining lifetime information of 100 persons from the population of couples in the age range of 29 – 70 years at an insurance company in Tehran. For this data set, we assume that parameter α has a prior Gamma distribution function. As mentioned, we have no information about the values of the hyper-parameters. Therefore, we should use the non-informative prior for the Bayesian estimation of the parameters. We use the importance sampling method to calculate parameter estimates using the Bayesian method. First, we need to produce observations from $\ell(\alpha|\mathcal{D}_1)$ using the method of [5]. Also, the histogram of the generated samples as well as the posterior function of α, are presented in Figure 1.

As mentioned, the purpose of this study is to estimate the parameters in this joint survival model for actuarial calculations in joint life insurance products. Now, using the Bayesian method, estimation of survival distribution parameters using the proposed algorithm in importance sampling method for $\alpha, \lambda_0, \lambda_1$ and λ_2 are obtained as 0.1053, 0.4850, 0.1034 and 0.1289, respectively. The 95% HPD credible intervals for $\alpha, \lambda_0, \lambda_1$ and λ_2 are also obtained as (0.0805, 0.1386), (0.2869, 0.5148), (0.0985, 0.1387) and (0.0879, 0.1538), respectively.

6. Conclusion

In this paper, Bayesian estimation in the dependent lives models was investigated based on the absolutely continuous bivariate Gopmertz (ACBGP). For this

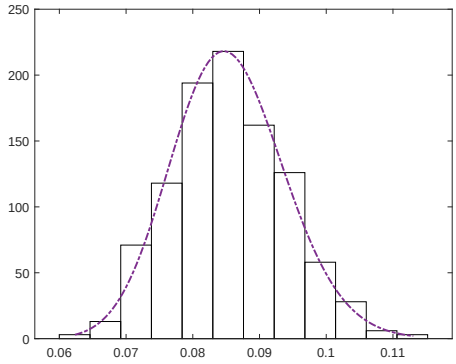


FIGURE 1. The histogram of the generated samples and the posterior density function of α.

purpose, a prior dependent distribution for the scale parameters and a prior distribution for the shape parameter was considered. Also, we assumed that prior distribution on the shape parameter is independent of the joint prior on λ_i. As can be seen, Bayes's estimates were not explicit in this case. Therefore, it was recommended to use the importance sampling method to estimate the parameters. We described in detail the structure of the importance sampling method for calculating their estimates and corresponding HPD credible intervals. Finally, one real data set were used to evaluate the performance of this method.

References

1. J. F. Carriere, *Bivariate survival models for coupled lives*, Scand. Actuar. J. **2000** (1) (2000) 17–32.
2. S. K. Iyer and D. Manjunath, *Correlated bivariate sequences for queueing and reliability applications*, Comm. Statist. Theory Methods **33** (2) (2004) 331–350.
3. C. Jagger and C. J. Sutton, *Death after marital bereavement is the risk increased?*, Stat. Med. **10** (3) (1991) 395–404.
4. S. Kotz, N. Balakrishnan and N. L. Johnson, *Continuous Multivariate Distributions: Models and Applications*, Vol. 1, 2nd ed., John Wiley and Sons, New York, 2000.
5. D. Kundu, *Bayesian inference and life testing plan for weibull distribution in presence of progressive censoring*, Technometrics **50** (2) (2008) 144–154.
6. D. Kundu and R. D. Gupta, *Modied sarhan-balakrishnan singular bivariate distribution*, Statist. Plann. Inference **140** (2) (2010) 526–538.
7. A. M. Sarhan and N. Balakrishnan, *A new class of bivariate distributions and its mixture*, J. Multivariate Anal. **98** (7) (2007) 1508–1527.

E-mail: sh_shoaee@sbu.ac.ir
E-mail: kohansal@sci.ikiu.ac.ir

Contributed Posters

Differential Equations and Dynamical Systems

The Fibering Method Approach to a Singular (p, q)-Laplacian Equation

Fereshteh Behboudi*
Department of Pure Mathematics, Faculty of Science, Imam Khomeini International University, Qazvin, Iran
and Abdolrahman Razani
Department of Pure Mathematics, Faculty of Science, Imam Khomeini International University, Qazvin, Iran

Abstract. In this paper, the existence of two weak non-negative non-trivial solutions of a nonlinear problem involving the (p, q)-Laplacian operator in a bounded domain with smooth boundary in $\mathbb{R}^N$ is proved via fibering method.
Keywords: (p, q)-Laplacian equation, Fibering method, Nehari manifold.
AMS Mathematical Subject Classification [2010]: 35J75, 35D30, 35P30.

1. Introduction

The Nehari manifold is closely related to the fibering maps, that is, maps of the form $t \to I(tu)$, where I is the energy functional associated with the problem. Brown [3] proved the existence of solutions for a semilinear elliptic equation involving the sign-changing weight functions via the Nehari manifold and the fibering method. Papageorgioua et al. [6] studied the existence of positive solutions for a weighted p-Laplacian problem

$$\begin{cases} -div(\xi(x)|\nabla u|^{p-2}\nabla u) = a(x)u^{-\gamma} + \lambda u^r, & x \in \Omega, \\ u = 0, & x \in \partial\Omega, \end{cases}$$

where Ω is a bounded domain with Lipschitz boundary in $\mathbb{R}^N$, $0 < \gamma < 1$ and the differential operator is a weighted p-Laplacian with a weight $\xi \in L^\infty(\Omega)$, $\xi \geq 0$.

Recently, the existence of two weak solutions for a singular (p, q)-Laplacian type equations with the singular terms is proved (See [1]).

In this paper, we are concerned with a quasi-linear problem, that is, a singular (p, q)-Laplacian elliptic problem

$$(1) \quad \begin{cases} -\Delta_p u - \Delta_q u + \theta(x)u^{p-1} = \beta(x)u^{p-1} + \lambda a(x)u^{-\gamma} + b(x)u^{r-1}, & x \in \Omega, \\ u \geq 0, & x \in \Omega, \\ u = 0, & x \in \partial\Omega, \end{cases}$$

where $\Omega \subset \mathbb{R}^N$ is a bounded domain with smooth boundary. The real numbers λ, p, q, γ and r are satisfying the assumptions

$$\lambda > 0, \ 0 < \gamma < 1, \ 1 < q < p < r < p^*,$$

*Presenter

where $p < N$ and $p^* := \frac{Np}{N-p}$. Here, $\Delta_m u = \operatorname{div}(|\nabla u|^{m-2}\nabla u)$ is the m-Laplacian operator for $m \in \{p, q\}$. Furthermore, $\theta \in L^m(\Omega)$ is an indefinite function, where $m > \frac{N}{p}$. The weight functions $a, b, \beta \in L^\infty(\Omega)$ and $a(x), b(x) > 0$ a.e. in Ω.

2. Preliminaries

In this section, we recall the necessary preliminaries and notations. Firstly, we recall that the norm in Lebesgue space $L^p(\Omega)$ is

$$\|u\|_{L^p(\Omega)} = \Big(\int_\Omega |u(x)|^p dx\Big)^{\frac{1}{p}},$$

and the Sobolev space $W_0^{1,p}(\Omega)$ is the closure of $C_0^\infty(\Omega)$ in $W^{1,p}(\Omega)$ endowed with the norm

$$\|u\|_p = \Big(\int_\Omega |\nabla u(x)|^p dx\Big)^{\frac{1}{p}},$$

for every u in $W_0^{1,p}(\Omega)$. Since Ω is bounded and $q < p$, we have the continuous embedding $W_0^{1,p}(\Omega) \hookrightarrow W_0^{1,q}(\Omega)$ such that

$$\|u\|_q \le C\|u\|_p,$$

where $u \in W_0^{1,p}(\Omega)$, for some positive constant $C := C(N, p, q, \Omega)$.

For simplicity, we set $X := W_0^{1,p}(\Omega)$ and $X^* := W_0^{-1,p'}(\Omega)$ (the dual space of X), where $\frac{1}{p} + \frac{1}{p'} = 1$.

The natural space to study (p, q)-Laplacian problems is Sobolev space $W_0^{1,p}(\Omega)$. Notice that $(X, \|\cdot\|_p)$ is a uniformly convex reflexive Banach space.

We consider the generalized eigenvalue problem

$$\begin{cases} -\Delta_p u - \Delta_q u + \theta(x)|u|^{p-2}u = \lambda|u|^{p-2}u, & x \in \Omega, \\ u = 0, & x \in \partial\Omega, \end{cases} \tag{2}$$

where $\Omega \subset \mathbb{R}^N$ is a bounded domain with smooth boundary, $\lambda \in \mathbb{R}$ and $1 < q < p$. Also $\theta \in L^m(\Omega)$ for $m > \frac{N}{p}$. Recently, the fractional form of the problem (2) is studied in [2].

Proposition 2.1. *The problem*

$$\begin{cases} -\Delta_p u + \theta(x)|u|^{p-2}u = \lambda|u|^{p-2}u, & x \in \Omega, \\ u = 0, & x \in \partial\Omega, \end{cases}$$

where $\lambda \in \mathbb{R}$, admits the first eigenvalue λ_1,

$$\lambda_1 := \inf\Big\{\int_\Omega |\nabla u|^p dx + \int_\Omega \theta(x)|u|^p dx : \int_\Omega |u|^p dx = 1\Big\}. \tag{3}$$

Definition 2.2. We say that $u \neq 0$, $u \in X$, is an eigenfunction of λ_1, if the following Euler-Lagrange equation holds for all functions $v \in X$

$$\int_\Omega |\nabla u|^{p-2}\nabla u \nabla v dx + \int_\Omega \theta(x)|u|^{p-2}uv dx = \lambda_1 \int_\Omega |u|^{p-2}uv dx.$$

Moreover, we define

$$\eta_1 := \inf\Big\{\|u\|_p^p + \frac{p}{q}\|u\|_q^q + \int_\Omega \theta(x)|u|^p dx : \int_\Omega |u|^p dx = 1\Big\}. \tag{4}$$

The number η_1 is called the first generalized eigenvalue of (2).

REMARK 2.3. Notice that $\lambda_1 = \eta_1$, where the values of λ_1 and η_1 are given by (3) and (4). In addition the infimum of η_1 is not attained.

PROPOSITION 2.4. [5] *Assume that the function $\beta \in L^\infty(\Omega)$ with $\beta(x) \leq \lambda_1$ a.e. in Ω and* $\text{meas}\{x : \ \beta(x) < \lambda_1\} > 0$, *then there exists $c > 0$ such that*

$$\int_\Omega |\nabla u|^p dx + \int_\Omega \theta(x)|u|^p dx - \int_\Omega \beta(x)|u|^p dx \geq c \int_\Omega |\nabla u|^p dx,$$

for each $u \in X$.

3. Two Solutions

The energy functional $J_\lambda : X \to \mathbb{R}$ associated with (1) is defined as follows

$$J_\lambda(u) := \frac{1}{p}\Big(\|u\|_p^p + \int_\Omega \theta(x)|u|^p dx - \int_\Omega \beta(x)|u|^p dx\Big) + \frac{1}{q}\|u\|_q^q - \frac{\lambda}{1-\gamma}\int_\Omega a(x)|u|^{1-\gamma} dx - \frac{1}{r}\int_\Omega b(x)|u|^r dx,$$

for every $u \in X$. Notice that the functional J_λ is unbounded from below on the space X. Nehari manifold is a good candidate for a subset of X such that the functional J_λ is bounded on it and is as follows

$$N_\lambda := \{u \in X \setminus \{0\} : \ \langle J'_\lambda(u), u\rangle = 0\}.$$

Clearly, critical points of J_λ must lie on N_λ. The fibering map $\phi : \mathbb{R}^+ \to \mathbb{R}$ for the functional J_λ is defined by $\phi_u(t) := J_\lambda(tu)$. These maps are introduced by Drabek and Pohozaev in [4]. For every $u \in X$, we have

$$\phi_u(t) = \frac{t^p}{p}\Big(\|u\|_p^p + \int_\Omega \theta(x)|u|^p dx - \int_\Omega \beta(x)|u|^p dx\Big) + \frac{t^q}{q}\|u\|_q^q - \lambda\frac{t^{1-\gamma}}{1-\gamma}\int_\Omega a(x)|u|^{1-\gamma} dx - \frac{t^r}{r}\int_\Omega b(x)|u|^r dx.$$

Notice that $tu \in N_\lambda$ if and only if $\phi'_u(t) = 0$ and especially $u \in N_\lambda$ if and only if $\phi'_u(1) = 0$. We decompose N_λ into three disjoint parts (See [7])

$$N_\lambda^+ = \{u \in N_\lambda : \ \phi''_u(1) > 0\}, \ \ N_\lambda^0 = \{u \in N_\lambda : \ \phi''_u(1) = 0\},$$
$$N_\lambda^- = \{u \in N_\lambda : \ \phi''_u(1) < 0\}.$$

A computation shows that J_λ is coercive and bounded below on N_λ. Thus we can prove the following lemmas.

LEMMA 3.1. *Suppose u is a maximum or minimum of J_λ on N_λ and $u \notin N_\lambda^0$. Then u is a critical point of J_λ.*

LEMMA 3.2. *There exists $\lambda_0 > 0$ such that for each $\lambda \in (0, \lambda_0)$, $N_\lambda^0 = \emptyset$.*

For applying the fibering method, we define the function $F_u : \mathbb{R}^+ \to \mathbb{R}$ by

$$F_u(t) := t^{p+\gamma-1}\Big(\|u\|_p^p + \int_\Omega \theta(x)|u|^p dx - \int_\Omega \beta(x)|u|^p dx\Big) + t^{q+\gamma-1}\|u\|_q^q - t^{r+\gamma-1}\int_\Omega b(x)|u|^r dx.$$

Clearly, $tu \in N_\lambda$ if and only if t is a solution of equation

$$F_u(t) = \lambda \int_\Omega a(x)|u|^{1-\gamma}dx.$$

Since $\int_\Omega b(x)|u|^r dx > 0$, it is clear that $F_u(t) \to -\infty$ as $t \to \infty$, $F'_u(t) > 0$ for t small enough and $F'_u(t) < 0$ for t large enough. We prove that there exists unique $t_{\max} > 0$ such that $F'_u(t_{\max}) = 0$. It is worth noting that F_u is increasing in $(0, t_{\max})$ and decreasing in $(t_{\max}, \infty)$. Then, there exist $t_1 < t_{\max}$ and $t_2 > t_{\max}$ such that $F_u(t_1) = F_u(t_2) = \lambda \int_\Omega a(x)|u|^{1-\gamma}dx$. That means $t_1 u, t_2 u \in N_\lambda$. Also, $F'_u(t_1) > 0$ and $F'_u(t_2) < 0$ leads to $t_1 u \in N^+_\lambda$ and $t_2 u \in N^-_\lambda$. We set

$$m_\lambda := \inf_{u \in N_\lambda} J_\lambda(u), \ m^+_\lambda := \inf_{u \in N^+_\lambda} J_\lambda(u), \ m^-_\lambda := \inf_{u \in N^-_\lambda} J_\lambda(u).$$

By using the properties of fibering maps, we can prove the existence of two positive solutions, which one of them is in N^+_λ and the other one is in N^-_λ.

PROPOSITION 3.3. *There exists $\hat{\lambda} \in (0, \lambda_0]$ and $u^* \in N^+_\lambda$ such that $J_\lambda(u^*) = m^+_\lambda = \inf_{N^+_\lambda} J_\lambda$, for every $\lambda \in (0, \hat{\lambda})$. Moreover, $u^*(x) \geq 0$ for every $x \in \Omega$.*

PROPOSITION 3.4. *If $\lambda \in (0, \hat{\lambda})$, then the problem (1) admits a weak positive solution $u^* \in X$ such that $u^* > 0$ in Ω and $J_\lambda(u^*) < 0$.*

We can minimize J_λ on the Nehari manifold N^-_λ and get the second non-negative solution.

PROPOSITION 3.5. *There exists $\tilde{\lambda} \in (0, \lambda_0]$ and $v^* \in N^-_\lambda$ such that $J_\lambda(v^*) = m^-_\lambda = \inf_{N^-_\lambda} J_\lambda$, for every $\lambda \in (0, \tilde{\lambda})$.*

PROPOSITION 3.6. *If $\lambda \in (0, \tilde{\lambda})$, then v^* is a weak solution of problem (1) such that $v^* > 0$ in Ω and $J_\lambda(v^*) > 0$.*

Now we state the main result of the paper.

THEOREM 3.7. *Assume that the function $\beta \leq \lambda_1$, where the value of λ_1 is given by (3) and $\text{meas}\{x : \beta(x) \leq \lambda_1\} > 0$. Then there exists $\lambda^* > 0$ such that for every $\lambda \in (0, \lambda^*)$, the problem (1) admits at least two weak non-negative non-trivial solutions.*

By the above propositions, one can prove the existence of two weak solutions. If we set $\lambda^* = \min\{\hat{\lambda}, \tilde{\lambda}\}$ and since $N^+_\lambda \cap N^-_\lambda = \emptyset$, one can conclude that u^* and v^* are distinct.

References

1. F. Behboudi and A. Razani, *Two weak Solutions for a singular (p, q)-Laplacian problem*, Filomat **33** (2019) 3399–3407.
2. F. Behboudi, A. Razani and M. Oveisiha, *Existence of a mountain pass solution for a nonlocal fractional (p, q)-Laplacian problem,* Bound. Value Probl. (2020) 149. DOI:10.1186/s13661-020-01446-w
3. K. J. Brown and Y. Zhang, *The Nehari manifold for a semilinear elliptic equation with a sign-changing weight function*, J. Differential Equations **193** (2003) 481–499.
4. P. Drabek and S. I. Pohozaev, *Positive solutions for the p-Laplacian: Application of the fibering method*, Proc. Roy. Soc. Edinburgh Sect. A **127** (1997) 703–726.

5. D. A. Kandilakis and M. Magiropoulos, *Existence of solutions for (p, q)-Laplacian equations with an indefinite potential*, Complex Var. Elliptic Equ. (2019). DOI:10.1080/17476933.2019.1631289
6. N. S. Papageorgiou and P. Winkert, *Positive solutions for weighted singular p-Laplace equations via Nehari manifolds*, Appl. Anal. (2019). DOI:10.1080/00036811.2019.1688791
7. G. Tarantello, *On nonhomogeneous elliptic equations involving critical Sobolev exponent*, Ann. Inst. H. Poincaré Anal. Non Linéaire **9** (1992) 281–304.

E-mail: f.behboudi@edu.ikiu.ac.ir
E-mail: razani@sci.ikiu.ac.ir

The Existence and Uniqueness of Solution for Fuzzy Differential Equations in Dual Form

Mehran Chehlabi*
Department of Mathematics, Savadkooh Branch, Islamic Azad University, Savadkooh, Iran

Abstract. In this paper, we introduce the dual form of a fuzzy differential equation, the so-called dual fuzzy differential equation. We obtain the results of existence and uniqueness of solution to a class of dual fuzzy differential equations from the point of view G-differentiability concept.
Keywords: Fuzzy, Fuzzy differential equations, Fuzzy dual differential equations.
AMS Mathematical Subject Classification [2010]: 34A07, 34K36, 35R13.

1. Introduction

One of the practical and important issues in various branches of science is solving the differential equations in conditions of uncertainty. The fuzzy concept is a powerful tool for expressing uncertainty in phenomena, such as the imprecise initial value or the boundary value problems. [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], for instance. In this paper, we will consider a class of fuzzy differential equations ($FDEs$) which appear in the dual form, named dual fuzzy differential equation ($DFDE$ for short).

Consider a FDE in the following general form

$$f_1\big(t, x(t), x'(t), x''(t), \ldots, x^{(n)}(t)\big) = f_2\big(t, x(t), x'(t), x''(t), \ldots, x^{(n)}(t)\big), \tag{1}$$

where t is the independent variable and $x'(t), x''(t), \ldots, x^{(n)}(t)$ are first n derivatives of the unknown fuzzy-valued function $x : I \subset \mathbb{R} \to \mathbb{R}_F$ and f_1 and f_2 are two continuous fuzzy-valued functions. Such an equation is said to be in dual form if the function x or at least one of the its derivatives up to order n appears on both sides of the equation.

We point out that in the fuzzy case the Eq. (1) may be not rewritable as the following equation

$$x^{(n)}(t) = f\big(t, x(t), x'(t), x''(t), \ldots, x^{(n-1)}(t)\big), \tag{2}$$

where f is a fuzzy-valued function. The first step in dealing with a FDE is to apply an appropriate derivative concept (See [3, 7] for instance). Some results of the existence and uniqueness of solutions to Eq. (2) as fuzzy initial value problem under Hukuhara derivative concept are obtained in [10]. And the results of existence and uniqueness of solutions under generalized differentiability concept (G-differentiability) to Eq. (2) of second order are found in [1]. Under G-differentiability, Bede et al. [4, 5] have obtained the explicit formulas of solutions

*Presenter

to first-order fuzzy linear differential equations in form (2) and the explicit formulas of solutions related to other forms of whose equations are obtained in [2, 6]. To the best of our knowledge, $FDEs$ in the dual case have not been studied so far. Accordingly, in order to develop the results on $FDEs$, we study the problem existence and uniqueness of solutions to a class of $DFDEs$ emerged as follows:

$$x^{(n)}(t) = f\big(t, x(t), x'(t), x''(t), \dots, x^{(n)}(t)\big). \tag{3}$$

We now give some definitions and introduce the necessary notation which will be used throughout the paper, see [2, 4, 9].

The symbol $\mathbb{R}_F$ denotes the set of all fuzzy numbers defined on real numbers $\mathbb{R}$. The metric structure $d : \mathbb{R}_F \times \mathbb{R}_F \to [0, +\infty)$ is given in terms of the Hausdorff distance by

$$d(u, v) = \sup_{\alpha\in[0,1]} \max\{|u_\alpha^- - v_\alpha^-|, |u_\alpha^+ - v_\alpha^+|\},\ u, v \in \mathbb{R}_{\mathbb{F}}.$$

In this paper, we fix $I = [t_1, t_2]$ and represent a fuzzy-valued function f on I in the parametric form as $[f(t)]^\alpha = [f_\alpha^-(t), f_\alpha^+(t)],\ \forall t \in I,\ \forall \alpha \in [0, 1]$.

Definition 1.1. Let $u, v \in \mathbb{R}_F$. If there exists $w \in \mathbb{R}_F$ such that, $u = v + w$ then w is called the H-difference of u, v and it is denoted as $u \ominus v$.

Definition 1.2. Let $f : I \to \mathbb{R}_F$ and $t_0 \in I$ and $I \subset \mathbb{R}$ is an open interval. We say that f is generalized Hukuhara differentiable (GH-differentiable) at t_0 if there exists an element $f'(t_0) \in \mathbb{R}_F$ such that either

1) for all $h > 0$ sufficiently small, $\exists\ f(t_0 + h) \ominus f(t_0)$, $f(t_0) \ominus f(t_0 - h)$ and

$$\lim_{h\to0^+} \frac{f(t_0 + h) \ominus f(t_0)}{h} = \lim_{h\to0^+} \frac{f(t_0) \ominus f(t_0 - h)}{h} = f'(t_0),$$

or

2) for all $h > 0$ sufficiently small, $\exists\ f(t_0) \ominus f(t_0 + h)$, $f(t_0 - h) \ominus f(t_0)$ and

$$\lim_{h\to0^+} \frac{f(t_0) \ominus f(t_0 + h)}{(-h)} = \lim_{h\to0^+} \frac{f(t_0 - h) \ominus f(t_0)}{(-h)} = f'(t_0).$$

Moreover, we say that f is (1)-differentiable at t_0, if it is GH-differentiable at t_0 in sense (1) and f is (2)-differentiable at t_0, if it is GH-differentiable at t_0 in sense (2).

In this paper, we assume that the fuzzy-valued functions f are (1)-differentiable for all $t \in I$ or are (2)-differentiable for all $t \in I$. Also, we denote by $C^n(I, \mathbb{R}_F)$ the set of continuous functions $f : I \to \mathbb{R}_F$ such that derivatives $f', f'', \dots, f^{(n)} : I \to \mathbb{R}_F$ exist as continuous functions.

2. Main Results

We know that the space $C(I, \mathbb{R}_F)$ of continuous functions $f : I \to \mathbb{R}_F$ is a complete metric space with the distance

$$D(f, g) = \sup_{t\in I}\{d(f(t), g(t))\},$$

or generally

$$H(x, y) = \sup_{t\in I}\{d(x(t), y(t))e^{-\rho t}\},$$

where $\rho \in \mathbb{R}$ is fixed.

THEOREM 2.1. *Let f be (2)-differentiable on $I = [t_1, t_2]$ and assume that the derivative f' is integrable over I. Then for each $t \in I$ we have*

$$f(t) = f(t_2) + (-1) \odot \int_t^{t_2} f'(s)ds.$$

PROOF. Let $[f(t)]^\alpha = [f_\alpha^-(t), f_\alpha^+(t)]$, for each $\alpha \in [0,1]$. Since f is (2)-differentiable,we have $[f'(t)]^\alpha = [f_\alpha^{+\prime}(t), f_\alpha^{-\prime}(t)]$. Then

$$\begin{aligned}\left[f(t_2) + (-1) \odot \int_t^{t_2} f'(s)ds\right]^\alpha &= [f_\alpha^-(t_2), f_\alpha^+(t_2)] + (-1) \odot \left[\int_t^{t_2} f_\alpha^{+\prime}(s)ds, \int_t^{t_2} f_\alpha^{-\prime}(s)ds\right] \\ &= \left[f_\alpha^-(t_2) - \int_t^{t_2} f_\alpha^{-\prime}(s)ds, f_\alpha^+(t_2) - \int_t^{t_2} f_\alpha^{+\prime}(s)ds\right] \\ &= [f_\alpha^-(t), f_\alpha^+(t)] = [f(t)]^\alpha.\end{aligned}$$

□

DEFINITION 2.2. We say that $x : I = [t_1, t_2] \to \mathbb{R}_F$ is $(1)^{(n)}$-solution ($(2)^{(n)}$-solution) to Eq. (1) on I, if $x(t)$ and its derivatives up to order $n-1$ are (1)-differentiable ((2)-differentiable, respectively) for each $t \in (t_1, t_2)$ and satisfy the Eq. (1), for each $t \in I$.

THEOREM 2.3.

i) *The function $x \in C_1^n(I, \mathbb{R})$ is $(1)^{(n)}$-solution to Eq. (2), if and only if the function $y(t) = x^{(n)}(t)$ satisfies the following integral equation*

$$y(t) = f(t, (I_n y)(t) + \varphi_{n-1}(t), (I_{n-1} y)(t) + \varphi_{n-2}(t), \ldots, (I_1 y)(t) + \varphi_0(t), y(t)), \tag{4}$$

where

$$(I_m y)(t) = \frac{1}{(m-1)!} \int_{t_1}^t (t-s)^{m-1} \odot y(s)ds, \ m = 1, 2, \ldots, n,$$

$$\varphi_m(t) = \frac{d^{n-m-1}}{dt} \left(\sum_{i=0}^{n-1} \frac{(t-t_1)^i}{i!} \odot x^{(i)}(t_1) \right), \ m = 0, 1, \ldots, n-1,$$

and $x^{(0)}(t) = x(t)$, $x^{(1)}(t) = x'(t)$ and $x^{(2)}(t) = x''(t)$.

ii) *The function $x \in C_2^n(I, \mathbb{R})$ is $(2)^{(n)}$-solution to Eq. (2), if and only if the function $y(t) = x^{(n)}(t)$ satisfies the following integral equation*

$$y(t) = f(t, (J_n y)(t) + \psi_{n-1}(t), (J_{n-1} y)(t) + \psi_{n-2}(t), \ldots, (J_1 y)(t) + \psi_0(t), y(t)), \tag{5}$$

where

$$(J_m y)(t) = \frac{(-1)^m}{(m-1)!} \int_t^{t_2} (s-t)^{m-1} \odot y(s)ds, \ m = 1, 2, \ldots, n,$$

and

$$\psi_m(t) = \frac{d^{n-m-1}}{dt} \left(\sum_{i=0}^{n-1} \frac{(-1)^i (t_2-t)^i}{i!} \odot x^{(i)}(t_2) \right), \ m = 0, 1, \ldots, n-1.$$

PROOF. Let us prove the case (ii), the proof is similar for case (i). At first it is easy to see that

$$\sum_{i=0}^{n-1} \frac{(-1)^i (t_2-t)^i}{i!} \odot x^{(i)}(t_2) \in C_2^n(I, \mathbb{R}),$$

and further

$$\psi_m(t) = \sum_{i=0}^{n-1} \frac{d^{n-m-1}}{dt} \left(\frac{(-1)^i (t_2 - t)^i}{i!} \right) \odot x^{(i)}(t_2), \ m = 0, 1, \ldots, n-1.$$

For $t \in [t_1, t_2]$, we get

$$\int_t^{t_2} x^{(n)}(s_n) ds_n = \int_t^{t_2} y(s_n) ds_n.$$

That gives, by Theorem 2.1,

$$x^{(n-1)}(t) = x^{(n-1)}(t_2) - \int_t^{t_2} y(s_n) ds_n = \psi_0(t) + (J_1 y)(t).$$

Similarly, from the last equality we get

$$\begin{aligned}
x^{(n-2)}(t) &= x^{(n-2)}(t_2) - \int_t^{t_2} (x^{(n-1)}(t_2) - \int_{s_{n-1}}^{t_2} y(s_n) ds_n) ds_{n-1} \\
&= x^{(n-2)}(t_2) - (t_2 - t) \odot x^{(n-1)}(t_2) + \int_t^{t_2} \int_{s_{n-1}}^{t_2} y(s_n) ds_n ds_{n-1} \\
&= \psi_1(t) + \int_{s_{n-2}}^{t_2} \int_{s_{n-2}}^{s_n} y(s_n) ds_{n-1} ds_n = \psi_1(t) + \int_{s_{n-2}}^{t_2} (s_n - s_{n-2}) \odot y(s_n) ds_n \\
&= \psi_1(t) + \int_t^{t_2} (s - t) \odot y(s) ds = \psi_1(t) + (J_2 y)(t).
\end{aligned}$$

By recurrence,

$$\begin{aligned}
x(t) = \ x^{(0)}(t) &= x^{(0)}(t_2) - \int_t^{t_2} x^{(1)}(s_1) ds_1 = x(t_2) - \int_t^{t_2} \left(x^{(1)}(t_2) - \int_{s_1}^{t_2} x^{(2)}(s_2) ds_2 \right) ds_1 \\
&= x(t_2) - (t_2 - t) \odot x^{(1)}(t_2) + \int_t^{t_2} \int_{s_1}^{t_2} \left(x^{(2)}(t_2) - \int_{s_2}^{t_2} x^{(3)}(s_3) ds_3 \right) ds_2 ds_1 \\
&= x(t_2) - (t_2 - t) \odot x^{(1)}(t_2) + \frac{(t_2 - t)^2}{2} \odot x^{(2)}(t_2) - \int_t^{t_2} \int_{s_1}^{t_2} \int_{s_2}^{t_2} x^{(3)}(s_3) ds_3 ds_2 ds_1 \\
&= \cdots \\
&= \sum_{i=0}^{n-1} \frac{(-1)^i (t_2 - t)^i}{i!} \odot x^{(i)}(t_2) + (-1)^n \odot \int_t^{t_2} \int_{s_1}^{t_2} \int_{s_2}^{t_2} \cdots \int_{s_{n-1}}^{t_2} x^{(n)}(s_n) ds_n ds_{n-1} \cdots ds_2 ds_1 \\
&= \psi_{n-1}(t) + (J_n y)(t).
\end{aligned}$$

□

THEOREM 2.4. *Let $f : I \times \mathbb{R}_F^{n+1} \to \mathbb{R}_F$ be continuous, and suppose that $M_0, M_1, \ldots, M_n > 0$ exist such that*

$$d\left(f(t, u_0, u_1, \ldots, u_n), f(t, v_0, v_1, \ldots, v_n)\right) < \sum_{i=0}^{n} M_i d(u_i, v_i),$$

for all $t \in I$, $u_0, u_1, \ldots, u_n, v_0, v_1, \ldots, v_n \in \mathbb{R}_F$ and $M_n < 1$. Then the integral Eq. (4) has a unique solution on I.

THEOREM 2.5. *Let $f : I \times \mathbb{R}_F^{n+1} \to \mathbb{R}_F$ be continuous, and suppose that $M_0, M_1, \ldots, M_n > 0$ exist such that*

$$d\left(f(t, u_0, u_1, \ldots, u_n), f(t, v_0, v_1, \ldots, v_n)\right) < \sum_{i=0}^{n} M_i d(u_i, v_i),$$

for all $t \in I$, $u_0, u_1, \ldots, u_n, v_0, v_1, \ldots, v_n \in \mathbb{R}_F$ *and* $M_n < 1$. *Then the integral Eq.* (5) *has a unique solution on* $I = [t_1, t_2]$ *with* t_2 *satisfying the following condition*

$$\sum_{i=0}^{n} M_i \frac{(t_2 - t_1)^{n-i}}{(n-i)!} < 1.$$

References

1. T. Allahviranloo, N. A. Kiani and M. Barkhordari, *Toward the existence and uniqueness of solutions of second-order fuzzy differential equations*, Inform. Sci. **179** (2009) 1207–1215.
2. T. Allahviranloo and M. Chehlabi, *Solving fuzzy differential equations based on the length function properties*, Soft Comput. **19** (2015) 307–320.
3. L. C. Barros, L. T. Gomes and P. A. Tonelli, *Fuzzy differential equations: An approach via fuzzification of the derivative operator*, Fuzzy Sets Syst. **230** (2013) 39–52.
4. B. Bede and S. G. Gal, *Solution of fuzzy differential equations based on generalized differentiability*, Commun. Math. Anal. **9** (2010) 22–41.
5. B. Bede and L. Stefanini, *Solution of fuzzy differential equations with generalized differentiability using LU-parametric representation*, Eusflat (2011) 785–790.
6. M. Chehlabi and T. Allahviranloo, *Positive or negative solutions to first-order fully fuzzy linear differential equations under generalized differentiability*, Appl. Soft Comput. **70** (2018) 359–370.
7. N. A. Gasilov, Ş. E. Amrahov and A. G. Fatullayev, *A new approach to fuzzy initial value problem*, Soft Comput. **18** (2014) 217–225.
8. A. Khastan and R. Rodríguez-López, *On the solutions to first order linear fuzzy differential equations*, Fuzzy Sets Syst. **295** (2016) 114–135.
9. A. Khastan and R. Rodríguez-López, *An existence and uniqueness result for fuzzy Goursat partial differential equation*, Fuzzy Sets Syst. **375** (2019) 141–160.
10. D.N. Georgioua, Juan J. Nieto and R. Rodríguez-López, *Initial value problems for higher-order fuzzy differential equations*, Nonlinear Anal. **63** (2005) 587–600.

E-mail: chehlabi@yahoo.com

Generalized Two-Sided Shift Map

Sanaz Lamei*
Faculty of Mathematical sciences, University of Guilan, Rasht, Iran
and Pouya Mehdipour
Federal University of Viçosa, Brasil

Abstract. The two-sided shift maps are automorphisms and one-sided shift maps are endomorphisms. These maps can be conjugate or semi-conjugate to some automorphisms or endomorphisms which admit appropriate Markov partitions. Here, we aim to introduce a generalized two-sided shift map for endomorphisms.
Keywords: Shift map, Endomorphism.
AMS Mathematical Subject Classification [2010]: 37B10, 08A35.

1. Introduction

Let $\mathcal{A} = \{a, \dots, b\}$ be the set of finite alphabets and $X = \{(x_i)_{i\in\mathbb{Z}}, x_i \in \mathcal{A}\}$ be a two-sided shift space. The shift map on Z is defined as $(\sigma(x))_i = x_{i+1}$ for $i \in \mathcal{Z}$. Since the two-sided shift maps are defined for automorphisms, we have $(\sigma^{-1}(x))_i = x_{i-1}$. Sometimes a shift map is conjugate to some map f on a space Y with an appropriate Markov partition. Then the orbit of the points in Z under f are in one-to-one correspondence to their itineraries in X. There are some texts describing different features of shift spaces and shift maps such as [1, 2, 3, 4, 5, 6].

If f is an endomorphism, then the map f and σ are semi-conjugate and the endomorphism shift map is defined on an one-sided shift space. So, σ loses the track of points in pre-images of f. Here we aim to introduce a generalized two-sided shift map which represents the full orbit of endomorphism maps. Also we generalize some concepts of classical shift maps to the generalized two-sided shift spaces.

2. Generalized Two-Sided Endomorphism Shift Map

Let $\mathcal{A} = \{a, \dots, b\}$ and $\mathcal{A}' = \{a', \dots, b'\}$ be two finite sets of alphabets and Y be the set of two-sided sequences over $\mathcal{A}'$ whose admissible words of length n is denoted by ${\mathcal{L}'}_n$. Let $\varphi_n : {\mathcal{L}'}_n \longrightarrow \mathcal{A}$ be a factor map. Define

$$x_i = \begin{cases} y_i \in \mathcal{A}', & \forall i \geq 0, \\ x_i = \varphi_n(y_i \dots y_{i+n-1}) \in \mathcal{A}, & \forall i < 0. \end{cases} \tag{1}$$

Now define the *generalized two-sided shift space* as the set

$$\Sigma_{\mathcal{A},\mathcal{A}'} := \{x = (x_i)_{i\in\mathbb{Z}} \ : x_i \text{ satisfies}(1)\},$$

*Presenter

and the *generalized two-sided shift map* on $\Sigma_{\mathcal{A},\mathcal{A}'}$ as

$$(\sigma(x))_i = \begin{cases} \varphi_n(x_0 \dots x_{n-1}), & \text{if } i = -1, \\ x_{i+1}, & \text{otherwise.} \end{cases}$$

THEOREM 2.1. *The generalized two-sided shift map is well-defined and the generalized shift space is closed under the generalized two-sided shift map.*

Define $\Sigma_{\mathcal{A}'} = \{(x_0, x_1, \dots) : (x_i)_{i\in\mathbb{Z}} \in \Sigma_{\mathcal{A},\mathcal{A}'}\}$ and $\Sigma_{\mathcal{A}} = \{(\dots, x_{-2}, x_{-1}) : (x_i)_{i\in\mathbb{Z}} \in \Sigma_{\mathcal{A},\mathcal{A}'}\}$. Let $\beta^{\mathcal{A},\mathcal{A}'}(\Sigma)$ be the set of admissible blocks in $\Sigma_{\mathcal{A},\mathcal{A}'}$ and $\beta_n^{\mathcal{A}}$ and $\beta_n^{\mathcal{A}'}$ be the subsets of $\beta^{\mathcal{A},\mathcal{A}'}(\Sigma)$ with alphabets in $\mathcal{A}$ and $\mathcal{A}'$ respectively.

Some of the generalized two-sided shift spaces can be specified by a list of forbidden blocks. It can be done by both specifying the forbidden blocks of the $\Sigma_{\mathcal{A}'}$ part or the forbidden blocks of the hole space. Denote them by $\mathcal{F}_{\mathcal{A}'}$ and $\mathcal{F}$ respectively. In the later case, if the number and the length of blocks in $\mathcal{F}$ are finite, then we call $\Sigma_{\mathcal{A},\mathcal{A}'}$ an M-step Markov space.

If the number and the length of blocks in $\mathcal{F}$ or $\mathcal{F}_{\mathcal{A}'}$ are finite, then the space $\Sigma_{\mathcal{A},\mathcal{A}'}$ can be represented by a graph. Each directed edge receives two labels, one from $\mathcal{A}$ and one from $\mathcal{A}'$. Walking in direction of edges and considering the labels from $\mathcal{A}'$, we get entries with non-negative indices and walking in opposite direction of edges and picking the labels from $\mathcal{A}$, we get the entries with negative indices of a point $x \in \Sigma_{\mathcal{A},\mathcal{A}'}$.

DEFINITION 2.2. A point $x = (x_i)_{i\in\mathbb{Z}} \in \Sigma$ is a periodic point with period m if for $i \geq 0$, $x_{mi} = a'_0$, $x_{mi+1} = a'_1$, …, $x_{mi+m-1} = a'_{m-1}$ and also, $x_{mi} = \varphi_n(x_0, \dots, x_{n-1})$, $x_{mi+1} = \varphi_n(x_1, \dots, x_n)$, …, $x_{mi+m-1} = \varphi_n(x_{m-1}, \dots, x_{m+n-2})$ for $i \leq -1$.

We represent the periodic orbit via its non-negative part.

EXAMPLE 2.3. Let $\mathcal{A}' = \{a, b, c, d\}$ and $\mathcal{A} = \{0, 1\}$ be two sets of alphabets and $\mathcal{F}_{\mathcal{A}'} = \{ac, ba, bb, bc, cb, cc, cd, da\}$ be the set of forbidden words. If $\varphi(a) = \varphi(b) = 1$ and $\varphi(c) = \varphi(d) = 0$, then

$$\begin{aligned} x &= (\dots, x_{-3}, x_{-2}, x_{-1}; x_0, x_1, x_2, x_3, x_4, \dots) \\ &= (\dots, \ 0, \ \ 1, \ \ 1; \ \ d, \ b, \ d, \ c, \ a, \ \dots), \end{aligned}$$

is a point in $\Sigma_{\mathcal{A},\mathcal{A}'}$. This space can be shown by the graph in Figure 1. Let $x = (\dots, 1, 0, 0, 1, 0, 0; a, d, c, a, d, c, \dots)$ and $y = (\dots, 1, 0, 0, 1, 0, 0; b, d, d, b, d, d, \dots)$ be two periodic points. The negative indecis for negative part for both points is 100. So, we represent the points via non-negative indices which are $x = (\overline{a, d, c})$ and $y = (\overline{b, d, d})$. The negative indices are determined by the non-negative part and the map ϕ.

Let $\mathcal{L}_n$ be the admissible words in $\Sigma \subseteq \Sigma_{\mathcal{A},\mathcal{A}'}$.

DEFINITION 2.4. For some natural number N, define the $h_N : \Sigma \to (\mathcal{L}_N)^{\mathbb{Z}}$ as

$$(h_N(x))_{[i]} := x_{[i,\, i+N-1]},$$

and the Nth generalized higher block shift as $h_N(\Sigma)$.

DEFINITION 2.5. For some natural number N, let $\gamma_N : \Sigma \to (\mathcal{L}_N)^{\mathbb{Z}}$ be defined as

$$(\gamma_N(x))_{[i]} := x_{[iN,\, iN+N-1]}$$

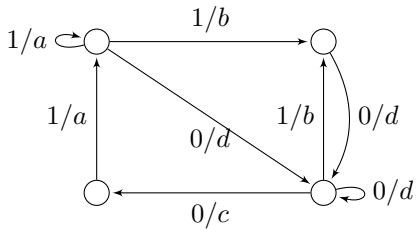


FIGURE 1. This graph represents generalized two-sided shift space $\Sigma_{\mathcal{A},\mathcal{A}'}$ with $\mathcal{A}' = \{a, b, c, d\}$ and $\mathcal{A} = \{0, 1\}$.

and define the Nth generalized higher power shift as $\gamma_N(\Sigma)$.

THEOREM 2.6. *The Nth generalized higher block shifts and Nth generalized higher power shifts of a generalized two-sided shift space are generalized two-sided shift spaces.*

DEFINITION 2.7. Let $\Sigma_{\mathcal{A},\mathcal{A}'}$ and $\Sigma_{\mathcal{C},\mathcal{C}'}$ be two generalized shift spaces with factor maps φ_n and φ_m respectively. For $k \geq n$ define the map $\Phi : \mathcal{L}_k \to \mathcal{C} \cup \mathcal{C}'$ as

$$\Phi(x_{[i,\, i+\ell-1]}) = \begin{cases} y_i' \in \mathcal{C}, & \text{if } x_i, \ldots, x_{i+\ell-1} \in \mathcal{A}', \\ y_i \in \mathcal{C}, & \text{if } x_i \in \mathcal{A}. \end{cases} \tag{2}$$

If there exists $\phi : \Sigma_{\mathcal{A},\mathcal{A}'} \to \Sigma_{\mathcal{C},\mathcal{C}'}$ defined by $\phi(x) = y$ satisfying (2), then it is called the *generalized sliding block code* induced by map Φ.

THEOREM 2.8. *The generalized sliding block codes are generalized two-sided shift spaces.*

THEOREM 2.9. *Let $\Sigma_{\mathcal{A},\mathcal{A}'}$ and $\Sigma_{\mathcal{C},\mathcal{C}'}$ be two generalized two-sided shift spaces with factor maps φ_n and φ_m and $\phi : \Sigma_{\mathcal{A},\mathcal{A}'} \to \Sigma_{\mathcal{C},\mathcal{C}'}$ be a generalized shift sliding block code. Then $\phi \circ \sigma_{\mathcal{A}'} = \sigma_{\mathcal{C}'} \circ \phi$.*

PROOF. Let $x = \ldots x_{-1}; x_0\, x_1 \ldots \in \Sigma_{\mathcal{A},\mathcal{A}'}$ and $y = \ldots y_{-1}; y_0\, y_1 \ldots \in \Sigma_{\mathcal{C},\mathcal{C}'}$. Suppose $\phi(x) = y$ according to equation $\Phi : \mathcal{L}_k \to \mathcal{C} \cup \mathcal{C}'$ defined in (2). For a word $x_{[i,\, i+k-1]} \in \mathcal{L}_k$),

$$\begin{aligned} ((\sigma_{\mathcal{C}'} \circ \phi)(x))_{[i]} &= \phi(\sigma_{\mathcal{C}'}(x))_{[i]} \\ &= \Phi((\sigma_{\mathcal{A}'}(x))_{[i,\, i+k-1]} \\ &= \Phi(x_{[i+1,\, i+k]}) \\ &= (\phi \circ \sigma_{\mathcal{A}'}(x))_{[i]}. \end{aligned}$$

□

THEOREM 2.10. (Generalization of Curtis-Hedlund-Lyndon Theorem) *Let $\Sigma_{\mathcal{A},\mathcal{A}'}$ and $\Sigma_{\mathcal{C},\mathcal{C}'}$ be generalized two-sided shift spaces with factor maps φ_n and φ_m respectively. A map ϕ is a generalized sliding block code if and only if $\phi \circ \sigma_{\mathcal{A}'} = \sigma_{\mathcal{C}'} \circ \phi$ and there exists $k \geq n$ such that $\phi(x)_i$ is a function of $x_{[i,\, i+k-1]}$ for $-k \leq i \leq 1$.*

PROOF. Since $\Sigma_{\mathcal{A},\mathcal{A}'}$ and $\Sigma_{\mathcal{C},\mathcal{C}'}$ are generalized two-sided shift spaces and ϕ is a generalized sliding block code, so the restriction of $\phi(x)_i$ to $x_{[i,\, i+k-1]}$ implies the map Φ over $x_{[i,\, i+k-1]}$ for $-k \leq i \leq 1$ and the diagram be commutative. We

prove the reverse direction. For the blocks $x_{[-m,\,-1]} \in \beta_n^{\mathcal{A}'}$ and $x_{[0,\,m]} \in \beta_n^{\mathcal{A}}$, the map Φ is defined and hence Φ is defined for all blocks in $\beta_n^{\mathcal{A}}$ and $\beta_n^{\mathcal{A}'}$. The map Φ is defined on the central blocks, so it is well-defined on all translations on them in $\beta^{\mathcal{A},\mathcal{A}'}(\Sigma)$, which completes the proof. □

THEOREM 2.11. *Let $\psi : \Sigma_{\mathcal{A},\mathcal{A}'} \to \Sigma_{\mathcal{C},\mathcal{C}'}$ be a generalized sliding block code with the block map $\Psi : \beta_\ell^{\mathcal{A},\mathcal{A}'} \to \mathcal{C} \cup \mathcal{C}'$. Then there exists a generalized higher block space $\widetilde{\Sigma_{\mathcal{A},\mathcal{A}'}}$ of $\Sigma_{\mathcal{A},\mathcal{A}'}$, which is conjugate to a generalized 1-block code $\widetilde{\psi} : \widetilde{\Sigma_{\mathcal{A},\mathcal{A}'}} \to \Sigma_{\mathcal{C},\mathcal{C}'}$ via the conjugacy map $h : \Sigma_{\mathcal{A},\mathcal{A}'} \to \widetilde{\Sigma_{\mathcal{A},\mathcal{A}'}}$. Also, $\widetilde{\psi} \circ h = \psi$.*

PROOF. Since $\Sigma_{\mathcal{A},\mathcal{A}'}$ is a generalized sliding block code, using Theorem 2.10, there exists a map ϕ and a number $k > 0$ such that $\phi(x)_i$ is a function of $x_{[i,\,i+k-1]}$ for $-k \leq i \leq 1$ and $\phi \circ \sigma_{\mathcal{A}'} = \sigma_{\mathcal{C}'} \circ \phi$. For this k, take $\widetilde{\Sigma_{\mathcal{A},\mathcal{A}'}}$ as $\Sigma^{[k]}_{\mathcal{A},\mathcal{A}'}$. Now let $h : \Sigma_{\mathcal{A},\mathcal{A}'} \to \widetilde{\Sigma_{\mathcal{A},\mathcal{A}'}}$ be the kth higher block code $(h(x))_{[i]} = x_{[i,\,i+k-1]}$. The map h is a conjugacy. Define $\widetilde{\phi}$ as $\phi \circ h^{-1}$ which becomes a 1-block map, see Figure 2. □

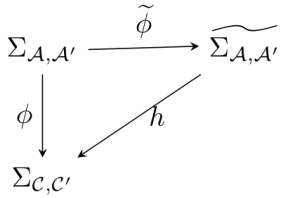


FIGURE 2.

References

1. M. Boyle, *Symbolic dynamics and matrices,* Combinatorial and graph-theoretical problems in linear algebra (Minneapolis, MN, 1991), 1–38, IMA Vol. Math. Appl., 50, Springer, New York, 1993.
2. M. Boyle, J. Franks and B. Kitchens, *Automorphisms of one-sided subshifts of finite type*, Ergod. Th. Dynam. Sys. **10** (1990) 421–449.
3. F. Blanchard and G. Hansel, *Systems codes*, Theoret. Comput. Sci. **44** (1986) 17–49.
4. G. A. Hedlund, *Endomorphisms and automorphisms of the shift dynamical system*, Math. Systems Theory **3** (1969) 320–375.
5. D. Lind, M. Marcus, *An Introduction to Symbolic Dynamics and Coding*, Cambridge University Press, Cambridge, 1995.
6. M. Morse and G. A. Hedlund, *Symbolic dynamics II. Sturmian trajectories*, Amer. J. Math. **62** (1940) 1–42.

E-mail: lamei@guilan.ac.ir
E-mail: Pouya@ufv.br

Chaotic Behaviour of Baker-Like Maps with One Discontinuity Point

Roya Makrooni*
Faculty of Mathematical Sciences, University of Sistan and Baluchestan, Zahedan, Iran
Mehdi Pourbarat
Faculty of Mathematical Sciences, Shahid Beheshti University, Tehran, Iran
and Neda Abbasi
Faculty of Mathematical sciences, Shahid Bahonar University, Kerman, Iran

Abstract. This paper deals with a family of one dimensional discontinuous maps known as Baker like maps. For this family it is studied the problem of existence of chaos according to the well known definition by Devaney. In fact, it is shown that if f is a generalized semi-baker map with two branches and its derivative greater than or equal to $\sqrt{2}$, then the dynamical system related to that is chaotic in the sense of Devaney.
Keywords: Dynamical system, Devaney chaos, Discontinuity point.
AMS Mathematical Subject Classification [2010]: 37B99, 37C70.

1. Introduction

Piecewise smooth dynamical systems have received a great deal of attention in recent years. The essential feature of a piecewise smooth (PWS for short) system, both continuous and discontinuous, which may greatly influence the dynamics, is the presence of a so-called *switching manifold* at which the system's function changes its definition [1]. These systems represent adequate mathematical models for many processes both in nature and engineering. Applications of these systems range from earthquake dynamics to nano-actuator and include electronic devices with relay or switching components (for example, buck/boost converters, mechanical systems with stick-slip or impact phenomena) and in general all switching systems occurring in various field as control theory, economics, biology and so on.

Nowadays many results related to dynamics of piecewise smooth systems and in particular about chaotic dynamics in such systems are already published, especially those related to piecewise linear maps. For example in [3] and [6] some relevant properties of the chaotic sets related to the piecewise linear maps with constant slope, known as $\beta-$Transformation, are determined.

This paper is devoted to an important class of piecewise smooth expanding maps of an interval into itself, constituted by $N \geq 2$ branches called Baker like maps. Regarding Baker like maps with $N \geq 2$ branches, in [2] the authors give the necessary and sufficient conditions for a discontinuous expanding map to be chaotic in the whole interval in terms of homoclinic bifurcations. Also in [7] for Baker like map with infinitely many branches the existence of full measure unbounded chaotic

*Presenter

attractors which are persistent under parameter perturbation (also called robust) has been proved.

Here we consider $N = 2$, that is a map from an interval into itself with one discontinuity point, and give analytical sufficient conditions under which the system is chaotic in the sense of Devaney. This subject is not new in the literature. The basic tools are related to a Baker like map on the interval, with two branches that corresponds to Lorenz maps, which has been deeply studied since many years and is nowadays of common knowledge (See for example [4, 5, 8, 10]). In the present paper a rigorous proof to the existence of chaos in the sense of Devaney in that system is presented in terms of the derivatives of the branches. Here we improve the results of chaotic dynamics of Baker like maps with two branches in [9], which already published by same authors.

In the following, some basic concepts and notations are given, chaos in the sense of Devaney is stated and Baker like map is introduced.

For convenience, suppose that $I = [0,\ 1]$, and $r \in \mathbb{N} \cup \{\infty\}$.

Definition 1.1. The map $f : I \longrightarrow I$ is called a piecewise C^r-smooth one dimensional dynamical system if there exist the points $0 = \xi_0 < \xi_1 < \cdots < \xi_{N-1} < \xi_N = 1$ such that $f|_{(\xi_i,\ \xi_{i+1})}$ is C^r-smooth.

Suppose that f is a piecewise C^r-smooth one dimensional dynamical system. The orbit of the point $x \in I$ is defined as

$$orb(x, f) := \{f^n(x) : n \in \mathbb{N} \cup \{0\}\}.$$

The point x is periodic, if there is a nutral number n such that $f^n(x) = x$. The set of all periodic points is called to $Per(f)$.

The PWS map f is topologically transitive in I, if for any pair of non-empty open sets U and V of I there exists a natural number n such that $f^n(U) \cap V \neq \varnothing$. It is topologically mixing in I if for any pair of non-empty open sets U and V in X there exists a natural number n such that for any $m > n,\ f^m(U) \cap V \neq \varnothing$.

The word *chaos* was first introduced into mathematics by Li and Yorke. Then in 1986 a precise mathematical definition of chaos was suggested by Devaney. In terms of the PWS maps we have: a PWS map $f : I \longrightarrow I$ is chaoticin the sense of Devaney if $Per(f)$ is dense in I, it is topologically transitive and has sensitive dependence on initial conditions in I.

The 1D piecewise C^1-smooth map which we are interested in defined as follows

Definition 1.2. Let N be a natural number and λ a real number both bigger than one and real numbers $\xi_0 = 0 < \xi_1 < \cdots < \xi_{N-1} < \xi_N = 1$ are given. For each $1 \leq i \leq N$ suppose that $I_i = [\xi_{i-1},\ \xi_i)$ and $f_i :\ I_i \longrightarrow [0,\ 1)$ is a differentiable map satisfying ${f_i}' > \lambda$. Also, suppose that f_i is surjective for all $1 < i < N$ and $f_{1-}(\xi_1) = 1$ and $f_{N+}(\xi_{N-1}) = 0$. Then map $f :\ I \longrightarrow I$ given by

$$f(x) := f_i(x) \quad x \in I_i, \quad f(1) := f_{N-}(1),$$

is called a Baker-like map with N branches and expanding rate λ.

The Baker-like maps are a straightforward generalization of Baker maps and Lorenz maps. It is worth to know that Baker like maps are conjugate with expanding maps on S^1 with one discontinuity point.

2. Main Results

We begin this section by proving the following proposition.

PROPOSITION 2.1. *Suppose that f is a Baker-like map with $\lambda > \sqrt{2}$ and $N = 2$. Then for each one of the intervals I_1 and I_2 and each interval $(a,\ b) \subset I_i$, there are one interval $J_i \subset (a,\ b)$ and one natural number k such that f^k is continuous on J_i and $f^k(J_i) = I_i$.*

PROOF. Without losing generality, we may assume that $(a,\ b) \cap \{\xi_1, f^{-1}(\xi_1)\}$ is an empty set. We prove this proposition via two following lemmas.

