Proceedings of the Conference on March 14 - International Day of Mathematics

Vuković, Mirjana, urednik; Nurkanović, Mehmed, urednik

2024-12-26

Academy of Sciences and Arts of Bosnia and Herzegovina

https://bastina.anubih.ba/handle/123456789/798

PROCEEDINGS OF THE CONFERENCE ON
MARCH 14 – INTERNATIONAL DAY OF MATHEMATICS

AKADEMIJA NAUKA I UMJETNOSTI BOSNE I HERCEGOVINE
АКАДЕМИЈА НАУКА И УМЈЕТНОСТИ БОСНЕ И ХЕРЦЕГОВИНЕ
ACADEMY OF SCIENCES AND ARTS OF BOSNIA AND HERZEGOVINA

Posebna izdanja
Knjiga CCXVI

Odjeljenje prirodnih i matematičkih nauka
Knjiga 30

ZBORNIK RADOVA S KONFERENCIJE
14. MART - INTERNACIONALNI DAN MATEMATIKE

Sarajevo, 14. mart 2024.

Urednici
Akademkinja Mirjana Vuković
Prof. dr. Mehmed Nurkanović

SARAJEVO, 2024.

DOI: 10.5644/PI2024.216.00

AKADEMIJA NAUKA I UMJETNOSTI BOSNE I HERCEGOVINE
АКАДЕМИЈА НАУКА И УМЈЕТНОСТИ БОСНЕ И ХЕРЦЕГОВИНЕ
ACADEMY OF SCIENCES AND ARTS OF BOSNIA AND HERZEGOVINA

Special Editions
Volume CCXVI

Department of Natural and Mathematical Sciences
Volume 30

PROCEEDINGS OF THE CONFERENCE ON MARCH 14 - INTERNATIONAL DAY OF MATHEMATICS

Sarajevo, 14th March 2024

Editors
Academician Mirjana Vuković
Mehmed Nurkanović, PhD

SARAJEVO 2024

PROCEEDINGS OF THE CONFERENCE ON MARCH 14 – INTERNATIONAL DAY OF MATHEMATICS
Sarajevo, 14th March 2024

Publisher
Academy of Sciences and Arts of Bosnia and Herzegovina

For the publisher
Academician Muris Čičić

Organizing committee
Academician Mirjana Vuković
Mehmed Nurkanović, PhD
Amela Muratović Ribić, PhD
Zehra Nurkanović, PhD
M. Sc. Amra Avdagić
M. Sc. Suada Alić
Naida Nišić

Editors
Academician Mirjana Vuković
Mehmed Nurkanović, PhD

Lecturer for English
Lejla Miller, PhD

DTP
Rasim Kovačević

The press
Dobra knjiga

Printing
100

Sarajevo 2024

EBSCO

ISBN 978-9926-574-07-9
CIP zapis dostupan u COBISS sistemu Nacionalne i univerzitetske biblioteke BiH pod
ID brojem 62808582

This proceedings book was published with the support of the Municipality of Centar, Sarajevo
Štampanje ovog zbornika podržano je od strane Općine Centar, Sarajevo

Općina **Centar** Sarajvo

THE PAPERS PRESENTED AT THE CONFERENCE ON MARCH 14 – INTERNATIONAL DAY OF MATHEMATICS DEDICATED TO THE JUBILEE OF ACADEMICIAN PROF. DR. MIRJANA VUKOVIĆ

CONTENTS

PROCEEDINGS OF THE CONFERENCE ON MARCH 14 – INTERNATIONAL DAY OF MATHEMATICS
Special Editions ANUBiH, Book CCXVI, OPMN Book 30
Sarajevo 2024, **p. 1–4** DOI: 10.5644/PI2024.216.01

GREETINGS TO THE PARTICIPANTS AND A NOTE ABOUT THE DAY OF MATHEMATICS AND MATHEMATICS

MIRJANA VUKOVIĆ

ABSTRACT. After a short greeting addressed to those present, the paper begins with a note about March 14 – International Day of Mathematics and ends with a panoramic view of mathematics – the queen of sciences by Gauss.

Dear colleagues and fellow mathematicians, above all dear friends,
Dear participants of *the Conference dedicated to March* 14^{th} *– the International Day of Mathematics*, *which we are celebrating for the first time at our Academy of Sciences and Arts of Bosnia and Herzegovina*,
Dear students, especially students of my high school – Third Gymnasium – I hope that at least some of you will become future mathematicians and thus succeed us as your teachers and professors,
Dear participants – Ladies and Gentlemen!

It is my great honor and pleasure to greet you on behalf of the *Academy of Sciences and Arts of Bosnia and Herzegovina* – the temple of all sciences, including mathematics, and especially the Department of Natural Sciences and Mathematics, which is the organizer of this Scientific Conference, as well as on my behalf. I am honored to be a part of this wonderful conference in which my former students participate – that is, my students with their students, and especially because the conference is dedicated to our dear mathematics. On this occasion, I must mention the first mathematicians – my great and dear teachers, academicians Mahmut Bajraktarević and Fikret Vajzović, to whom I was an assistant, Manojlo Maravić and Veselin Perić, who were my mentors, for my doctorate and master's degree.

Special Editions ANUBiH, Book CCXVI, OPMN Book 30, pp. 1–4

I will begin my presentation with a beautiful sentence by academician physicist Anatoly L. Bukhachenko: *"There is a real miracle in the world – it is Mathematics, a divine, royal science, a magical invention of people, created at the top of the mind and the tip of a pen. This miraculous science has the magical ability to predict the unpredictable and connect the unconnected. It amazes with its magical and mystical ability to predict properties and phenomena."*

So isn't it natural that a special day is dedicated to such a special science to emphasize its importance? Thus, at the proposal of the *International Mathematical Union*, the 40$^{\text{th}}$ General Assembly of UNESCO, in November 2019, declared *March* 14$^{\text{th}}$ *as the International Day of Mathematics*.

Since then, March 14$^{\text{th}}$ has been celebrated as *International Day of Mathematics*, while until 2020, that day was celebrated as π*-number Day* or π*-Day*. Even as π*-Day*, that date was very significant for mathematics. Namely, it is well known that the problem of squaring the circle – one of the oldest and best-known problems (negatively) was solved only when the transcendence of the number π was proven (*Ferdinand von Lindemann*, 1882), which also speaks of the importance of the number π. Proclaiming March 14$^{\text{th}}$ as International Day of Mathematics highlighted the importance of mathematics – *the "queen of science"*!

It is difficult to list all the reasons *why the International Day of Mathematics is important.*

First of all, mathematics is present in our everyday life. No matter how uninteresting, even boring it may seem to some, it is certainly one of the oldest and most important sciences, both for education and for life in general, because it develops the ability to think and teaches us an analytical way of thinking. Mathematics promotes wisdom and quickens our minds. But mathematics is also essential in a world of constant change.

History proves that mathematics is an old science, along with philosophy, one of the oldest. In ancient times, and even for a long time after that, there was almost no philosopher who was not at the same time a mathematician and vice versa.

Thus, the aforementioned problem of squaring the circle was being solved by numerous mathematicians for almost 2500 years. In the Rhind papyrus, a rule for approximately determining the side of a square whose area is equal to the area of a given circle is presented, but the Greeks were not satisfied with approximate solutions. The first surviving records of the problem of squaring the circle testify that it was already dealt with in the 5$^{\text{th}}$ century BC. *Anaxagoras* – the founder of the Athenian school of philosophy and *Antiphon* from Athens, as well as *Archimedes*, and much later *Leonardo da Vinci*. All attempts to solve this problem, as well as the attempts of Arab mathematicians who also dealt with this problem and determined the number π much more precisely (*Al Kashi* in the 14$^{\text{th}}$ century), remained without results. Much later, the German mathematician *F. Von Lindemann* showed that the number π cannot be elementary constructed, as well as that π is not a solution to any algebraic equation with integer coefficients. The problem of squaring the circle was practically reduced to the construction of the number π.

From everything that has been said so far, it can be seen that already with its paradigmatic place in the domain of human knowledge, independent of all other valid reasons,

mathematics deserves a special place.

The oldest known thinkers of the ancient civilization noticed the characteristic of the mathematical form of knowledge and since then it has served as a model of scientificity and a measure of exactness of the overall knowledge.

Thus, already in the Middle Ages, mathematics, in its division of that time, constituted two of the seven skills, for which study was dedicated the traditional university (geometry and arithmetic) in the quadrivium. And the third seventh – logic in the trivium, today, in the relevant part, in the form of mathematical logic, would also be classified in the domain of mathematics.

In the foundations of the new century, in the spiritual and material implications of which, by the way, we still live today, Gallilei's knowledge is incorporated, according to which *"the book of nature is written in the language of mathematics"*, while, according to Kant, *"individual scientific disciplines reach the level of scientificity as their need for the language of mathematics in the formulation and expression of one's knowledge"*.

Now we will say something more about mathematics and hierarchy in science in general, as seen by *Friedrick Turner*[1] a specialist in social sciences from the University of Texas:

"There is, he says, a pyramid of science. The base of that pyramid is mathematics, not because it is more abstract, or because it is in some sense *"better or more loved"* than other sciences, but because it does not have to rely on any other science, while physics, which belongs to the next layer (floor) of the pyramid, must inevitably relies on mathematics. The floor above physics is chemistry, which cannot do without its support, and its support is, of course, what is below it, that is physics. The next floor is biology, which must rely on a good knowledge of both physics and chemistry."

With an old saying, this could also be expressed oppositely: physicists admit only to mathematics, and mathematicians only to God, that it is above them.

I will end the story of mathematics with *Gauss's saying "mathematics is the queen of science", Gauss is the prince of mathematics*, and the saying of the great Swedish mathematician *Gösta Mittag-Leffler*: *"The best work of mathematicians is art, highly perfect art, like the most secret dreams of the imagination, clear and bright. Mathematical genius and artistic genius touch each other."*

In addition, March 14^{th} is celebrated to popularize mathematics, as a science that played a significant role in the development of civilization, and which is of great importance, both for the present and for the future of civilization.

Therefore, March 14^{th} – *International Day of Mathematics* is also important for raising the level of awareness of the importance of mathematics, not only because it contributes to the development of logical thinking through school, but also because it has an extremely wide application in all areas of everyday life, including, in addition to physics and technology, even medicine, music, sports, as well as numerous other fields.

And finally I would like to extend my heartfelt gratitude to all of you for this wonderful Conference, especially to my ex-students, to all who presented your papers, as well as

[1]Leon Lederman (with Dick Teresi), God Particle (In serbian), Series of Popular Science SFINGA, Beograd, 1998.

to students of my high school – Third Gymnasium.

I am especially thankful to two of my dear ex-students now colleagues – university professors Mehmed Nurkanović and Amela Muratović Ribić who initiated this event and have lightened up the spark within you and put in a lot of effort to organize this event and prepare this Proceedings dedicated to my jubilee.

Academician Mirjana Vuković, ANUBiH

SOME PHOTOS FROM THE CONFERENCE

Academicians M. Vuković and D. Milošević before the start of the Conference

Conference participants: academicians B. Peruničić and L. Lincender Cvijetić and professors V. Vladičić and M. Pikula

Students of the Third Gymnasium who participated with their presentation: Andrea Božinović, Ema Džafić, Amina Bahtanović, Fatima Fulurija, and Dženita Podžić

PROCEEDINGS OF THE CONFERENCE ON MARCH 14 – INTERNATIONAL DAY OF MATHEMATICS
Special Editions ANUBiH, Book CCXVI, OPMN Book 30
Sarajevo 2024, **p. 5–14** DOI: 10.5644/PI2024.216.02

HAPPY YUBILEE PROFESSOR MIRJANA!

