The 10th Seminar on

Linear Algebra

and Its Applications

10

Extended Abstracts

Faculty of Mathematics and Computer &
Mahani Mathematical Research Center
Shahid Bahonar University of Kerman, Kerman, Iran

16-19 August, 2020
Kerman, Iran

The 10th Seminar on

Linear Algebra and its Applications

Faculty of Mathematics and Computer & Mahani Mathematical Research Center

Shahid Bahonar University of Kerman, Kerman, Iran

16-19 August, 2020

Extended Abstracts

Dedicated to Alireza Afzalipour and Fakhereh Saba,
the Founders of Kerman University

Table of Contents

Abstract of Invited Speakers

List of Papers

Part 1: Talks

Part 2: Posters

Preface

It is our great pleasure and honor to welcome you at the 10^{th} Seminar on Linear Algebra and its Applications (SLAA10). The SLAA10 will be held on 16-19 August 2020, hosted by the Faculty of Mathematics and Computer & Mahani Mathematical Research Center, Shahid Bahonar University of Kerman, Iran. As a part of the series of the bi-annually held seminars of the Iranian Mathematical Society (IMS), this seminar aims to create a friendly discussion atmosphere for researchers in linear algebra and numerical linear algebra.

The SLAA10 received 130 submissions in that each submission was reviewed by three dedicated members of the Scientific Committee. According to a thorough discussion by the reviewers, 76 submissions were accepted for publication: 59 as oral presentations and 17 as posters. In addition, the third and fourth Mehdi Radjabalipour Prize for Linear Algebra and its Applications will be awarded during this seminar.

We are proud to present a very interesting program. The Seminar program included 10 plenary talks with distinguished invited speakers and two workshops: "Linear Algebra in Data Mining" and "Software in Numerical Linear Algebra (Chebfun Toolbox)".

Finally, we immensely thank the authors for submitting their research papers to the SLAA10, and are grateful to the members of the Scientific Committee for dedicating their attention and time to assessing the papers. We are also very thankful to the members of the Executive Committee for their efforts in the arrangement, promotion, and organization of the seminar.

Seminar Organizing Committee

Scientific Committee

1 Ali Armandnejad . . . . . Vali-e-Asr University of Rafsanjan
2 Gholamreza Aghamollaei . . . . . Shahid Bahonar University of Kerman
3 Morad Ahmadnasab . . . . . University of Kurdistan
4 Seyyed Alireza Ashrafi . . . . . University of Kashan
5 Mahdi Eftekhari . . . . . Shahid Bahonar University of Kerman
6 Hamidreza Afshin . . . . . Vali-e-Asr University of Rafsanjan
7 Saieed Akbari . . . . . Sharif University of Technology
8 Ghasem Barid Loghmani . . . . . Yazd University
9 Hamid Beigy . . . . . Sharif University of Technology
10 Azita Tajaddini . . . . . Shahid Bahonar University of Kerman
11 Ali Taghavi . . . . . University of Mazandaran
12 Faezeh Totounian . . . . . Ferdowsi University of Mashhad
13 Hassan Hajiabolhassan . . . . . Shahid Beheshti University
14 Ali Hamzeh . . . . . Shiraz University
15 Davod Khojasteh Salkuyeh . . . . . University of Guilan
16 Mohammad Ali Dehghan . . . . . Vali-e-Asr University of Rafsanjan
17 Mehdi Rajabalipour . . . . . Shahid Bahonar University of Kerman
18 Hyedar Radjavi . . . . . University of Waterloo
19 Azim Rivaz . . . . . Shahid Bahonar University of Kerman
20 Abbas Salemi Parizi . . . . . Shahid Bahonar University of Kerman
21 Mohammad Shahryari . . . . . Sultan Qaboos University
22 Farshid Abdollahi . . . . . Shiraz University
23 Ataollah Askari Hemmat . . . . . Shahid Bahonar University of Kerman
24 Farzad Fathi Zadeh . . . . . Swansea University

25 Kazem Ghanbari . . . . . Sahand University of Technology
26 Hossein Mohebi . . . . . Shahid Bahonar University of Kerman
27 Mahmoud Mohseni Moghadam . . . . . Shahid Bahonar University of Kerman
28 Mohammad Sal Moslehian . . . . . Ferdowsi University of Mashahad
29 Seyed Mahmoud Manjegani . . . . . Isfahan University of Technology
30 Hossein Momenaee . . . . . Shahid Bahonar University of Kerman
31 Hossein Nezamabadi-pour . . . . . Shahid Bahonar University of Kerman
32 Alimohammad Nazari . . . . . Arak University
33 Ahmad Nickabadi . . . . . Amirkabir University of Technology
34 Behnam Hashemi . . . . . Shiraz University of Technology
35 Seyyed Mansour Vaezpour . . . . . Amirkabir University of Technology
36 Bamdad Reza Yahaghi . . . . . Golestan University

Executive Committee

1 Gholamreza Aghamollaei . . . . . Shahid Bahonar University of Kerman
2 Mohammad Ebrahimi . . . . . Shahid Bahonar University of Kerman
3 Vahid Amirzadeh . . . . . Shahid Bahonar University of Kerman
4 Asma Ilkhanizadeh Manesh . . . . . Vali-e-Asr University of Rafsanjan
5 Azita Tajaddini . . . . . Shahid Bahonar University of Kerman
6 Ali Jabari Shahzadeh Mohammadi . . . . . Shahid Bahonar University of Kerman
7 Mina Jamshidi . . . . . Graduate University of Advanced Technology
8 Soodeh Hosseini . . . . . Shahid Bahonar University of Kerman
9 Seyed Mohammad Mehdi Hosseini . . . . . Shahid Bahonar University of Kerman
10 Fatemeh Khalooei . . . . . Shahid Bahonar University of Kerman
11 Maryam Khosravi . . . . . Shahid Bahonar University of Kerman
12 Farzad Dadipour . . . . . Graduate University of Advanced Technology
13 Mohammad Hossein Daryaee . . . . . Shahid Bahonar University of Kerman
14 Esmaeil Rostami . . . . . Shahid Bahonar University of Kerman
15 Sharifeh Rezagholi . . . . . Payame Noor University, Kerman
16 Abolfazl Rafiepour . . . . . Shahid Bahonar University of Kerman
17 Somayeh Zangoei Zadeh . . . . . Shahid Bahonar University of Kerman
18 Abbas Salemi Parizi . . . . . Shahid Bahonar University of Kerman
19 Habibollah Saeedi . . . . . Shahid Bahonar University of Kerman
20 Mohammad Soleymani Baghshah . . . . . Shahid Bahonar University of Kerman
21 Nosratollah Shajareh Porsalavati . . . . . Shahid Bahonar University of Kerman
22 Alemeh Sheikhhosseini . . . . . Shahid Bahonar University of Kerman
23 Ayyub Sheikhi . . . . . Shahid Bahonar University of Kerman
24 Farid Saberi-Movahed . . . . . Graduate University of Advanced Technology

Organizers

Shahid Bahonar
University of Kerman

IRANIAN MATHEMATICAL SOCIETY

Sponsors

ISC
Islamic World Science Citation Center

مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری
Regional Information Center for Science and Technology RICeST

The Office of President
Vice-Presidency for Science and Technology
Iran National Science Foundation

Ministry of Interior
Islamic Republic of Iran
Kerman Provincial Government

Vali-e-Asr University of Rafsanjan

Yazd University

Graduate University of Advanced Technology

University of Bam

Payame Noor University

TEJARAT BANK بانک تجارت

NICICO
NATIONAL IRANIAN COPPER INDUSTRIES CO.

BAREZ

Invited Speakers

1. Prof. Rajendra Bhatia, Ashoka University, India
2. Prof. Anne Greenbaum, University of Washington, USA
3. Prof. Chi-Kwong Li, College of William & Mary, USA
4. Prof. Michael L. Overton, New York University, USA
5. Dr. Fatemeh Panjeh Ali Beik, Vali-e-Asr University of Rafsanjan, Iran
6. Prof. Panayiotis Psarrakos, National Technical University of Athens, Greece
7. Prof. Mohammad Sal Moslehian, Ferdowsi University of Mashhad, Iran
8. Prof. Tin-Yau Tam, University of Nevada, USA
9. Prof. Nick Trefethen, University of Oxford, UK
10. Prof. Qingxiang Xu, Shanghai Normal University, China

A message from Chandler Davis

Dear Abbas Salemi,

It is kind and appropriate for you to take the occasion of this impressive gathering on Linear Algebra to celebrate the 75th birthday of Professor Mehdi Radjabalipour. I am eager to add my words of appreciation and gratitude to him.

He is the bond between us, being my doctoral student and your doctoral supervisor. But in the same sense, he is the bond from the long-past Operator Theory Seminar Peter Rosenthal and I ran at the University of Toronto, to the present and future flourishing community he led at Kerman after his return to his homeland. He gave us a glimpse of this community by inviting us to his international seminar in Kerman: not virtual, I was most grateful of the opportunity to take part in person, my first and only time in Iran. What a cheerful and impressive sight you all were! I was honoured to have this connection to such a healthy school, leading in addition to fruitful contacts since that time with you and other members of the Radjabalipour circle.

Long life to Mehdi Radjabalipour and his followers, and to his school, and to his science which is our science, and to his country in this world which is our world. Let no one divide us.

Chandler Davis
August 18, 2020

A message from Peter Rosenthal

I first met Mehdi shortly after he arrived in Toronto to begin his graduate studies. Heydar Radjavi had told me that Mehdi was a very strong student and it soon became apparent to me that Heydar was correct. Mehdi completed an excellent Ph.D. thesis under the supervision of Chandler Davis in 1973.

Over the many years since then, Mehdi oscillated between Canada and Iran.

I have been very privileged to be a co-author with Mehdi on a number of papers. Mehdi is an optimal co-author. He is very knowledgeable and very creative. He is also very generous; he contributed more than his share to each of the joint papers that he and I have been involved in. Thanks Mehdi.

Moreover, Mehdi is an extremely pleasant person to talk with, about mathematics and many other things.

Happy Birthday Mehdi !!!

Peter Rosenthal
August 15, 2020

Abstract of
Invited Speakers

Metrics and means on positive definite matrices

Rajendra Bhatia*

Ashoka University, Sonepat, India

Abstract

This lecture will be an introduction to the use of some geometric ideas in defining a mean (barycentre) of a collection of positive definite matrices. Developed over the last fifteen years, these ideas have found use in diverse areas, both theoretical and practical.

The lecture will be addressed to graduate students interested in operator theory, linear algebra and matrix analysis. It is recommended that they read up on the basic facts about positive definite matrices before the lecture (eg, from Horn and Johnson, Matrix Analysis, Chapter 7; or R. Bhatia, Positive Definite Matrices, Chapters 1 and 4.)

*Speaker. Email address: rajenbhatia@gmail.com

Crouzeix's conjecture, extremal Blaschke products, and K-spectral sets

Anne Greenbaum*

Department of Applied Mathematics, University of Washington

Joint with: Kelly Bickel, Michel Crouzeix, Pamela Gorkin, Kenan Li, Thomas Ransford, Felix Schwenninger, Elias Wegert

---

Abstract

In 2004, Michel Crouzeix conjectured that for any square matrix A and any polynomial (or analytic function) f,

$$\|f(A)\| \leq 2 \max_{z \in W(A)} |f(z)| \qquad \text{(Crouzeix's conjecture)},$$

where $W(A) := \{q^*Aq : q^*q = 1\}$ is the numerical range of A and $\|\cdot\|$ denotes the spectral norm. In 2017, Crouzeix and Palencia showed that the inequality holds if 2 is replaced by $1+\sqrt{2}$, but the original conjecture remains unproved.

The form of functions f that maximize $\|f(A)\| / \max_{z \in W(A)} |f(z)|$ is known: $f = B \circ \varphi$, where φ is any conformal mapping from $W(A)$ to the unit disk $\mathbb{D}$ and B is a finite Blaschke product of degree at most $n-1$, when A is an $n \times n$ matrix. For a given conformal mapping φ, the Blaschke product B_E that maximizes this ratio is referred to as an *extremal Blaschke product*. We discuss some known properties of extremal Blaschke products. For example, it is known that the left and right singular vectors corresponding to the largest singular value of $B_E \circ \varphi(A)$ are orthogonal to each other. An interesting property that has been observed numerically but has not been proved is that an extremal Blaschke product B_E corresponding to a given conformal mapping φ often has degree much less than $n-1$. We also do not know if/when the extremal Blaschke product is unique. I will give an example where there are two extremal Blaschke products.

Additional work has been aimed at showing that other sets Ω that do not necessarily contain $W(A)$ are K-spectral sets; that is, that for a given value K, $\|f(A)\| \leq K \max_{z \in \Omega} |f(z)|$ for all functions f analytic in Ω. We show that various annular regions are $(1+\sqrt{2})$-spectral sets and that a more general convex region with a circular hole or cutout is a $(3+2\sqrt{3})$-spectral set. I show how these results can be used to give bounds on the convergence of rational Krylov subspace methods.

---

*Speaker. Email address: greenbau@uw.edu

Joint numerical ranges and commutative matrices

Chi-Kwong Li*

Department of Mathematics, College of William & Mary, Institute for Quantum Computing, University of Waterloo

Joint with: Yiu-Tung Poon (Iowa State University) and Yashu Wang (National Chung Hsing University)

Abstract

The connection between the commutativity of a family of $n \times n$ matrices and their generalized joint numerical ranges is discussed. Implications of the results to representation theory and quantum information science will be mentioned.

*Speaker. Email address: ckli@math.wm.edu

Crouzeix's conjecture

Michael L. Overton*

Courant Institute of Mathematical Sciences, New York University, USA

Joint with: Anne Greenbaum and Adrian Lewis

Abstract

Crouzeix's conjecture is among the most intriguing developments in matrix theory in recent years. Made in 2004 by Michel Crouzeix, it postulates that, for any polynomial p and any matrix A, $\|p(A)\| \leq 2\max(|p(z)| : z \in W(A))$, where the norm is the 2-norm and $W(A)$ is the field of values (numerical range) of A, that is the set of points attained by v^*Av for some vector v of unit length. Crouzeix proved in 2007 that the inequality above holds if 2 is replaced by 11.08, and recently this was greatly improved by Palencia, replacing 2 by $1 + \sqrt{(2)}$. Furthermore, it is known that the conjecture holds in a number of special cases, including $n = 2$. We use nonsmooth optimization to investigate the conjecture numerically by locally minimizing the "Crouzeix ratio", defined as the quotient with numerator the right-hand side and denominator the left-hand side of the conjectured inequality. We also present local nonsmooth variational analysis of the Crouzeix ratio at conjectured global minimizers. All our results strongly support the truth of Crouzeix's conjecture.

*Speaker. Email address: mo1@nyu.edu

A survey on preconditioning techniques for double saddle point systems: spectral and field-of-values analyses

Fatemeh Panjeh Ali Beik[1,*] and Michele Benzi[2,†]

[1]Department of Mathematics, Vali-e-Asr University of Rafsanjan, P.O. Box 518, Rafsanjan, Iran

[2]Classe di Scienze, Scuola Normale Superiore, Piazza dei Cavalieri 7, 56126 Pisa, Italy

Abstract

In this talk some preconditioning techniques are presented for a class of linear systems with double Saddle point structure arising in finite element discretizations of coupled Stokes-Darcy flow [3, 4] and modeling of liquid crystals directors [5]. We investigate different preconditionering techniques including block preconditioners [1–3], constraint preconditioners [4] and augmented Lagrangian-based ones. We present spectral and field-of-value analyses of the exact versions of these preconditioners. Numerical experiments will be reported for test problems from two mentioned applications.

References

[1] F. A. P. Beik and M. Benzi. Iterative methods for double saddle point systems. *SIAM J. Matrix Anal. Appl.*, 39:902–921, 2018.

[2] F. A. P. Beik and M. Benzi. Block preconditioners for saddle point systems arising from liquid crystal directors modeling. *Calcolo.*, 55:29 2018.

[3] M. Cai, M. Mu and J. Xu. Preconditioning techniques for a mixed Stokes/Darcy model in porous media applications. *J. Comput. Appl. Math.*, 233:346–355, 2009.

[4] P. Chidyagwai, S. Ladenheim and D. B. Szyld. Constraint preconditioning for the coupled Stokes-Darcy system. *SIAM J. Sci. Comput.*, 38:A668–A690, 2016.

[5] A. Ramage and E. C. Jr. Gartland. A preconditioned nullspace method for liquid crystal director modeling. *SIAM J. Sci. Comput.*, 35:B226–B247, 2013.

*Speaker. Email address: f.beik@vru.ac.ir

†Email address: michele.benzi@sns.it

From matrices to matrix polynomials

Panayiotis J. Psarrakos*

Department of Mathematics,

School of Applied Mathematical and Physical Sciences,

National Technical University of Athens

Abstract

The study of matrix polynomials of higher degree has attracted considerable attention in recent decades. The interest has been motivated by a wide range of applications of polynomial eigenvalue problems in areas such as differential equations, systems theory, control theory, mechanics and vibrations. In this presentation, we will see how results of the standard matrix theory, concerning Jordan structure, pseudospectra, eigenvalue condition numbers, spectral distance problems, numerical ranges and (entry-wise) nonnegative matrices, have been extended to the setting of matrix polynomials in a natural way. In particular, basic matrix theory can be viewed as the study of a special case of matrix polynomials of first degree.

*Speaker. Email address: ppsarr@math.ntua.gr

Choi-Davis-Jensen inequality revisited

Mohammad Sal Moslehian*

Department of Pure Mathematics, Center of Excellence in Analysis on Algebraic Structures (CEAAS), Ferdowsi University of Mashhad, P. O. Box 1159, Mashhad 91775, Iran

Abstract

Let f be an operator convex function defined on an interval $J \subset \mathbb{R}$. Then the so-called Choi–Davis–Jensen inequality $f(\Phi(A)) \leq \Phi(f(A))$ holds for all self-adjoint operators A with spectrum in J and all unital positive linear maps Φ. The converse holds true. If f is convex but not operator convex, then it is known that the Choi–Davis-Jensen inequality remains valid for 2×2 Hermitian matrices A. Several variants and reverses of this inequality have been obtained by some mathematicians. In this talk, we explore recent results on this inequality as well as Kadison's inequality. In addition, some asymmetric Choi–Davis-Jensen inequalities are presented.

References

[1] R. Bhatia and R. Sharma, *Some inequalities for positive linear maps*, Linear Algebra Appl. **436** (2012), 1562–1571.

[2] J.-C. Bourin and É. Ricard, *An asymmetric Kadison's inequality*, Linear Algebra Appl. **433** (2010), 499–510.

[3] F. Hansen, H. Najafi, and M.S. Moslehian, Operator maps of Jensen-type, Positivity 22 (2018), no. 5, 1255–1263.

[4] M. Kian, M.S. Moslehian, and R. Nakamoto, Asymmetric Choi–Davis inequalities, preprint.

*Speaker. Email address: moslehian@um.ac.ir

Geometry and inequalities associated with symmetric space of noncompact type

Tin-Yau Tam*

University of Nevada, USA

Joint with: Luyining (Elaine) Gan and Xuhua (Roy) Liu

Abstract

Denote by $\mathbb{P}_n$ the space of $n \times n$ positive definite matrices. For $A, B \in \mathbb{P}_n$, the (metric) geometric mean was introduced by Pusz and Woronowicz (1975), while the spectral geometric mean by Fiedler and Pták (1997):

$$A\sharp B = A^{1/2}(A^{-1/2}BA^{-1/2})^{1/2}A^{1/2},$$
$$A\natural B = (A^{-1}\sharp B)^{1/2}A(A^{-1}\sharp B)^{1/2}.$$

where $\sharp$ and $\natural$ denote the (metric) geometric mean and spectral geometric mean.
The t-(metric) geometric mean and t-spectral geometric mean are paths joining A and B in $\mathbb{P}_n$, $t \in [0, 1]$:

$$A\sharp_t B = A^{1/2}(A^{-1/2}BA^{-1/2})^t A^{1/2},$$
$$A\natural_t B = (A^{-1}\sharp B)^t A(A^{-1}\sharp B)^t.$$

P_n can be equipped with a suitable Riemannian metric so that the curve $A\sharp_t B$ with $0 \le t \le 1$ is the unique geodesic joining A and B in P_n.
The t-spectral geometric mean was introduced by Ahn, Kim and Lim (2007). When $t = 1/2$, they are abbreviated as $A\sharp_{1/2}B = A\sharp B$ and $A\natural_{1/2}B = A\natural B$.
We shall discuss the (metric) geometric mean and spectral geometric mean, first in the space of P_n and then in the context of symmetric space associated with a noncompact semisimple Lie group.

*Speaker. Email address: ttam@unr.edu

Chebfun and continuous linear algebra

Nick Trefethen*

University of Oxford, UK

Abstract

At the heart of the Chebfun project is the realization of continuous analogues of the discrete structures and operations of numerical linear algebra. For example, there are continuous analogues of the QR and LU decompositions and the SVD. This talk will review the mathematics and the algorithms of continuous linear algebra with Chebfun demonstrations.

*Speaker. Email address: trefethen@maths.ox.ac.uk

Generalized parallel sum of adjointable operators on Hilbert C^*-modules

Qingxiang Xu*

Department of Mathematics, Shanghai Normal University
Shanghai 200234, P.R. China

Joint with: C. Fu, M.S. Moslehian and A. Zamani

Abstract

We introduce the notion of a tractable pair of operators as well as that of the generalized parallel sum in the setting of adjointable operators on Hilbert C^*-modules. Some significant results about the parallel sum known for matrices and Hilbert space operators are extended to the case of the generalized parallel sum. In particular, a factorization theorem on the parallel sum is proved, and a common upper bound of two positive operators is constructed in the Hilbert C^*-module case. The harmonic mean for positive operators on Hilbert C^*-modules is also dealt with.

References

[1] W. N. Anderson, Jr. and R. J. Duffin, Series and parallel addition of matrices, J. Math. Anal. Appl. 26 (1969), 576–594.

[2] R. G. Douglas, On majorization, factorization, and range inclusion of operators on Hilbert spaces, Proc. Amer. Math. Soc. 17 (1966), 413–415.

[3] X. Fang, M. S. Moslehian, and Q. Xu, On majorization and range inclusion of operators on Hilbert C^*-modules, Linear Multilinear Algebra 66 (2018), no. 12, 2493–2500.

[4] P. A. Fillmore and J. P. Williams, On operator ranges, Adv. Math. 7 (1971), 254–281.

[5] C. Fu, M. S. Moslehian, Q. Xu and A. Zamani, Generalized parallel sum of adjointable operators on Hilbert C^*-modules, Linear Multilinear Algebra, accepted

[6] W. Luo, C. Song, and Q. Xu, Perturbation estimation for the parallel sum of Hermitian positive semi-definite matrices, Linear Multilinear Algebra 67 (2019), no. 10, 1971–1984.

[7] W. Luo, C. Song, and Q. Xu, The parallel sum for adjointable operators on Hilbert C^*-modules, Acta Math. Sinica (Chinese Series) 62 (2019), no. 4, 541–552.

[8] S. K. Mitra and P. L. Odell, On parallel summability of matrices, Linear Algebra Appl. 74 (1986), 239–255.

[9] M. Uchiyama, Operator functions and the operator harmonic mean, Proc. Amer. Math. Soc. 148 (2020), no. 2, 797–809.

*Speaker. Email address: qingxiang_xu@126.com

Papers

Part 1: Talks

Numerically solving the singular semi-Sylvester equation[1]

Majid Adib and Alireza Movahedian*

Department of Mathematics, University of Zanjan, Zanjan, Iran

Abstract

Matrix equations are one of the most widely used equations in various sciences. The Sylvester equation is one of these important equations. In this paper we define singular semi-Sylvester equation and then solve it using the Drazin-inverse generalized minimum residual method. Finally, we show the efficiency of our method.

Keywords: Sylvester equations, Drazin-inverse

Mathematics Subject Classification [2010]: 15A03

1 Introduction

The semi-Sylvester equation $AX - EBX = C$ where $A \in \mathbb{R}^{n\times n}$, $E \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times n}$ and $C \in \mathbb{R}^{n\times s}$ are given and $X \in \mathbb{R}^{n\times s}$ is to be determined, is one of the most important matrix equations in theory and applications and appear frequently in many areas. Several direct and iterative methods are proposed for solving semi-Sylvester equation. During last years, sevral projection methods based on Krylov subspace methods have also been proposed [4]. Karimi and Attarzadeh showed that in a particular case, the semi-Sylvester equation $AX - EBX = C$ can be converted into the following multiple linear systems [3]

$$A^{(i)}x^{(i)} = b^{(i)}, \qquad i = 1, 2, ..., s. \tag{1}$$

Ton et al. presented the Galerkin projection method for solving multiple linear systems [2]. Karimi and Attarzadeh have considered a special case of the semi-Sylvester equation [3], in which the matrix B is normal. They studied the nonsingular case of multiple linear systems (1) by presenting the following proposition and in this case, they applied Galerkin projection method to solve the semi-Sylvester equation.

Proposition 1.1. *(a) Assume A and B are symmetric matrices and E is symmetric positive definite matrix and*

$$\lambda_j < \frac{\langle Ax, x\rangle}{\langle Ex, x\rangle}, \qquad j = 1, 2, ..., s, \tag{2}$$

where λ_j is the eigenvalues of B. Then $\hat{A}^{(i)} = A - \lambda_i E$ is symmetic positive definite.

(b) Let A, B and E be symmetric positive definite matrices and symmetric positive semi-definite matrix, respectively. Then $(A-\lambda_j E)$, $j = 1, 2, ..., s$ are symmetric positive definite, where λ_j is the eigenvalues of B.

[1] Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

* Speaker. Email address: ali.movahedian96@gmail.com

In this paper, we intend to consider a general case that the above propositions 1.1 dose not exist, that is, the multiple linear systems (1) be singular, so in this regard, we provide the following definition.

Definition 1.2. We say that the multiple linear systems (1) is singular, if at least one of the coefficients matrices is singular. Also we say that the semi-Sylvester equation is singular if the corresponding multiple linear systems (1) is singular.

Now assume that the semi-sylvester equation is singular. In this case, we apply the Drazin-inverse and $DGMRES(m)$ method for solving the multiple linear systems (1) and hence the semi-Sylvester equation. The results of this method will be compared with the results of Galerkin projection method [3], in point of view CPU-time, accurancy and iteration number. Note that the semi-Sylvester equation is the generalization of the standard Sylvester equation (this means that, if E is identity matrix I or an arbitary nonsingular matrix then the semi-Sylvester equation becomes the standard Sylvester equation).

2 Drazin-inverse generalized minimum residual method

Consider the following linear system $Ax = b$ where $A \in R^{n\times n}$ is a singular matrix, $b \in R^n$ and $ind(A)$ is α. Here $ind(A)$ is the smallest nonnegative number that satisfy in $rank(A^{\alpha+1}) = rank(A^{\alpha})$. The matrix $X \in R^{n\times n}$ satisfying the conditions

$$AX = XA, \quad A^{\alpha}XA = A^{\alpha}, \quad XAX = X,$$

is called the Drazin-inverse of the matrix A. The Drazin-inverse of A denoted by A^D. In [5] the author proposed an effective model of usage for $DGMRES$, denoted $DGMRES(m)$, which is analogous to the $GMRES(m)$ and requires a fixed amount of storage for its implementation. In restarted $DGMRES$ ($DGMRES(m)$) the method is restarted once Krylov subspace reachs dimension m, and the current approximate solution becomes the new initial guess for the next m iterations. The restart parameter m is generally chosen small relative to n to keep storage and computation requirments reasonable. In the sequel, we review the $DGMRES(m)$ method.
$DGMRES(m)$ method is a Krylov subspace method for computing the Drazin-inverse solution of consistent or inconsistent linear system $Ax = b$ [6]. In this method, there are not any restriction on the matrix A. Thus, in general, A is non-Hermitian,$\alpha = ind(A)$ is arbitrary, and the spectrum of A can be any shape. Thus, it is unnecessary for us to put any restriction on the linear system $Ax = b$. So the system may be consistent or inconsistent. We only assume that $ind(A)$ is known. $DGMRES(m)$ method starts with an initial vector x_0 and generates a sequence of vectors $x_1, x_2, \cdots$ as follows

$$x_m = x_0 + q_{m-1}(A)r_0, \qquad r_0 = b - Ax_0, \tag{3}$$

where $q_{m-1}(\lambda)$ is a polynomial in λ of degree at most $m-1$ defined as follows

$$q_{m-1}(\lambda) = \sum_{i=1}^{m-\alpha} c_i \lambda^{\alpha+i-1}, \qquad \alpha = ind(A). \tag{4}$$

We define $p_m(\lambda) = 1 - \lambda q_{m-1}(\lambda)$ and $r_m = p_m(A)r_0$. Thus we have

$$x_m = x_0 + \sum_{i=1}^{m-\alpha} c_i A^{\alpha+i-1}, \qquad r_m = b - Ax_m = r_0 - \sum_{i=1}^{m-\alpha} c_i A^{\alpha+i} r_0. \tag{5}$$

The Krylov subspace used is as follows

$$\mathcal{K}_{m-\alpha}(A, A^{\alpha}r_0) = span\{A^{\alpha}r_0, A^{\alpha+1}r_0, \cdots, A^{m-1}r_0\}. \tag{6}$$

We can orthogonize the Krylov vectors $\{A^{\alpha}r_0, A^{\alpha+1}r_0, \cdots, A^{m-1}r_0\}$ by the Arnoldi- modified Gram-Schmidt process [1,4]. Let we set resulting orthonormal vectors as the columns of the matrix $\hat{V}_k$ as follows

$$\hat{V}_k = [v_1|v_2\cdots|v_k], \qquad k = 1, 2, \cdots, m. \tag{7}$$

Thus we can write

$$x_m = x_0 + \hat{V}_{m-\alpha}\xi_m, \qquad \xi \in \mathbb{R}^{m-\alpha}, \tag{8}$$

which we need to determine ξ_m. First, note that $r_m = r_0 - A\hat{V}_{m-\alpha}\xi_m$, so we have

$$A^{\alpha}r_m = A^{\alpha}r_0 - A^{\alpha+1}\hat{V}_{m-\alpha}\xi_m = \beta v_1 - A^{\alpha+1}\hat{V}_{m-\alpha}\xi_m. \tag{9}$$

Next, we write $A\hat{V}_k = \hat{V}_{k+1}\bar{H}_k$ where

$$\bar{H}_k = \begin{bmatrix} h_{11} & h_{12} & \cdots & \cdots & h_{1k} \\ h_{21} & h_{22} & \cdots & \cdots & h_{2k} \\ 0 & h_{32} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & h_{kk} \\ 0 & \cdots & \cdots & 0 & h_{k+1,k}. \end{bmatrix}. \tag{10}$$

Note that $\bar{H}_k \in \mathbb{R}^{(k+1)\times k}$ and $rank(\bar{H}_k) = k$. If we apply (10) to $A^{\alpha+1}\hat{V}_{m-\alpha}$ we have

$$A^{\alpha+1}\hat{V}_{m-\alpha} = A^{\alpha}\hat{V}_{m-\alpha+1}\bar{H}_{m-\alpha} = A^{\alpha-1}\hat{V}_{m-\alpha+2}\bar{H}_{m-\alpha+1}\bar{H}_{m-\alpha} = \hat{V}_{m+1}\hat{H}_m,$$
$$\hat{H}_m = \bar{H}_m\bar{H}_{m-1}\cdots\bar{H}_{m-\alpha}.$$

Thus $A^{\alpha}r_m = \beta v_1 - \hat{V}_{m+1}\hat{H}_m\xi_m$. We also have $\hat{V}_{m+1}^T\hat{V}_{m+1} = I_{(m+1)\times(m+1)}$ and $rank(\hat{H}_m) = m - \alpha$. We finally have the $(m+1)\times(m-\alpha)$ least squares problem

$$\|A^{\alpha}r_m\| = \|\beta e_1 - \hat{m}\xi_m\| = \min_{\xi\in\mathbb{R}^{m-\alpha}} \|\beta e_1 - \hat{H}_m\xi\| \tag{11}$$

Note that n is normally very large and $m \ll n$, which implies that the problem in (11) is very small. Also, note that since $\hat{H}_m$ is a full rank, we can determine ξ_m by applying the QR decomposition on $\hat{H}_m$. Thus $\hat{H}_m = Q_mR_m$, where $Q_m \in \mathbb{R}^{(m+1)\times(m-\alpha)}$ is a unitary matrix, that is, $Q_m^TQ_m = I_{(m-\alpha)\times(m-\alpha)}$ and R_m is an upper triangular matrix. Since $\hat{H}_m$ is full rank, so R_m is nonsingular, therefore we can compute ξ_m by solving the upper triangular system as follows

$$R_m\xi_m = \beta(Q_m^Te_1), \qquad e_1 = [1, 0, \cdots, 0]^T. \tag{12}$$

Consequently, the algorithm of the $DGMRES(m)$ method is as follows

Algorithm 2.1. (DGMRES(m) algorithm).

1. Choose an initial guess $x_0 = 0$ and compute $r_0 = b - Ax_0$ and $A^{\alpha}r_0$.
2. Compute $\beta = \|A^{\alpha}r_0\|$ and set $v_1 = \beta_{-1}(A^{\alpha}r_0)$.

3. Orthogonalize the Krylov vectors $A^{\alpha}r_0, A^{\alpha+1}r_0, \cdots, A^{m+\alpha+1}r_0$ via the Arnoldi-Gram-Schmidt process carried out like the modified Gram-Schmidt process:
   For $j = 1, \cdots, m$ do
   $u = Av_j$
   For $i = 1, \cdots, j$ do
   $h_{i,j} = \langle u, v_j \rangle$
   $u = u - h_{i,j}v_i$
   end
   $h_{j+1,j} = \|u\|, \qquad v_{j+1} = u/h_{j+1,j}$
   end (The vectors $v_1, v_2, \cdots, v_{m+1}$ obtained by this way form an orthonormal set.)

4. For $k = 1 : m$ form the matrices $\hat{V}_k \in \mathbb{R}^{n\times k}$ and $\bar{H}_k \in \mathbb{R}^{(k+1)\times k}$

$$\hat{V}_k = [v_1|v_2|\cdots|v_k], \qquad \bar{H}_k = \begin{bmatrix} h_{11} & h_{12} & \cdots & \cdots & h_{1k} \\ h_{21} & h_{22} & \cdots & \cdots & h_{2k} \\ 0 & h_{32} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & h_{kk} \\ 0 & \cdots & \cdots & 0 & h_{k+1,k}. \end{bmatrix}$$

5. Form the matrix $\hat{H}_m = \bar{H}_m\bar{H}_{m-1}\cdots\bar{H}_{m-\alpha}$.

6. Compute the QR decomposition of $\hat{H}_m$: $\hat{H}_m = Q_mR_m; Q_m \in \mathbb{R}^{(m+1)\times(m-\alpha)}$ and $R_m \in \mathbb{R}^{(m-\alpha)\times(m-\alpha)}$. ($R_m$ is upper triangular.)

7. Solve the (upper triangular) system $R_m\xi_m = \beta(Q_m^Te_1)$, where $e_1 = [1, 0, \cdots, 0]^T$.

8. Compute $x_m = x_0 + \hat{V}_{m-\alpha}\xi_m$(then $\|A^{\alpha}r_m\| = \beta\sqrt{1 - \|Q_m^Te_1\|^2}$). If satisfied then stop.

9. Set $x_0 = x_m$, compute $r_0 = b - Ax_0$, and go to 2.

3 Numerically solving the semi-Sylvester equation and some experiments

In this section, we want to numerically solve the semi-Sylvester equation $AX - EBX = C$, by using the following theorem.

Theorem 3.1. *Let $A \in \mathbb{R}^{n\times n}$. Then A is a normal matrix if and only if it is unitarily similar to a diagonal matrix*

Now let in the semi-Sylvester equation, B is a normal matrix. So, according to Theorem 3.1 there are a unitary matrix Q_B and a diagonal matrix Λ_B such that

$$B = Q_B\Lambda_BQ_B^T, \tag{13}$$

where the diagonal components of Λ_B are eigenvalues of B and the columns of the unitary matrix Q_B are normalized eigenvectores of B. By substitution of (13) in $AX - EBX = C$,

we have $AXQ_B - EXQ_B\Lambda_B = CQ_B$. By taking $\hat{X} = XQ_B$ and $\hat{C} = CQ_B$, we obtain the following multiple linear systems

$$\hat{A}^{(i)}\hat{x}^{(i)} = \hat{c}^{(i)}, \qquad i = 1, 2, \cdots, s, \tag{14}$$

where $\hat{A}^{(i)} = (A - \lambda_i E)$, $\hat{x}^{(i)}$ is the i-th column of $\hat{X}$ and $\hat{c}^{(i)}$ is the i-th column of $\hat{C}$. Therefore, the semi-Sylvester equation is converted to s linear systems. Notice, in this paper we considered the general case; that is, we did not impose any conditions and constraints on coefficients matrices of the resulting system. Therefore, it is possible to solve the semi-Sylvester equation by using s-time of the $DGMRES(m)$ method.

Now we use the corresponding multiple linear systems form (form (14)) to solve the semi-Sylvester equation and we consider the singular case. In this case, we used the $DGMRES(m)$ method to solve these systems. The described method is written with MATLAB.. In the following, we give an example. In this example the coefficients matrices are singular and ill-conditioned. The initial matrix X_0, is the zero matrix and the stop condition is $\|A^{\alpha}r_i\|_2 \leq 1e-04$. The results obtained are presented in following table which are compared with Galerkin projection method in point of view CPU-time, iteration numbers and residuals norm. In the table the symbols **itration** and **time** are total iteration numbers and total CPU-time respectively.

Example 3.2. In this example we consider semi-Sylvester equation that coefficients matrices are singular, the maximum condition number is $3:36e+22$ and $ind(A(i))$ are all equal to 5. The matrices constituting the semi-Sylvester are as follows:

$$A = 5 * hilb(n,n), \qquad E = hilb(n,n),$$
$$B = tridiag\left(-1 + \frac{1}{1+s}, 5, -1 + \frac{1}{1+s}\right), \qquad C = ones(n,s),$$

where $n = 1000$ and $s = 4$. The numerical results obtained

method(1000,4,m)	problem	Tol	time(s)	iteration	min $\|A^{\alpha}r_i\|_2$	max $\|A^{\alpha}r_i\|_2$
Galerkin	exapmle 3.2	1e-04	2.17	430	2.2590e-19	7.6387e-05
DGMRES(10)	exapmle 3.2	1e-04	1.35	4	6.5855e-23	2.0287e-05

4 Conclusions

As the results of the table presented in the previous section show that when the coefficients matrices are singular and ill-conditioned, in point of view CPU-time, iteration numbers and residuals norm $\|A^{\alpha}r_i\|_2$, the $DGMRES(m)$ method has a more better performance than the Galerkin projection method.

References

[1] W. E. Arnoldi. The principle of minimized iterations in the solution of the matrix eigenvalue problem. *Quarterly of applied mathematics*, 9(1)(1951),17–29.

[2] T. F. Chan and M. K. Ng. Galerkin projection methods for solving multiple linear systems. *SIAM Journal on Scientific Computing,* 21(3)(1999),836–850.

[3] S. Karimi and F. Attarzadeh. A new iterative scheme for solving the semi sylvester equation. *Applied Mathematics,* 4(01)(2013):6.

[4] M. Robb'e and M. Sadkane. Use of near-breakdowns in the block arnoldi method for solving large sylvester equations. *Applied Numerical Mathematics,* 584(9)(2008),486–498.

[5] A. Sidi. A unified approach to krylov subspace methods for the drazin-inverse solution of singular nonsymmetric linear systems. *Linear algebra and its applications,* 298(1–3)(1990),99–113.

[6] A. Sidi. Dgmres: A gmres-type algorithm for drazin-inverse solution of singu- lar non-symmetric linear systems. *Linear algebra and its applications,* 335(1–3)(2001),189–204.

Implicitly restarted global GMRES for solving $AXB = C$[1]

Najmeh Azizizadeh[1,*], Azita Tajaddini[2] and Amin Rafiei[3]

[1,2]Department of Applied Mathematics, Faculty of Mathematics & Computer Sciences, Shahid Bahonar University of Kerman, Kerman, Iran

[3]Department of Applied Mathematics, Hakim Sabzevari University, Sabzevar, Iran

Abstract

Global Krylov subspace methods are generally used with restarting to reduce storage costs. At the time of restart, some information is lost and this slows down the convergence. Here, an implicitly restarted global GMRES method is proposed that uses the implicitly generalized global Arnoldi algorithm to retain this information. This method deflates the smallest eigenvalues and augments the approximate block harmonic Ritz vectors to the generalized Krylov subspace but not with the usual starting block vector. Ultimately, the efficiency of this method is evaluated by virtue of an example.

Keywords: Harmonic Ritz value, Deflation, GLGMRES-IR

Mathematics Subject Classification [2010]: 15A23

1 Introduction

Consider the following matrix equation

$$AXB = C \tag{1}$$

where $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{s\times s}$ are nonsingular and $C \in \mathbb{R}^{n\times s}$ $(s \ll n)$ are given matrices and $X \in \mathbb{R}^{n\times s}$ is an unknown matrix. Note that the matrix equation (1) can be reformulated by the following linear system

$$\mathcal{A}x = c,$$

where $\mathcal{A} = B^T \otimes A$ and the vectors $x = vec(X)$ and $c = vec(C)$. However, it seems quite costly and ill-conditioned to solve the above linear system of equations. The matrix equation (1) was put into quite a few applications such as control theory and image restoration. Over the last decade, several iterative methods have been proposed to solve the matrix equation (1), for example, the global GMRES methods [4], the NSCG method [1]. The convergence of the global FOM and global GMRES for solving $AXB = C$ are investigated in [2]. For solving matrix equations, we are able to make use of restarting techniques. Restarting is fundamentally needed to reduce storage requirements and orthogonalization costs. However, restarting slows down the convergence and makes the choice of the

[1]Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

[*]Speaker. Email address: nazizizadeh@yahoo.com

new starting vector difficult. To overcome these problems, implicit restarting has been presented as a new variant of restarting.

This new technique can be viewed as a truncated form of the implicitly shifted QR iteration. In [5], implicit restarting with Arnoldi iteration is applied and is showed that the rate of convergence improves. In [3], implicitly restarted GMRES for solving nonsymmetric equations has investigated.

In this paper, the restarting global Krylov subspace method is employed for solving the matrix equation (1). At the time of restart, some information is lost and this slows down the convergence. However, some important information is kept at the restart time. So to store this information, it should be updated a starting matrix of the generalized global Arnoldi algorithm. Accordance with this, the generalized global Arnoldi factorization is updated through QR iterations. This iteration is an extension of the implicit double- shift QR iteration. This iterative scheme is called the implicitly generalized global Arnoldi algorithm (IGGA). Eventually, the global GMRES is combined with IGGA Algorithm to solve matrix equation (1). This method is called implicitly restarted global GMRES (Gl-GMRES-IR).

This paper is organized as follows. In section 2, the generalized global Arnoldi process is reviewed. Subsection 2.1 is allocated to the implicitly generalized global Arnoldi process and its relations. In section 4, the Gl-GMRES method and its eigenvalue problem are investigated. Eventually, this method will be compared with another methods. Thoroughout this paper, the following notations are used. Let $\mathbb{R}^{m\times n}$ be the set of $m \times n$ real matrices. The symbols A^T, $\|A\|_2$ and $trace(A)$ will denote the transpose, 2-norm and trace of a matrix $A \in \mathbb{R}^{m\times n}$, respectively. For any two matrices A and B in $\mathbb{R}^{n\times s}$, $< A, B >_F = trace(A^T B)$ is defined as the inner product. The associated norm is the Frobenius norm obtained by $\|.\|_F$. Further, $vec(.)$ will stand for the vec operator, i.e. $vec(A) = (a_1^T, a_2^T, ..., a_n^T)^T$ for the matrix $A = (a_1, a_2, ..., a_s) \in \mathbb{R}^{n\times s}$, where $a_j, j = 1, 2, ..., s$ is the j-th column of A and $A \otimes B = (a_{ij}B)$ denotes the Kronecker product of the matrices A and B. Let $A = [A_1, ..., A_p] \in \mathbb{R}^{n\times ps}$ and $B = [B_1, ..., B_l] \in \mathbb{R}^{n\times ls}$, where $A_j, B_j \in \mathbb{R}^{n\times s}$. The matrix $A^T \diamond B$ is defined by $(A^T \diamond B)_{ij} =< A_i, B_j >_F$.

2 The generalized global Arnoldi process

Let V be a matrix of size $n \times s$. Then the generalized Krylov subspace is defined as

$$\begin{aligned}\mathcal{G}K_m(A, V, B) =& span\{V, AVB, A^2VB^2, \ldots, A^{m-1}VB^{m-1}\}\\ =&\{\sum_{i=1}^{m} \alpha_i A^{i-1}VB^{i-1} | \alpha_i \in \mathbb{R}, \quad i = 1, \ldots, m\},\end{aligned}$$

where $A^0 = B^0 = I$.
The generalized global Arnoldi process allows us to construct an F-orthonormal basis for the generalized Krylov subspace, for more details see [4]. This Algorithm uses the Gram-Schmidt process to compute an F-orthonormal basis $\mathcal{V}_m = [V_1, V_2, \ldots, V_m]$, $V_i \in \mathbb{R}^{n\times s}$ for the generalized Krylov subspace $\mathcal{G}K_m(A, V, B)$ and an upper Hessenberg matrix $H_m \in \mathbb{R}^{m\times m}$. Let $\overline{H}_m$ be the corresponding $m + 1$ by m matrix with last row having only the nonzero element $h_{m+1,m}$. The relations:

$$A\mathcal{V}_m(I_m \otimes B) = \mathcal{V}_m(H_m \otimes I_s) + \hat{R}_m(e_m^T \otimes I_s),$$

$$= \mathcal{V}_{m+1}(\overline{H}_m \otimes I_s), \tag{2}$$

$$\mathcal{V}_m^T \diamond (A\mathcal{V}_m(I_m \otimes B)) = H_m, \tag{3}$$

hold, where $\mathcal{V}_m \in \mathbb{R}^{n\times ms}$. One can also verify that $\mathcal{V}_m^T \diamond \mathcal{V}_m = I_m, \mathcal{V}_m^T \diamond \hat{R}_m = 0$. $\hat{R}_m \in \mathbb{R}^{n\times s}$ is called the residual matrix. An alternative way to write (2) is as follows:

$$A\mathcal{V}_m(I_m \otimes B) = (\mathcal{V}_m, V_{m+1}) \begin{pmatrix} H_m \\ h_{m+1,m} e_m^T \end{pmatrix} \otimes I_s, \tag{4}$$

where $h_{m+1,m} = \|\hat{R}_m\|_F$ and $V_{m+1} = \frac{\hat{R}_m}{h_{m+1,m}}$.
By this representation, it is obvious that (4) is just a truncation of the complete reduction

$$A(\mathcal{V}_m, \hat{\mathcal{V}}_{n-m})(I_m \otimes B) =$$
$$(\mathcal{V}_m, \hat{\mathcal{V}}_{n-m}) \begin{pmatrix} H_m & M \\ h_{m+1,m}(e_1 e_m^T) & \hat{H}_{n-m} \end{pmatrix} \otimes I_s, \tag{5}$$

where $(\mathcal{V}_m, \hat{\mathcal{V}}_{n-m}) \in \mathbb{R}^{n\times n}$ is F-orthonormal, and $\hat{H}_{n-m} \in \mathbb{R}^{(n-m)\times(n-m)}$ is an upper Hessenberg matrix. In the subsection 2.1, the generalized global Arnoldi Algorithm through a new version of implicit shifted QR iteration is updated and so it is called implicitly generalized global Arnoldi process.

2.1 The implicitly generalized global Arnoldi process

In this section, the generalized global Arnoldi factorization via QR iterations is updated. This will lead to an updating formula that may be used to implement iterative techniques to derive the residual matrix $\hat{R}_k = h_{k+1,k}V_{k+1}$ to zero. In the following, we describe one iteration step of the p shifts of the generalized implicit shifted QR iteration.

Let the positive integer k be a fixed pre-specified integer of the modest size. Let p be another positive integer and consider $k + p$ steps of the generalized global Arnoldi Algorithm. Therefore, the relation

$$A\mathcal{V}_{k+p}(I_{k+p} \otimes B) = \mathcal{V}_{k+p}(H_{k+p} \otimes I_s) + \hat{R}_{k+p}(e_{k+p}^T \otimes I_s),$$

holds, where $\hat{R}_{k+p} = h_{k+p+1,k+p}V_{k+p+1}$.

Let τ_1 be a shift. In addition, consider the QR factorization of $H_{k+p} - \tau_1 I_{k+p} = Q_1 R_1$, where $Q_1 \in \mathbb{R}^{k+p\times k+p}$ is an orthogonal and R_1 is an upper triangular matrix. Then it can be easily shown that

$$\begin{aligned} &A\mathcal{V}_{k+p}(Q_1 \otimes I_s)(I_{k+p} \otimes B) - \mathcal{V}_{k+p}(Q_1 \otimes I_s)((\tau_1 I_{k+p} + R_1 Q_1) \otimes I_s) \\ &= h_{k+p+1,k+p}V_{k+p+1}(e_{k+p}^T Q_1 \otimes I_s). \end{aligned} \tag{6}$$

Since $H_{k+p} - \tau_1 I_{k+p} = Q_1 R_1$ and $Q_1^T Q_1 = I$. Multiplying this relation from right side in Q_1 and left side in Q_1^T, it gets

$$Q_1^T H_{k+p} Q_1 = R_1 Q_1 + \tau_1 I_{k+p}. \tag{7}$$

Substituting (7) in (6), it yields

$$A\mathcal{V}_{k+p}(Q_1 \otimes I_s)(I_{k+p} \otimes B) = (\mathcal{V}_{k+p}(Q_1 \otimes I_s), V_{k+p+1}) \begin{pmatrix} Q_1^T H_{k+p} Q_1 \\ h_{k+p+1,k+p}(e_{k+p}^T Q_1) \end{pmatrix} \otimes I_s,$$

where $Q_1^T H_{k+p} Q_1$ is still an upper Hessenberg matrix.

Now, if we apply p shifts, as a result, we will have

$$A\mathcal{V}_{k+p}^{+}(I_{k+p}\otimes B)=(\mathcal{V}_{k+p}^{+},V_{k+p+1})\begin{pmatrix} H_{k+p}^{+} \\ h_{k+p+1,k+p}(e_{k+p}^{T}Q)\end{pmatrix}\otimes I_s, \tag{8}$$

where $\mathcal{V}_{k+p}^{+}=\mathcal{V}_{k+p}(Q\otimes I_s)$, $H_{k+p}^{+}=Q^T H_{k+p}Q$ and $Q=Q_1\dots Q_p$ with Q_j be the orthogonal matrix associated with the shifts $\tau_j, j=1,2,\dots,p$. Now partition $\mathcal{V}_{k+p}^{+}$ and H_{k+p}^{+} in the following form

$$\mathcal{V}_{k+p}^{+}=(\mathcal{V}_{k}^{+},\hat{\mathcal{V}}_p),\qquad H_{k+p}^{+}=\begin{pmatrix} H_k^{+} & M \\ h_{k+1,k}^{+}e_1e_k^T & \hat{H}_p\end{pmatrix}, \tag{9}$$

and consider

$$h_{k+p+1,k+p}e_{k+p}^{T}Q=(0,0,\dots,\tilde{h}_{k+p+1,k+p},b^T). \tag{10}$$

Substituting (9) and (10) into (8), it follows that

$$A(\mathcal{V}_{k}^{+},\hat{\mathcal{V}}_p)(I_{k+p}\otimes B)=(\mathcal{V}_{k}^{+},\hat{\mathcal{V}}_p,V_{k+p+1})\Big(\begin{pmatrix} H_k^{+} & M \\ h_{k+1,k}^{+}e_1e_k^T & \hat{H}_p \\ \tilde{h}_{k+p+1,k+p}e_k^T & b^T\end{pmatrix}\otimes I_s\Big). \tag{11}$$

Since the first k columns on both sides of (10) are equal, then it obtains

$$A\mathcal{V}_{k}^{+}(I_k\otimes B)=\mathcal{V}_{k}^{+}(H_k^{+}\otimes I_s)+R_k^{+}(e_k^T\otimes I_s),$$

where $R_k^{+}=h_{k+1,k}^{+}\hat{\mathcal{V}}_p(e_1\otimes I_s)+\tilde{h}_{k+p+1,k+p}V_{k+p+1}$ (in fact, R_k^{+} is a new version of $\hat{R}_k$). Hence

$$A\mathcal{V}_{k}^{+}(I_k\otimes B)=(\mathcal{V}_{k}^{+},V_{k+1}^{+})\Big(\begin{pmatrix} H_k^{+} \\ h_{k+1,k}^{+}e_k^T\end{pmatrix}\otimes I_s\Big), \tag{12}$$

where $V_{k+1}^{+}=\frac{R_k^{+}}{h_{k+1,k}^{+}}$ and $h_{k+1,k}^{+}=\|R_k^{+}\|_F$. Note $\mathcal{V}_k^{+}\diamond(\hat{\mathcal{V}}_p e_1)=0$ and $\mathcal{V}_k^{+}\diamond V_{k+p+1}=0$. Thus (12) is a logical generalized global Arnoldi factorization of $(B^T\otimes A)$. The above process run until $R_k^{+}=0$. The above process is called the implicitly generalized global Arnoldi (IGGA).

3 The global GMRES method and its eigenvalue problem

In this section, a brief description of the Gl-GMRES method for solving the matrix equation (1) is given. For further details, refer to [2, 4].

Let $X_0\in\mathbb{R}^{n\times s}$ be an initial guess for (1). At the mth step of the Gl-GMRES method find the approximation solution $X_m=X_0+\mathcal{V}_m(d\otimes I_s)$ such that $d\in\mathbb{R}^m$, and the cloumns of $\mathcal{V}_m$ are an F-orthonormal basis for $\mathcal{GK}_m(A,R_0,B)$. The corresponding residual matrix will be $R_m=R_0-A\mathcal{V}_m(I_m\otimes B)(d\otimes I_s)$, which should be F-orthogonal to $A\mathcal{GK}_m(A,R_0,B)B$. This orthogonality relation is equivaletn to the minimiztion problem $min\Big\|\|R_0\|_F e_1-\overline{H}_m d\Big\|_2$. Now, in order to accelerate the convergence of the Gl-GMRES method, it requires to compute k $(1\le k\le m)$ harmonic Ritz pairs. Let the columns of $\mathcal{V}_m$ be the F-orthonormal basis of $\mathcal{GK}_m(A,R_0,B)$, we look for k harmonic Ritz pairs $(\tilde{\theta}_i,\tilde{g}_i)$ that satisfy

$$A\mathcal{V}_m(\tilde{g}_i\otimes I_s)(I_m\otimes B)-\tilde{\theta}_i\mathcal{V}_m(\tilde{g}_i\otimes I_s)\perp_F A\mathcal{GK}_m(A,\tilde{R}_0,B)(I_m\otimes B), \tag{13}$$

for $i = 1, 2, \cdots, k$.
Here, we want to deflate k smallest of eigenvalues in magnitude. Hence, we can compute $(\tilde{\theta}_i, \tilde{g}_i)$ via solving the following generalized small-sized eigenvalue problem

$$\Big((A\mathcal{V}_m(I_m \otimes B))^T \diamond \mathcal{V}_m\Big)\tilde{g}_i = \frac{1}{\tilde{\theta}_i}\Big((A\mathcal{V}_m(I_m \otimes B))^T \diamond (A\mathcal{V}_m(I_m \otimes B))\Big)\tilde{g}_i. \tag{14}$$

By (2) and (3), the relation (14) can be rewritten as

$$H_m^T\tilde{g}_i = \frac{1}{\tilde{\theta}_i}\overline{H}_m^T\overline{H}_m\tilde{g}_i. \tag{15}$$

If H_m is nonsingular, then relation (15) can be written to the following form

$$(H_m + h_{m+1,m}^2 H_m^{-T} e_m e_m^T)\tilde{g}_i = \tilde{\theta}_i\tilde{g}_i. \tag{16}$$

Also, we define "the harmonic Ritz block vectors" as $\tilde{Y}_i = \mathcal{V}_m(\tilde{g}_i \otimes I_s)$, and the corresponding harmonic residual block vector is as $\dot{R}_i = A\tilde{Y}_i B - \tilde{\theta}_i\tilde{Y}_i$. In the following propositions, two important results for the harmonic residual block vector is mentioned.

Proposition 3.1. *The residual matrix for a harmonic Ritz block vector is F-orthogonal to* $A\mathcal{G}K_m(A, \tilde{R}_0, B)B$.

Proposition 3.2. *The residual harmonic Ritz block vector is a multiple of the residual matrix associated with Gl-GMRES method, i.e* $\dot{R}_j = \gamma_j R_0$, *where* $R_0 = R_m$ *at the time restart.*

We will develop an implicitly restarting Gl-GMRES algorithm that is called Gl-GMRES-IR algorithm. This Algorithm is described in 3.3.

Algorithm 3.3. The implicitly restarted Gl-GMRES (Gl-GMRES-IR).
Input: $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{s\times s}$, $C, X_0 \in \mathbb{R}^{n\times s}$ and $tol > 0$, m and p are integer numbers and $k = m - p$ is the desired number of shifts.

1. Compute $R_0 = C - AX_0B$. Set $V = R_0$.
2. Apply the generalized global Arnoldi in [4] with $m = k + p$ iteration and obtain $\bar{H}_m$ and $\mathcal{V}_{m+1}$.
3. Solve $min\|\mathcal{V}_{m+1}^T \diamond \tilde{R}_0 - \bar{H}_m d\|_2$ for d.
4. Compute $X_m = X_0 + \mathcal{V}_m(d \otimes I_s)$ and $\tilde{R}_m = C - AX_mB$. If $\|\tilde{R}_m\|_F < tol$ stop.
6. Set $X_0 = X_m$, then the implicitly generalized global Arnoldi Algorithm to obtain $\mathcal{V}_{m+1}, \bar{H}_m$, compute the k smallest eigenvalues of the matrix (16).Then go to step 3.3.

4 Numerical results

In this section, the numerical behavior of the Gl-GMRES-IR against the GlFOM and GlGMRES methods [4] is evaluated. Here, m and k show the number of iterations for each restart and the number of the harmonic Ritz block vectors, respectively. The initial guess is taken to be zero and the right hand side matrix of the matrix equation (1) is $C = rand(n, s)$. The condition $\|C - AX_lB\|_F < 10^{-6}\|C\|$ is considered as stopping criterion.

Example 4.1. In this example, the matrix A is selected from matrix market collection [1] and B is a bidiagonal matrix with $0.01, 0.2, 10, 20, \ldots, 10(s-2)$ and $0.1'$s on the main and the super diagonal, respectively. As observed from Table 1, Gl-GMRES-IR is faster than the other methods.

Table 1: Numerical results of the four methods on the cavity01 and the sherman4 matrices with $s = 25, m = 10, 20$ and $k = 6$. Here, iter., res.norm and CPU show the number of iterations, residual norm and run time.

	(cavity01, B)			(sherman4, B)		
Methods	iter	res.norm	CPU	iter	res.norm	CPU
Gl-GMRES-IR(10, 6)	700	9.9963e-07	20.1721	103	8.5244e-07	3.0247
Gl-FOM-DR(10, 6)	781	7.1374e-07	30.2738	192	1.6216e-07	10.6267
GlGMRES(10)	1273	9.9929e-07	54.4906	232	9.9045e-07	4.3000
GlFOM(10)	2001	1.5200e-06	35.5204	301	9.9780e-07	14.9061
Gl-GMRES-IR(20, 6)	126	9.9584e-07	38.9868	11	9.2388e-07	0.1391
Gl-FOM-IR(20, 6)	145	9.9341e-07	47.4272	31	6.6698e-07	9.8022
GlGMRES(20)	297	9.9980e-07	41.5393	6	6.5130e-07	0.2912
GlFOM(20)	848	9.9950e-07	49.8374	16	2.7356e-07	2.5751

References

[1] M. Khorsand Zak and F. Toutounian, Nested splitting conjugate gradient method for matrix equation AXB=C and preconditioning, *Comput. Math. Appl.*, 66 (2013), 269–278.

[2] M. Mohseni Moghadam, A. Rivaz, A. Tajaddini and F. Saberi Movahed, Convergence analysis of the global FOM and GMRES methods for solving matrix equations AXB = C with SPD coefficients, *Bull. Iranian Math. Soc.*, 41 (2015), no. 4, 981–1001.

[3] R. Morgan, Implicitly restarted GMRES and Arnoldi methods for nonsymmetric systems of equations, *SIAM J. Matrix Anal. Appl.* 21 (2000), No. 4, 1112–1135.

[4] F. Panjeh Ali Beik, Note to the global GMRES for solving the matrix equation AXB=F, *World Acad. Sci. Eng. Technol.*, 56 (2011), 08–23.

[5] D.C. Sorensen, Implicit application of polynomial filters in a k-step Arnoldi method, *SIAM J. Matrix Anal. Appl.*, 13 (1992), No. 1, 357–385.

[6] H. X. Zhong, G. Wu and G. L. Chen, A flexible and adaptive simpler block GMRES with deflated restarting for linear systems with multiple right-hand sides, *J. Comput. Appl. Math.*, 282 (2015), 139–156.

[1] https://math.nist.gov/MatrixMarket/

The relationship between n-positivity, linearity and continuity of positive mappings between C^*-algebras[1]

Ali Dadkhah* and Mohammad Sal Moslehian

Department of Pure Mathematics, Center Of Excellence in Analysis on Algebraic Structures (CEAAS), Ferdowsi University of Mashhad, P. O. Box 1159, Mashhad 91775, Iran

Abstract

We investigate some classes of positive mappings (not necessarily linear) in the setting of C^*-algebras. First, we give some results about the superadditivity and the starshapeness of such maps. Then, for a certain class of unital positive maps $\Phi : \mathscr{A} \longrightarrow \mathscr{B}$ between unital C^*-algebras, we present the relation between the n-positivity, the linearity and the continuity of Φ.

Keywords: C^*-algebra, n-positive map, Superadditive, Nonlinear positive map
Mathematics Subject Classification [2010]: 15A60, 47A63

1 Introduction

Let $\mathscr{H}$ and $\mathscr{K}$ be complex Hilbert spaces. Let us denote by $\mathbb{B}(\mathscr{H})$ and $\mathbb{B}(\mathscr{K})$ the algebras of all bounded linear operators on $\mathscr{H}$ and $\mathscr{K}$, respectively. In the case when $\mathscr{H} = \mathbb{C}^n$, we identify $\mathbb{B}(\mathbb{C}^n)$ with the matrix algebra of $n \times n$ complex matrices $M_n(\mathbb{C})$. Here we consider the usual Löwner order $\leq$ on the real space of self-adjoint operators. An operator A is said to be strictly positive (denoted by $A > 0$) if it is a positive invertible operator.Thanks to the Gelfand–Naimark–Segal theorem, we may assume that any C^*-algebra is a closed C^*-subalgebra of $\mathbb{B}(\mathscr{H})$ for some Hilbert space $\mathscr{H}$. We use $\mathscr{A}, \mathscr{B}, \cdots$ to denote C^*-algebras and $\mathscr{A}_+$ and $\mathscr{A}_{++}$ to denote the sets of all positive and positive invertible elements of $\mathscr{A}$, respectively. The geometric mean is defined by $A\sharp B = A^{\frac{1}{2}} \left(A^{-\frac{1}{2}} B A^{-\frac{1}{2}}\right)^{\frac{1}{2}} A^{\frac{1}{2}}$ for operators $A \in \mathscr{A}_{++}$ and $B \in \mathscr{A}_+$. If A commute with B, then $A\sharp B = (AB)^{\frac{1}{2}}$.
A map $\Phi : \mathscr{A} \to \mathscr{B}$ between C^*-algebras is said to be $*$-map or self-adjoint if it is $*$-preserving i.e. $\Phi(A^*) = \Phi(A)^*$ and it is called positive if it holds that $\Phi(\mathscr{A}_+) \subset \mathscr{B}_+$. It is called strictly positive, whenever $\Phi(\mathscr{A}_{++}) \subset \mathscr{B}_{++}$. We say that Φ is unital if $\mathscr{A}, \mathscr{B}$ are unital and Φ preserves the unit. For simplicity of notation, we denote both units of $\mathscr{A}$ and $\mathscr{B}$ by I. A map Φ is called n-positive if the map $\Phi_n : M_n(\mathscr{A}) \to M_n(\mathscr{B})$ defined by $\Phi_n([a_{ij}]) = [\Phi(a_{ij})]$ is positive, where $M_n(\mathscr{A})$ stands for the C^*-algebra of $n \times n$ matrices with entries in $\mathscr{A}$. A map Φ is said to be completely positive if it is n-positive for all $n \in \mathbb{N}$. We say a positive map $\Phi : \mathscr{A} \to \mathscr{B}$ is in the class $S^{(n)}_{\text{mon}+}$, whenever the map Φ_n is monotone on positive elements of $M_n(\mathscr{A})$.

[1] Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University
* Speaker. Email address: dadkhah61@yahoo.com

2 Main results

We start our work by giving some examples of positive maps in the class $S^{(n)}_{\text{mon}+}$.

Example 2.1. Some examples of positive maps in the class $S^{(n)}_{\text{mon}+}$:

- Every $2n$-positive map is in the class $S^{(n)}_{\text{mon}+}$.
- Power functions $\Phi_p : \mathbb{C} \to \mathbb{C}$ defined by $\Phi_p(x) = |x|^p$ $(1 \leq p < 2)$ are in the class $S^{(2)}_{\text{mon}+}$, however, Φ_p $(1 \leq p < 2)$ are only 3-positive but not 4-positive.
- Every positive semidefinite matrix $P \in M_n(\mathbb{C})$ induces a map $\phi_P : M_n(\mathbb{C}) \to \mathbb{C}$ defined by $\phi_P(A) = |\text{tr}(AP)|$, which is a 3-positive semi–norm on $M_n(\mathbb{C})$ belonging to the class $S^{(2)}_{\text{mon}+}$
- Every positive linear functional $\varphi : \mathscr{A} \longrightarrow \mathbb{C}$ on a C^*-algebra induces a non-linear 3-positive map $\Phi : \mathscr{A} \longrightarrow \mathbb{C}$ given by $\Phi(A) = |\varphi(A)|$ that belongs to the class $S^{(2)}_{\text{mon}+}$.

A map $\Phi : \mathscr{X} \subseteq \mathscr{A} \to \mathscr{B}$ is said to be superadditive on a subset $\mathscr{X}$ of $\mathscr{A}$, which is closed under addition, if

$$\Phi(A + B) \geq \Phi(A) + \Phi(B)$$

for every $A, B \in \mathscr{X}$ and it is strongly superadditive, if

$$\Phi(A + B + C) + \Phi(A) \geq \Phi(A + B) + \Phi(A + C)$$

for every $A, B, C \in \mathscr{X}$.
It is knwon that [1] if $\Phi : \mathscr{A} \to \mathscr{B}$ is in the class $S^{(2)}_{\text{mon}+}$, then it is strongly superadditive on $\mathscr{A}_+$.

A map $\Phi : \mathscr{X} \subseteq \mathscr{A} \to \mathscr{B}$ is called starshaped if $\Phi(\alpha A) \leq \alpha\Phi(A)$ for any $A \in \mathscr{X}$ and every $\alpha \in [0, 1]$.
It is known that every starshaped function $f : [0, \infty) \to [0, \infty)$ is superadditive. However, the converse of this statement is not true, in general. The next theorem shows that the strong superadditivity of a continuous positive map Φ (with $\Phi(0) = 0$) on a subset $\mathscr{X}$ (closed under addition) of $\mathscr{A}_+$ implies the starshapeness of Φ on the $\mathscr{X}$.

Theorem 2.2. *Let $\mathscr{A}, \mathscr{B}$ be two C^*-algebras. If $\Phi : \mathscr{A} \to \mathscr{B}$ is a continuous (non-linear) positive map, which is strongly superadditive on any subset $\mathscr{X}$ (closed under addition) of $\mathscr{A}_+$, then the following statements are equivalent:*

(i) $\Phi(0) = 0$,

(ii) $\Phi(\alpha A) \leq \alpha\Phi(A)$ *for every* $\alpha \in (0, 1]$ *and* $A \in \mathscr{X}$,

(iii) $\Phi(\alpha A) \geq \alpha\Phi(A)$ *for every* $\alpha \in [1, \infty)$ *and* $A \in \mathscr{X}$.

In [1], the authors presented the relation between the n-positivity, homogeneity and the linearity of some positive mappings between C^*-algebras. If $\Phi : \mathscr{A} \to \mathscr{B}$ is a continuous unital 3-positive map between unital C^*-algebras, which is in the class $S^{(2)}_{\text{mon}+}$ and $\Phi(0) = 0$, then

1. if $\Phi(\alpha I) = \alpha I$ for some $\alpha \in \mathbb{C}_1 \cup \mathbb{R}_+$, then $\Phi(\alpha A) = \alpha\Phi(A)$ for every $A \in \mathscr{A}$,

2. if $\Phi(\alpha I) = \alpha I$ for some $\alpha \in \mathbb{C}_1 \cup \mathbb{R}_+$ with $|\alpha| \neq 0, 1$, then $\Phi(\beta A + B) = \beta\Phi(A) + \Phi(B)$ for every $\beta \in \mathbb{R}_+$ and $A, B \in \mathscr{A}_+$,

in which $\mathbb{C}_1 = \{z \in \mathbb{C} : |z| \leq 1\}$. Moreover, if either Φ is 6-positive or in the class $S^{(4)}_{\text{mon}+}$, then

1. if $\Phi(\alpha I) = \alpha I$ for some $\alpha \in \mathbb{C}_1 \cup \mathbb{R}_+$ with $|\alpha| \neq 0, 1$, then $\Phi(\beta A + B) = \beta\Phi(A) + \Phi(B)$ for every $\beta \in \mathbb{R}$ and $A, B \in \mathscr{A}$,

2. if $\Phi\big(z\, I\big) = z\, I$ for some $z \in \mathbb{C}$ with $\text{Im}(z) \neq 0$ and $|z| < 1$, then Φ is linear on $\mathscr{A}$.

The following examples show the necessity of some hypotheses in the above facts.

Example 2.3. (1) Consider the map $\varphi : \mathbb{C} \to \mathbb{C}$ defined by $\varphi(z) = \dfrac{1}{3}(|z|^{\frac{3}{2}} + |z|^{\frac{4}{3}} + 1)$. It is known that φ is a 3-positive map and $\varphi \in S^{(2)}_{\text{mon}+}$ (see [4, Theorem 5.1]). Evidently, there exists a number $z \in [1.1, 2]$ such that $\varphi(z) = z$. However, φ is not additive on positive numbers.
(2) Let $(\mathscr{A}, \|\cdot\|)$ be a unital C^*-algebra. According to [1, Corollary 3.6], we see that $\|\cdot\|$ is not a 3-positive map in the most C^*-algebras. For every $\alpha > 0$, we have $\|\alpha I\| = \alpha$ while $\|\cdot\|$ is not additive on positive elements of $\mathscr{A}$, in general.
(3) The map $|\cdot| : \mathbb{C} \to \mathbb{C}$ is a 3-positive map in the class $S^{(2)}_{\text{mon}+}$. However, $|\cdot|$ is not 6-positive, nor is in the class $S^{(4)}_{\text{mon}+}$. For every $\alpha > 0$, we have $|\alpha I| = \alpha$, but $|\cdot|$ is not additive on $\mathbb{C}$, see [1].

We aim to give the relation between n-positivity and the continuity of a positive map between C^*-algebras. It is known that (see [5]) if $\Phi : \mathscr{A} \to \mathscr{B}$ is a 2-positive map between C^*-algebras, then

$$\Phi(A \sharp B) \leq \Phi(A) \sharp \Phi(B)$$

for every $A, B \in \mathscr{A}_{++}$.

Theorem 2.4. *Let $\mathscr{A}$ be a unital C^*-algebra and $\mathscr{B}$ be a C^*-algebra. If $\Phi : \mathscr{A} \to \mathscr{B}$ is a 2-positive map, then*

$$\text{s. o.} - \lim_{\varepsilon \to 0} \Phi(A + \varepsilon I) = \Phi(A)$$

for every $A \in \mathscr{A}_{++}$, where the convergence is in the strong operator topology.

We have the following theorem.

Theorem 2.5. *If $\Phi : \mathscr{A} \to \mathscr{B}$ is a positive map between C^*-algebras in the class $S^{(2)}_{\text{mon}+}$, then $\Phi(\mathscr{A}, \|\cdot\|) \to (\mathscr{B}, \|\cdot\|)$ is continuous.*

3 Conclusion

Some results about positive linear maps between C^*-algebras are still valid for non-linear positive maps if they are n-positive for some $n \in \mathbb{N}$. Moreover, for a certain class of positive maps, in order to be homogeneous and linear, it is sufficient to investigate the homogeneity at only one scalar.

References

[1] A. Dadkhah, M. S. Moslehian, *Non-linear positive maps between $C*$-algebras*, (2018) DOI: 10.1080/03081087.2018.1547357 .

[2] A. Dadkhah, M. S. Moslehian, and K. Yanagi *Noncommutative versions of inequalities in quantum information theory*, Anal. Math. Phys. **9** (2019), 2151–2169.

[3] A. Dadkhah, M. S. Moslehian, *More on non-linear positive maps* , (preprint).

[4] F. Hiai, *Monotonicity for entrywise functions of matrices*, Linear Algebra Appl. **431**, no. 8, (2009), 1125–1146.

[5] M. S. Moslehian, M. Kian, Q. Xu, *Positivity of 2×2 block matrices of operators*, Banach J. Math. Anal. **13** (2019), no. 3, 726–743.

[6] M. Günther and L. Klotz, *Lieb functions and m-positivity of norms*, Linear Algebra Appl. **456** (2014), 54–63.

Supervised feature selection via information gain, maximum projection and minimum redundancy[1]

Mahdi Eftekhari[1], Farid Saberi-Movahed[2,*] and Adel Mehrpooya[1]

[1]Department of Computer Engineering, Shahid Bahonar University of Kerman, Kerman, Iran

[2]Department of Applied Mathematics, Faculty of Sciences and Modern Technologies, Graduate University of Advanced Technology, Kerman, Iran

Abstract

Feature selection problem is an important issue in both data clustering and data classification. This paper introduces a supervised framework for the task of feature selection. The proposed method is built based on applying the information gain method into the framework of maximizing relevancy, and aims to reduce the redundancy between the selected features by using the idea of maximum projection and minimum redundancy. Several experimental results on seven well-known microarray datasets demonstrate the promising performance of the proposed method over some state-of-the-art methods in this area.

Keywords: Machine learning, Supervised feature selection, Information gain, Maximum projection, Minimum redundancy

Mathematics Subject Classification [2010]: 62H30

1 Introduction

Dimensionality reduction, known as a challenging subject in machine learning, has delivered innumerable magnificent achievements over the past decades. It is noticeable that a major category of problems regarding dimensionality reduction has formed based on feature selection methods and related conceptions. In specific, these methods determine the most representative features of the original feature space with respect to a selection criterion [1].

During past years, a variety of techniques have been developed to characterize feature selection problems. As an excellent example, the information gain (IG) method can be mentioned which is a particularly prevalent attribute evaluation method that has found widespread application in the context of feature selection [1, 2]. The IG method works as a univariate filter which ranks all the features in order of importance. In the next step, the archetypal features are selected by the IG method according to a certain threshold. Another notable example is matrix factorization-based approaches that have been widely employed in the frame of feature selection problem. In [4], Wang et al. have propounded a novel unsupervised feature selection method via matrix factorization (MFFS) which was

[1]Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

*Speaker. Email address: fdsaberi@gmail.com

based on the subspace distance and made use of a especial matrix factorization criterion. According to the notion of projection, Wang et al. [5] have introduced another feature selection method via maximum projection and minimum redundancy (MPMR) in such a way as to find non-redundant features from the whole feature set.

Most of the existing research works in the field of feature selection have been constructed by using some fundamental aspects such as the matrix factorization technique and the geometric information of data. In addition to these considerations, another viewpoint would be important to formalize the feature selection problem. This viewpoint is referred to the concept of linear independence. To be more specific, for a given set of features like microarray high-dimensional datasets in which the number of features is larger than the number of samples, it may be possible to express one feature as a linear combination of the other features. As a novel feature selection method, Ebrahimpour et al. [2] have applied the notion of linear independence to derive a supervised feature selection method for determining representative non-redundant subsets of features in microarray datasets.

In this paper, motivated by the idea of MPMR to eliminate redundant features, and taking advantage of the IG method to maximize relevancy, we propose a supervised feature selection method in which the impact of selecting linearly independent subsets of features on the feature selection problem is considered. In order to evaluate how well the theoretical results are effective to perform simultaneously dimensionality reduction and classification tasks, the algorithms are compared with several well-known supervised feature selection methods through microarray high-dimensional datasets.

2 Notations

In the present paper, we indicate scalers by italic lowercase letters such as x; vectors by bold lowercase letters such as $\mathbf{x}$; matrices by bold uppercase letters such as $\mathbf{X}$; and sets by italic uppercase letters such as X. Throughout this paper, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_d] \in \mathbb{R}^{n\times d}$ is the data matrix associated with the original feature set $\{\mathbf{x}_1, \ldots, \mathbf{x}_d\}$, where n is the number of samples, and d is the number of features. For simplicity, the submatrix of $\mathbf{X}$ associated with a feature subset X_I is denoted by the $n \times k$ matrix $\mathbf{X}_I$, where I is the index set of selected features, and $|I|$ is the number of elements in the set I. In addition, $\|\mathbf{X}\|_F$ denotes the well-known Frobenius norm of $\mathbf{X}$, the transpose of $\mathbf{X}$ is denoted by $\mathbf{X}^T$, $\mathrm{rank}(\mathbf{X})$ indicates the rank of $\mathbf{X}$, $\mathbf{X}^\dagger$ denotes the Moore-Penrose pseudo-inverse of $\mathbf{X}$, and $\mathrm{span}(\mathbf{X})$ is the set of all possible linear combinations of the columns of $\mathbf{X}$.

3 Maximum projection and minimum redundancy

In recent times, the concept of maximum projection and minimum redundancy (MPMR) has been introduced by Wang et al. [5] for constructing an unsupervised feature selection algorithm. The procedure of feature selection based on MPMR is described in Algorithm 3.1. It is worthwhile pointing out that the main idea behind of MPMR is that minimizing the redundancy of the selected features is equivalent to keeping them close to orthogonality; for more discussion on the mechanism of MPMR, the reader is referred to [5].

Algorithm 3.1. Feature selection based on MPMR.

Input. Data matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_d] \in \mathbb{R}^{n\times d}$, and the number of selected features k.

1. $X_I = \{\mathbf{x}_{i_1}\}$ and $I = \{i_1\}$ such that $i_1 = \mathrm{argmin}_i \|\mathbf{X} - \mathbf{x}_i(\mathbf{x}_i^T\mathbf{x}_i)^\dagger\mathbf{x}_i^T\mathbf{X}\|_F$;
2. **for** $j = 2, \ldots, k$ **do**

3. $X_I \leftarrow X_I \cup \{\mathbf{x}_{i_j}\}$ and $I \leftarrow I \cup \{i_j\}$ such that $i_j = \text{argmax}_i \|\mathbf{Y}_i\|_F$ in which $\mathbf{Y}_i$ is the ith column of $\mathbf{Y} = \mathbf{X} - \mathbf{X}_I(\mathbf{X}_I^T\mathbf{X}_I)^\dagger\mathbf{X}_I^T\mathbf{X}$.
4. **end for**

Output. An index set of the selected features $I \subseteq \{1, \ldots, d\}$ and $|I| = k$.

In the rest of this section, it can be demonstrated that if Algorithm 3.1 proceeds k steps, then the selected features form a linearly independent subset of original features. Here, it should also be noted that since microarray datasets used in this paper are of full row rank and $n \ll d$, it is assumed from now on that $\text{rank}(\mathbf{X}) = n$.

Theorem 3.2. *Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_d] \in \mathbb{R}^{n\times d}$ be a data matrix such that $\text{rank}(\mathbf{X}) = n$. Assume that Algorithm 3.1 proceeds k steps and $k \leq n$. Then, the set $\{\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_k}\}$ is a linearly independent subset of the original features.*

Proof. Suppose that $\mathbf{x}_{i_1} \neq 0$ has been chosen such that $i_1 = \text{argmin}_i \|\mathbf{X} - \mathbf{x}_i(\mathbf{x}_i^T\mathbf{x}_i)^\dagger\mathbf{x}_i^T\mathbf{X}\|_F$. Let $I = \{i_1\}$ and define $\mathbf{P}_I = \mathbf{X}_I(\mathbf{X}_I^T\mathbf{X}_I)^\dagger\mathbf{X}_I^T$ to be the orthogonal projection of $\mathbf{X}_I$. If we set $\mathbf{Y} = \mathbf{X} - \mathbf{P}_I\mathbf{X}$, then it can be verified that $\mathbf{Y}_{i_1} = 0$. From this fact and the assumption that $\text{rank}(\mathbf{X}) = n$, we can select a vector $\mathbf{x}_{i_2}$ not only $\mathbf{x}_{i_2} \notin \text{span}(\mathbf{x}_{i_1})$, but also $\mathbf{x}_{i_2} = \text{argmax}_i \|\mathbf{Y}_i\|_F$. Therefore, $X_I = \{\mathbf{x}_{i_1}, \mathbf{x}_{i_2}\}$ is a linearly independent subset of X. By continuing in this fashion, we are able to construct a linearly independent subset $X_I = \{\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_n}\}$ of X. □

4 Main results

In this section, the details of the proposed supervised feature selection via information gain, maximum projection and minimum redundancy (SF-IG-MPMR) are described. The proposed SF-IG-MPMR method consists of two steps:

1. **The pre-processing step.** As it can be seen from Algorithm 3.1, in case the value of d is large, then this algorithm may be very expensive. For instance, the first step of Algorithm 3.1 requires to compute the Moore-Penrose pseudoinverse of $\mathbf{x}_i^T\mathbf{x}_i$ for $i = 1, \ldots, d$. In order to get rid of this difficulty, Algorithm 3.1 can be modified in a way that not only does its computational complexity reduce, but also a subset of features is selected with the aim of maximizing the relevancy. To achieve this goal, in the pre-processing step, we propose that the information gain method should be applied so as to assign a specific rank to each feature and to select features according to an ordered ranking of all the features. Afterwards, in lieu of the original features, the top-p ranked features are selected. In fact, the following procedure is performed:

Input. Data matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_d] \in \mathbb{R}^{n\times d}$, and the parameter p.
1. **for** $i = 1, \ldots, d$ **do**
2. IG(i) $\longleftarrow$ Information gain (IG) score for the ith feature.
3. **end for**
4. Arranging the IG scores in descending order.
5. $\mathbf{X}^{\text{new}} \longleftarrow$ Rearranging the features in $\mathbf{X}$ in terms of sorting order. Set $\mathbf{X}^{\text{new}} = [\mathbf{x}_1^{\text{new}}, \mathbf{x}_2^{\text{new}}, \ldots, \mathbf{x}_p^{\text{new}}]$.

Here, it should be highlighted that according to the results given in [3], the value of p in this paper is selected to be equal to $2n$.

2. **The feature selection step.** In this step, a linearly independent subset of features is selected from the top-p ranked features that obtained from the pre-processing step. Note that this selection procedure is guided by Algorithm 3.1. In this way, it is expected that the redundancy among the selected features is significantly on the decline due to the fact that they are linearly independent.

5 Experimental results

In this section, some experiments are performed to study the effectiveness of the proposed SF-IG-MPMR method. This method is also compared with seven feature selection algorithms, and the obtained results of the empirical experiments on a set of seven widely-used binary microarray datasets are presented. The detailed characteristics of the datasets are summarized in Table 1.

Table 1: Statistics of the seven binary microarray datasets applied in the experiments [1].

	Brain	CNS	Colon	DLBCL	GLi85	Ovarian	SMK
# Samples	21	60	62	47	85	253	187
# Features	12625	7129	2000	4026	22283	15154	19993

5.1 Comparison methods

In recent years, some feature selection methods have been suggested in the subject of microarray datasets. Several commonly used instances of such methods applied in this paper are as follows: IG, ReliefF, MRMR, SVM-RFE [1], RREFS [2], SFS-BMF1 and SFS-BMF1 [3]. Furthermore, "No-FS", mentioned in our experiments, refers to the case in which no feature selection algorithm is used.

5.2 Experimental settings

There is a parameter that is required to be set. For the SF-IG-MPMR, SFS-BMF1 and SFS-BMF1 methods, the number of selected features, k, is tuned from $\{5, 10, 15, 20\}$ for all the datasets. For the rest of the feature selection methods IG, ReliefF, MRMR, SVM-RFE and RREFS, the values of the parameter k are considered as suggested in the original references [1, 2]. In addition to this, three well-known classifiers C4.5, Naïve-Bayes and SVM are adopted to assess the classification performance of the feature selection algorithms. Moreover, the 5-fold DOB-SCV technique is utilized in order to improve the performance of the algorithms. For all the feature selection methods considered in the experiments, the best results of the optimal parameters in terms of the classification performance are reported. Moreover, the experiments are run 10 times and averaged.

5.3 Measures for evaluation

In order to have a fair criterion to assess the performance of the feature selection algorithms, four well-known measures are occasionally applied: sensitivity (Se), specificity (Sp), G-mean and accuracy (Ac) criteria. These criteria are defined as:

$$\text{Se} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{Sp} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \quad \text{G-mean} = \sqrt{\text{Se} \times \text{Sp}}, \quad \text{Ac} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}},$$

where $\text{TP}, \text{TN}, \text{FN}$ and FP indicate true positives, true negatives, false negatives and false positives, respectively.

5.4 Results and analysis

This subsection discusses a number of empirical results associated with the performance of the SF-IG-MPMR algorithm and presents comparisons of these results with those of

the methods indicated in Subsection 5.1. It is notable that the number 50 assigned to IG, ReliefF, and SVM-RFE, and the number 10 assigned to MRMR refer respectively to the top 50 and 10 features in relation to the ranking techniques employed by those methods.

It is apparent that feature selection strategies specifically aim at deciding which features are the most representative of the whole feature set in terms of their information content. In this regard, the number of selected features do count for feature selection methods. The results of this measure for various methods are reported in Table 2. Considering the data in Table 2, it is crystal clear that the SF-IG-MPMR method has the ability to omit more than 99% of all the features. Furthermore, the number of features selected by SF-IG-MPMR is almost by far the smallest compared to that number for the other methods.

Table 2: Number of the selected features.

Methods	**Brain**	**CNS**	**Colon**	**DLBCL**	**Gli85**	**Ovarian**	**SMK**
IG-50	50	50	50	50	50	50	50
ReliefF-50	50	50	50	50	50	50	50
SVM-RFE-50	50	50	50	50	50	50	50
MRMR-10	10	10	10	10	10	10	10
RREFS	20	48	50	38	68	202	150
SFS-BMF1	5	20	5	20	20	20	20
SFS-BMF2	5	20	5	20	20	20	20
SF-IG-MPMR (Ours)	5	5	5	5	5	5	5

In order to illustrate the effectiveness of the proposed method, the average performance of the three classifiers for different feature selection methods is illustrated in terms of the evaluation measures Ac, Se, Sp and G-mean in Figures 1 and 2. It is worthwhile to note that the higher the Ac, Se, Sp and G-mean values are, the better the classification results will be.

In Figure 1, the x-axis indicates the measures Ac, Se, Sp and G-mean, and the y-axis represents the obtained values of these measures for the different feature selection methods. In Figure 2, the x-axis indicates the different feature selection methods, and the y-axis represents the values of the measures Ac, Se, Sp and G-mean which are indicated by the blue, purple, green and red colors, respectively.

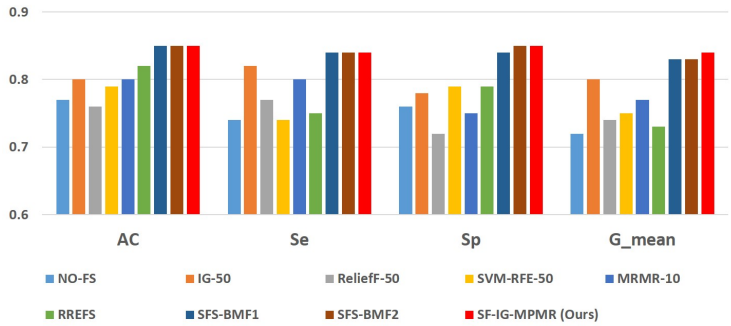


Figure 1: Bar chart of the average values of the measures Ac, Se, Sp and G-mean for the three classifiers C4.5, Naïve-Bayes and SVM on microarray datasets after performing the DOB-SCV method with 5 folds. Note that the higher the Ac, Se, Sp and G-mean values are, the better the classification performance will be.

Figures 1 and 2 demonstrate that except for the SFS-BMF1 and SFS-BMF1 methods, the SF-IG-MPMR method produces the better results in comparison with the other meth-

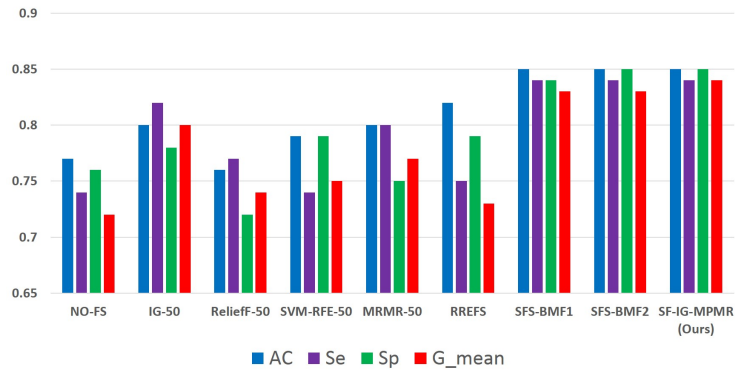


Figure 2: Bar chart of the average values of the measures Ac, Se, Sp and G-mean for the three classifiers C4.5, Naïve-Bayes and SVM on microarray datasets after performing the DOB-SCV method with 5 folds. Note that the higher the Ac, Se, Sp and G-mean values are, the better the classification performance will be.

ods in terms of the aforementioned measures. Moreover, the performance of SF-IG-MPMR is almost the same as that of SFS-BMF1 and SFS-BMF1.

6 Conclusion

In this paper, a novel supervised framework for feature selection has been proposed. This approach is grounded on maximizing relevancy and reducing redundancy among selected features according to the use of information gain method. Experimental results on seven well-known microarray datasets show the superiority of the suggested method over some baseline methods.

References

[1] V. Bolón-Canedo, N. Sánchez-Marono, A. Alonso-Betanzos, J.M. Benítez and F. Herrera, A review of microarray datasets and applied feature selection methods, *Inf. Sci.*, 282 (2014), 111–135.

[2] M.K. Ebrahimpour, M. Zare, M. Eftekhari and G. Aghamolaei, Occam's razor in dimension reduction: Using reduced row Echelon form for finding linear independent features in high dimensional microarray datasets, *Eng. Appl. Artif. Intell.*, 62 (2017), 214–221.

[3] F. Saberi-Movahed, M. Eftekhari and M. Mohtashami, Supervised feature selection by constituting a basis for the original space of features and matrix factorization, *Int. J. Mach. Learn. Cyber.*, 11 (2020), 1405–1421.

[4] S. Wang, W. Pedrycz, Q. Zhu and W. Zhu, Subspace learning for unsupervised feature selection via matrix factorization, *Pattern Recognit.*, 48 (2015), 10–19.

[5] S. Wang, W. Pedrycz, Q. Zhu and W. Zhu, Unsupervised feature selection via maximum projection and minimum redundancy, *Knowl.-Based Syst.*, 75 (2015), 19–29.

Using finite element method for solving weakly singular Volterra integral equations[1]

Majid Erfanian[1,*] and Hamed Zeidabadi[2]

[1]Department of Science, School of Mathematical Sciences, University of Zabol, Zabol, Iran

[2]Faculty of Engineering, Sabzevar University of New Technology, Sabzevar, Iran

Abstract

In this work, we applied a new method for solving linear weakly singular Volterra integral equations. We begin the theoretical study with the acquires of the variational form, and we also use the finite element method to approximate our problems. We estimate the error of the method by proving some theorems. Moreover, in the final section, we present some numerical examples.

Keywords: Singular integro-differential equation, Finite element, Error estimation
Mathematics Subject Classification [2010]: 45D05, 65L60

1 Introduction

In subject finite element methods for differential equations and integral equations, many authors have to work for example in [5]. In this paper we used of adaptive finite element method (FEM) and Lagrange polynomials to obtain an approximate solution for linear weakly singular Volterra integral equation as follow:

$$u(x) = f(x) + \int_a^x \frac{W(x,t)\,u(t)}{(x-t)^\alpha}, \qquad 0 < \alpha < 1, \tag{1}$$

that $W(x,t)$ and $f(x)$ are known continuous functions, and $u(x)$ is a unknown function.

2 Finite element method

First, we obtain weak and variational form of the equation (1), for this purpose we show bilinear form with $B : \mathbb{V} \times \mathbb{V} \to \mathbb{R}$ and $L : \mathbb{V} \to \mathbb{R}$ is a linear functional, and $\mathbb{V} = H^0(\Omega) = L_2(\Omega)$, that $\Omega = [a,b] \subset \mathbb{R}$ is an infinite dimensional space, and for all arbitrary function $v(x) \in \mathbb{V}$ we have where

$$\begin{aligned} B(u,v) =& L(v), \quad \text{and} \quad L(v) = \int_\Omega f(x)v(x)dx, \\ B(u,v) =& \int_\Omega u(x)v(x)dx - \int_\Omega v(x)\Big(\int_a^x \frac{W(x,t)\,u(t)}{(x-t)^\alpha}dt\Big)dx, \end{aligned} \tag{2}$$

[1]Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

*Speaker. Email address: erfaniyan@uoz.ac.ir

thus

$$B(\gamma u+\beta w,v)=\int_{\Omega}(\gamma u+\beta w)v(x)dx-\int_{\Omega}v(x)\int_{a}^{x}\frac{W(x,t)(\gamma u(t)+\beta w(t))}{(x-t)^{\alpha}}dtdx$$
$$=\gamma B(u,v)+\beta B(w,v),$$

then, B is a bilinear form. Consider $\{\phi_i\}_{i=1}^{n}$, is a set of basis continuous piecewise Lagrange polynomial functions of degree at most m, and $\mathbb{V}_h=span\{\phi_1,\phi_2,...,\phi_n\}$, and

$$\phi_i(x_j)=\delta_{ij},\qquad i,j=1,2,...,N.$$

For $u_h(x)$ and $v_h(x)$ we have

$$u_h(x)=\sum_{i=1}^{n}a_i\phi_i(x),\qquad v_h(x)=\sum_{j=1}^{n}b_j\phi_j(x),\tag{3}$$

hence, by substituting (3) in variational formulation we have

$$\sum_{j=1}^{n}b_j\{\sum_{i=1}^{n}a_i\{\int_{\Omega}\phi_i(x)\phi_j(x)dx-\int_{\Omega}\phi_j(x)(\int_{0}^{x}\frac{K(x,t)}{(x-t)^{\alpha}}\phi_i(t)\,dt)\,dx\}$$
$$-\int_{\Omega}f(x)\phi_j(x)dx\}=0.\tag{4}$$

Since, b_j, $j=1,2,\dots n$, are arbitrary, we have

$$\sum_{i=1}^{n}a_i\{\int_{\Omega}\phi_i(x)\phi_j(x)dx-\int_{\Omega}\phi_j(x)(\int_{0}^{x}\frac{W(x,t)}{(x-t)^{\alpha}}\phi_i(t)\,dt)\,dx\}$$
$$-\int_{\Omega}f(x)\phi_j(x)dx\}=0.\tag{5}$$

Now, we define

$$C_{i,j}=\int_{\Omega}\phi_i(x)\phi_j(x)dx-\int_{\Omega}\phi_j(x)\int_{0}^{x}\frac{W(x,t)}{(x-t)^{\alpha}}\phi_i(t)\,dt)\,dx,\quad i,j=1,2,...,n,\tag{6}$$

and

$$F_j=\int_{\Omega}f(x)\phi_j(x)dx,\qquad j=1,2,...,n,\tag{7}$$

thus

$$\sum_{i=1}^{n}C_{ij}a_i=F_j,\qquad j=1,2,...,n,\tag{8}$$

then from system (8) for $\mathbf{F}=[F_1,F_2,...,F_n]^T$, we have

$$C^T\mathbf{A}=\mathbf{F},\tag{9}$$

that,

$$\mathbf{A}=[a_1,a_2,...,a_n]^T,\quad C=[C_{ij}],\quad for\quad i,j=1,2,...,n.$$

By solving of the system (9), we can obtaine approximate solution of equation (1).

3 Error Analysis

In this section, by using the theorem, we get an upper bound for the error of our method, and we proved the order of convergence is a $O(h^{\zeta})$. For this purpose, suppose that $\mathbb{V}$ and B are a Hilbert space and symmetric respectively.

Definition 3.1. If B is a $\mathbb{V}$-elliptic bilinear form, then an inner product energy is a $(.,.) : \mathbb{V} \times \mathbb{V} \to \mathbb{R}$ and the energy norm as

$$||u||_E^2 = (u,u)_B = B(u,v),$$

Definition 3.2. For operator $\Pi : \mathbb{V} \to \mathbb{V}_h$, projection operators as $\Pi u = \tilde{u}_h = \sum_{i=1}^{n} \tilde{a}_i \phi_i(x)$.

Theorem 3.3. *Let $\alpha > 0$, then bilinear form B, defined by (2) is a $\mathbb{V}$-ellipticity and equation (1) has a unique solution, and order of convergence is a $O(h^{\zeta})$.*

Proof. From equation (2) we have

$$|B(u,v)| = |\int_{\Omega} u(x)v(x)dx - \int_{\Omega} v(x) \int_a^x \frac{W(x,t)u(t)}{(x-t)^{\alpha}} dtdx|,$$

with using of the Cauchy-Schwarz inequality and L_2-norm, we have

$$\begin{aligned}
|B(u,v)| &\leq ||u||_{L_2(\Omega)}||v||_{L_2(\Omega)} + W| \int_a^b v(x) \int_a^x \frac{u(t)}{(x-t)^{\alpha}} dtdx| \\
&= ||u||_{L_2(\Omega)}||v||_{L_2(\Omega)} + W| \int_a^b v(x)u(\eta_x) \int_a^b \frac{1}{(x-t)^{\alpha}} dtdx| \\
&\leq ||u||_{L_2(\Omega)}||v||_{L_2(\Omega)} + W| \int_a^b v(x)u(\eta_x) \frac{1}{1-\alpha}(x-t)^{1-\alpha} \, |_{t=a}^{t=x} \, dx| \\
&\leq ||u||_{L_2(\Omega)}||v||_{L_2(\Omega)} + \frac{W(b-a)^{1-\alpha}}{1-\alpha} | \int_a^b v(x)u(\eta_x)dx| \\
&\leq (1 + \frac{W(b-a)^{1-\alpha}}{1-\alpha})||u||_{L^2(\Omega)}||v||_{L^2(\Omega)},
\end{aligned}$$

where

$$W = \max |W(x,t)|, \quad x \in [a,b], and \;\; t \in [a,x], \tag{10}$$

then B is a continuous. Furthermore, we proved V-ellipticity of B, for this purpose we have

$$\begin{aligned}
B(v,v) &= \int_{\Omega} v(x)v(x)dx - \int_{\Omega} v(x) \int_a^x \frac{W(x,t)v(t)}{(x-t)^{\alpha}} dtdx \\
&\geq ||v||_{L_2(\Omega)}^2 - W(\frac{(b-a)^{1-\alpha}}{1-\alpha})||v||_{L_2}^2 = (\eta)||v||_{L_2(\Omega)}^2,
\end{aligned} \tag{11}$$

then $B(v,v) \geq (\eta)||v||_{L_2(\Omega)}^2$, or $B(v,v) \geq \alpha||v||_{L_2(\Omega)}^2$, where

$$\eta = 1 - W(\frac{(b-a)^{1-\alpha}}{1-\alpha}),$$

if $\eta > 0$, thus B is a $\mathbb{V}$-ellipticity, therefore, by using of Lax-Milgram theorem we proved the equation (1) has a unique solution. Suppose u_h is an approximate solution, so we have

$$B(u,v_h) = l(v_h), \qquad B(u_h,v_h) = l(v_h), \quad \forall v_h \in \mathbb{V}_h. \tag{12}$$

If $e = u - u_h$, that u are an exact solution of equation(1), then

$$B(e, v_h) = 0, \quad \forall v_h \in \mathbb{V}_h. \tag{13}$$

By Schwartz's inequality, and relation between energy norm and inner product we have

$$|B(v, w)| \leq ||v||_E ||w||_E, \qquad \forall v, w \in \mathbb{V}, \tag{14}$$

by using of (13) we have $(e, v_h)_B = B(e, v_h) = 0$. Therefore, e is an orthogonal for any v_h. By using of Cea's Lemma [**?**], and for each particular $\tilde{v}_h$ in V_h, and

$$||u - u_h||_E = min\{||u - v_h||_E;\ v_h \in \mathbb{V}_h\},$$

we have $\inf ||u - v_h||_{\mathbb{V}} \leq ||u - \tilde{v}_h||_{\mathbb{V}}$, if $\tilde{v}_h$ is equal to $\tilde{u}_h$, then $||u - u_h||_V \leq c||u - \tilde{u}_h||_{\mathbb{V}}$, if we get an upper bounded for the interpolation error, we have

$$||u - \tilde{u}_h||_V \leq cMh^{\beta}, \quad \beta > 0,$$

and c is independent of h, therefore

$$||u - u_h||_{\mathbb{V}} \leq \frac{CM}{\eta} h^{\beta}.$$

Thus $h \to 0$, and the order of method is a $O(h^{\beta})$. □

4 Numerical Examples

Example 4.1. Consider the linear weakly singular Volterra integral equation:

$$u(x) - \int_0^x \frac{1}{\sqrt{x-t}} u(t)dt = x - \frac{4}{3}x^{\frac{3}{2}}, \quad 0 < x \leq 1,$$

with the exact solution $u(x) = x$.

By using Lagrange polynomials of degree 2, and $M = 10$, the results obtained are presented in Table 1 and Figure 1. The global error of $2.37E-10$ is reported by authors (See [2] for more details).

x	$Exact$	$Numerical$	$RBFmethod$
0.1	0.1000000	0.1000000	0.0999932
0.2	0.2000000	0.2000000	0.1999937
0.3	0.3000000	0.3000000	0.2999908
0.4	0.4000000	0.4000000	0.3999884
0.5	0.5000000	0.5000002	0.4999836
0.6	0.6000000	0.6000005	0.5997819
0.7	0.7000000	0.7000006	0.6999698
0.8	0.8000000	0.8000017	0.7999591
0.9	0.9000000	0.9000040	0.8999434
1	1.0000000	1.0000091	0.9999249

Table 1: Numerical results for Example 4.1.

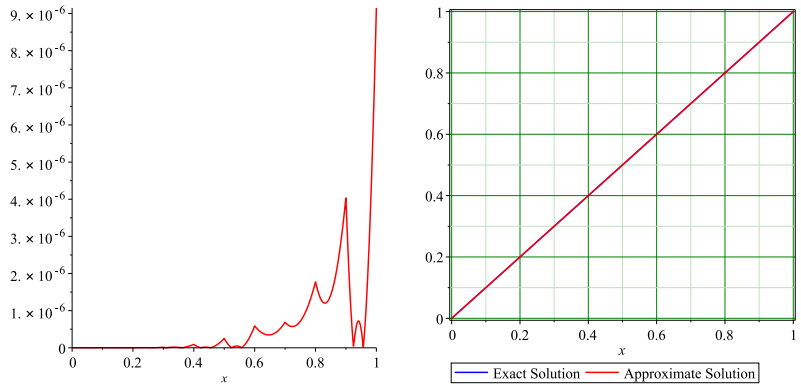


Figure 1: Diagrams of exact and numerical solutions and graph of error for Example 4.1.

Example 4.2. In this example, we consider

$$u(x) = f(x) + \int_0^x \frac{-u(t)}{4\sqrt{x-t}}dt, \quad 0 \le x \le 1,$$

corresponding to the following data

$$f(x) = (1+x)^{\frac{-1}{2}} + \frac{\pi}{8} - \frac{1}{4}\arcsin(\frac{1-x}{1+x}),$$

with exact solution $u(x) = \frac{1}{\sqrt{1+x}}$.

By using Lagrange polynomials of degree 2, and $M = 10$, the results obtained are presented in Table 2 and Figure 2. For comparison, we note that the error of approximation using product integration method with step size 0.05 is about $1.0E-7$ (See [4] for more details).

x	$Exact$	$Numerical$	$RBFmethod$
0.1	0.9534625	0.9534606	0.9535137
0.2	0.9128709	0.9128768	0.9128688
0.3	0.8770580	0.8770625	0.8770231
0.4	0.8451542	0.8451618	0.8451318
0.5	0.8164965	0.8165025	0.8165060
0.6	0.7905694	0.7905842	0.7905978
0.7	0.7669649	0.7669637	0.7669830
0.8	0.7453559	0.7453482	0.7453433
0.9	0.7254762	0.7254816	0.7254464
1	0.7071067	0.7071089	0.7071263

Table 2: Numerical results for Example 4.2.

5 Conclusions

In this paper, we used of adaptive finite element method and Lagrange polynomials to solving one of the most important linear weakly singular Volterra integral equations that

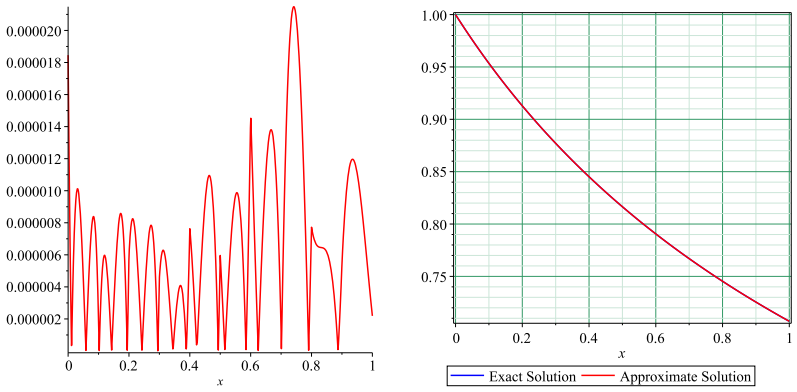


Figure 2: Diagrams of exact and numerical solutions and graph of error for Example 4.2.

very important in the concrete problem of mechanics or physics. First, we obtain a weak and variational form of the equation (1), and with using the system (9), we can obtain an approximate solution. In section Error analysis we proved B is a $\mathbb{V}$-ellipticity and equation (1) has a unique solution, and order of convergence is a $O(h^{\zeta})$. In section Numerical Examples, we have solved three problems considered from [2–4] the results obtained are presented in Table 1, 2 and Figure 1, 2, the comparison of results confirms the better accuracy with this method.

References

[1] K. E. Atkinson. *A Survey of Numerical Methods for the Solution of Fredholm Integral Equations of the Second Kind*, Society for Industrial and Applied Mathematics, Philadelphia, PA,1976.

[2] A. Galperin and E.J. Kansa, *Application of global optimazation and radial basis functions to numerical solution of weakly singular Volterra equations*, J. Comp. Appl. Math. (43)(2002), 491–499.

[3] L. Huang , Y. Huang, Xi. Li, *Approximate solution of Abel integral equation*, J. Comp. Appl. Math. (56)(2008), 1748–1757.

[4] P. Linz, *Analytical and numerical methods for Volterra equations*, SIAM, Philadelphia, 1985.

[5] M. Erfanian, H. Zeidabadi, *Using of Bernstein spectral Galerkin method for solving of weakly singular VolterraFredholm integral equations*, Mathematical Sciences, Vol(12), 103109, 2018.

Matrix flows maintaining positivity[1]

Kazem Ghanbari*

Department of Mathematics, Sahand University of Technology, Tabriz, Iran

Abstract

In this paper we present two types of positivity for matrices and the latest development in matrix differential equations maintaining the positivity of the initial condition. Two important types of positivity are *Comlpete Positivity* and *Total Positivity*. Matrix A is completely positive if it can be decomposed as $A = BB^T$, where B is a nonnegative matrix. A matrix is called totally positive if all minors of the matrix are positive. Two square matrices are said to be *Isospectral* if they have the same eigenvalues. In this paper we introduce a matrix differential equation that preserves the positivity property of the initial matrix.

Keywords: Isospectral flow, Completely positive matrix, Totally positive matrix
Mathematics Subject Classification [2010]: 58J53, 15A18, 15B35, 15A24

1 Introduction

An isospectral matrix flow is charachterized by the following matrix differential equation

$$\frac{dA(t)}{dt} = [F(A), A], \quad A(0) = A_0, \quad t \geq 0, \tag{1}$$

where $F(A)$ is a matrix function and $[X, Y] = XY - YX$ is known as Lie braket. It can be shown that the solution of this differental equations is

$$A(t) = U(t)A_0U(t)^{-1}. \tag{2}$$

Therefore the solution $A(t)$ and the initial matrix A_0 are isospectral. Since the eigenvalues of $A(t)$ remain invariant, this implies such flows are interesting in the context of numerical linear algebra. In fact, the suitably constructed isospectral flows give a continuous realization process for a discrete algorithm. It is proved that Toda lattice at each integer value of t gives the iterates of the QR algorithm [5].

Isospectral flows are also a useful tool in studying inverse eigenvalue problems, for example see Chu [3]: seeking a matrix of a given structure that possesses a specified set of eigenvalues. Such problems are important in a wide range of applications, ranging from the theory of vibrations to control theory, tomography, system identification, geophysics, and particle physics [3]. It is natural to ask for what matrix function $F(A)$ this flow also maintains the positivity properties of the initail matrix A_0?

[1] Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

*Speaker. Email address: kghanbari@sut.ac.ir

2 Main results

In this section we state some latest developments in isospectral matrix flows for an specific type of matrix function $F(A)$. For general type of $F(A)$ the problem is still open [3].