LEMMA 2.2. *There are one interval $J \subset (a,\ b)$, one natural number r and one index $i \in \{1, 2\}$, such that f^r is continuous on J and $f^r(J) = I_i$.*

PROOF. Let $U_0 := (a,\ b)$ and $k_1 \in \mathbb{N}$ be the smallest number such that

$$\xi_1 \in f^{k_1}(U_0).$$

Notice that there is such k_1 since f is an expanding map with $\lambda > \sqrt{2}$. Also here $k_1 \geq 2$, since $\xi_1 \notin U_0 \cup f(U_0)$. Moreover, f is a continuous and non decreasing map on the intervals $U_0, f(U_0), \cdots f^{k_1-1}(U_0)$. So $0, 1 \notin f^{k_1}(U_0)$. Let

$$U_1 := \begin{cases} \big(f^{k_1}(a),\ \xi_1\big),\ 2\Big|\big(f^{k_1}(a),\ \xi_1\big)\Big| > \Big|f^{k_1}(U_0)\Big|, \\ \big(\xi_1,\ f^{k_1}(b)\big),\ 2\Big|\big(f^{k_1}(a),\ \xi_1\big)\Big| \leq \Big|f^{k_1}(U_0)\Big|. \end{cases}$$

By using mean value theorem two times, we have $|U_1| \geq \frac{\left|f^{k_1}(U_0)\right|}{2} > \frac{\lambda^2}{2}|U_0|$. By induction method, we construct finite sequences of the numbers $\{k_i\}_{i=1}^n$ and intervals $\{U_i\}_{i=0}^n$ as follows. For $i \geq 1$ suppose that the number k_i and the interval U_i are characterized. Also suppose that $f^{-1}(\xi_1) \notin U_i$, then there is the smallest natural number $k_{i+1} \geq 2$ such that $\xi_1 \in f^{k_{i+1}}(U_i)$. Let

$$U_{i+1} := \begin{cases} \big(f^{k_{i+1}}(a_*),\ \xi_1\big),\ 2\Big|\big(f^{k_{i+1}}(a_*),\ \xi_1\big)\Big| > \Big|f^{k_{i+1}}(U_i)\Big|, \\ \big(\xi_1,\ f^{k_{i+1}}(b_*)\big),\ 2\Big|\big(f^{k_{i+1}}(a_*),\ \xi_1\big)\Big| \leq \Big|f^{k_{i+1}}(U_i)\Big|, \end{cases}$$

where $a_* := \inf U_i$ and $b_* := \sup U_i$. With this notation, we have $|U_{i+1}| > \frac{\lambda^2}{2}|U_i|$ and this led to $|U_{i+1}| > (\frac{\lambda^2}{2})^{i+1}|U_0|$. Thus there is a natural number n such that $f^{-1}(\xi_1) \in U_n$, since $\lambda > \sqrt{2}$. Consequently, $f(U_n)$ contains at least one of the intervals I_1 and I_2. Take index $i \in \{1, 2\}$, such that $f(U_n)$ contains I_i. There is an interval $\tilde{U} \subset U_n$, such that f is continuous on $\tilde{U}$ and $f(\tilde{U}) = I_i$. On the other hand, the function f^{k_n} is continuous and non decreasing on the interval U_n and so $V_{n-1} := f^{-k_n}(\tilde{U})$ is a non empty interval of U_{n-1}. For given $1 < j \leq n$, let $V_{n-j} := f^{-k_{n-j}}(V_{n-j+1})$ that is a non empty interval of U_{n-j}. Now take $r := \sum_{j=0}^n k_j + 1$ and $J := V_0$ that is a non empty interval of $(a,\ b)$. Above discussion says that f^r is continuous on J since $V_j \subset U_j$, and moreover $f^r(J) = I_i$ which completes the proof of the lemma. □

LEMMA 2.3. *For given $i \in \{1, 2\}$, there is one interval $C \subset I_i$ and one natural number s such that f^s is continuous on C and $f^s(C) = I_{3-i}$.*

PROOF. If $|I_1| \leq |I_2|$ then let $I_* := I_1$ and $I^* := I_2$. Otherwise, let $I_* := I_2$ and $I^* := I_1$.

Claim 1: There is one interval $C_1 \subset I^*$ such that $f(C_1) = I_*$.

By using mean value theorem on the intervals I_1 and I_2 and assumption $\lambda > \sqrt{2}$, one obtains $\sqrt{2} < \max\{\frac{1-\alpha}{\xi_1}, \frac{\beta}{1-\xi_1}\}$. For the intervals I_1 and I_2, two cases are possible to occur. Case 1. $|I_1| \leq |I_2|$. In this case $I_* = I_1$ and $I^* = I_2$. Thus, $I_* = (0,\ \xi_1) \subset (0,\ \beta) = f(I^*)$ since $\xi_1 \leq 1-\xi_1$ and $\sqrt{2} < \frac{\beta}{1-\xi_1}$. Case 2. $|I_1| > |I_2|$. In this case $I_* = I_2$ and $I^* = I_1$. Thus, $I_* = (\xi_1,\ 1) \subset (\alpha,\ 1) = f(I^*)$ since $\xi_1 > 1-\xi_1$ and $\sqrt{2} < \frac{1-\alpha}{\xi_1}$.
This completed the proof of the lemma since $I_* \subset f(I^*)$.

Claim 2: There are one interval $C_2 \subset I_*$ and one natural number t such that f^t is continuous on C_2 and $f^t(C_2) = I^*$.

We prove the claim in the case of $|I_1| \leq |I_2|$ and the other case will be proved in a similar way. In first case we have $I_* = I_1$ and $I^* = I_2$. If $f(I_1)$ contains ξ_1, then it contains I_2. Thus, there is one interval $C_2 \subset I_1$ such that $f(C_2) = I_1$ and we take $t = 1$. Otherwise, $I_1^\circ \cap \{\xi_1, f^{-1}(\xi_1)\} = \emptyset$. By using Lemma 2.2, there is a interval $C_2 \subset I_1$ and a natural number t such that f^t is continuous on C_2 and $f^t(C_2) = I_2$. Note that $f^t(C_2)$ is not the interval I_1, because the sequence $\{U_i\}_{i=0}^n$ with $U_0 = I_1^\circ$, that we construct among the proof of the lemma, satisfying $|U_i| < |U_{i+1}|$ and so $f(U_n)$ contains I_2. Depending on I_i is I_* or I^*, above claim gives the existence of a $C \subset I_i$ and a natural number s such that f^s is continuous on C and $f^s(C) = I_{3-i}$. □

To prove the Proposition 2.1, assume that interval I_i and $(a, b) \subset I_i$ are given. From Lemma 2.2, there are one interval $J \subset (a,\ b)$, one natural number r and one index $j \in \{1, 2\}$, such that f^r is continuous on J and $f^r(J) = I_j$. If j=i, we let $k = r$ and $J_i = J$. Otherwise, from Lemma 2.4, there is one interval $C \subset I_j$ and one natural number s such that f^s is continuous on C and $f^s(C) = I_{3-j} = I_i$. Here, we let $k = r + s$. Since, f^r is continuous on J and $f^r(J) = I_j$, there is one interval $J_i \subset J$, such that $f^r(J_i) = C$. Consequently, $f^k(J_i) = I_i$ that completed the proof of the theorem. □

The following theorem states that lower bound $\sqrt{2}$ for the derivation is enough for occurrence of a full chaos.

THEOREM 2.4. *Suppose that f is a Baker-like map with $\lambda > \sqrt{2}$. Then f is topologically mixing and $Per(f)$ is dense in I.*

PROOF. To the first assertion, let $U \subset [0, 1]$ be an open interval, ξ be the discontinuity point, $I_1 = (0, \xi)$ and $I_2 = (\xi, 1)$. Then there exist $n \in \mathbb{N}$ such that $\xi \in f^n(U)$. We can write $f^n(U) = V_1 \cup V_2$, where $V_1 \subset I_1$ and $V_2 \subset I_2$. By Proposition 2.1, there are open intervals $U_1 \subset V_1$ and $U_2 \subset V_2$ and positive numbers $k_1, k_2 \in \mathbb{N}$ such that $f^{n+k_1}(V_1) = I_1$ and $f^{n+k_2}(V_2) = I_2$. So for $n = \max\{k_1, k_2\}$ we have $f^{n+k}(U) = I_1 \cup I_2$. As long as $\alpha \leq \beta$, the map f is topologically mixing. The second assertion obtain since $J_i \subset I_i$, $f^k(J_i) = I_i$ and f^k is continuous on J_i. This completes the proof of the theorem. □

In the following example, map f has derivative smaller or equal to $\sqrt{2}$ and it loses topological transitivity on $[0, 1]$.

EXAMPLE 2.5. Let $\lambda \in (1,\ \sqrt{2})$ and Baker-like map f as follows:

$$f(x) = \begin{cases} f_1(x) = \lambda x - \frac{\lambda}{2} + 1, & 0 \leq x < \frac{1}{2}, \\ f_2(x) = \lambda x - \frac{\lambda}{2}, & \frac{1}{2} \leq x \leq 1. \end{cases}$$

Since $\alpha = 1 - \frac{\lambda}{2} < \frac{1}{2} < \frac{\lambda}{2} = \beta$, we have $f(\beta) < \frac{1}{2} < f(\alpha)$ and

$$f\big((\alpha,\ \beta)\big) = (1 - \frac{\lambda^2}{2} + \frac{\lambda}{2},\ 1) \cup (0,\ \frac{\lambda^2}{2} - \frac{\lambda}{2}).$$

By applying map f, we obtain

$$f^2\big((\alpha,\ \beta)\big) = (\frac{\lambda}{2} + \frac{\lambda^2}{2} - \frac{\lambda^3}{2},\ \beta) \cup (\alpha,\ \frac{\lambda^3}{2} - \frac{\lambda^2}{2} - \frac{\lambda}{2} + 1).$$

Since $1 < \lambda < \sqrt{2}$, then $f^2\big((\alpha,\ \beta)\big) \subset (\alpha,\ \beta)$. So $f(\beta) < \alpha < \beta < f(\alpha)$, hence f is not transitive in I.

References

1. M. Bernardo, C. Budd, A. R. Champneys and P. Kowalczyk, *Piecewise-Smooth Dynamical Systems: Theory and Applications*, Vol. 163, Springer Verlag, London, 2008.
2. L. Gardini and R. Makrooni, *Necessary and sufficient conditions of full chaos for expanding Baker-like maps and their use in non-expanding Lorenz maps*, Commun. Nonlinear Sci. Numer. Simul. **67** (2019) 272–289.
3. P. Glendinning, *Topological conjugation of Lorenz maps by $\beta-$transformations*, Math. Proc. Camb. Phil. Soc. Cambridge University Press, **107** (2) (1990) pp. 401–413.
4. P. Glendinning and C. Sparrow, *Prime and renormalisable kneading invariants and the dynamics of expanding Lorenz maps*, Phys. Nonlinear Phenom. **62** (1–4) (1993) 22–50.
5. P. Glendinning and T. Hall, *Zeros of the kneading invariant and topological entropy for Lorenz maps*, Nonlinearity **9** (4) (1996) 999–1014.
6. F. Hofbauer, *The maximal measure for linear mod one transformations*, J. London Math. Soc. **23** (1981) 92–112.
7. R. Makrooni, N. Abbasi, M. Pourbarat and L. Gardini, *Robust unbounded chaotic attractors in 1D discontinuous maps*, Chaos Solitons & Fractals **77** (2015) 310–318.
8. W. Parry, *The Lorenz Attractor and a Related Population Model*, in ergodic theory, Lecture Notes in Mathematics, Springer, Berlin, 1979.
9. M. Pourbarat, N. Abbasi and R. Makrooni, *Chaos in smooth piecewise dynamical systems with one discontinuous point*, JAMM, **9** (2) (1398) 93–105 (in persian).
10. D. Rand, *The topological classification of Lorenz attractors*, Math. Proc. Cambridge Philos. Soc. **83** (3) (1978) 451–460.

E-mail: royamakrooni@yahoo.com
E-mail: m-pourbarat@sbu.ac.ir
E-mail: n_abbasi@sbu.ac.ir

Symbolic Dynamics of All Degrees of Freedom Around Symmetric Homoclinics

Mahdiyeh Molaei Derakhtenjani*
Department of Mathematical Sciences, University of Birjand, Birjand, Iran
and Omid Rabiei Motlagh
Department of Mathematical Sciences, University of Birjand, Birjand, Iran

Abstract. In this paper, we consider a spatial symmetric system with double spiral homoclinic orbits. We construct the global Poincare map in the outer region of the homoclinic orbits and show that the system has symbolic dynamics of all degrees of freedom.
Keywords: Poincare map, Double symmetric homoclinic orbit, Symbolic dynamic of N_0 degrees of freedom.
AMS Mathematical Subject Classification [2010]: 37C29, 37B10.

1. Introduction

Symbolic dynamics and homoclinic orbits have been studied for the last decades (See for example [1, 3] and references therein). Today it is probably well known that such systems have symbolic dynamics with finite degrees of freedom in the inner region of the homoclinic orbits [2]. Here, we implement the global Poincare map to show that a spatial symmetric ODE with spiral double homoclinic has symbolic dynamics of all degrees of freedom in the outer region of the homoclinic orbits. The ODE is

$$\dot{x} = \rho x - \omega y + F(x,y,z), \dot{y} = \omega x + \rho y + G(x,y,z), \dot{z} = \lambda z + H(x,y,z),$$

where $\lambda > -\rho > 0$, $\omega > 0$ and F, G and H are C^2 and nonlinear at the origin. We also assume that the system posses a double symmetric (w.r.t. the origin) homoclinic orbits $\Gamma_{1,2}$, see Figure 1(A).

In [4, Chapt. 27], a Poincare map $\tilde{P} = \tilde{p}_1 \circ \tilde{p}_0$ was presented by combining two inner Poincare maps $\tilde{p}_0 : \Sigma_0 \to \Sigma_1$ and $\tilde{p}_1 : \Sigma_1 \to \Sigma_0$ based on the cross sections Σ_0, Σ_1 as below (See Figure 1(B)),

$$\Sigma_0 = \left\{ (x,0,z) \ : \ \begin{array}{l} \epsilon e^{2\pi\rho/\omega} < x < \epsilon, \\ 0 < z < \epsilon \end{array} \right\}, \quad \Sigma_1 = \left\{ (x,y,\epsilon) \ : \ \begin{array}{l} 0 < |x| < \epsilon, \\ 0 < |y| < \epsilon \end{array} \right\}.$$

Theorem 1.1. [4, Ch. 27] *The above Poincare map $\tilde{P}$ contains a chaotic behavior caused by a symbolic dynamics of two degrees of freedom.*

In what follows, we concern on constructing the outer Poincare map $P = p_3 \circ p_2 \circ p_1 \circ p_0$ by defining four cross sections Π_i, $i = 0, 1, 2, 3$, and local Poincare

*Presenter

maps $p_i : \Pi i \to \Pi_{i+1}$, $(\Pi_4 = \Pi_0)$ (See Figure 1(C))

$$\Pi_0 = \left\{(x,0,z) : \begin{array}{l} \epsilon e^{\frac{2\pi\rho}{\omega}} < x < \epsilon, \\ -\epsilon < z < 0 \end{array}\right\}, \ \Pi_1 = \{(x,y,-\epsilon) : |x|,|y| \le \epsilon\},$$

$$\Pi_2 = \left\{(x,0,z) : \begin{array}{l} -\epsilon < x < -\epsilon e^{\frac{2\pi\rho}{\omega}}, \\ 0 < z < \epsilon \end{array}\right\}, \ \Pi_3 = \{(x,y,\epsilon) : |x|,|y| \le \epsilon\}.$$

The main result of this paper is summarized as follows and the scheme of its proof will be arranged in the next sections.

THEOREM 1.2. *For any $N_0 \in \mathbb{N}$, The above Poincare map P contains a symbolic dynamic of N_0 degrees of freedom which is caused by a N_0-fold Smale Horseshoe.*

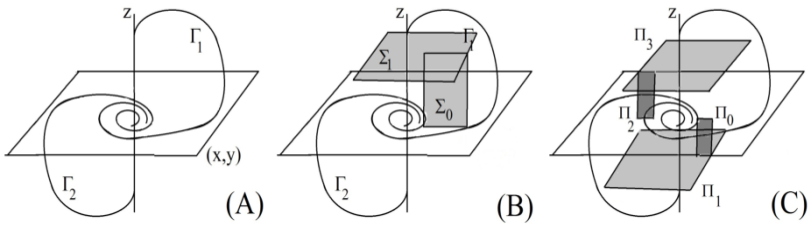


FIGURE 1. (A): The double symmetric homoclinic orbit. (B),(C): The corresponding cross sections.

2. Poincare Map P, Scheme of Construction

Since the ODE is symmetric w.r.t. the origin, so the Poincare maps p_0 and p_2 are symmetric too. They are also constructed by continuing the flow generated by the corresponding linearized system at the origin. It is easy to see that the time flight for points $(x,0,z) \in \Pi_i,\ i = 0,2$ to reach to $\Pi_j,\ j = 1,3$ is $t = \frac{1}{\lambda}\log|\epsilon/z|$. So that $p_i(x,0,z) = M_i(x,0,z)^T$, where

$$M_i = \begin{bmatrix} |\frac{\epsilon}{z}|^{\frac{\rho}{\lambda}}\cos(\frac{\omega}{\lambda}\log|\frac{\epsilon}{z}|) & -|\frac{\epsilon}{z}|^{\frac{\rho}{\lambda}}\sin(\frac{\omega}{\lambda}\log|\frac{\epsilon}{z}|) & 0 \\ |\frac{\epsilon}{z}|^{\frac{\rho}{\lambda}}\sin(\frac{\omega}{\lambda}\log|\frac{\epsilon}{z}|) & |\frac{\epsilon}{z}|^{\frac{\rho}{\lambda}}\cos(\frac{\omega}{\lambda}\log|\frac{\epsilon}{z}|) & 0 \\ 0 & 0 & |\frac{\epsilon}{z}| \end{bmatrix}, N = \begin{bmatrix} a & b & 0 \\ 0 & 0 & 0 \\ c & d & 0 \end{bmatrix}.$$

The Poincare maps $p_j,\ j = 1,3$ are symmetric and they are approximated by an affine map. Indeed Taylor expansion of p_1 about $(0,0,-\epsilon)$ yields to $p_1(x,y,-\epsilon) \approx (-\epsilon,0,0)^T + N.(x,y,0)^T$, where a,b,c and d are constants and, N is given above. Then, because of the symmetry, $p_3(x,y,\epsilon) \approx (\epsilon,0,0)^T + N.(x,y,0)^T$. Thus, finally we have the Poincare map P as below

$$P(x,0,z) = (\epsilon,0,0)^T + NM_2(-\epsilon,0,0)^T + NM_2NM_1(x,0,z)^T,$$

where M_2 is computed for

$$z := z_2 = \left(\frac{-\epsilon}{z}\right)^{\frac{\rho}{\lambda}}\left(c\cos\left(\frac{\omega}{\lambda}\log\frac{-\epsilon}{z}\right) + d\sin\left(\frac{\omega}{\lambda}\log\frac{-\epsilon}{z}\right)\right).$$

It is easy to see that, the Poincare map P is defined just if $z_2 > 0$, or equivalently, $c\tan^2(\frac{\omega}{2\lambda}\log\frac{-\epsilon}{z}) - 2d\tan(\frac{\omega}{2\lambda}\log\frac{-\epsilon}{z}) - c < 0$. If $c > 0$ (the case $c < 0$ is similar); then, for $\Delta = d/c$, we should have $\Delta - \sqrt{\Delta^2+1} < \tan(\frac{\omega}{2\lambda}\log\frac{-\epsilon}{z}) < \Delta+\sqrt{\Delta^2+1}$. Let $\phi_{1,2} = \tan^{-1}(\Delta\pm\sqrt{\Delta^2+1})$, then we find rectangles $\mathtt{R}^1_k$, $\mathtt{R}^2_k \subset \Pi_0$, $(k = 0, 1, 2, \ldots)$, given below such that for $(x, 0, z) \in \mathtt{R}^i_k$, $(i = 1, 2)$ we have $z_2 > 0$.

$$\mathtt{R}^1_k = \left\{(x,0,z)\in\Pi_0 \ : \ \begin{array}{c} \epsilon e^{\frac{2\pi\rho}{\omega}} < x < \epsilon, \\ -\epsilon e^{-\frac{\lambda}{\omega}(2k\pi)} < z < -\epsilon e^{-\frac{\lambda}{\omega}(2k\pi+2\phi_1)} \end{array}\right\},$$

$$\mathtt{R}^2_k = \left\{(x,0,z)\in\Pi_0 \ : \ \begin{array}{c} \epsilon e^{\frac{2\pi\rho}{\omega}} < x < \epsilon, \\ -\epsilon e^{-\frac{\lambda}{\omega}(2k\pi+2\phi_2)} < z < -\epsilon e^{-\frac{\lambda}{\omega}(2k\pi)} \end{array}\right\}.$$

LEMMA 2.1. *Let $c > 0$. The domain of the Poincare map P is*

$$Dom(P) = \bigcup_{k\geq 0, i=1,2} R^i_k.$$

Let $k = 0, 1, 2, \ldots$ be fixed, then $p_1 \circ p_0(R^i_k)$, $(i = 1, 2)$, are two distinct vertical rectangles in Π_2, see Figure 2(A).

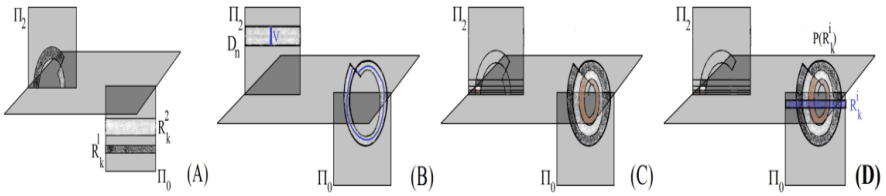


FIGURE 2. (A): The domain of the global poincare map and its image on Π_2. (B),(C),(D): The process scheme of the symbolic dynamic of N_0 degrees of freedom.

3. Scheme of Proof of Theorem 1.2

Consider the cross section Π_2 and for $n = 0, 1, 2, \ldots$ define the rectangles

$$D_n = \left\{(x,0,z) \ : \ \begin{array}{c} -\epsilon \leq x \leq -\epsilon e^{\frac{2\pi\rho}{\omega}}, \\ \epsilon e^{\frac{-2\pi(n+1)\lambda}{\omega}} \leq z < \epsilon e^{\frac{-2\pi n\lambda}{\omega}} \end{array}\right\} \subseteq \Pi_2.$$

It is easy to see that $\Pi_2 = \cup_{n\geq 0} D_n$.

LEMMA 3.1. *Let $a < 0$ and $k \geq 0$ be fixed. Then $p_3 \circ p_2(D_n)$ is a loop strip witch intersects Π_0 as Figure 2(B). Furthermore, any vertical strip $V \subset D_n$ completes a loop in $p_3 \circ p_2(D_n)$.*

Thus, if $k \geq 0$ is fixed, then from Lemma 2.1, there exists $n_0 \geq 0$ such that if $n \geq n_0$ then $p_1 \circ p_0(\mathtt{R}^i_k)$, $(i = 1, 2)$, intersects D_n in two distinct vertical strips. Therefore, from Lemma 3.1, $P(\mathtt{R}^i_k) = (p_3 \circ p_2)\big(p_1 \circ p_0(\mathtt{R}^i_k)\big)$, $(i = 1, 2)$, makes an infinitely many rounds loop which intersects Π_0 in infinitely many folds, see Figure 2(C). It must be noted that a similar argument as Lemmas 2.1 and 3.1 holds if

$c < 0$ and $a > 0$. Thus we may assume that the results above hold if $ac < 0$. The following lemma completes the proof of the Theorem 1.2.

LEMMA 3.2. *Let $1 \leq N_0 \in \mathbb{N}$ be fixed and $ac < 0$. There exists $k \geq 0$ such that $P(R_k^i)$, $(i = 1, 2)$, intersects R_k^i in $2N_0$ distinct vertical strips. These vertical strips make a N_0-fold Smale horseshoe which imposes a symbolic dynamic of N_0 degrees of freedom, see Figure 2*(D).

References

1. B. P. Kitchens, *Symbolic Dynamics*, Springer-Verlag Berlin Heidelberg, Berlin, 1998.
2. I. S. Labouriau and A. A. P. Rodrigues, *Global bifurcations close to symmetry*, J. Math. Anal. Appl. **444** (1) (2016) 648–671.
3. D. Lind and B. Marcus, *An Introduction to Symbolic Dynamics and Coding*, Cambridge University Press, Cambridge, 1995.
4. S. Wiggins, *Introduction to Applied Nonlinear Dynamical Systems and Chaos*, 2nd ed., Springer-Verlag, New York, 2003.

E-mail: m.molaei@birjand.ac.ir
E-mail: orabieimotlagh@birjand.ac.ir

Discontinuous Sturm-Liouville Problem and Prüfer Substitutions

Seyfollah Mosazadeh*
Department of Mathematics, Faculty of Mathematical Sciences, University of Kashan, Kashan, Iran
and Hikmet Koyunbakan
Department of Mathematics, Faculty of Science, Firat University, Elazig, Turkey

Abstract. In this work, we consider a discontinuous Sturm-Liouville problem with separated boundary conditions and discontinuity conditions in $\frac{b}{2} \in (0, b)$. First, we present new Prüfer substitutions for discontinuous case. Then, the asymptotic form of the eigenvalues and the nodal points are obtained. Finally, by using the nodal lengths, we obtain the solution of the inverse nodal problem.

Keywords: Discontinuous Sturm-Liouville problem, Prüfer substitutions, Nodal points, Inverse problem.
AMS Mathematical Subject Classification [2010]: 34A55, 34B24.

1. Introduction

We consider the Sturm-Liouville equation of the form

$$-y'' + q(t)y = \lambda^2 y, \qquad 0 < t < b, \tag{1}$$

with Dirichlet boundary conditions

$$y(0) = 0 = y(b), \tag{2}$$

and discontinuity conditions

$$\begin{cases} y(\frac{b}{2} + 0) = \alpha y(\frac{b}{2} - 0), \\ y'(\frac{b}{2} + 0) = \alpha^{-1} y'(\frac{b}{2} - 0), \end{cases} \tag{3}$$

where λ is the spectral parameter, $q(t)$ and α are real, $q(t) \in C^1([0, b])$, $\alpha > 0$ and $\alpha \neq 1$.

Equation (1) often appears in mathematics, mathematical physics, mathematical chemistry and other branches of natural sciences. Also, boundary value problems with discontinuities inside the intervals mostly appear in physics, electronics and mechanics [2, 10].

If the function $q(t)$ in (1) is known and we are going to find the eigenvalues and the eigenfunctions of the problem (1)-(3), then the problem is called *Direct*. If the function $q(t)$ is unknown and the question arises as to what extent $q(t)$ can be reconstructed from given spectral data, then the problem is called *Inverse*.

*Presenter

Inverse spectral problems, reconstructing the operators from their spectral characters, are divided into two categories: inverse eigenvalue problems and inverse nodal problems. Inverse eigenvalue problems were studied by many researchers. McLaughlin seems to have been the first researcher to consider the inverse nodal problem in 1986 [5]. In this kind of problem, the eigenvalues togerher with the roots of the eigenfunctions (nodal points) are used. After McLaughlin, other researches studied the inverse problem with diverse conditions by using the nodal points (For example, see [3, 6, 7]).

In recent years, some authors solved the inverse nodal problem with discontinuity conditions without using Prüfer substitutions (For example, see [8, 9]). Moreover, some authors used the Prüfer substitutions and solved the inverse nodal problem without discontinuity conditions [1, 4].

In this paper, we give new Prüfer substitutions for discontinuous Sturm-Liouville problem (1)-(3), differently from the classical Prüfer substitutions. Then, by using the modified Prüfer substitutions, we present the asymptotic estimates of the eigenvalues and the nodal points. Then, we present a constructive procedure for the solution of the inverse problem in two cases, $t < \frac{b}{2}$ and $t > \frac{b}{2}$.

2. Prüfer Substitutions and Preliminary Results

Let $y(t,\lambda)$ be the solution of the Eq. (1) under the initial conditions

$$y(0,\lambda) = 0, \qquad y'(0,\lambda) = 1.$$

Using the well-known method (See, e.g., [2]), one can obtain that

$$y(t,\lambda) = \frac{\sin \rho t}{\rho} + O\left(\frac{1}{\rho}\exp(|\tau|t)\right),$$

as $|\rho| \to \infty$, where $\rho = \sqrt{\lambda}$ and $\tau = Im\rho$. Therefore, we get the following estimate:

$$y(t,\lambda) = \frac{\sin \rho t}{\rho} - \frac{\cos \rho t}{2\rho^2}\int_0^t q(u)du + \frac{1}{2\rho^2}\int_0^t q(u)\cos\rho(t-2u)du + O\left(\frac{1}{\rho^3}\exp(|\tau|t)\right).$$

Let us introduce the aptitude function $R(t,\lambda)$ and the phase function $\Theta(t,\lambda)$, which are defined according to the given solution $y(t,\lambda)$ by

$$\begin{cases} y(t,\lambda) = R(t)\sin\Theta(\lambda t), \\ \\ y'(t,\lambda) = \lambda R(t)\cos\Theta(\lambda t), \end{cases}$$

for the Eq. (1). For non-trivial solution of (1), we assume that R is positive.

We are ready to define new Prüfer substitutions for the discontinuous problem (1)-(3) as follows:

$$y(t,\lambda) = R(t)\sin(\lambda\Theta(t)), \qquad \text{if and only if} \qquad t < \frac{b}{2},$$

$$y(t,\lambda) = R(b-t)\sin(\lambda\Theta(b-t)), \quad \text{if and only if} \qquad \frac{b}{2} < t < b.$$

In the following theorem, we present the asymptotic form of the eigenvalues.

THEOREM 2.1. *The eigenvalues of the problem (1)-(3) have the form*

$$\lambda_n = \frac{2n\pi}{b} + \frac{1}{4n\pi}\left\{\beta + (-1)^{n-1}\gamma + \frac{8\delta_n^2}{b}\right\} + O(\frac{1}{n^2}),$$

where

$$\beta = \int_0^b q(t)dt,$$

$$\gamma = \int_0^b q(t)dt - 2\int_0^{\frac{b}{2}} q(t)dt,$$

$$\delta_n = \begin{cases} \arcsin(\frac{1}{\sqrt{1+\alpha^2}}), & \text{if } n \text{ is even,} \\ -\arcsin(\frac{|\alpha|}{\sqrt{1+\alpha^2}}), & \text{if } n \text{ is odd.} \end{cases}$$

3. Main Results

In this section, we obtain the nodal points and present a constructive procedure for the solution of the inverse problem by using the nodal lengths.

Let λ_n be an eigenvalue of the problem (1)-(3), and $\{t_n^j\}_{j=1}^{n-1}$ be the zeros of the corresponding eigenfunction $y_n(t) = y(t, \lambda_n)$. In the following theorem, we give the nodal points of (1)-(3).

THEOREM 3.1. *The nodal points of the problem* (1)-(3) *have the following asymptotic estimates as* $n \to \infty$*:*

$$t_n^j = \frac{jb}{n} + \frac{b^2}{8n^2\pi^2}\int_0^{b_n^j} q(u)du + O(\frac{1}{n^3}), \quad t_n^j \in (0, \frac{b}{2}), \tag{4}$$

$$\begin{aligned} t_n^j = & \frac{jb}{n} - \frac{jb\delta_n}{n^2\pi} + \frac{b^2}{8n^2\pi^2}\int_0^{t_n^j} q(u)du \\ & - \frac{b^2}{8n^2\pi^2}\int_0^{\frac{b}{2}} q(u)du + O(\frac{1}{n^3}), \quad t_n^j \in (\frac{b}{2}, b). \end{aligned} \tag{5}$$

Denote the nodal lengths of $y(t, \lambda_n)$ as follow:

$$\ell_n^j = t_n^{j+1} - t_n^j, \quad 1 \le j \le n-1.$$

Applying (4)-(5), we have the following corollary.

COROLLARY 3.2. *The nodal lengths of the problem* (1)-(3) *have the following asymptotic estimates:*

$$\ell_n^j = \frac{2\pi}{\lambda_n} + \frac{1}{2\lambda_n^2}\int_{t_n^j}^{t_n^{j+1}} q(u)du + o(\frac{1}{\lambda_n^2}), \quad t_n^j < \frac{b}{2},$$

$$\ell_n^j = \frac{2\pi}{\lambda_n} + \frac{\delta_n}{\lambda_n} + \frac{1}{2\lambda_n^2}\int_{t_n^j}^{t_n^{j+1}} q(u)du + o(\frac{1}{\lambda_n^2}), \quad t_n^j > \frac{b}{2}.$$

Denote the nodal set $T_n = \{t_n^j\}_{j=1}^{n-1}$. Now, we provide a constructive procedure to obtain the solution of the inverse nodal problem.

THEOREM 3.3. *Fix* $\nu = 0 \vee 1$. *Given the nodal set* T_n. *For* $t \in (0, \frac{b}{2})$, *the function* $q(t)$ *of* (1) *satisfies the following formula:*

$$q(t) = \lim_{n\to\infty} \frac{8n^2\pi^2}{b^2}\Big\{\frac{n\ell_n^j}{b} - 1\Big\} + \frac{1}{b}(\beta + (-1)^{\nu-1}\gamma),$$

for $j = j_n(t) = \max\{j : \ t_n^j < t\}$.

THEOREM 3.4. *Fix $\nu = 0 \vee 1$ and $t \in (\frac{b}{2}, b)$. Let T be a subset of nodal points which is dense on $(0, b)$. Then*

$$q(t) = \lim_{n\to\infty} \frac{16n^2\pi^3}{b^2(2\pi+\delta_n)}\Big\{\frac{n\ell_n^j}{b} - \frac{\delta_n}{2\pi} - 1\Big\} + \frac{2\pi(\beta + (-1)^{\nu-1}\gamma)}{b(2\pi+\delta_\nu)},$$

for $j = j_n(t) = \max\{j : \ t_n^j < t\}$.

Acknowledgement

This research is partially supported by the University of Kashan under grant number 985969/5.

References

1. Y. -H. Cheng, C. -K. Law, W. -C. Lian and W. -C. Wang, *An inverse nodal problem and Ambarzumyan problem for the periodic p-Laplacian operator with integrable potentials*, Taiwanese. J. Math. **19** (2015) 1305–1316.
2. G. Freiling and V. Yurko, *Inverse Sturm-Liouville Problems and their Applications*, Nova Science Publishers Inc., New York, 2001.
3. H. Koyunbakan, *The inverse nodal problem for a differential operator with an eigenvalue in the boundary condition*, Appl. Math. Lett. **21** (2008) 1301–1305.
4. H. Koyunbakan, T. Gulsen and E. Yilmaz, *Inverse nodal problem for a p-Laplacian Sturm-Liouville equation with polynomially boundary condition*, Elect. J. Diff. Equ. **2018** (2018) 1–9.
5. J. R. McLaughlin, *Analytical methods for recovering coefficients in differential equations from spectral data*, SIAM Rev. **28** (1986) 53–72.
6. S. Mosazadeh, *The uniqueness theorem for inverse nodal problems with a chemical potential*, Iranian J. Math. Chem. **8** (2017) 403–411.
7. A. Neamaty, N. Yousefi and A. Dabbaghian, *The numerical values of the nodal points for the Sturm-Liouville equation with one turning point*, Comput. Methods Differ. Equ. **7** (2019) 124–137.
8. A. S. Ozkan and B. Keskin, *Inverse nodal problems for SturmLiouville equation with eigenparameter-dependent boundary and jump conditions*, Inv. Prob. Sci. Eng. **23** (2015) 1306–1312.
9. C. T. Shieh and V. A. Yurko, *Inverse nodal and inverse spectral problems for discontinuous boundary value problems*, J. Math. Anal. Appl. **347** (2008) 266–272.
10. A. N. Tikhonov and A. A. Samarskii, *Equations of Mathematical Physics*, Oxford Pergamon Press, New York, 1963.

E-mail: s.mosazadeh@kashanu.ac.ir
E-mail: hkoyunbakan@gmail.com

Contributed Posters

Interdisciplinary Mathematics

Coding Theory on the Generalized Balancing Sequence

Elahe Mehraban*
Department of Pure Mathematics, Faculty of Mathematical Sciences, University of Guilan, Rasht, Iran
and Mansour Hashemi
Department of Pure Mathematics, Faculty of Mathematical Sciences, University of Guilan, Rasht, Iran

Abstract. In this paper, we introduce the generalized balancing sequence and its matrix. Then, we get the nth power of its matrix denoted by Q_m^n. At last, by using Q_m^n, we give the coding and the decoding method.
Keywords: Generalized k-balancing number, Coding and decoding method.
AMS Mathematical Subject Classification [2010]: 11C20, 11B39, 68P30.

1. Introduction

A positive integer n is called a balancing number if $1+2+3+\cdots+(n-1) = (n+1)+(n+2)+(n+3)+\cdots+(n+r)$ for some nonnegative integer r and calling r as the balancer corresponding to n. Balancing numbers are studied and generalized in many ways (See [2, 3]). In [4], introduced k-balancing numbers, denoted by $\{B_n^k\}_{n=0}^{\infty}$, as follows:

For any positive number k, balancing numbers $\{B_n^k\}_{n=0}^{\infty}$ is defined by

$$B_{n+1}^k = 6kB_n^k - B_{n-1}^k, \quad n \geqslant 1, \tag{1}$$

with the initial conditions $B_0^k = 0, B_1^k = 1$. For example, let $k = 1$. We have $\{B_n^1\}_{n=0}^{\infty} = \{0, 1, 6, 37, \dots\}$.

One of the applications of sequences and their matrices is to use in coding theory. In [5], Apostolic introduced the Fibonacci code which is used in the source coding as well as in cryptography. Many authors study the generalized Fibonacci code (See [1]). Here, we define the generalized balancing sequence and its matrice. Then, we give the coding/decoding method.

2. Main Results

In this section, let $k = 1$. We generalize relation (1) as follows:

Definition 2.1. For $m \geqslant 3$, the generalized balancing sequence $\{B_{m,t}\}_{t=0}^{\infty}$ is defined by:

$$B_{m,t} = 6B_{m,t-1} - B_{m,t-2} - \cdots - B_{m,t-m}, \quad t \geqslant 1,$$

with the initial conditions $B_{m,0} = B_{m,1} = \cdots = B_{m,t-2} = 0$ and $B_{m,t-1} = 1$.

*Presenter

For example if $m = 3$, we have $\{B_{3,t}\}_{t=0}^{\infty} = \{0, 0, 1, 6, 35, \ldots\}$.

DEFINITION 2.2. For $m \geqslant 3$, let Q_m be an $m \times m$ generalized balancing matrix. We have

$$Q_m = \begin{bmatrix} 6 & -1 & -1 & \cdots & -1 & -1 \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix},$$

THEOREM 2.3. *For $n \geqslant 1$ and $m = 3$, we have*

$$Q_3^n = \begin{bmatrix} B_{3,n+2} & -B_{3,n+1} + B_{3,n} & -B_{3,n+1} \\ B_{3,n+1} & -B_{3,n} + B_{3,n-1} & -B_{3,n} \\ B_{3,n} & -B_{3,n-1} + B_{3,n-2} & -B_{3,n-1} \end{bmatrix},$$

and $B_{3,t}$ is the element of the generalized balancing sequence.

PROOF. By using an induction method on n, setting $m = 3$ and $n = 1$ and by Definition 2.2 we have

$$Q_3 = \begin{bmatrix} 6 & -1 & -1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

Suppose that the statement holds for $n = k$. Therefore, for $n = k + 1$ we have

$$\begin{aligned} (Q_3)^{k+1} &= \begin{bmatrix} 6 & -1 & -1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} B_{3,3+k-1} & -(B_{3,3+k-2} + B_{3,3+k-3}) & B_{3,3+k-2} \\ B_{3,3+k-2} & -(B_{3,3+k-3} + B_{3,3+k-4}) & B_{3,3+k-3} \\ B_{3,3+k-3} & -(B_{3,3+k-4} + B_{3,3+k-5}) & B_{3,3+k-4} \end{bmatrix} \\ &= \begin{bmatrix} B_{3,3+k} & -(B_{3,3+k-1} + B_{3,3+k-2}) & B_{3,3+k-1} \\ B_{3,3+k-1} & -(B_{3,3+k-2} + B_{3,3+k-3}) & B_{3,3+k-2} \\ B_{3,3+k-2} & -(B_{3,3+k-3} + B_{3,3+k-4}) & B_{3,3+k-3} \end{bmatrix}. \end{aligned}$$

□

Now, we are ready to generalize the idea of the nth power of the matrix Q_3 to the nth power of the matrix $Q_m (m > 3)$.

THEOREM 2.4. *For $n \geqslant 1$ and $m \geqslant 4$, we have*

$$Q_m^n = \begin{bmatrix} B_{m,n+m-1} & a & e & \cdots & -(B_{m,n+m-2} + B_{m,n+m-3}) & -B_{m,n+m-2} \\ B_{m,n+m-2} & b & f & \cdots & -(B_{m,n+m-3} + B_{m,n+m-4}) & -B_{n+m-3} \\ B_{m,n+m-3} & c & g & \cdots & -(B_{m,n+m-4} + B_{m,n+m-5}) & -B_{m,n+m-4} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ B_{m,n} & d & h & \cdots & -(B_{m,n-1} + B_{m,n-2}) & -B_{m,n-1} \end{bmatrix},$$

where

$$\begin{aligned}
a &= -(B_{m,n+m-2} + B_{m,n+m-3} + \cdots + B_{m,n}),\\
b &= -(B_{m,n+m-3} + B_{m,n+m-4} + \cdots + B_{m,n-1}),\\
c &= -(B_{m,n+m-4} + B_{m,n+m-5} + \cdots + B_{m,n-2}),\\
d &= -(B_{m,n-1} + B_{m,n-2} + \cdots + B_{m,n-m+1}),\\
e &= -(B_{m,n+m-2} + B_{m,n+m-2} + \cdots + B_{m,n+1}),\\
f &= -(B_{m,n+m-3} + B_{m,n+m-4} + \cdots + B_{m,n}),\\
g &= -(B_{m,n+m-3} + B_{m,n+m-4} + \cdots + B_{m,n-1}),\\
h &= -(B_{m,n-1} + B_{m,n-2} + \cdots + B_{m,n-m}),
\end{aligned}$$

and $B_{m,t}$ is the element of the generalized balancing sequence.

Proof. Similarly Theorem 2.3, we can prove. □

Example 2.5. We have

$$Q_4^3 = \begin{bmatrix} B_{4,6} & -(B_{4,5}+B_{4,4}+B_{4,3}) & (B_{4,5}+B_{4,4}) & -B_{4,5} \\ B_{4,5} & -(B_{4,4}+B_{4,3}+B_{4,2}) & (B_{4,4}+B_{4,3}) & -B_{4,4} \\ B_{4,4} & -(B_{4,3}+B_{4,2}+B_{4,1}) & (B_{4,3}+B_{4,2}) & -B_{4,3} \\ B_{4,3} & -(B_{4,2}+B_{4,1}+B_{4,0}) & (B_{4,2}+B_{4,1}) & -B_{4,2} \end{bmatrix} = \begin{bmatrix} 203 & -42 & -41 & -35 \\ 35 & -7 & -7 & -6 \\ 6 & -1 & -1 & -1 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

According to Theorem 2.4, we can obtain the following results.

Corollary 2.6. *For $n, t \geqslant 1$, we have*

$$(Q_m^n) \times (Q_m^t) = (Q_m^t) \times (Q_m^n) = (Q_m^{n+t}).$$

Corollary 2.7. *The determinant of Q_m^n is equal to $(-1)^{nm}$.*

Here, we find the coding and decoding on the generalized balancing matrix Q_m^n and get its error detection and correction.

For $m \geqslant 3$ and an initial message $M_{m\times m}$, we will name a transformation $E = M \times Q_m^n$ as the generalized balancing matrix coding and a transformation $M = E \times {Q_m^n}^{-1}$ as the generalized balancing matrix decoding. Also, the matrix E is a code matrix. Now, we explain the proposed the method by an example.

Example 2.8. Suppose $m = 3$, $n = 4$ and

$$M = \begin{bmatrix} 1 & 2 & 4 \\ 2 & 4 & 3 \\ 1 & 0 & 2 \end{bmatrix}.$$

We have,

$$Q_3^4 = \begin{bmatrix} B_{3,6} & -(B_{3,5}+B_{3,4}) & B_{3,5} \\ B_{3,5} & -(B_{3,4}+B_{3,3}) & B_{3,4} \\ B_{3,4} & -(B_{3,3}+B_{3,2}) & B_{3,3} \end{bmatrix} = \begin{bmatrix} 1177 & -238 & -203 \\ 203 & -41 & -35 \\ 35 & -7 & -6 \end{bmatrix}.$$

By the above notations, we have

$$E = M \times Q_3^4 = \begin{bmatrix} 1 & 2 & 4 \\ 2 & 4 & 3 \\ 1 & 0 & 2 \end{bmatrix} \times \begin{bmatrix} 1177 & -238 & -203 \\ 203 & -41 & -35 \\ 35 & -7 & -6 \end{bmatrix} = \begin{bmatrix} 1723 & -348 & -297 \\ 3271 & -661 & -564 \\ 1274 & -252 & -215 \end{bmatrix}.$$

Also, we obtain

$$M = E \times Q_3^{4^{-1}}$$
$$= \begin{bmatrix} 1723 & -348 & -297 \\ 3271 & -661 & -564 \\ 1274 & -252 & -215 \end{bmatrix} \times \begin{bmatrix} 1 & -7 & 7 \\ -7 & 43 & -14 \\ 14 & -91 & 57 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 2 & 4 \\ 2 & 4 & 3 \\ 1 & 0 & 2 \end{bmatrix}.$$

Now, we get the determinant of the code matrix E. The code matrix E is defined by $E = M \times Q_m^n$. According to Corollary 2.7, we have

$$detE = det(M \times Q_m^n) = detM \times detQ_m^n = detM \times (-1)^{mn}.$$

Therefore, it is clear that the determinant of the initial message M is connected with the determinant of the code message E. So, we obtain the determinant of the matrix M. $det(M)$ treats as a controller of entries of the code matrix E received from the communication channel. After receiving the code matrix E and computing the determinant of M, we will compute the determinant of E. Then, we will compare them together. If $detE = \pm detM$, this means the matrix E has passed from the communication channel without error. Otherwise, according to the matrix E of the order, we have $m \times m$ "single", "double", ..., "m^2-fold" errors. Thus,

$$1C_{m^2} + 2C_{m^2} + \cdots + m^2 C_{m^2} = 2^{m^2} - 1.$$

For example, let $m = 3$. According to the matrix E of the order 3×3, we have "single", "double",..., "nine-fold" errors. The first assumption is that there exists only one error in the matrix E received from the communication channel. It's clear that there are nine different cases for it as follows:

$$(1)\begin{bmatrix} a & e_2 & e_3 \\ e_4 & e_5 & e_6 \\ e_7 & e_8 & e_9 \end{bmatrix}, \quad (2)\begin{bmatrix} e_1 & b & e_3 \\ e_4 & e_5 & e_6 \\ e_7 & e_8 & e_9 \end{bmatrix}, \quad (3)\begin{bmatrix} e_1 & e_2 & c \\ e_4 & e_5 & e_6 \\ e_7 & e_8 & e_9 \end{bmatrix},$$
$$(4)\begin{bmatrix} e_1 & e_2 & e_3 \\ d & e_5 & e_6 \\ e_7 & e_8 & e_9 \end{bmatrix}, \quad (5)\begin{bmatrix} e_1 & e_2 & e_3 \\ e_4 & e & e_6 \\ e_7 & e_8 & e_9 \end{bmatrix}, \quad (6)\begin{bmatrix} e_1 & e_2 & e_3 \\ e_4 & e_5 & f \\ e_7 & e_8 & e_9 \end{bmatrix},$$
$$(7)\begin{bmatrix} e_1 & e_2 & e_3 \\ e_4 & e_5 & e_6 \\ g & e_8 & e_9 \end{bmatrix}, \quad (8)\begin{bmatrix} e_1 & e_2 & e_3 \\ e_4 & e_5 & e_6 \\ e_7 & h & e_9 \end{bmatrix}, \quad (9)\begin{bmatrix} e_1 & e_2 & e_1 \\ e_4 & e_5 & e_6 \\ e_7 & e_8 & i \end{bmatrix},$$

where $a, b, \ldots, i$ are possible "destroyed" entries. From $det(E) = (-1)^{nm} \times det(M)$, we have

$$(1)a(e_5e_9 - e_6e_8) - e_2(e_4e_9 - e_6e_7) + e_3(e_4e_8 - e_5e_7) = (-1)^{nm} \times detM,$$
$$(2)e_1(e_5e_9 - e_6e_8) - b(e_4e_9 - e_6e_7) + e_3(e_4e_8 - e_5e_7) = (-1)^{nm} \times detM,$$
$$\vdots$$
$$(9)e_1(e_5i - e_6e_8) - e_2(e_4i - e_6e_7) + e_3(e_4e_8 - e_5e_7) = (-1)^{nm} \times detM.$$

In a similar way, we will obtain a double error for the matrix E. For example, we consider a bivariate case for matrix E as follows:

$$\begin{bmatrix} a & b & e_3 \\ e_4 & e_5 & e_6 \\ e_7 & e_8 & e_9 \end{bmatrix},$$

which possible cases are $\binom{9}{2} = 36$. Similarly, we obtain "triple", "four-fold", ..., "nine-fold" errors, which the total number of cases is

$$\binom{9}{1} + \binom{9}{2} + \cdots + \binom{9}{9} = 2^9 - 1.$$

Therefore, there are $2^9 - 1 = 511$ errors.

The generalized balancing matrix coding/decoding is calculated very quickly by computer. Also, the correcting ability and detection ability of this coding method is very high in comparison with a classical algebraic coding-decoding method.

References

1. M. Esmaeili, M. Esmaeili and T. A. Gulliver, *A new class of Fibonacci sequence based error correcting codes*, Cryptogr. Commun. **9** (3) (2017) 379–396.
2. R. Gautam, *Balancing numbers and application*, J. Adv. College Eng. Manage. **4** (2018) 137–143.
3. K. Liptai, F. Luca, A. Pintér and L. Szalay, *Generalized balancing numbers*, Indag. Math. (N.S.) **20** (1) (2009) 87–100.
4. A. Özkoc, *Tridigonal matrices via k-balancing number*, Brit. J. Math. Comput. Sci. **10** (4) (2015) 1–11.
5. A. P. Stakhov, *Fibonacci matrices, a generalization of the cassini formula and new coding theory*, Chaos, Solitions & Fractals **30** (1) (2006) 56–66.

E-mail: e.mehraban.math@gmail.com
E-mail: m-hashemi@guilan.ac.ir

Mackey-Glass Time Series Prediction Using Rough-Neural Networks

Ghasem Ahmadi*
Department of Mathematics, Payame Noor University, P. O. Box 19395-3697, Tehran, Iran
and Mohammad Dehghandar
Department of Mathematics, Payame Noor University, P. O. Box 19395-3697, Tehran, Iran

Abstract. Due to the wonderful properties of artificial neural networks (ANNs) such as universal approximation, they have been used to approximate the nonlinearities in many disciplines of science and engineering. In this work, we propose the rough-neural networks (R-NNs) for the one-step ahead prediction of the Mackey-Glass time series (TS) as an important benchmark problem in TS forecasting. We train the R-NNs with a Lyapunov-based learning algorithm and we compare the simulation results with multilayer perceptron.

Keywords: Artificial neural networks, Time series prediction, Rough-neural networks, Mackey-Glass time series, Lyapunov-based learning algorithm.
AMS Mathematical Subject Classification [2010]: 68T05, 62M10.

1. Introduction

Time series (TS) prediction is an important field of science that has many applications in different aspects such as economic,medical sciences and engineering. Due to the essence of natural processes and the effects of uncertainties and noises, TS prediction is a challenging subject in the literature, and the researchers try to achieve the efficient methodologies in this context.

One of the most powerful approaches in the prediction of TS is the artificial neural networks (ANNs) [5, 6, 7]. ANNs are computational objects that can model the nonlinear processes. They are motivated from the nervous system of human beings. In fact, ANNs are networks of some computational units called neurons, where their connections contain some parameters. These parameters are adjusted with some learning algorithms. In recent years, ANNs are used in many applications and their usefulness have been proved.

Due to the existence of noises and uncertainties in the real TS, the conventional ANNs have some problems in their predictions. To cope with the uncertainties in the TS predictions, in this work the rough-neural networks (R-NNs) have been proposed to predict the nonlinear TS. R-NNs are proposed by Lingers on the basis of rough set theory [9]. They contain some rough neurons that help them to handel the uncertainties. A rough neuron is a pair of conventional neurons called

*Presenter

the upper and lower bound neurons, where the information exchange between them. In recent years they have been used in some applications such as traffic volume prediction [9] and system identification [1, 2, 3, 4] and electricity price forecasting [8].

This work concentrate on the prediction of Mackey-Glass (M-G) TS using R-NNs. M-G TS is a benchmark problem in this context. This TS is considered as the outputs of a continuous-time dynamic system and then, it is identified with R-NNs. An online Lyapunove-based learning algorithm is used to train R-NNs.

The reminder of this work is structured as follows. In Section 2, a brief introduction about the TS prediction is given. In Section 3, the R-NNs are introduced concisely. In Section 4, the problem of TS prediction using R-NNs is considered and a Lyapunov-based learning algorithm is proposed for training R-NNs. In Section 5, the M-G TS is predicted using R-NNs. The conclusions are stated in Section 6.

2. Time Series Prediction

A TS is a set of data $\{x(t)\}_{t=0}^{\infty}$, where t show the time index. Theoretically, $x(t)$ may be considered as a continuous function with the variable t [7]. In the real processes, the sampled data are used to achieve a discrete dataset. In the prediction of TS using neural networks, the past values are used to forecast the future values. In other words, we try to find the function g such that

$$x(t+d) = g(x(t), x(t-1), \ldots, x(t-T)),$$

where T is the number of time steps. If d is chosen 1, then we have the one-step ahead prediction and if $d > 1$, then we have the multi-step ahead prediction.

3. Rough-Neural Networks in TS Predictions

This section gives the structure of R-NN in TS prediction and computes its output. Consider the R-NN, where the hidden neurons are rough and the output neurons are conventional, as shown in Figure 1. We show the output vector of R-NN with $\widehat{x}(t+1)$ and the input vector of R-NN with

$$\mathbf{x} = [\overline{x}(t), \underline{x}(t), \overline{x}(t-1), \underline{x}(t-1), \ldots, \overline{x}(t-T), \underline{x}(t-T), 1]^T,$$

where $\underline{x}$ is the lower bound, and $\overline{x}$ is the upper bound of x, and T is the number of time steps. The component 1 in the vector $\mathbf{x}$ shows the input corresponding to the biases. Let $\underline{V}$ and $\overline{V}$ be the parameters between the inputs and hidden lower bound neurons and the parameters between the inputs and hidden upper bound neurons, respectively. Suppose that $\underline{W}$ and $\overline{W}$ be the parameters between the hidden lower bound neurons and output neurons and the parameters between the hidden upper bound neurons and output neurons, respectively.