MEHMED NURKANOVIĆ

ABSTRACT. This special edition of ANUBiH, the Collection of papers presented at the Conference on the occasion of World Mathematics Day, ANUBiH, Sarajevo, March 14, 2024, and the Conference itself, is dedicated to our dear Academician Prof. Dr. Mirjana Vuković in honor of her significant jubilee. Here, we will give a brief overview of her life and work.

1. EDUCATION AND ACADEMIC CAREER OF PROFESSOR MIRJANA VUKOVIĆ

As a student and long-time associate of the Editorial team of the Sarajevo Journal of Mathematics, I am familiar with many details about the life of Professor and Academician Mirjana Vuković, a person I hold dear. However, considering her exceptionally rich academic career, intertwined with teaching activities at the universities where she worked, her extensive and successful scientific and research contributions, and her significant social engagements, it is impossible to include all the details of her life in a short essay. Therefore, I have provided a brief overview of her life and work.

2010 *Mathematics Subject Classification.* 01A70, 01A65.

Key words and phrases. Graded and paragraded structures: groups, rings, modules, radicals of paragraded rings, Krull's theorem, Wedderburn-Artin theorem, ADS-theorem for paragraded rings.

Mirjana Vuković was born in 1948 in Fojnica and from the maternity hospital arrived in Sarajevo. Due to the nature of her father's career as a partisan and high-ranking officer, her family frequently moved. Thus, she attended the first six grades of elementary school in Varaždin and Maribor and completed her primary education in Sarajevo upon the family's return. She then attended the Braća Ribar High School (now Third Gymnasium), graduating as one of the school's top students. She went on to study mathematics at the Department of Mathematics of the Faculty of Natural Sciences and Mathematics in Sarajevo despite having the qualifications to pursue any field of study, including the arts. She completed her studies in record time, graduating as the top student in her class. Her exceptional academic achievements earned her numerous accolades: all silver medals (for individual academic years), a gold medal (for overall academic excellence), and scholarships from the Hasan Brkić University Fund, the first of which was awarded upon the recommendation of her professors, Academician Mahmut Bajraktarević and Professor Šefkija Raljević who was the Dean at the time.

Immediately after graduating, she began her academic career at the same Department, first as an Assistant (1972), then as an Assistant Professor (1979), an Associate Professor (1984), and finally as a Full Professor (1989). It is worth noting that Professor M. Vuković also pursued her education at some of the world's most prestigious universities, including the Moscow State University Lomonosov (1975/76 academic year) and the Pierre and Marie Curie University in Paris in December 1976, when her collaboration with the renowned French mathematician Marc Krasner began. Thus, referring to her educational journey, she often says, "my three universities," meaning the University of Sarajevo, Moscow State University Lomonosov, and Paris's Pierre and Marie Curie University.

At the Faculty of Natural Sciences and Mathematics in Sarajevo, she completed a two-year postgraduate program, thus earning a Master of Mathematical Sciences degree (1975) with a thesis entitled "Hensel fields and Henselisations" (written in Serbo-Croatian). She then earned her Ph.D. in Mathematical Sciences (1979) by defending her dissertation "Some Problems on Summability and Their Applications to Generalised Fourier Series" (in Serbo-Croatian), under the mentorship of Academician Manojlo Maravić.

During her illustrious academic career, Professor Mirjana held several significant positions at the Faculty of Natural Sciences and Mathematics and the University of Sarajevo. She served as Vice-Dean for Science and Teaching (1982–1984) and was elected, in two consecutive terms, as the youngest Vice-Rector of the University of Sarajevo for Science, Teaching and Scientific Research (1988–1993). A crowning achievement of her scientific research contributions and her impact on the development of mathematics in Bosnia and Herzegovina was her election as a corresponding member of the Academy of Sciences and Arts of Bosnia and Herzegovina in 2012 and a full member in 2018. With this election, she became the first female mathematician and the first woman elected to the Department of Natural and Mathematical Sciences of the Academy.

As I write this text, I recall my time as a student at the Department of Mathematics of the Faculty of Natural Sciences and Mathematics in Sarajevo. Professor Mirjana Vuković, then a young and much-loved professor, taught Analysis 3 to third-year students in the teaching course and Complex Analysis to fourth-year students in the general mathematics course. Taking over the challenging legacy of her professor and distinguished lecturer, Academician M. Bajraktarević, Professor Mirjana successfully maintained the high quality of lectures in Complex Analysis, becoming thus a worthy successor to Academician Bajraktarević. I can confidently say that Analysis 1, taught by Academician Bajraktarević, and Complex Analysis were my favorite subjects during my studies. Later, Professor Mirjana also taught in other various fields of mathematics, such as algebra. It is particularly noteworthy to mention the wartime period when, despite extremely difficult circumstances and even being injured, she, along with a few other professors, made significant efforts to keep the Department of Mathematics at the Faculty of Natural Sciences and Mathematics functioning. She also taught mathematics at technical faculties within the University of Sarajevo during that challenging time.

2. SCIENTIFIC RESEARCH WORK OF PROFESSOR MIRJANA VUKOVIĆ

2.1. **About Her Scientific Contributions**

Professor Mirjana Vuković's scientific interest is related to several important and contemporary areas of mathematical analysis and modern algebra. Her scientific papers are predominantly of a foundational theoretical nature.

Her work in mathematical analysis primarily focuses on summability theory and Fourier analysis. While preparing her doctoral dissertation under the mentorship of Academician Manojlo Maravić, she explored issues related to the summability of multiple Fourier series and the summability of expansions in terms of the eigenfunctions of the Laplace operator. Initially, she investigated some properties of the class G_h^κ – summability methods and problems of the inversion for this class of methods, proving, in doing so, several Tauberian-type theorems [1] and a Convexity Theorem for the G_h^κ – summability method [2]. At that time, only two convexity theorems for summability methods were known. The first such theorem was established by the famous Hungarian mathematician M. Riesz, whose summability method was named after him i. e. Riesz summability method. Academician M. Maravić proved the convexity theorem for the G_θ^κ – method, while M. Vuković proved the third theorem for the G_h^κ – summability method [2]. The results that M. Vuković obtained in this research formed a significant part of her doctoral dissertation. However, later, in a paper [28], M. Vuković with her students E. Ilić-Georgijević and O. Stevanović applying Parseval's formula proved a G_h^κ – summability analogue of Avadhani's theorem for the Riesz–summability of the eigenfunction expansion. A crucial step in the proof of this theorem was to find a function $g(x)$ that would lead them to the kernel of the G_h^κ – summability, which is more complex than the kernel of the Riesz summability.

Let us also point out that Prof. Mirjana participated in research in the fields of *Functional Analysis* and *Theory of Distributions*. In a joint paper, M. Vuković with academicians S. Pilipović and F. Vajzović in [25] studied various classes of distribution of

semigroups on the spaces of functions $\mathcal{F}_r$, $r \in \mathbb{R}$ distinguished by their behavior at the origin. In the paper [26], M. Vuković with S. Pilipović and A. Bučkovska proved an approximation result for the bilinear Hilbert transform and used it for the inversion of the bilinear Hilbert transform. Also, they analyzed p-Lebesgue points ($p \geq 1$).

A little later, in [27], M. Vuković, as a good expert in Fourier analysis, together with I. Zubac analyzes quasiasymptotic boundedness of distributions and their wavelet transforms in general, as well as for a class of exponentially bounded distributions and their wavelet transforms in particular. The main idea of this paper is to use, instead of the quasiasymptotic behaviour, the notion of quasiasymptotic boundedness. In this way, they obtain new Abelian type theorems for the wavelet transform of distributions with different growth.

The crowning achievement of Professor Mirjana Vuković's scientific work lies in her papers in the field of abstract algebra, to which she dedicated most of her time and where she achieved exceptional results. Her initial findings in this area emerged through collaboration with the renowned French mathematician Marc Krasner. In those papers a new abstract theory was established – *the theory of paragraded structures*, which are increasingly and rightfully referred to as the *Krasner-Vuković paragraded structures (groups, rings, modules)*. These represent fundamental theoretical results in algebra, which M. Vuković and M. Krasner obtained by addressing the long-standing question: *under what conditions are graded structures (groups, rings, modules) closed with respect to the direct sum and the direct product*? In this way, their solution led to structures more general than the Bourbaki-Krasner graded structures. These appeared for the first time in their joint papers, Paragraded Structures (in French) Parts I, II, and III [3–5], published in the prestigious *Proceedings of the Japan Academy, Japan Academy of Sciences*, based on reviews by one of the most prominent and influential Japanese mathematicians of the time, Academician Shokichi Iyanaga. Subsequently, their monograph Paragraded Structures (Groups, Rings, Modules) (in French) was published as part of the esteemed *Queen's Papers in Pure and Applied Mathematics* series at *Queen's University, Kingston*, Canada [6]. This paper gained significant attention, even during its announcement in the book *Il mondo Krasneriano* [7] by the renowned Canadian mathematician *Paolo Ribenboim*.

The significance of this monographic work is reflected in the fact that it, along with other papers by M. Vuković in this field, can be found in five world languages across 154 libraries worldwide, including: Université "Pierre et Marie Curie", "Sorbonne Universitée" (Paris); "Lomonosov – Moscow State University – MGU, Moscow"; "Universitée Joseph Fourier", Grenoble; "Karlshruer Institut für Technologie – KIT" (Karlsruhe); Mannheim Universität, and others. These works are also listed in libraries such as "Open Library" and "WorldCat." Additionally, they have been available for purchase online through the "Queen's University Bookstore" and "Amazon.com"...

Let us now elaborate further on the results of her subsequent research in the field of algebra. In [8], after proving the existence and uniqueness of the primary decomposition of moduloids, Prof. Mirjana Vuković and E. Ilić Georgievć briefly turned our attention to Krull's theorem and the existence of the primary decomposition of Krasner-Vuković

paragraded rings. In several articles, she then studied the primary and Jacobson radicals [9, 11, 14], as well as the general theory of radicals [10].

In a joint paper with her student E. Ilić Georgijević [11], they discuss the primary decomposition in the case of general graded modules – moduloids, a generalization of already done work for general graded rings-anneids. These structures, introduced by Marc Krasner are more general than the graded structures of Bourbaki since they do not require associativity, nor commutativity, nor unitarity in the set of grades. After proving the existence and uniqueness of the primary decomposition of moduloids, they briefly turn their attention to Krull's Theorem and the existence of the primary decomposition of Krasner-Vuković paragraded rings.

In paper [12], Prof. M. Vuković and E. Ilić Georgijević prove the paragraded version of the *Wedderburn-Artin Theorem*. Following the methods known from the abstract case, they first prove the Density Theorem and observe the matrix rings whose entries are from a paragraded ring. However, in order to arrive to the desired structure theorem, they introduce the notion of a Jacobson radical of a paragraded ring and prove some properties that are analogs of the abstract case. In the process, they study the faithful and irreducible paragraded modules over noncommutative paragraded rings and prove the paragraded version of the well-known Schur's Lemma.

In paper [13], Prof. Mirjana started with a short historical development of graduation which begins with Krasner's famous notion of a corpoid, introduced in the 1940s and general graded groups in Krasner's sense, which are more general than Bourbaki's. Also, she presented some results from the theory of Krasner-Vuković's para- and extra-graded groups including examples of paragraduations which are and which are not graduations, and some proofs of statements that were not given earlier, and finally provided the missing step in the proof of the result.

In paper [14], Prof. Vuković studied paragraded modules over noncommutative paragraded rings and, as a main result, proved the paragraded version of Schur's Lemma.