Theorem 2.1. *Suppose $F(A) = A^{+^T} - A^+$, where A^+ is upper triangular part of A. If A_0 is symmetric and totally positive, then $A(t)$ the solutions of the matrix flow (1) is symmetric and totally positive. Morover $\sigma(A(t)) = \sigma(A_0)$, [7].*

We developed this result to nonsymmetric matrices as follows:

Theorem 2.2. *Suppose $F(A) = A_u - A_l$, where A_u, A_l are upper and lower triangular part of A, respectivly. If A_0 is totally positive, not necessarily symmetric, then $A(t)$ the solutions of the matrix flow (1) is totally positive. Morover $\sigma(A(t)) = \sigma(A_0)$, [4].*

We also developed Theorem 2.1 for the set of completely positive matrices as follows:

Theorem 2.3. *$F(A) = A^{+^T} - A^+$, where A^+ is upper triangular part of A. If A_0 is symmetric and completely positive, then $A(t)$ the solutions of the matrix flow (1) is completely positive. Morover $\sigma(A(t)) = \sigma(A_0)$.*

3 Preliminaries and Definitions

In this section we introduce some preliminary materials used in the proof of the main results.

Definition 3.1. Denote the set of eigenvalues of a matrix A by $\sigma(A)$. If A and B are two square matrices with the same size such that $\sigma(A) = \sigma(B)$ then A and B are called isospectral.

For example if A is a given matrix then for eny nonsingular matrix P the matrices PAP^{-1} and A are isospectral. Note that the similarty transformation may change the positivity propert of A.

Definition 3.2. The determinat of a square submatrix of a given matrix A with row $\alpha = \{\alpha_1, \alpha_2, \cdots, \alpha_p\}$ and column $\beta = \{\beta_1, \beta_2, \cdots, \beta_p\}$ is denoted by $A(\alpha; \beta)$. The matrix A is called totally positive (TP) if all minors of A are positive. If all minors of A are nonnegative then A is called totally nonnegative (TN).

Totall positivity arise in the study of in-line systems, rods, beams, Sturm-Liouville differential equations, etc [5]. If we pick up a matrix randomly, it will NOT be TN or TP, most probably. But there are some construction algorithms that we can construct TP matrices using prescribed data (Inverse Eigenvalue Problems).

Definition 3.3. Matrix A is completely positive (CP) if it can be decomposed as $A = BB^T$, where B is a nonnegative matrix, i.e. all entries of B are nonnegative.

Definition 3.4. A semidefinite entrywise nonnegative matrix is called *doubly nonnegatve*. It is clear that every completely positive matrix is doubly nonnegatve by definition.

Completely positive matrices arise in the study of block designs in combinatorics, in probability, and in various applications of statistics, including a Markovian model for DNA evolution [2].

Definition 3.5. If A is an $n \times n$ matrix then comparision matrix of A is denoted by $M(A)$ and is defined as follows:

$$M(A) = \begin{cases} |a_{ij}|, & if \quad i = j \\ -|a_{ij}|, & if \quad i \neq j \end{cases},$$

It is clear that every TP matrix is CP therefore we can construct CP matrices from spectral data. But the reverse is not true. There are some criterion to detect TP and CP matrices. For detailed and comprehensive knowledge on this topics refer to nice books [5] for total positivity, and [2] for complete positivity. The number of minors to be checked for totall positivity is too much. Ando [1] found a criterion that needs to check much smaller set of minors for totall positivity. Let $Q_{p,n}$ denote the set of strictly increasing sequence $\alpha = \{\alpha_1, \alpha_2, \cdots, \alpha_p\}$ chosen from $1, 2, 3, \cdots, n$. Define

$$d(\alpha) = \sum_{i=1}^{n-1} (\alpha_{i+1} - \alpha_i - 1),$$

and note that if $\alpha \in Q_{p,n}$ then $d(\alpha) = 0$ iff $\alpha_{i+1} = \alpha_i + 1$, for $i = 1, 2, \cdots, p-1$; i.e., $d(\alpha) = 0$ iff $alpha_1, \alpha_2, \cdots, \alpha_p$ consists of consecutive integers. We define $Q^0_{k,n}$ as the subset of $Q_{k,n}$ consisting of those α with $d(\alpha) = 0$. We state the following theorem of Ando [1387].

Theorem 3.6. *$A \in M_n$ is TP if $A(\alpha; \beta) > 0$ for all $\alpha, \beta \in Q^0_{k,n}$, $k = 1, 2, \cdots, n$.*

Theorem 3.7. *If A is a symmetric and nonnegative and if its comparison matrix $M(A)$ is positive semidefinite, then A is completely positive*

It can be easily verified that a matrix isospectral flow may or may not preserve positivity property of the initial matrix A_0. We are interested in studying isospectral flows that maintain poitivity properties of the initial matrix A_0.

Example 3.8. Consider the following isospectral flow where N is a given constant matrix

$$\frac{dA(t)}{dt} = [A, N] = AN - NA, \quad A(0) = A_0, \quad t \geq 0.$$

It is clear that the solution is $A(t) = e^{-tN} A_0 e^{tN}$, so the flow is isospectral. If A_0 is positive semidefinite then so is $A(t)$. But for the case of A_0 to be TP or CP then $A(t)$ is not TP or CP in general.

Example 3.9. [**Toda Flow**] Let A_0 be a given completely positive having off-diagonal entries positive that appear in discretization of Sturm-Liouville differential equation and mass-spring vibrating system. By definition of completely positive matrix it is clear that A_0 is a doubly nonnegatve matrix. Consider the isospectral flow of the form (1) where A is a tridiagonal matrix of the following form

$$A = \begin{pmatrix} a_1 & b_1 & 0 & . & . & . \\ b_1 & a_2 & b_2 & 0 & . & . \\ 0 & b_2 & a_3 & b_3 & 0 & . \\ . & . & . & . & . & . \\ . & . & . & . & . & b_{n-1} \\ . & . & . & . & b_{n-1} & a_n \end{pmatrix}. \tag{3}$$

The corresponding differential equation will be as follows:

$$\begin{cases} \dot{a}_k = 2(b_{k-1}^2 - b_k^2 \\ \dot{b}_k = (a_{k+1} - a_k)b_k, \quad k = 1, 2, \cdots, n \end{cases}$$

where b_0 and b_n supposed to be zero. Let $\sigma(A_0) = \{\lambda_i\}_{i=1}^n$, thus all eigenvalues are positive, therefore $A(t)$ will be positive semidefinite since the flow is isospectral. Thus $a_k > 0$ for $k = 1, 2, \cdots, n$. Solving the second differential equation in the system above we find

$$b_k(t) = b_k(0)e^{(a_{k+1}(t) - a_k(t))}.$$

Computing the comparison matrix $M(A)$ shows that $M(A)$ is positive semeidefinite. Using Theorem 3.7 shows that $A(t)$ is completely positive. It can be checked that if A_0 is TP then $A(t)$ also will be TP.

4 Conclusion

In this paper we introduced the concept of isospectral matrix flows. We presented an isospectral flow that preserve the positivity property of the intial matix A_0.

References

[1] T. Ando, Totally positive matrices, *Linear Algebra and its Applications,* vol. 90 (1987) 165219.

[2] A. Berman, *Completely Positive Matrices*, World Scientific 2003

[3] M. T. Chu, Inverse eigenvalue problems, *SIAM review,* vol. 40 (1998) no. 1, 139.

[4] M. Moghaddam, K. Ghnbari and A. Mingarelli, Isospectral matrix flow maintaining staircase structure and total positivity of an initial matrix*Linear Algebra and its Applications,* Vol. 517, (2017) 134-147.

[5] G.M.L. Gladwell, *Inverse problems in vibration,* Kluwer Academic Publishers, 2004.

[6] G.M.L. Gladwell, Total positivity and the QR algorithm, *Linear Algebra and its Applications,* vol. 271 (1998) 257272.

[7] G.. M. L. Gladwell, Total positivity and Toda flow, *Linear Algebra and its Applications,* Vol. 350 (2002) 279-284.

Conditional square matrices of order 2 with given determinant[1]

Mehdi Hassani*

Department of Mathematics, University of Zanjan, University Blvd., Zanjan 45371-38791, Iran

Abstract

In search of a method to generate matrices of order 2 with large positive integer elements and having small determinant, we prove that for given positive integers d and M there exist many infinitely matrices $A = [a_{ij}]_{1\leqslant i,j\leqslant 2}$ with integer elements satisfying $a_{ij} \geqslant M$ and $\det A = d$. Our proof, which is based on the theory of linear Diophantine equations with two variables, has capacity to be followed numerically. Hence, we present several practical examples of these conditional square matrices running over a Maple code.

Keywords: Determinants, Linear Diophantine equation

Mathematics Subject Classification [2010]: 15A15, 15A12, 11Y50

1 Introduction

The question of the present paper arose to my mind when I was searching square matrices with large positive integer elements and having small positive integer determinant. A typical example is

$$\det \begin{bmatrix} 10888869450418352160891456789 & 40329146112660562323832109 0 \\ 20390342059236161031596777818 & 75519785404578371878408668 9 \end{bmatrix} = 1. \quad (1)$$

Focusing on the case that matrices under study are of order 2, we prove the following result.

Theorem 1.1. *Given positive integers d and M there exists many infinitely matrices $A = [a_{ij}]_{1\leqslant i,j\leqslant 2}$ with integer elements satisfying $a_{ij} \geqslant M$ and $\det A = d$.*

We give the proof of Theorem 1.1 in the next section. Our proof based on the theory of linear Diophantine equations with two variables, which is known in number theory literatures (for example see Theorem 2.9 of [2]). The proof has capacity to be followed numerically, hence we get a method to generate matrices with large positive integer elements and having small determinant. We will provide several methods of generation in Section 3, running over a Maple code.

[1] Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

*Speaker. Email address: mehdi.hassani@znu.ac.ir

2 Proof of Theorem 1.1

Let

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}.$$

We take $a_{11} = M$ and $a_{12} = M + 1$, for which we observe that $\gcd(M, M + 1) = 1$ for all integers M. Also, let $a_{21} = y$ and $a_{22} = x$. Thus, the equality $\det A = d$ reads as the following Diophantine equation

$$Mx - (M + 1)y = d,$$

which has many infinitely solutions $x = x_0 + (M + 1)t$ and $y = y_0 + Mt$ with $t \in \mathbb{Z}$, and (x_0, y_0) is any particular solution of the equation. We observe that $M(-d)-(M+1)(-d) = d$. Hence we may take $x_0 = y_0 = -d$. Thus, we obtain

$$x = (M + 1)t - d, \qquad y = Mt - d, \qquad (t \in \mathbb{Z}).$$

To realize the condition $a_{ij} \geqslant M$, we take $t \geqslant \frac{M+d}{M}$ to ensure that $x, y \geqslant M$. Hence, for each integer $t \geqslant 2 + \lfloor \frac{d}{M} \rfloor$ we have

$$\det \begin{bmatrix} M & M + 1 \\ Mt - d & (M + 1)t - d \end{bmatrix} = d,$$

providing many infinitely matrices $A = [a_{ij}]_{1 \leqslant i,j \leqslant 2}$ with integer elements satisfying $a_{ij} \geqslant M$ and $\det A = d$. Note that if $d < M$ we can also take $t = 1$, still keeping positivity of elements.

Remark 2.1. In the above proof, the generator elements $a_{11} = M$ and $a_{12} = M + 1$ satisfy $\gcd(a_{11}, a_{12}) = 1$, to make sure that $\gcd(M, M + 1)|d$, and hence existing many infinite solutions for the related Diophantine equation $a_{11}x - a_{12}y = d$. We may choose generator elements a_{11} and a_{12} in other ways, for example as follows

$$(a_{11} = 2M + 1, a_{12} = 9M + 4),$$
$$(a_{11} = 5M + 2, a_{12} = 7M + 3),$$
$$(a_{11} = 2M + 3, a_{12} = 4M + 5),$$

all satisfying the desired condition $\gcd(a_{11}, a_{12}) = 1$. Another option to choose generator elements a_{11} and a_{12} is to take two distinct large primes. Also, shifted factorials, $n! \pm m$ for practical values of $m, n \in \mathbb{N}$ are good choices to take large generator elements, but here we should make sure that they are coprime.

3 Numerical Results

A Maple code to generate matrices $A = [a_{ij}]_{1 \leqslant i,j \leqslant 2}$ with integer elements satisfying $a_{ij} \geqslant M$ and $\det A = d$ is as follows.

```
restart:
with(LinearAlgebra):
a11:="here put the element a11":
a12:="here put the element a12":
s:=[op(isolve(a11*X-a12*Y="here put d"))]:
```

```
t:="here put t":
a21:=rhs(eval(s[2],_Z1=t)):
a22:=rhs(eval(s[1],_Z1=t)):
A:=Matrix([[a11,a12],[a21,a22]]);
Determinant(A);
```

As some numerical examples, we run the above Maple code with $t = 1$, giving positive elements, on several generator elements, mentioned in Remark 2.1.

Example 3.1. Let $a_{11} = M$ and $a_{12} = M + 1$, with $M = 299792458$ and $d = 2$. We obtain

$$\det \begin{bmatrix} 299792458 & 299792459 \\ 599584914 & 599584916 \end{bmatrix} = 2.$$

Example 3.2. Let $a_{11} = 2M + 1$ and $a_{12} = 9M + 4$, with $M = 30! - 20! + 1398$ and $d = 1$. Note that we use factorials just to generate large numbers. We obtain

$$\det \begin{bmatrix} 530505719624377251468600606722797 & 2387275738309697631608702730252586 \\ 530505719624377251468600606722799 & 2387275738309697631608702730252595 \end{bmatrix} = 1.$$

Example 3.3. Let $a_{11} = 5M + 2$ and $a_{12} = 7M + 3$, with $M = 2^{97} - 1$ and $d = -1$. We mention that our method works for all integer values of d, including negative values. We obtain

$$\det \begin{bmatrix} 7922816251426433759354395033 57 & 1109194275199700726309615304700 \\ 7922816251426433759354395033 62 & 1109194275199700726309615304707 \end{bmatrix} = -1.$$

Example 3.4. Let $a_{11} = 2M + 3$ and $a_{12} = 4M + 5$, with $M = 2^{34}$ and $d = 1$. We obtain

$$\det \begin{bmatrix} 4835703278458516698824707 & 9671406556917033397649413 \\ 4835703278458516698824708 & 9671406556917033397649415 \end{bmatrix} = 1.$$

We observe that in all of the above examples, elements in twice are close. To get matrices with elements far from each other, we use shifted factorials $n! \pm m$ for practical values of $m, n \in \mathbb{N}$.

Example 3.5. Let $a_{11} = 29! + 139^8$ and $a_{12} = 30! - 13982020$. We obtain

$$\det \begin{bmatrix} 884176199373984130821082768 3681 & 265252859812191058636308466017980 \\ 121224144469037754199301453542 40 & 363672433407107530811944149365921 \end{bmatrix} = 1.$$

As another example, by taking $a_{11} = 27! + 123456789$ and $a_{12} = 26! - 12345678910$ we obtain the equation (1).

4 Conclusion

In this paper we consider a kind of finding conditional square matrices, under certain given conditions. Obtaining such conditional matrices seems to be useful to in specific problems, when we wish to get expected results. Although, we considered matrices of order 2 with large positive integer elements and having small determinant, the similar problem with higher orders seems interesting and hard.

References

[1] R. Bhatia, *Matrix Analysis*, Springer-Verlage, New York, 1997.

[2] D.M. Burton, *Elementary number theory*, McGraw-Hill, New York, 2007.

Slt-majorization and its linear preservers[1]

Asma Ilkhanizadeh Manesh*

Department of Mathematics, Vali-e-Asr University of Rafsanjan,
P.O. Box: 7713936417, Rafsanjan, Iran

Abstract

Let $\mathbf{M}_n$ be the algebra of all n-by-n real matrices. A matrix $R \in \mathbf{M}_n$ with nonnegative entries is called *row substochastic* if each row sum is at most 1. For $x, y \in \mathbb{R}^n$, we say that x is row substochastic lower triangular majorized by y (write as $x \prec_{slt} y$) if there exists a row substochastic lower triangular matrix R such that $x = Ry$. In this paper, the structure of all linear functions $T : \mathbb{R}^2 \to \mathbb{R}^2$, preserving (strongly preserving) $\prec_{slt}$ are characterized.

Keywords: (Strong) linear preserver, Row substochastic matrices, Slt-majorization

Mathematics Subject Classification [2010]: 15A04, 15A51

1 Introduction

Recently, the concept generalized stochastic matrices has been attended specially and many papers have been published in this topic. For example, one can see [1]- [3].

Throughout the article,
$\mathbb{R}^n$ denotes the set of all n-by-1 real vectors;
$\mathcal{RS}_n^{lt}$ denotes the collection of all n-by-n row substochastic lower triangular matrices;
$\{e_1, \ldots, e_n\}$ denotes the standard basis of $\mathbb{R}^n$;
$A(n_1, \ldots, n_l | n_1, \ldots, n_l)$ denotes the submatrix of A obtained from A by deleting rows and columns $n_1, \ldots, n_l$;
$\mathbb{N}_k$ denotes the set $\{1, \ldots, k\} \subset \mathbb{N}$;
A^t denotes the transpose of a given matrix A;
$[T]$ denotes the matrix representation of a linear function $T : \mathbb{R}^n \to \mathbb{R}^n$ with respect to the standard basis;
$\mathcal{C}(\mathcal{A})$ denotes the set $\{\sum_{i=1}^m \lambda_i a_i \mid m \in \mathbb{N},\ \lambda_i \geq 0,\ \sum_{i=1}^m \lambda_i \leq 1,\ a_i \in A,\ i \in \mathbb{N}_m\}$, where $A \subseteq \mathbb{R}^n$.

A linear function $T : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ is said to be a linear preserver (strong linear preserver) of $\sim$ if $T(x) \sim T(y)$ whenever $x \sim y$ ($T(x) \sim T(y)$ if and only if $x \sim y$).

[1]Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

*Speaker. Email address: a.ilkhani@vru.ac.ir

2 Main results

In this section, we will characterize all linear functions that preserves (strongly preserves) slt-majorization on $\mathbb{R}^2$.

Definition 2.1. A matrix R with nonnegative entries is called *row substochastic* if all its row sums is less than or equal to one.

Definition 2.2. Let $x, y \in \mathbb{R}^n$. We say that x *slt-majorized* by y (in symbol $x \prec_{slt} y$) if $x = Ry$, for some $R \in \mathcal{RS}_n^{lt}$.

We bring the followings with no proof.

Lemma 2.3. *Let $x = (x_1, \dots, x_n)^t$, $y = (y_1, \dots, y_n)^t \in \mathbb{R}^n$. Then $x \prec_{slt} y$ if and only if $x_i \in \mathcal{C}\{y_1, \dots, y_i\}$, for all $i \in \mathbb{N}_n$.*

Lemma 2.4. *Suppose $T : \mathbb{R}^n \to \mathbb{R}^n$ is a linear preserver of $\prec_{slt}$. Assume that $S : \mathbb{R}^k \to \mathbb{R}^k$ is the linear function with $[S] = [T](k+1, \dots, n)$. Then S preserves $\prec_{slt}$ on $\mathbb{R}^k$.*

Proof. Let $x' = (x_1, \dots, x_k)^t, y' = (y_1, \dots, y_k)^t \in \mathbb{R}^k$ and let $x' \prec_{slt} y'$. Then, by Lemma 2.3, $x := (x_1, \dots, x_k, 0, \dots, 0)^t \prec_{slt} y := (y_1, \dots, y_k, 0, \dots, 0)^t \in \mathbb{R}^n$ and hence $Tx \prec_{slt} Ty$. This implies that $Sx' \prec_{slt} Sy'$. Therefore, S preserves $\prec_{slt}$ on $\mathbb{R}^k$. □

Lemma 2.5. *Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be a linear preserver of $\prec_{slt}$. Then $[T]$ is lower triangular.*

Proof. Let $[T] = [a_{ij}]$. Use induction on n. For $n = 1$, there is nothing to prove. For $n \geq 2$, assume that the matrix representation of every linear preservers of $\prec_{slt}$ on $\mathbb{R}^{n-1}$ is an upper triangular matrix. Let $S : \mathbb{R}^{n-1} \to \mathbb{R}^{n-1}$ be the linear function with $[S] = [T](n)$. By Lemma 2.4, the linear function S preserves $\prec_{slt}$ on $\mathbb{R}^{n-1}$. The induction hypothesis insures that $[S]$ is an $n-1 \times n-1$ lower triangular matrix. So it is enough to show that $a_{1n} = a_{2n} = \cdots = a_{n-1n} = 0$. Put $x = e_n$ and $y = e_{n-1}$. Then $x \prec_{slt} y$ and hence $Tx = (a_{1n}, a_{2n}, \dots, a_{n-1n}, a_{nn})^t \prec_{slt} (0, \dots, 0, a_{n-1n-1}, a_{nn-1})^t = Ty$. By Lemma 2.3, it implies that $a_{1n} = a_{2n} = \cdots = a_{n-2n} = 0$. So it is enough to show that $a_{n-1n} = 0$. Assume, if possible, that $a_{n-1n} \neq 0$. Without loss of generality, suppose that $a_{n-1n} = 1$.

We consider two cases.

Case 1. $a_{n-1n-1} \neq 0$. Let $x = e_n$ and $y = \dfrac{-1}{a_{n-1n-1}} e_{n-1} + e_n$. So $x \prec_{slt} y$ and hence $Tx \prec_{slt} Ty$. It follows that $1 = 0$, which is a contradiction.

Case 2. $a_{n-1n-1}0$. Let $x = e_n$ and $y = e_{n-1}$. We see $x \prec_{slt} y$ and hence $Tx \prec_{slt} Ty$. It implies that $1 = 0$, a contradiction. Thus $a_{n-1n} = 0$ and hence the induction argument is completed. Therefore, $[T]$ is an lower triangular matrix. □

2.1 Slt-Majorization on $\mathbb{R}^2$

Here, we obtain the structure of all linear functions $T : \mathbb{R}^2 \to \mathbb{R}^2$, preserving $\prec_{slt}$.

Theorem 2.6. *Let $T : \mathbb{R}^2 \to \mathbb{R}^2$ be a linear function. Assume $[T] = [a_{ij}]$. Then T preserves $\prec_{slt}$ if and only if one of the following holds.*

(a) $[T] = \begin{pmatrix} a_{11} & 0 \\ a_{21} & 0 \end{pmatrix}$.

(b) $[T] = \begin{pmatrix} a_{11} & 0 \\ 0 & a_{22} \end{pmatrix}$, $a_{22} \neq 0$, *and* $(0, a_{22}, a_{11})^t$ *is monotone.*

Proof. If T preserves $\prec_{slt}$, Lemma 2.5 ensures that $[T]$ is lower triangular. If $a_{22} = 0$, then we have (a). If $a_{22} \neq 0$; Without loss of generality assume that $a_{22} = 1$. We cliam that $a_{21} = 0$. If $a_{21} \neq 0$; Set $x = e_2$ and $y = \frac{-1}{a_{21}}e_1 + e_2$. We observe that $x \prec_{slt} y$, and then $Tx \prec_{slt} Ty$. This shows that $ac \leq 0$.

We consider two steps.

Step 1. $a_{11} = 0$. By putting $x = e_2$ and $y = \frac{-1}{a_{21}}e_1 + e_2$, we obtain a contradiction.

Step 2. $a_{11} \neq 0$. If $a_{21} < 0 < a_{11}$, set $x = (\frac{1}{a_{21}} - 1)e_1 - e_2$ and $y = -e_1 + a_{21}e_2$. We deduce that $a_{11}a_{21} > 0$, a contradiction. If $a_{11} < 0 < a_{21}$, put $x = (\frac{-1}{a_{21}} + 1)e_1 + e_2$ and $y = e_1 - a_{21}e_2$. We conclude that $a_{11}a_{21} > 0$, a contradiction.

Thus, $a_{21} = 0$. On the other hand, since $e_2 \prec_{slt} e_1$, we find $1 \in \mathcal{C}\{0, a_{11}\}$, and hence $(0, 1, a_{11})^t$ is monotone. We have (b).

To prove the sufficiency, let $x = (x_1, x_2)^t, y = (y_1, y_2)^t \in \mathbb{R}^2$ and let $x \prec_{slt} y$. If (a) holds, we see $Tx \prec_{slt} Ty$. If (b) holds, we could suppose $a_{22=1}$. Then $Tx = (a_{11}x_1, x_2)^t$ and $Ty = (a_{11}y_1, y_2)^t$. As $x_2 \in \mathcal{C}\{y_1, y_2\}$, there exist $\alpha, \beta \geq 0$, $\alpha + \beta \leq 1$ such that $x_2 = \alpha y_1 + \beta y_2$. Thus, $x_2 = \alpha a_{11}(Ty)_1 + \beta(Ty)_2$. We see that $Tx \prec_{slt} Ty$. □

Lemma 2.7. *Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be a linear function that strongly preserves $\prec_{slt}$. Then T is invertible.*

Proof. Suppose that $T(x) = 0$, where $x \in \mathbb{R}^n$. Notice that since T is linear, we have $T(0) = 0 = T(x)$. Then it is obvious that $T(x) \prec_{slt} T(0)$. Therefore, $x \prec_{slt} 0$, because T strongly preserves slt-majorization. So $x = 0$, and hence T is invertible. □

The following theorem characterizes the linear functions $T : \mathbb{R}^2 \to \mathbb{R}^2$ which strongly preserves slt-majorization.

Theorem 2.8. *A linear function $T : \mathbb{R}^2 \to \mathbb{R}^2$ strongly preserves $\prec_{slt}$ if and only if $[T] = \alpha I_2$, for some $\alpha \in \mathbb{R} \setminus \{0\}$.*

Proof. First, suppose that T strongly preserves $\prec_{slt}$. Lemma 2.7 ensures that T is invertible and hence $Te_2 \neq 0$. Theorem 2.6 ensures that

$$[T] = \begin{pmatrix} a_{11} & 0 \\ 0 & a_{22} \end{pmatrix},$$

and the vector $(0, a_{22}, a_{11})^t$ is monotone.

One obtains

$$[T]^{-1} = \begin{pmatrix} \frac{1}{a_{11}} & 0 \\ 0 & \frac{1}{a_{22}} \end{pmatrix}.$$

Since T strongly preserves $\prec_{slt}$, we conclude T^{-1} is a linear preserver of $\prec_{slt}$, and hence the vector $(0, \frac{1}{a_{22}}, \frac{1}{a_{11}})^t$ is monotone. Therefore, $a_{11} = a_{22}$.

For the converse, assume that there exists $\alpha \in \mathbb{R}$ such that $\alpha \neq 0$ and $[T] = \alpha I_2$. Thus, $[T]^{-1} = \frac{1}{\alpha} I_2$. Then both of T and T^{-1} preserve $\prec_{slt}$, and therefore, T strongly preserves $\prec_{slt}$. □

3 Conclusion

Recently, the concept generalized stochastic matrices has been attended specially and many papers have been published in this topic. Due to the importance of the topic in this article, we have focused on this topic.

References

[1] A. Armandnejad and A. Ilkhanizadeh Manesh, Gut-majorization on $\mathbf{M}_{n,m}$ and its linear preservers, *Electronic Journal of Linear Algebra*, 23 (2012) 646-654.

[2] A. Ilkhanizadeh Manesh, Right gut-Majorization on $\mathbf{M}_{n,m}$, *Electronic Journal of Linear Algebra*, 31(2016) 646-654.

[3] A. Ilkhanizadeh Manesh and A. Armandnejad, Ut-Majorization on $\mathbb{R}^n$ and its Linear Preservers, *Operator Theory: Advances and Applications*, 242 (2014) 253–259.

On semi-convergence of the improved symmetric successive over-relaxation method for singular saddle point problems[1]

Mohammad Mahdi Izadkhah*

Department of Computer Science, Faculty of Computer and Industrial Engineering, Birjand University of Technology, Birjand, Iran

Abstract

In this paper, we analyze the semi-convergence of the improved symmetric successive over-relaxation method for singular saddle point problems. In fact, when the $(1,2)$-block of the coefficient matrix is rank deficient, the saddle point problem is singular. Here, we study sufficient conditions for semi-convergence of the improved symmetric successive over-relaxation method.

Keywords: Saddle point problem, Iterative method, Semi-convergence, Improved SSOR, Preconditioner

Mathematics Subject Classification [2010]: 65F08, 65F10

1 Introduction

Consider the following large and sparse saddle point problem

$$\mathcal{A}u \equiv \begin{pmatrix} A & B \\ -B^T & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} p \\ -q \end{pmatrix} \equiv b, \tag{1}$$

where $A \in \mathbb{R}^{m\times m}$ is a symmetric positive definite matrix, $B \in \mathbb{R}^{m\times n}$ is a rank deficient matrix, $p \in \mathbb{R}^m$, and $q \in \mathbb{R}^n$ with $n \leq m$. It follows that the coefficient matrix of the saddle point problem (1) is singular.

The saddle point problems is crucially important in a variety of scientific and engineering applications [3]. Here we mention some applications of the saddle point problems like mixed finite element approximation of elliptic partial differential equations, optimal control, computational fluid dynamics, weighted least-squares problems, electronic networks, computer graphics and nonlinearly constrained optimization, and so forth [1, 2, 5].

When B in (1) is of full rank, a number of iteration methods and their numerical properties have been discussed to solve the saddle point problem (1) in the literature, such as SOR-like method presented in [1] and improved SSOR method given in [2].

Though most often the matrix B occur in the form of full column rank, but not always in practise. For example, in the finite difference discretization of the Navier-Stokes equation with periodic boundary conditions, B in (1) becomes singular [5]. In recent years,

[1] Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

*Speaker. Email address: izadkhah@birjandut.ac.ir

there has been a surge of interest in solving singular saddle point problems (1), which some of them are Uzawa-type method proposed in [5] and Uzawa-HSS method given in [4].

In this paper, the idea of the improved symmetric successive over-relaxation method is established for the singular saddle point problems (ISSORS) of the form (1). We present the semi-convergence conditions of the proposed ISSORS method. Finally, some conclusions are presented.

2 Improved SSOR iteration method

In this section, we proposed improved symmetric successive over-relaxation method given in [2] for solving saddle point problem (1). So, we consider the following splitting

$$\mathcal{A} = \mathcal{D} - \mathcal{A}_l - \mathcal{A}_u, \tag{2}$$

where

$$\mathcal{D} = \begin{pmatrix} A & 0 \\ 0 & Q \end{pmatrix}, \quad \mathcal{A}_l = \begin{pmatrix} -\frac{1}{2}A & 0 \\ B^T & \frac{1}{2}Q \end{pmatrix}, \quad \mathcal{A}_u = \begin{pmatrix} \frac{1}{2}A & -B \\ 0 & \frac{1}{2}Q \end{pmatrix},$$

for nonsingular and symmetric matrix $Q \in \mathbb{R}^{n\times n}$. Set

$$\mathcal{L} = \mathcal{D}^{-1}\mathcal{A}_l = \begin{pmatrix} -\frac{1}{2}I & 0 \\ Q^{-1}B^T & \frac{1}{2}I \end{pmatrix}, \quad \mathcal{U} = \mathcal{D}^{-1}\mathcal{A}_u = \begin{pmatrix} \frac{1}{2}I & -A^{-1}B \\ 0 & \frac{1}{2}I \end{pmatrix}, \tag{3}$$

where I is the identity matrix of the appropriate dimension.

Suppose that $u^{(k)} = [x^{(k)T}, y^{(k)T}]^T$ is the k-th approximation of the exact solution of (1), then improved symmetric successive over-relaxation iterative method is obtained as

$$u^{(k+\frac{1}{2})} = (I - \omega\mathcal{L})^{-1}((1-\omega)I - \omega\mathcal{U})u^{(k)} + \omega(I - \omega\mathcal{L})^{-1}\mathcal{D}^{-1}b, \tag{4}$$

$$u^{(k+1)} = (I - \omega\mathcal{U})^{-1}((1-\omega)I - \omega\mathcal{L})u^{(k+\frac{1}{2})} + \omega(I - \omega\mathcal{U})^{-1}\mathcal{D}^{-1}b. \tag{5}$$

So, based on Eqs. (4) and (5) and assuming $\omega \neq \pm 2$, the ISSORS method can be considered as the following algorithm.

Algorithm 2.1. Improved SSOR iteration method

Given initial guess $u^{(0)} = [(x^{(0)})^T, (y^{(0)})^T]^T$. For $k = 0, 1, 2, \ldots$ until iteration sequence $\{[(x^{(k)})^T, (y^{(k)})^T]^T\}$ is convergent, compute

1. $y^{(k+1)} = y^{(k)} + \frac{4\omega}{2+\omega}Q^{-1}B^T\left\{x^{(k)} + \frac{2\omega}{2-\omega}A^{-1}(p - By^{(k)})\right\} - \frac{4\omega}{2-\omega}Q^{-1}q.$
2. $x^{(k+1)} = \frac{2-3\omega}{2+\omega}x^{(k)} - \frac{2\omega}{2-\omega}A^{-1}B\left\{y^{(k+1)} + \frac{2-3\omega}{2+\omega}y^{(k)}\right\} + \frac{4\omega}{2+\omega}A^{-1}p.$

3 Semi-convergence of the ISSORS

In this section, we discuss the semi-convergence of the ISSORS for solving singular saddle point problem (1). The following Lemma provides the necessary and sufficient conditions for semi-convergence of a stationary iterative method [4, 5].

Lemma 3.1. *Let G be a nonsingular matrix. Then, iteration scheme $x^{(k+1)} = x^{(k)} + G(d - Mx^{(k)})$ used for solving singular linear system $Mx = d$ is semi-convergent if and only if the following two conditions are fulfilled*

(i) $\text{index}(I-T)=1$, *or equivalently,* $\text{rank}(I-T)^2=\text{rank}(I-T)$, *where* $T=I-GM$ *is the iteration matrix.*

(ii) The pseudo-spectral radius of T *is less than 1, i.e.*

$$\nu(T)=\max\{|\lambda|:\lambda\in\sigma(T)\ \ and\ \ \lambda\neq 1\}<1,$$

where $\sigma(T)$ *is the spectrum of the matrix* T.

Here we denote the null space of the matrix A by $\text{null}(A)$. To analyze the semi-convergence properties of the ISSORS iteration method for the singular saddle point problem (1), we write the Algorithm 2.1 as

$$\begin{pmatrix} x^{(k+1)} \\ y^{(k+1)} \end{pmatrix}=\begin{pmatrix} x^{(k)} \\ y^{(k)} \end{pmatrix}+\mathcal{G}\left(b-\mathcal{A}\begin{pmatrix} x^{(k)} \\ y^{(k)} \end{pmatrix}\right), \tag{6}$$

where

$$\begin{aligned}\mathcal{G}&=\omega(2-\omega)(I-\omega\mathcal{U})^{-1}(I-\omega\mathcal{L})^{-1}\mathcal{D}^{-1}\\&=\begin{pmatrix} \frac{4\omega}{2+\omega}A^{-1}-\frac{16\omega^2}{(2-\omega)(4-\omega^2)}A^{-1}BQ^{-1}B^TA^{-1} & -\frac{8\omega^2}{(2-\omega)^2}A^{-1}BQ^{-1} \\ \frac{8\omega^2}{4-\omega^2}Q^{-1}B^TA^{-1} & \frac{4\omega}{2-\omega}Q^{-1}\end{pmatrix}.\end{aligned} \tag{7}$$

By what mentioned above, we can obtain the iteration matrix of the scheme (6) as

$$\begin{aligned}\mathcal{T}&=I-\mathcal{G}\mathcal{A}\\&=\begin{pmatrix} \frac{2-3\omega}{2+\omega}I-\frac{8\omega^2}{4-\omega^2}A^{-1}BQ^{-1}B^T & -\frac{4\omega}{2+\omega}A^{-1}B+\frac{16\omega^2}{(2-\omega)(4-\omega^2)}A^{-1}BQ^{-1}B^TA^{-1}B \\ \frac{4\omega}{2+\omega}Q^{-1}B^T & I-\frac{8\omega^2}{4-\omega^2}Q^{-1}B^TA^{-1}B\end{pmatrix}.\end{aligned} \tag{8}$$

Now, we prove that the ISSORS iteration method is semi-convergent for solving the singular saddle point problem (1) under some restrictions as in Lemma 3.1. It is necessary to mention that, in this algorithm, the matrix Q is an approximation of the Schur complement matrix $B^TA^{-1}B$ [1, 2].

Lemma 3.2. *Let* $A\in\mathbb{R}^{m\times m}$ *be symmetric positive definite,* $B\in\mathbb{R}^{m\times n}$ *be rank deficient, and* $Q\in\mathbb{R}^{n\times n}$ *be nonsingular and symmetric. Then iteration matrix* $\mathcal{T}$ *of the ISSORS method given in Eq.* (8) *satisfies*

$$\text{index}(I-\mathcal{T})=1. \tag{9}$$

Proof. Inasmuch as $\mathcal{T}=I-\mathcal{G}\mathcal{A}$ in (8), so Eq. (9) holds if

$$\text{null}(\mathcal{G}\mathcal{A})=\text{null}((\mathcal{G}\mathcal{A})^2),$$

where $\mathcal{G}$ is defined as in (7). It is obvious that $\text{null}(\mathcal{G}\mathcal{A})\subseteq\text{null}((\mathcal{G}\mathcal{A})^2)$.

Let $x=[x_1^T,x_2^T]^T\in\mathbb{R}^{m+n}$ satisfies $(\mathcal{G}\mathcal{A})^2x=0$. Denote $y=(\mathcal{G}\mathcal{A})x$. So, we have

$$y=\begin{pmatrix} y_1 \\ y_2\end{pmatrix}=\begin{pmatrix} \frac{4\omega}{2+\omega}A^{-1}-\frac{16\omega^2}{(2-\omega)(4-\omega^2)}A^{-1}BQ^{-1}B^TA^{-1} & -\frac{8\omega^2}{(2-\omega)^2}A^{-1}BQ^{-1} \\ \frac{8\omega^2}{4-\omega^2}Q^{-1}B^TA^{-1} & \frac{4\omega}{2-\omega}Q^{-1}\end{pmatrix}\times \tag{10}$$

$$\begin{pmatrix} A & B \\ -B^T & 0\end{pmatrix}\begin{pmatrix} x_1 \\ x_2\end{pmatrix}. \tag{11}$$

This results in

$$\begin{cases} y_1 = \left(\frac{4\omega}{2+\omega} I - \frac{8\omega^3}{(2-\omega)(4-\omega^2)} A^{-1}BQ^{-1}B^T\right) x_1 \\ \quad + \left(\frac{4\omega}{2+\omega} A^{-1}B - \frac{16\omega^2}{(2-\omega)(4-\omega^2)} A^{-1}BQ^{-1}B^TA^{-1}B\right) x_2, \\ y_2 = -\frac{4\omega}{2+\omega} Q^{-1}B^T x_1 + \frac{8\omega^2}{4-\omega^2} Q^{-1}B^TA^{-1}Bx_2. \end{cases} \tag{12}$$

Since $\mathcal{G}$ is invertible and $(\mathcal{GA})y = (\mathcal{GA})^2 x = 0$, it holds that $\mathcal{A}y = 0$, i.e.,

$$Ay_1 + By_2 = 0, \quad -B^T y_1 = 0. \tag{13}$$

It is easy to get $y_1 = -A^{-1}By_2$. Then we obtain $B^TA^{-1}By_2 = 0$. Therefore $By_2 = 0$ and $y_1 = 0$. From $By_2 = 0$, we attain $x_1 = \frac{2\omega}{2-\omega} A^{-1}Bx_2$, that means $y_2 = 0$. Thus $y = (\mathcal{GA})x = 0$, i.e.,

$$\text{null}((\mathcal{GA})^2) \subseteq \text{null}(\mathcal{GA}).$$

The proof is complete. □

Let the singular value decomposition of matrix B be as

$$B = U(B_r, 0)V^T, \quad B_r = \begin{pmatrix} \Sigma_r \\ 0 \end{pmatrix} \in \mathbb{R}^{m\times r}, \quad \Sigma_r = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r) \in \mathbb{R}^{r\times r}, \tag{14}$$

with $U \in \mathbb{R}^{m\times m}, V \in \mathbb{R}^{n\times n}$ being two orthogonal matrices and $\sigma_i (i = 1, 2, \ldots, r)$ being singular values of B. We define

$$P = \begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix}.$$

It is obvious that P is a $(m+n)\times(m+n)$ orthogonal matrix, and therefore we have the orthogonal similarities

$$\hat{\mathcal{T}} = P^T\mathcal{T}P, \quad \hat{A} = U^TAU, \quad \hat{Q} = V^TQV. \tag{15}$$

Hence $\mathcal{T}$ has the same spectrum as the matrix $\hat{\mathcal{T}}$. Let $[\hat{Q}^{-1}]_{i:j,k:s}$ stands for the submatrix of $\hat{Q}^{-1}$ by considering rows i to j and columns from k to s(we use the notation $[\cdot]_r$ for the case $[\cdot]_{1:r,1:r}$),

Lemma 3.3. *Suppose $A \in \mathbb{R}^{m\times m}$ be symmetric positive definite, $B \in \mathbb{R}^{m\times n}$ be rank deficient. Let $\hat{A}$ and $\hat{Q}$ are defined as in (15). Assume that for B_r defined in (14), all eigenvalues of $\hat{Q}^{-1}B_r^T\hat{A}^{-1}B_r$ are real and positive. Then, pseudo-spectral of the iteration matrix $\mathcal{T}$ of the ISSORS method is less that one if*

$$0 < \omega < \frac{2}{1+2\sqrt{\mu}}, \tag{16}$$

where $\mu = \rho(\hat{Q}^{-1}B_r^T\hat{A}^{-1}B_r)$.

Proof. Firstly, by definition of $\hat{\mathcal{T}}$ in (15), it holds that

$$\hat{\mathcal{T}} = \begin{pmatrix} \hat{\mathcal{T}}_{11} & \hat{\mathcal{T}}_{12} \\ \hat{\mathcal{T}}_{21} & \hat{\mathcal{T}}_{22} \end{pmatrix}, \tag{17}$$

where

$$\hat{\mathcal{T}}_{11} = \frac{2-3\omega}{2+\omega} I_m - \frac{8\omega^2}{(4-\omega^2)} U^TA^{-1}BQ^{-1}B^TU$$

$$= \frac{2-3\omega}{2+\omega} I_m - \frac{8\omega^2}{(4-\omega^2)} \hat{A}^{-1}(B_r, 0)\hat{Q}^{-1} \begin{pmatrix} B_r^T \\ 0 \end{pmatrix}$$
$$= \frac{2-3\omega}{2+\omega} I_m - \frac{8\omega^2}{(4-\omega^2)} \hat{A}^{-1} B_r [\hat{Q}^{-1}]_r B_r^T, \tag{18}$$

$$\hat{\mathcal{T}}_{12} = -\frac{4\omega}{2+\omega} U^T A^{-1} BV + \frac{16\omega^2}{(2-\omega)(4-\omega^2)} U^T A^{-1} B Q^{-1} B^T A^{-1} BV$$
$$= -\frac{4\omega}{2+\omega} \hat{A}^{-1}(B_r, 0) + \frac{16\omega^2}{(2-\omega)(4-\omega^2)} \hat{A}^{-1}(B_r, 0)\hat{Q}^{-1} \begin{pmatrix} B_r^T \\ 0 \end{pmatrix} \hat{A}^{-1}(B_r, 0)$$
$$= \left(-\frac{4\omega}{2+\omega} \hat{A}^{-1} B_r + \frac{16\omega^2}{(2-\omega)(4-\omega^2)} \hat{A}^{-1} B_r [\hat{Q}^{-1}]_r B_r^T \hat{A}^{-1} B_r, 0 \right), \tag{19}$$

$$\hat{\mathcal{T}}_{21} = \frac{4\omega}{2+\omega} V^T Q^{-1} B^T U$$
$$= \frac{4\omega}{2+\omega} \hat{Q}^{-1} \begin{pmatrix} B_r^T \\ 0 \end{pmatrix}$$
$$= \frac{4\omega}{2+\omega} \begin{pmatrix} [\hat{Q}^{-1}]_r B_r^T \\ [\hat{Q}^{-1}]_{r+1,n,1:r} \; \mathrm{B}_r^T \end{pmatrix}, \tag{20}$$

$$\hat{\mathcal{T}}_{22} = I_n - \frac{8\omega^2}{4-\omega^2} V^T Q^{-1} B^T A^{-1} BV$$
$$= I_n - \frac{8\omega^2}{4-\omega^2} \hat{Q}^{-1} \begin{pmatrix} B_r^T \\ 0 \end{pmatrix} \hat{A}^{-1}(B_r, 0)$$
$$= \begin{pmatrix} I_r - \frac{8\omega^2}{4-\omega^2} [\hat{Q}^{-1}]_r B_r^T \hat{A}^{-1} B_r & 0 \\ 0 & I_{n-r} \end{pmatrix}. \tag{21}$$

It follows from (18)-(21) that

$$\hat{\mathcal{T}} = \begin{pmatrix} \tilde{\mathcal{T}} & 0 \\ \begin{pmatrix} 0 \\ [\hat{Q}^{-1}]_{r+1,n,1:r} \; \mathrm{B}_r^T \end{pmatrix} & I_{n-r} \end{pmatrix}, \tag{22}$$

with

$$\tilde{\mathcal{T}} = \begin{pmatrix} \frac{2-3\omega}{2+\omega} I_m - \frac{8\omega^2}{(4-\omega^2)} \hat{A}^{-1} B_r [\hat{Q}^{-1}]_r B_r^T & -\frac{4\omega}{2+\omega} \hat{A}^{-1} B_r + \frac{16\omega^2}{(2-\omega)(4-\omega^2)} \hat{A}^{-1} B_r [\hat{Q}^{-1}]_r B_r^T \hat{A}^{-1} B_r \\ \frac{4\omega}{2+\omega} [\hat{Q}^{-1}]_r B_r^T & I_r - \frac{8\omega^2}{4-\omega^2} [\hat{Q}^{-1}]_r B_r^T \hat{A}^{-1} B_r \end{pmatrix}$$

Then from (22), $\nu(\mathcal{T}) = \nu(\hat{\mathcal{T}}) < 1$ holds if and only if $\rho(\tilde{\mathcal{T}}) < 1$. Note that $\tilde{\mathcal{T}}$ can be viewed as the iteration matrix of the ISSOR iteration method [2] applied to the nonsingular saddle point problem with the following coefficient matrix

$$\begin{pmatrix} \hat{A} & B_r \\ -B_r^T & 0 \end{pmatrix},$$

for the preconditioner $\hat{Q} = V^T Q V$. One can finish the proof off by Theorem 2 in [2]. □

To present the semi-convergence properties of the ISSORS method, by making use of the aforementioned Lemmas 3.2 and 3.3, we summarize and state following Theorem.

Theorem 3.4. *Let $A \in \mathbb{R}^{m\times m}$ be symmetric positive definite, $B \in \mathbb{R}^{m\times n}$ be rank deficient, and ω fulfilled in (16) in Lemma 3.3. Then the ISSORS iteration method (6) is semi-convergent for solving the singular saddle point problem (1).*

4 Conclusion

An extension of the improved symmetric successive over-relaxation method has been given for singular saddle point problems. The method involves one preconditioning matrix for clustering eigenvalues of the iteration matrix of ISSORS method. Furthermore, sufficient conditions has been given for the semi-convergence of the proposed ISSORS method.

References

[1] G.H. Golub, X. Wu, J.Y. Yuan, SOR-like methods for augmented systems, BIT 41 (2001), 71–85.

[2] D. K. Salkuyeh, S. Shamsi and A. Sadeghi, An improved symmetric SOR iterative method for augmented systems, *Tamkang J. Math.*, 43 (2012), No. 4, 479–490.

[3] Z. Li, R. Chu, H. Zhang, Accelerating the shift-splitting iteration algorithm, *Appl. Math. Comput.* 361 (2019) 421–429.

[4] A.-L. Yang, X. Li and Y.-J. Wu, On semi-convergence of the Uzawa-HSS method for singular saddle point problems, *Appl. Math. Comput.* 252 (2015) 88–98.

[5] N.-M. Zhang, T.-T. Lu and Y.-M. Wei, Semi-convergence analysis of Uzawa methods for singular saddle point problems, *J. Comput. Appl. Math.* 255 (2014) 334–345.

دانشگاه شهید باهنر کرمان، ایران
دهمین سمینار جبر خطی و کاربردهای آن
۲۶ لغایت ۲۹ مرداد ۱۳۹۹
10
Shahid Bahonar University of Kerman, Iran
Seminar on Linear Algebra and its Applications
August 16 - 19 , 2020

Quasi invariant polynomials of a matrix[1]

Amir Jafari* and Amin Najafi Amin

Department of Mathematical Sciences, Sharif University of Technology, Tehran, Iran

Abstract

Let n be a positive integer and A be a square matrix of size n, with entries in a field F. A polynomial $p \in F[x_1, \ldots, x_n]$ is called a quasi-invariant polynomial for A if $p(xA)$ is a constant multiple of $p(x)$ where x is the vector $(x_1, \ldots, x_n)$. In this article, we classify all quasi-invariant polynomials of a given non singular matrix A when F is algebraically closed and of characteristic zero. The classification is done by constructing a canonical basis that will be made precise in the text.

Keywords: Linear algebra, Invariant polynomial, Jordan normal form, Quasi invariant

Mathematics Subject Classification [2010]: 15A03, 15A23

1 Introduction

This article deals with the type of mathematics studied in 19th century by Cayely, Klein, Hilbert, etc. see for example [2] and for a modern treatment [3].

Let F be a field and A is a square matrix of size n and entries in F, a polynomial $p \in F[x_1, \ldots, x_n]$ is a quasi invariant for A if

$$p(xA) = cp(x)$$

for some fixed $c \in F$. For example a linear polynomial $p(x) = a_1x_1 + \cdots + a_nx_n$ is a quasi invariant polynomial for A if and only if $a = (a_1, \ldots, a_n)^T$ is an eigen-vector for A, that is $Aa = ca$ for some $c \in F$. When A is a diagonalizabe matrix over F, then we will get n linear quasi invariant polynomials $p_1, \ldots, p_n$ this way corresponding respectively to the eigen-values $c_1, \ldots, c_n$ (with required multiplicities). It is easy to show that any quasi invariant polynomial p, with $p(xA) = cp(x)$, is written uniquely as

$$p = \sum a_{i_1,\ldots,i_n} p_1^{i_1} \ldots p_n^{i_n}$$

where for all $(i_1, \ldots, i_n)$ that $a_{i_1,\ldots,i_n} \neq 0$, we have

$$c_1^{i_1} \ldots c_n^{i_n} = c.$$

To classify all quasi invariant polynomials of a matrix A, one may replace A with a matrix $B = SAS^{-1}$ similar to it. It is because, one has $p(xA) = cp(x)$ if and only if $q(xB) = cq(x)$

[1] Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

*Speaker. Email address: ajafari@sharif.ir

with $q(x) = p(xS)$. So if F is an algebraically closed field, we may without loss of generality assume that A is in normal Jordan form. The examples below show that in absence of diagonalizability, interesting phenomena can happen and we may get quasi invariants of genuinely higher degrees. We use $J_{\lambda,n}$ for the size n Jordan block with λ on its main diagonal and 1 on its off diagonal right below the main diagonal.

Example 1.1. If $A = J_{\lambda,3}$ then other than the obvious quasi invariant polynomial $p_1(x_1, x_2, x_3) = x_3$, the polynomial

$$p_2(x_1, x_2, x_3) = \lambda(x_2^2 - 2x_1x_3) - x_2x_3$$

is also quasi invariant with $p_2(xA) = \lambda^2 p_2(x)$.

Example 1.2. If $A = J_{\lambda,4}$ then other than the obvious quasi invariant polynomial $p_1 = x_4$ and the polynomial $p_2 = \lambda(x_3^2 - 2x_2x_4) - x_3x_4$ coming from Example 1.1, the polynomial

$$p_3 = \lambda(-x_3^3 + 3x_2x_3x_4 - 3x_1x_4^2) - 2x_2x_4^2 + x_3^2x_4$$

is quasi invariant with $p_3(xA) = \lambda^3 p_3(x)$.

Example 1.3. If $A = \mathrm{diag}(J_{\lambda_1,2}, J_{\lambda_2,2})$ is a square matrix of size 4 with two Jordan blocks, then other than the obvious quasi invariant polynomials $p_1 = x_2$ and $p_2 = x_4$, the polynomial $p_3 = \lambda_1 x_1x_4 - \lambda_2 x_2x_3$ is quasi invariant with $p_3(xA) = \lambda_1\lambda_2 p_3(x)$.

Example 1.4. In Example 1.2 above, any polynomial p of the form

$$\sum a_{i_1,i_2,i_3} p_1^{i_1} p_2^{i_2} p_3^{i_3}$$

where $i_1 + 2i_2 + 3i_3$ is a fixed integer k is quasi invariant with $p(xA) = \lambda^k p(x)$. One might conjecture that these are all quasi invariant polynomials for A, which is in fact wrong. The following degree 4 polynomial p given by

$$\lambda^3(-3x_3^2x_2^2 - 6x_1x_3^3 + 8x_2^3x_4 - 18x_1x_2x_3x_4 + 9x_1^2x_4^2) + 3\lambda^2(x_2x_3^3 - 2x_2^2x_3x_4 - 3x_1x_3^2x_4 + 6x_1x_2x_4^2)$$
$$+ \lambda(-5x_2x_3^2x_4 + 8x_2^2x_4^2 + 3x_1x_3x_4^2) + 2x_2x_3x_4^2$$

is quasi invariant with $p(xA) = \lambda^4 p(x)$. However it can not be written as a polynomial in terms of p_1, p_2 and p_3 . However it can be written as a rational function

$$p = \frac{-p_2^3 + p_1p_2p_3 + \lambda p_3^2}{x_4^2}.$$

The goal of this article is to show that for a non-singular matrix, which we may assume is in normal Jordan form

$$A = \mathrm{diag}(J_{\lambda_1,n_1}, \ldots, J_{\lambda_k,n_k})$$

with $n_1 \geq \cdots \geq n_l \geq 2 > n_{l+1} = \cdots = n_k = 1$ and $l \geq 1$, besides the obvious k quasi invariant polynomials $x_{n_1}, x_{n_1+n_2}, \ldots, x_{n_1+\cdots+n_k} = x_n$, one can add $k-1$ extra explicitly constructed quasi invariant polynomials of degrees 2 and 3 to get a set $p_1, \ldots, p_{n-1}$ of quasi invariant polynomials with $p_i(xA) = c_i p_i(x)$ for $i = 1, \ldots, n-1$ such that any quasi invariant polynomial $p(x)$ with $p(xA) = cp(x)$ can be uniquely expressed as a rational function

$$\frac{\sum a_{i_1,\ldots,i_{n-1}} p_1^{i_1} \ldots p_{n-1}^{i_{n-1}}}{x_{n_1}^{m_1} \ldots x_{n_1+\cdots+n_l}^{m_l}}$$

where for all $i_1, \ldots, i_{n-1}$ with non zero $a_{i_1,\ldots,i_{n-1}}$ we must have $c_1^{i_1} \ldots c_{n-1}^{i_{n-1}}$ are fixed independent of $i_1, \ldots, i_{n-1}$.

2 Main results

In this section a canonical basis for a matrix A in a normal Jordan form will be constructed. First we need few lemmas and conventions. The binomial coefficient $\binom{n}{k}$ is $\frac{n(n-1)\ldots(n-k+1)}{k!}$ which is defined for all real numbers and n and all integers $k \geq 0$, we set $\binom{n}{k}$ to be zero if $k < 0$.

Lemma 2.1. *Let $A = J_{\lambda,n}$ and $n > 1$ is an odd number. Let $m = \frac{n-1}{2}$. The polynomial of degree 2*

$$p_{n-1} = \sum_{i=0}^{m} \sum_{j=0}^{i} A_{i,j} \lambda^i x_{m+1-i+j} x_{n-j}$$

where

$$A_{i,j} = (-1)^{m-j} \left(2\binom{m-j-1}{i-j-1} + \binom{m-j-1}{i-j} \right)$$

is a quasi invariant polynomial with $p_{n-1}(xA) = \lambda^2 p_{n-1}(x)$.

Proof. When A is applied to x, x_i will be transformed to $\lambda x_i + x_{i+1}$ where $x_{n+1} := 0$ by convention. So in order to show $p_{n-1}(xA) = \lambda^2 p(x)$ we need to show the following relation holds for $A_{i,j}$

$$A_{i,j} + A_{i+1,j+1} + A_{i,j+1} = 0$$

which is an easy consequence of the Pascal's identity. □

Lemma 2.2. *Let $A = J_{\lambda,n}$ and $n > 2$ is an even number. Let $m = \frac{n-2}{2}$. The polynomial of degree 3*

$$p_{n-1} = \sum_{i=0}^{m} (x_{n-1} g_i - x_n h_i) \lambda^i$$

where

$$g_i = (-1)^i x_{m+2} x_{n-i} - \sum_{j=1}^{i} \left(\binom{m-j}{i-j+1} + 2\binom{m-j}{i-j} \right) x_{m+1-i+j} x_{n+1-j}$$

and

$$h_i = (m+1+i)\binom{m}{i} x_{m+1-i} x_n + \sum_{j=1}^{i} (m+1+i-2j)\binom{m-j}{i-j} x_{m+1-i+j} x_{n-j}$$

is quasi invariant for A, i.e. $p_{n-1}(xA) = \lambda^3 p_{n-1}(x)$.

Proof. This is similar and a little more tedious than the previous lemma and will be left to the reader. □

Lemma 2.3. *Let $A = diag(J_{\lambda_1,n_1}, \ldots, J_{\lambda_k,n_k})$ be a square matrix in Jordan normal form, with $n_1 \geq \cdots \geq n_l \geq 2 > n_{l+1} = \cdots = n_k = 1$, then for $i = i, \ldots, l-1$, the polynomials*

$$p_{n-l+i} = \lambda_1 x_{n_1-1} x_{n_1+\cdots+n_{i+1}} - \lambda_{i+1} x_{n_1} x_{n_1+\cdots+n_{i+1}-1}$$

are quasi invariant polynomials for A with $p_{n-l+i}(xA) = \lambda_1 \lambda_{i+1} p(x)$.

Proof. Easily checked by evaluating $p_{n-l+i}(xA)$. □

Remark 2.4. These polynomials are extensions of the examples 1.1,1.2 and 1.3 of the introduction.

To make the presentation a little easier, we will assume here after that all matrices are non-singular, i.e. all eigen-values are non-zero. The ground field F is also assumed to be algebraically closed and of characteristic zero.

Theorem 2.5. *If A is a Jordan block $J_{\lambda,n}$ with $\lambda \neq 0$, then if $n = 1$ or $n = 2$ and quasi invariant polynomial $p(x)$ with $p(xA) = cp(x)$ is othe form $\sum a_i x_n^i$ where for all i with $a_i \neq 0$, $\lambda^i = c$. If $n \geq 3$, then we have invariant polynomials $p_2(x_{n-2}, x_{n-2}, x_{n-3})$, $p_3(x_{n-3}, x_{n-2}, x_{n-1}, x_n), dots, p_{n-1}(x_1, \ldots, x_n)$ of degrees 2 and 3 alternatively constructed from lemma 2.1 and lemma 2.2. for lower corner submatrices of A of sizes $3, 4, \ldots, n$ that together with $p_1 = x_n$ form a basis in the sense that any invariant polynomial p is uniquely written as*

$$\frac{\sum a_{i_1,\ldots,i_{n-1}} p_1^{i_1} \cdots p_{n-1}^{i_{n-1}}}{x_n^k}$$

where $-k + i_1 + 2i_2 + 3i_3 + 2i_4 + 3i_5 + \ldots$ is fixed.

Proof. The case $n = 1$, is trivial. Now let $n = 2$. Assume that $f(x_1, x_2) = \sum_{i=0}^{m} a_i x_1^i x_2^{m-i}$ is a quasi invariant polynomial of degree m, with $f(\lambda x_1 + x_2, \lambda x_2) = cf(x_1, x_2)$. Assume that the highest power of x_1 in f is k. By comparing the coefficients of $x_1^k x_2^{m-k}$ of both sides it follows that $c = \lambda^m$. By comparing the coeffiecients of $x_1^{k-1} x_2^{m-k+1}$ of both sides it follows that $a_{k-1}\lambda^m + k a_k \lambda^{m-1} = a_{k-1}\lambda^m$ and hence $k = 0$. This shows that x_2 is a basis for the space of all quasi invariant polynomials for A. Now we prove the theorem by induction on n. It is easy to check that $p_1, \ldots, p_{n-1}$ are algebraically independent using the fact each time a new variable appear in them. Let $p(x_1, \ldots, x_n)$ be a quasi invariant polynomial, if x_1 does not appear in p, then by induction p has the required represention in term of $p_1, \ldots, p_{n-2}$. If the power of x_1 in p is $m \geq 1$, write $p = \sum_{i=0}^{m} h_i(x_2, \ldots, x_n) x_1^m$. Then by comparing the highest power of x_1 in both sides of $p(xA) = cp(x)$ it follows that $h_m(xA) = c\lambda^{-m} h_m(x)$. Note by the construction of p_{n-1} in lemmas 2.1 and 2.2, one has

$$p_{n-1} = a x_1 x_n^r + s(x_2, \ldots, x_n)$$

where $r = 1$ if n is odd and $r = 2$ if n is even, a is a non-zero element (since characteristic is zero) and $p_{n-1}(xA) = \lambda^{r+1} p_{n-1}(x)$. It follows that

$$q(x) = (a x_n^r)^m p(x) - h_m(x)(p_{n-1}(x))^m$$

is quasi invariant with $q(xA) = c\lambda^{mr} q(x)$ and the power of x_1 is at most $m - 1$ in $q(x)$. So by induction (on highest exponent of x_1), we can write $q(x)$ in terms of $p_1, \ldots, p_{n-1}$ as required in the theorem and then solving for $p(x)$, the theorem will be proved. □

Remark 2.6. There are many more examples of basis for A besides the one given in Theorem 2.5. It can be shown that any basis can not have more than $\lfloor \frac{n-1}{2} \rfloor$ forms of degree 2. In the construction given in the Theorem 2.5 this maximum number is achieved.

Theorem 2.7. *Let $A = diag(J_{\lambda_1,n_1}, \ldots, J_{\lambda_k,n_k})$ be a square matrix in Jordan normal form, with $n_1 \geq \cdots \geq n_l \geq 2 > n_{l+1} = \cdots = n_k = 1$ and $l > 0$. Then by the previous theorem each Jordan block of size $n_i > 1$ will give $n_i - 1$ quasi invariant polynomials and so together we get $n - l$ quasi invariant polynomials $p_1, \ldots, p_{n-l}$ and then by using lemma 2.3 we construct $l - 1$ quasi invariant polynomials $p_{n-l+1}, \ldots, p_{n-1}$ of degree 2. Let*

$p_i(xA) = c_i p(x)$. *Any quasi invariant polynomial* $p(x)$ *with* $p(xA) = cp(x)$ *can be written uniquely as*

$$\frac{\sum a_{i_1,\ldots,i_{n-1}} p_1^{i_1} \ldots p_{n-1}^{i_{n-1}}}{x_{n_1}^{m_1} \ldots x_{n_1+\cdots+n_l}^{m_l}}$$

where for all $i_1, \ldots, i_{n-1}$ *with non zero* $a_{i_1,\ldots,i_{n-1}}$ *we must have* $c_1^{i_1} \ldots c_{n-1}^{i_{n-1}} = c\lambda_1^{m_1} \ldots \lambda_l^{m_l}$.

Proof. The proof of algebraic independence of $p_1, \ldots, p_{n-1}$ is skipped. This can be done using Jacobian criterion in [1]. Assume $p(x)$ is a quasi invariant polynomial with $p(xA) = cp(x)$. If $p(x)$ does not have x_1, then by induction on n (size of the matrix), $p(x)$ has the desired rational expression. If $n_1 > 2$, then the same proof as before using p_{n_1-1} will reduce the power of x_1 in $p(x)$ and induction on the exponent of x_1 will finish the proof. If $n_1 = 2$, and all $n_2 = \cdots = n_k = 1$ then the statement of the theorem becomes trivial, using a similar technique used in the previous theorem for $n = 2$. Finally if $n_2 = 2$ (note that $n_1 \geq n_2$, so $n_2 \leq 2$) then one can use $r = \lambda_1 x_1 x_4 - \lambda_2 x_2 x_3$ to reduce the power of x_1 in $p(x)$. Let

$$p(x) = \sum_{i=0}^{m} h_i(x_2, \ldots, x_n) x_1^i$$

$$q(x) = (\lambda_1 x_4)^m p(x) - (\lambda_1 x_1 x_4 - \lambda_2 x_2 x_3)^m h_m(x)$$

is a quasi invariant polynomial whose x_1 variable has degree less than m. Due to the restriction in space, we leave the details for a full version of this paper.

□

Remark 2.8. The introduction of quasi invariant instead of invariant has the benefit of making the space in some sense finitely generated. Hilbert proved that for a finite group of matrices the space of all invariant polynomials is finitely generated, however if the group is not finite, for example a cyclic infinite order group generated by a non-singular matrix A, this space is not in general finitely generated. A quasi invariant homogeneous polynomial of degree d, is in fact an eigen-vector of the k fold Kronecker product of A with itself. So the theory of quasi invariant polynomials is in fact an eigen-vector problem for all Kronecker products of A with itself. In the literature there are results about the eigen-values of the Kronecker products but very few results about the eigen-vector problem.

References

[1] M.Beecken, J.Mittmann, N.Saxena, Algebraic independence and blackbox identity testing, Hausdorff center for Mathematics, Bonn, Germany , 2011.

[2] D. Hilbert, Über die vollen invarientensysteme (On full invariant systems), *Math Annalen*, 42(3), 313. 1893.

[3] B. Sturmfels, Algorithms in invariant theory, Springer Wien New York, second edition, 2008.

Normalization method on max-plus algebra and its application[1]

Sedighe Jmashidvand*, Fateme Olia and Shaban Ghalandarzadeh

Faculty of Mathematics, K. N. Toosi University of Technology, Tehran, Iran

Abstract

In this paper, we introduce and analyze the normalization method for solving a system of linear equations over max-plus algebra. We use this method to construct an associated normalized matrix, which gives a technique for solving the linear system. We present a procedure to determine the column rank and the row rank of a matrix.

Keywords: Semiring, Max-plus algebra, System of linear equations, Column rank, Row rank

Mathematics Subject Classification [2010]: 16Y60, 65F05, 15A03

1 Introduction

Systems of linear equations play a fundamental role in mathematics problems. Solving these systems is among the important tasks of linear algebra. We intend to present a method for examining the behavior of linear systems and solving them over Max-plus algebra. The first notion of a semiring was given by Vandiver [3] in 1934. A semiring $(S, +, ., 0, 1)$ is an algebraic structure in which $(S, +)$ is a commutative monoid with an identity element 0 and $(S, .)$ is a monoid with an identity element 1, connected by ring-like distributivity. The additive identity 0 is multiplicatively absorbing, and $0 \neq 1$. Note that for convenience, we mainly consider $S = (\mathbb{R} \cup \{-\infty\}, max, +, -\infty, 0)$ which is called "max $-$plus algebra" in this work. We want to solve the system $AX = b$, where $A = (a_{ij}) \in M_{m\times n}(S)$, $b \in S^m$ and X is an unknown vector of size n. To this end, we present a necessary and sufficient condition based on the associated normalized matrix, which is obtained from a proposed normalization method. Additionally, we introduce an equivalent relation over matrices that implies the associated normalized matrix of a linear system and each of its equivalent systems should be the same. As such, the solvability of a linear system and its equivalent system depend on each other. Determining the column rank and the row rank of a matrix is of particular interest in studying the behavior of matrices. As a result of the normalization method, we are able to find the column rank and the row rank of matrices over tropical semirings.

[1] Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

*Speaker. Email address: sjamshidvand@mail.kntu.ac.ir

2 Definitions and Preliminaries

Definition 2.1. (See [1]) Let S be a semiring. A left S-semimodule is a commutative monoid $(\mathcal{M}, +)$ with identity element $0_{\mathcal{M}}$ for which we have a scalar multiplication function $S \times \mathcal{M} \longrightarrow \mathcal{M}$, denoted by $(s, m) \mapsto sm$, which satisfies the following conditions for all $s, s' \in S$ and $m, m' \in \mathcal{M}$:

1. $(ss')m = s(s'm)$;
2. $s(m + m') = sm + sm'$;
3. $(s + s')m = sm + s'm$;
4. $1_S m = m$;
5. $s0_{\mathcal{M}} = 0_{\mathcal{M}} = 0_S m$.

Right semimodules over S are defined in an analogous manner.

Definition 2.2. A nonempty subset $\mathcal{N}$ of a left S-semimodule $\mathcal{M}$ is a subsemimodule of $\mathcal{M}$ if $\mathcal{N}$ is closed under addition and scalar multiplication. The rank of a left S-semimodule $\mathcal{M}$ is the smallest n for which there exists a set of generators of $\mathcal{M}$ with cardinality n.

Definition 2.3. (See [2]) Let $\mathcal{M}$ be a left S-semimodule. A nonempty subset $\mathcal{A}$ of $\mathcal{M}$ is called linearly independent if $\alpha \notin Span(\mathcal{A} \setminus \{\alpha\})$ for any $\alpha \in \mathcal{A}$. If $\mathcal{A}$ is not linearly independent then it is called linearly dependent.

Definition 2.4. (See [4]) Let $A \in M_{m\times n}(S)$. The right S-subsemimodule of $M_{m\times 1}(S)$ generated by the columns of A is called column space and denoted by $colrank(A)$. Similarly, The left S-subsemimodule of $M_{1\times n}(S)$ generated by the rows of A is called row space and denoted by $rowrank(A)$.

The set of all $m \times n$ matrices over S denotes by $M_{m\times n}(S)$. The matrix operations for any $A, B \in M_{m\times n}(S)$, $C \in M_{n\times l}(S)$ and $\lambda \in S$ can be considered as follows.

$$A + B = (\max(a_{ij}, b_{ij}))_{m\times n}, \quad AC = (\max_{k=1}^{n}(a_{ik} + c_{kj}))_{m\times l}, \quad and \quad \lambda A = (\lambda + a_{ij})_{m\times n}.$$

For convenience, we can denote the scalar multiplication λA by $\lambda + A$. Moreover, max $-$plus algebra is a commutative semiring, which implies $\lambda + A = A + \lambda$.

we study the system of linear equations $AX = b$ where $A \in M_{m\times n}(S)$, $b \in S^m$ and X is an unknown column vector of size n over max $-$pluse algebra, whose $i-$th equation is

$$\max(a_{i1} + x_1, a_{i2} + x_2, \cdots, a_{in} + x_n) = b_i.$$

Definition 2.5. Let $A, B \in M_n(S)$ such that $A = (a_{ij})$ and $B = (b_{ij})$. We say $A \leq B$ if and only if $a_{ij} \leq b_{ij}$ for every $i \in \underline{m}$ and $j \in \underline{n}$ where $\underline{n} = \{1, \cdots, n\}$ and $\underline{m} = \{1, \cdots, m\}$.

Definition 2.6. A solution X^* of the system $AX = b$ is called maximal, if $X \leq X^*$ for any solution X.

Definition 2.7. A vector $b \in S^m$ is called regular if $b_i \neq -\infty$ for any $i \in \underline{m}$.

3 Main results

In this section, we introduce a method, which we call the normalization method, for solving a system of linear equations. Consider the system of linear equations $AX = b$, where $A = (a_{ij}) \in M_{m\times n}(S)$, $b = (b_i)$ is a regular $m-$vector over S and X is an unknown $n-$vector. Let the j-th column of the matrix A be denoted by A_j .

Definition 3.1. (Normalization Method) Let $A \in M_{m\times n}(S)$ and $A_j \in S^m$ be a regular vector for any $j \in \underline{n}$. Then the normalized matrix of A is denoted by

$$\tilde{A} = \left[\ A_1 - \hat{A}_1 \ \middle|\ A_2 - \hat{A}_2 \ \middle|\ \cdots \ \middle|\ A_n - \hat{A}_n \ \right],$$

where $\hat{A}_j = \frac{a_{1j}+a_{2j}+\cdots+a_{mj}}{m}$ for every $j \in \underline{n}$.
Similarly, the normalized vector of the regular vector $b \in S^m$ is

$$\tilde{b} = b - \hat{b},$$

where $\hat{b} = \frac{b_1+b_2+\cdots+b_m}{m}$.
As such, we can rewrite the system $AX = b$ as the normalized system $\tilde{A}Y = \tilde{b}$, where $Y = (\hat{A}_j - \hat{b}) + X = (\hat{A}_j - \hat{b} + x_j)_{j=1}^n$, as follows.

$$\begin{aligned}
AX = b \ &\Rightarrow \max(A_1 + x_1, A_2 + x_2, \cdots, A_n + x_n) = b \\
&\Rightarrow \max((A_1 - \hat{A}_1) + \hat{A}_1 + x_1, (A_2 - \hat{A}_2) + \hat{A}_2 + x_2, \cdots, (A_n - \hat{A}_n) + \hat{A}_n + x_n) = (b - \hat{b}) + \hat{b} \\
&\Rightarrow \max(\tilde{A}_1 + \hat{A}_1 + x_1, \tilde{A}_2 + \hat{A}_2 + x_2, \cdots, \tilde{A}_n + \hat{A}_n + x_n) = \tilde{b} + \hat{b} \\
&\Rightarrow \max(\tilde{A}_1 + (\hat{A}_1 - \hat{b} + x_1), \tilde{A}_2 + (\hat{A}_2 - \hat{b} + x_2), \cdots, \tilde{A}_n + (\hat{A}_n - \hat{b} + x_n)) = \tilde{b} \\
&\Rightarrow \max(\tilde{A}_1 + y_1, \tilde{A}_2 + y_2, \cdots, \tilde{A}_n + y_n) = \tilde{b} \\
&\Rightarrow \tilde{A}Y = \tilde{b}.
\end{aligned}$$

Hence $y_j \leq \tilde{b}_i - \tilde{a}_{ij}$ for every $i \in \underline{m}$ and $j \in \underline{n}$. Now, we define the associated normalized matrix $Q = (q_{ij}) \in M_{m\times n}(S)$ where $q_{ij} = \tilde{b}_i - \tilde{a}_{ij}$. We choose y_j as the minimum element of Q_j (the j-th column of Q), which we call the "j-th column minimum element".
It should be noted that if $a_{ij} = -\infty$ for some $i \in \underline{m}$ and $j \in \underline{n}$, then we will not count a_{ij} in the normalization process of column A_j, i.e.

$$\hat{A}_j = \frac{a_{1j} + a_{2j} + \cdots + a_{(i-1)j} + a_{(i+1)j} + \cdots + a_{mj}}{m-1}.$$

As such, $\tilde{a}_{ij} = -\infty$ and we set $q_{ij} := (-\infty)^-$ such that $s < (-\infty)^-$ for any $s \in S$. Thus, q_{ij} does not affect the $j-$th column minimum element. Consequently and without loss of generality, we assume that every column of the system matrix is regular.

Theorem 3.2. *The linear system of equations $AX = b$ has solutions if and only if there exists at least one column minimum element in every row of Q.*

Proof. Let the system $AX = b$ has solutions. Suppose the i-th row of Q has no column minimum element for some $i \in \underline{m}$. That is $y_j < \tilde{b}_i - \tilde{a}_{ij}$ for every $j \in \underline{n}$, therefore the i-th equation of the system $\tilde{A}Y = \tilde{b}$ is

$$\max(\tilde{a}_{i1} + y_1, \tilde{a}_{i2} + y_2, \cdots, \tilde{a}_{in} + y_n) < \tilde{b}_i.$$

Hence, the system $\tilde{A}Y = \tilde{b}$ and a fortiori the system $AX = b$ have no solution, which is a contradiction.

Conversely, suppose that every row of the matrix Q contains at least one column minimum element, so for any $i \in \underline{m}$ there is some $j \in \underline{n}$ such that $y_j = \tilde{b}_i - \tilde{a}_{ij}$. Then

$$\max(\tilde{a}_{i1} + y_1, \tilde{a}_{i2} + y_2, \cdots, \tilde{a}_{ij} + y_j, \cdots, \tilde{a}_{in} + y_n) = \tilde{b}_i$$

for every $i \in \underline{m}$. Thus, the system $\tilde{A}Y = \tilde{b}$ and consequently the system $AX = b$ have solutions. □

Remark 3.3. The solution of the system $AX = b$ that is obtained from Theorem 3.2 is maximal.

Example 3.4. Let $A \in M_{4\times5}(S)$. Consider the following system $AX = b$:

$$\begin{bmatrix} 165 & 57 & 72 & -7 & 0 \\ 141 & 64 & 48 & 3 & -1 \\ 137 & 101 & 46 & 0 & 2 \\ -243 & 98 & -206 & 156 & -5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 102 \\ 78 \\ 76 \\ 160 \end{bmatrix}.$$

By Definition 3.1, the system $AX = b$ is rewritten as the normalized system $\tilde{A}Y = \tilde{b}$:

$$\begin{bmatrix} 115 & -23 & 82 & -45 & 1 \\ 91 & -16 & 58 & -35 & 0 \\ 87 & 21 & 56 & -38 & 3 \\ -293 & 18 & -196 & 118 & -4 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = \begin{bmatrix} -2 \\ -26 \\ -28 \\ 56 \end{bmatrix}.$$

Note that the j-th column of $\tilde{A}$ is $\tilde{A}_j = (a_{ij} - \hat{A}_j)_{i=1}^4$, for any $1 \le j \le 5$ and $\tilde{b} = (b_i - \hat{b})_{i=1}^4$, where $\hat{A}_1 = 50$, $\hat{A}_2 = 80$, $\hat{A}_3 = -10$, $\hat{A}_4 = 38$, $\hat{A}_5 = -1$, $\hat{b} = 104$. Now, we can build the matrix $Q = (q_{ij}) \in M_{4\times5}(S)$, with $q_{ij} = \tilde{b}_i - \tilde{a}_{ij}$ as follows.

$$\begin{bmatrix} \boxed{-117} & 21 & \boxed{-84} & 43 & -3 \\ \boxed{-117} & -10 & \boxed{-84} & 9 & -26 \\ -115 & \boxed{-49} & \boxed{-84} & 10 & \boxed{-31} \\ 349 & 38 & 252 & \boxed{-62} & 60 \end{bmatrix};$$

where the minimum column elements are boxed. Since every row of Q contains at least one of these minimum column elements, due to Theorem 3.2, the system $\tilde{A}Y = \tilde{b}$ has the maximal solution Y^*:

$$Y^* = \begin{bmatrix} -117 \\ -49 \\ -84 \\ -62 \\ -31 \end{bmatrix}.$$

Hence, the system $AX = b$ has the maximal solution X^*:

$$X^* = \begin{bmatrix} -63 \\ -25 \\ 30 \\ 4 \\ 74 \end{bmatrix};$$

where $x_j^* = y_j^* - \hat{A}_j + \hat{b}$, for any $1 \le j \le 5$.

3.1 Solving equivalent systems of linear equations

Definition 3.5. Let $A, A' \in M_{m\times n}(S)$.We say A is equivalent to A' if there exist nonzero coefficients $\alpha_1, \alpha_2, \cdots, \alpha_n \in S$ such that $A'_j = A_j + \alpha_j$ for any $j \in \underline{n}$, and we write

$$A \sim A' \Longleftrightarrow A' = [A_1 + \alpha_1 | \cdots | A_n + \alpha_n]$$

for some $\alpha_1, \alpha_2, \cdots, \alpha_n \in S\backslash\{-\infty\}$.

The equivalence class of A is defined as follows.

$$[A] = \{A' \in M_{m\times n}(S) | A \sim A'\}$$

Note that this equivalence relation also holds for vectors.

Theorem 3.6. *Let $A \in M_{m\times n}(S)$ and $b \in S^m$ be a regular vector. Then the system $AX = b$ has solutions if and only if the equivalent system $A'X' = b'$ has solutions for any $A' \in [A]$ and $b' \in [b]$.*

Proof. Suppose $AX = b$ has solutions. By theorem 3.2, every row of its associated normalized matrix, $Q = (q_{ij})$, contains at least one column minimum element, where

$$q_{ij} = \tilde{b_i} - \tilde{a}_{ij} = (b_i - \hat{b}) - (a_{ij} - \hat{A}_j).$$

On the other hand, since $A' = (a'_{ij}) \in [A]$ and $b' = (b'_i) \in [b]$, there exist coefficients $\alpha_1, \alpha_2, \cdots, \alpha_n, \beta \in S\backslash\{-\infty\}$ such that $a'_{ij} = a_{ij} + \alpha_j$ and $b'_i = b_i + \beta$. Now, consider the associated normalized matrix $Q' = (q'_{ij})$ of the system $A'X' = b'$ such that

$$\begin{aligned} q'_{ij} = \tilde{b'}_i - \tilde{a'}_{ij} &= (b' - \hat{b'}) - (a'_{ij} - \hat{A'}_j) \\ &= (b_i + \beta - \hat{b'}) - (a_{ij} + \alpha_j - \hat{A'}_j) \\ &= (b_i - \hat{b}) - (a_{ij} - \hat{A}_j) \qquad (3.1) \\ &= q_{ij}, \end{aligned}$$

for any $i \in \underline{m}$ and $j \in \underline{n}$. It should be noted that the equality (3.1) is obtained from:

$$\begin{aligned} \hat{b'} = \frac{b'_1 + \cdots + b'_m}{m} &= \frac{(b_1 + \beta) + \cdots + (b_m + \beta)}{m} \\ &= \frac{(b_1 + \cdots + b_m)}{m} + \beta \\ &= \hat{b} + \beta \end{aligned}$$

and similarly, $\hat{A'} = \hat{A}_j + \alpha_j$. This means $Q = Q'$ and consequently, the column minimum elements of Q and Q' are the same. Hence, the proof is complete. Similarly, we can prove the converse. □

3.2 Determining the column rank by normalization method

We consider the following arbitrary matrix A:

$$A = \left[\ A_1 \ \middle|\ A_2 \ \middle|\ \cdots \ \middle|\ A_n \ \right],$$

where A_j is the j-th column of A.
We check the existence of solutions of the following system by the normalization method:

$$\left[\ A_1 \ \middle|\ A_2 \ \middle|\ \cdots \ \middle|\ A_{n-1} \ \right] X = A_n. \qquad (1)$$

Here, we have two cases:

(a) If the system (1) has no solution, we conclude that A_n is an independent column of A. In this case, A_n can not be removed from the set of generators of $Col(A)$. As such, we consider the following system by setting A_n as the first column of the coefficient matrix:

$$\left[\ A_n \ \middle|\ A_1 \ \middle|\ A_2 \ \middle|\ \cdots \ \middle|\ A_{n-2}\ \right] X = A_{n-1}, \tag{2}$$

(b) If the system (1) has solutions, then A_n is dependent on the other columns of matrix A. Hence, we remove the column A_n from the set of generators of $Col(A)$, and $colrank(A) \leq n-1$. Now, we can consider the new system as follows.

$$\left[\ A_1 \ \middle|\ A_2 \ \middle|\ \cdots \ \middle|\ A_{n-2}\ \right] X = A_{n-1}, \tag{3}$$

Next, we check both cases (a) and (b) for the systems (2) or (3) depending on which one has happened. We repeat this until we get a linear system whose vector is A_1 and whose matrix is the independent columns of matrix A which are obtained from the procedure. Finally, we check both cases 1 and 2 for this last system. At this point, we can completely determine the independent columns and the column rank of A.

Remark 3.7. Note that we can obtain the row rank of A by applying the above method to the matrix A^T and finding the column rank of A^T, i.e., $rowrank(A) = colrank(A^T)$.

Example 3.8. Consider the following matrix $A \in M_{4\times 5}(S)$;

$$\begin{bmatrix} 4 & -4 & 2 & 3 & 3 \\ 5 & 7 & 7 & 2 & 6 \\ 10 & 12 & 12 & 8 & 11 \\ 4 & -3 & 2 & 3 & 3 \end{bmatrix},$$

by applying the above method, we can conclude that $colrank(A) = 2$.

4 Conclusion

In this paper, applying the normalization method to a linear system, we presented a necessary and sufficient condition for the system to have a solution. In order to determine the column rank and the row rank of an arbitrary matrix.

References

[1] J. S. Golan, *Semirings and their Applications*, Klumer Academic, Dordrecht,1999.

[2] Y. J. Tan, *Inner products on semimodules over a commutative semiring*, Linear Algebra and its Applications. 460, (2017), 151-173.

[3] H. S. Vandiver, *Note on a simple type of algebra in which the cancellation law of addition does not hold*, Bulletin of the American Mathematical Society. 40(12) (1934), 914-920.

[4] D. Wilding, *Linear algebra over semirings*, Doctoral dissertation, The University of Manchester, United Kingdom, 2015.

Is the origin included in the polynomial numerical hulls of direct sum of two Jordan blocks?[1]

Saeed Karami*

Department of Mathematics, Institute for Advanced Studies in Basic Sciences (IASBS), Zanjan, Iran

Abstract

Let $J_n(\lambda)$ be the $n \times n$ Jordan block with a positive real eigenvalue λ. In this note, we give some sufficient conditions on λ so that the origin is not included in the polynomial numerical hull of degree 2 for the matrix $J_n(\lambda) \oplus J_n(-\lambda)$.

Keywords: Jordan block, Polynomial numerical hull, Stagnation, GMRES method
Mathematics Subject Classification [2010]: 15A03, 15A23, 15B36

1 Introduction

Let $M_n(\mathbb{C})$ be the set of all $n \times n$ complex matrices. The numerical range (or field of values) of a matrix $A \in M_n(\mathbb{C})$ is a convex and compact subset of the complex plane: $W(A) = \{x^*Ax : \ x \in \mathbb{C}^n, \|x\| = 1\}$, where x^* stands for transpose of the complex conjugate of the vector x and $\|x\|$ and $\|A\|$ represent the Euclidean 2-norm of a vector x and a matrix A, respectively. By a Jordan block $J_n(\lambda)$, we mean an $n \times n$ bidiagonal upper triangular Toeplitz matrix with λ on its main diagonal and 1 on its superdiagonal. We use the notation J_n instead of $J_n(0)$. It is known that the numerical range of a Jordan block $J_n(\lambda)$ is a closed circular disk with the center at λ and the radius $r = \cos(\frac{\pi}{n+1})$ i.e. $W(J_n(\lambda)) = \mathcal{D}(\lambda, \cos(\frac{\pi}{n+1}))$ [3].

For any $1 \leq k \leq n$, the polynomial numerical hull of degree k for $A \in M_n(\mathbb{C})$ was introduced and defined by Nevanlinna [6]

$$\mathscr{H}_k(A) := \{z \in \mathbb{C} : |p(z)| \leq \|p(A)\|, \ \forall p \in \mathscr{P}_k\}, \tag{1}$$

where $\mathscr{P}_k$ denotes the set of all polynomials of degree at most k, and $\|.\|$ denotes the $2-$ matrix norm. Polynomial numerical hulls can be considered as a generalization of the numerical range; $\mathscr{H}_1(A) = W(A)$. These sets have many useful properties in the studying iterative methods such as Krylov subspace methods. Th following lemma, gives us some basic properties of the polynomial numerical hulls.

Lemma 1.1. *[2, 6] Let $A \in M_n(\mathbb{C})$. Then for any $1 \leq k \leq n$ the following properties hold:*

[1] Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University
*Speaker. Email address: s.karami@iasbs.ac.ir

1. $\sigma(A) = \mathscr{H}_n(A) \subseteq \mathscr{H}_{n-1}(A) \subseteq \cdots \subseteq \mathscr{H}_2(A) \subseteq \mathscr{H}_1(A) = W(A)$.
2. $\mathscr{H}_k(U^*AU) = \mathscr{H}_k(A)$ *for any unitary matrix* $U \in M_n(\mathbb{C})$.
3. $\mathscr{H}_k(\alpha A + \beta I) = \alpha\mathscr{H}_k(A) + \beta$ *for all* α *and* β *in the complex plane* $\mathbb{C}$.
4. *If* A *is a Hermitian matrix, then* $\mathscr{H}_2(A) = \sigma(A)$.

Polynomial numerical hulls of a Jordan block has been studied in [1]. It is shown that the polynomial numerical hull of degree k, $1 \leq k \leq n-1$, is a circular disk with a positive radius around the eigenvalue of Jordan block. In the paper, we consider the polynomial numerical hull of degree 2 for the matrix A, where $A = J_n(\lambda) \oplus J_n(\lambda)$. This subject is related to the stagnation of order 2 of the Generalized Minimal Residual (GMRES) method for solving the linear system $Ax = b$. For these types of matrices, $0 \in \mathscr{H}_m(A)$ if and only if there exists a right hand side vector b such that the GMRES method for solving the linear system $Ax = b$ (with the initial guess $x_0 = 0$) has stagnation of order m [4]. Therefore, if $0 \notin \mathscr{H}_2(A)$, then for any right hand side vector b, the GMRES method does not stagnate. This provides a motivation for our work.

2 Main results

In this section, we study the conditions on λ for which we have $0 \notin \mathscr{H}_2(A)$, where $A = J_n(\lambda) \oplus J_n(-\lambda)$ and $\lambda > 0$. Please, note that the results of this section are different of the ones stated in [5, Theorem 2.6]. According to Theorem 2.6 of [5], if $\lambda > \sqrt{2}$, then $0 \notin \mathscr{H}_2(J_n(\lambda) \oplus J_n(-\lambda))$, for any $n \geq 2$. However in this section, for any $n \geq 2$ we give a positive scalar a_n such that if $\lambda > a_n$ then $0 \notin \mathscr{H}_2(J_n(\lambda) \oplus J_n(-\lambda))$. Our method, here, is giving an including region for the numerical range of the matrices $J_n(\lambda)^k$, $k = 2, \ldots$ independent of [5, Remark 4].

At the first, by using an orthogonal transformation, we determine the Jordan canonical form of the matrix J_n^k. Actually, for any $2 \leq k \leq n-1$, we use a permutation to gathering Jordan sub-blocks of J_n^k. This permutation is determined in terms of remainders of division of $n-1$ by k. Note that, for any $k \geq n$, $J_n^k = 0$ and for any $1 \leq k \leq n-1$, the matrix J_n^k is an upper triangular matrix with 1's on its k^{th} superdiagonal and 0 in other places.

Lemma 2.1. *Let* $n \in \mathbb{N}$. *Then for any* $1 \leq k \leq n-1$, *the matrix* J_n^k *is orthogonal similar to the matrix*

$$\underbrace{J_{m_1} \oplus \ldots \oplus J_{m_1}}_{r_1 \, times} \oplus \underbrace{J_{m_2} \oplus \ldots \oplus J_{m_2}}_{r_2 \, times},$$

where $m_1 = [\frac{n-1}{k}] + 1, m_2 = [\frac{n-1}{k}]$, $r_1 = ((n-1) \mod k) + 1$ *and* $r_2 = k - r_1$.

Example 2.2. The matrix J_{16}^3 is orthogonal similar to the matrix $J_6 \oplus J_5 \oplus J_5$ and the matrix J_{16}^{14} is orthogonal similar to $J_2 \oplus J_2 \oplus 0_{12}$. Also the matrix J_{21}^8 is orthogonal similar to the matrix $J_3 \oplus J_3 \oplus J_3 \oplus J_3 \oplus J_3 \oplus J_2 \oplus J_2 \oplus J_2$.

Next, we turn to our main result in this section. For this aim, we recall that for any two matrices A and B, $W(A \oplus B) = \text{Conv}(W(A) \cup W(B))$ [3]. Since, $m_1 > m_2$ we get $W(J_{m_1}) \supseteq W(J_{m_2})$. Therefore $W(J_n^k) = W(J_{m_1}) = \mathcal{D}(0, \cos(\frac{\pi}{m_1+1})) = \mathcal{D}(0, \cos(\frac{\pi}{[\frac{n-1}{k}]+2}))$.

Now, we consider polynomial numerical hulls of the matrix $J_n(a) \oplus J_n(b)$, where $a, b \in \mathbb{C}$. By rotation and translation we need only to study these sets for the matrices of the form $J_n(\lambda) \oplus J_n(-\lambda)$, where $\lambda \in \mathbb{R}$, (see Lemma 1.1). We remember that for any $1 \leq k \leq n-1$,

$\mathscr{H}_k(J_n(\lambda)) = D(\lambda, r_{k,n})$, the circle disk with the center λ, and radius $r_{k,n}$ [1, 2]. It is known that $0 < r_{n-1,n} \leq r_{n-2,n} \leq \cdots \leq r_{1,n} = \cos(\frac{\pi}{n+1}) < 1$. Since $J_n(\lambda)$ and $J_n(-\lambda)$ are principle sub-matrices of A, we obtain that

$$D(\lambda, r_{k,n}) \cup D(-\lambda, r_{k,n}) \subseteq \mathscr{H}_k\left(J_n(\lambda) \oplus J_n(-\lambda)\right), \quad k = 2, \ldots, n-1. \tag{2}$$

Remark 2.3. By using the fact that the matrix $A = J_n(\lambda) \oplus J_n(-\lambda)$ is unitary equivalent to the matrix $\tilde{A} = J_n(\lambda) \oplus -J_n(\lambda)$, we obtain that the sets $\mathscr{H}_k(A), k = 1, \ldots, 2n$ are symmetric with respect to x and y axises.

Lemma 2.4. *Let $A = J_2(\lambda) \oplus J_2(-\lambda)$. Then the following statements are equivalent:*

(i) $|\lambda| \leq 1$,

(ii) $0 \in \mathscr{H}_2(A)$,

(iii) $0 \in \mathscr{H}_3(A)$.

Proof. The equivalence of (i) and (iii) is due to [5, Theorem 2.1] and we need to prove the equivalence of (i) and (ii). If $0 \in \mathscr{H}_2(A)$, then

$$\begin{aligned} 0 \in W(A^2) &= F\big(J_2(\lambda)^2 \oplus J_2(-\lambda)^2\big) \\ &= F(J_2(\lambda)^2) \subseteq D(\lambda^2, |\lambda|). \end{aligned} \tag{3}$$

Thus $|\lambda| \leq 1$. For converse assume that $|\lambda| \leq 1$ and let $X = (x_1, \ldots, x_4)^T \in \mathbb{C}^4$, where $x_1 = x_3 = \frac{\sqrt{1+\sqrt{1-|\lambda|^2}}}{2}$ and $x_2 = -x_4 = \frac{-\lambda}{4x_1}$. It is readily seen that $\|X\| = 1$ and $X^*AX = X^*A^2X = 0$. Therefore $(0,0) \in W(A, A^2)$ and hence $0 \in \mathscr{H}_2(A)$. □

Now, we study polynomial numerical hull of degree 2 for general n. Note that as we say in the introduction analytical computing of polynomial numerical hull for general n represent a very difficult problem which leads to computing zeros of an complex polynomial of order $2n$. However, in the following we give an analytical conditions on λ such that origin lie in the polynomial numerical hull of degree 2.

Theorem 2.5. *Let $A = J_n(\lambda) \oplus J_n(-\lambda)$ where $\lambda \in (0, +\infty)$ and $n \geq 3$. Let $m = [\frac{n+1}{2}]$ and $r_1 = \cos(\frac{\pi}{n+1})$,$r_2 = \cos(\frac{\pi}{m+1})$. Then:*

i) If $0 \leq \lambda \leq 1$, then $0 \in \mathscr{H}_2(A)$. Moreover $[-\sqrt{\lambda^2+\lambda}, \sqrt{\lambda^2+\lambda}] \subseteq \mathscr{H}_2(A)$.

ii) If $\lambda > r_1 + \sqrt{r_1^2 + r_2}$, then $\mathscr{H}_2(A) \cap \{z : |z| < \alpha\} = \emptyset$, where $\alpha = (\lambda^2 - 2\lambda r_1 - r_2)^{\frac{1}{2}}$ and therefore $0 \notin \mathscr{H}_2(A)$.

Proof. Since the matrix $J_2(\lambda) \oplus J_2(-\lambda)$ is a principle sub matrix of the matrix A, the assertion in (i) follows by using Lemma 2.4. Now, we consider (ii). An observation shows that $A^2 = B \oplus C$, where $B = J_n(\lambda)^2 = \lambda^2 I_n + 2\lambda J_n + J_n^2$ and $C = \lambda^2 I_n - 2\lambda J_n + J_n^2$. The matrix C is unitary similar to the matrix B. Therefore $W(A^2) = W(B) = W(J_n(\lambda)^2)$. So

$$\begin{aligned} W(A^2) &= W(\lambda^2 + 2\lambda J_l + J_l^2) \\ &\subseteq \lambda^2 + 2\lambda \mathcal{D}(0, r_1) + \mathcal{D}(0, r_2) = \mathcal{D}(\lambda^2, 2\lambda r_1 + r_2). \end{aligned} \tag{4}$$

Since $\lambda > r_1 + \sqrt{r_1^2 + r_2}$, we have $\lambda^2 - 2\lambda r_1 - r_2 > 0$. If $|z| < \alpha$, then $|z^2 - \lambda^2| \geq \lambda^2 - |z|^2 > 2\lambda r_1 + r_2$ and hence $z^2 \notin W(A^2)$. Therefore $z \notin \mathscr{H}_2(A)$ and the proof is completed. □

Remark 2.6. Let $A = J_{n_1}(\lambda) \oplus J_{n_2}(-\lambda)$. By choosing $n = \max\{n_1, n_2\}$ and letting $\hat{A} = J_n(\lambda) \oplus J_n(-\lambda)$, we know that A is a principle sub-matrix of $\hat{A}$. So if $0 \notin \mathscr{H}_2(\hat{A})$, then $0 \notin \mathscr{H}_2(A)$. Hence we can use Theorem 2.5 for $\hat{A}$ instead of A.

Example 2.7. Let $A = J_4(2) \oplus J_3(-2)$. Then $n = 4$ and by Theorem 2.5, we obtain that $\mathscr{H}_2(A) \cap \{z \in \mathbb{C} : |z| < \alpha\} = \emptyset$, where $\alpha = \sqrt{4 - 4\cos\frac{\pi}{5} - \cos\frac{\pi}{3}} = 0.5137$. Therefore $0 \notin \mathscr{H}_2(A)$. Also note that by equation (2), $D(2, r_{2,4}) \cup D(-2, r_{2,3}) = D(2, 0.7) \cup D(-2, 0.6) \subseteq \mathscr{H}_2(A)$. In Figure 1, we plot the set $\mathscr{H}_2(A)$ together with the circle $\{z : |z| = 0.5137\}$(which is shown by dashed curve).

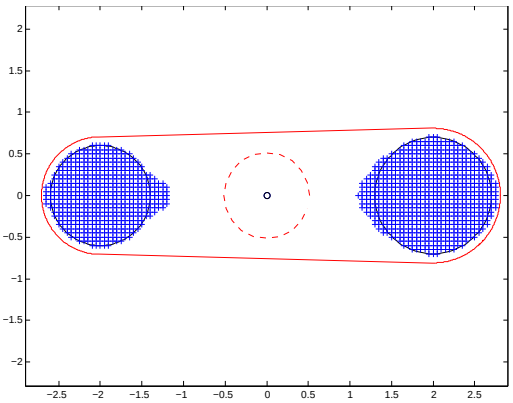


Figure 1: Polynomial numerical hull of degree 2

Now, let $A = J_{n_1}(\lambda_1) \oplus J_{n_2}(\lambda_2)$, where $\lambda_i \in \mathbb{R}, n_i \geq 2, i = 1, 2$ and without loss of generality assume that $\lambda_2 < \lambda_1$. Here we investigate when $0 \in \mathscr{H}_2(A)$. If $\lambda_1 < -\cos(\frac{\pi}{n_1+1})$ or $\lambda_2 > \cos(\frac{\pi}{n_2+1})$, then $0 \notin W(A) = \mathscr{H}_1(A)$ and so $0 \notin \mathscr{H}_2(A)$. Also if $|\lambda_i| \leq r_{2,n_i}$ for $i = 1$ or $i = 2$, then $0 \in \mathscr{H}_2(A)$, where $r_{2,n}$ denotes the radius of the circular disk $\mathscr{H}_2(J_n)$ (see eqaution (2)). The crucial case is when $\lambda_2 < 0 < \lambda_1$. By using Theorem 2.5 and translation property of polynomial numerical hulls, in the following we give some statements for this case, which help us to investigate whether $0 \in \mathscr{H}_2(A)$.

Theorem 2.8. *Let $A = J_n(\lambda_1) \oplus J_m(\lambda_2)$ where $\lambda_1, \lambda_2 \in \mathbb{R}$ and $\lambda_1 \geq \lambda_2$. Let $l = \max\{n, m\}, m_l = [\frac{l+1}{2}]$, $r_1 = \cos(\frac{\pi}{l+1}), r_2 = \cos(\frac{\pi}{m_l+1})$, $\overline{\lambda} = \frac{\lambda_1+\lambda_2}{2}$ and $d = \frac{\lambda_1-\lambda_2}{2}$. If $d > r_1 + \sqrt{r_1^2 + r_2}$, then $\mathscr{H}_2(A) \cap \{z \in \mathbb{C} : |z - \overline{\lambda}| < \alpha\} = \emptyset$ where $\alpha = \sqrt{d^2 - 2dr_1 - r_2}$.*

Example 2.9. Let $A = J_3(5) \oplus J_2(-1)$. It is readily seen that $0 \in \mathscr{H}_1(A) = W(A)$. By the previous theorem notations we have: $d = 3, \overline{\lambda} = 2, r_1 = 0.707, r_2 = 0.5$ and $\alpha = 2.0635$. Thus $\mathrm{Re}(\mathscr{H}_2(A)) \cap (-0.0635, 4.0635) = \emptyset$. In particular we obtain that $0 \notin \mathscr{H}_2(A)$(see Figure 2).

3 Conclusion

For the matrix $A = J_n(\lambda) \oplus J_n(-\lambda)$ the origin lies between λ and $-\lambda$. Thus $0 \in W(A)$, the numerical range of A. In this note, we gave some sufficient conditions on the positive scalar λ such that the origin is not included in the polynomial numerical hull of degree 2 of the matrix A.

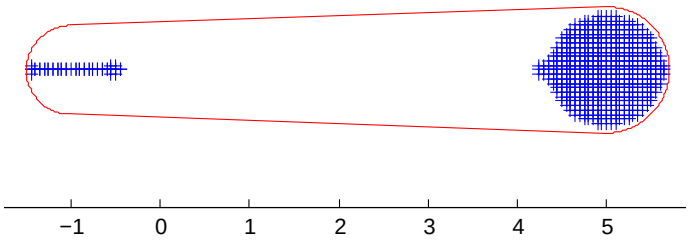


Figure 2

References

[1] V. Faber, A. Greenbaum and D. E. Marshall, The polynomial numerical hulls of Jordan blocks and related matrices, *Linear Algebra Appl.*, 374 (2003), 231–246.

[2] A. Greenbaum, Generalizations of the field of values useful in the study of polynomial functions of a matrix, *Linear Algebra Appl.*, 347 (2002), 233–249.

[3] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, 1994.

[4] K. Jubilou and H. Sadok, Analysis of some vector extrapolation methods for solving systems of linear equations, *Numer. Math.*, 70 (1995), 73–89.

[5] S. Karami and A. Salemi, Polynomial numerical hulls of the direct sum of two Jordan blocks, *Linear Algebra Appl.*, 585 (2020), 209–226.

[6] O. Nevalinna, *Convergence of iterations for linear equation*, Basel: Birkhauser, 1993.

Majorization relation and linear preservers[1]

Fatemeh Khalooei*

Department of Pure Mathematics, Faculty of Mathematics and Computer,
Shahid Bahonar University of Kerman, Kerman, Iran

Abstract

In this article we define and study a majorization relation on $\mathbb{R}^n$, also we study its linear preservers. Let $x,\ y$ be in $\mathbb{R}^n$, we say x is majorized by y and write $x \prec y$ when $x = Ty$ for some t-transform T. We say a linear transformation f is a linear preserver of $\prec_t$ if $x \prec_t y$ implies that $f(x) \prec_t f(y)$.

Keywords: Majorization, T-transform, Linear preserver

Mathematics Subject Classification [2010]: 15A04, 15A21, 15A51

1 Introduction

We call a linear map T on $\mathbb{R}^n$ a t-transform if there exists $0 \leq t \leq 1$ and indices $1 \leq j,\ k \leq n$ such that

$$Ty = (y_1, \ldots, y_{j-1}, ty_j + (1-t)y_k, y_{j+1}, \ldots, (1-t)y_j + y_k, y_{k+1}, \ldots, y_n),$$

for all $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$.
If $[T]$ is the matrix representation of a t-transform T with respect to the stundard basis of $\mathbb{R}^n$ then

$$[T] = \begin{pmatrix} 1 & & & & & & & 0 \\ & \ddots & & & & & & \\ & & 1-t & & t & & & \\ & & & 1 & & & & \\ & & & & \ddots & & & \\ & & t & & 1-t & & & \\ & & & & & & \ddots & \\ 0 & & & & & & & 1 \end{pmatrix}$$

Also we can write $[T] = tI + (1-t)Q$, where I is the $n \times n$ identity matrix and Q is a permutation matrix that $Qe_j = e_k$, $Qe_k = e_j$ and $Qe_i = e_i$ for all $i \neq j,\ k$, where $\{e_1, \ldots, e_n\}$ is the standard basis on $\mathbb{R}^n$. It is trivial that a t-transform is singular if and only if $t = 1/2$.

[1] Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

*Speaker. Email address: f_khalooei@uk.ac.ir

If x and y are nonincreasing vectors in $\mathbb{R}^n$ such that $\sum_{i=1}^{k} x_i \leq \sum_{i=1}^{k} y_i$ for $k = 1, \ldots, n$ with equality for $k = n$, then we say that x is multivariate majorized by y and write $x \prec y$. A linear operator T on $\mathbb{R}^n$ is said to be a linear preserver of a given relation $\prec$ on $\mathbb{R}^n$ if $x \prec y$ implies that $Tx \prec Ty$. For more information about other majorization and their linear preservers we refer the reader to [3], [4] and [5].
An $n \times n$ matrix $D = [d_{ij}]$ is called doubly stochastic if $d_{ij} \geq 0$, $\sum_{k=1}^{n} d_{ik}$ and $\sum_{k=1}^{n} d_{kj}$ are equal to 1 for all $i,\ j$. The set of doubly stochastic matrices is denoted by $\mathcal{DS}(n)$. Also we can describe doubly stochastic matrices by

$$\mathcal{DS} = \{D \in M_n : D \geq 0,\ De = e,\ D^t e = e\},$$

where $e \in \mathbb{R}^n$ is the vector whose components are all $+1$.

Theorem 1.1. *[2, Birkhoff's Theorem] The set of $n \times n$ doubly stochastic matrices is a convex set whose extreme points are the permutation matrices.*

Theorem 1.2. *[2]For $x,\ y \in \mathbb{R}^n$, the following statements are equivalent*

1. *$x \prec y$,*
2. *x is obtained from y by a finite number of t-transforms,*
3. *$x = Dy$ for some doubly stochastic matrix D.*

2 Main results

In this section we define a majorization relation on $\mathbb{R}^n$ and study some of its properties. Also by an example we show that $\prec$ dose not imply it.

Definition 2.1. For $x,\ y \in \mathbb{R}^n$ we say x is t-majorized by y and write $x \prec_t y$ when $x = Ty$ for some t-transform T.

Corollary 2.2. *On $\mathbb{R}^2$ the following statemants are true*

1. *A 2×2 matrix is doubly stochastic if and only if it is a t-transform.*
2. *If $x,\ y \in \mathbb{R}^2$ then $x \prec y$ if and only if $x \prec_t y$.*

Theorem 2.3. *For $x,\ y \in \mathbb{R}^n$*

1. *if $n \geq 2$, $x \prec_t y$ and $y \prec_t x$ if and only if $x = Py$ for some $n \times n$ permutation matrix P, which is the identity matrix or a permutation matrix that just interchanges two coordinates.*
2. *for $n = 2$, $x \prec_t y$ and $y \prec_t x$ if and only if $x = Py$ for some 2×2 permutation matrix P.*

Example 2.4. Multivariate majorization does not imply t-majorization on $\mathbb{R}^n$. $n \geq 3$. $(3, 2.5, 1.5) \prec (4, 2, 1)$ but $(3, 2.5, 1.5) \not\prec_t (4, 2, 1)$.

Theorem 2.5. *[1] A linear map $T : \mathbb{R}^n \to \mathbb{R}^n$ preserves $\prec$ if and only if one of the following holds*

1. *$Tx = (trx)a$ for some $a \in \mathbb{R}^n$,*
2. *$Tx = \alpha Px + \beta Jx$, for some $\alpha,\ \beta \in \mathbb{R}$ and $n \times n$ permutation matrix P.*

Theorem 2.6. *If $T : \mathbb{R}^n \to \mathbb{R}^n$ has the form $Tx = (trx)a$ for some $a \in \mathbb{R}^n$ or $Tx = \alpha Px + \beta Jx$, for some $\alpha,\ \beta \in \mathbb{R}$ and $n \times n$ permutation matrix P, then T is a linear preserver of $\prec_t$.*

References

[1] T. Ando, Majorization, Doubly stochastic matrices, and comparison of eigenvalues, *Linear Algebra and its Applications*, 118 (1989), 163–248.

[2] R. Bhatia, *Matrix Analysis*, Springer-Verlage, New York, 1997.

[3] F. Khalooei and A. Salemi, Linear preservers of majorization, *Iranian Journal of Mathematical Sciences and informatics*, 6, No. 2 (2011), 43-50.

[4] F. Khalooei, Linear maps which preserve or strongly preserve majorization on matrices, *Bulletin of the Iranian Mathematical Society*,41, No. 7 (2015), 77-83.

[5] F. Khalooei, Linear preservers of two sided matrix majorization, *Wavelets and Linear Algebra*, 1(2014), 33-38.

On the preconditioning of generalized saddle point problems using symmetric and skew-symmetric iteration method[1]

Davod Khojasteh Salkuyeh*

Faculty of Mathematical Sciences, University of Guilan, Rasht, Iran

Abstract

In the implementation of the symmetric and skew-symmetric splitting preconditioner in a Krylov subspace method for generalized saddle point problems, a shifted skew-symmetric system should be solved. In this paper, we propose an efficient iterative method for solving this system and investigate its convergence properties. Numerical results are given to show the efficiency of the method.