Further, let $\underline{O}$, and $\overline{O}$ the outputs of lower bound neurons in the hidden layer and the outputs of upper bound neurons in the hidden layer, respectively. Besides, ϕ shows the activation function in the hidden layer. Then, according to the definition of rough neurons we have [2, 9]:

$$\underline{O} = \min\left(\underline{\phi}, \overline{\phi}\right), \ \overline{O} = \max\left(\underline{\phi}, \overline{\phi}\right),$$

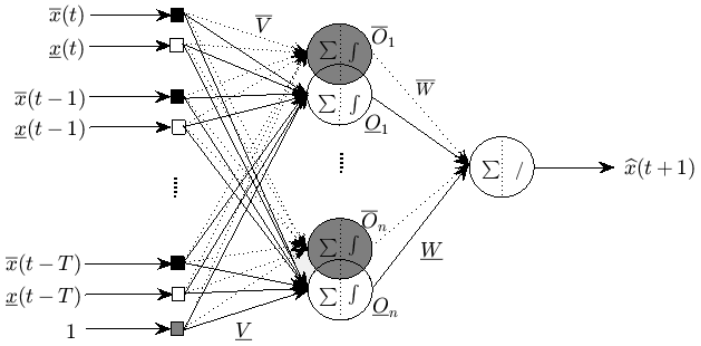


FIGURE 1. Structure of R-NN in TS prediction.

where $\underline{\phi} = \phi(\underline{V}\mathbf{x})$, $\overline{\phi} = \phi(\overline{V}\mathbf{x})$. The output $\widehat{x}(t+1)$ of R-NN is given by

$$\widehat{x}(t+1) = \underline{WO} + \overline{WO} = \underline{W}\min\left(\underline{\phi}, \overline{\phi}\right) + \overline{W}\max\left(\underline{\phi}, \overline{\phi}\right).$$

Consider the n-vectors $\underline{\delta}$ and $\overline{\delta}$ such that $\underline{\delta}^j + \overline{\delta}^j = 1$, $\underline{\delta}^j, \overline{\delta}^j = 0$ *or* 1 $(j = 1, 2, \ldots, n)$, and

$$\underline{\delta}^j \phi^j(\underline{I}) + \overline{\delta}^j \phi^j(\overline{I}) \leq \phi^j(\underline{I}),\ \phi^j(\overline{I}),\ \overline{\delta}^j \phi^j(\underline{I}) + \underline{\delta}^j \phi^j(\overline{I}) \geq \phi^j(\underline{I}),\ \phi^j(\overline{I}).$$

Then, we can write the the output of R-NN in the form

$$\widehat{x}(t+1) = \mathcal{C}\phi(\underline{V}x) + \mathcal{D}\phi(\overline{V}x), \tag{1}$$

where $\mathcal{C} = \underline{W}diag(\underline{\delta}) + \overline{W}diag(\overline{\delta})$ and $\mathcal{D} = \underline{W}diag(\overline{\delta}) + \overline{W}diag(\underline{\delta})$.

4. Time Series Prediction Using Rough-Neural Networks

According to the nature of TS, it can be considered as the outputs of a dynamic system (without inputs) $\dot{\mathbf{z}}(t) = f(\mathbf{z}(t))$ where $\mathbf{z}(t)$ is the state vector of this system. Let A is a Hurwitz matrix and $f(\mathbf{z}(t)) = A\mathbf{z}(t) + g(\mathbf{z}(t))$, where g is the nonlinear part of f. Now, assume that R-NN can model g using the parameters $\widehat{\mathcal{C}}$, $\widehat{\mathcal{D}}$, $\underline{\widehat{V}}$ and $\overline{\widehat{V}}$. Thus, using the Eq. (1), we can write

$$\dot{\widehat{\mathbf{z}}} = A\widehat{\mathbf{z}} + \widehat{\mathcal{C}}\phi\left(\underline{\widehat{V}}x\right) + \widehat{\mathcal{D}}\phi\left(\overline{\widehat{V}}x\right),$$

where

$$\dot{\underline{\widehat{W}}} = e\left[\min(\underline{\phi}, \overline{\phi})\right]\Gamma_1^{-1},\ \dot{\overline{\widehat{W}}} = e\left[\max(\underline{\phi}, \overline{\phi})\right]\Gamma_2^{-1},$$
$$\dot{\underline{\widehat{V}}} = \Gamma_3^{-1}(\underline{\phi}')^T\widehat{\mathcal{C}}^T e x^T,\ \dot{\overline{\widehat{V}}} = \Gamma_4^{-1}(\overline{\phi}')^T\widehat{\mathcal{D}}^T e x^T.$$

These online learning laws have been derived on the basis of Lyapunov stability theory. In [2], a Lyapunov-based learning algorithm is proposed for R-NNs in the identification of nonlinear discrete-time dynamic systems, where in the present work, the derivation of learning laws is done for the continuous-time processes. In the applications the continuous-time models are more reliable than discrete-time models [1].

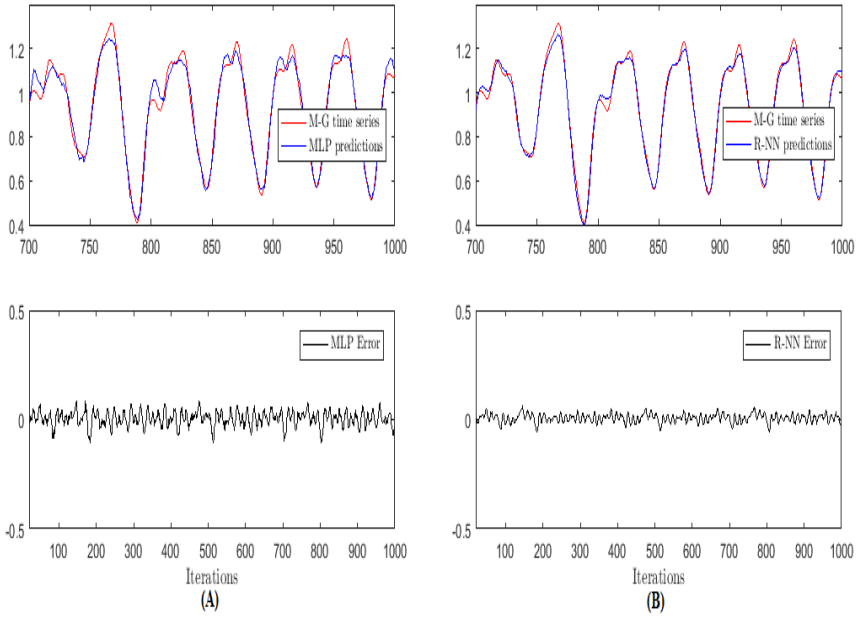


FIGURE 2. The M-G TS, their predictions and the errors in the testing of MLP with 64 hidden neurons (part (A)) and and R-NN with 20 hidden rough neurons (part (B)), in the presence of noises (SNR=20).

5. Mackey-Glass TS Prediction

The M-G TS is generated using the equation

$$\dot{z}(t) = \frac{0.2z(t-\tau)}{1+z^{10}(t-\tau)} - 0.1z(t), \tag{2}$$

where $\tau = 17$. In this work, we used the four existing values $z(t)$, $x(t-6)$, $x(t-12)$, $x(t-18)$ for the one-step ahead prediction of this TS. The one-step ahead prediction of (2) is done by MLP and R-NN, where the activation functions of hidden neurons are sinusoidal.

TABLE 1. MSEs of MLP and R-NN models in the prediction of M-G TS. The column n_h denotes the number of hidden (rough) neurons.

Model	n_h	Parameters	MSE
MLP	32	160	0.0011
MLP	64	320	0.0009
R-NN	10	160	0.0007
R-NN	20	320	0.0005

The initial values of parameters in the models are some random numbers between -0.5 and 0.5. The input vector of MLP is $x = [z(t), z(t-6), z(t-12), 1]^T$

and the input vector of R-NN is

$$x = [\overline{z}(t), \underline{z}(t), \overline{z}(t-6), \underline{z}(t-6), \overline{z}(t-12), \underline{z}(t-12), 1]^T .$$

The design matrix A is chosen as $[-25]$. The algorithm design parameters for MLP are chosen as $n_h = 32,\ 64,\ \Gamma_1, \Gamma_2 = 100 I_{n_h \times n_h}$,where Γ_1 and Γ_2 denote the learning rates. The algorithm design parameters for R-NN are chosen as $n_h = 10,\ 20,\ \Gamma_1, \Gamma_2, \Gamma_3, \Gamma_4 = 100 I_{n_h \times n_h}$,where Γ_1, Γ_2, Γ_3 and Γ_4 denote the learning rates. The sampling time for this simulation is chosen as 0.02.

The MSEs of one-step ahead prediction of M-G TS with MLP and R-NN models are listed in Table 1. According to the number of parameters given in the column "Parameters" of the Table 1, the performance of MLP with 32 and 64 hidden neurons are comparable with the performance of R-NN with 10 and 20 hidden rough neurons, respectively. The M-G TS, their predictions and the errors in the testing of MLP with 64 hidden neurons and and R-NN with 20 hidden rough neurons (in the presence of noises (SNR=20)), are shown in Figure 2. From the Table 1 and Figure 2, we can conclude that the performance of R-NN in the prediction of M-G TS is better than MLP.

6. Conclusion

In this work, we propose the rough-neural networks (R-NNs) for the prediction of Mackey-Glass time series, where the model are trained with a Lyapunov-based learning algorithm. Simulation results show the efficiencies of R-NNs. Future works focus on the usage of R-NNs for the prediction of real-world time series data. In addition, we try to utilize the R-NNs to solve the nonlinear regression problems in the other fields of science and engineering.

References

1. G. Ahmadi, *Stable rough extreme learning machines for the identification of uncertain continuous-time nonlinear systems*, COAM **4** (1) (2019) 83–101.
2. G. Ahmadi and M. Teshnehlab, *Designing and implementation of stable sinusoidal rough-neural identifier*, IEEE Trans. Neural Netw. Learn. Syst. **28** (8) (2017) 1774–1786.
3. G. Ahmadi and M. Teshnehlab, *Identification of multiple input-multiple output non-linear system cement rotary kiln using stochastic gradient-based rough-neural network*, JAIDM **8** (3) (2020) 417–425.
4. G. Ahmadi, M. Teshnehlab and F. Soltanian, *A higher order online Lyapunov-based emotional learning for rough-neural identifiers*, COAM **3** (1) (2018) 87–108.
5. Y. Chen, B. Yang and J. Dong, *Time-series prediction using a local linear wavelet neural network*, Neurocomputing, **69** (4) (2006) 449–465.
6. D. O. Faruk, *A hybrid neural network and ARIMA model for water quality time series prediction*, Eng. Appl. Artif. Intell. **23** (4) (2010) 586–594.
7. R. J. Frank, N. Davey and S. P. Hunt, *Time series prediction and neural networks*, J. Intell. Robot. Syst. **31** (2001) 91–103.
8. H. Jahangir, H. Tayarani, S. Baghali, A. Ahmadian, A. Elkamel and M. A. Golkar, *A novel electricity price forecasting approach based on dimension reduction strategy and rough artificial neural networks*, IEEE Trans. Industr. Inform. **16** (4) (2020) 2369–2381.
9. P. Lingras, *Rough neural networks*, In: Proc. of the 6th Int. conf. on Information Processing and Management of uncertainty in Knowledgebased Systems, Granada, Spain, (1996) pp. 1445–1450.

E-mail: g.ahmadi@pnu.ac.ir
E-mail: m_dehghandar@pnu.ac.ir

Some Families of Composite Graphs and Distance-Based Invariants

Mahdieh Azari*
Department of Mathematics, Kazerun Branch, Islamic Azad University, P. O. Box 73135-168, Kazerun, Iran

Abstract. In this paper, we consider some families of composite graphs such as double graphs, extended double covers, and strong double graphs, and study the relation between some distance-based graph invariants of the resulting graphs with the corresponding invariants of the parent graph.
Keywords: Eccentricity of a vertex, Graph invariant, Composite graph.
AMS Mathematical Subject Classification [2010]: 05C12, 05C76.

1. Introduction

Assume that G is a connected finite graph not containing loop or multiple edges. Its vertex and edge sets are represented by $V(G)$ and $E(G)$, respectively. The greatest distance between a vertex $u \in V(G)$ and other vertices of G is called the eccentricity of u.

A *graph invariant* is a numerical value associated to a graph that is preserved under graph isomorphisms. In recent years, several graph invariants based on the graph theoretical notion of eccentricity have been proposed, most of which are successfully applied in QSAR and QSPR studies in chemistry.

The *eccentric connectivity index* is one of the best-known vertex-eccentricity-based graph invariants. This invariant was suggested in 1997 by Sharma et al. [9] as

$$\xi^c(G) = \sum_{v \in V(G)} d_G(v)\varepsilon_G(v) = \sum_{vu \in E(G)} \big(\varepsilon_G(v) + \varepsilon_G(u)\big),$$

where $d_G(v)$ and $\varepsilon_G(v)$ denote the degree and eccentricity of v, respectively. The *total eccentricity* $\zeta(G)$ is the sum of all vertex eccentricities of G.

Malik [7] put forward the *inverse total eccentricity index* of G as

$$\zeta^{-1}(G) = \sum_{v \in V(G)} \frac{1}{\varepsilon_G(v)}.$$

Hua and Miao [6] proposed the *eccentric connectivity coindex* as

$$\overline{\xi}^c(G) = \sum_{vu \notin E(G)} \big(\varepsilon_G(v) + \varepsilon_G(u)\big).$$

The bound of the above summation is on non-adjacent vertex pairs of G.

*Presenter

The *connective eccentricity index* of G was introduced by Gupta et al. [4] as

$$\xi^{ce}(G) = \sum_{v\in V(G)} \frac{d_G(v)}{\varepsilon_G(v)} = \sum_{vu\in E(G)} \big(\frac{1}{\varepsilon_G(v)} + \frac{1}{\varepsilon_G(u)}\big).$$

The *inverse connective eccentricity index* of G was proposed by Malik [7] as

$$\xi_{ce}^{-1}(G) = \sum_{v\in V(G)} \frac{\varepsilon_G(v)}{d_G(v)}.$$

Ashrafi and Ghorbani [2] introduced the *modified eccentric connectivity index* of G as

$$\xi_c(G) = \sum_{v\in V(G)} \delta_G(v)\varepsilon_G(v),$$

in which $\delta_G(v) = \sum_{vu\in E(G)} d_G(u)$.

Gupta et al. [5] considered the *eccentric adjacency index* of G as

$$\xi^{ad}(G) = \sum_{v\in V(G)} \frac{\delta_G(v)}{\varepsilon_G(v)}.$$

The *eccentric harmonic index* of G was defined by Ediz [3] as

$$H_e(G) = \sum_{vu\in E(G)} \frac{2}{\varepsilon_G(v) + \varepsilon_G(u)}.$$

The *non-self-centrality number* was put forward by Xu et al. [10] as

$$N(G) = \frac{1}{2} \sum_{v,u\in V(G)} |\varepsilon_G(v) - \varepsilon_G(u)|.$$

Many graphs are composed of simpler ones by the use of operations on graphs and, as a consequence, the properties of the resulting graphs are strongly connected to the properties of their building blocks. From that fact, the study of composite graphs is an important research subject in graph theory. In this paper, we consider some distance-based graph invariants including the above- mentioned ones, and study them for some families of composite graphs.

2. Main Results

Here, we consider some families of composite graphs such as double graphs, extended double covers, and strong double graphs, and study the relation between some distance-based graph invariants of the resulting graphs with the corresponding invariants of the parent graph. In this section, we assume that G has n vertices, m edges, and υ vertices of degree $n-1$. We also assume that, $V(G) = \{u_1, u_2, \ldots, u_n\}$, $V = \{v_1, v_2, \ldots, v_n\}$ and $W = \{w_1, w_2, \ldots, w_n\}$.

2.1. Double Graph. The *double graph* $D[G]$ of G is a graph with vertex set $V \cup W$ and edge set $\{v_iv_j, w_iw_j, v_iw_j, v_jw_i : u_iu_j \in E(G)\}$ (See [8]).

Theorem 2.1.

1) $\overline{\xi}^c(D[G]) = 2\big(2\overline{\xi}^c(G) + \zeta(G) + \upsilon\big)$.
2) $\xi^{ce}(D[G]) = 2\big(2\xi^{ce}(G) - \upsilon(n-1)\big)$.
3) $\xi_{ce}^{-1}(D[G]) = \xi_{ce}^{-1}(G) + \frac{\upsilon}{n-1}$.

4) $\xi_c(D[G]) = 8\Big(\xi_c(G) + \sum_{\varepsilon_G(u_i)=1} \delta_G(u_i)\Big)$.
5) $\xi^{ad}(D[G]) = 4\big(2\xi^{ad}(G) - \sum_{\varepsilon_G(u_i)=1} \delta_G(u_i)\big)$.
6) $H_e(D[G]) = 4H_e(G) - \frac{v}{3}(v + 2n - 3)$.
7) $N(D[G]) = 4\big(N(G) - v(n - v)\big)$.

2.2. Extended Double Cover. The *extended double cover* $ED[G]$ of G is a graph with vertex set $V \cup W$ and edge set $\{v_iw_j, v_jw_i : u_iu_j \in E(G)\} \cup \{v_iw_i : 1 \le i \le n\}$ (See [1]).

THEOREM 2.2.

$$\overline{\xi}^c(ED[G]) = 2\big(\overline{\xi}^c(G) + (n-1)\zeta(G) + 2(n(n-1) - m)\big).$$

THEOREM 2.3. *If $G \ncong K_n$, then*

$$\xi^{ce}(ED[G]) < \frac{1}{2}\big(\xi^{ce}(G) + \zeta^{-1}(G) + 2m + n\big).$$

Let $ID(G) = \sum_{i=1}^{n} \frac{1}{d_G(u_i)}$. This invariant is known as the *inverse degree* of G.

THEOREM 2.4. *If $G \ncong K_2$, then*

$$\xi_{ce}^{-1}(ED[G]) < \frac{1}{2}\big(\xi_{ce}^{-1}(G) + ID(G) + \zeta(G) + n\big).$$

Let $M_1(G) = \sum_{i=1}^{n} d_G(u_i)^2$. This invariant is called the *first Zagreb index*.

THEOREM 2.5.

$$\xi_c(ED[G]) = 2\big(\xi_c(G) + 2\xi^c(G) + \zeta(G) + M_1(G) + n + 4m\big).$$

THEOREM 2.6. *If $G \ncong K_n$, then*

$$\xi^{ad}(ED[G]) < \frac{1}{2}\big(\xi^{ad}(G) + 2\xi^{ce}(G) + M_1(G) + \zeta^{-1}(G) + 4m + n\big).$$

THEOREM 2.7. *If $G \ncong K_n$, then*

$$H_e(ED[G]) < \frac{1}{4}\big(2H_e(G) + \zeta^{-1}(G) + 2m + n\big).$$

THEOREM 2.8.

$$N(ED[G]) = 4N(G).$$

2.3. Strong Double Graph. The *strong double graph* $SD[G]$ of G is a graph with vertex set $V \cup W$ and edge set $\{v_iv_j, w_iw_j, v_iw_j, v_jw_i : u_iu_j \in E(G)\} \cup \{v_iw_i : 1 \le i \le n\}$ (See [8]).

THEOREM 2.9.

1) $\overline{\xi}^c(SD[G]) = 4\overline{\xi}^c(G)$.
2) $\xi^{ce}(SD[G]) = 4\big(\xi_{ce}(G) + \zeta^{-1}(G)\big)$.
3) $\xi_{ce}^{-1}(SD[G]) < \frac{1}{4}\big(\xi_{ce}^{-1}(G) + 2\zeta(G)\big)$.
4) $\xi_c(SD[G]) = 2\big(4\xi_c(G) + 4\xi^c(G) + \zeta(G)\big)$.
5) $\xi^{ad}(SD[G]) = 2\big(4\xi^{ad}(G) + 4\xi^{ce}(G) + \zeta^{-1}(G)\big)$.
6) $H_e(\mathcal{SD}[G]) = 4H_e(G) + \zeta^{-1}(G)$.
7) $N(SD[G]) = 4N(G)$.

References

1. N. Alon, *Eigenvalues and expanders*, Combinatorica **6** (1986) 83–96.
2. A. R. Ashrafi and M. Ghorbani, *A study of fullerenes by MEC polynomials*, Electron. Mater. Lett. **6** (2) (2010) 87–90.
3. S. Ediz, *On the Ediz eccentric connectivity index of a graph*, Optoelectron. Adv. Mat. **5** (2011) 1263–1264.
4. S. Gupta, M. Singh and A. K. Madan, *Connective eccentricity index: A novel topological descriptor for predicting biological activity*, J. Mol. Graph. Model. **18** (2000) 18–25.
5. S. Gupta, M. Singh and A. K. Madan, *Predicting anti-HIV activity: Computational approach using a novel topological descriptor*, J. Comput. Aided. Mol. Des. **15** (7) (2001) 671–678.
6. H. Hua and Z. Miao, *The total eccentricity sum of non-adjacent vertex pairs in graphs*, Bull. Malays. Math. Sci. Soc. **42** (3) (2019) 947–963.
7. M. A. Malik, *Two degree-distance based topological descriptors of some product graphs*, Discrete Appl. Math. **236** (C) (2018) 315–328.
8. E. Munarini, C. Perelli Cippo, A. Scagliola and N. Zagaglia Salvi, *Double graphs*, Discrete Math. **308** (2008) 242–254.
9. V. Sharma, R. Goswami and A. K. Madan, *Eccentric connectivity index: A novel highly discriminating topological descriptor for structure-property and structure-activity studies*, J. Chem. Inf. Comput. Sci. **37** (1997) 273–282.
10. K. Xu, K. C. Das and A. D. Maden, *On a novel eccentricity-based invariant of a graph*, Acta Math. Sin. (Engl. Ser.) **32** (1) (2016) 1477–1493.

E-mail: mahdie.azari@gmail.com; azari@kau.ac.ir

Graph Theoretical Models for Genome Rearrangements Analysis

Nafiseh Jafarzadeh*
Faculty of Mathematical Sciences, Tarbiat Modares University, Tehran, Iran
and Ali Iranmnaesh
Faculty of Mathematical Sciences,Tarbiat Modares University, Tehran, Iran

Abstract. The computational study of genome rearrangements is one of the most important research area in computational biology and bioinformatics. In this paper, we define a novel graph data structure as a rearrangement model for whole genome alignment in large scales. This model is capable of realizing non-collinear changes as well as collinear changes. Also we apply our rearrangement graphical model to present a dynamic programing method for alignment of an arbitrary sequence to a pan-genome reference which is encoded as an outerplanar graph. In this method, a gapped alignment is considered, where the gaps could be affine, linear or constant.
Keywords: Genome analysis, Graph theory, Multiple alignment, Genome rearrangement.
AMS Mathematical Subject Classification [2010]: 92B05, 92C05, 92C42.

1. Introduction

The analysis of genome rearrangements has started from 1983, when Dobzhansky and Sturtevant [1] observed that the evolution of certain Drosophila species could be explained using a sequence of reversals. In 1988, Jeffrey Palmer [2] observed some interesting patterns in the evolution of plant organelles and he compared the mitochondrial genomes of cabbages and turnips. The aim of genome rearrangement is to investigate the order of homologous segments and infer genomic distances based on the number of breakpoints or predict scenarios of evolutionary changes. Graphs can assist in improving genome comparison through multiple alignments and also analysis of rearrangements. In addition, graphs provide an intuitive representation of similarities and changes between genomes, and so visualize alignment structures.

2. Method

To represent the common structure between homologous segments in a set of whole genome sequences, we define the concept of "Alignment-set" as follows: An "Alignment-set" is a set of maximal homologous segments with maximal length and denoted by A-set. The size of "A-set" is the number of aligned segments. Note that each A-set may contain multiple segments of the same genome when there is some duplication in a genome. The essential information about possible inversions is the orientation of segments with respect to each other and not the

*Presenter

orientation of the A-set representation. An adjacency of two A-sets $\partial_1, \partial_2 \in \sum_S$ is called a breakpoint if they are adjacent in at least two segments but not in all their segments. Figure 1, is shown an example of 4 breakpoints in multiple alignment of 3 sequences.

3. Discussion

We construct a graph G which every A-set is a vertex of G and there is an edge between two A-sets if there adjacency between two segments of them. In fact, the graph G is an adjacency graph. Since one A-set may contain more than one segment from the same genome, each A-set can be adjacent to itself and also there may be multiple edges between two vertices. Using $pDFS$ Algorithm [3], we compute biconnected components of G. Since in [4] has shown that a graph is outerplanar if and only if every one of its biconnected components is outerplanar, we restrict the outerplanarity to biconnected subgraphs. In [5] a conceptually simple algorithm is presented to determine if a graph is a maximal outerplanar or outerplanar graph. So we apply $MOP - TEST$ Algorithm [5] to recognize outerplanar and non-outerplanar subgraphs. If all of the connected components of G are outerplanar graphs, then we do not need to third step and we can skip that. But if there is one or more non-outerplanar components, they contain some minors isomorphic to K_4 or $K_{2,3}$. By merging two adjacency-connected vertices of K_4 minors and two non-adjacency vertices of $K_{2,3}$ minors, we form graph G to an outerplanar graph G_1. Finally, to make our graph biconnected, we need to make it bridgeless. In the second step, if there is any bridge as a biconnected component, we easily merge vertices of that bridge just like you see in Figure 2.

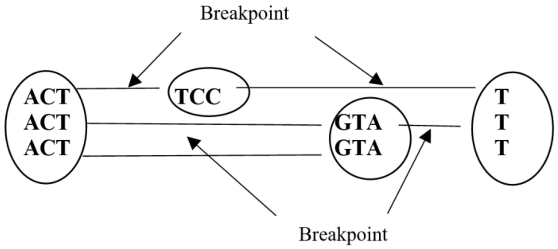


FIGURE 1. Breakpoints in multiple alignment.

4. Conclusion

Genome rearrangements problem consists of finding the evolution between genomes by solving a combinatorial puzzle to find the shortest sequence of rearrangements that can transform each genome into another. In this paper, we described a new graph theoretical data structure as a genome rearrangement model which represents and analyss repeat segments in a set of related genomes. Our method determines an outer planar graph model for multiple alignment of whole genome sequences. Also, this graph representation provides a circular visualization to

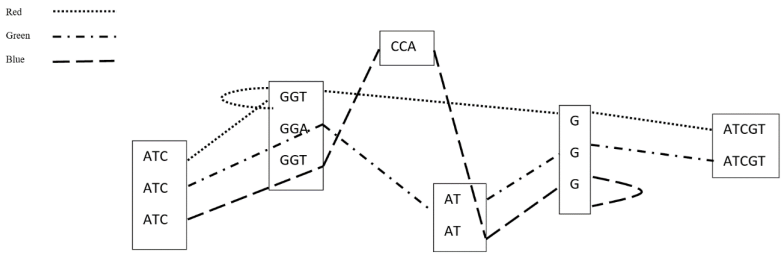


FIGURE 2. The graph G according to 3 sequences ATCGGTTGGGATCGT (Red), ATCAGGATGATCGT (Green) and ATCTGGCCATAGG (Blue).

simplify the study of evolutionary relationships between aligned genomes. Comparing with traditional alignment matrix or partial order alignment graph, our model is more flexible by classification non-collinear structural changes like inversion, translocations and duplications as well as collinear changes like insertion and deletion. This rearrangement model can be use in computational analysis of cancer genomic data and other chromosomal aberrations.

References

1. T. Dobzhansky and A. H. Sturtevant, *Inversions in the chromosomes of Drosophila pseudoobscura*, Genetics. **23** (1938) 28–64.
2. J. D. Palmer and L. A. Herbon, *Plant mitochondrial DNA evolved rapidly in structure, but slowly in sequence*, J. Mol. Evol. **28** (1988) 87–97.
3. J. Fostier, S. Proost, B. Dhoedt, Y. Saeys, P. Demeester, Y. Van de Peer and K. Vandepoele, *A greedy graph-based algorithm for the alignment of multiple homologous gene lists*, Bioinformatics **27** (2011) 749–756.
4. F. Harary, *Graph Theory*, ddison-Wesley, Boston, 1969.
5. S. L. Mitchell, *Linear algorithms to recognize outerplanar and maximal outerplanar graphs*, Infor. Process. Lett. **9** (1979) 229–232.

E-mail: nafise.jafarzadeh@modares.ac.ir
E-mail: iranmanesh@modares.ac.ir

Nonstandard Finite Difference Scheme to Approximate the Coronavirus Disease Model

Mehdi Karami
Department of Mathematics, Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran
Mehran Namjoo*
Department fo Mathematics, Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran
and Mehran Aminian
Department of Mathematics, Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran

Abstract. In this paper, numerical solution of the Coronavirus disease 2019 (COVID-19) model is presented on the basis of nonstandard finite difference (NSFD) scheme. At first, the positivity and boundedness of the model are discussed. Afterwards, the stability analysis of the equilibrium point model is discussed in detail. The nonstandard finite difference scheme is implemented to study the dynamic behaviours COVID-19 model. Numerical results show that the NSFD scheme approach is easy to be implemented and accurate when applied to COVID-19 model.
Keywords: Boundedness, COVID-19, Nonstandard finite difference scheme, Positivity, Stability.
AMS Mathematical Subject Classification [2010]: 34D05, 92D30.

1. Introduction

Modeling and simulation are important decision tools that can be useful to control human diseases [1]. Since each disease exhibits its own particular biological characteristics, the model need to be adapted to each specific case in order to be able to tackle real situations. Coronavirus disease 2019 is an infections disease emerging in China that has rapidly spread in other countries. This is a new virus and a completely new situation. In March 2020, the disease was confirmed in more than 118000 cases reported in 114 countries [2]. The virus spreads from human-to-human via droplets or through contaminated surfaces which in turn enter the nasal mucosa or mucosa eyes through touch. Disease is characterized by cough, fever and sore throat and may result in virus induced pneumonia and progressive respiratory failure owing to alveolar damage caused by the virus. As a result, the mortality rates from the illness also relatively high. Mathematical modeling provides a tool to better understand the transmission dynamics of infectious diseases [3]. These models are one of the most widely used to describe the dynamics of epidemics. In this framework, the changes in the populations of several classes of interacting individuals are described using ordinary differential equations (ODEs). It is not always possible to find the exact solution of the nonlinear models that have at least these ODEs. Therefore, it is sometimes more useful to find numerical solutions of these type systems in order to program easily and visualize the results.

*Presenter

Whenever continuous dynamic systems have been converted into discrete system, the properties of continues systems are not transferred fully to the the discrete systems in the case of large stepsize in the discrete systems. However, if we use NSFD scheme, the properties of the continuous dynamic systems can be preserved into its discrete counterpart. The NSFD scheme is developed for compensating the weakness, such as numerical instability that may be caused by standard finite difference (SFD) scheme. In this research, we apply susceptible, exposed, infected and recovered (SEIR) framework to model the dynamic of COVID-19. A sensible model for the COVID-19 at time t is a system of form

$$
\begin{cases}
S'(t) = -\beta SI, \\
E'(t) = \lambda + \beta SI - \sigma E, \\
I'(t) = \sigma E - \gamma I - \mu I, \\
R'(t) = \gamma I, \\
S(0) = S_0,\ E(0) = E_0,\ I(0) = I_0,\ R(0) = R_0\,.
\end{cases} \tag{1}
$$

In this model, $S(t)$ is the number of susceptible population in the community of time t, $E(t)$ is the number of exposed (but not infectious) people at time t, $I(t)$ stands the number of infected people at time t and $R(t)$ denotes the number of recovered people at time t. Here β is rate the susceptible individual may become exposed to the Coronavirus, σ is passing the time to show the symptoms of COVID-19 or time to identify clinically as positive to the virus, γ is rate the patient can be recovered, μ is rate the patients whose condition is critical can end up losing their lives and λ is rate of virus infection from the exposed people who arrived from the other countries.

Since, the exact solution of the model (1) cannot be easily derived, therefore a numerical approach is used. A number of numerical methods has been developed to solve systems of ordinary differential equations (ODEs). However, the available methods such as standard finite difference schemes and Runge-Kutta sometimes fail to generate the main properties of the model such as stability and positivity. In fact, this can lead to the incorrect interpretation of studied phenomena. The NSFD scheme is the numerical scheme that can be used to simulate the solutions of mathematical model. It is found that the NSFD schemes overcome the weakness of the standard finite difference and Runge-Kutta methods. The reminder of the paper is organized as follows. In Section 2, we prove positivity and boundedness of the solution model of (1). Section 3 deals with stability analysis of COVID-19 model. In Section 4 we provide a summary of the important feature of the procedures for constructing NSFD scheme for systems of ODEs. In continuation, we formulated a NSFD scheme for the COVID-19 model. Finally, Numerical experiments are included to show the efficiency of the NSFD scheme.

2. Positivity and Boundedness

In this section, we prove positivity and boundedness of the solution model of (1). We remark that, the 4th-equation of system (1) is not coupled with the other equations. Thus, we can consider the first three equations of that system.

THEOREM 2.1. *If $S(0)$, $I(0)$ and $E(0) > 0$, then $S(t)$, $I(t) \geq 0$ and $E(t) \geq 0$.*

PROOF. Since the EI-coordinate plane is invariant under the flows of system, therefore for all $t \geq 0$, $S(t) > 0$. Let $A = \{t \geq 0 : E(t) < 0\}$ and $B = \{t \geq$

$0 : I(t) < 0\}$, we will show that $C = A \cup B = \varnothing$. Suppose that $C \neq \varnothing$ and $t_0 = \inf(C)$. So, $t_0 > 0$ and for all $t \in (0, t_0]$, $I(t) \geq 0$ and $E(t) \geq 0$. Now, we show that $E(t_0) \neq 0$. Suppose that $E(t_0) = 0$, it follows that from the third equation of system (1) $\frac{\mathrm{d}E(t_0)}{\mathrm{d}t} = \lambda + \beta S(t_0)I(t_0) > 0$. Hence, there is $\varepsilon > 0$ such that $\frac{\mathrm{d}E(t)}{\mathrm{d}t} > 0$, for all second $t \in (t_0 - \varepsilon, t_0 + \varepsilon) \subseteq (0, +\infty)$. Therefore $E(t_0) > E(t_0 - \varepsilon) \geq 0$. This clearly forces $E(t_0) > 0$. By a similar argument, we conclude that $I(t_0) > 0$. Therefore $t_0 \neq \inf(C)$ and we conclude that C is empty. □

LEMMA 2.2. *Let $K(t)$ be a derivative function from $[0, +\infty]$ to $\mathbb{R}$ such that $K(t) \geq 0$ for every $t \geq 0$. If $\alpha > 0$, $\beta \in \mathbb{R}$, such that $K'(t) + \alpha K(t) \leq \beta$, for every $t \geq 0$, then $K(t) \leq K(0) + \frac{\beta}{\alpha}$.*

LEMMA 2.3. *If $S(0), E(0) \geq 0$ and $I(0) > 0$, then $S(t) \leq S(0)$, for every $t \geq 0$.*

PROOF. By (1) it is obvious that $\frac{\mathrm{d}S}{\mathrm{d}t} = -\beta SI \leq 0$. So, for all $t \geq 0$, $S(t) \leq S(0)$. □

LEMMA 2.4. *If $K(t) = S(t) + E(t)$, then $K(t) \leq L$, where $L = S(0) + E(0) + \frac{\lambda + \sigma S(0)}{\sigma}$.*

PROOF. Our proof starts with the observation that $K'(t) = \lambda - \sigma E$. So $K'(t) + \sigma K(t) = \lambda + \sigma S(t) \leq \lambda + \sigma S(0)$. Therefore by the previous lemma $K(t) \leq K(0) + \frac{\lambda + \sigma S(0)}{\sigma}$. This establishes the formula. □

LEMMA 2.5. *If $G(t) = S(t) + E(t) + I(t)$ then $G(t) \leq M$, where $M = G(0) + \frac{\lambda + (\gamma + \mu)L}{\gamma + \mu}$.*

PROOF. Our proof is based on the fact that $G'(t) = \lambda - (\gamma + \mu)I(t)$. Consequently, $G'(t) + (\gamma + \mu)G(t) = \lambda + (\gamma + \mu)K(t) \leq \lambda + (\gamma + \mu)L$. It follows that $G(t) \leq G(0) + \frac{\lambda + (\gamma + \mu)L}{\gamma + \mu}$. □

3. Stability Analysis of the COVID-19 Model

To evaluate the equilibrium points of the system Eq. (1), let

$$\begin{aligned} -\beta SI &= 0, \\ \lambda + \beta SI - \sigma E &= 0, \\ \sigma E - \gamma I - \mu I &= 0. \end{aligned}$$

Then the equilibrium point is $E = (0, \frac{\lambda}{\sigma}, \frac{\lambda}{\gamma + \mu})$.

THEOREM 3.1. *System (1) is always locally asymptotically stable around E.*

PROOF. At the equilibrium point E, the Jacobian matrix is

$$J(E) = \begin{pmatrix} -\beta \dfrac{\lambda}{\gamma + \mu} & 0 & 0 \\ \beta \dfrac{\lambda}{\gamma + \mu} & -\sigma & 0 \\ 0 & \sigma & -\gamma - \mu \end{pmatrix}.$$

The corresponding eigenvalues are $\lambda_1 = -\beta\frac{\lambda}{\gamma+\mu}$, $\lambda_2 = -\sigma$ and $\lambda_3 = -\gamma-\mu$. Since $\lambda_i < 0$, $i = 1, 2, 3$, therefore the equilibrium point E is asymptotically stable. □

4. A Nonstandard Finite Difference Scheme for the COVID-19 Model

The NSFD schemes were firstly proposed by Mickens for ODEs. To describe a NSFD scheme, we consider an ODE such as

$$x'(t) = f(t, x, \lambda),\ x(0) = x_0,\ t \in [0, t_f], \tag{2}$$

where λ is a parameter. Given a discretization $t_n = nh$.

NSFD scheme is constructed by following two main steps. First, the derivative of the left-hand side of Eq. (2) is replaced by a discrete from $x'(t_k) \approx \frac{x_{k+1}-x_k}{\Phi(h,\lambda)}$, where x_k is an approximation of $x(t_k)$ and $0 < \Phi(h) < 1$ with $\Phi(h) = h + O(h^2)$. Second, the nonlinear term in the (2) is replaced by a nonlocal discrete approximation $F(t, x_{k+1}, x_k, \ldots, \lambda)$ depending on some of the previous approximation [4, 5, 6, 7]. Hence, the gained scheme is described as follows

$$\frac{\Phi(x_{k+1}) - \Phi(x_k)}{\Phi(h, \lambda)} = F(t, x_{k+1}, x_k, \ldots, \lambda)\,.$$

Examples of the denominator function that satisfy the above condition are h, $\sin(h)$, $1 - e^{-h}$, $\frac{1-e^{-h}}{\lambda}$. The second NSFD scheme requirement is that the dependent functions should be modeled on the discrete time computational grid. For example, the terms x^2 and xy can be approximated using $x_n x_{n+1}$ and $x_{n+1}y_n$, respectively. Applying the NSFD scheme, we obtain the following discrete model for model (1)

$$\begin{cases} \dfrac{S_{k+1} - S_k}{\Phi_1(h)} = -\beta S_{k+1} I_k\,, \\ \dfrac{E_{k+1} - E_k}{\Phi_2(h)} = \lambda + \beta S_{k+1} I_k - \sigma E_{k+1}\,, \\ \dfrac{I_{k+1} - I_k}{\Phi_3(h)} = \sigma E_{k+1} - (\gamma+\mu) I_{k+1}\,. \end{cases} \tag{3}$$

Rearranging the Eq. (3), we get

$$\begin{cases} S_{k+1} = \dfrac{S_k}{1 + h\beta I_k}\,, \\ E_{k+1} = \dfrac{E_k + \lambda\Phi_2(h) + \beta\Phi_2(h) S_{k+1} I_k}{1 + \sigma\Phi_2(h)}\,, \\ I_{k+1} = \dfrac{\sigma\Phi_3(h) E_{k+1} + I_k}{1 + (\gamma+\mu)\Phi_3(h)}\,, \end{cases} \tag{4}$$

where

$$\Phi_1(h) = h,\ \ \Phi_2(h) = \frac{e^{\sigma h} - 1}{\sigma},\ \ \Phi_3(h) = \frac{e^{(\gamma+\mu)h} - 1}{\gamma+\mu}\,.$$

The Eq. (4) should be computed in sequence, because the value of S_{k+1} is used for calculating the value E_{k+1}, which is then used to calculate the value of I_{k+1}. Observe that right hand side of (4) is always positive for all stepsize h. Therefore, solution of (4) is always positive with any positive initial value and stepsize.

5. Numerical Results

In this section, the numerical solutions of the proposed NSFD scheme on three cases are presented. In the first simulation we consider the parameter values $\beta = 0.07$, $\gamma = 0.24$, $\mu = 0.02$, $\sigma = 1.4$ and $\lambda = 0.000205$ with initial condition value $S_0 = 20$, $E_0 = 30$ and $I_0 = 25$ for simulating time $1000s$ and stepsize $h = 0.2$. Figure 1 shows that the NSFD scheme (4) converges to the equilibrium point $E = (0, 0.000146, 0.000125)$. In Figure 2 the numerical solutions of NSFD scheme (4) are depicted by choosing $\beta = 0.7$, $\gamma = 0.14$, $\mu = 0.06$, $\sigma = 0.13$ and $\lambda = 3$ with initial condition $S_0 = 20$, $E_0 = 30$, $I_0 = 25$ and stepsize $h = 0.5$. The Figure 2 confirms that (S_k, E_k, I_k) converges to the equilibrium point $E = (0, 23.07, 15)$. Finally, In Figure 3 we have plotted the behaviour of the NSFD scheme for the value parameters $\beta = 0.7$, $\gamma = 0.14$, $\mu = 0.06$, $\sigma = 0.13$ and $\lambda = 100$ for simulating time $t = 1000s$ and stepsize $h = 0.2$. The Figure 3 confirms that (S_k, E_k, I_k) approaches the equilibrium point solution $E = (0, 789, 23, 500)$. The results show that the numerical solutions of the proposed NSFD scheme preserves the main properties of the COVID-19 model such positivity and stability.

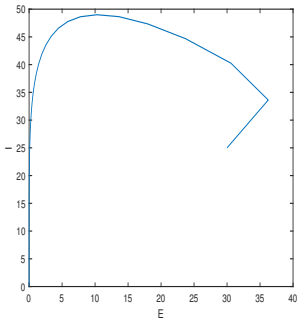


FIGURE 1. Numerical simulation with $h = 0.2$ for NSFD scheme in $E - I$ plane.

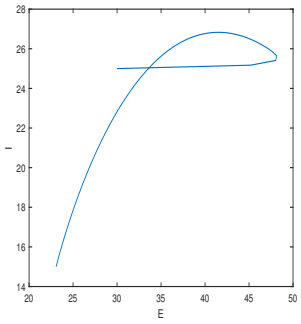


FIGURE 2. Numerical simulation with $h = 0.5$ for NSFD scheme in $E - I$ plane.

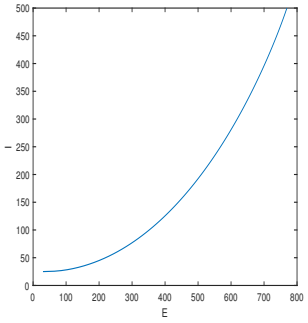


FIGURE 3. Numerical simulation with $h = 0.2$ for NSFD scheme in $E - I$ plane.

References

1. M. Anderson, *Population biology of infectious diseases: Part I.*, Nature **280** (5721) (1979) 361–367.
2. L. Wang, Y. Wang, Y. Chen and Q. Qin, *Unique epidemiological and clinical features of the emerging* 2019 *novel Coronavirus phenomena (COVID-*19*) implicate special control measures*, J. Med. virol. (2020). DOI: 10.1002/jmv.25748
3. A. J. Kucharski, T. W. Russell, C. Diamond, Y. Liu, J. Edmunds, S. Funk and R. M. Eggo, *Early dynamics of transmission and control of Covid-*19*: A mathematical modelling study*, Lancet. Infect. Dis. **20** (5) (2020) 553–558.
4. R. E. Mickens, *Advances in the Applications of Nonstandard Finite Difference Schemes*, Wiley-Interscience, Singapore, 2005.
5. S. Zibaei and M. Namjoo, *A non-standard finite difference scheme for solving fractional-order model of HIV-1 infection of CD* 4^+ *T-cells*, Iranian J. Math. Chem. **6** (2) (2015) 145–160.
6. S. Zibaei and M. Namjoo, *A nonstandard finite difference scheme for solving three-species food chain with fractional-order Lotka-Volterra model*, Iranian J. Numer. Anal. Optim. **6** (1) (2016) 53–78.
7. S. Zibaei and M. Namjoo, *A NSFD scheme for Lotka-Volterra food web model*, Iran. J. Sci. Technol. Trans. A Sci. **38** (4) (2014) 399–414.

E-mail: m.karami@vru.ac.ir
E-mail: namjoo@vru.ac.ir
E-mail: mehran.aminian@vru.ac.ir

On Neutro Quadruple Groups

Florentin Smarandache
Department of Mathematics and Science, University of New Mexico, Gallup, NM 87301, USA
Akbar Rezaei*
Department of Mathematics, Payame Noor University, P. O. Box 19395-3697, Tehran, Iran
Adesina Abdul Akeem Agboola
Department of Mathematics, College of Physical Sciences, Director ICTREC, Federal University of Agriculture, PMB 2240, Abeokuta, Ogun State, Nigeria
Young Bae Jun
Gyeongsang National University, South Korea
Rajab Ali Borzooei
Department of Mathematics, Shahid Beheshti University, Tehran, Iran
Bijan Davvaz
Department of Mathematics, Yazd University, Yazd, Iran
Arsham Borumand Saeid
Department of Pure Mathematics, Faculty of Mathematics and Computer, Shahid Bahonar University of Kerman, Kerman, Iran
Mohammad Akram
University of the Punjab, New Campus, Lahore, Pakistan
Mohammad Hamidi and Saeed Mirvakili
Department of Mathematics, Payame Noor University, P. O. Box 19395-3697, Tehran, Iran

Abstract. As generalizations and alternatives of classical algebraic structures there have been introduced in 2019 the Neutro Algebraic Structures (or Neutro Algebras) and Anti Algebraic structures (or Anti Algebras). Unlike the classical algebraic structures, where all operations are well-defined and all axioms are totally true, in Neutro Algebras and Anti Algebras the operations may be partially well-defined and the axioms partially true or respectively totally outer-defined and the axioms totally false. These Neutro Algebras and Anti Algebras form a new field of research, which is inspired from our real world. In this paper, we study neutrosophic quadruple algebraic structures and Neutro Quadruple Algebraic Structures. Neutro Quadruple Group is studied in particular and several examples are provided. It is shown that $(NQ(\mathbb{Z}), \div)$ is a Neutro Quadruple Group. Substructures of Neutro Quadruple Groups are also presented with examples.

Keywords: Neutrosophic quadruple number, Neutro Quadruple Group, Neutro Quadruple Subgroup.

AMS Mathematical Subject Classification [2010]: 03E72, 06F35, 08A72.

*Presenter

1. Introduction

It was started from Paradoxism, then to Neutrosophy, and afterwards to Neutrosophic Set and Neutrosophic Algebraic Structures. Paradoxism [10] is an international movement in science and culture, founded by Smarandache in 1980s, based on excessive use of antitheses, oxymoron, contradictions, and paradoxes. During the three decades (1980-2020) hundreds of authors from tens of countries around the globe contributed papers to 15 international paradoxist anthologies. In 1995, Smarandache extended the paradoxism (based on opposites) to a new branch of philosophy called neutrosophy (based on opposites and their neutrals), that gave birth to many scientific branches, such as: neutrosophic logic, neutrosophic set, neutrosophic probability and statistics, neutrosophic algebraic structures, and so on with multiple applications in engineering, computer science, administrative work, medical research etc. Neutrosophy is an extension of Yin-Yang Ancient Chinese Philosophy and of course of Dialectics. From Classical Algebraic Structures to Neutro Algebraic Structures and Anti Algebraic Structures. In 2019 Smarandache [8] generalized the classical glgebraic structures to Neutro Algebraic Structures (or Neutro Algebras) whose operations and axioms are partially true, partially indeterminate, and partially false as extensions of Partial Algebra, and to Anti Algebraic Structures (or AntiAlgebra) whose operations and axioms are totally false. "Algebra" can be: groupoid, semigroup, monoid, group, commutative group, ring, field, vector space, BCK-Algebra, BCI-Algebra, K-algebra, BE-algebra, etc. (See [1]-[7]).

In the present paper, we study neutrosophic quadruple algebraic structures and Neutro Quadruple Algebraic Structures. Neutro Quadruple Group is studied in particular and several examples are provided. It is shown that $(NQ(\mathbb{Z}), \div)$ is a Neutro Quadruple Group. Substructures of Neutro Quadruple Groups are also presented with examples.

The sets of natural/integer/rational/real/complex numbers are respectively denoted by $\mathbb{N} \subseteq \mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R} \subseteq \mathbb{C}$.

The Neutrosophic Quadruple Numbers and the Absorbance Law were introduced by Smarandache in 2015 [9]; they have the general form:
$N = a + bT + cI + dF$, where a, b, c, d may be numbers of any type (natural, integer, rational, irrational, real, complex, etc.), where "a" is the known part of the neutrosophic quadruple number N, while "$bT + cI + dF$" is the unknown part of the neutrosophic quadruple number N; then the unknown part is split into three subparts: degree of confidence (T), degree of indeterminacy of confidence (non-confidence) (I), and degree of non-confidence (F). N is a four-dimensional vector that can also be written as: $N = (a, b, c, d)$.

There are transcendental, irrational etc. numbers that are not well known, they are only partially known and partially unknown, they may have infinitely many decimals. Not even the most modern supercomputers can compute more than a few thousands decimals, but the infinitely many left decimals still remain unknown. Therefore, such numbers are very little known (because only a finite number of decimals are known), and infinitely unknown (because an infinite number of decimals are unknown). Take for example: $\sqrt{2} = 1.4142\ldots$.

2. Arithmetic Operations on the Neutrosophic Set of Quadruple Numbers

DEFINITION 2.1. A neutrosophic set of quadruple numbers denoted by $NQ(X)$ is a set defined by

$$NQ(X) = \{(a, bT, cI, dF) : a, b, c, d \in \mathbb{R} \text{ or } \mathbb{C}\},$$

where T, I, F have their usual neutrosophic logic meanings.

DEFINITION 2.2. A neutrosophic quadruple number is a number of the form $(a, bT, cI, dF) \in NQ(X)$. For any neutrosophic quadruple number (a, bT, cI, dF) representing any entity which may be a number, an idea, an object, etc, a is called the known part and (bT, cI, dF) is called the unknown part. Two neutrosophic quadruple numbers $x = (a, bT, cI, dF)$ and $y = (e, fT, gI, hF)$ are said to be equal written $x = y$ if and only if $a = e, b = f, c = g, d = h$.

Multiplication of two neutrosophic quadruple numbers cannot be carried out like multiplication of two real or complex numbers. In order to multiply two neutrosophic quadruple numbers $a = (a_1, a_2T, a_3I, a_4F), b = (b_1, b_2T, b_3I, b_4F) \in NQ(X)$, the prevalence order of $\{T, I, F\}$ is required.

Two neutrosophic quadruple numbers $m = (a_1, b_1T, c_1I, d_1F)$ and $n = (a_2, b_2T, c_2I, d_2F)$ cannot be divided as we do for real and complex numbers. Since the literal neutrosophic components T, I and F are not invertible, the inversion of a neutrosophic quadruple number or the division of a neutrosophic quadruple number by another neutrosophic quadruple number must be carried out a systematic way. Suppose we are to evaluate m/n. Then we must look for a neutrosophic quadruple number $p = (x, yT, zI, wF)$ equivalent to m/n. In this way, we write

$$\begin{aligned} m/n &= p \\ \Rightarrow \frac{(a_1, b_1T, c_1I, d_1F)}{(a_2, b_2T, c_2I, d_2F)} &= (x, yT, zI, wF) \\ \Leftrightarrow (a_2, b_2T, c_2I, d_2F)(x, yT, zI, wF) &\equiv (a_1, b_1T, c_1I, d_1F). \end{aligned} \tag{1}$$

Assuming the prevalence order $T \succ I \succ F$ and from the equality of two neutrosophic quadruple numbers, we obtain from Eq. (1)

$$\begin{aligned} a_2x &= a_1, \\ b_2x + (a_2 + b_2 + c_2 + d_2)y + b_2z + b_2w &= b_1, \\ c_2x + (a_2 + c_2 + d_2)z + c_2w &= c_1, \\ d_2x + (a_2 + d_2)w &= d_1, \end{aligned}$$

a system of linear equations in unknowns x, y, z and w. By similarly assuming the prevalence order $T \prec I \prec F$, we obtain from Eq. (1)

$$\begin{aligned} a_2x &= a_1, \\ b_2x + (a_2 + b_2)y &= b_1, \\ c_2x + c_2y + (a_2 + b_2 + c_2)z &= c_1, \\ d_2x + d_2y + d_2z + (a_2 + b_2 + c_2 + d_2)w &= d_1, \end{aligned}$$

a system of linear equations in unknowns x, y, z and w.

3. Neutrosophic Quadruple Algebraic Structures, Neutrosophic Quadruple Algebraic Hyper Structures and Neutro Quadruple Algebraic Structures

3.1. Neutrosophic Quadruple Algebraic Structures and Neutrosophic Quadruple Algebraic Hyper Structures. Let $NQ(X)$ be a neutrosophic quadruple set and let $* : NQ(X) \times NQ(X) \to NQ(X)$ be a classical binary operation on $NQ(X)$. The couple $(NQ(X), *)$ is called a neutrosophic quadruple algebraic structure. The structure $(NQ(X), *)$ is named according to the classical laws and axioms satisfied or obeyed by $*$.

If $* : NQ(X) \times NQ(X) \to \mathbb{P}(NQ(X))$ is the classical hyper operation on $NQ(X)$. Then the couple $(NQ(X), *)$ is called a neutrosophic quadruple hyper algebraic structure; and the hyper structure $(NQ(X), *)$ is named according to the classical laws and axioms satisfied by $*$.

3.2. Neutro Quadruple Algebraic Structures. In this section unless otherwise stated, the optimistic prevalence order $T \succ I \succ F$ will be assumed.

Definition 3.1. Let $NQ(G)$ be a nonempty set and let $* : NQ(G) \times NQ(G) \to NQ(G)$ be a binary operation on $NQ(G)$. The couple $(NQ(G), *)$ is called a neutrosophic quadruple group if the following conditions hold:

(QG1) $x * y \in G \ \forall x, y \in NQ(G)$ [closure law].
(QG2) $x * (y * z) = (x * y) * z \ \forall x, y, z \in G$ [axiom of associativity].
(QG3) There exists $e \in NQ(G)$ such that $x * e = e * x = x \ \forall x \in NQ(G)$ [axiom of existence of neutral element].
(QG4) There exists $y \in NQ(G)$ such that $x * y = y * x = e \ \forall x \in NQ(G)$ [axiom of existence of inverse element], where e is the neutral element of $NQ(G)$.
If in addition $\forall x, y \in NQ(G)$, we have
(QG5) $x * y = y * x$, then $(NQ(G), *)$ is called a commutative neutrosophic quadruple group.

Definition 3.2. [Neutro Sophication of the law and axioms of the neutrosophic quadruple]

(NQ(G)1) There exist some duplets $(x, y), (u, v), (p, q), \in NQ(G)$ such that $x*y \in G$ (inner-defined with degree of truth T) and $[u * v =$ indeterminate (with degree of indeterminacy I) or $p * q \notin NQ(G)$ (outer-defined/falsehood with degree of falsehood F)] [Neutro Closure Law].
(NQ(G)2) There exist some triplets $(x, y, z), (p, q, r), (u, v, w) \in NQ(G)$ such that $x * (y * z) = (x * y) * z$ (inner-defined with degree of truth T) and $[[p * (q * r)]$or $[(p * q) * r] =$ indeterminate (with degree of indeterminacy I) or $u * (v * w) \neq (u * v) * w$ (outer-defined/falsehood with degree of falsehood F)] [NeutroAxiom of associativity (Neutro Associativity)].
(NQ(G)3) There exists an element $e \in NQ(G)$ such that $x * e = e * x = x$ (inner-defined with degree of truth T) and $[[x * e]$or$[e * x] =$ indeterminate (with degree of indeterminacy I) or $x * e \neq x \neq e * x$ (outer-defined/falsehood with degree of falsehood F)] for at least one $x \in NQ(G)$ [Neutro Axiom of existence of neutral element (Neutro Neutral Element)].

(NQ(G)4) There exists an element $u \in NQ(G)$ such that $x * u = u * x = e$ (inner-defined with degree of truth T) and [[$x*u$]or[$u*x$)] = indeterminate (with degree of indeterminacy I) or $x * u \neq e \neq u * x$ (outer-defined/falsehood with degre of falsehood F)] for at least one $x \in G$ [Neutro Axiom of existence of inverse element (Neutro Inverse Element)], where e is a Neutro Neutral Element in $NQ(G)$.

(NQ(G)5) There exist some duplets $(x, y), (u, v), (p, q) \in NQ(G)$ such that $x * y = y * x$ (inner-defined with degree of truth T) and [[$u * v$]or[$v * u$] = indeterminate (with degree of indeterminacy I) or $p * q \neq q * p$ (outer-defined/falsehood with degree of falsehood F)] [Neutro Axiom of commutativity (Neutro Commutativity)].