Her paper [15] is concerned with the theory of paragraded rings, which begins with a series of Krasner and Vuković's notes in Proceedings of the Japan Academy, which first appeared in the late 1980s. Prof. Vuković presented prime and Jacobson radicals, discussed the general Kurosh-Amitsur theory of radicals of paragraded rings, established that the theorem of Anderson, Divinsky, and Sulinski holds for paragraded rings, and characterized paragraded normal radicals. She also proved that all special paragraded radicals of paragraded rings can be described by appropriate classes of their graded modules.

Paper [16] by Prof. Vuković begins with a note about Aleksander V. Mikhalev and a short introduction to some historical facts about graded structures that are old and new by M. Krasner. Later, she gave a panoramic view of more general Krasner's graded groups, introduced Krasner-Vuković's paragraded groups, and concluded with some results in the theory of paragraded groups.

The aim of paper [17] was to introduce two versions of paragraded Brown-McCoy radicals, the Brown-McCoy radical and the large Brown-McCoy radical of paragraded rings, and then, using inspiration from Halberstadt's results on Jacobson radicals of graded rings, to prove that the large Brown-McCoy radical of paragraded rings coincides

with the largest homogeneous ideal contained in the classical Brown-McCoy radical ring.

Although her scientific work is primarily focused on fundamental and theoretical research in the aforementioned fields of analysis and algebra, it is also important to highlight her contribution to the field of applied mathematics. Together with R. C. Hrosik, M. Tuba, and M. Pikula, she published three papers in 2014 on facial recognition using neural networks.

2.2. **Other publications**

Prof. Mirjana is the author of more than 10 books and university textbooks. However, among them, the following stand out specifically —chronologically: *Differential Equations* 1 [23] and *Differential Equations* 2 [24], *Group Theory and Representations with Applications in Physics* [18], *Algebra I – Group Theory* (An Overview of the Theory and Problems) [19] (co-authored with Acad. V. Perić), and her most recent work, *Mathematicians – Academicians* [21].

The books (textbooks) Differential Equations 1 & 2 – Theory and Problems cover a wide range of carefully selected problems with detailed solutions in ordinary differential equations, systems of differential equations, partial differential equations, and equations of mathematical physics.

These books are based on many years of Prof. Mirjana's experience in teaching the courses Differential Equations and Analysis 3 to Mathematics students at the Department of Mathematics, University of Sarajevo. Most of the problems included in these books she set in written examinations in the courses Differential Equations and Analysis 3, during 1972-79 when she was an assistant to Academician Fikret Vajzović. Apart from these, she has included a certain number of challenging problems earlier set by her professors and academicians: Mahmut Bajraktarević in examinations in the course Differential Equations for undergraduate students of mathematics, and Manojlo Maravić in the course Equation of Mathematical Physics for postgraduates students of different technical faculties.

As a result of Prof. Mirjana's extensive experience in teaching the course Introduction to Mathematics at the Department of Physics, where group theory and the theory of representations of finite groups were taught, the book (textbook) Group Theory and Representations with Applications in Physics [18] was created. This book is also directly related to her long-standing scientific work in the field of abstract algebra. It provides an excellent presentation of group theory and its applications, especially in physics. *"It should certainly be emphasized that, in the presentation and selection of material, the author's solid mathematical background was usefully complemented by her remarkable education in the field of physics,"* and that *"this textbook will happily fill a gap in the textbook literature in South Slavic languages, particularly in the area related to the theory of group representations and the theory of continuous groups."* (From the review by Academician Veselin Perić).

Also, as a culmination of the long-standing teaching experience of Academician Mirjana Vuković and Academician Veselin Perić in various algebra courses, at both under-

graduate and postgraduate levels, at the Department of Mathematics of the Faculty of Natural Sciences and Mathematics of the University of Sarajevo, and the Department of Mathematics and Computer Science of the Faculty of Philosophy of the University of East Sarajevo, a very high-quality book was created: *Algebra I – Group Theory (Overview of Theory and Problems)* [19]. As indicated by the title, the book is divided into two parts: a theoretical part and a part with problems, each consisting of five chapters. M. Džamonja, as a reviewer, wrote the following about this book: *"The theoretical part of the book reminds me of the legendary book Algebra 1 by Prof. Perić, which served as the main textbook for generations of mathematicians at the Faculty of Natural Sciences and Mathematics in Sarajevo and beyond."* By the way, the theory was written by M. Vuković.

Thus, this manuscript represents an ideal combination of an excellently executed theoretical section and an incredibly abundant and interesting selection of solved problems, making it a complete, useful, and engaging book that can be used both as a textbook and as a valuable addition to any bibliography.

One special book, of a slightly different nature than the previous ones, is Mathematicians – Academicians [21], in which Prof. Mirjana subtly writes about the academicians - her professors: Mahmut Bajraktarević, Manojlo Maravić, Branislav Martić, Veselin Perić, and Fikret Vajzović, but also about professors Vera Šnajder and Šefkija Raljević, who all made significant contributions to the development of mathematics in Bosnia and Herzegovina. In the introduction to the book, the author, Prof. Mirjana, wrote: "The goal of this book was not only to better familiarize the public with and rescue from oblivion the Bosnian-Herzegovinian mathematicians who became the first members of the Academy of Sciences and Arts of Bosnia and Herzegovina, but also, at the same time, to present in a certain manner the history of Bosnian-Herzegovinian mathematics through their stories." The reviewer of the book, Academician Dejan Milošević, emphasized the following: *"The main, third chapter, forms the backbone of the book. By skillfully combining biographical data, personal memories, and a knowledgeable description of academicians' scientific contribution, the author, academician Mirjana Vuković, has created a work of special significance, one which will be read with pleasure even by those for whom mathematics is not the main preoccupation in life. And it was precisely the extremely broad knowledge of mathematics, which the author of this book chose as her life path, that enabled academician Mirjana Vuković to write expertly about the various areas of mathematics that these academicians dealt with."*

Thus, Prof. Mirjana, alongside providing biographical data, highlighted the scientific activities of each of the academicians she wrote about in this book, thereby contributing to the history of mathematics in this region [20], [21].

Additionally, Prof. Mirjana has participated as the principal researcher or project leader in around twenty scientific research projects, including four international ones: two in the Central European Research Support Scheme, and, during her time as a visiting professor in Grenoble and Maribor, co-leader of the French project "Paragraded Structures and their Applications to Non-archimedian Analysis" at the Joseph Fourier Institute in Grenoble (with Prof. A. Panchishkin) and the project "Connections be-

tween Krasner-Vuković paragraded structures (groups, rings, modules) and Lie superalgebras" in Maribor (with Prof. D. Pagon), which were funded by the Rhône Alpes TEMPRA-PECO region (2001), i.e. from the funds of the JoinEU-SEE - ERASMUS Mundus Project Partnership (2013). This latter project aimed to connect the results of two magnificent algebraic schools: the French school of algebra on one side and the Russian—Moscow school on the other.

Thanks to her new theory and interest in the unknown and still insufficiently researched paragraded structures, M. Vuković has been invited to numerous important conferences and symposia.

Thus, by invitation, she spent a month at the renowned *"Fields Institute"* (ICRA X, Toronto, 2002, Canada) and gave lectures at numerous other prestigious European universities, such as: *"Charles University"* (Prague, 1999), *"Joseph Fourier"* University (Grenoble, 2000 and 2001), *"Johannes Kepler"* University (Linz, 2001), and participated with a paper or poster at all post-war world congresses up to 2012. She also took part in the *"Third Croatian Congress of Mathematicians"* in Split (2004); the Mathematical Institute of the Serbian Academy of Sciences and Arts (*Mathematical Colloquium*, Belgrade, 2009); the *"International Algebra Conference dedicated to the 100th anniversary of the birth of the great mathematician A. G. Kurosh and the 250th anniversary of Moscow State University – Lomonosov,"* as well as at the conferences: *Международная конференция "Современные проблемы математики механики и их приложений"* (Moscow, 2009) and *Международный алгебраический симпозиум посвящен 80-летию Кафедры высшей алгебры Механико-математического факультета МГУ и 70-летию профессора А.В. Михалева* (Moscow, 2010), etc.

3. Awards and Recognitions

Prof. Vuković has received numerous awards and recognitions for her scientific and overall work. Some of these include: the highest Republic Award *"Veselin Masleša"* (1987) for scientific work in the field of mathematics; the *Memorial Plaque of the City of Sarajevo* on the occasion of the 40th anniversary of its liberation (1985); the *Order of Labor with a Silver Wreath* from the presidency of former Yugoslavia (1987); the *Charter on the occasion of the 50th anniversary of the founding of the University of Sarajevo* (1999), etc.

4. Happy Jubilee, Professor Mirjana!

Finally, it is important to highlight the human side of Prof. Mirjana. As her student and collaborator for the past ten years, I can say that Prof. Mirjana is a kind person, an outstanding intellectual, and especially a mathematician with all her being. She has dedicated her whole life to mathematics and often proudly emphasizes how much she loves it. In recent years, we have witnessed her enthusiasm as she strives, as the editor-in-chief of the scientific journal Sarajevo Journal of Mathematics, published by the Department of Natural Sciences and Mathematics of ANUBiH, not only to keep it alive but to elevate it to the highest possible level.

I must also point out that she is always ready to help others whenever she can. I still remember with gratitude how she helped me to obtain literature for an exam during my postgraduate studies, especially when it was very difficult to find appropriate literature at the time. But she also helped many other math students and was always willing to assist anyone in need. On behalf of generations of her students, on behalf of generations of mathematicians, dear Professor Mirjana, thank you so much.

And especially, I congratulate our dear Professor Mirjana on this jubilee, wishing her many more years of health, happiness, and success in all areas, particularly in mathematics.

REFERENCES (a selection)

[1] M. Vuković, *On an O – inverse Theorem*, Radovi Odjeljenja Prirodnih i Matematičkih Nauka ANUBiH, Knj. LXIX/20, 55-61, (1982).

[2] M. Vuković, *A convexity Theorem for Ghk –summability*, Radovi Odjeljenja prirodnih i matematičkih nauka ANUBiH, Knj. LXXXIVI/22, 133-139 (1983).

[3] M. Vuković (with M. Krasner), *Structures paragraduées (groupes, anneaux, modules) I*, Proc. Japan Acad., Ser. A, 62, No. 9, 350-352 (1986).
https://projecteuclid.org/euclid.pja/1195514122

[4] M. Vuković (with M. Krasner), *Structures paragraduées (groupes, anneaux, modules) II*, Proc. Japan Acad., Ser. .A, 62, No. 10, 389-391, (1986).
https://projecteuclid.org/euclid.pja/1195514064

[5] M. Vuković (with M. Krasner), *Structures paragraduées (groupes, anneaux, modules) III*, Proc. Japan Acad., Ser. A, 63, No. 1, 10-12 (1987).
https://projecteuclid.org/euclid.pja/1195514019

[6] M. Vuković (with M. Krasner), *Structures paragraduées (groupes, anneaux, modules) (scientific monograph)*, Queen's Papers in Pure and Applied Mathematics, No.77, viii +163, 1987.
https://www.anubih.ba/images/clanovi/redovni/biografije/20240718_COVER_-_VIAF_MV.pdf

[7] P. Ribenboim, *Il mondo Krasneriano*, Queen's preprint, No. 1983-12, Queen's University, Kingston, ON., Canada, pp. 158.