Keywords: SSS, Iterative method, Preconditioner, Skew-symmetric, Generalized saddle point

Mathematics Subject Classification [2010]: 65F10, 65F50

1 Introduction

Benzi and Golub in [2] proposed using the symmetric and skew-symmetric (SSS) iteration method for solving the generalized saddle point problems

$$\mathcal{A}\mathbf{x} = \begin{pmatrix} A & B^T \\ -B & C \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix} \equiv \mathbf{b}, \tag{1}$$

where $A \in \mathbb{R}^{n\times n}$, $C \in \mathbb{R}^{m\times m}$ and $B \in \mathbb{R}^{m\times n}$, $f \in \mathbb{R}^n$, $g \in \mathbb{R}^m$ and $m \leq n$. According to the SSS iteration, the matrix $\mathcal{A}$ is split as $\mathcal{A} = \mathcal{H} + \mathcal{S}$, where

$$\mathcal{H} = \frac{1}{2}\left(\mathcal{A} + \mathcal{A}^T\right) = \begin{pmatrix} H & 0 \\ 0 & C \end{pmatrix} \quad \text{and} \quad \mathcal{S} = \frac{1}{2}\left(\mathcal{A} - \mathcal{A}^T\right) = \begin{pmatrix} S & B^T \\ -B & 0 \end{pmatrix},$$

in which $H = (A + A^T)/2$ and $S = (A - A^T)/2$. For $\alpha > 0$, both of the matrices $\alpha\mathcal{I} + \mathcal{H}$ and $\alpha\mathcal{I} + \mathcal{S}$ are nonsingular, where $\mathcal{I}$ is the identity matrix of order $m + n$. In this case, the SSS iteration method for the saddle point problem (1) is written as

$$\begin{cases} (\alpha\mathcal{I} + \mathcal{H})\,\mathbf{x}^{k+\frac{1}{2}} &= (\alpha\mathcal{I} - \mathcal{S})\,\mathbf{x}^k + \mathbf{b}, \\ (\alpha\mathcal{I} + \mathcal{S})\,\mathbf{x}^{k+1} &= (\alpha\mathcal{I} - \mathcal{H})\,\mathbf{x}^{k+\frac{1}{2}} + \mathbf{b}, \end{cases} \tag{2}$$

where $\mathbf{x}^0$ is an initial guess. Computing $\mathbf{x}^{k+\frac{1}{2}}$ from the first equation and substituting it in the second one, gives the iteration $\mathbf{x}^{k+1} = \mathcal{T}_\alpha \mathbf{x}^k + \mathbf{c}$, where

$$\mathcal{T}_\alpha = (\alpha\mathcal{I} + \mathcal{S})^{-1}(\alpha\mathcal{I} - \mathcal{H})(\alpha\mathcal{I} + \mathcal{H})^{-1}(\alpha\mathcal{I} - \mathcal{S}),$$

[1] Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

* Speaker. Email address: khojasteh@guilan.ac.ir

and $\mathbf{c} = 2\alpha\,(\alpha\mathcal{I}+\mathcal{S})^{-1}\,(\alpha\mathcal{I}+\mathcal{H})^{-1}\,\mathbf{b}$. In [2], it was shown that if A is positive real ($x^TAx > 0$, for every nonzero vector $x \in \mathbb{R}^n$), C is symmetric positive semidefinite and B has full rank, then the iteration (2) is unconditionally convergent.

As the authors of [2] mentioned the iteration method (2) is typically show for the method to be competitive and proposed using a nonsymmetric Krylov subspace method like GMRES or its restarted version GMRES(m) [6] in conjunction with the SSS preconditioner induced by the iteration method. It is known that there is a unique splitting $\mathcal{A} = \mathcal{M}_\alpha - \mathcal{N}_\alpha$, with $\mathcal{M}_\alpha$ being nonsingular and $\mathcal{T}_\alpha = \mathcal{M}_\alpha^{-1}\mathcal{N}_\alpha = \mathcal{I} - \mathcal{M}_\alpha^{-1}\mathcal{A}$, where

$$\mathcal{M}_\alpha = \frac{1}{2\alpha}\,(\alpha\mathcal{I}+\mathcal{H})\,(\alpha\mathcal{I}+\mathcal{S})\,, \quad \mathcal{N}_\alpha = \frac{1}{2\alpha}\,(\alpha\mathcal{I}-\mathcal{H})\,(\alpha\mathcal{I}-\mathcal{S})\,. \tag{3}$$

If the SSS method is convergent then the eigenvalues of $\mathcal{A}\mathcal{M}_\alpha^{-1}$ are included in the unit circle centered at $(1,0)$. Hence, it is expected that a Krylov subspace method like GMRES or its restarted version will be suitable for solving the right-preconditioned system $\mathcal{A}\mathcal{M}_\alpha^{-1}\mathbf{y} = \mathbf{b}$ with $\mathbf{x} = \mathcal{M}_\alpha^{-1}\mathbf{y}$. The pre-factor $\frac{1}{2\alpha}$ in $\mathcal{M}_\alpha$ has no effect on the preconditioned system, hence it can be omitted and the matrix $\mathcal{M}_\alpha = (\alpha\mathcal{I}+\mathcal{H})\,(\alpha\mathcal{I}+\mathcal{S})$ can be used as a preconditioner.

Application of the preconditioner $\mathcal{M}_\alpha$ within the GMRES method requires solving linear systems of the form $\mathcal{M}_\alpha\mathbf{z} = \mathbf{r}$ which can be done by first solving $(\alpha\mathcal{I}+\mathcal{H})\mathbf{v} = \mathbf{r}$, for $\mathbf{v}$ and then $(\alpha\mathcal{I}+\mathcal{S})\mathbf{z} = \mathbf{v}$. System $(\alpha\mathcal{I}+\mathcal{H})\mathbf{v} = \mathbf{r}$ can be reduced to two sub-systems of the form $(\alpha I_n + H)v_1 = r_1$ and $(\alpha I_m + C)v_2 = r_2$, where $v_1, r_1 \in \mathbb{R}^n$, $v_2, r_2 \in \mathbb{R}^m$, $v = (v_1; v_2)$ and $r = (r_1; r_2)$. Here, I_r denotes the identity matrix of order r. Obviously, the coefficient matrices of these systems are SPD. Hence, they can be solved exactly using the Cholesky factorization or inexactly by the conjugate gradient (CG) method [6]. However, solving Eq. $(\alpha\mathcal{I}+\mathcal{S})\mathbf{z} = \mathbf{v}$ is not trivial. This system can be equivalently rewritten as

$$\mathcal{R}\mathbf{z} = \begin{pmatrix} \alpha I_n + S & B^T \\ -B & \alpha I_m \end{pmatrix}\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \mathbf{v}, \tag{4}$$

where $z_1 \in \mathbb{R}^n$, $z_2 \in \mathbb{R}^m$ and $z = (z_1; z_2)$. As the authors of [2] suggested, the vector z can be computed by first solving

$$\left(\alpha^2 I_m + B(I_n + \frac{1}{\alpha}S)^{-1}B^T\right) z_2 = B(I_n + \frac{1}{\alpha}S)^{-1}v_1 + \alpha v_2, \tag{5}$$

for z_2, followed by $(\alpha I_n + S)z_1 = v_1 - B^T v_2$. Both of these systems can be solved directly using the LU factorization or inexactly using a Krylov subspace matrix like GMRES.

In this paper, we present an efficient iterative method for solving the system (4) which is unconditionally convergent. So, in the implementation of the SSS preconditioner within a Krylov subspace method we can apply the proposed iteration method.

2 The new method

We split the coefficient matrix of Eq. (4) as $\mathcal{R} = \mathcal{R}_1 + \mathcal{R}_2$, where

$$\mathcal{R}_1 = \begin{pmatrix} \alpha I_n + S & 0 \\ 0 & \alpha I_m \end{pmatrix} \quad \text{and} \quad \mathcal{R}_2 = \begin{pmatrix} 0 & B^T \\ -B & 0 \end{pmatrix},$$

Obviously, $\mathcal{R}_1$ and $\mathcal{R}_2$ are shifted skew-symmetric and skew-symmetric matrices, respectively. In fact, Eq. $\mathcal{R} = \mathcal{R}_1 + \mathcal{R}_2$ presents a shifted skew-symmetric and skew-symmetric splitting (SSSS). For $\beta > 0$, using the splittings

$$\mathcal{R} = (\beta\mathcal{I} + \mathcal{R}_1) - (\beta\mathcal{I} - \mathcal{R}_2) = (\beta\mathcal{I} + \mathcal{R}_2) - (\beta\mathcal{I} - \mathcal{R}_1),$$

we propose the SSSS iteration method for solving Eq. (4) as following

$$\begin{cases} (\beta\mathcal{I}+\mathcal{R}_1)\,\mathbf{z}^{k+\frac{1}{2}} &= (\beta\mathcal{I}-\mathcal{R}_2)\,\mathbf{z}^{k}+\mathbf{v}, \\ (\beta\mathcal{I}+\mathcal{R}_2)\,\mathbf{z}^{k+1} &= (\beta\mathcal{I}-\mathcal{R}_1)\,\mathbf{z}^{k+\frac{1}{2}}+\mathbf{v}. \end{cases} \tag{6}$$

The SSSS iteration method can be written as $\mathbf{z}^{k+1}=\mathcal{G}_{\alpha,\beta}\mathbf{z}^{k}+\mathbf{d}$, where

$$\mathcal{G}_{\alpha,\beta}=(\beta\mathcal{I}+\mathcal{R}_2)^{-1}(\beta\mathcal{I}-\mathcal{R}_1)(\beta\mathcal{I}+\mathcal{R}_1)^{-1}(\beta\mathcal{I}-\mathcal{R}_2),$$

and $\mathbf{d}=2\beta\,(\beta\mathcal{I}+\mathcal{R}_2)^{-1}(\beta\mathcal{I}+\mathcal{R}_1)^{-1}\mathbf{v}$. On the other hand, we have $\mathcal{R}=\mathcal{P}_{\alpha,\beta}-\mathcal{Q}_{\alpha,\beta}$, where

$$\mathcal{P}_{\alpha,\beta}=\frac{1}{2\beta}(\beta\mathcal{I}+\mathcal{R}_2)(\beta\mathcal{I}+\mathcal{R}_1),\quad \mathcal{Q}_{\alpha,\beta}=\frac{1}{2\beta}(\beta\mathcal{I}-\mathcal{R}_2)(\beta\mathcal{I}-\mathcal{R}_1).$$

Therefore, $\mathcal{P}_{\alpha,\beta}$ can be used as a preconditioner for the system (4). We now state the convergence of the SSSS iteration method for solving (4).

Theorem 2.1. *Let $S\in\mathbb{R}^{n\times n}$ be skew-symmetric matrix and $B\in\mathbb{R}^{m\times n}$. Then, the SSSS iteration method is unconditionally convergent, i.e., $\rho(\mathcal{G}_{\alpha,\beta})<1$ for all $\alpha,\beta>0$.*

Proof. Evidently, the matrix $\mathcal{G}_{\alpha,\beta}$ is similar to

$$\tilde{\mathcal{G}}_{\alpha,\beta}=(\beta\mathcal{I}-\mathcal{R}_1)(\beta\mathcal{I}+\mathcal{R}_1)^{-1}(\beta\mathcal{I}-\mathcal{R}_2)(\beta\mathcal{I}+\mathcal{R}_2)^{-1}=\mathcal{U}\mathcal{V},$$

where $\mathcal{U}=(\beta\mathcal{I}-\mathcal{R}_1)(\beta\mathcal{I}+\mathcal{R}_1)^{-1}$ and $\mathcal{V}=(\beta\mathcal{I}-\mathcal{R}_2)(\beta\mathcal{I}+\mathcal{R}_2)^{-1}$. Since $\mathcal{R}_2$ is a skew-symmetric matrix, we deduce that the matrix $\mathcal{V}$ is orthogonal and as a result we have $\|\mathcal{V}\|_2=1$ (see [4, p. 68]). On the other hand, we have

$$\mathcal{U}=((\beta-\alpha)\mathcal{I}-\mathcal{J})((\beta+\alpha)\mathcal{I}+\mathcal{J})^{-1}, \tag{7}$$

where $\mathcal{J}=\mathrm{bldiag}(S,0)$. Clearly, the matrix $\mathcal{J}$ is skew-symmetric and there is an orthogonal matrix $\mathcal{W}$, such that $\mathcal{J}=\mathcal{W}\mathcal{D}\mathcal{W}^T$, where $\mathcal{D}=\mathrm{bldiag}(D,0)$ with $D=\mathrm{diag}(\lambda_1,\ldots,\lambda_n)$ and $\lambda_i\in\sigma(S)$, $i=1,\ldots,n$. It is well-known that the eigenvalues of the matrix S can be written as $\lambda_i=\mathrm{i}\mu_i$, where $\mu_i\in\mathbb{R}$, $i=1,2,\ldots,n$ and $\mathrm{i}=\sqrt{-1}$ (see [5, p. 101]). Therefore, it follows from (7) that

$$\mathcal{U}=\mathcal{W}((\beta-\alpha)\mathcal{I}-\mathcal{D})((\beta+\alpha)\mathcal{I}+\mathcal{D})^{-1}\mathcal{W}^T. \tag{8}$$

Hence,

$$\begin{aligned}
\rho(\mathcal{G}_{\alpha,\beta}) &= \rho(\tilde{\mathcal{G}}_{\alpha,\beta})\le\|\mathcal{U}\mathcal{V}\|_2\le\|\mathcal{U}\|_2\|\mathcal{V}\|_2=\|\mathcal{U}\|_2 \\
&= \left\|\mathcal{W}((\beta-\alpha)\mathcal{I}-\mathcal{D})((\beta+\alpha)\mathcal{I}+\mathcal{D})^{-1}\mathcal{W}^T\right\|_2 \\
&= \left\|((\beta-\alpha)\mathcal{I}-\mathcal{D})((\beta+\alpha)\mathcal{I}+\mathcal{D})^{-1}\right\|_2 \\
&= \left\|\begin{pmatrix}((\beta-\alpha)I_n-D)((\beta+\alpha)I_n+D)^{-1} & 0\\ 0 & \frac{\beta-\alpha}{\beta+\alpha}I_m\end{pmatrix}\right\|_2 \\
&= \left\|\mathrm{diag}\left(\frac{(\beta-\alpha)-\mathrm{i}\mu_1}{(\beta+\alpha)+\mathrm{i}\mu_1},\ldots,\frac{(\beta-\alpha)-\mathrm{i}\mu_n}{(\beta+\alpha)+\mathrm{i}\mu_n},\frac{\beta-\alpha}{\beta+\alpha},\ldots,\frac{\beta-\alpha}{\beta+\alpha}\right)\right\|_2 \\
&= \max\left\{\left|\frac{\beta-\alpha}{\beta+\alpha}\right|,\sqrt{\frac{(\beta-\alpha)^2+\mu_i^2}{(\beta+\alpha)^2+\mu_i^2}}:i=1,\ldots,n\right\} \\
&= \sqrt{\frac{(\beta-\alpha)^2+\rho(S)^2}{(\beta+\alpha)^2+\rho(S)^2}}=:\delta_{\alpha,\beta}.
\end{aligned}$$

Obviously, $\delta_{\alpha,\beta}<1$, for all $\alpha,\beta>0$, which proves the convergence of the SSSS iteration method. □

Theorem 2.2. *Under the assumptions of Theorem 2.1 and for a fixed value of α, we have*

$$\beta^* = \operatorname*{argmin}_{\beta} \delta_{\alpha,\beta} = \sqrt{\alpha^2 + \rho(S)^2}.$$

Proof. Letting $g(\beta) = \delta_{\alpha,\beta}^2$, we get

$$g'(\beta) = \frac{4\alpha\left(\beta^2 - (\alpha^2 + \rho(S)^2)\right)}{\left((\beta+\alpha)^2 + \rho(S)^2\right)^2}.$$

Therefore, the minimizer of $\delta_{\alpha,\beta}$ is given by $\beta^* = \sqrt{\alpha^2 + \rho(S)^2}$. □

3 Inexact version of SSSS and its implementation issues

To compute $\mathbf{z}^{k+1}$ from (6), we need to solve two subsystems with the coefficient matrices $\beta\mathcal{I} + \mathcal{R}_1$ and $\beta\mathcal{I} + \mathcal{R}_2$, which are very costly. To improve the implementation of the SSSS iteration method, we can employ iteration methods for solving the two subsystems. This, results in the inexact version of the SSSS (ISSSS) iteration method.

Let $\gamma^k = \mathbf{z}^{k+\frac{1}{2}} - \mathbf{z}^k$. In this case, we have $\mathbf{z}^{k+\frac{1}{2}} = \mathbf{z}^k + \gamma^k$. Substituting $\mathbf{z}^{k+\frac{1}{2}}$ in the first relation in Eq. (6), gives

$$(\beta\mathcal{I} + \mathcal{R}_1)\,\gamma^k = \mathbf{v} - (\mathcal{R}_1 + \mathcal{R}_2)\,\mathbf{z}^k = \mathbf{v} - \mathcal{R}\mathbf{z}^k =: \mathbf{r}^k. \tag{9}$$

This system is equivalent to

$$\begin{pmatrix} (\beta+\alpha)I_n + S & 0 \\ 0 & (\beta+\alpha)I_m \end{pmatrix} \begin{pmatrix} \gamma_1^k \\ \gamma_2^k \end{pmatrix} = \begin{pmatrix} \mathbf{r}_1^k \\ \mathbf{r}_2^k \end{pmatrix},$$

where $\gamma_1^k, \mathbf{r}_1^k \in \mathbb{R}^n$ and $\gamma_2^k, \mathbf{r}_2^k \in \mathbb{R}^m$. To solve the above system we first solve the system

$$((\beta+\alpha)I_n + S)\,\gamma_1^k = \mathbf{r}_1^k, \tag{10}$$

for computing γ_1^k, using a Krylov subspace method like GMRES or its restarted version GMRES(ℓ). Then the vector γ_2^k is simply computed via $\gamma_2^k = \mathbf{r}_2^k/(\beta+\alpha)$.

Similarly, by setting $\gamma^{k+\frac{1}{2}} = \mathbf{z}^{k+1} - \mathbf{z}^{k+\frac{1}{2}}$, from the second equation in (6) we get

$$(\beta\mathcal{I} + \mathcal{R}_2)\,\gamma^{k+\frac{1}{2}} = \mathbf{v} - (\mathcal{R}_1 + \mathcal{R}_2)\,\mathbf{z}^{k+\frac{1}{2}} = \mathbf{v} - \mathcal{R}\mathbf{z}^{k+\frac{1}{2}} =: \mathbf{r}^{k+\frac{1}{2}}. \tag{11}$$

After computing the vector $\gamma^{k+\frac{1}{2}}$ from the latter equation, the vector $\mathbf{z}^{k+1}$ is computed via $\mathbf{z}^{k+1} = \mathbf{z}^{k+\frac{1}{2}} + \gamma^{k+\frac{1}{2}}$. In the ISSSS algorithm, the systems (9) and (11) are solved inexactly using the iterative methods. System (11) is equivalent to

$$\begin{pmatrix} \beta I_n & B^T \\ -B & \beta I_m \end{pmatrix} \begin{pmatrix} \gamma_1^{k+\frac{1}{2}} \\ \gamma_2^{k+\frac{1}{2}} \end{pmatrix} = \begin{pmatrix} \mathbf{r}_1^{k+\frac{1}{2}} \\ \mathbf{r}_2^{k+\frac{1}{2}} \end{pmatrix},$$

where $\gamma_1^{k+\frac{1}{2}}, \mathbf{r}_1^{k+\frac{1}{2}} \in \mathbb{R}^n$ and $\gamma_2^{k+\frac{1}{2}}, \mathbf{r}_2^{k+\frac{1}{2}} \in \mathbb{R}^m$. For solving the above system, we first solve the system

$$(\beta^2 I_m + BB^T)\gamma_2^{k+\frac{1}{2}} = B\mathbf{r}_1^{k+\frac{1}{2}} + \beta\mathbf{r}_2^{k+\frac{1}{2}} =: \tilde{\mathbf{r}}^{k+\frac{1}{2}}, \tag{12}$$

for computing $\gamma_2^{k+\frac{1}{2}}$ using the CG method, and then simply compute $\gamma_1^{k+\frac{1}{2}}$ via $\gamma_1^{k+\frac{1}{2}} = (\mathbf{r}_1^{k+\frac{1}{2}} - B^T\gamma_2^{k+\frac{1}{2}})/\beta$. The resulting algorithm is summarized as follows.

Algorithm 3.1. The ISSSS iteration method

1. *Choose an initial guess* $\mathbf{z}^0$.
2. *For* $k = 0, 1, 2, \ldots,$ *until convergence, Do*
3. *Compute* $\mathbf{r}^k = \mathbf{v} - \mathcal{R}\mathbf{z}^k$.
4. *Solve the system* (10) *approximately for* γ_1^k *using GMRES.*
5. *Compute* $\gamma_2^k = \mathbf{r}_2^k/(\beta + \alpha)$.
6. *Set* $\gamma^k = (\gamma_1^k; \gamma_2^k)$ *and* $\mathbf{z}^{k+\frac{1}{2}} = \mathbf{z}^k + \gamma^k$.
7. *Compute* $\mathbf{r}^{k+\frac{1}{2}} = \mathbf{v} - \mathcal{R}\mathbf{z}^{k+\frac{1}{2}}$.
8. *Solve the system* (12) *approximately for* $\gamma_2^{k+\frac{1}{2}}$ *using CG.*
9. *Compute* $\gamma_1^{k+\frac{1}{2}} = (\mathbf{r}_1^{k+\frac{1}{2}} - B^T\gamma_2^{k+\frac{1}{2}})/\beta$.
10. *Set* $\gamma^{k+\frac{1}{2}} = (\gamma_1^{k+\frac{1}{2}}; \gamma_2^{k+\frac{1}{2}})$ *and* $\mathbf{z}^{k+1} = \mathbf{z}^{k+\frac{1}{2}} + \gamma^{k+\frac{1}{2}}$.
11. *EndDo*

4 Numerical experiments

In order to show the effectiveness of the ISSSS iteration method we solve some generalized saddle point problems by the flexible GMRES (FGMRES) method [6] in conjunction with the SSS preconditioner. All numerical experiments were performed in MATLAB 2013a on an Intel core i7 CPU (3.50 GHz) 16G RAM Windows 7 system. A zero vector was always used as an initial guess and the stopping criterion $\|\mathbf{b} - \mathcal{A}\mathbf{x}\|_2 < 10^{-6}\|\mathbf{b}\|_2$ was used.

We consider the Oseen problem

$$\begin{cases} -\nu\Delta\mathbf{u} + \mathbf{w}\cdot\nabla\mathbf{u} + \nabla p = \mathbf{f} & \text{in} \quad \Omega, \\ \nabla\cdot\mathbf{u} = 0 & \text{in} \quad \Omega, \end{cases} \tag{13}$$

with suitable boundary conditions on $\partial\Omega$, where $\Omega \subset \mathbb{R}^2$ is a bounded domain and $\mathbf{w}$ is a given divergence free field. The parameter $\nu > 0$ is the viscosity, the vector field $\mathbf{u}$ stands for the velocity and p represents the pressure. The Oseen problem (13) is obtained from the linearization of the steady-state Navier-Stokes equation by the Picard iteration where the vector field $\mathbf{w}$ is the approximation of $\mathbf{u}$ from the previous Picard iteration. Nine iterations of the Picard iteration were used and the generated generalized saddle point in the ninth iteration was used. We use the stabilized Q1-P0 finite element method for the leaky lid driven cavity problems on stretched grids on the unit square, with the viscosity parameter $\nu = 0.01$. The stabilization parameter ($\beta = 0.25$) was used in all cases. We use the IFISS software package [3] to generate the linear systems corresponding to 32×32, 64×64 and 128×128 grids. We mention that the matrix A is non-symmetric positive definite, however the matrix B and has rank $m - 2$. In this case the matrix $\mathcal{A}$ is singular. To get a nonsingular matrix $\mathcal{A}$ we drop last two rows of B and last two rows and columns of C. For all test problems the right-hand side vector $\mathbf{b}$ is set to be $\mathbf{b} = \mathcal{A}[1, 1, \ldots, 1]^T$.

In the implementation of the SSS preconditioner we used the following two methods. In both methods and for all the subsystems a zero vector was used as an initial guess and the maximum number of iterations were set to be 20.

Method 1: The subsystems $(\alpha I_n + H)v_1 = r_1$ and $(\alpha I_m + C)v_2 = r_2$ were solved using the CG method. We also solved the system $\mathcal{R}z = \mathbf{v}$ using the GMRES(10). The iterations of CG and GMRES(10) were stopped as soon as the residual 2-norm was reduced by a factor of 10^3.

Method 2: Similar to Method 1, the systems $(\alpha I_n + H)v_1 = r_1$ and $(\alpha I_m + C)v_2 = r_2$ were solved using the CG method. For solving the system $\mathcal{R}z = \mathbf{v}$, the ISSSS algorithm

was employed. For both the CG method and the ISSSS algorithm for solving the above systems, the iterations were stopped as soon as the residual 2-norm was reduced by a factor of 10^3. In the kth iteration of the ISSSS algorithm the system $((\beta+\alpha)I_n + S)\,\gamma_1^k = \mathbf{r}_1^k$ was solved using GMRES(10) with the stopping criterion $\|\mathbf{r}_1^k - ((\beta+\alpha)I_n + S)\,\gamma_1^k\|_2 < \epsilon_k\|\mathbf{r}_1^k\|_2$, where $\epsilon_k = \max\{10^{-3}, 0.1 \times 0.9^k\}$ (see [1]). Also, the system (12) was solved using CG and the stopping criterion $\|\tilde{\mathbf{r}}^{k+\frac{1}{2}} - (\beta^2 I_m + BB^T)\gamma_2^{k+\frac{1}{2}}\|_2 < \epsilon_k\|\tilde{\mathbf{r}}^{k+\frac{1}{2}}\|_2$ was used.

For all the test problems we first choose an appropriate value of the parameter α for Method 1 and the same value of α along with β^* was used for Method 2. Numerical results are given in Tables 1. As we observe there is no significant difference between the number of iterations (Iter) of two methods. However, the elapsed CPU time for Method 2 is always less than those of Method 1.

Table 1: Numerical results for the stretched grid and $\nu = 0.01$.

				Metod 1			Metod 2			
grid	n	m	cond($\mathcal{A}$)	α	Iter	CPU	α	β^*	Iter	CPU
5×5	2178	1022	9.20×10^4	0.1	257	2.96	0.1	0.105	257	1.92
6×6	8450	4094	3.06×10^6	0.01	135	17.87	0.01	0.025	135	9.88
7×7	33282	16382	1.32×10^8	0.01	167	54.09	0.01	0.018	167	24.84

5 Conclusion

We have presented the shifted skew-symmetric splitting (SSSS) method and its inexact version, ISSSS, for solving the shifted skew-symmetric system appeared in the implementation of the SSS preconditioner for the generalized saddle point problems. We have shown that the SSSS iteration method is unconditionally convergent. Numerical results have shown that the new implementation of the SSS preconditioner significantly reduces the CPU time of the classical implementation of the SSS preconditioner.

References

[1] Z.-Z. Bai, G.H. Golub, M.K. Ng, Hermitian and skew-Hermitian splitting methods for non-Hermitian positive definite linear systems, *SIAM J. Matrix Anal. Appl.* 24 (2003), 603-626.

[2] M. Benzi and G.H. Golub, A preconditioner for generalized saddle point problems, *SIAM J. Matrix Anal. Appl.* 26 (2004), 20-41.

[3] H.C. Elman, A. Ramage and D.J. Silvester, IFISS: A Matlab toolbox for modelling incompressible flow, *ACM Trans. Math. Softw.* 33 (2007) Article 14.

[4] G.H. Golub and C.F. Van Loan, *Matrix computations*, 4th Edition, Johns Hopkins University Press, Baltimore, 2013.

[5] R.A. Horn and C.R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, 1990.

[6] Y. Saad, *Iterative methods for sparse linear systems*, 2nd Edition, SIAM, Philadelphia, 2003.

Minimal residual HSS iteration method for the Sylvester equation[1]

Mohammad Khorsand Zak*

Department of Applied Mathematics, Aligudarz Branch, Islamic Azad University, Aligudarz, Iran

Abstract

By applying the minimal residual technique to the Hermitian and skew-Hermitian (HSS) iteration scheme, we introduce a non-stationary iteration method named minimal residual Hermitian and skew-Hermitian (MRHSS) iteration method, to solve the continuous Sylvester equation. Numerical results verify the effectiveness and robustness of the MRHSS iteration method for the Sylvester equation.

Keywords: Sylvester equation, Hermitian and skew-Hermitian method, Minimal residual

Mathematics Subject Classification [2010]: 65F10, 65F30, 65F50

1 Introduction

In many problems in scientific computing we encounter with matrix equations. Nowadays, the continuous Sylvester equation is possibly the most famous and the most broadly employed linear matrix equation, and is given as

$$AX + XB = C, \tag{1}$$

where $A \in \mathbb{C}^{n\times n}$, $B \in \mathbb{C}^{m\times m}$ and $C \in \mathbb{C}^{n\times m}$ are defined matrices and $X \in \mathbb{C}^{n\times m}$ is an unknown matrix. Equation (1) has a unique solution if and only if A and $-B$ have no common eigenvalues, which will be assumed throughout this paper. The Sylvester equation appears frequently in many areas of applied mathematics and plays vital roles in a number of applications such as control theory, model reduction and image processing, see [1–3] and their references.

The matrix equation (1) is mathematically equivalent to the linear system of equations

$$\mathcal{A}x = c, \tag{2}$$

where the matrix $\mathcal{A}$ is of dimension $nm \times nm$ and is given by

$$\mathcal{A} = I_m \otimes A + B^T \otimes I_n, \tag{3}$$

where $\otimes$ denotes the Kronecker product ($A \otimes B = [a_{ij}B]$) and

$$c = vec(C) = (c_{11}, c_{21}, \cdots, c_{n1}, c_{12}, c_{22}, \cdots, c_{n2}, \cdots, c_{nm})^T,$$
$$x = vec(X) = (x_{11}, x_{21}, \cdots, x_{n1}, x_{12}, x_{22}, \cdots, x_{n2}, \cdots, x_{nm})^T.$$

[1]Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

*Speaker. Email address: mo.khorsand@mail.um.ac.ir

Of course, this is a numerically poor way to determine the solution X of the Sylvester equation (1), as the linear system of equations (2) is costly to solve and can be ill-conditioned.

When both coefficient matrices are (non-Hermitian) positive semi-definite, and at least one of them is positive definite, the Hermitian and skew-Hermitian splitting (HSS) method [1] and the nested splitting conjugate gradient (NSCG) method [2] are often the methods of choice for efficiently and accurately solving the Sylvester equation (1).

Motivated by [4, 5], we apply the minimal residual technique to the Hermitian and skew-Hermitian iteration scheme and introduce a non-stationary iteration method named minimal residual Hermitian and skew-Hermitian (MRHSS) iteration method to solve the continuous Sylvester equation.

In the remainder of this paper, we use $||M||_2$, $||M||_F$ and I_n to denote the spectral norm, the Frobenius norm of a matrix $M \in \mathbb{C}^{n\times n}$, and the identity matrix with dimension n, respectively. Note that $||.||_2$ is also used to represent the 2-norm of a vector. Furthermore, we have the following equivalent relationships between the Frobenius norm of a matrix R and the 2-norm of a vector $r = vec(R)$:

$$||r||_2 = \sqrt{\sum_{i=1}^{mn} |r_i|^2} \Leftrightarrow ||R||_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |R_{ij}|^2}.$$

2 Main results

For the linear system of equations (2), we consider the Hermitian and skew-Hermitian splitting $\mathcal{A} = \mathcal{H} + \mathcal{S}$, where

$$\mathcal{H} = \frac{\mathcal{A} + \mathcal{A}^T}{2}, \quad \mathcal{S} = \frac{\mathcal{A} - \mathcal{A}^T}{2}, \tag{4}$$

are the Hermitian and skew-Hermitian parts of matrix $\mathcal{A}$, respectively. Then, the iteration scheme of the MRHSS iteration method [4, 5] for system of linear equations (2) is

$$\begin{cases} x^{(k+\frac{1}{2})} &= x^{(k)} + \beta_k \delta^{(k)} \\ x^{(k+1)} &= x^{(k+\frac{1}{2})} + \gamma_k \delta^{(k+\frac{1}{2})}, \end{cases} \tag{5}$$

where, $\delta^{(k)} = (\hat{\alpha} I + \mathcal{H})^{-1} r^{(k)}$, $\delta^{(k+\frac{1}{2})} = (\hat{\alpha} I + \mathcal{S})^{-1} r^{(k+\frac{1}{2})}$, $r^{(k)} = c - \mathcal{A}x^{(k)}$ and $r^{(k+\frac{1}{2})} = c - \mathcal{A}x^{(k+\frac{1}{2})}$. Let $M_1 = \mathcal{A}(\hat{\alpha} I + \mathcal{H})^{-1}$ and $M_2 = \mathcal{A}(\hat{\alpha} I + \mathcal{S})^{-1}$. The residual form of iteration scheme (5) can be written as

$$\begin{cases} r^{(k+\frac{1}{2})} &= r^{(k)} - \beta_k M_1 r^{(k)} \\ r^{(k+1)} &= r^{(k+\frac{1}{2})} - \gamma_k M_2 r^{(k+\frac{1}{2})}. \end{cases} \tag{6}$$

Denote $M = (\hat{\alpha} I + \mathcal{H})^{-1}$. Then, an inner product can be defined as

$$(x, y)_M = (Mx, My), \qquad \forall x, y \in \mathbb{C}^{nm}, \tag{7}$$

where $(\cdot, \cdot)$ denotes the l^2 inner product of two vectors. Thus, for $x \in \mathbb{C}^{nm}$ and $X \in \mathbb{C}^{nm\times nm}$, the induced vector and the induced matrix norms can be defined as $||x||_M = ||Mx||_2$ and $||X||_M = ||MXM^{-1}||_2$, respectively. Now, the parameter β_k is determined by the 2-norm of the residual, and we have

$$\beta_k = \frac{(r^{(k)}, M_1 r^{(k)})}{||M_1 r^{(k)}||_2^2}. \tag{8}$$

However, the parameter γ_k will be determined by minimizing the M-norm of the residual rather than the 2-norm, see [4]. Therefore, we have

$$\gamma_k = \frac{(Mr^{(k+\frac{1}{2})}, MM_2r^{(k+\frac{1}{2})})}{||MM_2r^{(k+\frac{1}{2})}||_2^2}. \tag{9}$$

The iteration scheme (5) is an unconditionally convergent MRHSS iteration method [4].

For the Sylvester equation (1), according to iterative scheme (5), we have the following iteration scheme

$$\begin{cases} X^{(k+\frac{1}{2})} &= X^{(k)} + \beta_k\Delta^{(k)} \\ X^{(k+1)} &= X^{(k+\frac{1}{2})} + \gamma_k\Delta^{(k+\frac{1}{2})}, \end{cases} \tag{10}$$

where, $\Delta^{(0)}$ obtain from the Sylvester equation

$$H_A(\alpha)\Delta^{(0)} + \Delta^{(0)}H_B(\alpha) = R^{(0)}, \tag{11}$$

and $\Delta^{(k+\frac{1}{2})}$ obtain from the Sylvester equation

$$S_A(\alpha)\Delta^{(k+\frac{1}{2})} + \Delta^{(k+\frac{1}{2})}S_B(\alpha) = R^{(k+\frac{1}{2})}, \tag{12}$$

with $R^{(0)} = C - AX^{(0)} - X^{(0)}B$ and $R^{(k+\frac{1}{2})} = C - AX^{(k+\frac{1}{2})} - X^{(k+\frac{1}{2})}B$. We state how to update $\Delta^{(k+1)}$ a few later.

If the Sylvester equation (1) has a unique solution, then under the assumption A and B are positive semi-definite and at last one of them is positive definite, we can easily see that there is no common eigenvalue between the matrices H_A and $-H_B$ (also for S_A and $-S_B$), so the Sylvester equations (11) and (12) have unique solution for all given right hand side matrices.

Let $H_A(\alpha) = \alpha I_n + H_A, S_A(\alpha) = \alpha I_n + S_A, H_{B^T}(\alpha) = \alpha I_m + H_{B^T}, S_{B^T}(\alpha) = \alpha I_m + S_{B^T}$ and $H_A, S_A, H_{B^T}, S_{B^T}$ are the Hermitian and skew-Hermitian parts of A and B^T, respectively. From (3) and (4), by using the Kronecker product's properties, we have

$$\hat{\alpha}I + \mathcal{H} = I_m \otimes H_A(\alpha) + H_{B^T}(\alpha) \otimes I_n \tag{13}$$

$$\hat{\alpha}I + \mathcal{S} = I_m \otimes S_A(\alpha) + S_{B^T}(\alpha) \otimes I_n \tag{14}$$

where $\alpha = \frac{\hat{\alpha}}{2}$. Form relations (6), we can obtain

$$\begin{cases} R^{(k+\frac{1}{2})} &= R^{(k)} - \beta W^{(k)} \\ R^{(k+1)} &= R^{(k+\frac{1}{2})} - \gamma W^{(k+\frac{1}{2})} \end{cases} \tag{15}$$

where $W^{(k)} = A\Delta^{(k)} + \Delta^{(k)}B$ and $W^{(k+\frac{1}{2})} = A\Delta^{(k+\frac{1}{2})} + \Delta^{(k+\frac{1}{2})}B$. Moreover, similar to (8) and (9), we can obtain

$$\beta = \frac{\langle R^{(k)}, W^{(k)}\rangle_F}{\langle W^{(k)}, W^{(k)}\rangle_F}, \tag{16}$$

and

$$\gamma = \frac{\langle V^{(k+\frac{1}{2})}, U^{(k+\frac{1}{2})}\rangle_F}{\langle U^{(k+\frac{1}{2})}, U^{(k+\frac{1}{2})}\rangle_F}, \tag{17}$$

where, $V^{(k+\frac{1}{2})}$ obtain from the Sylvester equation

$$H_A(\alpha)V^{(k+\frac{1}{2})} + V^{(k+\frac{1}{2})}H_B(\alpha) = R^{(k+\frac{1}{2})},$$

and $U^{(k+\frac{1}{2})}$ obtain from the Sylvester equation

$$H_A(\alpha)U^{(k+\frac{1}{2})} + U^{(k+\frac{1}{2})}H_B(\alpha) = W^{(k+\frac{1}{2})}$$

On the surface, four systems of linear equations should be solved at each step of the MRHSS method for system of linear equations (2). But it can be reduced to three. Denote $\zeta^{(k+\frac{1}{2})} = (\hat{\alpha}I + \mathcal{H})^{-1}r^{(k+\frac{1}{2})}$ and $v^{(k+\frac{1}{2})} = (\hat{\alpha}I + \mathcal{H})^{-1}\mathcal{A}\delta^{(k+\frac{1}{2})}$, the vector $\delta^{(k+1)}$ in Step $k+1$ can be calculated as follows

$$\begin{aligned}
\delta^{(k+1)} &= (\hat{\alpha}I + \mathcal{H})^{-1}(c - \mathcal{A}x^{(k+1)}) \\
&= (\hat{\alpha}I + \mathcal{H})^{-1}(c - \mathcal{A}(x^{(k+\frac{1}{2})} + \gamma_k\delta^{(k+\frac{1}{2})})) \\
&= (\hat{\alpha}I + \mathcal{H})^{-1}(r^{(k+\frac{1}{2})} - \gamma_k\mathcal{A}\delta^{(k+\frac{1}{2})}) \\
&= \zeta^{(k+\frac{1}{2})} - \gamma_k v^{(k+\frac{1}{2})},
\end{aligned}$$

where the $\zeta^{(k+\frac{1}{2})}$ and $v^{(k+\frac{1}{2})}$ have been calculated in Step k. Therefore, in (10) we can update $\Delta^{(k+1)}$ as

$$\Delta^{(k+1)} = V^{(k+\frac{1}{2})} - \gamma U^{(k+\frac{1}{2})}.$$

In addition, we choose the value of parameter α as in [1].

Therefore, an implementation of the MRHSS method for the continuous Sylvester equation can be given by the following algorithm.

Algorithm 2.1. The MRHSS algorithm for the Sylvester equation

1. Select an initial guess $X^{(0)}$, compute $R^{(0)} = C - AX^{(0)} - X^{(0)}B$
2. Solve $H_A(\alpha)\Delta^{(0)} + \Delta^{(0)}H_B(\alpha) = R^{(0)}$
3. For $k = 0, 1, 2, \cdots$, until convergence, Do:
4. $\quad W^{(k)} = A\Delta^{(k)} + \Delta^{(k)}B$
5. $\quad \beta = \frac{\langle R^{(k)}, W^{(k)}\rangle_F}{\langle W^{(k)}, W^{(k)}\rangle_F}$
6. $\quad X^{(k+\frac{1}{2})} = X^{(k)} + \beta\Delta^{(k)}$
7. $\quad R^{(k+\frac{1}{2})} = R^{(k)} - \beta W^{(k)}$
8. $\quad$ Solve $S_A(\alpha)\Delta^{(k+\frac{1}{2})} + \Delta^{(k+\frac{1}{2})}S_B(\alpha) = R^{(k+\frac{1}{2})}$
9. $\quad$ Solve $H_A(\alpha)V^{(k+\frac{1}{2})} + V^{(k+\frac{1}{2})}H_B(\alpha) = R^{(k+\frac{1}{2})}$
10. $\quad W^{(k+\frac{1}{2})} = A\Delta^{(k+\frac{1}{2})} + \Delta^{(k+\frac{1}{2})}B$
11. $\quad$ Solve $H_A(\alpha)U^{(k+\frac{1}{2})} + U^{(k+\frac{1}{2})}H_B(\alpha) = W^{(k+\frac{1}{2})}$
12. $\quad \gamma = \frac{\langle V^{(k+\frac{1}{2})}, U^{(k+\frac{1}{2})}\rangle_F}{\langle U^{(k+\frac{1}{2})}, U^{(k+\frac{1}{2})}\rangle_F}$
13. $\quad X^{(k+1)} = X^{(k+\frac{1}{2})} + \gamma\Delta^{(k+\frac{1}{2})}$
14. $\quad R^{(k+1)} = R^{(k+\frac{1}{2})} - \gamma W^{(k+\frac{1}{2})}$
15. $\quad \Delta^{(k+1)} = V^{(k+\frac{1}{2})} - \gamma U^{(k+\frac{1}{2})}$
16. End Do

Theorem 2.2. *Suppose that the coefficient matrices A and B in the continuous Sylvester equation (1) are non-Hermitian positive semi-definite, and at least one of them is positive definite. Then the MRHSS iteration method (10) for solving the Sylvester equation (1) is unconditionally convergent for any $\alpha > 0$ and any initial guess $X^{(0)} \in \mathbb{C}^{n\times m}$.*

Proof. The continuous Sylvester equation (1) is mathematically equivalent to the linear system of equations (2). Therefore, the proof is similar to that of Theorem 3.3 in [4] with only technical modifications. □

3 Numerical results

All numerical experiments presented in this section were computed in double precision with a number of MATLAB codes. All iterations are started from the zero matrix for initial $X^{(0)}$ and terminated when the current iterate satisfies $\frac{\|R^{(k)}\|_F}{\|R^{(0)}\|_F} \leq 10^{-8}$, where $R^{(k)} = C - AX^{(k)} - X^{(k)}B$ is the residual of the kth iterate. Also we use the tolerance $\varepsilon = 0.001$ for inner iterations in corresponding methods. For each experiment we report the CPU time, the number of total outer iteration steps and the norm of residual $\|R^{(k)}\|_F$, and compare the HSS [1] iterative method with the MRHSS iterative method for solving the continuous Sylvester equation (1).

Example 3.1. For this example, we use the matrices

$$A = M + 2rN + \frac{100}{(n+1)^2}I, \qquad \text{and} \qquad B = M + 2rN + \frac{100}{(m+1)^2}I$$

where $M = \text{tridiag}(-1, 2, -1)$, $N = \text{tridiag}(0.5, 0, -0.5)$ from suitable dimensions, and $r = 0.01$ [2]. For this problem we consider $n = 2048$ and $m = 8$.

Example 3.2. We consider the continuous Sylvester equation (1) with $n = m = 512$ and the coefficient matrices

$$\begin{cases} A = \text{diag}(1, 2, \cdots, n) + rL^T, \\ B = 2^{-t}I_n + \text{diag}(1, 2, \cdots, n) + rL^T + 2^{-t}L, \end{cases}$$

with L the strictly lower triangular matrix having ones in the lower triangle part [1].

Example 3.3. For this example, we use $A = B = \text{tridiag}(-1, 4, -2)$ of dimension 1000×1000 instead the coefficient matrices A and B [2, 3].

We apply the iteration methods to these problems and the results are given in Table 1. Comparing the results in the Table 1, Shows that the MRHSS method is more efficient

Table 1: The results for the problems

	HSS			MRHSS		
	CPU	iteration	res-norm	CPU	iteration	res-norm
Example 3.1	0.65	20	1.44109e-5	0.48	12	8.4888e-6
Example 3.2	210.01	99	0.0302	143.06	49	0.0304
Example 3.3	247.26	21	1.7697e-4	36.29	12	1.0527e-4

versus the HSS method.

4 Conclusion

In this paper, we have proposed an efficient iterative method, which named the MRHSS method, for solving the continuous Sylvester equation $AX + XB = C$. We have compared the MRHSS method with the HSS method for some problems. We have observed that, for these problems the MRHSS method is more efficient versus the HSS method.

References

[1] Z. Z. Bai, *On Hermitian and skew-Hermitian splitting iteration methods for continuous Sylvester equations*, J. Comput. Math., 29:2 (2011) 185–198.

[2] M. Khorsand Zak and F. Toutounian, *Nested splitting CG-like iterative method for solving the continuous Sylvester equation and preconditioning*, Adv. Comput. Math., 40 (2014) 865–880.

[3] M. Khorsand Zak and F. Toutounian, *An iterative method for solving the continuous Sylvester equation by emphasizing on the skew-Hermitian parts of the coefficient matrices*, Intern. J. Computer Math., 94 (2017) 633–649.

[4] A.-L Yang, *On the convergence of the minimum residual HSS iteration method*, Appl. Math. Lett., 94 (2019) 210–216.

[5] A.-L Yang, Y. Cao, and Y.-J. Wu, *Minimum residual Hermitian and skew-Hermitian splitting iteration method for non-Hermitian positive definite linear systems*, BIT Numer. Math., 59 (2019) 299–319.

Some applications of linear algebra in combinatorics[1]

Maryam Khosravi*

Faculty of Mathematics and computer, Shahid Bahonar University of Kerman, Kerman, Iran

Abstract

In this paper, using some linear algebraic methods, we show that every latin tade can be produced by intercalates (i.e. latin trades of volume 4). A similar result is true for 4-cycle systems. That is, every 4-cycle system can be generated by double-diamonds (4-cycle systems of volume 2).

Keywords: Latin square, 4-cycle system, Trade, Nulity of matrix, Basis

Mathematics Subject Classification [2010]: 05B20, 05B30, 15A03

1 introduction

An interesting problem in combinatorics is that whether there can be defined some moves (using trades of small volume or something else) between different elements of a class of combinatorial objects with the same parameters, such as latin squares, Steiner triple systems, etc. These moves must be such a way that each element has chance to be produced by these moves.

By simulating an ergodic Markov chain whose stationary distribution is uniform over the space of $n \times n$ latin squares, Mark T. Jacobson and Peter Matthews [4], have discussed elegant method by which they generate latin squares with a uniform distribution (approximately). The central issue is construction of moves that connect the squares.

There does not exist a known move between Steiner triple systems as yet. Steiner triple systems are 3-cycle systems.

In this note, we investigate two classes: latin squares and 4-cycle systems.

2 Latin square

A *latin square* L of order n is an $n \times n$ array with entries chosen from an n-set $N = \{0, 1, \ldots, n-1\}$ in such a way that each element of N occurs precisely once in each row and in each column of the array. For ease of exposition, a latin square L will be represented by a set of ordered triples $\{(i, j; L_{ij}) : \text{ element } L_{ij} \text{ accures in cell } (i, j) \text{ of the array}\}$.

A *partial latin square* P of order n is an $n \times n$ array in which some of the entries are filled with elements from N in such a way that each element of N occurs at most once in each

[1] Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

* Email address: khosravi_m@uk.ac.ir

row and at most once in each column of the array. The set $S_P = \{(i,j) : \ (i,j;P_{ij} \in P\}$ is called the shape of P and the number of elements in S_P is called the volume of P.

In a latin square P, we define $R^r_P(C^r_P)$ as the set of entries occurring in row (column) r of P.

A *latin trade* $T = (P,Q)$ of volume s is an ordered set of two partial latin squares of volumes s and orders n, such that

1. $S_P = S_Q$.
2. for each $(i,j) \in S_p$, $P_{ij} \neq Q_{ij}$.
3. for each r, $0 \leq r \leq n-1$, $R^r_P = R^r_Q$ and $C^r_P = C^r_Q$.

For example, the following tables show a latin trade of order 5 and volume 19:

.	.	2	3	1
.	2	.	1	4
1	.	0	4	3
0	4	1	.	2
4	1	3	2	0

.	.	1	2	3
.	1	.	4	2
4	.	3	1	0
1	2	0	.	4
0	4	2	3	1

$\Rightarrow$

.	.	2_1	3_2	1_3
.	2_1	.	1_4	4_2
1_4	.	0_3	4_1	3_0
0_1	4_2	1_0	.	2_4
4_0	1_4	3_2	2_3	0_1

A latin trade of volume 4 which is unique (up to isomorphism), is called an *intercalate.*

An interesting result about latin trades is as follows.

Theorem 2.1. *Every latin trade can be written as a sum of intercalates.*

In [4], the authors found a method by which they generate latin squares uniformly. Although, most of their lengthy paper is to construct a latin square from another. This method was rewitten by Aryapour and Mahmoodian [1] in a simple way using trades. They noted that in this constraction, sometimes we generate an improper latin trade.

Finally, in [5], the authors defined an inclusion matrix M such that every latin trade can be considered as a vector in kenel of M. They show that there exists a basis of kernel of M consist of intercalates.

Of course, for each $1 \leq k \leq n-1$ and $2 \leq i,j \leq n$, they defined an intercalate $P = \{(1,1;0),(i,1;k),(1,j;k),(i,j;0)\}$ and $Q = \{(1,1;k),(i,1;0),(1,j;0),(i,j;k)\}$ and show that these intrcalates form a basis for latin trades.

3 4-cycle systems

Let T_1 be a set of edge-disjoint 4–cycles on the vertex set $\{1,2,\ldots,v\}$. Then T_1 is called a 4–cycle trade, if there exists a set, T_2, of edge-disjoint 4–cycles on the same vertex set $\{1,2,\ldots,v\}$, such that $T_1 \cap T_2 = \emptyset$ and $\bigcup_{C\in T_1} E(C) = \bigcup_{C\in T_2} E(C)$.

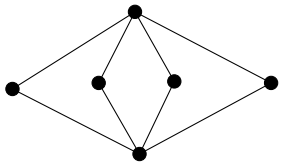
We call T_2 a disjoint mate of T_1 and the pair (T_1,T_2) is called a 4–cycle bitrade of volume $s = |T_1|$ and foundation $v = |\bigcup_{C\in T_1} V(C)|$. Here for them we use the term "trade".

A μ-way 4-cycle trade is a collection of μ disjoint collections $\{T_1,\ldots,T_\mu\}$ such that (T_i,T_j) forms a 4- cycle bitrade for each $i \neq j$.

The following well-known theorem states that for which values of n a 4-CS(n) exists.

Theorem 3.1. [2, Page 266] *A necessary and sufficient condition for the existence of a* 4-CS(n) *is that* $n \equiv 1 (\bmod 8)$

A double-diamond is a trade of volume 2 which can be seen in the following graph:



Let M be a pair inclusion matrix whose rows are corresponded to the edges of the complete graph K_n and its columns are corresponded to all possible 4-cycles of the complete graph K_n.

Since by every four vertices a, b, c and d, we can construct three different 4-cycles (a, b, c, d), (a, c, b, d) and (a, b, d, c), the matrix M has exactly $3\binom{n}{4}$ columns. Thus, the matrix M is of size $\binom{n}{2} \times 3\binom{n}{4}$.

Also, for each trade (T_1, T_2), we consider a "frequency" vector X with $3\binom{n}{4}$ components with 1 for each cycle in T_1 and -1 for each cycle in T_2, other cycles are corresponded with a 0 component. It is easy to see that every vector X corresponding to a trade, is a vector in the kernel of M. We show that there exists a bases for the kernel of M, containing only double-diamonds.

Theorem 3.2. *The pair inclusion matrix M is a full rank matrix.*

Corollary 3.3. *The nullity of M is $3\binom{n}{4} - \binom{n}{2}$.*

Theorem 3.4. *The set of vectors corresponding with the double-diamonds is a generating set for the kernel of M. In other words, for each n, there exists a set of $3\binom{n}{4} - \binom{n}{2}$ linearly independent vectors in the kernel of M, where each of them is corresponded to a double-diamond.*

4 Conclusion

It seems that the linear algebra provides some useful tools to study the trades in different combinatorial designs. Generating these designs with small trade, help us to know more about big trades by studying small ones.

References

[1] M. Aryapoor and E.S. Mahmoodian, On uniformly generating Latin squares, *Bull. Inst. Combin. Appl.* 62 (2011), 48–68.

[2] C. J. Colbourn and J. H. Dinitz, *The CRC handbook of combinatorial designs*, CRC Press, 2010.

[3] D. Donovan and E. S. Mahmoodian, An algorithm for writing any Latin interchange as a sum of intercalates, *Bull. Inst. Combin. Appl.*, 34 (2002), 90–98. Corrigendum: *Bull. Inst. Combin. Appl.* 37(2003), 44.

[4] M.T. Jacobson and P. Mattews, Generating uniformly distributed random Latin squares, *J. Combin. Des.*, 4 (1996), no. 6, 405–437.

[5] A. A. Khanban, M. Mahdian, and E. S. Mahmoodian, A linear algebraic approach to orthogonal arrays and Latin squares, *Ars Combinatoria*, 105 (2012), 15–22.

[6] M. Khosravi, E.S. Mahmoodian and S. Rashidi, A linear algebraic approach to 4-cycle systems, in preperation.

Norm inequalities related to the Kadison inequality[1]

Mohsen Kian*

Department of Mathematics, University of Bojnord, Iran

Abstract

We present some norm inequalities related to the Kadison inequality. In particular, we utilize the famous Furuta inequality to obtain some complements to the asymmetric Kadison inequality for matrices.

Keywords: Positive linear map, Kadison inequality, Unitary matrix

Mathematics Subject Classification [2010]: 15A60, 15A45

1 Introduction

In the present article, $\mathbb{M}_n$ denotes the algebra of all $n\times n$ matrices with complex entries and I denotes the identity matrix. We write $A \geq 0$ when A is a positive semidefnite matrix. The well-known (Löwner) partial order on the real space of all Hermitian matrices is defined by $A \leq B$ if and only if $B - A \geq 0$. When $mI \leq A \leq MI$, we simply write $m \leq A \leq M$. For $J \subseteq \mathbb{R}$, we denoted by $\sigma(J)$ the set of all Hermitian matrices, whose eigenvalues are contained in J. If $f : J \to \mathbb{R}$ is a continuous function, then $f(A)$ is defined by the functional calculus for every $A \in \sigma(J)$. A continuous function $f : J \to \mathbb{R}$ is called matrix convex (concave) if $f\left(\frac{A+B}{2}\right) \leq (\geq) \frac{f(A)+f(B)}{2}$ for all $A, B \in \sigma(J)$.

A map Φ defined on $\mathbb{M}_n$ is called *positive* whenever it preserve the Löwner order. Φ is called unital if $\Phi(I) = I$.

The celebrated Kadison's inequality asserts that $\Phi(A^2) \geq \Phi(A)^2$ holds for every unital positive linear map Φ and every Hermitian matrix A. It provides a non-commutative extension for the positivity of the famous variance quantity of a random variable X

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2,$$

where $\mathbb{E}$ is the expectation value.

In application of the probability theory, it is useful to have some lower bounds for the variance. In the non-commutative setting, it is known that if $m \leq A \leq M$ for two positive scalars $m < M$, then

$$\Phi(A^2) - \Phi(A)^2 \leq \frac{(M-m)^2}{4} \quad \text{and} \quad \Phi(A^2) \leq \frac{(M+m)^2}{4mM}\Phi(A)^2. \tag{1}$$

[1]Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

*Speaker. Email address: kian@ub.ac.ir

The Kadison inequality is equivalent to the positivity of the block matrix

$$\begin{bmatrix} I & \Phi(A) \\ \Phi(A) & \Phi(A^2) \end{bmatrix}.$$

The authors of [5] present a generalization of Kadison inequality by proving that the operator matrix

$$\begin{bmatrix} I & \Phi(A) & \cdots & \Phi\left(A^r\right) \\ \Phi(A) & \Phi(A^2) & \cdots & \Phi\left(A^{r+1}\right) \\ \vdots & \vdots & \ddots & \vdots \\ \Phi\left(A^r\right) & \Phi\left(A^{r+1}\right) & \cdots & \Phi\left(A^{2r}\right) \end{bmatrix}$$

is positive.

Furuta gave a variant of the Kadison inequality as

$$|\Phi(X^p)^r \Phi(X^q)^r| \leq \Phi(X^{(p+q)r}), \tag{2}$$

when $0 \leq p \leq q$ and $\frac{q}{p+q} \leq r \leq \frac{2q}{p+q}$. In particular,

$$|\Phi(X^p)\Phi(X^q)| \leq \Phi(X^{p+q}) \tag{3}$$

holds for every $0 \leq p \leq q$ or

$$|\Phi(X^r)\Phi(X)| \leq \Phi(X)^{1+r} \tag{4}$$

for every $r \in [0,1]$.

In this paper, we present some complementary inequalities to (3) and (4).

2 Main results

First we note that inequality (4) is equivalent to

$$\left\| \Phi(X)^{-\frac{1+r}{2}} \, |\Phi(X^r)\Phi(X)| \, \Phi(X)^{-\frac{1+r}{2}} \right\| \leq 1, \tag{5}$$

that is valid for every $r \in [0,1]$.

We need some facts about the Furuta inequality. If $A \geq B \geq 0$, the so-called Löwner–Heinz inequality implies that $A^r \geq B^r$ for every $r \in [0,1]$. If $r \notin [0,1]$, then this is not true in general. The celebrated Furuta inequality provides an analogue order preserving result by showing that if $A \geq B \geq 0$, then

$$A^{1+r} \geq \left(A^{\frac{r}{2}} B^p A^{\frac{r}{2}}\right)^{\frac{1+r}{p+r}} \qquad (r \geq 0, \quad p \geq 1). \tag{6}$$

M. Fujii et. all presented the next result regarding the Furuta inequality in [3].

Lemma 2.1. *Let A and B be positive definite matrices such that $0 < m \leq B \leq M$ for two positive real numbers $m < M$. Then*

$$\left\| A^{\frac{1}{2}} \left(A^{\frac{s}{2}} B^{p+s} A^{\frac{s}{2}}\right)^{\frac{1}{p}} A^{\frac{1}{2}} \right\| \leq K\left(h^{1+r}, \frac{p+s}{1+r}\right)^{\frac{1}{p}} \left\| A^{\frac{1+r}{2}} B^{1+r} A^{\frac{1+r}{2}} \right\|^{\frac{p+s}{p(1+r)}} \tag{7}$$

holds for all $p \geq 1$ and $s \geq r > -1$, in which

$$K(h,p) = \frac{h^p - 1}{(p-1)(h-1)} \left(\frac{p-1}{p} \frac{h^p - 1}{h^p - h}\right)^p, \quad h = \frac{M}{m}$$

is the generalized Kantorovich constant.

Now we present a complementary result to (4).

Theorem 2.2. *Suppose that Φ is a positive linear map and X is a positive definite matrix. If $m \leq |\Phi(X^q)\Phi(X^p)|^{\frac{q}{p}} \leq M$ for two positive scalars m, M, then*

$$K\left(h^{\frac{p}{q}}, 2\right)^{-\frac{1}{2}} \leq \left\| \Phi(X^p)^{-\frac{q+p}{2p}} \left|\Phi(X^q)\Phi(X^p)\right| \Phi(X^p)^{-\frac{q+p}{2p}} \right\|. \tag{8}$$

holds for all $p \leq q \leq 2p$.

Remark 2.3. Note that if we put $r = \frac{q}{p}$ and put $X^{\frac{1}{p}}$ instead of X in (8), then we have

$$K\left(h^{\frac{1}{r}}, 2\right)^{-\frac{1}{2}} \leq \left\| \Phi(X)^{-\frac{1+r}{2}} \left|\Phi(X^r)\Phi(X)\right| \Phi(X)^{-\frac{1+r}{2}} \right\| \tag{9}$$

which gives obviously a counterpart to (5). Moreover, with $r = 1$ this gives a converse to the Kadison inequality as

$$K(h, 2)^{-1/2} \leq \left\|\Phi(X)^{-1}\Phi(X)^2\Phi(X)^{-1}\right\| \leq 1.$$

In the next theorem, we give a difference counterpart to (3).

Theorem 2.4. *Let A be a positive definite matrix such that $\mathrm{sp}(A) \subseteq [m, M]$. If $p, q \geq 0$, then there exists a unitary matrix U such that*

$$\Phi(A^{p+q}) - U\left|\Phi(A^p)\Phi(A^q)\right|U^* \leq C(h^q, 1 + \frac{p}{q}) + (C(h^p, 2) - C(h^{2q}, \frac{p}{q}))^{\frac{1}{2}} M^q$$

for every unital positive linear mapping Φ, in which

$$C(h, p) = \frac{Mm^p - mM^p}{M - m} + (p - 1)\left(\frac{1}{p}\frac{M^p - m^p}{M - m}\right)^{\frac{p}{p-1}}.$$

Remark 2.5. It proof of Theorem 2.4, we use the fact that if X, Y are positive definite matrices, then

$$(X + Y)^{\frac{1}{2}} \leq UX^{\frac{1}{2}}U^* + VY^{\frac{1}{2}}X^* \tag{10}$$

for some unitaries U and V. In [4], it was shown that the inequality (10) can be stated without the presence of unitaries U and V if we restrict the spectrum of matrices X, Y:
Theorem. [4, Corollary 2.9] If f is a continuous concave function with $f(0) \geq 0$, then $f(X+Y) \leq f(X)+f(Y)$ for all positive definite matrices X, Y for which $X \leq aI \leq X+Y$ and $Y \leq aI \leq X + Y$ for some scalar $a > 0$.

Using this fact, Theorem 2.4 can be stated without presence of unitaries U and V.

References

[1] M. Kian, M.S. Moslehian and R. Nakamoto, *Non-commutative Chebyshev inequality*, submitted.

[2] M. Kian, M.S. Moslehian and Yuki Seo, *Variants of Ando–Hiai inequality for operator power means*, Linear and Multilinear Algebra (2019), DOI: 10.1080/03081087.2019.1635981.

[3] M. Fujii, R. Nakamoto and M. Tominaga *Reverse of the grand Furuta inequality and its application*, Banach J. Math. Anal. **2** (2008), 23–30.

[4] M.S. Moslehian, J. Mićić and M. Kian, *An operator inequality and its consequences*, Linear Algebra Appl. **439** (2013), 584–591.

[5] R. Sharma, P. Devi and R. Kumari, *A note on inequalities for positive linear maps*, Linear Algebra Appl. **528** (2017), 113–123.

More accurate generalizations of Berezin number inequalities[1]

Rahmatollah Lashkaripour, Mojtaba Bakherad and Monire Hajmohamadi*

Department of Mathematics, Faculty of Mathematics, University of Sistan and Baluchestan, Zahedan, Iran

Abstract

In this paper, we generalize several Berezin number inequalities involving product of operators, which act on a Hilbert space $\mathcal{H}(\Omega)$.

Keywords: Numerical Berezin, Heinz mean, 2×2 matrices

Mathematics Subject Classification [2010]: 47A30, 47A12, 15A60

1 Introduction

Let $\mathbb{B}(\mathcal{H})$ denote the C^*-algebra of all bounded linear operators on a complex Hilbert space $\mathcal{H}$ with an inner product $\langle\,.\,,\,.\,\rangle$ and the corresponding norm $\|\,.\,\|$. In the case when $\dim\mathcal{H} = n$, we identify $\mathbb{B}(\mathcal{H})$ with the matrix algebra $\mathbb{M}_n$ of all $n \times n$ matrices with entries in the complex field. An operator $A \in \mathbb{B}(\mathcal{H})$ is called positive if $\langle Ax, x\rangle \geq 0$ for all $x \in \mathcal{H}$, and then we write $A \geq 0$.

A functional Hilbert space $\mathcal{H} = \mathcal{H}(\Omega)$ is a Hilbert space of complex valued functions on a (nonempty) set Ω, which has the property that point evaluations are continuous i.e. for each $\lambda \in \Omega$ the map $f \mapsto f(\lambda)$ is a continuous linear functional on $\mathcal{H}$. The Riesz representation theorem ensure that for each $\lambda \in \Omega$ there is a unique element $k_\lambda \in \mathcal{H}$ such that $f(\lambda) = \langle f, k_\lambda\rangle$ for all $f \in \mathcal{H}$. The collection $\{k_\lambda : \lambda \in \Omega\}$ is called the reproducing kernel of $\mathcal{H}$. If $\{e_n\}$ is an orthonormal basis for a functional Hilbert space $\mathcal{H}$, then the reproducing kernel of $\mathcal{H}$ is given by $k_\lambda(z) = \sum_n \overline{e_n(\lambda)}e_n(z)$; (see [4, Problem 37]). For $\lambda \in \Omega$, let $\hat{k_\lambda} = \frac{k_\lambda}{\|k_\lambda\|}$ be the normalized reproducing kernel of $\mathcal{H}$. For a bounded linear operator A on $\mathcal{H}$, the function $\widetilde{A}$ defined on Ω by $\widetilde{A}(\lambda) = \langle A\hat{k_\lambda}, \hat{k_\lambda}\rangle$ is the Berezin symbol of A, which firstly have been introduced by Berezin [2]. The Berezin set and the Berezin number of the operator A are defined by

$$\mathbf{Ber}(A) := \{\widetilde{A}(\lambda) : \lambda \in \Omega\} \qquad \text{and} \qquad \mathbf{ber}(A) := \sup\{|\widetilde{A}(\lambda)| : \lambda \in \Omega\},$$

respectively,(see [5]). The numerical radius of $A \in \mathbb{B}(\mathcal{H})$ is defined by $w(A) := \sup\{|\langle Ax, x\rangle| : x \in \mathcal{H}, \|x\| = 1\}$. It is clear that

$$\mathbf{ber}(A) \leq w(A) \leq \|A\| \tag{1}$$

[1]Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

*Speaker. Email address: monire.hajmohamadi@yahoo.com

for all $A \in \mathbb{B}(\mathcal{H})$. Moreover, The Berezin number of an operator A, B satisfies the following properties:
(i) $\mathbf{ber}(\alpha \mathrm{A}) = |\alpha|\mathbf{ber}(\mathrm{A})$ for all $\alpha \in \mathbb{C}$.
(ii) $\mathbf{ber}(A+B) \leq \mathbf{ber}(A) + \mathbf{ber}(B)$.
The authors in [1] showed some Berezin number inequalities as follows:

$$\mathbf{ber}(A^*XB) \leq \frac{1}{2}\mathbf{ber}(B^*|X|B + A^*|X^*|A), \tag{2}$$

$$\mathbf{ber}(AX \pm XA) \leq \mathbf{ber}^{\frac{1}{2}}(A^*A + AA^*)\mathbf{ber}^{\frac{1}{2}}(X^*X + XX^*),$$

and

$$\mathbf{ber}(A^*XB + B^*YA) \leq 2\sqrt{\|X\|\|Y\|}\mathbf{ber}^{\frac{1}{2}}(B^*B)\mathbf{ber}^{\frac{1}{2}}(AA^*) \tag{3}$$

for any $A, B, X, Y \in \mathbb{B}(\mathcal{H}(\Omega))$.
In this paper, we would like to state more extensions of Berezin number inequalities. Moreover, we obtain several Berezin number inequalities based on the 2×2 operator matrices.

2 Main results

Lemma 2.1. *Let $T \in \mathbb{B}(\mathcal{H})$ and $x, y \in \mathcal{H}$ be any vectors.*
(a) *If $0 \leq \alpha \leq 1$, then*

$$|\langle Tx, y\rangle|^2 \leq \langle |T|^{2\alpha}x, x\rangle\langle |T^*|^{2(1-\alpha)}y, y\rangle,$$

where $|T| = (T^*T)^{\frac{1}{2}}$ is the absolute value of T.
(b) *If f, g are nonnegative continuous functions on $[0,\infty)$ which are satisfying the relation $f(t)g(t) = t$ $(t \in [0,\infty))$, then*

$$|\langle Tx, y\rangle| \leq \|f(|T|)x\|\,\|g(|T^*|)y\|$$

for all $x, y \in \mathcal{H}$.

Lemma 2.2. *Let $A \in \mathbb{B}(\mathcal{H}_1(\Omega))$, $B \in \mathbb{B}(\mathcal{H}_2(\Omega), \mathcal{H}_1(\Omega))$, $C \in \mathbb{B}(\mathcal{H}_1(\Omega), \mathcal{H}_2(\Omega))$ and $D \in \mathbb{B}(\mathcal{H}_2(\Omega))$. Then the following statements hold:*
(a) $\boldsymbol{ber}\left(\begin{bmatrix} A & 0 \\ 0 & D \end{bmatrix}\right) \leq \max\{\boldsymbol{ber}(A), \boldsymbol{ber}(D)\}$;
(b) $\boldsymbol{ber}\left(\begin{bmatrix} 0 & B \\ C & 0 \end{bmatrix}\right) \leq \frac{1}{2}(\|B\| + \|C\|)$.

Theorem 2.3. *Let $A, B, X \in \mathbb{B}(\mathcal{H}(\Omega))$. Then*
(i) $\boldsymbol{ber}^{\mathrm{r}}(\mathrm{A^*XB}) \leq \|\mathrm{X}\|^{\mathrm{r}}\boldsymbol{ber}\left(\frac{1}{\mathrm{p}}(\mathrm{A^*A})^{\frac{\mathrm{pr}}{2}} + \frac{1}{\mathrm{q}}(\mathrm{B^*B})^{\frac{\mathrm{qr}}{2}}\right)$ *for $r \geq 0$ and $p, q > 1$ with $\frac{1}{p} + \frac{1}{q} = 1$ and $pr, qr \geq 2$.*
(ii) $\boldsymbol{ber}(\mathrm{A^*XB}) \leq \frac{1}{2}\boldsymbol{ber}(\mathrm{B^*|X|^{2\alpha}B + A^*|X^*|^{2(1-\alpha)}A})$ *for every $0 \leq \alpha \leq 1$.*

Theorem 2.4. *Suppose that $A, B, X \in \mathbb{B}(\mathcal{H}(\Omega))$ such that A, B are positive. Then*

$$\boldsymbol{ber}^r(A^\alpha XB^{1-\alpha}) \leq \|X\|^r\left(\boldsymbol{ber}(\alpha A^r + (1-\alpha)B^r) - \inf_{\|\hat{k}_\lambda\|=1}\eta(\hat{k}_\lambda)\right), \tag{4}$$

in which $\eta(\hat{k}_\lambda) = r_0(\langle A^r\hat{k}_\lambda, \hat{k}_\lambda\rangle^{\frac{1}{2}} - \langle B^r\hat{k}_\lambda, \hat{k}_\lambda\rangle^{\frac{1}{2}})^2$, $r_0 = \min\{\alpha, 1-\alpha\}$, $r \geq 2$ and $0 \leq \alpha \leq 1$.

Remark 2.5. Putting $A = B = I$ in inequality (4), we get a generalization of the inequality (1).

The Heinz mean is defined as $H_\alpha(a,b) = \frac{a^{1-\alpha}b^\alpha + a^\alpha b^{1-\alpha}}{2}$ for $a, b > 0$ and $0 \leq \alpha \leq 1$. The function H_α is symmetric about the point $\alpha = \frac{1}{2}$ and $\sqrt{ab} \leq H_\alpha(a,b) \leq \frac{a+b}{2}$ for all $\alpha \in [0,1]$.

Theorem 2.6. *Suppose that $A, B, X \in \mathbb{B}(\mathcal{H}(\Omega))$ such that A, B are positive. Then*

$$\begin{aligned}\boldsymbol{ber}^r\Big(\frac{A^\alpha X B^{1-\alpha} + A^{1-\alpha} X B^\alpha}{2}\Big) &\leq \frac{\|X\|^r}{2}\left(\boldsymbol{ber}(A^r + B^r) - 2\inf_{\|\hat{k}_\lambda\|=1}\eta(\hat{k}_\lambda)\right)\\ &\leq \frac{\|X\|^r}{2}\Big(\boldsymbol{ber}(\alpha A^r + (1-\alpha)B^r) + \boldsymbol{ber}((1-\alpha)A^r + \alpha B^r)\\ &\qquad - 2\inf_{\|\hat{k}_\lambda\|=1}\eta(\hat{k}_\lambda)\Big),\end{aligned}$$

in which $\eta(\hat{k}_\lambda) = r_0(\langle A^r\hat{k}_\lambda, \hat{k}_\lambda\rangle^{\frac{1}{2}} - \langle B^r\hat{k}_\lambda, \hat{k}_\lambda\rangle^{\frac{1}{2}})^2$, $r_0 = \min\{\alpha, 1-\alpha\}$, $r \geq 2$ and $0 \leq \alpha \leq 1$.

For positive operators $X, Y \in \mathcal{L}(\mathcal{H})$, the operator geometric mean is the positive operator $X\sharp Y = X^{\frac{1}{2}}\left(X^{-\frac{1}{2}}YX^{-\frac{1}{2}}\right)^{\frac{1}{2}}X^{\frac{1}{2}}$, In the next theorem we can obtain an upper bound for the Berezin number involving power geometric mean.

Theorem 2.7. *Let $X, Y, Z \in \mathbb{B}(\mathcal{H})$ be operators such that X, Y are positive. If $p \geq q > 1$ with $\frac{1}{p} + \frac{1}{q} = 1$, then*

$$ber^r\left((X\sharp Y)Z\right) \leq ber\left(\frac{X^{\frac{rp}{2}}}{p} + \frac{(Z^\star YZ)^{\frac{rq}{2}}}{q}\right) - \frac{1}{p}\inf_{\lambda\in\Omega}\left(\left[\widetilde{X}(\lambda)\right]^{\frac{rp}{4}} - \left[\widetilde{(Z^\star YZ)}(\lambda)\right]^{\frac{rq}{4}}\right)^2$$

for all $r \geq \frac{2}{q}$.

Corollary 2.8. *Let $X, Y \in \mathbb{B}(\mathcal{H})$ be positive operators and let $p \geq q > 1$ with $\frac{1}{p} + \frac{1}{q} = 1$. Then*

$$ber^r(X\sharp Y) \leq ber\left(\frac{X^{\frac{rp}{2}}}{p} + \frac{Y^{\frac{rq}{2}}}{q}\right) - \frac{1}{p}\inf_{\lambda\in\Omega}\left(\left[\widetilde{X}(\lambda)\right]^{\frac{rp}{4}} - \left[\widetilde{Y}(\lambda)\right]^{\frac{rq}{4}}\right)^2$$

for all $r \geq \frac{2}{q}$.

Corollary 2.9. *Let $X, Y \in \mathbb{B}(\mathcal{H})$ be positive operators. Then*

$$\sqrt{2}ber(X\sharp Y) \leq ber_2(X,Y) \leq ber^{\frac{1}{2}}\left(X^2 + Y^2\right).$$

Proposition 2.10. *Let $X, Y, Z \in \mathbb{B}(\mathcal{H})$ such that X, Y are positie and let $\frac{1}{p} + \frac{1}{q} = 1$. Then*

$$\begin{aligned}\|(X\sharp Y)Z\|_{ber}^r \leq &\left\|\frac{X^{\frac{rp}{2}}}{p}\right\|_{ber} + \left\|\frac{(Z^\star YZ)^{\frac{rq}{2}}}{q}\right\|_{ber}\\ &- \frac{1}{p}\inf_{\mu,\lambda\in\Omega}\left(\left\langle X\widehat{k}_\mu, \widehat{k}_\mu\right\rangle^{\frac{rp}{4}} - \left\langle Z^\star YZ\widehat{k}_\lambda, \widehat{k}_\lambda\right\rangle^{\frac{rq}{4}}\right)^2\end{aligned}$$

for all $r \geq \frac{2}{q}$.

In the following we state some Berezin number inequalities for 2×2 matrices.

Proposition 2.11. *Let* $T = \begin{bmatrix} 0 & B \\ C & 0 \end{bmatrix} \in \mathbb{B}(\mathcal{H}_1(\Omega) \oplus \mathcal{H}_2(\Omega))$ *and* f, g *be nonnegative continuous functions on* $[0,\infty)$ *satisfying the relation* $f(t)g(t) = t$ $(t \in [0,\infty))$. *Then*

$$\boldsymbol{ber}^r(T) \leq \max\left\{ \boldsymbol{ber}\left(\frac{1}{p}f^{pr}(|\ C\ |) + \frac{1}{q}g^{qr}(|\ B^*\ |)\right), \boldsymbol{ber}\left(\frac{1}{p}f^{pr}(|\ B\ |) + \frac{1}{q}g^{qr}(|\ C^*\ |)\right)\right\}, \tag{5}$$

in which $r \geq 1$, $p \geq q > 1$ *such that* $\frac{1}{p} + \frac{1}{q} = 1$ *and* $pr \geq 2$.

Proposition 2.12. *Let* $T = \begin{bmatrix} A & 0 \\ 0 & D \end{bmatrix} \in \mathbb{B}(\mathcal{H}_1(\Omega) \oplus \mathcal{H}_2(\Omega))$. *Then*

$$\boldsymbol{ber}^r(T) \leq \frac{1}{2}\max\{\boldsymbol{ber}(|A|^r + |A^*|^r), \boldsymbol{ber}(|D|^r + |D^*|^r)\} \tag{6}$$

for $r \geq 1$.

Corollary 2.13. *Let* $T = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$ *with* $A, B, C, D \in \mathbb{B}(\mathcal{H})$. *Then*

$$\boldsymbol{ber}(T) \leq \frac{1}{2}\max\{\boldsymbol{ber}(|C| + |B^*|), \boldsymbol{ber}(|B| + |C^*|)\} + \frac{1}{2}\max\{\boldsymbol{ber}(|A| + |A^*|), \boldsymbol{ber}(|D| + |D^*|)\}.$$

In particular,

$$\boldsymbol{ber}\left(\begin{bmatrix} A & B \\ B & A \end{bmatrix}\right) \leq \frac{1}{2}(\boldsymbol{ber}(|A| + |A^*|) + \boldsymbol{ber}(|B| + |B^*|)).$$

References

[1] M. Bakherad and M.T. Karaev, *Berezin number inequalities for Hilbert space operators*, Concrete Operator (to appear).

[2] F.A. Berezin, *Covariant and contravariant symbols for operators,* Math. USSR-Izv. **6** (1972), 1117–1151.

[3] M. Hajmohamadi, R. Lashkaripour and M. Bakherad, *Improvements of Berezin number inequalities*, Linear and Multilinear Algebra Journal doi:10.1080/03081087.2018.1538310.

[4] P.R. Halmos, *A Hilbert Space Problem Book*, 2nd ed., springer, New York, 1982.

[5] M.T. Karaev, *Berezin symbol and invertibility of operators on the functional Hilbert spaces*, J. Funct. Anal. **238** (2006), 181–192.

Max-spectral radius of products for non-negative matrices[1]

S. Mahmoud Manjegani and Hojr Shokooh Saljooghi*

Department of Mathematical Sciences at Isfahan University of Technology, Iran

Abstract

Several authors have proved inequalities on the spectral radius, operator norm and numerical radius of Hadamard products and ordinary products of non-negative matrices. The aim of this paper is to investigate and study the max-spectral radius inequalities for Hadamard, conventional and max-products of non-negative matrices.

Keywords: Max-algebra, Max-eigenvalue, Max-spectral radius

Mathematics Subject Classification [2010]: 15A18, 15A48, 15A80

1 Introduction

In recent years both industry and the academic world have become more and more interested in techniques to model, to analyse problems. One of these tools is max-algebra. The max-algebra is a subdivision of mathematics, which has many applications. max-algebra system has been studied in research papers and books from the early 1960's. The *max algebra system* consists of the non-negative real numbers $\mathbb{R}_+$ equipped with the operation of multiplication $a \otimes b = ab$, and maximization $a \oplus b = \max\{a, b\}$. Furthermore, the max algebra is isomorphic to the *max-plus algebra*, which consists of the set $\mathbb{R}_+\{-\infty\}$ with operations of maximization and addition [2, 4, 6].This algebra system and its isomorphic version raise the possibility of changing the non-linear phenomena in different areas such as parallel computation, transportation networks, timetabled programs, IT, dynamic systems, combinatorial optimization, and mathematical physics to linear-algebra. Furthermore, this algebra system has been used directly in areas such as algorithm,Vetrbi, analysing DNA and in AHP for ranking matrices. In this algebra system, there is no deduction but many of appeared problems in linear algebra like equation systems, eigenvalue, projections, subspaces, singular value decomposition, duality theory have developed and have reached other areas like functional analysis, algebra topology and combinatorial optimization.

Let $M_n(\mathbb{R})$ be the set of all $n \times n$ real matrices and $M_n(\mathbb{R}_+)$ be the set of all $n \times n$ non-negative matrices. Let A and B be two matrices in $M_n(\mathbb{R}_+)$. We say that $A \leq B$ if $a_{ij} \leq b_{ij}$ for all $i, j = 1, 2, \ldots, n$. The max-product $A \otimes B$ and max-sum $A \oplus B$ defined as follows

$$(A \otimes B)_{ij} = \max_k a_{ik} b_{kj}, \quad (A \oplus B)_{ij} = \max\{a_{ij}, b_{ij}\}.$$

[1]Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

*Speaker. Email address: h.shokooh@math.iut.ac.ir

The notation $A_{\otimes}^{k}$ denotes the k^{th} power of A. For non-negative vector $x \in \mathbb{R}^n$, the notation $A \otimes x$ means $(A \otimes x)_i = \max\limits_{1\leq k\leq n} a_{ij}x_j$. A non-negative scalar λ is called a max-eigenvalue of A if $A \otimes x = \lambda x$ for some non-negative vector $x \neq 0$. The set of all max-eigenvalues of A is denoted by $\sigma_{\otimes}(A)$. let $\|A\| = \|A\|_{\infty} = \max_{i,j} a_{ij}$ and $\|x\| = \|x\|_{\infty} = \max_i x_i$. In linear algebra, the spectral radius plays a key role in a variety of areas, including the stability theory of difference and differential inclusions and wavelet analysis. In this paper, we extend the work described and study properties of max-spectral radius inequality.

The max-spectral radius of $A \in M_n(\mathbb{R}_+)$ is denoted by $r_{\otimes}(A)$ and defined by the maximum cycle geometric mean $r_{\otimes}(A)$, which is defined by

$$r_{\otimes}(A) = \max\Big\{ \sqrt[k]{a_{i_1 i_k} \cdots a_{i_3 i_2} a_{i_2 i_1}} : k \leq n \ \text{ and } \ i_1, \ldots, i_k \in \{1, \ldots, n\} \ \text{ mutually distinct} \Big\}.$$

It is known that $r_{\otimes}(A)$ is the largest max-eigenvalue of A, that is

$$r_{\otimes}(A) = \max\{\lambda : \lambda \in \sigma_{\otimes}(A)\}.$$

The max version of the Gelfand formula holds for any $A \in M_n(\mathbb{R}_+)$ which is

$$r_{\otimes}(A) = \lim_{j\to\infty} \|A_{\otimes}^{j}\|^{1/j} = \inf_{j\in\mathbb{N}} \|A_{\otimes}^{j}\|^{1/j}.$$

For vector $x \in \mathbb{R}_+^n$, the local max-spectral radius of A at x is defined by

$$r_x(A) = \lim_{j\to\infty} \|A_{\otimes}^{j} \otimes x\|^{1/j}.$$

Lemma 1.1. *If A and B are two non-negative matrices with $A \leq B$ and $x \in \mathbb{R}_+^n$, then*

$$r_{\otimes}(A) \leq r_{\otimes}(B), \quad r_x(A) \leq r_x(B).$$

Proof. The proof is straightforward. □

Lemma 1.2. *[5] Let $A \in \mathbb{R}_+^{n\times n}$, $j \in \{1, \ldots, n\}$. Then $r_{e_j}(A)$ is maximum of all $t \geq 0$ with the following property $(*)$:*
there exist $a \geq 0$, $b \geq 1$ and mutually distinct indices $i_0 := j, i_1, \ldots, i_a, i_{a+1}, \ldots, i_{a+b-1} \in \{1, \ldots, n\}$ such that

$$\prod_{s=0}^{a-1} A_{i_{s+1},i_s} \neq 0 \ \text{ and } \ \prod_{s=a}^{a+b-1} A_{i_{s+1},i_s} = t^b$$

where we set $i_{a+b} = i_a$.

Theorem 1.3. *[5] Let $A \in M_n(\mathbb{R}_+)$ and $x \in \mathbb{R}_+^n$ be a non-zero vector. Then*

1. $r_{\otimes}(A) = \max\{r_{e_j}(A) : \quad 1 \leq j \leq n, \quad x_j \neq 0\}$.
2. $\sigma_{\otimes}(A) = \{r_{e_j}(A) : \quad 1 \leq j \leq n\}$.

2 Main results

The spectral radius $\rho(A)$ of $A \in M_n(\mathbb{C})$ is the largest modulus of $\sigma(A)$. The spectral radius is not sub-multiplicative that is $\rho(AB) \leq \rho(A)\rho(B)$ does not hold in general, not even for non-negative matrices [1]. On the other hand, for non-negative A and B, the spectral radius is sub-multiplicative with respect to the Hadamard product: $\rho(A \circ B) \leq \rho(A)\rho(B)$.

In 2009, X. Zhan conjectured that for non-negative $n \times n$ matrices A and B, the spectral radius $\rho(A \circ B)$ of the Hadamard product satisfies

$$\rho(A \circ B) \leq \rho(AB),$$

where AB denotes the conventional matrix product of A and B. This conjecture was confirmed by K.M.R. Audenaert as follows

$$\rho(A \circ B) \leq \rho^{\frac{1}{2}}((A \circ A)(B \circ B)) \leq \rho(AB).$$

These inequalities were established via a trace description of the spectral radius. Using the fact that the Hadamard product is a principal sub matrix of the Kronecker product. R.A. Horn and F. Zhang proved in 2010 the inequalities

$$\rho(A \circ B) \leq \rho^{\frac{1}{2}}(AB \circ BA) \leq \rho(AB).$$

We are interested in answering the following questions. From now we will use $\mu(A)$ for $r_{\otimes}(A)$.

1. $\mu(AB) \leq \mu(A)\mu(B)$?
2. $\mu(A \circ B) \leq \mu(A)\mu(B)$?
3. $\mu(A \otimes B) \leq \mu(A)\mu(B)$?
4. $\mu(A \otimes B) \leq \mu(AB)$?
5. $\mu(A \circ B) \leq \mu(AB)$?
6. $\mu(A \circ B) \leq \mu(A \otimes B)$?
7. $\mu(A \circ B) \leq \mu(A \otimes B \circ B \otimes A)^{1/2} \leq \mu(A \otimes B)$?
8. $\mu(A \circ B) \leq \mu(AB \circ BA)^{1/2} \leq \mu(AB)$?

It is well known that $\rho(AB) = \rho(BA)$ for every two $n \times n$ matrices A and B. The following example shows that this not true for max-spectral radius.

Example 2.1. Let $A = \begin{bmatrix} 1 & 2 \\ 0 & 0 \end{bmatrix}$, and $B = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$. Then $AB = \begin{bmatrix} 5 & 8 \\ 0 & 0 \end{bmatrix}$, and $BA = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$. Thus, $\mu(AB) = 5 \neq \mu(BA) = 4$.

Conjecture: Let $A, B \in M_n(\mathbb{R}^n_+)$. Then $\mu(A \otimes B) = \mu(B \otimes A)$.

In the following example we show that $\mu(A \otimes B) \leq \mu(A)\mu(B)$ does not hold. Of course, $\mu(A^k_{\otimes}) = \mu(A)^k$ for every $k \in \mathbb{N}$.

Example 2.2. Let $A = \begin{bmatrix} 1 & 25 \\ 1 & 1 \end{bmatrix}$, $B = \begin{bmatrix} 4 & 1 \\ 1 & 1 \end{bmatrix}$ be given. Then $A \otimes B = \begin{bmatrix} 25 & 25 \\ 4 & 1 \end{bmatrix}$, So

$$\mu(A)\mu(B) = 5 \times 4 < 25 = \mu(A \otimes B).$$

Example 2.3. Let $A = \begin{bmatrix} 5 & 1 \\ 1 & 1 \end{bmatrix}$, $B = \begin{bmatrix} 2 & 2 \\ 1 & 3 \end{bmatrix}$ be given. Then $A \otimes B = \begin{bmatrix} 10 & 10 \\ 2 & 3 \end{bmatrix}$, So

$$\mu(A)\mu(B) = 5 \times 3 > 5 = \mu(A \otimes B).$$

The following examples show that there is no relation between $\mu(AB)$ and $\mu(A)\mu(B)$.

Example 2.4. Let $A = \begin{bmatrix} 1 & 2 \\ 0 & 0 \end{bmatrix}$, $B = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ be given. Then $AB = \begin{bmatrix} 7 & 10 \\ 0 & 0 \end{bmatrix}$, So

$$\mu(A)\mu(B) = 1 \times 4 < 7 = \mu(AB)$$

Example 2.5. Let $A = \begin{bmatrix} \frac{1}{2} & \frac{1}{3} \\ 0 & 0 \end{bmatrix}$, $B = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ be given. Then $AB = \begin{bmatrix} \frac{1}{3} & \frac{1}{2} \\ 0 & 0 \end{bmatrix}$, So

$$\mu(AB) = \frac{1}{3} < \frac{1}{2} \times 1 = \mu(A)\mu(B)$$

Theorem 2.6. *Let A, B be $n \times n$ non-negative matrices, then*

$$\mu(A \otimes B) \leq \mu(AB).$$

Proof. The entries of $A \otimes B$ are as $\max\{a_{ik}b_{kj}\}$ while entries of AB are as liner combination of a_{ik}, b_{kj}. Since a_{ik}, b_{kj} are non-negative, so $a_{ik}b_{kj} \geq \max\{a_{ik}, b_{kj}\}$.Therefore $\mu(A \otimes B) \leq \mu(AB)$ by Lemma 1.1. □

Theorem 2.7. *Let A, B be $n \times n$ non-negative matrices. Then*

$$\mu(A \circ B) \leq \mu(A \otimes B).$$

Proof. It is direct result of definition $A \circ B$, $A \otimes B$ and definition of max-spectral radius. □

Corollary 2.8. *Let $A, B \in M_n(\mathbb{R}^n_+)$. Then*

$$\mu(A \circ B) \leq \mu(AB).$$

Theorem 2.9. *[3] If $A, B \in M_n(\mathbb{R}_+)$, then*

$$\mu(A \circ B) \leq \mu^{\frac{1}{2}}(A \otimes B \circ B \otimes A) \leq \mu(A \otimes B).$$

Question: Let A, B be $n \times n$ two non-negative matrices, then

$$\mu(A \circ B) \leq \mu^{\frac{1}{2}}(AB \circ BA) \leq \mu(AB)?$$

Example 2.10. $A = \begin{bmatrix} 4 & 2 \\ 1 & 1 \end{bmatrix}$, $B = \begin{bmatrix} 4 & 1 \\ 2 & 1 \end{bmatrix}$ be given. Then $A \otimes B = \begin{bmatrix} 16 & 4 \\ 4 & 1 \end{bmatrix}$, $AB = \begin{bmatrix} 20 & 6 \\ 6 & 2 \end{bmatrix}$, and $AB \circ BA = \begin{bmatrix} 340 & 54 \\ 54 & 10 \end{bmatrix}$. Thus

$$\mu(A \circ B) = 16 < \sqrt{340} = \mu^{\frac{1}{2}}(AB \circ BA) < 20 = \mu(AB)$$

Example 2.11. $A = \begin{bmatrix} 4 & 2 \\ 3 & 1 \end{bmatrix}$, $B = \begin{bmatrix} 4 & 3 \\ 2 & 1 \end{bmatrix}$ be given. Then $A \otimes B = \begin{bmatrix} 16 & 12 \\ 12 & 1 \end{bmatrix}$, $AB = \begin{bmatrix} 20 & 14 \\ 14 & 10 \end{bmatrix}$, and $AB \circ BA = \begin{bmatrix} 500 & 154 \\ 154 & 50 \end{bmatrix}$. Thus

$$\mu(A \circ B) = 16 < 20 = \mu(AB) < \sqrt{500} = \mu^{\frac{1}{2}}(AB \circ BA),$$

but $\mu(AB) = 20$.

Lemma 2.12. *Let A, B be $n \times n$ two non-negative matrices and $x \in \mathbb{R}^n_+$, then*

$$\mu(A \circ B) \leq \mu^{\frac{1}{2}}(AB \circ BA), \quad r_x(A \circ B) \leq r_x^{\frac{1}{2}}(AB \circ BA).$$

Proof. Since A and B are non-negative matrices, by definition Hadamard and conventional products of matrices and definition max-spectral radius, we get the result by Lemma 1.1. □

Acknowledgement

This work was supported by the Department of Mathematical Sciences at Isfahan University of Technology, Iran.

References

[1] Koenraad M.R. Audenaert, Spectral radius of Hadamard product versus conventional product for non-negative matrices. Linear Algebra and its Appl., 432.1 (2010): 366-368.

[2] R.B. Bapat,D.P. Stanford, P. van den Driessche, Pattern properties and spectral inequalities in max algebra, SIAM J. of Matrix Analysis and Appl., 16 (1995) 964-976.

[3] Peperko, Aljoa. Bounds on the generalized and the joint spectral radius of Hadamard products of bounded sets of positive operators on sequence spaces. Linear Algebra and its Applications 437.1 (2012): 189-201.

[4] L. Elsner, P. Van Driessche, Modifying the method in max algebra, Linear Algebra Appl., 332-334 (2001) 3-13.

[5] V. Muller, A. Peperko, On the spectrum in max-algebra, Linear Algebra Appl., 485 (2015), 250-266.

[6] H. Shokoh Saljooghi, Spectral Properties of Matrix Polynomials in the Max Algebra [master thesis], Shahid Bahonar University of Kerman, 2012.

Constructing cross sectional area of vibrating rod using two spectra[1]

Hanif Mirzaei*

Department of Mathematics, Faculty of basic sciences, Sahand University of Technology, Tabriz, Iran

Abstract

In this research, the construction of non-symmetric cross sectional area of vibrating rod is proposed. For this purpose, using finite difference method, we discretized the rod equation to a Jacobi matrix eigenvalue problem. Then, with correction of given spectra and using Lancsoz method, we construct the Jacobi matrix. Finally, according to the relation between the entries of Jacobi matrix and cross sectional area, we obtain the cross sectional area at different points. Some numerical examples are given.

Keywords: Rod equation, Lancsoz method, Discretization, Jacobi matrix

1 Introduction

Free vibration of a free-fix straight rod of unit length is governed by the following eigenvalue problem:

$$\begin{cases} (a(x)y'(x))' + \lambda a(x)y(x) = 0, & x \in (0,1), \\ y'(0) = 0, \ y(1) = 0, \end{cases} \tag{1}$$

where $a(x) > 0$ is the cross sectional area at point x, λ is the eigenvalue, $y(x)$ is the displacement of an element dx [5]. It is proved that, problem (1) has infinite number of eigenvalues which are distinct, nonnegative and can be ordered as follows

$$0 \leq \lambda_1 \leq \lambda_2 \leq \cdots, \lim_{k\to\infty} \lambda_k = \infty. \tag{2}$$

see [5]. The set of all eigenvalues of the problem (1) is called the spectrum and denoted by $\sigma(a(x), \infty, 0)$. The construction of $a(x)$ from spectral data (eigenvalues, eigenfunctions or both) are studied in different papers [2–4]. In general, if $a(x)$ is symmetric respect to mid point, then $a(x)$ can be constructed using one spectrum. Otherwise, two spectra corresponding to, two set of boundary conditions are needed. On the other words, for constructing a non-symmetric $a(x)$, we need two spectra $\sigma(a(x), \infty, 0)$ and $\sigma(a(x), 0, 0)$. The spectrum of the problem (1) is an infinite sequence of nonnegative real numbers, but in practice, only the first few eigenvalues are given. In this paper, we try to solve the following problem:

Main problem: Given two finite sequence $\{\lambda_i\}_{i=1}^{N}$ and $\{\mu_i\}_{i=1}^{N-1}$. Construct the cross sectional area $a(x)$ such that, $\{\lambda_i\}_{i=1}^{N}$ and $\{\mu_i\}_{i=1}^{N-1}$ are primitive eigenvalues of the spectra $\sigma(a(x), \infty, 0)$ and $\sigma(a(x), 0, 0)$, respectively.

[1] Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

*Speaker. Email address: h_mirzaei@sut.ac.ir

2 Main results

In this section, we try to solve the **Main problem**. For this purpose, using the idea of [2,3], we obtain the required data of well known Lancsoz method for approximating $a(x)$. Using finite difference method, problem (1) reduces to the following matrix eigenvalue problem:

$$KY = \lambda MY, \quad Y = [y_1, \cdots, y_N], \tag{3}$$

where,

$$K = \begin{bmatrix} a_1 & -a_1 & & & & \\ -a_1 & a_1+a_2 & a_2 & & & \\ & \ddots & \ddots & \ddots & & \\ & & -a_{N-2} & a_{N-2}+a_{N-1} & -a_{N-1} & \\ & & & -a_{N-1} & a_{N-1}+a_N \end{bmatrix},$$

$$M = diag(\frac{a_1}{2}, \frac{a_1+a_2}{2}, \cdots, \frac{a_{N-1}+a_N}{2}).$$

and $h = \frac{1}{N}$, $x_i = ih$, $a_i = a(x_i - \frac{h}{2})$, $y_i = y(x_i)$. The cross sectional area $a(x)$ is nonzero thus, M is nonsingular and problem (4) can be written as follows:

$$JX = \lambda X, \quad J = D^{-1}KD^{-1}, \quad X = DY, D = M^{\frac{1}{2}}. \tag{4}$$

For solving **Main problem** first, we construct the Jacobi matrix J, then according to the relations between the entries of J and cross sectional area, we obtain a_i.
We denote the spectrum of J by $\sigma(J)$. For constructing J, two spectra $\sigma(J)$ and $\sigma(J_1)$ or $\sigma(J_N)$ are needed [5], where J_i is the matrix obtained from J by deleting ith row and column. Now, this question arise that: How we can obtain the spectra $\sigma(J)$ and $\sigma(J_1)$ from the given data $\{\lambda_i\}_{i=1}^N$ and $\{\mu_i\}_{i=1}^{N-1}$?
Problem (4) is the approximation of problem (1) thus, the eigenvalues of J can be approximated by the first N eigenvalues of the problem (1). But, it is observed in Tables 1 and 2 that only a few lower eigenvalues of the problem (1) are a good approximation for the eigenvalues of matrices J and J_1. On the other words, if we denote the eigenvalues of J and problem (1) by $\lambda_i^*(a)$ and $\lambda_i(a)$, respectively, then $|\lambda_i(a) - \lambda_i^*(a)|$ is an increasing sequence.

Let $\epsilon_i = \lambda_i(\mathbf{a}) - \lambda_i^*(\mathbf{a})$ and $\delta_i = \mu_i(\mathbf{a}) - \mu_i^*(\mathbf{a})$, where $a(x) = \mathbf{a}$ is a constant and μ_i, μ_i^* are the eigenvalues of the problem (1) with fix-fix boundary conditions and corresponding matrix J_1, respectively. We observe that $\lambda_i(a(x)) - \epsilon_i$ and $\mu_i(a(x)) - \delta_i$ are good approximations for $\lambda_i^*(a(x))$ and $\mu_i^*(a(x))$, respectively via [1] and references there in, such that we have

$$|\lambda_i - \epsilon_i - \lambda_i^*| = O(ih^2), \qquad |\mu_j - \delta_j - \mu_j^*| = O(jh^2).$$

Thus, we can construct the matrix J using Lancsoz method [5] such that $\sigma(J) = \{\lambda_i(a(x)) - \epsilon_i\}_{i=1}^N$ and $\sigma(J_1) = \{\mu_i(a(x)) - \delta_i\}_{i=1}^{N-1}$. In the following example, we solve the **Main problem** by our method for the cases $a_1(x) = e^x$ and $a_2(x) = (1+x)^2$.

Example 2.1. Suppose that $a_1(x) = e^x$ and $a_2(x) = (1+x)^2$. We compute the eigenvalues $\{\lambda_i\}_{i=1}^N$ and $\{\mu_i\}_{i=1}^{N-1}$ using Matslise package [6]. For the case $a(x) = constant$, we have

$$\lambda_i = (i-0.5)^2\pi^2, \quad \lambda_i^* = \frac{2}{h^2}(1 - cos((i-0.5)\pi h)), \quad \mu_i = i^2\pi^2, \quad \mu_i^* = \frac{2}{h^2}(1 - cos(i\pi h)).$$

Thus we can compute $\sigma(J) = \{\lambda_i - \epsilon_i\}$ and $\sigma(J_1) = \{\mu_i - \delta_i\}$. The numerical results of our method for $a_1(x)$ and $a_2(x)$ are given in the Table 3 and Figures 1 and 2.

Table 1: Errors of the eigenvalues of problem (1) for $a(x) = e^x$ with $N = 20$

λ_i	$\lvert\lambda_i - \lambda_i^*\rvert$	$\lvert\lambda_i - \epsilon_i - \lambda_i^*\rvert$	μ_i	$\lvert\mu_i - \mu_i^*\rvert$	$\lvert\mu_i - \delta_i - \mu_i^*\rvert$
3.6231	5.9e-3	4.6-3	10.1196	2.3e-2	3.1-3
23.4423	1.3e-1	3.5e-2	39.7284	3.3e-1	1.2-2
62.9297	8.8e-1	9.5e-2	89.0764	1.6e+0	2.7-2
122.1499	3.1e+0	1.8e-1	158.1637	5.1e+0	4.7-2
201.1078	8.4e+0	2.0e-1	246.9901	1.2e+1	7.3-2
299.8044	1.8e+1	4.3e-1	355.5558	2.5e+1	1.0-1
418.2400	3.5e+1	5.9e-1	483.8606	4.6e+1	1.4-1
556.4146	6.2e+1	7.7e-1	631.9046	7.9e+1	1.7-1
714.3284	1.0e+2	9.5e-1	799.6879	1.2e+2	2.1-1
891.9814	1.5e+2	1.1e+0	987.2104	1.8e+2	2.4-1
1089.3736	2.2e+2	1.3e+0	1194.4721	2.6e+2	2.8-1
1306.5049	3.2e+2	1.5e+0	1421.4730	3.7e+2	3.2-1
1543.3755	4.3e+2	1.7e+0	1668.2131	5.1e+2	3.6-1
1799.9852	5.8e+2	1.9e+0	1934.6924	6.6e+2	3.9-1
2076.3342	7.5e+2	2.0e+0	2220.9109	8.5e+2	4.2-1
2372.4223	9.6e+2	2.1e+0	2526.8687	1.1e+3	4.5-1
2688.2497	1.2e+3	2.3e+0	2852.5656	1.3e+3	4.7-1
3023.8162	1.5e+3	2.4e+0	3198.0018	1.6e+3	4.8-1
3379.1220	1.8e+3	2.4e+0	3563.1771	1.9e+3	4.9-1
3754.1670	2.1e+3	2.4e+0			

Table 2: Errors of the eigenvalues of problem (1) for $a(x) = (1 + x)^2$ with $N = 20$

λ_i	$\lvert\lambda_i - \lambda_i^*\rvert$	$\lvert\lambda_i - \epsilon_i - \lambda_i^*\rvert$	μ_i	$\lvert\mu_i - \mu_i^*\rvert$	$\lvert\mu_i - \delta_i - \mu_i^*\rvert$
4.1158	8.7e-3	7.4-3	9.8696	2.0e-2	1.4-4
24.1393	1.5e-1	5.6e-2	39.4784	3.2e-1	1.6e-4
63.6591	9.4e-1	1.5e-1	88.8264	1.6e+0	1.5e-4
122.8891	3.3e+0	2.9e-1	157.9136	5.1e+0	1.4e-4
201.8512	8.6e+0	4.7e-1	246.74011	1.2e+1	1.2e-4
300.5499	1.8e+1	7.0e-1	355.3057	2.5e+1	1.1e-4
418.9868	3.5e+1	9.5e-1	483.6106	4.6e+1	8.1e-5
557.1622	6.2e+1	1.2e+0	631.6546	7.8e+1	5.5e-5
715.076	1.0e+2	1.5e+0	799.4379	1.2e+2	2.8e-5
892.7299	1.5e+2	1.8e+0	986.9604	1.8e+2	0
1090.1223	2.2e+2	2.1e+0	1194.2221	2.6e+2	2.8e-5
1307.2539	3.2e+2	2.4e+0	1421.2230	3.7e+2	5.5e-5
1544.1246	4.4e+2	2.7e+0	1667.9631	5.0e+2	8.1e-5
1800.7344	5.8e+2	3.0e+0	1934.4424	6.6e+2	1.1e-4
2077.0835	7.6e+2	3.3e+0	2220.6609	8.5e+2	1.2e-4
2373.1717	9.7e+2	3.5e+0	2526.6187	1.1e+3	1.4e-4
2688.9991	1.2e+3	3.7e+0	2852.3156	1.2e+3	1.5e-4
3024.5657	1.5e+3	3.8e+0	3197.7518	1.6e+3	1.6e-4
3379.8716	1.8e+3	3.9e+0	3562.9271	2.0e+3	1.4e-4
3754.9166	2.1e+3	3.6e+0			

Table 3: Numerical errors for $a_1(x) = e^x$ and $a_2(x) = (1+x)^2$

N	$\| a_1(x) - a_i \|_\infty$	$\| a_2(x) - a_i \|_\infty$
10	5.28e-02	1.06e-01
20	2.57e-02	5.16e-02
30	2.00e-02	3.40e-02
40	1.47e-03	2.53e-02

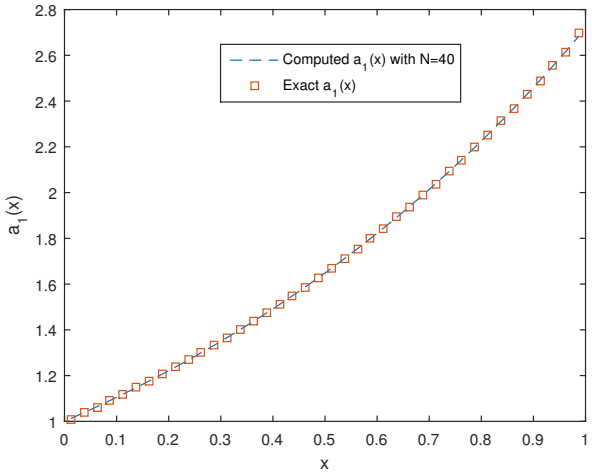


Figure 1: Results for $a_1(x)$ with $N = 40$.

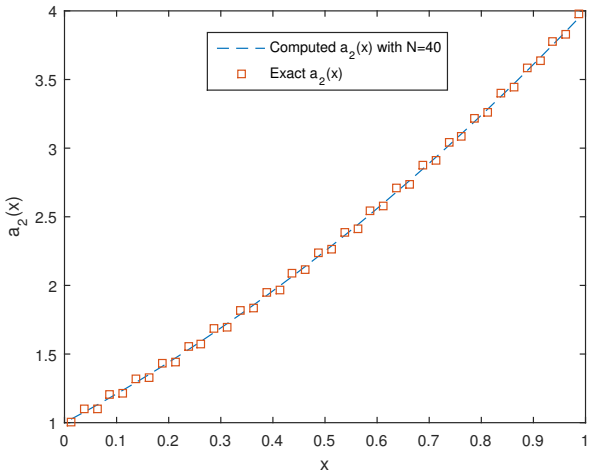


Figure 2: Results for $a_2(x)$ with $N = 40$.

3 Conclusion

In this paper, the cross sectional area of vibrating rod using two spectra and Lancsoz method is constructed. We observe that the correction terms ϵ_i and δ_i play an important role in the construction procedure.

References

[1] A. L. Andrew, *Asymptotic correction of more Sturm-Liouville eigenvalue estimates*, BIT Numerical Mathematics, 43 (2003), 485503.

[2] Q. Gao, Z. Huang and X. Cheng, *A finite difference method for an inverse Sturm-Liouville problem in impedance form*, Numer. Algor., 70 (2015), 669-690.

[3] Q. Gao, Q. Zhao, X. Zheng and Y. Ling, *Convergence of Numerovs method for inverse SturmLiouville problems*, Applied Mathematics and Computation, 293 (2017), 1-17.

[4] K. Ghanbari, H. Mirzaei, M. G. Gladwell, *Reconstructing of a Rod using one spectrum and minimal mass condition*, Inverse problem in science and engineering, 22 (2014), 325-333.

[5] G.M.L. Gladwell, *Inverse problem in vibration*, Kluwer academic publishers, New York, 2004.

[6] V. Ledoux, M.V. Daele and G.V. Berghe, *Matslise: A Matlab package for the numerical solution of Sturm-Liouville and Schrodinger equations*, ACM Translations on Mathematical Software, 31 (2005), 532-554.