Definition 3.3. A Neutro Quadruple Group $NQ(G)$ is an alternative to the neutrosophic quadruple group $Q(G)$ that has at least one NeutroLaw or at least one of $\{NQ(G)1, NQ(G)2, NQ(G)3, NQ(G)4\}$ with no Anti Law or Anti Axiom.

Definition 3.4. A Neutro Commutative Quadruple Group $NQ(G)$ is an alternative to the commutative neutrosophic quadruple group $Q(G)$ that has at least one Neutro Law or at least one of $\{NQ(G)1, NQ(G)2, NQ(G)3, NQ(G)4\}$ and $NQ(G)5$ with no Anti Law or Anti Axiom.

NeutroClosure of $\div$ over $NQ(\mathbb{Z})$

For the degree of truth, let $a = (0, 0T, I, 0F) \in NQ(\mathbb{Z})$. Then

$$a \div a = \frac{(0, 0T, I, 0F)}{(0, 0T, I, 0F)} = (1 - k_1 - k_2, 0T, k_1 I, k_2 F) \in NQ(\mathbb{Z}), k_1, k_2 \in \mathbb{Z}.$$

For the degree of indeterminacy, let $a = (4, 5T, -2I, -7F), b = (0, -6T, I, 3F) \in NQ(\mathbb{Z})$. Then

$$a \div b = \frac{(4, 5T, -2I, -7F)}{(0, -6T, I, 3F)} = \left(\frac{4}{0}, ?T, ?I, ?F\right) \notin NQ(\mathbb{Z}).$$

For the degree of falsehood, let $a = (0, 0T, 0I, F), b = (0, 0T, 0I, 2F) \in NQ(\mathbb{Z})$. Then

$$a \div b = \frac{(0, 0T, 0I, F)}{(0, 0T, 0I, 2F)} = \left(\frac{1}{2} - k, 0T, 0I, kF\right) \notin NQ(\mathbb{Z}), k \in \mathbb{Z}.$$

Neutro Associativity of $\div$ over $NQ(\mathbb{Z})$

For the degree of truth, let $a = (6, 6T, 6I, 6F), b = (2, 2T, 2I, 2F)$, $c = (-1, 0T, 0I, 0F) \in NQ(\mathbb{Z})$. Then

$$\begin{aligned}
a \div (b \div c) &= (6, 6T, 6I, 6F) \div ((2, 2T, 2I, 2F) \div (-1, 0T, 0I, 0F)) \\
&= (6, 6T, 6I, 6F) \div (-2, 0T, 0I, 0F) \\
&= (-3, 0T, 0I, 0F). \\
(a \div b) \div c &= ((6, 6T, 6I, 6F) \div (2, 2T, 2I, 2F)) \div (-1, 0T, 0I, 0F) \\
&= (3, 0T, 0I, 0F) \div (-1, 0T, 0I, 0F) \\
&= (-3, 0T, 0I, 0F).
\end{aligned}$$

For the degree of indeterminacy, let $a=(4,-T,2I,-7F), b=(0,T,0I,-8F)$, $c=(0,0T,9I,-F)\in NQ(\mathbb{Z})$. Then

$$\begin{aligned}
a\div(b\div c) &= (4,-T,2I,-7F)\div((0,T,0I,-8F)\div(0,0T,9I,-F))\\
&= (4,-T,2I,-7F)\div\left(8-k,\frac{1}{8}T,-9I,kF\right),k\in\mathbb{Z}\\
&= (?,?T,?I,?F).\\
(a\div b)\div c &= ((4,-T,2I,-7F)\div(0,T,0I,-8F))\div(0,0T,9I,-F)\\
&= \left(\frac{4}{0},?T,?I,?F\right)\div(0,0T,9I,-F)\\
&= (?,?T,?I,?F).
\end{aligned}$$

For the degree of falsehood, let $a=(0,5T,0I,0F), b=(0,T,0I,0F)$, $c=(5,0T,0I,0F)\in NQ(\mathbb{Z})$. Then

$$\begin{aligned}
a\div(b\div c) &= (0,5T,0I,0F)\div((0,T,0I,0F)\div(5,0T,0I,0F))\\
&= (0,5T,0I,0F)\div\left(0,\frac{1}{5}T,0I,0F\right)\\
&= (25-k_1-k_2-k_3,k_1T,k_2I,k_3F)\in NQ(\mathbb{Z}),k_1,k_2,k_3\in\mathbb{Z}.\\
(a\div b)\div c &= ((0,5T,0I,0F)\div(0,T,0I,0F))\div(5,0T,0I,0F)\\
&= (5-k_1-k_2-k_3,k_1T,k_2I,k_3F)\div(5,0T,0I,0F),k_1,k_2,k_3\in\mathbb{Z}\\
&= \left(\frac{1}{5}(5-k_1-k_2-k_3),\frac{1}{5}k_1T,\frac{1}{5}k_2I,\frac{1}{5}k_3F\right)\notin NQ(\mathbb{Z}).
\end{aligned}$$

Existence of Neutro Unitary Element and Neutro Inverse Element in $NQ(\mathbb{Z})$ w.r.t. $\div$

Let $a=(0,T,0I,0F), b=(0,0T,I,0F), c=(0,0T,0I,F)\in NQ(\mathbb{Z})$. Then

$$\begin{aligned}
(2)\,a\div a &= \frac{(0,T,0I,0F)}{(0,T,0I,0F)}=(1-k_1-k_2-k_3,k_1T,k_2I,k_3F),k_1,k_2,k_3\in\mathbb{Z}.\\
(3)\,b\div b &= \frac{(0,0T,I,0F)}{(0,0T,I,0F)}=(1-k_1-k_2,0T,k_1I,k_2F),k_1,k_2\in\mathbb{Z}.\\
(4)\,c\div c &= \frac{(0,0T,0I,F)}{(0,0T,0I,F)}=(1-k,0T,0I,kF),k\in\mathbb{Z}.\\
(5)\,a\div b &= \frac{(0,T,0I,0F)}{(0,0T,I,0F)}=(-(k_1+k_2),T,k_1I,k_2F),k_1,k_2\in\mathbb{Z}.\\
(6)\,b\div a &= \frac{(0,0T,I,0F)}{(0,T,0I,0F)}=(-(k_1+k_2+k_3),k_1T,k_2I,k_3F),k_1,k_2,k_3\in\mathbb{Z}.
\end{aligned}$$

For the degree of truth, putting $k_1=1,k_2=k_3=0$ in Eq. (2), $k_1=1,k_2=0$ in Eq. (3) and $k=1$ in Eq. (4) we will obtain $a\div a=a, b\div b=b$ and $c\div c=c$. These show that a,b,c are respectively Neutro Unitary Elements and Neutro Inverse Elements in $NQ(\mathbb{Z})$.

For the degree of falsehood, putting $k_1\neq 1,k_2\neq k_3\neq 0$ in Eq. (2), $k_1\neq 1,k_2\neq 0$ in Eq. (3) and $k\neq 1$ in Eq. (4) we will obtain $a\div a\neq a, b\div b\neq b$ and $c\div c\neq c$. These show that a,b,c are respectively not Neutro Unitary Elements and Neutro Inverse Elements in $NQ(\mathbb{Z})$.

Neutro Commtativity of $\div$ over $NQ(\mathbb{Z})$

For the degree of truth, putting $k_1 = 1, k_2 = k_3 = 0$ in Eq. (2), $k_1 = 1, k_2 = 0$ in Eq. (3) and $k = 1$ in Eq. (4) we will obtain $a \div a = a, b \div b = b$ and $c \div c = c$. These show the commutativity of $\div$ wrt a, b and c $NQ(\mathbb{Z})$.

For the degree of falsehood, putting $k_1 = k_2 = k_3 = 1$ in Eq. (5) and Eq. (6), we will obtain $a \div b = (-2, T, I, F)$ and $b \div a = (-3, T, I, F) \neq a \div b$. Hence, $\div$ is Neutro Commutative in $NQ(\mathbb{Z})$.

Definition 3.5. Let $(NQ(G), *)$ be a neutrosophic quadruple group. A nonempty subset $NQ(H)$ of $NQ(G)$ is called a Neutro Quadruple Subgroup of $NQ(G)$ if $NQ(H), *)$ is a neutrosophic quadruple group of the same type as $(NQ(G), *)$.

Example 3.6.

i) For $n = 2, 3, 4, \ldots$ $(NQ(n\mathbb{Z}), -)$ is a Neutro Quadruple Subgroup of $(NQ(\mathbb{Z}), -)$.
ii) For $n = 2, 3, 4, \ldots$ $(NQ(n\mathbb{Z}), \times)$ is a Neutro Quadruple Subgroup of $(NQ(\mathbb{Z}), \times)$.

Example 3.7.

i) Let $NQ(H) = \{(a, bT, cI, dF) : a, b, c, d \in \{1, 2, 3\}\}$ be a subset of the Neutro Quadruple Group $(NQ(\mathbb{Z}_4), -)$. Then $(NQ(H), -)$ is a Neutro Quadruple Subgroup of $(NQ(\mathbb{Z}_4), -)$.
ii) Let $NQ(K) = \{(w, xT, yI, zF) : a, b, c, d \in \{1, 3, 5\}\}$ be a subset of the Neutro Quadruple Group $(NQ(\mathbb{Z}_6), \times)$. Then $(NQ(H), \times)$ is a Neutro Quadruple Subgroup of $(NQ(\mathbb{Z}_6), \times)$.

4. Conclusion

We have in this paper studied neutrosophic quadruple algebraic structures and Neutro Quadruple Algebraic Structures. Neutro Quadruple Group was studied in particular and several examples were provided. It was shown that $(NQ(\mathbb{Z}), \div)$ is a Neutro Quadruple Group. Substructures of Neutro Quadruple Groups were also presented with examples.

Acknowledgement

The authors would like to express their thanks to the referees for their valuable suggestions and comments.

References

1. A. A. A. Agboola, M. A. Ibrahim and E. O. Adeleke, *Elementary examination of NeutroAlgebras and AntiAlgebras viz-a-viz the classical number systems,* Int. J. Neutrosophic Sci. **4** (1) (2020) 16–19.
2. M. Akram and K. -P. Shum, *A survey on single-valued neutrosophic K-algebras*, J. Math. Res. Appl. **40** (3) (2020). DOI: 10.3770/j.issn:2095-2651.2020.03.000
3. M. Al-Tahan and B. Davvaz, *On some properties of neutrosophic quadruple Hv-rings*, Neutrosophic Sets Sys. **36** (2020) 256–270.
4. R. A. Borzooei, M. Mohseni Takallo, F. Smarandache and Y. B. Jun, *Positive implicative BMBJ-neutrosophic ideals in BCK-algebras*, Neutrosophic Sets Sys. **23** (2018) 126–141.

5. M. Hamidi and F. Smarandache, *Neutro-BCK-Algebra*, Int. J. Neutrosophic Sci. **8** (2020) 110–117.
6. Y. B. Jun, S. Z. Song, F. Smarandache and H. Bordbar, *Neutrosophic Quadruple BCK/BCI-Algebras*, Axioms **7** (41) (2018). DOI:10.3390/axioms7020041.
7. A. Rezaei and F. Smarandache, *On Neutro-BE-Algebras and Anti-BE-Algebras*, Int. J. Neutrosophic Sci. **4** (2020) 8–15.
8. F. Smarandache, *Introduction to NeutroAlgebraic structures and AntiAlgebraic structures, in advances of standard and nonstandard neutrosophic theories,* Pons Publishing House Brussels, Belgium, Chapter 6, pages 240–265, 2019.
9. F. Smarandache, *Neutrosophic quadruple numbers, refined neutrosophic quadruple numbers, absorbance law, and the multiplication of neutrosophic quadruple numbers*, Neutrosophic Sets Sys. **10** (2015) 96–98.
10. UNM, *Paradoxism, the last vanguard of second millennium*, http://fs.unm.edu/a/paradoxism.htm.

E-mail: smarand@unm.edu
E-mail: rezaei@pnu.ac.ir
E-mail: agboolaaaa@funaab.edu.ng
E-mail: skywine@gmail.com
E-mail: borzooei@sbu.ac.ir
E-mail: davvaz@yazd.ac.ir
E-mail: arsham@uk.ac.ir
E-mail: m.akram@pucit.edu.pk
E-mail: m.hamidi@pnu.ac.ir
E-mail: saeed_mirvakili@pnu.ac.ir

Contributed Posters

Numerical Analysis

A Note on Family of Additive Semi-Implicit Runge-Kutta Schemes

Sadegh Amiri*
Department of Basic Sciences, Shahid Sattari Aeronautical University of Science and Technology, P. O. Box 13846-63113, Tehran, Iran

Abstract. In this paper, we deal with the order conditions of a family of an additive semi-implicit Runge-Kutta schemes for solving ordinary differential equations (ODEs). It is shown that for the multi-dimensional case, some of extracting order conditions must be added to the order conditions obtained from these methods in the one-dimensional case.
Keywords: Additive semi-implicit Runge-Kutta, Order conditions.
AMS Mathematical Subject Classification [2010]: 65Nxx, 65N06.

1. Introduction

The effects of viscosity, heat conduction, diffusion and hypersonic flows often contain non equilibrium processes of thermal excitations and chemical reactions because of high gas temperature and high speeds. One of the major difficulties in computing such flows is the stiffness of the governing equations in temporal integrations. So, additive semi-implicit Runge-Kutta methods for stiff semi-discrete systems of ordinary differential equations for transient hypersonic flows with thermo-chemical non-equilibrium systems are considered by some authors [1, 2]. For example, the researchers in [2] have studied on three different semi-implicit Runge-Kutta methods for additively split differential equations in the form of $u' = f(u) + g(u)$, where f is treated by explicit Runge-Kutta methods and g is simultaneously treated by three implicit Runge-Kutta methods. They have formulated parameter identification as a multi-dimensional problem. They have derived the coefficients of up to third-order accurate additive semi-implicit Runge-Kutta methods. Also, they have considered a general s-stage additive semi-implicit Runge-Kutta methods as follows:

$$\begin{aligned} u^{n+1} &= u^n + \sum_{i=1}^{s} w_i K_i, \\ K_i &= h\left(f(u^n + \sum_{j=1}^{i-1} b_{ij} K_j) + g(u^n + \sum_{j=1}^{i-1} c_{ij} K_j + a_i K_i) \right). \end{aligned} \tag{1}$$

The parameters w_i, b_{ij}, c_{ij} and a_i are coefficients were be determined from accuracy and stability conditions. Authors in [2] obtained the order conditions of methods (1) such that these methods are of up to third-order and then authors in [1] used this order conditions for nonstiff systems. Authors in [2] declared that these

*Presenter

order conditions are obtained based on Taylor expansion. Therefore, in this work, we focus on the specialized way which exploit the order conditions for the multi-dimensional ODEs. In the following, we proof that based on the multi-dimensional Taylor expansion one must has 9 order conditions instead of 8 order conditions, when we want to obtain the third order method.

2. Extracting Order Conditions for the Multi-Dimensional ODEs

At first we consider ODE system

$$x'(t) = F(t, x(t)), \quad t \in [t_0, T], \ x(t_0) = x_0, \tag{2}$$

where $x_0 \in \mathbb{R}^d$ and $d \geq 1$ is positive integer and $F : [t_0, T] \times \mathbb{R}^d \longrightarrow \mathbb{R}$. Without lose of generality, instead of system (2) we can consider autonomous form

$$x'(t) = F(x(t)), \quad t \in [0, T], \ x(t_0) = x_0,$$

where $x_0 \in \mathbb{R}^d$ and $d \geq 1$ is positive integer and $F : \mathbb{R}^d \longrightarrow \mathbb{R}$. Therefore in this case for stepsize h, according to multi-dimensional Taylor expansion, for the Ith component of $x(t_0 + h)$ we have

$$x^I(t_0 + h) = x_0^I + hF^I + \frac{h^2}{2!}(F'F)^I + \frac{h^3}{3!}(F''(F, F) + F'(F'(F)))^I + O(h^4),$$

where the components of vectors are denoted by superscript indices which chosen as capitals and

$$(F'F)^I = \sum_{J=1}^{d} F^J \frac{\partial F^I}{\partial x^J}, \tag{3}$$

$$(F''(F, F))^I = \sum_{J,L=1}^{d} F^J F^L \frac{\partial^2 F^I}{\partial x^J \partial x^L}, \tag{4}$$

$$(F'(F'(F)))^I = \sum_{J,L=1}^{d} F^L \frac{\partial F^I}{\partial x^J} \frac{\partial F^J}{\partial x^L}. \tag{5}$$

Now if we put $F = f + g$ then we have

$$\begin{aligned} x^I(t_0 + h) = x_0^I + h(f + g)^I + \frac{h^2}{2!}((f + g)'(f + g))^I \\ + \frac{h^3}{3!}((f + g)''((f + g), (f + g)) + (f + g)'((f + g)'((f + g))))^I + O(h^4), \end{aligned}$$

and so according to (3)-(5) we get,

$$\begin{aligned} x^I(t_0 + h) = x_0^I + h(f + g)^I + \frac{h^2}{2!}\left(\sum_{J=1}^{d}(f + g)^J \frac{\partial f^I}{\partial x^J} + \sum_{J=1}^{d}(f + g)^J \frac{\partial g^I}{\partial x^J}\right) \\ + \frac{h^3}{3!}\left(\sum_{J,L=1}^{d}(f + g)^J (f + g)^L \frac{\partial^2 f^I}{\partial x^J \partial x^L} + \sum_{J,L=1}^{d}(f + g)^J (f + g)^L \frac{\partial^2 g^I}{\partial x^J \partial x^L}\right. \\ \left. + \sum_{J,L=1}^{d}(f + g)^L \frac{\partial f^I}{\partial x^J} \frac{\partial f^J}{\partial x^L} + \sum_{J,L=1}^{d}(f + g)^L \frac{\partial f^I}{\partial x^J} \frac{\partial g^J}{\partial x^L}\right. \end{aligned}$$

$$+\sum_{J,L=1}^{d}(f+g)^L\frac{\partial g^I}{\partial x^J}\frac{\partial f^J}{\partial x^L}+\sum_{J,L=1}^{d}(f+g)^L\frac{\partial g^I}{\partial x^J}\frac{\partial g^J}{\partial x^L}\Bigg)+O(h^4).$$

Therefore based on multi-dimensional Taylor expansion of $x^I(t_0+h)$ the method is of first order if:

$$w_1+w_2+w_3=1,$$

second order if

$$w_2b_{21}+w_3(b_{31}+b_{32})=\frac{1}{2},$$
$$w_1a_1+w_2(a_2+c_{21})+w_3(a_3+c_{31}+c_{32})=\frac{1}{2},$$

third order if

$$w_2b_{21}^2+w_3(b_{31}+b_{32})^2=\frac{1}{3},$$
$$w_3b_{21}b_{32}=\frac{1}{6},$$
$$w_2a_2b_{21}+w_3(b_{21}c_{32}+a_3(b_{31}+b_{32}))=\frac{1}{6},\tag{6}$$
$$w_2b_{21}a_1+w_3(a_1b_{31}+a_2b_{32}+c_{21}b_{32})=\frac{1}{6},\tag{7}$$
$$w_1a_1^2+w_2(a_2^2+c_{21}(a_1+a_2))+w_3(a_1c_{31}+c_{32}(a_2+c_{21})+a_3(a_3+c_{31}+c_{32}))=\frac{1}{6},$$
$$w_1a_1^2+w_2(a_2+c_{21})^2+w_3(a_3+c_{31}+c_{32})^2=\frac{1}{3},$$

while authors in [1, 2] obtained 8 order conditions. This is due to the fact that in one dimensional case ($d=1$) the terms

$$\sum_{J,L=1}^{d}(f+g)^L\frac{\partial f^I}{\partial x^J}\frac{\partial g^J}{\partial x^L},$$

and

$$\sum_{J,L=1}^{d}(f+g)^L\frac{\partial g^I}{\partial x^J}\frac{\partial f^J}{\partial x^L},$$

be equal while for multi-dimensional case ($d>1$) the situation is different and these are not equal. Indeed, in the case of $d=1$ instead of order conditions (6)-(7), we have the following order condition

$$w_2(b_{21}a_2+b_{21}a_1)+w_3(a_1b_{31}+a_2b_{32}+c_{21}b_{32}+b_{21}c_{32}+a_3b_{31}+a_3b_{32})=\frac{1}{3}.$$

3. Conclusion

In this study, we investigate the order conditions of a family of an additive semi-implicit Runge-Kutta schemes for solving ordinary differential equations. Also, we consider some recent attempts about these methods. Finally, we proved that in the multi-dimensional problem ($d>1$) we must have 9 order conditions instead of 8 order conditions obtained by [1, 2].

References

1. C. Pantano, *An additive semi-implicit Runge-Kutta family of scheme for nonstiff systems*, Appl. Numer. Math. **57** (2007) 297–303.
2. X. Zhong, *Additive semi-implicit Runge-Kutta methods for computing high-speed nonequilibrium reactive flows*, J. Comput. Phys. **128** (1996) 19–31.

E-mail: s.amiri@ssau.ac.ir
E-mail: amirimath@yahoo.com

An Inverse Problem for an Equation Modeling Shallow Water under Small Rotation

Fatemeh Ghanadian*
School of Mathematics and Computer Science, Damghan University, Damghan 36715-364, Iran
Reza Pourgholi
School of Mathematics and Computer Science, Damghan University, Damghan 36715-364, Iran
and Seyed Hashem Tabasi
School of Mathematics and Computer Science, Damghan University, Damghan 36715-364, Iran

Abstract. This article we consider a nonlinear inverse problem related to an equation modeling shallow water under small rotation. By using noisy data, we apply two B-Splines with different levels, the quitic B-spline and septic B-spline, to study this problem. For both levels, we prove the stability and convergence analysis. The results show that an excellent estimation of the unknown functions of the nonlinear inverse problem.
Keywords: Shallow water, Inverse problem, Quartic B-spline, Stability.
AMS Mathematical Subject Classification [2010]: 35Q53, 35B35, 68W25, 35R30, 65M70.

1. Introduction

It is known that the KdV-Burgers equation

$$u_t + bu_{xxx} + uu_x - au_{xx} = f(x,t),$$

was derived in [10] as a dissipated version of the KdV equation

$$u_t + bu_{xxx} + uu_x = f(x,t), \tag{1}$$

to describe the propagation of undular bores in shallow water, and weakly nonlinear plasma waves with certain dissipative effects. Ignoring the dissipation term the KdV equation, has solitary wave solutions. In this paper, we consider the Ostrovsky-Burgers equation

$$(u_t + bu_{xxx} + uu_x - au_{xx})_x = \gamma u + f(x,t). \tag{2}$$

Equation (2) was appeared in modeling internal waves in the ocean or surface waves in a shallow channel with an uneven bottom under the effects of the interfacial friction (See [5, Chapter 1] and [6, 11]). In (2), and the positive constants γ and a are the rotation and friction coefficients, respectively. The function f denotes the external force and b is the dispersion coefficient which its sign is related to the type of dispersion. Ignoring the dissipation term u_{xx}, (2) leads to the

*Presenter

Ostrovsky equation which was derived by Ostrovsky in 1978 [9] to model weakly nonlinear surface and internal waves in a rotating ocean; see also [3, 4]. It was also demonstrated in [8] that the nonlinear oblique magneto-acoustic waves in a rotating plasma can be described by (1). A model of the propagation of long internal waves in a deep rotating fluid can be found in [2]. If one considers the limit of no high-frequency dispersion $b = 0$, the resulting equation is called the OstrovskyHunter equation [1]. It is worth noting that in spite of similarity of structures of (1) and the KdV equation, the Ostrovsky equation, unlike the KdV equation, is evidently nonintegrable by the method of inverse scattering transform [4].

In the present paper, we study numerically Eq. (2) in the domain $(x,t) \in [0,1] \times [0,T]$ with the final time T, the initial condition

$$u(x,0) = p(x), \quad x \in [0,1], \tag{3}$$

and boundary conditions

$$u(0,t) = f_1(t), \quad u_x(0,t) = f_2(t), \quad t \in [0,T], \tag{4}$$

$$u(1,t) = g_1(t), \quad u_x(1,t) = g_2(t), \quad u_{xx}(1,t) = g_3(t), \quad t \in [0,T], \tag{5}$$

where $p(x)$, $g_1(t)$, $g_2(t)$, $g_3(t)$ and $f(x,t)$ are continuous known functions, while $f_1(t)$, $f_2(t)$, and the wave amplitude $u(x,t)$ are unknown which remain to be determined. Here, we consider two numerical methods to find the solutions of (2), Collocation method based on septic B-spline basis functions and quintic B-spline basis functions. It is known that the use of B-splines have many different features and are effective in numerical works. One of the most important feature is that the conditions on the continuity of functions are built-in and have the smooth interpolation functions. On the other hand, as the support of each B-spline is embedded only on a few sub-intervals, the resulting matrix related to the discretized equation will be tightly banded.

Moreover, if one combine with collocation, the solution procedure will be clear and shorten.

2. Main Results

We first use the Septic B-spline collection method, region of the solution of the problem is restrained over $0 \leq x \leq 1$. The Eqs. (2)-(5) will be solved with the over-specified conditions

$$u(a,t) = h_1(t), \quad u_x(a,t) = h_2(t), \quad u_{xx}(a,t) = h_3(t), \tag{6}$$

where $t \in [0,T]$ and $0 < a < 1$ is a fixed point. We define the septic B-spline $B_j(x)$ for $j = -3(1)N+3$ as in [7]. Now let $U_m(x,t) \in \zeta$ be the B-spline approximation to the exact solution $u(x,t)$ in the form $U_m(x,t) = \sum_{j=-3}^{m+3} c_j(t) B_j(x)$. By substituting the trial functions B_j into the above identity, the nodal values of $U, U^{'}, U^{''}, U^{'''}, U^{(4)}$ and $U^{(5)}$ are obtained in terms of the element parameters c_m

by

$$
\begin{aligned}
U_m &= c_{m-3} + 120c_{m-2} + 1191c_{m-1} + 2416c_m + 1191c_{m+1} + 120c_{m+2} + c_{m+3},\\
U'_m &= \frac{7}{h}(-c_{m-3} - 56c_{m-2} - 245c_{m-1} + 2451c_{m+1} + 56c_{m+2} + c_{m+3}),\\
U''_m &= \frac{42}{h^2}(c_{m-3} + 24c_{m-2} + 15c_{m-1} - 80c_m + 15c_{m+1} + 24c_{m+2} + c_{m+3}),\\
U'''_m &= \frac{210}{h^3}(-c_{m-3} - 8c_{m-2} + 19c_{m-1} - 19c_{m+1} + 8c_{m+2} + c_{m+3}),\\
U^{(4)}_m &= \frac{840}{h^4}(c_{m-3} - 9c_{m-1} + 16c_m - 9c_{m+1} + c_{m+3}),\\
U^{(5)}_m &= \frac{2520}{h^5}(-c_{m-3} + 4c_{m-2} - 5c_{m-1} + 5c_{m+1} - 4c_{m+2} + c_{m+3}).
\end{aligned}
$$

First we use the following finite difference approximation to discretize $u^n_t = \dfrac{u^{n+1} - u^n}{k}$ and $u^n = \dfrac{u^{n+1} + u^n}{2}$, where $u^n = u(x, t_n)$ and $u^0 = u(x, 0) = p(x)$. The nonlinear term is linearized by using the quasi-linearization formula: $f(u^{n+1}, u^{n+1}_x) = f(u^n, u^n_x) + (u^{n+1} - u^n)\frac{\partial f^n}{\partial u} + (u^{n+1}_x - u^n_x)\frac{\partial f^n}{\partial u_x}$. By replacing the approximate solution U with u, and using the nodal values U and the derivatives of U, System, consists of $(N+1)$ linear equation with $(N+7)$ unknowns. To have a unique solution of the above system we are required the over-specified condition (6). Suppose that $a = x_s$, $1 \le s \le N-1$, thusly we have $u(x_s, t) = h_1(t)$, $u_x(x_s, t) = h_2(t)$ and $u_{xx}(x_s, t) = h_3(t)$, where $t \in [0, T]$ Hence, we derive that $AC = B$ is a system of $(N+7)$ linear equations with $(N+7)$ unknowns.

We notice that the matrix A is ill-condition, so we obtain solution of system $AC = B$ by using the Tikhonov regularization method.we check the convergence of our algorithm Suppose that $U(x) = \sum_{j=-3}^{N+3} c_j B_j(x)$ is the B-spline collocation approximation of $u(x)$. The following lemma and theorem will be important in our analysis that proofs of them have been done.

LEMMA 2.1. *If* $\{B_{-3}, B_{-2}, B_{-1}, B_0, \ldots, B_N, B_{N+1}, B_{N+2}, B_{N+3}\}$ *be the septic B-spline, then* $\left|\sum_{j=-3}^{N+3} B_j(x)\right| \le 7456$ *for* $x \in [0, 1]$.

THEOREM 2.2. *Let* $u \in C^8[0, 1]$ *be an exact solution of* (2) *such that* $\left|\frac{\partial^8 u(x,t)}{\partial x^8}\right| \le L$ *for all* x *and* t. *If* $U(x, t)$ *is the numerical approximation by our method of* u, *then* $\|u(x) - U(x)\| \le O(k + h^4)$.

Also we will solve the inverse problem (2) by a new modification of the quintic Bsplines collocation method with the over-specified conditions $u(a, t) = h_1(t)$ and $u_x(a, t) = h_2(t)$, where $t \in [0, t_f]$, $0 < a < 1$ is a fixed point. We consider the quintic B-splines [12]. Let $U_m(x, t) \in \zeta$ be the B-spline approximation to the exact solution $u(x, t)$ in the form $U_m(x, t) = \sum_{j=-2}^{m+2} c_j(t) B_j(x)$, where $c_j(t)$ are time dependent parameters determined by the boundary and collocation conditions. Substituting the trial functions B_j into the above equation, the nodal values of U,

U', U'', U''' and $U^{(4)}$ are obtained in terms of the element parameters c_m by

$$
\begin{aligned}
U_m &= c_{m-2} + 26c_{m-1} + 66c_m + 26c_{m+1} + c_{m+2},\\
U'_m &= \frac{5}{h}(-c_{m-2} - 10c_{m-1} + 10c_{m+1} + c_{m+2}),\\
U''_m &= \frac{20}{h^2}(c_{m-2} + 2c_{m-1} - 6c_m + 2c_{m+1} + c_{m+2}),\\
U'''_m &= \frac{60}{h^3}(-c_{m-2} + 2c_{m-1} - 2c_{m+1} + c_{m+2}),\\
U^{(4)}_m &= \frac{120}{h^4}(c_{m-2} - 4c_{m-1} + 6c_m - 4c_{m+1} + c_{m+2}).
\end{aligned} \tag{7}
$$

System, contains $(N+1)$-linear equation with $(N+5)$ unknowns. To have a unique solution of the above system, we are required the above over-specified condition. Assume that $a = x_s$, $1 \le s \le N-1$. Equation (5) holds and moreover $u(x_s,t) = h_1(t), u_x(x_s,t) = h_2(t)$, where $t \in [0,T]$. If we consider $m = s$ in (7), then we have

$$
\begin{aligned}
h_1(t_{n+1}) &= c^{n+1}_{s-2} + 26c^{n+1}_{s-1} + 66c^{n+1}_s + 26c^{n+1}_{s+1} + c^{n+1}_{s+2},\\
h_2(t_{n+1}) &= \frac{5}{h}(c^{n+1}_{s-2} + 10c^{n+1}_{s-1} - 10c^{n+1}_{s+1} - c^{n+1}_{s+2}),\\
g_1(t_{n+1}) &= c^{n+1}_{N-2} + 26c^{n+1}_{N-1} + 66c^{n+1}_N + 26c^{n+1}_{N+1} + c^{n+1}_{N+2},\\
g_2(t_{n+1}) &= \frac{5}{h}(c^{n+1}_{N-2} + 10c^{n+1}_{N-1} - 10c^{n+1}_{N+1} + c^{n+1}_{N+2}).
\end{aligned}
$$

Consequently, $AC = B$ is a system of $(N+5)$ linear equations with $(N+5)$-unknown functions. We notice that the matrix A is ill-condition, so we obtain solution of system $AC = B$ by using the Tikhonov regularization method. Similar to the convergence of the previous part, we need to recall the following lemma and theorem that proofs of them have been done,

LEMMA 2.3. *The B-splines $\{B_{-2}, B_{-1}, B_0, \ldots, B_N, B_{N+1}, B_{N+2}\}$ satisfies the following inequality $\left|\sum_{j=-2}^{N+2} B_j(x)\right| \le 186$ for $x \in [0,1]$.*

THEOREM 2.4. *Let $u(x,t) \in C^6[0,1]$ be the exact solution of (2) such that $\left|\frac{\partial^6 u(x,t)}{\partial x^6}\right| \le L$. Assume that $U(x,t)$ is the numerical approximation by our methods, then $\|u(x) - U(x)\| \le O(k + h^2)$.*

We investigated the stability for both methods by applying Von-Neuman stability analysis.

EXAMPLE 2.5. In our first example, we consider the nonlinear inverse problem (2) and (3)-(5), where $a = 1$, $b = 5$ and $\gamma = 3$ with the initial data $u(x,0) = \sin(x)$, and the external force $f(x,t) = \cos(t+x) + \frac{3}{2}\sin(t+x) + \cos^2(t+x) - \sin^2(t+x)$. An exact solutions of this problem is $u(x,t) = \sin(x+t)$ with $u(0,t) = f_1(t) = \sin(t), u_x(0,t) = f_2(t) = \cos(t)$. Tables 1-3 show total error S for some values of N for each method.

We have employed successfully the septic B-spline and the quintic B-spline method to estimate unknown boundary conditions in an inverse problem related to the Ostrovsky-Burgers equations (2) and (3)-(5). By comparing the numerical

TABLE 1. The comparison between exact and numerical solutions of Example 2.5 for $u(0.1,t)$, $|u(0.1,t)-u^*(0.1,t)|$, by the Quintic and Septic B-spline methods with $N=30,50,100$.

	N=30		N=50		N=100	
t	quintic	Septic	quintic	Septic	quintic	Septic
0.1	$4.8e-04$	$7.2e-06$	$8.7e-04$	$6.5e-05$	$1.4e-03$	$4.5e-05$
0.5	$1.7e-03$	$1.5e-05$	$2.7e-03$	$1.8e-08$	$3.5e-03$	$2.2e-05$
1	$3.0e-03$	$2.5e-06$	$4.2e-03$	$3.3e-06$	$5.2e-03$	$2.1e-05$
S_{f_1}	$5.9758e-05$	$1.1963e-06$	$8.7071e-05$	$1.4328e-06$	$1.1311e-04$	$1.3533e-06$

TABLE 2. The comparison between exact and numerical solutions of Example 2.5 for $f_1(t)$, $|f_1(t)-f_1^*(t)|$, by employing the Quintic and Septic B-spline methods with $N=30,50,100$.

	N=30		N=50		N=100	
t	quintic	Septic	quintic	Septic	quintic	Septic
0.1	$1.9e-05$	$1.9e-05$	$1.1e-04$	$5.9e-05$	$1.2e-04$	$1.9e-06$
0.5	$8.2e-04$	$2.8e-05$	$6.2e-04$	$1.3e-05$	$4.0e-04$	$2.0e-07$
1	$1.6e-03$	$1.5e-06$	$1.1e-03$	$3.7e-06$	$6.5e-04$	$4.6e-07$
S_{f_1}	$3.1833e-05$	$1.1633e-06$	$2.3106e-05$	$1.1575e-06$	$1.3381e-05$	$9.8944e-07$

TABLE 3. The comparison between exact and numerical solutions of Example 2.5 for $f_2(t)$, $|f_2(t)-f_2^*(t)|$, by the Quintic and Septic B-spline methods with $N=30,50,100$.

	N=30		N=50		N=100	
t	quintic	Septic	quintic	Septic	quintic	Septic
0.1	$4.9e-03$	$1.9e-03$	$3.6e-04$	$6.3e-03$	$1.2e-02$	$7.7e-04$
0.5	$2.8e-02$	$3.7e-03$	$2.8e-02$	$2.4e-03$	$4.0e-02$	$1.5e-04$
1	$5.2e-02$	$3.5e-03$	$5.5e-02$	$3.0e-04$	$6.5e-02$	$1.1e-04$
S_{f_1}	$9.6783e-04$	$7.7698e-05$	0.0012	$8.3921e-05$	0.0013	$3.1064e-05$

results, we showed that the accuracy and stability of the septic B-spline method is more than ones for the the quintic B-spline method. Since, the associated coefficient matrix in the septic B-spline method and quintic B-spline method are usually ill-conditioned,we have used the Tikhonov regularization method to obtain a stable numerical approximation of solution.

References

1. G. U. Chen and J. P. Boyd, *Analytical and numerical studies of weakly nonlocal solitary waves of the rotation-modified Kortewegde Vries equation,* Phys. D **155** (2001) 201–222.
2. A. Esfahani and S. Levandosky, *Solitary waves of the rotation-generalized Benjamin-Ono equation*, Discrete Contin. Dyn. Syst. **33** (2013) 663–700.
3. V. N. Galkin and Y. A. Stepanyants, *On the existence of stationary solitary waves in a rotating field,* Appl. Math. Mech. **55** (1991) 939–943.
4. O. A. Gilman, R. Grimshaw and Y. A. Stepanyants, *Approximate and numerical solutions of the stationary Ostrovsky equation,* Stud. Appl. Math. **95** (1995) 115–126
5. R. Grimshaw, *Internal Solitary Waves*, Environmental Stratified Flows, Kluwer Academic Publishers, Dordrecht, 2001.

6. H. Mitsudera and R. Grimshaw, *Effects of friction on a localized structure in a baroclinic current*, J. Physical Oceanography **23** (1993) 2265–2292.
7. R. Mohammadi *Numerical approximation for viscous CahnHilliard equation via septic B-spline*, Appl. Anal. **100** (1) (2021) 93–115.
8. M. A. Obregon and Y. A. Stepanyants, *Oblique magneto-acoustic solitons* 460 *in rotating plasma*, Phys. Lett. A **249** (1998) 315–323.
9. L. A. Ostrovsky, *Nonlinear internal waves in a rotating ocean,* Oceanologia **18** (1978) 181–191.
10. E. Ott and R. N. Sudan, *Damping of solitary waves*, Phys. Fluids **13** (1970) 1432–1434.
11. H. Wang and A. Esfahani, *Well-posedness and asymptotic behavior of the dissipative Ostrovsky equation*, Evol. Equ. Control Theory **8** (2019) 709–735.
12. H. Zhang, X. Han and X. Yang, *Quintic B-spline collocation method for fourth order partial integro-differential equations with a weakly singular kernel*, Appl. Math. Comput. **219** (2013) 6565–6575.

E-mail: ghanadian85@gmail.com
E-mail: pourgholi@du.ac.ir
E-mail: tabasi@du.ac.ir

Local RBF-PUM for the Steady-State Diffusion-Reaction System with Discontinuous Coefficients

Faranak Gholampour*
Department of Mathematics, Shiraz University of Technology, Shiraz, Iran
Esmail Hesameddini
Department of Mathematics, Shiraz University of Technology, Shiraz, Iran
and Ameneh Taleei
Department of Mathematics, Shiraz University of Technology, Shiraz, Iran

Abstract. In this work, we propose the radial basis function (RBF) partition of unity method (PUM) for system of steady-state diffusion-reaction equations with discontinuous coefficients in 2D. The collocation based RBF-PUM is a local mesh-free method that reduces the computational cost of the global versions. To ensure the stability of the solution, as the shape parameter ε goes to zero, the RBF-QR algorithm is employed. This algorithm bypasses troubles associated with the determination of ε and enables us to get higher accuracy. Our results show the potential of proposed method in handling arbitrary interfaces and relatively large scale domains.
Keywords: Radial basis function (RBF), Partition of unity method (PUM), RBF-QR algorithm, Steady-state diffusion-reaction system.
AMS Mathematical Subject Classification [2010]: 35J57, 65N35, 82B24.

1. Introduction

In the past decades, research utilizing mesh-free methods have gained significant attention in solving engineering problems. Especially, in the simulation of problems consisting of different materials or the same material but at different states, in which the coefficients of the partial differential equations (PDEs) might have discontinuities along the material interfaces. Such problems are known as interface problems. For many applications in real-world problems, the material interfaces and boundaries can be complicated and very irregular. However for geometrically large-scale problems, the computational cost related to dense matrices is the main issue of the global methods [1]. This research is devoted to a localized RBF method, the collocation based RBF partition of unity method (RBF-PUM) [2, 3]. This method leads to well-conditioned discrete systems which allow us to address large-scale problems. In this method, the domain is covered by a collection of overlapping patches.

However, we face with numerical ill-conditioning for small values of shape parameter ε that correspond to increasing flatness of C^∞ RBFs. Note that the near-flat parameter regime is often of particular computational interest in terms of accuracy. It has been found out that this difficulty can be bypassed by applying proper algorithms [3, 4]. The stable RBF algorithms make the choice of ε

*Presenter

much less critical. One type of these algorithms is the RBF-QR algorithm which includes a change from the C^∞ Gaussian basis function to a better-conditioned basis. Therefore, we employ the RBF-QR algorithm that allows stable evaluations for any small ε [5, 6]. Our aim of this paper is to investigate the local RBF-PUM in solving the steady-state coupled diffusion-reaction system in 2D involving complicated interface within relatively large domain sizes. Let $\Omega \subset \mathbb{R}^2$ is separated into two disjoint subdomains Ω^- and Ω^+ by an interface Γ such that $\Omega = \Omega^+ \cup \Omega^- \cup \Gamma$. Consider the following general steady-state coupled diffusion-reaction system with Dirichlet boundary conditions,

$$-\mathrm{div}(\alpha(\mathbf{x})\nabla u(\mathbf{x})) + \beta_1(\mathbf{x})u(\mathbf{x}) + \beta_2(\mathbf{x})v(\mathbf{x}) = f_1(\mathbf{x}), \tag{1}$$

$$-\mathrm{div}(\sigma(\mathbf{x})\nabla v(\mathbf{x})) + \gamma_1(\mathbf{x})v(\mathbf{x}) + \gamma_2(\mathbf{x})u(\mathbf{x}) = f_2(\mathbf{x}), \quad \mathbf{x} \in \Omega^{\pm}, \tag{2}$$

where the coefficients may have finite discontinuities along the interface. The jumps in solutions and their derivatives can be specified as jump conditions across the interface, i.e.,

$$[u(\mathbf{x})]_\Gamma = h_1(\mathbf{x}), \quad [\alpha u_{\mathbf{n}}(\mathbf{x})]_\Gamma = h_2(\mathbf{x}), \tag{3}$$

$$[v(\mathbf{x})]_\Gamma = k_1(\mathbf{x}), \quad [\sigma v_{\mathbf{n}}(\mathbf{x})]_\Gamma = k_2(\mathbf{x}), \quad \mathbf{x} \in \Gamma, \tag{4}$$

where vector $\mathbf{n}$ is the unit normal direction pointing to Ω^+ side.

1.1. The RBF-QR Partition of Unity Method. To apply the partition of unity method for a problem, a set of overlapping patches $\{\Omega_j\}_{j=1}^{N_p}$ are constructed to cover the domain Ω, i.e., $\Omega \subseteq \bigcup_{j=1}^{N_p} \Omega_j$. Note that there should be an upper bound K for the number of patches that overlap at any $\mathbf{x} \in \Omega$. The global approximant $\widehat{u}(\mathbf{x})$ in Ω is created as follow

$$\widehat{u}(\mathbf{x}) = \sum_{j=1}^{N_p} w_j(\mathbf{x})\widehat{u}_j(\mathbf{x}), \tag{5}$$

where $\widehat{u}_j(\mathbf{x})$ are the local approximants on the patches $\Omega_j, j = 1, \ldots, N_p$. The weight functions $\{w_j(\mathbf{x})\}_{j=1}^{N_p}$ are non-negative and compactly supported on $\Omega_j, j = 1, \ldots, N_p$ that satisfy $\sum_{j=1}^{N_p} w_j(\mathbf{x}) = 1$ for $\mathbf{x} \in \Omega$. Also, the weights must be sufficiently smooth according to the differential operators of the problem to be solved. We use the Shepard's method applied to the compactly supported C^2 Wendland function to construct the weight functions, i.e.,

$$w_j(\mathbf{x}) = \frac{\varphi_j(\mathbf{x})}{\sum_{i=1}^{N_p} \varphi_i(\mathbf{x})}, \quad j = 1, \ldots, N_p,$$

where the function $\varphi(r)$ need to be mapped in Ω_j [2]. Let ϕ be a RBF and $\left\{\mathbf{x}_k^j\right\}_{k=1}^{N_j}$ be the local set of nodes in patch Ω_j. The local approximant $\widehat{u}_j(\mathbf{x})$ for every patch Ω_j is defined as follow

$$\widehat{u}_j(\mathbf{x}) = \sum_{k=1}^{N_j} \alpha_k^j \phi(\|\mathbf{x} - \mathbf{x}_k^j\|), \tag{6}$$

where $\left\{\alpha_k^j\right\}_{k=1}^{N_j}$ denotes the unknowns. By evaluating (6) at $\left\{\mathbf{x}_k^j\right\}_{k=1}^{N_j}$ we have

$$\widehat{\mathbf{U}}_j = A^j \boldsymbol{\alpha}^j, \tag{7}$$

in which $\widehat{\mathbf{U}}_j = [\widehat{u}_j(\mathbf{x}_1^j), \widehat{u}_j(\mathbf{x}_2^j), \ldots, \widehat{u}_j(\mathbf{x}_{N_j}^j)]^T$, $\boldsymbol{\alpha}^j = [\alpha_1^j, \alpha_2^j, \ldots, \alpha_{N_j}^j]^T$, and $(A^j)_{i,k} = \phi(\|\mathbf{x}_i^j - \mathbf{x}_k^j\|),\ i,k = 1,\ldots,N_j$. In this work, the Gaussian RBF is used to guarantee the invertibility of A^j for distinct nodes. Using (7), the local approximant $\widehat{u}_j$ (6) is written as

$$\widehat{u}_j(\mathbf{x}) = \Phi^j(\mathbf{x})\boldsymbol{\alpha}^j = \Phi^j(\mathbf{x})(A^j)^{-1}\widehat{\mathbf{U}}_j, \tag{8}$$

where $\Phi^j(\mathbf{x}) = [\phi(\|\mathbf{x} - \mathbf{x}_1^j\|), \phi(\|\mathbf{x} - \mathbf{x}_2^j\|), \ldots, \phi(\|\mathbf{x} - \mathbf{x}_{N_j}^j\|)]$. Now, the local differentiation matrices are computed in order to obtain derivatives of $\widehat{u}_j(\mathbf{x})$ in patch Ω_j. To do this, we apply the differential operator $\mathcal{L}$ to (8) and evaluate it at $\left\{\mathbf{x}_k^j\right\}_{k=1}^{N_j}$. For each patch Ω_j, the differentiation matrix for operator $\mathcal{L}$ is obtained as

$$A_{\mathcal{L}}^j = \boldsymbol{\Phi}_{\mathcal{L}}^j (A^j)^{-1}, \tag{9}$$

where $(\boldsymbol{\Phi}_{\mathcal{L}}^j)_{i,k} = \mathcal{L}\phi(\|\mathbf{x}_i{}^j - \mathbf{x}_k^j\|),\ i,k = 1,\ldots,N_j$. By applying operator $\mathcal{L}$ to (5) and using (8), we have

$$\mathcal{L}\widehat{u}(\mathbf{x}) = \sum_{j=1}^{N_p} \mathcal{L}(w_j(\mathbf{x})\Phi^j(\mathbf{x}))(A^j)^{-1}\widehat{\mathbf{U}}_j.$$

Consider the global set of nodes as $\{\mathbf{x}_k\}_{k=1}^N$. To evaluate $\mathcal{L}\widehat{u}(\mathbf{x})$ at these nodes, i.e., $\mathcal{L}\widehat{\mathbf{U}} = [\mathcal{L}\widehat{u}(\mathbf{x}_1), \mathcal{L}\widehat{u}(\mathbf{x}_2), \ldots, \mathcal{L}\widehat{u}(\mathbf{x}_N)]^T$, we compute $\mathcal{L}(w_j(\mathbf{x})\Phi^j(\mathbf{x}))(A^j)^{-1}$ in patch Ω_j. We apply the product derivative rule and Eq. (9) to get the local matrices as follow

$$\begin{aligned} D_{\mathcal{L}_1}^j &= W^j A_{\mathcal{L}_1}^j + W_{\mathcal{L}_1}^j I; & \mathcal{L}_1 &= \tfrac{\partial}{\partial x}, \\ D_{\mathcal{L}_2}^j &= W^j A_{\mathcal{L}_2}^j + 2W_{\mathcal{L}_1}^j A_{\mathcal{L}_1}^j + W_{\mathcal{L}_2}^j I; & \mathcal{L}_2 &= \tfrac{\partial^2}{\partial x^2}, \end{aligned}$$

where $W^j = \mathrm{diag}(w_j(\mathbf{x}_1^j), \ldots, w_j(\mathbf{x}_{N_j}^j)), W_{\mathcal{L}}^j = \mathrm{diag}(\mathcal{L}w_j(\mathbf{x}_1^j), \ldots, \mathcal{L}w_j(\mathbf{x}_{N_j}^j))$, and I is an $N_j \times N_j$ identity matrix. The global differentiation matrix $D_{\mathcal{L}}$ is a sparse matrix assembled by the computed local matrices. This sparsity is important to make a reduction in the computer time and memory usage for large-scale problems [2]. The RBF-QR algorithm is performed by a change from the Gaussian basis to a more stable one. This new basis is a finite expansion of nearly flat RBFs, that is truncated at i_{max}; for details on the selected i_{max} see [5]. Now, the calculation of RBF-QR basis ψ at $\mathbf{x} = (r,\theta)$ is breifly expressed. First, the matrix C is defined which is of size $N \times M$, $M = \frac{(i_{\max}+1)(i_{\max}+2)}{2}$ and k-th row of this matrix is as follow

$$[c_{0,0}(\mathbf{x}_k), c_{1,0}(\mathbf{x}_k), s_{1,0}(\mathbf{x}_k), c_{2,0}(\mathbf{x}_k), c_{2,1}(\mathbf{x}_k), s_{2,1}(\mathbf{x}_k), \ldots, s_{i_{\max},m}(\mathbf{x}_k)]\,,$$

where $c_{i,j}(\mathbf{x}_k) = \upsilon(r_k)\cos(\Theta_k)$, $s_{i,j}(\mathbf{x}_k) = \upsilon(r_k)\sin(\Theta_k)$ and

$$\upsilon(r) = b_{2j+m}t_{i-2j}e^{-\varepsilon^2 r^2}\, r^i\ {}_1F_2(\zeta_{i,j}, \eta_{i,j}, \varepsilon^4 r^2), \quad \Theta = (2j+m)\theta,$$

$$m=\begin{cases}0, & i=2k,\\ 1, & i=2k+1,\end{cases}\qquad t_i=\begin{cases}\frac{1}{2}, & i=0,\\ 1, & i>0,\end{cases}\qquad b_j=\begin{cases}1, & j=0,\\ 2, & j>0.\end{cases}$$

The arguments of function ${}_1F_2$ are $\zeta_{i,j}=\frac{i-2j+m+1}{2}$ and $\eta_{i,j}=[i-2j+1,\frac{i+2j+m+2}{2}]$.

We also have an $M\times M$ matrix as

$$E=\mathrm{diag}(E_{0,0},E_{1,0},E_{1,0},E_{2,0},E_{2,1},E_{2,1},\ldots,E_{i_{\max},m}),$$

with entries

$$E_{i,j}=\frac{\varepsilon^{2i}}{2^{i-2j-1}(\frac{i+2j+m}{2})!(\frac{i-2j-m}{2})!}.$$

We QR-factorize matrix C as $C=Q[R_1\quad R_2]$, where Q is an unitary matrix, R_1 is an $N\times N$ upper triangular matrix. We define $\tilde{R}=E_1^{-1}R_1^{-1}R_2E_2$, where E_1 and E_2 are diagonal blocks of sizes $N\times N$ and $(M-N)\times(M-N)$, respectively. Then, we compute an $M\times 1$ vector as

$$P(\mathbf{x})=\left[P^c_{0,0}(\mathbf{x}),P^c_{1,0}(\mathbf{x}),P^s_{1,0}(\mathbf{x}),P^c_{2,0}(\mathbf{x}),P^c_{2,1}(\mathbf{x}),P^s_{2,1}(\mathbf{x}),\ldots,P^s_{i_{\max},m}(\mathbf{x})\right]^T,$$

where for $i=0,\ldots,i_{\max}$, we have

$$\begin{cases}P^c_{i,j}(\mathbf{x})=\omega(r)\cos(\Theta), & j=0,\ldots,\frac{i-m}{2},\\ P^s_{i,j}(\mathbf{x})=\omega(r)\sin(\Theta), & j=1-m,\ldots,\frac{i-m}{2},\quad \Theta\neq 0,\end{cases}$$

and $\omega(r)=e^{-\varepsilon^2r^2}r^{2j}P_{i-2j}(r)$, where $P_l(r)$ is the l-th order Chebyshev polynomial of the first kind. The RBF-QR basis is finally calculated as

$$[\psi_1(\mathbf{x}),\psi_2(\mathbf{x}),\ldots,\psi_N(\mathbf{x})]^T=\begin{bmatrix} I_N & \tilde{R}\end{bmatrix}P(\mathbf{x}).$$

To solve the problem under consideration by using the proposed method, the Eqs. (1)-(4) are discretized at the corresponding collocation nodes and RBF-QR partition of unity approximation is used for all derivatives.

2. Numerical Results

Example 2.1. The parameter values are chosen to be $(\alpha^-,\alpha^+)=(\sin(2x-y)+3,1+e^{x+2y})$, $(\sigma^-,\sigma^+)=(x+y+3,2+\sin(x+y))$, $(\beta_1,\beta_2)=(x^2+y^2+1,\cos(xy)+2)$, $(\gamma_1,\gamma_2)=(x^2+y^2+1,e^{x+y})$. The exact solutions are given as

$$\begin{cases}u^-(x,y)=\sin(x+y),\\ u^+(x,y)=x^3-y^3,\end{cases}\qquad\begin{cases}v^-(x,y)=\cos(x+y),\\ v^+(x,y)=x^3+y^3.\end{cases}$$

We consider a complex interface geometry given as $r(\theta)=2+\frac{2}{5}\cos(\theta)+\frac{3}{10}\sin(6\theta)$, $\theta\in[0,2\pi]$ within $[-4,4]^2$. Here, the tuple (N^-,N^+,np^-,np^+) is defined, where $N^\pm$ is the number of Halton-type nodes in $\overline{\Omega^\pm}$, and $np^\pm$ is the patch numbers. In this example, the obtained results of the RBF partition of unity method (PUM) with RBF-QR partition of unity method (PUM-QR) are compared and the abilities of proposed method (PUM-QR) are shown in solving the coupled system (1)-(4) involving discontinous variable coefficients and complex interface. Figure 1(a), shows the problem domain and the corresponding patches for $(866,1905,26,52)$. Figure 1(b) displays the L_∞ absolute error versus ε for PUM and PUM-QR methods. We observe the PUM-QR method not only improves the

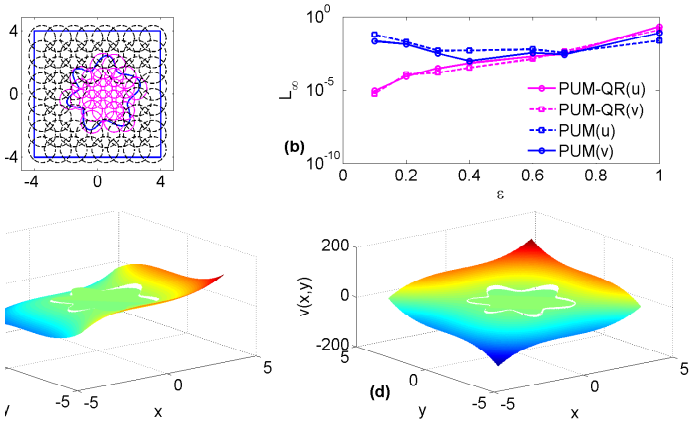


FIGURE 1. (a) The distribution of corresponding patches in domain, (b) the behavior of error as a function of ε, (c) the numerical solution for u (d) and for v using PUM-QR for $\varepsilon = 0.01$.

accuracy still but also implements in a numerically stable way. While a breakdown occurs in PUM error curve for small ε, the PUM-QR is able to overcome this instability. The graphs of numerical solutions are shown in Figures 1(c) and 1(d) using PUM-QR method for $\varepsilon = 0.01$ and $(866, 1905, 26, 52)$.