[8] M. Vuković, *Structures graduées et paragraduées*, Prepublication de l'Institut Fourier, Universit´e de Grenoble I (CNRS), No. 536, pp. 1-40 (2001).
https://www-fourier.univ-grenoble-alpes.fr/sites/default/files/ref_536.pdf

[9] M. Vuković (with E. Ilić Georgijević), *Primary Decomposition of General Graded Structures*, Buletunul Acad. de S¸tiint¸e a Republicii Moldova, Matematica, 1, 77, pp. 87- 96 (2015).

[10] M. Vuković (with E. Ilić Georgijević), *A Note on Radicals of Paragraded Rings*, Sarajevo J. Math., Vol. 12 (25), No. 2, Suppl., pp. 307- 316 (2016).

[11] M. Vuković (with E. Ilić Georgijević), *A Note on General Radicals of Paragraded Rings*, Sarajevo, J. Math. Vol. 12 (25), No. 2, Suppl., 317-324 (2016).

[12] M. Vuković (with E. Ilić Georgijević), *The Wedderburn–Artin Theorem for Paragraded Rings*, J. Math. Sci., 221, No. 3, 391- 400 (2017) (Translat. from Fundam. Prikl. Mat., Moscow, T. 19, No. 6, 125-139 (2014)).

[13] M. Vuković, *From Krasner's Corpoid and Bourbaki's Graduations to Krasner's Graduations and Krasner-Vuković's Paragraduations*, Sarajevo J. Math. Vol.14 (27), No.2, pp.175-190 (2018).

[14] M. Vuković, *On noncommutative paragraded rings*, Sarajevo J. Math. Vol.16, No.1, pp. 5-11 (2020).

[15] M. Vuković, *Radicals of paragraded rings*, J. Math. Sci., 275, No. 4, 379-392 (2023) (Translat. from Fund. Prikl. Mat., Vol. 24, No. 2, pp. 3-22 (2022).

[16] M. Vuković, *Panoramic view of graded structures from Euler and Bourbaki–Krasner to Krasner–Vuković*, Fund. Prik. Mat., Vol. 24, No. 3, 23-37(2023), in Russian (and in J. Math. Sci. 283, 838-848 (2024), in English).

[17] M. Vuković, *Brown-McCoy and large Brown-McCoy radicals of paragraded rings* (accepted for publication).
[18] M. Vuković, *Teorija grupa i reprezentacija s primjenama u fizici,* Sarajevo Publishing & Prirodno-matemat. fakultet, Sarajevo, pp. 384 (2003).
[19] M. Vuković (with V. Perić), *Algebra – Teorija grupa* (Pregled teorije i zadaci), Univerzitet u Istočnom Sarajevu, Trebinje, pp. 6+365 (2021).
[20] M. Vuković, *From the Belgrade School of Mihajlo Petrović Alas to the Sarajevo School of Analysis* (in Serbian), Scientific Meetings Serbian Academy of Science and Arts, on October 2-3, 2018, Book CLXXXII, Presidency Book 12, Mihajlo Petrović Alas, pp. 161-172.(2019).
https://dais.sanu.ac.rs/bitstream/handle/123456789/9392/rad11.pdf?sequence=1&isAllowed=y
[21] M. Vuković, *Mathematicians – academicians*, ANUBiH - Posebna izdanja, Knj. CCVII, Odjeljenje prirod. i mat. nauka, Knj. 28, pp. 150 (2023)
[22] M. Vuković, *Curriculum Vitae in pictures*, ANUBiH, Sarajevo (2024).
https://www.anubih.ba/wp-content/uploads/O-MENI-11-SHB-9.pdf
[23] M. Vuković, *Diferencijalne jednadžbe 1 – Teorija i zadaci*, Univerzietska knjiga, Sarajevo, pp 536 (2000).
[24] M. Vuković, *Diferencijalne jednadžbe 2 – Teorija i zadaci*, Univerzietska knjiga, Sarajevo, pp 258 (2001).
[25] M. Vuković (with S. Pilipović and F. Vajzović) *Distribution semigroups on function spaces with singularities at zero*, Novi Sad Journal of Mathematics, 38, No. 1, 127-135, (2008).
[26] M. Vuković (with A. Bučkovska and S. Pilipović) *Inversion Theorem for Bilinear Hilbert Transform*, Integral Transforms and Special Functions,19, No. 1, 3177-325, (2008).
[27] M. Vuković (with I. Zubac), Abel type theorems for the wavelet transform through the quasiasymptotic bondedness Novi Sad Journal of Mathematics, 45, No. 1, 201-206, (2015).
[28] M. Vuković (with E. Ilić-Georgijević and O. Stevanović), *On an application of Parseval's formula to problems of G^{κ}_{θ}- sumability of eigenfunction expansion of the Laplacian operator*, Sarajevo Journal of Mathematics, 12 (25), No. 2, 267-276, (2016).

Mehmed Nurkanović
University of Tuzla
Department of mathematics
U. Vejzagića 4, 75 000 Tuzla
Bosnia and Herzegovina
e-mail: *mehmed.nurkanovic@untz.ba*

PROCEEDINGS OF THE CONFERENCE ON MARCH 14 – INTERNATIONAL DAY OF MATHEMATICS
Special Editions ANUBiH, Book CCXVI, OPMN Book 30
Sarajevo 2024, **p. 15–20** DOI: 10.5644/PI2024.216.03

PERFECT NONLINEAR FUNCTIONS AND THEIR APPLICATIONS

AMELA MURATOVIĆ-RIBIĆ

ABSTRACT. Perfect nonlinear functions are closely related to cryptoanalysis. They are defined on finite groups and represent a connection between computer science, algebra, number theory and combinatorics. In addition to the importance for formation of secure cryptological tools, they are used both in the theory of coding, and in pure mathematics. The most important related results of this significant subfield of mathematics, so far, are presented.

1. INTRODUCTION

Cryptology, due to the intensive development of information technologies, has a very important role and is intensively developing, adapting to new trends. For this reason,we also witness also the development of mathematics that has applications in cryptology, and these two lines of research are closely related. This is the motivation for the article about perfectly nonlinear functions that play a significant role in the development of symmetric cryptosystems.

The derivative over real and complex functions is very significant and represents the best affine approximation of functions in the neighborhood of a given point. On the other hand, when it is defined over finite groups, its meaning is somewhat different and has an application in combinatorics, in designs and other combinatorial structures such as differential sets. Cryptology, as a science, consists of encryption, decryption and cryptanalysis, i.e. attacks on encryption systems. In the late 1980s, Sean Murphy studied the FEAL encryption algorithm considering equations of the form $G(x+a) - G(x) = b$, and at the same time Eli Biham and Adi Shamir concluded that in DES equal differences in the plaintext produce equal differences in the chipertext more often than usual, which initiated development of differential cryptanalysis. APN over the field of characteristic 2 was found in 2009 which led to a new direction of development in this field of mathematics, see [1].

Definition 1.1. *Let A and B be finite Abelian groups i $F : A \to B$ a function. For a given $a \in A$, a function defined by*

$$D_a F : A \to B, \qquad x \mapsto F(x+a) - F(x)$$

2020 *Mathematics Subject Classification.* 68P25.
Key words and phrases. perfect nonlinear functions, bent functions, S-box.

is called the ***derivative of*** F.

For given $a \in A$ and $b \in B$ the relation

$$F(x+a) - F(x) = b. \tag{1.1}$$

it is called ***the derivative with input difference*** a ***and output difference*** b.

In computer science, it is common to write data using strings, i.e. zero and one strings of length n. Transformations are performed on these strings, especially during encryption and coding, and for this reason it is necessary to use the mathematical tools of first linear algebra and then finite fields for easier manipulation of strings.

Let q be a positive integer and let $\mathbb{Z}_q$ be a ring. Since we are looking at strings, we denote the function from $\mathbb{Z}_q^n$ to $\mathbb{Z}_q$ with the lowercase letter f, that is, $f : \mathbb{Z}_q^n \to \mathbb{Z}_q$, and the functions with the domain of several variables with the uppercase letter $F : \mathbb{Z}_q^n \to \mathbb{Z}_q^m$. The introduction to differential cryptanalysis led to the study of differentials for nonlinear functions and to the study of the number of solutions of equation (1.1).

Definition 1.2. *Let $F : A \to B$ be a function. Denote by*

$$\delta(a,b) = |\{x | F(x+a) - F(x) = b\}|.$$

Let $\Delta_F = \max_{a \in A, a \neq 0} \delta(a,b)$. We say that F is Δ_F uniform.

It is easy to see that

$$|A| = \sum \delta(a,b) \leq \sum \Delta_F = \Delta_F |B| \quad \text{and thus} \quad \Delta_F \geq |B|/|A|.$$

Definition 1.3. *We call the function F perfectly nonlinear (PN) if $\Delta_F = |B|/|A|$.*

A function f is said to be Boolean if $f : \mathbb{Z}_2^n \to \mathbb{Z}_2$. Perfect nonlinear functions in this case of Boolean functions were studied by Willi Meier and Othmar Staffelbach and in the case $\Delta_F = 2^{n-1}$.

Theorem 1.1. *Let the function $F : \mathbb{Z}_2^n \to \mathbb{Z}_2^m$ be perfect nonlinear. Then, for every y from the domain of F $|F^{-1}(y)| = |\{x \in \mathbb{Z}_2^n | F(x) = y\}| = b_y 2^{\frac{n}{2}-m}$, holds where b_y is an odd integer.*

As $|F^{-1}(y)|$ is a positive integer for at least one $y \in \mathbb{Z}_2^m$ it follows that PN functions from $\mathbb{Z}_2^n$ in $\mathbb{Z}_2^m$ exist only if $\frac{n}{2} - m \geq 0$, i.e. $n \geq 2m$. Although the research of these functions is related to cryptography, so $q = 2$ is mostly taken, the terms are extended to other values of q.

In the case of Abelian groups, where $|A| = |B|$, PN functions are called planar functions. As the derivative is then a bijective function, for each $a \in A, \quad a \neq 0$, each equation $F(x+a) - F(x) = b$ has exactly one solution, so especially for fixed a there is exactly one x such that $F(x+a) - F(x) = 0$. Hence $F(x+a) = F(x)$ where $x + a \neq x$. Therefore, the planar function is not bijective. However, planar functions can be used for purposes other than cryptology. Let a finite additive Abelian group $A = \{a_1, a_2, \ldots, a_n\}$ be given. The Latin square over the elements of the group can be defined by $L_{ij} = (a_i + f(a_j)), 1 \leq i, j \leq n$ where $f : A \to A$ is a bijection. A Latin square is a matrix in which the elements in each row and each column are different from each other. Kelly's

table of groups gives some of the Latin squares, but not all. If $A = \mathbb{Z}_3, \quad f(x) = 2x$ we have

$$L = \begin{pmatrix} 0 & 2 & 1 \\ 1 & 0 & 2 \\ 2 & 1 & 0 \end{pmatrix}.$$

Mutually orthogonal Latin squares L and H are Latin squares of equal dimensions and such that for $1 \leq i, j, s, l \leq n$,$(L_{ij}, H_{ij}) \neq (L_{sl}, H_{sl})$, for $(i, j) \neq (s, l)$. They are important for making schedules, designing experiments, etc. The maximal family of mutually orthogonal Latin squares is obtained for $n = p^s$ where p is a prime number and the largest number of mutually orthogonal pairs is equal to $n - 1$. If we have a planar mapping F on A, then the Latin squares defined by $L_{ij}^a = (a_i + F(a_j + a) - F(a_j))$ for all $a \in A, \quad a \neq 0$ form the maximal family of mutually orthogonal Latin squares.