Inverse problem for H-symmetric pentadiagonal matrices[1]

Hanif Mirzaei and Kazem Ghanbari*

Faculty of Basic Sciences, Department of Mathematics, Sahand University of Technology, Tabriz, Iran

Abstract

The set of eigenvalues of a square matrix P is denoted by $\sigma(P)$, and the set of eigenvalues of the submatrix obtained from P by deleting the first i rows and columns of P is denoted by $\sigma_i(P)$. In this paper we solve the inverse eigenvalue problem for H-Symmetric pentadiagonal matrices, not necessarily symmetric. Using $\sigma, \sigma_1, \sigma_2$, not necessarily interlacing, we construct the solution by modified Lancsoz algorithm.

Keywords: Inverse eigenvalue problem, Modified Lanczos algorithm, H-symmetric matrices

Mathematics Subject Classification [2010]: 15A18, 65F18

1 Introduction

Let $H = \text{diag}(\delta_1, \delta_2, \cdots, \delta_n)$ be a diagonal matrix such that $H^2 = I$, where I is identity matrix. In $\mathbb{C}^n$ we define the inner product $[x, y]_H = y^* H x$. An square real matrix P is said to be H-Symmetric if $HP^T H = P$. The matrix $HP^T H$ is called $H-$adjoint of P and it is denoted by $P^\sharp$. It is proved that if P is H-Symmetric and $\sigma(P)$ is real and disjoint, then the eigenvectors are H-Orthonormal, i.e. there are eigenvectors $u_1, u_2, \ldots, u_n$ of P such that

$$HU^T HU = I, \qquad \text{where } U = [u_1, u_2, \cdot, u_n].$$

See [1, 2]. It can be easily verified that the pentadiagonal real matrix P of the form

$$P = \begin{pmatrix} a_1 & \epsilon_1 b_1 & \epsilon_1\epsilon_2 c_1 & & & \\ b_1 & a_2 & \epsilon_2 b_2 & \epsilon_2\epsilon_3 c_2 & & \\ c_1 & b_2 & a_3 & \cdots & & \\ & \ddots & \ddots & \ddots & \ddots & \epsilon_{n-2}\epsilon_{n-1}c_{n-2} \\ & & & & & \epsilon_{n-1}b_{n-1} \\ & & & c_{n-2} & b_{n-1} & a_n \end{pmatrix}, \tag{1}$$

for $H = \text{diag}(1, \epsilon_1, \epsilon_1\epsilon_2, \cdots, \prod_{i=1}^{n-1} \epsilon_i)$ is a H-Symmetric matrix. If $H = I$ then P is pentadiagonal symmetric matrix. Pentadiagonal symmetric matrices arise in discrete vibrating beams, see [4] for more detals. Inverse eigenvalue problems for symmetric pentadiagonal

[1] Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

*Speaker. Email address: kghanbari@sut.ac.ir

matices are studied by many authors, for example see [3,4]. In most of the cases the reconstruction procedure require three interlacing real spectra. Using finite difference metheod for discretizing non-smooth beams may lead to a nonsymmetric stiffness matrix, see [6]. H-Symmetric pentadiagonal matrices of the form (1) appear in non-hermitian quantum mechanics [5]. The objective of this paper is to study the inverse eigenvalue problem for H-Symmetric pentadiagonal matrices. Indeed using three given spectrum that may or may not have interlacing property, we construct H-Symmetric pentadiagonal matrices of the form (1) such that $\sigma(P), \sigma_1(P)$ and $\sigma_2(P)$ are the prescribed spectrums. We use the modified form of Lancsoz algoritm to construct the solution and we prove that the solution is not unique. The solution obtained by this algorithm produce eigenvectors that for large size matrices may not be H-Orthonormal. To resolve this case we use a modified gram schmidt orthogonalization procedure to make eigenvactors to be H-Orthonormal.

2 Construction of the solution

In this section, we state the main inverse eigenvalue problem and construct the solution. We consider conditions on the given data for which this problem has solution.

Inverse Problem. Given three real spectrum $\{\lambda_i\}_{i=1}^n, \{\mu_i\}_{i=1}^{n-1}, \{\nu_i\}_{i=1}^{n-2}$, construct a H-Symmetric pentadiagonal matrix P of the form (1) such that

$$\sigma(P) = \{\lambda_i\}_{i=1}^n, \quad \sigma_1(P) = \{\mu_i\}_{i=1}^{n-1}, \quad \sigma_2(P) = \{\nu_i\}_{i=1}^{n-2}.$$

Theorem 2.1. *Let P be a H-Symmetric matrix with $H = diag(\delta_1, \delta_2, \cdots, \delta_n)$. Let $\sigma(P) = \{\lambda_i\}_{i=1}^n$ and*

$$u_1 = [u_{11}, u_{12}, \cdots, u_{1n}], \quad u_2 = [u_{21}, u_{22}, \cdots, u_{2n}],$$

are the vectors of first and second components of the eigenvectors of P, respectively, such that

$$[u_1, u_2]_H = 0, \quad [u_1, u_1]_H = \delta_1, \quad [u_2, u_2]_H = \delta_2.$$

Then the entries of P can be constructed as follows:

$$\begin{aligned}
a_i &= \delta_i \sum_{j=1}^{n} \lambda_j \delta_j u_{1j}^2, \quad b_i = \varepsilon_i \delta_{i+1} \sum_{j=1}^{n} \lambda_j \delta_j u_{1j} u_{i+1,j}, \quad c_i = \sqrt{D_i}, \\
D_i &= \delta_{i+2} \sum_{j=1}^{n} \delta_j [(\lambda_j - a_i) u_{ij} - c_{i-2} u_{i-2,j} - b_{i-1} u_{i-1,j} - \epsilon_i b_i u_{i+1,j}]^2, \\
u_{i+2,j} &= \frac{1}{\epsilon_i \epsilon_{i+1} c_i} [(\lambda_j - a_i) u_{ij} - c_{i-2} u_{i-2,j} - b_{i-1} u_{i-1,j} - \epsilon_i b_i u_{i+1,j}],
\end{aligned}$$

where $c_0 = c_{-1} = b_0 = 0$.

Proof. Suppose U is a matrix that its columns are eigenvectors of P. If $\Lambda = \text{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n)$ then

$$AU = U\Lambda \tag{2}$$

Writting the first row of the equation (2) impies that

$$a_1 u_{1j} + \epsilon_1 b_1 u_{2j} + \epsilon_1 \epsilon_2 c_1 u_{3j} = \lambda_j u_{1j}. \tag{3}$$

Multiplying both sides of (3) by $\delta_j u1j$ and summing up from $j=1$ to $j=n$ and using H-Orthonormal properties of the vectors u_1 and u_2 we find

$$a_1 = \delta_1 \sum_{j=1}^{n} \lambda_j \delta_j u_{1j}^2. \tag{4}$$

Similarly, multiplying both sides of (3) by $\delta_j u1j$ and summing up from $j=1$ to $j=n$ we find

$$\epsilon_1 b_1 \delta_2 = \sum_{j=1}^{n} \lambda_j \delta_j u_{1j} u_{2j},$$

Using the equation (3) implies that

$$\epsilon_1 \epsilon_2 u_{3j} = (\lambda_j - a_1) u_{1j} - \epsilon_1 b_1 u_{2j}. \tag{5}$$

Taking to the power 2 and multiplying the last eqauation by δ_j and summing up from $j=1$ to $j=n$ implies that

$$c_1^2 \delta_3 = \sum_{j=1}^{n} \delta_j [(\lambda_j - a_1) u_{1j} - \epsilon_1 b_1 u_{2j}]^2, \tag{6}$$

that concludes $c_1 = \sqrt{D_1}$ where $D_1 = \delta_3 \sum_{j=1}^{n} \delta_j [(\lambda_j - a_1) u_{1j} - \epsilon_1 b_1 u_{2j}]^2$. Again using the equation (3) we find

$$u_{3j} = \frac{1}{\epsilon_1 \epsilon_2 c_1} [(\lambda_j - a_1) u_{1j} - \epsilon_1 b_1 u_{2j}. \tag{7}$$

Continuing this procedure will produce all entries of the matrix P. Note that for solvibility of the inverse problem by this Theorem the eigendata must be chosen such that D_i given by theorem to be nonnegative, otherwise the problem has no solution. □

Now we are ready to construct the solution of the inverse problems by three given spectra. First, we compute the first entries of the eigenvectors of P by two given spectra.

Theorem 2.2. *Let P be H-Symmetric and*

$$\sigma(P) = \{\lambda_i\}_{i=1}^{n}, \quad \sigma_1(P) = \{\mu_i\}_{i=1}^{n-1}$$

are real and distinct such that there exists a permutation of $\sigma(P)$, say $\{\lambda_{k_i}\}_{i=1}^{n}$ such that

$$\delta_i \frac{\Pi_{j=1}^{n-1}(\mu_j - \lambda_{k_i})}{\Pi_{j=1, j\neq i}^{n}(\lambda_{k_j} - \lambda_{k_i})} > 0, \quad i = 1, 2, \cdots, n.$$

Then the first entries of the eigenvectors of P i.e., u_{1i} are computed as follows:

$$u_{1i} = \sqrt{\delta_i \frac{\Pi_{j=1}^{n-1}(\mu_j - \lambda_{k_i})}{\Pi_{j=1, j\neq i}^{n}(\lambda_{k_j} - \lambda_{k_i})}}. \tag{8}$$

Theorem 2.3. *Let $H_1 = diag(\delta_1^1, \delta_2^1, \cdots, \delta_{n-1}^1)$ and P_1 be H_1-Symmetric matrix. Suppose $V = [v_1, v_2, \cdots, v_{n-1}]$ is the matrix with columns consisting the eigenvectors of P_1. Then the eigenvalues of P are the roots of the following equation:*

$$(a_1 - \lambda) - \sum_{i=1}^{n-1} \frac{\delta_i^1 \epsilon_1 (b_1 v_{1i} + c_1 v_{2i})^2}{\mu_i - \lambda} = 0. \tag{9}$$

Theorem 2.4. *Let P be H-Symmetric and P_1 be H_1-Symmetric such that $\sigma_2(P) = \{\nu_i\}_{i=1}^{n-2}$. Then*

$$u_{2i} = v_{1i} \sum_{j=1}^{n-1} \frac{\delta_i^1 \epsilon_1 \delta_j^1 (b_1 v_{1i} + c_1 v_{2i})}{(\mu_j - \lambda_i)} u_{1i},$$

where $b_1 v_{1i} + c_1 v_{2i}$ is computed by

$$(b_1 v_{1j} + c_1 v_{2j})^2 = -\delta_j^1 \epsilon_1 \frac{\Pi_{i=1}^{n} (\lambda_i - \mu_j)}{\Pi_{i=1, i\neq j}^{n-1} (\mu_i - \mu_j)}, \quad j = 1, 2, \ldots, n-1.$$

Now given the spectrum $\{\lambda_i\}_{i=1}^n$ and having the first and second components of eigenvectors, i.e., u_{1i} and u_{2i} we can construct the matrix P using modified Lanczos algorithm (Theorem 2.1).

3 Numerical Examples

Example 3.1. Cosider the H- symmetric pentadiagonal matrices of the form (1) with entries

$$a_i = \begin{cases} 20, & i = 1, \\ 6, & i = 2, \ldots, n-1, \\ 5, & i = n, \end{cases} \qquad b_i = -4, \quad c_i = 1, \quad \epsilon_i = \begin{cases} -1, & i = 1, \\ 1, & i = 2, \ldots, n-1. \end{cases}$$

In Tables 1 and 2, for different values of n, we compared the spectra of computed matrix $P^{(c)}$ with the initial given matrix P.

Table 1: Numerical results for example 3.1 without Modified Gram-Schmit method

n	$\|\sigma(P^{(c)}) - \sigma(P)\|$	$\|\sigma_1(P^{(c)}) - \sigma_1(P)\|$	$\|\sigma_2(P^{(c)}) - \sigma_2(P)\|$
10	1.41e-07	1.37e-07	1.45e-07
50	165.59	166.77	166.88
100	337.07	338.06	338.20
200	374.86	375.59	375.72

Table 2: Numerical results for example 3.1 with applying Modified Gram-Schmit method

n	$\|\sigma(P^{(c)}) - \sigma(P)\|$	$\|\sigma_1(P^{(c)}) - \sigma_1(P)\|$	$\|\sigma_2(P^{(c)}) - \sigma_2(P)\|$
10	4.06e-14	1.19e-14	9.74e-15
50	1.35e-13	3.12e-14	3.50e-14
100	1.55e-13	4.12e-14	4.35e-14
200	4.30e-13	5.51e-14	5.71e-14

4 Concluding Remark

Using Lanczos algorithm we construct pentadiagonal matrix P and H-Orthonormal eigenvectors. For large scale matrices the constructed eigenvectors might not be H-Orthonormal. To overcome this difficulty we use the modified Gram Schmidt orthogonalization. This algorithm transforms the vectors $u_1, u_2, \cdots, u_n$ into H-Orthonormal vectors as follows:

1. Put $u_1 = \frac{v_1}{\sqrt{u_1^T H u_1}}$,
2. For $i = 2, 3, \cdots, n$ define $S_i = u_i - \sum_{j=1}^{i-1} \delta_j [u_i, u_j] u_j$,
3. Set $u_i = \frac{S_i}{\sqrt{[u_i, u_i]_H}}$.

Due to the fact that the components of u_{1i} and v_{1i} have signs $+, -$, also $b_1 v_{1j} + c_1 v_{2j}$ have signs $+, -$, therefore the solution matrix is not unique. To illustrate the efficiency of the numerical examples we compare the prescribed eigenvalues with the eigenvalues of the constructed matrix.

References

[1] N. Bebiano, J. da Providencia, Inverse spectral problems for structured pseudo-symmetric matrices, Linear Algebra Appl, 2013;438:40624074.

[2] N. Bebiano, C.M. Fonseca, J. da Providencia, An inverse eigenvalue problem for periodic Jacobi matrices in Minkowski spaces, Linear Algebra Appl, 2011;435:20332045.

[3] K. Ghanbari and H. Mirzaei, Inverse eigenvalue problem for pentadiagonal matrices, Inverse Problems in Science and Engineering, 2014; 22(4): 530-542.

[4] G.M.L. Gladwell: Inverse problems in vibration. *Kluwer Academic Publishers*, 2004.

[5] R.P.S. Han, J.W. Zu, Pseudo non-selfadjoint and non-selfadjoint systems in atructural dynamics, Journal of Sound and Vibration, 1995;184(4): 725-742.

[6] M. Metrovic, Finite Difference Method for Non-smooth Beam Bending, Int'l Conf. Scientific Computing, CSC'17

Reflection matrices and linear preservers of majorization[1]

Ahmad Mohammadhasani* and Yamin Sayyari

Department of Mathematics, Sirjan University Of Technology, Sirjan, Iran

Abstract

A reflection is a mapping from an Euclidean space to itself that is an isometry. In this paper, we have outlined the concept of left majorization (resp. right majorization) of the group of reflection matrices to the line passing through the origin of the coordinates, and we have found all linear preserver transformations of this kind of majorization.

Keywords: Reflection matrix, Majorization, Linear preserver

Mathematics Subject Classification [2010]: 15A04, 15A21, 15A30, 47B49

1 Introduction

Let X be a real vector space, $W \subseteq X$, $conv(W)$ be the convex hull of W and G be a left action (right action) on X. The group G induces an equivalence relation on X, defined by $x \simeq y$ if and only if $x = gy$ ($x = yg$) for some $g \in G$. The equivalence classes of this relation are called the orbits of G. for each $y \in X$ the orbit of y is as follows:

$$O_G(y) = \{gy | g \in G\} \ (O_G(y) = \{yg | g \in G\}).$$

A vector x is said to be G-majorized of the left (of the right) by y and we write $x \prec_{lG} y$ ($x \prec_{rG} y$) if $x \in conv(O_G(y))$. Let $T : X \longrightarrow X$ be a mapping and $\sim$ be a relation on X. We say T is a preserver of $\sim$ if $Tx \sim Ty$ whenever $x \sim y$, it is called a strong preserver of $\sim$ if it further satisfies $x \sim y$ whenever $Tx \sim Ty$.

2 Main results

In this section section, the concept of majorization is studied and then the linear preservers of this concept are characterized.

Definition 2.1. Let θ be a real number, define

$$P_\theta = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ \sin(\theta) & -\cos(\theta) \end{bmatrix}$$

and $G_\theta = \{I_2, P_\theta\}$. Its obvious that G_θ is a group.

[1] Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

*Speaker. Email address: a.mohammadhasani53@gmail.com

For each z the orbit of $z = (x, y)^t \in \mathbb{R}^2$ is a follows:

$$O_{G_\theta}(z) = \{gz : g \in G_\theta\}.$$

Let $z_1, z_2 \in \mathbb{R}^2$. We say that $z_1 = (x_1, y_1)^t$ G-majorized of the left by $z_2 = (x_2, y_2)^t$ (denote by $z_1 \prec_{l\theta} z_2$) if $z_1 \in conv(O_{G_\theta}(z_2))$.

Theorem 2.2. *Let z_1 and z_2 are two vectors of the $\mathbb{R}^2$.*

1. *$z_1 \prec_{l\theta} z_2$ if and only if the $z_1 = tz_2 + (1-t)P_\theta z_2$, for some $0 \leq t \leq 1$.*
2. *$z_1 \sim_{l\theta} z_2$ if and only if $z_1 = z_2$ or $z_1 = P_\theta z_2$.*

Proof. □

Theorem 2.3. *Let T be a linear operator on M_2. Then T preserves G-majorized $\prec_{l\theta}$ if and only if one of the following holds:*

1. *$\theta \neq n\pi$ and*

$$[T] = \begin{bmatrix} a\sin\theta & b\sin\theta \\ b\sin\theta & a\sin\theta - 2b\cos\theta \end{bmatrix} \text{ or } [T] = \begin{bmatrix} a\sin\theta & a(1-\cos\theta) \\ b\sin\theta & b(1-\cos\theta) \end{bmatrix}$$

   for some real numbers a, b.

2. *$\theta = 2n\pi$ and*

$$[T] = \begin{bmatrix} a & 0 \\ b & 0 \end{bmatrix} \text{ or } [T] = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}$$

   for some real numbers a, b.

3. *$\theta = (2n+1)\pi$ and*

$$[T] = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix} \text{ or } [T] = \begin{bmatrix} 0 & a \\ 0 & b \end{bmatrix}$$

   for some real numbers a, b.

Proof. 1. If

$$A = [T] = \begin{bmatrix} a\sin\theta & b\sin\theta \\ b\sin\theta & a\sin\theta - 2b\cos\theta \end{bmatrix}$$

then

$$A(ty + (1-t)P_\theta y) = tAy + (1-t)P_\theta Ay \prec_{l\theta} Ay$$

for each $y \in \mathbb{R}^2$, and if

$$A = [T] = \begin{bmatrix} a\sin\theta & a(1-\cos\theta) \\ b\sin\theta & b(1-\cos\theta) \end{bmatrix}$$

then

$$A(ty + (1-t)P_\theta y) = Ay \prec_{l\theta} Ay$$

for each $y \in \mathbb{R}^2$, so T preservers $\prec_{l\theta}$.

Now Let $\theta \neq n\pi$ and T preservers $\prec_{l\theta}$ and

$$[T] = A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

Since $e_1 \sim_\theta (\cos\theta, \sin\theta)^t$,

$$(a_{11}, a_{21})^t \sim_\theta (a_{11}\cos\theta + a_{12}\sin\theta, a_{21}\cos\theta + a_{22}\sin\theta)^t \tag{1}$$

therefor

$$(a_{11}\cos\theta + a_{12}\sin\theta, a_{21}\cos\theta + a_{22}\sin\theta)^t = (a_{11}, a_{21})^t$$

or

$$\begin{aligned}&(a_{11}\cos\theta + a_{12}\sin\theta, a_{21}\cos\theta + a_{22}\sin\theta)^t = P_\theta(a_{11}, a_{21})^t \\ &= (a_{11}\cos\theta + a_{21}\sin\theta, a_{11}\sin\theta - a_{21}\cos\theta)^t.\end{aligned}$$

Similarly, since $e_2 \sim_\theta (\sin\theta, -\cos\theta)^t$,

$$(a_{12}, a_{22})^t \sim_\theta (a_{11}\sin\theta - a_{12}\cos\theta, a_{21}\sin\theta - a_{22}\cos\theta)^t \tag{2}$$

therefor

$$(a_{11}\sin\theta - a_{12}\cos\theta, a_{21}\sin\theta - a_{22}\cos\theta)^t = (a_{12}, a_{22})^t$$

or

$$\begin{aligned}&(a_{11}\sin\theta - a_{12}\cos\theta, a_{21}\sin\theta - a_{22}\cos\theta)^t = P_\theta(a_{12}, a_{22})^t \\ &= (a_{12}\cos\theta + a_{22}\sin\theta, a_{12}\sin\theta - a_{22}\cos\theta)^t.\end{aligned}$$

We consider four cases.

Case1: Let

$$(a_{11}\cos\theta + a_{12}\sin\theta, a_{21}\cos\theta + a_{22}\sin\theta)^t = (a_{11}, a_{21})^t$$

and

$$\begin{aligned}&(a_{11}\sin\theta - a_{12}\cos\theta, a_{21}\sin\theta - a_{22}\cos\theta)^t \\ &= (a_{12}\cos\theta + a_{22}\sin\theta, a_{12}\sin\theta - a_{22}\cos\theta)^t\end{aligned}$$

in this case $a_{12} = a_{21}$ and $a_{22} = a_{11} - 2a_{12}\cot\theta$ so

$$A = \begin{bmatrix} a\sin\theta & b\sin\theta \\ b\sin\theta & a\sin\theta - 2b\cos\theta \end{bmatrix}$$

where $a = \frac{a_{11}}{\sin\theta}$ and $b = \frac{a_{12}}{\sin\theta}$.

Case2: Let

$$(a_{11}\cos\theta + a_{12}\sin\theta, a_{21}\cos\theta + a_{22}\sin\theta)^t = (a_{11}, a_{21})^t$$

and

$$(a_{11}\sin\theta - a_{12}\cos\theta, a_{21}\sin\theta - a_{22}\cos\theta)^t = (a_{12}, a_{22})^t$$

in this case $a_{12} = \frac{1-\cos\theta}{\sin\theta}a_{11}$ and $a_{22} = \frac{1-\cos\theta}{\sin\theta}a_{21}$ so

$$A = \begin{bmatrix} a\sin\theta & a(1-\cos\theta) \\ b\sin\theta & b(1-\cos\theta) \end{bmatrix}$$

where $a = \frac{a_{11}}{\sin\theta}$ and $b = \frac{a_{21}}{\sin\theta}$.

Case3: Let

$$\begin{aligned}&(a_{11}\cos\theta + a_{12}\sin\theta, a_{21}\cos\theta + a_{22}\sin\theta)^t \\ &= (a_{11}\cos\theta + a_{21}\sin\theta, a_{11}\sin\theta - a_{21}\cos\theta)^t\end{aligned}$$

and

$$\begin{aligned}&(a_{11}\sin\theta - a_{12}\cos\theta, a_{21}\sin\theta - a_{22}\cos\theta)^t \\ &= (a_{12}\cos\theta + a_{22}\sin\theta, a_{12}\sin\theta - a_{22}\cos\theta)^t\end{aligned}$$

in this case $a_{12} = a_{21}$ and $a_{22} = a_{11} - 2a_{12}\cot\theta$ so

$$A = \begin{bmatrix} a\sin\theta & b\sin\theta \\ b\sin\theta & a\sin\theta - 2b\cos\theta \end{bmatrix}$$

where $a = \frac{a_{11}}{\sin\theta}$ and $b = \frac{a_{12}}{\sin\theta}$.

Case4: Let

$$\begin{aligned}&(a_{11}\cos\theta + a_{12}\sin\theta, a_{21}\cos\theta + a_{22}\sin\theta)^t \\ &= (a_{11}\cos\theta + a_{21}\sin\theta, a_{11}\sin\theta - a_{21}\cos\theta)^t\end{aligned}$$

and

$$(a_{11}\sin\theta - a_{12}\cos\theta, a_{21}\sin\theta - a_{22}\cos\theta)^t = (a_{12}, a_{22})^t$$

in this case $a_{12} = a_{21}$ and $a_{22} = a_{11} - 2a_{12}\cot\theta$. so

$$A = \begin{bmatrix} a\sin\theta & b\sin\theta \\ b\sin\theta & a\sin\theta - 2b\cos\theta \end{bmatrix}$$

where $a = \frac{a_{11}}{\sin\theta}$ and $b = \frac{a_{11}}{1+\cos\theta}$.

2. We only prove the necessary condition. The Relation (2) results that

$$(a_{12}, a_{22})^t \sim_\theta (-a_{12}, -a_{22})^t$$

so

$$\begin{aligned}&(-a_{12}, -a_{22})^t = (a_{12}, a_{22})^t \text{ or} \\ &(-a_{12}, -a_{22})^t = P_\theta(a_{12}, a_{22})^t = (a_{12}, -a_{22})^t.\end{aligned}$$

If $(-a_{12}, -a_{22})^t = (a_{12}, a_{22})^t$ have $a_{12} = a_{22} = 0$. So

$$A = \begin{bmatrix} a_{11} & 0 \\ a_{21} & 0 \end{bmatrix}$$

If

$$(-a_{12}, -a_{22})^t = P_\theta(a_{12}, a_{22})^t = (a_{12}, -a_{22})^t$$

we have $a_{12} = 0$. On the other hand $(1, 1) \sim_\theta (1, -1)$ consequence that

$$(a_{11}, a_{21} + a_{22}) \sim_\theta (a_{11}, a_{21} - a_{22})$$

so $(a_{11}, a_{21} - a_{22}) = (a_{11}, a_{21} + a_{22})$ or $(a_{11}, a_{21} - a_{22}) = (a_{11}, -a_{21} - a_{22})$. Hence $a_{22} = 0$ or $a_{21} = 0$. Thus

$$[T] = \begin{bmatrix} a_{11} & 0 \\ a_{21} & 0 \end{bmatrix} \text{ or } [T] = \begin{bmatrix} a_{11} & 0 \\ 0 & a_{22} \end{bmatrix}$$

3. We only prove the necessary condition. The Relation (1) results that

$$(a_{11}, a_{21})^t \sim_\theta (-a_{11}, -a_{21})^t$$

so

$$\begin{aligned} &(-a_{11}, -a_{21})^t = (a_{11}, a_{21})^t \text{ or} \\ &(-a_{11}, -a_{21})^t = P_\theta (a_{11}, a_{21})^t = (-a_{11}, a_{21})^t. \end{aligned}$$

If $(-a_{11}, -a_{21})^t = (a_{11}, a_{21})^t$ have $a_{11} = a_{21} = 0$. So

$$A = \begin{bmatrix} 0 & a_{12} \\ 0 & a_{22} \end{bmatrix}$$

If

$$(-a_{11}, -a_{21})^t = (-a_{11}, a_{21})^t$$

we have $a_{21} = 0$. On the other hand $(1, 1) \sim_\theta (-1, 1)$ consequence that

$$(a_{11} + a_{12}, a_{22}) \sim_\theta (-a_{11} + a_{12}, a_{22})$$

so

$$\begin{aligned} &(-a_{11} + a_{12}, a_{22}) = (a_{11} + a_{12}, a_{22}) \\ &\text{or } (-a_{11} + a_{12}, a_{22}) = (-a_{11} - a_{12}, a_{22}). \end{aligned}$$

Hence $a_{11} = 0$ or $a_{12} = 0$. Thus

$$[T] = \begin{bmatrix} 0 & a_{12} \\ 0 & a_{22} \end{bmatrix} \text{ or } [T] = \begin{bmatrix} a_{11} & 0 \\ 0 & a_{22} \end{bmatrix}$$

□

Similarly, For each vector $z = (x, y) \in \mathbb{R}_2$ the orbit of z is a follows:

$$O_{G_\theta}(z) = \{zg : g \in G_\theta\}.$$

Let $z_1, z_2 \in \mathbb{R}_2$. We say that $z_1 = (x_1, y_1)^t$ G-majorized of the right by $z_2 = (x_2, y_2)^t$ (denote by $z_1 \prec_{r\theta} z_2$) if $z_1 \in conv(O_{G_\theta}(z_2))$.

Theorem 2.4. *Let z_1 and z_2 are two vectors of the $\mathbb{R}_2$.*

1. *$z_1 \prec_{r\theta} z_2$ if and only if the $z_1 = tz_2 + (1 - t)z_2P_\theta$, for some $0 \leq t \leq 1$.*

2. $z_1 \sim_{r\theta} z_2$ *if and only if* $z_1 = z_2$ *or* $z_1 = z_2 P_\theta$.

Theorem 2.5. *Let* T *be a linear operator on* M_2. *Then* T *preserves* G*-majorized* $\prec_{r\theta}$ *if and only if one of the following holds:*

1. $\theta \neq n\pi$ *and*

$$[T] = \begin{bmatrix} a\sin\theta & b\sin\theta \\ b\sin\theta & a\sin\theta - 2b\cos\theta \end{bmatrix} \text{ or } [T] = \begin{bmatrix} a\sin\theta & b\sin\theta \\ a(1-\cos\theta) & b(1-\cos\theta) \end{bmatrix}$$

2. $\theta = 2n\pi$ *and*

$$[T] = \begin{bmatrix} a & 0 \\ b & 0 \end{bmatrix} \text{ or } [T] = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}$$

3. $\theta = (2n+1)\pi$ *and*

$$[T] = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix} \text{ or } [T] = \begin{bmatrix} 0 & 0 \\ a & b \end{bmatrix}$$

for some real numbers a, b.

References

[1] A. Armandnejad and A. Salemi, On linear preservers of lgw-majorization on $M_{n,m}$. *Bulletin of the Malaysian Mathmatical Society*, 35 (3):755-764, 2012.

[2] R. Bahatia, Matrix Analysis. *Springer-Verlag, New York*, 1997.

[3] G. Dahl, Matrix majorization, Linear Algebra Appl., 288 (1999), 53-73.

[4] Francisco D. Martnez Pera, Pedro G. Massey, Luis E. Silvestre, Weak matrix majorization, Linear Algebra and its Applications 403 (2005) 343368.

[5] A. M. Hasani and M. Radjabalipour, On linear preservers of (right) matrix majorization. *Linear Algebra and its Applications*, 423:255-261, 2007.

[6] A. W. Marshall, I. Olkin, and B. C. Arnold, *Inequalities: Theory of majorization and its applications*. Springer, New York, 2011.

A matrix approach for time fractional option pricing[1]

Nasibeh Mollahasani* and Habibollah Saeedi

Department of Applied Mathematics, Shahid Bahonar University of Kerman, Keman, Iran

Abstract

In this paper triangular functions (TF) and the related operational matrices of fractional integration are applied to solve time-fractional Black-Sholes equation. By using this equation, we are able to price an option, which is one of the most important derivatives in financial markets. The numerical result confirms the efficiency and accuracy of the proposed method.

Keywords: Operational matrix, Triangular functions, Option pricing
Mathematics Subject Classification [2010]: 91G80, 62P05, 26A33

1 Introduction

Options are usually used for risk reduction in stock markets and has been widely used by traders and practitioners. One of the most important financial derivative pricing models is BlackScholes model, which has been the basis of current financial models [1].
In recent years, many financial studies have been conducted on markets with fractional models due to better and more acceptable results in real markets [2]. In the early 1970s, Fisher Black, Miren Scholes and Robert Merton took a big step in option pricing. The result was a model named Black-Scholes model. This model has had a great impact on the option pricing and risk hedge methods. In this paper, we introduce a new operational method to solve a special differential equation of fractional order. The aim of this work is to present an operational method (operational triangular functions method) for approximating the solution of fractional Black-Scholes equation [3]:

$$\frac{\partial^\alpha c(s,t)}{\partial t^\alpha} + \frac{1}{2}\sigma^2 s^2 \frac{\partial^2 c(s,t)}{\partial s^2} + (r-D)\frac{\partial c(s,t)}{\partial s} = rc(s,t), \quad (s,t) \in (B_d, B_u) \times (0,T), \quad (1)$$

such that when $\alpha = 1$ is the classic Black-Scholes formula. For an option pricing problem, the initial and boundary conditions of this equation are:

$$c(B_d,t) = P(t), \qquad c(B_u,t) = Q(t), \quad (2)$$

$$c(s,T) = V(s), \qquad B_d < s < B_u, \quad (3)$$

where r is the rate of interest, σ is volatility of stock price, D is the dividend yield, T is the expire time, B_d is the lowest stock price in time interval $[0,T]$ with d probability and

[1]Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University
*Speaker. Email address: n.mollahasani@math.uk.ac.ir

B_u is the highest stock price in time interval $[0,T]$ with u probability such that $u+d=1$. The conditions (2) and (3) for European call option are as:

$$c(s,T)=max\{0,s_{max}-k\},\quad 0\leq s<\infty,\quad c(0,t)=0,\quad 0\leq t\leq T,$$
$$c(s,t)=s_{max}-ke^{-rt},\qquad s\to\infty.$$

Also, the mentioned conditions for European put option are as follows:

$$c(s,T)=max\{0,k-s_{max}\},\quad 0\leq s<\infty,\quad c(0,t)=ke^{-rt},\quad 0\leq t\leq T,$$
$$c(s,t)=0,\quad s\to\infty.$$

2 Fractional Calculus

There are several definitions of a fractional derivative of order $\alpha>0$. The two most commonly used definitions are the Riemann- Liouville and Caputo. Each definition uses Riemann- Liouville fractional integration and derivatives of whole order. The Riemann-Liouville fractional integration of order $\alpha\geq 0$ of function f is defined as:

$$I^{\alpha}_{0,t}f(t)=\begin{cases}\frac{1}{\Gamma(\alpha)}\int_0^t(t-\tau)^{\alpha-1}f(\tau)d\tau, & \alpha>0,\\ f(t), & \alpha=0,\end{cases}$$

and the Caputo fractional derivative of order α is defined as ${}_cD^{\alpha}_{0,t}f(x)=I^{m-\alpha}_{0,t}D^m f(x)$, where D^m is the usual integer differential operator of order m, $I^{m-\alpha}_{0,t}$ is the Riemann-Liouville integral operator of order $m-\alpha$ and $m-1<\alpha\leq m$.
The relation between Riemann- Liouville operator and Caputo operator is given by the following lemma:

Lemma 2.1. *If $m-1<\alpha\leq m$, $m\in\mathbb{N}$, then ${}_cD^{\alpha}_{0,t}I^{\alpha}_{0,t}f(x)=f(x)$, and:*

$$I^{\alpha}_{0,t}\ {}_cD^{\alpha}_{0,t}f(x)=f(x)-\sum_{k=0}^{m-1}f^{(k)}(0^+)\frac{x^k}{k!},\ x>0.$$

3 Triangular Functions

In this section, first we introduce triangular functions, then using them and operational matrices, the fractional Black-Scholes equation will be solved .

Definition 3.1. Divide $[0,L)$ into m equal parts. Suppose that $h=\frac{L}{m}$. Triangular functions are defined as:

$$T^1_i(t)=\begin{cases}1-\frac{t-ih}{h}, & ih\leq t<(i+1)h;\\ 0, & \text{otherwise},\end{cases}$$

and:

$$T^2_i(t)=\begin{cases}\frac{t-ih}{h}, & ih\leq t<(i+1)h;\\ 0, & \text{otherwise},\end{cases}$$

for $i=0,1,...,m-1$.

Let's define the TF-vector as the following:

$$\mathbf{T}(t) = \begin{bmatrix} T_1(t) \\ T_2(t) \end{bmatrix},$$

such that:

$$T_1(t) = [T_0^1(t), T_1^1(t), T_2^1(t), ..., T_{m-1}^1(t)]^T, \quad T_2(t) = [T_0^2(t), T_1^2(t), T_2^2(t), ..., T_{m-1}^2(t)]^T.$$

For each continuous function $f \in L^2[0, L]$, we have:

$$\begin{aligned} f(x) &\simeq \sum_{i=0}^{m-1} c_i T_i^1(t) + \sum_{i=0}^{m-1} d_i T_i^2(t) = F_1^T.T_1(t) + F_2^T.T_2(t) \\ &= \begin{bmatrix} F_1 \\ F_2 \end{bmatrix}^T \begin{bmatrix} T_1(t) \\ T_2(t) \end{bmatrix} = \mathbf{F}^T \mathbf{T}(t) = f_m(t), \end{aligned}$$

where $c_i = f(t_i)$ and $d_i = f(t_{i+1})$ for $i = 0, 1, ..., m-1$ [4].

3.1 Operational Matrices of Triangular Functions

Now we investigate the operational matrix of fractional integral of triangular functions (of order α).

Definition 3.2. Fractional integration of triangular functions of order α is defined as:

$$I^\alpha \mathbf{T}(t) = \frac{1}{\Gamma(\alpha)} \int_0^t (t-\tau)^{(\alpha-1)} T(\tau) d\tau \simeq \mathbf{P}^\alpha \mathbf{T}(t),$$

where $\mathbf{P}^\alpha$ is called the operational matrix of fractional integration of triangular functions and derived according to the following theorem.

Theorem 3.3. *[4] The operational matrix of fractional integration of triangular functions, P^α, is defined as the following:*

$$I^\alpha \mathbf{T}(t) \simeq \mathbf{P}^\alpha \mathbf{T}(t), \quad \mathbf{P}^\alpha = \begin{bmatrix} P_{11}^\alpha & P_{12}^\alpha \\ P_{21}^\alpha & P_{22}^\alpha \end{bmatrix}, \tag{4}$$

where:

$$P_{11}^\alpha = \begin{bmatrix} 0 & \xi_0 & \xi_1 & \dots & \xi_{m-2} \\ 0 & 0 & \xi_0 & \dots & \xi_{m-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \xi_1 \\ 0 & 0 & 0 & \dots & \xi_0 \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}, P_{12}^\alpha = \begin{bmatrix} \xi_0 & \xi_1 & \xi_2 & \dots & \xi_{m-1} \\ 0 & \xi_0 & \xi_1 & \dots & \xi_{m-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \xi_2 \\ 0 & 0 & 0 & \dots & \xi_1 \\ 0 & 0 & 0 & \dots & \xi_0 \end{bmatrix},$$

such that $\xi_j = \dfrac{h^\alpha}{\Gamma(\alpha+2)}[(j+1)^\alpha(\alpha-j) + j^{\alpha+1}]$ *for* $j = 0, 1, 2, ..., m-1$ *and:*

$$P_{21}^\alpha = \begin{bmatrix} 0 & \eta_0 & \eta_1 & \dots & \eta_{m-2} \\ 0 & 0 & \eta_0 & \dots & \eta_{m-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \eta_1 \\ 0 & 0 & 0 & \dots & \eta_0 \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}, P_{22}^\alpha = \begin{bmatrix} \eta_0 & \eta_1 & \eta_2 & \dots & \eta_{m-1} \\ 0 & \eta_0 & \eta_1 & \dots & \eta_{m-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \eta_2 \\ 0 & 0 & 0 & \dots & \eta_1 \\ 0 & 0 & 0 & \dots & \eta_0 \end{bmatrix},$$

such that $\eta_j = \dfrac{h^\alpha}{\Gamma(\alpha+2)}[(j+1)^{\alpha+1} - j^\alpha(\alpha+j+1)]$ *for* $j = 0, 1, 2, ..., m-1$.

4 TF-Method for Solving Fractional Black-Scholes Equation

Suppose that $c(s,t)$ which shows option pricing in time t by the price of s, can be written as:

$$\frac{\partial^2 c(s,t)}{\partial s^2} \simeq \sum_{i=0}^{n}\sum_{j=0}^{m} c_{ij} T_i^{\alpha}(s) T_j^{\alpha}(t) = \mathbf{T}^T(t)\mathbf{C}\mathbf{T}(s),$$

where $\mathbf{C}$ is the coefficient matrix which is unknown.
Applying the fractional integration operator $I_{0,s}^{\alpha}$ to the both sides of (5) and using the operational matrix of (4), we get:

$$\frac{\partial c(s,t)}{\partial s} \simeq \mathbf{T}^T(t)\mathbf{C}\mathbf{P}^1\mathbf{T}(t) + L(t), \tag{5}$$

where $L(t)$ will be calculated later. Again by applying the integration operator $I_{0,s}^{2}$ to the both sides of (5) and using the operational matrix of (4), we have:

$$c(s,t) \simeq \mathbf{T}^T(t)\mathbf{C}\mathbf{P}^2\mathbf{T}(s) + sL(t) + Z(t). \tag{6}$$

By the assumption $B_d = 0$, it is obvious that $Z(t) = P(t)$. On the other hand, $L(t)$ is unknown in (5) and (6), which can be obtained by applying the condition $c(B_u,t) = Q(t)$ in (6):

$$c(B_u,t) = \mathbf{T}^T(t)\mathbf{C}\mathbf{P}^2\mathbf{T}(B_u) + B_u L(t) + P(t).$$

Therefore:

$$L(t) \simeq \frac{1}{B_u}(Q(t) - P(t)) - \frac{1}{B_u}\mathbf{T}^T(t)\mathbf{C}\mathbf{P}^2\mathbf{T}(B_u). \tag{7}$$

Substituting (7) in (5) and (6), we get respectively:

$$\frac{\partial c(s,t)}{\partial s} \simeq \mathbf{T}^T(t)\mathbf{C}\mathbf{P}^1\mathbf{T}(s) + \frac{1}{B_u}(Q(t) - P(t)) - \frac{1}{B_u}\mathbf{T}^T(t)\mathbf{C}\mathbf{P}^2\mathbf{T}(B_u), \tag{8}$$

$$c(s,t) \simeq \mathbf{T}^T(t)\mathbf{C}\mathbf{P}^2\mathbf{T}(s) + \frac{s}{B_u}(Q(t) - P(t) - \mathbf{T}^T(t)\mathbf{C}\mathbf{P}^2\mathbf{T}(B_u)) + P(t). \tag{9}$$

Now by replacing (5), (8) and (9) in (1), we have:

$$\begin{aligned}\frac{\partial^{\alpha} c(s,t)}{\partial t^{\alpha}} \cong\ & \mathbf{T}^T(t)[r\mathbf{C}\mathbf{P}^2\mathbf{T}(s) - r\frac{s}{h}\mathbf{C}\mathbf{P}^2\mathbf{T}(B_u) - \frac{1}{2}\sigma^2 s^2 \mathbf{C}\mathbf{T}(s) \\ & -rs\mathbf{C}\mathbf{P}^1\mathbf{T}(s) + r\frac{s}{h}\mathbf{C}\mathbf{P}^2\mathbf{T}(B_u)] + r\mathbf{T}^T(t)P,\end{aligned}$$

where $P(t) \cong \mathbf{T}^T(t)P$. For simplicity, we define:

$$X := r\mathbf{C}\mathbf{P}^2\mathbf{T}(s) - r\frac{s}{h}\mathbf{C}\mathbf{P}^2\mathbf{T}(B_u) - \frac{1}{2}\sigma^2 s^2 \mathbf{C}\mathbf{T}(s) - rs\mathbf{C}\mathbf{P}^1\mathbf{T}(s) + r\frac{s}{h}\mathbf{C}\mathbf{P}^2\mathbf{T}(B_u)] + r\mathbf{T}^T(t)P,$$

therefore:

$$\frac{\partial^{\alpha} c(s,t)}{\partial t^{\alpha}} \cong \mathbf{T}^T(t)X + r\mathbf{T}^T(t)P. \tag{10}$$

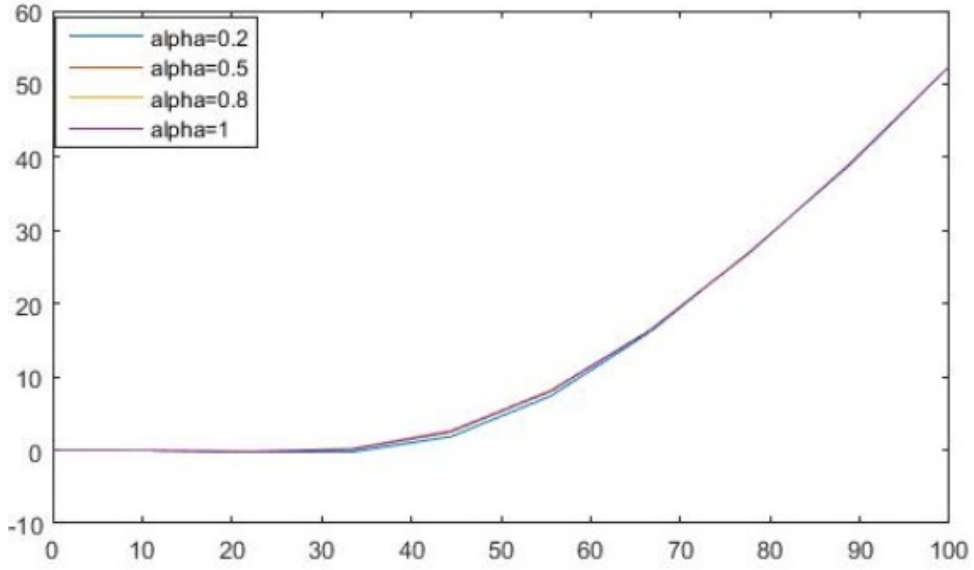


Figure 1: Fractional European Option Pricing for Different α and $N = M = 50$.

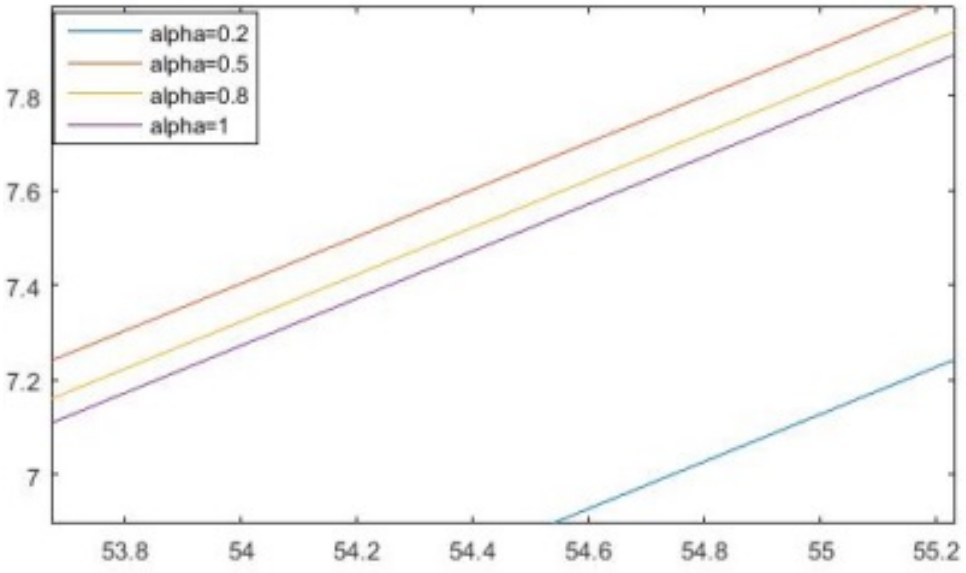


Figure 2: The restriction of Figure 1 for $53.5 \leq s \leq 55.5$.

Applying $I^{\alpha}_{0,t}$ on both sides of (10) and using the operational matrix of triangular functions, we get:

$$c(s,t) \simeq \mathbf{T}^T(t)(\mathbf{P}^{\alpha})^T X + r\mathbf{T}^T(t)(\mathbf{P}^{\alpha})^T P + \omega(s). \tag{11}$$

To calculate $\omega(s)$, we use the terminal condition $c(s,T) = V(s)$ in (3):

$$\omega(s) = V(s) - \mathbf{T}^T(T)(\mathbf{P}^{\alpha})^T X - r\mathbf{T}^T(T)(\mathbf{P}^{\alpha})^T P. \tag{12}$$

substituting (12) in (11), we have:

$$c(s,t) \simeq (\mathbf{T}^T(t) - \mathbf{T}^T(T))(\mathbf{P}^{\alpha})^T X + r(\mathbf{T}^T(t) - \mathbf{T}^T(T))(\mathbf{P}^{\alpha})^T P + V(s). \tag{13}$$

By equating (11) and (13), we get:

$$\begin{aligned}&(\mathbf{T}^T(t) - \mathbf{T}^T(T))(\mathbf{P}^{\alpha})^T X + r(\mathbf{T}^T(t) - \mathbf{T}^T(T))(\mathbf{P}^{\alpha})^T P + V(s) \\ &= \mathbf{T}^T(t)\mathbf{C}\mathbf{P}^2\mathbf{T}(s) + \frac{s}{B_u}(Q(t) - P(t) - \mathbf{T}^T(t)\mathbf{C}\mathbf{P}^2\mathbf{T}(B_u)) + P(t).\end{aligned} \tag{14}$$

In order to find the unknown $\mathbf{C}$ in (14), the collocation method is utilized. Thus, by putting the collocation points (s_i, t_j) for $i = 1, 2, ..., 2n$ and $j = 1, 2, ..., 2m$ in (14), we have a system of $2n \times 2m$ equations with an unknown matrix $\mathbf{C}_{2n\times 2m}$.
After finding the coefficient matrix $\mathbf{C}$, using function approximation for two variables case, European option pricing is as follows:

$$c(s,t) \cong \mathbf{T}^T(t)\mathbf{C}\mathbf{T}(s).$$

5 Numerical results

Example 5.1. Consider the following fractional Black-Scholes equation respect to time for European option:

$$\frac{\partial^{\alpha} c(s,t)}{\partial t^{\alpha}} + \frac{1}{2}\sigma^2 s^2 \frac{\partial^2 c(s,t)}{\partial s^2} + r\frac{\partial c(s,t)}{\partial s} - rc(s,t) = 0, \quad (s,t) \in (B_d, B_u) \times (0,T),$$
$$c(B_d,t) = P(t), \qquad c(B_u,t) = Q(t), \qquad c(s_{max},T) = V(s),$$

where $V(s) = max\{0, s_{max} - k\}$, $P(t) = Q(t) = 0$ for $B_d = 0$, $B_u = 100$, $k = 50$, $r = 0.05$, $\sigma = 0.25$, for the expire time $T = 1$. Figures 1 and 2 demonstrate the application of the presented method for $N = M = 50$ and some $0 < \alpha < 1$.

References

[1] T. Bjork, *Arbitrage Theory in Continuous Time*, Oxford University Press, New York, 2009.

[2] V. Daftardar, *Fractional Calculus*, Narosa Publishing House, 2014.

[3] M. N. Koleva,L. G. Vulkov, Numerical solution of time-fractional Black-Scholes equation, *SBMAC-Sociedade Brasileira de Matematica Aplicada Computacional*, 36 (2016), 1699–1715.

[4] H. Saeedi, A fractional-order operational method for numerical treatment of multi-order fractional partial differential equation with variable coefficients, *Sema Journal*, 75 (2018), No. 3, 421–433.

On Laplacian spectral characterization of generalized sun graphs[1]

Fatemeh Motialah* and Mohammad Hassan Shirdareh Haghighi

Department of Mathematics, Shiraz University, Shiraz, Iran

Abstract

A generalized sun graph $S(n,p)$ is the corona product of the cycle C_n and the empty graph of order p. We study Laplacian spectral characterization of generalized signed sun graphs and show that balanced generalized singed sun graphs can not be characterized by their Laplacian spectra.

Keywords: Generalized sun graph, Laplacian spectrum
Mathematics Subject Classification [2010]: 05C50

1 Introduction

In this paper we assume that all graphs are simple, i.e. without any loops or multiple edges.
Recall that a signed graph $\Lambda = (G, \sigma)$ is a simple graph $G = (V(G), E(G))$ equipped with a signed function $\sigma : E(G) \to \{+, -\}$.

The adjacency matrix of a signed graph $\Lambda = (G, \sigma)$ is defined as $A(\Lambda) = (a^{\sigma}_{ij})$ with $a^{\sigma}_{ij} = \sigma(ij)a_{ij}$ where $A(G) = (a_{ij})$ is the usual adjacency matrix of G. Also the Laplacian matrix of Λ is defined as $L(\Lambda) = D(G) - A(\Lambda)$. For a signed graph $\Lambda = (G, \sigma)$ and $U \subseteq V(G)$, let Λ^U be the signed graph obtained from Λ by reversing the signatures of the edges in the cut $[U, V(G) \setminus U]$. Namely $\sigma_{\Lambda^U}(e) = -\sigma_{\Lambda}(e)$ for any edge e between U and $V(G) \setminus U$ and $\sigma_{\Lambda^U}(e) = \sigma_{\Lambda}(e)$ otherwise. The signed graph Λ^U is called a switching of Λ, and Λ and Λ^U are called switching equivalent, for the following well-known and easy-to-prove theorem.

Theorem 1.1. *The adjacency (Laplacian) matrices of Λ and Λ^U are similar.*

Corollary 1.2. *The adjacency (Laplacian) matrices of Λ and Λ^U have the same characteristic polynomials.*

Definition 1.3. Two signed graphs are said to be A-cospectral (L-cospectral) if they have the same adjacency (Laplacian) characteristic polynomials. Also we say that a signed graph Λ is determined by its adjacency (Laplacian) spectrum if every graph that is A-cospectral (L-cospectral) to Λ is switching isomorphic to Λ.

[1] Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

*Speaker. Email address: fmotialah2011@yahoo.com

The sign of a cycle in a signed graph is positive, if it contains an even number of negative edges, otherwise it is negative.

Definition 1.4. A signed graph is said to be balanced if all of its cycles (if any) are positive, otherwise it is unbalanced.

Therefore, by this definition and the notion of switching isomorphism, basically there exists two non-switching isomorphic structures for a cycle C_n. One of them is $(C_n, +)$ in which all edges are positive, the other is $(C_n, -)$ in which just one edge is negative. Also note that the signs of the edges of an induced tree in a signed graph is irrelevant, since such edges can be changed to be positive by suitable switching. It follows that for a unicyclic graph G,there exist two non-switching isomorphic signed G which is determined by the sign of its cycle. We denote them by $(G, +)$ (balanced) and $(G, -)$ (unbalanced).

Lemma 1.5. *[2, Lemma 4.4] For the cycle C_n we have:*

$$\begin{aligned} Spec_A(C_n, +) &= \{2\cos\frac{2k}{n}\pi, k = 0, 1, \ldots, n-1\}, \\ Spec_A(C_n, -) &= \{2\cos\frac{2k+1}{n}\pi, k = 0, 1, \ldots, n-1\}, \end{aligned}$$

Definition 1.6. The corona of two graphs G_1 and G_2, denoted by $G_1 \circ G_2$ is the graph obtained from G_1 and n disjoint copies of G_2, where n is the order of G_1, such that each vertex of G_1 is adjacent to all vertices of a corresponding copy of G_2.

For example let $G_1 = C_4$ and $G_2 = K_2$. The two different coronas $G_1 \circ G_2$ and $G_2 \circ G_1$ are shown in Figure 1.

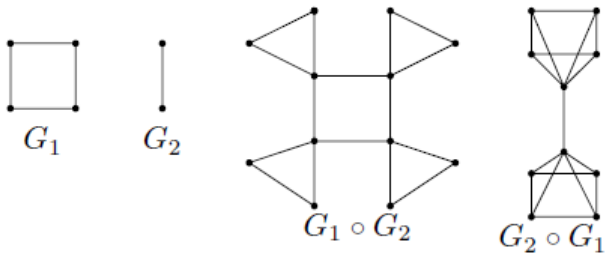


Figure 1: Coronas of two graphs

The corona of a cycle and a single vertex is called a sun. Laplacian spectra characterization of signed sun graphs is given in [4].

A generalized sun graph $S(n, p)$ is the corona product of the cycle C_n and the empty graph (graph with no edge) of order p.

When we have two signed graphs, for their corona to become a signed graph, different ways exist to give signs to the additional connecting edges [1]. The graph $S(n, p)$ is a unicyclic graph, when it is signed it has only two non-switching isomorphic structures, $(S(n, p), +)$, the balanced one; and $(S(n, p), -)$, the unbalanced one, in which in the latter only one edge of the cycle (which can be any of its edges) is negative.

As in [4], we now consider generalized signed sun graphs and show that the balanced generalized sun graph $(S(n, p), +)$ can not be characterized by its Laplacian spectrum

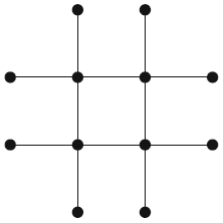


Figure 2: $S(4,2)$

whenever n is even. Therefore we can say generalized signed sun graphs can not be characterized by their Laplacian spectra.

We need the following well-known lemma, which can be proved plainly.

Lemma 1.7. *Let G be a graph with p adjacent pendant edges (pendant edges with a common vertex). Then it has 1 as a Laplacian eigenvalue with multiplicity at least $p-1$.*

By employing the proof of the Lemma above, we can easily conclude the following proposition.

Proposition 1.8. *The signed graph $\Lambda = (S(n,p),\sigma)$ has 1 as a Laplacian eigenvalue with multiplicity at least $n(p-1)$.*

2 Main results

First we derive the eigenvalues of $\Lambda = (S(n,p),\sigma)$, where $\sigma \in \{+,-\}$.

As we have seen in proposition 1.8, at least $n(p-1)$ eigenvalues of the Laplacian matrix of Λ are 1. Now we have to compute other Laplacian eigenvalues.

We consider the following labeling of vertices of Λ. First the vertices of the cycle come as $1,2,\dots n$. Then we pick one vertex attached to the vertex 1, one vertex attached to the vertex 2, and so on; repeating this process p times to complete labeling. The Laplacian matrix of Λ is then:

$$L(\Lambda) = \begin{pmatrix} (p+2)I_n - A(C_n,\sigma) & -I_n & -I_n & \dots & -I_n \\ -I_n & I_n & 0_n & \dots & 0_n \\ \vdots & & & & \\ -I_n & 0_n & 0_n & \dots & I_n \end{pmatrix}.$$

Theorem 2.1. *Let $\Lambda = (S(n,p),\sigma)$. The Laplacian eigenvalues of Λ are*

$$\frac{(3+p-\mu_i) \pm \sqrt{(3+p-\mu_i)^2 - 4(2-\mu_i)}}{2}$$

for $i = 0,\dots,n-1$; where $\mu_i = 2\cos\frac{2i}{n}\pi$ if $\sigma = +$ and $\mu_i = 2\cos\frac{2i+1}{n}\pi$ if $\sigma = -$.

Proof. Let $\psi(\Lambda,x)$ be the characteristic polynomial of $L(\Lambda)$ and $\phi(C_n,x)$ be the characteristic polynomials of $A(C_n)$. If $x \neq 1$, then by an elementary row operation we get

$$\psi(\Lambda,x) = \det(xI_{n+np} - L(\Lambda)) = \det(T),$$

where

$$T = \begin{pmatrix} ((x-(p+2)) - \frac{p}{(x-1)})I_n + A(C_n) & pI_n & (p-1)I_n & \dots & I_n \\ 0_n & (x-1)I_n & 0_n & \dots & 0_n \\ \vdots & & & & \\ 0_n & 0_n & 0_n & \dots & (x-1)I_n \end{pmatrix}.$$

Therefore, $\psi(\Lambda, x) = (x-1)^{np} \det(((x-(p+2)) - \frac{p}{(x-1)})I_n + A(C_n))$.

Since $x \neq 1$, $\frac{p}{(x-1)} - (x-(p+2))$ is an eigenvalue of $A(C_n)$, provided that x is a root of $\psi(\Lambda, x)$. If $\frac{p}{(x-1)} - x + p + 2 = \alpha$, we have $x^2 - (3+p-\alpha)x - (\alpha - 2) = 0$ and hence $x = \frac{(3+p-\alpha) \pm \sqrt{(3+p-\alpha)^2 - 4(2-\alpha)}}{2}$.

Thus, by Lemma 1.5, $\frac{(3+p-\mu_i) \pm \sqrt{(3+p-\mu_i)^2 - 4(2-\mu_i)}}{2}$ are the Laplacian eigenvalues of Λ, where $\mu_i = 2\cos\frac{2i+1}{n}\pi$ for $\sigma = -$, and $\mu_i = 2\cos\frac{2i}{n}\pi$ for $\sigma = +$, $i = 0, \ldots, n-1$. □

Corollary 2.2. *The Laplacian eigenvalues of signed graph $(S(n,p), \sigma)$ are 1 with multiplicity $n(p-1)$ and*

$$\frac{(3+p-\mu_i) \pm \sqrt{(3+p-\mu_i)^2 - 4(2-\mu_i)}}{2}$$

for $i = 0, \ldots, n-1$; where $\mu_i = 2\cos\frac{2i}{n}\pi$ if $\sigma = +$ and $\mu_i = 2\cos\frac{2i+1}{n}\pi$ if $\sigma = -$.

Proof. By Proposition 1.8 and Theorem 2.1, it is obvious. □

Theorem 2.3. *Let $\Lambda = (S(n,p), +)$ be a signed graph, where $n \geq 6$ is even. Then Λ is L-cospectral with $(S(\frac{n}{2}, p), +) \cup (S(\frac{n}{2}, p), -)$.*

Proof. By Theorem 2.1, Laplacian eigenvalues of Λ are $\frac{(3+p-\mu_i) \pm \sqrt{(3+p-\mu_i)^2 - 4(2-\mu_i)}}{2}$ for $i = 0, \ldots, n-1$; where $\mu_i = 2\cos\frac{2i}{n}\pi$. For i even, they are the eigenvalues of $(S(\frac{n}{2}, p), +)$, and for i odd, they are the eigenvalues of $(S(\frac{n}{2}, p), -)$. □

3 Conclusion

We showed that a generalized balanced sun graph $S(n,p)$ cannot be characterized by its Laplacian spectrum when n is even. This extends some results of [4].

References

[1] B. Adhikari, A. Singh, S. K. Yadav, Corona product of signed graphs and its application to signed network modelling, *arXiv: 1908.10018vl*, (2019).

[2] F. Belardo, P. Petecki, Spectral characterizations of signed lollipop graphs, *Linear Algebra Appl.*, 480 (2015), 144–167.

[3] F. Belardo, S. K. Simić, On the Laplacian coefficient of signed graphs, *Linear Algebra Appl.*, 475 (2015), 94–113.

[4] F. Motialah, M. H. Shirdareh Haghighi, Laplacian spectral characterization of signed sun graphs, *Theory and Applications of Graphs*, 6 (2019), Iss. 2, Art. 3.

Non-standard finite difference scheme for a fractional-order chaotic system[1]

Mehran Namjoo*, Mehdi Karami and Mehran Aminian

Department of Mathematics, Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran

Abstract

In this paper, we introduce a novel fractional-order chaotic autonomous system. Stability analysis of the fractional–order system is studied using the fractional Routh–Hurwitz criteria. The nonstandard finite difference (NSFD) scheme is implemented to study the dynamic behaviors in the novel fractional-order chaotic autonomous system. The lowest order for the system to remain chaotic is found. The numerical results show that the NSFD approach is easy and accurate when applied to fractional-order chaotic system.

Keywords: Chaos, Grunwald-Letnikov derivative, Stability, Fractional calculus, Nonstandard finite difference scheme

Mathematics Subject Classification [2010]: 60H15, 65M12

1 Introduction

In the recent years there is increasing interest in fractional calculus which deals with integration or differentiation of arbitrary orders. The list of applications of fractional calculus has been evergrowing and includes control theory, viscoelasticity, diffusion, turbulence, biology, economics, electromagnetism and many other physical processes. The interest in the study of fractional-order nonlinear systems lies in the fact that fractional derivatives provide an excellent tool for the description of memory and hereditary properties, which are not taken into account in the classical integer-order models. Studying dynamics in fractional-order nonlinear systems has become an interesting topic and the fractional calculus is playing a more and more important role for analysis of the nonlinear dynamical systems.

This paper is organized as follows: In next section, we give some basic definitions and properties of the Grünwald–Letnikov (GL) approximation and provide a brief overview of the important feature of the procedures for constructing NSFD schemes for ODEs. In section 3, we introduce a novel fractional-order chaotic autonomous system and also fractional Routh–Hurwitz stability conditions are given for the local asymptotic stability of the fractional systems. In section 4, we will discuss the stability analysis of fractional system. In addition, we present the idea of NSFD scheme for solving the novel fractional-order chaotic autonomous system. Numerical results show that the NSFD approach is easy to be implemented and accurated when applied to novel fractional-order autonomous system.

[1]Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

*Speaker. Email address: namjoo@vru.ac.ir

2 NSFD scheme and Grünwald–Letnikov approximation

The initial foundation of NSFD schemes came from the exact finite difference schemes. These schemes are well developed by Mickens [3] in the past decades. These schemes are developed for compensating the weaknesses such as numerical instabilities that may be caused by standard finite difference methods. Consider the autonomous ODE is given by

$$x' = f(t, x, \lambda), \qquad x(0) = x_0, \qquad t \in [0, t_f],$$

where λ is a parameter and $f(t, x, \lambda)$ is, in general, a nonlinear function. For a discrete-time grid with step size, $\triangle t = h$, we replace the independent variable t by $t \approx t_n = nh$, for $n = 0, 1, 2, \ldots, N$ where $h = \frac{t_f}{N}$. The dependent variable $x(t)$ is replaced by $x(t) \approx x_n$, where x_n is the approximation of $x(t_n)$. The NSFD scheme requires that x' has the more general representation $x' \cong \frac{x_{n+1}-x_n}{\phi}$, where the denominator function, i.e. ϕ has the property $\phi(h) = h + O(h^2)$. Derivatives of fractional–order have been introduced in several ways. In this paper we consider GL approach. The GL method of approximation for the one–dimensional fractional derivative is as follows:

$$D^\alpha x(t) = f(t, x(t)), \qquad x(0) = x_0, \qquad t \in [0, t_f], \tag{1}$$

$$D^\alpha x(t) = \lim_{h \to 0} h^{-\alpha} \sum_{j=0}^{[\frac{t}{h}]} (-1)^j \binom{\alpha}{j} x(t - jh),$$

where $0 < \alpha \leq 1$, D^α denotes the fractional derivative and h is the step size and $[\frac{t}{h}]$ denotes the integer part of $\frac{t}{h}$. Therefore, Eq. (1) is discretized as follows:

$$\sum_{j=0}^{n} c_j^\alpha x_{n-j} = f(t_n, x_n), \qquad n = 1, 2, 3, ...$$

where $t_n = nh$ and c_j^α are the GL coefficients defined as:

$$c_j^\alpha = (1 - \frac{1+\alpha}{j}) c_{j-1}^\alpha, \qquad c_0^\alpha = h^{-\alpha}, \qquad j = 1, 2, 3, ...$$

By applying this technique and using the GL discretization method, it yields the following relations

$$x_{n+1} = \frac{-\sum_{j=1}^{n+1} c_j^\alpha x_{n+1-j} + f(t_{n+1}, x_{n+1})}{c_0^\alpha}, \qquad n = 0, 1, 2, ...$$

where $c_0^\alpha = \phi(h)^{-\alpha}$.

3 The novel fractional-order chaotic autonomous system

Ref. [2] reported a three-dimensional autonomous system which relies on two multipliers and one quadratic term to introduce the nonlinearity necessary for folding trajectories. The chaotic attractor obtained from the new system according to the detailed numerical as well as theoretical analysis is also the butterfly shaped attractor, exhibiting the abundant and complex chaotic dynamics. This chaotic system is a new attractor which is similar

to Lorenz chaotic attractor. The chaotic system is described by the following non-linear integer-order differential equations

$$\begin{aligned} x^{'} &= -ax + fyz, \\ y^{'} &= cy - dxz, \\ z^{'} &= -bz + ey^2, \end{aligned} \tag{2}$$

where x, y, and z are the state variables, and a, b, c, d, e, and f are positive constant parameters. Now we introduce fractional–order into the system (2) of chaotic system. The new system is described by the following set of fractional ODEs of order $\alpha_1, \alpha_2, \alpha_3 > 0$, in the following form

$$\begin{aligned} &D^{\alpha_1} x = -ax + fyz, \\ &D^{\alpha_2} y = cy - dxz, \\ &D^{\alpha_3} z = -bz + ey^2, \\ &0 < \alpha_i \leq 1, \qquad i = 1, 2, 3. \end{aligned} \tag{3}$$

with initial condition

$$x(0) = x_0, \qquad y(0) = y_0, \qquad z(0) = z_0.$$

In order to analyze the stability of the system, fractional Routh–Hurwitz stability conditions for fractional–order differential equations are introduced. The Jacobian matrix J of the system Eqs. (3) at the equilibrium point $E = (x^*, y^*, z^*)$ is computed as

$$J(E) = \begin{pmatrix} -a & fz^* & fy^* \\ -dz^* & c & -dx^* \\ 0 & 2ey^* & -b \end{pmatrix}, \tag{4}$$

The existence and local stability conditions of these equilibrium point are as follows:

Let $D(P)$ denotes the discriminant of a polynomial P

$$P(\lambda) = \lambda^3 + a_1\lambda^2 + a_2\lambda + a_3 = 0, \tag{5}$$

and

$$D(P) = 18a_1a_2a_3 + (a_1a_2)^2 - 4a_3(a_1)^3 - 4(a_2)^3 - 27(a_3)^2,$$

using the results of [1], we have the following Routh–Hurwitz stability conditions for FDEs:

(i) If $D(P) > 0$, then the necessary and sufficient condition for the equilibrium point E to be locally asymptotically stable is $a_1 > 0, a_3 > 0, a_1a_2 - a_3 > 0$.

(ii) If $D(P) < 0$, $a_1 \geq 0, a_2 \geq 0, a_3 > 0$, then the equilibrium point E is locally asymptotically stable for $\alpha < 2/3$. However, if $D(P) < 0, a_1 < 0, a_2 < 0, \alpha > 2/3$, then all roots of polynomial (5) satisfy the condition $|arg(\lambda)| < \frac{\alpha\pi}{2}$.

(iii) If $D(P) < 0$, $a_1 > 0, a_2 > 0, a_1a_2 - a_3 = 0$, then the equilibrium point E is locally asymptotically stable for all $\alpha \in [0, 1)$.

(iv) The necessary condition for the equilibrium point E to be locally asymptotically stable is $a_3 > 0$.

In the next section we discuss the asymptotic stability of the equilibrium point E of the system Eqs. (3).

4 Stability analysis of the fractional–order chaotic autonomous system

When $a = 16$, $b = 5$, $c = 10$, $d = 6$, $e = 18$, and $f = 0.5$, the new system (2) has three real equilibrium points $E_1(0, 0, 0)$, $E_2(0.325, 1.4243, 7.303)$, $E_3(-0.325, -1.4243, 7.303)$. The local stability conditions of these equilibrium points are as follows.

Theorem 4.1. *For the parameters $a = 16$, $b = 5$, $c = 10$, $d = 6$, $e = 18$, and $f = 0.5$, the equilibrium point E_1 of system Eqs. (3) is unstable for any $\alpha \in (0, 1)$.*

Theorem 4.2. *When the parameters $a = 16$, $b = 5$, $c = 10$, $d = 6$, $e = 18$, and $f = 0.5$, if $\alpha < 0.89$, then equilibrium points E_2 and E_3 of system Eqs. (3) are stable.*

4.1 NSFD scheme for fractional–order system

By using definition of GL derivative and use NSFD for the system Eqs. (3) we have:

$$\begin{aligned}
\sum_{j=0}^{n+1} c_j^{\alpha_1} x_{n+1-j} &= -a x_{n+1} + f y_n z_n, \\
\sum_{j=0}^{n+1} c_j^{\alpha_2} y_{n+1-j} &= c y_n - d x_{n+1} z_n, \\
\sum_{j=0}^{n+1} c_j^{\alpha_3} z_{n+1-j} &= -b z_{n+1} + e y_{n+1}^2.
\end{aligned} \tag{6}$$

Doing some algebraic manipulation to Eqs. (6) yields the following relations

$$\begin{aligned}
x_{n+1} &= \frac{-\sum_{j=1}^{n+1} c_j^{\alpha_1} x_{n+1-j} + f y_n z_n}{c_0^{\alpha_1} + a}, \\
y_{n+1} &= \frac{-\sum_{j=1}^{n+1} c_j^{\alpha_2} y_{n+1-j} + c y_n - d x_{n+1} z_n}{c_0^{\alpha_2}}, \\
z_{n+1} &= \frac{-\sum_{j=1}^{n+1} c_j^{\alpha_3} z_{n+1-j} + e y_{n+1}^2}{c_0^{\alpha_3} + b},
\end{aligned}$$

where

$$c_0^{\alpha_1} = \phi_1(h)^{-\alpha_1}, \qquad c_0^{\alpha_2} = \phi_2(h)^{-\alpha_2}, \qquad c_0^{\alpha_3} = \phi_3(h)^{-\alpha_3},$$

with [4–6]

$$\phi_1(h) = \frac{e^{ah} - 1}{a}, \qquad \phi_2(h) = \frac{e^{ch} - 1}{c}, \qquad \phi_3(h) = \frac{e^{bh} - 1}{b}.$$

5 Simulation and results

Analytical studies always remain incomplete without numerical verification of the results. In this section, numerical results from the implementation of NSFD scheme for the novel

fractional-order autonomous system are presented. Using NSFD scheme, when $\alpha_1 = \alpha_2 = \alpha_3 = \alpha$, the simulation results demonstrate that the lowest order for the system (3) to remain chaotic is $\alpha = 0.89$. The approximate solutions are displayed in Figs. 1-3 for different $0 < \alpha_i \leq 1$, $i = 1, 2, 3$.

In Fig. 1 is depicted phase trajectory of the classical novel autonomous system (2) for commensurate order $\alpha = 1$ and parameters $a = 16$, $b = 5$, $c = 10$, $d = 6$, $e = 18$, and $f = 0.5$ with the initial conditions $(x(0), y(0), z(0)) = (0.05.0.05, 0.0001)$, for simulation time 20s and step size $h = 0.01$.

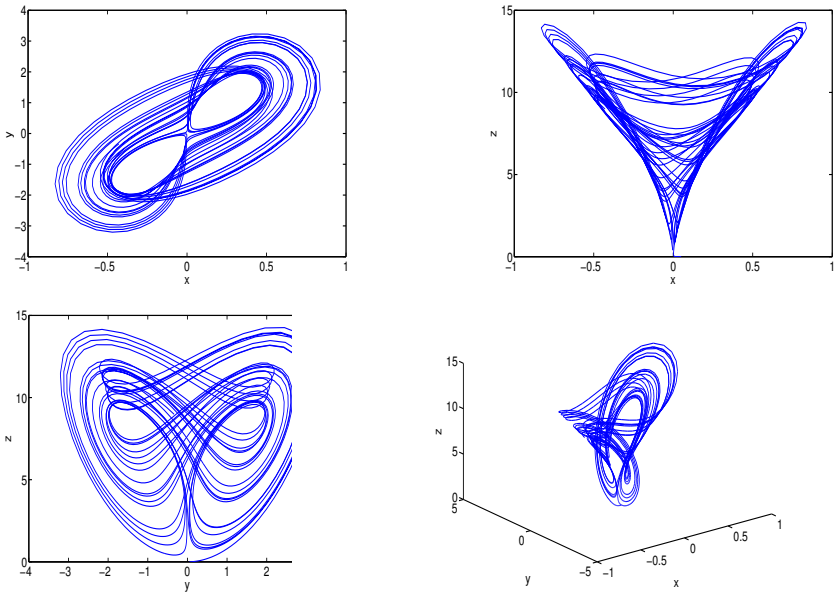


Figure 1: The chaotic attractor and the equilibrium point of system (3)when $\alpha = 1$.

In Fig. 2 is depicted phase trajectory of the fractional–order novel autonomous chaotic system (3) for commensurate order $\alpha = 0.88$ and parameters $a = 16$, $b = 5$, $c = 10$, $d = 6$, $e = 18$, and $f = 0.5$ with the initial condition $(x(0), y(0), z(0)) = (0.05.0.05, 0.0001)$, for simulation time 20s and step size $h = 0.01$.

In Fig. 3 is depicted phase trajectory of the fractional–order novel autonomous chaotic system (3) for incommensurate order $\alpha_1 = 0.88$, $\alpha_2 = 0.85$, $\alpha_3 = 0.80$ and parameters $a = 16$, $b = 5$, $c = 10$, $d = 6$, $e = 18$, and $f = 0.5$ with the initial condition $(x(0), y(0), z(0)) = (0.05.0.05, 0.0001)$, for simulation time 20s and step size $h = 0.01$.

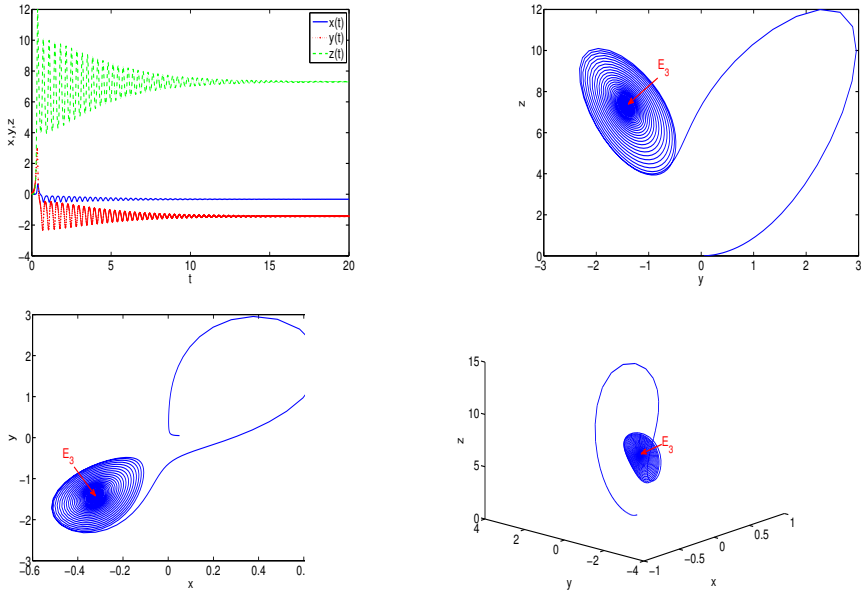


Figure 2: Phase trajectory of the equilibrium point E_3 when when $\alpha = 0.88$.

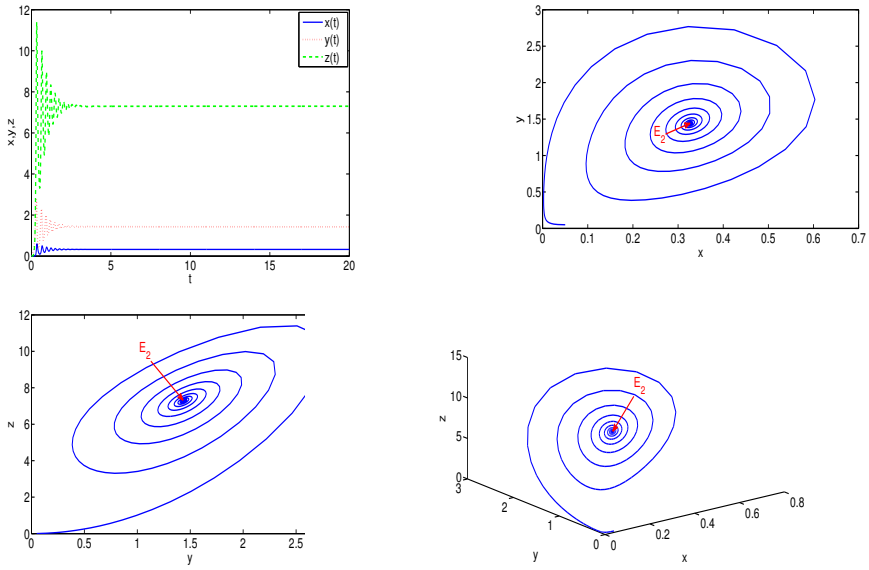


Figure 3: Phase trajectory of the equilibrium point E_2 when $\alpha_1 = 0.88$, $\alpha_2 = 0.85$ and $\alpha_3 = 0.80$.

References

[1] E. Ahmed, A. M. A. El-Sayed and H. A. A. El-Saka, *On some Routh-Hurwitz conditions for fractional order differential equations and their applications in Lorenz, Rössler, Chua and Chen systems*, Phys Lett A 358 (2006), pp. 1-4.

[2] A. Gholizadeh, H. Saberi Nik and A. Jajarmi, *Analysis and control of a three-dimensional autonomous chaotic system*, Appl. Math. Inf. Sci. 9 (2015), No. 2, 739-

747.

[3] R. E. Mickens, *Advances in the Applications of Nonstandard Finite Difference Schemes*, Wiley-Interscience, Singapore, (2005).

[4] 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 Journal of Mathematical Chemistry 6 (2015), No. 2, pp. 145-160.

[5] S. Zibaei and M. Namjoo, *A Nonstandard Finite Difference Scheme for Solving Three-Species Food Chain with Fractional-Order Lotka-Volterra Model*, Iranian Journal of Numerical Analysis and Optimization 6 (2016), No. 1, pp. 53-78.

[6] S. Zibaei and M. Namjoo, *A NSFD scheme for Lotka-Volterra food web model*, Iran. J. Sci. Technol. Trans. A Sci. 38 (2014), No. 4, pp. 399-414.

The trace class on the proper H^*-algebra structure[1]

Akbar Nazari and Mohammad Amin Moarrefi*

Department of Pure Mathematices, Faculty of Mathematics & Computer,
Shahid Bahonar University of Kerman, Kerman, Iran

Abstract

Warren Ambrose defined a general structure, namely H^*-algebra. Saworotnow discussed trace class on these structures. In this work, we investigate this algebra and this trace class. Also, we mention some properties of trace functional. Finally, a norm, based on the trace functional, and some examples on these subjects are described.

Keywords: H^*-algebra, Trace class

Mathematics Subject Classification [2010]: 06F25, 16R30

1 Introduction

The word Algebra was first used by al-Khwarizmi in 780-850. Different structures are derived from Algebra.

An algebra over field $\mathbb{F}$ is a vector space $\mathcal{A}$ over $\mathbb{F}$ that also has a multiplication defined on it that makes $\mathcal{A}$ into a ring such that if $\alpha \in \mathbb{F}$ and $a, b \in \mathcal{A}$, $\alpha(ab) = (\alpha a)b = a(\alpha b)$.

A Banach algebra is an algebra $\mathcal{A}$ over field $\mathbb{F}$ that has a norm $\|\cdot\|_{\mathcal{A}}$ relative to which $\mathcal{A}$ is a Banach space and such that for all $a, b \in \mathcal{A}$,

$$\|ab\|_{\mathcal{A}} \leq \|a\|_{\mathcal{A}} \|b\|_{\mathcal{A}} \qquad \text{(Algebra norm)}$$

For a Banach algebra $\mathcal{A}$, an involution is a map $a \mapsto a^*$ from $\mathcal{A}$ into $\mathcal{A}$ such that the following properties hold for $a, b \in \mathcal{A}$ and $\alpha \in \mathbb{C}$:

$$(a^*)^* = a, \qquad (ab)^* = b^* a^*, \qquad (\alpha a + b)^* = \bar{\alpha} a^* + b^*.$$

Each Banach algebra equipped with an involution is called Banach $*$-algebra or B^*-algebra.

Definition 1.1. A C^*-algebra is a Banach algebra $\mathcal{A}$ with an involution such that for every $a \in \mathcal{A}$,

$$\|a^* a\|_{\mathcal{A}} = \|a\|_{\mathcal{A}}^2 .$$

[1] Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

* Speaker. Email address: ma_moarrefi@yahoo.com

2 H^*-Algebra

Warren Ambrose defined an H^*-algebra structure in [1]. He defined H^*-algebras as a generalization of L_2-algebras. In continue, we survey on the definition and some properties of these structures. Also present some examples for this algebra. Then checking the difference between that and the similar algebra. Also, discuss on involution $*$ and proper H^*-algebras. Finally express structure theorem for H^*-algebra.

Definition 2.1 ([1]). The Banach algebra $\mathcal{A}$ is called H^*-algebra, if satisfies the following conditions:

i. The underlying Banach space of $\mathcal{A}$ is Hilbert space of arbitrary dimention.

ii. For each $a \in \mathcal{A}$, there exist adjoint $a^* \in \mathcal{A}$ such that

$$\langle ab, c\rangle_{\mathcal{A}} = \langle b, a^*c\rangle_{\mathcal{A}} \quad \text{and} \quad \langle ab, c\rangle_{\mathcal{A}} = \langle a, cb^*\rangle_{\mathcal{A}} \tag{1}$$

for all $a, b, c \in \mathcal{A}$.

This means, an H^*-algebra $\mathcal{A}$ is a (real or complex) Banach algebra and a (real or complex) Hilbert space with an inner product $\langle a, b\rangle_{\mathcal{A}}$ such that the algebra norm $\|a\|_{\mathcal{A}}$ and the Hilbert space norm $\langle a, a\rangle_{\mathcal{A}}^{1/2}$ are equal for all $a \in \mathcal{A}$. And also, for every element $a \in \mathcal{A}$ there is **at least** one adjoint element a^* satisfies in (1).

Example 2.2. We explain two examples for H^*-algebras and express relation between these algebras and C^*-algebras:

i. Complex number, $\mathbb{C}$, with inner product $\langle \alpha, \beta\rangle = \mathrm{Re}\,(\alpha\beta^*)$ and induced norm by this inner product, is an H^*-algebra and C^*-algebra, because $\langle \alpha\beta, \gamma\rangle = \mathrm{Re}\,(\alpha\beta\overline{\gamma}) = \langle \beta, \alpha^*\gamma\rangle = \langle \alpha, \gamma\beta^*\rangle$ and $\|\alpha^*\alpha\| = \|\alpha\|^2$ where for every complex number η, $\eta^* = \overline{\eta}$ and $\|\eta = \eta_1 + \eta_2 i\| = \sqrt{\langle \eta, \eta\rangle} = \sqrt{\mathrm{Re}\,(\eta\eta^*)} = \eta_1^2 + \eta_2^2$.

ii. The Clifford algebra $\mathcal{A} = Cl_{0,n}$ is a finite-dimension real H^*-algebra with respect to $\langle \lambda, \mu\rangle := 2^n[\lambda\overline{\mu}]_0 = 2^n \sum_A \lambda_A \mu_A$.

Let $|\lambda|_0^2 := \langle \lambda, \lambda\rangle_0 = 2^n \sum_A \lambda_A^2$ be an induced norm by above inner product on $C\ell_{02}$ (i.e. $i^2 = j^2 = -1$). Consider $\lambda = i + j$, then $\lambda\overline{\lambda} = 2$, $|\lambda\overline{\lambda}|_0 = 4$ and $|\lambda|_0^2 = 2^2(1^2 + 1^2) = 8$. Hence $|\lambda\overline{\lambda}|_0 \neq |\lambda|_0^2$.

Therefore $Cl_{0,n}$ with this inner product and induced norm is an H^*-algebra and is not a C^*-algebra (For more information see [4]).

The adjoint a^* of a is not unique. Consider any Hilbert space and make it into an algebra by defining the $\langle a, b\rangle := 0$. It is trivial that this is an H^*- algebra in which every element is an adjoint of every element.

For every H^*-algebra $\mathcal{A}$, $a\mathcal{A} = (0)$ is equivalent to $\mathcal{A}a = (0)$, for all $a \in \mathcal{A}$.

Definition 2.3 ([1]). An H^*-algebra is proper or semi-simple if the only $a \in \mathcal{A}$ satisfies in $a\mathcal{A} = (0)$ is $a = 0$.

Theorem 2.4 ([1]). *An H^*-algebra is proper if and only if every element has a unique adjoint.*

Definition 2.5. Let $\mathcal{A}$ be a proper H^*-algebra and $a, e, f \in \mathcal{A}$. Then

i. a is self-adjoint member of $\mathcal{A}$ if $a^* = a$;

ii. a is positive member of $\mathcal{A}$ if $\langle ax, x\rangle_{\mathcal{A}} \geq 0$ for all $x \in \mathcal{A}$;

iii. a is normal element if $a^*a = aa^*$;

iv. e is idempotent if $e^2 = e \neq 0$;

v. e is sa-idempotent or projection if e be an idempotent and self-adjoint, i.e. $e^2 = e = e^* \neq 0$;

vi. The non-zero idempotents e, f are called doubly orthogonal if $ef = fe = 0$ and $\langle e, f\rangle_{\mathcal{A}} = 0$;

vii. An idempotent is primitive if it can not be expressed as the sum of two doubly orthogonal idempotents.

Theorem 2.6 ([1]). *Every proper H^*-algebra contains an sa-idempotent.*

Theorem 2.7 ([1]). *Every proper H^*-algebra contains a (non-empty) maximal family of doubly orthogonal primitive sa-idempotents*

Theorem 2.8 (First structure theorem, [[5]). *] Let $\{e_i\}$ be a maximal family of doubly orthogonal primitive sa-idempotents in a proper H^*-algebra $\mathcal{A}$. Then*

$$\mathcal{A} = \sum_i e_i A = \sum_i A e_i,$$

that is, $\mathcal{A}$ is the direct sum of the minimal left ideals $\mathcal{A}e_i$ and $\mathcal{A}$ is a direct sum of the minimal right ideals $e_i\mathcal{A}$.

2.1 Trace-Class for H^*-Algebras

In this section, $\mathcal{A}$ is a proper H^*-algebra. We describe definition of trace–class and trace–functional for H^*-algebras. After that, we explain some consequences and relations about these.

Lemma 2.9 ([6]). *Let b be a normal element in $\mathcal{A}$. Then there exists a projection base $\{e_\alpha\}_{\alpha\in\Lambda}$ for $\mathcal{A}$ and a family $\{\lambda_\alpha\}_{\alpha\in\Lambda}$ of scalars such that $b = \sum_{\alpha\in\Lambda} \lambda_\alpha e_\alpha$. The nonzero numbers λ_α are nonzero numbers in the spectrum of b.*

If $b = a^*a$ for some $a \in \mathcal{A}$ then every $\lambda_\alpha \geq 0$.

Corollary 2.10 ([6]). *For each $a \neq 0$ in $\mathcal{A}$ there exists a sequence $\{e_n\}$ of mutually orthogonal projections and a sequence $\{\lambda_n\}$ of positive numbers such that*

$$a^*a = \sum_n \lambda_n e_n. \tag{2}$$

Note also that $a^*ae_n = e_n a^*a = \lambda_n e_n$ for each n.

Define $[a] := \sum_n \mu_n e_n$, where $\mu_n := \sqrt{\lambda_n} \geq 0$, For every $n \in \mathbb{N}$, in Equation (2) of Corollary 2.10. Then for each k:

$$\begin{aligned}\sum_{n=1}^{k} \|\mu_n e_n\|^2_{\mathcal{A}} = \sum_{n=1}^{k} \mu_n^2 \langle e_n, e_n\rangle_{\mathcal{A}} &= \sum_{n=1}^{k} \langle \lambda_n e_n, e_n\rangle_{\mathcal{A}} \\ &= \sum_{n=1}^{k} \langle a^*ae_n, e_n\rangle_{\mathcal{A}} = \sum_{n=1}^{k} \langle ae_n, ae_n\rangle_{\mathcal{A}}\end{aligned}$$

$$= \sum_{n=1}^{k} \|ae_n\|_{\mathcal{A}}^2 \leq \|a\|_{\mathcal{A}}^2 .$$

Therefore $\sum_n \mu_n e_n$ is converges. Hence $[a] = \sum_n \mu_n e_n$ is well-define.

Lemma 2.11 ([6]). *For each $a \in \mathcal{A}$ there exists a unique positive member $[a]$ of $\mathcal{A}$ such that $[a]^2 = a^*a$ (note that $[a]^* = [a]$).*

The trace–class for $\mathcal{A}$ is the set

$$\tau(\mathcal{A}) = \{xy | x, y \in \mathcal{A}\},$$

i.e. $\tau(\mathcal{A})$ be the set of all products xy of members x, y of $\mathcal{A}$. Every trace–class is non–empty, because each idempotent element $e^2 = e \in \mathcal{A}$ is belong to $\tau(\mathcal{A})$, see Theorem 2.6.

If $a \in \tau(\mathcal{A})$, then $a = xy$ for some $x, y \in \mathcal{A}$. We define

$$\operatorname{tr} a := \langle y, x^* \rangle_{\mathcal{A}} .$$

In every proper H^*-algebra $\operatorname{tr}(ab) = \operatorname{tr}(ba)$, for each $a, b \in \mathcal{A}$. The trace tr is a positive functional, i.e. $\operatorname{tr}(a) \geq 0$, for every positive element $a \in \mathcal{A}$. There exists $b \in \mathcal{A}$ such that $\operatorname{tr}(b) < 0$. Therefore in follow use "$[a]$" to build a norm according to the trace functional, because $[a]$ is a positive member of $\mathcal{A}$, for every a.

Definition 2.12 ([6]). We define $\tau(a) := \operatorname{tr}([a]) = \operatorname{tr}(\sum_{n=1}^{\infty} \mu_n e_n) = \sum_{n=1}^{\infty} \mu_n$ for every $a \in \mathcal{A}$.

Corollary 2.13 ([6]). *Suppose $\mathcal{A}$ be an H^*-algebra. Then*

i. $\tau(a^*a) = \operatorname{tr}(a^*a) = \|a\|_{\mathcal{A}}^2$, *for all* $a \in \mathcal{A}$;

ii. $|tra| \leq \tau(a)$, *for all* $a \in \tau(\mathcal{A})$;

iii. $\|a\|_{\mathcal{A}} \leq \tau(a)$, *for all* $a \in \tau(\mathcal{A})$;

iv. $\tau(ab) \leq \|a\|_{\mathcal{A}} \cdot \|b\|_{\mathcal{A}}$, *for all* $a, b \in \mathcal{A}$;

v. $\tau(ab) \leq \tau(a)\tau(b)$, *for all* $a, b \in \tau(\mathcal{A})$.

3 Some Examples

Example 3.1. Consider a structure $(\ell_2(\mathbb{N}), +, \cdot, \langle \cdot, \cdot \rangle)$ with the standard addition and scalar product. Suppose $a = (a_1, a_2, \cdots) = \{a_i\}_{i=1}^{\infty}, b = (b_1, b_2, \cdots) = \{b_i\}_{i=1}^{\infty} \in \ell_2(\mathbb{N})$ where $a_i, b_i \in \mathbb{F} = \mathbb{C}$. We define

$$a \cdot b = (a_1 \cdot b_1, a_2 \cdot b_2, \cdots) = \{a_i \cdot b_i\}_{i=1}^{\infty} \qquad \text{(product)}$$

$$\langle a, b \rangle_{\mathcal{A}} = \sum_{i=1}^{\infty} a_i \overline{b_i} \qquad \text{(inner product)}$$

$$a^* = (\overline{a_1}, \overline{a_2}, \cdots) \qquad \text{(adjoint)}$$

where $\bar{\cdot}$ is conjugate of a complex number. Then we know that $\langle \cdot, \cdot \rangle_{\mathcal{A}}$ is an inner product and $\|a\|_{\mathcal{A}} = \langle a, a \rangle_{\mathcal{A}}^{\frac{1}{2}} = (\sum_{n=1}^{\infty} |a_n|^2)^{\frac{1}{2}}$ is induced norm. So $\mathcal{A}$ is a Hilbert space and a Banach algebra. Also $*$ is an involution.

$\mathcal{A}$ is an H^*-algebra, because

$$\begin{aligned}\langle ab, c\rangle_{\mathcal{A}} &= \sum a_i b_i \overline{c}_i,\\ \langle b, a^*c\rangle_{\mathcal{A}} &= \sum b_i \overline{\overline{a_i} c_i} = \sum a_i b_i \overline{c}_i,\\ \langle a, cb^*\rangle_{\mathcal{A}} &= \sum a_i \overline{c_i \overline{b_i}} = \sum a_i b_i \overline{c}_i,\end{aligned}$$

are equal. But it has not C^*-algebra structure. For check this, let $a = (2, 3, 0, 0, \cdots)$. Then $\|a\|^2_{\mathcal{A}} \neq \|a^*a\|_{\mathcal{A}}$.

It is easy to show that $1_{\mathcal{A}}$ must be $\{1_{\mathbb{R}}\}_{i=1}^{\infty}$ and it is not belong to $\ell_2(\mathbb{N})$.

Let $\delta_i = \{\delta_{ij}\}_{j=1}^{\infty}$, $i \in \mathbb{N}$, where δ_{ij} is the Kronecker delta. Then $\{\delta_i\}$ is family of doubly orthogonal primitive *sa*-idempotents.

For every $a = \{a_n\}_{n=1}^{\infty}$, $[a] = \{|a_n|\}_{n=1}^{\infty}$. Then $\operatorname{tr}(a) = \operatorname{tr}(\{a_n\}_{n=1}^{\infty}) = \sum_{n=1}^{\infty} a_n$ and $\tau(a) = \operatorname{tr}([a]) = \sum_{n=1}^{\infty} |a_n|$. Therefore $\|a\|_{\mathcal{A}} = \langle a, a\rangle_{\mathcal{A}}^{\frac{1}{2}} = \sqrt{\operatorname{tr}(a^*a)} = \sqrt{\tau(a^*a)}$.

Example 3.2. Consider a structure $(M_2(\mathbb{C}), +, \times, \text{scalar product})$ with the common matrix addition, matrix product and the scalar product. Also, involution on a matrix A be conjugate transpose of A, i.e. $A^* = \overline{A^T}$. Suppose $A, B \in M_2(\mathbb{C})$ where $a_{ij}, b_{ij} \in \mathbb{F} = \mathbb{C}$. We define inner product

$$\langle A, B\rangle_{\mathcal{A}} = \sum_{i=1}^{2}\sum_{j=1}^{2} a_{ij} \cdot \overline{b_{ij}}.$$

Then $\langle \cdot, \cdot\rangle_{\mathcal{A}}$ is an inner product and $\mathcal{A}$ is a Hilbert space.

Induced norm is $\|\cdot\|^2_{\mathcal{A}} = \langle A, A\rangle_{\mathcal{A}} = \sum_{i=1}^{2}\sum_{j=1}^{2} a_{ij} \cdot \overline{a_{ij}} = \sum_{i=1}^{2}\sum_{j=1}^{2} |a_{ij}|^2$, and it is Frobenius norm. So, $\mathcal{A}$ is a Banach algebra.

$\mathcal{A}$ is an H^*-algebra, becuse

$$\begin{aligned}\langle A \times B, C\rangle &= \langle B, A^* \times C\rangle = \langle A, C \times B^*\rangle\\ &= a_{11}b_{11}\overline{c_{11}} + a_{12}b_{21}\overline{c_{11}} + a_{11}b_{12}\overline{c_{12}} + a_{12}b_{22}\overline{c_{12}}\\ &\quad + a_{21}b_{11}\overline{c_{21}} + a_{22}b_{21}\overline{c_{21}} + a_{21}b_{12}\overline{c_{22}} + a_{22}b_{22}\overline{c_{22}},\end{aligned}$$

but it is not a C^*-algebra, because $\|I^*I\| \neq \|I\|^2$.

$1_{\mathcal{A}} = I_{2\times 2} \in \mathcal{A}$, but $\|1_{\mathcal{A}}\|_{\mathcal{A}} \neq 1$. Then $\mathcal{A}$ is not unital Banach algebra.

Now calculate trace functional tr for each $A \in \mathcal{A}$

$$\operatorname{tr}(A) = \operatorname{tr}(AI) = \langle A, I\rangle = a_{11} + a_{22}.$$

For check the part(i) of Proposition 2.13, let $A = \begin{pmatrix} a & b\\ c & d\end{pmatrix}$. Then $A^*A = \begin{pmatrix} \bar{a}a + \bar{c}c & \bar{a}b + \bar{c}d\\ \bar{b}a + \bar{d}c & \bar{b}b + \bar{d}d\end{pmatrix}$. And $\operatorname{tr}(A^*A) = |a|^2 + |c|^2 + |b|^2 + |d|^2 = \|A\|^2$.

4 Conclusion

An H^*-algebra structure is a complete vector space with a multiplication on its elements such that it has a particular inner product. Then investigate a trace-class sub-algebra of this and a trace functional and a trace norm. Finally, we mentioned some properties of these.

References

[1] W. Ambrose, Structure theorems for a special class of Banach algebras, *Trans. Amer. Math. Soc.*, 57 (1945), 364–386.

[2] J. B. Conway, *A Course in Functional Analysis*, Springer, New York, 1990.

[3] C. N. Kellogg, Centralizers and H^*-algebras, *Pacific Journal of Mathematics*, 17 (1966), 121–129.

[4] M. A. Moarrefi and A. Nazari, Investigate of the C^*-Algebraic Structure of the Clifford Algebra, 49th Annual Iranian Mathematics Conference, Elm-o-Sanat University, Tehran, Iran, pp 3069–3075, 2018.

[5] L. Molnar, Modular bases in a hilbert A-module, *Czechoslovak Mathematical Journal*, 42 (1992), 649–656.

[6] P. P. Saworotnow and J. C. Friedell, Trace-class for an arbitrary H^*-algebra, *Proc. Amer. Math. Soc.*, 26 (1970), 95–100.

A remarkable solution to symmetric inverse eigenvalue problem[1]

Alimohammad Nazari* 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 numbers σ via a special unit lower triangular matrix, we find a symmetric matrix such that σ is its spectrum and in continue we bring a conditon for solving symmetric inverse eigenvalue problem(SNIEP).

Keywords: Symmetric nonnegative inverse eigenvalue problem, Spectrum of matrix, Perron eigenvalue

Mathematics Subject Classification [2010]: 15A29, 15A18

1 Introduction

The nonnegative inverse eigenvalue problem (NIEP) asks for necessary and sufficient conditions on a list $\sigma = (\lambda_1, \lambda_2, \ldots, \lambda_n)$ of real or complex numbers in order to σ be a spectrum of a nonnegative matrix A, we will say that σ is realizable and that it is realization of σ. Since all eigenvalues of symmetric matrices are real, then if $\sigma = (\lambda_1, \lambda_2, \ldots, \lambda_n)$ of real numbers then (SNIEP) is symmetric nonnegative inverse eigenvalue problem and if we find a symmetric matrix C with eigenvalues σ, then we will say that σ is symmetrically realizable and that it is symmetric realization of σ.

In [1] Fiedler obtained some necessary and some sufficient conditions for a set of n real numbers $\sigma = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ that it to be the set of eigenvalues of $n \times n$ symmetric nonnegative matrix. Although nonsymmetric inverse eigenvalue problem (NIEP) has not been solved in general, however sufficient conditions for the SNIEP have been obtained in [3].

Some necessary conditions on the list of real number $\sigma = (\lambda_1, \lambda_2, \ldots, \lambda_n)$ to be the spectrum of a nonnegative matrix are listed below.
(1) The Perron eigenvalue $\max\{|\lambda_i|; \lambda_i \in \sigma\}$ belongs to σ (Perron-Frobenius theorem).
(2) $s_k = \sum_{i=1}^{n} \lambda_i^k \geq 0$.
(3)$s_k^m \leq n^{m-1} s_{km}$ for $k, m = 1, 2, \ldots$ (JLL inequality) [2, 3]. One of the special and interesting cases of SNIEP is inverse eigenvalues of Euclidean distance matrix (EDM). For instance, T.L. Hayden, R. Reams and J. Wells have solved the inverse eigenvalue problem for Euclidean distance matrices of order $n = 3, 4, 5, 6$, and any n for which there exists a Hadamard matrix and also they solved this problem: If for $n \in \mathbb{N}$ there exists a Hadamard matrix of order n, then there is an $(n+1) \times (n+1)$ and an $(n+2) \times (n+2)$ distance

[1]Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University
*Speaker. Email address: a-nazari@araku.ac.ir

matrix with eigenvalues which hold under special conditions for $n \leqslant 16$ [5]. Nazari and Mahdinasab solved this problem without using any Hadamard matrix [4].

A matrix L is called unit lower triangular if it is lower triangular matrix and all entries on its main diagonal are one. The inverse such a matrix also is unit lower triangular. In Gaussian elimination method and LU factorization unit lower triangular matrices play an important role.

Recently, Nazari and Nezami were able to solve the inverse eigenvalue problem in general by using unit lower triangular matrices [6]. In this paper for a given set of real numbers by helping a special unit lower triangular matrix we find a symmetric matrix C, such that σ is its spectrum and in continue we find a sufficient condition that C is nonnegative symmetric matrix.

2 Main results

In this paper for a given set of real numbers σ by helping of similarity of matrices and via a special unit lower triangular matrices, we find a symmetric matrix that σ is its spectrum.

Lemma 2.1. *If*

$$L = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & & 0 \\ \vdots & & & \\ 1 & 1 & \cdots & 1 \end{pmatrix}$$

is $k \times k$ unit lower triangular matrix, then

$$L^{-1} = \begin{pmatrix} 1 & 0 & & 0 \\ -1 & 1 & & 0 \\ 0 & & & \\ \vdots & & & \\ 0 & \cdots & -1 & 1 \end{pmatrix}.$$

Proof. It is easy to see that $LL^{-1} = LL^{-1} = I_k$, that I_k is $k \times k$ identity matrix. □

Theorem 2.2. *Let $\sigma = \{\lambda_1, \lambda_2, \cdots, \lambda_n\}$ is set of real numbers, then there exist a symmetric matrix that σ is it's spectrum.*

Proof. Let

$$A = \begin{pmatrix} \lambda_1 & a_{12} & a_{13} & \cdots & a_{1n} \\ 0 & \lambda_2 & a_{23} & & a_{2n} \\ 0 & 0 & \lambda_3 & & a_{3n} \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_{nn} \end{pmatrix},$$

and

$$L = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & & 0 \\ 1 & 1 & 1 & & 0 \\ \vdots & & & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & 1 \end{pmatrix},$$

is unit lower triangular matrix. Now by Lemma 2.1, we compute

$$C = L^{-1}AL = \begin{pmatrix} 1 & 0 & & 0 \\ -1 & 1 & & 0 \\ 0 & & & \\ \vdots & & & \\ 0 & \cdots & -1 & 1 \end{pmatrix} \begin{pmatrix} \lambda_1 & a_{12} & a_{13} & \cdots & a_{1n} \\ 0 & \lambda_2 & a_{23} & & a_{2n} \\ 0 & 0 & \lambda_3 & & a_{3n} \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_{nn} \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & & 0 \\ 1 & 1 & 1 & & 0 \\ \vdots & & & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & 1 \end{pmatrix}. \tag{1}$$

Finally we find the entries $a_{12}, \cdots, a_{1n}, a_{23}, \cdots, a_{2n} \cdots, a_{nn-1}$ such that the matrix C to be symmetric. For this we select the elements on upper triangular matrix A as following

$$a_{ij} = \frac{i}{j}(\lambda_j - \lambda_{j-1}), \qquad i = 1, \cdots, n-1, \ j = i+1, \cdots, n. \tag{2}$$

Then the matrix C is similar to matrix A and is symmetric because without lose of generality we assume that $i < j$, then

$$C_{ij} = (a_{ij} - a_{i-1j}) + (a_{ij+1} - a_{i-1j+1}) + \cdots + (a_{in} - a_{i-1n}), \tag{3}$$

if we substitute (2) in to (3) then after simplification we have

$$C_{ij} = -\frac{1}{j}\lambda_{j-1} + \frac{1}{j(j+1)}\lambda_j + \frac{1}{(j+1)(j+2)}\lambda_{j+1} + \cdots + \frac{1}{(n-1)(n-2)}\lambda_{n-1} + \frac{1}{n}\lambda_n. \tag{4}$$

On the other hand we have

$$C_{ji} = (a_{ji} - a_{j-1i}) + (a_{ji+1} - a_{j-1i+1}) + \cdots + (a_{jj} - a_{j-1j}) + (a_{jj+1} - a_{j-1j+1}) + \cdots + (a_{jn} - a_{j-1n}) \tag{5}$$

Since $i < j$ and A is upper triangular matrix, then from (5) we have

$$C_{ji} = (a_{jj} - a_{j-1j}) + (a_{jj+1} - a_{j-1j+1}) + \cdots + (a_{jn} - a_{j-1n}). \tag{6}$$

It is easy to see that ralation (6) after replacing (2) and relation (4) are equal. Then $C_{ij} = C_{ji}$ for all $i < j$. Therefore the matrix C is symmetric and has eigenvalues σ. □

Theorem 2.3. *The necessary and sufficient condition for the matrix C in (1) to be nonnegative is that*

$$\frac{1}{n}(\lambda_n + (n-1)\lambda_{n-1}) \geq 0.$$

Corollary 2.4. *If $\sigma = \{\lambda_1, \lambda_2, \ldots, \lambda_n\} \subset \mathbb{Q}$, then the symmetric matrix C with eigenvalues σ lies in $\mathbb{Q}^{n\times n}$.*

Proof. Since $\sigma = \{\lambda_1, \lambda_2, \ldots, \lambda_n\} \subset \mathbb{Q}$, then by (2) the matrix $A \in \mathbb{Q}^{n\times n}$, and due to the structure of the matrix C, this matrix has rational entries. □

Example 2.5. Consider $\sigma = \{12, -4, -3, -2\}$ then

$$L^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix},$$

$$A := \begin{bmatrix} -2 & -1/2 & -1/3 & 4 \\ 0 & -3 & -2/3 & 8 \\ 0 & 0 & -4 & 12 \\ 0 & 0 & 0 & 12 \end{bmatrix}$$

and

$$L := \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

then

$$C := L^{-1}AL = \begin{bmatrix} 7/6 & \frac{19}{6} & 11/3 & 4 \\ \frac{19}{6} & 7/6 & 11/3 & 4 \\ 11/3 & 11/3 & 2/3 & 4 \\ 4 & 4 & 4 & 0 \end{bmatrix}.$$

References

[1] M. Fiedler, Eigenvalues of nonnegative symmetric matrices, Linear Algebra Appl. 9 (1974) 119-142.

[2] R. Lowey, D. London, A note on an inverse problem for nonnegative matrices, Linear and Multilinear Algebra 6(1978)83–90.

[3] C.R. Johnson, Row stochastic matrices similar to doubly stochastic matrices, Linear and Multilinear Algebra 10 (2) (1981) 113-130.

[4] A.M. Nazari, F. Mahdinasab, Inverse eigenvalue problem of distance matrix via orthogonal matrix, Linear Algebra and its Applications 450 (2014) 202-216.

[5] T.L. Hayden, R. Reams, J. Wells, Methods for constructing distance matrices and the inverse eigenvalues problem, Linear Algebra and its Application 295 (1999) 97-112.

[6] A.M. Nazari, A. Nezami, Inverse eigenvalues problem of nonnegative matrices via unit lower triangular matrices, ELA, 2019, Accepted paper.