References

1. S. DeMarchi, A. Martínez, E. Perracchione and M. Rossini, *RBF-based partition of unity methods for elliptic PDEs: Adaptivity and stability issues via variably scaled kernels*, J. Sci. Comput. **79** (2019) 321–344.
2. A. Safdari-Vaighani, A. Heryudono and E. Larsson, *A radial basis function partition of unity collocation method for convection-diffusion equations arising in financial applications*, J. Sci. Comput., **64** (2015) 341–367.
3. E. Larsson, V. Shcherbakov and A. Heryudono, *A least square radial basis function partition of unity methods for solving PDEs*, SIAM J. Sci. Comput. **39** (6) (2017) A2538–A2563.
4. A. Heryudono, E. Larsson, A. Ramage and L. von Sydow, *Preconditioning for radial basis function partition of unity methods*, J. Sci. Comput. **67** (3) (2016) 1089–1109.
5. E. Larsson, E. Lehto, A. Heryudono and B. Fornberg, *Stable computation of differentiation matrices and scattered node stencils based on Gaussian radial basis functions*, SIAM J. Sci. Comput. **35** (4) (2013) A2096–A2119.
6. F. Gholampour, E. Hesameddini and A. Taleei, *A stable RBF partition of unity local method for elliptic interface problems in two dimensions*, Eng. Anal. Bound. Elem. **123** (2021) 220–232.

E-mail: F.Gholampour@sutech.ac.ir
E-mail: hesameddini@sutech.ac.ir
E-mail: a.taleei@sutech.ac.ir

The Three-Term Recurrence Variant of the Conjugate Gradient Squared Method to Solve the Non-Symmetric Linear System $Ax = b$

Eisa Khosravi Dehdezi*
Department of Mathematics, Persian Gulf University, Bushehr, Iran

Abstract. Here, the three-term recurrence variant of the conjugate gradient squared method to solve the non-symmetric linear system $Ax = b$ is obtained. Numerical experiments are provided to show the superiority of the new method with the common CGS method.
Keywords: CGS, CGS-TTRV, Iterative methods.
AMS Mathematical Subject Classification [2010]: 65F22, 65F25, 65L80.

1. Introduction

As we know, the common CGS method is one of the best methods to solve the non-symmetric linear system

$$Ax = b, \quad A \in \mathbb{R}^{n\times n},\ x \in \mathbb{R}^n. \tag{1}$$

The bi-conjugate gradient or Bi-CG method is an iterative method which can be apply to solve the large sparse non-symmetric linear system $Ax = b$. This method based on the non-symmetric Lanczos procedure [1]. To remove the transpose of A (A^T) in the Bi-CG method and to gain faster convergence, the common conjugate gradient squared algorithm was developed by Sonneveld, where in exact arithmetic, terminates with true solution after $j \leq n$ steps [2]. The common CGS method to solve (1) is expressed in Algorithm 1.

Algorithm 1. Common CGS algorithm for solving (1)
Set $r_0 = b - Ax_0$ and choose $\tilde{r}_0$ so that $\rho_0 = \tilde{r}_0^T r_0 \neq 0$.
$p_0 = u_0 = r_0;\ v_0 = Ap_0;$
for $j = 0, 1, 2, \ldots$
$v_j = Ap_j;\ \sigma_j = \tilde{r}_0^T v_j;$
$\alpha_j = \rho_j/\sigma_j;$
$q_j = u_j - \alpha_j Ap_j;$
$x_{j+1} = x_j + \alpha_j(u_j + q_j);$
$r_{j+1} = r_j - \alpha_j A(u_j + q_j);$
$\rho_{j+1} = \tilde{r}_0^T r_{j+1};\ \beta_j = \rho_{j+1}/\rho_j;$
$u_{j+1} = r_{j+1} + \beta_j q_j;$
$p_{j+1} = r_{j+1} + \beta_j(q_j + \beta_j p_j);$

*Presenter

end

Here, we obtain a new variant of the common CGS method by using the Lanczos algorithm and the recurrence property of the orthogonal vectors.

The rest of the paper is as follows. In the next section, the three-term recurrence variant of the conjugate gradient squared algorithm for solving (1) is obtained and in the final section, we show the comparative results.

2. Three-Term Recurrence Variant of Conjugate Gradient Squared Algorithm

By using the recurrence property of the orthogonal vectors, the residual polynomials $r_j(t)$ and $\tilde{r}_j(t)$ associated with the j-th Bi-CG iterative method should satisfy a three-term recurrence. So we consider a three-term recurrence of the form

$$r_{j+1}(t) = \rho_j(r_j(t) - \gamma_j t r_j(t)) + \delta_j r_{j-1}(t), \tag{2}$$

and

$$\hat{r}_{j+1}(t) = \hat{\rho}_j(\tilde{r}_j(t) - \tilde{\gamma}_j t \tilde{r}_j(t)) + \tilde{\delta}_j \tilde{r}_{j-1}(t). \tag{3}$$

Also, using the consistency conditions $r_j(0) = 1$ and $\tilde{r}_j(0) = 1$ for every j, we have

$$r_{j+1}(t) = \rho_j(r_j(t) - \gamma_j t r_j(t)) + (1 - \rho_j) r_{j-1}(t), \tag{4}$$

and

$$\tilde{r}_{j+1}(t) = \tilde{\rho}_j(\tilde{r}_j(t) - \tilde{\gamma}_j t \tilde{r}_j(t)) + (1 - \tilde{\rho}_j) \tilde{r}_{j-1}(t). \tag{5}$$

If $r_j(0) = 1,\ \tilde{r}_j(0) = 1$ and $r_{j-1}(0) = 1,\ \tilde{r}_{j-1}(0) = 1$, then $r_{j+1}(0) = 1,\ \tilde{r}_{j+1}(0) = 1$. Applying the Eqs. (4) and (5) for the sequence of residual vectors,

$$r_{j+1} = \rho_j(r_j - \gamma_j A r_j) + (1 - \rho_j) r_{j-1}, \tag{6}$$

and

$$\tilde{r}_{j+1} = \tilde{\rho}_j(\tilde{r}_j - \tilde{\gamma}_j A^T \tilde{r}_j) + (1 - \tilde{\rho}_j) \tilde{r}_{j-1}. \tag{7}$$

Using inherent property of Bi-CG method which implies that $< r_i, \tilde{r}_j >= 0,\ i \neq j$, Eqs. (6) and (7) we have

$$< r_{j+1}, \tilde{r}_j >= \rho_j(< r_j, \tilde{r}_j > - \gamma_j < A r_j, \tilde{r}_j >) + (1 - \rho_j) < r_{j-1}, \tilde{r}_j >= 0,$$

thus

$$\gamma_j =< r_j, \tilde{r}_j > / < A r_j, \tilde{r}_j >,\ j = 0, 1, \ldots.$$

Similarly

$$\tilde{\gamma}_j =< r_j, \tilde{r}_j > / < A r_j, \tilde{r}_j >= \gamma_j,\ j = 0, 1, \ldots.$$

For computing ρ_j and $\tilde{\rho}_j$

$$< r_{j+1}, \tilde{r}_{j-1} >= \rho_j(< r_j, \tilde{r}_{j-1} > - \gamma_j < A r_j, \tilde{r}_{j-1} >) + (1 - \rho_j) < r_{j-1}, \tilde{r}_{j-1} >= 0,$$

thus

$$-\rho_j \gamma_j < A r_j, \tilde{r}_{j-1} > + (1 - \rho_j) < r_{j-1}, \tilde{r}_{j-1} >= 0,$$

or equivalently

$$-\rho_j \gamma_j < r_j, A^T \tilde{r}_{j-1} > + (1 - \rho_j) < r_{j-1}, \tilde{r}_{j-1} >= 0. \tag{8}$$

Calculating $< r_j, A^T\tilde{r}_{j-1} >$ from Eq. (7), we have

$$< r_j, A^T\tilde{r}_{j-1} >= -\frac{1}{\tilde{\rho}_{j-1}\gamma_{j-1}} < r_j, \tilde{r}_j > . \tag{9}$$

Substituting Eq. (9) in Eq. (8) and after some calculating

$$\rho_j = (1 - \frac{\gamma_j}{\gamma_{j-1}} \frac{< r_j, \tilde{r}_j >}{< r_{j-1}, \tilde{r}_{j-1} >} \frac{1}{\rho^*_{j-1}})^{-1},$$

and similarly

$$\tilde{\rho}_j = (1 - \frac{\gamma_j}{\gamma_{j-1}} \frac{< r_j, \tilde{r}_j >}{< r_{j-1}, \tilde{r}_{j-1} >} \frac{1}{\rho_{j-1}})^{-1}.$$

Like CGS algorithm, we want to omit the A^T in the Bi-CG and to gain faster convergent method. For this purpose we define the auxiliary vector

$$p_j(t) = \frac{r_j(t) - r_{j+1}(t)}{\rho_j \gamma_j t}, \tag{10}$$

so

$$r_{j+1} = r_j - \alpha_j A p_j, \quad \alpha_j = \rho_j \gamma_j. \tag{11}$$

Substituting r_{j+1} from Eq. (11) in Eq. (10) and after some calculating, we obtain

$$p_{j+1} = r_{j+1} + \beta_j p_j, \quad \beta_j = \frac{\alpha_j}{\alpha_{j+1}}(\rho_{j+1} - 1).$$

The residual vector at every step j of the Bi-CG method can be considered as

$$r_j = \phi_j(A) r_0,$$

where ϕ_j is a polynomial of degree j satisfying the condition $\phi_j(0) = 1$. Similarly, the conjugate-direction polynomial $\pi_j(t)$ can be expressed by

$$p_j = \pi_j(A) r_0,$$

where the ϕ_j is a polynomial of degree j. From the Bi-CG method, the directions $\tilde{r}_j$ and $\tilde{\phi}_j$ can be defined as the same recurrences as r_j and π_j, where A is replaced by A^T, that means

$$\tilde{r}_j = \phi_j(A^T)\tilde{r}_0, \ \tilde{p}_j = \pi_j(A^T)\tilde{r}_0.$$

So the scalar γ_j, ρ_j and $\tilde{\rho}_j$ are given by

$$\gamma_j = \frac{< r_j, \tilde{r}_j >}{< Ar_j, \tilde{r}_j >} = \frac{< \phi_j^2(A) r_0, \tilde{r}_0 >}{< A\phi_j^2(A) r_0, \tilde{r}_0 >},$$

$$\rho_j = (1 - \frac{\gamma_j}{\gamma_{j-1}} \frac{< r_j, \tilde{r}_j >}{< r_{j-1}, \tilde{r}_{j-1} >} \frac{1}{\tilde{\rho}_{j-1}})^{-1} = (1 - \frac{\gamma_j}{\gamma_{j-1}} \frac{< \phi_j^2(A) r_0, \tilde{r}_0 >}{< \phi_{j-1}^2(A) r_0, \tilde{r}_0 >} \frac{1}{\tilde{\rho}_{j-1}})^{-1},$$

$$\tilde{\rho}_j = (1 - \frac{\gamma_j}{\gamma_{j-1}} \frac{< r_j, \tilde{r}_j >}{< r_{j-1}, \tilde{r}_{j-1} >} \frac{1}{\rho_{j-1}})^{-1} = (1 - \frac{\gamma_j}{\gamma_{j-1}} \frac{< \phi_j^2(A) r_0, \tilde{r}_0 >}{< \phi_{j-1}^2(A) r_0, \tilde{r}_0 >} \frac{1}{\rho_{j-1}})^{-1},$$

which shows that if we get a recursion for the vectors $\phi_j^2(A) r_0$, then computing γ_j, ρ_j and $\tilde{\rho}_j$ are not difficult.

The derivation of the method is as follows. Very similarity to the obtaining CGS method (See [1], pages 681-683) and due to existing restrictions, we ignore details. The resulting algorithm is named by CGS - Three Term Recurrence Variant and is given follows.

Algorithm 2. CGS-Three Term Recurrence Variant (CGS-TTRV) for solving (1).
Set $r_0 = b - Ax_0$ and choose $\tilde{r}_0$ so that $t_0 = \tilde{r}_0^T r_0 \neq 0$.
$p_0 = u_0 = r_0;\ \rho_0 = \tilde{\rho}_0 = 1;\ v_0 = Ar_0;\ \sigma_0 = \tilde{r}_0^T v_0;\ \gamma_0 = t_0/\sigma_0;$
for $j = 0, 1, 2, \ldots,$
$v_j = Ar_j;\ \sigma_j = \tilde{r}_0^T v_j;$
$\alpha_j = \rho_j/\sigma_j;$
$q_j = u_j - \alpha_j A p_j;$
$x_{j+1} = x_j + \alpha_j(u_j + q_j);$
$r_{j+1} = r_j - \alpha_j A(u_j + q_j);$
$t_{j+1} = \tilde{r}_0^T r_{j+1};\ v_{j+1} = Ar_{j+1};\ \sigma_{j+1} = \tilde{r}_0^T v_{j+1};$
$\gamma_{j+1} = t_{j+1}/\sigma_{j+1};$
$\rho_{j+1} = (1 - \frac{t_{j+1}\gamma_{j+1}}{t_j\gamma_j}\frac{1}{\tilde{\rho}_j})^{-1};$
$\tilde{\rho}_{j+1} = (1 - \frac{t_{j+1}\gamma_{j+1}}{t_j\gamma_j}\frac{1}{\rho_j})^{-1};$
$\alpha_{j+1} = \rho_{j+1}\gamma_{j+1};$
$\beta_j = \frac{\alpha_j}{\alpha_{j+1}}(\rho_{j+1} - 1);$
$u_{j+1} = r_{j+1} + \beta_j q_j;$
$p_{j+1} = r_{j+1} + \beta_j(q_j + \beta_j p_j);$
end

The new method occupies the same storage as the common formulation because of no more matrix-vector product than that.

TABLE 1. The computational time(s) and the number of iterations.

Matrix name	Cond	nnz	CGS		CGS-TTRV	
			Iter	Time	Iter	Time
MHD416A	4.1600e+02	416	58	0.0135	58	0.0131
ADD20	1.7637e+04	13151	289	1.3188	276	1.8673
IBM32	3.4242e+07	935	1	0.0023	1	0.0022
LOP163	4.3782e+4	224	34	0.0117	1	0.0034
ABB313	1.5439e+0	313	8	0.0101	8	0.0110
ADD32	2.1363e+02	19848	57	1.0535	57	1.5407
CK656	1.1802e+07	3884	353	0.0774	353	0.0982
JGL009	1.9267e+50	50	6	0.0104	6	0.0091
TOLS90	2.4913e+04	1746	101	0.0108	106	0.0101
BFW782A	4.6250e+03	7514	292	0.1413	292	0.1927
PLAT1919	4.9059e+02	17159	118	0.3418	119	0.5170
IMPCOL-B	2.6670e+05	271	402	0.0168	391	0.0101

3. Numerical Examples

Consider the linear system $Ax = b$, where the coefficient matrix A chosen form [3]. We considered the zero vector as an initial vector and for simplicity, the vector b is chosen such that $x = (1, 1, \ldots, 1)^T$ is the exact solution. The stopping criterion $\|r_j\|_2 < 10^{-10}$ is used, where $r_j = b - Ax_j$ is the j-th residual. Table 1 presents CPU Times in seconds (Time), the iteration numbers (Iter), the number of nonzero entries (nnz) and the condition number of the coefficient matrix (Cond).

References

1. Â. Björck, *Numerical Methods in Matrix Computations,* Springer, Cham, 2015.
2. P. Sonneveld, *CGS, a fast Lanczos-type solver for nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput. **10** (1) (1989) 36–52.
3. Matrix Market, `URL:http://math.nist.gov/MatrixMarket`, October 2002.

E-mail: esakhosravidehdezi@gmail.com

Simulation of Some Numerical Methods for RODEs Driven by Fractional Brownian Motion

Azar Mirzaei*
Department of Mathematics, Faculty of Science, Razi University, Kermanshah, Iran
and Minoo Kamrani
Department of Mathematics, Faculty of Science, Razi University, Kermanshah, Iran

Abstract. Similar to the deterministic calculus, most of the stochastic differential equations and random ordinary differential equations (RODEs) do not have explicit analytical solutions and numerical methods are important tools to investigate these equations. The aim of this paper is to investigate simulation of some numerical methods for RODEs which are derived by fractional brownian motions with Hurst Parameter H.
Keywords: Random ordinary differential equations, Fractional Brownian motion, Implicit methods.
AMS Mathematical Subject Classification [2010]: 60G22, 37H10, 65C30.

1. Introduction

Random ordinary differential equations (RODEs) have a stochastic process in their vector field functions and can be investigated pathwise as deterministic ODEs. They have been used in a wide range of applications such as biology, medicine, population dynamics and engineering [3] and play an important role in the theory of random dynamical systems, however, they have been long overshadowed by SDEs.

When the noise is regular noise, there is, in fact, a close connection between RODEs and SDEs. On the other hand, Doss and Sussmann proved that any finite dimensional SDE with commutative noise can be transformed to a RODE and it was later generalized to all SDEs [1]. In this paper we consider simulation of some numerical methods for RODEs which are derived by fractional brownian motion.

Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ be a filtered probability space, and $B^H = (B_t^H)_{t\geq 0}$ be a one-dimensional standard $(\mathcal{F}_t)$-adapted fractional Brownian motion with Hurst parameter $H \in (\frac{1}{2}, 1)$ i.e., B^H is a continuous centered Gaussian process with covariance function

$$\mathbb{E}(B_t^H B_s^H) = \frac{1}{2}(t^{2H} + s^{2H} - |t-s|^{2H}).$$

For $H = \frac{1}{2}$, B^H is a standard Brownian motion, while for $H \neq \frac{1}{2}$, it is neither a semi martingale nor a Markov process and it does not have independent increments. Moreover the increment of the process in an interval $[s, t]$ has a normal

*Presenter

distribution with zero mean and variance

$$\left(\mathbb{E}\left(B_t^H - B_s^H\right)^2\right)^{\frac{1}{2}} = |t-s|^H.$$

As a consequence, the process B^H has α-Hölder continuous paths for all $\alpha \in (0, H)$. If $H > 1/2$, then the process $(B_t^H, t \geq 0)$ exhibits a long-range dependence, that is, if

$$r(n) = \mathbb{E}(B_1^H(B_{n+1}^H - B_n^H)),$$

then $\sum_{n=1}^{\infty} r_n = \infty$.

A fractional Brownian motion is also self-similar, that is, $(B_{\alpha t}^H, t \geq 0)$ has the same probability law as $(\alpha^H B_t^H, t \geq 0)$ [2, 5]. A process satisfying this property is called a self-similar process with the Hurst parameter H. Since in many problems related to network traffic analysis, mathematical finance, and many other fields the processes under study seem empirically to exhibit the selfsimilar properties, and the long-range dependent properties, and since the fractional Brownian motions are the simplest processes of this kind, it is important to have a systematic study of these processes and to use them to construct other stochastic processes.

The stochastic (Wiener) integral with respect to fractional Brownian motions for deterministic kernels is easily defined as follows.

LEMMA 1.1. [4] *Let $L^2(0,T)$ denotes the space of equivalence classes of measurable functions f such that*

$$\int_0^T \int_0^T f(s)f(t)|s-t|^{2H-2}dsdt < \infty.$$

If $f, g \in L^2(0,T)$ then

$$\mathbb{E}\Big(\int_0^T f(u)dB_u^H \int_0^T g(v)dB_v^H\Big) = H(2H-1)\int_0^T \int_0^T f(u)g(v)|u-v|^{2H-2}dudv.$$

2. Main Results

Suppose $T > 0$, $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ be a probability space. Consider the following random ordinary differential equation

$$\frac{dX(t)}{dt} = f\left(B_t^H, X(t)\right), \; X(0) = x_0. \tag{1}$$

Our aim is to obtain the numerical solution of the above equation by some implicit methods. Consider the following implicit averaged Euler scheme (IAES):

$$X_{n+1} = X_n + f\left(I_n, X_{n+1}\right)\Delta, \quad n = 0, 1, \ldots,$$

and the implicit averaged midpoint scheme (IAMS) given by

$$X_{n+1} = X_n + f\left(I_n, \frac{1}{2}(X_n + X_{n+1})\right)\Delta, \quad n = 0, 1, \ldots,$$

where

$$I_n = \frac{1}{\Delta}\int_{n\Delta}^{(n+1)\Delta} B_s^H(\omega)ds, \quad \omega \in \Omega. \tag{2}$$

Convergence of the above methods has been investigated in [1], for the case ζ_t is Hölder continuous, i.e, a real number $\theta \in (0,1)$ and a random variable $\Theta : \Omega \to [0,\infty)$ are available, such that

$$\|\zeta_t(\omega) - \zeta_s(\omega)\| \leqslant \Theta(\omega)|t-s|^\theta, \quad \omega \in \Omega.$$

Since fractional brownian motion is continuous stochastic process with Hölder continuous sample paths, therefore it satisfies the necessary condition of the stochastic process for the convergence of the above methods. Therefore we can consider the above numerical methods for the solution of (1). But because the increments of the fractional Brownian motions need not be independent, simulation of I_n it not so easy.

Our aim is to explain simulation of the proposed methods with the stochastic process defined in (2). For this aim we need to obtain the distribution of I_n. Trivially

$$\mathbb{E}(I_n) = 0,$$

and we have

$$I_n = \frac{1}{\Delta}\int_0^\Delta B_s^H(\omega)ds = \frac{1}{\Delta}\left(B_\Delta^H(\omega)\Delta - \int_0^\Delta s dB_s^H(\omega)\right).$$

Let

$$P_n = \int_0^\Delta s dB_s^H(\omega).$$

Therefore by Lemma 1.1

$$\begin{aligned}
\mathbb{E}(P_n^2) &= \mathbb{E}\left\{\left(\int_0^\Delta s dB_s^H\right)\left(\int_0^\Delta t dB_t^H\right)\right\} = \gamma_H\int_0^\Delta\int_0^\Delta st|s-t|^{2H-2}dsdt \\
&= \gamma_H\int_0^\Delta\int_0^s ts(s-t)^{2H-2}dtds + \gamma_H\int_0^\Delta\int_0^t st(t-s)^{2H-2}dsdt \qquad (3)\\
&= \frac{\Delta^{2H+2}}{2H+2},
\end{aligned}$$

where $\gamma_H = H(2H-1)$. Also

$$\begin{aligned}
\mathbb{E}(B_\Delta^H(\omega)P_n) &= \mathbb{E}\left\{\left(\int_0^\Delta s dB_s^H\right)\left(\int_0^\Delta dB_t^H\right)\right\} = \gamma_H\int_0^\Delta\int_0^\Delta s|s-t|^{2H-2}dtds \\
&= \gamma_H\int_0^\Delta\int_0^s s(s-t)^{2H-2}dtds + \gamma_H\int_0^\Delta\int_0^t s(t-s)^{2H-2}dsdt \qquad (4)\\
&= \frac{1}{2}\Delta^{2H+1}.
\end{aligned}$$

Therefore by (3) and (4) we get

$$\begin{aligned}\mathbb{E}(I_n^2) &= \frac{1}{\Delta^2}\mathbb{E}\left(B_\Delta^H(\omega)\Delta - P_n\right)^2 \\ &= \frac{1}{\Delta^2}\left(\Delta^2\mathbb{E}(B_\Delta^H(\omega))^2 + \mathbb{E}(P_n^2) - 2\Delta\mathbb{E}(B_\Delta^H(\omega)P_n)\right) \\ &= \frac{1}{\Delta^2}\left(\Delta^{2H+2} + \frac{\Delta^{2H+2}}{2H+2} - \Delta^{2H+2}\right) \\ &= \frac{\Delta^{2H}}{2H+2}.\end{aligned}$$

Therefore $I_n, n = 0, \ldots, N-1$ are $\mathcal{N}(0, \frac{\Delta^{2H}}{2H+2})$-distributed random variables.

Also because the increments of the fractional brownian motions are not necessarily independent we should obtain $\mathbb{E}(I_n I_m)$, for $n, m = 0, \ldots, N-1$ which can be obtained similarly to the above calculations.

For simulating of $I_n, n = 0, \ldots, N-1$ let

$$V = (I_0, \ldots, I_{N-1}),$$

by Cholesky decomposition the covariance matrix $VV^\top$ can be written as $LL^\top$, where L is a $N \times N$ lower triangular matrix. It can be proved that such a decomposition exists since $VV^\top$ is a symmetric positive definite matrix. Afterwards, the samples of I_n for $n = 0, \ldots, N-1$ are generated by multiplying a vector with N independent and identically distributed standard normal components with a square lower triangular matrix L.

References

1. Y. Asai, *Numerical Methods for Random Ordinary Differential Equations and their Applications in Biology and Medicine*, Ph.D. Thesis, Dissertation, Institute für Mathematik Goethe Universität Frankfurt am Main, 2016.
2. T. E. Duncan, Y. Hu and B. Pasic-Duncan, *Stochastic calculus for fractional Brownian motion*, SIAM J. Control Optim. **38** (2000) 582–612.
3. X. Han and P. E. Kloeden, *Random Ordinary Differential Equations and their Numerical Solution*, Springer Natur Singapore, 2017.
4. S. Huang and S. Cambanis, *Stochastic and multiple Wiener integrals for Gaussian processes*, Ann. Probab. **6** (1978) 585–614.
5. D. Nualart, *The Malliavin Calculus and Related Topics*, Springer-Verlag Berlin Heidelberg, 2006.

E-mail: a.mirzaei9495@yahoo.com
E-mail: m.kamrani@razi.ac.ir

The Local Meshless Collocaion Method for Solving 2D Fractional Klein-Kramers Dynamics Equation on Irregular Domains

Hosein Pourbashash*
Faculty of Engineering, University of Garmsar, Garmsar, Iran
and Mahmood Khaksar-e Oshagh
Mosaheb Institute of Mathematics, Kharazmi University, Tehran, Iran

Abstract. Here, we propose a local meshless collocation method to solve two-dimensional (2D) Klein-Kramers equation with a fractional derivative in the Riemann-Liouville sense, in the time term. The radial basis function-differential quadrature method (RBF-DQ) has been employed to estimate the spatial directions. To discrete the time-variable, we employ two different strategies with convergence orders $\mathcal{O}(\tau^{1+\alpha})$ and $\mathcal{O}(\tau^{2-\alpha})$ for $0 < \alpha < 1$.
Keywords: Fractional Klein-Kramers, RBF-differential quadrature method, Local meshless collocation method, Riemann-Liouville fractional derivatives.
AMS Mathematical Subject Classification [2010]: 26A33, 34K37, 35R11.

1. Introduction

1.1. Considered Equation. The time fractional Klein-Kramers equation can be derived based on the generalized Chapman-Kolmogorov equation for a Markovian process as follows [3]

$$
\begin{aligned}
&\frac{\partial v}{\partial t} = {}_0D_t^{1-\alpha}\left[-\gamma y\frac{\partial}{\partial x} + \gamma\frac{\partial}{\partial y}\left(\eta y - \frac{F(x)}{m}\right) + \frac{\gamma\eta}{m\beta}\frac{\partial^2}{\partial y^2}\right] v + f, \\
&0 < x,\ y < L,\ \ 0 < t \le T, \\
&v(x,y,t) = \psi(x,y), \qquad (x,y) \in \partial\Omega, \qquad 0 < t < T, \\
&v(x,y,0) = \omega(x,y), \qquad 0 < x,\ y < L,
\end{aligned} \tag{1}
$$

where $0 < \alpha < 1$. The Rieman-Liouville fractional partial derivative of order $1-\alpha$ is defined by

$${}_0D_t^{1-\alpha}u(x,y,t) = \frac{1}{\Gamma(\alpha)}\frac{\partial}{\partial t}\int_0^t \frac{u(x,y,\eta)}{(t-\eta)^{1-\alpha}}d\eta.$$

2. Main Result

2.1. The Meshless Local RBF-DQ (LRBF-DQ) Technique. The local RBF-DQ method is a meshfree RBF collocation base method. In this method the unknown function v can be approximated by direct use of the RBFs and the m-th derivative can be calculated by applying the differential quadrature (DQ)

*Presenter

technique. Let the computational domain $\Omega \subset \mathbb{R}^2$. We consider a set of distinct points such as $\{(x_1^i, y_1^i), (x_2^i, y_2^i), \ldots, (x_{n_i}^i, y_{n_i}^i)\}$ in each support domain related to $(x_i, y_i) \in \Omega$ with n_i nodes. In the DQ method, the m-th derivative at a reference point can be approximate based on the linear combination of function values at all nodes into its support domain. Here, the weight coefficients can be determined by the following smooth function

$$\left.\frac{\partial^m v(x,y)}{\partial x^m}\right|_{(x,y)=(x_i,y_i)} = \sum_{j=0}^{n_i} w_{i,j}^{m,x} v(x_j^i, y_j^i), \quad i = 0, 1, 2, \ldots, N.$$

Since in this method RBFs are used as a basis functions, we have

$$\left.\frac{\partial^m \phi_j(x,y)}{\partial x^m}\right|_{(x,y)=(x_i,y_i)} = \sum_{k=0}^{n_i} w_{i,k}^{m,x} \phi_j(x_k^i, y_k^i), \quad i, j = 0, 1, 2, \ldots, n_i, \quad i \neq j,$$

By collocating the nodes $\{(x_1^i, y_1^i), (x_2^i, y_2^i), \ldots, (x_{n_i}^i, y_{n_i}^i)\}$, we arrive at the following system

$$\underbrace{\begin{bmatrix} \frac{\partial^m \phi_0(x_i,y_i)}{\partial x^m} \\ \frac{\partial^m \phi_1(x_i,y_i)}{\partial x^m} \\ \vdots \\ \frac{\partial^m \phi_{n_i}(x_i,y_i)}{\partial x^m} \end{bmatrix}}_{\left[\frac{\partial^m \phi(x_i,y_i)}{\partial x^m}\right]} = \underbrace{\begin{bmatrix} \phi_0(x_0^i, y_0^i) & \phi_0(x_1^i, y_1^i) & \ldots & \phi_0(x_{n_i}^i, y_{n_i}^i) \\ \phi_1(x_0^i, y_0^i) & \varphi_1(x_1^i, y_1^i) & \ldots & \phi_1(x_{n_i}^i, y_{n_i}^i) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{n_i}(x_0^i, y_0^i) & \phi_{n_i}(x_1^i, y_1^i) & \ldots & \phi_{n_i}(x_{n_i}^i, y_{n_i}^i) \end{bmatrix}}_{[A]} \underbrace{\begin{bmatrix} w_{i,0}^{(m,x)} \\ w_{i,1}^{(m,x)} \\ \vdots \\ w_{i,n_i}^{(m,x)} \end{bmatrix}}_{[w^x]},$$

$$\underbrace{\begin{bmatrix} \frac{\partial^m \phi_0(x_i,y_i)}{\partial y^m} \\ \frac{\partial^m \phi_1(x_i,y_i)}{\partial y^m} \\ \vdots \\ \frac{\partial^m \phi_{n_i}(x_i,y_i)}{\partial y^m} \end{bmatrix}}_{\left[\frac{\partial^m \phi(x_i,y_i)}{\partial y^m}\right]} = \underbrace{\begin{bmatrix} \phi_0(x_0^i, y_0^i) & \phi_0(x_1^i, y_1^i) & \ldots & \phi_0(x_{n_i}^i, y_{n_i}^i) \\ \phi_1(x_0^i, y_0^i) & \varphi_1(x_1^i, y_1^i) & \ldots & \phi_1(x_{n_i}^i, y_{n_i}^i) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{n_i}(x_0^i, y_0^i) & \phi_{n_i}(x_1^i, y_1^i) & \ldots & \phi_{n_i}(x_{n_i}^i, y_{n_i}^i) \end{bmatrix}}_{[A]} \underbrace{\begin{bmatrix} w_{i,0}^{(m,y)} \\ w_{i,1}^{(m,y)} \\ \vdots \\ w_{i,n_i}^{(m,y)} \end{bmatrix}}_{[w^y]},$$

we obtain

$$[w^{x,m}] = [A]^{-1}[\frac{\partial^m \varphi(x_i, y_i)}{\partial x^m}],$$

$$[w^{y,m}] = [A]^{-1}[\frac{\partial^m \varphi(x_i, y_i)}{\partial y^m}].$$

2.2. Time Discretization Schemes. Let $t_k = k\tau$, for $k = 0, 1, \ldots, N$, where $\tau = T/N$ is the time step. By integrating from Eq. (1) on the $[t_k, t_{k+1}]$, we

have

$$v(x,y,t_{k+1}) - v(x,y,t_k) = \int_{t_k}^{t_{k+1}} f(x,y,s)ds$$
$$+I_{0+}^{\alpha}\left[-\gamma y\frac{\partial}{\partial x}+\gamma\frac{\partial}{\partial y}\left(\eta y-\frac{F(x)}{m}\right)+\frac{\gamma\eta}{m\beta}\frac{\partial^2}{\partial y^2}\right]v(x,y,t_{k+1})$$
$$-I_{0+}^{\alpha}\left[-\gamma y\frac{\partial}{\partial x}+\gamma\frac{\partial}{\partial y}\left(\eta y-\frac{F(x)}{m}\right)+\frac{\gamma\eta}{m\beta}\frac{\partial^2}{\partial y^2}\right]v(x,y,t_k).$$

Now, for $0\le k\le N-1$, we have

$$(1-\eta\mu_1)\,v^{k+1}+\varpi_1 y\frac{\partial v^{k+1}}{\partial x}-\varpi_1\left(\eta y-\frac{F(x)}{m}\right)\frac{\partial v^{k+1}}{\partial y}-\varpi_2\frac{\partial^2 v^{k+1}}{\partial y^2}$$
$$=v^k+\sum_{j=0}^{k-1}(\lambda_{j+1}-\lambda_j)\{-y\varpi_1\frac{\partial v^{k-j}}{\partial x}+\eta\varpi_1 v^{k-j}$$
$$+\varpi_1(\eta y-\frac{F(x)}{m})\frac{\partial v^{k-j}}{\partial y}+\varpi_2\frac{\partial^2 v^{k-j}}{\partial y^2}\}+\tau f^{k+1}+\mathcal{E}^{\alpha},$$

in which $|\mathcal{E}^{\alpha}|<C\tau^{1+\alpha}$ and $\varpi_1=\frac{\gamma\tau^{\alpha}}{\Gamma(2-\alpha)}$, $\varpi_2=\frac{\gamma\eta}{m\beta}\frac{\tau^{\alpha}}{\Gamma(2-\alpha)}$. Removing the small term $\mathcal{E}^{\alpha}$, yields

$$(2)\qquad (1-\varpi_1\eta)\,V^{k+1}+\varpi_1 y\frac{\partial V^{k+1}}{\partial x}-\varpi_1\left(\eta y-\frac{F(x)}{m}\right)\frac{\partial V^{k+1}}{\partial y}-\varpi_2\frac{\partial^2 V^{k+1}}{\partial y^2}$$
$$=V^k+\tau f^{k+1}+\sum_{j=0}^{k-1}(\lambda_{j+1}-\lambda_j)\{-\varpi_1 y\frac{\partial V^{k-j}}{\partial x}+\eta\varpi_1 V^{k-j}$$
$$+\varpi_1\left(\eta y-\frac{F(x)}{m}\right)\frac{\partial V^{k-j}}{\partial y}+\varpi_2\frac{\partial^2 V^{k-j}}{\partial y^2}\},$$

where $0\le k\le N-1$. The convergence order of the presented first time-discrete scheme (FTDS), in Eq. (2) is $\mathcal{O}(\tau^{1+\alpha})$ in time variable [4].

Now, by multiplying both sides of (1) by the fractional Riemann-Liouville integral operator ${}_0D_t^{\alpha-1}$, and using the properties of fractional operators [2], yield

$$(3)\qquad {}_0^cD_t^{\alpha}v(x,y,t)=\gamma\left[-y\frac{\partial}{\partial x}+\frac{\partial}{\partial y}\left(\eta y-\frac{F(x)}{m}\right)+\frac{\eta}{m\beta}\frac{\partial^2}{\partial y^2}\right]v(x,y,t)+H(x,y,t),$$

in which ${}_0D_t^{\alpha-1}(f(x,y,t))=H(x,y,t)$, Also, ${}_0D_t^{\alpha}$ is the Caputo fractional operator is as follows

$$\frac{\partial^{\alpha}u(x,y,t)}{\partial t^{\alpha}} = \frac{1}{\Gamma(1-\alpha)}\int_0^t\frac{\partial u(x,y,s)}{\partial s}\frac{ds}{(t-s)^{\alpha}}.$$

LEMMA 2.1. *Let $0<\beta<1$ and $u(t)\in C^2[0,t_k]$. Then*

$$\left|\frac{1}{\Gamma(1-\beta)}\int_0^{t_k}\frac{u'(t)}{(t_k-t)^{\beta}}dt-\frac{\tau^{-\beta}}{\Gamma(2-\beta)}\left[b_0u(t_k)-\sum_{m=1}^{k-1}(b_{k-m-1}-b_{k-m})u(t_m)-b_{k-1}u(t_0)\right]\right|$$
$$\le\frac{1}{\Gamma(2-\beta)}\left[\frac{1-\beta}{12}+\frac{2^{2-\beta}}{2-\beta}-(1+2^{-\beta})\right]\max_{0\le t\le t_k}|u''(t)|\tau^{2-\beta},$$

where $b_m=(m+1)^{1-\beta}-m^{1-\beta}$.

Now, consider Eq. (3) at point (x, y, t_k)

$${}_0^c D_t^\alpha v(x,y,t_k) = \left[-\gamma y\frac{\partial}{\partial x} + \gamma\frac{\partial}{\partial y}\left(\eta y - \frac{F(x)}{m}\right) + \frac{\gamma\eta}{m\beta}\frac{\partial^2}{\partial y^2}\right] v(x,y,t_k) + G(x,y,t_k).$$

Employing Lemma 2.1, yields

$$\begin{aligned}&(\mu - \eta\gamma)v^k + \gamma y\frac{\partial v^k}{\partial x} - \gamma\left(\eta y - \frac{F(x)}{m}\right)\frac{\partial v^k}{\partial y} - \frac{\gamma\eta}{m\beta}\frac{\partial^2 v^k}{\partial y^2}\\ &= \mu\sum_{j=1}^{k-1}(b_{k-j-1} - b_{k-j})v^k + \mu b_{k-1}v^0 + G^k + \mathcal{R}^\alpha,\end{aligned}$$

where $1 \le k \le N$, in which $|\mathcal{R}^\alpha| \le C\tau^{2-\alpha}$, $\mu = \frac{\tau^{-\alpha}}{\Gamma(2-\alpha)}$. By deleting small term $\mathcal{R}^\alpha$, we can write

$$\begin{aligned}&(\mu - \eta\gamma)V^k + \gamma y\frac{\partial V^k}{\partial x} - \gamma\left(\eta y - \frac{F(x)}{m}\right)\frac{\partial V^k}{\partial y} - \frac{\gamma\eta}{m\beta}\frac{\partial^2 V^k}{\partial y^2}\\ &= \mu\sum_{j=1}^{k-1}(b_{k-j-1} - b_{k-j})V^j + \mu b_{k-1}V^0 + G^k,\end{aligned} \tag{4}$$

where $1 \le k \le N$.The convergence order in time variable of the second time-discrete scheme (STDS) presented in (4) is $\mathcal{O}(\tau^{2-\alpha})$.

3. Numerical Results

The numerical experiments carried out by utilizing MATLAB R2017b on a Core i5 (3.6 GHz) PC with 8 Gigabyte of RAM.

Example 3.1. Consider the 2D fractional Klein-Kremers Eq. (1) with $\alpha = 0.75$, where $f(x, y, t)$ has been chosen such that the exact solution is a Gaussian pulse with the following form

$$u(x,y,t) = t^3 \exp\left(\frac{(x-0.5)^2 + (y-0.5)^2}{-\beta}\right).$$

Numerical solution of this problem in an irregular domain has been considered in [1]. Dehghan et al. have proposed two meshless methods, MLPG and RBF collocation methods, for this problem. Table 1 shows the results of the presented LRBF-DQ method against their results in an irregular domain with the same discritization parameters. The considered irregular domain is shown in Figure 1, clearly. Figure 1 shows the graphs of approximate solution and absolute error at different final times with 400 nodes and $\tau = 10^{-3}$ using the LRBF-DQ method and the first time discrete scheme.

TABLE 1. Comparison between absolute errors of presented LRBF-DQ method and MLPG and RBF collocation methods [1] with $h = 1/10$ and different τ at $T = 1$.

	LRBF-DQ		RBF collocation		MLPG	
τ	$\|U^* - U\|_\infty$	$\|U^* - U\|_2$	$\|U^* - U\|_\infty$	$\|U^* - U\|_2$	$\|U^* - U\|_\infty$	$\|U^* - U\|_2$
1/10	$4.5183e-3$	$3.0501e-2$	$3.5125e-3$	$2.5947e-2$	$1.8311e-1$	$8.8324e-1$
1/20	$1.4125e-3$	$5.2732e-3$	$1.5704e-3$	$5.1632e-3$	$6.1827e-2$	$1.1062e-1$
1/40	$5.1571e-4$	$9.0811e-4$	$7.3764e-4$	$1.5063e-3$	$2.0411e-2$	$7.3348e-2$
1/80	$9.1266e-5$	$3.1782e-4$	$1.6752e-4$	$6.2702e-4$	$4.7921e-3$	$2.6172e-2$
1/160	$6.0729e-5$	$1.8105e-4$	$8.5202e-5$	$3.1429e-4$	$2.7421e-3$	$6.0258e-3$

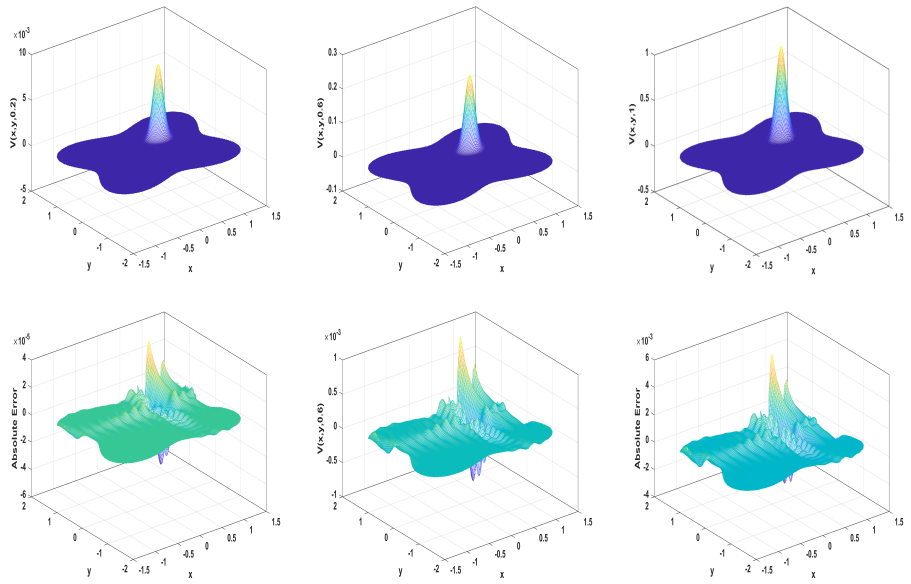


FIGURE 1. Approximate solution and absolute error at different final times with 400 nodes and $\tau = 10^{-2}$ using the LRBF-DQ method.

References

1. M. Dehghan, M. Abbaszadeh and A. Mohebbi, *Meshless local Petrov-Galerkin and RBFs collocation methods for solving 2D fractional Klein-Kramers dynamics equation on irregular domains*, Comput. Model Eng. Sci. **107** (2015) 481–516.
2. H. Pourbashash and M. Khaksar-e Oshagh, *Local RBF-FD technique for solving the two-dimensional modified anomalous sub-diffusion equation*, Appl. Math. Comput. **339** (2018) 144–152.
3. R. Metzler and J. Klafter, *From a generalized chapman- kolmogorov equation to the fractional klein- kramers equation*, J. Phys. Chem B **104** (2000) 3851–3857.

4. A. Mohebbi, M. Abbaszadeh and M. Dehghan, *The meshless method of radial basis functions for the numerical solution of time fractional telegraph equation*, Int. J. Numer. Methods Heat & Fluid Flow **24** (2014) 1636–1659.

E-mail: h.pourbashash@ugsr.ir
E-mail: mkhaksar@aut.ac.ir

Existence Theorem of a Quasi Solution to Inverse Source Problem in a Space Fractional Diffusion Equation

Amir Hossein Salehi Shayegan
Department of Mathematics, Faculty of Basic Science, Khatam-ol-Anbia (PBU) University, Tehran, Iran
Mohammad Shahriari
Department of Mathematics, Faculty of Basic Science, University of Maragheh, Maragheh, Iran
and Ali Safaie*
Department of Mathematics, Faculty of Basic Science, University of Maragheh, Maragheh, Iran

Abstract. In this paper, the existence solution of an inverse source problem related to a space fractional diffusion equation is studied. To this end, we consider a methodology, involving minimization of a cost functional to identify the unknown source function $f = f(x,t)$. Firstly, the stability of the corresponding direct problem is proved and then the continuity of the cost functional is concluded. Using these results the existence solution of the inverse source problem is given in an appropriate compact subset of admissible functions.
Keywords: Minimization of a cost functional, Inverse source problem, Space fractional diffusion equation.
AMS Mathematical Subject Classification [2010]: 35R30.

1. Introduction

We study the inverse problem associated with the following space fractional diffusion problem

$$u_t(x,t) - \frac{1}{2}{}^{R}D_x^{\alpha}u(x,t) - \frac{1}{2}{}_{x}^{R}D^{\alpha}u(x,t) = f(x,t), \quad (x,t) \in Q_T, \tag{1}$$

$$u(0,t) = u(\ell,t) = 0, \qquad t \in (0,T), \tag{2}$$

$$u(x,0) = \phi(x), \qquad x \in \Lambda, \tag{3}$$

where $u_t := \frac{\partial u}{\partial t}$, $\Lambda = (0,\ell)$, $Q_T = \Lambda \times (0,T)$ and $1 < \alpha < 2$. Here ${}^{R}D_x^{\alpha}u(x,t)$ and ${}_{x}^{R}D^{\alpha}u(x,t)$ denote the left and right Riemann-Liouville fractional derivatives, respectively, which are defined for $x \in (0,\ell)$ by

$${}^{R}D_x^{\alpha}u(x,t) = \frac{1}{\Gamma(2-\alpha)}\frac{d^2}{dx^2}\int_0^x \frac{u(\xi,t)}{(x-\xi)^{\alpha-1}}d\xi,$$

$${}_{x}^{R}D^{\alpha}u(x,t) = \frac{1}{\Gamma(2-\alpha)}\frac{d^2}{dx^2}\int_x^{\ell} \frac{u(\xi,t)}{(\xi-x)^{\alpha-1}}d\xi.$$

*Presenter

The inverse problem here consists of determining the source term $f = f(x,t)$ from the measured data at the final time

$$u(x,T) = \psi(x). \tag{4}$$

The function $\psi(x)$ is assumed to be the measured output data and also the functions f and ϕ are the inputs data. In this context, the inverse source problem (1)-(4) and the problem (1)-(3) for a given f will be referred as the problem (ISP) and the direct problem, respectively.

It is worth to point out that for $\alpha = 1$ and $\alpha = 2$, the ISP (1)-(4) is a classical ISP and has been studied by some researchers [1, 2]. But to our knowledge, there are a few works on inverse source space fractional diffusion equations.

Let us denote by $\chi := L_2(Q_T)$, the set of admissible unknown source functions f. Evidently, the set χ is closed and convex. The weak solution of the direct problem (1)-(3) will be defined as the function $u \in B^{\alpha/2}(Q_T)$ satisfying the integral identity

$$\Pi(u,v) = F(v), \qquad \forall v \in B^{\alpha/2}(Q_T),$$

where the bilinear form $\Pi(\cdot,\cdot)$ is defined by

$$\Pi(u,v) := (u_t, v)_{L_2(Q_T)} - \frac{1}{2}\left({}^{R}D_x^{\frac{\alpha}{2}} u, {}_x^{R}D^{\frac{\alpha}{2}} v\right)_{L_2(Q_T)} - \frac{1}{2}\left({}_x^{R}D^{\frac{\alpha}{2}} u, {}^{R}D_x^{\frac{\alpha}{2}} v\right)_{L_2(Q_T)},$$

and the functional $F(\cdot)$ is given by $F(v) := (f,v)_{L_2(Q_T)}$. Here

$$B^{\alpha}(Q_T) := L_\infty\left((0,T), L_2(\Lambda)\right) \cap L_2\left((0,T), H_0^{\alpha}(\Lambda)\right),$$

is a Banach space with respect to the norm

$$\|v\|_{B^{\alpha}(Q_T)} = \left(\max_{0\le t\le T} \|v(\cdot,t)\|^2_{L_2(\Lambda)} + \|v\|^2_{L_2((0,T),H_0^{\alpha}(\Lambda))}\right)^{1/2},$$

where $L_2((0,T), H_0^{\alpha}(\Lambda)) = \left\{v \mid \|v(\cdot\,,t)\|_{H_0^{\alpha}(\Lambda)} \in L_2(0,T)\right\}$, endowed with the norm

$$\|v\|_{L_2((0,T),H_0^{\alpha}(\Lambda))} = \left\|\|v(\cdot,t)\|_{H_0^{\alpha}(\Lambda)}\right\|_{L_2(0,T)}.$$

In the above definition $H_0^{\alpha}(\Lambda)$ denotes the usual fractional Sobolev space with respect to the norm $\|\cdot\|_{H_0^{\alpha}(\Lambda)}$ (For more details see [1]). Now suppose that $f \in \chi$ and $\phi \in L_2(\Lambda)$. Then it is proved that the weak solution $u \in B^{\alpha/2}(Q_T)$ of the direct problem (1)-(3) exists and is unique [3, 4]. We denote this weak solution by $u(x,t;f)$ corresponding to a given $f \in \chi$. If this function satisfies the additional condition (4), then it must satisfy the equation

$$u(x,t;f)\,|_{t=T} = \psi(x), \qquad x \in \Lambda. \tag{5}$$

However, due to measurement errors in practice, exact equality in the above equation is usually not achieved [1]. For this reason, we define a quasi solution of the inverse problem as a solution of a minimization problem. In doing so, find $f_* \in \chi$ such that

$$J(f_*) = \inf_{f\in\chi} J(f),$$

where

$$J(f) = \int_0^\ell (u(x,t;f)\,|_{t=T} - \psi(x))^2 dx.$$

Clearly, if $J(f_*) = 0$, then the quasi solution $f_* \in \chi$ is a strict solution of the inverse problem (1)-(4) and also $f_* \in \chi$ satisfies the functional Eq. (5). To prove the existence of the quasi solution, one needs to have, existence and uniqueness of the weak solution of direct problem (See [4]), stability of the problem (See Lemmas 2.1 and Corollary 2.2), continuity of the functional J (See Lemma 2.3), having a compact subset of admissible functions χ (See Theorem 2.4).

In next section, we give some explanations about above requirements.

2. Existence Theorem of the Quasi Solution

Let f and $f+\delta f \in \chi$ be source functions. We denote by $u(x,t;f)$ and $u(x,t;f+\delta f)$ the corresponding solutions of the problem (1)-(3). Then

$$\delta u(x,t;f) := u(x,t;f+\delta f) - u(x,t;f),$$

is the solution of the following problem

$$\delta u_t(x,t) - \frac{1}{2}\,{}^R D_x^\alpha \delta u(x,t) - \frac{1}{2}\,{}_x^R D^\alpha \delta u(x,t) = \delta f(x,t), \tag{6}$$

$$\delta u(0,t) = \delta u(\ell,t) = 0, \qquad t \in (0,T), \tag{7}$$

$$\delta u(x,0) = 0, \qquad x \in \Lambda. \tag{8}$$

The first variation ΔJ of the cost functional J is

$$\begin{aligned}\Delta J(f) &:= J(f+\delta f) - J(f)\\ &= 2\int_0^\ell (u(x,t;f)\,|_{t=T} - \psi(x))\delta u(x,t;f)\,|_{t=T}\, dx\\ &+ \int_0^\ell (\delta u(x,t;f)\,|_{t=T})^2 dx,\end{aligned}$$

where $\delta u(x,t;f)$ is the solution of (6)-(8).

Lemma 2.1. *Let $f, f+\delta f \in \chi$ be given source functions. If $u = u(x,t;f)$ is the solution of direct problem (1)-(3) and $p = p(x,t)$ is the solution of adjoint problem*

$$p_t(x,t) + \frac{1}{2}\,{}^R D_x^\alpha p(x,t) + \frac{1}{2}\,{}_x^R D^\alpha p(x,t) = 0, \qquad (x,t) \in Q_T, \tag{9}$$

$$p(0,t) = p(\ell,t) = 0, \qquad t \in (0,T), \tag{10}$$

$$p(x,T) = q(x), \qquad x \in \Lambda, \tag{11}$$

with an arbitrary function $q = q(x) \in L_2(\Lambda)$, then the following integral identity holds

$$\int_0^\ell q(x)\delta u(x,t;f)\,|_{t=T} dx = \int_0^T \int_0^\ell \delta f(x,t)p(x,t)dxdt. \tag{12}$$

Proof. Multiply (6) by p and integrate over Q_T to get

$$(\delta u_t, p)_{L_2(Q_T)} - \frac{1}{2}\big({}^R D_x^\alpha\, \delta u, p\big)_{L_2(Q_T)} - \frac{1}{2}\big({}_x^R D^\alpha\, \delta u, p\big)_{L_2(Q_T)} = (\delta f, p)_{L_2(Q_T)}. \tag{13}$$

According to [5], we have

$$\begin{aligned} \left({}^{R}D_{x}^{\alpha}\delta u, p\right)_{L_2(Q_T)} &= \left(\delta u, {}_{x}^{R}D^{\alpha}p\right)_{L_2(Q_T)}, \\ \left({}_{x}^{R}D^{\alpha}\delta u, p\right)_{L_2(Q_T)} &= \left(\delta u, {}^{R}D_{x}^{\alpha}p\right)_{L_2(Q_T)}. \end{aligned} \tag{14}$$

Now, consider the first term on the left hand side (13). Applying integration by parts, we get

$$\begin{aligned} (\delta u_t, p)_{L_2(Q_T)} &= \int_0^{\ell}\int_0^{T} \delta u_t(x,t;f)p(x,t)dtdx \\ &= \int_0^{\ell} \delta u(x,t;f)\,|_{t=T}\; p(x,T)dx - \int_0^{\ell} \delta u(x,t;f)\,|_{t=0}\; p(x,0)dx \\ &- (\delta u, p_t)_{L_2(Q_T)}. \end{aligned}$$

So, we obtain

$$(\delta u_t, p)_{L_2(Q_T)} = \int_0^{\ell} \delta u(x,t;f)\,|_{t=T}\; q(x)dx - (\delta u, p_t)_{L_2(Q_T)}. \tag{15}$$

For the second and third terms on the left hand side (13), using (14) we have

$$\begin{aligned} &-\frac{1}{2}\left({}^{R}D_{x}^{\alpha}\delta u, p\right)_{L_2(Q_T)} - \frac{1}{2}\left({}_{x}^{R}D^{\alpha}\delta u, p\right)_{L_2(Q_T)} \\ &= -\frac{1}{2}\left(\delta u, {}_{x}^{R}D^{\alpha}p\right)_{L_2(Q_T)} - \frac{1}{2}\left(\delta u, {}^{R}D_{x}^{\alpha}p\right)_{L_2(Q_T)}. \end{aligned} \tag{16}$$

Applying (15) and (16) in (13), we can obtain

$$\begin{aligned} &\int_0^{\ell} \delta u(x,t;f)\,|_{t=T}\; q(x)dx - (\delta u, p_t)_{L_2(Q_T)} \\ &-\frac{1}{2}\left(\delta u, {}_{x}^{R}D^{\alpha}p\right)_{L_2(Q_T)} - \frac{1}{2}\left(\delta u, {}^{R}D_{x}^{\alpha}p\right)_{L_2(Q_T)} = (\delta f, p)_{L_2(Q_T)}, \end{aligned}$$

and

$$\begin{aligned} \int_0^{\ell} \delta u(x,t;f)\,|_{t=T}\; q(x)dx &+ \left(\delta u, -p_t - \frac{1}{2}\,{}_{x}^{R}D^{\alpha}p - \frac{1}{2}\,{}^{R}D_{x}^{\alpha}p\right)_{L_2(Q_T)} \\ &= (\delta f, p)_{L_2(Q_T)}, \end{aligned}$$

which leads to

$$\int_0^{\ell} \delta u(x,t;f)\,|_{t=T}\; q(x)dx = \int_0^{T}\int_0^{\ell} \delta f(x,t)p(x,t)dxdt.$$

□

Corollary 2.2. *Let us choose an arbitrary control function $q = q(x)$ in (12) as $q(x) := \frac{\delta u(x,t;f)|_{t=T}}{\|\delta u(x,t;f)|_{t=T}\|_{L_2(\Lambda)}}$. Then we obtain*

$$\|\delta u(x,t;f)\,|_{t=T}\,\|_{L_2(\Lambda)} \le \|p\|_{L_2(Q_T)}\|\delta f\|_{L_2(Q_T)}, \tag{17}$$

where $\delta u = \delta u(x,t;f)$ is the solution of (6)-(8) and $p = p(x,t)$ is defined in Lemma 2.1. We note that the existence and uniqueness of (9)- (11) are the straight forward results of [5].