Planar functions do not exist for $q = 2$, because $F(x + a) - F(x) = F((x + a) + a) - F(x + a)$. In the case of vector Boolean functions, equation (1.1), if it is consistent, has at least two solutions $(x, x + a)$. Therefore, the minimum value for differential uniformity is 2. If this minimum value is reached, the function is said to be almost perfectly nonlinear (APN).

Definition 1.4. *A function $F : A \to B$ is called APN or almost perfectly nonlinear if it is differentially $(2|B|)/(|A|)$ uniform.*

2. BENT AND ALMOST BENT FUNCTIONS

Let $q > 1$ be an integer and denote by $\omega \in \mathbb{C}$ the q-th root of unity, i.e. $\omega = e^{(2\pi i/q)}$.

Definition 2.1. *The Walsh transform of the q-ary function $f : \mathbb{Z}_q^n \to \mathbb{Z}_q$ computed in $\mathbb{C}$ is defined by*

$$\hat{f}(a) = \sum_{x \in \mathbb{Z}_q^n} \omega^{f(x) - \langle a, x \rangle}$$

where $a \in \mathbb{Z}_q^n$ and for $a = (a_1, a_2, \ldots, a_n)$, $x = (x_1, x_2, \ldots, x_n) \in \mathbb{Z}_q^n$

$$\langle a, x \rangle = a_1 x_1 + a_2 x_2 + \cdots + a_n x_n.$$

If $q = 2$, then $\omega = -1$. If $q = p$ is a prime number, then $\mathbb{Z}_p^n$ is a vector space over the field $\mathbb{Z}_p$ and $\langle a, x \rangle$ is a scalar product. Now let $q = p^n$ and $\mathbb{F}_q$ be a finite field. Then $\mathbb{F}_q$ is a vector space over $\mathbb{Z}_p$ with base $(\beta_1, \beta_2 \ldots, \beta_n)$. All elements from $\mathbb{F}_q$ can be uniquely represented by $x = a_1\beta_1 + a_2\beta_2 + \cdots + a_n\beta_n$ where the coefficients are $a_1, a_2, \ldots, a_n \in \mathbb{Z}_p$.

Then, the correspondence

$$(a_1, a_2 \ldots, a_n) \to a_1\beta_1 + a_2\beta_2 + \cdots + a_n\beta_n$$

is an isomorphism of the vector spaces $\mathbb{Z}_p^n$ and $\mathbb{F}_q$ over the field $\mathbb{Z}_p$. The scalar product is a linear function and can be defined in $\mathbb{F}_q$ via the absolute trace function with $\langle a, x \rangle = Tr(ax) = \sum_{i=0}^{n-1} (ax)^{p^i}$.

The values of the Walsh transformation for a fixed a, denoted by $\hat{f}(a)$, are called the Walsh coefficients of f. They are used to measure the distance from f to the function

$x \mapsto \langle a, x\rangle$. For a prime number p these are the only linear functions in the vector space $\mathbb{Z}_p^n$.

Definition 2.2. *Linearity $\mathcal{L}(f)$ of function f is defined by*

$$\mathcal{L}(f) = max_{a\in\mathbb{Z}_q^n}|\hat{f}(a)|.$$

For vector functions $F : \mathbb{Z}_q^n \to \mathbb{Z}_q^m$ linearity is defined by the linearity of non-trivial linear combinations of its coordinate functions. For a given $\lambda \in \mathbb{Z}_q^m$ the q-ary function $f_\lambda(x) = \langle \lambda, F(x)\rangle$ is called the λ-component of F.

The Walsh spectrum of F is the set of all values $\hat{f_\lambda}(a)$ for all $a \in \mathbb{Z}_q^n$ and all $\lambda \in \mathbb{Z}_q^m \setminus \{0\}$.

The linearity of the function F is defined by

$$\mathcal{L}(F) = max_{\lambda\in\mathbb{Z}_q^m\setminus\{0\}}\mathcal{L}(f_\lambda).$$

Using Parseval's theorem, we obtain that $q^{\frac{n}{2}} \leq \mathcal{L}(F)$, and for the functions $x \mapsto \langle a, \cdot\rangle$, $\mathcal{L}(F) = q^n$ holds, so we have $q^{n/2} \leq \mathcal{L}(F) \leq q^n$ for all functions F.

Definition 2.3. *We call function F for $q = 2$ bent, (and for $q > 2$ generalized bent functions) if $q^{n/2} = \mathcal{L}(F)$ holds.*

If $m = n$ bent functions do not exist. When n is odd, the lower limit for $\mathcal{L}(F)$ is $2^{(n+1)/2}$.

Definition 2.4. *For $q = 2$, the vector Boolean function for which $\mathcal{L}(F) = 2^{(n+1)/2}$ holds is called is almost-bent (AB).*

For even values of n, functions are known for which $\mathcal{L}(F) = 2^{\frac{n}{2}+1}$ is valid, but no function with minimal linearity has yet been found. For bijective functions $F : \mathbb{Z}_2^n \to \mathbb{Z}_2^n$ the smallest known linearity is $2^{(\lfloor n/2\rfloor+1)}$. Chabaud and Vaudenay proved a result that gives a connection between perfectly nonlinear functions and bent functions. A Boolean function is a bent function if and only if it is perfectly nonlinear. For the q-ary case, the perfectly nonlinear function is bent, and the reverse holds if q is a prime number. The same is true for near-PN and near-bent functions. Let F be a vector Boolean function. If F is almost-bent, then it is almost perfectly nonlinear. If F is almost perfectly nonlinear and the Walsh coefficients of $\langle 1, F(x)\rangle$ are divisible by 2^{m+1}, then F is almost-bent. Almost perfectly non-linear vector Boolean functions are also used in coding theory, to form linear codes with a specific weight. More details can be found in [1].

3. Equivalence of bent-functions

Due to the bijective affine transformations $G(x) = L_1(F(L_2(x))) + L_3(x)$ where $L_1(x)$ and $L_2(x)$ are bijections, the linearity of the Boolean vector functions F remains preserved and we call such functions EA-equivalent. So, we study the classes of equivalent bent functions. The concept of CCZ-equivalent functions is used in coding theory. Namely, the functions F and G are CCZ-equivalent if the corresponding codes C_F and

C_G are equivalent, where the code parity check matrix is defined by

$$H_F = \begin{pmatrix} \cdots & 1 & \cdots \\ \cdots & x & \cdots \\ \cdots & F(x) & \cdots \end{pmatrix}.$$

CCZ equivalence is very hard to establish, but it coincides with EA-equivalence for planar, Boolean functions, vector bent functions if $q = p = 2$ and vector bent functions if $p = q$ is an odd prime and $m = n$.

4. KNOWN CONSTRUCTIONS OF PN MAPPING

In this section we denote the monomial as the function $F_d : \mathbb{F}_{p^n} \to \mathbb{F}_{p^n}$, $x \mapsto x^d$. Monomials are bijective if and only if the exponent d is relatively prime with $p^n - 1$. The mappings $w \to w^{p^n}$ are linear isomorphisms such that all monomials $F_e(x) = x^e$, when e is in the same cyclotomic class $\{p^i \cdot d | 0 \le i < n\}$ have the same differential uniformity and nonlinearity as well as the monomial $F_d(x) = x^d$ and therefore the monomial with the smallest exponent in the class is usually studied. Further, for $a \neq 0, \delta(a,b) = |\{x|F(x+a) - F(x) = b\}| = |\{x|(x=a)^d - x^d = b\}| = |\{x|(x=1)^d - x^d = \frac{b}{a^d}\}| = \delta(1, \frac{b}{a^d})$.

If $b = 0$, then $\delta(1,0) = \gcd(d, p^n - 1) - 1$, and in particular the monomial is bijective if and only if $\delta(1,0) = 0$.

There are plenty of monomials that are APN in fields of characteristic 2, and for odd characteristic the known PN monomials are x^2, x^{p^t+1} where $\frac{n}{\gcd(n,t)}$ is odd, $x^{(p^t=1)/2}$ where $p = 3$, t is odd and $\gcd(n,t) = 1$. Bent functions for infinitely many fields and are called exceptional. Many polynomials with these properties have also been found. An interesting example is the mapping $G : x \mapsto c^x$ because for each $a \neq 0$ we have the difference $G(x+a) - G(x) = c^{x+a} - c^x = (c^a - 1)G(x)$.

When observing functions from $\mathbb{Z}_2^m$ in $\mathbb{Z}_2$ strings can be divided into several parts whose sum length is m. So, for example, functions $F(X,Y) = g(X) + h(Y)$are bent on $\mathbb{Z}_2^{2m}$, if g and h are bent on $\mathbb{Z}_2^m$. The same is true for symmetric functions, but such decompositions are not interesting in practice.

One of the more significant constructions is Maiorana McFarland: Let g and π be permutations on $\mathbb{Z}_2^m$, then$f(X,Y) = \pi(X)Y + g(X)$ is bent on $\mathbb{Z}_2^{2m}$.

The other significant construction is using partial spreads. Let $\mathbb{Z}_2^{2m}$ be a vector space.

Let's define the family $\mathcal{PC}^-$: Let $H_1, H_2, ..., H_{2m-1}$ be vector spaces such that $H_i \cap H_j = 0, \quad i \neq j$. Let $H^* = \cup_{i=1}^{2m-1}(H_i \setminus \{0\})$. Then the characteristic function of H^* is bent on $\mathbb{Z}_2^{2m}$.

Now define the family $\mathcal{PC}^+$: The union of any $2m+1$ subspaces with $H_i \cap H_j = 0, \ i \neq j$ is called a cover and its characteristic function is bent on $\mathbb{Z}_2^{2m}$.

These construction can be found in [2]. There are many known bent functions and classes of planar functions, but this area continues to attract the attention of mathematicians and is intensively developed.

5. Applications

The most significant application of APN functions is in the construction of S-boxes in symmetric cryptosystems. Encryption is a mapping $S_k : \mathbb{F}_q \to \mathbb{F}_q$ where k is the key, i.e. field element on which the encryption function depends and which is secret. In general, all parts of symmetric cryptosystems are linear except for S-boxes. The good construction of S-boxes guarantees the security of the cryptosystem and the avalanche effect which means that small changes in the input data cause large changes in the ciphertext. In his hardware-oriented MISTY design, Matsui split the 16-bit state into two odd parts of different lengths so that he could use APN permutations. The designers Daemen and Rijmen of the AES algorithm, which was aimed at software implementation, could not go that far, so they settled on a suboptimal choice which is a differentially 4-uniform inverse function. It is considered that large S-boxes provide greater security against cryptographic attacks, and this area is intensively studied in the wider mathematical community, where the APN functions are of greatest importance.

References

[1] Celine Blondean, Kaisa Nyberg, *Perfect nonlinear functions and cryptography*, Finite Fields and Their Applications 32, 120-147, 2015.

[2] John F. Dillon, E.M. Wright *Elementary Hadamard difference sets*, PhD thesis, University of Maryland, 1974.