Solving linear systems over max-plus algebra through pseudo-inverse method

Fateme Olia[1,*], Sedighe Jamshidvand[1] and Amirhossein Amiraslani[2]

[1]Faculty of Mathematics, K. N. Toosi University of Technology, Tehran, Iran

[2]School of STEM, Department of Mathematics, Capilano University, North Vancouver, BC, Canada

Abstract

Nowadays, certain problems in automata theory, control theory, manufacturing systems and parallel processing systems are intimately linked to linear systems over max-plus algebra. The main purpose of this paper is to introduce a method based on the pseudo-inverse of a matrix for solving a linear system of equations over max-plus algebra. To this end, we present a necessary and sufficient condition for the system to have a maximal solution.

Keywords: Semiring, max-plus algebra, System of linear equations, Pseudo-inverse
Mathematics Subject Classification [2010]: 16Y60, 65F05, 15A06

1 Introduction

Solving systems of linear equations is an important aspect of linear algebra. We propose a systematic method to understand the behavior of linear systems over max $-$plus algebra. Systems of linear equations over semirings find applications in various areas of engineering, computer science, optimization theory, control theory, etc (see e.g. [1, 3]). Semirings are algebraic structures similar to rings, but subtraction and division can not necessarily be defined for them. The notion of a semiring was first introduced by Vandiver [5] in 1934. A semiring $(S, +, ., 0, 1)$ is an algebraic structure in which $(S, +)$ is a commutative monoid with an identity element 0 and $(S, .)$ is a monoid with an identity element 1, connected by ring-like distributivity. The additive identity 0 is multiplicatively absorbing, and $0 \neq 1$. For convenience, we mainly consider $S = (\mathbb{R} \cup \{-\infty\}, max, +, -\infty, 0)$, which is called max $-$plus algebra.

We intend to solve the system of linear equations $AX = b$, where $A = (a_{ij}) \in M_n(S)$, $b \in S^n$ and X is an unknown vector over S. To this end, we present a necessary and sufficient condition based on the "pseudo-inverse", A^-, of matrix A with determinant $det_\varepsilon(A) \in U(S)$ to solve the system, where $U(S)$ is the set of the unit elements of S. It is shown that the proposed method is not limited to square matrices, and can be extended to arbitrary matrices of size $m \times n$ as well. In such cases, we try to convert the non-square system to a square one of size $\min\{m, n\}$.

*Speaker. Email address: folya@mail.kntu.ac.ir

1.1 Definitions and preliminaries

Definition 1.1. (See [2]) A semiring $(S, +, ., 0, 1)$ is an algebraic system consisting of a nonempty set S with two binary operations, addition and multiplication, such that the following conditions hold:

1. $(S, +)$ is a commutative monoid with identity element 0;
2. $(S, \cdot)$ is a monoid with identity element 1;
3. Multiplication distributes over addition from either side, that is $a(b + c) = ab + ac$ and $(b + c)a = ba + ca$ for all $a, b, c \in S$;
4. The neutral element of S is an absorbing element, that is $a \cdot 0 = 0 = 0 \cdot a$ for all $a \in S$;
5. $1 \neq 0$.

A semiring is called commutative if $a \cdot b = b \cdot a$ for all $a, b \in S$.

Let S be the max $-$plus algebra. We denote the set of all $m \times n$ matrices over S by $M_{m\times n}(S)$. For any $A = (a_{ij}) \in M_{m\times n}(S)$, $B = (b_{ij}) \in M_{m\times n}(S)$, $C = (c_{ij}) \in M_{n\times l}(S)$ and $\lambda \in S$, we define the matrix operations as follows.

$$A + B = (\max(a_{ij}, b_{ij})),$$

$$AC = (\max_{k=1}^{n}(a_{ik} + c_{kj})),$$

and

$$\lambda A = (\lambda + a_{ij}).$$

For convenience, we can denote the scalar multiplication λA by $\lambda+A$. Moreover, max $-$plus algebra is a commutative semiring which implies $\lambda + A = A + \lambda$. It is easy to verify that $M_n(S) := M_{n\times n}(S)$ forms a semiring with respect to the matrix addition and the matrix multiplication.

The concept of the determinant of a matrix over a commutative semiring requires the definition of an ε-function. (See [4] for more details.)

Definition 1.2. Let $(S, +, ., 0, 1)$ be a commutative semiring. A bijection ε on S is called an ε-function of S, if $\varepsilon(\varepsilon(a)) = a$, $\varepsilon(a + b) = \varepsilon(a) + \varepsilon(b)$, and $\varepsilon(ab) = \varepsilon(a)b = a\varepsilon(b)$ for all $a, b \in S$. Consequently, $\varepsilon(a)\varepsilon(b) = ab$ and $\varepsilon(0) = 0$.
The identity mapping: $a \mapsto a$ is an ε-function of S.

Definition 1.3. Let $A \in M_n(S)$, S be the max $-$plus algebra and $\mathcal{S}_n$ be the symmetric group of degree $n \geq 2$. The ε-determinant of A, denoted by $det_\varepsilon(A)$, is defined by

$$det_\varepsilon(A) = \max_{\sigma\in\mathcal{S}_n} \varepsilon^{\tau(\sigma)}(a_{1\sigma(1)} + a_{2\sigma(2)} + \cdots + a_{n\sigma(n)}),$$

where $\tau(\sigma)$ is the number of the inversions of the permutation σ, and $\varepsilon^{(k)}$ is defined by $\varepsilon^{(0)}(a) = a$ and $\varepsilon^{(k)}(a) = \varepsilon^{(k-1)}(\varepsilon(a))$ for all positive integers k.
In particular, let ε be the identity function, since $\varepsilon^{(2)}(a) = a$, we then have:

$$det_\varepsilon(A) = \max_{\sigma\in\mathcal{S}_n}(a_{1\sigma(1)} + a_{2\sigma(2)} + \cdots + a_{n\sigma(n)}).$$

Definition 1.4. Let $A \in M_n(S)$ and ε be an ε-function of S. The ε-adjoint matrix A, denoted by $adj_\varepsilon(A)$, is defined as follows.

$$adj_\varepsilon(A) = (\varepsilon^{(i+j)}(det_\varepsilon(A(i|j)))_{n\times n})^T,$$

where $A(i|j)$ denotes the $(n-1)\times(n-1)$ submatrix of A obtained from A by removing the i-th row and the j-th column.

Theorem 1.5. *(See [4]) Let $A \in M_n(S)$. We have*

1. $Aadj_\varepsilon(A) = (det_\varepsilon(A_r(i \Rightarrow j)))_{n\times n}$,
2. $adj_\varepsilon(A)A = (det_\varepsilon(A_c(i \Rightarrow j)))_{n\times n}$,

where $A_r(i \Rightarrow j)$ ($A_c(i \Rightarrow j)$) denotes the matrix obtained from A by replacing the j-th row (column) of A by the i-th row (column) of A.

Definition 1.6. Let $A \in M_n(S)$ and $det_\varepsilon(A) \in U(S)$. The pseudo-inverse of A, denoted by A^-, is defined as $A^- = (a_{ij}^-)$ where $a_{ij}^- = (adj_\varepsilon(A))_{ij} - det_\varepsilon(A)$.

Corollary 1.7. *Let $A \in M_n(S)$. Then the elements of the multiplication matrix AA^- are*

$$(AA^-)_{ij} = det_\varepsilon(A_r(i \Rightarrow j)) - det_\varepsilon(A).$$

In particular, the diagonal entries of the matrix AA^- are 0. Furthermore, the entries of the matrix A^-A are defined analogusly.

Proof. Clearly, the diagonal entries of the matrix AA^- are:

$$\begin{aligned}(AA^-)_{ii} &= (Aadj_\varepsilon(A))_{ii} - det_\varepsilon(A)\\ &= det_\varepsilon(A_r(i \Rightarrow i)) - det_\varepsilon(A)\\ &= det_\varepsilon(A) - det_\varepsilon(A)\\ &= 0\end{aligned}$$

□

Consider the system of linear equations $AX = b$ where $A \in M_n(S)$, $b \in S^n$ and X is an unknown column vector of size n over S, whose $i-$th equation is

$$\max(a_{i1} + x_1, a_{i2} + x_2, \cdots, a_{in} + x_n) = b_i.$$

Definition 1.8. Let $A, B \in M_{m\times n}(S)$ such that $A = (a_{ij})$ and $B = (b_{ij})$. We say $A \leq B$ if and only if $a_{ij} \leq b_{ij}$ for every $i \in \underline{m}, j \in \underline{n}$.

Definition 1.9. A solution X^* of the system $AX = b$ is called maximal if $X \leq X^*$ for any solution X.

Definition 1.10. Let $b \in S^m$. Then b is called a regular vector if $b_i \neq -\infty$ for any $i \in \underline{m}$.

2 Main results

In this section, we present the pseudo-inverse method for solving a linear system of equations over max $-$plus algebra. The pseudo-inverse method determines the maximal solution of a linear system if solutions exist.

Theorem 2.1. *Let $A \in M_n(S)$ and $b \in S^n$ be a regular vector. Then the system $AX = b$ has the maximal solution $X^* = A^- b$ with $X^* = (x_i^*)_{i=1}^n$ if and only if $(AA^-)_{ij} \leq b_i - b_j$ for any $i, j \in \{1, \cdots, n\}$.*

Proof. Suppose that $(AA^-)_{ij} \leq b_i - b_j$ for any $i, j \in \{1, \cdots, n\}$. First, we show that the system $AX = b$ has the solution $X^* = A^- b$. Clearly, $AX^* = AA^- b$, so for any $i \in \{1, \cdots, n\}$:

$$\begin{aligned}(AX^*)_i = (AA^- b)_i &= \max_{j=1}^{n}((AA^-)_{ij} + b_j) \\ &= \max((AA^-)_{ii} + b_i, \max_{i \neq j}((AA^-)_{ij} + b_j)).\end{aligned}$$

Since for any $i, j \in \{1, \cdots, n\}$, $(AA^-)_{ij} + b_j \leq b_i$, and $(AA^-)_{ii} + b_i = b_i$, we have $(AX^*)_i = b_i$. As such, X^* is a solution of the system $AX = b$.
Now, we prove $X^* = A^- b$ is a maximal solution. Since, $AX^* = b$, then $A^- AX^* = X^*$. As such, the k-th equation of the system $A^- AX^* = X^*$ is

$$max((A^- A)_{k1} + x_1^*, \cdots, x_k^*, \cdots, (A^- A)_{kn} + x_n^*) = x_k^*,$$

that implies

$$(A^- A)_{kl} \leq x_k^* - x_l^* \ for\ any\ l \neq k. \tag{1}$$

Now, suppose that $Y = (y_i)_{i=1}^n$ is another solution of the system $AX = b$. This means $AY = b$, and $(A^- A)Y = X^*$. Without loss of generality, we can assume there exists only $j \in \{1, \cdots, n\}$ such that $y_j \neq x_j^*$, i.e., $y_i = x_i^*$ for any $i \neq j$. The j-th equation of the $A^- AY = X^*$ is

$$max((A^- A)_{j1} + x_1^*, \cdots, (A^- A)_{jj} + y_j, \cdots, (A^- A)_{jn} + x_n^*) = x_j^*.$$

This means $(A^- A)_{jj} + y_j \leq x_j^*$ which implies $y_j < x_j^*$. Moreover, if all inequalities (1) for $k = j$ are proper, then

$$max((A^- A)_{j1} + x_1^*, \cdots, y_j, \cdots, (A^- A)_{jn} + x_n^*) < x_j^*.$$

Hence, Y is not the solution of the system $AX = b$. That leads to a contradiction.
This happens if all inequalities in (1) are proper, so we can conclude that X^* is a unique solution of the system $AX = b$. Otherwise, if some of the inequalities are not proper, i.e., $(A^- A)_{jl} = x_j^* - x_l^*$ for some $l \neq j$, then Y is a solution of the system $AX = b$ such that $Y \leq X^*$. Consequently, X^* is a maximal solution.
Conversely, suppose that $X^* = A^- b$ is a maximal solution of the system $AX = b$. Then $AA^- b = b$. That implies $(AA^-)_{ij} \leq b_i - b_j$ for any $i, j \in \underline{n}$. □

In the following example, we show that $(AA^-)_{ij} \leq b_i - b_j$ is a sufficient condition for the system $AX = b$ to have the maximal solution $X^* = A^- b$.

Example 2.2. Let $A \in M_4(S)$. Consider the following system $AX = b$:

$$\begin{bmatrix} 1 & -6 & 2 & -5 \\ 4 & 5 & 1 & -2 \\ 7 & -1 & 3 & 0 \\ -2 & -9 & -5 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 2 \\ 7 \\ 3 \\ -4 \end{bmatrix},$$

where $det_\varepsilon(A) = 14$. Due to Theorem 2.1, we must check the condition $(AA^-)_{ij} \leq b_i - b_j$ for any $i, j \in \{1, \cdots, 4\}$ where $(AA^-)_{ij} = (Aadj_\varepsilon(A))_{ij} - det_\varepsilon(A) = det_\varepsilon(A_r(i \Rightarrow j)) - det_\varepsilon(A)$ (see Theorem 1.5). As such, AA^- is

$$AA^- = \begin{bmatrix} 0 & -11 & -6 & -5 \\ -1 & 0 & -3 & -2 \\ 1 & -6 & 0 & 0 \\ -7 & -14 & -9 & 0 \end{bmatrix}.$$

Indeed, it is easier to check $(AA^-)_{ij} \leq b_i - b_j \leq -(AA^-)_{ji}$ for any $1 \leq i \leq j \leq 4$.
Since these inequalities hold, for instance $(AA^-)_{12} \leq 2 - 7 \leq -(AA^-)_{21}$, the system $AX = b$ has the maximal solution $X^* = A^-b$:

$$X^* = \begin{bmatrix} -6 & -13 & -7 & -7 \\ -6 & -5 & -8 & -7 \\ -2 & -13 & -8 & -7 \\ -7 & -14 & -9 & 0 \end{bmatrix} \begin{bmatrix} 2 \\ 7 \\ 3 \\ -4 \end{bmatrix} = \begin{bmatrix} -4 \\ 2 \\ 0 \\ -4 \end{bmatrix}.$$

2.0.1 Extension of the method to non-square linear systems

We are interested in studying the solution of a non-square linear system of equations as well. Let $A \in M_{m\times n}(S)$ with $m \neq n$, and $b \in S^m$ be a regular vector. For solving the non-square system $AX = b$ by Theorem 2.1, we must consider a square linear system of order $min\{m, n\}$ corresponding to it. Since $m \neq n$, we have the following two cases:

1. If $m < n$, then we consider the square linear system of order m corresponding to the system $AX = b$. Let $X = A^TY$ where Y is an unknown vector of size m. Then the square linear system $AA^TY = b$ is obtained from replacing X in $AX = b$. Suppose that the conditions of Theorem 2.1 hold for the system $AA^TY = b$, so the system $AA^TY = b$ has the maximal solution $Y^* = (AA^T)^-b$. If so, the system $AX = b$ has (at least) a solution in the form of $X = A^TY^* = A^T(AA^T)^-b$, which is not necessarily maximal.

2. If $n < m$, then we consider the square linear system of size n corresponding to the system $AX = b$. Clearly, we have the square linear system $A^TAX = A^Tb$ of size n. Assume that the conditions of Theorem 2.1 hold for the system $A^TAX = A^Tb$. If so, it has the maximal solution $X^* = (A^TA)^-A^Tb$. Note further that $X^* = (A^TA)^-A^Tb$ is not necessarily the solution of the system $AX = b$ unless b is an eigenvector of $A(A^TA)^-A^T$ corresponding to the eigenvalue 0, i.e.; $AX^* = A(A^TA)^-A^Tb = b$.

Example 2.3. Let $A \in M_{4\times5}(S)$. Consider the following system $AX = b$:

$$\begin{bmatrix} -4 & 7 & 12 & -3 & 0 \\ 3 & 2 & 8 & 3 & -1 \\ -9 & 1 & 6 & 0 & 2 \\ 2 & 8 & -5 & 1 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 14 \\ 10 \\ 8 \\ 11 \end{bmatrix}.$$

Due to the extension method, the non-square system $AX = b$ can be converted into the following square system $AA^TY = b$, cosidering $X = A^TY$:

$$\begin{bmatrix} 24 & 20 & 18 & 15 \\ 20 & 16 & 14 & 10 \\ 18 & 14 & 12 & 9 \\ 15 & 10 & 9 & 16 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 14 \\ 10 \\ 8 \\ 11 \end{bmatrix}.$$

The conditions of Theorem 2.1 hold for the system $AA^TY = b$, that is $((AA^T)(AA^T)^-)_{ij} \leq b_i - b_j$ for any $i, j \in \{1, \cdots, 4\}$, where $(AA^T)(AA^T)^-$ is the following matrix:

$$\begin{bmatrix} 0 & 4 & 6 & -1 \\ -4 & 0 & 2 & -5 \\ -6 & -2 & 0 & -7 \\ -9 & -5 & -3 & 0 \end{bmatrix}.$$

As such, the system $AA^TY = b$ has the maximal solution $Y^* = (AA^T)^-b$:

$$Y^* = \begin{bmatrix} -24 & -20 & -18 & -25 \\ -20 & -16 & -14 & -21 \\ -18 & -14 & -12 & -19 \\ -25 & -21 & -19 & -16 \end{bmatrix} \begin{bmatrix} 14 \\ 10 \\ 8 \\ 11 \end{bmatrix} = \begin{bmatrix} -10 \\ -6 \\ -4 \\ -5 \end{bmatrix}.$$

Hence, $X = A^TY^*$ is a solution of the non-square system $AX = b$:

$$X = \begin{bmatrix} -3 \\ 3 \\ 2 \\ -3 \\ -2 \end{bmatrix},$$

which is not necessarily maximal solution.

3 Conclusion

In this paper, we presented necessary and sufficient conditions for the linear systems of equations to have a maximal solution using the pseduo-inverse of system matrices. We also extended the idea to nonsquare systems.

References

[1] F. Baccelli, G. Cohen, G. J. Olsder and J. P. Quadrat, *Synchronization and linearity: an algebra for discrete event systems*, Wiley, New York, 1992.

[2] J. S. Golan, *Semirings and their Applications*, Kluwer Academic, Dordrecht, 1999.

[3] U. Hebish, H. J. Weinert, *Semirings: algebraic theory and applications in computer science*, World Scientific, Singapore, 1998.

[4] Y. J. Tan, Determinants of matrices over semirings, *Linear and Multilinear Algebra*, 62 (2014), No. 4, 498–517.

[5] H. S. Vandiver, Note on a simple type of algebra in which the cancellation law of addition does not hold, *Bulletin of the American Mathematical Society*, 40 (1934), No. 12, 914–920.

A note to preconditioners extracted from majorization matrix for multi-linear systems[1]

Fatemeh Panjeh Ali Beik and Mehdi Najafi-Kalyani*

Department of Mathematics, Vali-e-Asr University of Rafsanjan, PO Box 518, Rafsanjan, Iran

Abstract

In this work, we study the performance of a general class of preconditioners to accelerate the convergence speed of iterative schemes for solving multi-linear systems whose coefficient tensor is a strong $\mathcal{M}$-tensor. Some comparison results are presented between preconditioners extracted from the majorization matrix associated with the coefficient tensor. Numerical experiments are reported for a test example to illustrate the validity of theoretical discussions.

Keywords: Iterative method, Multi-linear system, Convergence, $\mathcal{M}$-tensor, Preconditioner

Mathematics Subject Classification [2010]: 65F10, 15A69

1 Introduction

Consider the following multi-linear system

$$\mathcal{A}x^{m-1} = b, \tag{1}$$

where $\mathcal{A} = (a_{i_1\dots i_m})$ is an order m dimension n real tensor, x and b are n dimensional real vectors. Here the n dimensional vector $\mathcal{A}x^{m-1}$ is given by [6]:

$$(\mathcal{A}x^{m-1})_i = \sum_{i_2,\dots,i_m=1}^{n} a_{ii_2\dots i_m} x_{i_2}\cdots x_{i_m}, \quad i = 1, 2, \dots, n.$$

In the sequel, we use $\mathbb{R}^{[m,n]}$ to denote the set of all order m dimension n real tensors for notational simplicity. The following definition for the product between a matrix and tensor is used throughout the paper which is a special case of the product between two tensors given in [7].

Definition 1.1. If $A \in \mathbb{R}^{[2,n]}$ and $\mathcal{B} = (b_{i_1\dots i_m}) \in \mathbb{R}^{[m,n]}$, then the tensor $\mathcal{C} = A\mathcal{B}$ belongs to $\mathbb{R}^{[m,n]}$ and its entries are given as follows:

$$c_{ji_2\dots j_m} = \sum_{j_2=1}^{n} a_{jj_2} b_{j_2 i_2\dots i_m}, \quad 1 \le j, i_\ell \le n,$$

for $j = 2, \dots, n$.

[1] Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

*Speaker. Email address: m.najafi.uk@gmail.com

In this work, we consider the case that $\mathcal{A}$ is a strong $\mathcal{M}$-tensor. In order to recall the definition of an $\mathcal{M}$-tensor, we need the following definition of tensor eigenvalues and eigenvectors; see [6].

Definition 1.2. Let $\mathcal{A} \in \mathbb{R}^{[m,n]}$. A pair $(\lambda, x) \in \mathbb{C} \times (\mathbb{C}^n \backslash \{0\})$ is called an eigenpair of $\mathcal{A}$ if they satisfy the equation

$$\mathcal{A}x^{m-1} = \lambda x^{[m-1]},$$

where $x^{[m-1]} = (x_1^{m-1}, \ldots, x_n^{m-1})^T$. The eigenpair (λ, x) is called and H-eigenpair if both λ and x are real.

The spectral radius of $\mathcal{A}$ is defined by $\rho(\mathcal{A}) = \max\{|\lambda| \mid \lambda \in \sigma(\mathcal{A})\}$ in which $\sigma(\mathcal{A})$ stands for the set of eigenvalues of $\mathcal{A}$. In the sequel, the unit tensor in $\mathbb{R}^{[m,n]}$ is denoted by $\mathcal{I}_m$ where $\mathcal{I}_m = (\delta_{i_1 \ldots i_m})$ such that

$$\delta_{i_1 \ldots i_m} = \begin{cases} 1, & i_1 = \cdots = i_m \\ 0, & \text{otherwise} \end{cases}.$$

Definition 1.3. Let $\mathcal{A} \in \mathbb{R}^{[m,n]}$. The tensor $\mathcal{A}$ is called a $\mathcal{Z}$-tensor if its off-diagonal entries are non-positive. If there exists a nonnegative tensor $\mathcal{B}$ and a positive real number $\eta \geq \rho(\mathcal{B})$ such that

$$\mathcal{A} = \eta \mathcal{I}_m - \mathcal{B},$$

then $\mathcal{A}$ is an $\mathcal{M}$-tensor. If $\eta > \rho(\mathcal{B})$, then $\mathcal{A}$ is called a strong $\mathcal{M}$-tensor.

It is known that if $\mathcal{A}$ is a strong $\mathcal{M}$-tensor then for every positive vector b the multilinear system $\mathcal{A}x^{m-1} = b$ has a unique positive solution [4, Lemma 4.1].

For $\mathcal{A} \in \mathbb{R}^{[m,n]}$, the majorization matrix $M(\mathcal{A})$ of $\mathcal{A}$ is the $n \times n$ matrix with the entries $M(\mathcal{A})_{ij} = a_{ij\ldots j}$ for $i, j = 1, 2, \ldots, n$.

For a general $\mathcal{M}$-tensor, the following lemma is proved by Liu et al. [4, Lemma 3.6].

Lemma 1.4. *If $\mathcal{A}$ is a strong $\mathcal{M}$-tensor, then $M(\mathcal{A})$ is a nonsingular M-matrix.*

Definition 1.5. Let $\mathcal{A} \in \mathbb{R}^{[m,n]}$. If $M(\mathcal{A})$ is a nonsingular matrix and $\mathcal{A} = M(\mathcal{A})\mathcal{I}_m$, the matrix $M(\mathcal{A})^{-1}$ is called the order 2 left-inverse of $\mathcal{A}$.

In [4], Liu et al. defined the concepts of left-invertibility of a tensor and tensor splitting. We also utilize the same definitions during this work given as follows:

Definition 1.6. Let $\mathcal{A} \in \mathbb{R}^{[m,n]}$. If $\mathcal{A}$ has an order 2 left-inverse, $\mathcal{A}$ is called a left-invertible tensor or a left-nonsingular tensor.

The decomposition $\mathcal{A} = \mathcal{E} - \mathcal{F}$ is called tensor splitting if $\mathcal{E}$ is left-nonsingular. The splitting $\mathcal{A} = \mathcal{E} - \mathcal{F}$ is said to be a regular splitting of $\mathcal{A}$ if $M(\mathcal{E})^{-1} \geq 0$ and $\mathcal{F} \geq 0$; a weak regular splitting if $M(\mathcal{E})^{-1} \geq 0$ and $M(\mathcal{E})^{-1}\mathcal{F} \geq 0$; a convergent splitting if $\rho(M(\mathcal{E})^{-1}\mathcal{F}) < 1$.

A generic tensor splitting iterative scheme is given by

$$x_k = [M(\mathcal{E})^{-1}\mathcal{F}x_{k-1}^{m-1} + M(\mathcal{E})^{-1}b]^{[\frac{1}{m-1}]}, \qquad k = 1, 2, \ldots, \tag{2}$$

where x_0 is given. The tensor $M(\mathcal{E})^{-1}\mathcal{F}$ is called the iteration tensor of iterative scheme (2). Liu et al. [4] showed that $\rho(M(\mathcal{E})^{-1}\mathcal{F})$ can be seen as an approximate convergence rate of (2).

Throughout this paper, we assume that each diagonal entry of the tensor $\mathcal{A}$ in (1) is equal to one. Also, we consider the decomposition $\mathcal{A} = \mathcal{I}_m - \mathcal{L} - \mathcal{F}$ where $\mathcal{L} = L\mathcal{I}_m$ in which $-L$ is the strictly lower triangular part of $M(\mathcal{A})$.

2 A class of preconditioners

In order to accelerate the asymptotic convergence rate of (2), one can use preconditioners. In fact, instead of (1), we solve the following preconditioned multi-linear

$$P\mathcal{A}x^{m-1} = Pb,$$

for a given preconditioner $P \in \mathbb{R}^{n\times n}$.

More recently, the performance of the preconsitioner $P_{\max} = I + S_{\max}$ was studied in [1] in which,

$$S_{\max} = (s^m_{ik_i}) = \begin{cases} -a_{ik_i\dots k_i}, & i = 1,\dots,n-1, k_i > i \\ 0, & \text{otherwise} \end{cases}$$

where $k_i = \min\{j | \max_j |a_{ij\dots j}|, i < n, j > i\}$.

In this paper, we consider a class of preconditioners in the form

$$\tilde{P} = I + \tilde{S}, \tag{3}$$

where

$$\tilde{S} = (\tilde{S}_{ij}) = \begin{cases} -\alpha_{ij}a_{ij\dots j}, & i,j = 1,\dots,n, \ (i \neq j) \\ 0, & i = j, \end{cases}$$

here the constants $\alpha_{ij} \in [0,1]$ are given for $i,j = 1,2,\dots,n$. Evidently, preconditioner $\tilde{P}$ reduces to $P_{\max}$ for proper choices of α_{ij} $(1 \leq i,j \leq n)$.

For a given tensor $\mathcal{A} \in \mathbb{R}^{[m,n]}$, the following comparison result between two weak regular splittings $\mathcal{A} = \mathcal{E}_1 - \mathcal{F}_1 = \mathcal{E}_2 - \mathcal{F}_2$ is proved in [2, Lemma 5.3].

Lemma 2.1. *Let $\mathcal{A} \in \mathbb{R}^{[m,n]}$ be a strong $\mathcal{M}$-tensor and $\mathcal{A} = \mathcal{E}_1 - \mathcal{F}_1 = \mathcal{E}_2 - \mathcal{F}_2$ be two weak regular splittings with $M(\mathcal{E}_2)^{-1} \geq M(\mathcal{E}_1)^{-1}$. If the Perron vector x of $M(\mathcal{E}_2)^{-1}\mathcal{F}_2$ satisfies $\mathcal{A}x^{m-1} \geq 0$ then $\rho(M(\mathcal{E}_2)^{-1}\mathcal{F}_2) \leq \rho(M(\mathcal{E}_1)^{-1}\mathcal{F}_1)$.*

We can show that the above result remains valid, if the assumption $\mathcal{A}x^{m-1} \geq 0$ hold for the Perron vector x of $M(\mathcal{E}_1)^{-1}\mathcal{F}_1$. We state this fact as the following lemma which its proof is omitted.

Lemma 2.2. *Let $\mathcal{A} \in \mathbb{R}^{[m,n]}$ be a strong $\mathcal{M}$-tensor and $\mathcal{A} = \mathcal{E}_1 - \mathcal{F}_1 = \mathcal{E}_2 - \mathcal{F}_2$ be two weak regular splittings with $M(\mathcal{E}_2)^{-1} \geq M(\mathcal{E}_1)^{-1}$. If the Perron vector x of $M(\mathcal{E}_1)^{-1}\mathcal{F}_1$ satisfies $\mathcal{A}x^{m-1} \geq 0$ then $\rho(M(\mathcal{E}_2)^{-1}\mathcal{F}_2) \leq \rho(M(\mathcal{E}_1)^{-1}\mathcal{F}_1)$.*

Now we present the following lemma without proof. Then a theorem is proved which reveals that except the assumption of being strong $\mathcal{M}$-tensor for $\mathcal{A}$, other hypotheses in [1, Theorems 1 and 2] are not required to be assumed. Basically, they could be concluded from the fact that $\mathcal{A}$ is a strong $\mathcal{M}$-tensor.

Lemma 2.3. *Let $\mathcal{A} \in \mathbb{R}^{[m,n]}$ be a $\mathcal{Z}$-tensor. Assume that $\mathcal{A} = \mathcal{I}_m - \mathcal{L} - \mathcal{F}$ where $\mathcal{L} = L\mathcal{I}_m$ in which $-L$ is the strictly lower part of $M(\mathcal{A})$. The tensor $\mathcal{A}$ is a strong $\mathcal{M}$-tensor if and only if $\tilde{\mathcal{A}} = \tilde{P}\mathcal{A}$ is a strong $\mathcal{M}$-tensor.*

Theorem 2.4. *Let $\mathcal{A} \in \mathbb{R}^{[m,n]}$ be a strong $\mathcal{M}$-tensor. If $\tilde{\mathcal{A}} = \tilde{\mathcal{E}} - \tilde{\mathcal{F}}$ such that $\tilde{\mathcal{E}} = \mathcal{I}_m - \tilde{\mathcal{D}} - \mathcal{L} - \tilde{\mathcal{L}}$ where $\tilde{\mathcal{D}} = \tilde{D}\mathcal{I}_m$ and $\tilde{\mathcal{L}} = \tilde{L}\mathcal{I}_m$ in which $\tilde{D}$ and $\tilde{L}$ are the diagonal and strictly lower triangular parts of $M(\tilde{S}\mathcal{L})$. Then $M(\tilde{\mathcal{E}})$ is an M-matrix.*

Proof. It can be seen that $M(\tilde{\mathcal{E}}) = I - \tilde{D} - L - \tilde{L}$. Evidently,

$$M(\tilde{S}\mathcal{L})_{ij} = (\tilde{S}M(\mathcal{L}))_{ij} \qquad 1 \leq i, j \leq n,$$

which results in

$$M(\tilde{S}\mathcal{L})_{ij} = \sum_{j_2=1}^{j-1} \alpha_{ij_2} a_{ij_2\ldots j_2} a_{j_2 j\ldots j}. \tag{4}$$

Therefore the diagonal part of $M(\tilde{\mathcal{E}})$ is given by

$$M(\tilde{\mathcal{E}})_{ii} = 1 - \sum_{j_2=1}^{i-1} \alpha_{ij_2} a_{ij_2\ldots j_2} a_{j_2 i\ldots i} \geq 1 - \sum_{\substack{j_2=1 \\ (j_2\neq i)}}^{n} \alpha_{ij_2} a_{ij_2\ldots j_2} a_{j_2 i\ldots i} = M(\tilde{\mathcal{A}})_{ii},$$

for $i = 1, 2, \ldots, n$. By Lemma 2.3, $\tilde{\mathcal{A}}$ is a strong $\mathcal{M}$-tensor which ensures that $M(\tilde{\mathcal{A}})$ is an M-matrix by Lemma 1.4. Consequently, the diagonal entries of $M(\tilde{\mathcal{E}})$ are positive. Also, since $0 \leq \alpha_{ij} \leq 1$, it is observed that

$$(L + \tilde{L})_{ij} = (\alpha_{ij} - 1)a_{ij\ldots j} + \sum_{\substack{j_2=1 \\ (j_2\neq j)}}^{i-1} \alpha_{ij_2} a_{ij_2\ldots j_2} a_{j_2 j\ldots j} \geq 0,$$

for $i > j$ reminding that $a_{ij\ldots j} \leq 0$ when $i \neq j$. It is not difficult to verify that $M(\tilde{\mathcal{E}}) = M_1 - N_1$ is a regular convergent splitting with $M_1 = I - \tilde{D}$ and $N_1 = L + \tilde{L} \geq 0$. This ensures that $M(\tilde{\mathcal{E}})$ is an M-matrix which completes the proof. □

Remark 2.5. Let $\mathcal{A} \in \mathbb{R}^{[m,n]}$ be a strong $\mathcal{M}$-tensor. If $\tilde{\mathcal{A}} = \tilde{\mathcal{E}} - \tilde{\mathcal{F}}$ such that $\tilde{\mathcal{E}} = \mathcal{I}_m - \tilde{\mathcal{D}} - \mathcal{L} - \tilde{\mathcal{L}}$ where $\tilde{\mathcal{D}} = \tilde{D}\mathcal{I}_m$ and $\tilde{\mathcal{L}} = \tilde{L}\mathcal{I}_m$ in which $\tilde{D}$ and $\tilde{L}$ are the diagonal and strictly lower triangular parts of $M(\tilde{S}\mathcal{L})$. From [4, Theorem 3.18], Lemma 2.3 and the above theorem, we deduce that if $M(\tilde{\mathcal{E}})^{-1}\tilde{\mathcal{F}} \geq 0$, then $\tilde{\mathcal{A}} = \tilde{\mathcal{E}} - \tilde{\mathcal{F}}$ is a weak regular convergent splitting.

Proposition 2.6. *Let $\tilde{P} = I + \tilde{S}$ and $\tilde{S}$ be defined such that the nonzero entries of $\tilde{S}$ and $S_{\max}$ are equal. Assume that the remaining nonzero elements of $\tilde{S}$ are defined as before such that $\tilde{S} \geq S_{\max}$. Then $M(\tilde{\mathcal{E}})^{-1} \geq M(\mathcal{E}_{\max})^{-1} \geq 0$ where $\tilde{\mathcal{E}}$ defined as in Theorem 2.4, $\mathcal{E}_{\max} = \mathcal{I}_m - \mathcal{D}_{\max} - \mathcal{L} - \mathcal{L}_{\max}$ in which $\mathcal{D}_{\max} = D_{\max}\mathcal{I}_m$ and $\mathcal{L}_{\max} = L_{\max}\mathcal{I}_m$ such that $D_{\max}$ and $L_{\max}$ are the diagonal and strictly lower triangular parts of $M(S_{\max}\mathcal{L})$.*

Proof. It can be observed that

$$M(\tilde{S}\mathcal{L}) \geq M(S_{\max}\mathcal{L}). \tag{5}$$

Let $M(\tilde{\mathcal{E}}) = (I - \tilde{D}) - (\tilde{L} + L)$ and $M(\mathcal{E}_{\max}) = (I - D_{\max}) - (L_{\max} + L)$ where $\tilde{D}$ ($D_{\max}$) and $\tilde{L}$ ($L_{\max}$) are respectively the diagonal and strictly lower part of $M(\tilde{S}\mathcal{L})$ ($M(S_{\max}\mathcal{L})$). Here $-L$ denotes the strictly lower part of $M(\mathcal{A})$. Evidently, $I - \tilde{D} \geq 0$ ($I - D_{\max} \geq 0$) which can be concluded from the fact that $M(\tilde{\mathcal{A}})$ ($M(\mathcal{A}_{\max})$) is an M-matrix. From (5), we have $\tilde{D} \geq D_{\max}$ and $\tilde{L} \geq L_{\max} \geq 0$. Hence it can be verified that

$$(I - \tilde{D})^{-1}(\tilde{L} + L) \geq (I - D_{\max})^{-1}(L_{\max} + L) \geq 0,$$

which implies that

$$M(\tilde{\mathcal{E}})^{-1} = (I - \tilde{D})^{-1} \sum_{l=0}^{n-1} (I - \tilde{D})^{-l}(\tilde{L} + L)^l$$

$$\geq (I - D_{\max})^{-1} \sum_{l=0}^{n-1} (I - D_{\max})^{-l} (L_{\max} + L)^l = M(\mathcal{E}_{\max})^{-1}.$$

Similar to the proof of Theorem 2.4, it can be observed that $M(\tilde{\mathcal{E}})^{-1} \geq 0$ and $M(\mathcal{E}_{\max})^{-1} \geq 0$ which completes the proof. □

Proposition 2.7. *Let $\mathcal{A} \in \mathbb{R}^{[m,n]}$ be a strong $\mathcal{M}$-tensor and the eigenvalues of $\tilde{S}$ are all real. Let (ρ, x) be the Perron eigenpair of $M(\tilde{\mathcal{E}})^{-1}\tilde{\mathcal{F}}$. If $\alpha_{ij} \in [0,1]$ and $\frac{\tilde{\rho}^2}{1-\tilde{\rho}^2} \leq \frac{\rho}{1-\rho}$, then $\mathcal{A}x^{m-1} \geq 0$ where $\tilde{\rho} = \rho(\tilde{S})$.*

Proof. The proof follows from similar strategy used in [2, Lemma 5.4]. □

Remark 2.8. In addition to the assumption of Proposition 2.6, if $\tilde{\mathcal{F}} - \tilde{S}\mathcal{I}_m \geq 0$ then $\tilde{\mathcal{A}} = \tilde{\mathcal{E}} - \tilde{\mathcal{F}}$ is a convergent regular splitting. As a result, the splitting $\mathcal{A} = (I+\tilde{S})^{-1}\tilde{\mathcal{E}} - (I+\tilde{S})^{-1}\tilde{\mathcal{F}}$ is weak regular. Let $\mathcal{A} = (I + S_{\max})^{-1}\mathcal{E}_{\max} - (I + S_{\max})^{-1}\mathcal{F}_{\max}$. It is known that for the Perron vector $x_{\max}$ of $M(\mathcal{E}_{\max})^{-1}\mathcal{F}_{\max}$, by Proposition 2.7, we have $\mathcal{A}x_{\max} \geq 0$. Using Lemma 2.2, for the weak regular splittings

$$\mathcal{A} = (I + \tilde{S})^{-1}\tilde{\mathcal{E}} - (I + \tilde{S})^{-1}\tilde{\mathcal{F}} = (I + S_{\max})^{-1}\mathcal{E}_{\max} - (I + S_{\max})^{-1}\mathcal{F}_{\max},$$

we deduce that $\rho(M(\tilde{\mathcal{E}})^{-1}\tilde{\mathcal{F}}) \leq \rho(M(\mathcal{E}_{\max})^{-1}\mathcal{F}_{\max}) < 1$.

3 A test example

Numerical results in this part were computed using MATLAB version 9.4 (R2018a) running on an Intel Core i5 CPU at 2.50 GHz with 8 GB of memory. We report total required number of iterations and consumed CPU-time (in seconds) under "Iter" and "CPU–times(s)", respectively. We set the maximum iteration number as 1000 and the stopping criterion is

$$\|\mathcal{A}x_k^{m-1} - b\|_2 \leq \varepsilon,$$

where x_k is the kth approximate solution, $\varepsilon = 10^{-10}$ and the initial vector is taken to be zero. The spectral radius of the nonnegative iteration tensors are computed by the power method given in [5].

Example 3.1. Let $\mathcal{A} \in \mathbb{R}^{[3,n]}$ and $b \in \mathbb{R}^n$ with

$$\begin{cases} a_{111} = a_{nnn} = 1, & \\ a_{iii} = 4, & i = 2,3,\ldots,n-1, \\ a_{i,i-1,i} = -1/2, & i = 2,3,\ldots,n-1, \\ a_{i,i-1,i-1} = -1/2, & i = 2,3,\ldots,n-1, \\ a_{i,i+1,i+1} = -1/2, & i = 2,3,\ldots,n-1, \\ a_{i,i-2,i-2} = -1/2 & i = 3,4,\ldots,n-2, \\ a_{i,i+2,i+2} = -1/2 & i = 3,4,\ldots,n-2, \\ a_{i,i-5,i-5} = -1/2 & i = 6,7,\ldots,n-5, \\ a_{i,i+5,i+5} = -1/2 & i = 6,7,\ldots,n-5, \end{cases} \quad \text{and} \quad \begin{cases} b_1 = c_0^2, \\ b_i = \frac{a}{(n-1)^2}, \quad i = 2,3,\ldots,n-1, \\ b_n = c_1^2, \end{cases}$$

where $c_0 = 1/2, c_1 = 1/3$ and $a = 2$. Since $a_{iii} \geq 1$ for $i = 1,2,\ldots,n$, we solve the multilinear system $D^{-1}\mathcal{A}x^2 = D^{-1}b$ by iterative method (2); here $D = \text{diag}(a_{111},\ldots,a_{nnn})$. The corresponding results are reported in Table 1. We set $\tilde{P} = I + \tilde{S}$ with $\tilde{S} = I - \text{triu}(M(\mathcal{A}))$. As seen, the obtained results show the validity of our discussions in Remark 2.8 and $\tilde{P} = I + \tilde{S}$ outperforms $P_{\max} = I + S_{\max}$. We comment that for $n = 300, 350$, the corresponding spectral radii of each (preconditioned) method are equal up to 4 digits.

Table 1: Example 3.1: Numerical results for applying iterative method (2).

n	Preconditioner P	$\mathcal{E}$ $(\mathcal{F}=\mathcal{E}-P\mathcal{A})$	$\rho(M(\mathcal{E})^{-1}\mathcal{F})$	CPU–times(s)	Iter
150	I	$\mathcal{I}_m-\mathcal{L}$	0.7921	0.3156	73
	$P_{\max}$	$\mathcal{I}_m-\mathcal{L}-\mathcal{D}_{\max}-\mathcal{L}_{\max}$	0.7477	0.2509	59
	$\tilde{P}$	$\mathcal{I}_m-\mathcal{L}-\tilde{\mathcal{D}}-\tilde{\mathcal{L}}$	0.6684	0.1723	42
300	I	$\mathcal{I}_m-\mathcal{L}$	0.7923	2.8292	72
	$P_{\max}$	$\mathcal{I}_m-\mathcal{L}-\mathcal{D}_{\max}-\mathcal{L}_{\max}$	0.7482	2.2002	58
	$\tilde{P}$	$\mathcal{I}_m-\mathcal{L}-\tilde{\mathcal{D}}-\tilde{\mathcal{L}}$	0.6693	1.4740	41
350	I	$\mathcal{I}_m-\mathcal{L}$	0.7923	4.2838	72
	$P_{\max}$	$\mathcal{I}_m-\mathcal{L}-\mathcal{D}_{\max}-\mathcal{L}_{\max}$	0.7482	3.9541	58
	$\tilde{P}$	$\mathcal{I}_m-\mathcal{L}-\tilde{\mathcal{D}}-\tilde{\mathcal{L}}$	0.6693	2.7067	41

4 Conclusions

In this work, we proposed a general class of preconditioners which incorporate some of the recently examined preconditioners in the literature for solving multi-linear systems. It should be commented that the idea of constructing such a kind of preconditioners is taken from [3]. Numerical experiments were reported for a test problem to numerically confirm the validity of presented results. The performance of preconditioners in conjunction with Krylov subspace methods for solving the mentioned multi-linear systems is a project to be currently undertaken.

Acknowledgment

The above presented results are some parts of a project which is currently in progress.

References

[1] L. B. Cui, M. H. Li and Y. Song, Preconditioned tensor splitting iteration method for solving multi-linear systems, *Appl. Math. Lett.*, 96 (2019), 89–94.

[2] W. Li, D. Liu and S.W. Vong, Comparison results for splitting iterations for solving multi-linear systems, *Appl. Numer. Math.*, 134 (2018), 105–124.

[3] D. K. Salkuyeh, M. Hasani and F. P. A. Beik, On the preconditioned AOR iterative method for Z-matrices, *J. Comput. Appl. Math.*, 36 (2017), 877–883.

[4] D. Liu, W. Li and S. W. Vong, The tensor splitting with application to solve multi-linear systems, *J. Comput. Appl. Math.*, 330 (2018), 75–94.

[5] M. Ng, L. Qi and G. Zhou, Finding the largest eigenvalue of a nonnegative tensor, *SIAM J. Matrix Anal. Appl.*, 31 (2010), 1090–1099.

[6] L. Qi, Eigenvalues of a real supersymmetric tensor, *J. Symbolic Comput.*, 40 (2005), 1302–1324.

[7] J. Shao, A general product of tensors with applications, *Linear Algebra Appl.*, 439 (2013), 2350–2366.

Matrix forms of the Black-Scholes equation with boundary conditions[1]

Mahdi Razavi*, Mohammad Mehdi Hosseini and Abbas Salemi

Department of Applied Mathematics, Faculty of Mathematics and Computer, Shahid Bahonar University of Kerman, Kerman, Iran

Abstract

In mathematical finance, the Black-Scholes equation is a backward parabolic partial differential equation finding the price evolution of a European call/put options. There are numerous numerical and analytical methods to solve Black-Scholes equation, but most of these methods have computational complexity and so far these methods could not present a general form to solve the Black-Scholes equation. In this paper by using the spectral method and special linear operators, we obtain matrix form of the Black-Scholes equation and matrix form of boundary conditions. Moreover, by using these matrix forms, we present a linear system of equations which approximate the solutions of the Black-Scholes equation.

Keywords: Black-Scholes equation, Finance mathematic, Spectral methods, Chebyshev polynomials

Mathematics Subject Classification [2010]: 91G60, 65M70, 65F05, 15A06

1 Introduction

Black-Scholes equation is one of the most important differential equations with second-order partial derivation in financial mathematics for estimating option price [1–4]. According the two variables of time t and price s this equation is defined as:

$$V_t + \frac{1}{2}\sigma^2 s^2 V_{ss} + rsV_s - rV = 0 \tag{1}$$

where $V(s,t)$ is the price of the option, r is the risk-free interest rate and σ is the volatility of the stock. Given that the Black-Scholes equation is a backward parabolic equation for a unique solution, we must specify final and boundary conditions in as follows:
Typically we must pose two conditions in s, as we have a V_{ss} term in the equation, but only one in t, as we only have a V_t term in it. For example, we could specify that $V(s,t) = V_a(t)$ *on* $s = a$ and $V(s,t) = V_b(t)$ *on* $s = b$, where $V_a(t)$ and $V_a(t)$ are two given functions of t. Also due to the backward of the equation, we must also impose a final condition such as $V(s,T) = V_T(s)$, where $V_T(s)$ is a known function. One of the best boundary conditions induced for the Black-Scholes equation for call and put options is as follows:

[1]Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

*Speaker. Email address: mrazavi@math.uk.ac.ir

European call option:

$$c(0,t)=0 \quad 0\leqslant t\leqslant T, \quad c(M,t)=M-ke^{-r(T-t)} \quad 0\leqslant t\leqslant T, \quad c(s,T)=\max\{s-k,0\} \quad 0\leqslant s\leqslant M.$$

European put option:

$$p(0,t)=ke^{-rt} \quad 0\leqslant t\leqslant T, \quad p(M,t)=0 \quad 0\leqslant t\leqslant T, \quad p(s,T)=\max\{k-s,0\} \quad 0\leqslant s\leqslant M.$$

In this paper by using the spectral method and special linear operators, we obtain matrix forms of the Black-Scholes equation and boundary conditions. Moreover, by using these matrix forms, we present a linear system of equations which approximate the solutions of Black-Scholes equation.

2 Preliminaries

In this section, we state some useful definitions and well-known results. Also, in this section we assume that the function of approximation of the Black Scholes equation is $v=\sum_{i,j=0}^{m}\alpha_i\beta_jT_i(s)T_j(t)$.

Definition 2.1. The Chebyshev polynomial $T_n(x)$ of the first kind is a polynomial in x of degree n, defined by the relation

$$T_n(x)=cosn\theta \quad \text{where} \quad x=cos\theta.$$

Lemma 2.2. *If* $v(x)=\sum_{k=0}^{m}\alpha_kT_k(x)$, *then* $v'(x)=\sum_{k=0}^{m}\alpha_kT_k^{'}(x)=\sum_{k=0}^{m}\hat{\alpha}_kT_k(x)$ *and* $v^{''}(x)=\sum_{k=0}^{m}\alpha_kT_k^{''}(x)=\sum_{k=0}^{m}\hat{\hat{\alpha}}_kT_k(x)$. *Assume that* $\alpha:=[\alpha_0\ \alpha_1\ \cdots\ \alpha_m]^t$, $\hat{\alpha}:=[\hat{\alpha}_0\ \hat{\alpha}_1\ \cdots\ \hat{\alpha}_m]^t$ *and* $\hat{\hat{\alpha}}:=[\hat{\hat{\alpha}}_0\ \hat{\hat{\alpha}}_1\ \cdots\ \hat{\hat{\alpha}}_m]^t$. *Therefore, there exists derivative matrix* D *such that* $\alpha^{(1)}=D\alpha$ *and* $\alpha^{(2)}=D^2\alpha$, *where*

$$D_{ij}=\begin{cases} j-1 & i+j \text{ is odd}, \quad j>i=1,\\ 2j-2 & i+j \text{ is odd}, \quad j>i>1,\\ 0 & \text{other wise,}\end{cases}$$

$$(D^2)_{ij}=\begin{cases} \frac{((j-1)^2-(i-1)^2)(j-1)}{2} & i+j \text{ is even}, \quad j>i=1,\\ ((j-1)^2-(i-1)^2)(j-1) & i+j \text{ is even}, \quad j>i>1,\\ 0 & \text{other wise.}\end{cases}$$

Lemma 2.3. *Let* $u_s=\sum_{i,j=0}^{m}\alpha_i\beta_jT_i^{'}(s)T_j(t)=\sum_{i,j=0}^{m}\hat{\alpha}_i\beta_jT_i(s)T_j(t)$, $u_{ss}=\sum_{i,j=0}^{m}\alpha_i\beta_jT_i^{''}(s)T_j(t)=\sum_{i,j=0}^{m}\hat{\hat{\alpha}}_i\beta_jT_i(s)T_j(t), u_t=\sum_{i,j=0}^{m}\alpha_i\beta_jT_i(s)T_j^{'}(t)=\sum_{i,j=0}^{m}\alpha_i\hat{\beta}_jT_i(s)T_j(t)$ *and* D, D^2, *be the matrices as in Lemma 2.2. Then*

$$a)[\hat{\alpha}\beta]=(D\otimes I))[\alpha\beta]. \quad b)[\alpha\hat{\beta}]=(I\otimes D)[\alpha\beta]. \quad c)[\hat{\hat{\alpha}}\beta]=(D^2\otimes I)[\alpha\beta].$$

where

$$[\alpha\beta]:=[\alpha_0\beta_0\ \ \alpha_0\beta_1\dots\alpha_0\beta_m\mid\dots\mid\alpha_m\beta_0\ \ \alpha_m\beta_1\dots\alpha_m\beta_m], \quad [\hat{\alpha}\beta]:=[\hat{\alpha}_0\beta_0\ \ \hat{\alpha}_0\beta_1\dots\hat{\alpha}_0\beta_m\mid\dots\mid\hat{\alpha}_m\beta_0\ \ \hat{\alpha}_m\beta_1\dots\hat{\alpha}_m\beta_m],$$

$$[\alpha\hat{\beta}]:=[\alpha_0\hat{\beta}_0\ \ \alpha_0\hat{\beta}_1\dots\alpha_0\hat{\beta}_m\mid\dots\mid\alpha_m\hat{\beta}_0\ \ \alpha_m\hat{\beta}_1\dots\alpha_m\hat{\beta}_m], \quad [\hat{\hat{\alpha}}\beta]:=[\hat{\hat{\alpha}}_0\beta_0\ \ \hat{\hat{\alpha}}_0\beta_1\dots\hat{\hat{\alpha}}_0\beta_m\mid\dots\mid\hat{\hat{\alpha}}_m\beta_0\ \ \hat{\hat{\alpha}}_m\beta_1\dots\hat{\hat{\alpha}}_m\beta_m].$$

3 Matrix forms of the Black-Scholes equation with boundary conditions

In this section, by using Lemmas 2.2 and 2.3, we present the matrix forms of the Black-Scholes equation and boundary conditions. Given that Chebyshev polynomials are defined in interval $[-1,1]$, we first need to transfer the variables of the Black-Scholes equation to interval $[-1,1]$. So, we do the following steps respectively

step 1: In regard with the fact that variable t is bounded in $[0,T]$, by changing variable $\tilde{t}=\frac{2}{T}t-1$, we convey variable t to $[-1,1]$.

step 2: In regard with the fact that variable s is bounded in $[0,M]$, by changing variable $\tilde{s}=\frac{2}{M}s-1$, we convey variable s to $[-1,1]$.

step 3: By exerting the changing variable of Steps (1) and (2) in the Black-Scholes equation by using chain derivation and relations:

$$V_t=\frac{2}{T}V_{\tilde{t}}\ ,\quad V_s=\frac{2}{M}V_{\tilde{s}}\ ,\quad V_{ss}=\frac{4}{M^2}V_{\tilde{s}\tilde{s}}\ ,$$

we have:

$$4V_{\tilde{t}}+\sigma^2T(\tilde{s}+1)^2V_{\tilde{s}\tilde{s}}+2rT(\tilde{s}+1)V_{\tilde{s}}-2rTV=0, \tag{2}$$

$$\begin{aligned} &v(\tilde{s},1)=\max\{\frac{M}{2}(\tilde{s}+1)-k,0\} \quad -1\leqslant\tilde{s}\leqslant 1, \quad v(-1,\tilde{t})=0 \quad -1\leqslant\tilde{t}\leqslant 1 \\ &v(1,\tilde{t})=M-ke^{-r(T-\frac{T}{2}(\tilde{t}+1))} \qquad -1\leqslant\tilde{t}\leqslant 1 \end{aligned} \tag{3}$$

Now, by using the spectral method, we present the matrix forms of the Black-Scholes equation and boundary conditions in Theorem 3.1 and 3.2.

Theorem 3.1. *If $v=\sum_{i,j=0}^{m}\alpha_i\beta_jT_i(\tilde{s})T_j(\tilde{t})$, then matrix implementation collocation spectral method of the Black-Scholes equation (2) is*

$$\hat{A}[\alpha\beta]=0, \tag{4}$$

where

$$\hat{A}=\Lambda_B\Pi_B,\quad \Pi_B:=4(I\otimes D)+\sigma^2T(\tilde{s}+1)^2(D^2\otimes I)+2rT(\tilde{s}+1)(D\otimes I)-2rT(I\otimes I),$$

and

$$\Lambda_B=[T_0(\tilde{s})T_0(\tilde{t})\ T_0(\tilde{s})T_1(\tilde{t})\dots\ T_0(\tilde{s})T_m(\tilde{t})\mid\dots\mid\ T_m(\tilde{s})T_0(\tilde{t})\ T_m(\tilde{s})T_1(\tilde{t})\ \dots T_m(\tilde{s})T_m(\tilde{t})].$$

Theorem 3.2. *If $v=\sum_{i,j=0}^{m}\alpha_i\beta_jT_i(\tilde{s})T_j(\tilde{t})$, then the matrix implementation collocation spectral method of the boundary conditions in Black-Scholes equation (3) is equal to*

$$\tilde{A}[\alpha\beta]=\tilde{G}, \tag{5}$$

where

$$\tilde{A}=\begin{bmatrix}(vec(\mathbb{T}^t(\tilde{s})\times e))^t\\ (vec(\tilde{e}^t\times\mathbb{T}(\tilde{t}))^t\\ (vec(e^t\times\mathbb{T}(\tilde{t}))^t\end{bmatrix}\in M_{3,(m+1)^2},\qquad \tilde{G}=\begin{bmatrix}\max\{\frac{M}{2}(\tilde{s}+1)-k,0\}\\ 0\\ M-ke^{-r(T-\frac{T}{2}(\tilde{t}+1))}\end{bmatrix},$$

$$\mathbb{T}(t):=[T_0(t)\ T_1(t)\ \cdots T_m(t)],\quad \tilde{e}:=[1\quad -1\quad \cdots(-1)^m],\quad e:=[1\quad 1\quad\cdots\quad 1].$$

Remark 3.3. If $p_i, i = 1, 2, ..., m+1$ are the roots of $T_{m+1}(x)$, by replacing p_i in relations (4) and (5), we obtain two linear systems. By choosing $(m+1)^2$ independent linear rows from these systems and solving the new linear system, we can find unknown coefficients $\alpha_i\beta_j$ for $i, j = 0, 1, 2, ..., m$.

In the following, we obtain the matrix forms for $m = 2$. Then $v = \sum_{i,j=0}^{2} \alpha_i\beta_j T_i(\tilde{s})T_j(\tilde{t})$. We consider $T_3(x) = 4x^3 - 3x$, with roots $p_1 = -\frac{\sqrt{3}}{2}$, $p_2 = 0$, $p_3 = \frac{\sqrt{3}}{2}$. Now, by using Theorem 3.1 and Theorem 3.2, we obtain the following:

$$\Pi_B = \begin{bmatrix} -2rT & 4 & 0 & 2r(\tilde{s}+1)T & 0 & 0 & 4\sigma^2(\tilde{s}+1)^2 & 0 & 0 \\ 0 & -2rT & 16 & 0 & 2r(\tilde{s}+1)T & 0 & 0 & 4\sigma^2(\tilde{s}+1)^2 & 0 \\ 0 & 0 & -2rT & 0 & 0 & 2r(\tilde{s}+1)T & 0 & 0 & 4\sigma^2(\tilde{s}+1)^2 \\ 0 & 0 & 0 & -2rT & 4 & 0 & 8r(\tilde{s}+1)T & 0 & 0 \\ 0 & 0 & 0 & 0 & -2rT & 16 & 0 & 8r(\tilde{s}+1)T & 0 \\ 0 & 0 & 0 & 0 & 0 & -2rT & 0 & 0 & 8r(\tilde{s}+1)T \\ 0 & 0 & 0 & 0 & 0 & 0 & -2rT & 4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -2rT & 16 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -2rT \end{bmatrix},$$

$$\Lambda_B = [T_0(\tilde{s})T_0(\tilde{t})\ T_0(\tilde{s})T_1(\tilde{t})\ T_0(\tilde{s})T_2(\tilde{t})\ T_1(\tilde{s})T_0(\tilde{t})\ T_1(\tilde{s})T_1(\tilde{t})\ T_1(\tilde{s})T_2(\tilde{t})\ T_2(\tilde{s})T_0(\tilde{t})\ T_2(\tilde{s})T_1(\tilde{t})\ T_2(\tilde{s})T_2(\tilde{t})]$$
$$= [1\ \ \tilde{t}\ \ 2\tilde{t}^2-1\ \ \tilde{s}\ \ \tilde{s}\tilde{t}\ \ \tilde{s}(2\tilde{t}^2-1)\ \ 2\tilde{s}^2-1\ \ \tilde{t}(2\tilde{s}^2-1)\ \ (2\tilde{s}^2-1)(2\tilde{t}^2-1)],$$
$$\hat{A} = \Lambda_B\Pi_B, \quad [\alpha\beta] = [\alpha_0\beta_0\ \alpha_0\beta_1\ \alpha_0\beta_2, \alpha_1\beta_0\ \alpha_1\beta_2\ \alpha_2\beta_0\ \alpha_2\beta_1\ \alpha_2\beta_2]^t,$$

$$\tilde{A} = \begin{bmatrix} 1 & 1 & 1 & \tilde{s} & \tilde{s} & \tilde{s} & 2\tilde{s}^2-1 & 2\tilde{s}^2-1 & 2\tilde{s}^2-1 \\ 1 & \tilde{t} & 2\tilde{t}^2-1 & -1 & -\tilde{t} & -(2\tilde{t}^2-1) & 1 & \tilde{t} & 2\tilde{t}^2-1 \\ 1 & \tilde{t} & 2\tilde{t}^2-1 & 1 & \tilde{t} & 2\tilde{t}^2-1 & 1 & \tilde{t} & 2\tilde{t}^2-1 \end{bmatrix}, \quad \tilde{G} = \begin{bmatrix} \max\{\frac{M}{2}(\tilde{s}+1)-k, 0\} \\ 0 \\ M - ke^{-r(T-\frac{T}{2}(\tilde{t}+1))} \end{bmatrix}.$$

4 Conclusion

Implementation of the spectral method on Black-Scholes equation is complicated. Here we are going to improve this problem by introducing matrix forms for Black-Scholes equation and boundary conditions, which is an important step in reducing the calculation and complexity of the implementation of the spectral method for this equation.

References

[1] P. Bradimarte, *Numerical Method in Finance and Economics*, John Wiley Press, 2006.

[2] B. Haidvogel, *The solution of Poisson's Equation by Expansion in Chebychev polynomials*, J. Comput. Phys, 1979.

[3] F. Chen, J. Shen and H. Yu, A New Spectral Element Method for Pricing European Options Under the Black-Scholes and Merton Jump Diffusion Models, *Journal of Scientific Computing*, 52 (2012), 499–518.

[4] A. Ishtiaq, A semi-analytic spectral method for elliptic partial differential equations, *Electronic Journal of Differential Equations*, 43 (2017), 1-11.

The triangle inequality with n-elements in quasi Banach spaces[1]

Asiyeh Rezaei* and Farzad Dadipour

Department of Mathematics, Gradute University of Advanced Technology, Kerman, Iran

Abstract

We investigate all n-tuples which satisfy the generalized triangle inequality of the second type in quasi Banach spaces. As applications, we get some new results associated with generalizations of the triangle inequality in quasi Banach spaces and we confirm some already known results in a new approach.

Keywords: Triangle inequality of the second type, Generalized triangle inequality, Quasi Banach space, Norm inequalities

Mathematics Subject Classification [2010]: 46A16, 47A30, 46B20

1 Introduction

The triangle inequality is considered to be one of the most fundamental inequalities in mathematics. There are many interesting generalizations, refinements and reverses of the triangle inequality in normed spaces, quasi normed spaces, inner product spaces, pre-Hilbert $C^\star$ moduals by some authors [3]. Some generalizations of the triangle inequality are profitable to study the geometrical structure of Banach spaces. Espacially, based on the triangle inequality of the second type

$$\|x+y\| \leq 2\left(\|x\|^2+\|y\|^2\right) \tag{1}$$

and its generalizations in normed spaces. Takahasi et al. [6] obtained some conditions for which the inequality

$$\frac{\|ax+by\|^q}{\lambda} \leq \frac{\|x\|^q}{\mu}+\frac{\|y\|^q}{\nu} \quad (\lambda=\mu a^2+\nu b^2,\ \lambda\mu\nu>0)$$

holds For $q \geq 1$. In [2] Dadipour et al. discussed the generalized triangle inequality of the second type and its reverse in normed spaces. Also Izumida et al. presented another approach to characterizations of the generalized triangle inequality by using ψ-direct sums of Banach spaces.

In this talk, we investigate all n-tuples which satisfy the generalized triangle inequality of the second type

$$\|x_1+\cdots+x_n\|^q \leq \frac{\|x_1\|^q}{\mu_1}+\cdots+\frac{\|x_n\|^q}{\mu_n}, \quad (\text{for all } x_1,\ldots,x_n \in X,\ q>1), \tag{2}$$

[1]Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

*Speaker. Email address: a.rezaei@student.kgut.ac.ir

where $(X, \|.\|)$ is a quasi Banach space. As applications, we get some new results associated with generalizations of the triangle inequality in quasi Banach spaces and we confirm some already known results due to Belbachir et al. [1] and Dadipour et al. [2] in a new approach.

In the remainder of this section we recall some basic concepts, preliminary results and symbols that are used throughout this note.
A quasi norm on a vector space X is a real valued function $\|\cdot\| : X \to \mathbb{R}$ with the following properties:

(i) $\|x\| \geq 0$, for all $x \in X$ and $\|x\| = 0$ if and only if $x = 0$,

(ii) $\|\lambda x\| = |\lambda|\|x\|$, for all $\lambda \in \mathbb{R}$ and all $x \in X$,

(iii) There is a constant $C \geq 1$ such that $\|x + y\| \leq C\,(\|x\| + \|y\|)$, for all $x, y \in X$.

The smallest possible C in (iii) is called the modulus of concavity of $\|\cdot\|$ and the pair $(X, \|\cdot\|)$ is called a quasi normed space. If it is possible to take $C = 1$ we obtain a norm. A quasi norm $\|\cdot\|$ is called a p-norm $(0 < p \leq 1)$ if it satisfies

$$\|x + y\|^p \leq \|x\|^p + \|y\|^p \quad (x, y \in X).$$

In this case, a quasi normed space is called a p-normed space.
There are many different equivalent metrics on a quasi normed space, one of these, is given by Aoki and Rolewicz. The Aoki-Rolewicz theorem [4] states that if $(X, \|.\|)$ is a quasi normed space with the modulus of concavity C, then there is $p \in (0, 1]$ such that the following

$$|||x||| := \inf\left\{\left(\sum_{i=1}^{n}\|x_i\|^p\right)^{\frac{1}{p}} : n > 0,\ x_1, \ldots, x_n \in X,\ x = \sum_{i=1}^{n} x_i\right\},$$

defines a p-norm equivalent to quasi norm $\|.\|$. Moreover $|||x||| \leq \|x\| \leq 2C\,|||\,x\,|||$ and $2^{\frac{1}{p}-1} \leq C$. So every quasi norm is equivalent to some p-norms $(0 < p \leq 1)$ and $d(x, y) := |||x - y|||^p$ defines a metric topology on X. A quasi normed space (p-normed space) is called a quasi Banach space (p-Banach space) if every Cauchy sequence converges.
The notion of q-norm is a specification of a quasi norm that Belbachir et al. [1] introduced it as follows:
A real valued function $\|\cdot\|$ on a vector space X is called a q-norm $(q \geq 1)$ if it satisfies (i), (ii) in the above and the following inequality

$$\|x + y\|^q \leq 2^{q-1}(\|x\|^q + \|y\|^q) \quad (x, y \in X). \tag{3}$$

Considering the inequality $\|x\|^q + \|y\|^q \leq (\|x\| + \|y\|)^q$, we deduce that every q-norm is a quasi norm with the modulus of concavity $C \leq 2^{\frac{q-1}{q}}$.

Let $(X, \|.\|)$ be a quasi Banach space and $q > 1$. By $F(q)$ we denote all n-tuples $(\mu_1, \ldots, \mu_n) \in \mathbb{R}^n$ with positive coordinates for which inequality (2) holds for all $x_1, \ldots, x_n \in X$. Inequality (2) is also called the characteristic inequality of $F(q)$. We should notice that there is no n-tuple $(\mu_1, \ldots, \mu_n) \in \mathbb{R}^n$ with some negative coordinates satisfying inequality (2) (To see this, assume that there exists $(\mu_1, \ldots, \mu_n) \in \mathbb{R}^n$ such that $\mu_j < 0$ for some $j = 1, \ldots, n$ and inequality (2) holds for all $x_1, \ldots, x_n \in X$. One can take $x_j \in X \setminus \{0\}$ and $x_i = 0$ $(i = 1, \ldots, n,\ i \neq j)$ and get a contradiction.). So our main aim is to investigate $F(q)$ for all $q > 1$.

2 Main results

We obtain some regions of $\mathbb{R}^n$ which are contained in $F(q)$ for all $q > 1$ with the most accurate as possible. So we can state the following theorem.

Theorem 2.1 ([5, Theorem 2]). *Let $(X, \|.\|)$ be a quasi Banach space with the modulus of cancavity C and $q > 1$. Then the following hold:*

(i) $F(q) \supseteq \left\{(\mu_1, \dots, \mu_n) : \mu_1, \dots, \mu_n > 0 \text{ and } \left(\sum_{i=1}^n \mu_i^{\frac{1}{q-1}}\right)^{q-1} \leq C^{-q(1+\log_2^{n-1})}\right\}$;
(*the case where* $n \neq 4, 6$);

(ii) $F(q) \supseteq \left\{(\mu_1, \dots, \mu_n) : \mu_1, \dots, \mu_n > 0 \text{ and } \left(\sum_{i=1}^n \mu_i^{\frac{1}{q-1}}\right)^{q-1} \leq C^{-\frac{nq}{2}}\right\}$;
(*the case where* $n = 4, 6$).

In the next, as a reverse inclusion of the last result, we can get a region of $\mathbb{R}^n$ which contains $F(q)$.

Proposition 2.2 ([5, Proposition 1]). *Let $(X, \|.\|)$ be a quasi Banach space and $q > 1$. Then the following inclusion holds:*

$$F(q) \subseteq \{(\mu_1, \dots, \mu_n) : \mu_1, \dots, \mu_n > 0, \sum_{i=1}^n \mu_i^{\frac{1}{q-1}} \leq 1\}.$$

The results in the following corollaries are derived from Theorem 2.1 and Proposition 2.2 as some special cases.
Taking $C = 1$ and by using Theorem 2.1 and Proposition 2.2, we have the following corollary which was proved by Dadipour et. al [2, Theorem 2.4(i)].

Corollary 2.3 ([2, Theorem 2.4(i)]). *Let $(X, \|.\|)$ be a normed space and $q > 1$. Then*

$$F(q) = \left\{(\mu_1, \dots, \mu_n) : \mu_1, \dots, \mu_n > 0, \sum_{i=1}^n \mu_i^{\frac{1}{q-1}} \leq 1\right\}.$$

Finally with connection to the notion of q-norms, we get the following result which was proved by Belbachir et. al [1, Proposition 2.1]

Corollary 2.4 ([1, Proposition 2.1]). *Every norm in a usual sense is a q-norm for all $q > 1$.*

3 Conclusion

In quasi Banach spaces, by using the well-known Holder inequality, some regions of $\mathbb{R}^n$ which are contained in the set of all n-tuples satisfying the generalized triangle inequality are obtained. The results provide a better understanding of the behaviors of some inequalities with the source of the triangle inequality in some vector spaces such as $\mathbb{R}^n, l^p, \dots$.

References

[1] H. Belbachir, M. Mirzavaziri, M.S. Moslehian, q-norms are really norms, *Aust. J. Math. Anal. Appl.*, 3 (2006) 1-3. Article 2.

[2] F. Dadipour, M.S. Moslehian, J.M. Rassias, S.E. Takahasi, Characterization of a generalized triangle inequality in normed spaces, *Nonlinear Anal.*, 75 (2012), No. 2, 735–741.

[3] N. Minculete, R. Pltnea, Improved estimates for the triangle inequality, J. Inequal. Appl. 17(1) (2017) 1-12.

[4] A. Pietsch, *History of Banach spaces and linear operators*, Springer, Birkhuser Publisher, 2007.

[5] A. Rezaei, F. Dadipour, On generalized triangle inequality of the second type in quasi normed spaces, submitted.

[6] S.-E. Takahasi, J.M. Rassias, S. Saitoh, Y. Takahashi, Refined generalizations of the triangle inequality on Banach space, *Math. Inequal. Appl.*, 13 (2010) 733-741.

دانشگاه شهید باهنر کرمان، ایران
دهمین سمینار جبر خطی و کاربردهای آن
۲۶ لغایت ۲۹ مرداد ۱۳۹۹
Shahid Bahonar University of Kerman, Iran
Seminar on Linear Algebra and its Applications
August 16 - 19 , 2020

Some results on the higher rank numerical ranges of $(A-\lambda I)^{(\alpha+1)}$[1]

Sharifeh Rezagholi*

Department of Basic Science, Payame Noor University, Tehran, Iran

Abstract

In this paper, for any $n \times n$ matrix A with index α, the rank$-k$ numerical range of the matrix polynomial $(A-\lambda I)^{(\alpha+1)}$ is investigated. Also, some of algebraic and geometrical properties of them, by focus on the nilpotent and Jordan matrices, are studied.

Keywords: Index numerical range, Index higher rank numerical range, Nilpotent matrices

Mathematics Subject Classification [2010]: 15A03, 15B36

1 Introduction

Let $M_{n,k}$ be the set of $n \times k$ matrices with complex entries and

$$L(\lambda) = A_m\lambda^m + A_{m-1}\lambda^{m-1} + \cdots + A_1\lambda + A_0 \tag{1}$$

be a matrix polynomial with $A_i \in M_{n,n}$ and $A_m \neq 0$. For a positive integer $k \geq 1$, the rank$-k$ numerical range of $L(\lambda)$ is defined as

$$\Lambda_k(L(\lambda)) = \{\lambda \in \mathbb{C} : Q^*L(\lambda)Q = 0_k \text{ for some } Q \in M_{n,k} \text{ with } Q^*Q = I_k\}.$$

If $\text{Ł}(\lambda) = \lambda I - A$, this set reduces to $\Lambda_k(A)$. When $k = 1$, the rank$-k$ numerical range is $W(\text{Ł}(\lambda))$, the classical numerical range of the polynomial $\text{Ł}(\lambda)$. By $\alpha = ind(A)$, the index of the matrix $A \in M_{n,n}$, we mean the size of the largest Jordan block corresponding to the zero eigenvalue of A. It is obvious that if A is nonsingular, then $ind(A) = 0$. Recently in [4], using the matrix polynomial $(A-\lambda I)^{\alpha+1}$ the index numerical range of the matrix A is defined and denoted by $IW(A) = W((A-\lambda I)^{\alpha+1}) = \{z \in \mathbb{C} : x^*(A - zI_n)^{\alpha+1}x = 0, \text{ for some } x \in \mathbb{C}^n \setminus \{0\}\}$. In the following proposition, we list some properties of the index numerical range useful in this paper; To see the proofs and more results see [4].

Proposition 1.1. *Let $A \in M_{n,n}$. Then*
(i) If A is nonsingular, then $IW(A) = W(A)$;
(ii) $IW(A)$ is compact and connected subset of $\mathbb{C}$;
(iii) $\sigma(A) \subseteq IW(A)$;
(iv) If $\beta, \gamma \in \mathbb{C}$, then $IW(\beta I_n + \gamma A) = \beta + \gamma IW(A)$.

In this paper, we introduce the notion of index rank$-k$ numerical range of matrices. In special case, we study some algebraic and geometrical properties of the Jordan matrix J_n($n \times n$ Jordan block with zero eigenvalue).

[1] Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

*Speaker. Email address: sh_rezagholi79@yahoo.com

2 Main results

Using the same way of definition of the index numerical range, we have the following definition.

Definition 2.1. Let $A \in M_{n,n}$ and $\alpha = ind(A)$. The index rank$-k$ numerical range of A is defined and denoted by

$$I\Lambda_k(A) = \{\lambda \in \mathbb{C} : Q^*(A - \lambda I)^{\alpha+1}Q = 0I_k \text{ for some } Q \in M_{n,k} \text{ with } Q^*Q = I_k\}.$$

It is obvious that $I\Lambda_1(A) = IW(A)$, and so it is a generalization of index numerical range. The following proposition lists some basic and useful properties of the rank$-k$ numerical range.

Proposition 2.2. *Let $A \in M_{n,n}$. Then the following statements are true*
(i) If A is a nonsingular matrix, then $I\Lambda_k(A) = \Lambda_k(A)$;
(ii) If $U \in M_n$ is a unitary matrix, then $I\Lambda_k(U^*AU) = I\Lambda_k(A)$;
(iii) $I\Lambda_k(A) \subseteq I\Lambda_{k-1}(A) \subseteq \cdots \subseteq I\Lambda_1(A)$;
(iv) If $\beta, \gamma \in \mathbb{C}$, then $I\Lambda_k(\beta I_n + \gamma A) = \beta + \gamma I\Lambda_k(A)$.

Proof. If A is nonsingular, then $ind(A) = 0$ and this shows that $I\Lambda_k(A) = \Lambda_k[A - \lambda I] = \Lambda_k(A)$, where $A - \lambda I$ is considered as a matrix polynomial. This shows (i); Let $\alpha = ind(A)$ and $X \in M_{n,k}$ be such that $X^*X = I_k$ and $X^*(A - \lambda I)^{\alpha+1}X = 0I_k$. If U is a unitary matrix, then the isometry matrix $U^*X \in M_{n,k}$ helps us to conclude (ii); (iii) is a coclusion of [1, Proposition 3]; To see (iv), suppose that $\gamma \neq 0$. $I\Lambda_k(\beta I_n + \gamma A) = \{\lambda \in \mathbb{C} : 0 \in \Lambda_k((\lambda - \beta)I_n - \gamma A)^{\alpha+1}\} = \{\lambda \in \mathbb{C} : 0 \in \Lambda_k(\gamma^{\alpha+1}(\frac{\lambda-\beta}{\gamma} - A)^{\alpha+1})\} = \{\lambda \in \mathbb{C} : 0 \in \gamma^{\alpha+1}\Lambda_k((\frac{\lambda-\beta}{\gamma} - A)^{\alpha+1})\}$. So, $\lambda \in I\Lambda_k(\beta I_n + \gamma A)$ if and only if $\frac{\lambda-\beta}{\gamma} \in I\Lambda_k(A)$. This shows that $I\Lambda_k(\beta I_n + \gamma A) = \beta + \gamma I\Lambda_k(A)$. For $\gamma = 0$, the equality holds obviously □

Since $(A - \lambda I)^{\alpha+1}$ is a monic matrix polynomial, using [1, Propositions 1 and 10] we have the following proposition.

Proposition 2.3. *Let $A \in M_{n,n}$. Then $I\Lambda_k(A)$ is a compact subset of $\mathbb{C}$.*

If $k = 1$, we have the following corollary which is proved in [4] by another way.

Corollary 2.4. *Let $A \in M_{n,n}$. Then $IW(A)$ is a compact subset of $\mathbb{C}$.*

For nilpotent matrices, if we use first k columns of I_n respectively to costruct an isometry matrix, one can see that zero is a member of index rank-k numerical range. In the next theorem we see that index rank-k numerical range of nilpotent matrices is connected.

Theorem 2.5. *Let $A \in M_{n,n}$ be a nilpotent matrix. Then $I\Lambda_k(A)$ is connected.*

Remark 2.6. Although we know that $IW(A)$ is connected(Proposition 2.2(ii)), one can use the above theorem in case $k = 1$, to give another proof for connectedness of index numerical range of nilpotent matrices.

To find the index higher numerical ranges of matrices, finding this set for J_n can be useful. The next theorem show the shape of the index higher numerical ranges of Jordan matrices;

Theorem 2.7. *Let J_n be the $n \times n$ Jordan matrix with zero eigenvalue. Then $I\Lambda_k(J_n)$ is a closed disk.*

Proof. Let $U = diag(1, e^{-i\theta}, e^{-2i\theta}, \ldots, e^{-(n-1)i\theta})$, where $\theta \in \mathbb{R}$. It's obvious that U is a unitary matrix with the property $U^*J_nU = e^{i\theta}J_n$. So, by Proposition 2.2(ii) and (iv), $I\Lambda_k(e^{i\theta}J_n) = e^{i\theta}I\Lambda_k(J_n)$, for all $\theta \in \mathbb{R}$. By Proposition 2.5, $I\Lambda_k(J_n)$ is connected and [4, Theorem 2] shows that it is compact. These shows that $I\Lambda_k(A)$ is a closed disk around the origin. The special case is derived by choosing $k = 1$ and using the fact that $0 \in \sigma(J_n) \subseteq IW(A)$. □

By setting $k = 1$ in the above theorem and using proposition 2.2(iii), we have the following corollary.

Corollary 2.8. *Let J_n be the $n \times n$ Jordan matrix with zero eigenvalue. Then $IW(A)$ is a closed disk around the origin.*

In [4, Theorem 7], we saw that $\{z \in \mathbb{C} : |z| \leq (n+1)/2\} \subseteq IW(J_n)$. Moreover, [4, Example 1] shows that $IW(J_2) = \{z \in \mathbb{C} : |z| \leq 3/2\}$, i.e., the radious of the mentioned disk in the above theorem is exactly $(2+1)/2$. The following example show that the mentioned disk radious may be bigger than $(n+1)/2$.

Example 2.9. Let $z \in IW(J_3)$. So, $z^4 - 4z^3x^*J_3x + 6z^2x^*J_3^2x = 0$. Let $x = (1/2, 1/\sqrt{2}, 1/2)^t$. Then $x^*J_3x = \sqrt{2}/2$ and $x^*J_3^2x = 1/4$. So, $z = \sqrt{2} + 1/\sqrt{2} \in IW(J_3)$, while $|z| > (3+1)/2$.

Corollary 2.10. *Let J_n be the $n \times n$ Jordan matrix with zero eigenvalue. Then $I\Lambda_k(J_n)$ is convex. In special case, $IW(J_n)$ is convex.*

In the following example, we see that if $A \neq J_n$, then $IW(A)$ may not be convex.

Example 2.11. ([4, Lemma2]) Let $A = diag(\lambda, 0)$ where $0 \neq \lambda \in \mathbb{R}$. Then

$$IW(A) = \{z : |z - (\lambda/2)| = \lambda/2\},$$

which is not convex.

3 Conclusion

The index rank$-k$ numerical range of matrices may not be connected. We can find the number of connected components of it for some special matrices exactly. The number of connected components for nilpotent matrices is one, i.e., the index rank$-k$ numerical range of nilpotent matrices are connected. Using this fact, we find that the rank$-k$ numerical range of Jordan matrices is a closed disk around the origin.

References

[1] A. Aretaki and J. Maroulas, The higher rank numerical range of matrix polynomials, *Central uropian Journal of Mathematics*, 11(2013) , 435-446.

[2] R. Horn and C. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, New York, 1991.

[3] C.K. Li and H. Nakazato, Some results on the q-numerical range, *Linear and Multilinear algebra*, 43(1998), 385-409.

[4] M. Safarzadeh and A. Salemi, DGMRES and index numerical range of matrices, *Journal of Computational and Applied Mathematics*, 335(2018) 349-360.

An operational wavelet approach for 2D Abel integral equation[1]

Sayed Amjad Samareh Hashemi[1,2*] and Mostafa Poursharifi[2]

[1]Department of Applied Mathematics, Faculty of Mathematics and Computer, Shahid Bahonar University of Kerman, Kerman, Iran

[2]Department of Basic Sciences, School of Mathematical Sciences, PO BOX 19395-3697, Payame Noor University, Tehran, Iran

Abstract

In this paper, an operational wavelet method is introduced for finding an approximate solution of a class of two-dimensional Volterra weakly integral equations (two-dimensional Abel integral equations of the second kind). The presented method is a spectral method based on Chelyshkov wavelets from operational matrices. Perspective numerical examples show the efficiency and applicability of the proposed method in smooth and nonsmooth cases.

Keywords: 2D Abel integral equation, Fractional integral, Chelyshkov wavelet, Operational matrix

Mathematics Subject Classification [2010]: 15A60, 46N40, 47N40

1 Introduction

Wavelet constitutes a family of functions constructed from dialation and translation of a single function called the mother wavelet. When the dialation parameter a and the translation parameter b vary continuously, we have the following family of continuous wavelets as [2].

$$\psi_{a,b}(t) = |a|^{-\frac{1}{2}}\psi(\frac{t-b}{a}), \quad a, b \in \mathbb{R},$$

where ψ is the mother wavelet.

Chelyshkov Wavelets (ChWs), $\psi_{n,m}(x) = \psi(k, n, m, x)$, are defined on the interval $[0, L)$ by [2]:

$$\psi_{n,m}(t) = \begin{cases} \sqrt{2^k(2m+1)}\, P_m(2^k \frac{t}{L} - n), & \frac{n}{2^k}L \le t < \frac{n+1}{2^k}L, \\ 0, & \text{otherwise.} \end{cases}$$

where $P_m(t)$ is the Chelyshkov polynomial, which is defined as follows:

$$P_m(t) := \rho_{m,M}(t) = \sum_{j=0}^{M-m} a_{j,m}\, t^{m+j}, \quad m = 0, 1, \ldots, M, \tag{1}$$

[1]Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

*Speaker. Email address: amjad.hashemi@gmail.com

where:

$$a_{j,m} = (-1)^j \binom{M-m}{j} \binom{M+m+j+1}{M-m}.$$

These polynomials are orthogonal over the interval $[0,1]$ with respect to the weight function $w(t) = 1$, i.e. :

$$\int_0^1 P_n(t)\, P_m(t)\, dt = \frac{\delta_{mn}}{m+n+1},$$

where δ_{mn} is the Kronecker delta. According to the definition (1) it is obvious that for a fixed integer M, the polynomials $P_m(t)\,, \quad m = 0, 1, \ldots, M$ are polynomials exactly of degree M.

The ChWs $\{\psi_{n,m}(t)\,|\,n = 0, 1, \ldots 2^k - 1,\ m = 0, 1, \ldots M\}$ forms an orthonormal basis for $L^2[0, L]$. By using the orthogonality of ChWs, any function $f(t) \in L^2[0, L]$ can be expanded in terms of ChWs as:

$$f(t) = \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} c_{n,m}\, \psi_{n,m}(t), \tag{2}$$

where $c_{n,m} = \langle f(t), \psi_{n,m} \rangle = \int_0^L f(t)\, \psi_{n,m}(t)\, dt$. If the infinite series in Eq. (2) is truncated, then it can be written as:

$$f(t) \simeq \sum_{n=0}^{2^k-1} \sum_{m=0}^{M} c_{n,m}\, \psi_{n,m}(t) = \mathbf{C}^T \mathbf{\Psi}(t),$$

where $\mathbf{C}$ and $\mathbf{\Psi}$ are $\hat{m} = 2^k(M+1)$-vectors, given by:

$$\begin{aligned} \mathbf{C}^T &= [c_{0,0}, c_{0,1}, \ldots, c_{0,M}, c_{1,0}, \ldots, c_{1,M}, \ldots, c_{2^k-1,0} \ldots, c_{2^k-1,M}] \\ &= [c_1, c_2, \ldots, c_{\hat{m}}], \end{aligned} \tag{3}$$

$$\begin{aligned} \mathbf{\Psi}(t)^T &= [\psi_{0,0}(t), \ldots, \psi_{0,M}(t), \psi_{1,0}(t), \ldots, \psi_{1,M}(t), \ldots, \psi_{2^k-1,0}(t), \ldots, \psi_{2^k-1,M}(t)] \\ &= [\psi_1(t), \psi_2(t), \ldots, \psi_{\hat{m}}(t)]. \end{aligned} \tag{4}$$

The purpose of this paper is introducing an operational method for solving the following two dimensional Abel equation, by using ChWs:

$$u(x,y) = f(x,y) + \int_0^x \int_0^y \frac{u(s,t)}{(x-s)^{\alpha}(y-t)^{\beta}}\, ds\, dt, \tag{5}$$

where $0 < \alpha < 1$, $0 < \beta < 1$, and f is a given function.

2 Main results

2.1 The fractional integration in the Riemann-Liouville sense

There are several definitions of a fractional integration of order $\alpha \geq 0$, and not necessarily equivalent to each other, [?]. The most used definition is due to Riemann-Liouville, which is defined as:

$$I^{\alpha}_{0,t} f(t) = \begin{cases} \frac{1}{\Gamma(\alpha)} \int_0^t (t-\tau)^{\alpha-1} f(\tau)\, d\tau\ , & \alpha > 0\ ,\ t > 0, \\ f(t), & \alpha = 0. \end{cases} \tag{6}$$

One of the basic properties of the operator $I^{\alpha}_{0,t}$ is:

$$I^{\alpha}_{0,t}\, x^{\beta} = \frac{\Gamma(\beta+1)}{\Gamma(\beta+\alpha+1)}\, x^{\beta+\alpha}.$$

2.2 Fractional Integration of ChW Vector $\mathbf{\Psi(t)}$

Let $\mathbf{\Psi(t)}$ be the ChW vector of size $\hat{m} = 2^k(M+1)$ defined in (4). The Reimann-Liouville fractional integral of order α for vector $\mathbf{\Psi}(t)$ can be approximated by:

$$I^{\alpha}\mathbf{\Psi}(t) \simeq \mathbf{P}^{(\alpha)}\,\mathbf{\Psi}(t), \tag{7}$$

where $\mathbf{P}^{(\alpha)} = [p_{i,j}^{(\alpha)}]$ is an $\hat{m} \times \hat{m}$ matrix, known as the fractional operational matrix for the ChW, defined by:

$$p_{ij}^{(\alpha)} = \langle I^{\alpha}\psi_i(t), \psi_j(t)\rangle.$$

After some calculations and simplifications we will have:

$$I^{\alpha}\psi_{n,m}(t) = A_m \Bigg[u_{b_n}(t)\,(t-b_n)^{\alpha} \sum_{j=0}^{M-m} \frac{a_j\,(m+j)!}{\Gamma(m+j+\alpha+1)}\,(2^k t - n)^{m+j} - u_{b_{n+1}}(t)\,(t-b_{n+1})^{\alpha} \sum_{j=0}^{M-m}\sum_{l=0}^{m+j} \frac{a_j\,(m+j)!}{(m+j-l)!\,\Gamma(l+\alpha+1)}\,(2^k t - n - 1)^{l} \Bigg], \tag{8}$$

in which $u_a(t) = u(t-a)$ and $u(t)$ is the unit step function and $b_n = \frac{n}{2^k}L$.

For example for $k=1$, $M=2$, and $\alpha = \frac{1}{2}$, $P^{(\alpha)}$ can be obtain as:

$$\begin{pmatrix} 0.21415 & 0.21085 & 0.08753 & 0.05698 & 0.07211 & 0.07628 \\ -0.05531 & 0.33159 & 0.33295 & 0.05732 & 0.11957 & 0.13728 \\ 0.00824 & -0.03329 & 0.38685 & 0.22626 & 0.25063 & 0.22944 \\ 0. & 0. & 0. & 0.21415 & 0.21085 & 0.08753 \\ 0. & 0. & 0. & -0.05531 & 0.33159 & 0.33295 \\ 0. & 0. & 0. & 0.00824 & -0.03329 & 0.38685 \end{pmatrix}.$$

If we consider definition (4) and define

$$\mathbf{\Psi}(x,y) = \mathbf{\Psi}(x) \otimes \mathbf{\Psi}(y), \tag{9}$$

where, $\otimes$ is the Kronecker product[?],

Now, let have a closer look at equation (5) and its terms and use some other useful formulas. First, note that, according to Eqs. (2) and (7), we have:

$$\int_0^x \frac{f(t)}{(x-t)^{\alpha}}\,dt = \Gamma(1-\alpha) I_{0,x}^{1-\alpha}(x).$$

Now, by using Eqs. (9) and (7) we get:

$$\begin{aligned} \int_0^x\int_0^y \frac{\psi(s,t)}{(x-s)^{\alpha}(y-t)^{\beta}}\,ds\,dt &= \int_0^x\int_0^y \frac{\psi(s)\otimes\psi(t)}{(x-s)^{\alpha}(y-t)^{\beta}}\,ds\,dt \\ &= \left(\int_0^x \frac{\psi(s)}{(x-s)^{\alpha}}\,ds\right) \otimes \left(\int_0^y \frac{\psi(t)}{(y-t)^{\beta}}\,dt\right) \\ &= \left(\Gamma(1-\alpha)\, I_{0,x}^{1-\alpha}\psi(x)\right) \otimes \left(\Gamma(1-\beta)\, I_{0,y}^{1-\beta}\psi(y)\right) \\ &= \left(\Gamma(1-\alpha)\,\Gamma(1-\beta)\right) \left(I_{0,x}^{1-\alpha}\psi(x)\right) \otimes \left(I_{0,y}^{1-\beta}\psi(y)\right) \\ &= \left(\Gamma(1-\alpha)\,\Gamma(1-\beta)\right) \left(P^{(1-\alpha)}\,\psi(x)\right) \otimes \left(P^{(1-\beta)}\psi(y)\right) \end{aligned}$$

$$= (\Gamma(1-\alpha)\,\Gamma(1-\beta))\left(P^{(1-\alpha)} \otimes P^{(1-\beta)}\right)\left(\psi(x) \otimes \psi(y)\right)$$
$$= (\Gamma(1-\alpha)\,\Gamma(1-\beta))\,P^{(1-\alpha,1-\beta)}\,\psi(x,y). \tag{10}$$

Let $u(x,y) \simeq \mathbf{C}^T\,\mathbf{\Psi}(x,y)$ and $f(x,y) \simeq \mathbf{F}^T\,\mathbf{\Psi}(x,y)$.

According to Eq. (10) it can be seen that Eq. (5) transforms to the following matrix relation:

$$\mathbf{C}^T\,\mathbf{\Psi}(x,y) \simeq \mathbf{F}^T\,\mathbf{\Psi}(x,y) + (\Gamma(1-\alpha)\,\Gamma(1-\beta))\,\mathbf{C}^T\,P^{(1-\alpha,1-\beta)}\,\mathbf{\Psi}(x,y). \tag{11}$$

Hence, Eq. (11) coverts to a linear system of equations, as:

$$\mathbf{C}^T = \mathbf{F}^T + (\Gamma(1-\alpha)\,\Gamma(1-\beta))\,\mathbf{C}^T\,P^{(1-\alpha,1-\beta)},$$

or:

$$\mathbf{C} = \mathbf{F} + (\Gamma(1-\alpha)\,\Gamma(1-\beta))\,(P^{(1-\alpha,1-\beta)})^T\,\mathbf{C},$$

and equivalently

$$\left(I - (\Gamma(1-\alpha)\,\Gamma(1-\beta))\,(P^{(1-\alpha,1-\beta)})^T\right)\mathbf{C} = \mathbf{F}. \tag{12}$$

By solving the system of linear equations (12), for unknown vector $\mathbf{C}$, the approximate solution of the main Eq. (5) can be obtained as: $u(x,y) \simeq \mathbf{C}^T\,\mathbf{\Psi}(x,y)$.

3 Numerical results

In this section, an example presented to verify the capability and efficiency of the proposed method. In this example, we consider $L_1 = L_2 = 1$. To show the error, we use

$$e(x,y) = |u(x,y) - \hat{u}(x,y)|, \tag{13}$$

in which $u(x,y)$ is the exact solution and $\hat{u}(x,y)$ is the approximate solution given by the suggested method.

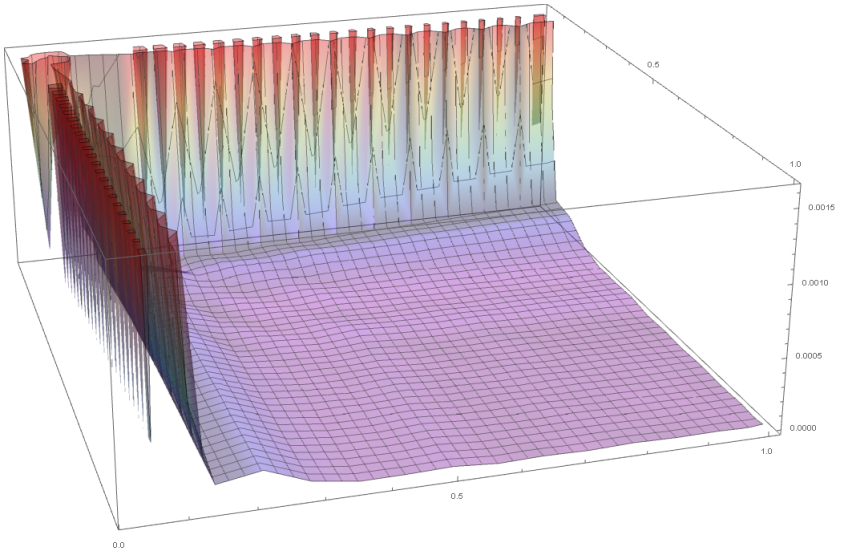


Figure 1: Error Function of Example (3.1) for $M = 2$ and $k = 3$

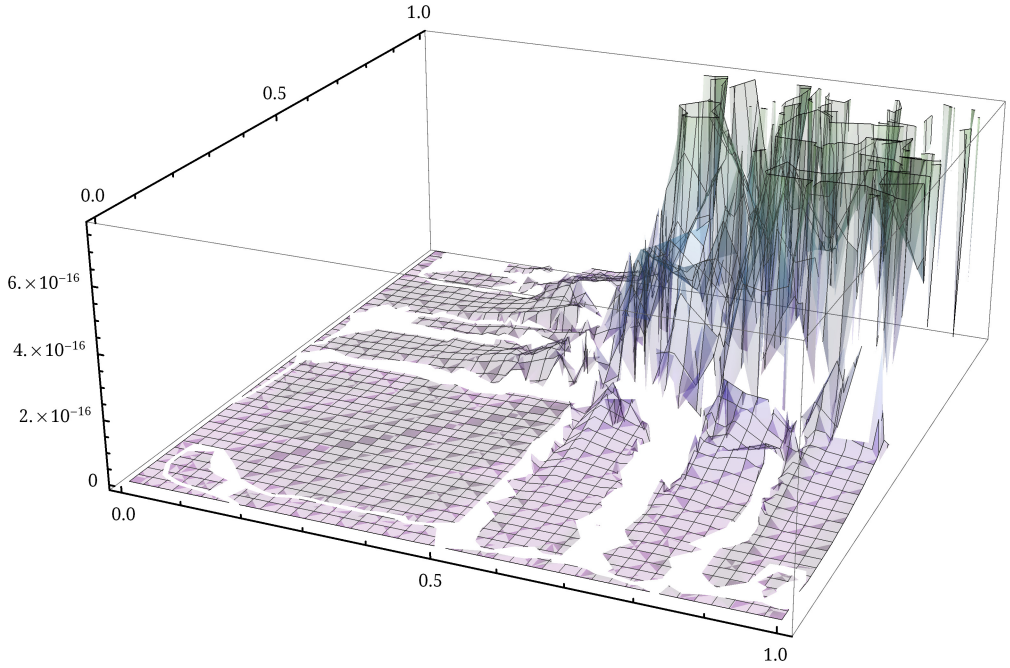


Figure 2: Error Function of Example (3.2) for $M = 3$ and $k = 1$

Example 3.1. Cosider the following two dimensional Abel equation:

$$u(x,y) = f(x,y) + \int_0^x \int_0^y \frac{u(\xi,\zeta)}{(x-\xi)^{\alpha}(y-\zeta)^{\beta}}\, d\xi\, d\zeta, \tag{14}$$

in which $\alpha = \frac{1}{2}$, $\beta = \frac{3}{4}$ and $f(x,y) = \sqrt{x\,y} - \mathcal{B}(\frac{1}{2},\frac{3}{2})\,\mathcal{B}(\frac{1}{4},\frac{3}{2})\,x\,y^{\frac{5}{4}}$, where $\mathcal{B}(a,b)$ is the Beta Function. The exact solution of the problem is $u(x,y) = \sqrt{x\,y}$. We implement the proposed method with $M = 2$ and $k = 3$. Fig. (1) shows the error function introduced in (13)

Example 3.2. Consider the following problem,

$$u(x,y) = f(x,y) + \int_0^x \int_0^y \frac{u(\xi,\zeta)}{(x-\xi)^{\alpha}(y-\zeta)^{\beta}}\, d\xi\, d\zeta,$$

where $f(x,y) = x^3y^3 - \frac{36x^{4-\alpha}y^{4-\beta}}{(1-\alpha)(2-\alpha)(3-\alpha)(4-\alpha)(1-\beta)(2-\beta)(3-\beta)(4-\beta)}$, and $\alpha = 0.3$, and $\beta = 0.5$. The exact solution is $u(x,y) = x^3y^3$.

As can be seen, in comparison with example 1 of [4], for (relatively small) $M = 3$ and $k = 1$, approximate solution is very accurate.

References

[1] Leslie Hogben, *Handbook of Linear Algebra*, Chapman and Hall/CRC, 2013.

[2] Fakhrodin Mohammadi, Numerical solution of systems of fractional delay differential equations using a new kind of wavelet basis,*Comp. Appl. Math.*, (2018) 37: 4122. https://doi.org/10.1007/s40314-017-0550-x

[3] Esmail Hesameddini, Mehdi Shahbazi, Two-dimensional shifted Legendre polynomials operational matrix method for solving the two-dimensional integral equations of fractional order, *Applied Mathematics and Computation*, 322 (2018) 40-54

[4] Yubin Pan, Jin Huang, Yanying Ma, Bernstein series solutions of multidimensional linear and nonlinear Volterra integral equations with fractional order weakly singular kernels, *Applied Mathematics and Computation*, 347 (2019) 149–161

Convex analysis and spectral functions[1]

Alireza Sattarzadeh[1,*] and Hossein Mohebi[2]

[1]Department of Mathematics, Graduate University of Advanced Technology, Kerman, Iran

[2]Department of Mathematics, Shahid Bahonar University of Kerman, Kerman, Iran

Abstract

In this paper, we investigate some properties of spectral functions from convex analysis and monotone operator theory point of view. Indeed, we study ε-subdifferential of spectral functions. Also, we present the Fitzpatrick function of the subdifferential of a spectral function in terms of the Fitzpatrick function of the subdifferential of corresponding symmetric function.

Keywords: Spectral function, Convex analysis, Monotone operator, Fitzpatrick function

Mathematics Subject Classification [2010]: 15A18, 49J52, 47A75

1 Introduction and Preliminaries

There has been growing interest in the variational analysis of spectral functions. This growing trend is due to spectral functions that have important applications to some fundamental problems in applied mathematics such as semi-definite programming and engineering problems (see [2, 3], and references therein).
A function F defined on $\mathcal{S}_n$ is called spectral if

$$F(U^T AU) = F(A),\ \forall\ A \in \mathcal{S}_n,\ \forall\ U \in \mathcal{O}_n,$$

where $\mathcal{S}_n$ is the vector space of all $n \times n$ real symmetric matrices and $\mathcal{O}_n$ is the group of all real orthogonal matrices.
One can easily see [3] that every spectral function is the composition of a symmetric function f defined on $\mathbb{R}^n$ and the eigenvalue function $\lambda : \mathcal{S}_n \longrightarrow \mathbb{R}^n$, i.e.,

$$F(A) = (f \circ \lambda)(A),\ \forall\ A \in \mathcal{S}_n.$$

Hence there exists a one-to-one correspondence between the spectral functions F defined on $\mathcal{S}_n$ and the symmetric functions f defined on $\mathbb{R}^n$. In recent years a lot of research shows that the properties of F are inherited from the properties of f, and vice versa [2–5].
The notion of a maximal monotone operator is crucial in optimization as it captures both the subdifferential operator of a convex, lower semicontinuous, and proper function and any continuous linear positive operator. It was recently discovered that most fundamental

[1]Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

[*]Speaker. Email: arsattarzadeh@gmail.com

results on maximal monotone operators allow simpler proofs utilizing Fitzpatrick functions.
We consider the Euclidean space $\mathbb{R}^n$ with the inner product $\langle .,. \rangle$ and the induced norm $\|.\|$. For a function $f : \mathbb{R}^n \longrightarrow \bar{\mathbb{R}} := [-\infty, +\infty]$, define the domain of f by

$$dom(f) := \{x \in \mathbb{R}^n : f(x) < +\infty\}.$$

We say that f is proper if $dom(f) \neq \emptyset$ and $f(x) > -\infty$ for all $x \in \mathbb{R}^n$. The set of all proper lower semi-continuous (l.s.c) and convex functions defined on $\mathbb{R}^n$ with values in $\bar{\mathbb{R}}$ is denoted by $\Gamma_0(\mathbb{R}^n)$. The Fenchel-Moreau conjugate of a function $f : \mathbb{R}^n \longrightarrow \bar{\mathbb{R}}$ is defined by $f^* : \mathbb{R}^n \longrightarrow \bar{\mathbb{R}}$

$$f^*(x) := \sup_{y \in \mathbb{R}^n} \{\langle x, y \rangle - f(y)\}, \ \forall \ x \in \mathbb{R}^n,$$

and the second conjugate (or bi-conjugate) of f is defined by

$$f^{**}(x) := \sup_{y \in \mathbb{R}^n} \{\langle x, y \rangle - f^*(y)\}, \ \forall \ x \in \mathbb{R}^n.$$

Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a function and $x_0 \in dom(f)$. Recall [1] that the subdifferential of f is the set valued mapping $\partial f : \mathbb{R}^n \rightrightarrows R^n$ defined by

$$\partial f(x_0) := \{u \in \mathbb{R}^n : \langle u, x - x_0 \rangle \leq f(x) - f(x_0), \quad \forall \ x \in \mathbb{R}^n\},$$

and for given $\varepsilon \geq 0$, the ε-subdifferential of f is the set valued mapping defined by

$$\partial_\varepsilon f(x_0) := \{u \in \mathbb{R}^n : \langle u, x - x_0 \rangle \leq f(x) - f(x_0) + \varepsilon, \quad \forall \ x \in \mathbb{R}^n\},$$

For set valued mapping $T : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$, we consider the graph of T by

$$G(T) := \{(x, u) \in \mathbb{R}^n \times \mathbb{R}^n : u \in Tx\}.$$

and T is called monotone, if

$$\langle x - y, u - v \rangle \geq 0, \quad \forall \ (x, u) \in G(T), \quad \forall \ (y, v) \in G(T).$$

A set valued mapping $T : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is called maximal monotone, if T is monotone and $T = T'$ for any monotone mapping $T' : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ such that $G(T) \subseteq G(T')$.
Let $T : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be monotone. Correspondence to the mapping T, the Fitzpatrick function $\varphi_T : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}$ is defined by [1]

$$\varphi_T(x, u) = \sup_{(y,v) \in G(T)} \{\langle x, v \rangle + \langle y, u \rangle - \langle y, v \rangle\}, \quad \forall \ (x, u) \in \mathbb{R}^n \times \mathbb{R}^n. \tag{1}$$

The following theorem is well known in convex analysis and monotone operator theory [1, 2].

Theorem 1.1. Let $f \in \Gamma_0(\mathbb{R}^n)$. Then, $\partial f : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a maximal monotone operator. Also, we have

$$\varphi_{\partial f}(x, u) \geq \langle x, u \rangle, \quad \forall \ (x, u) \in \mathbb{R}^n \times \mathbb{R}^n,$$

with equality holds if and only if $(x, u) \in G(\partial f)$.

We endow $\mathcal{S}_n$ with the trace inner product [2,3]:

$$\langle A, B\rangle := tr(AB),\ \forall\ A, B \in \mathcal{S}_n.$$

This inner product induces the Frobenius norm [3], i.e., $\|A\|_F = \sqrt{tr(A^2)}$. For any $x \in \mathbb{R}^n$, we denote by the symbol $Diag(x)$ the $n \times n$ matrix with components of x on its diagonal and with zero off the diagonal. For $x \in \mathbb{R}$, we denote the vector $x^\dagger \in \mathbb{R}^n$, with $x^\dagger := (x, x, \cdots, x)$.
Define the eigenvalue function $\lambda : \mathcal{S}_n \longrightarrow \mathbb{R}^n$ by $\lambda(A) := (\lambda_1(A), \lambda_2(A), \cdots, \lambda_n(A))$ for each $A \in \mathcal{S}_n$, where $\lambda_1(A), \lambda_2(A), \cdots, \lambda_n(A)$ are the eigenvalues of A and ordered in a non-increasing order, i.e., $\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_n(A)$. The following theorem due to von Neumann plays a central role in the spectral variation analysis.

Theorem 1.2. [2,3] For any $A, B \in \mathcal{S}_n$, we have

$$\|\lambda(A) - \lambda(B)\| \leq \|A - B\|_F,$$

and

$$\langle A, B\rangle \leq \langle \lambda(A), \lambda(B)\rangle. \tag{2}$$

Every $A \in \mathcal{S}_n$ admits a spectral decomposition of the form $A = UDiag(\lambda(A))U^T$ for some $U \in \mathcal{O}_n$. For each $A \in \mathcal{S}_n$, define the set of all orthogonal matrices giving the ordered spectral decomposition of A by

$$\mathcal{O}_A := \{U \in \mathcal{O}_n : U^T AU = Diag(\lambda(A))\}.$$

It is clear that $\mathcal{O}_A$ is non-empty for each $A \in \mathcal{S}_n$.
A function $F : \mathcal{S}_n \longrightarrow \bar{\mathbb{R}}$ is called spectral if F is $\mathcal{O}_n$-invariant, i.e.,

$$F(U^T AU) = F(A),\ \forall\ A \in dom(F),\ \forall\ U \in \mathcal{O}_n.$$

It is not difficult to see [3] that any spectral function F defined on $\mathcal{S}_n$ can be written as a composition $f \circ \lambda$ for some symmetric function f defined on $\mathbb{R}^n$ (a function $f : \mathbb{R}^n \longrightarrow \bar{\mathbb{R}}$ is called symmetric if $f(x) = f(Px)$ for all permutation matrices P and for all $x \in \mathbb{R}^n$). For instance, it is well-known that for each $A \in \mathcal{S}_n$,

$$\|A\|_F^2 = \sum_{i=1}^{n} [\lambda_i(A)]^2 = \|\lambda(A)\|^2,$$

i.e.,

$$\|A\|_F = (\|.\| \circ \lambda)(A).$$

The above relation shows that the Frobenius norm is a spectral function defined on $\mathcal{S}_n$ associated with the standard Euclidean norm on $\mathbb{R}^n$.
The following theorems present some properties of spectral functions in point of view convex analysis.

Theorem 1.3. [2,3] Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a symmetric function. Then, $f \in \Gamma_0(\mathbb{R}^n)$ if and only if $f \circ \lambda \in \Gamma_0(\mathcal{S}_n)$. Also, one has

$$(f \circ \lambda)^*(A) = f^* \circ \lambda(A),\ \forall\ A \in \mathcal{S}_n. \tag{3}$$

Theorem 1.4. [2,3] Let $f \in \Gamma_0(\mathbb{R}^n)$ be symmetric function. Let $A \in \mathcal{S}_n$ be arbitrary. Then,

$$\partial(f \circ \lambda)(A) = \{UDiag(v)U^T : v \in \partial f(\lambda(A)),\quad U \in \mathcal{O}_A\}.$$

Also, if $B \in \partial(f \circ \lambda)(A)$, Then A and B are simultaneously diagonalizable.

2 Main results

We first present some properties of subdifferential of the spectral function. This properties are immediate consequence of Theorem 1.4.