Next, in order to prove that the functional $J(\phi)$ is continuous, we will use the stability estimate (17).

LEMMA 2.3. *The functional $J(f)$, is continuous on χ in the sense that if $\|f_n - f\|_{L_2(\Lambda)} \to 0$, then $|J(f_n) - J(f)| \to 0$ as $n \to \infty$.*

PROOF. Let $\{f_n\}_{n=1}^{\infty} \in \chi$ be a sequence of initial data which converges to $f \in \chi$. Thus we have

$$\begin{aligned}
|J(f_n) - J(f)| &= \left| \int_0^\ell \left(u(x,t;f_n)\,|_{t=T} - \psi(x)\right)^2 dx - \int_0^\ell \left(u(x,t;f)\,|_{t=T} - \psi(x)\right)^2 dx \right| \\
&= \left| \|u(\cdot,T;f_n) - \psi(x)\|^2_{L_2(\Lambda)} - \|u(\cdot,T;f) - \psi(x)\|^2_{L_2(\Lambda)} \right| \\
&= \left| \|u(\cdot,T;f_n) - \psi(x)\|_{L_2(\Lambda)} + \|u(\cdot,T;f) - \psi(x)\|_{L_2(\Lambda)} \right| \\
&\times \left| \|u(\cdot,T;f_n) - \psi(x)\|_{L_2(\Lambda)} - \|u(\cdot,T;f) - \psi(x)\|_{L_2(\Lambda)} \right| \\
&\leq C\|u(\cdot,T;f_n) - u(\cdot,T;f)\|_{L_2(\Lambda)} \\
&\leq C\|p\|_{L_2(Q_T)}\|f_n - f\|_{L_2(\Lambda)},
\end{aligned}$$

The Lemma 2.1, Corollary 2.2 and the above inequality show that $|J(f_n) - J(f)|$ goes to zero as $n \to \infty$. This completes the proof. □

THEOREM 2.4. *Let $\chi_c \subset \chi$ be a compact subset of source functions. Then the ISP (1)-(4) has at least one quasi solution in χ_c.*

PROOF. By using Weierstrass theorem and Lemma 2.3, existence result for the ISP is proved. □

References

1. A. Hasanov, *Simultaneous determination of source terms in a linear parabolic problem from the final overdetermination: Weak solution approach*, J. Math. Anal. Appl. **330** (2) (2007) 766–779.
2. A. Hasanov Hasanoğlu, V. Romanov and G. Vladimir, *Introduction to Inverse Problems for Differential Equations*, Springer, New York, 2017.
3. A. H. Salehi Shayegan and A. Zakeri, *A numerical method for determining a quasi solution of a backward time-fractional diffusion equation*, Inverse Probl. Sci. Eng. **26** (8) (2018) 1130–1154.
4. A. H. Salehi Shayegan and A. Zakeri, *Quasi solution of a backward space fractional diffusion equation*, J. Inverse Ill-Posed Probl. **27** (6) (2019) 795–814.
5. A. H. Salehi Shayegan, R. Bayat Tajvar, A. R. Ghanbari and A. Safaie, *Inverse source problem in a space fractional diffusion equation from the final overdetermination*, Appl. Math. **64** (4) (2019) 469–484.

E-mail: ah.salehi@mail.kntu.ac.ir
E-mail: shahriari@maragheh.ac.ir
E-mail: alisafaie549@gmail.com

The Spectral Element Method for the Solution of Two Dimensional Telegraph Equation

Marziyeh Saffarian*
Faculty of Mathematical Sciences, University of Kashan, Kashan, Iran
and Akbar Mohebbi
Faculty of Mathematical Sciences, University of Kashan, Kashan, Iran

Abstract. In this paper, we present a numerical scheme for the solution of two dimensional telegraph equation. We use spectral element method in spatial direction and Crank-Nicolson method in temporal direction. The unconditional stability of the semi discrete scheme is proved and error estimate of the fully discrete method is presented. Finally, we consider a test problem to demonstrate the accuracy and applicability of the proposed method.
Keywords: Two dimensional telegraph equation, Spectral element method, Crank-Nicolson scheme.
AMS Mathematical Subject Classification [2010]: 65M06, 65M60, 65M12.

1. Introduction

At the present work, we propose a numerical scheme for the solution of two dimensional telegraph equation as [4]

$$(1)\qquad \begin{cases} u_{tt} + 2\alpha u_t + u = \Delta u + f(\mathbf{x}, t), & (\mathbf{x}, t) \in \Omega \times (0, T], \\ u(\mathbf{x}, 0) = g_1(\mathbf{x}), \quad u_t(\mathbf{x}, 0) = g_2(\mathbf{x}), & \mathbf{x} \in \Omega, \\ u(\mathbf{x}, t) = h(\mathbf{x}, t), & (\mathbf{x}, t) \in \partial\Omega \times (0, T], \end{cases}$$

where $\Omega \subset \mathbf{R}^2$.

At the first time, Patera used the spectral element method to solve problems of computational fluid dynamics in 1984. This method is combination of the finite element method and the spectral method. Therefore, this method have the high order accuracy of the spectral method and the geometric flexibility of the finite element method [3]. This method is used to diversity of the partial differential equations [2, 5].

In this work, we use Crank-Nicolson scheme to discretize Eq. (1) in temporal direction and apply the spectral element method to approximate this equation in spatial direction. Then we obtain an uncondition stable scheme of order $\mathcal{O}(\tau^2)$ for Eq. (1).

The rest of this work is orgonized as: In Section 2, we use Crank-Nicolson scheme to discretize Eq. (1) in temporal direction. We introduce the spectral element method and implement this method on telegraph equation in Section 3. Also, we present the error estimate of the fully discrete scheme in this section. We consider a test problem to illustrate the efficiency and accuracy of the proposed

*Presenter

scheme in Section 4. In Section 5, a brief conclusion is dedicated. At the end, some references are introduced.

2. Semi-Discrete Scheme

Suppose $L_2(\Omega)$ is the space of measurable functions whose square is Lebesgue integrable in Ω. Define the inner product and norm in this space as

$$(u,v)=\int_{\Omega} uvd\Omega, \quad \|v\|=(v,v)^{\frac{1}{2}}.$$

Also define sobolev space with the inner product and norm as

$$H^1(\Omega)=\{v\in L^2(\Omega),\ \nabla v\in L^2(\Omega)\}, \quad H_0^1(\Omega)=\{v\in H^1(\Omega),\ v|_{\partial\Omega}=0\},$$

$$(u,v)_1=(u,v)+(\nabla u,\nabla v), \quad \|v\|_1=(v,v)_1^{\frac{1}{2}}, \quad |v|_1=(\nabla v,\nabla v)^{\frac{1}{2}}.$$

Consider the following notations

$$t_n=n\tau,\ n=0,1,\ldots,N, \qquad T=N\tau,$$

$$u(\mathbf{x},t_n)=u^n, \quad u_{tt}^n=(u^{n+1}-2u^n+u^{n-1})/\tau^2, \quad u_t^n=(u^{n+1}-u^{n-1})/2\tau.$$

Applying the Crank-Nicolson scheme on the Eq. (1) gives

$$u_{tt}^n+2\alpha u_t^n+\frac{1}{2}(u^{n+1}+u^{n-1})=\frac{1}{2}(\Delta u^{n+1}+\Delta u^{n-1})+f^n+R,$$

or

$$(1+\alpha\tau+\frac{\tau^2}{2})u^{n+1}-\frac{\tau^2}{2}(\Delta u^{n+1})=2u^n-(1-\alpha\tau+\frac{\tau^2}{2})u^{n-1} + \frac{\tau^2}{2}\Delta u^{n-1}+\tau^2 f^n+\tau^2 R,$$

where $|R|\le C_1\tau^2$. Let U^n is an approximation of the exact solution u^n. Omitting the small term R gives

$$(2)\qquad \begin{aligned}(1+\alpha\tau+\frac{\tau^2}{2})U^{n+1}-\frac{\tau^2}{2}\Delta U^{n+1} &= 2U^n-(1-\alpha\tau+\frac{\tau^2}{2})U^{n-1}\\ &+ \frac{\tau^2}{2}\Delta U^{n-1}+\tau^2 f^n.\end{aligned}$$

Theorem 2.1. *The semi-discrete scheme (2) is unconditionally stable for all $U\in H_0^1$. That is*

$$\|U^{n+1}\|^2\le C\Big((1+\alpha\tau+\frac{\tau^2}{2})^2\|U^0\|^2 + \tau^2\|\nabla U^0\|^2+\tau^2(1+\alpha\tau+\frac{\tau^2}{2})\|g_2\|^2+\tau^4\|\nabla g_2\|^2+T^4\max_{1\le k\le N}\|f^k\|^2\Big),$$

where C is positive constant.

3. Spectral Element Method

In this section, we use the Legendre spectral element method to obtain fully discrete scheme. The Galerkin formulation of the Eq. (2) as:

Find $U^{n+1} \in H_0^1(\Omega)$ such that for all $v \in H_0^1$

$$
\begin{aligned}
(3) \qquad (1+\alpha\tau+\frac{\tau^2}{2})(U^{n+1},v) \quad &+ \quad \frac{\tau^2}{2}(\nabla U^{n+1},\nabla v) \\
&= \quad 2(U^n,v)-(1-\alpha\tau+\frac{\tau^2}{2})(U^{n-1},v) \\
&- \quad \frac{\tau^2}{2}(\nabla U^{n-1},\nabla v)+\tau^2 f^n.
\end{aligned}
$$

We divide the domain into N_e non-overlapping subdomains. Then we can write the solution approximation of order M for function U in per element as

$$U^e(x,t_k)=\sum_{i=0}^{M} U(x_i,t_k)\varphi_i(x), \quad 1\le e\le N_e, \quad 1\le k\le N,$$

where N_e denotes the number of elements, $U(x_i,t_k)$ is the ith local degree freedom and $\varphi_i(x)$ is the ith Lagrange polynomial of order M as

$$\varphi_i(\eta)=\frac{1}{M(M+1)L_M(\eta_i)}\frac{(\eta^2-1)L_M^{'}(\eta)}{\eta-\eta_i}, \quad 0\le i\le M,\ -1\le\eta\le 1,$$

where $\{\eta_i\}_{i=0}^M$ are Gauss-Lobatto-Legendre (GLL) points and L_M is the Legendre polynomial of order M.

Let x_{e-1} and x_e denote the boundary points of each element and the length of element is $h_e=x_e-x_{e-1}$. The entries of the element mass matrix and the element stiffness matrix, respectively, are presented by

$$(4) \qquad B_{ij}^e=\int_{x_{e-1}}^{x_e}\varphi_i(x)\varphi_j(x)dx=\frac{h_e}{2}\int_{-1}^{1}\varphi_i(\eta)\varphi_j(\eta)d\eta=\frac{h_e}{2}\delta_{ij}w_i,$$

$$(5) \quad D_{ij}^e=\int_{x_{e-1}}^{x_e}\frac{\varphi_i(x)}{dx}\frac{\varphi_j(x)}{dx}dx=\frac{2}{h_e}\int_{-1}^{1}\frac{\varphi_i(\eta)}{d\eta}\frac{d\varphi_j}{d\eta}d\eta=\frac{2}{h_e}\sum_{l=0}^{M}d_{il}d_{jl}w_l,$$

To evaluate the integrals in (4) and (5), we use Gaussian quadrature with the $M+1$ GLL points. $\{w_l\}_{l=0}^M$ are the GLL quadrature weights and d_{ij} are the entires of transpose of matrix K as

$$K_{ij}^e=\begin{cases}\dfrac{L_M(\eta_i)}{L_M(\eta_j)}\dfrac{1}{\eta_i-\eta_j}, & i\neq j,\\[2ex] -\dfrac{M(M+1)}{4}, & i=j=0,\\[2ex] \dfrac{M(M+1)}{4}, & i=j=M,\\[2ex] 0, & otherwise.\end{cases}$$

3.1. Error Estimate. Suppose $\mathbb{P}_M(\Omega)$ is the space of polynomials defined on Ω with the degree no greater than $M \in \mathbb{N}$. Define

$$\mathcal{H}_M^0 = \{v \in H_0^1 : v|_{\Omega_e} \in \mathbb{P}_M(\Omega)\}.$$

Define the Ritz projection $\Re_h : H_0^1 \to \mathcal{H}_M^0$, as

$$(\nabla(u - \Re_h u), \nabla v) = 0, \quad u \in H_0^1, \quad v \in \mathcal{H}_M^0.$$

LEMMA 3.1. [1] *Let $u \in H^\varsigma$, and h_k is the diameter of element k, then*

$$||u - \Re_h u|| \leq C h_k^{((\min(M+1),\varsigma)-1)} M^{1-\varsigma} ||u||_\varsigma,$$

THEOREM 3.2. *Suppose u^n be the exact solution of the Eq. (1), U^n be the solution of the full discrete scheme (3) and $e^n = u^n - U^n$. Then the followimg error estimate holds*

$$||e^n|| \leq C(\tau^2 + M^{1-\varsigma}), \quad C = \max\left\{2C_{u1}, (1 + \alpha\tau + \frac{\tau^2}{2})C_u, \sqrt{3C_1}\tau\right\}.$$

4. Numerical Results

In this section, we report the numerical experiment of presented method. Let E_1 and E_2 are error correspond to time steps τ_1 and τ_2, then computational order of the presented method can be calculate by

$$C - order = \frac{\log \frac{E_1}{E_2}}{\log \frac{\tau_1}{\tau_2}}.$$

EXAMPLE 4.1. Consider Eq. (1) with $\alpha = 1$ and the exact solution $u(\mathbf{x}, t) = \cos(t)\sin(x)\sin(y)$ [4]. We use the proposed scheme $\Omega = [0, 1] \times [0, 1]$ and present the L_∞ and L_2 norm of errors and computational order of the proposed method with $N_e = 5$ and $M = 5$ in Table 1. The graphs of numerical solution and absolute error for this problem are presented in Figure 1. In Figure 2, we present the graph of error as a function of M with $Ne = 3$, and the error as a function of Ne with $M = 3$.

TABLE 1. Errors and computational orders at $T = 1$ for Test problem 1.

τ	L_∞	C-order	L_2	C-order
1/20	3.4751×10^{-5}		4.9933×10^{-4}	
1/40	8.6523×10^{-6}	2.0059	1.2240×10^{-4}	2.0284
1/80	2.1510×10^{-6}	2.0081	3.0411×10^{-5}	2.0089
1/160	5.3861×10^{-7}	1.9980	7.5940×10^{-6}	2.0017
1/320	1.3456×10^{-7}	2.0010	1.8979×10^{-6}	2.0005

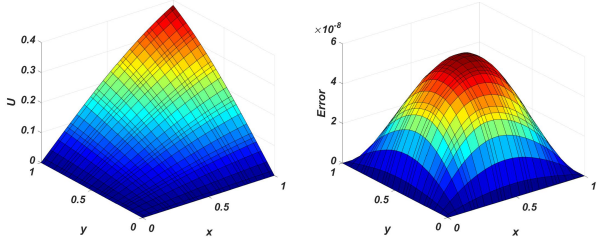


FIGURE 1. Numerical solution (left side) and absolute error (right side) with $\tau = 0.002$.

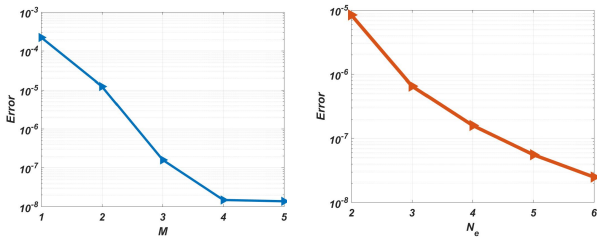


FIGURE 2. Error as function of M (left panel) and Ne (right panel) with $\tau = 0.001$.

5. Conclution

In this article, we studied the spectral element method for the solution of two dimensional telegraph equation. A finite difference scheme of order $\mathcal{O}(\tau^2)$ is presented for discretizing temporal direction. The accuracy of the present method is shown by an example.

References

1. C. Canuto, M. Y. Hussaini, A. Quarteroni and T. A. Zang, *Spectral Methods Fundamentals in Single Domains*, Springer, New York, 2006.
2. M Dehghan, M. Abbaszadeh and A. Mohebbi, *Legendre spectral element method for solving time fractional modified anomalous sub-diffusion equation,* Appl. Math. Model. **40** (2016) 3635–3654.
3. F. X. Giraldo, *Strong and weak Lagrange-Galerkin spectral element methods for the shallow water equations*, Compu. Math. Appl. **45** (2003) 97–121.
4. S. Yuzbasi and M. Karakayir, *A Galerkin-like scheme to solve two-dimensional telegraph equation using collocation points in initial and boundary conditions*, Comput. Math. Appl. **74** (2017) 3242–3249.
5. M. Saffarian and A. Mohebbi, *The Galerkin spectral element method for the solution of two-dimensional multiterm time fractional diffusion-wave equation*, Math. Meth. Appl. Sci. (2019) 1–17.

E-mail: m.Saffarian11@grad.kashanu.ac.ir
E-mail: a_mohebbi@kashanu.ac.ir

Approximation of Wiener Integrals via Rationalized Haar Functions

Nasrin Samadyar*
Department of Mathematics, Faculty of Mathematical Sciences, Alzahra University, Tehran, Iran

Abstract. In this paper, we present a suitable numerical technique to approximate the Wiener integrals which either their exact values are not available or finding their exact values are complicated. This suggested method is based on rationalized Haar functions which form a orthogonal basis for Hilbert space $L^2[0,1]$. Finally, we estimate some numerical examples to indicate the high accuracy and efficiency of the suggested technique.
Keywords: Brownian motion process, Wiener integrals, Rationalized Haar functions.
AMS Mathematical Subject Classification [2010]: 60J65, 60H05.

1. Introduction

Definition 1.1. Brownian motion process $\{B(t)\}$ is a stochastic process which satisfies in the following properties [5]

1) $B(t) - B(s)$ for $t > s$ is independent of the past. That means for $0 < u < v < s < t < T$, the increments $B(t) - B(s)$ and $B(v) - B(u)$ are independent.
2) $B(t)-B(s)$ for $t > s$ has Normal distribution with mean zero and variance $t-s$. In other words, $B(t)-B(s) \sim \sqrt{t-s}N(0,1)$, where $N(0,1)$ denotes Normal distribution with zero mean and unit variance.
3) $B(t)$, $t \geq 0$ are continuous functions of t.

The general form of Wiener integral is as follows

$$\int_0^T f(t)dB(t),$$

such that $f(t)$ is a deterministic function (it dose not depend on $B(t)$) and $B(t)$ denotes a standard Brownian motion process.

Remark 1.2. If $f(t)$ be a differentiable function (more generally, a function of finite variation), then the Wiener integral $\int_0^T f(t)dB(t)$ can be defined by formally using the integration by parts formula as follows [2]

$$\int_0^T f(t)dB(t) = f(T)B(T) - \int_0^T B(t)df(t).$$

*Presenter

2. Preliminaries

DEFINITION 2.1. The functions $RH(r,t), r = 1,2,3,\ldots$ are composed of three values $+1$, -1 and 0 and are defined on the interval $[0,1)$ as follows [4]

$$RH(r,t) = \begin{cases} +1, & J_1 \le t < J_{\frac{1}{2}}, \\ -1, & J_{\frac{1}{2}} \le t < J_0, \\ 0, & \text{otherwise}, \end{cases}$$

where

$$J_u = \frac{j-u}{2^i}, \qquad u = 0, \frac{1}{2}, 1,$$
$$r = 2^i + j - 1, \qquad i = 0,1,2,\ldots, \quad j = 1,2,3,\ldots,2^i.$$

The function $RH(0,t)$ is defined for $i = j = 0$ as follows

$$RH(0,t) = 1, \qquad 0 \le t < 1.$$

The orthogonality property of these functions is given by

$$\int_0^1 RH(r,t)RH(s,t) = \begin{cases} 2^{-i}, & r = s, \\ 0, & r \ne s. \end{cases}$$

THEOREM 2.2. *The sequence $\{RH(r,t)\}_{r=0}^{\infty}$ is a complete orthogonal basis for Hilbert space $L^2[0,1)$. So, we can approximate every function $f(t) \in C[0,1])$ via the following series* [1]

$$f(t) \simeq f_n(t) = \sum_{r=0}^{n} a_r RH(r,t), \tag{1}$$

where $a_r = 2^i \int_0^1 f(t)RH(r,t)dt$, $r = 0,1,2,\ldots$.

3. Approximation of Wiener Integrals

THEOREM 3.1. *Suppose that $\{\varphi_i\}_{i=0}^{\infty}$ be an orthonormal basis for Hilbert space $L^2[0,T]$ and $f(t) \in L^2[0,T]$. Every square integrable function $f(t)$ can be expanded as follows*

$$f(t) = \sum_{i=0}^{\infty} \langle f, \varphi_i \rangle \varphi_i(t), \tag{2}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product in Hilbert space $L^2[0,T]$ and is defined as

$$\langle f, g \rangle = \int_0^T f(t)g(t)dt.$$

THEOREM 3.2. The Wiener integral $\int_0^T f(t)dB(t)$ is provided by stochastic integrating from both sides of Eq. (2) over the interval $[0,T)$. Thus, we obtain

$$\int_0^T f(t)dB(t) = \sum_{i=0}^{\infty} \langle f, \varphi_i \rangle \int_0^T \varphi_i(t)dB(t), \tag{3}$$

such that the existed series in Eq. (3) has almost surely convergence [3].

The functions $RH(r,t), r = 0,1,2,\ldots$ introduced in Definition 2.1 are not orthonormal. But, we can construct a orthonormal basis by dividing them by their norm. So, we obtain orthonormal rationalized Haar functions $ORH(r,t)$ as follows

$$ORH(r,t) = \frac{1}{\sqrt{2^{-i}}} RH(r,t).$$

From orthonormality property of functions $\{ORH(r,t)\}_{r=0}^{\infty}$, we can write the approximation of every square integrable function $f(t) \in L^2[0,T)$ as follows

$$f(t) \simeq f_n(t) = \sum_{r=0}^{n} \langle f(t), ORH(r,t)\rangle ORH(r,t).$$

To approximate the Winner integrals over the interval $[0,T)$ using rational Haar functions, Eq. (1) yields

$$\begin{aligned}
\int_0^T f(t)dB(t) &\simeq \int_0^T \sum_{r=0}^{n} \Big(2^i \int_0^T f(t)RH(r,t)dt\Big) RH(r,t)dB(t) \\
&= \sum_{r=0}^{n} \Big(2^i \int_0^T f(t)RH(r,t)dt\Big)\Big(\int_0^T RH(r,t)dB(t)\Big) \\
&= \Big(\int_0^T f(t)RH(0,t)dt\Big)\Big(\int_0^T RH(0,t)dB(t)\Big) \\
&+ \sum_{r=1}^{n} \Big(2^i \int_0^T f(t)RH(r,t)dt\Big)\Big(\int_0^T RH(r,t)dB(t)\Big) \\
&= B(1)\int_0^1 f(t)dt \\
&+ \sum_{r=1}^{n} 2^i \Big(\int_{J_1}^{J_{\frac{1}{2}}} f(t)dt - \int_{J_{\frac{1}{2}}}^{J_0} f(t)dt\Big)\Big(2B(J_{\frac{1}{2}}) - B(J_1) - B(J_0)\Big).
\end{aligned}$$

4. Test Problems

In this section, we apply the present method for different values of n to estimate some Wiener integrals and report obtained results in tables. All the numerical calculations have achieved by running MATLAB code on an Intel COREi3 laptop.

Example 4.1. Approximate the values of Wiener integral $\int_0^1 \sin(t)dB(t)$. To obtain the exact values of this Wiener integral, we utilize integration by parts formula and get

$$\int_0^1 \sin(t)dB(t) = \sin(1)B(1) - \int_0^1 \cos(t)B(t)dt,$$

where $\int_0^1 \cos(t)B(t)dt$ is approximated via Riemann sum idea as follows:

$$\int_0^1 \cos(t)B(t)dt \simeq \sum_{i=0}^{N-1} \cos(t_i)B(t_i)(t_{i+1} - t_i),$$

where $0 = t_0 < t_1 < t_2 < \cdots < t_N = 1$. On the other hand, we approximate it via the explained method as follows:

$$\begin{aligned}\int_0^1 \sin(t)dB(t) &\simeq B(1)\int_0^1 \sin(t)dt \\ &+ \sum_{r=0}^{n} 2^i\Big(\int_{J_1}^{J_{\frac{1}{2}}} \sin(t)dt - \int_{J_{\frac{1}{2}}}^{J_0} \sin(t)dt\Big)\Big(2B(J_{\frac{1}{2}}) - B(J_1) - B(J_0)\Big).\end{aligned}$$

The exact values, approximate values and absolute error values for different amounts of n and N have been reported in Table 1. This results confirm that our method is very accurate and efficient.

TABLE 1. Results of Example 4.1.

	$N = 25$			$N = 50$		
n	Exact	Approximate	Error	Exact	Approximate	Error
7	0.4156	0.3841	0.0318	-0.3814	-0.3734	0.0079
15	-0.2845	-0.2934	0.0089	0.1415	0.1466	0.0051
31	0.6903	0.6823	0.0079	-0.3599	-0.3639	0.0039
63	-0.3632	-0.3571	0.0061	-0.6481	-0.6468	0.0012
127	-0.4484	-0.4468	0.0015	-0.3182	-0.3187	0.0005

EXAMPLE 4.2. Approximate the values of Wiener integral $\int_0^1 \cos(t)dB(t)$. To obtain the exact values of this Wiener integral, we utilize integration by parts formula and get

$$\int_0^1 \cos(t)dB(t) = \cos(1)B(1) + \int_0^1 \sin(t)B(t)dt,$$

where $\int_0^1 \sin(t)B(t)dt$ is approximated via Riemann sum idea as follows:

$$\int_0^1 \sin(t)B(t)dt \simeq \sum_{i=0}^{N-1} \sin(t_i)B(t_i)(t_{i+1} - t_i),$$

where $0 = t_0 < t_1 < t_2 < \cdots < t_N = 1$. On the other hand, we approximate it via the explained method as follows:

$$\begin{aligned}\int_0^1 \cos(t)dB(t) &\simeq B(1)\int_0^1 \cos(t)dt \\ &+ \sum_{r=0}^{n} 2^i\Big(\int_{J_1}^{J_{\frac{1}{2}}} \cos(t)dt - \int_{J_{\frac{1}{2}}}^{J_0} \cos(t)dt\Big)\Big(2B(J_{\frac{1}{2}}) - B(J_1) - B(J_0)\Big).\end{aligned}$$

The exact values, approximate values and absolute error values for different amounts of n and N have been reported in Table 2. This results confirm that our method is very accurate and efficient.

TABLE 2. Results of Example 4.2.

	$N = 25$			$N = 50$		
n	Exact	Approximate	Error	Exact	Approximate	Error
7	0.6773	0.6873	0.0099	0.3691	0.3615	0.0076
15	-0.4226	-0.4268	0.0041	-0.5209	-0.5156	0.0053
31	-0.5522	-0.5547	0.0024	0.6114	0.6157	0.0043
63	0.2301	0.2319	0.0018	0.0098	0.0091	0.0007
127	-0.6837	-0.6848	0.0010	-0.1667	-0.1662	0.0004

5. Conclusion

In this paper, an efficient algorithm has been applied to approximate the value of Wiener integrals. Some test problems have been included to demonstrate the efficiency and accuracy of suggested method. From obtained results, we conclude that more accurate numerical results can be provided by

- increasing the number of used basis n,
- considering a larger value for N.

Acknowledgement

The authors would like to express our very great appreciation to reviewers for their valuable comments which have improved the quality of our paper.

References

1. M. Erfanian, M. Gachpazan and H. Beiglo, *Rationalized Haar wavelet bases to approximate solution of nonlinear Fredholm integral equations with error analysis*, Appl. Math. Comput. **265** (2015) 304–312.
2. F. C. Klebaner, *Introduction to Stochastic Calculus with Applications*, 2nd ed., Imperial College Press, London, 2005.
3. H. -H. Kuo, *Introduction to Stochastic Integration*, Springer-Verlag, New York, 2006.
4. K. Maleknejad and F. Mirzaee, *Using rationalized Haar wavelet for solving linear integral equations*, Appl. Math. Comput. **160** (2005) 579–587.
5. F. Mirzaee and N. Samadyar, *Using radial basis functions to solve two dimensional linear stochastic integral equations on non-rectangular domains*, Eng. Anal. Bound. Elem. **92** (2018) 180–195.

E-mail: nas.samadyar@gmail.com

Construction of a New Family of Optimal Fourth Order Methods without Derivative for Solving Nonlinear Equations

Samaneh Saneifar*
Department of Mathematics, Yazd University, Yazd, Iran
and Mohammad Heydari
Department of Mathematics, Yazd University, Yazd, Iran

Abstract. In this investigation, a new one-parameter family of derivative free two-point methods of the optimal order four to find simple roots of nonlinear equations is proposed and analyzed. The new scheme is constructed using the idea of rational interpolation. Several numerical examples are given to illustrate the performance of the presented method.
Keywords: Derivative free method, Rational interpolation, Order of convergence, Iterative methods.
AMS Mathematical Subject Classification [2010]: 65H05, 65B99, 65G30.

1. Introduction

There are several iterative methods for the numerical solution of nonlinear equations, see for example Ostrowski [5], Traub [9], Petković et al. [6], Heydari et al. [1] and references therein. Let α be a simple real root of a real function $f : D \subset \mathbb{R} \to \mathbb{R}$ and let x_0 be an initial approximation to α.

Newton's method is the well-known iterative method for finding simple root α and it is given by

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad k = 0, 1, 2, \ldots. \tag{1}$$

The sequence of successive iterates $\{x_k\}$ generated from (1) converges quadratically to α. According to the conjecture of Kung and Traub [3], the order of convergence of any multipoint method requiring $n+1$ evaluations cannot exceed the bound 2^n. Multipoint methods with this property are called optimal methods.

Kung and Traub obtained two-point method of fourth order [3],

$$\begin{cases} y_k = x_k - \dfrac{f(x_k)}{f'(x_k)}, \\ x_{k+1} = y_k - \dfrac{f(x_k)^2 f(y_k)}{f'(x_k)(f(y_k) - f(x_k))^2}, \end{cases} \quad k = 0, 1, 2, \ldots. \tag{2}$$

*Presenter

Another fourth order method is proposed by Ostrowski [5] as follows:

$$\begin{cases} y_k = x_k - \dfrac{f(x_k)}{f'(x_k)}, \\ x_{k+1} = y_k - \dfrac{f(y_k)}{f'(x_k)} \dfrac{f(x_k)}{f(x_k) - 2f(y_k)}, \end{cases} \qquad k = 0, 1, 2, \ldots. \tag{3}$$

The iterative methods (2) and (3) require two evaluations of the given function and one of its first derivative per iteration and therefore these methods support the Kung-Traub conjecture.

In recent years, some fourth order iterative methods without derivative have been proposed and analyzed for solving nonlinear equations [4, 7]. Recently Ren et al. [7] proposed the following one-parameter family of derivative free two-point methods of the four order,

$$\begin{cases} y_k = x_k - \dfrac{f(x_k)}{f[x_k, z_k]}, \\ x_{k+1} = y_k - \dfrac{f(y_k)}{f[x_k, y_k] + f[y_k, z_k] - f[x_k, z_k] + a(y_k - x_k)(y_k - z_k)}, \end{cases} \tag{4}$$

where $k = 0, 1, 2, \ldots$, $z_k = x_k + f(x_k)$, $f[x_k, y_k] = \dfrac{f(x_k) - f(y_k)}{x_k - y_k}$ is divided difference and a is a real parameter.

Liu et al. [4] have shown that the derivative free two-point method

$$\begin{cases} y_k = x_k - \dfrac{f(x_k)}{f[z_k, x_k]}, \\ x_{l+1} = y_k - \dfrac{f(y_k)(f[x_k, y_k] - f[y_k, z_k] + f[x_k, z_k])}{(f[x_k, y_k])^2}, \end{cases} \tag{5}$$

where $k = 0, 1, 2, \ldots$ has order four.

The main aim of this paper is to present a new two-point method without derivative based on rational interpolation. In this technique the optimal order of convergence will be achieved by only three function evaluations.

2. Methodology and Convergence Analysis

We start from the fourth order Steffensen-Newton scheme [6]

$$\begin{cases} y_k = x_k - \dfrac{\lambda f(x_k)^2}{f(x_k + \lambda f(x_k)) - f(x_k)}, \\ x_{k+1} = y_k - \dfrac{f(y_k)}{f'(y_k)}, \end{cases} \qquad k = 0, 1, 2, \ldots, \tag{6}$$

and replacing $f'(y_k)$ by a suitable approximation which does not require new information. Here, we apply the following rational interpolation formula [8]

$$q(x) = \phi(x_k) + \frac{x - x_k}{\phi(x_k, y_k) + \dfrac{x - y_k}{\phi(x_k, y_k, z_k)}}, \tag{7}$$

where the inverse differences $\phi(x_k)$, $\phi(x_k, y_k)$ and $\phi(x_k, y_k, z_k)$ are determined from the conditions

$$q(x_k) = f(x_k), \quad q(y_k) = f(y_k), \quad q(z_k) = f(z_k), \quad z_k = x_k + \lambda f(x_k). \tag{8}$$

According to (8) we find

$$\phi(x_k) = f(x_k), \quad \phi(x_k, y_k) = \frac{1}{f[x_k, y_k]}, \tag{9}$$
$$\phi(x_k, y_k, z_k) = \frac{(y_k - z_k) f[x_k, y_k] f[x_k, z_k]}{f[x_k, z_k] - f[x_k, y_k]}.$$

Now, by substituting (9) in (7), we have

$$f'(y_k) \approx q'(y_k) = f[x_k, y_k] - (f[x_k, y_k])^2 \Big(\frac{1}{f[x_k, y_k]} - \frac{1}{f[x_k, z_k]}\Big)\Big(\frac{y_k - x_k}{y_k - z_k}\Big).$$

Replacing this approximation of $f'(y_k)$ in the second step of (6), we obtain

$$\begin{cases} y_k = x_k - u(x_k), \\ x_{k+1} = y_k - \dfrac{f(y_k)}{f[x_k, y_k] - (f[x_k, y_k])^2 \Big(\frac{1}{f[x_k, y_k]} - \frac{1}{f[x_k, z_k]}\Big)\Big(\frac{u(x_k)}{u(x_k) + \lambda f(x_k)}\Big)}, \end{cases} \tag{10}$$

where $k = 0, 1, 2, \ldots$, $\lambda \in \mathbb{R} - \{0\}$ and $u(x_k) = \dfrac{f(x_k)}{f[x_k, z_k]}$. According to the above analysis, we can provide the following convergence theorem.

THEOREM 2.1. *Let $\alpha \in I_f \subset D$ be a simple zero of a sufficiently differentiable function $f : D \subset \mathbb{R} \to \mathbb{R}$ for an open interval I_f. If x_0 is sufficiently close to α, then the new method defined by (10) is of fourth order, and satisfies the error relation*

$$\varepsilon_{k+1} = [(f'(\alpha)\lambda + 1)^2 c_2 (2c_2^2 - c_3)]\varepsilon_k^4 + O(\varepsilon_k^5),$$

where $\varepsilon_k = x_k - \alpha$ and $c_k = \dfrac{f^{(k)}(\alpha)}{k! f'(\alpha)}$, $k = 1, 2, \ldots$.

PROOF. Let α be a simple zero of f, $\tilde{\varepsilon}_k = y_k - \alpha$ and $d_k = z_k - \alpha$. Using the Taylor expansion around α, we obtain

$$f(x_k) = f'(\alpha)[\varepsilon_k + c_2\varepsilon_k^2 + c_3\varepsilon_k^3 + O(\varepsilon_k^4)],$$
$$f(y_k) = f'(\alpha)[\tilde{\varepsilon}_k + c_2\tilde{\varepsilon}_k^2 + c_3\tilde{\varepsilon}_k^3 + O(\tilde{\varepsilon}_k^4)],$$
$$f(z_k) = f'(\alpha)[d_k + c_2 d_k^2 + c_3 d_k^3 + O(d_k^4)],$$

$$\begin{aligned} u(x_k) &= \varepsilon_k - [f'(\alpha)\lambda + 1]c_2\varepsilon_k^2 \\ &+ [f'(\alpha)^2(c_2^2 - c_3)\lambda^2 + 2f'(\alpha)(c_2^2 - \frac{3}{2}c_3)\lambda + 2c_2^2 - 2c_3]\varepsilon_k^3 + O(\varepsilon_k^4), \end{aligned} \tag{11}$$

and therefore, we can get

$$\tilde{\varepsilon}_k = [f'(\alpha)\lambda + 1]c_2\varepsilon_k^2 + [-f'(\alpha)^2(c_2^2 - c_3)\lambda^2 - 2f'(\alpha)(c_2^2 - \frac{3}{2}c_3)\lambda - 2c_2^2 + 2c_3]\varepsilon_k^3 + O(\varepsilon_k^4),$$
$$d_k = [1 + \lambda f'(\alpha)]\varepsilon_k + \lambda f'(\alpha)c_2\varepsilon_k^2 + \lambda f'(\alpha)c_3\varepsilon_k^3 + O(\varepsilon_k^4).$$

Also, by using Taylor expansion, we have

$$\begin{aligned} f[x_k, z_k] &= f'(\alpha) + f'(\alpha)[f'(\alpha)\lambda + 2]c_2\varepsilon_k \\ &+ f'(\alpha)[f'(\alpha)^2\lambda^2 c_3 + \lambda(c_2^2 + 3c_3)f'(\alpha) + 3c_3]\varepsilon_k^2 \\ &+ [(f'(\alpha)^2\lambda^2 c_4 + 2\lambda(c_2c_3 + c_4)f'(\alpha) + 2c_4)(f'(\alpha)\lambda + 2)f'(\alpha)]\varepsilon_k^3 \\ &+ O(\varepsilon_k^4), \end{aligned} \tag{12}$$

$$
\begin{aligned}
(13) \qquad f[x_k, y_k] &= f'(\alpha) + f'(\alpha)c_2\varepsilon_k + f'(\alpha)[(f'(\alpha)\lambda + 1)c_2^2 + c_3]\varepsilon_k^2 \\
&- f'(\alpha)[(f'(\alpha)^2\lambda^2 + 2f'(\alpha)\lambda + 2)c_2^3 \\
&- (f'(\alpha)^2\lambda^2 c_3 + 4f'(\alpha)\lambda c_3 + 3c_3)c_2 - c_4]\varepsilon_k^3 + O(\varepsilon_k^4).
\end{aligned}
$$

Substituting (11)-(13) into (10), we get the following error equation:

$$\varepsilon_{k+1} = [(f'(\alpha)\lambda + 1)^2 c_2(2c_2^2 - c_3)]\varepsilon_k^4 + O(\varepsilon_k^5).$$

This means that the method defined by (10) is of fourth order. This proof is completed. □

3. Numerical Experiments

In this section, we present some numerical experiments using the presented method (10) and compare the obtained results to Kung and Traub method (2), Ostrowski method (3), Ren et al. method (4) and Liu et al. method (5). All computation were done using the MAPLE package using 256 digit floating point arithmetic. Here, we used the following test functions [2, 6]

$$
\begin{aligned}
f_1(x) &= (x-2)(x^{10} + x + 1)e^{-x-1}, \ \alpha = 2, \\
f_2(x) &= (2 + x^3)\cos(\frac{\pi x}{2}) + \log(x^2 + 2x + 2), \ \alpha = -1, \\
f_3(x) &= xe^x + \log(1 + x + x^4), \ \alpha = 0.
\end{aligned}
$$

Tables 1-3 display the absolute values of the errors of approximations x_k in the first four iterations. The numerical results presented in Tables 1-3 show that the proposed methods in this contribution have better performance as compared with the methods (2)-(5).

TABLE 1. Numerical results for $f_1(x)$ with $x_0 = 2.1$.

error	(2)	(3)	(4) $a = 0$	(5)	(10) $\lambda = -0.01$
$\lvert x_1 - \alpha\rvert$	3.45(−3)	1.72(−3)	2.66(−2)	2.66(−2)	3.09(−4)
$\lvert x_2 - \alpha\rvert$	1.36(−8)	3.13(−10)	2.09(−3)	2.10(−3)	2.15(−13)
$\lvert x_3 - \alpha\rvert$	3.38(−30)	3.49(−37)	1.26(−6)	1.31(−6)	5.14(−50)
$\lvert x_4 - \alpha\rvert$	1.31(−116)	5.43(−145)	2.53(−19)	3.02(−19)	1.66(−196)

TABLE 2. Numerical results $f_2(x)$ with $x_0 = -0.93$.

error	(2)	(3)	(4) $a = 0$	(5)	(10) $\lambda = -0.64$
$\lvert x_1 - \alpha\rvert$	7.01(−4)	4.87(−4)	1.94(−3)	2.12(−3)	9.66(−5)
$\lvert x_2 - \alpha\rvert$	2.58(−11)	3.37(−12)	5.40(−9)	9.96(−9)	3.03(−19)
$\lvert x_3 - \alpha\rvert$	4.78(−41)	7.83(−45)	3.39(−31)	5.16(−30)	2.57(−77)
$\lvert x_4 - \alpha\rvert$	5.66(−160)	2.28(−175)	5.31(−120)	3.71(−115)	5.00(−256)

TABLE 3. Numerical results $f_3(x)$ with $x_0 = -0.5$.

error	(2)	(3)	(4) $a = 0$	(5)	(10) $\lambda = -0.5$
$\lvert x_1 - \alpha\rvert$	$6.52(-2)$	$3.80(-2)$	1.20	0.68	$4.66(-5)$
$\lvert x_2 - \alpha\rvert$	$1.57(-6)$	$2.16(-7)$	1.12	1.14	$4.65(-29)$
$\lvert x_3 - \alpha\rvert$	$4.47(-25)$	$1.92(-28)$	1.12	1.12	$4.59(-173)$
$\lvert x_4 - \alpha\rvert$	$2.91(-99)$	$1.21(-112)$	1.12	1.12	$1.43(-342)$

References

1. M. Heydari, S. M. Hosseini and G. B. Loghmani, *On two new families of iterative methods for solving nonlinear equations with optimal order*, Appl. Anal. Discrete Math. **5** (2011) 93–109.
2. M. Junjua, F. Zafar and N. Yasmin, *Optimal derivative-free root finding methods based on inverse interpolation*, Mathematics **7** (2019) 164–173.
3. H. T. Kung and J. F. Traub, *Optimal order of one-point and multipoint iteration*, J. ACM **21** (1974) 643–651.
4. Z. Liu, Q. Zheng and P. Zhao, *A variant of Steensens method of fourth-order convergence and its applications*, Appl. Math. Comput. **216** (2010) 1978–1983.
5. A. M. Ostrowski, *Solution of Equations and Systems of Equations*, Academic Press, New York, 1960.
6. M. S. Petković, B. Neta, L. D. Petković and J. Dzunic, *Multipoint Methods for Solving Nonlinear Equations*, Elsevier, Amsterdam, 2012.
7. H. Ren, Q. Wu and W. Bi, *A class of two-step Steensen type methods with fourth-order convergence*, Appl. Math. Comput. **209** (2009) 206–210.
8. J. Stoer, *Introduction to Numerical Analysis*, Springer, New York, 2002.
9. J. F. Traub, *Iterative Methods for the Solution of Equations*, Prentice-Hall, Englewood Clis, New Jersey, 1964.

E-mail: samane.saneifar@gmail.com
E-mail: m.heydari@yazd.ac.ir

An Operational Matrix Based-Method Using the Barycentric Basis Functions to Solve the Model of HIV Infection of CD4$^+$ T-cells

Soraya Torkaman*
Department of Mathematics, Yazd University, Yazd, Iran
Ghasem Barid Loghmani
Department of Mathematics, Yazd University, Yazd, Iran
and Mohammad Heydari
Department of Mathematics, Yazd University, Yazd, Iran

Abstract. In this study, a class of nonlinear ordinary differential equation systems that arising in the HIV infection model of CD4$^+$ T-cells is approximated by the numerical method based on operational matrices of the barycentric rational basis functions. Applying the proposed method, the nonlinear governing ordinary differential equations are reduced to a system of nonlinear algebraic equations. In the end, the efficiency of the proposed method is illustrated with some numerical examples and compared with some existing numerical methods.
Keywords: HIV Infection of CD4$^+$ T-cells, Barycentric rational basis functions, Operational matrices.
AMS Mathematical Subject Classification [2010]: 13F55, 05E40, 05C65.

1. Introduction

Most of the practical problems arising in physics, chemistry and biology can be described as time dependent phenomena. Mathematical models forthese phenomena often lead to nonlinear differential equations. In the present research, we study a system of nonlinear differential equations that predict the spread of HIV infection. The HIV infection in a human body depends on three major components T, I and V that represent the concentration of susceptible CD4$^+$ T-cells, CD4$^+$ T-cells infected by the HIV virus and free HIV virus particles in the blood at time t, respectively. So, their acts can be mathematically described as follows [1]

$$\begin{cases} \frac{dT}{dt} = q - \alpha T + rT(1 - \frac{T+I}{T_{max}}) - \kappa VT, \\ \frac{dI}{dt} = \kappa VT - \beta I, \\ \frac{dV}{dt} = \mu\beta I - \gamma V, \end{cases} \tag{1}$$

with the initial conditions:

$$T(0) = r_1,\ I(0) = r_2,\ V(0) = r_3,\ 0 \le t \le R. \tag{2}$$

The parameters α, β and γ stand for natural turnover rates of uninfected T-cells, infected T-cells and virusparticles, respectively, T_{max} is the maximum level of CD4$^+$ T-cells, $\kappa > 0$ is the infection rate, μ is the number of virus particles that

*Presenter

produced by each infected CD4$^+$ T-cells, q is a constant rate to produce CD4$^+$ T-cells, and r is a rate of T-cells multiplication.

2. Linear Barycentric Rational Interpolation

Let $f : [a, b] \to \mathbb{R}$, $n \in \mathbb{N}$, $\{t_j\}_{j=0}^n$ be a set of equidistant nodes, where $t_j = a + jh$, $j = 0, 1, \ldots, n$, $hn = b - a$ and $\{f(t_j)\}_{j=0}^n$ be the corresponding $n + 1$ given data. A linear barycentric rational interpolation for this data can be written as follows:

$$r_n(t) = \sum_{j=0}^{n} f(t_j)\phi_j(t),\ \ \phi_j(t) = \Big(\sum_{i=0}^{n} \frac{w_i}{t - t_i}\Big)^{-1} \frac{w_j}{t - t_j},\ \ j = 0, 1, \ldots, n, \tag{3}$$

where $\{w_j\}_{j=0}^n$ are arbitrary nonzero set of weights. The basis functions $\{\phi_j(t)\}_{j=0}^n$ have the following properties:

- **Partition of Unity:** $\sum_{j=0}^n \phi_j(t) = 1$.
- **Lagrange Property:** $\phi_j(t_i) = \delta_{ij},\ \ i, j = 0, 1, \ldots, n$, where δ_{ij} is the Kronecker delta.

Let $0 \le d \le n$ be an integer parameter. Floater and Hormann [2] proposed a family of rational interpolants based on the barycentric weights w_j as follows:

$$w_j = (-1)^j \sum_{i=\max(j-d,0)}^{\min(j,n-d)} \Big(\prod_{k=i,k\neq j}^{i+d} \frac{1}{|\, t_j - t_k \,|} \Big), \qquad j = 0, 1, \ldots, n. \tag{4}$$

By choosing the interpolation weights (4), interpolant (3) has no real poles [2]. Due to the Lagrange property of the basis functions (3), any function f on the interval $[a, b]$ can be approximated as:

$$f(t) \simeq \sum_{j=0}^{n} f(t_j)\phi_j(t). \tag{5}$$

By considering the barycentric weights (4), the approximation error of (5) can be estimated as follows:

THEOREM 2.1. [2] *Consider $f \in C^{d+2}[a, b]$. Then*

$$\| f(t) - \sum_{j=0}^{n} f(t_j)\phi_j(t) \|_\infty \le \begin{cases} h^{d+1}(1 + \gamma\mu)(b - a)\frac{\|f^{(d+2)}\|_\infty}{d+2}, & (n - d)\ odd, \\ h^{d+1}(1 + \gamma\mu)\Big((b - a)\frac{\|f^{(d+2)}\|_\infty}{d+2} + \frac{\|f^{(d+1)}\|_\infty}{d+1}\Big), & (n - d)\ even, \end{cases}$$

where

$$\gamma = \begin{cases} 1, & d = 0, \\ 0, & d \ge 1, \end{cases}$$

$$h = \max_{0\le i\le n-1}(t_{i+1} - t_i),\ \ \mu = \max_{1\le i\le n-2} \min\left(\frac{t_{i+1} - t_i}{t_i - t_{i-1}}, \frac{t_{i+1} - t_i}{t_{i+2} - t_{i+1}}\right).$$

3. Operational Matrices of Integration and Product

In this section, the operational matrices of integration and product for the barycentric rational basis functions are determined. Let $\{\phi_j(t)\}_{j=0}^n$ be the basis functions in (3) and $\Phi(t)$ be a $(n + 1) \times 1$ vector as follows:

$$\Phi(t) = [\phi_0(t), \phi_1(t), \ldots, \phi_n(t)]^T. \tag{6}$$

LEMMA 3.1. Let $\Phi(t)$ be the vector defined in (6), then

$$\int_a^t \Phi(s)ds \simeq \mathbf{P}\Phi(t),$$

where $\mathbf{P} = (p_{ij})$ is the $(n+1)\times(n+1)$ operational matrix of integration in which p_{ij} can be computed as:

$$p_{ij} = \int_a^{t_j} \phi_i(s)ds,\ \ i,j = 0,1,\ldots,n.$$

PROOF. It is clear that

$$\int_a^t \Phi(s)ds = \Big[\int_a^t \phi_0(s)ds, \int_a^t \phi_1(s)ds, \ldots, \int_a^t \phi_n(s)ds\Big]^T. \tag{7}$$

By using (5), any components of the vector (7) can be approximated as:

$$\int_a^t \phi_i(s)ds \simeq \sum_{j=0}^{n} p_{ij}\phi_j(t),\ \ i = 0,1,\ldots,n. \tag{8}$$

Hence, by substituting (8) in (7), the desirable result can be obtained. □

LEMMA 3.2. Let $F = [f_0, f_1, \ldots, f_n]^T$ be a column vector. Then

$$\Phi(t)\Phi^T(t)F \simeq \tilde{F}\Phi(t),$$

where $\tilde{F}$ is a $(n+1)\times(n+1)$ product operational matrix as $\tilde{F} = diag[f_0, f_1, \ldots, f_n]$.

PROOF. The proof is similar to [3, Lemma 3.3]. □

4. Description of the Numerical Method

In this section, a numerical approach based on the operational matrices of integration and product for solving the nonlinear system of ODEs (1) is described. For this purpose, by using (5), we approximate the functions $\frac{dT}{dt}$, $\frac{dI}{dt}$ and $\frac{dV}{dt}$ as

$$\frac{dT}{dt} \simeq \tilde{T}^T\Phi(t), \qquad \frac{dI}{dt} \simeq \tilde{I}^T\Phi(t), \qquad \frac{dV}{dt} \simeq \tilde{V}^T\Phi(t), \tag{9}$$

where

$$\tilde{T} = [T_0, T_1, \ldots, T_n]^T,\ \ \tilde{I} = [I_0, I_1, \ldots, I_n]^T,\ \ \tilde{V} = [V_0, V_1, \ldots, V_n]^T,$$

are unknown vectors. By integrating from (9) on the interval $[0,t]$, using Lemma 3.1 and the initial conditions (2), one can get

$$T(t) \simeq \tilde{T}^T\mathbf{P}\Phi(t) + r_1,\ \ I(t) \simeq \tilde{I}^T\mathbf{P}\Phi(t) + r_2,\ \ V(t) \simeq \tilde{V}^T\mathbf{P}\Phi(t) + r_3.$$

Again, by applying (5), the functions $T(t)$, $I(t)$, $V(t)$ and q can be approximated as

$$T(t) \simeq \mathcal{T}\Phi(t), \qquad I(t) \simeq \mathcal{I}\Phi(t), \qquad V(t) \simeq \mathcal{V}\Phi(t), \qquad q \simeq \mathcal{Q}^T\Phi(t), \tag{10}$$

where

$$\mathcal{T} = \tilde{T}^T\mathbf{P} + R_1^T, \quad \mathcal{I} = \tilde{I}^T\mathbf{P} + R_2^T, \quad \mathcal{V} = \tilde{V}^T\mathbf{P} + R_3^T,$$

TABLE 1. Comparison between the proposed method and other numerical results on the interval $[0, 1]$.

$T(t)$	Numerical	OBCM [1]	LWM [4]	PM	$E_T(t)$
0.2	0.2088080786	0.2129281262	0.2088073215	0.2088079910	$8.75e-08$
0.6	0.7644237744	0.7757846339	0.7641476415	0.7644231669	$6.07e-07$
1.0	2.5915941109	2.7432245704	2.5571462314	2.5915926461	$1.46e-06$
$I(t)$	Numerical	OBCM [1]	LWM [4]	PM	$E_I(t)$
0.2	$0.6032701156e-5$	$0.5903681847e-5$	$0.6032704663e-5$	$0.6032695618e-5$	$5.54e-12$
0.6	$0.2122378271e-4$	$0.2123339357e-4$	$0.2112628765e-4$	$0.2122376415e-4$	$1.86e-11$
1.0	$0.4003780390e-4$	$0.3943044147e-4$	$0.3287654321e-4$	$0.4003777698e-4$	$2.69e-11$
$V(t)$	Numerical	OBCM [1]	LWM [4]	PM	$E_V(t)$
0.2	0.0618798419	0.0616038027	0.06187990765	0.0618798413	$6.15e-10$
0.6	0.0237045487	0.0236278850	0.02381098734	0.0237045499	$1.22e-09$
1.0	0.0091008437	0.0081082206	0.01605042314	0.0091008460	$2.25e-09$

and R_1, R_2, R_3 and $\mathcal{Q}$ are $(n+1)\times 1$ vectors as follows

$$R_1 = [r_1, r_1, \ldots, r_1]^T,\ R_2 = [r_2, r_2, \ldots, r_2]^T,$$
$$R_3 = [r_3, r_3, \ldots, r_3]^T,\ \mathcal{Q} = [q, q, \ldots, q]^T.$$

According to (10) and employing Lemma 3.2, we obtain

$$V(t)T(t) \simeq \mathcal{V}\Phi(t)\mathcal{T}\Phi(t) = \mathcal{V}\Phi(t)\Phi^T(t)\mathcal{T}^T = \mathcal{V}\mathcal{B}_1\Phi(t), \tag{11}$$
$$T(t)T(t) \simeq \mathcal{T}\Phi(t)\mathcal{T}\Phi(t) = \mathcal{T}\Phi(t)\Phi^T(t)\mathcal{T}^T = \mathcal{T}\mathcal{B}_1\Phi(t), \tag{12}$$
$$I(t)T(t) \simeq \mathcal{I}\Phi(t)\mathcal{T}\Phi(t) = \mathcal{I}\Phi(t)\Phi^T(t)\mathcal{T}^T = \mathcal{I}\mathcal{B}_1\Phi(t), \tag{13}$$

where $\mathcal{B}_1$ is $(n+1)\times(n+1)$ diagonal matrix. By substituting (9), (10) and (11)-(13) in the nonlinear system of ordinary differential equations (1) and canceling $\Phi(t)$, we obtain the nonlinear system of algebraic equations as

$$\begin{cases} \tilde{T}^T - \mathcal{Q} + (\alpha - r)\mathcal{T} + \frac{r}{T_{max}}(\mathcal{T}+\mathcal{I})\mathcal{B}_1 + \kappa\mathcal{V}\mathcal{B}_1 = 0, \\ \tilde{I}^T - \kappa\mathcal{V}\mathcal{B}_1 + \beta\mathcal{I} = 0, \\ \tilde{V}^T - \mu\beta\mathcal{I} + \gamma\mathcal{V} = 0. \end{cases} \tag{14}$$

By solving the nonlinear system of algebraic equations (14), the unknown vectors $\tilde{T}$, $\tilde{I}$ and $\tilde{V}$ are determined and by substituting in (9), the unknown functions $T(t)$, $I(t)$ and $V(t)$ are computed.