(Received: May 15, 2024)
(Revised: 11 September, 2024)

Amela Muratović-Ribić
University of Sarajevo
Faculty of science and mathematics
Department of mathematics and computer sciences
Zmaja od Bosne 33-35
71 000 Sarajevo
Bosnia and Herzegovina
e-mail: *amela@pmf.unsa.ba*

PROCEEDINGS OF THE CONFERENCE ON MARCH 14 – INTERNATIONAL DAY OF MATHEMATICS
Special Editions ANUBiH, Book CCXVI, OPMN Book 30
Sarajevo 2024, **p. 21–32** DOI: 10.5644/PI2024.216.04

IMPROVING THE SDA ALGORITHMS FOR SOLVING THE T-PALINDROMIC QEPS

ALEKSANDRA KOSTIĆ, VALENTINA TIMOTIĆ, AND IZET HORMAN

Dedicated to the 75th birthday of our dear Professor Mirjana Vuković

ABSTRACT. The T-palindromic quadratic eigenvalue problem (QEP) $(\lambda^2 B + \lambda C + A)x = 0$, with $A, B, C \in \mathbb{C}^{n\times n}$, $C^T = C$ and $B^T = A$, belongs to the structured quadratic eigenvalue problems. These problems occur in solving fast train vibration problems. Vibration is produced from the interaction between the wheels of trains and the rails underneath. To solve these problems finite element packages can not be used, because of poor accuracy. Standard methods for solving the T-palindromic QEPs are SDAs (the structure-preserving doubling algorithms) methods. The structure of the T-palindromic QEPs allows the improvement of SDAs methods.

1. INTRODUCTION

The T-palindromic quadratic eigenvalue problem (QEP)

$$(\lambda^2 B + \lambda C + A)x = 0,\ x \neq 0, \tag{1.1}$$

with $A, B, C \in \mathbb{C}^{n\times n}$, $C^T = C$ and $B^T = A$, belongs to the structured quadratic eigenvalue problems. For simpler notation, equation (1.1) can be written in the following form

$$(\lambda^2 A_1 + \lambda A_0 + A_1^T)x = 0,\ x \neq 0, \tag{1.2}$$

where $A_0^T = A_0$. Let us define

$$P(\lambda) := \lambda^2 A_1 + \lambda A_0 + A_1^T. \tag{1.3}$$

These problems occur in solving fast train vibration problems. Vibration is produced from the interaction between the wheels of trains and the rails underneath. Matrices A_0 and A_1 are dependent on some parameter ω associated with the speed of the train. The eigenvalues λ are related to the vibration frequencies and the corresponding eigenvectors x reflect the shape of the vibration [5, 11]. Palindromic eigenvalue problems are also used in many other applications such as surface acoustic wave filters, which have wide application in the telecommunication industry. Additional applications of the palindromic eigenvalue problems are given in the papers [1, 13]. The special structure of the palindromic eigenvalue problem (1.2) gives specific spectrum properties that we

2020 *Mathematics Subject Classification.* 15A18, 65F15.
Key words and phrases. The T-palindromic QEPs, fast train vibration problems, the SDA methods.

will look back at. Transposing (1.2) implies an important reciprocity property of the spectrum of the palindromic eigenvalue problem,

$$\lambda \in \sigma(P(\lambda)) \Rightarrow \frac{1}{\lambda} \in \sigma(P(\lambda)), \tag{1.4}$$

with $\sigma(\cdot)$ denoting the spectrum, and the convention that 0 and ∞ are considered to be mutually reciprocal. The polynomials

$$\lambda^2 A_1 + \lambda A_0 + A_1^T$$

and

$$\nu^2 A_1 - \nu A_0 + A_1^T$$

define the same palindromic eigenvalue problem ($\nu = -\lambda$).

Spectral symmetries for various types of palindromic eigenvalue problems are given in the paper [13].

An excellent overview of the methods for solving the eigenvalue problems is given in the papers [5, 6].

Finite element packages can not be used to solve these problems, due to their poor accuracy. There are two well-known tools in literature for solving QEPs, linearization and methods based on variational characterization. Applying linearization, we obtain a generalized eigenvalue problem to which the QZ algorithm can be applied. The disadvantage of linearization is that the QZ algorithm does not preserve the palindromic structure. To avoid this problem, in papers [2, 8, 13–16] a palindromic linearization of the form

$$\lambda Z + Z^T \tag{1.5}$$

with

$$Z = \begin{bmatrix} A_1^T & A_0 - A_1 \\ A_1^T & A_1^T \end{bmatrix} \tag{1.6}$$

was presented.

Standard methods for solving the T-palindromic QEPs are SDA (the structure preserving doubling algorithms) methods [5]. The structure of the T-palindromic QEPs allows improvement of SDA methods. In this paper, we will deal with the improvement of this algorithm and its numerical stabilization.

The paper is organized in the following way: In Section 2 the basic idea of SDA algorithms (SDA1 and SDA2) and deflation will be presented. In Section 3 new results related to the stability of the algorithm will be considered. In Section 4 we applied some properties of the T- quadratic eigenvalue problem in order to stabilize the algorithm. The conclusion is given in Section 5.

2. Structure preserving algorithms (SDA)

For greater transparency of the paper we will separate this section into two subsections: Deflation and SDA algorithms.

2.1. **Deflation**

We have seen that eigenvalues 0 and ∞ come in pairs in palindromic eigenvalue problems. Numerical experiments in [5] show that there are below 1.5 % finite nonzero eigenvalues whose absolute value belongs to the segment $[10^{-14}, 10^{14}]$. Therefore deflation has great significance and let us look at it first.

The idea of deflation is to find infinite eigenvalues or those equal to 0, that are a consequence of the singularity of the matrix A_1, and to suppress them (deflation) before applying methods for calculating eigenvalues.

From the mathematical and physical model it is obtained that matrices A_1 and A_0 have the following form:

$$A_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ L & 0 & 0 \end{bmatrix} \in \mathbb{C}^{n\times n},\ A_0 = \begin{bmatrix} C_{11} & C_{12} & 0 \\ C_{12}^T & C_{22} & C_{23} \\ 0 & C_{23}^T & C_{33} \end{bmatrix} \in \mathbb{C}^{n\times n},$$

where

$$L \in \mathbb{C}^{n_m\times n_1},\ C_{11} = C_{11}^T \in \mathbb{C}^{n_1\times n_1},\ C_{33} = C_{33}^T \in \mathbb{C}^{n_m\times n_m}, C_{22} = C_{22}^T \in \mathbb{C}^{l\times l}$$

and

$$l = n - n_1 - n_m.$$

Assume that C_{22} is nonsingular. Let

$$\Theta = \begin{bmatrix} I_{n_1} & -C_{12}C_{22}^{-1} & 0 \\ 0 & I_l & 0 \\ 0 & -C_{23}^T C_{22}^{-1} & I_{n_m} \end{bmatrix},\ \Pi = \begin{bmatrix} I_{n_1} & 0 & 0 \\ 0 & 0 & I_{n_m} \\ 0 & I_l & 0 \end{bmatrix}.$$

Using a similarity transformation, $P(\lambda)$ can be transferred to the following form

$$\begin{aligned} \Pi\Theta P(\lambda)\Theta^T\Pi^T &= \begin{bmatrix} \lambda(C_{11} - C_{12}C_{22}^{-1}C_{12}^T) & L^T - \lambda C_{12}C_{22}^{-1}C_{23} & 0 \\ \lambda(\lambda L - C_{23}^T C_{22}^{-1}C_{12}^T) & \lambda(C_{33} - C_{23}^T C_{22}^{-1}C_{23}) & 0 \\ 0 & 0 & \lambda C_{22} \end{bmatrix} \\ &= \operatorname{diag}(I_{n_1}, \lambda I_{n_m}, I_l) \begin{bmatrix} S(\lambda) & 0 \\ 0 & \lambda C_{22} \end{bmatrix}, \end{aligned}$$

where

$$S(\lambda) = \begin{bmatrix} \lambda\tilde{C}_{11} & L^T - \lambda\tilde{C}_{12} \\ \lambda L - \tilde{C}_{12}^T & \tilde{C}_{22} \end{bmatrix}$$

and

$$\begin{aligned} \tilde{C}_{11} &\equiv C_{11} - C_{12}C_{22}^{-1}C_{12}^T, \\ \tilde{C}_{12} &\equiv C_{12}C_{22}^{-1}C_{23}, \\ \tilde{C}_{22} &\equiv C_{33} - C_{23}^T C_{22}^{-1}C_{23}. \end{aligned}$$

Lemma 2.1. *Let* $[x^T, y^T]^T$ *be an eigenvector of* $S(\lambda)$*. Then*

$$\Theta_1^T \Pi_2^T \begin{bmatrix} x \\ y \\ 0 \end{bmatrix} = \begin{bmatrix} x \\ -C_{22}^{-1}(C_{12}^T x + C_{23} y) \\ y \end{bmatrix}$$

is an eigenvector of $P(\lambda)$*. Furthermore,* $\sigma(P(\lambda)) = \sigma(S(\lambda)) \cup \{0, \infty\}$.

2.2. **SDA algorithms**

It is important to preserve the palindromic structure at all times. Therefore we use the well-known SDA1 and SDA2 algorithms. Let us look at the SDA1 algorithm. The equation (1.2) can be written in a factored form. By replacing the row-blocks, we obtain that the pencil $S(\lambda)$ is equivalent to

$$\lambda \begin{bmatrix} L & 0 \\ \tilde{C}_{11} & -\tilde{C}_{12} \end{bmatrix} + \begin{bmatrix} -\tilde{C}_{12}^T & \tilde{C}_{22} \\ 0 & L^T \end{bmatrix}, \tag{2.1}$$

which is a generalized standard symplectic form. The structure-preserving doubling algorithm SDA1 can then be applied to solve the corresponding eigenvalue problem.

Let us introduce now the SDA2 algorithm. The T-palindromic pencil for nonsingular X can be written in the following form:

$$P(\lambda) = (\lambda A_1 - X)X^{-1}(\lambda X - A_1^T) + A_1 X^{-1} A_1^T + X + A_0. \tag{2.2}$$

The main idea is to write the T-palindromic pencil (2.2) in the factored form:

$$P(\lambda) = (\lambda A_1 - X)X^{-1}(\lambda X - A_1^T) \tag{2.3}$$

for some nonsingular X if and only if X satisfies the following nonlinear matrix equation with the plus sign:

$$A_1 X^{-1} A_1^T + X + A_0 = 0. \tag{2.4}$$

We apply the SDA algorithm on the equation (2.4), which preserves the structure of the problem [4, 7, 9, 10, 12].