Lemma 2.1. Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a symmetric function. Let $A, B \in \mathcal{S}_n$. Then the following assertions are true:

1) If $B \in \partial(f \circ \lambda)(A)$, then $\lambda(B) \in \partial f(\lambda(A))$.

2) If $B \in \partial(f \circ \lambda)(A)$ and $U \in \mathcal{O}_n$, then $UBU^T \in \partial(f \circ \lambda)(UAU^T)$.

3) $y \in \partial f(x)$ if and only if $Diag(y) \in \partial(f \circ \lambda)(Diag(x))$.

The following theorem states properties of the ε-subdifferential of the spectral function.

Theorem 2.1. Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a symmetric function. Let $A, B \in \mathcal{S}_n$. Then the following assertions hold:

1) If $B \in \partial_\varepsilon(f \circ \lambda)(A)$, then $\lambda(B) \in \partial_\varepsilon f(\lambda(A))$.

2) Let $v \in \partial_\varepsilon f(\lambda(A))$ and $U \in \mathcal{O}_A$. Then, $UDiag(v)U^T \in \partial_\varepsilon(f \circ \lambda)(A)$.

3) Suppose that $\lambda(B) \in \partial_\varepsilon f(\lambda(A))$, and A, B are simultaneously diagonalizable. Then, $B \in \partial_\varepsilon(f \circ \lambda)(A)$.

The following lemma is an immediate consequence of Theorem 1.1 and Theorem 1.3.

Lemma 2.2. Let $f \in \Gamma_0(\mathbb{R}^n)$ be a symmetric function. Then, $\partial(f \circ \lambda)$ is a maximal monotone operator on $\mathcal{S}_n$.

Now, we investigate the Fitzpatrick function of the subdifferential of the spectral function.

Theorem 2.2. Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a symmetric function. Let $A, B \in \mathcal{S}_n$ be arbitrary. Then

$$\varphi_{\partial(f\circ\lambda)}(A, B) \leq \varphi_{\partial f}(\lambda(A), \lambda(B)). \tag{4}$$

Furthermore, suppose that one of the following assertions holds:

(i) A and B are simultaneously diagonalizable.

(ii) $G(\partial f) = \{(x^\dagger, y^\dagger) : x, y \in \mathbb{R}\}$.

Then, equality holds in (4).

Proof. First, note that it follows from (3) that

$$\varphi_{\partial(f\circ\lambda)}(A, B) = \sup_{Y \in \partial(f\circ\lambda)(X)} \{\langle A, Y\rangle + \langle X, B\rangle - \langle X, Y\rangle\},$$

and

$$\varphi_{\partial f}(\lambda(A), \lambda(B)) = \sup_{y \in \partial f(x)} \{\langle x, \lambda(B)\rangle + \langle \lambda(A), y\rangle - \langle x, y\rangle\}.$$

Let $Y \in \partial(f \circ \lambda)(X)$ be arbitrary. Theorem 1.4 implies that there exists $U \in \mathcal{O}_X \cap \mathcal{O}_Y$ such that

$$X = UDiag(\lambda(X))U^T, \qquad Y = UDiag(\lambda(Y))U^T.$$

Now, consider

$$\begin{aligned}\langle A,Y\rangle+\langle X,B\rangle-\langle X,Y\rangle &= \langle A,Y\rangle+\langle X,B\rangle-\langle UDiag(\lambda(X))U^T,UDiag(\lambda(Y))U^T\rangle\\ &= \langle A,Y\rangle+\langle X,B\rangle-\langle Diag(\lambda(X)),Diag(\lambda(Y))\rangle\\ &\leq \langle\lambda(A),\lambda(Y)\rangle+\langle\lambda(X),\lambda(B)\rangle-\langle\lambda(X),\lambda(Y)\rangle\\ &\leq \varphi_{\partial f}(\lambda(A),\lambda(B)).\end{aligned}$$

By taking supremum over all $(X,Y)\in G(\partial(f\circ\lambda))$, we get

$$\varphi_{\partial(f\circ\lambda)}(A,B)\leq\varphi_{\partial f}(\lambda(A),\lambda(B)). \tag{5}$$

Now, suppose that assertion (i) holds. Let $U\in\mathcal{O}_n$ be such that

$$A=UDiag(\lambda(A))U^T,\qquad B=UDiag(\lambda(B))U^T.$$

Let $y\in\partial f(x)$ be arbitrary. Consider

$$\begin{aligned}&\langle x,\lambda(B)\rangle+\langle\lambda(A),y\rangle-\langle x,y\rangle\\ &=\langle Diag(x),Diag(\lambda(B))\rangle+\langle Diag(\lambda(A)),Diag(y)\rangle-\langle Diag(x),Diag(y)\rangle\\ &=\langle Diag(x),U^TBU\rangle+\langle U^TAU,Diag(y)\rangle-\langle Diag(x),Diag(y)\rangle\\ &=\langle UDiag(x)U^T,B\rangle+\langle A,UDiag(y)U^T\rangle-\langle UDiag(x)U^T,UDiag(y)U^T\rangle\\ &\leq\varphi_{\partial(f\circ\lambda)}(A,B).\end{aligned}$$

Taking supremum over all $(x,y)\in\partial f$. We conclude that the reverse of the inequality (4) holds.
Now, assume that assertion (ii) holds. Let $(x^\dagger,y^\dagger)\in G(\partial f)$ be arbitrary. Since $\sum_{i=1}^n c_{ii}=\sum_{i=1}^n\lambda_i(C)$, for each $C=(c_{ij})\in\mathcal{S}_n$. Hence

$$\begin{aligned}&\langle x^\dagger,\lambda(B)\rangle+\langle\lambda(A),y^\dagger\rangle-\langle x^\dagger,y^\dagger\rangle\\ &=\langle Diag(x^\dagger),Diag(\lambda(B))\rangle+\langle Diag(\lambda(A)),Diag(y^\dagger)\rangle-\langle Diag(x^\dagger),Diag(y^\dagger)\rangle\\ &=\langle Diag(x^\dagger),B\rangle+\langle A,Diag(y^\dagger)\rangle-\langle Diag(x^\dagger),Diag(y^\dagger)\rangle\\ &\leq\varphi_{\partial(f\circ\lambda)}(A,B).\end{aligned}$$

Now, by taking supremum over all $(x^\dagger,y^\dagger)\in G(\partial f)$, we have

$$\varphi_{\partial f}(\lambda(A),\lambda(B))\leq\varphi_{\partial(f\circ\lambda)}(A,B),$$

which completes the proof. □

Corollary 2.1. Let $f:\mathbb{R}^n\longrightarrow\mathbb{R}$ be a symmetric and sublinear function. Let A and B be simultaneously diagonalizable. Then,

$$\varphi_{\partial(f\circ\lambda)}(A,B)=(f\circ\lambda)(A)+(f\circ\lambda)^*(B).$$

References

[1] H.H. Bauschke and P.L. Combettes, Convex Analysis and Monotone Operators Theory in Hilbert Spaces, Springer, New York, 2011.

[2] J.M. Borwein and Q. Zhu, Techniques of Variational Analysis, CMS/Springer, New York, 2005.

[3] A.S. Lewis, Convex analysis on the Hermitian matrices, SIAM J. Optim., 6 (1996), 164-177.

[4] H. Mohebi and A. Salemi, Analysis of symmetric matrix valued functions, Numer. Funct. Anal. Optim., 28 (5-6) (2007), 691-715.

[5] A.R. Sattarzadeh and H. Mohebi, Some results on convex spectral functions:I, Wavelets and Linear Algebra 5(1) (2018), 49- 56.

Rational rotation matrices and linear preservers of majorization[1]

Yamin Sayyari* and Ahmad Mohammadhasani

Department of Mathematics, Sirjan University of Technology, Sirjan, Iran

Abstract

A rotation group is a group in which the elements are orthogonal matrices with determinant 1. In this paper, we study the majorization of the group of rational rotation around the origin of coordinate and identify the linear preserver transformations of this type of majorization.

Keywords: Rotation matrix, G-majorization, Linear preserver

Mathematics Subject Classification [2010]: 15A04, 15A21, 15A30, 47B49

1 Introduction

In this section we have defined the action on a group and expressed its relation to majorization.

Definition 1.1. Let G be a group (or semigroup) and X a set. Then G is said to act on X on the left if there is a mapping $\theta : G \times X \longrightarrow X$ satisfying two conditions:

1. If e is the identity element of G, then

$$\theta(e, x) = x \quad \text{for all } x \in X.$$

2. If $g_1, g_2 \in G$, then

$$\theta(g_1, \theta(g_2, x)) = \theta(g_1 g_2, x) \quad \text{for all } x \in X.$$

Similarly, G is said to act on X on the right if there is a mapping

$$\theta : X \times G \longrightarrow X$$

satisfying two conditions:

1. If e is the identity element of G, then

$$\theta(x, e) = x \quad \text{for all } x \in X.$$

[1]Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

*Speaker. Email address: y.sayyari@gmail.com

2. If $g_1, g_2 \in G$, then

$$\theta(\theta(x, g_1), g_2) = \theta(x, g_1 g_2) \quad \text{for all } x \in X.$$

If X is a real vector space, then each left action G (resp. right action G) creates a left majorization relation $\prec_{lG}$ (resp. right majorization relation $\prec_{rG}$) on X, which we will describe below.

Let X be a real vector space, $W \subseteq X$, $conv(W)$ be the convex hull of W and G be a left action (right action) on X. The group G induces an equivalence relation on X, defined by $x \simeq y$ if and only if $x = gy$ $(x = yg)$ for some $g \in G$. The equivalence classes of this relation are called the orbits of G. for each $y \in X$ the orbit of y is as follows:

$$O_G(y) = \{gy | g \in G\} \ (O_G(y) = \{yg | g \in G\}).$$

A vector x is said to be G-majorized of the left (of the right) by y and we write $x \prec_{lG} y$ $(x \prec_{rG} y)$ if $x \in conv(O_G(y))$. Let $T : X \longrightarrow X$ be a mapping and $\sim$ be a relation on X. We say T is a preserver of $\sim$ if $Tx \sim Ty$ whenever $x \sim y$, it is called a strong preserver of $\sim$ if it further satisfies $x \sim y$ whenever $Tx \sim Ty$.

2 Main results

In this section section, the concept of majorization is studied and then the linear preservers of this concept are characterized.

Definition 2.1. Let n be a natural number, define

$$R_{(n,k)} = \begin{bmatrix} \cos(\frac{2k\pi}{n}) & -\sin(\frac{2k\pi}{n}) \\ \sin(\frac{2k\pi}{n}) & \cos(\frac{2k\pi}{n}) \end{bmatrix}$$

and $G_n = \{R_{(n,k)} | 0 \leq k \leq n-1\}$. Its obvious that G_n is a group.

We use the vectors symbol $z = (x, y)^t$ or complex numbers symbol $z = x + iy$ as needed for each point on the xy-plane. For each z the orbit of $z = (x, y)^t$ is a follows:

$$O_{G_n}(z) = \{gz : g \in G_n\}.$$

We say that $z_1 = (x_1, y_1)^t$ G-majorized by $z_2 = (x_2, y_2)^t$ (denote by $z_1 \prec_n z_2$) if $z_1 \in conv(O_{G_n}(z_2))$.

Theorem 2.2. *Let z_1 and z_2 are two members of the xy-plane.*

1. *$z_1 \prec_n z_2$ if and only if the z_1 is located on or inside the regular n-polygon bound to the origin center of the coordinates and the z_2 corner.*

2. *$z_1 \sim_n z_2$ if and only if $z_1^n = z_2^n$.*

Proof. Since $R_{(n,1)}$ rotates points in the xy-plane counterclockwise through an angle $\frac{2\pi}{n}$ with respect to the x axis about the origin of a two-dimensional Cartesian coordinate system, Parts 1 and 2 are easily proven. □

Corollary 2.3. *Let z_1 and z_2 are two members of the xy-plane and $z_1 \sim_n z_2$ then $|z_1| = |z_2|$.*

Theorem 2.4. *Let T be a linear operator on $\mathbb{R}^2$. Then T preserves $\sim_n$ if and only if one of the following holds:*

1. *$n = 2$ and T is an arbitrary linear operator T on M_2.*
2. *$n \neq 2$ and*

$$[T] = \begin{bmatrix} a & b \\ -b & a \end{bmatrix} \text{ or } [T] = \begin{bmatrix} a & b \\ b & -a \end{bmatrix}$$

for some real numbers a, b. (i.e. $T(z) = Az$ or $T(z) = \overline{Az}$ for some complex number $A = a - ib$).

Proof. If

$$[T] = \begin{bmatrix} a & b \\ -b & a \end{bmatrix}$$

then

$$T(z) = (a - ib)z$$

so $z_1^n = z_2^n$ results that $((a - ib)z_1)^n = ((a - ib)z_2)^n$. If

$$[T] = \begin{bmatrix} a & b \\ b & -a \end{bmatrix}$$

then

$$T(z) = (a + ib)\overline{z}$$

so $z_1^n = z_2^n$ results that $((a + ib)\overline{z_1})^n = ((a + ib)\overline{z_2})^n$. So, T preserves $\sim_n$.

If $n = 2$, $G_2 = \{I_2, -I_2\}$ and every linear operator T preserves $\sim_2$.

Conversly, let T preserves $\sim_n$ and

$$[T] = A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

Since $(1, 0)^t \sim_n (\cos(\frac{2k\pi}{n}), \sin(\frac{2k\pi}{n}))^t = R_{(n,1)}(1, 0)^t$ for every $0 \leq k \leq n - 1$, so

$$A(1, 0)^t \sim_n A(\cos(\frac{2k\pi}{n}), \sin(\frac{2k\pi}{n}))^t$$

Hence,

$$a_{11}^2 + a_{21}^2 = (a_{11}\cos(\frac{2k\pi}{n}) + a_{12}\sin(\frac{2k\pi}{n}))^2 + (a_{21}\cos(\frac{2k\pi}{n}) + a_{22}\sin(\frac{2k\pi}{n}))^2$$

so

$$(a_{11}^2 + a_{21}^2 - a_{12}^2 - a_{22}^2)\sin^2(\frac{2k\pi}{n}) = 2(a_{11}a_{12} + a_{21}a_{22})\sin(\frac{2k\pi}{n})\cos(\frac{2k\pi}{n}) \tag{1}$$

for every $0 \leq k \leq n - 1$. Since for $n \neq 2, 4$ we can choise two $k_1 \neq k_2$ that $\cot(\frac{2k_1\pi}{n}) \neq \cot(\frac{2k_2\pi}{n})$, thus

$$\begin{cases} a_{11}^2 + a_{21}^2 - a_{12}^2 - a_{22}^2 = 0 \\ a_{11}a_{12} + a_{21}a_{22} = 0 \end{cases} \tag{2}$$

Case 1: If $a_{21} = a_{11} = 0$ so (2) results that $a_{12} = a_{22} = 0$.
Case 2: If $a_{21} = a_{12} = 0$ so (2) results that $a_{11} = \pm a_{22}$.
Case 3: If $a_{21} \neq 0$ then (2) results that

$$a_{22} = -\frac{a_{11}a_{12}}{a_{21}}$$

and

$$a_{11}^2 + a_{21}^2 - a_{12}^2 - (-\frac{a_{11}a_{12}}{a_{21}})^2 = 0$$

thus

$$(a_{11}^2 + a_{21}^2)(a_{21}^2 - a_{12}^2) = 0$$

so $a_{21} = \pm a_{12}$. $a_{21} = +a_{12}$ and (2) results that $a_{11} = -a_{22}$, and $a_{21} = -a_{12}$ and (2) results that $a_{11} = a_{22}$.
Case 4: If $n = 4$, (1) results that

$$a_{11}^2 + a_{21}^2 - a_{12}^2 - a_{22}^2 = 0 \tag{3}$$

On the other hand have

$$(a_{11}, a_{12})^t \sim_4 (a_{12}, -a_{11})^t$$

thus

$$(a_{11}^2 + a_{12}^2, a_{21}a_{11} + a_{22}a_{12})^t \sim_4 (0, a_{21}a_{12} - a_{22}a_{11})^t \tag{4}$$

also

$$(a_{21}, a_{22})^t \sim_4 (-a_{22}, a_{21})^t$$

and

$$(a_{11}a_{21} + a_{12}a_{22}, a_{21}^2 + a_{22}^2)^t \sim_4 (-a_{22}a_{11} + a_{12}a_{21}, 0)^t \tag{5}$$

of the (4) and (5) have

$$(a_{11}^2 + a_{12}^2, a_{21}a_{11} + a_{22}a_{12})^t \sim_4 (a_{11}a_{21} + a_{12}a_{22}, a_{21}^2 + a_{22}^2)^t$$

therfore

$$a_{11}^2 + a_{12}^2 = a_{21}^2 + a_{22}^2 \tag{6}$$

Equalities (3) and (6) yields that

$$a_{11}^2 = a_{22}^2 \text{ and } a_{12}^2 = a_{21}^2$$

so $a_{11} = \pm a_{22}$ and $a_{12} = \pm a_{21}$.
Now we prove that $a_{11} = -a_{22}$ and $a_{12} = +a_{21}$ or $a_{11} = +a_{22}$ and $a_{12} = -a_{21}$.
If $a_{11} = +a_{22} = a \neq 0$ and $a_{12} = +a_{21} = b \neq 0$, so

$$A = \begin{bmatrix} a & b \\ b & a \end{bmatrix}$$

Since $(1,1)^t \sim_4 (1,-1)^t$, $(a+b,b+a)^t \sim_4 (a-b,b-a)^t$. Thus $|a+b|^4 = |a-b|^4$ so $a+b = \pm(a-b)$, this is a contradiction. Similarly, if $a_{11} = -a_{22} = a \neq 0$ and $a_{12} = -a_{21} = b \neq 0$, so

$$A = \begin{bmatrix} a & b \\ -b & -a \end{bmatrix}$$

Since $(1,1)^t \sim_4 (1,-1)^t$, $(a+b,-b-a)^t \sim_4 (a-b,-b+a)^t$. Thus $|a+b|^4 = |a-b|^4$ so $a+b = \pm(a-b)$, this is a contradiction. Thus $a_{11} = -a_{22}$ and $a_{12} = +a_{21}$ or $a_{11} = +a_{22}$ and $a_{12} = -a_{21}$. □

In this section (m,n) is the largest divisor the common the two integers m,n. Let $\theta = \frac{m}{n}$ be a rational number with $(m,n) = 1$, define

$$R_{(\theta,k)} = \begin{bmatrix} \cos(2k\pi\theta) & -\sin(2k\pi\theta) \\ \sin(2k\pi\theta) & \cos(2k\pi\theta) \end{bmatrix}$$

and $G_\theta = \{R_{(\theta,k)} | k = 0,1,2,...\}$. Its obvious that G_θ is a group. For each z the orbit of $z = (x,y)^t$ is a follows:

$$O_{G_\theta}(z) = \{gz : g \in G_\theta\}.$$

We say that $z_1 = (x_1,y_1)^t$ G_θ-majorized by $z_2 = (x_2,y_2)^t$ (denote by $z_1 \prec_\theta z_2$) if $z_1 \in conv(O_{G_\theta}(z_2)$, where the notion $conv(A)$ is the convex hull of a set A.

Theorem 2.5. *Let z_1 and z_2 are two members of the xy-plane and $\theta = \frac{m}{n}$ be rational number with $(m,n) = 1$.*

1. *$z_1 \prec_\theta z_2$ if and only if the z_1 is located on or inside the regular n-polygon bound to the origin center of the coordinates and the z_2 corner.*
2. *$z_1 \sim_\theta z_2$ if and only if $z_1^n = z_2^n$.*

Corollary 2.6. *Let z_1 and z_2 are two members of the xy-plane and $z_1 \sim_\theta z_2$ then $|z_1| = |z_2|$.*

Corollary 2.7. *Let T be a linear operator on $\mathbb{R}^2$ and $\theta = \frac{m}{n}$. Then T preserves G_θ-majorized $\sim_\theta$ if and only if one of the following holds:*

1. *Any linear operator T preserves G_θ-majorized $\sim_2$.*
2. *$n \neq 2$ and*

$$[T] = \begin{bmatrix} a & b \\ -b & a \end{bmatrix} \text{ or } [T] = \begin{bmatrix} a & b \\ b & -a \end{bmatrix}$$

i.e. $T(z) = Az$ or $T(z) = A\overline{z}$ for some complex number A.

References

[1] A. Armandnejad and Z. Gashool, Strong linear preservers of g-tridiagonal majorization on $\mathbb{R}^n$. *Electronic Journal of Linear Algebra*, 123:115-121, 2012.

[2] R. Bahatia, Matrix Analysis. *Springer-Verlag, New York*, 1997.

[3] R. A. Brualdi and G. Dahl, An extension of the polytope of doubly stochastic matrices, *Linear and Multilinear Algebra*, 6(3):393-408, 2013 .

[4] G. Dahl, Matrix majorization, Linear Algebra Appl., 288 (1999), 53-73.

[5] A. M. Hasani and M. Radjabalipour, The structure of linear operators strongly preservingmajorizations of matrices. *Electronic Journal of Linear Algebra*, 15:260-268, 2006.

Electronic Journal of Linear Algebra, 31(1):13-26, 2016.

[6] A. W. Marshall, I. Olkin, and B. C. Arnold, *Inequalities: Theory of majorization and its applications*. Springer, New York, 2011.

دانشگاه شهید باهنر کرمان، ایران
دهمین سمینار جبر خطی
و کاربردهای آن
۲۶ لغایت ۲۹ مرداد ۱۳۹۹
10
Shahid Bahonar University of Kerman, Iran
Seminar on Linear Algebra
and its Applications
August 16 - 19 , 2020

Anti-pentadiagonal block band matrices with perturbed corners[1]

Maryam Shams Solary*

Department of Mathematics, Payame Noor University, Po Box 19395-3697, Tehran, Iran

Abstract

In this paper, using a suitable modification technique, an orthogonal block diagonalization and a number of formulas for anti-pentadiagonal block band persymmetric Hankel matrices with perturbed corners are shown. These formulas include block diagonalization, determinant, inverse, and eigenvalues of these matrices. Also, using an orthogonal block diagonalization for Toeplitz-plus-Hankel matrices, these results are presented. The validity of the approaches is illustrated by numerical experiments.

Keywords: Anti-pentadiagonal block band persymmetric Hankel matri, Inverse, Determinant, Eigenvalues

Mathematics Subject Classification [2010]: 15A18, 15B05

1 Introduction

Spectral and computational properties of persymmetric Hankel matrices have been studied by several authors such as Bini, Fasino and Lita da Silva in [1,3,5].
The proposed algorithms construct the fast computational methods for the evaluation of the block diagonalization, the determinant and the characteristic for anti-pentadiagonal block band persymmetric Hankel matrices with perturbed corners.
Block band symmetric Toeplitz matrices (BBST-matrices) and block band persymmetric Hankel matrices (BBPSH-matrices) arise in a wide variety of applications such as the finite difference approximation, linear dynamical systems, multigrid techniques, algorithms based on the cyclic reduction among others.
Here, we try to explain some statements that make fast computational methods for the evaluation of block diagonalizations, determinants, and characteristic polynomial of persymmetric Hankel matrices with perturbed corners.
Let $R,\ A_1,\ A_2,\ A_3$ be defined real matrices $m \times m$ and $\mathbf{H}_N$ be an $N-block \times N-block$ anti-pentadiagonal BBPSH-matrices with perturbed corners:

$$\mathbf{H}_N = \begin{pmatrix} & & & & A_3 & A_2 & R \\ & & & A_3 & A_2 & A_1 & A_2 \\ & & A_3 & A_2 & A_1 & A_2 & A_3 \\ & \iddots & \iddots & \iddots & \iddots & \iddots & \\ A_3 & A_2 & A_1 & A_2 & A_3 & & \\ A_2 & A_1 & A_2 & A_3 & & & \\ R & A_2 & A_3 & & & & \end{pmatrix}. \tag{1}$$

[1]Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University
*Speaker. Email address: shamssolary@pnu.ac.ir, shamssolary@gmail.com

Throughout, $\mathbf{I}$ is the identity matrix and $\mathbf{0}$ is the zero matrix of any size to satisfy the conformability requirement of a particular operation; the transpose of a matrix $\mathbf{S}$ is denoted by $\mathbf{S}^T$.
The following $N \times N$ symmetric matrix

$$[\mathbf{S}]_{ij} = \sqrt{\frac{2}{N+1}} \sin\left[\frac{ij\pi}{N+1}\right], \tag{2}$$

is essential in the procedure due to its special properties, particularly $\mathbf{S}^T = \mathbf{S} = \mathbf{S}^{-1}$.
The symbol $\otimes$ will denote the Kronecker product [2] and $\mathbf{I}_m$ is an $m \times m$ identity matrix and $C = A_3 + R - A_1$.

2 Main results

Theorem 2.1. *Let $\mathbf{H}_N$ be an $N-block \times N-block$ anti-pentadiagonal BBPSH-matrices with perturbed corners is given by (1) and*

$$F_i = -A_3\nu_i^2 - A_2\nu_i - (A_1 - 2A_3), \tag{3}$$

$\nu_i = 2\cos\left(\frac{Ni\pi}{N+1}\right)$, $i = 1, 2, \ldots, N$ *that F_i's are invertible matrices with simple eigenvalues.*

$$\mathbf{H}_N = [\mathbf{S} \otimes \mathbf{I}_m]\, \mathbf{P} \left(\begin{array}{c|c} \mathbf{D}_3 + \mathbf{u}\mathbf{u}^T \otimes C & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{D}_4 - \mathbf{v}\mathbf{v}^T \otimes C \end{array}\right) \mathbf{P}^T\, [\mathbf{S} \otimes \mathbf{I}_m] \tag{4}$$

that the following relation holds:

(a) If N is even,
$\mathbf{D}_3 = diag(F_1, F_3, \ldots, F_{N-1})$, $\mathbf{D}_4 = diag(F_2, F_4, \ldots, F_N)$,
$\mathbf{P}$ *is the $N-block \times N-block$ block permutation matrix and $\mathbf{u}$, $\mathbf{v}$ are defined by (19).*

(b) If N is odd,
$\mathbf{D}_3 = diag(F_1, F_3, \ldots, F_N)$, $\mathbf{D}_4 = diag(F_2, F_4, \ldots, F_{N-1})$,
$\mathbf{P}$ *is the $N-block \times N-block$ block permutation matrix and $\mathbf{u}$, $\mathbf{v}$ are defined by (21).*

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_3 \\ \vdots \\ u_{N-1} \end{pmatrix}, \quad \mathbf{v} = \begin{pmatrix} v_2 \\ v_4 \\ \vdots \\ v_N \end{pmatrix}, \tag{5}$$

where

$$u_{2i-1} = \frac{2}{\sqrt{N+1}} \sin\left(\frac{(2i-1)\pi}{N+1}\right), \quad v_{2i} = \frac{2}{\sqrt{N+1}} \sin\left(\frac{2i\pi}{N+1}\right) \tag{6}$$

$i = 1, 2, \ldots, \frac{N}{2}$, *when N is even or (b).*

Proof. We can find a class of simultaneously diagonalizable matrices which have a suitable block submatrix generating by block band persymmetric Hankel matrices by bordering

technique in [1]. Suppose that an $N-block \times N-block$ anti-pentadiagonal BBPSH-matrices with perturbed corners similar (7) and a sparse matrix $\hat{\mathbf{E}}_N$ similar (8):

$$\hat{\mathbf{H}}_n = \begin{pmatrix} & & & & A_3 & A_2 & A_1 - A_3 \\ & & & A_3 & A_2 & A_1 & A_2 \\ & & A_3 & A_2 & A_1 & A_2 & A_3 \\ & \iddots & \iddots & \iddots & \iddots & \iddots & \\ A_3 & A_2 & A_1 & A_2 & A_3 & & \\ A_2 & A_1 & A_2 & A_3 & & & \\ A_1 - A_3 & A_2 & A_3 & & & & \end{pmatrix}, \tag{7}$$

and

$$\hat{\mathbf{E}}_N = \begin{pmatrix} & & & \mathbf{0} & \mathbf{0} & A_3 + R - A_1 \\ & & & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ & & \iddots & \iddots & \iddots & \\ & \mathbf{0} & \mathbf{0} & \mathbf{0} & & \\ A_3 + R - A_1 & \mathbf{0} & \mathbf{0} & & & \end{pmatrix}. \tag{8}$$

Then $\mathbf{H}_N = \hat{\mathbf{H}}_N + \hat{\mathbf{E}}_N$, by Proposition 3.1 in [1] $[\mathbf{S} \otimes \mathbf{I}_m]\, \mathbf{P}\hat{\mathbf{E}}_N\mathbf{P}^T\, [\mathbf{S} \otimes \mathbf{I}_m]$, we get

$$[\mathbf{S} \otimes \mathbf{I}_m]\, \mathbf{P}\mathbf{H}_N\mathbf{P}^T\, [\mathbf{S} \otimes \mathbf{I}_m] = [\mathbf{S} \otimes \mathbf{I}_m]\, \mathbf{P}(\hat{\mathbf{H}}_N + \hat{\mathbf{E}}_N)\mathbf{P}^T\, [\mathbf{S} \otimes \mathbf{I}_m]\,.$$

If $\mathbf{E_N} = [E_{ij}]$

$$[E]_{ij} = \frac{2}{N+1} \sin\left(\frac{i\pi}{N+1}\right) \sin\left(\frac{Nj\pi}{N+1}\right) \left[1 + (-1)^{i+j}\right],$$

then

$$\mathbf{P}^T\, [\mathbf{E}_N \otimes C]\, \mathbf{P} - \left(\begin{array}{c|c} \mathbf{u}\mathbf{u}^T \otimes C & \mathbf{0} \\ \hline \mathbf{0} & -\mathbf{v}\mathbf{v}^T \otimes C \end{array}\right) = \mathbf{0},$$

and

$$[\mathbf{S} \otimes \mathbf{I}_m]\, \mathbf{H}_N\, [\mathbf{S} \otimes \mathbf{I}_m] = diag(F_1, F_2, \ldots, F_N) + \mathbf{E}_N.$$

By permuting rows and columns of $\hat{\mathbf{H}}_N + \hat{\mathbf{E}}_N$ according to the permutation matrices which yields

$$[\mathbf{S} \otimes \mathbf{I}_m]\, \mathbf{P}\hat{\mathbf{H}}_N\mathbf{P}^T\, [\mathbf{S} \otimes \mathbf{I}_m] = \left(\begin{array}{c|c} \mathbf{D}_3 & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{D}_4 \end{array}\right), \tag{9}$$

and

$$[\mathbf{S} \otimes \mathbf{I}_m]\, \mathbf{P}\hat{\mathbf{E}}_N\mathbf{P}^T\, [\mathbf{S} \otimes \mathbf{I}_m] = \left(\begin{array}{c|c} \mathbf{u}\mathbf{u}^T \otimes C & \mathbf{0} \\ \hline \mathbf{0} & -\mathbf{v}\mathbf{v}^T \otimes C \end{array}\right). \tag{10}$$

From Equations (9) and (10) and by adding them with together, then deduce Equation (4). □

Theorem 2.1 allows us to deduce the inverse and the determinant of the anti-pentadiagonal block band persymmetric Hankel matrices with perturbed corners.

Theorem 2.2. *Let $\mathbf{H}_N$ be an $N-block \times N-block$ anti-pentadiagonal BBPSH-matrices with perturbed corners (1) and F_i, $i = 1, 2, \ldots, N$ be given by (3), and relations in Theorem 2.1, we can find $\mathbf{H}_N^{-1}$:*

$$\mathbf{H}_N^{-1} = [\mathbf{S} \otimes \mathbf{I}_m]\, \mathbf{P} \left(\begin{array}{c|c} \mathbf{D}_3^{-1} - \mathbf{U}_1(\mathbf{I} + \mathbf{V}_1\mathbf{U}_1)^{-1}\mathbf{V}_1\mathbf{D}_3^{-1} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{D}_4^{-1} + \mathbf{U}_0(\mathbf{I} - \mathbf{V}_0\mathbf{U}_0)^{-1}\mathbf{V}_0\mathbf{D}_4^{-1} \end{array} \right) \mathbf{P}^T\, [\mathbf{S} \otimes \mathbf{I}_m]\,. \tag{11}$$

Also, we have:

Theorem 2.3. *Let $\mathbf{H}_N$ be an $N-block \times N-block$ anti-pentadiagonal BBPSH-matrices with perturbed corners (1) and F_i, $i = 1, 2, \ldots, N$ be given by (3), then by some relations in Theorem 2.2, we have:*

$$det(\mathbf{H}_N) = \left(\prod_{i=1}^{N} det(F_i) \right) det \left[I + \frac{4C}{N+1} \sum_{i=1}^{N} \sin^2 \left(\frac{(2i-1)\pi}{N+1} \right) F_{2i-1}^{-1} \right] \left[I - \frac{4C}{N+1} \sum_{i=1}^{N} \sin^2 \left(\frac{2i\pi}{N+1} \right) F_{2i}^{-1} \right]. \tag{12}$$

From the generalized of the some results of this Section and the some results of Chapter 6 in [4] the following formula is explained:

$$(\lambda \mathbf{I} - F_i)^{-1} = \sum_{j=1}^{m} F_i^{(j)} \sum_{k=0}^{r_j - 1} \frac{1}{(\lambda - \nu_j)^{k+1}} (F_i - \nu_j \mathbf{I})^k, \tag{13}$$

whenever $\lambda \notin \sigma(F_i)$ and the minimal polynomial of F_i is $p(t) = (t - \nu_1)^{r_1} \ldots (t - \nu_m)^{r_m}$, $\nu_i \neq \nu_j$ when $i \neq j$.
Jordan canonical form of F_i is $F_i = S_i J_i S_i^{-1}$, then $F_i^{(j)} = S_i D_{ij} S_i^{-1}$, where D_{ij} is a block diagonal matrix that is conformal with J_i; every block of J_i that has eigenvalue ν_j corresponds to an identity block in D_{ij} and all other blocks of D_{ij} are zero.

Theorem 2.4. *Let $\mathbf{H}_N$ be an $N-block \times N-block$ anti-pentadiagonal BBPSH-matrices with perturbed corners (16), F_i, $i = 1, 2, \ldots, N$ be invertible matrices with simple eigenvalues are given by (3), then:*
The eigenvalues of matrix $\mathbf{H}_N$ can be found by the roots of the following functions:

$$r(\lambda) = \mathbf{I} + \frac{4C}{N+1} \sum_{i=1}^{N} \sum_{j=1}^{m} \sin^2 \left(\frac{(2i-1)\pi}{N+1} \right) \frac{1}{\lambda - \nu_j^{(2i-1)}} F_{2i-1}^{(j)}, \tag{14}$$

$$s(\lambda) = \mathbf{I} - \frac{4C}{N+1} \sum_{i=1}^{N} \sum_{j=1}^{m} \sin^2 \left(\frac{2i\pi}{N+1} \right) \frac{1}{\lambda - \nu_j^{(2i)}} F_{2i}^{(j)}, \tag{15}$$

where $\nu_j^{(2i-1)}$ are eigenvalues of F_{2i-1} and $\nu_j^{(2i)}$ are eigenvalues of F_{2i}, $i = 1, 2, \ldots, \frac{N}{2}$ for N when N is even, and $i = 1, 2, \ldots, \frac{N-1}{2}$ for N when N is odd, whenever $\lambda \notin \sigma(F_i)$.

3 Final Comments

An orthogonal block diagonalization of pentadiagonal BBST-matrices and anti-pentadiagonal BBPSH-matrices both having perturbed corners are valuable. So, solution of the inverse, determinant and the characteristic polynomial is a fast way for next computations by this

block diagonalization.
Let $\mathbf{T}_N$ be an $N-block \times N-block$ pentadiagonal BBST-matrices with perturbed corners:

$$\mathbf{T}_N = \begin{pmatrix} R & A_2 & A_3 & & & & \\ A_2 & A_1 & A_2 & A_3 & & & \\ A_3 & A_2 & A_1 & A_2 & A_3 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & A_3 & A_2 & A_1 & A_2 & A_3 \\ & & & A_3 & A_2 & A_1 & A_2 \\ & & & & A_3 & A_2 & R \end{pmatrix}. \tag{16}$$

In [6], we have:

Theorem 3.1. *Let $\mathbf{T}_N$ be an $N-block \times N-block$ pentadiagonal BBST-matrices with perturbed corners is given by (16) and*

$$B_i = A_3\mu_i^2 + A_2\mu_i + A_1 - 2A_3, \tag{17}$$

$\mu_i = 2\cos\left(\frac{i\pi}{N+1}\right)$, $i = 1, 2, \ldots, N$ *that B_i's are matrices with simple eigenvalues, $\mathbf{S}$ is the $N \times N$ symmetric matrix (2) and $C = A_3 + R - A_1$, then:*

$$\mathbf{T}_N = [\mathbf{S} \otimes \mathbf{I}_m]\, \mathbf{P} \left(\begin{array}{c|c} \mathbf{D}_1 + \mathbf{u}\mathbf{u}^T \otimes C & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{D}_2 + \mathbf{v}\mathbf{v}^T \otimes C \end{array}\right) \mathbf{P}^T\, [\mathbf{S} \otimes \mathbf{I}_m] \tag{18}$$

where (a) $\mathbf{D}_1 = diag(B_1, B_3, \ldots, B_{N-1})$, $\mathbf{D}_2 = diag(B_2, B_4, \ldots, B_N)$,
$\mathbf{P}$ is the $N-block \times N-block$ permutation matrix.

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_3 \\ \vdots \\ u_{N-1} \end{pmatrix}, \quad \mathbf{v} = \begin{pmatrix} v_2 \\ v_4 \\ \vdots \\ v_N \end{pmatrix}, \tag{19}$$

where

$$u_{2i-1} = \frac{2}{\sqrt{N+1}} \sin\left(\frac{(2i-1)\pi}{N+1}\right), \quad v_{2i} = \frac{2}{\sqrt{N+1}} \sin\left(\frac{2i\pi}{N+1}\right) \tag{20}$$

$i = 1, 2, \ldots, \frac{N}{2}$, *when N is even or (b),*

$\mathbf{D}_1 = diag(B_1, B_3, \ldots, B_N)$, $\mathbf{D}_2 = diag(B_2, B_4, \ldots, B_{N-1})$,
$\mathbf{P}$ is the $N-block \times N-block$ permutation matrix.

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_3 \\ \vdots \\ u_N \end{pmatrix}, \quad \mathbf{v} = \begin{pmatrix} v_2 \\ v_4 \\ \vdots \\ v_{N-1} \end{pmatrix}, \tag{21}$$

where

$$u_{2i-1} = \frac{2}{\sqrt{N+1}} \sin\left(\frac{(2i-1)\pi}{N+1}\right), \quad v_{2i} = \frac{2}{\sqrt{N+1}} \sin\left(\frac{2i\pi}{N+1}\right) \tag{22}$$

$i = 1, 2, \ldots, \frac{N-1}{2}$, *whenever N is odd.*

The same conditions apply in the case of block $\mathbf{T}_N + \mathbf{H}_N$ matrices with perturbed corners:

$$\mathbf{T}_N + \mathbf{H}_N = \begin{pmatrix} R & A_2 & A_3 & & & & & & & A_3 & A_2 & R \\ A_2 & A_1 & A_2 & A_3 & & & & & A_3 & A_2 & A_1 & A_2 \\ A_3 & A_2 & A_1 & A_2 & A_3 & & & A_3 & A_2 & A_1 & A_2 & A_3 \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots & \iddots & \iddots & \iddots & \iddots & \iddots & \\ & \iddots & \iddots & \iddots & \iddots & \iddots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \\ A_3 & A_2 & A_1 & A_2 & A_3 & & & A_3 & A_2 & A_1 & A_2 & A_3 \\ A_2 & A_1 & A_2 & A_3 & & & & & A_3 & A_2 & A_1 & A_2 \\ R & A_2 & A_3 & & & & & & & A_3 & A_2 & R \end{pmatrix}. \tag{23}$$

These implements are constructive for finding the functions of matrices, eigenvalues, eigenvectors, integer powers and parallel computations in similar cases.
Now by Theorem 2.1 in the last section, we can derive:

Theorem 3.2. *Let $\mathbf{T}_N + \mathbf{H}_N$ be an $N-$block $\times$ $N-$block pentadiagonal Toeplitz-plus-Hankel matrices with perturbed corners is given by (23), then:*

$$\mathbf{T}_N + \mathbf{H}_N = [\mathbf{S} \otimes \mathbf{I}_m]\, \mathbf{P} \left(\begin{array}{c|c} \mathbf{D}_1 + \mathbf{D}_3 + 2\mathbf{u}\mathbf{u}^T \otimes C & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array} \right) \mathbf{P}^T\, [\mathbf{S} \otimes \mathbf{I}_m]\,, \tag{24}$$

that is used by the notations introduced in Theorem 2.1 and Theorem 3.1.

References

[1] D. Bini, M. Capovani, Sepctral and computational properties of band symmetric Toeplitz matrices, *Linear Algebra Appl.*, 52/53 (1983) 99–126.

[2] D. Bini, M. Capovaui, Tensor rank and border rank of band Toeplitz matrices, *SIAM Journal on Computing*, 16 (2) (1987) 252–258.

[3] D. Fasino, Spectral and structural properties of some pentadiagonal symmetric matrices, *Calcolo*, 25 (4) (1988) 301–310.

[4] C. R. Johnson, R. A. Horn, *Topics in Matrix Analysis*, Cambridge University Press 1991.

[5] J. Lita da Silva, On anti-pentadiagonal persymmetric Hankel matrices with perturbed corners, *Comput. Math. Appl.*, 72 (2016) 415–426.

[6] M.S. Solary, Computational properties of pentadiagonal and anti-pentadiagonal block band matrices with perturbed corners, *Soft Computing*, Accepted for publication October (2019).

Multivariate group majorization on $M_{n,m}$ and its linear preservers[1]

Mohammad Soleymani*

Department of Mathematics, Shahid Bahonar University of Kerman, Kerman, Iran

Abstract

For $X, Y \in M_{n,m}$, X is said to be multivariate majorized by Y, denoted by $X \prec_m Y$, if there exists a doubly stochastic matrix $D \in M_n$ such that $X = DY$. In this paper, we extend multivariate majorization as a group majorization on $M_{n,m}$. Let G be a subgroup of orthogonal group $O(\mathbb{R}^n)$. We say that X is GM-majorized by Y (written as $X \prec_{GM} Y$), if $X = \sum_{i=1}^{k} c_i g_i Y$ for some $g_i \in G$, $c_i \geq 0$, and $\sum_{i=1}^{k} c_i = 1$. We state equivalent conditions for linear preservers of multivariate group majorization.

Keywords: Matrix majorization, Group majorization, Linear preserver

Mathematics Subject Classification [2010]: 15A86, 15A39, 15B51

1 Introduction

Definition 1.1. For $x, y \in \mathbb{R}^n$, we say that y majorizes x and write $x \prec y$, if

$$\sum_{i=1}^{k} x_i^{\downarrow} \leqslant \sum_{i=1}^{k} x_i^{\downarrow}$$

for $k = 1, \ldots, n-1$ and equality holds for $k = n$, where $x^{\downarrow} = (x_1^{\downarrow}, \ldots, x_n^{\downarrow})$ is arrangement of x in non-increasing order.

We say that a linear operator $A : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ preservers majorization, if $Ax \prec Ay$ whenever $x \prec y$. The following theorem has an essential role to characterize linear preservers of majorization, see [1].

Theorem 1.2. *[1, Theorem 2.6] Let A be a linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$. Then the following conditions are mutually equivalent:*

a *A preserves majorization.*

b *$Ax \sim Ay$ whenever $x \sim y$.*

c *For any permutation matrix $\Pi \in M_n$ there exists a permutation matrix $\widehat{\Pi} \in M_m$ such that $\widehat{\Pi} A = A\Pi$.*

[1] Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

*Speaker. Email address: m.soleymani@uk.ac.ir

By above theorem, we can characterize all linear preservers of majorization.

Theorem 1.3. *[1, Corollary 2.7] Any linear operator $A : \mathbb{R}^n \to \mathbb{R}^n$ preserving majorization has one of the following forms:*

a $A = \boldsymbol{a}e^t$ *for some* $\boldsymbol{a} \in \mathbb{R}^n$.

b $A = \alpha\Pi + \beta J_n$ *for some* $\alpha, \beta \in \mathbb{R}$ *and* $\Pi \in \mathbb{P}_n$.

A matrix $D \in M_n$ is called doubly stochastic if $De = e$ and $D^te = e$. We know that $x \prec y$ if and only if $x = Dy$ for some doubly stochastic matrix D. Birkhoff theorem [3, Theorem II.2.3] says that the set of all $n \times n$ doubly stochastic matrices is the convex hull of $\mathbb{P}_n$. On the other word, $x \prec y$ if and only if $x \in \text{conv}\{Px : P \in \mathbb{P}_n\}$. By replacing $\mathbb{P}_n$ with any subgroup of orthogonal group $O(\mathbb{R}^n)$, we can define a new concept of majorization on $\mathbb{R}^n$ which is called group majorization.

Definition 1.4. Let V be a finite dimensional inner product space and G be a subgroup of orthogonal group $O(V)$. We say that x is group majorized by y, write $x \prec_G y$, if $x \in \text{conv}\{gy : g \in G\}$.

In this paper, as same as Theorem 1.2, we state an equivalent condition for matrix representations of linear preservers $T : M_{n,m} \to M_{n,m}$ of G-majorizations, where G is a finite subgroup of $O(\mathbb{R}^n)$.

2 Main results

The concept of matrix majorization is defined by directional majorization [5] or multivariate majorization [2] as follows:

Definition 2.1. For $X, Y \in M_{n,m}$, we say that X is directional majorized by Y and write $X \prec_d Y$ if $Xv \prec Yv$ for every $v \in \mathbb{R}^m$.

Definition 2.2. For $X, Y \in M_{n,m}$, we say that X is multivariate majorized by Y and write $X \prec_m Y$ if there exists doubly stochastic matrix $D \in M_n$ such that $X = DY$.

In [2], multivariate majorization defined as $X = YD$. Since D^t is doubly stochastic, The definition of Beasley means $X^t \prec_m Y^t$ with the above definition. The concept of group majorization can be extended for matrices as follows.

Definition 2.3. For $X, Y \in M_{n,m}$, X is said to be multivariate group majorized by Y (written as $X \prec_{GM} Y$), if $X \in \text{conv}\{gY : g \in G\}$ and G is a subgroup of $O(\mathbb{R}^n)$.

On the other word, $X \prec_{GM} Y$ if $X = \sum_{i=1}^k c_i g_i Y$ where $g_i \in G$, $c_i \geq 0$, $\sum_{i=1}^k c_i = 1$. Now, we prove an equivalent condition for linear preservers of GM-majorization as Theorem 1.2. To do this, we need some preliminaries.
For every $A = (a_{ij}) \in M_{n,m}$, we associate the vector $\text{vec}(A) \in \mathbb{R}^{nm}$ defined by

$$\text{vec}(A) = [a_{11}, \ldots, a_{n1}, a_{12}, \ldots, a_{n2}, \ldots, a_{1m}, \ldots, a_{nm}]^t.$$

Let $\mathcal{B}$ be the standard basis for $M_{n,m}$. On the other word

$$\mathcal{B} = \{E_{11}, \ldots, E_{n1}, E_{12}, \ldots, E_{n2}, \ldots, E_{1m}, \ldots, E_{nm}\}.$$

Also let $[T]_{\mathcal{B}}$ be representation of T with respect to $\mathcal{B}$. Then

$$[T]_{\mathcal{B}} = \begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1m} \\ B_{21} & B_{22} & \cdots & B_{2m} \\ \vdots & \vdots & & \vdots \\ B_{m1} & B_{m2} & \cdots & B_{mm} \end{pmatrix}, \tag{1}$$

where each $B_{ij} \in M_n$ and $\mathrm{vec}(T(X)) = [T]_{\mathcal{B}}(\mathrm{vec}(X))$. Let $A \in M_{n,m}$, $X \in M_{m,p}$, $B \in M_{p,q}$ and $C \in M_{n,q}$. By [4, Lemma 4.3.1], $AXB = C$ if and only if

$$\mathrm{vec}(C) = \mathrm{vec}(AXB) = (B^t \otimes A)\mathrm{vec}(X). \tag{2}$$

To verify linear preservers of multivariate group majorization, we deal with $x \sim_{GM} y$ means $x \prec_{GM} y$ and $y \prec_{GM} x$. The following theorem gives an equivalent condition for $\sim_{GM}$.

Theorem 2.4. *Let $X, Y \in M_{n,m}$. Then $X \sim_{GM} Y$ if and only if $X = gY$ for some $g \in G$.*

Proof. By the definition of multivariate group majorization, $X \prec_{GM} Y$ means that $X = \sum_{t=1}^{k} \alpha_t g_t Y$. Since $g_t \in O(\mathbb{R}^n)$,

$$\|X\|_2 = \|\sum_{t=1}^{k} \alpha_t g_t Y\|_2 \leq \sum_{t=1}^{k} \alpha_t \|g_t Y\|_2 = \sum_{t=1}^{k} \alpha_t \|Y\|_2 = \|Y\|_2. \tag{3}$$

On the other hand, $Y \prec_{GM} X$ and then $\|Y\| \leq \|X\|$. Hence, equality holds in (3). If $\alpha_{t'} \neq 0$ for some $1 \leq t' \leq k$, then

$$\|\alpha_{t'} g_{t'} Y + Z\|_2 = \|\alpha_{t'} g_{t'} Y\|_2 + \|Z\|_2,$$

where $Z = \sum_{t=1, t\neq t'}^{k} \alpha_t g_t Y$. Since equality holds in triangle inequality(cauchy-schwarz inequality), $Z = \lambda \alpha_{t'} g_{t'} Y$ for some $\lambda \in \mathbb{R}$. Therefore, $X = (1+\lambda)\alpha_{t'} g_{t'} Y$. Since $\|X\|_2 = \|Y\|_2$, $(1+\lambda)\alpha_{t'} = 1$. □

The following theorem states an equivalent condition for matrix representations of linear operator $T : M_{n,m} \to M_{n,m}$ which preserves GM-majorization, where G is a finite subgroup of $O(\mathbb{R}^n)$.

Theorem 2.5. *Let G be a finite subgroup of $O(\mathbb{R}^n)$, $T : M_{n,m} \to M_{n,m}$ be a linear operator and*

$$[T]_{\mathcal{B}} = \begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1m} \\ B_{21} & B_{22} & \cdots & B_{2m} \\ \vdots & \vdots & & \vdots \\ B_{m1} & B_{m2} & \cdots & B_{mm} \end{pmatrix}.$$

Then T preserves $\sim_{GM}$ if and only if for every $g \in G$ there exists a matrix $\widehat{g} \in G$ such that $\widehat{g} B_{ij} = B_{ij} g$ for each $i = 1, \ldots, n$ and $j = 1, \ldots, m$.

Now, we will prove the follwing extention of [5, Theorem 2] as a result of Theorem 2.5.

Corollary 2.6. *Let T be a linear operator on $M_{n,m}$. The following are equivalent:*

1 *T preserves multivariate majorization.*

2 *T preserves directional majorization.*

3 $TX \prec_d TY$ *whenever* $X \prec_m Y$.

4 $TX \sim_d TY$ *whenever* $X \sim_d Y$.

5 $TX \sim_m TY$ *whenever* $X \sim_m Y$.

6 *One of the following holds :*

a *There exist* $R, S \in M_m$ *and* $P \in \mathbb{P}_n$ *such that* $T(X) = PXR + J_nXS$.

b *There exist* $A_1, \ldots, A_m \in M_{n,m}$ *such that* $T(X) = \sum_{j=1}^{m} tr(x_j)A_j$.

References

[1] T. Ando, Majorization, doubly stochastic matrices and comparison of eigenvalues, *Linear Algebra Appl.*, 118, (1989), 163–248.

[2] L.B. Beasley, S.G. Lee, Linear operators preserving multivariate majorization , *Linear Algebra Appl.*, 304 (2000) 141–159.

[3] R. Bhatia, *Matrix Analysis*, Springer-Verlag, New York, 1997.

[4] R.A. Horn, C.R. Johnson, *Topics in matrix analysis*, Cambridge University Press, 1994.

[5] C.K. Li, E. Poon, Linear operators preserving directional majorization, *Linear Algebra Appl.*, 325, (2001), 141–146.

Maps preserving strong Jordan multiple $*$-product on $*$-algebras[1]

Ali Taghavi*

Department of Mathematics, Faculty of Mathematical Sciences,
University of Mazandaran, P. O. Box 47416-1468, Babolsar, Iran

Abstract

Let $\mathcal{A}$ be an arbitrary $*$-algebra with unit I over the real or complex field $\mathbb{F}$ that contains a nontrivial idempotent P_1 and $n \geq 1$ be a natural number. It is shown that if a surjective map $\varphi : \mathcal{A} \longrightarrow \mathcal{A}$ satisfies

$$\varphi(P) \bullet_{n-1} \varphi(P) \bullet \varphi(A) = P \bullet_{n-1} P \bullet A,$$

for every $A \in \mathcal{A}$ and projection $P \in \{P_1,\ I - P_1\}$, where $A \bullet_{n-1} A$ denotes the Jordan multiple $*$-product of $n-1$ A's, then $\varphi(A) = \varphi(I)A$ for all $A \in \mathcal{A}$ and $\varphi(I)^2 = I$.

Keywords: Maps preserving, Strong Jordan multiple, $*$-product
Mathematics Subject Classification [2010]: 15A03, 15A23, 15B36

1 Introduction

Let $\mathcal{A}$ be a $*$-algebra. For $A, B \in \mathcal{A}$, we define Jordan $*$-product and Lie $*$-product of A, B respectively by $A \bullet B = AB + BA^*$ and $[A, B]_* = AB - BA^*$, which are two different kinds of new products. The products are found playing a more and more important role in some researches (see [1-3]). Recently, many mathematicians focused on the study of the new products. In [1] which $\mathcal{M}$ and $\mathcal{N}$ are two von Neumann algebras, it is proved that a not necessarily linear bijective map $\varphi : \mathcal{M} \longrightarrow \mathcal{N}$ satisfies $\varphi([S, T]_*) = [\varphi(T), \varphi(S)]_*$ for all $T, S \in \mathcal{M}$ if and only if φ is the direct sum of a linear $*$-isomorphism and a conjugate linear $*$-isomorphism. Also in [4] where $\mathcal{A}$ and $\mathcal{B}$ are two factor von Neumann algebras, it is characterized that a not necessarily linear bijective map $\Phi : \mathcal{A} \longrightarrow \mathcal{B}$ satisfies $\Phi(A \bullet B) = \Phi(A) \bullet \Phi(B)$ for all $A, B \in \mathcal{A}$ if and only if Φ is a $*$-ring isomorphism.

Let $\mathcal{R}$ be an associative ring (or an associative algebra over a field $\mathbb{F}$). Then recall a map $\varphi : \mathcal{R} \longrightarrow \mathcal{R}$ preserves strong commutativity or strong Lie Product if $[\varphi(A), \varphi(B)] = [A, B]$, for each $A, B \in \mathcal{A}$ that $[A, B]$ is Lie product i.e. $[A, B] = AB - BA$. Similarly φ preserves strong Jordan product if $\varphi(A) \circ \varphi(B) = A \circ B$, for each $A, B \in \mathcal{A}$ that $A \circ B$ is Jordan product i.e. $A \circ B = AB + BA$. The structure of linear (or nonlinear) maps that preserve strong commutativity and strong Jordan product have been investigated in [3]. Gonga et al [3] proved that every nonlinear map φ that preserves strong Jordan product on any algebra $\mathcal{R}$ with unit I over a field $\mathbb{F}$, has the form of $\varphi(A) = \varphi(I)A$, for all $A \in \mathcal{R}$, where $\varphi(I) \in \mathcal{R}$ and $\varphi(I)^2 = I$.

[1] Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

*Speaker. Email address: taghavi@umz.ac.ir

For a ring $\mathcal{R}$ and a positive integer k, recall that the k-commutator of elements $A, B \in \mathcal{R}$ is defined by $[A,B]_k = [[A,B]_{k-1}, B]$ with $[A,B]_0 = A$ and $[A,B]_1 = [A,B] = AB - BA$; similarly we define $A \circ_k B = (A \circ_{k-1} B) \circ B$ with $A \circ_0 B = A$ and $A \circ_1 B = A \circ B = AB + BA$. A map $\varphi : \mathcal{R} \longrightarrow \mathcal{R}$ is called strong k-commutativity preserver if $[\varphi(A), \varphi(B)]_k = [A,B]_k$ for all $A, B \in \mathcal{R}$ and φ is called strong k-Jordan product if $\varphi(A) \circ_k \varphi(B) = A \circ_k B$ for each $A, B \in \mathcal{R}$. Qi [2], characterizes the structure of a strong 2-commutativity preserving map on prime algebra. Also Lin and Hou [5] characterize the structure of a map that preserves Strong 3-commutativity on standard algebras. Moreover recently in [6] authors proved the concrete form of a map that preserves strong 2-Jordan product on standard operator algebras, properly infinite von Neumann algebras and nest algebras.

The aim of this paper is to extend this work by studying surjective maps that preserves strong skew Jordan multiple $*$-product on general $*$-algebras. We prove that if $\mathcal{A}$ be an aribtrary $*$-algebra (with identity I) over the real or complex field $\mathbb{F}$ that contains a nontrivial idempotent P_1 and $\varphi : \mathcal{A} \longrightarrow \mathcal{A}$ satisfies condition

$$\varphi(P) \bullet_{n-1} \varphi(P) \bullet \varphi(A) = P \bullet_{n-1} P \bullet A,$$

for every $A \in \mathcal{A}$ and projection $P \in \{P_1,\ I - P_1\}$, then $\varphi(A) = \varphi(I)A$ for all $A \in \mathcal{A}$ and $\varphi(I)^2 = I$. Where, $n \geq 1$ a natural number and $A \bullet_{n-1} A$ with repeat $n-1$ times A is the Jordan multiple $*$-product.

We are now ready to state the main results of the paper.

2 Main results

We begin by showing a preliminary lemma.

Lemma 2.1. *Let $\mathcal{A}$ be an arbitrary $*$-algebra over the real or complex field $\mathbb{F}$ that contains a nontrivial idempotent P and $n \geq 1$ a natural number. If $P \bullet_{n-1} P \bullet A = 0$, then $PA = 0 = AP$.*

Following, we will state the main results and proofs.

Theorem 2.2. *Let $\mathcal{A}$ be an arbitrary $*$-algebra with unit I over the real or complex field $\mathbb{F}$ that contains a nontrivial idempotent P_1 and $n \geq 1$ a natural number. Assume that $\varphi : \mathcal{A} \longrightarrow \mathcal{A}$ is a surjective map satisfying the condition*

$$\varphi(P) \bullet_{n-1} \varphi(P) \bullet \varphi(A) = P \bullet_{n-1} P \bullet A, \tag{1}$$

for all $A \in \mathcal{A}$ and projection $P \in \{P_1,\ I - P_1\}$. Then $\varphi(A) = \varphi(I)A$ for all $A \in \mathcal{A}$ and $\varphi(I)^2 = I$.

Proof. We assume $P_2 = I - P_1$ and organize the proof into several steps.

Step 1. φ is injective.

Step 2. i) $\varphi(A^*) = \varphi(A)^*$ for all $A \in \mathcal{A}$.
ii) $\varphi(P)^{n+1} = P$ for every $P \in \{P_1,\ P_2\}$.

Step 3. For every $A \in \mathcal{A}$ and $P \in \{P_1,\ P_2\}$, we have

$$\varphi(P)\varphi(A) + \varphi(A)\varphi(P) = PA + AP. \tag{2}$$

We prove the result in two cases.

Case 1. Let $n = 2k - 1$ and $k \in \mathbb{N}$.

Case 2. Let $n = 2k$ and $k \in \mathbb{N}$.

Step 4. $PA\varphi(P) = \varphi(P)AP$ for all $A \in \mathcal{A}$ and $P \in \{P_1,\ P_2\}$.

Step 5. $\varphi(A) = \varphi(I)A$ for all $A \in \mathcal{A}$.

□

Acknowledgment

This research work has been supported by a research grant from the University of Mazandaran.

References

[1] ZF. Bai, SP. Du, *Maps preserving product $XY - YX^*$ on von Neuman algebras*, J. Math. Anal. Appl. 2012; 386:103-109.

[2] X. Fang. Qi, *Strong 2-commutativity preserving maps on prime rings.*

[3] L. Gonga, X. Qi, J. Shao and F. Zhang, *Strong (skew) ξ-Lie commutativity preserving maps on algebras*, Cogent Mathematics. 2(1) (2015).

[4] C. Li, F. Lu, X. Fang, *Nonlinear mappings preseving product $XY + YX^*$ on factor von Neuman algebra.* Linear Algebra Appl.438 (2013), 2339-2345.

[5] M. Y. Liu, J. c. Hou, *Strong 3-commutativity preserving maps on standard algebras*, Acta Math. Sin., Engl. Ser., 2017, 33(12), 16591670 .

[6] A. Taghavi, F. Kolivand. *Maps preserving strong 2-Jordan product on some algebras*, Asian-European Journal of Mathematics, 10(3) (2017), 1750044, 1-10.

A new feedback approach to increase consensus based on experts' self confidence, trust and similarity[1]

Atefeh Taghavi[1,*], Esfandiar Eslami[2], Enrique Herrera Viedma[3] and Raquel Ureña[4]
[1]Department of Mathematics, Graduate University of Advanced Technology, Kerman, Iran
[2]Faculty of Mathematics and computer Science, Shahid Bahonar University of Kerman, Iran
[3]Department of Computer Science and Artificial Intelligence, University of Granada, Spain
[4]Institute of Artificial Intelligence, De Montfort University, Leicester, UK

Abstract

In decision-making problems, the obvious or implicit effects of the experts on each other can be used such that the final solution has a higher degree of consensus. Here we introduce a consensus model based on the similarity between the experts' preferences and trust degree on each other, to reach a higher consensus. First, each experts' profile is specified based on trust degree, self-confidence, and consistency. Then, the experts are clustered by using the cosine similarity measure. In a feedback mechanism, the experts with the low-rank profile are receiving some advice from those experts who are similar to and have a higher profile rank.

Keywords: Group decision making, Consensus, Feedback mechanism, Intuitionistic fuzzy preference relation, Similarity

Mathematics Subject Classification [2010]: 15B15, 90B50

1 Introduction

In a group decision making scenario, GDM, a group of experts need to evaluate a set of alternatives. One of the challenges is how to reach a solution with the maximum individual consistency and global consensus. Taking into account, the interpersonal relationships of experts and their impact on each other's views, we can increase the level of final consensus [1, 2]. In [3], a new feedback mechanism that works in large scale decision making processes is proposed by bringing together both decision making approaches and opinion dynamics. This mechanism involves experts' classification based on Jaccard similarity and also an inter-agent influence. As noted in [1], the Jaccard similarity function is slower than other functions. It also reaches a lower level of consensus than others. Therefore, to improve this, we could use the cosine similarity function, which is much faster than Jaccard and also it could reach a higher level of consensus with almost a stable process. In this contribution, according to the trust between the experts, their self-confidence and their individual consistency, the experts' profiles are identified; Then by using the cosine similarity (which is stable in measuring consensus, regardless of the number of experts [1, 4]), the experts are clustered. This approach helps in reaching a higher final consensus by offering recommendations for experts with low consensus degree by their higher level neighbors.

[1]Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University
[*]Speaker. Email address: taghavi.atefe@gmail.com

2 Background

Normally, in decision making problems, a set of experts, $E = \{e_1, ..., e_m\}$, are asked to declare their preferences on the set of available alternatives, $X = \{x_1, ..., x_n\}$. It is shown that the most effective method to express preferences is the pairwise comparison.

Definition 2.1 (Intuitionistic Fuzzy Preference Relation)**.** "An intuitionistic fuzzy preference relation B on a finite set of alternatives $X = \{x_1, \ldots, x_n\}$ is characterised by a membership function $\mu_B \colon X \times X \to [0, 1]$ and a non-membership function $\nu_B \colon X \times X \to [0, 1]$ such that $0 \leq \mu_B(x_i, x_j) + \nu_B(x_i, x_j) \leq 1$ for all $(x_i, x_j) \in X \times X$, with $\mu_B(x_i, x_j) = \mu_{ij}$ interpreted as the certainty degree up to which x_i is preferred to x_j; and $\nu_B(x_i, x_j) = \nu_{ij}$ interpreted as the certainty degree up to which x_i is non-preferred to x_j [5]."

In the case $\mu_{ii} = \nu_{ii} = 0.5\ \forall i \in \{1, \ldots, n\}$ and $\mu_{ji} = \nu_{ij} \forall i, j \in \{1, \ldots, n\}$ then B is reciprocal.

To estimate the preference value between a pair of alternatives, (x_i, x_j) with $(i < j)$, when an intermediate alternative x_k $(k \neq i, j)$ is available, multiplicative consistency property could be used; $mr_{ij}^k = \dfrac{r_{ik} \cdot r_{kj} \cdot r_{ji}}{r_{jk} \cdot r_{ki}}$ whereas the denominator should not be zero.

The total estimated value based on multiplicative transitivity is assessed by the average of all possible mr_{ij}^k of the pair of alternatives (x_i, x_j): $mr_{ij} = \dfrac{\sum\limits_{k \in R_{ij}^{01}} mr_{ij}^k}{\#R_{ij}^{01}}$; in which $R_{ij}^{01} = \{k \neq i, j | (r_{ik}, r_{kj}) \notin R^{01}\}$, $R^{01} = \{(1, 0), (0, 1)\}$, and $\#R_{ij}^{01}$ is the cardinality of R_{ij}^{01}. Therefor, $MR = (mr_{ij})$, can be constructed.

Definition 2.2 (Multiplicative Consistency [5])**.** A fuzzy preference relation $R = (r_{ij})$ is multiplicative consistent if and only if $R = MR$.

The consistency of a fuzzy preference relation is measured at three different levels [5]:

Consistency Index of pair of alternatives: $CL_{ij} = 1 - d(r_{ij}, mr_{ij}) \quad \forall i, j.$

Consistency Level of alternatives: $CL_i = \dfrac{\sum\limits_{j=1;\ i \neq j}^{n} CL_{ij}}{n-1}$.

Consistency Level of a fuzzy preference relation.

$$CL = \frac{\sum\limits_{i=1}^{n} CL_i}{n}. \tag{1}$$

Given a reciprocal intuitionistic fuzzy preference relation, $B = (b_{ij}) = (\langle \mu_{ij}, \nu_{ij} \rangle)$, Ureña et al. in [5] introduce the concept of experts' confidence degree with three different levels:

Definition 2.3 (Self-Confidence Degree [5])**.**

- For an given intuitionistic preference value b_{ij} the confidence degree is measured as: $CFL_{ij} = 1 - \tau_{ij}$, where $\tau_{ij} = 1 - \mu_{ij} - \nu_{ij}$ is the hesitancy degree associated to b_{ij}.
- The confidence degree associated to the alternative x_i is defined as:

$$CFL_i = \frac{\sum_{j=1, j \neq i}^{n} (CFL_{ij} + CFL_{ji})}{2(n-1)}.$$

- For a reciprocal intuitionistic fuzzy preference relation B, the confidence degree is:

$$CFL_B = \frac{\sum_{i=1}^{n} CFL_i}{n} \tag{2}$$

Normally in a group decision-making real problem, the experts' opinions affected with the others based on their trust in them. Taking into account the trust between the experts, the final solution would have higher consensus.

Definition 2.4 (Trust Function (TF) [2])**.** An ordered tuple $\gamma = (t, d)$ where $t, d \in [0, 1]$ and t, d are representing the trust and distrust degrees respectively, will be referred to as a trust function value. The set of trust function values (TFs), or trust function, will be denoted by $\Gamma = \{\gamma = (t, d) | t, d \in [0, 1]\}$.

Intuitionistic trust function (ITFs), which is more natural in real world, is defined by adding the extra condition $0 \leq t + d \leq 1$ to the TFs' definition.

Definition 2.5 (Trust Score (TS) [2])**.** The trust score associated to an ordered pair of trust/distrust values $\gamma = (t, d)$ is:

$$TS(\gamma) = \frac{t - d + 1}{2}. \tag{3}$$

3 Proposed algorithm

In Figure 1 a simple scheme of the algorithm is illustrated.

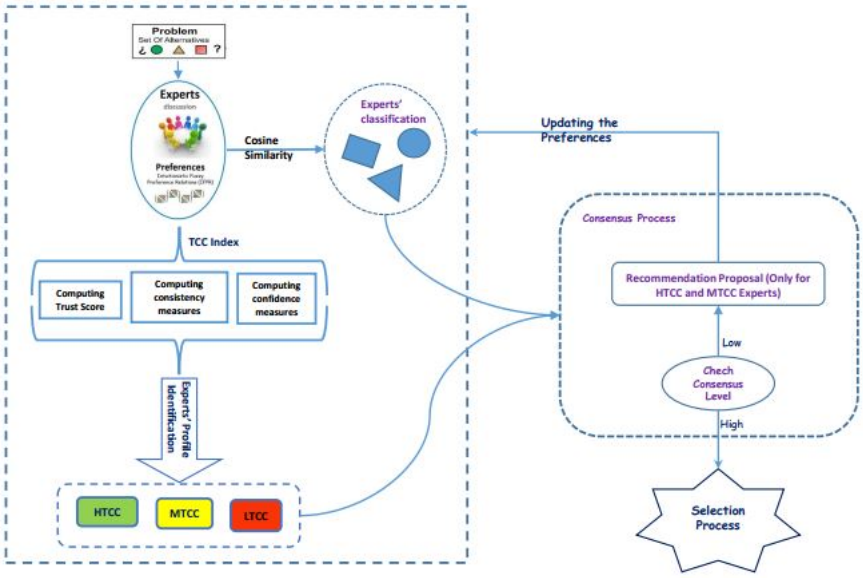


Figure 1: The Proposal Algorithm Flowchart

Definition 3.1. Experts' awareness degree: For each $e_h \in E$

$$AD^h = (\delta_1).TS^h + (\delta_2).CL^h + (\delta_3).CFL^h, \tag{4}$$

In which the parameters TS^h, is the trust score; CL^h, is the consistency degree and CFL^h is the confidence degree associated to B_h. Also, the parameters to control the weights of those three criteria in the considered variable are $\delta_i \in [0, 1], i = 1, 2, 3$ and $\sum_{i=1}^{3} \delta_i = 1$.

Given a Minimum AD Threshold $AD_{THmin} \in [0,1]$ and a superior AD threshold $AD_{THsup} \in [0,1]$, the experts can be classified in the following profiles:

- Profile 1: The experts with high degree of awareness, $HTCC$ experts, Influencers. An expert e_h is considered as a $HTCC$ expert if and only if $AD^h > AD_{THsup}$.
- Profile 2: Experts with medium level of awareness, $MTCC$ experts: $AD_{THmin} \leq AD^h \leq AD_{THsup}$.
- Profile 3: Experts with low degree of awareness, $LTCC$ experts: $AD^h \leq AD_{THmin}$.

By using a suitable distance measure, we can evaluate the similarities between experts. As investigated in [1] the cosine and dice distance functions result in fairly similar and stable global consensus levels regardless of the number of experts. When the number of experts is eight or higher, the Manhattan and the Euclidean distance functions tend to produce higher values of consensus than the previous distance functions; while for lower numbers of experts it is reverse. Besides all of these, it has been shown that the Jaccard distance function always yields the lowest global consensus values. Ureña et al. in [3], used a similarity measure based on the Jaccard distance function which considered the intersection between two experts' preference relation as the number of same preferences that both agents prefer. In [4] a cluster-based consensus measure with feedback mechanism is proposed in which the experts are clustered by using the cosine similarity of preferences. So, regard to the advantages of cosine similarity the experts are classified based on their intuitionistic fuzzy preference similarity. Here, inspired by this similarity measure, we introduce a new similarity measure between a pair of experts as follows:

Definition 3.2. Given two experts, e^p, e^q, with reciprocal intuitionistic fuzzy preference relations R^p, R^q, the similarity matrix is:

$$Sim(R^p, R^q) = CS^{pq} = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(\mu_{ij}^p.\mu_{ij}^q)}{\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(\mu_{ij}^p.\mu_{ij}^p)}.\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(\mu_{ij}^q.\mu_{ij}^q)}}$$

After computing the similarity matrix, CS^{pq}, by using previous definition, the experts could be clustered.

Definition 3.3 (AD-IOWA operator)**.** Using the reciprocal intuitionistic fuzzy preference relations $\{B^1, ..., B^m\}$, an awareness degree IOWA (AD-IOWA) operator of dimension m; Φ_w^{tcc} , is an IOWA operator whose set of order inducing values is the set of awareness degree index values, $\{AD^1, ..., AD^m\}$, associated with the set of experts. Then, the collective reciprocal intuitionistic fuzzy preference relation $B^{ad} = (b_{ij}^{ad}) = (\langle \mu_{ij}^{ad}, \nu_{ij}^{ad} \rangle)$ is computed as follows: $\mu_{ij}^{ad} = \Phi_w^{ad}(\langle AD^1, \mu_{ij}^1 \rangle, ..., \langle AD^m, \mu_{ij}^m \rangle) = \sum_{h=1}^{m} w_h.\mu_{ij}^{\sigma(h)}$ and, $\nu_{ij}^{ad} = \Phi_w^{ad}(\langle AD^1, \nu_{ij}^1 \rangle, ..., \langle AD^m, \nu_{ij}^m \rangle) = \sum_{h=1}^{m} w_h.\nu_{ij}^{\sigma(h)}$.

The weights of the AD-IOWA operator are obtained as follows: $w_h = Q(\frac{\sum_{i=1}^{h} AD^{\sigma(i)}}{T}) - Q(\frac{\sum_{i=1}^{h-1} AD^{\sigma(i)}}{T})$ in which $T = \sum_{i=1}^{m} AD^i$ and Q is the membership function of the linguistic quantifier.

The average similarity between each expert's preference relation and the global aggregate one is considered as the global consensus. The global matrix G is computed using the $AD - IOWA$ operator. So the global consensus is defined as follows:

Definition 3.4. For m experts involving in the decision making process, the overall consensus level C_s, is: $C_s = \dfrac{\sum_{h=1}^{m} CS^{hG}}{m}$ where $CS^{hG} = Sim(R^h, G)$.

If this level of consensus does not reach the minimum threshold, θ, then an iterative feedback process is activated. While C_s satisfy the minimum threshold, $C_s \geqslant \theta$, the consensus reaching process ends, and the selection process is activated to find the final solution.

In the case $C_s < \theta$, each expert with the $LTCCprofile$ will receive some recommendation from its neighbors' with both $HTCCprofile$ and $MTCCprofile$. Each experts with $MTCCprofile$ will received some advice from its neighbor with $HTCCprofile$ and $MTCCprofile$. The feedback recommendation spread is shown in Figure 2.

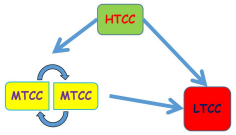


Figure 2: Feedback Spreading Scheme

4 Example

Suppose eight experts, $E = \{e_1, ..., e_8\}$, express their preferences to six alternatives, $X = \{x_1, ..., x_6\}$ by using the reciprocal intuitionistic fuzzy preference relations, which are converted to the following fuzzy preference relations according to [5]:

$$P^1 = \begin{pmatrix} 0.5 & 0.4 & 0.3 & 0.6 & 0.65 & 0.8 \\ 0.5 & 0.5 & 0.2 & 0.7 & 0.65 & 0.6 \\ 0.75 & 0.7 & 0.5 & 0.4 & 0.2 & 0.25 \\ 0.35 & 0.25 & 0.5 & 0.5 & 0.45 & 0.3 \\ 0.25 & 0.1 & 0.7 & 0.5 & 0.5 & 0.7 \\ 0.1 & 0.25 & 0.6 & 0.65 & 0.2 & 0.5 \end{pmatrix} \quad P^2 = \begin{pmatrix} 0.5 & 0.2 & 0.3 & 0.4 & 0.5 & 0.55 \\ 0.7 & 0.5 & 0.45 & 0.6 & 0.3 & 0.35 \\ 0.65 & 0.5 & 0.5 & 0.5 & 0.4 & 0.1 \\ 0.55 & 0.3 & 0.4 & 0.5 & 0.6 & 0.7 \\ 0.45 & 0.7 & 0.55 & 0.3 & 0.5 & 0.4 \\ 0.4 & 0.6 & 0.8 & 0.25 & 0.5 & 0.5 \end{pmatrix}$$

$$P^3 = \begin{pmatrix} 0.5 & 0.5 & 0.6 & 0.4 & 0.15 & 0.35 \\ 0.45 & 0.5 & 0.35 & 0.5 & 0.3 & 0.65 \\ 0.4 & 0.6 & 0.5 & 0.7 & 0.1 & 0.6 \\ 0.5 & 0.4 & 0.25 & 0.5 & 0.7 & 0.4 \\ 0.8 & 0.6 & 0.85 & 0.2 & 0.5 & 0.2 \\ 0.6 & 0.3 & 0.3 & 0.5 & 0.8 & 0.5 \end{pmatrix} \quad P^4 = \begin{pmatrix} 0.5 & 0.25 & 0.2 & 0.3 & 0.55 & 0.6 \\ 0.7 & 0.5 & 0.1 & 0.7 & 0.2 & 0.45 \\ 0.75 & 0.8 & 0.5 & 0.6 & 0.3 & 0.35 \\ 0.6 & 0.2 & 0.3 & 0.5 & 0.85 & 0.5 \\ 0.4 & 0.65 & 0.7 & 0.1 & 0.5 & 0.65 \\ 0.3 & 0.4 & 0.6 & 0.45 & 0.35 & 0.5 \end{pmatrix}$$

$$P^5 = \begin{pmatrix} 0.5 & 0.62 & 0.25 & 0.5 & 0.6 & 0.68 \\ 0.35 & 0.5 & 0.2 & 0.7 & 0.81 & 0.3 \\ 0.7 & 0.69 & 0.5 & 0.3 & 0.25 & 0.2 \\ 0.45 & 0.25 & 0.68 & 0.5 & 0.45 & 0.15 \\ 0.4 & 0.1 & 0.7 & 0.4 & 0.5 & 0.75 \\ 0.25 & 0.65 & 0.78 & 0.7 & 0.2 & 0.5 \end{pmatrix} \quad P^6 = \begin{pmatrix} 0.5 & 0.35 & 0.12 & 0.5 & 0.55 & 0.5 \\ 0.6 & 0.5 & 0.3 & 0.6 & 0.15 & 0.4 \\ 0.85 & 0.6 & 0.5 & 0.45 & 0.4 & 0.2 \\ 0.45 & 0.35 & 0.5 & 0.5 & 0.5 & 0.6 \\ 0.37 & 0.8 & 0.55 & 0.4 & 0.5 & 0.45 \\ 0.4 & 0.55 & 0.75 & 0.35 & 0.5 & 0.5 \end{pmatrix}$$

$$P^7 = \begin{pmatrix} 0.5 & 0.6 & 0.4 & 0.5 & 0.2 & 0.65 \\ 0.3 & 0.5 & 0.25 & 0.6 & 0.35 & 0.5 \\ 0.55 & 0.7 & 0.5 & 0.5 & 0.15 & 0.6 \\ 0.45 & 0.35 & 0.4 & 0.5 & 0.7 & 0.5 \\ 0.7 & 0.6 & 0.8 & 0.25 & 0.5 & 0.25 \\ 0.2 & 0.4 & 0.35 & 0.45 & 0.72 & 0.5 \end{pmatrix} \quad P^8 = \begin{pmatrix} 0.5 & 0.3 & 0.4 & 0.2 & 0.55 & 0.38 \\ 0.65 & 0.5 & 0.25 & 0.6 & 0.28 & 0.3 \\ 0.6 & 0.72 & 0.5 & 0.65 & 0.15 & 0.25 \\ 0.75 & 0.35 & 0.3 & 0.5 & 0.7 & 0.6 \\ 0.45 & 0.7 & 0.82 & 0.25 & 0.5 & 0.55 \\ 0.55 & 0.63 & 0.7 & 0.35 & 0.4 & 0.5 \end{pmatrix}$$

To classify the experts by using cosine similarity we have following similarity matrix:

$$CS = \begin{pmatrix} 0.5 & 0.90 & 0.823 & 0.913 & 0.971 & 0.919 & 0.891 & 0.89 \\ 0.9 & 0.5 & 0.85 & 0.938 & 0.848 & 0.975 & 0.89 & 0.953 \\ 0.824 & 0.849 & 0.5 & 0.862 & 0.767 & 0.841 & 0.963 & 0.889 \\ 0.913 & 0.938 & 0.862 & 0.5 & 0.864 & 0.952 & 0.906 & 0.967 \\ 0.971 & 0.848 & 0.767 & 0.864 & 0.5 & 0.868 & 0.85 & 0.852 \\ 0.919 & 0.975 & 0.84 & 0.952 & 0.868 & 0.5 & 0.901 & 0.937 \\ 0.891 & 0.89 & 0.963 & 0.906 & 0.85 & 0.901 & 0.5 & 0.9 \\ 0.89 & 0.953 & 0.889 & 0.967 & 0.852 & 0.937 & 0.9 & 0.5 \end{pmatrix}$$

Based on this similarity we obtain three clusters: $\{e_1, e_5\}$, $\{e_2, e_4, e_6, e_8\}, \{e_3, e_7\}$.

Now, for identifying experts profiles, it need to have consistency, confidence and trust degree, which are computed as follows: The consistency Level of the fuzzy preference relations; using formula (1); first multiplicative transitivity matrices, MR, are obtained. So

we have: $CL^1 = 0.53$, $CL^2 = 0.32$, $CL^3 = 0.64$, $CL^4 = 0.59, CL^5 = 0.68$, $CL^6 = 0.25$, $CL^7 = 0.34$, $CL^8 = 0.32$.

The confidence degree associated to each reciprocal intuitionistic fuzzy preference relation; formula (2) $CFL^1 = 0.91$, $CFL^2 = 0.93$, $CFL^3 = 0.94$, $CFL^4 = 0.93, CFL^1 = 0.94$, $CFL^2 = 0.94$, $CFL^3 = 0.93$, $CFL^4 = 0.96$. Now, by considering the following trust/distrust matrix, the trust score (TS) of each expert is computed.

$$TdT = \begin{pmatrix} - & (0.5, 0.43) & (0.6, 0.4) & (0.8, 0.19) & (0.7, 0.2) & (0.34, 0.45) & (0.73, 0.21) & (0.6, 0.3) \\ (0.8, 0.17) & - & (0.44, 0.55) & (0.7, 0.3) & (0.6, 0.35) & (0.73, 0.2) & (0.58, 0.4) & (0.76, 0.2) \\ (0.53, 0.4) & (0.7, 0.3) & - & (0.32, 0.65) & (0.8, 0.1) & (0.59, 0.35) & (0.9, 0.1) & (0.64, 0.32) \\ (0.62, 0.3) & (0.44, 0.5) & (0.29, 0.7) & - & (0.7, 0.2) & (0.8, 0.1) & (0.6, 0.3) & (0.9, 0.1) \\ (0.7, 0.25) & (0.8, 0.15) & (0.5, 0.4) & (0.45, 0.5) & - & (0.65, 0.3) & (0.7, 0.15) & (0.68, 0.27) \\ (0.45, 0.5) & (0.9, 0.1) & (0.7, 0.2) & (0.5, 0.4) & (0.4, 0.55) & - & (0.8, 0.15) & (0.55, 0.34) \\ (0.66, 0.3) & (0.5, 0.45) & (0.8, 0.15) & (0.7, 0.2) & (0.55, 0.35) & (0.47, 0.5) & - & (0.7, 0.27) \\ (0.8, 0.3) & (0.62, 0.53) & (0.6, 0.29) & (0.5, 0.2) & (0.7, 0.25) & (0.68, 0.22) & (0.4, 0.55) & - \end{pmatrix}$$

By using formula (3), we have: $TS^1 = 0.65$, $TS^2 = 0.67$, $TS^3 = 0.66$, $TS^4 = 0.65 TS^5 = 0.68$, $TS^6 = 0.65$, $TS^7 = 0.65$, $TS^8 = 0.64$.

Now, by considering $\delta_1 = 0.4$, $\delta_2 = 0.3$, $\delta_1 = 0.3$, and using formula (4), the trust/confidence/consistency index will be: $AD^1 = 0.69$, $AD^2 = 0.65$, $AD^3 = 0.74$, $AD^4 = 0.72 AD^5 = 0.76$, $AD^6 = 0.61$, $AD^7 = 0.64$, $AD^8 = 0.64$.

Given the minimum and superior thresholds for AD; $AD_{THmin} = 0.65, AD_{THsup} = 0.7\}$, the profiles will be: (i) Profile 1= $HTCC\ experts = \{e_5, e_3, e_4\}$, (ii) Profile 2= $MTCC\ experts = \{e_1, e_2\}$, (iii) Profile 3= $LTCC\ experts = \{e_7, e_8, e_6\}$.

So, based on this classification and the similarity between the experts, we have:

(I) e_1 receives advice from e_5, (II) e_2 receives advice from e_4, (III) e_6 receives advice from e_2, e_4, (IV) e_7 receives advice from e_3, (V) e_8 receives advice from e_2, e_4,

5 Conclusion

In this contribution we present a new consensus approach that includes a feedback mechanism in which, based on the similarity between the experts, and their interpersonal relationships some recommendations are provided to the experts.

References

[1] F. Chiclana, J. T. Garca, M. del Moral, and E. Herrera-Viedma, A statistical comparative study of different similarity measures of consensus in group decision making, *Information Sciences*, (2013) 221, 110–123.

[2] J. Wu, F. Chiclana, H. Fujita, and E. Herrera-Viedma, A visual interaction consensus model for social network group decision making with trust propagation, *Knowledge-Based Systems*, 000 (2017), 1–12.

[3] R. Ureña, F. Chiclana, G. Melancon, and E. Herrera-Viedma, A social network based approach for consensus achievement in multiperson decision making, *Information Fusion*, 47 (2019), 72–87.

[4] N. H. Kamis, F. Chiclana, and J. Levesley, Preference similarity network structural equivalence clustering based consensus group decision making model, *Applied Soft Computing*, 67 (2018), 706–720.

[5] R. Ureña, F. Chiclana, H. Fujita, and E. Herrera-Viedma, Confidence consistency driven group decision making approach with incomplete reciprocal intuitionistic preference relations, *Knowledge-Based Systems*, 89 (2015), 86–96.

Generalized Drazin inverses for linear operators[1]

Farzaneh Tayebi Semnani[1,*] and Marjan Sheibani Abdolyousefi[2]

[1]Department of Mathematics, Statistics and Computer Science,
University of Semnan, Semnan, Iran

[2]Women's University of Semnan (Farzanegan), Semnan, Iran

Abstract

Additive results for the generalized Drazin inverse of Banach space operators are presented. Under some conditions on generalized Drazin invertible operators a and b, we give explicit representations of $(a+b)^d$. Then we apply our results to 2×2 operator matrices.

Keywords: g-Drazin inverse, Additive property, Operator matrix

Mathematics Subject Classification [2010]: 15A09, 32A65, 16E50

1 Introduction

Let X be an arbitrary complex Banach space and $\mathcal{A}$ denote the Banach algebra $\mathcal{L}(X)$ of all bounded operators on X. The commutant of $a\in\mathcal{A}$ is defined by $comm(a)=\{x\in\mathcal{A}\mid xa=ax\}$. Here, $\mathcal{A}^{qnil}=\{a\in\mathcal{A}\mid 1+ax\in U(\mathcal{A}) \text{ for every } x\in comm(a)\}$. As is well known, $a\in\mathcal{A}^{qnil}\Leftrightarrow \lim\limits_{n\to\infty}\parallel a^n\parallel^{\frac{1}{n}}=0$. An element a in $\mathcal{A}$ has g-Drazin inverse, i.e., generalized Drazin inverse, provided that there exists $b\in\mathcal{A}$ such that

$$b=bab, ab=ba, a-a^2b\in\mathcal{A}^{qnil}.$$

Such b, if exists, is unique, and is called the g-Drazin inverse of a, and denote it by a^d. We use $\mathcal{A}^d$ to stand for the set of all g-Drazin invertible $a\in\mathcal{A}$. The g-Drazin inverses have various applications in singular differential and difference equations, Markov chains, and iterative methods (see [1–3]).

Suppose the bounded linear operators a and b on an arbitrary complex Banach space have g-Drazin inverses. In Section 2, we present new conditions on a,b, and prove that $a+b$ has g-Drazin inverse. These extend the results of Djordjevic and Wei [3, Theorem 2.3] and Yang and Liu [6, Theorem 2.1]. They are also the main tool in our following development. We next consider the g-Drazin inverse of a 2×2 operator matrix

$$M=\begin{pmatrix} A & B\\ C & D\end{pmatrix} \qquad (*)$$

where $A\in\mathcal{L}(X), D\in\mathcal{L}(Y)$ are GD-invertible and X,Y are complex Banach spaces. Here, M is a bounded operator on $X\oplus Y$. In Section 3, we present some g-Drazin inverses for a 2×2 operator matrix M under a number of different conditions.

[1]Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

*Speaker. Email address: ftayebis@gmail.com

2 Main results

The purpose of this section is to establish new conditions under which the sum of two g-Drazin invertible operators has g-Drazin inverse. We begin with

Lemma 2.1. *Let $a, b \in \mathcal{A}$ and $ab = 0$. If $a, b \in \mathcal{A}^d$, then $a + b \in \mathcal{A}^d$ and*

$$(a+b)^d = (1-bb^d)\Big(\sum_{n=0}^{\infty} b^n (a^d)^n\Big)a^d + b^d\Big(\sum_{n=0}^{\infty}(b^d)^n a^n\Big)(1-aa^d).$$

Lemma 2.2. *Let $a \in \mathcal{A}$ and $n \in \mathbb{N}$. Then $a^n \in \mathcal{A}^d$ if and only if $a \in \mathcal{A}^d$.*

Lemma 2.3. *Let $A \in M_{m\times n}(\mathcal{A}), B \in M_{n\times m}(\mathcal{A})$ and $k \in \mathbb{N}$. Then $(AB)^k \in M_n(\mathcal{A})^d$ if and only if $(BA)^k \in M_m(\mathcal{A})^d$.*

We are now ready to extend [6, Theorem 2.1 and Theorem 2.2] and prove:

Theorem 2.4. *Let $a, b \in \mathcal{A}^d$. If $aba = 0$ and $ab^2 = 0$, then $a + b \in \mathcal{A}^d$ and*

$$(a+b)^d = (1,b)M^d \begin{pmatrix} a \\ 1 \end{pmatrix}, M^d = F^d + G(F^d)^2,$$

where

$$F^d = (I - KK^d)\Big[\sum_{n=0}^{\infty} K^n (H^d)^n\Big]H^d + K^d\Big[\sum_{n=0}^{\infty}(K^d)^n H^n\Big](I - HH^d);$$
$$H^d = \begin{pmatrix} (a^d)^2 & 0 \\ (a^d)^3 & 0 \end{pmatrix}, K^d = \begin{pmatrix} 0 & 0 \\ (b^d)^3 & (b^d)^2 \end{pmatrix}, G^2 = 0.$$

Proof. Set

$$M = \begin{pmatrix} a^2 + ab & a^2 b \\ a + b & ab + b^2 \end{pmatrix}.$$

Then

$$M = \begin{pmatrix} ab & a^2 b \\ 0 & ab \end{pmatrix} + \begin{pmatrix} a^2 & 0 \\ a + b & b^2 \end{pmatrix} := G + F.$$

We see that $G^2 = 0$ and $GF = 0$.

$$F = \begin{pmatrix} a^2 & 0 \\ a + b & b^2 \end{pmatrix} = \begin{pmatrix} a^2 & 0 \\ a & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ b & b^2 \end{pmatrix} := H + K.$$

One easily check that

$$H = \begin{pmatrix} a^2 & 0 \\ a & 0 \end{pmatrix} = \begin{pmatrix} a \\ 1 \end{pmatrix}(a, 0).$$

Since $(a, 0)\begin{pmatrix} a \\ 1 \end{pmatrix} = a^2 \in \mathbb{A}^d$, it follows by Cline's formula (see [5, Theorem 2.1]), we see that

$$\begin{aligned} H^d &= \begin{pmatrix} a \\ 1 \end{pmatrix}((a^2)^d)^2(a, 0) = \begin{pmatrix} a \\ 1 \end{pmatrix}(a^d)^4(a, 0) \\ &= \begin{pmatrix} a(a^d)^4 P & 0 \\ (a^d)^4 a & 0 \end{pmatrix} = \begin{pmatrix} (a^d)^2 & 0 \\ (a^d)^3 & 0 \end{pmatrix}. \end{aligned}$$

Likewise, We have

$$K^d = \begin{pmatrix} 0 \\ b \end{pmatrix}(b^d)^4(1, Q) = \begin{pmatrix} 0 & 0 \\ (b^d)^3 & (b^d)^2 \end{pmatrix}.$$

Clearly, $HK = 0$. In light of Lemma 2.1,

$$F^d = (I - KK^d)\big[\sum_{n=0}^{\infty} K^n(H^d)^n\big]H^d + K^d\big[\sum_{n=0}^{\infty}(K^d)^n H^n\big](I - HH^d)$$

In light of [6, Theorem 2.1], we see that

$$M^d = F^d + G(F^d)^2.$$

Clearly, $M = (\begin{pmatrix} a \\ 1 \end{pmatrix}(1, b))^2$. By virtue of Lemma 2.1,

$$(a+b)^d = \big((1,b)\begin{pmatrix} a \\ 1 \end{pmatrix}\big)^d = (1,b)M^d\begin{pmatrix} a \\ 1 \end{pmatrix}.$$

as asserted. □

Corollary 2.5. *Let $a, b \in \mathcal{A}^{qnil}$. If $aba = 0$ and $ab^2 = 0$, then $a + b \in \mathcal{A}^{qnil}$.*

Proof. Since $a, b \in \mathcal{A}^{qnil}$, we see that $a^d = b^d = 0$. In light of Theorem 2.4, $(a+b)^d = 0$, and therefore $a + b \in \mathcal{A}^{qnil}$, as required. □

In [6], Sun et al. the Drazin inverse of $P + Q$ in the case of $PQ^2 = 0, P^2QP = 0, (QP)^2 = 0$ for two square matrices over a skew field. As is well known, every square matrix over skew fields has Drazin inverse. We are now ready to extend [6, Theorem 3.1] to g-Drazin inverses of bounded linear operators and prove:

Theorem 2.6. *Let $a, b \in \mathcal{A}^d$. If $ab^2 = 0, a^2ba = 0$ and $(ba)^2 = 0$, then $a + b \in \mathcal{A}^d$ and*

$$(a+b)^d = (1,b)M^d\begin{pmatrix} a \\ 1 \end{pmatrix}, M^d = F^d + G(F^d)^2 + G^2(F^d)^3 + G^3(F^d)^4,$$

where

$$F^d = (I - KK^d)\big[\sum_{n=0}^{\infty} K^n(H^d)^n\big]H^d + K^d\big[\sum_{n=0}^{\infty}(K^d)^n H^n\big](I - HH^d);$$
$$H^d = \begin{pmatrix} (a^d)^2 & 0 \\ (a^d)^3 & 0 \end{pmatrix}, K^d = \begin{pmatrix} 0 & 0 \\ (b^d)^3 & (b^d)^2 \end{pmatrix}, G^4 = 0.$$

Proof. Set

$$M = \begin{pmatrix} a^3 + a^2b + aba & a^3b + abab \\ a^2 + ab + ba + b^2 & a^2b + bab + b^3 \end{pmatrix}.$$

Then

$$M = \begin{pmatrix} a^2b + aba & a^3b + abab \\ 0 & a^2b + bab \end{pmatrix} + \begin{pmatrix} a^3 & 0 \\ a^2 + ab + ba + b^2 & b^3 \end{pmatrix}$$
$$:= G + F.$$

We see that $G^4 = 0, FGF = 0$ and $FG^2 = 0$. Moreover, we have

$$\begin{aligned} F &= \begin{pmatrix} a^3 & 0 \\ a^2 + ba & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ b^2 + ab & b^3 \end{pmatrix} \\ &:= H + K. \end{aligned}$$

One easily check that

$$H = \begin{pmatrix} a^3 & 0 \\ a^2 + ba & 0 \end{pmatrix} = \begin{pmatrix} a^2 \\ a + b \end{pmatrix}(a, 0).$$

Since $(a,0)\begin{pmatrix} a^2 \\ a+b \end{pmatrix} = a^3 \in \mathcal{A}^d$, it follows by Cline's formula, we see that

$$\begin{aligned} H^d &= \begin{pmatrix} a^2 \\ a+b \end{pmatrix}((a^3)^d)^2(a,0) = \begin{pmatrix} a^2 \\ a+b \end{pmatrix}(a^d)^6(a,0) \\ &= \begin{pmatrix} (a^d)^3 & 0 \\ (a^d)^4+b(a^d)^5 & 0 \end{pmatrix}. \end{aligned}$$

Likewise, We have

$$K^d = \begin{pmatrix} 0 \\ b \end{pmatrix}(b^d)^4(1,b) = \begin{pmatrix} 0 & 0 \\ (b^d)^3 & (b^d)^2 \end{pmatrix}.$$

Clearly, $HK = 0$. In light of Lemma 2.1,

$$F^d = (I-KK^d)\big[\sum_{n=0}^{\infty} K^n(H^d)^n\big]H^d + K^d\big[\sum_{n=0}^{\infty}(K^d)^nH^n\big](I-HH^d)$$

By Lemma 2.1 again, we have

$$M^d = F^d + G(F^d)^2 + G^2(F^d)^3 + G^3(F^d)^4.$$

Obviously, $M = \big(\begin{pmatrix} a \\ 1 \end{pmatrix}(1,b)\big)^3$. By virtue of Cline's formula,

$$(a+b)^d = \big((1,b)\begin{pmatrix} a \\ 1 \end{pmatrix}\big)^d = (1,b)M^d\begin{pmatrix} a \\ 1 \end{pmatrix},$$

as desired. □

Let $a,b \in \mathcal{A}^d$. If $a^2b = 0, aba^2 = 0$ and $(ba)^2 = 0$, then $a+b \in \mathcal{A}^d$. This can be proved in a symmetric way as in Theorem 2.6.

3 g-Drazin inverse of an operator matrix

Let $A \in \mathcal{L}(X), D \in \mathcal{L}(Y)$ be GD-invertible and M be given by $(*)$. The aim of this section is to consider a GD-invertible 2×2 operator matrix M. Using different splitting of the operator matrix M as $M = p+q$, we will apply Theorem 2.4 to obtain various conditions for a GD-invertible M, which extend [6, Theorem 2.1 and Theorem 2.2].

Theorem 3.1. *If $BCA = 0, BCB = 0, DCA = 0$ and $DCB = 0$, then M is GD-invertible.*

Proof. We easily see that

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix} = p+q,$$

where

$$p = \begin{pmatrix} A & B \\ 0 & D \end{pmatrix}, q = \begin{pmatrix} 0 & 0 \\ C & 0 \end{pmatrix}.$$

By virtue of [3, Lemma 2.2] p and q are GD-invertible. □

Corollary 3.2. *If $BC = 0$ and $DC = 0$, then M is GD-invertible.*

Proof. If $BC = 0$ then $BCA = 0$ and $BCB = 0$. If $DC = 0$, then $DCA = 0$ and $DCB = 0$. So we get the result by Theorem 3.1. □

Corollary 3.3. *If $CA = 0$ and $CB = 0$, then M is GD-invertible.*

Proof. If $CA = 0$ then $BCA = 0$ and $DCA = 0$. If $CB = 0$, then $DCB = 0$ and $BCB = 0$. So we get the result by Theorem 3.1 □

Theorem 3.4. *If $ABC = 0, ABD = 0, CBC = 0, CBD = 0$, then M is GD-invertible.*

Proof. Clearly, we have

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix} = p + q,$$

where

$$p = \begin{pmatrix} A & 0 \\ C & D \end{pmatrix}, q = \begin{pmatrix} 0 & B \\ 0 & 0 \end{pmatrix}.$$

Then by Theorem 2.4, we complete the proof as in Theorem 3.1. □

Corollary 3.5. (1) *If $BC = 0$ and $BD = 0$, then M is GD-invertible.*
(2) *If $AB = 0$ and $CB = 0$, then M is GD-invertible.*

Example 3.6. Let A, B, C be operators, acting on separable Hilbert space $l_2(\mathbb{N})$, defined as follows respectively:

$$\begin{aligned} A(x_1, x_2, x_3, x_4, \cdots) &= (x_1, x_1, x_3, x_4, \cdots), \\ B(x_1, x_2, x_3, x_4, \cdots) &= (x_1, -x_1, x_3, x_4, \cdots), \\ C(x_1, x_2, x_3, x_4, \cdots) &= (x_1 + x_2, x_1 - x_2, 0, 0, \cdots), \\ D(x_1, x_2, x_3, x_4, \cdots) &= (-x_2, x_2, 0, 0, \cdots). \end{aligned}$$

Set $M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$. Then $BCA = 0, BCB = 0, DCA = 0$ and $DCB = 0$. By virtue of Theorem 3.4, M is GD-invertible.

It is convenient this stage to include the following spiliting Theorem.

Theorem 3.7. *If $BCA = 0, BCB = 0, BDC = 0$ and $BD^2 = 0$, then M is GD-invertible.*

Proof. Let

$$p = \begin{pmatrix} A & B \\ 0 & 0 \end{pmatrix}, q = \begin{pmatrix} 0 & 0 \\ C & D \end{pmatrix}.$$

Then $M = p + q$. In view of [3, Lemma 2.2,] p and q are GD-invertible. By hypothesis, we easily verify that $pqp = 0$ and $pq^2 = 0$. This completes the proof, by Theorem 2.4. □

Theorem 3.8. *If $ABC = 0, ABD = 0, DCB = 0, BCBC = 0$ and $BCBD = 0$, then M has g-Drazin inverse.*

Proof. Write $M = p + q$, where

$$p = \begin{pmatrix} A & 0 \\ C & D \end{pmatrix}, q = \begin{pmatrix} 0 & B \\ 0 & 0 \end{pmatrix}.$$

By using [3, Lemma 2.2] it is clear that p, q have g-Drazin inverses. Obviously, $pq^2 = 0$. Also by the assumptions $ABC = 0, ABD = 0, DCB = 0$ we have $p^2qp = 0$. By using $BCBC = 0$ and $BCBD = 0$, we have $(qp)^2 = 0$. Then we get the result by Theorem 2.6. □

Corollary 3.9. *If $ABC = 0, ABD = 0, BCB = 0$ and $DCB = 0$, then M has g-Drazin inverse.*

Proof. It is special case of Theorem 3.8. □

If $AB = 0$ and $CB = 0$, we claim that M has g-Drazin inverse (see [2, Theorem 2]). This is a direct consequence of Corollary 3.9.

Example 3.10. Let $M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$, where

$$A = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{pmatrix}, B = \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}, C = \begin{pmatrix} 1 & 0 & 1 \end{pmatrix} \text{ and } D = 0$$

be complex matrices. Then $ABC = 0, ABD = 0, BCB = 0$ and $DCB = 0$. In this case, $AB, CB \neq 0$.

4 Conclusion

The g-Drazin inverse of the sum of two GD-invertible operators was presented under some conditions. These results are applied to obtain the g-Drazin inverse of 2×2 operator matrices.

References

[1] S. L. Campbell and C. D. Mayer, Generalized Inverse of Linear Transformations *Pitman*, London, 1979.

[2] C. Deng; D. S. Cvetcovic-Ilic and Y. Wei, Some results on the genrealized Derazin inverse of operator matrices, *Linear and Multilinear Algebra* **58**(2010), 503-521.

[3] D.S. Djordjevic and Y. Wei, Additive results for the generalized Drazin inverse, *J. Austral. Math. Soc.*, **73**(2002), 115–125.

[4] Y. Jiang, Y. Wen and Q. Zeng, Generalizations of Cline's formula for three generalized inverses, *Revista. Un. Mat. Argentina*, **58**(2017), 127–134.

[5] Y. Liao; J. Chen and J. Cui, Cline's formula for the generalized Drazin inverse, *Bull. Malays. Math. Sci. Soc.*, **37**(2014), 37–42.

[6] H. Yang and X. Liu, The Drazin inverse of the sum of two matrices and its applications, *J. Comput. Appl. Math.*, **235**(2011), 1412–1417.

A new bound for the Perron vector of weakly irreducible nonnegative tensors[1]

Mohsen Tourang* and Mostafa Zangiabadi

Department of Mathematics, University of Hormozgan, P. O. Box 3995, Bandar Abbas, Iran

Abstract

In this study, we obtain new lower bound for the ratio of the largest and smallest components in a Perron vector for the weakly nonnegative irreducible tensors and compare this bound to the known bounds. Numerical experiment are given to validate the efficiency of our new bound.

Keywords: Weakly irreducible nonnegative tensors, Perron-Frobenius theorem, Perron vector

Mathematics Subject Classification [2010]: 15A18, 15A69, 15A42

1 Introduction

Let $\mathbb{C}(\mathbb{R})$ be the set of all complex (real) numbers, $\mathbb{R}_+(\mathbb{R}_{++})$ be the set of all nonnegative (positive) numbers, $\mathbb{C}^n(\mathbb{R}^n)$ be the set of all dimension n complex (real) vectors, and $\mathbb{R}^n_+(\mathbb{R}^n_{++})$ be the set of all dimension n nonnegative (positive) vectors. An order m dimension n complex (real) tensor $\mathcal{A} = (a_{i_1 i_2 \ldots i_m})$, denoted by $\mathcal{A} \in \mathbb{C}^{[m,n]}(\mathcal{A} \in \mathbb{R}^{[m,n]}$, respectively), consists of n^m entries:

$$a_{i_1 i_2 \ldots i_m} \in \mathbb{C}(\mathbb{R}), \quad \forall\, i_j = 1, \ldots, n, \quad j = 1, \ldots, m.$$

A tensor $\mathcal{A} = (a_{i_1 i_2 \ldots i_m}) \in \mathbb{R}^{[m,n]}$ is called nonnegative (positive) if

$$a_{i_1 i_2 \ldots i_m} \geq 0 \ (a_{i_1 i_2 \ldots i_m} > 0), \quad \forall\, i_j = 1, \ldots, n, \quad j = 1, \ldots, m.$$

Tensors have many similarities with matrices and many related results of matrices such as eigenvector and eigenvalue can be extended to higher order tensors [3]. Furthermore, structured matrices such as nonnegative matrices and weakly irreducible matrices can also be extended to higher order tensors and these are becoming the focus of recent tensor research [3]. In recent years, the maximal eigenvalue problem and the Perron vector for nonnegative tensors has attracted special attention because it has many important applications such as positive definiteness of a multivariate form, multilinear pagerank, hypergraphs, higher-order Markov chains [3]. Chang et al. [1] generalized the Perron-Frobenius theorem from irreducible nonnegative matrices to irreducible nonnegative tensors. Friedland et al. [2] introduced weakly irreducible nonnegative tensors and

[1] Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

*Speaker. Email address: mohsentourang@gmail.com

established the Perron-Frobenius theorem for them. A Perron vector can be used for co-ranking schemes for objects and relations in multi-relational or tensor data, and higher-order Markov chains [3]. In this note, we propose a new lower bound for the ratio of the largest component and thesmallest component of a Perron vector for the weakly non-negative irreducible tensors. And we show that the proposed result improves the bounds in [4, 5].

We continue this section with some fundamental notions and properties developed in tensor analysis [3], which are needed in the subsequent section.

Definition 1.1. For a vector $x \in \mathbb{C}^n$, we use x_i to denote its components and $x^{[m-1]}$ to denote a vector in $\mathbb{C}^n$ such that

$$x_i^{[m-1]} = x_i^{m-1} \quad \textit{for all } i.$$

$\mathcal{A}x^{m-1}$ denotes a vector in $\mathbb{C}^n$, whose i th component is

$$(\mathcal{A}x^{m-1})_i = \sum_{i_2,i_3,\ldots,i_m=1}^{n} a_{ii_2\ldots i_m} x_{i_2}\ldots x_{i_m}.$$

A pair $(\lambda, x) \in \mathbb{C} \times (\mathbb{C}^n \setminus \{0\})$ is called an eigenpair (eigenvalue - eigenvector pair) of $\mathcal{A}$, if they satisfy

$$\mathcal{A}x^{m-1} = \lambda x^{[m-1]}.$$

Specifically, (λ, x) is called an H-eigenpair if $(\lambda, x) \in \mathbb{R} \times \mathbb{R}^n \setminus \{0\}$.

Definition 1.2. Let $\mathcal{A}$ be an $m-$order $n-$dimensional tensor.

(i) We call $\sigma(\mathcal{A})$ as the set of all eigenvalues of $\mathcal{A}$. Assume $\sigma(\mathcal{A}) \neq \emptyset$. Then the spectral radius of $\mathcal{A}$ is denoted by

$$\rho(\mathcal{A}) = \max\{|\lambda| \ : \ \lambda \in \sigma(A)\}.$$

(ii) We call a tensor $\mathcal{A}$ reducible if there exists a nonempty proper index subset $I \subset \langle n\rangle := \{1, 2, ..., n\}$ such that

$$a_{i_1 i_2 \ldots i_m} = 0, \quad \forall\, i_1 \in I, \quad i_2, ..., i_m \notin I.$$

If $\mathcal{A}$ is not reducible, then we call $\mathcal{A}$ irreducible.

(iii) We call a tensor $\mathcal{A}$ nonnegative weakly irreducible, if for any nonempty proper index subset $I \subset \langle n\rangle$, there is at least an entry $a_{i_1 i_2 \ldots i_m} > 0$, where $i_1 \in I$, and at least an $i_j \notin I, \quad j = 2, ..., m$.

(iv) We denote by $\delta_{i_1 i_2 \ldots i_m}$, the Kronecker symbol for the case of m indices, that is,

$$\delta_{i_1 i_2 \ldots i_m} = \begin{cases} 1, & i_1 = i_2 = \ldots = i_m, \\ 0, & \text{otherwise.} \end{cases}$$

Let us recall the Perron-Frobenius theorem for irreducible nonnegative tensors given in [1].