5. Numerical Results and Conclusion

In this section, we apply the proposed method (PM) to approximate the solution of the system of ODEs (1) with the following parameters:

$$r_1 = r_3 = 0.1, r_2 = 0, \alpha = 0.02, \beta = 0.3, \gamma = 2.4, \mu = 10, \kappa = 0.0027,$$
$$r = 3, R = 2, q = 0.1, T_{max} = 1500.$$

To demonstrate the efficiency of the proposed method, we compare the results obtained by this method with the numerical method based on fourth-order Runge-Kutta method (RK4), orthonormal Bernstein collocation method (OBCM) [1] and Legendre wavelet method (LWM) [4]. Figure 1 illustrates the approximation and numerical results for $T(t)$, $I(t)$ and $V(t)$ with $n = 20$ and $d = 19$. Figure 2 provides the absolute error of the proposed method, where $E_f(t) = |f(t)_{RK4} - f(t)_{PM}|$.

Table 1 gives a comparison between the proposed method, OBCM [1] and LWM [4] for $n = 8$ and $d = 7$.

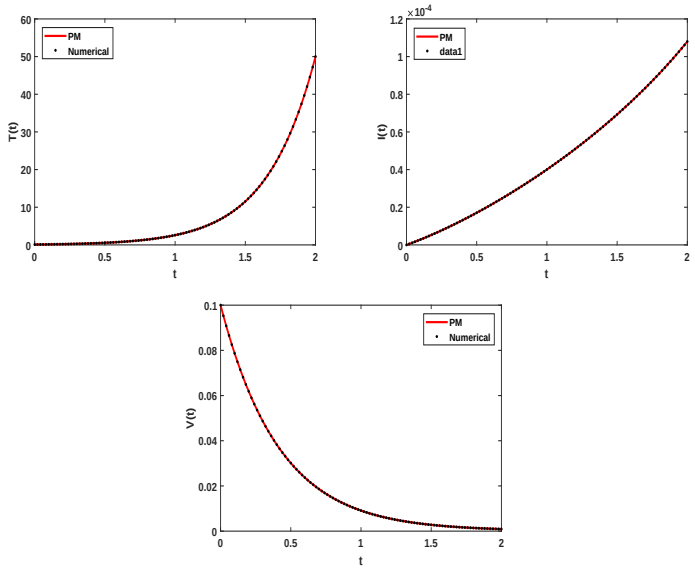


FIGURE 1. Graphs of the approximation and numerical results for $T(t)$, $I(t)$ and $V(t)$ on th interval $[0, 2]$.

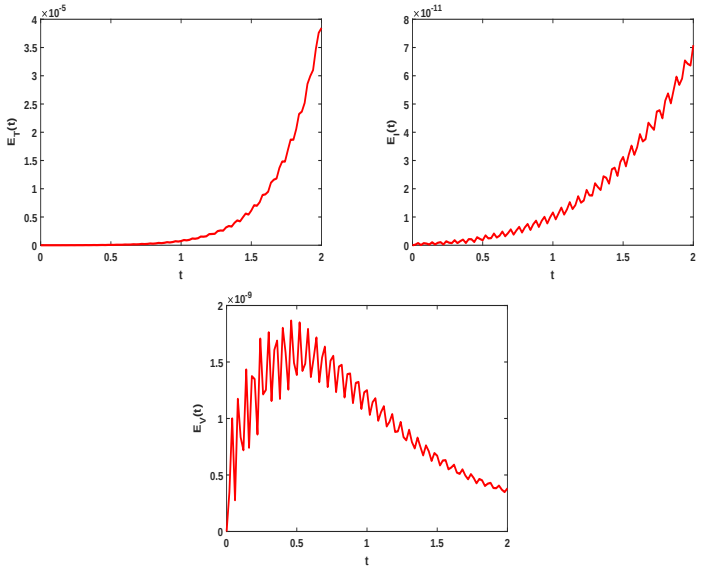


FIGURE 2. Graphs of the absolute error of the proposed method with $n = 20$ and $d = 19$ on the interval $[0, 2]$.

Graphical and tabulated results confirm that the proposed method can be successfully applied for solving the model of HIV Infection of $CD4^+$ T-cells.

References

1. F. Mirzaee and N. Samadyar, *Parameters estimation of HIV infection model of $CD4^+$ T-cells by applying orthonormal Bernstein collocation method*, J. Biomath. **11** (2) (2018). DOI: 10.1142/S1793524518500201
2. M. S. Floater and K. Hormann, *Barycentric rational interpolation with no poles and high rates of approximation*, Numer. Math. **107** (2007) 315–331.
3. M. Heydari, G. B. Loghmani and S. M. Hosseini, *Operational matrices of Chebyshev cardinal functions and their application for solving delay differential equations arising in electrodynamics with error estimation*, Appl. Math. Model. **37** (14–15) (2013) 7789–7809.
4. S. G. Venkatesh, S. R. Balachandar, S. K. Ayyaswamy and K. Balasubramanian, *A new approach for solving a model for HIV infection of $CD4^+$ T-cells arising in mathematical chemistry using wavelets*, J. Math. Chem. **54** (2016) 1072–1082.

E-mail: torkaman@stu.yazd.ac.ir
E-mail: loghmani@yazd.ac.ir
E-mail: m.heydari@yazd.ac.ir

Contributed Posters

Optimization

A Modified Conjugate Gradient Method for Nonsmooth Optimization Problems

Fahimeh Abdollahi*
Department of Mathematics, K. N. Toosi University of Technology, Tehran, Iran
and Masoud Fatemi
Department of Mathematics, K. N. Toosi University of Technology, Tehran, Iran

Abstract. In this paper, we introduce an efficient conjugate gradient method for solving nonsmooth optimization problems by using the Moreau-Yosida regularization approach. The search directions generated by our proposed procedure satisfy the sufficient descent property, and more importantly, belong to a suitable trust region. Our proposed method is globally convergent under mild assumptions. The numerical comparative results on a collection of test problems show the efficiency and superiority of our proposed method.
Keywords: Conjugate gradient method, Nonsmooth optimization, Global convergence.
AMS Mathematical Subject Classification [2010]: 65K10, 65Kxx.

1. Introduction

Consider the following optimization problem

$$\min_{x\in\mathbb{R}^n} f(x). \tag{1}$$

In this paper, we investigate optimization problems like (1), where f is a nonsmooth convex function.

Optimization problems are appeared in many research fields such as engineering, management, economics, medicine, pharmacy, astronomy, etc. They are highly regarded and also there exist many effective ways to solve them. A challenging issue regarding to (1) is solving large-scale nonsmooth problems.

There are a lot of methods for solving (1) with continuously differentiable objective function such as Newton and Quasi-Newton methods, trust region methods and conjugate gradient methods. All the introduced methods find the optimal solution by generating descent directions using exact or inexact line search procedures.

Conjugate gradient methods have been extensively employed by researchers in recent decades due to their strong local and global convergence properties and also low memory requirements for solving large-scale problems. Conjugate gradient iterations are generally defined as follows.

$$x_{k+1} = x_k + \alpha_k d_k, \tag{2}$$

where $\alpha_k > 0$ is a step length and the search direction d_k is computed by

$$d_{k+1} = -g_{k+1} + \beta_k d_k, \quad (d_0 = -g_0),$$

*Presenter

recursively, where $g_k =: \nabla f(x_k)$. The scalar β_k known as the conjugate gradient parameter is indeed one of the most important parameters of these methods. In fact, various choices of β_k lead to different conjugate gradient algorithms. There are many articles that addressed different types of these algorithms and compared them numerically. Although these methods originally have been developed for solving smooth optimization problems, some researchers have recently used them to solve nonsmooth convex optimization problems.

Some well-known methods available for solving nonsmooth convex optimization problems are subgradient and bundle methods. But, our approach here to cast these problems is transforming them to equivalent smooth problems via Moreau-Yosida regularization technique.

Consider

$$\min_{x\in\mathcal{R}^n} F(x), \tag{3}$$

where $F : \mathcal{R}^n \longrightarrow \mathcal{R}$ is the so-called Moreau-Yosida regularization of f, which is defined by

$$F(x) = \min_{z\in\mathcal{R}^n}\{f(z) + \frac{1}{2\lambda}||z - x||^2\}, \tag{4}$$

where λ is a positive parameter. The set of minimizers of (1) coincide with the set of minimizers of (3). Fortunately, F is a differentiable convex function even when the function f is nondifferentiable [5]. There are various iterative methods for solving (3) in many articles.

The good features of conjugate gradient methods for smooth problems encouraged us to modify these methods for nonsmooth problems.

2. A Brief Review

Note that the right hand side of (4) is well-defined in the convex case and while it is strongly convex, it has a unique minimizer which is denoted by

$$p(x) = arg\min_{z\in\mathcal{R}^n}\{f(z) + \frac{1}{2\lambda}||z - x||^2\}.$$

Therefore, F can be expressed by

$$F(x) = f(p(x)) + \frac{1}{2\lambda}||p(x) - x||^2,$$

which its gradient is

$$g(x) = \frac{x - p(x)}{\lambda}.$$

$F(x)$ and $g(x)$ have remarkable properties, as follows.

THEOREM 2.1. [3] *The function F in (4) is finite-valued, convex, everywhere differentiable and its gradient is*

$$g(x) = \frac{x - p(x)}{\lambda}.$$

g is globally Lipschitz continuous with the constant term $\frac{1}{\lambda}$, namely

$$\|g(x_1) - g(x_2)\| \leq \frac{1}{\lambda}||x_1 - x_2||.$$

Finally, the following statements are equivalent.

i) x *is the minimizer of* f.
ii) x *is the minimizer of* F.
iii) $g(x) = 0$.
iv) $x = p(x)$.

In order to provide an efficient conjugate gradient algorithm, Fatemi [1] introduced an optimization problem by combining the three conditions

$$d_{k+1}^T y_k = 0,\ d_i^T g_k = 0\ (i = 0, 1, \ldots, k-1),\ d_k^T g_k \leq \eta ||g_k||^2\ (\eta > 0 \text{ is a constant}),$$

where $y_k := g_{k+1} - g_k$ that are familiar in the linear conjugate gradient theory. The problem was

$$\min_{\beta_k} \big[g_{k+1}^T d_{k+1} + M\big(\big(g_{k+2}^T s_k\big)^2 + \big(d_{k+1}^T y_k\big)^2 \big) \big],$$

where $s_k := x_{k+1} - x_k$ and M is a penalty parameter. By solving this problem and using the secant condition $B_{k+1} s_k = y_k$, a new β_k was proposed as follows.

$$\beta_k = \frac{-1}{2M(1+t^2)} \frac{g_{k+1}^T d_k}{(y_k^T d_k)^2} + \frac{y_k^T g_{k+1}}{y_k^T d_k} - \frac{t}{(1+t^2)} \frac{s_k^T g_{k+1}}{y_k^T d_k}, \tag{5}$$

where $t > 0$ is a suitable approximation of the step length α_k. The author showed that the resulting method is globally converged for general functions and also is highly efficient than some other methods.

3. Preliminary Results

The good features of the method presented in [1], inspired us to modify it for solving nonsmooth problems.

Considering $g_k = \nabla F(x_k)$, we define

$$d_{k+1} = -g_{k+1} + \beta_k^N d_k, \quad (d_0 = -g_0), \tag{6}$$

where

$$\beta_k^N = (y_k - \frac{1}{2M(1+t^2)} \frac{d_k}{T_k} - \frac{t}{1+t^2} s_k)^T \frac{g_{k+1}}{T_k}, \tag{7}$$

and

$$T_k = \max\{\gamma \|d_k\|.\|y_k\|, \|d_k\|.\|s_k\|, |d_k^T y_k|\} \geq 0, \tag{8}$$

for some constant $\gamma > 0$.

Equality (7) is our modified version of (5) suitable for nonsmooth problems as we will show in the following.

LEMMA 3.1. *Consider conjugate gradient iterations based on* (2) *and* (6) *with any step length* $\alpha_k > 0$ *and* β_k *in* (7). *Then, for a positive scalar* $0 < c < 1$, *we have*

$$d_{k+1}^T g_{k+1} \leq -(1-c)||g_{k+1}||^2,$$

where

$$M = \frac{2c}{(1+t^2)\|y_k\|^2}. \tag{9}$$

It is easy to see that replacing (9) in (7) yields

$$\beta_k^N = (y_k - \frac{\|y_k\|^2}{4c}\frac{d_k}{T_k} - \frac{t}{1+t^2}s_k)^T \frac{g_{k+1}}{T_k}. \tag{10}$$

LEMMA 3.2. *For the search direction d_k introduced by (6) and (10), we have*

$$\|d_k\| \leq (2 + \frac{1}{\gamma} + \frac{1}{4c\gamma^2})\|g_k\|.$$

4. Global Convergence

We now sum up the contents of the previous sections to introduce our new conjugate gradient algorithm.

Algorithm 1. New Conjugate Gradient Algorithm (NCG).

- Step 1. Choose a starting point $x_0 \in \mathbb{R}^n$ and a suitable value for positive parameters λ, γ, δ, $0 < \sigma < 1$, $0 < c < 1$ and $0 < \epsilon < 1$. Compute $g_0 = \nabla F(x_0)$, set $d_0 = -g_0$ and $k = 0$.
- Step 2. Check the stopping condition. if $\|g_k\| < \epsilon$ then stop; else go to step 3.
- Step 3. Compute the step length α_k using the following Armijo-type line search.
$$F(x_k + \alpha_k d_k) - F(x_k) \leq \sigma \alpha_k g_k^T d_k,$$
where $\alpha_k = \delta \times 2^{-i_k}$ for $i_k \in \{0, 1, 2, \dots\}$.
- Step 4. Compute $x_{k+1} = x_k + \alpha_k d_k$, $g_{k+1} = \nabla F(x_{k+1})$, $s_k = x_{k+1} - x_k$ and $y_k = g_{k+1} - g_k$.
- Step 5. Compute the conjugate gradient parameter β_k^N using (10) and (8).
- Step 6. Compute the search direction $d_{k+1} = -g_{k+1} + \beta_k^N d_k$.
- Step 7. Set $k = k + 1$ and go to step 2.

In order to prove the global convergence of the Algorithm 1, we consider the following necessary assumptions.

1. The function F is bounded from below.
2. The sequence $\{V_i\}$ is bounded, i.e. there is a constant L such that for each i, $\|V_i\| \leq L$.

LEMMA 4.1. *Let $\{x_k, \alpha_k\}$ be the sequence generated by the Algorithm 1 and the above assumptions 1 and 2 hold. Then, for sufficiently large k, there exists constant $\alpha_0 > 0$ such that*

$$\alpha_k \geq \alpha_0.$$

THEOREM 4.2. *Assume that the conditions in Lemma 4.1 hold. Then we have*

$$\lim_{k \longrightarrow \infty} \inf ||g_k|| = 0.$$

and any accumulation point of x_k is an optimal solution of (1).

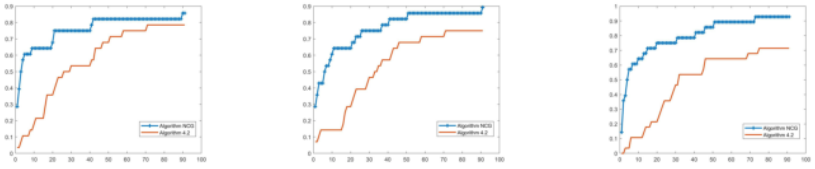
5. Numerical Experiments and Comparisons

In this section, we investigate the numerical performance of the Algorithm 1, by solving some unconstrained nonsmooth test problems and comparing its results by the algorithm 4.2 presented in [6].

The algorithms are terminated if either $||g_k|| \leq 10^{-7}$ or the number of iterations are exceeded 3000.

We considered small and large scale problems reported by [2] and [4] in our numerical tests.

The reported results express that the proposed algorithm can successfully solve all test problems. We can see, by Figure 1, that the Algorithm 1 acts better than the algorithm 4.2 in the sense of Dolan-Moré performance profile. Therefore, it can be introduced as an acceptable and efficient way to solve nonsmooth problems.



(a) Total number of iterations (k).

(b) Total number of function evaluations (i).

(c) CPU time (t).

FIGURE 1. Performance profiles of these methods.

Acknowledgement

The authors thank the Research Council of K.N. Toosi University of Technology for supporting this work.

References

1. M. Fatemi, *A new efficient conjugate gradient method for unconstrained optimization*, J. Comput. Appl. Math. **300** (2016) 207–216.
2. M. Haarala, K. Miettinen and M. M. Mäkelä, *New limited memory bundle method for large-scale nonsmooth optimization*, Optim. Methods Softw. **19** (6) (2004) 673–692.
3. J. B. Hiriart-Urruty and C. Lemaréchal, *Acceleration of the cutting-plane algorithm: Primal forms of bundle methods*, In: Convex Analysis and Minimization Algorithms II, Springer, Berlin, (1993) pp. 275–330.
4. L. Lukšan and J. Vlcek, *Test Problems for Nonsmooth Unconstrained and Linearly Constrained Optimization*, Technical report No. 798, Institute of Computer Science, Academy of Sciences of the Czech Republic Pod vodrenskou v 2, 182 07 Prague 8, Czech Republic, 2000.
5. N. Parikh and S. Boyd, *Proximal algorithms*, Found. Trends Optim. **1** (3) (2014) 127–239.
6. G. Yuan, Z. Meng and Y. Li, *A modified hestenes and stiefel conjugate gradient algorithm for large-scale nonsmooth minimizations and nonlinear equations*, J. Optim. Theory Appl. **168** (1) (2016) 129–152.

E-mail: fabdollahi@email.kntu.ac.ir

E-mail: smfatemi@kntu.ac.ir

Minimal Zero Norm Solution for Quadratic Programming Problem

Hajar Alimorad*
Department of Mathematics, Jahrom University, P. O. Box 74135-111, Jahrom, Iran

Abstract. The Quadratic Programming (QP) is used in many important issues in our lives, such as finance, agriculture, economics, and marketing. So far, a variety methods have been presented to solve this problem and each method has its own advantages and disadvantages. In this article, we will reach the minimal zero norm solution of the non-linear problem equal with QP, using the Karush-Kuhn-Tucker (KKT) method. Since the conditions in KKT method are the sufficient conditions required for solving the problem, with the new method the general optimal would be found. In the last part, there would be numerical examples solved and the results would be compared with other resources, to study the efficiency of the method.
Keywords: Karush-Kuhn-Tucker conditions, Minimal zero norm, Non-linear programming, Quadratic Programming.
AMS Mathematical Subject Classification [2010]: 90C20, 90C29, 65F35, 65F99.

1. Introduction and Introducing the Problem

A Quadratic Programming (QP) would be stated as

$$\begin{aligned} \min \quad & f(X) = C^T X + \tfrac{1}{2} X^T Q X \\ S.\,to: \quad & AX \leq b, \\ & X \geq 0. \end{aligned} \tag{1}$$

In which

$$C = (c_1, c_2, \ldots, c_n), \quad b = (b_1, b_2, \ldots, b_m), \quad X = (x_1, x_2, \ldots, x_n),$$

are vectors and $A_{m\times n}, Q_{n\times n}$ are matrix. Since the constraints in problem (1) are linear, the defined region by these conditions is convex. Therefore, if the objective function $f(X)$ is convex and there is one local solution for problem (1), this solution will be global. If Q is positive definite (or positive semi-definite), the function $f(X)$ is a convex one. When Q is not a positive semi-definite matrix (either indefinite or negative semi-definite), the objective function is nonconvex and may have local minimizers that are not global.

In linear programming if there is one optimal solution, there is always one vertex optima in the admissible region. In QP, the mentioned conditions are not necessary and the optimal solution might be positioned in an inner point of the admissible region.

QP has been very successful for modeling many real life problems. Most applications of QP have been in finance, agriculture, economics, production operations,

*Presenter

marketing and public policy [1, 4, 7]. In economics, it is used in demand-supply response and enterprise selection. In finance, it is used in portfolio analysis, in agriculture, in crop selection [1].

2. Solving Method of QP

Several different methods have been presented for solving problem (1). One of the most important methods is using Karush-Kuhn-Tucker (KKT) theorem.

2.1. KKT Optimality Conditions. Suppose that X is a local optimal solution of the QP such that it satisfies $AX = b, \quad X \geq 0$ and assume that Q is a positive semi-definite matrix. Then, there exist vectors Y, U and V such that the following conditions hold

$$\begin{aligned} & A^T U + QX - Y = C^T, \\ & AX + V = b, \\ & X^T Y + U^T V = 0, \\ & X \geq 0, \quad U \geq 0, \quad Y \geq 0, \quad V \geq 0. \end{aligned} \tag{2}$$

Furthermore, X is a global optimal solution [8].

Two last cases of these conditions are indicative of complementary relationships which are satisfied for $x_j = 0$ or $y_j = 0, \quad (j = 1, 2, \ldots, n)$ (or both of them). Also, which are satisfied for $u_i = 0$ or $v_i = 0 \quad (i = 1, 2, \ldots, m)$ (or both of them).

Problem (2) is solvable through phase I of the two-phase Simplex method. Non-basic variables whose complementary variable is currently basic should not be chosen as an input variable. Interior-Point Method (IPM) is the latter method for solving problem (2).

IPM finds primal-dual solution (X, Y, U, V) by applying variants of Newton's method to the optimality conditions and modifying the search directions and step lengths so that $X \geq 0, \ Y \geq 0, \ U \geq 0$ and $V \geq 0$ are satisfied strictly at every iteration [2, 9]. In this method, the existence of nonnegative constraints creates a difficulty and regarding solutions, complementary condition should be examined at each phase. Moreover, this method starts with a strictly feasible iterate that is not always a trivial task [9].

3. Outline of Work

In this paper, for solving non-linear system (2), a simple and effective method is presented based on minimal zero norm. Then, for examining the effectiveness of the method, the numerical examples solved through previously presented methods, will be solved using the new method. Minimal zero norm solution are often desired in some real applications such as bimatrix game and portfolio selection [6].

Zero norm of a vector is defined by

$$\|X\|_0 = \lim_{p \to 0^+} \|X\|_p = \sum_{i=1}^{n} sign(|x_i|),$$

which is equal to the number of non-zero elements of vector X.

If $p \geq 1$, it is usual to refer to $\|X\|_p$ as the ℓ_p norm of $X \in R^n$. The notation $sign(.)$ represents the operator such that, for any real number a, $sign(a) = 1$ if

$a > 0$, $sign(a) = 0$ if $a = 0$ and $sign(a) = -1$ if $a < 0$ (See [6]). Non-linear system (2), can turn into a problem of non-linear programming which is easily solved by MATLAB software, by introducing a proper objective function. Since in solving the non-linear programming problem, not all the conditions necessarily satisfied and some tolerance is accepted, the objective function must be chosen the way that the non-linear conditions satisfy.

If we consider the problem with the objective function of minimal zero norm of the unknown vectors, the optimal solution would be easily found. Here, considering the complement non-linear conditions and target function of minimal zero norm, the general optimal solution is found. The function $\min \quad \|(X, Y, U, V)\|_0$, minimizes the number of non-zero elements of the unknown vector.

3.1. Suggestions for Continuing the Process. As it was mentioned in the previous section, we can solve the problem by defining the objective function $\min \quad \|(X, Y, U, V)\|_0$. If we can separately consider the objective function as the minimal zero norm for the complement conditions, means, every complement condition is a function of the minimizing zero norm, then we have a problem of multi-objective programming as follow:

$$\begin{aligned} &\min \quad z_i = |sign(x_i)| + |sign(y_i)|, \quad i = 1, 2, \ldots, n, \\ &\min \quad z_j^{'} = |sign(u_j)| + |sign(v_j)|, \quad j = 1, 2, \ldots, m. \end{aligned}$$

The multi-objective programming problem with linear constraints, is solvable with the help of inventive algorithms such as Genetic algorithm. Unfortunately, if the complement conditions are not considered for all constraints, this method would not work for all examples, since there is no effective and efficient method for solving multi-objective programming with non-linear conditions [5]. Usually constraints would be added to target functions using the penalty functions. Finding a method for solving multi-objective problems with linear and non-linear constraints be studied as a suggestion.

4. Numerical Examples

In this section, to study the efficiency of this method, there are examples from the specified reference. These examples are solved using the mentioned method and the answers are satisfying in comparison with previous work.

EXAMPLE 4.1. Solve the following quadratic programming problem [3]:

$$\begin{aligned} &\min \quad f(X) = 15x_1 + 30x_2 + 4x_1x_2 - 2x_1^2 + 4x_2^2 \\ &S.to: \quad x_1 + 2x_2 \leq 30, \\ &\qquad x_1 \geq 0, \quad x_2 \geq 0. \end{aligned}$$

Considering the KKT proposition, the problem would be rewritten as follow:

$$\begin{aligned} &4x_1 - 4x_2 + u_1 - y_1 = 15, \\ &-4x_1 + 8x_2 + 2u_1 - y_2 = 30, \\ &x_1 + 2x_2 + v_1 = 30, \\ &x_iy_i = 0, \quad u_1v_1 = 0, \\ &x_i, y_i, u_1, v_1 \geq 0, \quad i = 1, 2. \end{aligned}$$

By defining the objective function

$$\min \quad Z = sign(|x_1|) + sign(|x_2|) + sign(|y_1|) + sign(|y_2|) + sign(|u_1|) + sign(|v_1|),$$

(the minimal zero norm of the variables) we have a non-linear programming problem which is solvable with MATLAB software and the fmincon formula. The optimal solution of the problem, with the new method is

$$(x_1^*, x_2^*, u_1^*) = (12, 9, 3).$$

This result satisfies in all constraints and complement conditions and has the best value of the objective function.

EXAMPLE 4.2. Solve the following problem [3]:

$$\begin{aligned} \min \quad & f(X) = 4x_1 + 6x_2 - 2x_1^2 - 2x_1x_2 - 2x_2^2 \\ S.to: \quad & x_1 + 2x_2 \leq 2, \\ & x_1 \geq 0, \quad x_2 \geq 0. \end{aligned}$$

Considering KKT constraints and the objective function of zero norm we have the following:

$$\begin{aligned} \min \quad & Z = \|(X, Y, U, V)\|_0 \\ S.to: \quad & 4x_1 + 2x_2 + u_1 - y_1 = 4, \\ & 2x_1 + 4x_2 + 2u_1 - y_2 = 6, \\ & x_1 + 2x_2 + v_1 = 2, \\ & x_i y_i = 0, \quad u_1 v_1 = 0, \\ & x_i, y_i, u_1, v_1 \geq 0, \quad i = 1, 2. \end{aligned}$$

The optimal solution is

$$(x_1^*, x_2^*, u_1^*) = (\frac{1}{3}, \frac{5}{6}, 1).$$

EXAMPLE 4.3. Solve the following problem [8]

$$\begin{aligned} \min \quad & f(X) = -4x_1 + x_1^2 - 2x_1x_2 + 2x_2^2 \\ S.to: \quad & 2x_1 + x_2 \leq 6, \\ & x_1 - 4x_2 \leq 0, \\ & x_1 \geq 0, \quad x_2 \geq 0. \end{aligned}$$

Considering KKT conditions, we have

$$\begin{aligned} \min \quad & Z = \|(X, Y, U, V)\|_0 \\ S.to: \quad & 2x_1 - 2x_2 + 2u_1 + u_2 - y_1 = 4, \\ & -2x_1 + 4x_2 + u_1 - 4u_2 - y_2 = 0, \\ & 2x_1 + x_2 + v_1 = 6, \\ & x_1 - 4x_2 + v_2 = 0, \\ & x_i y_i = 0, \quad u_i v_i = 0, \\ & x_i, y_i, u_i, v_i \geq 0, \quad i = 1, 2. \end{aligned}$$

The optimal solution is

$$(x_1^*, x_2^*, u_1^*, v_2^*) = (\frac{32}{13}, \frac{14}{13}, \frac{8}{13}, \frac{24}{13}).$$

EXAMPLE 4.4. Solve the following problem [8]

$$\begin{aligned} \min \quad & f(X) = -8x_1 - 16x_2 + x_1^2 + 4x_2^2 \\ S.to: \quad & x_1 + x_2 \leq 5, \\ & x_1 \leq 3, \\ & x_1 \geq 0, \quad x_2 \geq 0. \end{aligned}$$

Considering KKT conditions, we have

$$\begin{aligned} \min \quad & Z = \|(X, Y, U, V)\|_0 \\ S.to: \quad & 2x_1 + u_1 + u_2 - y_1 = 8, \\ & 8x_1 + u_1 - y_2 = 16, \\ & x_1 + x_2 + v_1 = 5, \\ & x_1 + v_2 = 3, \\ & x_i y_i = 0, \quad u_i v_i = 0, \\ & x_i, y_i, u_i, v_i \geq 0, \quad i = 1, 2. \end{aligned}$$

The optimal solution is

$$(x_1^*, x_2^*, u_1^*, u_2^*) = (3, 2, 0, 2).$$

5. Conclusion

In this article, a simple and effective method is introduced for solving quadratic programming problem, with the help of KKT constraints and definition of the objective function. In non-linear problems, the result might be approximate. Although, in the new method we have a non-linear programming problem, with the help of KKT constraints, the global optimal result would be found. With regard to the complement conditions in non-linear constraints, the objective function of zero norm, is the most suitable function for solving this problem. The results are completely satisfactory comparing to the others represented for QP problem.

References

1. B. A. McCarl, H. Moskowitz and H. Furtan, *Quadratic programming applications*, Omega **5** (1) (1977) 43–55.
2. G. Cornuéjols and R. Tütüncü, *Optimization Methods in Finance*, Cambridge University Press, New York, 2007.
3. F. Hillier and G. Lieberman, *Introduction to Operation Research*, 3rd ed., Holden-Day, Inc., San Francisco, 1980.
4. J. K. Shim, *A Survey of quadratic programming applications to business and economics*, Int. J. Syst. Sci. (2007) 105–115.
5. J. Matejaš and T. Perić, *A new iterative method for solving multiobjective linear programming problem*, Appl. Math. and Comput. **243** (2014) 746–754.
6. M. Shang, C. Zhang and N. Xiu, *Minimal zero norm solution of linear complementarity problems*, J. Optim. Theory Appl. **163** (2014) 795–814.

7. O. K. Gupta, *Applications of quadratic programming*, J. Info. Optim. Sci. **16** (1) (1995) 177–194.
8. S. S. Rao, *Optimization: Theory and Application*, 2nd ed., Wiley, New York, 1984.
9. S. J. Wright, *Primal-Dual Interior-Point Method*, SIAM, Philadelphia, PA, USA, 1997.

E-mail: h.alimorad@jahromu.ac.ir

A Novel Scaled Conjugate Gradient Method for Large Scale Unconstrained Optimization Problems

Fatemeh Nikzad*
Department of Applied Mathematics, Payame Noor University, Tehran, Iran
Saeed Nezhadhosein
Department of Applied Mathematics, Payame Noor University, Tehran, Iran
and Aghileh Heydari
Department of Applied Mathematics, Payame Noor University, Tehran, Iran

Abstract. Here, a new spectral conjugate gradient method, based on a modified secant equation, is proposed for solving large scale unconstrained optimization problems. The new method has two main features contain sufficient descent and conjugacy conditions, which are essential for the global convergence. Numerical experiments are done on a set of test functions of the CUTEr collection and the results are compared with some well-known methods.
Keywords: Large scale unconstrained optimization (LUO), Spectral conjugate gradient (SCG), Modified secant equations.
AMS Mathematical Subject Classification [2010]: 90C06, 90C26, 65Y20.

1. Introduction

Large scale unconstrained optimization (LUO) problems are famous due to the widespread applications in science and engineering [15], which are formulated as follows:

$$\min_{x\in\mathbb{R}^n} f(x),$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a multivariable continuously differentiable function and its gradient, g, is available. As an important class of iterative methods, scaled or spectral conjugate gradient (SCG) methods are popular for solving UO problems. Starting from an initial solution $x_0 \in \mathbb{R}^n$, we update it by following sequential expression:

$$x_{k+1} = x_k + \alpha_k d_k,$$

where α_k is the step length and d_k is the direction at k-th iteration. The search direction can be calculated by following recursive formula [7]

$$d_{k+1} = -\theta_{k+1} g_{k+1} + \beta_k d_k, \tag{1}$$

where $g_k = g(x_k)$. In Eq. (1), θ_{k+1} and β_k are the scaled and conjugate parameters, respectively, and the initial search direction is set as $d_0 = -g_0$. SCGs are an extend family of the conjugate gradient (CG) methods, which have advantages such as simplicity iterations, low memory requirements and strong global

*Presenter

convergence properties [4]. For $\theta_{k+1} = 1$, the SCG methods are converted to CG methods, and for $\beta_k = 0$, they converted to spectral gradient method (SGM) [14]. Moreover, the step length α_k usually chosen by Wolfe conditions [15], requiring that:

$$f(x_k + \alpha_k d_k) - f(x_k) \leq \delta \alpha_k g_k^T d_k, \tag{2}$$

$$g(x_k + \alpha_k d_k)^T d_k \geq \sigma g_k^T d_k, \tag{3}$$

where $0 < \delta < \sigma < 1$, $f_k = f(x_k)$.

Bergin and Martinze [7], proposed the first SCG algorithm based on SGM, which introduced by Raydan [14] as an extended Barzilai and Borwein idea in [2]. The SGM scaled parameter, which is the inverse of the Raydan quotient and is grounded on the eigenvalue of the average Hessian matrix, is defined as follows [2]

$$\theta_{k+1} = \frac{s_k^T s_k}{s_k^T y_k}, \tag{4}$$

where $s_k = x_{k+1} - x_k$ and $y_k = g_{k+1} - g_k$. The spectral parameter for SGM in Eq. (4), is determined using a two-point approximation of the standard secant equation.

Using the SGM scaled parameter in Eq. (4) and the classic CG parameters, Bergin and Martinz [7] proposed standard SCGs with following CG parameters:

$$\beta_k^{SHS} = \frac{\theta_{k+1} g_{k+1}^T y_k}{y_k^T d_k}, \ \beta_k^{SFR} = \frac{\theta_k g_{k+1}^2}{\theta_{k+1} g_k^2},$$

$$\beta_k^{SPR} = \frac{\theta_k g_{k+1}^T y_k}{\theta_{k+1} \|g_k\|}, \ \beta_k^{SP} = \frac{g_{k+1}^T (\theta_{k+1} y_k - s_k)}{y_k^T d_k}.$$

It is clear that, in the case of $\theta_{k+1} = \theta_k = 1$, these formulas are reduced to the corresponded on the classical CGs [11].

Setting the SCG parameters, (θ_{k+1}, β_k) in Eq. (1), is an interesting issue in literatures, which affects on numerical performance and global convergence properties. These parameters deal with satisfying some essential properties in SCGs, contains descent and conjugacy. There are two groups of strategies to set the SCG parameters. In the first group, the parameters are set such that the search directions satisfies descent or sufficient descent conditions (For instance, see [10, 16]). In the second group, using the secant equation (or modified version of it), the conjugacy conditions are applied to set the SCG parameters (For instance, see the references [1, 6, 10, 13]). Here, using both strategies, we propose a new SCG method, which has both features: sufficient descent and conjugacy.

The remainder of this paper is organized as follows. In Section 2, the new SCG algorithm is proposed. In Section 3, we numerically compare our methods with SCGs in [9].

2. A Modified SCG Method

In this section, inspired by the *JC* SCG method introduced by Jian et al. [9], a modified SCG method is proposed for solving LUO problems, which the conjugate parameter is set based on Dai and Kou (DK) [5] CG family and the scaled parameter is set based on quasi-Newton (QN) approach.

The QN methods are based on secant equation as $B_{k+1}s_k = y_k$, where B_{k+1} is the approximation of the Hessian matrix in x_{k+1}. To improve the quality of these methods modified versions of them are introduced in literatures. For example, Li and Fukushima [12], proposed a modified secant equation as $B_{k+1}s_k = z_k$, where

$$z_k = y_k + h_k||g_k||^r s_k, \quad h_k = D + \max\left\{-\frac{s_k^T y_k}{||s_k||^2}, 0\right\}||g_k||^{-r}, \tag{5}$$

where r is a positive constant parameter. This equation in leads to global convergence property even without convexity assumption on the objective function for QN methods [12]. Moreover, it achieves a high-order accuracy in approximating the second-order curvature of the objective function. Based on the modified secant equation in Eq. (5) and inspired by [9], we set the CG parameter as follows:

$$\beta_k^H = \frac{g_{k+1}^T z_k}{d_k^T z_k} - \frac{\|z_k\|^2 g_{k+1}^T d_k}{(d_k^T z_k)^2}, \tag{6}$$

where z_k defined in (5). The conjugacy parameter β_k^H in Eq. (6) is a especial version of DK conjugate parameter. A main feature of the CG parameter β_k^H is that the corresponding direction has sufficient descent condition property [9]. Also, from Eq. (5), the following inequality holds:

$$s_k^T z_k \geq D||g_k|^r||s_k||^2 > 0,$$

which is necessary to global convergence property. Now, to set the scaled parameter, we use QN search direction as $d_{k+1} = -B_{k+1}^{-1}g_{k+1}$, which with comparing with SCG search direction in Eq. (1) leads to new scaled parameter. Therefore, similar to [9], based on double-truncating technique, which insure both the sufficient descent property and bounded property of the sequence of spectral parameters, the scale parameter is defined as follows:

$$\theta_{k+1}^{M\pm} = \begin{cases} \theta_{k+1}^{M\pm}, & \theta_{k+1}^{M\pm} \in [\frac{1}{4} + \eta, \tau], \\ 1, & otherwise, \end{cases} \tag{7}$$

where $\theta_{k+1}^{M\pm}$ is the notation of θ_{k+1}^{M-} or θ_{k+1}^{M+}, with the following definitions:

$$\theta_{k+1}^{M+} = 1 - \frac{1}{g_{k+1}^T z_k}\left(\frac{\|y_k\|^2 d_k^T g_{k+1}}{d_k^T z_k} - s_k^T g_{k+1}\right), \qquad \theta_{k+1}^{M-} = 1 - \frac{\|y_k\|^2 d_k^T g_{k+1}}{(d_k^T z_k)(g_{k+1}^T z_k)},$$

where τ is a suitable large upper bound and positive η is a suitable small and positive constant. The SCG method with the parameters $(\beta_k^H, \theta_{k+1}^{M\pm})$, is named modified JC method, with the following search direction:

$$d_{k+1} = -\theta_{k+1}^{M\pm} g_{k+1} + \beta_k^H d_k.$$

The global convergence of the modified JC method is very similar to JC methods.

3. Numerical Experiments

In this section, we present some numerical experiments, obtained by applying a MATLAB 8.8.0.1 (R2013a). The numerical results are compared with two version of the JC SCG methods, contain $JC+$ and $JC-$, proposed in [9]. The implementations were performed on a computer, Intel(R) Core (TM) A10-8700P CPU 3.20 Gigahertz 64-bit desktop with 8 Gigabyte RAM. Our experiments have been done

on the test problems of unconstrained optimization problems of CUTEr collection [8].

In all the methods, we used the effective approximate Wolfe conditions described in Eqs. (2)-(3) with parameters $\sigma = 0.9$ and $\delta = 10^{-4}$. For the modified JC method, we use the parameters $D = 10^{-4}$ and $r = 3$, in Eq. (5), and $\eta = 10^{-4}$ and $\tau = 10$, in Eq. (7). Moreover, the same stop condition is considered for all methods which is $\|g_k\|_\infty \leq 10^{-6}$ and the maximum number of iterations is limited to 1000.

The comparing data contain the CPU time and the number of evaluations for function, n_f, and gradient, n_g as $n_f + 3n_g$. To approximately assess the performance of different algorithms, we use the performance profile introduced by Dolan and More [3].

As shown in Figure 1, with respect to CPU time and the number of evaluations for function and gradient, the proposed methods, modified $JC+$ SCG method is the best method among modified $JC-$ and $JC\pm$ SCG methods.

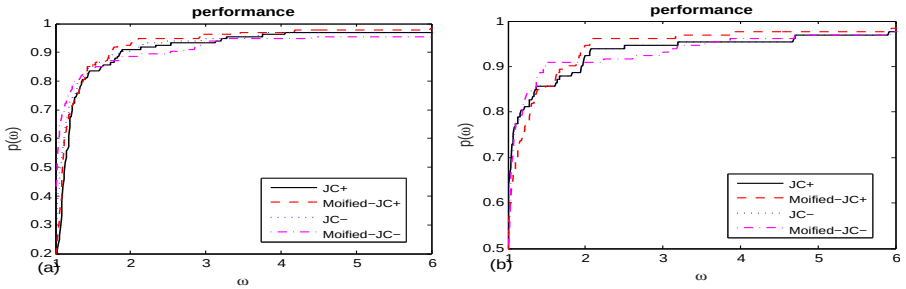


FIGURE 1. Performance profiles based on number iteration for $JC\pm$ and the proposed modified $JC\pm$, (a): CPU time and (b): number of evaluations, $n_f + 3n_g$.

References

1. N. Andrei, *Accelerated scaled memoryless BFGS preconditioned conjugate gradient algorithm for unconstrained optimization*, European J. Oper. Res. **204** (2010) 410 –420.
2. J. Barzilai and J. Borwein, *Two-point step size gradient methods*, IMA J. Numer. Anal. **8** (1988) 141–148.
3. E. D. Dolan and J. J. Moré, *Benchmarking optimization software with performance profiles*, Math. Program **91** (2002) 201–213.
4. Y. H. Dai, J. Han, G. Liu, D. Sun, H. Yin and Y. X. Yuan, *Convergence properties of nonlinear conjugate gradient methods*, Appl. Math. Optim. **10** (2000) 345–358.
5. Y. H Dai and C. X. Kou, *A nonlinear conjugate gradient algorithm with an optimal property and an improved wolfe line search*, SIAM J. Optim. **23** (2013) 296–320.
6. P. Faramarzi and K. Amini, *A modified spectral conjugate gradient method with global convergence*, J. Optim. Theory Appl. **182** (2019) 667–690.
7. E. G. Birgin and J. M. Martínez, *A spectral conjugate gradient method for unconstrained optimization*, Appl. Math. Optim. **43**(2001) 117–128.
8. B. Ingrid, C. Andrew, G. Nicholas and T. Philippe, *CUTE: Constrained and unconstrained testing environment*, ACM Trans. Math. Softw. **21** (1995) 123–160.
9. J. Jian, Q. Chen, X. Jiang, Y. Zeng and J. Yin, *A new spectral conjugate gradient method for large-scale unconstrained optimization*, Optim. Methods Softw. **32** (2017) 503–515.

10. J. K. Liu, Y. M. Feng and L. M. Zou, *A spectral conjugate gradient method for solving large-scale unconstrained optimization*, Comput. Math. Appl. **77** (2019) 731–739.
11. I. E. Livieris and P. Pintelas, *Spectral conjugate gradient methods, Sufficient descent property, Modified secant equation*, J. Comput. Appl. Math. **239** (2013) 396–405.
12. D. H. Li and M. Fukushima, *A modified BFGS method and its global convergence in non-convex minimization*, J. Comput. Appl. Math. **129** (2001) 15–35.
13. J. K. Liu, Y. M. Feng and L. M. Zou, *Scaled conjugate gradient algorithms for unconstrained optimization*, Compu. Optim. Appl. **38** (2007) 401–416.
14. M. Raydan, *The Barzilai and Borwein gradient method for the large scale unconstrained minimization problem*, IAM J. Optim. **7** (1997) 265–292.
15. W. Sun and Y. X. Yuan, *Optimization Theory and Methods Nonlinear Programming*, Springer US, New York, 2006.
16. L. Zhang, W. Zhou and L. Donghui, *Some descent three-term conjugate gradient methods and their global convergence*, Optim. Methods Softw. **22** (2007) 697–711.

E-mail: fatemenikzad2764@gmail.com
E-mail: s_nejhadhosein@pnu.ac.ir
E-mail: a_heidari@pnu.ac.ir

Function Approximation Using Feed-Forward Neural Networks

Saeed Nezhadhosein*
Department of Applied Mathematics, Payame Noor University, Tehran, Iran
and Fatemeh Nikzad
Department of Applied Mathematics, Payame Noor University, Tehran, Iran

Abstract. Here, a three layer backpropagation feed-forward neural network with batch updating approach, is proposed for function approximation. The training process is considered as different conjugate gradient (CG) algorithms. Numerical experiments show that the Fletcher Reeves CG algorithm is the most accurate than other methods.
Keywords: Feed-forward neural network, Function approximation, Conjugate gradient.
AMS Mathematical Subject Classification [2010]: 13F55, 05E40, 05C65.

1. Introduction

Artificial neural networks (ANNs), as a major topic in computational intelligence, have many applications in science and engineering such as pattern recognition, function approximation and classification [3, 5]. As the most broadly case, feed-forward ANNs have been widely used in many researches. Backpropagation (BP) algorithm is the most famous method used for learning these ANNs, which is implemented by two practical ways: batch updating approach and online updating approach [6]. These methods differ in the ways of weight corrections and training schemes of data set.

There are two main groups for training ANNs in BP methods, supervised and unsupervised approaches. Gradient descent method is the most famous approach in supervised methods. The main drawback of gradient descent method is low convergence. To accelerate the convergence, some authors applied second order algorithms in training process of for BP algorithm such as Newton, Levenberg-Marquardt and CG. Function approximation is a well-known problem, which can be solved by ANNs. For example, Yang et al. [7] proposed three types of neural networks contain Radial Basis Function (RBF), BP and Generalized Regression neural network (GRNN) for solving these problems.

Here, using a batch updating approach in BP algorithm, we apply some CGs for training phase. Also similar to [5], to more suitable learning rate in each training epoch the generalized Armijo method is used.

The remainder of this paper is organized as follows. In Section 2, a new ANN algorithm is proposed. In Section 3, we present some numerical experiments for function approximation.

*Presenter

2. A New BP Neural Network Algorithm

In this section a new BP feedforward ANN is proposed for function approximation, which used CGs for training phase. Also, the learning rate is set based on the line search approach proposed in [5], which is a generalized Armijo procedure.

At first, we construct a three layer BP network with m input, n hidden and one output neurons, respectively. Given a data set of size J, as a couple $(X^i, O^i),\ i = 1,\ 2, \ldots, J$, where $X^i \in \mathbb{R}^m$ and $O^i \in \mathbb{R}$ are the input and ideal output of the i-th sample. Let $u \in \mathbb{R}^n$ be the connecting weight vector between hidden and output layers and $V = (w_{ij})_{n\times m}$ be the matrix of weights connecting input and hidden layers. Also we denote the connecting weights between the input layer and the i-th neuron of the hidden layer as a column vector $v_i \in \mathbb{R}^m$, where $v_i = [w_{1i}, w_{2i}, \ldots, w_{mi}]^T$, as the i-th row of the weight matrix V. Now, for simplicity, the all the weights of the network can be combined in a vector as $W = [u^T, v_1^T, v_2^T, \ldots, v_n^T]^T \in \mathbb{R}^{n(m+1)}$. For each input, such as X^i, the output of hidden layer is as $z = G(VX^i)$, where G is a vector valued function as $G(z) = [g(z_1), g(z_2), \ldots, g(z_n)]^T$ and g is a smooth active function. Now, the output of the ANN is evaluated as follows:

$$y = f(u^T z) = f(u^T G(VX)),$$

where f is a real smooth function as an active function of output layer. For the specific weight vector W, the error function, as the mean square error between real and ideal outputs, can be experted as follows:

$$E(W) = \frac{1}{2}\sum_{k=0}^{J}\left(O^k - f(u^T G(VX^k))\right)^2. \tag{1}$$

As a second order algorithm, CG methods can be used to train the ANNs. Using W_0 as a initial weight vector, CGs construct a sequence $\{W_k\}$, as follows:

$$W_{k+1} = W_k + \alpha_k d_k, \tag{2}$$

where α_k is step length or learning rate and d_k is search direction defined by following recursive formula:

$$d_{k+1} = -E_w^{k+1} + \beta_k d_k, \qquad d_0 = -E_w(W_0), \tag{3}$$

where E_w^k is the gradient vector of the error function, defined in Eq. (1), in W_k, which is $E_w^k = E_w(W^k)$ and $E_w = [E_u, E_{v_1}, \ldots, E_{v_n}]^T$. The partial derivations of the error function with respect to u and $v_i,\ i = 1, 2, \ldots, n$ can be calculated as follows:

$$\begin{aligned} E_u &= \sum_{k=0}^{J}(O^k - y^k) f'(u^T G(VX^k)) G(VX^k), \\ E_{v_i} &= \sum_{k=0}^{J}(O^k - y^k) f'(u^T G(VX^k)) u_i g'(v_i^T X^k) X^k. \end{aligned}$$

In Eq. (3), β_k is conjugate parameter of CG, called CG parameter. There are different CG methods which distinguished by definition of CG parameters.

Some of well-known CGs are Fletcher-Reeves (FR) [1], Hestenes-Stiefel (HS) [2], Polak-Ribiere-Polyak (PRP)[4], with following CG parameter:

$$\beta_k^{FR} = \frac{\|E_w^{k+1}\|^2}{\|E_w^k\|^2}, \qquad \beta_k^{PR} = \frac{y_k^T E_w^{k+1}}{\|E_w^k\|^2}, \qquad \beta_k^{HS} = \frac{y_k^T E_w^{k+1}}{y_k^T d_k}. \tag{4}$$

In order to compare the efficiency of different CGs for training ANNs, we apply three FR, PR and HS CGs for them. Therefore, three BP algorithms are introduced named: BPFR, BPPR and BPHS.

To set the learning rate we use the generalized Armijo technique, proposed in [5]. Let μ_1, μ_2 be in $(0,1)$, $\mu_1 \leq \mu_2$, and γ_1 and γ_2 are positive constant values. To set the learning rate α_k, at first a parameter α^* is evaluated so that the following inequality satisfies:

$$E(W^k + \alpha^* d_k) > E(W^k) + \mu_2 \alpha^* d_k^T E_W^k.$$

Next, the learning rate α_k is calculated by following inequalities:

$$E(W_k + \alpha_k d_k) \leq E(W_k) + \mu_1 \alpha_k d_k^T E_W^k,$$
$$\alpha_k \geq \gamma_1, \qquad \text{or} \qquad \alpha_k \geq \gamma_2 \alpha_k^* > 0.$$

Now with the iterative method in (2) and different CGs, as training procedures, we can construct BP algorithms for function approximation. Algorithm 1, shows the steps of three BP algorithms.

Algorithm 1. The BP algorithms (BPFR, BPPR, BPHS).

- **Initialization:** Input the number neurons for hidden layer, n, the number of inputs, m, the number of samples, J, the dataset (X^i, O^i), $i = 1, 2, \ldots, J$, W_0 as a random weight vector and α_0 as arbitrary positive constant.
- Step 1. Let $d_0 = -E_W(W_0)$ and $k = 0$.
- Step 2. Let $W_{k+1} = W_k + \alpha_k d_k$.
- Step 3. Evaluate α_{k+1} with the generalized Armijo procedure.
- Step 4. Let $d_{k+1} = -E_W^{k+1} + \alpha_k d_k$.
- Step 5. Update β_k, based on β_k^{FR}, β_k^{PRP} or β_k^{HS} in Eq. (4).
- Step 6. Let $k = k + 1$.
- Step 7. If stop conditions are not satisfied, go to Step 2.

3. Numerical Experiments

In this section, we present numerical experiments to approximate three test functions, obtained by applying a MATLAB 8.8.0.1 (R2013a). The implementations were performed on a computer, Intel(R) Core (TM) A10-8700P CPU 3.20 Gigahertz 64-bit desktop with 8 Gigabyte RAM. For this purpose, we consider three following functions:

$$f_1(x_1, x_2) = \sin(x_1) + \sin(x_2),$$
$$f_2(x_1, x_2) = x_1 \exp(x_2),$$
$$f_3(x_1, x_2, x_3) = x_1 x_2 \cos(x_3) + x_2 \exp(x_1 x_3).$$

To set up the training data set, 1000 input data are selected uniformly in $[0,1]^d$, where d is the dimension of the function. The structure of the ANN contains input

nodes, 10 hidden nodes and one output node. Also, the active function for the hidden layer is "tanh", while for the output layer, the identical active function is selected. For the initial weights in Algorithm 1, we choose random vector in $(0, 1)$ with normal distribution. The training procedure will be terminated when the number of epochs reached to 1000 or the norm of objective function gradient less than 10^{-3}.

To make clear comparison of BPFR, BPPR and BPHS methods, we graphically plot the norm of gradient and function evaluation in Figure 1. It is shown that with respect to the rate of convergent, BPFR method is the best method among BPHS and BPPR methods.

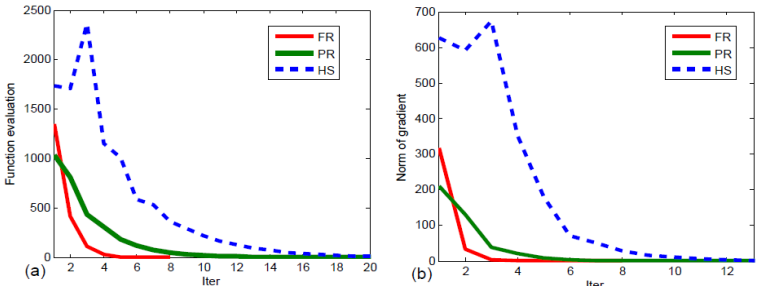


FIGURE 1. Performance profiles of BPFR, BPPR and BPHS methods for function evaluation (a) and norm of function (b).

References

1. R. Fletcher and C. M. Reeves, *Function minimization by conjugate gradients*, Comput. J. **7** (1964) 149–154.
2. M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Research Nat. Bur. Standards **49** (1952) 409–436.
3. A. D. Papalexopoulos, S. Y. Hao and T. M. Peng, *An implementation of a neural network-based load forecasting-model for the EMS*, IEEE Trans, Power Syst. **9** (1994) 1956–1962.
4. E. Polak and G. Ribiere, *Note sur la convergence de mèthodes de directions conjuguèes*, Rev. Franaise Informat. Recherche Oprationnelle **3** (1969) 35–43.
5. J. Wang, B. Zhang, Z. Sun, W. Hao and Q. Sun, *A novel conjugate gradient method with generalized Armijo search for efficient training of feedforward neural networks*, Neurocomputing **275** (2018) 308–316.
6. W. Wu, J. Wang, M. CHeng and Z. Li, *Convergence analysis of online gradient method for BP neural network*, Neural netw. **24** (2011) 91-98.
7. S. Yang, T. O. Ting, K. L. Man and S. Guan, *Investigation of neural networks for function approximation*, Procedia Comput. Sci. **17** (2013) 586–594.

E-mail: fatemenikzad2764@gmail.com
E-mail: s_nejhadhosein@pnu.ac.ir

The Minimax Location Problem with Closest Distance with Circle Demand Regions

Ahmadreza Raeisi Dehkordi*
Faculty of Mathematical Sciences, University of Isfahan, Isfahan, Iran
and Ali Ansari Ardali
Faculty of Mathematical Sciences, University of Shahrekord, Shahrekord, Iran

Abstract. We consider the constrained minimax location problem with closest distance and circle demand regions. Some properties concerning existence and uniqueness of the optimal solution are provided. The existence and uniqueness of the optimal solution are investigated. Moreover, we develop an efficient algorithm for solving this class of problems and we provide its convergence under some mild assumptions.
Keywords: Minimax, Location, Algorithm, Optimality condition, Uniqueness.
AMS Mathematical Subject Classification [2010]: 49J52, 65K10, 90C26.

1. Introduction

Let S be a closed subset in $\mathbb{R}^n$, the *distance function* is defined as $d_S(x) = \min_{s\in S} \|x-s\|$. The mathematical model of the constrained single facility minimax location problem with closest distance is as follows:

$$\min_{x\in S} F(x) := \max_{i\in\mathcal{M}} w_i d_{A_i}(x), \tag{1}$$

where S is a closed set in $\mathbb{R}^n$, w_i $(i \in \mathcal{M})$ are positive weights, and $A_i = \{x \in \mathbb{R}^n \mid \|x-a_i\| \leqslant r_i\}$ $(i \in \mathcal{M})$ are the circles in $\mathbb{R}^n$, called the demand regions. When S is convex (nonconvex), then we say that problem (1) is a convex (nonconvex) problem.