Assume that the matrix $\tilde{C}_{22}$ is invertible. Define $\tilde{S}(\lambda)$ as follows:

$$\begin{aligned} \tilde{S}(\lambda) &\equiv \begin{bmatrix} I_{n_1} & -L^T \tilde{C}_{22}^{-1} \\ 0 & I_{n_3} \end{bmatrix} S(\lambda) \begin{bmatrix} I_{n_1} & 0 \\ \tilde{C}_{22}^{-1} \tilde{C}_{12}^T & I_{n_3} \end{bmatrix} \\ &= \begin{bmatrix} \lambda(\tilde{C}_{11} - L^T \tilde{C}_{22}^{-1} L - \tilde{C}_{12} \tilde{C}_{22}^{-1} \tilde{C}_{12}^T) + L^T \tilde{C}_{22}^{-1} \tilde{C}_{12}^T & -\lambda \tilde{C}_{12} \\ \lambda L & \tilde{C}_{22} \end{bmatrix} \end{aligned} \tag{2.5}$$

and let $[\tilde{x}^T, \tilde{y}^T]^T$ be an eigenvector of $\tilde{S}(\lambda)$, i.e.

$$\lambda[(\tilde{C}_{11} - L^T\tilde{C}_{22}^{-1}L - \tilde{C}_{12}\tilde{C}_{22}^{-1}\tilde{C}_{12}^T)\tilde{x} - \tilde{C}_{12}\tilde{y}] + L^T\tilde{C}_{22}^{-1}\tilde{C}_{12}^T\tilde{x} = 0, \tag{2.6}$$

$$\lambda L\tilde{x} + \tilde{C}_{22}\tilde{y} = 0. \tag{2.7}$$

Since the matrix $\tilde{C}_{22}$ is invertible, from (2.7) $\tilde{y}$ can be represented as

$$\tilde{y} = -\lambda\tilde{C}_{22}^{-1}L\tilde{x}. \tag{2.8}$$

Let us denote

$$A_{d_1} = \tilde{C}_{12}\tilde{C}_{22}^{-1}L,$$
$$A_{d_0} = \tilde{C}_{11} - L^T\tilde{C}_{22}^{-1}L - \tilde{C}_{12}\tilde{C}_{22}^{-1}\tilde{C}_{12}^T.$$

Suppose that X is nonsingular. Rewrite $P_d(\lambda)$ as

$$P_d(\lambda) = (\lambda A_{d_1} - X)X^{-1}(\lambda X - A_{d_1}^T) + \lambda(A_{d_1}X^{-1}A_{d_1}^T + X + A_{d_0}).$$

Let us apply (2.3) and (2.4) on $P_d(\lambda)$. It follows that $P_d(\lambda)$ can be factorized (or square-rooted) as

$$P_d(\lambda) = (\lambda A_{d_1} - X)X^{-1}(\lambda X - A_{d_1}^T),$$

for some nonsingular X if and only X satisfies the following nonlinear matrix equation with the plus sign:

$$A_{d_1}X^{-1}A_{d_1}^T + X + A_{d_0} = 0.$$

Algorithm 2.1. *(SDA for Palindromic QEP)*
Input: $C_{11}, C_{22}, C_{33}, C_{12}, C_{23}, L; \tau$ *(a small tolerance)*;
Output: an eigenpar $(\lambda, [x^T, z^T, y^T]^T)$*of Palindromic QEP.*
Compute
$\tilde{C}_{11} = C_{11} - C_{12}C_{22}^{-1}C_{12}^T$,
$\tilde{C}_{12} = C_{12}C_{22}^{-1}C_{23}$,
$\tilde{C}_{22} = C_{33} - C_{23}^TC_{22}^{-1}C_{23}$,
$A_{d_1} = \tilde{C}_{12}\tilde{C}_{22}^{-1}L$,
$A_{d_0} = \tilde{C}_{11} - L^T\tilde{C}_{22}^{-1}L - \tilde{C}_{12}\tilde{C}_{22}^{-1}\tilde{C}_{12}^T$;
Set $k = 0, R_k = A_{d_1}^T, Q_k = -A_{d_0}$ *and* $P_k = 0$;
Do until convergence:
Compute $R_{k+1} = R_k(Q_k - P_k)^{-1}R_k$,
$Q_{k+1} = Q_k - R_k^T(Q_k - P_k)^{-1}R_k$,
$P_{k+1} = P_k + R_k(Q_k - P_k)^{-1}R_k^T$, $k = k + 1$;
If $\|Q_k - Q_{k-1}\| \leqslant \tau\|Q_k\|$, *Stop* ;
End;
Compute the left/right eigenpairs $(\lambda_u, \tilde{x}_s)$, $(\lambda_u, \hat{x}_r)$ *of* $Q_k\hat{x} = \lambda A_{d_1}\hat{x}$;
Solve $(\lambda_u Q_k - A_{d_1}^T)\tilde{x}_u = Q_k\hat{x}_r$;
Set $\lambda_s = \lambda_u^{-1}$;
Solve $\tilde{C}_{22}\tilde{y} = -\lambda L\tilde{x}$ *with* $(\lambda, \tilde{x}) = (\lambda_s, \tilde{x}_s)$ *or* $(\lambda, \tilde{x}) = (\lambda_u, \tilde{x}_u)$;
Set $x = \tilde{x}$; *Compute* $y = \tilde{C}_{22}^{-1}\tilde{C}_{12}^T\tilde{x} + \tilde{y}$, $z = -C_{22}^{-1}(C_{12}^Tx + C_{23}y)$;

3. New ideas for improving the stability of the algorithm

In addition to preserving the structure of the eigenvalue problem during the implementation of the algorithm, it is very important to pay attention to the stability of the algorithm. The analysis of Algorithm 2.1 clearly shows the points at which the problem may occur:

- invertibility of the matrix $Q_k - P_k$ in the Algorithm 2.1;
- efficient preprocessing of problems when C_{22} is ill-conditioned;
- efficient preprocessing of problems when L is ill-conditioned.

3.1. **Invertibility of significant matrices**

After deflation, the palindromic eigenvalue problem

$$(\lambda^2 A_{d_1} + \lambda A_{d_0} + A_{d_1}^T)v = 0,\ v \neq 0 \tag{3.1}$$

is considered, which is a lower dimension problem.

It is clear that the singularity of the matrix A_{d_0} can cause a problem in Algorithm 2.1, because for $k = 0,\ Q_0 = -A_{d_0},\ P_0 = 0$, and the invertibility of the matrix $Q_0 - P_0$ in the first step of Algorithm 2.1 is required.

Let

$$A_{d_0} = Q A'_{d_0} Q^T$$

be the Schur decomposition of a matrix A_{d_0}. Then A'_{d_0} is a diagonal matrix. Elements of the matrix A'_{d_0} are eigenvalues of the matrix A_{d_0}. Matrices Q and Q^T are orthogonal. Without loss of generality let us assume that the first p elements on the diagonal of the matrix A'_{d_0} are zeros.

The equation (3.1) can be written in the following form

$$(\lambda^2 A_{d_1} + \lambda Q A'_{d_0} Q^T + A_{d_1}^T)v = 0. \tag{3.2}$$

After multiplying (3.2) with Q^T on the left we obtain an equivalent eigenvalue problem

$$(\lambda^2 Q^T A_{d_1} Q + \lambda A'_{d_0} + Q^T A_{d_1}^T Q) Q^T v = 0, \tag{3.3}$$

which is the T-palindromic eigenvalue problem, with eigenvector $w = Q^T v = \begin{bmatrix} a \\ b \end{bmatrix}$.

For simpler notation, the problem can be presented in the following form

$$(\lambda^2 B_{d_1} + \lambda B_{d_0} + B_{d_1}^T)w = 0, \tag{3.4}$$

respectively in the block matrix form

$$\left(\lambda^2 \begin{bmatrix} B_{d_{11}} & B_{d_{12}} \\ B_{d_{21}} & B_{d_{22}} \end{bmatrix} + \lambda \begin{bmatrix} 0 & 0 \\ 0 & B_{d_{02}} \end{bmatrix} + \begin{bmatrix} B_{d_{11}}^T & B_{d_{21}}^T \\ B_{d_{12}}^T & B_{d_{22}}^T \end{bmatrix} \right) \begin{bmatrix} a \\ b \end{bmatrix} = 0. \tag{3.5}$$

It is clear that we obtain the quadratic T-palindromic eigenvalue problem

$$(\lambda^2 B_{d_{22}} + \lambda B_{d_{02}} + B_{d_{22}}^T)b = 0, b \neq 0.$$

Suppose that $a = 0$ and that the eigenvector b satisfies

$$(\lambda^2 B_{d_{12}} + B_{d_{21}}^T)b = 0.$$

In this case we obtain the T- palindromic eigenvalue problem of a lower dimension then the one we were dealing with, and w is an eigenvector of the eigenvalue problem (3.5).

It is also clear in the case that a is an eigenvector of the T-palindromic linear eigenvalue problem

$$B_{d_{11}}^T a = -pB_{d_{11}}a,\ p = \lambda^2,$$

and $b = 0$ and the eigenvector a satisfies the additional condition

$$(\lambda^2 B_{d_{21}} + B_{d_{12}}^T)a = 0.$$

Then $w = \begin{bmatrix} a \\ 0 \end{bmatrix}$ is an eigenvector of the T-palindromic quadratic eigenvalue problem (3.4).

If the Schur decomposition of the matrix A does not lead to the T-palindromic quadratic eigenvalue problem of lower dimension or to the linear T-palindromic eigenvalue problem of lower dimension, then in Algorithm 2.1 the singular matrix $(-A_{d_0})^{-1}$ which does not exist needs to be replaced with the pseudoinverse matrix $(-A_{d_0})^{+}$, which is the best approximation of the inverse matrix.

If Q_k and P_k from Algorithm 2.1 have the property that $Q_k - P_k$ is a singular matrix, let us write

$$Q_k - P_k = \bar{Q}_k (Q_k - P_k)' \bar{Q}_k^T,$$

where $(Q_k - P_k)'$ is a diagonal matrix which has eigenvalues of the matrix $(Q_k - P_k)$ and the first s diagonal elements are equal to zero.

Respectively,

$$(Q_k - P_k)' = \begin{bmatrix} 0 & 0 \\ 0 & (Q_k - P_k)'' \end{bmatrix},$$

where $(Q_k - P_k)''$ is an invertible diagonal matrix, which has eigenvalues of the matrix $(Q_k - P_k)$ different from zero. In this case it is suggested to use the matrix

$$\begin{bmatrix} 0 & 0 \\ 0 & ((Q_k - P_k)'')^{-1} \end{bmatrix},$$

instead of the matrix $(Q_k - P_k)^{-1}$ which does not exist.

3.2. **ill-conditioned** C_{22}

In Subsection 2.1

$$\begin{aligned} \tilde{C}_{11} &\equiv C_{11} - C_{12}C_{22}^{-1}C_{12}^T, \\ \tilde{C}_{12} &\equiv C_{12}C_{22}^{-1}C_{23}, \\ \tilde{C}_{22} &\equiv C_{33} - C_{23}^T C_{22}^{-1} C_{23}, \end{aligned}$$

were defined, where C_{22} is invertible. If the matrix C_{22} is ill-conditioned, the calculation of the inverse matrix C_{22}^{-1} is numerically unstable. In order to stabilize this process QR-factorization with Givens rotation is used. Thus,

$$C_{22} = Q_{22}R_{22},$$
$$C_{22}^{-1} = R_{22}^{-1}Q_{22}^{T}.$$

It follows that

$$\tilde{C}_{11} \equiv C_{11} - C_{12}R_{22}^{-1}Q_{22}^{T}C_{12}^{T},$$
$$\tilde{C}_{12} \equiv C_{12}R_{22}^{-1}Q_{22}^{T}C_{23},$$
$$\tilde{C}_{22} \equiv C_{33} - C_{23}^{T}R_{22}^{-1}Q_{22}^{T}C_{23}.$$

The algorithm cost is $\frac{4}{3}l^3$ flops.

If QR-factorization shows instability, the matrix C_{22} can be replaced by the pseudoinverse (Moore-Penrose inverse) C_{22}^{+}.

3.3. **ill-conditioned** L

Numerical results prove that L is ill-conditioned. The conditional $\kappa(L)$ is equal to

$$\kappa(L) = \frac{\max_{\|x\|=1} \|Lx\|}{\min_{\|x\|=1} \|Lx\|} \approx 10^{20}.$$

In previous papers, preconditioning was not performed. For larger dimensions preconditioning must be done and our proposal is to apply QR-factorization of the matrix L.