Theorem 1.3. *(see Theorem 1.4 of [1]) Suppose that $\mathcal{A}$ is an irreducible nonnegative tensor of order m dimension n. Then $\rho(\mathcal{A}) > 0$ is an eigenvalue of $\mathcal{A}$ with a positive eigenvector x corresponding to it.*

Remark 1.4. It is noted that the spectral radius $\rho(\mathcal{A})$ is the largest H-eigenvalue for the nonnegative tensor [1].

Note that $\rho(\mathcal{A})$ and x in Theorem 1.3 are called the Perron root and the Perron vector of $\mathcal{A}$, respectively, and $(\rho(\mathcal{A}), x)$ is regarded as a Perron eigenpair. Subsequently, Friedland et al. [2] generalized the result in Theorem 1.3 to weakly irreducible nonnegative tensors as follows:

Theorem 1.5. *Suppose that $\mathcal{A}$ is a weakly irreducible nonnegative tensor of order m dimension n. Then $\rho(\mathcal{A})$ is a positive H-eigenvalue λ, with a positive H-eigenvector x. Furthermore, λ is the unique H-eigenvalue of with a positive H-eigenvector, and x is the unique positive H-eigenvector associated with λ, up to a multiplicative constant.*

2 Main results

In the fields of numerical analysis and social networks (for example, see [6]), it is important to obtain estimates of the ratio of components of a Perron vector of $\mathcal{A}$. The problem of estimating the ratio

$$\gamma = \max_{1\le i,j\le n} \frac{x_i}{x_j}$$

for a maximal eigenvector of a positive tensor has been examined theoretically in [4, Theorem 3.5] as follows:

Theorem 2.1. *Let $\mathcal{A}$ be a positive tensor of order m dimension n with maximal eigenvector $x = (x_1, x_2, ..., x_n)^T$. Then*

$$\sqrt{\frac{R}{r}} \le \left(\max_{p,q} \frac{x_p}{x_q}\right)^{m-1}, \tag{1}$$

where $r_i(\mathcal{A}) = \sum_{i_2,\dots i_m=1}^{n} a_{ii_2\dots i_m}$, $R := \max_i r_i(\mathcal{A})$, $r := \min_j r_j(\mathcal{A})$.

Recently in [5, Theorem 3.2], by estimating the ratio of the largest component and the smallest component of a Perron vector, Wang et al. gave the following bound for γ of a weakly irreducible nonnegative tensor and proved it is better than the bound in (1).

Theorem 2.2. *Let $\mathcal{A}$ be a weakly irreducible nonnegative tensor of order m dimension n with the spectral radius $\rho(\mathcal{A})$ and the Perron vector x. Then*

$$\sqrt{\frac{R - \min(a_{i\dots i}, a_{j\dots j})}{r - \min(a_{i\dots i}, a_{j\dots j})}} \le \left(\max_{p,q} \frac{x_p}{x_q}\right)^{m-1}.$$

Now, we make a new lower bound of γ for the weakly irreducible nonnegative tensor, and show by example that the lower bound on γ is sharp.

Theorem 2.3. *Suppose that $\mathcal{A}$ is a weakly irreducible nonnegative tensor of order m dimension n. Let $T = \{t \in N : r_t(\mathcal{A}) < \rho(\mathcal{A})\}$ and $S = \{s \in N : r_s(\mathcal{A}) > \rho(\mathcal{A})\}$. Also, let $i \in S$ such that $r_i(\mathcal{A}) = R$ and $j \in T$ such that $r_j(\mathcal{A}) = r$. Then*

$$\gamma^{m-1} \ge \max\left(\frac{R - a_{ii\dots i} + \sum_{k\in S\setminus\{i\}} a_{ik\dots k} \frac{r_k(\mathcal{A}) - \rho(\mathcal{A})}{\rho(\mathcal{A}) - a_{kk\dots k}}}{\rho(\mathcal{A}) - a_{ii\dots i}}, \frac{\rho(\mathcal{A}) - a_{jj\dots j}}{r - a_{jj\dots j} - \sum_{k\in T\setminus\{j\}} a_{jk\dots k} \frac{\rho(\mathcal{A}) - r_k(\mathcal{A})}{\rho(\mathcal{A}) - a_{kk\dots k}}} \right).$$

Proof. Let $\lambda = \rho(\mathcal{A})$ be the Perron root with the Perron vector x, i.e.

$$\mathcal{A}x^{m-1} = \rho(\mathcal{A})x^{[m-1]}.$$

Since $\mathcal{A}$ is a weakly irreducible, x is positive from Theorem 1.5 (Perron-Frobenius theorem for weakly irreducible nonnegative tensor). By choosing $x_s = \max_i x_i$ and $x_t = \min_i x_i$, for each k, we have

$$\begin{aligned}(\rho(\mathcal{A}) - a_{kk\ldots k})x_k^{m-1} &= \sum_{\substack{i_2,\ldots i_m=1\\ \delta_{ki_2\ldots i_m}=0}}^{n} a_{ki_2\ldots i_m}x_{i_2}\ldots x_{i_m}\\ &\geq x_t^{m-1}\left(r_k\left(\mathcal{A}\right) - a_{kk\ldots k}\right) \gneqq 0\end{aligned}$$

where the final expression is positive since $\mathcal{A}$ is nonnegative weakly irreducible and x is positive. Thus, for each k,

$$\left(\frac{x_k}{x_t}\right)^{m-1} \geq \frac{r_k\left(\mathcal{A}\right) - a_{kk\ldots k}}{\rho(\mathcal{A}) - a_{kk\ldots k}}. \tag{2}$$

Since $R = r_i(\mathcal{A}) := \max_p r_p(\mathcal{A})$, then

$$R - \sum_{k\in S\setminus\{i\}} a_{ik\ldots k} = \sum_{\substack{i_2,\ldots i_m=1\\ \delta_{ki_2\ldots i_m}=0}}^{n} a_{ii_2\ldots i_m} + \sum_{k\notin S\setminus\{i\}} a_{ik\ldots k}. \tag{3}$$

Now by using (2) and (3), one has

$$\begin{aligned}(\rho(\mathcal{A}) - a_{ii\ldots i})x_i^{m-1} &= \sum_{\substack{i_2,\ldots i_m=1\\ \delta_{ii_2\ldots i_m}=0}}^{n} a_{ii_2\ldots i_m}x_{i_2}\ldots x_{i_m}\\ &\geq \sum_{k\in S\setminus\{i\}} a_{ik\ldots k}x_k^{m-1} + x_t^{m-1}\sum_{\substack{i_2,\ldots i_m=1\\ \delta_{ki_2\ldots i_m}=0\\ \delta_{ii_2\ldots i_m}=0}} a_{ii_2\ldots i_m} + x_t^{m-1}\sum_{k\notin S\setminus\{i\}} a_{ik\ldots k}\\ &\geq x_t^{m-1}\left(\sum_{k\in S\setminus\{i\}} a_{ik\ldots k}\frac{r_k(\mathcal{A}) - a_{kk\ldots k}}{\rho(\mathcal{A}) - a_{kk\ldots k}} + R - a_{ii\ldots i} - \sum_{k\in S\setminus\{i\}} a_{ik\ldots k}\right).\end{aligned}$$

Putting $r_k(\mathcal{A}) - a_{kk\ldots k} = (r_k(\mathcal{A}) - \rho(\mathcal{A})) + (\rho(\mathcal{A}) - a_{kk\ldots k})$ and simplifying, we obtain the first lower bound

$$\gamma^{m-1} \geq \left(\frac{x_i}{x_t}\right)^{m-1} \geq \frac{(R - a_{ii\ldots i}) + \sum_{k\in S\setminus\{i\}} a_{ik\ldots k}\frac{r_k(\mathcal{A})-\rho(\mathcal{A})}{\rho(\mathcal{A})-a_{kk\ldots k}}}{\rho(\mathcal{A}) - a_{ii\ldots i}}. \tag{4}$$

To obtain the second lower bound on γ, we note first that for each k,

$$\begin{aligned}0 < (\rho(\mathcal{A}) - a_{kk\ldots k})x_k^{m-1} &= \sum_{\substack{i_2,\ldots i_m=1\\ \delta_{ki_2\ldots i_m}=0}}^{n} a_{ki_2\ldots i_m}x_{i_2}\ldots x_{i_m}\\ &\leq (r_k(\mathcal{A}) - a_{kk\ldots k})x_s^{m-1}\end{aligned}$$

and so

$$\left(\frac{x_s}{x_k}\right)^{m-1} \geq \frac{\rho(\mathcal{A}) - a_{kk\ldots k}}{r_k(\mathcal{A}) - a_{kk\ldots k}} > 0. \tag{5}$$

Since $r = r_j(\mathcal{A}) := \min_p r_p(\mathcal{A})$, similarly by (5) we have

$$\begin{aligned}
(\rho(\mathcal{A}) - a_{jj\ldots j})\, x_j^{m-1} &= \sum_{\substack{i_2,\ldots i_m=1 \\ \delta_{ji_2\ldots i_m}=0}}^{n} a_{ji_2\ldots i_m} x_{i_2}\ldots x_{i_m} \\
&\leq \sum_{k\in T\setminus\{j\}} a_{jk\ldots k} x_k^{m-1} + x_s^{m-1} \sum_{k\notin T\setminus\{j\}} a_{jk\ldots k} + x_s^{m-1} \sum_{\substack{i_2,\ldots i_m=1 \\ \delta_{ki_2\ldots i_m}=0 \\ \delta_{ji_2\ldots i_m}=0}}^{n} a_{ji_2\ldots i_m} \\
&\leq x_s^{m-1} \left(\sum_{k\in T\setminus\{j\}} a_{jk\ldots k} \frac{r_k(\mathcal{A}) - a_{kk\ldots k}}{\rho(\mathcal{A}) - a_{kk\ldots k}} + r - a_{jj\ldots j} - \sum_{k\in T\setminus\{j\}} a_{jk\ldots k} \right).
\end{aligned}$$

Putting $r_k(\mathcal{A}) - a_{kk\ldots k} = (r_k(\mathcal{A}) - \rho(\mathcal{A})) + (\rho(\mathcal{A}) - a_{kk\ldots k})$ and simplifying, we obtain

$$\begin{aligned}
(\rho(\mathcal{A}) - a_{jj\ldots j})\, x_t^{m-1} &\leq (\rho(\mathcal{A}) - a_{jj\ldots j})\, x_j^{m-1} \\
&\leq x_s^{m-1} \left(\sum_{k\in T\setminus\{j\}} a_{jk\ldots k} \frac{r_k(\mathcal{A}) - \rho(\mathcal{A})}{\rho(\mathcal{A}) - a_{kk\ldots k}} + r - a_{jj\ldots j} \right).
\end{aligned}$$

The previous inequality implies that

$$\gamma^{m-1} = \left(\frac{x_s}{x_t}\right)^{m-1} \geq \frac{\rho(\mathcal{A}) - a_{jj\ldots j}}{\sum_{k\in T\setminus\{j\}} a_{jk\ldots k} \frac{r_k(\mathcal{A}) - \rho(\mathcal{A})}{\rho(\mathcal{A}) - a_{kk\ldots k}} + r - a_{jj\ldots j}}. \tag{6}$$

Finally, by (4) and (6), we have

$$\gamma^{m-1} \geq \max\left(\frac{R - a_{ii\ldots i} + \sum_{k\in S\setminus\{i\}} a_{ik\ldots k} \frac{r_k(\mathcal{A}) - \rho(\mathcal{A})}{\rho(\mathcal{A}) - a_{kk\ldots k}}}{\rho(\mathcal{A}) - a_{ii\ldots i}}, \frac{\rho(\mathcal{A}) - a_{jj\ldots j}}{r - a_{jj\ldots j} - \sum_{k\in T\setminus\{j\}} a_{jk\ldots k} \frac{\rho(\mathcal{A}) - r_k(\mathcal{A})}{\rho(\mathcal{A}) - a_{kk\ldots k}}} \right).$$

The proof is completed. □

We now show the efficiency of the new bounds in Theorem 2.3 by the following example which is considered in [5, Example 3.4].

Example 2.4. Let $\mathcal{A} = (a_{ijk})$ be an order 3 dimension 3 tensor with

$$a_{ijk} = \begin{cases}
a_{111} = 1\,; & a_{112} = 1\,; & a_{121} = 1\,; & a_{122} = 1\,; & a_{133} = 1\,; \\
a_{211} = 1\,; & a_{212} = 1\,; & a_{213} = \frac{3}{2}\,; & a_{221} = 1\,; & a_{231} = \frac{3}{2}\,; \quad a_{233} = 1\,; \\
a_{311} = 1\,; & a_{313} = 3\,; & a_{322} = 1\,; & a_{331} = 3\,; & a_{333} = 1\,; \\
a_{ijk} = 0\,, & \textit{otherwise}.
\end{cases}$$

For this tensor, it can be verified that

$$(\rho(\mathcal{A}), x) = (6.6575, (0.5756, 0.6826, 0.7890))\,.$$

That is, the Perron root is 6.6575 and the Perron vector is $(0.5756, 0.6826, 0.7890)^T$. We compute the lower bound of γ, (i.e. the ratio of the largest and smallest entries in a Perron vector) for $\mathcal{A}$ given by Theorem 2.1, Theorem 2.2, and Theorem 2.3

$$\begin{aligned}
&\textit{Thoerem } 2.1: && \gamma^2 \geq \sqrt{\frac{9}{5}} \approx 1.34164078649.\\
&\textit{Thoerem } 2.2: && \gamma^2 \geq \sqrt{2} \approx 1.41421356237.\\
&\textit{Thoerem } 2.3: && \gamma^2 \geq 1.42314551266.\\
&\textit{Actual value}: && \gamma^2 = 1.87893793997.
\end{aligned}$$

This example shows that the lower bound of Theorem 2.3 is better than those of Theorem 2.1 in [4] and Theorem 2.2 in [5].

3 Conclusion

Here, a new ratio of the largest and smallest values of the Perron vector for the weakly irreducible nonnegative tensors is presented. We demonstrated that the lower bound is sharper than the conclusions of [4, 5] by a running example.

References

[1] K.C. Chang, P. Kelly and Z. Tan, Perron–Frobenius theorem for nonnegative tensors, *Communications in Mathematical Sciences*, 6(2008), No. 2, 507–520.

[2] S. Friedland, G. Stphane and H. Lixing, Perron–Frobenius theorem for nonnegative multilinear forms and extensions, *Linear Algebra and its Applications*, 438 (2013), No. 2, 738–749.

[3] L. Qi and L. Ziyan, *Tensor analysis: spectral theory and special tensors*, Vol. 151, Siam, 2017.

[4] Z. Wang and W. Wei, Bounds for the greatest eigenvalue of positive tensors, *J. Ind. Manage. Optim*, 10 (2014), 1031–1039.

[5] G. Wang, W. Yanan and W. Yiju, Some Ostrowski-type bound estimations of spectral radius for weakly irreducible nonnegative tensors, *Linear and Multilinear Algebra*, (2018), 1–18.

[6] S. Wasserman and F. Katherine, *Social Network Analysis. Structural Analysis in the Social Sciences*, Cambridge University Press,(1994).

A new projection method for solving large Sylvester equations[1]

Faezeh Toutounian[1], Zahra Asgari[2,*] and Esmail Babolian[2]

[1]Department of Applied Mathematics, Ferdowsi University of Mashhad, Mashhad, Iran

[2]Department of Mathematics, Kharazmi University, Tehran, Iran

Abstract

In this paper, we propose a new projection method to solve large Sylvester matrix equations. The new approach projects the problem onto extended global Krylov subspace and gets a low dimensional equation. We use the global Golub-Kahan bidiagonalization procedure to construct the F-orthonormal basis for the extended Krylov subspaces. Finally, we give some theoretical results and present numerical experiments.

Keywords: Matrix equations, Golub-Kahan bidiagonalization, Extended global Krylov subspace

Mathematics Subject Classification [2010]: 65F15, 65F10

1 Introduction

In this paper we will consider the Sylvester matrix equation of the form

$$AX + XB + CD^T = 0, \tag{1}$$

where $A \in \mathbb{R}^{N\times N}$ is assumed to be large, $B \in \mathbb{R}^{M\times M}$, $C \in \mathbb{R}^{N\times s}$, $D \in \mathbb{R}^{M\times s}$ and $X \in \mathbb{R}^{N\times M}$ is unknown matrix for equation (1). The matrix equation (1) plays the fundamental role in many areas such as control and communications theory. Direct methods for solving the matrix equation (1), are attractive if the matrices are of small size. These methods are based on the Schur decomposition, by which the original equation is transformed into a form that is easy to be solved by a forward substitution. Iterative projection methods for solving large Sylvester matrix equations have been developed during the past years.

In this paper we present a new projection method that projects the initial problem onto an extended global Krylov subspace. The new projection method builds the F-orthonormal basis of enriched global Krylov subspaces and allows us to compute low rank approximations to the solution of (1). The extended global Krylov subspaces are generated by means of the new extended global Golub and Kahan procedure. We mention that the Golub and Kahan process first introduced in [1]. In [4], the authors defined the global bidiagonalization based on Golub and Kahan procedure.

[1]Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

*Speaker. Email address: za.asgari93@gmail.com

The outline of this paper is as follows. In Section 2, we give a quick overview of the global Golub-Kahan procedure and its properties. In Section 3, we present the extended version of global Golub-Kahan procedure and its properties. In Section 4, we show how to apply the extended global Golub-Kahan procedure to obtain low rank approximate solutions to the Sylvester equation (1). Section 5 is devoted to some numerical experimants. Finally, we make some concluding remarks in Section 6.

2 The global Golub-Kahan procedure

In this section, we present a brief description of the Global Bidiag 1 algorithm [4]. This algorithm is the basis for the extended global Golub-Kahan procedure.

The global Bidiag 1 procedure constructs the sets of the $n\times p$ block vectors $V_1, V_2, \ldots, V_k$ and $U_1, U_2, \ldots, U_k$ such that $\langle V_i^T, V_j\rangle_F = 0, \langle U_i^T, U_j\rangle_F = 0$, for $i \neq j$, and $\|V_i\|_F = \|U_i\|_F = 1$ and after k steps they form the F-orthonormal bases of $\mathbb{R}^{n\times kp}$.

Global Bidiag 1 (*Starting matrix G; reduction to lower bidiagonal form*)

$$\beta_1 U_1 = G, \quad \alpha_1 V_1 = A^T U_1, \tag{2}$$

$$\left.\begin{array}{l} \beta_{i+1}U_{i+1} = AV_i - \alpha_i U_i, \\ \alpha_{i+1}V_{i+1} = A^T U_{i+1} - \beta_{i+1}V_i, \end{array}\right\} \quad i = 1, 2, \ldots k, \tag{3}$$

The scalars $\alpha_i \geq 0$ and $\beta_i \geq 0$ are chosen so that $\|U_i\|_F = \|V_i\|_F = 1$. With the definitions

$$\overline{U}_k \equiv [U_1, U_2, \ldots, U_k], \qquad \overline{V}_k \equiv [V_1, V_2, \ldots, V_k], \qquad T_k \equiv \begin{bmatrix} \alpha_1 & & & & \\ \beta_2 & \alpha_2 & & & \\ & \ddots & \ddots & & \\ & & \beta_k & \alpha_k & \\ & & & \beta_{k+1} & \end{bmatrix}, \tag{4}$$

and using the Kronecker product $\otimes$, the recurrence relations (2) and (3) may be rewritten as:

$$\begin{aligned} \overline{U}_{k+1}(\beta_1 e_1 \otimes I_p) &= G, \\ A^{-T}\overline{V}_k &= \overline{U}_{k+1}(T_k \otimes I_p), \\ A^{-1}\overline{U}_{k+1} &= \overline{V}_k(T_k^T \otimes I_p) + \alpha_{k+1}V_{k+1}(e_{k+1}^T \otimes I_p), \end{aligned} \tag{5}$$

where e_j is the jth column of the identity matrix. We have $\langle U_i, U_j\rangle_F = \langle V_i, V_j\rangle_F = 0$ for $i \neq j$ and $\|U_i\|_F = \|V_i\|_F = 1$. We can easily show that $[U_1, U_2, \cdots, U_k]$ and $[V_1, V_2, \cdots, V_k]$ are the F-orthonormal basis of the subspaces $\mathcal{K}_k(AA^T, U_1)$ and $\mathcal{K}_k(A^TA, V_1)$, respectively. More details about the global Golub-Kahan process can be found in [4].

3 The extended global Golub-Kahan process

In this section we present the extended version of global Golub-Kahan procedure applied to the pair (A, C) where the matrix $A \in \mathbb{R}^{n\times n}$ is assumed to be nonsingular and $C \in \mathbb{R}^{n\times p}$. The algorithm proceeds by first running k steps of the global Golub-Kahan process [4] with A^{-T}, and then continuing with m iterations of the global Golub-Kahan process with A, while maintaining F-orthogonalization among all generated vectors in the sequence. By performing k steps of the Golub-Kahan procedure to the pair (A^{-T}, C), we have

$$\beta_1 U_1 = C, \quad \alpha_1 V_1 = A^{-1} U_1, \tag{6}$$

$$\left.\begin{array}{l}\beta_{i+1}U_{i+1} = A^{-T}V_i - \alpha_i U_i,\\ \alpha_{i+1}V_{i+1} = A^{-1}U_{i+1} - \beta_{i+1}V_i,\end{array}\right\} \quad i = 1, 2, \ldots, k, \tag{7}$$

where the scalars $\alpha_i \geq 0$, and $\beta_i \geq 0$ are chosen so that $\|U_i\|_F = \|V_i\|_F = 1$. With the definitions

$$\overline{U}_k \equiv [U_1, U_2, \ldots, U_k], \qquad \overline{V}_k \equiv [V_1, V_2, \ldots, V_k], \qquad T_k \equiv \begin{bmatrix} \alpha_1 & & & \\ \beta_2 & \alpha_2 & & \\ & \ddots & \ddots & \\ & & \beta_k & \alpha_k \\ & & & \beta_{k+1} \end{bmatrix}, \tag{8}$$

and using the Kronecker product $\otimes$, the recurrence relations (6) and (7) may be rewritten as:

$$\begin{array}{rl} \overline{U}_{k+1}(\beta_1 e_1 \otimes I_p) = & C, \\ A^{-T}\overline{V}_k = & \overline{U}_{k+1}(T_k \otimes I_p), \\ A^{-1}\overline{U}_{k+1} = & \overline{V}_k(T_k^T \otimes I_p) + \alpha_{k+1}V_{k+1}(e_{k+1}^T \otimes I_p), \\ A^{-1}\overline{U}_k = & \overline{V}_k(\overline{T}_k^T \otimes I_p), \end{array} \tag{9}$$

where $\overline{T}_k$ obtained from T_k by deleting its last row and e_j is the jth column identity matrix. We have $\langle U_i, U_j\rangle_F = \langle V_i, V_j\rangle_F = 0$ for $i \neq j$ and $\|U_i\|_F = \|V_i\|_F = 1$. We can easily show that $[U_1, U_2, \cdots, U_k]$ and $[V_1, V_2, \cdots, V_k]$ are the F-orthonormal basis of the subspaces $\mathcal{K}_k((AA^T)^{-1}, C)$ and $\mathcal{K}_k((A^TA)^{-1}, A^{-1}C)$, respectively. Now we again use the global Golub and Kahan bidiagonalization applied to the pair (A, U_1) in order to construct the matrices $Q_1, Q_2, \cdots, Q_m$ and $P_1, P_2, \cdots, P_{m+1}$ such that

$$\mathcal{U}_{k+1,m} = [U_1, U_2, \cdots, U_{k+1}, Q_1, Q_2, \cdots, Q_m] \quad \text{and} \quad \mathcal{V}_{k,m+1} = [V_1, V_2, \cdots, V_k, P_1, P_2, \cdots, P_{m+1}]$$

form the F-orthonormal basis of the subspaces

$$\begin{array}{l} \mathcal{K}^e_{k+1,m}(AA^T, U_1) = \text{span}\{(AA^T)^{-k}U_1, \ldots, (AA^T)^{-1}U_1, U_1, (AA^T)U_1, \ldots, (AA^T)^{m-1}U_1\}, \\ \mathcal{K}^e_{k,m+1}(A^TA, V_1) = \text{span}\{(A^TA)^{-k+1}V_1, \ldots, (A^TA)^{-1}V_1, A^TC, (A^TA)V_1, \ldots, (A^TA)^mV_1\}, \end{array}$$

respectively. In order to have the F-orthonormal basis $\mathcal{U}_{k+1,m}, \mathcal{V}_{k,m+1}$, first we generate the matrix P_1 satisfying

$$\tilde{\alpha}_1 P_1 = A^T U_1 - \sum_{i=1}^{k} h_{i1} V_i, \tag{10}$$

where the scalars $\tilde{\alpha}_1 \geq 0$ and $h_{i1}, i = 1, 2, \cdots, k$, are chosen so that $\langle P_1, V_i\rangle_F = 0$ and $\|P_1\|_F = 1$. Then, we generate the matrix Q_1 satisfying

$$\tilde{\beta}_1 Q_1 = AP_1 - \tilde{\alpha}_1 U_1 - \sum_{i=2}^{k+1} g_{i1} U_i, \tag{11}$$

where the scalars $\tilde{\beta}_1 \geq 0$ and g_{i1} are chosen so that $\langle Q_1, U_i\rangle_F = 0$ and $\|Q_1\|_F = 1$. Now we construct $Q_2, Q_3, \ldots, Q_{m+1}$ and $P_2, P_3, \ldots, P_{m+1}$ with the recurrence relations:

$$\left.\begin{array}{l}\tilde{\alpha}_i P_i = A^T Q_{i-1} - \tilde{\beta}_{i-1} P_{i-1},\\ \tilde{\beta}_i Q_i = AP_i - \tilde{\alpha}_i Q_{i-1},\end{array}\right\} \quad i = 2, 3, \ldots m+1. \tag{12}$$

With the definitions $g_{11} = \tilde{\alpha}_1$ and

$$\overline{Q}_m \equiv [Q_1, Q_2, \cdots Q_m], \quad \overline{P}_m \equiv [P_1, P_2, \cdots P_m], \quad \tilde{T}_m \equiv \begin{bmatrix} \tilde{\beta}_1 & & & \\ \tilde{\alpha}_2 & \tilde{\beta}_2 & & \\ & \ddots & \ddots & \\ & & \tilde{\alpha}_m & \tilde{\beta}_m \\ & & & \tilde{\alpha}_{m+1} \end{bmatrix},$$

the recurrence relations (12) may be rewritten as

$$\begin{aligned} A^T\overline{Q}_m &= \overline{P}_{m+1}(\tilde{T}_m \otimes I_p), \\ A\overline{P}_m &= \overline{Q}_m(\overline{\tilde{T}}_m^T \otimes I_p) + \textstyle\sum_{i=1}^{k+1} g_{i1}U_i(e_1^T \otimes I_p), \end{aligned} \tag{13}$$

where $\overline{\tilde{T}}_m$ is the matrix obtained from $\tilde{T}_m$ by deleting its last row.

The main steps of the extended global Golub-Kahan algorithm to generate $\mathcal{U}_{k+1,m}$ and $\mathcal{V}_{k,m+1}$ may be summarized as follows.

Algorithm 3.1. The extended global Golub-Kahan algorithm

1. Inputs: $A \in \mathbb{R}^{n\times n}, C \in \mathbb{R}^{n\times p}, k$, and m.
2. Compute $\beta_1 = \|C\|_F, \quad U_1 = C/\beta_1, \quad \alpha_1 = \|A^{-1}U_1\|_F, \quad V_1 = (A^{-1}U_1)/\alpha_1,$
3. For $i = 1, \ldots, k,$
   compute $W = A^{-T}V_i - \alpha_i U_i, \quad \beta_{i+1} = \|W\|_F, \quad U_{i+1} = W/\beta_{i+1},$
   compute $W = A^{-1}U_{i+1} - \beta_{i+1}V_i, \quad \alpha_{i+1} = \|W\|_F, \quad V_{i+1} = W/\alpha_{i+1},$
   End For.
4. Compute $W = A^T U_1,$
   For $i = 1, \ldots, k,$ compute $h_{i1} = \langle V_i, W\rangle_F, \quad W = W - h_{i1}V_i,$ End For.
   Compute $\tilde{\alpha}_1 = \|W\|_F, \quad P_1 = W/\tilde{\alpha}_1.$
5. Compute $W = AP_1 - \tilde{\alpha}_1 U_1,$
   For $i = 2, \ldots, k+1,$ compute $g_{i1} = \langle U_i, W\rangle_F, \quad W = W - g_{i1}U_i, \quad$ End For.
   Compute $\tilde{\beta}_1 = \|W\|_F, \quad Q_1 = W/\tilde{\beta}_1,$
6. For $i = 2, \ldots, m+1,$
   compute $W = A^T Q_{i-1} - \tilde{\beta}_{i-1}P_{i-1}, \quad \tilde{\alpha}_i = \|W\|_F, \quad P_i = W/\tilde{\alpha}_i,$
   compute $W = AP_i - \tilde{\alpha}_i Q_{i-1}, \quad \tilde{\beta}_i = \|W\|_F, \quad Q_i = W/\tilde{\beta}_i$
   End For.

In implementation of Algorithm 3.1, instead of using the matrix-vector products with A^{-1}, we use the LU-decomposition of A for computing V_1 and W in steps 2 and 3.

For the extended global Golub-Kahan Algorithm, we have the following proposition.

Proposition 3.2. *Suppose that (k,m) steps of Algorithm 1 have been carried out. Let*

$$F_{k+1} = \begin{bmatrix} 1 & \alpha_1\beta_2^{-1} & & & \\ & 1 & \alpha_2\beta_3^{-1} & & \\ & & 1 & \ddots & \\ & & & \ddots & \alpha_k\beta_{k+1}^{-1} \\ & & & & 1 \end{bmatrix}, \quad J_k = \begin{bmatrix} h_{11} & \beta_2^{-1} & & & \\ h_{21} & & \beta_3^{-1} & & \\ \vdots & & & \ddots & \\ h_{k1} & & & & \beta_{k+1}^{-1} \end{bmatrix}.$$

Then we have

$$A^T\mathcal{U}_{k+1,m} = \mathcal{V}_{k,m+1}(\mathcal{F}_{k+1,m} \otimes I_p), \qquad \text{with } \mathcal{F}_{k+1,m} = \left[\begin{array}{c|c} J_k F_{k+1}^{-1} & 0_{k\times m} \\ \hline \tilde{\alpha}_1 e_1^T F_{k+1}^{-1} & \tilde{\beta}_1 e_1^T \\ \hline 0_{m\times(k+1)} & \underline{\tilde{T}}_m \end{array}\right], \tag{14}$$

where $\underline{\tilde{T}}_m$ *is the matrix obtained from* $\tilde{T}_m$ *by deleting its first row and* $e_1^T = [1, 0, \ldots, 0] \in \mathbb{R}^{1\times m}$.

Proof. From (10) and (7), we have

$$\begin{aligned} &A^T U_1 = \tilde{\alpha}_1 P_1 + \textstyle\sum_{i=1}^{k} h_{i1} V_i, \\ &A^T U_{i+1} + \alpha_i \beta_{i+1}^{-1} A^T U_i = \beta_{i+1}^{-1} V_i, \qquad \text{for } i = 1, \ldots, k. \end{aligned}$$

By using the definition of matrices F_{k+1} and J_k, these equations can be written as follows:

$$A^T U_{k+1} = [\overline{V}_k, P_1] \left(\begin{bmatrix} J_k \\ \tilde{\alpha}_1 e_1^T \end{bmatrix} \otimes I_p \right) (F_{k+1} \otimes I_p)^{-1} = [\overline{V}_k, P_1] \left(\begin{bmatrix} J_k F_{k+1}^{-1} \\ \tilde{\alpha}_1 e_1^T F_{k+1}^{-1} \end{bmatrix} \otimes I_p \right).$$

This together with the first relation of (13) implies the desired relation (14). □

4 Low rank approximation solution to the Sylvester equation

As in [2], we define the $\diamond$-product $A^T \diamond B$ of the matrices A and B. The matrix $A = [A_1, A_2, \ldots, A_r]$ is F-orthonormal if and only if $A^T \diamond A = I_r$. Using the $\diamond$-product we have $\mathcal{U}_{k+1,m}^T \diamond \mathcal{U}_{k+1,m} = I_{m+k+1}$ and $\mathcal{V}_{k,m+1}^T \diamond \mathcal{V}_{k,m+1} = I_{m+k+1}$.

Now, we use the extended global Golub-Kahan Algorithm 1 to extract low rank approximate solution to the equation (1). Let $U_1^A = C/\|C\|_F$, $U_1^B = D/\|D\|_F$. Applying the extended global Golub-Kahan Algorithm 1 to the pairs (A^T, C) and (B^T, D) gives us the F-orthonormal basis $\mathcal{U}_{k+1,m}^A = [U_1^A, \cdots, U_{k+1}^A, Q_1^A, \cdots, Q_m^A]$,$\mathcal{V}_{k,m+1}^A = [V_1^A, \cdots, V_k^A, P_1^A, \cdots, P_{m+1}^A]$, $\mathcal{U}_{k+1,m}^B = [U_1^B, \cdots, U_{k+1}^B, Q_1^B, \cdots, Q_m^B]$ and $\mathcal{V}_{k,m+1}^B = [V_1^B, \cdots, V_k^B, P_1^B, \cdots, P_{m+1}^B]$ of the extended global Krylov subspaces $\mathcal{K}_{k+1,m}^e(AA^T, C)$, $\mathcal{K}_{k,m+1}^e(A^TA, A^TC)$, $\mathcal{K}_{k+1,m}^e(BB^T, D)$ and $\mathcal{K}_{k,m+1}^e(B^TB, B^TD)$, respectively. In addition, by using Proposition 2, we can define the matrices

$$\begin{aligned} \mathcal{T}_{k+1,m}^A &= (\mathcal{U}_{k+1,m}^A)^T \diamond A\mathcal{U}_{k+1,m}^A = ((\mathcal{U}_{k+1,m}^A)^T \diamond \mathcal{V}_{k,m+1}^A)\mathcal{F}_{k+1,m}^A, \\ \mathcal{T}_{k+1,m}^B &= (\mathcal{U}_{k+1,m}^B)^T \diamond B\mathcal{U}_{k+1,m}^B = ((\mathcal{U}_{k+1,m}^B)^T \diamond \mathcal{V}_{k,m+1}^B)\mathcal{F}_{k+1,m}^B, \end{aligned}$$

where the matrices $\mathcal{F}_{k+1,m}^A$ and $\mathcal{F}_{k+1,m}^B$ can be obtained through the Algorithm 1. Using the F-orthonormal basis $\mathcal{U}_{k+1,m}^A$ and $\mathcal{U}_{k+1,m}^B$, as in [3], we look for low-rank approximate solution that have the form

$$X_{k+1,m} = \mathcal{U}_{k+1,m}^A \, (Y_{k+1,m} \otimes I_p) \, (\mathcal{U}_{k+1,m}^B)^T, \tag{15}$$

where $Y_{k+1,m} \in \mathbb{R}^{(k+1+m)\times(k+1+m)}$. Let $R_{k+1,m} = AX_{k+1,m} + X_{k+1,m}B + CD^T$, be the residual associated with the approximate solution $X_{k+1,m}$. By incorporating (15) in $R_{k+1,m}$ and using the equation (14), we get

$$R_{k+1,m} = \mathcal{V}_{k,m+1}^A \, (\mathcal{F}_{k+1,m}^A \otimes I_p) \, (Y_{k+1,m} \otimes I_p) \, (\mathcal{U}_{k+1,m}^B)^T$$

$$+\mathcal{U}^A_{k+1,m}\left(Y_{k+1,m}\otimes I_p\right)\left((\mathcal{F}^B_{k+1,m})^T\otimes I_p\right)(\mathcal{V}^B_{k,m+1})^T+CD^T.$$

The matrix $Y_{k+1,m}$ can be obtained by imposing the orthogonality condition

$$(\mathcal{U}^A_{k+1,m})^T\diamond R_{k+1,m}\diamond\mathcal{U}^B_{k+1,m}=0. \tag{16}$$

Using (16) and the fact that $\mathcal{U}^A_{k+1,m}$ and $\mathcal{U}^B_{k+1,m}$ are F-orthonormal, we have

$$\begin{aligned}0&=(\mathcal{U}^A_{k+1,m})^T\diamond R_{k+1,m}\diamond\mathcal{U}^B_{k+1,m}\\&=\left((\mathcal{U}^A_{k+1,m})^T\diamond\mathcal{V}^A_{k,m+1}\right)\mathcal{F}^A_{k+1,m}\left((\mathcal{U}^B_{k+1,m})^T\diamond\mathcal{U}^B_{k+1,m}\right)Y_{k+1,m}\\&\quad+\left((\mathcal{U}^A_{k+1,m})^T\diamond\mathcal{U}^A_{k+1,m}\right)Y_{k+1,m}\left(\left((\mathcal{U}^B_{k+1,m})^T\diamond\mathcal{V}^B_{k,m+1}\right)\mathcal{F}^B_{k+1,m}\right)^T+(\mathcal{U}^A_{k+1,m})^T\diamond CD^T\diamond\mathcal{U}^B_{k+1,m}\\&=\mathcal{T}^A_{k+1,m}Y_{k+1,m}+Y_{k+1,m}(\mathcal{T}^B_{k+1,m})^T+\beta^A_1\beta^B_1e_1e_1^T,\end{aligned} \tag{17}$$

where e_1 is the first column of the identity matrix of order $m+k+1$. Assuming that the matrices $\mathcal{T}^A_{k+1,m}$ and $\mathcal{T}^B_{k+1,m}$ have no eigenvalue in common, then the unique solution of the low-dimentional Sylvester equation (17) can be obtained by applying a direct solver such as the Hessenberg-Schur method.

5 Numerical results

In this section we report some numerical results obtained by executing the new method for computing the solution of the equation (1). The stop criterion is $\|R_k\|_F/\|R_0\|_F\leq 10e-8$, where $R_0=CD^T$ is the initial residual matrix and R_k is the kth residual matrix.

Example 5.1. In this example, we use the matrices $A=tridiag(-1+10/(N+1),2,-1+10/(N+1))$, and $B=tridiag(-1+10/(M+1),2,-1+10/(M+1))$, where N and M are the order of matrices A and B, respectively. The entries of the matices C and D are random values uniformly distributed on $[0,1]$. The results were reported in Table 1.

Table 1: Numerical results for Example 1.

(N,M)	k,m	Residual norm
(800, 500)	$k=40,\ m=60$	8.1933e-06
(1000,500)	$k=40,\ m=60$	8.1933e-06
(1500,800)	$k=40,\ m=100$	2.7758e-05

6 Conclusion

In this paper, we have described the extended version of global Golub-Kahan procedure. By using this procedure, we have presented a new projection method for computing low rank approximate solutions for Sylvester matrix equations. Finally, some numerical experiments were given in order to show the efficiency of the proposed method.

References

[1] G. H. Golub, W. Kahan, Algorithm LSQR is based on the Lanczos process and bidiagonalization procedure, *SIAM J. Numer. Anal.* 2 (1965) 205-224.

[2] M. Heyouni, Extended Arnoldi methods for large low- rank Sylvester matrix equations, *Applied Numerical Mathematics*, 60 (2010) 1171-1182.

[3] L. Bio, Y. Lin, Y. Wei, A new projection method for solving large Sylvester equations, *Applied Numerical Mathematics*, 57 (2007) 521-532.

[4] F. Toutounian, S. Karimi, Global least squares method (Gl-LSQR) for solving general linear systems with several right hand sides, *Applied Mathematics and Computation*, 178 (2006) 452-460.

Numerical solution of a class of fractional optimal control problems using the fractional-order Bernoulli wavelet functions[1]

Forugh Valian[1,*], Mohammad Ali Vali[1,] and Yadollah Ordokhani[2]

[1]Department of Mathematics, Faculty of Mathematics and Computer, Shahid Bahonar University of Kerman, Iran

[2]Department of Mathematics, Faculty of Mathematical Sciences, Alzahra University, Tehran, Iran

Abstract

In this paper, an efficient method based on fractional-order Bernoulli wavelet functions (FBWFs) is presented for the numerical solution of a class of the fractional optimal control problems (FOCPs). To solve the problem, first the FOCP is transformed into an equivalent variational problem, then with the aid of FBWFs, operational matrix of RiemannLiouville fractional integration and Gauss quadrature formula, the problem is solved approximately.

Keywords: Fractional optimal control problems, Fractional-order Bernoulli wavelets, Operational matrix

Mathematics Subject Classification [2010]: 15A03, 15A23, 15B36

1 Introduction

In the last decades, many numerical techniques have been developed for solving the fractional optimal control problems. A fractional optimal control problem is a generalization that requires minimizing a performance index governed by a fractional differential equations. The fractional optimal control problems have been applied in transportation, electronic, chemical and biological systems. Because of its importance, the numerical solution of the FOCPs was investigated. In this paper, we consider a class of the fractional optimal control problem and solve this problem.

2 Main results

2.1 Definitions and mathematical preliminaries

Definition 2.1. The Caputo fractional derivative of order ν, when $q-1<\nu\leq q$, of $f(t)$ is defined by [1]

$$ {}_0^C D_t^\nu f(t) = \begin{cases} \dfrac{1}{\Gamma(q-\nu)} \displaystyle\int_0^t \frac{f^{(q)}(s)}{(t-s)^{(\nu+1-q)}} ds, & q-1<\nu<q, \quad q\in\mathbb{N}, \\ \dfrac{d^q f(t)}{dt^q}, & \nu=q, \end{cases} \tag{1} $$

[1]Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

*Speaker. Email address: F.valian@math.uk.ac.ir

where $\Gamma(\cdot)$ denotes the gamma function.

Definition 2.2. The Riemann-Liouville fractional integral operator of order $\nu \geq 0$ of $f(t)$ is defined by [2]

$$I_t^\nu f(t) = \begin{cases} \frac{1}{\Gamma(\nu)} \int_0^t \frac{f(s)}{(t-s)^{1-\nu}} ds, & \nu > 0, \quad t \geq 0, \\ f(t), & \nu = 0. \end{cases} \tag{2}$$

The useful relation between the Riemann-Liouville operator and Caputo operator is given by the following expression

$$I_t^\nu \, {}_0^C D_t^\nu f(t) = f(t) - \sum_{i=0}^{n-1} f^{(i)}(0) \frac{t^i}{i!}, \quad t \geq 0, \quad n-1 < \nu \leq n, \tag{3}$$

where n is an integer, and $f \in C_1^n$.

2.2 Fractional-order Bernoulli wavelets

The fractional-order Bernoulli wavelets of order $\alpha \in \mathbb{R}_+$, $\psi_{n,m}^\alpha$, $n = 1, 2, \ldots, 2^{k-1}, m = 0, 1, \ldots, M$, on the interval [0,1) defined by [3]

$$\psi_{n,m}^\alpha(t) = \begin{cases} 2^{\frac{k-1}{2}} \tilde{\beta}_m(2^{k-1}t^\alpha - n + 1), & \frac{n-1}{2^{k-1}} \leq t^\alpha < \frac{n}{2^{k-1}}, \\ 0, & otherwise, \end{cases} \tag{4}$$

that k can assume any positive integer and

$$\tilde{\beta}_m(2^{k-1}t^\alpha - n + 1) = \begin{cases} 1, & m = 0, \\ \dfrac{1}{\sqrt{\frac{(-1)^{(m-1)}(m!)^2}{(2m)!}\beta_{2m}}} \beta_m(2^{k-1}t^\alpha - n + 1), & m > 0, \end{cases} \tag{5}$$

where $\beta_m(t)$ are Bernoulli polynomials of order m on $[0, 1]$.

2.3 The Functions approximation

An arbitrary function $f(t)$ which is square integrable in the interval $[0, 1]$ can be expanded by FBWFs as [3]

$$f(t) = \sum_{n=1}^{\infty} \sum_{m=0}^{\infty} c_{n,m} \psi_{n,m}^\alpha(t). \tag{6}$$

The infinite series in Eq. (6) is truncated to approximate $f(t)$ in terms of the FBWFs as

$$f(t) \simeq \sum_{n=1}^{2^{k-1}} \sum_{m=0}^{M} c_{n,m} \psi_{n,m}^\alpha(t) = C^T \Psi^\alpha(t), \tag{7}$$

where the unknown vector C and $\Psi^\alpha(t)$ are $2^{k-1}(M+1)$ column vectors and given by [3]

$$\begin{aligned} C &= [c_{1,0}, c_{1,1}, \ldots, c_{1,M}, c_{2,0}, c_{2,1}, \ldots, c_{2,M}, \ldots, c_{2^{k-1},0}, c_{2^{k-1},1}, \ldots, c_{2^{k-1},M}]^T, \\ \Psi^\alpha(t) &= [\psi_{1,0}^\alpha(t), \psi_{1,1}^\alpha(t), \ldots, \psi_{1,M}^\alpha(t), \psi_{2,0}^\alpha(t), \psi_{2,1}^\alpha(t), \ldots, \psi_{2,M}^\alpha(t), \ldots, \\ &\quad \psi_{2^{k-1},0}^\alpha(t), \psi_{2^{k-1},1}^\alpha(t), \ldots, \psi_{2^{k-1},M}^\alpha(t)]^T, \end{aligned} \tag{8}$$

and

$$\begin{aligned} &C^T = F^T D^{-1}, \\ &D = \langle \Psi^\alpha, \Psi^\alpha \rangle = \int_0^1 \Psi^\alpha(t) \Psi^{\alpha \mathrm{T}}(t) t^{\alpha-1} dt, \\ &F = [f_{1,0}, f_{1,1}, \ldots, f_{1,M}, f_{2,0}, f_{2,1}, \ldots, f_{2,M}, \ldots, f_{2^{k-1},0}, f_{2^{k-1},1}, \ldots, f_{2^{k-1},M}]^T, \\ &f_{ij} = \langle f, \psi_{ij}^\alpha \rangle = \int_0^1 f(t), \psi_{ij}^\alpha(t) t^{\alpha-1} dt, \quad i = 1, 2, \ldots, 2^{k-1}, \quad j = 1, 2, \ldots, M. \end{aligned}$$

2.4 The operational matrix of fractional integration

The Riemann-liouville fractional integral of the vector $\Psi^\alpha(t)$ defined in Eq. (8) can be obtained as

$$I_t^\nu \Psi^\alpha(t) \simeq P^{(\nu,\alpha)} \Psi^\alpha(t), \tag{9}$$

where $P^{(\nu,\alpha)}$ denotes the $2^{k-1}(M+1) \times 2^{k-1}(M+1)$ operational matrix for Riemann-liouville fractional integration defined by

$$I_t^\nu \Psi^\alpha(t) = \begin{bmatrix} I_t^\nu \psi_{1,0}^\alpha(t) \\ I_t^\nu \psi_{1,1}^\alpha(t) \\ \vdots \\ I_t^\nu \psi_{1,M}^\alpha(t) \\ I_t^\nu \psi_{2,0}^\alpha(t) \\ I_t^\nu \psi_{2,1}^\alpha(t) \\ \vdots \\ I_t^\nu \psi_{2,M}^\alpha(t) \\ \vdots \\ I_t^\nu \psi_{2^{k-1},0}^\alpha(t) \\ \psi_{2^{k-1},1}^\alpha(t) \\ \vdots \\ \psi_{2^{k-1},M}^\alpha(t) \end{bmatrix} \simeq \begin{bmatrix} \sum_{n=1}^{2^{k-1}} \sum_{m=0}^{M} E_{n,m}^{1,0} \psi_{n,m}^\alpha(t) \\ \sum_{n=1}^{2^{k-1}} \sum_{m=0}^{M} E_{n,m}^{1,1} \psi_{n,m}^\alpha(t) \\ \vdots \\ \sum_{n=1}^{2^{k-1}} \sum_{m=0}^{M} E_{n,m}^{1,M} \psi_{n,m}^\alpha(t) \\ \sum_{n=1}^{2^{k-1}} \sum_{m=0}^{M} E_{n,m}^{2,0} \psi_{n,m}^\alpha(t) \\ \sum_{n=1}^{2^{k-1}} \sum_{m=0}^{M} E_{n,m}^{2,1} \psi_{n,m}^\alpha(t) \\ \vdots \\ \sum_{n=1}^{2^{k-1}} \sum_{m=0}^{M} E_{n,m}^{2,M} \psi_{n,m}^\alpha(t) \\ \vdots \\ \sum_{n=1}^{2^{k-1}} \sum_{m=0}^{M} E_{n,m}^{2^{k-1},0} \psi_{n,m}^\alpha(t) \\ \sum_{n=1}^{2^{k-1}} \sum_{m=0}^{M} E_{n,m}^{2^{k-1},1} \psi_{n,m}^\alpha(t) \\ \vdots \\ \sum_{n=1}^{2^{k-1}} \sum_{m=0}^{M} E_{n,m}^{2^{k-1},M} \psi_{n,m}^\alpha(t) \end{bmatrix} = \begin{bmatrix} (E^{1,0})^T \Psi^\alpha(t) \\ (E^{1,1})^T \Psi^\alpha(t) \\ \vdots \\ (E^{1,M})^T \Psi^\alpha(t) \\ (E^{2,0})^T \Psi^\alpha(t) \\ (E^{2,1})^T \Psi^\alpha(t) \\ \vdots \\ (E^{2,M})^T \Psi^\alpha(t) \\ \vdots \\ (E^{2^{k-1},0})^T \Psi^\alpha(t) \\ (E^{2^{k-1},1})^T \Psi^\alpha(t) \\ \vdots \\ (E^{2^{k-1},M})^T \Psi^\alpha(t) \end{bmatrix} = P^{(\nu,\alpha)} \Psi^\alpha(t)$$

where

$$\begin{aligned} &E^{i,j} = \hat{E}^{i,j} D^{-1}, \\ &\hat{E}^{i,j} = [\hat{E}_{1,0}^{i,j}, \hat{E}_{1,1}^{i,j}, \ldots, \hat{E}_{1,M}^{i,j}, \ldots, \hat{E}_{2,0}^{i,j}, \hat{E}_{2,1}^{i,j}, \ldots, \hat{E}_{2,M}^{i,j}, \ldots, \hat{E}_{2^{k-1},0}^{i,j}, \hat{E}_{2^{k-1},1}^{i,j}, \ldots, \hat{E}_{2^{k-1},M}^{i,j}]^T, \\ &\hat{E}_{n,m}^{i,j} = \langle I_t^\nu \psi_{i,j}^\alpha(t), \psi_{n,m}^\alpha(t) \rangle, \quad n = 1, \ldots, 2^{k-1}, \quad m = 0, \ldots, M. \end{aligned}$$

2.5 Numerical Method

In this study, we focus on the following fractional optimal control problems

$$\begin{aligned} &\min J = \int_0^1 \mathcal{F}(t, x(t), u(t))\, dt, \\ &{}_0^C D_t^\nu x(t) = \mathcal{G}(t, x(t)) + b(t) u(t) \\ &\nu > 0, \quad t \in [0, 1], \end{aligned} \tag{10}$$

where $x(t)$ and $u(t)$ are state and control functions, respectively.
From the above equation, we can write

$$u(t) = \frac{1}{b(t)} \left({}_0^C D_t^\nu x(t) - \mathcal{G}(t, x(t)) \right). \tag{11}$$

Now, for solving our problem, we expand ${}_0^C D_t^\nu x(t)$ by the FBWFs as

$${}_0^C D_t^\nu x(t) \simeq C^T \Psi^\alpha(t). \tag{12}$$

Using operational matrix of fractional integration and the property of Riemann-Liouville of integration, we have

$$x(t) = I_t^{\nu}(C^T\Psi^{\alpha}(t)) + \sum_{i=0}^{[\nu]} \frac{x_i t^i}{i!} = C^T P^{(\nu,\alpha)}\Psi^{\alpha}(t) + \sum_{i=0}^{[\nu]} \frac{x_i t^i}{i!} = C^T P^{(\nu,\alpha)}\Psi^{\alpha}(t) + d^T\Psi^{\alpha}(t), \tag{13}$$

where

$$\sum_{i=0}^{[\nu]} \frac{x_i t^i}{i!} \simeq d^T\Psi^{\alpha}(t).$$

By substituting Eqs. (12), (13) in Eqs. (11), (10) our problem is converted to following problem

$$\min \tilde{J} = \int_0^1 \mathcal{F}\left(t, (C^T P^{(\nu,\alpha)} + d^T)\Psi^{\alpha}(t), \frac{1}{b(t)}\left(C^T\Psi^{\alpha}(t) - \mathcal{G}\left(t, (C^T P^{(\nu,\alpha)} + d^T)\Psi^{\alpha}(t)\right)\right)\right) dt. \tag{14}$$

Using Gauss-Legendre quadrature rule to approximate integration on $[0,1]$ in Eq. (14) and following necessary conditions of optimization

$$\frac{\partial \tilde{J}}{\partial C} = 0, \tag{15}$$

we can determine C by means of packages such as Matlab, and we obtain the approximate solution of Eq. (10).

3 Numerical results

In this section, some numerical examples are provided to demonstrate the efficiency and reliability of the proposed method.

Example 3.1.

$$\begin{aligned} &\min J = \int_0^1 (x^2(t) - 2t^{\frac{3}{2}}x(t) + u^2(t) - \tfrac{3\sqrt{\pi}}{4}e^{-t}u(t) + e^{-t+t^{\frac{3}{2}}}u(t) + t^3 + \tfrac{9\pi}{64}e^{-2t} - \tfrac{3\sqrt{\pi}}{8}e^{-2t+t^{\frac{3}{2}}} + \\ &\qquad \tfrac{1}{4}e^{-2t+2t^{\frac{3}{2}}} + e^{2t})dt, \\ &{}_0^cD_t^{\frac{3}{2}}x(t) = e^{x(t)} + 2e^t u(t), \\ &x(0) = \dot{x}(0) = 0. \end{aligned} \tag{16}$$

The exact solution is $J^* = 3.194528049$, $x^*(t) = t^{\frac{3}{2}}$ and $u^*(t) = \frac{1}{2}e^{-t}(-e^{\frac{3}{2}} + \frac{3\sqrt{\pi}}{4})$. Table 1 shows the results for J of the present method with $k = 1, 2$ and $M = 1, 2, 3$. In Table 2, the results for J of the Legendre functions [4] and the present method are compared.

Example 3.2.

$$\begin{aligned} &\min J = \int_0^1 \left(e^t(x(t) - t^4 + t - 1)^2 + (1 + t^2)(u(t) + 1 - t + t^4 - \tfrac{8000t^{\frac{21}{10}}}{77\Gamma(\frac{1}{10})})^2\right) dt, \\ &{}_0^cD_t^{1.9}x(t) = x(t) + u(t), \\ &x(0) = 1, \quad \dot{x}(0) = -1, \end{aligned} \tag{17}$$

The exact solution is $J^* = 0$, $x^*(t) = t^4 - t + 1$ and $u^*(t) = -t^4 + t - 1 + \frac{8000t^{\frac{21}{10}}}{77\Gamma(\frac{1}{10})}$. In Table 3, the results for J of the present method with $k = 1, 2$ and $M = 1, 2, 3$ are listed.

Table 1: The estimated values of J for Example 3.1

	$\alpha = 1$	$\alpha = 1.5$
k=1, M=1	3.1963968	3.1948096
k=1, M=2	3.1945620	3.1945342
k=1, M=3	3.1945413	3.1945321
k=2, M=1	3.1946701	3.1945823
k=2, M=2	3.1945401	3.1945324
k=2, M=3	3.1945301	3.1945298

Table 2: The comparison of the estimated values of J with the other methods for Example 3.1

Legendre functions with $n = m = 3$	Present method with with $k = 1$, $M = 2$	Present method with with $k = 2$, $M = 3$	The exact value
3.19453	3.194534	3.194529	3.194528

Table 3: The estimated values of J for Example 3.2

	$\alpha = 1$	$\alpha = 1.9$
k=1, M=1	5.1E-3	2.1E-3
k=1, M=2	3.4E-3	1.8E-3
k=1, M=3	1.07E-4	3.1E-5
k=2, M=1	7.6E-4	4.7E-4
k=2, M=2	8.6E-5	6.4E-6
k=2, M=3	5.3E-7	6.7E-8

4 Conclusion

In this paper, by using the operational matrix of fractional integration and the fractional-order Bernoulli wavelet functions, the fractional optimal control problems were reduced to an equivalent variational problem. Then the approximate solution of variational problem obtained approximately by Gauss quadrature formula and solving the nonlinear system of equations. The numerical results demonstrate the validity and applicability of this method.

References

[1] 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, 1998.

[2] P. Rahimkhani, Y. Ordokhani, Generalized fractional-order Bernoulli-Legendre functions: an effective tool for solving two-dimensional fractional optimal control problems, *IMA J Math Control I*, 36 (2019), No. 1, 185–212.

[3] P. Rahimkhani, Y. Ordokhani, E. Babolian, Fractional-order Bernoulli wavelets and their applications, *Appl. Math. Model.*, 40 (2016), 8087-8107.

[4] A. Lotfi, S.A. Yousefi, M. Dehghan, Numerical solution of a class of fractional optimal control problems via the Legendre orthonormal basis combined with the operational matrix and the Gauss quadrature rule, *J. Comput. Appl. Math.*, 250 (2013), 143–160.

Using goal programming to solve least squares problems[1]

Mohammad Ali Yaghoobi*

Department of Applied Mathematics, Faculty of Mathematics and Computer,
Shahid Bahonar University of Kerman, Kerman, Iran

Abstract

This paper considers well-known least squares and multiobjective least squares problems. It is shown that those problems can be solved equivalently by goal programming models. Moreover, the proposed goal programming models are approximated by linear optimization problems.

Keywords: Least squares problems, Goal programming, Linear optimization
Mathematics Subject Classification [2010]: 93E24, 90C29, 90C05

1 Introduction

Least squares problems are very famous and have many applications in many fileds [1, 2, 5, 6]. The method of least squares was first introduced by Legendre and Gauss more than two hundred years ago. It has been one of the most used techniques in many application fileds such as statistics [6], finance [5], machine learning [1], etc. Recently, because of encountering with big data and big systems, multiobjective least squares are also investigated [1, 6]. In the case of multiobjective problems, different methods exist for dealing with more than one objective function [4]. In this area, goal programming is one of the popular methods which is widely used [3]. This paper introduces goal programming models for solving least squares as well as multiobjective least squares problems. Moreover, for better computational models, we approximate the goal programming models by linear optimization ones. Linear optimization problems can be solved effectively in polynomial times [5] by related softwares such as MATLAB, GAMS, LINGO, etc.

2 Goal Programming

Consider a multiobjective optimization problem as follows [4]:

$$\begin{array}{ll} \min & (f_1(x), f_2(x), ..., f_k(x)) \\ s.t. & x \in \mathcal{X}, \end{array} \tag{1}$$

where $f_i : \mathbb{R}^n \to \mathbb{R}$ is a real valued function for $i = 1, ..., k$, and $\mathcal{X} \subset \mathbb{R}^n$ is called the feasible region. The set $\mathcal{X}$, usually, is given by $\mathcal{X} = \{x \in \mathbb{R}^n \ : \ g_j(x) \leq b_i, \ i = 1, ..., m_1; \ h_j(x) = b_j, \ j = 1, ..., m_2\}$, where g_j $(j = 1, ..., m_1)$ and h_j $(j = 1, ..., m_2)$ are real valued functions.

[1] Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University
*Speaker. Email address: yaghoobi@uk.ac.ir

Goal programming conversts the problem (1) to the following optimization problem [3]:

$$\begin{array}{ll} \min & F(n_1,...,n_k,p_1,...,p_k) \\ s.t. & f_i(x)+n_i-p_i=t_i, \quad i=1,...,k, \\ & x\in\mathcal{X}; n_i,p_i\geq 0, \qquad i=1,...,k, \end{array} \tag{2}$$

where n_i and p_i $(i=1,...,k)$ are called deviational variables, t_i $(i=1,...,k)$ is a targret value related to the function f_i, and F is a function of deviational variables that should be minimized.

In goal programming the function F has different structures and is called the achievement function. One of the famous achievement functions is as follows [3]:

$$F(n_1,...,n_k,p_1,...,p_k)=\sum_{i=1}^{k} w_i n_i + w_i' p_i, \tag{3}$$

where w_i and w_i' $(i=1,...,k)$ are the relative weights assigned to the deviational variables n_i and p_i, respectively, according to the importance of the function f_i.

3 Solving least squares problems by goal programming

Consider a system of linear equations as follows [1]:

$$Ax=b, \tag{4}$$

where A is an $m\times n$ matrix, and $x\in\mathbb{R}^n, b\in\mathbb{R}^m$ are the vectors of variables and right hand side, respectively. The system (4) can be also written as $a^i x = b_i,\ i=1,...,m$, whenever a^i is the i-th row of the matrix A.

In least squares problems, we seek a solution x for the system (4) such that $\|Ax-b\|_2^2$ is as small as possible. Thus, the following optimization problem should be solved:

$$\begin{array}{ll} \min & \|Ax-b\|_2^2 \\ s.t. & x\in\mathbb{R}^n, \end{array} \tag{5}$$

which is a nonlinear (quadratic) optimization problem. Instead of solving the problem (5), we can solve the following optimization problem, by use of goal programming problem (2):

$$\begin{array}{ll} \min & \displaystyle\sum_{i=1}^{m} n_i^2+p_i^2 \\ s.t. & a^i x+n_i-p_i=b_i, \quad i=1,...,m, \\ & n_i,p_i\geq 0, \qquad i=1,...,m, \end{array} \tag{6}$$

where all weights w_i and w_i' $(i=1,...,m)$ are set equal to one, $t_i=b_i$ for $i=1,...,m$, and $F(n_1,...,n_m,p_1,...,p_m)=\displaystyle\sum_{i=1}^{m} n_i^2+p_i^2$.

Theorem 3.1. *The optimal solution of the optimization problems (5) and (6) are the same.*

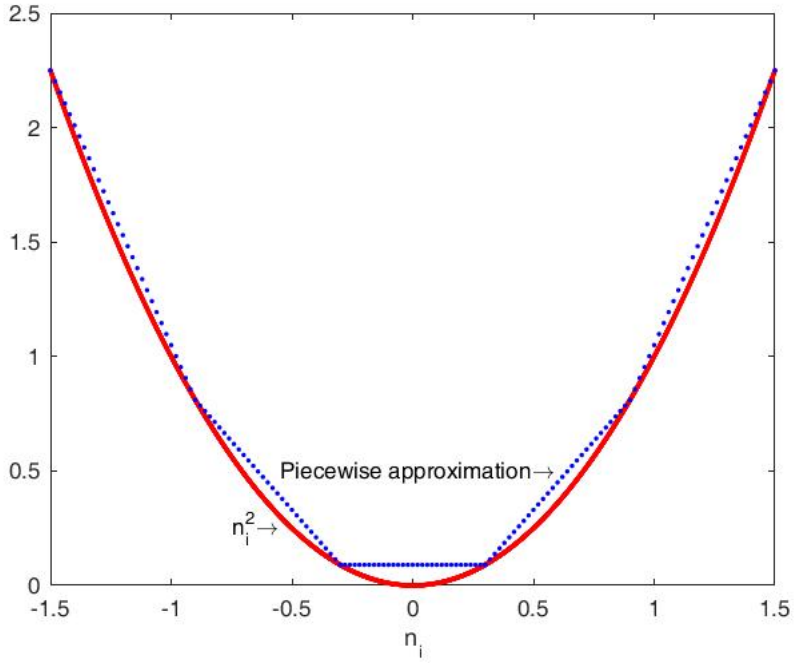


Figure 1: A piecewise approximation to the funtion n_i^2.

Although Theorem 3.1 states that we can solve the problem (6) instead of the problem (5), but the former problem is a nonlinear optimization problem. Indeed, the objective function of the problem (6) is sum of quadratic functions n_i^2 and p_i^2. The function n_i^2 can be appoximated by a piecewise linear function. Figure 1 shows one of such approximations. In fact, the piecewise linear function can be written as:

$$\max\{a_{i1}n_i + b_{i1}, a_{i2}n_i + b_{i2}, ..., a_{iv}n_i + b_{iv}\}, \tag{7}$$

where each $a_{ij}n_i + b_{ij}$ $(j = 1, ..., v)$ is, in general, an affine function. The function p_i^2 can be approximated similarly as:

$$\max\{c_{i1}p_i + d_{i1}, c_{i2}p_i + d_{i2}, ..., c_{iv}p_i + d_{iv}\}. \tag{8}$$

By using (7) and (8), the problem (6) is equivalent to the following linear programming problem:

$$\begin{aligned}
\min \quad & \sum_{i=1}^{m} y_i + z_i \\
s.t. \quad & a^i x + n_i - p_i = b_i, \quad i = 1, ..., m, \\
& a_{ij}n_i + b_{ij} \le y_i, \quad i = 1, ..., m, j = 1, ..., v, \\
& c_{ij}p_i + d_{ij} \le z_i, \quad i = 1, ..., m, j = 1, ..., v, \\
& n_i, p_i, y_i, z_i \ge 0, \quad i = 1, ..., m,
\end{aligned} \tag{9}$$

where y_i and z_i are the new continuous variables for $i = 1, ..., m$. Since linear programming problems can be solved by polynomial algorithms, the problem (9) can be solved effectively using commercial softwares such as MATLAB, GAMES, etc.

3.1 Multiobjective least squares

In multiobjective least squares problems, we consider the following systems of linear equations simultaneously [1, 6]:

$$A_1x = b_1,\ A_2x = b_2, ..., A_kx = b_k,$$

where A_l is an $m_l \times n$ matrix $(l = 1, ..., k)$, $b_l \in \mathbb{R}^{m_l}$ $(l = 1, ..., k)$, and $x \in \mathbb{R}^n$. The goal is finding a solution x such that all

$$\|A_1x - b_1\|_2^2,\ \|A_2x - b_2\|_2^2, ..., \|A_kx - b_k\|_2^2,$$

are as small as possible. By applying the results of Section 3, we can use the following multiobjective goal programming problem for solving this problem:

$$\begin{array}{lll} \min & (\sum_{i=1}^{m_1} n_{1i}^2 + p_{1i}^2, \sum_{i=1}^{m_2} n_{2i}^2 + p_{2i}^2, ..., \sum_{i=1}^{m_k} n_{ki}^2 + p_{ki}^2) & \\ s.t. & a^{li}x + n_{li} - p_{li} = b_{li}, & l = 1, ..., k, i = 1, ..., m_1, \\ & n_{li}, p_{li} \geq 0, & l = 1, ..., k, i = 1, ..., m_l, \end{array} \tag{10}$$

where a^{li} is the i-th row of the matrix A_l, and b_{li} is the i-th component of the vector b_l for $l = 1, ..., k$.

By using the well-known weighted sum method [4], the problem (10) can be converted to a single optimization problem as follows:

$$\begin{array}{lll} \min & \sum_{l=1}^{k}\sum_{i=1}^{m_l} w_{li}n_{li}^2 + w'_{li}p_{li}^2 & \\ s.t. & a^{li}x + n_{li} - p_{li} = b_{li}, & l = 1, ..., k, i = 1, ..., m_l, \\ & n_{li}, p_{li} \geq 0, & l = 1, ..., k, i = 1, ..., m_l, \end{array} \tag{11}$$

where w_{li} and w'_{li} are the weights assigned to the deviational variables n_{li} and p_{li}, respectively.

Finally, we use the linear approximation of Section 3 and propose a linear programming problem related to the problem (11) as:

$$\begin{array}{lll} \min & \sum_{l=1}^{k}\sum_{i=1}^{m_l} y_{li} + z_{li} & \\ s.t. & a^{li}x + n_{li} - p_{li} = b_{li}, & l = 1, ..., k, i = 1, ..., m_l, \\ & a_{lij}n_{li} + b_{lij} \leq y_{li} & l = 1, ..., k, i = 1, ..., m_l, j = 1, ..., v, \\ & c_{lij}p_{li} + d_{lij} \leq z_{li} & l = 1, ..., k, i = 1, ..., m_l, j = 1, ..., v, \\ & n_{li}, p_{li}, y_{li}, z_{li} \geq 0, & l = 1, ..., k, i = 1, ..., m_l. \end{array} \tag{12}$$

4 Numerical results

Example 4.1. Consider the following system of linear equations:

$$\begin{array}{rcl} 3x_1 + x_2 & = & 4, \\ x_1 + 2x_2 & = & 3, \\ -2x_1 + x_2 & = & 2. \end{array} \tag{13}$$

The exact solution of least squares problem for the system (13) is $x_1 = 0.4$ and $x_2 = 1.8$ with $\|Ax - b\|_2^2 = 3$. Table 1 shows the implementation results of the model (9).

Table 1: Results of the model (9) for the system (13).

# linear pieces	x_1	x_2	$\|Ax-b\|_2^2$	CPU time (seconds)
1	0.4002	1.7996	6	0.17
2	0.4	1.8	3	0.18

The model (9) was coded in MATLAB. In the next example, we consider a multiobjective least squares problem and solve it with the model (12).

Example 4.2. Consider a multiobjective least squares problem with two systems of linear equations. The first system is the same as the system (13) and the second one is as follows:

$$\begin{aligned} 4x_1 + x_2 &= 2, \\ x_1 - 2x_2 &= -1, \\ 3x_1 + x_2 &= 3. \end{aligned} \tag{14}$$

The exact solution of this multiobjective least squares problem with weights 0.6 and 0.4 for the systems (13) and (14), respectively, is $x_1 = 0.4066$ and $x_2 = 1.4091$ with $\|Ax-b\|_2^2 = 3.6222$. Table 2 shows the implementation results of the model (12) in MATLAB.

Table 2: Results of the model (12) for the systems (13) and (14).

# linear pieces	x_1	x_2	$\|Ax-b\|_2^2$	CPU time (seconds)
1	0.6	1.2	5.52	0.18
20	0.4	1.4	3.624	0.21
200	0.4071	1.4086	3.622	0.54
400	0.4070	1.4090	3.6222	1.39

5 Conclusion

This paper proposed two goal programming models for solving least squares and multiobjective least squares problems. Linear programming approximation of those models were also introduced. Numerical examples showed that the approximated solutions were acceptable. Moreover, the CPU times for computing the solutions were reasonable.

References

[1] S. Boyd and L. Vandenberghe, *Introduction to Applied Linear Algebra: Vectors, Matrices, and Least Squares*, Cambridge University Press, Cambridge, 2018.

[2] A. Charnes, W. W. Cooper and T. Sueyoshi, Least squares/ridge regression/ and goal programming/constrained regression alternatives, *European Journal of Operational Research*, 27 (1986), 146–157.

[3] D. F. Jones and M. Tamiz, *Practical Goal Programming*, Springer, Berlin, 2010.

[4] M. Ehrgott, *Multicriteria Optimization*, Springer, Berlin, 2005.

[5] S. Nadarajah and N. Secomandi, Relationship between least squares Monte Carlo and approximate linear programming, *Operations Research Letters*, 45 (2017), 409–414.

[6] E. G. Nepomuceno, R. H. C. Takahashi and L. A. Aguirre, Multiobjective parameter estimation for nonlinear systems: affine information and least squares formulation, *International Journal of Control*, 80 (2007), 863–871.

Papers

Part 2: Posters

Lower bounds for the energy of bipartite graphs[1]

Vahid Adish* and Maryam Khosravi

Faculty of Mathematics and computer, Shahid Bahonar University of Kerman, Kerman, Iran

Abstract

The energy $E(G)$ of a graph G is the sum of the absolute values of all eigenvalues of G. In this note, the authors are interested in the relation between the energy of a graph G and the matching number $\mu(G)$ of G. It is well-known that $E(G) \geq 2\mu(G)$. In this paper for a category of graphs, we improve the lower bound of the energy of graphs to $E(G) \geq 2\mu(G) + 2$.

Keywords: Energy of graphs, Matching number, Bipartite graph, Eigenvalue
Mathematics Subject Classification [2010]: 05C50, 15A18

1 Introduction

Let $G = (V(G), E(G))$be a simple graph. The order of G denotes the number of vertices of G. For two vertices x and y by $e = xy$ we mean the edge e between x and y. For every vertex $v \in V(G)$, the degree of v is the number of edges incident with v and is denoted by $deg_G(v)$. A k-regular graph is a graph such that every vertex of that has degree k. By $K_n, K_{p,q}$ and C_n, we mean the complete graph with n vertices, the complete bipartite graph with parts of sizes p, q, and the cycle with n vertices respectively.

For a subset U of $V(G)$, denote by $G - U$ the graph obtained from G by deleting the vertices of U together with all edges incident to them. If H is an induced subgraph of G, we will use $G - H$ to denote the induced subgraph $G - V(H)$. The subgraph $G - H$ of G is also called the complement of H in G. If F is a set of edges of G such that $G - F$ is the disjoint union of two complementary induced subgraphs H and K, then F is called a cut set of G, and we write $G - F = H \oplus K$. A cycle of G is called a Hamiltonian cycle if it contains all vertices of G, and G is called a Hamiltonian graph if a Hamiltonian cycle lies in G. Let G and H be graphs with vertex sets $V(G)$ and $V(H)$, respectively.

The Kronecker product of G and H, denoted by $G \otimes H$, is the graph defined as follows. The vertex set of $G \otimes H$ is $V(G) \times V(H)$. The vertices (u, v) and (u', v') are adjacent if u is adjacent to u' in G and v is adjacent to v' in H.

A matching M in G is a set of pairwise non-adjacent edges, that is, no two edges in M share a common vertex. A Graph G has perfect matching if it has a matching in which the edgesare collectively incident with all the vertices. A maximum matching is a matching that contains the largest possible number of edges. The matching number of G, denoted by $\mu(G)$, is the edge number of a maximum matching.

[1]Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University
*Speaker. Email address: vahidadish1@gmail.com

The energy of graphs was defined by Ivan Gutman in 1978. The energy $E(G)$ of a graph G is defined to be the sum of the absolute values of all eigenvalues of $A(G)$. The motivation for the definition of $E(G)$ comes from Chemistry, where the first results on $E(G)$ were obtained as early as the 1940s [3]. However, in the last two decades, research on graph energy has much intensified, resulting in a very large number (over 150) of publications.

Since the eigenvalues of the complete graph K_n are $n-1$ (with multiplicity 1) and -1 (with multiplicity $n-1$), so $E(K_n) = 2n-2$. Also $E(K_{p,q}) = 2\sqrt{pq}$, since the eigenvalues of the complete bipartite graph $K_{p,q}$ are $\sqrt{pq}$ (with multiplicity 1), 0 (with multiplicity $p+q-2$) and $-\sqrt{pq}$ (with multiplicity 1).

In [5] the energy and the matching number of a graph were compared and it was proved that $E(G) \geq 2\mu(G)$ for every graph G. In some special cases the equality condition was mentioned.

Recently, in [1], the author show that the necessary and sufficient condition for equality is that G is union of some complete bipartition graph with equal parts and some isolated vertices. Also, it was proved that for connected graphs without perfect matching, $E(G) \geq 2\mu(G)+1$, except for complete bipartite graphs of the form $K_{r,r+1}$. Furthuremore, we have the following bound for energy of graphs.

Theorem 1.1 ([1, Theorem 12]). *Let G be a graph whose cycles are vertex-disjoint. If we denote the number of odd cycles of G with length at least 5 by $c_0(G)$, then*

$$E(G) \geq 2\mu(G) + c_0(G).$$

2 Main results

Given a graph G, its bipartite double $G \otimes K_2$ is the Kronecker product of G and K_2. If G is connected, then its bipartite double is connected and bipartite, and if G has spectrum Φ, then $G \otimes K_2$ has spectrum $\Phi \cup -\Phi$. Thus $E(G \otimes K_2) = 2E(G)$. using this fact, we have the following lemma.

Lemma 2.1 ([2, Lemma 3.25]). *Let A and B be symmetric matrices of order $m \times m$ and $n \times n$, respectively. If $\lambda_1, ..., \lambda_m$ and $\mu_1, ..., \mu_n$ are the eigenvalues of A and B, respectively, then the eigenvalues of $A \otimes B$ are given by $\lambda_i \mu_j; i = 1, ..., m; j = 1, ..., n$.*

Using this lemma and the fact that $n-1$-regular bipartite graph with $2n$ vertices can be written as $K_{n,n} - T$ for some perfect matching T of $K_{n,n}$, we can compute the energy of $n-1$-regular bipartite graph.

Theorem 2.2. *Let G be $n-1$-regular bipartite graph with $2n$ vertices and $n \geq 3$, then we have $E(G) = 4n-4$.*

In [1], the author, prove that $E(C_n) \geq n+1$ for odd n. The following lemma, gives a simillar result for $n \equiv 2 \pmod 4$.

Lemma 2.3. *Let $G = C_n$ be a cycle of order n such that $n \equiv 2 \pmod 4$ then*

$$E(C_n) \geq n+2. \tag{1}$$

Note that in this case C_n has perfect maching and the previous lower bound for it was n.

As an anothor result, we obtain a refined lower bounds for the energy of Hamiltonian graph. We need the following Lemma.

Lemma 2.4 ([4, Theorem 3.1]). *Let H be an induced subgraph of a simple graph G. Then $E(H) \leq E(G)$ and equality holds if and only if $E(H) = E(G)$.*

One can easily see that deleting some edges of graph can decrease or increase the energy of graph. Although, the following lemma shows the graph energy changes when we delete a cut set of edges from a graph.

Lemma 2.5 ([4, Theorem 3.4]). *If E is a cut set of a simple graph G then $E(G - E) \leq E(G)$.*

Finally, our main result is as follows.

Theorem 2.6. *If G is a (non-complete) Hamiltonian bipartite graph with n vertices and $n \equiv 2 \pmod 4$ then $E(G) \geq n + 2$.*

Since bipartite graphs have no odd cycle, the lower bound which is obtained from Theorem 1.1, is n and this theorem improves this bound.

3 Conclusion

For some bipartite graph, we can improved the lower bound of energy of graph to $2\mu(G)+2$.

References

[1] F. Ashraf, Energy, matching number and odd cycles of graphs, *Linear Algebra Appl.*, 577 (2019), 159–167.

[2] R.B. Bapat, *Graphs and Matrices*, Springer-Verlage, London, 2010.

[3] C.A. Coulson, On the calculation of the energy in unsaturated hydrocarbon molecules, Proc. Cam-bridge Philos. Soc. 36 (1940), 201–203.

[4] J.Day, W.So, Graph energy change due to edge deletion, *Linear Algebra Appl.*, 428 (2008), 2070–2078.

[5] D.Wong, X.Wang, R.Chu, Lower bounds of graph energy in terms of matching number, *Linear Algebra Appl.*, 549 (2018), 276–286.

Some results on pseudospectrum of matrices[1]

Gholamreza Aghamollaei[1] and Sharifeh Rezagholi[2,*]

[1]Department of Pure Mathematics, Faculty of Mathematics and Computer, Shahid Bahonar University of Kerman, Kerman, Iran

[2]Department of Basic Science, Payame Noor University, Tehran, Iran

Abstract

In this paper, using the polynomial numerical hulls, a new upper bound for the pseudospectrum of matrices is obtained. Also, some properties of the pseudospectrum of matrices are investigated.

Keywords: Pseudospectrum, Spectrum, Polynomial numerical hulls
Mathematics Subject Classification [2010]: 15A60, 15A18

1 Introduction and preliminaries

Let $\mathbb{M}_n(\mathbb{C})$ be the algebra of all $n \times n$ complex matrices, and $A \in \mathbb{M}_n(\mathbb{C})$. The *field of values* or the *numerical range* of A is defined as $W(A) = \{x^*Ax \ : \ x \in \mathbb{C}^n, \ x^*x = 1\}$ which is useful in studying and understanding of matrices and operators, and has many applications in numerical analysis, quantum theory, etc; e.g., see [5] and its references. It is known, e.g., see [1, Lemma 6.22.1], that

$$W(A) = \{\lambda \in \mathbb{C} \ : \ |\lambda - \mu| \leq \|A - \mu I\|, \ \ \forall \mu \in \mathbb{C}\}, \tag{1}$$

where $\|.\|$ is the matrix norm subordinate to the Euclidean vector norm, and I is the $n \times n$ identity matrix. For a positive integer k, we denote by $\mathbb{P}_k$ the set of all scalar polynomials of degree k or less. Put $\mathbb{P} = \cup_{k=1}^{\infty} \mathbb{P}_k$ which is the set of all scalar polynomials. By this idea, the notion of numerical range $W(A)$ as in (1) can be generalized to the notion of *polynomial numerical hull of order* k of A, which is defined and denoted, e.g., see [4], by

$$V^k(A) = \{\lambda \in \mathbb{C} : \ |p(\lambda)| \leq \|p(A)\| \text{ for all } p \in \mathbb{P}_k\}.$$

This set has many applications in the study of convergence of iterative methods in solving linear systems. In the following proposition, we state some properties of polynomial numerical hulls of matrices which will be useful in our discussion. For more information, see [3, 4].

Proposition 1.1. *Let $A \in \mathbb{M}_n(\mathbb{C})$. Then*

[1]Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University
[*]Speaker. Email address: sh_rezagholi79@yahoo.com

1. $V^k(A)$ *is a compact set in* $\mathbb{C}$;
2. $\sigma(A) = V^m(A) \subseteq \cdots \subseteq V^{k+1}(A) \subseteq V^k(A) \subseteq \cdots \subseteq V^1(A) = W(A)$, *where* $m \geq n$;
3. $V^k(\alpha A + \beta I) = \alpha V^k(A) + \beta$, *where* $\alpha, \beta \in \mathbb{C}$;
4. $V^k(U^*AU) = V^k(A)$, *where* $U \in \mathbb{M}_n$ *is unitary;*
5. $V^k(A^T) = V^k(A)$ *and* $V^k(A^*) = \overline{V^k(A)} := \{\overline{\lambda} \ : \ \lambda \in V^k(A)\}$;
6. $V^k(A) = \{\lambda \in \mathbb{C} : (\lambda, \lambda^2, ..., \lambda^k) \in conv(W(A, A^2, ..., A^k))\}$, *where* $conv(.)$ *denotes the convex hull, and* $W(A_1, ..., A_k) := \{(x^*A_1x, ..., x^*A_kx) : x \in \mathbb{C}^n, \ x^*x = 1\}$ *is the joint numerical range of* $(A_1, ..., A_k) \in \mathbb{M}_n(\mathbb{C})^k$;
7. *If* A *is Hermitian, then* $V^k(A) = \begin{cases} conv(\sigma(A)) & for \ k = 1, \\ \sigma(A) & for \ k \geq 2; \end{cases}$
8. *If* B *is a principal submatrix of* A, *then* $V^k(B) \subseteq V^k(A)$;
9. $V^k(A) = \{\lambda \in \mathbb{C} \ : \ p(\lambda) \in W(p(A)) \ \ \forall p \in \mathbb{P}_k\}$.

For a given $\epsilon > 0$ and a matrix $A \in \mathbb{M}_n(\mathbb{C})$, the $\epsilon-$*pseudospectrum* (*pseudospectrum* for short) of A is defined and denoted, see [6], by

$$\sigma_\epsilon(A) = \bigcup_{E \in \mathbb{M}_n, \|E\| \leq \epsilon} \sigma(A + E), \tag{2}$$

where the matrix $A + E$ is a perturbation of A, and $\sigma(.)$ denotes the spectrum, i.e., the set of all eigenvalues. During this paper, we denote by $D_\epsilon(a) = \{\mu \in \mathbb{C} : |\mu - a| \leq \epsilon\}$ the closed disk at centered $a \in \mathbb{C}$ with radius $\epsilon > 0$. The following properties (see [6]) are useful in our discussion.

Proposition 1.2. *Let* $A \in \mathbb{M}_n(\mathbb{C})$ *and* $\epsilon > 0$. *Then*

1. $\sigma_\epsilon(\alpha A + \beta I) = \alpha \sigma_{\epsilon/|\alpha|}(A) + \beta$, *where* $\alpha, \beta \in \mathbb{C}$ *and* $\alpha \neq 0$;
2. $\sigma_\epsilon(A) = D_\epsilon(\mu)$ *if and only if* $A = \mu I$, *where* $\mu \in \mathbb{C}$;
3. *If* $A = \begin{pmatrix} A_1 & B \\ 0 & A_2 \end{pmatrix}$, *where* A_1 *and* A_2 *are square matrices, then* $\sigma_\epsilon(A_1) \cup \sigma_\epsilon(A_2) \subseteq \sigma_\epsilon(A)$. *The equality holds if* $B = 0$; *i.e.,* $\sigma_\epsilon(A_1 \oplus A_2) = \sigma_\epsilon(A_1) \cup \sigma_\epsilon(A_2)$.

Proposition 1.3. *Let* $A \in \mathbb{M}_n(\mathbb{C})$. *Then*

1. *for every* $\epsilon > 0$, $\sigma(A) + D_\epsilon(0) \subseteq \sigma_\epsilon(A)$;
2. A *is normal if and only if* $\sigma_\epsilon(A) = \sigma(A) + D_\epsilon(0)$ *for every* $\epsilon > 0$.

In this paper, we are going to study some algebraic and geometrical properties of pseudospectrum of matrices. Using the polynomial numerical hulls, we find a new upper bound for pseudospectrum. We also give some facts about the pseudospectrum of 2×2 block triangular matrices.

2 Main results

Let $A \in \mathbb{M}_n(\mathbb{C})$. In [2, Theorem 2.1], the author found the following cover for the pseudospectrum:

$$\sigma_\epsilon(A) \subseteq \{\lambda \in \mathbb{C} \ : \ dist(p(\lambda), W(p(A))) < \epsilon t_p\}, \tag{3}$$

where $p \in \mathbb{P}$ is an arbitrary polynomial, and t_p is a positive constant depending on p. We now are going to improve the bound in (3). For this, by the idea used in (2), we introduce the notion of extended polynomial numerical hull of order k as follows:

$$V_\epsilon^k(A) = \bigcup_{E \in \mathbb{M}_n, \|E\| \leq \epsilon} V^k(A+E).$$

Note, by Proposition 1.1(9), that:

$$\sigma_\epsilon(A) \subseteq \bigcup_{\|E\| \leq \epsilon} \bigcap_{p \in \mathbb{P}_k} \{\lambda \in \mathbb{C} \ : \ dist(p(\lambda), W(p(A+E))) = 0\} = V_\epsilon^k(A). \tag{4}$$

Obviously and by a simple computation, we see, for every $p \in \mathbb{P}_k$, $\lambda \in \mathbb{C}$, and any $E \in \mathbb{M}_n$ with $\|E\| \leq \epsilon$, that

$$dist(p(\lambda), W(p(A))) \leq dist(p(\lambda), W(p(A+E))) + \epsilon t_p.$$

This shows that the upper bound in (4) is an improvement of the bound mentioned in (3).

Theorem 2.1. *Let $\epsilon > 0$ and $A \in \mathbb{M}_n(\mathbb{C})$. Then*

1. *$V_\epsilon^k(U^*AU) = V_\epsilon^k(A)$, where $U \in \mathbb{M}_n(\mathbb{C})$ is unitary;*
2. *$\sigma_\epsilon(A) = V_\epsilon^m(A) \subseteq \cdots \subseteq V_\epsilon^{k+1}(A) \subseteq V_\epsilon^k(A) \subseteq \cdots \subseteq V_\epsilon^1(A) = W(A) + D_\epsilon(0)$, where $m \geq n$;*
3. *$V_\epsilon^k(\alpha A + \beta I) = \alpha V_{\epsilon/|\alpha|}^k(A) + \beta$, where $\alpha \neq 0$ and β are complex scalars;*
4. *$V_\epsilon^k(A)$ is a nonempty and compact set in $\mathbb{C}$;*
5. *$V_\epsilon^k(A^T) = V_\epsilon^k(A)$ and $V_\epsilon^k(A^*) = \overline{V_\epsilon^k(A)}$. Consequently, if A is Hermitian, then $V_\epsilon^k(A)$ is symmetric with respect to the real axis;*
6. *$V_\epsilon^k(A) = D_\epsilon(\mu)$ if and only if $A = \mu I$, where $\mu \in \mathbb{C}$;*
7. *If $A = A_1 \oplus A_2$, where $A_i \in M_{n_i}(\mathbb{C})$ with $n_1 + n_2 = n$, then $V_\epsilon^k(A_1) \cup V_\epsilon^k(A_2) \subseteq V_\epsilon^k(A)$. The set equality holds if $k = n$.*

In the following, we state a result about the pseudospectrum of matrices, i.e., for $V_\epsilon^n(.)$.

Theorem 2.2. *Let $\epsilon > 0$, and $A, B \in M_n$ be such that $AB = BA$. If A or $A + B$ is normal, then*

$$\sigma_\epsilon(A+B) \subseteq \sigma(A) + \sigma_\epsilon(B).$$

The following example shows that the condition "A or $A+B$ is normal" in Theorem 2.2 is necessary.

Example 2.3. Let $\epsilon > 0$, and $A = B = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$. Clearly, A and $A+B$ are not normal. Also, we have

$$\sigma_\epsilon(A+B) = 2\sigma_{\epsilon/2}(A) = D(0, \sqrt{2\epsilon + \epsilon^2}) \nsubseteq D(0, \sqrt{\epsilon + \epsilon^2}) = \sigma(A) + \sigma_\epsilon(B).$$

References

[1] F.F. Bonsall and J. Duncan, *Numerical Ranges II*, London Mathematical Society Lecture Notes Series, Cambridge University Press, UK, 1973.

[2] E.B. Davies, Spectral bounds using higher order numerical ranges, *London Math. Soc. J. Comput. Math.*, 8 (2005), 17-45.

[3] V. Faber, A. Greenbaum and D.E. Marshall, The polynomial numerical hulls of Jordan blocks and related matrices, *Linear Algebra Appl.*, 374 (2003), 231–246.

[4] A. Greenbaum, Generalizations of field of values useful in the study of polynomial functions of a matrix, *Linear Algebra Appl.*, 347 (2002), 233–249.

[5] R. Horn and C. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, New York, 1991.

[6] L.N. Trefethen and M. Embree, *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*, Princeton University Press, USA, 2005.

Modeling of leukemia B cells using system of fractional ordinary differential equations[1]

Yasin Fadaei[1,*], Ali Ahmadi[1] and Ali Ansari Ardali[2]

[1]Modeling in Health Research Center, Shahrekord University of Medical Sciences, Shahrekord, Iran

[2]Department of Applied Mathematics and Computer Sciences, Shahrekord University, P. O. Box 115, Shahrekord, Iran

Abstract

In this paper, a mathematical model has been developed for describing the behavior and interactions between B leukemia cells and three components of immune system. The model has been presented using a system of fractional ordinary differential equations (FODEs). Dynamics of the system are studied by determining eigenvalues of Jacobian matrix of the system at equilibrium points and the stability status of them. Bifurcation analysis showed that the use of the fractional-order model figures out unpredictable behaviors of the system such as *bistability* and *Hysteresis Phenomenon*.

Keywords: Fractional ordinary differential equations, Chronic lymphocytic leukemia, Stability analysis, Bifurcation

Mathematics Subject Classification [2010]: 92XX, 92Bxx

1 Introduction

Since the early 1990s, mathematical models have been studied in the form of a variety of ordinary differential equations in describing different aspects of cancer, such as tumor growth dynamics and the interaction of immune cells and tumor cells [4, 5].

Another category of mathematical models is differential, fractional-order equations. These fractional models are preferred over classic models for various reasons. Integrals and fractional derivatives describe dynamic systems that have memory and inherited properties. Because of the integral in the definition of fractional derivatives, these derivatives are non-local, that is, to calculate the fractional derivative at a given point, the points that are located in the neighborhood of that poin are used. Therefore, fractional differential equations are more suitable for models with rugged domains. The human immune system is a biological system that has memory and rugged cell population. Fractional differential equations act better than classical differential ones in explaining the behavior of immune system diseases such as B cell chronic lymphocytic leukemia (B–CLL). The derivative order in these equations serves as the memory of parameter of the system. These are some of

[1]Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

*Speaker. Email address: fadaei.yasin@gmail.com

the primary reasons why fractional differential equations models are increasingly applied to dynamical systems.
Due to advantages of fractional-order equations, in this study we will develop the model of Nanda et al. into a model that includes a system including fractional differential equations. We will show that the fractional model is better than the classical ODE model in understanding of the interactions between immune system and B–CLL cells.

Definition 1.1. The fractional integral of order $\alpha \in \mathbb{R}^+$ of the function $g(t)$, $t > 0$ is defined by

$$I^{\alpha} g(t) = \int_0^t \frac{(t-s)^{\alpha-1}}{\Gamma(\alpha)} g(s) ds \tag{1}$$

where the gamma function is defined, as usual, as $\Gamma(z) = \int_0^{\infty} e^{-t} t^{z-1} dt$, and the Caputo fractional derivative of order $\alpha \in (n-1, n)$ of $g(t)$, $t > 0$ is defined by

$$D^{\alpha} = I^{n-\alpha} D^n g(t), \quad D^{\alpha} = \frac{d^{\alpha}}{dt^{\alpha}}, \quad n = 1, 2, \ldots. \tag{2}$$

Theorem 1.2. *[6] Fractional–order ordinary differential equations system*

$$D^{\alpha} x_i(t) = g_i\left(x_1(t), x_2(t), \ldots, x_n(t)\right) \tag{3}$$
$$x_i(0) = c_i, \quad i = 1, 2, \ldots, n, \quad \alpha \in (0, 1), \tag{4}$$

is called asymptotically stable if all the eigenvalues λ_i, of the Jacobian matrix $J = \frac{\partial g_i}{\partial x_i}$ computed at the equilibrium points satisfy $|\arg(\lambda_i)| > \frac{\alpha\pi}{2}$, $i = 1, 2, \ldots, n$.

2 The fractional–order mathematical model

Here, we consider the B–CLL cell population, a cell population of the innate immune system, natural killer (NK) cells, and two cell populations of the adaptive immune system: cytotoxic T cells (C) and T-helper (H) cells. We also assume that t represents the variable time (day). We consider the following populations of cells, measured as concentrations of cells per μliter are denoted by:
B(t) = B–CLL cells population,
N(t) = total NK cells population,
C(t) = total cytotoxic T cells ($CD8^+T$) population,
H(t) = total helper T cells ($CD4^+T$) population.
The system of fractional differential equations is given by

$$\begin{cases} D^{\alpha} B &= s_B + (a-b)B - cBN - dBC, \\ D^{\alpha} N &= s_N - eN - fNB, \\ D^{\alpha} C &= s_C - gC - iCB + kr\frac{B^m}{\eta + B^m} HC, \\ D^{\alpha} H &= s_H - jH + r\frac{B^m}{\eta + B^m} H, \end{cases} \tag{5}$$

with initial conditions $B(0) \geq 0,\ N(0) \geq 0,\ C(0) \geq 0,\ H(0) \geq 0$ and $0 < \alpha < 1$.

3 Analysis of the model

To analyze the model, first, using a non-dimensionalization technique, we transform the model into a simpler form with fewer parameters:

$$\tilde{t} = et, \quad \tilde{B} = \frac{e}{s_B}B, \quad \tilde{N} = \frac{e}{s_N}N, \quad \tilde{C} = \frac{e}{s_C}C, \quad \tilde{H} = \frac{e}{s_H}H$$

and the corresponding paremeters are:

$$a_1 = \frac{a-b}{e}, \quad a_2 = \frac{cs_N}{e^2}, \quad a_3 = \frac{ds_C}{e^2}, \quad a_4 = \frac{fs_B}{e^2}, \quad a_5 = \frac{g}{e},$$

$$a_6 = \frac{is_B}{e^2}, \quad a_7 = k\frac{rs_H}{e^2}, \quad a_8 = \frac{r}{e}, \quad a_9 = \frac{j}{e}, \quad a_{10} = \eta\frac{e^m}{s_B^m}.$$

Dropping the tilde for notational clarity, the resulting system is given by

$$\begin{cases} D^\alpha B &= 1 - a_1 B - a_2 BN - a_3 BC, \\ D^\alpha N &= 1 - N - a_4 NB, \\ D^\alpha C &= 1 - a_5 C - a_6 CB + a_7 \frac{B^m}{a_{10}+B^m} HC, \\ D^\alpha H &= 1 - a_8 H + a_9 \frac{B^m}{a_{10}+B^m} H, \end{cases} \tag{6}$$

3.1 Equilibrium points and their stability

By studying the long-term behavior of the fractional differential equations system, we can obtain useful qualitative information from the progression of the disease. To better understand the dynamics of the system, we first detect the equilibrium points, that is, we should find values for which the changes of system are zero. To do this, we setting each of the four equations in the system (6) equal to zero:

$$D^\alpha B = D^\alpha N = D^\alpha T = D^\alpha H = 0,$$

now, from the NK cell population equation, we obtain:

$$N = \frac{1}{1 + a_4 B}, \tag{7}$$

similarly, the fixed point of H cells

$$H = \frac{a_{10} + B^m}{a_8 a_{10} + (a_8 - a_9)B^m}, \tag{8}$$

and, setting third equation of the (6) to zero gives

$$C = \frac{a_{10} + B^m}{a_5 a_{10} + a_5 B^m - a_7 H B^m + a_6 a_{10} B + a_6 B^{m+1}}, \tag{9}$$

finally, we set the first equation of (6) equal to zero and by inserting the relationships (7) and (9) in it, we have the following equation:

$$B(a_1 + a_2 N + a_3 C) - 1 = 0. \tag{10}$$

By numerical solving of Eq. (10), equilibrium points of the system will be obtained. For biological considerations, we only consider the non-negative equilibrium points. To determine the behavior of the cells near the equilibrium points, we examine their stability.

Suppose that $E = (\hat{B}, \hat{N}, \hat{C}, \hat{H})$ be an equilibrium point of system 6. We linearize the system by obtaining the Jacobian matrix of the system (6). The Jacobian matrix at E is computed by

$$J|_E = \begin{pmatrix} -a_1 - a_2\hat{N} - a_3\hat{C} & a_2\hat{B} & -a_3\hat{B} & 0 \\ -a_4\hat{N} & 1 - a_4\hat{B} & 0 & 0 \\ -a_6\hat{C} + \frac{a_7 a_{10} m \hat{H}\hat{C}\hat{B}^{m-1}}{(a_{10}+\hat{B}^m)^2} & 0 & -a_5 - a_6\hat{B} + \frac{a_7\hat{H}\hat{B}^m}{a_{10}+\hat{B}^m} & \frac{a_7\hat{C}\hat{B}^m}{a_{10}+\hat{B}^m} \\ \frac{a_9 a_{10}\hat{H}\hat{B}^{m-1}}{(a_{10}+\hat{B}^m)^2} & 0 & 0 & -a_8 + \frac{a_9\hat{B}^m}{a_{10}+\hat{B}^m} \end{pmatrix}. \tag{11}$$

Now, we calculate the eigenvalues of this matrix.
Let J_1 be a submatrix which is given by

$$J_1 = \begin{pmatrix} -a_1 - a_2\hat{N} - a_3\hat{C} & a_2\hat{B} \\ -a_4\hat{N} & 1 - a_4\hat{B} \end{pmatrix}, \tag{12}$$

two eigenvalues $\lambda_1,\ \lambda_2$ of the matrix J are equal to eigenvalues of the J_1 which are obtained by solving the characteristic equation

$$P(\lambda) = \lambda^2 - \tau\lambda + \Delta = 0,$$

where τ and Δ are trace and determinant of the matrix J_1, respectively. Other eigenvalues of the J are

$$\lambda_3 = a_8 + \frac{a_9\hat{B}^m}{a_{10} + \hat{B}^L}, \qquad \lambda_4 = -a_5 - a_6\hat{B} + \frac{a_7\hat{H}\hat{B}^m}{a_{10} + \hat{B}^m}.$$

If $a_7 < \frac{(a_5+a_6\hat{B})(a_{10}+\hat{B}^m)}{\hat{H}\hat{B}^m}$, then $\lambda_3 < 0$, also if $a_8 > \frac{a_9\hat{B}^m}{a_{10}+\hat{B}^m}$, then $\lambda_4 < 0$.

Theorem 3.1. *Assume that E be a non–negative equilibrium point of system 6 and let $\lambda_3 < 0,\ \lambda_4 < 0$*

1. *If $a_4 < \frac{1}{\hat{B}}$, then $\Delta < 0$ and the equilibrium point of the E is saddle–node.*
2. *If $a_4 > \frac{1}{\hat{B}}$, $|\arg(\lambda_3)| > \frac{\alpha\pi}{2}$ and $|\arg(\lambda_4)| > \frac{\alpha\pi}{2}$ then the equilibrium E is asymptotically stable.*

4 Data fitting and estimation of parameters

The model (5) contains 17 parameters that should be determined. We obtain the values of some of the parameters from available resources. However, their values may vary in different types of cancers and in each patient. We use data fitting to estimate the values of parameters that are unknown. To do this, we will use the data obtained from [4]. We perform the estimation of the parameters based on the data of CLL109 patient that have already been studied in [5]. By running the `fminsearch`, a MATLAB function, which is a least squares algorithm, we determine the values of the parameters for the values of $\alpha \in (0, 1)$. The estimated values of the parameters are presented in Table 1.

5 Numerical Simulation

The fractional form of the Adams-Bashforth-Moulton method will be used to numerically solve the system. We use the FDE12 function for the different values of the estimated

parameters corresponding to the patient CLL109.
For patient CLL109 with initial value $(B_0, N_0, C_0, H_0) = (38, 517, 2, 1)$, the solutions of the system (5) for different values of α is shown in Figure 1 (top-left). As shown in the curves with decreasing the value of α, the convergence rate also decreases. For $\alpha = 1$, the number of B–CLL cell population reaches 780 $cells/\mu l$ by 145th day, but then decreases. For values of $\alpha < 1$, this occurs later.

6 Bifurcation

The system (5) shows a substantial sensitivity to the values of certain parameters, that is, changing the values of the parameters leads to significant changes in the behavior of the system. With change in one parameter, the number of equilibrium points in the system may decrease or increase, or location and the stability of these points may change. In Section 3, the stability of equilibrium points was theoretically discussed. In this section, we interpret the bifurcation analysis of several parameters.
For CLL109 patient, we performed the bifurcation analysis with respect to the parameter r (recruitment rate of H cell by B–CLL cells). Figure 1 illustrates that *bistability* occurs with changing the value of the parameter r. In Figure 1 (right-top), there are two stable branches and one unstable branch. For $r \leq 0.00206$, the B–CLL cell population converges to a larger stable equilibrium point. For the values of $0.00206 < r \leq 0.00279$, the system contains an unstable equilibrium point and a small stable equilibrium point, in addition to the large equilibrium point. For values of $r < 0.00279$, there is a jump to a small stable equilibrium point, which shows the existence of *Hysteresis* in the model.
Hysteresis is a phenomenon that shows the dependence of the current state of a system on its previous state (pathway of changes). According to some evidence, re-activation of T-cells that have already experienced active state requires a lower signal threshold. In order to initiate the functions of the T-cell effector, first, a Reticular Activating System (RAS) should be activated. It is also required to activate RAS at high levels to launch T cell receptors [1]. This example illustrates the importance and application of the hysteresis phenomenon in the immune system. The use of memory-containing models such as model (5) clearly reveals the existence of this phenomenon.
This phenomenon has many uses in various fields such as physics, chemistry, engineering, biology and economics.
As can be seen in the Figure 1(bottom), the solutions of B–CLL cell population for values $r = 0.005$ and $r = 0.0014$ converges to stable equilibrium points $B_e = 3.77$ and $B_e = 2366$, respectively.
Determining the parameters that lead to decrease in B–CLL cell population and their bifurcation points are very effective in the treatment phase of the disease. In fact, in treatments need to move the effective parameter values across the bifurcation values.

References

[1] A. K. Abbas, A. H. Lichtman and S. Pillai, *Cellular and Molecular Immunology*, Saunders, 7 edition, 2007.

[2] R. J. DeBoer, H. Mohri, D. D. Ho and A. S. Perelson, Turnover rates of B cells, T cells, and NK cells in simian immunodeciency virus-infected and uninfected rhesus macaques, *The Journal of Immunology*, 170 (2003), 2479–2487.

Table 1: Parameter Values in the Model.

Parameter	Description	Units	Value	Source
s_B	Source term for new B–CLL	(cells/μl) /day	[6,2475]	Fitted
$(a-b)$	B-CLL density change rate	1/day	[0.002,130]	Estimated
c	B–CLL kill by NK cells	1/(cells/μl)(day)	[1.05e-06,9.56e-04]	Estimated
d	B–CLL kill by T cells	1/(cells/μl)(day)	[3.07e-05,0.1320]	Estimated
s_N	Source term for new NK cells	1/(cells/μl)(day)	[1.56,7.632]	[5]
e	Natural death rate of NK cells	1/day	0.0159	[2, 5]
f	Deactivation rate of NK by B–CLL	1/(cells/μl)(day)	0.0001	[5]
s_C	Source term for new T cells	1/(cells/μl)(day)	[0,12.8]	Estimated
g	Natural death rate of T cells	1/day	[0.000102,0.04]	Estimated
i	Inactivation of T-cells by B–CLL	1/day	0.0001	[5]
k	Scaling factor<1	–	0.55	Estimated
r	Activation rate of H cells by B–CLL	1/day	[0.0015,0.0075]	Estimated
m	Power of B–CLL in term of T and H rec.	–	2	Estimated
η	Half–saturation level in the rec.	cells/μl	10000	Estimated
s_H	Source term for new H cells	1/(cells/μl)(day)	[3.9,16.9]	[3]
j	Natural death rate of H cells	1/day	[0.00135,0.0338]	[3]

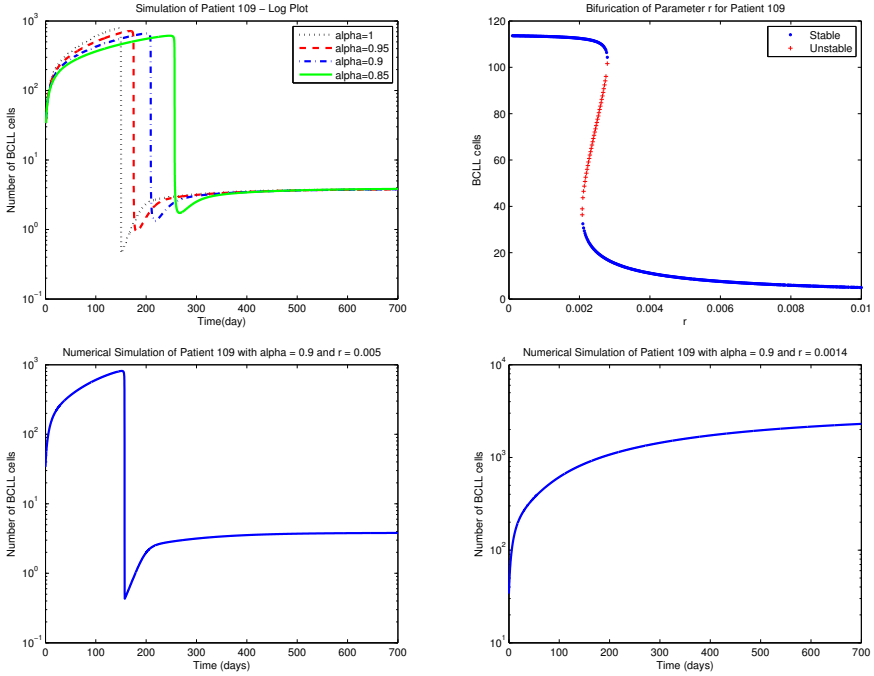


Figure 1: Numerical simulation result of cells population for Patient 109 (left-top), bifurcation analysis for parameter r (right-top), simulation of B–CLL for Patient 109 with $\alpha = 0.9$ and $r = 0.005$ (left-bottom), simulation of B–CLL for Patient 109 with $\alpha = 0.9$ and $r = 0.0014$ (right-bottom).

[3] M. Hellerstein, M. Hanley, D. Cesar, S. Siler, C. Papageorgopoulos, E. Wieder, et al., Directly measured kinetics of circulating T lymphocytes in normal and HIV-1-infected humans, *Nature Medicine*, 5 (1999), 83–89.

[4] B. T. Messmer, D. Messmer, S. L. Allen, J. E. Kolitz, P. Kudalkar, D. Cesar, et al., In vivo measurements document the dynamic cellular kinetics of chronic lymphocytic leukemia B cells, *The Journal of Clinic Investig.*, 115 (2005), 755–764.

[5] S. Nanda, L. de Pillis, and A., Radunskaya , B cell chronic lymphocytic leukemia: a model with immune response. *Discrete Continuous Dyn. Syst. Ser. B* 18, (2013) 1053–1076.

[6] I. Petras, Fractional-Order Nonlinear Systems: Modeling, Analysis and Simulation. Springer, Berlin, 2011.

Operational equations systems[1]

Javad Farokhi Ostad*

Department of Basic Sciences, Birjand University of Technology, Birjand, Iran

Abstract

In this paper, using operator matrices representation, we investigate the explicit solution of the operator equations and operational equations systems. In the general setting of the adjointable modular operators between Hilbert C-modules framework, this solution is expressed in terms of the Moore-Penrose inverses of the operators. The obtained results extend and generalize some known operator equations studied previously by a number of mathematicians.

Keywords: Hilbert C-module, Moore-Penrose inverse, Operator equation, Operator matrix

Mathematics Subject Classification [2010]: 15A03, 15A23, 15B36

1 Introduction

The solving operator equations has recently been found in the work of many researchers and even finding the exact solution for these equations has been interest. Since the matrix space is finite dimensional, so if we extend these equations to the space of operators, then also it will be able to solve equations in infinite dimensional cases. It may be assumed that the applied equations have a matrix representation, and the generalization of equations to higher spaces is studied merely for the abstract mathematical discussion. But there are many cases of functional equations appeared in physics, for examples Kadomtsev-Petviashvili operator equation$v_{xt} = \frac{1}{4}(v_{xxx} + 6v_x^2)_x + \frac{3}{4}v_{yy} + \frac{3}{2}(v_y v_x v_x v_y)$, which is different from the finite case. Similar to the matrix form, here too many of the solutions are found using general inverses, and in particular the Moore-Penrose inverse. In this paper, we try to solve a some of equations in the Hilbert C^*-modules. As we know, Hilbert C^*-modules are extension of Hilbert spaces with the same properties, nevertheless there exist some basic differences. However, some well known properties of Hilbert spaces like Pythagoras' equality, self-duality, and even decomposition into orthogonal complements do not hold in the framework Hilbert modules. Suppose that $\mathcal{A}$ is an arbitrary C*-algebra and $\mathcal{X}$ is a linear space which is a right $\mathcal{A}$-module and the scalar multiplication satisfies $\lambda(xa) = x(\lambda a) = (\lambda x)a$ for all $x \in \mathcal{X}$, $a \in \mathcal{A}, \lambda \in \mathbb{C}$. The $\mathcal{A}$-module $\mathcal{X}$ is called a *pre-Hilbert $\mathcal{A}$-module* if there exists an $\mathcal{A}$-valued map $\langle .,. \rangle : \mathcal{X} \times \mathcal{X} \to \mathcal{A}$ with the following properties:

(i) $\langle x, y + \lambda z \rangle = \langle x, y \rangle + \lambda \langle x, z \rangle$; for all $x, y, z \in E, \lambda \in \mathbb{C}$,

[1]Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

*Speaker. Email address: j.farrokhi@birjandut.ac.ir

(ii) $\langle x, ya\rangle = \langle x, y\rangle a$; for all $x, y \in \mathcal{X}$ and $a \in A$,

(iii) $\langle x, y\rangle^* = \langle y, x\rangle$; for all $x, y \in \mathcal{X}$,

(iv) $\langle x, x\rangle \geq 0$ and $\langle x, x\rangle = 0$ if and only if $x = 0$.