The concept of closest distance between sets is well known in mathematics and have received considerable attention in the facility location problems. Brimberg and Wesolowsky discussed in [2] the case, where the demand regions are some polygons. In particular, they present some applications for their model. In the unconstrained case, a graphical approach is provided in [2]. In addition, two efficient algorithms are investigated in [1] for solving problem (1), when the weights are equal to 1 and the demand regions are some polyhedral sets. In the constrained case, in [5] an algorithm is proposed for finding the optimal solution of problem (1) with planar polyhedral sets. Nobakhtian and Raeisi Dehkordi discussed in [6] the minimum of the sum of weighted Euclidean distances to the closest points of the demand regions and an algorithm is established for solving the proposed problem. An efficient algorithm is developed in [7] for finding the optimal solution set of the rectilinear distance location problem with box constraints.

*Presenter

At the primary contribution of this paper, we consider the single facility minimax location problem (1) and a geometric approach is proposed for seeking the optimal solution. However, finding the optimal solution of a nonconvex optimization problem is very difficult, and the proposed geometric condition intends to present an algorithm for finding an optimal solution in the nonconvex case. Moreover, the global convergence of the presented algorithm is proved. In particular, we show that the optimal solution is unique in the convex case.

The advantages of the proposed algorithm are as follows: (1) it is simple and easy to implement, (2) it solves a simple subproblem in each iteration, (3) it uses a few parameters, (4) it reduces the value of the objective function in each iteration, (5) the sequence generated by the proposed algorithm converges to a Clarke stationary point under mild assumptions.

Our second contribution is to present the existence and uniqueness results.

This paper is organized as follows. Section 2 proposes a necessary and sufficient condition of optimality for problemsfmc. In Section 3, we present a new algorithm for solving the problem in the nonconvex case.

1.1. The Existence and Uniqueness Results. The existence and uniqueness of the optimal solution of optimization problems play a key role for improving the numerical methods and solving these problems.

In the following theorem, we prove the existence of an optimal solution for problem (1) under a compactness hypothesis.

THEOREM 1.1. *Suppose that one of the sets $A_1, A_2, \ldots, A_m$ or S is compact, then problem (1) has at least one optimal solution.*

We now proceed to discuss the uniqueness of the optimal solution for problem (1).

THEOREM 1.2. *Suppose that S is compact and convex. Then the optimal solution of problem (1) is unique.*

The following theorem provides a geometric condition for the optimal solution of problem (1).

THEOREM 1.3. *If x^* is an optimal solution of problem (1), then*

$$x^* \in \cap_{i=1}^m \mathbb{B}_{\frac{r^*}{w_i}}(a_i) \cap S, \text{ and } \cap_{i=1}^m \mathbb{B}^\circ_{\frac{r^*}{w_i}}(a_i) \cap S = \emptyset,$$

where $r^* = f(x^*)$. Conversely, suppose that there exist $\tilde{x} \in \mathbb{R}^n$ and $\tilde{r} > 0$ such that

$$\tilde{x} \in \cap_{i=1}^m \mathbb{B}_{\frac{\tilde{r}}{w_i}}(a_i) \cap S, \text{ and } \cap_{i=1}^m \mathbb{B}^\circ_{\frac{\tilde{r}}{w_i}}(a_i) \cap S = \emptyset,$$

then $\tilde{r} = f(\tilde{x})$ and $\tilde{x}$ is an optimal solution of problem (1).

2. Algorithm

An algorithm for finding the optimal solution of problem (1) proceeds as follows.

Algorithm 1.
Input: Given the tolerance $\epsilon > 0$, choose the initial bracket $[L^0, U^0]$ containing the minimum value r^* of the objective function of problem (1) such that $L^0, U^0 \geqslant 0$,

and set $k = 0$.

Step 1. Set $r^k = \dfrac{U^k + L^k}{2}$.

Step 2. **If** $\cap_{i=1}^{m} \mathbb{B}_{\frac{r^k}{w_i}}(a_i) \cap S \neq \emptyset$, **then** Set $L^{k+1} = L^k$ and $U^{k+1} = r^k$, **else** set $L^{k+1} = r^k$ and $U^{k+1} = U^k$.

Step 3. **If** $U^{k+1} - L^{k+1} \leqslant \epsilon$, **then** Set $\tilde{x} \in \cap_{i=1}^{m} \mathbb{B}_{\frac{U^{k+1}}{w_i}}(a_i) \cap S$ as an ϵ-approximated solution of problem (1); **Otherwise** Set $k = k + 1$ and go to Step 1.

The initial bracket $[L^0, U^0]$ can be given by $L^0 = 0$ and $U^0 = f(x_0)$ for a feasible solution $x_0 \in S$. Also the established problem in Step 2 is a feasibility problem, and one can determine whether balls and the set S intersect by using the cyclic projection algorithm [3, 4].

The following theorem presents the convergence analysis for Algorithm 1.

THEOREM 2.1. *Algorithm* 1 *stops at an* ϵ*-approximated solution* $\tilde{x}$ *in at most* $\lfloor \log_2 \dfrac{U^0 - L^0}{\epsilon} \rfloor + 1$ *iterations.*

References

1. E. A. Alper Yildirim, *Two algorithms for the minimum enclosing ball problem*, SIAM J. Optim. **19** (3) (2008) 1368–1391.
2. J. Brimberg and G. O. Wesolowsky, *Locating facilities by minimax relative to closest points of demand areas,* Comput. Oper. Res. **29** (6) (2002) 625–636.
3. J. M. Borwein, G. Li and L. Yao, *Analysis of the convergence rate for the cyclic projection algorithm applied to basic semialgebraic convex sets,* SIAM J. Optim. **24** (1) (2014) 498–527.
4. F. Deutsch and H. Hundal, *The rate of convergence for the cyclic projections algorithm III: regularity of convex sets,* J. Approx. Theory **155** (2008) 155–184.
5. S. Nickel, J. Puerto and A. M. Rodriguez-Chia, *An approach to location models involving sets as existing facilities*, Math. Oper. Res. **28** (4) (2003) 693–715.
6. S. Nobakhtian and A. Raeisi Dehkordi, *An algorithm for generalized constrained multi-source Weber problem with demand substations*, 4OR-Q J. Oper. Res. **16** (2018) 343–377.
7. S. Nobakhtian and A. Raeisi Dehkordi, *A fast algorithm for the rectilinear distance location problem*, Math. Meth. Oper. Res. **88** (2018) 81–98.

E-mail: a.raisi@sci.ui.ac.ir
E-mail: ali.ansari9286@gmail.com

Contributed Posters

Probability and Statistical Processes

On the Tsallis Entropy Rate of Hidden Markov Chains

Zohre Nikooravesh*
Department of Basic Sciences, Birjand University of Technology, Birjand, Iran

Abstract. We study the Tsallis entropy rate of a hidden Markov process, defined by observing the output of a symmetric channel whose input is a first order Markov process. Although this definition is very simple, obtaining the exact amount of entropy rate in calculation is very difficult. We introduce some probability matrices based on Markov chain's and channel's parameters. Then, we try to obtain an estimate for the Tsallis entropy rate of hidden Markov chain by matrix algebra and its spectral representation. To do so, we use the Taylor expansion, and calculate some estimates for the first terms, for the entropy rate of the hidden Markov process.
Keywords: Perron-Frobenius theorem, Probability matrices, Spectral representation, Taylor expansion.
AMS Mathematical Subject Classification [2010]: 60J10, 94A17.

1. Introduction

Suppose that there is a first-order stationary Markov process as an input of a symmetric channel with a noisy process. The output of this channel is a hidden Markov chain. In recent years, the Shannon entropy rate of hidden Markov chain studied by different scientists. The Tsallis entropy rate of hidden Markov chain by an special noisy process will be studied in this paper. To reach this goal, we will use the Taylor expansion and Perron-Frobenius theorem for stochastic matrices.

Computing the Shannon entropy (here it is called entropy) of a hidden Markov process was studied by Blackwell [1], which is based on the intrinsic complexity of expressing the hidden Markov process entropy as a function of the process parameters.

Zuk et al. [9] showed formulas for higher-order coefficients of the Taylor expansion in the symmetric case for binary Hidden Markov Chain.

Tsallis [7] proposed the generalization of the entropy by postulating a non-extensive entropy, (i.e., Tsallis entropy), which covers Shannon entropy in particular cases. This measure is non-logarithmic. Vila et al. [8] investigated the application of three different Tsallis-based generalizations of mutual information to analyze the similarity between scanned documents. Another paper by Castello et al. [3] presented a study and a comparison of the use of different information-theoretic measures for polygonal mesh simplification by applying generalized measures from Information Theory such as Havrda-Charvát-Tsallis entropy and mutual information.

For Shannon and Tsallis entropies, Nikooravesh [6] applied the problem of the maximum entropy for generalization of a direct method for quantile estimation,

*Presenter

which used the integral-order probability weighted moments of in place of the product moments.

Our study will focus on the estimation of the entropy rate of the hidden Markov chain, where the channel parameters are small.

2. Preliminaries

Let $X = \{X_k\}_{k\geq 1}$ be a first-order stationary Markov process on $\mathbf{X} = \{0, 1, \ldots, m-1\}$, with transition matrix $\mathbf{P} = \{p_{ab}\}$ such that for every $k \geq 1$, $p_{ab} = P_X(X_k = b|X_{k-1} = a)$, where $a, b \in \mathbf{X}$. Also the initial distribution of the Markov chain is denoted by the vector $\mathbf{\Pi}_0$ such that $\mathbf{\Pi}_0(i) = Pr\{X_0 = i\}$ for $i \in \mathbf{X}$. Consider also a noise process $E = \{E_k\}_{k\geq 1}$, independent of X, such that $P(E_i = l) = \varepsilon_l$, where $l \in \mathbf{X}$ and $\sum_{l=0}^{m-1} \varepsilon_l = 1$. Now, define the process $Z = \{Z_k\}_{k\geq 1}$, with $Z_k = X_k \oplus E_k, k \geq 1$, where $\oplus$ denotes addition modulo m. Consider a stochastic process $\{Y_k\}_{k\geq 1}$ with state space $\mathbf{Y}$. The Tsallis entropy rate of the stochastic process $\{Y_k\}_{k\geq 1}$ is

$$S_q(\mathcal{Y}) = \lim_{n\to\infty} \frac{S_q(Y_1, Y_2, \ldots, Y_n)}{n},\ q > 0,\ q \neq 1,$$

where Y_t is a random variable demonstrating the state at time t, and

$$S_q(Y_1, Y_2, \ldots, Y_n) = -\sum_{y_1\in\mathbf{Y}} \sum_{y_2\in\mathbf{Y}} \cdots \sum_{y_n\in\mathbf{Y}} P^q(y_1, y_2, \ldots, y_n) \ln_q P(y_1, y_2, \ldots, y_n), \tag{1}$$

where $\ln_q(x) = (x^{1-q} - 1)/(1 - q)$. The process $\{Z_k\}_{k\geq 1}$ is a stochastic process, also it is an example of a hidden Markov process. Let

$$\mathbf{P}_n := [P(Z_1^n, E_n = 0), P(Z_1^n, E_n = 1), \ldots, P(Z_1^n, E_n = m - 1)],$$

and get $\mathbf{M}(Z_{n-1}, Z_n)$ as a probability matrix with dimension $m \times m$ and entries $\varepsilon_{j-1}P_X(Z_n \oplus (j - 1)|Z_{n-1} \oplus (i - 1))$ in i^{th} row and j^{th} column. So it is easy to show

$$\mathbf{P}_n = \mathbf{P}_{n-1}\mathbf{M}(Z_{n-1}, Z_n), P_Z(Z_1^n) = \mathbf{P}_1\mathbf{M}(Z_1, Z_2)\ldots\mathbf{M}(Z_{n-1}, Z_n)\mathbf{1}^t, \quad n > 1,$$

where $\mathbf{1} = [1, 1, \ldots, 1]_{1\times m}$ and superscript t denotes transposition.

We construct these matrices for a given realization z_1^n of Z_1^n. Using the notation $\mathbf{M}_i = \mathbf{M}(z_i, z_{i+1})$ we get

$$\mathbf{M}_i = \mathbf{M}_i^{(0)} + \varepsilon_1\mathbf{M}_i^{(1)} + \cdots + \varepsilon_{m-1}\mathbf{M}_i^{(m-1)}, \quad 1 \leq i \leq n - 1. \tag{2}$$

Similarly, one can show $\mathbf{P}_1 = \mathbf{P}_1^{(0)} + \varepsilon_1\mathbf{P}_1^{(1)} + \cdots + \varepsilon_{m-1}\mathbf{P}_1^{(m-1)}$, and

$$\begin{aligned} P_Z(z_1^n) &= \mathbf{P}_1\mathbf{M}_1\mathbf{M}_2\ldots\mathbf{M}_{n-1}\mathbf{1}^t \\ &= (\mathbf{P}_1^{(0)} + \sum_{i=1}^{m-1} \varepsilon_i\mathbf{P}_1^{(i)}) \prod_{k=1}^{n-1} (\mathbf{M}_k^{(0)} + \sum_{i=1}^{m-1} \varepsilon_i\mathbf{M}_k^{(i)})\mathbf{1}^t. \end{aligned} \tag{3}$$

3. The Tsallis Entropy Rate of a Hidden Markov Chain

The following formula will be useful in computing the Tsallis entropy of Z i.e., $R_n(q,\boldsymbol{\Upsilon}) = \sum_{z_1^n} P_Z^q(z_1^n)$, where the exponent s of P_Z is a complex variable, and the summation is over all n_tuples of $\mathbf{X}$. Note that by the Eq. (3), for $\boldsymbol{\Upsilon} = \mathbf{0}$, we can write $R_n(q,\mathbf{0}) = \sum_{z_1^n} P_X^q(z_1^n)$. Now by different form of both sides of Eq. (1), we have

$$S_q(X_1^n) = \frac{1}{q-1}(1 - \sum_{x_1^n} P_X^q(x_1^n)) = \frac{1}{q-1}(1 - R_n(q,\mathbf{0})),$$

and similarly $S_q(Z_1^n) = \dfrac{1}{q-1}(1 - R_n(q,\boldsymbol{\Upsilon}))$. Using Taylor expansion near $\boldsymbol{\Upsilon} = \mathbf{0}$, we have

$$R_n(q,\boldsymbol{\Upsilon}) = R_n(q,\mathbf{0}) + \sum_{k=1}^{m-1} \varepsilon_k \frac{\partial}{\partial \varepsilon_k} R_n(q,\boldsymbol{\Upsilon})|_{\boldsymbol{\Upsilon}=\mathbf{0}} + o(\varepsilon_{max}^2),$$

where $\varepsilon_{max} = \max\{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{m-1}\}$. Now by noting on both sides of above formula, we can implies

$$S_q(Z_1^n) = S_q(X_1^n) - \frac{1}{q-1} \sum_{k=1}^{m-1} \varepsilon_k \frac{\partial}{\partial \varepsilon_k} R_n(q,\boldsymbol{\Upsilon})|_{\boldsymbol{\Upsilon}=\mathbf{0}} + o(\varepsilon_{max}^2)).$$

For our aims, we must compute $\partial R_n(q,\boldsymbol{\Upsilon})/\partial \varepsilon_k$ at $\boldsymbol{\Upsilon} = \mathbf{0}$,

$$\frac{\partial}{\partial \varepsilon_k} R_n(q,\boldsymbol{\Upsilon})|_{\boldsymbol{\Upsilon}=\mathbf{0}} = \sum_{z_1^n} q P_Z^{q-1}(z_1^n) \frac{\partial}{\partial \varepsilon_k} P_Z(z_1^n)|_{\boldsymbol{\Upsilon}=\mathbf{0}}.$$

Using Eq. (2), the derivative of $P_Z(z_1^n)$ at $\boldsymbol{\Upsilon} = \mathbf{0}$ can be calculated as

$$\begin{aligned}\frac{\partial}{\partial \varepsilon_k} P_Z(z_1^n)|_{\boldsymbol{\Upsilon}=\mathbf{0}} &= -nP_X(z_1^n) + P_X(z_1 \oplus k\ z_2^n) \\ &+ \sum_{i=1}^{n-2} P_X(z_1^i\ z_{i+1} \oplus k\ z_{i+2}^n) + P_X(z_1^{n-1} z_n \oplus k).\end{aligned}$$

Now we can compute $S_q(\mathcal{Z})$(the entropy rate of the hidden Markov chain $\{Z_i\}_{i\geq 1}$), i.e.,

$$S_q(\mathcal{Z}) = \lim_{n\to\infty} \frac{1}{n} S_q(Z_1^n).$$

THEOREM 3.1. *Suppose $p_{ab} > 0$ for any $a, b \in \mathbf{X}$. The first order term in the entropy rate of the hidden Markov chain Z is converges to $S_q(\mathcal{X})$ exponentially for $q > 1$ and is divergent for $q < 1$.*

To prove the Theorem 3.1, it is necessary to express spectral representation of matrices and Perron-Frobenius theorem. We use the spectral representation [5] of the matrix $\mathbf{P}(q)$. Since $p_{ab} > 0$, for any $a, b \in \mathbf{X}$, the Perron-Frobenius theorem [2] applies. So there exists a real eigenvalue $\lambda_1(q)$ with algebraic geometric multiplicity one such that $\lambda_1(q) > 0$, and $\lambda_1(q) > |\lambda_j(q)|$ for any other eigenvalue

$\lambda_j(q)$. Moreover the left eigenvector $l_1(q)$ and the right eigenvector $r_1(s)$ associated with $\lambda_1(q)$ can be chosen positive and such that $l_1(q)r_1^t(q) = 1$.

Let $\lambda_2(q), \lambda_3(s), \ldots, \lambda_m(q)$ be the eigenvalues of the $\mathbf{P}(q)$ other than λ_1 ordered in such a way that $\lambda_1(q) > |\lambda_2(q)| > |\lambda_3(q)| > \cdots > |\lambda_m(q)|$ and we know that the vectors r_1 and l_1 are real-valued with nonnegative components. The matrix spectral representation yields

$$\mathbf{P}^k(q) = \lambda_1^k(q)(r_1^t(q)l_1(q)) + o(|\lambda_2|^k).$$

In addition if $\mathbf{P} \geq 0$ is irreducible (which of course includes $\mathbf{P} > 0$)

$$\begin{cases} \lambda_1(\mathbf{P}') > \lambda_1(\mathbf{P}), & \mathbf{P}' \geq \mathbf{P}, \mathbf{P}' \neq \mathbf{P}, \\ \lambda_1(\mathbf{P}') < \lambda_1(\mathbf{P}), & \mathbf{P}' \leq \mathbf{P}, \mathbf{P}' \neq \mathbf{P}, \end{cases}$$

where $\lambda_1(\mathbf{P})$ and $\lambda_1(\mathbf{P}')$ are the greatest eigenvalue of matrices $\mathbf{P}$ and $\mathbf{P}'$, respectively [4].

4. A Numerical Example

In Section 3 we proved that the first order term in the entropy rate of the hidden Markov chain $\{Z_n\}$ is converges to $S_q(\mathcal{X})$ exponentially for $q > 1$ and is divergent for $q < 1$. Now, we want to calculate the Tsallis entropy rate for a hidden Markov chain $\{Z_n\}$, whit stat space $S = \{0, 1\}$, transition probability matrix $\mathbf{P}$ for Markov chain $\{X_n\}$ whit stat space S and noise process $E = \{E_k\}_{k\geq 1}$, independent of X, such that $P(E_i = 1) = \varepsilon$ and $P(E_i = 0) = 1 - \varepsilon$ for $0 \leq \varepsilon \leq 1$. Now, define the process $Z = \{Z_k\}_{k\geq 1}$, with

$$Z_k = X_k \oplus E_k, \qquad k \geq 1,$$

We obtained the Tsallis entropy $S_q(Z_1^n)$ by using Eq. (2), for $\mathbf{P}$ with initial distribution $\mathbf{\Pi}_0$ and noise parameter $\varepsilon = 0.5$ for $q = 2$ and $q = 0.2$, that one can see in Table 1, where

$$\mathbf{P} = \begin{bmatrix} 0.75 & 0.25 \\ 0.4 & 0.6 \end{bmatrix}, \ \mathbf{\Pi}_0 = [0.5, 0.5].$$

TABLE 1. $S_q(Z_1^n)$ for $q = 2$ and $q = 0.2$ and $n = 2, 3, \ldots, 23$.

n	$S_2(Z_1^n)$	$S_{0.2}(Z_1^n)$	n	$S_2(Z_1^n)$	$S_{0.2}(Z_1^n)$
2	0.3569	1.2453	13	0.0768	110.6876
3	0.2774	1.7256	14	0.0714	176.6506
4	0.2251	2.4479	15	0.0666	283.3263
5	0.1881	3.5441	16	0.0625	456.4097
6	0.1607	5.2237	17	0.0588	738.0745
7	0.1398	7.8209	18	0.0556	1.1977e+03
8	0.1234	11.8695	19	0.0526	1.9494e+03
9	0.1102	18.2267	20	0.0500	3.1819e+03
10	0.0995	28.2731	21	0.0476	5.2064e+03
11	0.0906	44.2411	22	0.0455	8.5385e+03
12	0.0832	69.7504	23	0.0435	1.4032e+04

The results are shown in the Figure 1.

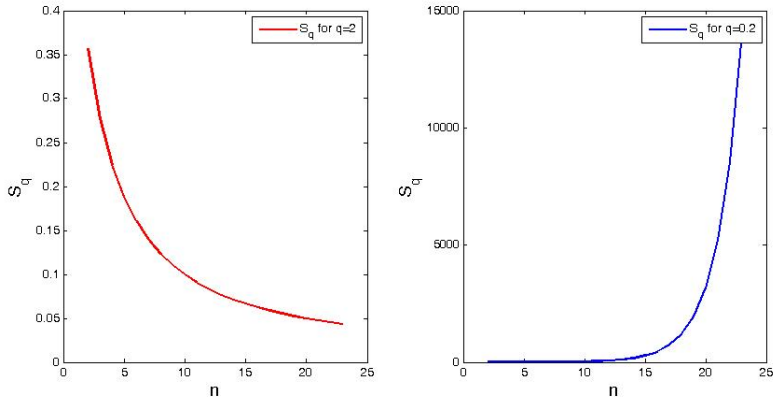


FIGURE 1. $S_q(Z_1^n)$ for $q = 2$ (the left) and $q = 0.2$ (the right) and $n = 2, 3, \ldots, 23$.

5. Conclusions

We studied the Tsallis entropy rate of a hidden Markov process defined as the output of a binary symmetric channel whose input is a binary Markov process. We first expressed the entropy rate of the hidden Markov process as a well-defined product of random matrices. These exponents are notoriously difficult to compute. Therefore, we turned our attention to asymptotic expansions, and derived a Taylor expansion of Tsallis entropy rate of the hidden Markov process when the probability of error is small. We observed that the first order term in the Tsallis entropy rate of the hidden Markov chain is converges to Tsallis entropy rate of the input Markov chain exponentially for $q > 1$, and is divergent for $q < 1$.

References

1. D. Blackwell, *The entropy of functions of finite-state Markov chains,* Trans. First Prague Conf. Information Theory, Statistical Decision Functions, Random Processes, Prague, (1957) pp. 13–20.
2. P. Bremaud, *Markov Chains*, Springer Verlag, New York, Berlin, Heidelberg, 1998.
3. P. Castello, C. González, M. Chover, M. Sbert and M. Feixas, *Tsallis entropy for geometry simplification,* Entropy **13** (2011) 1805–1828.
4. G. Frobenius, *Über matrizen aus nichtnegativen elemente,* Sitzungsber Kon Preuss Acad. Wiss. Berlin, (1912) 456–457.
5. S. Karlin and H. Taylor, *A first Course in Stochastic Processes,* Academic Press, New York, 1975.
6. Z. Nikooravesh, *Estimation of the probability function under special moments conditions using the maximum Shannon and Tsallis entropies,* Chil. J. Stat. **9** (2) (2018) 55–64.
7. C. Tsallis, *Possible generalization of Boltzmann-Gibbs statistics,* J. Stat. Phys. **52** (1988) 479–487.
8. M. Vila, A. Bardera and M. Feixas, *Tsallis mutual information for document classification,* Entropy **13** (2011) 1694–1707.
9. O. Zuk, I. Kanter and E. Domany, *Asymptotic of the entropy rate for a hidden Markov process,* J. Stat. Phys. **121** (2005) 343–360

E-mail: nikooravesh@birjandut.ac.ir

Generalized Entropy for Super Diffusion Walks in Graphs

Zohre Nikooravesh*
Department of Basic Sciences, Birjand University of Technology, Birjand, Iran

Abstract. In this paper, the entropy of the stochastic processes created by the movement of a walker in a graph is investigated. The Shannon-Khinchin entropy has four axioms that ignore one of them can make the generalized entropy. Here, we investigate the number of different finite paths asymptotically, for determining a generalized entropy. Then, we will study a special graph with finite nodes, with two different types of motion.
Keywords: Generalized entropies, Khinchins axioms, Random walks, Perron-Frobenius theorem.
AMS Mathematical Subject Classification [2010]: 60J10, 94A17.

1. Introduction

With recent surge of interest in complex networks in various fields including statistical physics and mathematical physics, many quantities have been proposed to characterize the structural properties of graphs, [3] and [5]. The study of a graph invariant in one field may also be a result of relevant importance in other areas of physics. This is because graphs are nowadays ubiquitous in many areas of physics such as in problems associated with the Ising, Potts and Hubbard models, in the solution of Feynman integrals in perturbative field theory, in quantum information theory such as quantum error correcting codes (graph states) or arrangements of interacting quantum mechanical particles (spin networks) and in many other fields, [2] and [7]. Among various graph invariants, a special role has been played by the concept of entropy. Dehmer and Mowshowitz [4] have used entropy measures for graphs for a long time in different fields. Inspired by connections between quantum information and graph theory, Passerini and Severini [12] have defined the von Neumann entropy for graphs, which in general depends on the regularity, the number of connected components, the shortest-path distance and nontrivial symmetries in the graph. Here, we define graph entropies based on walks in a graph. Walks in graphs play a fundamental role in the analysis of the structure and dynamical processes in networks [6]. The walk entropies thereby characterize the spread of a walk among the vertices or edges of the graph; in other words, we understand by the walk entropies how much the walk is Before proceeding, we summarize a few definitions which are necessary to make this paper self-contained. Let us consider here simple graphs $G = (V, E)$ with $|V| = n$ nodes and $|E| = m$ edges. A walk of length k is a sequence of (not necessarily distinct) nodes $v_0, v_1, \ldots, v_k$ such that for each $i = 1, 2, \ldots, k$ there is a link from $v_{(i-1)}$ to v_i. The number of walks of length N from node p to node q is given by $[\mathbf{A}^N]_{pq}$,

*Presenter

where **A** is the adjacency matrix of the graph. The degree of the node p, denoted by p_k, is the number of edges incident to it.

In order to define graph entropies based on the walks, we consider a random walker which walks from one node to another by using the edges of the graph.

This paper is organized as follows: Section 2 discusses extensive or generalized entropies that one can see more details in [10]. In this section, the four axioms of Khinchin, what the unique result is Shannon's entropy, are outlined. By ignoring the fourth axiom, one can obtain the general form of extensive entropies that depend on two parameters (c, d). Section 3 contains two cases as the main results. In this section, we examine two different types of motion in graphs. In the first case, at each step there is a choice of a new direction for the walker, while in the second case, after selecting a direction for walking, the change of direction is not possible for a finite number of next steps.

2. Review of Generalized Entropies

Shannon and Khinchin showed that, assuming four information theoretic axioms, the entropy must be of the Boltzmann-Gibbs type, $S = -\sum_i p_i \log p_i$. In many physical systems, one of these axioms may be violated. For non-ergodic systems, the so-called separation axiom is not valid.

Scientifics proved there are some entropies that not necessarily satisfies in all of Shannon and Khinchin axioms. These entropic forms are called generalized entropies and usually assume trace form for example in [13]

$$S_g(p) = \sum_i^W g(p_i), \tag{1}$$

where W is the number of states. Obviously not all generalized entropic forms are of this type. Renyi entropy, for example, is of the form, $G(\sum_i^W g(p_i))$, with G a monotonic function. We use trace forms Eq. (1) for simplicity. Renyi forms can be studied in exactly the same way, as will be shown, however, at more technical cost.

As mentioned, if all of Shannon and Khinchin axioms hold, the only possible entropy is the Boltzmann-Gibbs-Shannon (BGS) entropy. The generalized entropy for (large) admissible statistical systems (all of Shannon and Khinchin axioms except separability axiom hold) is derived from two hitherto unexplored fundamental scaling laws of extensive entropies [8]. Both scaling laws are characterized by exponents c and d, respectively, which allow one to uniquely define equivalence classes of entropies, meaning that two entropies are equivalent in the thermodynamic limit if their exponents (c, d) coincide. Each admissible system belongs to one of these equivalence classes (c, d), [8]. In terms of the exponents (c, d), Hanel and Thurner [8] showed that all generalized entropies have the form

$$S_{(c,d)} \propto \sum_{i=1}^{W} \Gamma(d+1, 1 - c \log p_i),$$

with

$$\Gamma(\mu, t) = \int_t^\infty y^{\mu-1} e^{-y} dy = \int_0^{e^{-t}} (-\ln x)^{\mu-1} dx.$$

Also, $\Gamma(\mu, t)$ named the incomplete Gamma-function.

2.1. Determining the Exponents, c and d. Consider a system with N elements. The number of system configurations (microstates) as a function of N are denoted by $W(N)$. Hanel and Thurner [9] said

$$\frac{1}{1-c} = \lim_{N\to\infty} N\frac{W'(N)}{W(N)},$$

and

$$d = \lim_{N\to\infty}\left[\frac{W(N)}{NW'(N)} + c - 1\right]\log W(N),$$

Here, W' means the derivative with respect to N.

3. Random Walks in Graphs without Self-Loop and with Self-Loop

In this section, we have two cases as the main results of this paper. We examine two different types of motion in graphs. In the first case, at each step there is a choice of a new direction for the walker, while in the second case, after selecting a direction for walking, the change of direction is not possible for a finite number of next steps.

Case 1: We now focus our discussion on random walks in undirected graphs with uniform edge weights, with no multi edges or self-loops. At each node, the random walk is equally likely to take any connected edges. Assume the graph $G(V, E)$ with $|V| = n$ nodes and $|E| = m$ edges, is connected. On the other hand, since the graph G is connected and has a cycle of odd length, one can find an integer k such that all of entries $\mathbf{A}^k$ are positive, where $\mathbf{A}$ is the adjacency matrix of the graph. We know the number of walks of length N i.e. $W(N)$ from node p to node q is given by $[\mathbf{A}^N]_{pq}$.

Now it is necessary to express spectral representation of matrices and Perron-Frobenius theorem. We use the spectral representation of the matrix $\mathbf{A}$ [11]. Since $a_{ij} \geq 0$, and there is an integer k such that $[\mathbf{A}^k]_{ij} > 0$, the Perron-Frobenius theorem applies [1]. So there exists a real eigenvalue λ_1 with algebraic geometric multiplicity one such that $\lambda_1 > 0$, and $\lambda_1 > |\lambda_j|$ for any other eigenvalue λ_j. Moreover the left eigenvector l_1 and the right eigenvector r_1 associated with λ_1 can be chosen positive and such that $l_1 r_1^t = 1$. Let $\lambda_2, \lambda_3, \ldots, \lambda_m$ be the eigenvalues of the $\mathbf{A}$ other than λ_1 ordered in such a way that $\lambda_1 > |\lambda_2| \geq |\lambda_3 \geq | \cdots \geq |\lambda_m|$ and we know that the vectors r_1 and l_1 are real-valued with nonnegative components. The matrix spectral representation yields

$$\mathbf{A}^N = \lambda_1^N(r_1^t l_1) + |\lambda_2|^N(r_2^t l_2) \Rightarrow \mathbf{A}^N = \lambda_1^N(r_1^t l_1) + o(|\lambda_2|^N).$$

We can consider

$$\lambda_1^N(r_1^t l_1) = o(\lambda_1^N), |\lambda_2|^N(r_2^t l_2) = o(|\lambda_2|^N).$$

So

$$A^N = \lambda_1^N(r_1^t l_1)(1 + o(\frac{|\lambda_2|^N}{\lambda_1^N})).$$

Now we know the number of walks of length N from node p to node q is given by $[\mathbf{A}^N]_{pq}$, so $W(N) \sim \lambda_1^N (r_1^t l_1)_{pq}(1+o(\rho))$, where $\rho = \frac{|\lambda_2|^N}{\lambda_1^N} < 1$. One can obtain

$$\frac{1}{1-c} = \lim_{N\to\infty} N \log \lambda_1 = \infty \Rightarrow c = 1,$$

and

$$d = \lim_{N\to\infty} \left[N \log \lambda_1 + \log (r_1^t l_1)_{pq}(1+o(\rho))\right] (\frac{1}{N \log \lambda_1} + c - 1) = 1.$$

So $(c, d) = (1, 1)$ and this random walk in graphs has Shannon entropy.

Case 2: We now focus our discussion on random walks in undirected graphs with uniform edge weights. At each node, the random walk is equally likely to take any edge. Now suppose the graph has self-loop with probability weight zero. A super diffusion walk in this graph is described as remaining in the same node for $[N^\beta]_+$ timesteps after selecting an edge in step N. In other words, the walker moves on self-loop without making any new decisions and the next free decision is possible at time step $N + [N^\beta]_+$. Clearly, the number of decision grows like $N^{1-\beta}$, and the number of possible sequences without considering self-loops, is related to $\mathbf{A}^{N^{1-\beta}}$, therefore

$$W(N) \sim \lambda_1^{N^{1-\beta}} (r_1^t l_1)_{pq}(1+o(\rho)),$$

where here $\rho = \frac{|\lambda_2|^{N^{1-\beta}}}{\lambda_1^{N^{1-\beta}}} < 1$. Consequently the associated extensive entropy is of class $(c, d) = (1, \frac{1}{1-\beta})$, because

$$\frac{1}{1-c} = \lim_{N\to\infty} (1-\beta) N^{1-\beta} \log \lambda_1 = \infty \Rightarrow c = 1,$$

and

$$d = \lim_{N\to\infty} \left(\frac{1}{((1-\beta) N^{1-\beta} \log \lambda_1)} + c - 1\right) [N^{1-\beta} \log \lambda_1$$

$$+(1-\beta) \log (r_1^t l_1)_{pq}(1+o(\rho))] = \tfrac{1}{1-\beta}.$$

4. Conclusion

We studied the relationship between the volume of state space of a stochastic process and its extensive (generalized) entropy. If the volume of state space ω is given as a function of system size, we know that how to determine the associated generalized entropy by computing the parameters (c, d). We demonstrated in two concrete cases how statistical systems determine their own extensive entropies. These cases examine the motion of a walker in undirected and connected graphs. In the first case, the walker selects the next node for displacement, from set of the possible nodes, at any time-step uniformly. In the second case, the walker selects a new node for movement and stays on his new place for a certain number of time-steps moving on its self-loop, meaning that after several time-steps, the walker goes to another node. In the first case that the certain kind of the Markov chains were shown, we obtained their generalized entropies as the same as the Shannon entropy. Whereas in the second case, where non-Markovian processes were investigated, their generalized entropies were not the Shannon entropy.

References

1. P. Bremaud, *Markov Chains*, Springer Verlag, New York, Berlin, Heidelberg, 1998.
2. G. Chiribella, G. M. D'Ariano and P. Perinott, *Theoretical framework for quantum networks*, Phys. Rev. A **80** (2) (2009) 022339.
3. L. da F. Costa, F. A. Rodrigues, G. Travieso and P. R. Villas Boas, *Characterization of complex networks*, A survey of measurements. Adv. Phys. **56** (1) (2007) 167–242.
4. M. Dehmer and A. Mowshowitz, *A history of graph entropy measures*, Inform. Sci. **181** (1) (2011) 57–78.
5. E. Estrada, *The Structure of Complex Networks: Theory and Applications*, Oxford University Press, Oxford, UK, 2011.
6. E. Estrada, N. Hatano and M. Benzi, *The physics of communicability in complex networks*, Phys. Rep. **514** (3) (2012) 89–119.
7. C. Godsil, S. Kirkland, S. Severini and J. Smith, *Number-theoretic nature of communication in quantum spin systems*, Phys. Rev. Lett. **109** (5) (2012) 050502.
8. R. Hanel and S. Thurner, *A comprehensive classification of complex statistical systems and an axiomatic derivation of their entropy and distribution functions*, Europhys. Lett. **93** (2011) 20006.
9. R. Hanel and S. Thurner, *When do generalized entropies apply? How phase space volume determines entropy*, Europhys. Lett. **96** (2011). DOI: 10. 1209/0295-5075/96/50003
10. R. Hanel and S. Thurner, *Generalized* (c, d) *entropy and aging random walks*, Entropy **15** (12) (2013) 5324–5337.
11. S. Karlin and H. Taylor, *A First Course in Stochastic Processes*, Academic Press, New York, 1975.
12. F. Passerini and S. Severini, *Quantifying complexity in networks: The von Neumann entropy*, Int. J. Agent Tech. Syst. **1** (4) (2009) 58–67.
13. C. Tsallis, *Possible generalization of Boltzmann-Gibbs statistics*, J. Stat. Phys. **52** (1988) 479–487.

E-mail: nikooravesh@birjandut.ac.ir

A New Wrapped Probability Distribution with Application in Weather Studies

Sajjad Piradl*
Department of Statistics, Payame Noor University, Tehran, Iran

Abstract. Some important variables such as wind directions are plays major role in the weather studies. Given the widespread use of the gamma variance probability distribution in the study of circular data, in this paper, we have proposed a generalization of this probability distribution named as the wrapped variance gamma probability distribution along with its probability density function. We also have studied some important features of this probability distribution. In practice, we have applied this probability distribution to a data set which consists of the wind directions data at a site on the Black mountain in the Australian Capital Territory. Because it has been made clear that wind directions and its characteristics are important for the maintenance of climate change and wind energy functioning.
Keywords: Circular data, Wrapped probability distribution, Wrapped variance gamma probability density function, Moments, Wind directions.
AMS Mathematical Subject Classification [2010]: 60E05.

1. Introduction

An axis is an undirected line, where there is no reason to distinguish one end of the line from the other. Phenomena in nature that can be described as axial data are numerous such as dance direction of insects, movement of sea creatures, etc. (Godfroy-Cooper and et al. [7]). Wells and SenGupta [21] modeled this kind of data by introduce the method of construction for axial distributions. This method was the wrapping of the circular probability distribution. Madan and Seneta [12] first was proposed the variance gamma probability distribution. This probability distribution is uses in weather studies, financial fields, pricing, etc (Tadikamalla [19], Fragiadakis and et al. [5], Mastrantonio and Calise [15]). Study of variance gamma probability distribution in the case of circular data can play a key role, because some important variables are axial in weather studies such as wind directions. Wrapped probability distribution first was introduced by Levy [11] and studied by the other researchers such as Mardia [13, 14], Jammaladaka and SenGupta [9], Jammalamadaka and Kozubowski [8], Gatto [6], Choelo [2], Umbach and Jammakadaka [20], Roy and Adnan [16, 17], Biswas and et al. [1] and Joshi and Jose [10]. Their studies include several wrapped probability distributions such as wrapped exponential probability distribution, wrapped gamma probability distribution, wrapped chi-square probability distribution, wrapped weighted exponential probability distribution, Wrapped Linley probability distribution, etc. However, it seems that the wrapped gamma variance distribution is better one

*Presenter

to model the directional data for the weather studies. Therefore, we define the variance gamma probability distribution and then study concept of the wrapped probability density function.

A random variable X has variance gamma probability distribution, if its density function is as follows:

$$f(x) = \frac{a^{2b}.e^{c.(x-d)}.|x-d|^{b-0.5}.H_{b-0.5}(m.|x-d|)}{\sqrt{\pi}.\Gamma(b).(2m)^{b-0.5}}, \qquad x \in \mathbb{R},$$

where $a > 0$, $b > 0$, $c > 0$, $d \in \mathbb{R}$ is the location parameter, H is the modified Bessel function of the third kind, $m > 0$, $a = \sqrt{m^2 - c^2}$ and $0 \leq |d| < m$.

On the other hand, the methods of create a circular model are:

1) Wrapping a linear probability distribution around a unit circle,
2) Specify properties such as maximum entropy, etc,
3) One may start with a probability distribution on the real line and apply a stereographic projection that identifies points Xon $\mathbb{R}$ with those on the circumference of the circle, say α.

Which a circular probability distribution is a probability distribution whose total probability is focused on the unit circle in the plane $\{(\cos\alpha, \sin\alpha) \mid 0 \leq \alpha < 2\pi\}$, with properties:

1) $\forall\alpha,\ f(\alpha) \geq 0$,
2) $\int_{\alpha=0}^{2\pi} f(\alpha)\, d\alpha = 1$,
3) $f(\alpha) = f(\alpha + 2\pi k),\ k \in \mathbb{Z}$,

where $f(\alpha)$ is the probability density function.

Therefore, if X is a random variable defined on $\mathbb{R}$, then the corresponding wrapped random variable is defined as $X_w = x \mod 2\pi$. Where this random variable is a many-valued function as follows:

$$X_w(\alpha) = \{f(\alpha + 2\pi k),\ k \in \mathbb{Z}\}.$$

So, given a wrapped random variable X_w defined on $[0, 2\pi)$, by the transformation $(\alpha + 2\pi k)$, $k \in \mathbb{Z}$, we expand the support of the random variable X_w to $\mathbb{R}$ such that we can apply an in line probability density function $h(x)$ to the argument $(\alpha + 2\pi k)$.

Also the wrapped probability density function $f(\alpha)$ related to the probability density function $h(x)$ of a linear random variable X is defined as follow:

$$f(\alpha) = \sum_{k=-\infty}^{\infty} f(\alpha + 2\pi k);\ \alpha \in [0, 2\pi),$$

The order of contents of this paper is so that the form of the probability distribution function of a wrapped variance gamma probability distribution is obtained through the concern wrapping in Section 2. In Section 3, some important features from this probability distribution are proposed [3]. In Section 4, the maximum likelihood estimation method has been used [18]. In Section 5, this estimation method is used for a real data set. Such a way, that as an example of axial data, we use a data set which consists of the wind directions data at a site on the Black mountain in the Australian Capital Territory. Finally, the conclusion appears in Section 6.

2. Probability Density Function of the Wrapped Variance Gamma Probability Distribution

If we consider

$$\alpha \equiv \alpha(x) = x \mod 2\pi,$$

then α is wrapped around the circle or is a wrapped variance gamma random variable with the probability density function as follows:

$$\begin{aligned} f(\alpha) &= \sum_{k=-\infty}^{\infty} f(\alpha + 2\pi k) \\ &= \frac{a^{2b}}{\sqrt{\pi}.\Gamma(b).(2m)^{b-0.5}} \cdot \sum_{k=-\infty}^{\infty} e^{c.(\alpha+2k\pi-d)}.|\alpha+2k\pi-d|^{b-0.5}.H_{b-0.5}(m.|\alpha+2k\pi-d|) \\ &= \frac{a^{2b}.e^{c.(\alpha-d)}}{\sqrt{\pi}.\Gamma(b).(2m)^{b-0.5}} \times \sum_{k=-\infty}^{\infty} \frac{e^{c.k2\pi}.H_{b-0.5}(m.|\alpha+2k\pi-d|)}{|\alpha+2k\pi-d|^{b-0.5}}, \qquad \alpha \in [0, 2\pi), \end{aligned}$$

where $a > 0$, $b > 0$, $c > 0$, $d \in \mathbb{R}$ is the location parameter, H is the modified Bessel function of the third kind, $m > 0$, $a = \sqrt{m^2 - c^2}$, $0 \leq |d| < m$and we say that random variable α has a wrapped variance gamma probability distribution with parameters a, b, c, d and m.

In The Figure 1 we show the probability density functions of some wrapped variance gamma distributions.

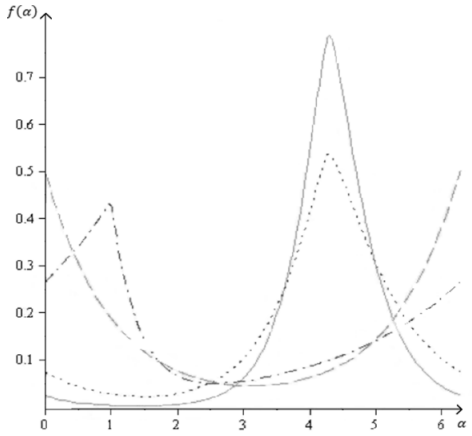


FIGURE 1. Some wrapped variance gamma distributions with solid line for$a = 2.49$, $b = 1.50$, $c = 0.20, d = -2.00$ and $m = 2.50$, dot line for $a = 1.49$, $b = 1.30$, $c = 0.20, d = -2.00$ and $m = 1.50$, dash line for $a = 1.00$, $b = 1.00$, $c = 0, d = 0$ and $m = 1.00$ and dash dot line$a = 1.18$, $b = 1.00$, $c = -1.00, d = 1.00$ and $m = 1.50$.

3. Some Features of the Wrapped Variance Gamma Probability Distribution

In this section, we propose some main properties of thewrapped variance gamma probability distribution. These characteristics are as follows:

- Characteristics function:

$$\varphi_{\alpha}(r) = E\left(e^{i.r.\alpha}\right) = \left\{\frac{a}{\left[m^2-(c+i.r)^2\right]^{0.5}}\right\}^{2b}.e^{i.d.r}; r = \pm 1, \pm 2, \ldots.$$

- Non-central trigonometric moments:

$$\mu_r = \left\{\frac{a}{\left[m^2-(c+i.r)^2\right]^{0.5}}\right\}^{2b}.\cos(d.r) + i.\left\{\frac{a}{\left[m^2-(c+i.r)^2\right]^{0.5}}\right\}^{2b}.\sin(d.r),$$

where $r = \pm 1, \pm 2, \ldots$.

- Alternative probability density function obtained under the non-central trigonometric moments:

$$f(\alpha) = \frac{1}{2\pi}.\left\{1 + 2\sum_{r=1}^{\infty}\left\{\left\{\frac{a}{\left[m^2-(c+i.r)^2\right]^{0.5}}\right\}^{2b}.\cos\left[r.(\alpha-d)\right]\right\}\right\}, \quad \alpha \in [0, 2\pi).$$

- Central trigonometric moments:

$$\mu'_r = \left\{\frac{a}{\left[m^2-(c+i.r)^2\right]^{0.5}}\right\}^{2b}, \; r = \pm 1, \pm 2, \ldots.$$

- Circular mean:

$$\mu = d.$$

- Circular variance:

$$\sigma^2 = 1 - \left\{\frac{a}{\left[m^2-(c+i)^2\right]^{0.5}}\right\}^{2b}.$$

- Circular standard deviation:

$$C.S.D = \left|\sqrt{-2\log\left\{\left\{\frac{a}{\left[m^2-(c+i)^2\right]^{0.5}}\right\}^{2b}\right\}}\right|.$$

- Circular dispersion:

$$\delta = \frac{1 - \left\{\frac{a}{\left[m^2-(c+2i)^2\right]^{0.5}}\right\}^{2b}}{2\left\{\left\{\frac{a}{\left[m^2-(c+i)^2\right]^{0.5}}\right\}^{2b}\right\}^{2}}.$$

- Kurtosis:

$$K = \frac{\left\{\frac{a}{\left[m^2-(c+2i)^2\right]^{0.5}}\right\}^{2b} - \left\{\left\{\frac{a}{\left[m^2-(c+i)^2\right]^{0.5}}\right\}^{2b}\right\}^{4}}{\left\{1-\left\{\frac{a}{\left[m^2-(c+2i)^2\right]^{0.5}}\right\}^{2b}\right\}^{2}}.$$

In Table 1, we obtained values of some features of the wrapped variance gamma probability distribution based on the values of the parameters that we used in the drawing of the Figure 1.

TABLE 1. Values of some features of the wrapped variance gamma probability distribution for various values of the parameters a, b, c, d, m.

Parameters values Features	$a = 2.49$ $b = 1.50$ $c = 0.20$ $d = -2.00$ $m = 2.50$	$a = 1.49$ $b = 1.30$ $c = 0.20$ $d = -2.00$ $m = 1.50$	$a = 1.00$ $b = 1.00$ $c = 0$ $d = 0$ $m = 1.00$	$a = 1.18$ $b = 1.00$ $c = -1.00$ $d = 1.00$ $m = 1.50$
Circular variance	0.22	0.41	0.50	0.74
Circular standard deviation	0.70	1.02	1.17	1.51
Kurtosis	2.49	0.89	0.55	0.38

4. Maximum Likelihood Estimation for the Wrapped Variance Gamma Probability Distribution Parameters

Let that $\alpha_1, \alpha_2, \ldots, \alpha_n$ be a random sample of size n from a wrapped variance gamma probability distribution. Then the likelihood function is as follows:

$$\begin{aligned} L(a,b,c,d,m;\alpha_1,\alpha_2,\ldots,\alpha_n) &= \prod_{i=1}^{n} \frac{a^{2b}.e^{c.(\alpha_i-d)}}{\sqrt{\pi}.\Gamma(b).(2m)^{b-0.5}} \\ &\cdot \sum_{k=-\infty}^{\infty} \frac{e^{c.k2\pi}.H_{b-0.5}(m.|\alpha_i+2k\pi-d|)}{|\alpha_i+2k\pi-d|^{b-0.5}}. \end{aligned} \tag{1}$$

With taking log in the relation (1) we achieve:

$$\begin{aligned} \log L(a,b,c,d,m,\alpha_1,\alpha_2,\ldots,\alpha_n) &= n.[2b.\log(a) - 0.5\log(\sqrt{\pi}) - 0.5\log(b) \\ &- (b-0.5).\log(2m)] + \sum_{i=1}^{n}[c.(\alpha_i-d)] \\ &+ \sum_{i=1}^{n}\sum_{k=-\infty}^{\infty}\log\left[e^{c.k2\pi}.H_{b-0.5}(m.|\alpha_i+2k\pi-d|)\right] \\ &- \sum_{i=1}^{n}\sum_{k=-\infty}^{\infty}[(b-0.5).\log(\alpha_i+2k\pi-d)]. \end{aligned} \tag{2}$$

Here estimates of the parameters can be achieved from the relation (2) by applying the numerical methods.

5. Real Data Study

The following data set are the wind directions that contain hourly measurements of three days at a site on the Black mountain in the Australian Capital Territory [4]:

$$0, 15, 50, 90, 150, 180, 220, 235, 240, 245, 250, 255,$$
$$265, 270, 280, 285, 300, 315, 330, 335, 340, 345.$$

Our purpose is to show that the wrapped variance gamma distribution is effective in the study of wind directions. To do this, we first find the maximum likelihood estimates of the wrapped variance gamma probability distribution parameters. Next, we find the maximum likelihood estimates of the generalized von-Mises probability distribution as another famous circular probability distribution applicable in the weather studies. Then we compare the goodness of fit of two probability distributions based on the calculation of the Kuiper statistic, Watson's U^2 statistic, maximized log-likelihood criterion, Akaike's information criterion and Bayesian information criterion for the each probability distribution. Numerical results for this comparison are summarized in Table 2.

TABLE 2. Summarized results for the comparison between the two probability distributions according to the wind directions data set.

Probability distributions	Maximum likelihood estimates	Kuiper statistic	Watson's U^2 statistic	Maximized log-likelihood criterion	Akaike's information criterion	Bayesian information criterion	P-value
Wrapped variance gamma	$a = 4.07$ $b = 2.00$ $c = 0.98$ $d = 2.10$ $m = 0.50$	4.35	1.28	-63.40	136.80	133.51	0.120
Generalized von-Mises	$a = 4.07$ $b = 2.00$ $c = 0.98$ $d = 2.10$ $m = 0.50$	4.66	1.71	-67.20	144.40	141.11	0.001

Results of Table 2 show that the generalized von-Mises probability distribution is not suitable to model this data set. The smaller values of the criterions for the wrapped variance gamma probability distribution, indicates a better fit. Thus, the goodness of fit test confirms the superiority of our proposed probability distribution for fit this data set than the generalized von-Mises probability distribution.

6. Conclusion

What we have done in this paper, is to introduce a new probability distribution named as wrapped variance gamma probability distribution and study some of its important features. We have also applied this new probability distribution to real data set of wind directions in meteorology science and demonstrated the superiority of its in modelling of this kind of data in comparison with generalized von-Mises probability distribution as another applicable well-known probability distribution. Finally, we can say that the methods and results of this paper can also be applied to climate change studies.

References

1. A. Biswas, J. Jha and S. Dutta, *Modelling circular random variables with a spike at zero,* Statist. Prob. Lett. **109** (2016) 194–201.
2. C. A. Coelho, *The wrapped Gamma distribution and wrapped sums and linear combinations of independent Gamma and Laplace distributions*, J. Statist. Theory Practice **1** (1) (2007) 1–29.
3. J. J. Fernández-Durán and M. M. Gregorío-Dominguez, *Bayesian analysis of circular distributions based on non-negative trigonometric sums*, J. Statist. Comput. Sim. **86** (16) (2016) 3175–3187.
4. N. I. Fisher, *Statistical Analysis of Circular Data*, Cambridge University Press, Cambridge, UK, 1995.
5. K. Fragiadakis, D. Karlis and S. G. Meintanis, *Inference procedures for the variance gamma model and applications,* J. Statist. Comput. Sim. **83** (3) (2013) 555–567.
6. R. Gatto, *A bootstrap test for circular data*, Commun. Statist.-Theory Methods **35** (2) (2006) 281–292.
7. M. Godfroy-Cooper, P. M. B. Sandor, J. D. Miller and R. B. Welch, *The interaction of vision and audition in two-dimensional space*, Front. Neurosci. **9** (2015) 311.
8. R. S. Jammalamadaka and T. J. Kozubowski, *New families of wrapped distributions foe modeling skew circular data*, Commun. Statist.-Theory Methods **33** (9) (2004) 2059–2074.
9. R. S. Jammaladaka and A. SenGupta, *Predictive inference for directional data*, Statist. Prob. Lett. **40** (3) (1998) 247–257.
10. S. Joshi and K. K. Jose, *Wrapped Lindley distribution*, Commun. Statist.-Theory Methods **47** (5) (2018) 1013–1021.
11. P. L. Levy, *L'addition des variables aléatoires définies sur une circonférence*, Bull. Soc. Math. France **67** (1939) 1–41.
12. D. B. Madan and E. Seneta, *The variance gamma (V.G.) model for share market returns,* J. Business **63** (4) (1990) 511–521.
13. K. V. Mardia, *Statistics of Directional Data*, Probability and Mathematical Statistics, Academic Press, London, 1972.
14. K. V. Mardia, *Statistics of directional data*, J. Royal Statist. Soc. Series B (Methodological) **37** (3) (1975) 349–393.
15. G. Mastrantonio and G. Calise, *Hidden Markov model for discrete circular-linear wind data time series*, J. Statist. Comput. Sim. **86** (13) (2016) 2611–2624.
16. S. Roy and M. A. S. Adnan, *Wrapped three parameter gamma distribution*, Proceedings of JSM 2010, Statistical Computing Section, American Statistical Association, (2010) pp. 4001–4014.
17. S. Roy and M. A. S. Adnan, *Wrapped weighted exponential distributions*, Statist. Prob. Lett. **82** (1) (2012) 77–83.
18. A. SenGupta and A. K. Laha, *A likelihood integrated method for exploratory graphical analysis of change point problem with directional data*, Commun. Statist.-Theory Methods **37** (11) (2008) 1783–1791.
19. P. R. Tadikamalla, *On a family of distributions obtained by the transformation of the gamma distribution*, J. Statist. Comput. Sim. **13** (3-4) (1981) 209–214.
20. D. Umbach and S. R. Jammalamadaka, *Building asymmetry into circular distributions*, Statist. Prob. Lett. **79** (5) (2009) 659–663.
21. M. T. Wells and A. SenGupta, *Advances in Directional and Linear Statistics*, A Festschrift for Sreenivasa Rao Jammalamadaka, Springer, New York, 2011.

E-mail: sajjadpiradl@yahoo.com

Author Index

Author Index

Author Index

www.aimc51.kashanu.ac.ir