The $\mathcal{A}$-module $\mathcal{X}$ is called a *Hilbert C^*-module* if $\mathcal{X}$ is complete with respect to the norm $\|x\| = \|\langle x, x\rangle\|^{1/2}$. Throughout this paper we assume that $\mathcal{A}$ is an arbitrary C^*-algebra. The notations $Ker(\cdot)$ and $ran(\cdot)$ stand for kernel and range of operators, respectively. The set of all bounded adjointable operator from $\mathcal{X}$ to $\mathcal{Y}$ is shown by $\mathcal{L}(\mathcal{X}, \mathcal{Y})$.

Theorem 1.1. ([4, Theorem 3.2]) Suppose that $A \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ has closed range. Then $A^* \in \mathcal{L}(\mathcal{Y}, \mathcal{X})$ has closed range, and

[(i)]

1. $\ker(A)$ is orthogonally complemented in $\mathcal{X}$, with $(\ker(A))^{\perp} = \mathrm{ran}(\mathrm{A}^*)$.
2. ran(A) is orthogonally complemented in $\mathcal{Y}$, with $(\mathrm{ran}(\mathrm{A}))^{\perp} = \ker(\mathrm{A}^*)$.

A generalized inverse of $A \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ is an operator $A^{\times} \in \mathcal{L}(\mathcal{Y}, \mathcal{X})$ such that

$$A\,A^{\times}A = A \quad \text{and} \quad A^{\times}A\,A^{\times} = A^{\times}. \tag{1}$$

If the first part hold it is called inner inverse and if both equations hold $A^{\times}$ called outer inverse of A.

Definition 1.2. Let $A \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$. The Moore-Penrose inverse $A^{\dagger}$ of A is unique solution X of the following equvalent operational systems:

$$\begin{cases} AXA = A \\ X\,A\,X = X \\ (A\,X)^* = A\,X \\ (X\,A)^* = XA \end{cases}$$

or

$$\begin{cases} XAA^* = A^* \\ XX^*A^* = X \end{cases}$$

Moreover, we know that a bounded adjointable operator may admit an unbounded operator as its Moore-Penrose, see [4] for more detailed information. A matrix form of a bounded adjointable operator $A \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ can be induced by some natural decompositions of Hilbert C^*-modules. Indeed, if $\mathcal{M}$ and $\mathcal{N}$ are closed orthogonally complemented submodules of $\mathcal{X}$ and $\mathcal{Y}$, respectively, and $\mathcal{X} = \mathcal{M} \oplus \mathcal{M}^{\perp}$, $\mathcal{Y} = \mathcal{N} \oplus \mathcal{N}^{\perp}$, then T can be written as the following 2×2 matrix

$$A = \begin{bmatrix} A_1 & A_2 \\ A_3 & A_4 \end{bmatrix} \tag{2}$$

where, $A_1 \in \mathcal{L}(\mathcal{M}, \mathcal{N})$, $A_2 \in \mathcal{L}(\mathcal{M}^{\perp}, \mathcal{N})$, $A_3 \in \mathcal{L}(\mathcal{M}, \mathcal{N}^{\perp})$ and $A_4 \in \mathcal{L}(\mathcal{M}^{\perp}, \mathcal{N}^{\perp})$. Note that $P_{\mathcal{M}}$ denotes the projection corresponding to $\mathcal{M}$.

In fact $A_1 = P_{\mathcal{N}}AP_{\mathcal{M}}$, $A_2 = P_{\mathcal{N}}A(1 - P_{\mathcal{M}})$ $A_3 = (1 - P_{\mathcal{N}})AP_{\mathcal{M}}$ and $A_4 = (1 - P_{\mathcal{N}})A(1 - P_{\mathcal{M}})$. The proof of the next important and widely used theorem can be found in many articles including [2] and [5].

Theorem 1.3. *Let $A \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ has closed range. Then A has the following matrix decomposition with respect to the orthogonal decompositions of submodules*

(a) If $\mathcal{X} = \mathrm{ran(A^*)} \oplus \mathrm{ker(A)}$ and $\mathcal{Y} = \mathrm{ran(A)} \oplus \mathrm{ker(A^*)}$, then

$$A = \begin{bmatrix} A_1 & 0 \\ 0 & 0 \end{bmatrix} : \begin{bmatrix} \mathrm{ran(A^*)} \\ \mathrm{ker}(A) \end{bmatrix} \to \begin{bmatrix} \mathrm{ran(A)} \\ \mathrm{ker}(A^*) \end{bmatrix},$$

where A_1 is invertible. Moreover,

$$A^\dagger = \begin{bmatrix} A_1^{-1} & 0 \\ 0 & 0 \end{bmatrix} : \begin{bmatrix} \mathrm{ran(A)} \\ \mathrm{ker}(A^*) \end{bmatrix} \to \begin{bmatrix} \mathrm{ran(A^*)} \\ \mathrm{ker}(A) \end{bmatrix}.$$

(b) If $\mathcal{X} = \mathcal{X}_1 \oplus \mathcal{X}_2$ and $\mathcal{Y} = \mathrm{ran(A)} \oplus \mathrm{ker(A^*)}$, then:

$$A = \begin{bmatrix} A_1 & A_2 \\ 0 & 0 \end{bmatrix} : \begin{bmatrix} \mathcal{X}_1 \\ \mathcal{X}_2 \end{bmatrix} \to \begin{bmatrix} \mathrm{ran(A)} \\ \mathrm{ker}(A^*) \end{bmatrix}, \tag{3}$$

and in this case,

$$A^\dagger = \begin{bmatrix} A_1^* D^{-1} & 0 \\ A_2^* D^{-1} & 0 \end{bmatrix}, \tag{4}$$

where $D = A_1 A_1^* + A_2 A_2^* \in \mathcal{L}(\mathrm{ran(A)})$ is positive and invertible.

(c) If $\mathcal{X} = \mathrm{ran(A^*)} \oplus \mathrm{ker(A)}$ and $\mathcal{Y} = \mathrm{ran(A)} \oplus \mathrm{ker(A^*)}$, then:

$$A = \begin{bmatrix} A_1 & 0 \\ A_2 & 0 \end{bmatrix} : \begin{bmatrix} \mathrm{ran(A^*)} \\ \mathrm{ker}(A) \end{bmatrix} \to \begin{bmatrix} \mathcal{Y}_1 \\ \mathcal{Y}_2 \end{bmatrix}, \tag{5}$$

and

$$A^\dagger = \begin{bmatrix} \mathfrak{D}^{-1} A_1^* & \mathfrak{D}^{-1} A_2^* \\ 0 & 0 \end{bmatrix}, \tag{6}$$

where $\mathfrak{D} = A_1^* A_1 + A_2^* A_2 \in \mathcal{L}(\mathrm{ran(A^*)})$ is positive and invertible.

Recall that if $A \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ has closed range, then $AA^\dagger = P_{\mathrm{ran(A)}}$ and $A^\dagger A = P_{\mathrm{ran(A^*)}}$.

An operator A with closed range is called EP if $\mathrm{ker}(A) = \mathrm{ker}(A^*)$. It is easy to see that,

$$A \text{ is EP} \Leftrightarrow \mathrm{ran(A)} = \mathrm{ran(A^*)} \Leftrightarrow \mathrm{AA^\dagger} = \mathrm{A^\dagger A}.$$

The interested reader, for more detail and informations of this section, can be referred to [6].

2 Main results

In this section, we investigate the explicit solution of the operational equations and also modular operator equation systems.

Theorem 2.1. *Let $\mathcal{X}$ be Hilbert C^*-module and A and $B \in \mathcal{L}(X)$ have closed ranges. Then, $AA^\dagger B = A = BA^\dagger A$ if and only if $AA^* = AB^*, \quad A^*A = A^*B$.*

Theorem 2.2. *Let $\mathcal{X}$ be Hilbert C^*-module and A, B and $C \in \mathcal{L}(X)$ have closed ranges, with the following factorization $A = BC$. Then A is idempotent if and only if $CB = I$.*

Corollary 2.3. *Let $\mathcal{X}$ be Hilbert C^*-module and $A \in \mathcal{L}(X)$ has inner inverse, then the equation $X(AX - I) = 0$ has unique solution.*

Theorem 2.4. *Let $\mathcal{X}$ be Hilbert C^*-module and A, B and $C \in \mathcal{L}(X)$ have closed ranges. If A, B is self-adjonit operators, then the modular operator equation $AXB - X = C$ has unique solution.*

Theorem 2.5. *Let $\mathcal{X}$ be Hilbert C^*-module and A, B, D and $E \in \mathcal{L}(X)$ have closed ranges. The modular operator equations system $\begin{cases} AX = B, \\ XD = E \end{cases}$ have common solution if and only if $AE = BD$. Moreover, in this case, $X = X_0 + (I - A^\dagger A)Y(I - DD^\dagger)$ for arbiterary $Y \in \mathcal{L}(X)$.*

Theorem 2.6. *Let $\mathcal{X}$ be Hilbert C^*-module and A, B and $D \in \mathcal{L}(X)$ have closed ranges. The modular operator equation $AXB = D$ has unique solution if and only if $AA^\dagger DB^\dagger B = D$. Moreover, in this case, $X = A^\dagger DB^\dagger + Y - A^\dagger AYBB^\dagger$, for arbiterary $Y \in \mathcal{L}(X)$.*

Example 2.7. Let $\mathcal{S}$ = the real space $L^2[0, 2\pi]$ of real valued functions and $\mathcal{S}^1$ = the absolutly continuous functions $A(t), 0 \leq t \leq \pi$, whose derivatives A' are in $\mathcal{S}$; and $\mathcal{S}^2 = \{A \in \mathcal{S}^1; A' \in \mathcal{S}^1\}$. If L be the differential operator with $D(L) = \{A \in \mathcal{S}^1; A(0) = A(2\pi) = 0\}$, then the equation $X(I - X^*X^\dagger) = 0$ has unique solution.

Theorem 2.8. *Let $\mathcal{X}$ be Hilbert C^*-module and $A, B \in \mathcal{L}(X)$ have closed ranges, whit $Ran(B) \subseteq Ran(A)$. If $A^\dagger B$ is invertible, then $(A + B)^\dagger = (I + A^\dagger B)^{-1} A^\dagger$.*

Proof. Since $Ran(B) \subseteq Ran(A)$, $A + B = A + AA^\dagger B = A(I + A^\dagger B)$. So, it is suffices to show that, $A, (I + A^\dagger B)$ have the reverse order law property. i.e. $(A(I + A^\dagger B))^\dagger = (I + A^\dagger B)^{-1} A^\dagger$.

From the fact that $Ran(A^\dagger A(I + A^\dagger B)) \subseteq Ran(I + A^\dagger B)$ and $Ran((I + A^\dagger B)(I + A^\dagger B)^\dagger) \subseteq Ran(A^\dagger)$, which is complete the proof. □

References

[1] D. S. Djordjević and N. Č. Dinčić, Reverse order law for the Moore–Penrose inverse, *J. Math. Anal. Appl.* 361 (2010) 252-261.

[2] J. Farokhi-Ostad and A. Janfada, Product of Ep operators on Hilbert C^*-modulrs, *Sahand Communications in Mathematical Analysis (SCMA)*, (10) (1) (2018), pp. 61-71.

[3] J. Farokhi-Ostad and A. Janfada, On closed range C^*-modular operators, *The Australian Journal of Mathematical Analysis and Applications (AJMAA)*, (15) (2) (2018), pp. 1-9.

[4] E. C. Lance, *Hilbert C*-Modules*, LMS Lecture Note Series 210, Cambridge Univ. Press, 1995.

[5] M Mohammadzadeh Karizaki, Z. N. Moghani, M. Khanehgir, Explicit solution to the operator equation $AXD + FX^*B = C$ Over Hilbert C^*-modules, *Journal of Mathematical Analysis* 10 (1) (2019), 52-64.

[6] M. Vosough and M.S. Moslehian, *Operator and Matrix Equations*, Ph.d thesis, Ferdowsi University of Mashhad, (2017).

[7] Q. Xu and L. Sheng, Positive semi-definite matrices of adjointable operators on Hilbert C*-modules, *Linear Algebra Appl.* ,428 (2008), 992-1000.

A remark on Smith's determinant[1]

Mehdi Hassani*

Department of Mathematics, University of Zanjan, University Blvd., Zanjan 45371-38791, Iran

Abstract

In this paper we obtain a full asymptotic expansion for the logarithm of the generalized Smith's determinant $\Delta_k(n) = \det\left[(\gcd(i,j))^k\right]_{1\leqslant i,j\leqslant n}$ where $\gcd(i,j)$ denotes the greatest common divisor of i and j. For any integer $k \geqslant 2$ we obtain the following Stirling type approximation

$$\Delta_k(n) = \left(\frac{n}{e}\right)^{kn} \beta_k^n \sqrt{(2\pi n)^k}\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right),$$

where β_k is an absolute constant defined by

$$\beta_k = \prod_p \left(1-\frac{1}{p^k}\right)^{\frac{1}{p}},$$

and p runs over all primes.

Keywords: Determinants, Arithmetic functions, Prime numbers
Mathematics Subject Classification [2010]: 15A15, 11C20, 11A25

1 Introduction

In 1875 H.J.S. Smith[2] [4] considered the determinant of the matrix $[a_{ij}]_{1\leqslant i,j\leqslant n}$ with $a_{ij} = \gcd(i,j)$, the greatest common divisor of i and j. He proved that

$$\det\left[\gcd(i,j)\right]_{1\leqslant i,j\leqslant n} = \prod_{m=1}^{n} \varphi(m), \tag{1}$$

where $\varphi(m)$ denotes the Euler function of m, counting the number of positive integers not exceeding m and coprime to m. The above mentioned determinant is known as *Smith's determinant*. Among several generalizations of Smith's determinant, it is known [3] that if f is an arithmetic function (a function defined over $\mathbb{N}$) then

$$\det\left[f(\gcd(i,j))\right]_{1\leqslant i,j\leqslant n} = \prod_{m=1}^{n} \sum_{d\mid m} \mu(d)\, f\left(\frac{m}{d}\right), \tag{2}$$

[1] Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

*Speaker. Email address: mehdi.hassani@znu.ac.ir

[2] Henry John Stephen Smith, 2 November 1826, Dublin, Ireland – 9 February 1883, Oxford, England

where $\mu(d)$ denotes the Möbius function of d, which is 1 if $d=1$, is $(-1)^k$ if d is equal to the product of k distinct primes, and is 0 otherwise. In this paper we are motivated by the asymptotic growth of generalized Smith's determinant (2) for $f(n)=n^k$ with $k\in\mathbb{N}$. The case $k=1$, admitting Smith's determinant (1), has been studied in [1], where we have proved that

$$\log\left(\det\left[\gcd(i,j)\right]_{1\leqslant i,j\leqslant n}\right)=n\log n+(\alpha_1-1)n+\frac{1}{2}\log n+\mathcal{O}(\log\log n),$$

such that $\alpha_1=\sum_p\frac{1}{p}\log(1-\frac{1}{p})$ with p running over all primes, is an absolute constant. In this paper we consider the case $k\geqslant 2$ by proving the following result.

Theorem 1.1. *Let $k\geqslant 2$ is fixed integer, and $\Delta_k(n)=\det\left[(\gcd(i,j))^k\right]_{1\leqslant i,j\leqslant n}$. Define the absolute constant α_k by*

$$\alpha_k=\sum_p\frac{1}{p}\log\left(1-\frac{1}{p^k}\right),$$

where p runs over all primes. Then, as $n\to\infty$ we have

$$\log\Delta_k(n)=kn\log n+(\alpha_k-k)\,n+\frac{k}{2}\log n+k\log\sqrt{2\pi}+\sum_{1\leqslant j\leqslant\frac{k}{2}}\frac{kB_{2j}}{(2j)(2j-1)n^{2j-1}}+\mathcal{O}\Big(\frac{1}{n^{k-1}}\Big),$$

where B_i denotes the i-th Bernoulli number.

By taking exponent we obtain the following Stirling type approximation for $\Delta_k(n)$ for each integer $k\geqslant 2$.

Corollary 1.2. *Let $k\geqslant 2$ be a fixed integer, and $\Delta_k(n)$ is defined as in Theorem 1.1. Then, as $n\to\infty$ we have*

$$\Delta_k(n)=\Big(\frac{n}{e}\Big)^{kn}\,\beta_k^n\,\sqrt{(2\pi n)^k}\left(1+\mathcal{O}\Big(\frac{1}{n}\Big)\right),$$

where β_k is an absolute constant defined by

$$\beta_k=\prod_p\left(1-\frac{1}{p^k}\right)^{\frac{1}{p}},$$

and p runs over all primes.

2 Proofs

Proof of Theorem 1.1. By using the relation (2) we have

$$\Delta_k(n)=\prod_{m=1}^{n}m^k\,g_k(m)=n!^k\prod_{m=1}^{n}g_k(m),$$

where $g_k(m)=\sum_{d|m}\mu(d)d^{-k}$. Since g_k is multiplicative, we get

$$g_k(m)=\prod_{p^a\|m}g_k(p^a)=\prod_{p^a\|m}\left(1-\frac{1}{p^k}\right)=\prod_{p|m}\left(1-\frac{1}{p^k}\right),$$

where $p^a \| m$ means that a is the largest power of the prime p for which $p^a | m$. Thus,

$$\Delta_k(n) = n!^k \prod_{m=1}^{n} \prod_{p|m} \left(1 - \frac{1}{p^k}\right).$$

We take logarithm to get

$$\log \Delta_k(n) = k \log n! + \sum_{m=1}^{n} \sum_{p|m} \log \left(1 - \frac{1}{p^k}\right). \tag{3}$$

Stirling's approximation [2] for $\log n!$ asserts that given any positive integer r, as $n \to \infty$ we have

$$\log n! = n \log n - n + \frac{1}{2} \log n + \log \sqrt{2\pi} + \sum_{j=1}^{r} \frac{B_{2j}}{(2j)(2j-1)n^{2j-1}} + \mathcal{O}\Big(\frac{1}{n^{2r+1}}\Big). \tag{4}$$

To approximate the double sum in (3), we change the order of summations to get

$$\begin{aligned}
\sum_{m=1}^{n} \sum_{p|m} \log \left(1 - \frac{1}{p^k}\right) &= \sum_{p \leqslant n} \log \left(1 - \frac{1}{p^k}\right) \sum_{\substack{m \leqslant n \\ p|m}} 1 = \sum_{p \leqslant n} \log \left(1 - \frac{1}{p^k}\right) \left\lfloor \frac{n}{p} \right\rfloor \\
&= \sum_{p \leqslant n} \log \left(1 - \frac{1}{p^k}\right) \left(\frac{n}{p} + \mathcal{O}(1)\right) \\
&= n \sum_{p \leqslant n} \frac{1}{p} \log \left(1 - \frac{1}{p^k}\right) + \mathcal{O}\left(\sum_{p \leqslant n} \log \left(1 - \frac{1}{p^k}\right)\right) \\
&= \alpha_k \, n + n \sum_{p > n} \frac{1}{p} \log \left(1 - \frac{1}{p^k}\right)^{-1} + \mathcal{O}\left(\sum_{p \leqslant n} \log \left(1 - \frac{1}{p^k}\right)\right)
\end{aligned}$$

Since $-\log(1-t) \sim t$ as $t \to 0$, we have

$$\sum_{p>n} \frac{1}{p} \log \left(1 - \frac{1}{p^k}\right)^{-1} \ll \sum_{p>n} \frac{1}{p^{k+1}} \ll \int_n^{\infty} \frac{dx}{x^{k+1}} \ll \frac{1}{n^k},$$

where $f \ll g$ has same meaning as $f = \mathcal{O}(g)$. Also,

$$\sum_{p \leqslant n} \log \left(1 - \frac{1}{p^k}\right) \ll \sum_{p \leqslant n} \frac{1}{p^k} \ll \int_2^n \frac{dx}{x^k} \ll \frac{1}{n^{k-1}}.$$

Thus, we obtain

$$\sum_{m=1}^{n} \sum_{p|m} \log \left(1 - \frac{1}{p^k}\right) = \alpha_k \, n + \mathcal{O}\left(\frac{1}{n^{k-1}}\right). \tag{5}$$

We take $r = \lfloor \frac{k}{2} \rfloor$ in (4), and we note that $2\lfloor \frac{k}{2} \rfloor + 1 \geqslant k - 1$. Considering (3) and (5) completes the proof. □

Proof of Corollary 1.2. Theorem 1.1 implies that

$$\log \Delta_k(n) = kn \log n + (\alpha_k - k)\, n + \frac{k}{2} \log n + k \log \sqrt{2\pi} + \mathcal{O}\Big(\frac{1}{n}\Big).$$

Taking exponent and considering the approximation $e^{\mathcal{O}(\frac{1}{n})} = 1 + \mathcal{O}(\frac{1}{n})$ completes the proof. □

3 Conclusion

Smith's determinant and its generalization are interesting examples of number theoretic determinants. Since the values of such determinants usually are given in terms of arithmetic functions, it is interesting to approximate true order of them. In this paper we obtain a Stirling type approximation for the generalization of Smith's determinant.

Acknowledgment

I express my gratitude to the anonymous referee(s) for careful reading of the manuscript and giving the many valuable suggestions and corrections, which improved the presentation of the paper.

References

[1] M. Hassani, Approximation of the growth of a number theoretic determinant, In Proceedings of the 6th Seminar on Linear Algebra and its Applications, 2011.

[2] F.W.J. Olver, *Asymptotics and Special Functions*, Academic Press, New York, 1974.

[3] H.N. Shapiro, *Introduction to the Theory of Numbers*, Dover, New York, 2008.

[4] H.J.S. Smith, On the value of a certain arithmetical determinant, *Proc. Lond. Math. Soc.*, 7 (1875/76), 208-212.

Conditions of reflexivity for multiplication operators on Banach function spaces on a plane domain[1]

Parastoo Heiatian Naeini*

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

Abstract

In this paper, we give conditions under which the powers of the multiplication operator M_z are reflexive on a Banach space of functions analytic on a plane domain.

Keywords: Multiplication operators, Bounded point evaluation, Reflexive operator
Mathematics Subject Classification [2010]: 47B37, 47A25

1 Introduction

For any set E and any function $f : E \to \mathbb{C}$, define $\|f\|_E$ by

$$\|f\|_E = sup\{|f(z)| : z \in E\}.$$

If B is bounded domain in the plane, then the Caratheodory hull($\mathbb{C}-$hull) of B is the complement of the closure of the unbounded componet of the complement of the closure of B. The $\mathbb{C}-$hull of B is denoted by B^*. Intuitively, B^* can be described as the interior of the outer boundary of B, and in analytic terms it can be defind as the interior of the set of all points z_0 in the plane such that $|p(z_0)| \leq sup\{|p(z)| : z \in B\}$ for all polynomails p. The componets of B^* are simply connected; in fact, one can easily see the each of these components has a connected complement. The componets of B^* that contains B is denoted by B_1. Note that for all polynomials p, $\|p\|_B = \|p\|_{B_1}$. Since B_1 is a Caraththeodory domain, so by the Farrel-Rubel-Shields theorem [2], each bounded analytic function on B_1 can be approximated by a sequence of polynomials pointwise boundedly.

For the algebra $\mathcal{B}(\mathcal{X})$ of all bounded linear operators on a Banach space $\mathcal{X}$, the weak operator topology (WOT) is the one in which a net A_α converges to A if $A_\alpha \to Ax$ weakly, $x \in \mathcal{X}$. Also, the strong operator topology (SOT) is the one in which a net A_α converges to A if $A_\alpha \to Ax$, $x \in \mathcal{X}$.

Recall that if $A \in \mathcal{B}(\mathcal{X})$, then $Lat(A)$ is by definition the lattice of all invariant subspaces of A, and $AlgLat(A)$ is the algebra of all operators B in $\mathcal{B}(\mathcal{X})$ such that $Lat(A) \subset Lat(B)$. An operator A in $\mathcal{B}(\mathcal{X})$ is said to be reflexive if $AlgLat(A) = W(A)$, where $W(A)$ is the smallest subalgebra of $\mathcal{B}(\mathcal{X})$ that contains A and the identity I is closed in the weak operator topology.

In [1], it is shown that any powers of the operator M_z is reflexive on Banach spaces of formal Laurent series. Also, reflexivity of the operator M_z on Hilbert function spaces

[1]Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University
*Speaker. Email address: p.heiatian.n@gmail.com

has been investigated in [3,5] and for the case of Banach function spaces, see [6]. Here we give some sufficient conditions so that the powers of the operator M_z, acting on Banach function spaces becomes reflexive. As, usual, for a good basic source of reflexivity we refer to [1].

Consider a Banach space $\mathcal{X}$ of function analytic on a plane domain G, such that for each $\lambda \in G$, the linear functional e_λ of evaluation at λ (defined by $e_\lambda(f) = f(\lambda)$) is bounded on $\mathcal{X}$. A complex-valued function φ on G for which $\varphi f \in \mathcal{X}$ for every $f \in \mathcal{X}$ is called a multiplier of $\mathcal{X}$ and the collection of all these multipliers is denoted by $\mathcal{M}(\mathcal{X})$. Each multiplier φ on $\mathcal{X}$ determines a multiplication operator M_φ on $\mathcal{X}$ by $M_\varphi f = \varphi f$, $f \in \mathcal{X}$. It is well-known that each multiplier is bounded analytic function on G, in fact $\|\varphi\|_G \leq \|M_\varphi\|$. The notation $\|\varphi\|_\infty = \|M_\varphi\|$ is usually used for the norm of the operator M_φ.

By $H(G)$ and $H^\infty(G)$ we will mean respectively the set of analytic functions on a plane domain G and the set of bounded analytic functions on G. Also, by $\mathcal{P}(G)$ we mean the uniform closure in $C(G, \mathbb{C})$(the space of continuous functions from G into $\mathbb{C}$) of the polynomials. Note that $f \in \mathcal{P}(G)$ if and only if there exists a sequence of polynomials $\{p_n\}_n$ that converges uniformly to f on every compact subset of G.

2 Main results

We investigate the reflexivity of the powers of the multiplication operator M_z acting on a Banach function space.

Recal that a sequence $\{x_n\}_n$ in a Banach space $\mathcal{X}$ is called a Schauder basis of $\mathcal{X}$ if for every $x \in \mathcal{X}$ there is a unique sequence of scalars $\{a_n\}_n$ so that $x = \sum_n a_n x_n$. In this case, the closed linear span of $\{x_n\}_n$ in of all $\mathcal{X}$. Also, for every integer n, the linear functional x_n^* on $\mathcal{X}$ defined by $x_n^*(\sum_i a_i x_i) = a_i$ is a bounded linear functional. These functional $\{x_n^*\}_n$, which are characterized by the relation $x_n^*(x_m) = \delta_m(n)$, are called the biorthogonal functional associated to the basis $\{x_n\}$. in the weak* topology, $x^* = \sum_n x^*(x_n)x_n^*$ for $x^* \in \mathcal{X}^*$, and we have convergence in norm for every $x^* = \sum_n x^*(x_n)x_n^*$ if and inly if the sequence $\{x_n^*\}_n$ is a Schauder basis of $\mathcal{X}^*$. For this to happen, $\mathcal{X}^*$ must, in particular be separable. On the other hand this is always the case if $\mathcal{X}$ is reflexive.

From now on, let Ω be a domain in the complex plane such that Ω_1 is equal to the open unit disc $\mathbb{D}$. Also suppose that the Banach space $\mathcal{X}$ under consideration satisfy the following axioms:

Axiom (1). $\mathcal{X}$ is a subspace of the space of all analytic functions on Ω that are continuous on $\overline{\Omega}$.

Axiom (2). For each $\lambda \in \overline{\Omega}$, the linear functional of evaluation at λ, e_λ, is bounded on $\mathcal{X}$.

Axiom (3). The sequences $\{f_k\}_k$ and $\{f_k^*\}_k$ are Schauder basis for $\mathcal{X}$ and $\mathcal{X}^*$ respectively, where $f_k(z) = z^k$ for all integers k and $\{f_k^*\}_k$ is also the biorthogonal functionals associated to $\{f_k\}_k$.

For $h = \sum_n \widehat{h}(n)z^n \in H(\mathbb{D}) \cap \mathcal{M}(\mathcal{X})$ and $\omega \in \partial\mathbb{D}$, define h_ω by $h_\omega(z) = h(\omega z)$. Then $h_\omega = \sum_n \widehat{h_\omega}(n)z^n$ where $\widehat{h_\omega}(n) = \omega^n \widehat{h}(n)$ for all n. Note that $H(\mathbb{D}) \cap \mathcal{M}(\mathcal{X})$ is nonempty since $1, z \in H(\mathbb{D}) \cap \mathcal{M}(\mathcal{X})$.

Definition 2.1. We say that $H(\mathbb{D}) \cap \mathcal{M}(\mathcal{X})$ is bi-isometrically rotation invariant whenever $\varphi \in H(\mathbb{D}) \cap \mathcal{M}(\mathcal{X})$, then $\varphi_{e^{-i\theta}} \in H(\mathbb{D}) \cap \mathcal{M}(\mathcal{X})$, $\|\varphi\|_\infty = \|\varphi_{e^{-i\theta}}\|_\infty$ and $\|\varphi\| = \|\varphi_{e^{-i\theta}}\|$ for all $\theta \in \mathbb{R}$.

Furthermore, we assume that $\mathcal{X}$ holds in the following axioms:
Axiom (4). $z \in \mathcal{X}$ and $H(\mathbb{D}) \cap \mathcal{M}(\mathcal{X})$ is bi-isometrically rotation invariant.

The following Lemma extends a result obtained by Allen Shields [4] that have been proved only for the special case where $\mathcal{X}$ is $H^2(\beta)$, the Hilbert space of formal power series.

Lemma 2.2. *Let $\varphi \in H(\mathbb{D}) \cap \mathcal{M}(\mathcal{X})$. Then for the sequence $\{r_n = \sum_j \widehat{r}_n(j) z^j\}$ such that $\widehat{r}_n(j) = (1 - \frac{j}{n+1})\widehat{\varphi}(j)$ whenever $j = o, ..., n$ and is 0 otherwise, we have $M_{r_n} \to M_\varphi$ in the weak operator topology.*

Theorem 2.3. *If $\mathcal{P}(\Omega) \subset \mathcal{M}(\mathcal{X})$, then M_{z^k} is reflexive on $\mathcal{X}$ for all $k \geq 1$.*

In the proof of Theorem 2.3, we used the assumption $\mathcal{P}(\Omega) \subset \mathcal{M}(\mathcal{X})$ to show that $H^\infty(\Omega_1) \cap \mathcal{X} \subset \mathcal{M}(\mathcal{X})$. So the following corollary is an immediate consequence of the proof of Theorem 2.3.

Corollary 2.4. *If $z \in \mathcal{M}(\mathcal{X})$ and $H^\infty(\Omega_1) \cap \mathcal{X} \subset \mathcal{M}(\mathcal{X})$, then M_{z^k} is reflexive on $\mathcal{X}$ for all $k \geq 1$.*

Recall that M_z is called polynomially bounded on $\mathcal{X} \subset H(\Omega)$ if there exists $c > 0$ such that $\|p(M_z)\| \leq c\|p\|_\Omega$ for all pollynomials p.

Theorem 2.5. *If M_z is polynomially bounded on $\mathcal{X}$, then M_{z^k} is reflexive on $\mathcal{X}$ for all $k \geq 1$.*

Proof. Since M_z is polynomially bounded, there exists $c > 0$ such that $\|p(M_z)\| \leq c\|p\|_\Omega$ for all polynomials p. By corollary 2.4, it is sufficient to show that $H^\infty(\Omega_1) \cap \mathcal{X} \subset \mathcal{M}(\mathcal{X})$. For this, let $f \in H^\infty(\Omega_1) \cap \mathcal{X}$. By the Farrel-Rubel-Shields theorm, there exists a sequence $\{p_n\}$ of polynomials converging to f such that for all n, $\|p_n\|_\Omega = \|p_n\|_{\Omega_1} \leq d.$ for some $d > 0$. So we obtain

$$\|M_{p_n}\| \leq c\|p_n\|_{\Omega_1} \leq cd$$

for all n. Since $\mathcal{X}$ is reflexive, the unit ball of $\mathcal{X}$ is weakly compact. Therefore ball $B(\mathcal{X})$ is compact in the weak operator topology and so by passing to a subsequence, if necessary, we may assume that for some $A \in B(\mathcal{X})$, $M_{p_n} \to A$ in the weak operator topology. Using the fact that $M^*_{p_n} \to A^*$ in the weak operator topology and acting these operators on e_λ we obtain

$$p_n(\lambda)e_\lambda = M^*_{p_n} e_\lambda \to A^* e_\lambda$$

wekly. Since $p_n(\lambda) \to f(\lambda)$, we see that

$$A^* e_\lambda = f(\lambda)e_\lambda.$$

Because the closed linear span of $\{e_\lambda : \lambda \in \Omega\}$ is weak star dense in $\mathcal{X}^*$, we conclude that $A = M_f$ and so $f \in \mathcal{M}(\mathcal{X})$. Thus indeed $H^\infty(\Omega_1) \cap \mathcal{X} \subset \mathcal{M}(\mathcal{X})$.

□

References

[1] J. B. Conwey, *A Course in Operator Theory* Amer. Math. Soc. (2000).

[2] T. Gamelin, *Uniform Algebra*, (1984), (NY: Chelsea)

[3] k. Jahedi, *On the multiplication operators on analytic function spaces*, Iranian J Sci. Technol., 37(4) (2013) 449-452.

[4] A. L. Shields, *Weighted shift operators and analytic functions theory*, Math. Surveys, A. M. S. Providence 13 (1974) 49-128.

[5] B. Yousefi, *Multiplication operators on Hilbert spaces of analytic functions*, Archiv der Mathe- matik, 83(6) (2004) 536-539.

[6] B. Yousefi and A. Khaksari, *Multiplication operators on analytic functional spaces* Taiwanese J. Math., 13(4) (2009) 1159-1165.

.

Ub-majorization on $\mathbf{M}_{m,n}$ and its linear preservers[1]

Asma Ilkhanizadeh Manesh*

Department of Mathematics, Vali-e-Asr,
University of Rafsanjan, P.O. Box: 7713936417, Rafsanjan, Iran

Abstract

An n-by-n real matrix (not necessarily nonnegative) R is g-row balanced (generalized row balanced) if all its row sums are zero. Let A, $B \in \mathbf{M}_{n,m}$. Then A is said to be ub-majorized by B (denoted by $A \prec_{ub} B$) if $A = RB$, for some n-by-n upper triangular g-row balanced matrix R. We wish to find the structure of all (strong) linear preservers of $\prec_{ub}$ on $\mathbb{R}^n$ and strong linear preservers of $\prec_{ub}$ on $\mathbf{M}_{n,m}$.

Keywords: Doubly stochastic matrix, Matrix majorization, Row stochastic matrix
Mathematics Subject Classification [2010]: 15A04, 15A51

1 Introduction

Let $\mathbf{M}_{n,m}$ be the set of all n-by-m real matrices, and let $\mathbb{R}^n$ be the set n-by-1 real vectors. A matrix $R = [r_{ij}] \in \mathbf{M}_n$ is called g-row balanced if $\sum_{j=1}^{n} r_{ij} = 0$, for all i $(1 \leq i \leq n)$. The collection of all n-by-n upper triangular g-row balanced matrices is denoted by $\mathcal{R}_n^{ub}$. The standard basis of $\mathbb{R}^n$ is denoted by $\{e_1, \ldots, e_n\}$, and $e = (1, 1, \ldots, 1)^t \in \mathbb{R}^n$. Let $[T]$ be the matrix representation of a linear function $T : \mathbf{M}_{n,m} \to \mathbf{M}_{n,m}$ with respect to the standard basis. For a subset $\mathcal{A} \subset \mathbb{R}^n$ $\mathcal{B}(\mathcal{A}) := \{\sum_{\mathrm{i}=1}^{\mathrm{m}} \lambda_{\mathrm{i}} \mathrm{a}_{\mathrm{i}} \mid \mathrm{m} \in \mathbb{N},\ \sum_{\mathrm{i}=1}^{\mathrm{m}} \lambda_{\mathrm{i}} = 0,\ \mathrm{a}_{\mathrm{i}} \in \mathcal{A},\ \ \mathrm{i} \in \mathbb{N}_{\mathrm{m}}\}$. Let $A(n_1, \ldots, n_l | m_1, \ldots, m_k)$ be the submatrix of A obtained from A by deleting rows $n_1, \ldots, n_l$ and columns $m_1, \ldots, m_k$, let $A(n_1, \ldots, n_l)$ be the abbreviation of $A(n_1, \ldots, n_l | n_1, \ldots, n_l)$.
Let $\mathcal{V}$ be a linear space of matrices, T be a linear function on $\mathcal{V}$, and $\mathcal{R}$ be a relation on $\mathcal{V}$. The linear function T is said to preserve $\mathcal{R}$, if $\mathcal{R}(T\mathcal{X}, T\mathcal{Y})$ whenever $\mathcal{R}(\mathcal{X}, \mathcal{Y})$. Also, T is said to strongly preserve $\mathcal{R}$, if

$$\mathcal{R}(T\mathcal{X}, T\mathcal{Y}) \Leftrightarrow \mathcal{R}(\mathcal{X}, \mathcal{Y}).$$

Definition 1.1. For $A, B \in \mathbf{M}_{n,m}$, it is said that A is ub-majorized by B, and denoted by $A \prec_{ub} B$, if there exists $R \in \mathcal{R}_n^{ub}$ such that $A = RB$.

In this paper, the linear preservers and strong preservers of $\prec_{ub}$ on $\mathbb{R}^n$ and $\mathbf{M}_{n,m}$, respectively, are fully characterized. For some deeper discussions of majorization and linear preservers of majorization we refer the reader to [1]- [3].

[1] Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University
*Speaker. Email address: a.ilkhani@vru.ac.ir

An n-by-n real matrix (not necessarily nonnegative) A is *g-row stochastic* (generalized row stochastic) if all its row sums are one. The collection of all n-by-n upper triangular g-row stochastic matrices is denoted by $\mathcal{R}_n^{gut}$.

The present paper continues in two further sections. Section 2 is devoted to a study of $\prec_{ub}$ on $\mathbb{R}^n$. In this section an equivalent condition for ub-majorization on $\mathbb{R}^n$ is obtained and some preliminaries are presented. In particular, the structure of all linear functions $T : \mathbb{R}^n \to \mathbb{R}^n$ preserving (strongly preserving) ub-majorization are characterized. Section 3 is assigned to investigate this relation on $\mathbf{M}_{n,m}$. In this section the strong linear preservers of $\prec_{ub}$ on $\mathbf{M}_{n,m}$ is stated.

2 G-row balanced on $\mathbb{R}^n$

This section studies some facts of $\prec_{ub}$ that are necessary for studying the linear preservers of $\prec_{ub}$ on $\mathbb{R}^n$. Also, we characterize the (strong) linear preservers of this relation $T : \mathbb{R}^n \to \mathbb{R}^n$.

The following proposition gives an equivalent condition for ub-majorization on $\mathbb{R}^n$.

Proposition 2.1. *Let $x = (x_1, \ldots, x_n)^t$, $y = (y_1, \ldots, y_n)^t \in \mathbb{R}^n$. Then $x \prec_{ub} y$ if and only if for each i $(1 \leq i \leq n)$ $x_i \in \mathcal{B}\{\mathrm{y_i}, \ldots, \mathrm{y_n}\}$.*

Now we assert some preliminaries to express our main results.

Lemma 2.2. *Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be a linear preserver of $\prec_{ub}$. Assume $S : \mathbb{R}^{n-k} \to \mathbb{R}^{n-k}$ is a linear function such that $[S] = [T](1, 2, \ldots, k)$. Then S preserves $\prec_{ub}$ on $\mathbb{R}^{n-k}$.*

Proof. Let $x' = (x_{k+1}, \ldots, x_n)^t, y' = (y_{k+1}, \ldots, y_n)^t \in \mathbb{R}^{n-k}$, and let $x' \prec_{ub} y'$. Then $x := (0, \ldots, 0, x_{k+1}, \ldots, x_n)^t \prec_{ub} y := (0, \ldots, 0, y_{k+1}, \ldots, y_n)^t$, by Proposition 2.1, and hence $Tx \prec_{ub} Ty$. That is, $(*, Sx')^t \prec_{ub} (*, Sy')^t$. It ensures that $Sx' \prec_{ub} Sy'$. Therefore, S preserves $\prec_{ub}$ on $\mathbb{R}^{n-k}$. □

Lemma 2.3. *Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be a linear preserver of $\prec_{ub}$. Then $[T]$ is upper triangular.*

Proof. Let $[T] = [a_{ij}]$. The proof is by induction on n. For $n = 1$, there is nothing to prove. If $n > 1$; Assuming the statement to hold for $n - 1$, we will prove it for n. Let $S : \mathbb{R}^{n-1} \to \mathbb{R}^{n-1}$ be the linear function with $[S] = [T](1)$. Lemma 2.2 ensures that the linear function S preserves $\prec_{ub}$ on $\mathbb{R}^{n-1}$. The induction hypothesis insures that $[S]$ is an upper triangular matrix. We only need to show that $a_{21} = a_{31} = \cdots = a_{n1} = 0$. As $e_1 \prec_{ub} e_2$, it shows that $Te_1 \prec_{ub} Te_2$. Hence $a_{31} = a_{41} = \cdots = a_{n1} = 0$. Therefore, $[T]$ is an upper triangular matrix. □

Lemma 2.4. *Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be a linear function such that $a_{kt} \neq 0$ for some k, t $(1 \leq k, t \leq n)$ where $[T] = [a_{ij}]$. Suppose that $a_{k+1t} = \cdots = a_{nt} = 0$ and there exists some j $(t + 1 \leq j \leq n - 1)$ such that $a_{k+1j} = \cdots = a_{nj} = 0$. Then T does not preserve $\prec_{ub}$.*

Proof. There is no loss of generality in assuming $a_{kt} = 1$. Fix $x = e_t$ and $y = -a_{kj}e_t + e_j$. One can easily see that $x \prec_{ub} y$ but $Tx \not\prec_{ub} Ty$. It follows that T does not preserve $\prec_{ub}$, as desired. □

The following theorem characterizes the structure of all linear functions $T : \mathbb{R}^n \to \mathbb{R}^n$, preserving $\prec_{ub}$.

Theorem 2.5. *Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be a linear function. Then T preserves $\prec_{ub}$ if and only if one of the following assertions holds.*

(i) $Te_1 = \cdots = Te_{n-1} = 0$. *In other words*

$$[T] = \begin{pmatrix} 0 & \dots & 0 & a_{1n} \\ 0 & \dots & 0 & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & a_{nn} \end{pmatrix}.$$

(ii) There exist t $(1 \leq t \leq n-1)$ and $1 \leq i_1 < \cdots < i_m \leq n-1$ such that $a_{i_1 t}, a_{i_2 t+1}, \dots, a_{i_m n-1} \neq 0$, and

$$[T] = \begin{pmatrix} 0 & * & & & \\ & a_{i_1 t} & & * & \\ & & \ddots & & \\ & & & a_{i_2 t+1} & \\ & & & & \ddots \\ & 0 & & & a_{i_m n-1} \\ & & & & 0 \quad * \end{pmatrix},$$

where we have one of the following conditions.
$\mathrm{card}\{a_{i_m+1n}, \dots, a_{nn}\} \geq 2$;
There are some k $(2 \leq k \leq m)$ and some i $(i_{k-1} < i < i_k)$ such that $r_i \neq r_{i_k} = \cdots = r_n$;
$r_{i_1} = r_{i_1+1} = \cdots = r_n$, *where r_i is the sum of entries on the ith row of $[T]$.*

Proof. We first suppose that T preserves $\prec_{ub}$ and (i) does not hold. As T preserves $\prec_{ub}$ on $\mathbb{R}^n$, Lemma 2.3 shows that $[T]$ is upper triangular. To prove (ii) we proceed by induction on n. First, let $n = 2$. We just prove $r_1 = r_2$. Without loss of generality assume that $a_{11} = 1$. Choose $x = e_1$ and $y = (a_{22} - a_{12})e_1 + e_2$. We see that $x \prec_{ub} y$, hence that $Tx \prec_{ub} Ty$, and finally that $r_1 = r_2$. Now assume that $n \geq 3$ and the statement holds for all linear preservers of $\prec_{ub}$ on $\mathbb{R}^{n-1}$. Let $S : \mathbb{R}^{n-1} \to \mathbb{R}^{n-1}$ be the linear function with $[S] = [T](1)$. Lemma 2.2 ensures that S preserves $\prec_{ub}$ on $\mathbb{R}^{n-1}$. By applying the induction hypothesis for S, we need only consider two steps.
Step 1. S satisfies (i). Lemma 2.4 ensures that the first nonzero column of T should be its $(n-1)st$ column. So we have to just show that $\mathrm{card}\{a_{i_m+1n}, \dots, a_{nn}\} \geq 2$ or $r_1 = r_2 = \cdots = r_n$. We may assume without loss of generality that $a_{1n-1} = 1$. If $\mathrm{card}\{a_{i_m+1n}, \dots, a_{nn}\} = 1$, we claim that $r_1 = r_n$. Consider $x = (a_{nn} - a_{1n-1})e_{n-1}$ and $y = (a_{nn} - a_{1n})e_{n-1} + e_n$. Observe that $x \prec_{ub} y$, and then $Tx \prec_{ub} Ty$. It implies that $r_1 = r_n$.
Step 2. S satisfies (ii). If the columns $1, \dots, t-1$ of T are zero, then there is nothing to prove. Otherwise, Lemma 2.4 ensures that the first nonzero column of T should be its $(t-1)st$ column. If $\mathrm{card}\{a_{i_m+1n}, \dots, a_{nn}\} \geq 2$, there is nothing to prove. If $\mathrm{card}\{a_{i_m+1n}, \dots, a_{nn}\} = 1$; If for $[S]$ there are some k $(3 \leq k \leq m)$ and some i $(i_{k-1} < i < i_k)$ such that $r_i \neq r_{i_k} = \cdots = r_n$, then (ii) holds for $[T]$. Otherwise, we have $r_{i_2} = r_{i_2+1} = \cdots = r_n$. If there is some i $(1 < i < i_2)$ such that $r_i \neq r_{i_2}$, then (ii) holds for $[T]$. If not; Then $r_2 = r_3 = \cdots = r_n$. We should prove $r_1 = r_n$. Without loss of generality we may assume that $a_{1t-1} = 1$. If $r_1 \neq r_n$; Define $x = e_{t-1}$ and $y = (a_{nn} - \sum_{j=t}^{n} a_{1j})e_{t-1} + \sum_{j=t}^{n} e_j$. As $x \prec_{ub} y$, we see that $Tx \prec_{ub} Ty$, which would be a contradiction. Hence $r_1 = r_n$.

For the converse, we prove the sufficiency of the conditions. Let $x = (x_1, \dots, x_n)^t$, $y = (y_1, \dots, y_n)^t \in \mathbb{R}^n$ such that $x \prec_{ub} y$. We claim that $Tx \prec_{ub} Ty$. If (i) holds, then

$Tx = (0, \dots, 0)^t$, and so $Tx \prec_{ub} Ty$. If (ii) holds; We proceed by induction on n. Suppose that $n \geq 2$ and that the assertion has been established for all linear functions on $\mathbb{R}^{n-1}$ with the conditions described in the hypothesis. Let $S : \mathbb{R}^{n-1} \to \mathbb{R}^{n-1}$ be the linear function with $[S] = [T](1)$. Set $x' = (x_2, \dots, x_n)^t$, $y' = (y_2, \dots, y_n)^t$. Then $x' \prec_{ub} y'$, and the induction hypothesis for S ensures that $Sx' \prec_{ub} Sy'$. That is, $((Tx)_2, \dots, (Tx)_n)^t \prec_{ub} ((Ty)_2, \dots, (Ty)_n)^t$. It remains to prove that $(Tx)_1 \in \mathcal{B}\{(\mathrm{Ty})_1, \dots, (\mathrm{Ty})_\mathrm{n}\}$. Notice that $\mathcal{B}\{(\mathrm{Ty})_1, \dots, (\mathrm{Ty})_\mathrm{n}\} = \{\beta_1((\mathrm{Ty})_1 - \mathrm{a_{nn}y_n}) + \beta_2((\mathrm{Ty})_2 - \mathrm{a_{nn}y_n}) + \dots + \beta_{\mathrm{i_m}}((\mathrm{Ty})_{\mathrm{i_m}} - \mathrm{a_{nn}y_n}) + \beta_{\mathrm{i_m+1}}(\mathrm{a_{i_m+1n}} - \mathrm{a_{nn}})\mathrm{y_n} + \dots + \beta_{\mathrm{n-1}}(\mathrm{a_{n-1n}} - \mathrm{a_{nn}})\mathrm{y_n} : \beta_1, \dots, \beta_{\mathrm{n-1}} \in \mathbb{R}\}$.
If $\mathrm{card}\{\mathrm{a_{i_m+1n}}, \dots, \mathrm{a_{nn}}\} \geq 2$; Without loss of generality $a_{n-1n} \neq a_{nn}$.
If $y_n \neq 0$; Set $\beta_{n-1} = \frac{(Tx)_1}{(a_{n-1n} - a_{nn})y_n}$, and the other $\beta_i = 0$. Then $(Tx)_1 \in \mathcal{B}\{(\mathrm{Ty})_1, \dots, (\mathrm{Ty})_\mathrm{n}\}$.
If $y_n = 0$; In this case if $y_{n-1} \neq 0$, then by choosing $\beta_{i_m} = \frac{(Tx)_1}{a_{i_m n-1} y_{n-1}}$, and the other $\beta_i = 0$, we obtain desired conclusion. By continuing this process, if $y_n = y_{n-1} = \dots = y_t = 0$, then $x_n = x_{n-1} = \dots = x_t = 0$, and so $Tx = (0, \dots, 0)^t$. It follows that $(Tx)_1 \in \mathcal{B}\{(\mathrm{Ty})_1, \dots, (\mathrm{Ty})_\mathrm{n}\}$.
If $\mathrm{card}\{\mathrm{a_{i_m+1n}}, \dots, \mathrm{a_{nn}}\} = 1$; With an argument almost identical to that of the above, the theme can be proved.
Therefore, T preserves $\prec_{ub}$. □

Lemma 2.6. *Let $T : \boldsymbol{M}_{n,m} \to \boldsymbol{M}_{n,m}$ be a linear function that strongly preserves ub-majorization. Then T is invertible.*

Proof. Suppose that $T(A) = 0$, where $A \in \mathbf{M}_{n,m}$. Notice that since T is linear, we have $T(0) = 0 = T(A)$. Then it is obvious that $T(A) \prec_{ub} T(0)$. Therefore, $A \prec_{ub} 0$, because T strongly preserves ub-majorization. So $A = 0$, and hence T is invertible. □

The following theorem characterizes all the linear functions $T : \mathbb{R}^n \to \mathbb{R}^n$ which strongly preserve ub-majorization.

Theorem 2.7. *A linear function $T : \mathbb{R}^n \to \mathbb{R}^n$ strongly preserves $\prec_{ub}$ if and only if $[T] = \alpha A$, for some $\alpha \in \mathbb{R} \setminus \{0\}$, and an invertible matrix $A \in \mathcal{R}_n^{gut}$.*

Proof. First, assume that T strongly preserves $\prec_{ub}$. Lemma 2.6 ensures that T is invertible. Let $[T] = [a_{ij}]$. Theorem 2.5 ensures that $[T]$ is upper triangular, $r_1 = \dots = r_n$, and $a_{11}, \dots, a_{nn} \neq 0$. So there exist an invertible matrix $A \in \mathcal{R}_n^{gut}$ and $\alpha \in \mathbb{R} \setminus \{0\}$ such that $[T] = \alpha A$.

Next, assume that there exist an invertible matrix $A \in \mathcal{R}_n^{gut}$ and $\alpha \in \mathbb{R} \setminus \{0\}$ such that $[T] = \alpha A$. Then both of T and T^{-1} preserve $\prec_{ub}$ on $\mathbb{R}^n$ and, therefore, T strongly preserves $\prec_{ub}$. □

3 G-row balanced on $\mathbf{M}_{n,m}$

In this section, we characterize strong linear preservers ub-majorization $T : \mathbf{M}_{n,m} \to \mathbf{M}_{n,m}$.

Let E be the n-by-n matrix with all of the entries of the last column equal to one and the other entries equal to zero. Notice that for each $R \in \mathcal{R}_n^{ub}$ we have $ER = RE = 0$.

We need the following lemmas to prove the last result of the paper.

Lemma 3.1. *Let $A \in \boldsymbol{M}_n$. Then the following conditions are equivalent.*
a) For each matrix $D \in \mathcal{R}_n^{ub}$ $AD = DA$.
b) For some $\alpha, \beta \in \mathbb{R}$ $A = \alpha I + \beta E$.
c) For each matrix $D \in \mathcal{R}_n^{ub}$ and for all $x, y \in \mathbb{R}^n$ $(Dx + ADy) \sim_{ub} (x + Ay)$.

Proof. $(a \to b)$ First by considering

$$D = \begin{pmatrix} 1 & -1 & & & & \\ & 1 & -1 & & 0 & \\ & & & \ddots & & \\ & 0 & & & 1 & -1 \\ & & & & & 0 \end{pmatrix},$$

and next

$$D = \begin{pmatrix} 0 & & & & & \\ & 0 & & & & \\ & & \ddots & & 0 & \\ & 0 & & & & \\ & & & 2 & 0 & -2 \\ & & & & 1 & -1 \\ & & & & & 0 \end{pmatrix},$$

we see that there exist some $\alpha, \beta \in \mathbb{R}$ such that $A = \alpha I + \beta E$.
$(b \to c)$ If $D \in \mathcal{R}_n^{ub}$ and $x, y \in \mathbb{R}^n$; Then $Dx + ADy = D(x + Ay)$, and hence $(Dx + ADy) \sim_{ub} (x + Ay)$.
$(c \to a)$ Fix i $(1 \leq i \leq n)$. Set $x = e - Ae_i$ and $y = e_i$. By the hypothesis, $(-DA + AD)e_i \sim_{ub} e$, and then $(-DA + AD)e_i = 0$. Thus $AD = DA$. □

For each i, j $(1 \leq i, j \leq m)$ consider the embedding $E^j : \mathbb{R}^n \to \mathbf{M}_{n,m}$ and the projection $E_i : \mathbf{M}_{n,m} \to \mathbb{R}^n$, where $E^j(x) = xe_j^t$ and $E_i(A) = Ae_i$. It is easy to show that for every linear function $T : \mathbf{M}_{n,m} \to \mathbf{M}_{n,m}$, $TX = T[x_1 \mid \ldots \mid x_m] = [\sum_{j=1}^m T_1^j x_j \mid \ldots \mid \sum_{j=1}^m T_m^j x_j]$, where $T_i^j = E_i T E^j$.

It is easy to see that if $T : \mathbf{M}_{n,m} \to \mathbf{M}_{n,m}$ is a linear preserver of $\sim_{ub}$, then T_i^j preserves $\sim_{ub}$ on $\mathbb{R}^n$, for all i, j $(1 \leq i, j \leq m)$.

Lemma 3.2. *Let $T : \boldsymbol{M}_{n,m} \to \boldsymbol{M}_{n,m}$ be a preserving of $\sim_{ub}$. If for some i $(1 \leq i \leq m)$ there exist some k $(1 \leq k \leq m)$ such that T_i^k is invertible, then $\sum_{j=1}^m A_i^j x_j = A_i^k \sum_{j=1}^m \alpha_i^j x_j + E \sum_{j=1}^m \beta_i^j x_j$ for some $\alpha_i^j, \beta_i^j \in \mathbb{R}$, where $A_i^j = [T_i^j]$.*

Proof. We may assume without loss of generality that $i, k = 1$ and $j = 2$. Let $D \in \mathcal{R}_n^{ub}$ and $x, y \in \mathbb{R}^n$. Then $D[x|y|0|\ldots|0] \sim_{ub} [x|y|0|\ldots|0]$, and so
$T[Dx|Dy|0|\ldots|0] \sim_{ub} T[x|y|0|\ldots|0]$. It implies that $[A_1^1 Dx + A_1^2 Dy \mid * \mid *] \sim_{ub} [A_1^1 x + A_1^2 y \mid * \mid *]$, and thus $A_1^1 Dx + A_1^2 Dy \sim_{ub} A_1^1 x + A_1^2 y$. Lemma 3.1 ensures that there exist $\alpha_1^2, \beta_1^2 \in \mathbb{R}$ such that $A_1^2 = \alpha_1^2 A_1^1 + \beta_1^2 E$. □

Lemma 3.3. *If $T : \boldsymbol{M}_{n,m} \to \boldsymbol{M}_{n,m}$ strongly preserves $\sim_{ub}$, then for each i $(1 \leq i \leq m)$ there exists some j $(1 \leq j \leq m)$ such that T_i^j is invertible.*

Proof. Consider $I = \{1 \leq i \leq m \mid T_i^j e_1 = 0,\ \forall 1 \leq j \leq m\}$. We claim that I is empty. If I is not empty; Without loss of generality $I = \{1, 2, \ldots, k\}$, where $1 \leq k \leq m$. If $k = m$, choose $X = [e_1 \mid 0 \mid \ldots \mid 0] \in \mathbf{M}_{n,m}$ and conclude that $X \neq 0$ but $TX = 0$, which is a contradiction, by Lemma 2.6. If $k < m$, by Lemma 3.3, for i $(k + 1 \leq i \leq m)$ and j $(1 \leq j \leq m)$, there exist invertible matrices A_i and $\alpha_i^j, \beta_i^j \in \mathbb{R}$ such that $\sum_{j=1}^m A_i^j x_j = A_i \sum_{j=1}^m \alpha_i^j x_j + E \sum_{j=1}^m \beta_i^j x_j$. Then there exist $\gamma_1, \ldots, \gamma_m \in \mathbb{R}$, not all zero, such that $\gamma_1(\alpha_{k+1}^1, \ldots, \alpha_m^1)^t + \cdots + \gamma_m(\alpha_{k+1}^m, \ldots, \alpha_m^m)^t = 0$. Let $x_j = \gamma_j e_1$ for each j $(1 \leq j \leq m)$ and $X = [x_1 \mid \ldots \mid x_m] \in \mathbf{M}_{n,m}$. We observe that $X \neq 0$, and $TX = 0$, a contradiction. Therefore, I is empty. □

Theorem 3.4. *Let $T : \boldsymbol{M}_{n,m} \to \boldsymbol{M}_{n,m}$ be a linear function. Then T strongly preserves $\sim_{ub}$ if and only if there exist $R, S \in \boldsymbol{M}_m$, $R(R+S)$ is invertible, and invertible matrix $A \in \mathcal{R}_n^{gut}$ such that $TX = AXR + EXS$.*

Proof. First, we prove the sufficiency of the conditions. Let $X, Y \in \mathbf{M}_{n,m}$ such that $X \sim_{ub} Y$. It means that $X = DY$ for some $D \in \mathcal{R}_n^{ub}$. Then $TX = AXR + EXS = A(DY)R + E(DY)S = A(DY)R = (ADA^{-1})(AYR + EYS) = (ADA^{-1})TY$. As $ADA^{-1} \in \mathcal{R}_n^{ub}$, we see that $TX \sim_{ub} TY$. On the other hand, if $TX \sim_{ub} TY$, then there exists some $D \in \mathcal{R}_n^{ub}$ such that $TX = DTY$. So $AXR + EXS = D(AYR + EYS)$, and hence $XR + EXS = (A^{-1}DA)YR$, because of A is invertible. Multiply this relation by E, and since $R + S$ is invertible, we conclude that $EX = 0$. Substitute $EX = 0$ in the relation $AXR + EXS = D(AYR + EYS)$, and as R is invertible, conclude that $X = (A^{-1}DA)Y$. Thus $X \sim_{ub} Y$. Therefore, T strongly preserves $\sim_{ub}$.

Next, assume that T strongly preserves $\sim_{ub}$. For $m = 1$ see Theorem 2.7. Let $m > 1$. Lemma 3.3 enures that for each i $(1 \leq i \leq m)$ there exists some j $(1 \leq j \leq m)$ such that T_i^j is invertible. Lemma 3.2 ensures that there exist invertible matrices $A_1, \ldots, A_m \in \mathbf{M}_n$, vectors $a_1, \ldots, a_m \in \mathbb{R}^m$, and a matrix $S' \in \mathbf{M}_m$ such that $TX = [A_1Xa_1 \mid \ldots \mid A_mXa_m] + EXS'$. One can prove $\operatorname{rank}\{a_1, \ldots, a_m\} \geq 2$. Without loss of generality, assume that $\{a_1, a_2\}$ is a linearly independent set. It implies that for every $x, y \in \mathbb{R}^n$, there exists $B_{x,y} \in \mathbf{M}_{n,m}$ such that $B_{x,y}a_1 = x$ and $B_{x,y}a_2 = y$. Let $X \in \mathbf{M}_{n,m}$ and invertible matrix $D \in \mathcal{R}_n^{ub}$. So $DX \sim_{ub} X$, and then $TDX \sim_{ub} TX$. Thus $[A_1DXa_1 \mid \ldots \mid A_mDXa_m] + EDXS \sim_{ub} [A_1Xa_1 \mid \ldots \mid A_mXa_m] + EXS$. Clearly, $A_1DXa_1 + A_2DXa_2 \sim_{ub} A_1Xa_1 + A_2Xa_2$. So for each $X \in \mathbf{M}_{n,m}$ and each invertible matrix $D \in \mathcal{R}_n^{ub}$ we have

$$DXa_1 + A_1^{-1}A_2DXa_2 \sim_{ub} Xa_1 + A_1^{-1}A_2Xa_2. \qquad (1)$$

By replacing $X = B_{x,y}$ in (1) $Dx + A_1^{-1}A_2Dy \sim_{gut} x + A_1^{-1}A_2y$, for each invertible matrix $D \in \mathcal{R}_n^{ub}$, and for each $x, y \in \mathbb{R}^n$. Lemma 3.1 states that $A_2 = \alpha A_1 + \beta E$ for some $\alpha, \beta \in \mathbb{R}$. For every $i \geq 3$ if $a_i = 0$ we can choose $A_i = A_1$. If $a_i \neq 0$, then $\{a_1, a_i\}$ or $\{a_2, a_i\}$ is linearly independent. Similar to above $A_i = \gamma_i A_1 + \delta_i E$ for some $\gamma_i, \delta_i \in \mathbb{R}$. Define $A := A_1$. Then for every $i \geq 2$, $A_i = \alpha_i A + \beta_i E$, for some $\alpha_i, \beta_i \in \mathbb{R}$. It implies that $TX = [AXa_1 \mid AX(r_2a_2) \mid \ldots \mid AX(r_ma_m)] + EXS = AXR + EXS$, in which $R = [a_1 \mid r_2a_2 \mid \ldots \mid r_ma_m]$, for some $r_2, \ldots, r_m \in \mathbb{R}$ and $S = S' + [0 \mid \beta_2a_2 \mid \ldots \mid \beta_ma_m]$. □

4 Conclusion

We know that majorization and linear preservers of a relationship are of particular importance. For this reason, in this article we define a new kind of relationship and we found its linear preservers.

References

[1] L. B. Beasley, S-G. Lee and Y-H Lee, A characterization of strong preservers of matrix majorization, *Linear Algebra and its Applications,* 367 (2003) 341–346.

[2] A. Ilkhanizadeh Manesh, On linear preservers of sgut-majorization on $\mathbf{M}_{n,m}$, *Bulletin of the Iranian Mathmatical Society,* 42 (2) (2016) 470-481.

[3] A. Ilkhanizadeh Manesh, Right gut-Majorization on $\mathbf{M}_{n,m}$, *Electronic Journal of Linear Algebra,* 31 (1) (2016) 13–26.

An extension of the numerical radius[1]

Mohsen Kian*

Department of Mathematics, University of Bojnord, Iran

Abstract

The matricial range has been introduced by W. Areveson as a matrix valued extension of the numerical range. Noting the connection between the numerical range and numerical radius, we introduce a quantity related to the matricial range of a matrix. We present some of its properties which are extensions of the results about the numerical radius.

Keywords: Numerical range, Numerical radius, Matricial range

Mathematics Subject Classification [2010]: 15A60, 47A12

1 Introduction

Throughout this paper assume that $\mathbb{M}_n$ is the algebra of all $n \times n$ matrices with complex entries and I denotes the identity matrix in any size. We write $A \geq 0$ ($A > 0$), when A is a positive semidefnite (positive definite) matrix. The well-known (Löwner) partial order on the real space of all Hermitian matrices is defined by $A \leq B$ if and only if $B - A \geq 0$.

A well-known concept in the matrix theory is the numerical range. The numerical range of a matrix $A \in \mathbb{M}_n$ is defined by $W(A) = \{x^*Ax;\ x \in \mathbb{C}^n,\ \|x\| = 1\}$. This set has many applications, for example in numerical analysis and differential equations. The numerical radius of a matrix $A \in \mathbb{M}_n$ is defined by

$$\omega(A) = \sup\{|z|;\ z \in W(A)\} = \sup\{|x^*Ax|\,;\ x \in \mathbb{C}^n,\ \|x\| = 1\}.$$

Some basic properties of the numerical radius are as follows:

Theorem A. For every $A, B \in \mathbb{M}_n$

(i) $\omega(A) = \omega(A^*)$ and $\omega(U^*AU) = \omega(A)$ for every unitary $U \in \mathbb{M}_n$;

(ii) $\frac{1}{2}\|A\| \leq \omega(A) \leq \|A\|$ and $\omega(A) = \|A\|$ if A is normal;

(iii) $\omega(A) \geq \frac{1}{2}\left\||A|^2 + |A^*|^2\right\|^{1/2} \geq \frac{1}{2}\|A\|$.

Moreover, it is known that $\omega(A) \leq 1$ if and only if there exits a Hermitian matrix H such that $\begin{bmatrix} I+H & T \\ T^* & I-H \end{bmatrix}$ is positive semi-definite.

[1]Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

*Speaker. Email address: kian@ub.ac.ir

A map $\Phi : \mathbb{M}_n \to \mathbb{M}_m$ is called *positive* if $\Phi(\mathbb{M}_n^+) \subseteq \mathbb{M}_m^+$, in which $\mathbb{M}_n^+$ is the set of all positive semi-definite matrices in $\mathbb{M}_n$. Moreover, for $k \in \mathbb{N}$, the mapping Φ is called k-positive if the mapping $\Phi_k : \mathbb{M}_k(\mathbb{M}_n) \to \mathbb{M}_k(\mathbb{M}_m)$ defined by $\Phi_k([A_{ij}]) = [\Phi(A_{ij})]$ is positive. If $\Phi : \mathbb{M}_n \to \mathbb{M}_m$ is k-positive for every $k \in \mathbb{N}$, then Φ is called completely positive (CP for short). Φ is called unital if $\Phi(I) = I$.

As a matrix valued extension of the numerical range, W. Arveson introduced the k'th matricial range of $A \in \mathbb{M}_n$ by

$$W^k(A) = \{\Phi(A);\ \Phi : C^*(A) \to \mathbb{M}_k \text{ is a unital CP map}\},$$

where $C^*(A)$ is the unital C^*-algebra generated by A. The matricial range has favourite properties, some of them like those of the numerical range. As a well-known property of the numerical range, the Toeplitz-Hausdorff result says that it is a convex set. Fortunately, the matricial range enjoys this property in a stronger manner. It is known that the matricial range of a matrix is C^*-convex. A set $\mathcal{K} \subseteq \mathbb{M}_n$ is called C^*-convex, if $X_1, \ldots, X_m \in \mathcal{K}$ and $C_1, \ldots, C_m \in \mathbb{M}_n$ with $\sum_{j=1}^m C_j^* C_j = I$ imply that $\sum_{j=1}^m A_j^* X_j A_j \in \mathcal{K}$. Indeed, this is a noncommutative generalization of linear convexity. Some basic properties of the matricial range reads as follows:

Theorem B. If $T \in \mathbb{M}_n$ and $k \in \mathbb{N}$, then

(i) $W^k(T^*) = W^k(T)$;
(ii) $W^k(U^*TU) = W^k(T)$ for each unitay $U \in \mathbb{M}_n$;
(iii) $W^k(\alpha I_n) = \{\alpha I_k\}$ and $W^k(\alpha T + \beta I) = \alpha W^k(T) + \beta I_k$ for all $\alpha, \beta \in \mathbb{C}$.

It should be remarked that except in some special cases, it is not routine to obtain the matricial ranges of a matrix.

The main aim of the preset work is to consider a related concept to the matricial range in parallel to the connection of the numerical range and numerical radius.

2 Main results

Minding the matricial range, it is natural to think about a possible extension of the numerical radius. However, a direct extension as $\max\{\|X\|;\ X \in W^k(A)\}$ would not be interesting, because it is exactly equal to $\|A\|$.

We introduced the following quantity related to the matricial range.

Definition 2.1. For $A \in \mathbb{M}_n$, we define

$$\omega^k(A) = \sup\{|\mathrm{Tr}\Phi(A)|;\ \Phi : C^*(A) \to \mathbb{M}_k \text{ is unital CP map}\}.$$

where $\mathrm{Tr}(\cdot)$ denotes the canonical trace.

It is easy to see that for every $k \in \mathbb{N}$,

(i) $\omega^k(A^*) = \omega^k(A)$;
(ii) $\omega^k(U^*AU) = \omega^k(A)$ for every unitary U;
(iii) $\omega^k(A) \leq k\|A\|$ equality holds if A is normal.

We present some other properties of $\omega^k(A)$ which are extension of facts in Theorem A.

Theorem 2.2. *Let $A \in \mathbb{M}_n$ and $k \in \mathbb{N}$. Then $\omega^k(A) \leq k$ if and only if there exits a Hermitian matrix $H \in \mathbb{M}_k$ such that $\begin{bmatrix} I+H & T \\ T^* & I-H \end{bmatrix}$ is positive semi-definite.*

Theorem 2.3. *Let $A, B \in \mathbb{M}_n$. For every $k \in \mathbb{N}$*

$$\omega^k(AB) \leq \frac{k}{2} \left\| |A^*|^2 + |B|^2 \right\|.$$

The next theorem provides an extension of (iii) of Theorem A.

Theorem 2.4. *Let $A \in \mathbb{M}_n$ and $k \in \mathbb{N}$. Then*

$$\omega^k(A) \geq \frac{k}{2} \left\| |A|^2 + |A^*|^2 \right\|^{1/2}.$$

References

[1] M. Dehghani and M. Kian, *On Matricial Ranges of Some Matrices*, J. Math. Ext. **13** (2019), 83–102.

[2] D.R. Farenick, *Matricial extension of the numerical range: A brief survey*, Linear Multilinear algebra, **34** (1993), 197–211.

[3] M. Kian, *C^*-convexity of norm unit balls*, J. Math. Anal. Appl. **445** (2017), 1417–1427.

[4] M. Kian, M. Dehghani and M. Sattari, *matricial radius*, submitted.

[5] S.-H. Tso and P.Y. Wu, *Matricial ranges of quadratic operators*, Rocky Mountain J. Math., **29** (1999), 1139–1152.

Steepest descent NSCG iteration method for solving non-symmetric positive definite linear systems[1]

Mohammad Khorsand Zak*

Department of Applied Mathematics, Aligudarz Branch, Islamic Azad University, Aligudarz, Iran

Abstract

By applying the steepest descent technique to the nested splitting conjugate gradient (NSCG) iteration scheme, we introduce a non-stationary iteration method named steepest descent nested splitting conjugate gradient (SDNSCG) iteration method to solve non-symmetric positive definite linear systems. Numerical results verify the effectiveness and robustness of the SDNSCG iteration method.

Keywords: Steepest descent, Linear system, Iterative method, NSCG method
Mathematics Subject Classification [2010]: 65F10, 65F15, 65F35

1 Introduction

In many problems in scientific computing we encounter with a system of linear equations such as

$$Ax = b, \tag{1}$$

where $A \in \mathbb{R}^{n\times n}$ is a nonsingular matrix, $x \in \mathbb{R}^n$ is an unknown vector and $b \in \mathbb{R}^n$ is a given vector. We consider a symmetric positive definite splitting

$$A = B - C, \tag{2}$$

where B is a symmetric positive definite matrix and $\rho(B^{-1}C) < 1$. Then the system of linear equations (1) is equivalent to the fixed-point equation

$$Bx = Cx + b.$$

For a given initial guess $x^{(0)}$, suppose that we have computed approximations $x^{(0)}$, $x^{(1)}$, $\cdots$, $x^{(k)}$ to the solution x^* of the system of linear equations (1). Then the next approximation $x^{(k+1)}$ is obtain by solving the system of linear equations

$$Bx = Cx^{(k)} + b, \tag{3}$$

by the conjugate gradient method. It is actually inner/outer iterations, which employees the conjugate gradient method as inner iteration to approximate each outer iterate, while

[1]Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University
*Speaker. Email address: mo.khorsand@mail.um.ac.ir

outer iteration is induced by the (2) splitting. This is the initial idea of the nested splitting conjugate gradient (NSCG) method.

Axelsson et al. [1] proposed a class of the NSCG method for solving linear systems with a coefficient matrix with a dominant positive definite symmetric part. They proposed some implementation strategies for choosing matrices B and C respect to the coefficient matrix A. In the special case for nonsymmetric and positive definite coefficient matrix A, they choose $B = H$ and $C = S$, When

$$H = \frac{A^T + A}{2}, \qquad S = \frac{A^T - A}{2}, \tag{4}$$

are the symmetric and skew-symmetric parts of matrix A, respectively. Moreover, the regularizing technique was proposed by using a quasi-Hermitian splitting

$$A = B(\alpha) - C(\alpha), \tag{5}$$

where

$$B(\alpha) = B + \alpha I, \qquad C(\alpha) = C + \alpha I, \tag{6}$$

and $\alpha \geq 0$ is a regularizing parameter [1].

For non-Hermitian positive definite coefficient matrix A, Li and Wu [3] proposed a single step Hermitian and skew-Hermitian (SHSS) method described as

$$(\alpha I + H)x^{(k+1)} = (\alpha I + S)x^{(k)} + b, \qquad k = 0, 1, \cdots, \tag{7}$$

where H and S are as in (4).

Wang et al. [4] proposed a single-step iteration method for non-Hermitian positive definite linear systems which described as

$$(P + H)x^{(k+1)} = (P + S)x^{(k)} + b, \qquad k = 0, 1, \cdots, \tag{8}$$

where H and S are as in (4) and P is a given Hermitian positive definite matrix. When $P = \alpha I$, this method reduced to the SHSS method [3]. As an other choice, we have $P = \alpha H$, see [4] for more details.

We can consider (7) and (8) as classes of the regularized NSCG method.

Recently, Yang et al. [6], by applying the minimum residual technique to the Hermitian and skew-Hermitian (HSS) iteration scheme, proposed a non-stationary iteration method named minimum residual HSS (MRHSS).

Motivated by [5,6], we apply the steepest descent technique to the nested splitting conjugate gradient iteration scheme and introduce a non-stationary iteration method named steepest descent nested splitting conjugate gradient (SDNSCG) iteration method to solve non-symmetric positive definite linear systems.

2 Main results

The linear system (3) can be rewritten as

$$x = x^{(k)} + B^{-1}r^{(k)},$$

where $r^{(k)} = b - Ax^{(k)}$ is the kth residual of the linear system (1). Thus, the NSCG iteration scheme (3) can be rewritten as

$$x^{(k+1)} = x^{(k)} + \delta^{(k)}, \tag{9}$$

where

$$\delta^{(k)} = B^{-1}r^{(k)}, \tag{10}$$

can be consider as the search direction from $x^{(k)}$ to $x^{(k+1)}$. The step size in (9) is unitary. Thus, for improving the efficiency of the iteration scheme (9), we introduce an arbitrary positive parameter β_k to control the step sizes, which leads to the following new iteration scheme:

$$x^{(k+1)} = x^{(k)} + \beta_k \delta^{(k)}. \tag{11}$$

The residual form of the iteration (11) can be written as

$$r^{(k+1)} = r^{(k)} - \beta_k A\delta^{(k)}. \tag{12}$$

From the steepest descent algorithm and the Petrov-Galerkin condition we have

$$\langle r^{(k)} - \beta_k A\delta^{(k)}, r^{(k)}\rangle = 0,$$

and this yields

$$\beta_k = \frac{\langle r^{(k)}, r^{(k)}\rangle}{\langle A\delta^{(k)}, r^{(k)}\rangle}. \tag{13}$$

Therefore, the SDNSCG algorithm can be describe as follows:

Algorithm 2.1. The steepest descent NSCG algorithm

1. Select an initial guess $x^{(0)}$, compute $r^{(0)} = b - Ax^{(0)}$
2. For $k = 0, 1, 2, \cdots$, until convergence, Do:
3. Solve system $B\delta^{(k)} = r^{(k)}$ by the CG method
4. $w = A\delta^{(k)}$
5. $\beta = \frac{\langle r^{(k)}, r^{(k)}\rangle}{\langle w, r^{(k)}\rangle}$
6. $x^{(k+1)} = x^{(k)} + \beta\delta^{(k)}$
7. $r^{(k+1)} = r^{(k)} - \beta w$
8. End Do

Remark 2.2. From relation (11), the iteration sequence of the SDNSCG method yields

$$x^{(j+1)} = x^{(j)} + \beta_j B^{-1} r^{(j)}. \tag{14}$$

Therefore, we can obtain

$$x^{(m)} = x^{(0)} + B^{-1} \sum_{j=0}^{m-1} \beta_j r^{(j)}. \tag{15}$$

Applying (12), with respect to (10), recursively for $k = j-1, \cdots, 1, 0$, we obtain

$$r^{(j)} = \left(\prod_{i=j-1}^{0} (I - \beta_i B^{-1}) \right) r^{(0)}. \tag{16}$$

By substituting (16) into (15), the approximate solution $x^{(m)}$ yields as

$$x^{(m)} = x^{(0)} + B^{-1} \sum_{j=0}^{m-1} \beta_j \left(\prod_{i=j-1}^{0} (I - \beta_i B^{-1}) \right) r^{(0)}.$$

So, the approximate solution $x^{(m)}$ does not belong to the affine space $x_0 + \mathcal{K}_m$, where $\mathcal{K}_m$ is a Krylov subspace. This means the SDNSCG iteration method is not a standard Krylov subspace method.

3 Numerical results

In this section, we give two examples to demonstrate the performance of the SDNSCG method for solving the linear system (1). Numerical comparisons with steepest descent (SD) and NSCG methods are also presented to show the advantage of the SDNSCG method.

Each iterations process is started from an initial vector having all entries equal to zero, and terminated once either the iteration number is over 10000 or the current iteration residual $r^{(k)} = b - Ax^{(k)}$ satisfies $||r^{(k)}||_2/||r^{(0)}||_2 \leq 10^{-10}$, where $r^{(0)} = b - Ax^{(0)}$ is the initial residual. In addition, we take the right hand side vector b such that the exact solution of the system of linear equations (1) is $x^* = (1, 1, \cdots, 1)^T$.

Example 3.1. Consider the non-symmetric positive definite linear system (1) with $A = \text{tridiag}(-2, 4, -1)$ as the coefficient matrix. We apply the methods for different dimensions n to the linear system. For each method, we reported the number of iterations and CPU time in second (in parentheses) in the Table 1. In the Table 1, we observe that for this

Table 1: Results for the Example 3.1

Method	$n = 1000$	$n = 2000$	$n = 3000$	$n = 4000$	$n = 5000$
SD	65(0.187)	63(0.594)	62(1.375)	61(2.218)	60(3.753)
NSCG	19(0.531)	18(2.102)	18(4.943)	18(6.531)	18(9.671)
SDNSCG	19(0.328)	18(1.843)	18(4.328)	18(5.963)	18(8.937)

example the SD method performs better than the other two methods in term of the CPU time. Moreover, the NSCG and SDNSCG methods give similar results for this example.

Example 3.2. For the second example, we consider test matrix **nos1** of dimension $n = 237$ from Harwell-Boeing collection as the coefficient matrix A in the linear system (1). The results of this problem presented in the Table 2. In Table 2, the † sign for the CPU time

Table 2: Results for the Example 3.2

	SD	NSCG	SDNSCG
CPU-time	†	11.265	0.062
iterations	> 10000	1753	2
$\|\|r^{(k)}\|\|_2$	1.3387e+5	0.7697	0.0104
$\|\|x^{(k)} - x^*\|\|_2$	11.9520	9.5606e-4	1.56658e-11

means that the method was not converged in 10000 iterations. The results in the Table 2 show the efficiency and advantage of the SDNSCG method versus the other methods.

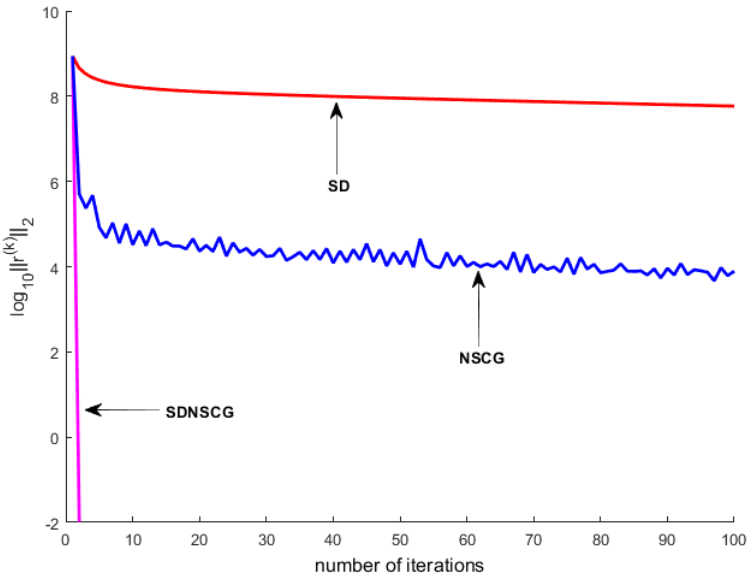


Figure 1: Convergence history of the methods for 100 first iterations

Moreover, we compare the convergence history of the methods for 100 first iterations in the Figure 1. Figure 1 shows, for this example the residual norm of the SDNSCG method decreases faster and sharper versus the other methods.

4 Conclusion

For non-symmetric positive definite system of linear equations, we present a kind of steepest descent NSCG (SDNSCG) iteration method to approximate its solution. Numerical results showed that the SDNSCG iteration method is very efficient and robust, especially for the second example.

References

[1] O. Axelsson, Z.-Z. Bai, and S.-X. Qiu, *A class of nested iterative schemes for linear systems with a coefficient matrix with a dominant positive definite symmetric part*, Numer. Algorithms, 35 (2004) 351–372.

[2] 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.

[3] C.-X. Li and S.-L. Wu, *A single-step HSS method for non-Hermitian positive definite linear systems*, Appl. Math. Lett., 44 (2015) 26–29.

[4] X. Wang, X.-Y. Xiao and Q.-Q. Zheng, *A single-step iteration method for non-Hermitian positive definite linear systems*, J. Comput. Appl. Math., 346 (2019) 471–482.

[5] A.-L Yang, *On the convergence of the minimum residual HSS iteration method*, Appl. Math. Lett., 94 (2019) 210–216.

[6] A.-L Yang, Y. Cao, and Y.-J. Wu, *Minimum residual Hermitian and skew-Hermitian splitting iteration method for non-Hermitian positive definite linear systems*, BIT Numer. Math., 59 (2019) 299–319.

Some generalizations of the numerical radius inequalities[1]

Rahmatollah Lashkaripour, Fatemeh Goli* and Monire Hajmohamadi

Department of Mathematics, Faculty of Mathematics,
University of Sistan and Baluchestan, Zahedan, Iran

Abstract

The main aim of this article is to obtain numerical radius inequalities for the Young and Heinz types of positive matrices A and B.

Keywords: Numerical radius, Operator matrix, Positive definite matrix
Mathematics Subject Classification [2010]: 47A12, 47A30, 47A63

1 Introduction

Suppose that $(\mathscr{H}, \langle\,.\,,\,.\,\rangle)$ is a complex Hilbert space and $\mathbb{B}(\mathscr{H})$ denotes the C^*-algebra of all bounded linear operators on $\mathscr{H}$. In the case when $\dim\mathscr{H} = n$, we identify $\mathbb{B}(\mathscr{H})$ with the matrix algebra $\mathbb{M}_n(\mathbb{C})$ of all $n \times n$ matrices with entries in the complex field, $\mathbb{M}_n^+(\mathbb{C})$ and $\mathbb{M}_n^{++}(\mathbb{C})$ are the cones of positive semidefinite and strictly positive semidefinite matrices in $\mathbb{M}_n(\mathbb{C})$. An operator $T \in \mathbb{B}(\mathscr{H})$ is called positive(positive semidefinite for a matrix) if $\langle Tx, x\rangle \geq 0$ for all $x \in \mathscr{H}$, then write $T \geq 0$. In $\mathbb{M}_n(\mathbb{C})$, beside the usual matrix product, the entrywise product of two matrices $T = (t_{ij})$ and $S = (s_{ij})$ is called their Hadamard(Schur) product $T \circ S = (t_{ij}s_{ij})$, a principal submatrix of the tensor product $T \otimes S = (t_{ij}S)_{1\leq i,j\leq n}$. The Schur Theorem says that the Hadamard product of two positive semidefinite matrices is positive semidefinite. Therefore, if $T = (t_{ij})$ is positive semidefinite and $x_1, x_2, ..., x_n$ are real numbers, then the matrices $(t_{ij})^k$ and $(x_ix_jt_{ij}) = \mathrm{diag}(x_1, x_2, ..., x_n)T\mathrm{diag}(x_1, x_2, ..., x_n)$ are positive semidefinite for any positive integer k. We shall say two matrices X and Y are congruent if $Y = S^*XS$ for some nonsingular matrix S(i.e. $\det S \neq 0$). Note that congruence is an equivalence relation, so if X is positive, then every congruence matrix to X is also positive.
The numerical range of an operator $T \in \mathbb{B}(\mathscr{H})$ is the subset of the complex number $\mathbb{C}$, given by

$$W(T) = \{\langle Tx, x\rangle \ : \ x \in H, \| \, x \, \| = 1\}.$$

The following properties of $W(T)$ are immediate:
(i) $W(\alpha I + \beta T) = \alpha + \beta W(T)$ for $\alpha, \beta \in \mathbb{C}$;
(ii) $W(T^*) = \{\overline{\lambda}; \lambda \in W(T)\}$;
(iii) $W(U^*TU) = W(T)$ for any unitary U.
The following lemma and theorems can be found in [1].

[1]Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University
*Speaker. Email address: f.goli@ pgs.usb.ac.ir

Lemma 1.1. *Let T be an operator on a two-dimensional space. Then $W(T)$ is an ellipse whose foci are the eigenvalues of T.*

Theorem 1.2. *The numerical range of an operator is convex.*

Theorem 1.3. *The spectrum of an operator is contained in the closure of its numerical range.*

The numerical radius $w(T)$ of an operator $T \in \mathbb{B}(\mathscr{H})$ is given by

$$w(T) = \sup\{|\lambda| \ : \ \lambda \in W(T)\}.$$

The numerical radius is not unitarily invariant but it is weakly unitarily invariant. This means that $w(UTU^*) = w(T)$ for any unitary matrix U. The numerical radius, $w(\cdot)$ is a norm.
This norm is equivalent to the operator norm $\|\cdot\|$. Recall that the operator norm of an operator $T \in \mathbb{B}(\mathscr{H})$ is defined as $\|T\| = \sup_{\|x\|=1} \|Tx\|$. In fact for any $T \in \mathbb{B}(\mathscr{H})$, we have

$$\frac{\|T\|}{2} \leq w(T) \leq \|T\|. \tag{1}$$

The inequalities in (1) are sharp: The second inequality becomes an equality if T is normal, while the first inequality becomes an equality if $T^2 = 0$.

For any operator $T \in \mathbb{B}(\mathscr{H})$ we have the following refinement of the seconed inequality in (1):

$$w(T) \leq \frac{1}{2}(\|T\| + \|T^2\|^{\frac{1}{2}}).$$

Also, a considerable improvement of the first inequality in (1) is given as follows:

$$\frac{\|T\|}{2} + \frac{|\|ReT\| - \frac{\|T\|}{2}|}{4} + \frac{|\|ImT\| - \frac{\|T\|}{2}|}{4} \leq w(T).$$

Kittaneh proved that for any operator $T \in \mathbb{B}(\mathscr{H})$

$$\frac{1}{4}\|T^*T + TT^*\| \leq w^2(T) \leq \frac{1}{2}\|T^*T + TT^*\|. \tag{2}$$

Since

$$\frac{1}{4}\|T\|^2 \leq \frac{1}{4}\|T^*T + TT^*\| \leq w^2(T) \leq \frac{1}{2}\|T^*T + TT^*\| \leq \|T\|^2,$$

then inequalities (2) improve the inequalities (1).
Similarly in [3] is shown for $T \in \mathbb{B}(\mathscr{H})$, $r \geq 1$ and $0 < \alpha < 1$, we have

$$w^r(T) = \frac{1}{2}\||T|^{2\alpha r} + |T^*|^{2(1-\alpha)r}\|,$$

$$w^{2r}(T) = \|\alpha|T|^{2r} + (1-\alpha)|T^*|^{2r}\|.$$

We recall the following result.

$$w(Y \circ Z) \leq \max_i y_{ii} w(Z), \tag{3}$$

where $Y \in \mathbb{M}_n^+(\mathbb{C})$ and $Z \in \mathbb{M}_n(\mathbb{C})$. The numerical radius for 2×2 operator matrices has the following well know properties.

$$w(\begin{pmatrix} A & 0 \\ 0 & D \end{pmatrix}) = \max(w(A), w(D)),$$

$$w(\begin{pmatrix} 0 & B \\ C & 0 \end{pmatrix}) = w(\begin{pmatrix} 0 & C \\ B & 0 \end{pmatrix}),$$

$$w(\begin{pmatrix} 0 & B \\ C & 0 \end{pmatrix}) = \frac{1}{2} \sup \|e^{i\theta}B + e^{-i\theta}C^*\|,$$

$$w(\begin{pmatrix} A & B \\ B & A \end{pmatrix}) = \max\{w(A+B), w(A-B)\}, \tag{4}$$

and from (4), we get

$$w(\begin{pmatrix} 0 & B \\ B & 0 \end{pmatrix}) = w(B).$$

Also, we have

$$w \begin{pmatrix} A & B \\ C & D \end{pmatrix} \geq w \begin{pmatrix} A & 0 \\ 0 & D \end{pmatrix},$$

and

$$w \begin{pmatrix} A & B \\ C & D \end{pmatrix} \geq w \begin{pmatrix} 0 & B \\ C & 0 \end{pmatrix}.$$

Also, we have a numerical radius inequalities involving off diagonal operator matrix as follows:

$$\frac{\max\{w(B+C), w(B-C)\}}{2} \leq w(\begin{pmatrix} 0 & B \\ C & 0 \end{pmatrix}) \leq \frac{w(B+C) + w(B-C)}{2}.$$

Consequently, we have

$$\frac{1}{2}w(B) \leq w(\begin{pmatrix} 0 & B \\ 0 & 0 \end{pmatrix}) \leq w(B).$$

Recall that the celebrated Heinz inequality states that for $A, B \in \mathbb{M}_n^+(\mathbb{C})$, $X \in \mathbb{M}_n(\mathbb{C})$ and $0 \leq \nu \leq 1$, we have

$$2|||A^{1/2}XB^{1/2}||| \leq |||A^{\nu}XB^{1-\nu} + A^{1-\nu}XB^{\nu}||| \leq |||AX + XB||| \tag{5}$$

for any unitarily invariant norm $|||\cdot|||$ on $\mathbb{M}_n(\mathbb{C})$.
Some mathematicians proved several refinements and extensions of this inequality. Sababheh in [4] showed for any $A > 0$, $X \in \mathbb{M}_n(\mathbb{C})$, $a > 0$, $\frac{2-a}{4} \leq \nu \leq \frac{2+a}{4}$ and $-2 < t \leq 2$ the following relation holds:

$$w(A^{\nu}XA^{1-\nu} + A^{1-\nu}XA^{\nu}) \leq \frac{2w(A^{1-a})}{2+t} w(A^aX + tA^{\frac{a}{2}}XA^{\frac{a}{2}} + XA^a). \tag{6}$$

Also, he showed the Young's version of the numerical radius inequality for $A > 0$, $X \in \mathbb{M}_n(\mathbb{C})$, $\nu \in \mathbb{R}$ and $-2 < t \leq 2$ as follows:

$$w(A^{\nu}XA^{1-\nu}) \leq \frac{w(A^{2\nu-1})}{t+2} w(AXA^{1-2\nu} + tA^{\frac{1}{2}}XA^{\frac{3}{2}-2\nu} + XA^{2-2\nu}). \tag{7}$$

In this paper, we extend this inequalites for positive matrices A and B.

2 Main results

In the section we give some extensions of the previous inequalities which are given by some authors.

Theorem 2.1. *Let $A, B > 0$ and $X \in \mathbb{M}_n(\mathbb{C})$. Then for $a > 0$, $\frac{2-a}{4} \leq \nu \leq \frac{2+a}{4}$ and $-2 < t \leq 2$, we have*

$$w(A^\nu X B^{1-\nu} + A^{1-\nu} X B^\nu) \leq \frac{4}{2+t} \max\left(w(A^{1-a}), w(B^{1-a})\right) w(A^a X + t A^{\frac{a}{2}} X B^{\frac{a}{2}} + X B^a).$$

Corollary 2.2. *Let $A, B > 0$. Then*

$$w(A+B) \leq \max\left(w(A^{\frac{1}{2}}), w(B^{\frac{1}{2}})\right) w(A^{\frac{1}{2}} + 2A^{\frac{1}{4}} B^{\frac{1}{4}} + B^{\frac{1}{2}}).$$

Corollary 2.3. *Suppose that $A, B > 0$ and $X \in \mathbb{M}_n(\mathbb{C})$. Then*

$$w(A^\nu X B^{1-\nu} + A^{1-\nu} X B^\nu) \leq 2w(AX + XB).$$

In particular,

$$w(A^{\frac{1}{2}} B^{\frac{1}{2}}) \leq w(A+B).$$

Theorem 2.4. *Let $A, B > 0$ and $X \in \mathbb{M}_n(\mathbb{C})$. Then for $\nu \in \mathbb{R}$ and $-2 < t \leq 2$,*

$$w(A^\nu X B^{1-\nu}) \leq \frac{2}{t+2} \max\left(w(A^{2\nu-1}), w(B^{2\nu-1})\right) w(AXB^{1-2\nu} + tA^{\frac{1}{2}} X B^{\frac{3}{2}-2\nu} + XB^{2-2\nu}).$$

Corollary 2.5. *Suppose $A, B > 0$ and $X \in \mathbb{M}_n(\mathbb{C})$. For $p, q > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$,*

$$w(A^{\frac{1}{p}} X B^{\frac{1}{q}}) \leq \max\left(w(A^{\frac{2}{p}-1}), w(B^{\frac{2}{p}-1})\right) w(AXB^{1-\frac{2}{p}} + XB^{\frac{2}{q}}).$$

In particular,

$$w(A^{\frac{1}{2}} X B^{\frac{1}{2}}) \leq w(AX + XB).$$

3 Conclusion

By an example we show that the inequality $w(A^\nu X B^{1-\nu} + A^{1-\nu} X B^\nu) \leq w(AX + XB)$ is not true in general. In this paper, we give an upper bound for $w(A^\nu X B^{1-\nu} + A^{1-\nu} X B^\nu)$.

References

[1] K.E. Gustafson and D.K.M. Rao, *Numerical Range, The Field of Values of Linear Operators and Matrices*, Springer, New York, 1997.

[2] F. Goli, R. Lashkaripour and M. Hajmohamadi, Some extentions of the numerical radius for the Young and Heinz types of operators, *Linear and Multilinear Algebra*, submited.

[3] M.El-Haddad and F. Kittaneh, Numerical radius inequalities for Hilbert Space Operator, *Studia Math*, 182 (2007), No. 20, 133–140.

[4] M. Sababheh, Heinz-type numerical radii inequalities, *Linear and Multilinear Algebra*, 347 (2019), No. 2, 953–964.

Classification of handwritten digits using Singular Value Decomposition[1]

Abdolhossein Naserasadi[1,*], Ali Hassani[2] and Faranges Kyanfar[1]

[1]Department of Applied Mathematics, Shahid Bahonar University of Kerman, Kerman, Iran

[2]Department of Computer Science, Shahid Bahonar University of Kerman, Kerman, Iran

Abstract

In this paper, Singular Value Decomposition is used to classify handwritten digits. Handwritten digit classification is a subarea of pattern recognition. Digital images have a basic matrix representation, which is used to detect different patterns of the same digits using SVD. Afterwards, any new sample images can be classified with a high accuracy level for recognition, which is reported.

Keywords: Singular Value Decomposition, Handwritten digits, Pattern recognition, MNIST dataset

Mathematics Subject Classification [2010]: 15A18, 68T10

1 Introduction

A classic problem in the field of pattern recognition is that of handwritten digit recognition. Suppose that we have an image of a digit submitted by a user via a scanner, a tablet, or other digital device. The goal is to design an algorithm that can correctly identify the digit. The applications of such an algorithm are far reaching. With this technology, the post office would be able to scan envelopes and effectively sort them by zip code and banks would be able to process checks more efficiently.

In this paper, we present a simple yet effective algorithm which assumes that each set of digits lies in subspace whose basis is obtained via the idea of Singular Value Decomposition (SVD). When an unknown digit is read in, we project the digit onto each of the ten subspaces and classify the digit according to the smallest residual vector under the 2-norm.

Here vectors are used to represent digits. The image of one digit is a 28×28 matrix of numbers, representing gray scale. It can also be represented as a vector in $\mathbb{R}^{784}$, by stacking the columns of the matrix. A set of n digits (handwritten 3's, say) can then be represented by matrix $A^{784\times n}$, and the columns of A can be thought of as a cluster. They also span a subspace of $\mathbb{R}^{784}$.

[1]Dedicated to Alireza Afzalipour and Fakhereh Saba, the founders of Kerman University

*Speaker. Email address: hosseinnaserasadi74@gmail.com

2 Classification of Handwritten Digits using SVD bases

The Singular Value Decomposition is a standard technique used in data analysis for the purpose of dimensionality reduction. Here it will be used as a tool for classification. Before we delve into the details of its application, let us first review some of the theoretical background about singular value decomposition.

Theorem 2.1. *Let A be an $m\times n$ matrix $(m \geq n)$ with nonzero singular values $\sigma_1, \sigma_2, \cdots, \sigma_r$ then there exist orthogonal matrix $U \in \mathbb{R}^{m\times m}$ also $V \in \mathbb{R}^{n\times n}$ such that:*

$$A = U\Sigma V^T, \Sigma = \begin{pmatrix} \Sigma_0 \\ 0 \end{pmatrix} \in \mathbb{R}^{m\times n}, \Sigma_0 = diag(\sigma_1, ..., \sigma_n), \tag{1}$$

therefore rank of A is r and we can write:

$$A = U\Sigma V^T = (U_r \hat{U}_r) \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} V_r^T \\ \hat{V}_r^T \end{pmatrix} = U_r \Sigma_r V_r^T, \tag{2}$$

where

$$U_r \in \mathbb{R}^{m\times r},\ \Sigma_r = diag(\sigma_1, ..., \sigma_r) \in \mathbb{R}^{r\times r},\ V_r \in \mathbb{R}^{n\times r}. \tag{3}$$

The SVD can be used to compute the rank of a matrix. However, the zero singular values usually appear as small numbers. Similarly, if A is made up from a rank k matrix and additive noise of small magnitude, then it will have k singular values that will be significantly larger than the rest. If trailing small diagonal elements of Σ are replaced by zeros, then a rank k approximation A_k of A is obtained as

$$A = (U_k \hat{U}_k) \begin{pmatrix} \Sigma_k & 0 \\ 0 & \hat{\Sigma}_k \end{pmatrix} \begin{pmatrix} V_k^T \\ \hat{V}_k^T \end{pmatrix} \approx (U_k \hat{U}_k) \begin{pmatrix} \Sigma_k & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} V_k^T \\ \hat{V}_k^T \end{pmatrix} = U_k \Sigma_k V_k^T = A_k,$$

such that

$$\Sigma_k \in \mathbb{R}^{k\times k}, \|\hat{\Sigma}_k\| < \epsilon.$$

Theorem 2.2. *Let $\|.\|$ denote any orthogonally invariant norm, and let the SVD of $A \in \mathbb{R}^{m\times n}$ be given as in Theorem 2.1 Assume that an integer k is given with $0 < k \leq r = rank(A)$. Then*

$$\min_{rank(B)=k} \|A - B\| = \|A - A_k\|,$$

where

$$A_k = U_k \Sigma_k V_k^T = \sum_{i=1}^{k} \sigma_i u_i v_i^T.$$

2.1 Theory and Algorithm

The problem we face in this section is:

Given a set of of manually classified digits (the training set), classify a set of unknown digits (the test set).

Each image of a handwritten digit can be considered as an $m \times m$ matrix where each entry in the matrix is a gray scale pixel value. The columns of each image are stacked to form a column vector of size $m^{2\times 1}$. All the stacked images of a single digit are concatenated to form a matrix $A_j \in \mathbb{R}^{m^2 \times n}$, with n being the number of training images for a particular digit and $j = 0, 1, \cdots, 9$ being the particular digit.

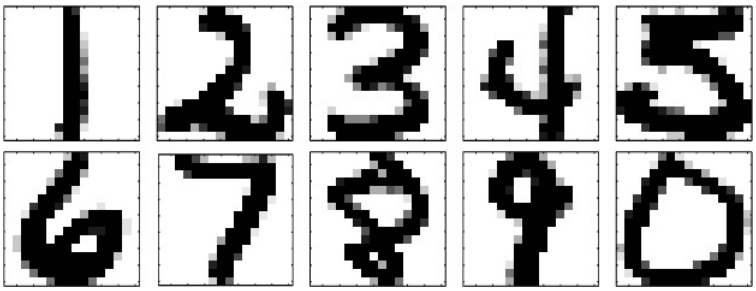


Figure 1: Handwritten digits from the US Postal Service Database.

The idea now is to model the variation within the set of training digits of one kind using an orthogonal basis of the subspace. An orthogonal basis can be computed using the SVD, and A can be approximated by a sum of rank 1 matrices Theorem2.2,

$$A \approx \sum_{i=1}^{k} \sigma_i u_i v_i^T,$$

for some value of k. Each column in A is an image of a digit 3, and therefore the left singular vectors u_i are an orthogonal basis in the image space of 3's. We will refer to the left singular vectors as singular images. From the matrix approximation properties of the SVD (Theorem 2.2) we know that the first singular vector represents the dominating direction of the data matrix. Therefore, if we fold the vectors u_i back to images, we expect the first singular vector to look like a 3, and the following singular images should represent the dominating variations of the training set around the first singular image.

The SVD basis classification algorithm will be based on the following assumptions.

1. Each digit (in the training and test sets) is well characterized by a few of the first singular images of its own kind.

2. An expansion in terms of the first few singular images discriminates well between the different classes of digits.

3. If an unknown digit can be better approximated in one particular basis of singular images, the basis of 3's say, than in the bases of the other classes, then it is likely that the unknown digit is a 3.

Thus we should compute how well an unknown digit can be represented in the ten different bases. This can be done by computing the residual vector in least squares problems of the type

$$\min_{\alpha_i} \|P - \sum_{i=1}^{k} \alpha_i u_i\|,$$

where P represents an unknown digit, and u_i the singular images. We can write this problem in the form

$$\min_{\alpha} \|P - U_k\alpha\|_2,$$

where $U_k = (u_1, u_2, \cdots, u_k)$. Instead of solving this minimization problem, we can equivalently solve for the square of the 2-norm.

$$\begin{aligned}
&\min_{\alpha} \|P - U_k\alpha\|_2^2 \\
&= \min_{\alpha} (P - U_k\alpha)^T (P - U_k\alpha) \\
&= \min_{\alpha} (P^T - \alpha^T U_k^T)(P - U_k\alpha) \\
&= \min_{\alpha} (U^T P - P^T U_k\alpha - \alpha^T U_k^T P + \alpha^T\alpha)
\end{aligned}$$

Taking the derivative of the last expression with respect to α and setting it equal to zero we get

$$\begin{aligned}
2\alpha^T - 2P^T U_k &= 0 \\
\alpha^T &= P^T U_k \\
\alpha &= U_k^T P
\end{aligned}$$

and the norm of the residual vector of the least squares problems is

$$\|(I - U_k U_k^T)P\|_2.$$

Intuitively, $U_k^T P$ is the projection of P onto the digit space so the distance is just the 2-norm of the residual vector. In figure2, we have a geometric illustration of the scenario where $S = span(Uk)$.

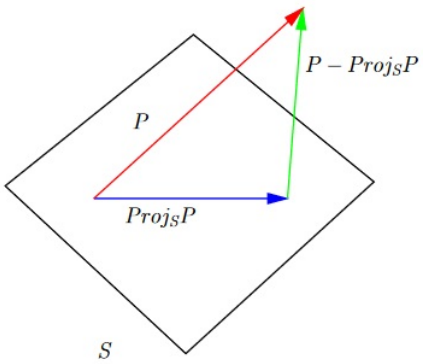


Figure 2: The probe P, its projection onto $S = span(Uk)$, and the residual.

The proposed algorithm is described in Algorithm 2.3. The key advantages of this algorithm are simplicity and lower complexity in large sets of data, which is further discussed in the next section.

Algorithm 2.3. SVD Basis Classification Algorithm

1. Create matrices $X_0, X_1, ..., X_9$ from $TrainingSet$ where
   X_i is a column matrix of all records belonging to class i, $i = 0, 1, ..., 9$.
2. For $i \in \{0, 1, ..., 9\}$ do
3. Compute the SVD of X_i, $U, \Sigma, V^T = X_i$
4. $D_i \leftarrow I - U_k U_k^T$
5. end for
6. For test vector $P \in TestSet$ do
7. $j = \text{argmin}_{i \in 0,1,...,9} \|D_i P\|_2$
8. Classify vector P as belonging to class j
9. end for
10. return $TestSet$ Labels

3 Numerical results

The results on a partial subset of the MNIST Digits dataset, containing 42,000 sample 28×28 dimensional gray scale images, are provided below. The method is compared to three other classifiers: K-Nearest Neighbors classifier, a Linear Kernel Support Vector Machine classifier and a Neural Network classifier. In order to evaluate the performance of this classification algorithm, we used 5-fold cross-validation for train/test split. This means that the entire dataset was shuffled and partitioned into 5 subsets. Then, for each subset, a new model is created in which the rest of the subsets are the training data and the subset itself is the test data. The parameter k was also set to 20 in these experiments. The neural network was trained using Adam optimizer and consists of 1 hidden layer with 100 neurons. The parameter K for the nearest neighbors classifier was set to 5. The linear SVM was set to $l2$ norm penalty and the regularization parameter was set to 1.0. All implementations were done using Python and the three other classifiers were already implemented by Scikit-Learn library. The SVD classifier was implemented using Numpy's SVD module. The results are provided in Table 1.

Table 1: Results of the 5-fold cross-validation on MNIST Digits

Method	Mean Accuracy	Computational Time (seconds)
SVD Classifier	95.49	51
KNN Classifier	96.68	895
Support Vector Machine Classifier	90.60	231
Neural Network Classifier	95.67	96

As it can be seen, the proposed classifier comes very close to KNN and the neural network which are some of the most widely used classifiers in terms of accuracy, while carrying less computational burden. The proposed approach exceeds the linear SVM classifier in terms of accuracy as well. The computational time provided is the mean time taken at each fold to train on the training data and classify the test data. It is worth

noting that while KNN has the longest computational time, almost 98 percent of this time was spent for classifying test data, which points to its weakness. The neural network on the other hand took less than a second to classify test data, while taking over 90 seconds to train, which also points out its slower training phase when compared to the proposed method. The same standard holds for the linear SVM which took over 230 seconds to train on the dataset, but took only a fraction of a second to evaluate the test data. This method however took only 21 seconds to train and 30 seconds to classify the test data, which shows a more moderate time in both overall.

4 Conclusion

In this paper, a basic representation for digital images was presented, and singular value decomposition was employed to extract patterns between images of the same class. These patterns can help classify new images, and as the test results show, is quite accurate in doing so, when provided with enough training data. When compared to other classifiers, it can be seen that this method is more efficient in terms of complexity when considering a large number of samples, which in this case was a total of 42,000 images. Possible future extensions of this method may be an extension to other types of numerical data, as well as colored images.

References

[1] Lars Elden, *Matrix methods in data mining and pattern recognition*, volume 15. SIAM, 2019.

[2] Lars Elden, Numerical linear algebra in data mining, *Acta Numerica*, 15:327-384, 2006.

[3] Andy Lassiter and Serkan Gugercin, Handwritten digit classification and reconstruction of marred image using singular value decomposition, 2013.

[4] Pacheco, Jose Israel, *A comparative study for the handwritten digit recognition problem*, California State University, Long Beach, 2011.

Author Index

*Speaker

S

T

U

V

X

Y

Z