4. Application of some properties of spectrum in the case of numerical instability of the SDA algorithm

In the previous section, we saw some of the classical ideas that can help to stabilize Algorithm 2.1. In this section we consider some properties of the quadratic palindromic eigenvalue problems, so we can replace the SDA2 algorithm, in the case of some of the above problems, with an appropriate algorithm for the linear eigenvalue problem, in the sense of the following two theorems:

Theorem 4.1. *If* $\pm 1 \notin \sigma(P(\lambda))$ *then the eigenvector* x *of the palindromic eigenvalue problem*

$$(\lambda^2 A_{d_1} + \lambda A_{d_0} + A_{d_1}^{T})x = 0,\ A_{d_0}^{T} = A_{d_0},\ x \neq 0, \tag{4.1}$$

satisfies the equation

$$(\lambda + \frac{1}{\lambda})x^{T}A_{d_1}x + x^{T}A_{d_0}x = 0. \tag{4.2}$$

Proof 4.1. *Let us multiply the equation* (4.1) *with* x^T *on the left. We obtain:*

$$\lambda^2 x^{T}A_{d_1}x + \lambda x^{T}A_{d_0}x + x^{T}A_{d_1}^{T}x = 0. \tag{4.3}$$

Thus for $x^{T}A_{d_1}x \in \mathbb{C}$,

$$x^{T}A_{d_1}^{T}x = (x^{T}A_{d_1}^{T}x)^{T},$$

$$x^T A_{d_1}^T x = x^T (A_{d_1}^T)^T (x^T)^T = x^T A_{d_1} x. \tag{4.4}$$

From the equation (4.3) *we have*

$$\lambda^2 x^T A_{d_1} x + \lambda x^T A_{d_0} x + x^T A_{d_1} x = 0. \tag{4.5}$$

If λ is an eigenvalue, and x is the right eigenvector of the eigenproblem (4.3) *then $\frac{1}{\lambda}$ is an eigenvalue and x^T is the left eigenvector of the eigenproblem* (4.3)*. This means that*

$$x^T (\frac{1}{\lambda^2} A_{d_1} + \frac{1}{\lambda} A_{d_0} + A_{d_1}^T) = 0. \tag{4.6}$$

If we multiply the equation (4.6) *on the right with the eigenvector x we obtain*

$$\frac{1}{\lambda^2} x^T A_{d_1} x + \frac{1}{\lambda} x^T A_{d_0} x + x^T A_{d_1}^T x = 0. \tag{4.7}$$

According to (4.4) *and* (4.7) *we get*

$$\frac{1}{\lambda^2} x^T A_{d_1} x + \frac{1}{\lambda} x^T A_{d_0} x + x^T A_{d_1} x = 0. \tag{4.8}$$

Subtracting (4.8) *from* (4.5) *we obtain*

$$(\lambda^2 - \frac{1}{\lambda^2}) x^T A_{d_1} x + (\lambda - \frac{1}{\lambda}) x^T A_{d_0} x = 0,$$

respectively

$$(\lambda - \frac{1}{\lambda})((\lambda + \frac{1}{\lambda}) x^T A_{d_1} x + x^T A_{d_0} x) = 0.$$

Since $\lambda \neq \pm 1$,

$$(\lambda + \frac{1}{\lambda}) x^T A_{d_1} x + x^T A_{d_0} x = 0. \tag{4.9}$$

holds.

It is interesting that the following holds:

Theorem 4.2. *If $\pm i \notin \sigma(P(\lambda))$ then the eigenvector x of the palindromic eigenvalue problem*

$$(\lambda^2 A_{d_1} + \lambda A_{d_0} + A_{d_1}^T) x = 0,\ A_{d_0}^T = A_{d_0},\ x \neq 0, \tag{4.10}$$

satisfies the equation

$$(\lambda + \frac{1}{\lambda}) x^T A_{d_1} x + x^T A_{d_0} x = 0. \tag{4.11}$$

Proof 4.2. *In the Proof 4.1 it is proved that equations* (4.5) *and* (4.8) *hold. By adding these two equations we obtain*

$$(\lambda^2 + 2 + \frac{1}{\lambda^2}) x^T A_{d_1} x + (\lambda + \frac{1}{\lambda}) x^T A_{d_0} x = 0,$$

i.e.

$$(\lambda + \frac{1}{\lambda})((\lambda + \frac{1}{\lambda}) x^T A_{d_1} x + x^T A_{d_0} x) = 0.$$

Since $\lambda \neq \pm i$,

$$(\lambda + \frac{1}{\lambda}) x^T A_{d_1} x + x^T A_{d_0} x = 0$$

holds.

Lemma 4.1. *For the invertible matrix A_{d_1} eigenvector of the linear eigenvalue problem*

$$(A_{d_1}^{-1}A_0)x = \mu x \tag{4.12}$$

satisfies the equation (4.9)*, where $\mu = -(\lambda + \frac{1}{\lambda})$.*

If matrix A_{d_1} is non-invertible then the eigenvector of the linear generalized eigenvalue problem

$$A_{d_0}x = -\mu A_{d_1}x$$

satisfies the equation (4.9)*.*

Lemma 4.2. *The eigenvalue of the eigenvalue problem* (1.2) *is* 0 *respectively* ∞ *if and only if the matrix A_1 is singular.*

Since the deflation of the palindromic eigenvalue problem was done first, it is clear that if the problem is reduced to a linear eigenvalue problem then it is not a generalized eigenvalue problem. In this case we apply the first part of Lemma 4.1.

Proposition 4.1. *If $\pm 1 \notin \sigma(P(\lambda))$ or $\pm i \notin \sigma(P(\lambda))$, the eigenvectors of the eigenvalue problem* (4.1) *are obtained as eigenvectors of the linear eigenvalue problem* (4.12) *or as a vector x which is normal to the vector*

$$(\lambda + \frac{1}{\lambda})xA_{d_1}x + A_{d_0}x = 0.$$

Proposition 4.2. *If the eigenvector of the eigenvalue problem* (4.1) *and eigenvalue problem* (4.12) *match, then the eigenvalue λ of the problem* (4.1) *and it's reciprocal eigenvalue $\frac{1}{\lambda}$ are obtained as solutions of the following equation*

$$-\lambda - \frac{1}{\lambda} = \mu,$$

where μ is the eigenvalue of the problem (4.12)*.*

Theorem 4.3. *If the matrix $A_{d_1}^{-1}$ can be diagonalized, then eigenvectors of the eigenvalue problem* (4.1) *are obtained according to Proposition 4.2.*

Proof 4.3. *Since the matrix $A_{d_1}^{-1}$ is diagonalized its eigenvectors are linearly independent and form the base of the space $\mathbb{C}^n$. Thus, only the zero vector is normal to the vector $(\lambda + \frac{1}{\lambda})xA_{d_1}x + A_{d_0}x = 0$.*

Remark 4.1. In the case $\pm 1 \in \sigma(P(\lambda))$ the deflation of eigenvalues is given in the paper [13].

From the above it can be seen that in the case of the problem of the SDA2 algorithm, it is better to try to reduce the problem to a linear eigenvalue problem than to apply classical stabilization methods.

5. Conclusion

In this paper important issues that can affect the convergence and stability of the algorithm are discussed. In previous papers these problems were overcome by the combination of SDA1 and SDA2 algorithms. Significant improvements were obtained. Pre-

conditioning of significant matrices was used, which is numerically better than the combination of SDA1 and SDA2 algorithms. Also, we improved the algorithm using some significant properties of the spectrum of the T- quadratic palindromic eigenvalue problem. Thus the properties of the palindromic pencil and the QEP structure was preserved. Further research: performing more extensive numerical tests, as well as the consideration of problems of higher dimensions and expanding the consideration of spectrum properties.

REFERENCES

[1] P. Benner, M. Bolhoefer, D. Kressner, C. Mehl and T. Stykel, editors, *Numerical Algebra, Matrix Theory, Differential-Algebraic Equations and Control Theory*, Festrschrift in Honor of Volker Mehrmann, Springer, 2015.

[2] R. Byers, D.S. Mackey, V. Mehrmann and H. Xu, *Symplectic, BVD, palindromic approaches to discrete-time control problems*, Technical report, TU Berlin, MATHEON, Germany, 2008.

[3] E. K.-W. Chu, H.-Y. Fan, and W.-W. Lin, *A structure-preserving doubling algorithm for continuous-time algebraic Riccati equations*, Linear Algebra Appl., 396:55-80, 2005.

[4] E. K.-W. Chu, H.-Y. Fan, W.-W. Lin, and C.-S. Wang, *Structure-preserving algorithms for periodic discrete-time algebraic Riccati equations*, Int. J. Control, 77:767-788, 2004.

[5] E. K.-W. Chu, T.-M. Huang, W.-W. Lin, and C.-T. Wu, *Vibration of Fast Trains,Palindromic Eigenvalue Problems and Structure-Preserving Doubling Algorithms*, J. Comput. Appl. Math., 219(1):237-252, 2008.

[6] E. K.-W. Chu, T.-M. Huang, W.-W. Lin, and C.-T. Wu, *Palindromic eigenvalue problems: a brief survey*, Taiwanese J. Math., 14(3A):743-779, 2010.

[7] X.-X. Guo, W.-W. Lin, and S.-F. Xu, *A structure-preserving doubling algorithm for nonsymmetric algebraic Riccati equation*, Num. Math., 103:393-412, 2006.

[8] N.J. Higham, D.S. Mackey, N. Mackey and F. Tisseur, *Symmetric linearizations of matrix polynomials*, SIAM J. Matrix Anal. Appl., 29:143-159, 2006.

[9] T.-M. Hwang, E.K.-W. Chu, and W.-W. Lin, *A generalized structure-preserving doubling algorithm for generalized discrete-time algebraic Riccati equations*, Int. J. Control, 78:1063-1075, 2005.

[10] T.-M. Hwang and W.-W. Lin, *Structured doubling algorithms for weak Hermitian solutions of algebraic Riccati equations*, Technical report, NCTS Preprints in Mathematics, National Tsing Hua University, Hsinchu, Taiwan, 2006-7-009, 2006.

[11] C.F. Ipsen, *Accurate eigenvalues for fast trains*, SIAM News, 37, 2004.

[12] W.-W. Lin and S.-F. Xu, *Convergence analysis of structure-perserving doubling algorithms for Riccati-tye matrix equations*, SIAM J. Matrix Anal. Appl., 28(1):26-39, 2006.

[13] D.S. Mackey, N. Mackey, C. Mehl and V. Mehrmann, *Palindromic polynomial eigenvalue problems: Good vibrations from good linearizations*, Technical report, DFG Research Center, Technische Universitaet, Berlin, Germany, 2005.

[14] D.S. Mackey, N. Mackey, C. Mehl and V. Mehrmann, *Structured polynomial eigenvalue problems: Good vibrations from good linearizations*, SIAM J. Matrix Anal. Appl., 28:1029-1051, 2006.

[15] D.S. Mackey, N. Mackey, C. Mehl and V. Mehrmann, *Vector spaces of linearizations for matrix polynomials*, SIAM J. Matrix Anal. Appl., 28:971-1004, 2006.

[16] D.S. Mackey, N. Mackey, C. Mehl and V. Mehrmann, *Numerical methods for palindromic eigenvalue problems*, Technical report, Technische Universitaet, Berlin, MATHEON, Germany, 2007.

Special Editions ANUBiH, Book CCXVI, OPMN Book 30, pp. 21–32

(Received: May 17, 2024)
(Revised: June 07, 2024)

Aleksandra Kostić
University of Sarajevo
Faculty of Mechanical Engineering
Vilsonovo šetalište 9, 71000 Sarajevo, BiH
e-mail: *kostic@mef.unsa.ba*
and
Valentina Timotić
University of East Sarajevo
Faculty of Philosophy
Alekse Šantića 1, 71420 Pale, BiH
e-mail: *valentina.timotic@ffuis.edu.ba*
and
Izet Horman
University of Sarajevo
Faculty of Mechanical Engineering
Vilsonovo šetalište 9, 71000 Sarajevo, BiH
e-mail: *horman@mef.unsa.ba*

PROCEEDINGS OF THE CONFERENCE ON MARCH 14 – INTERNATIONAL DAY OF MATHEMATICS
Special Editions ANUBiH, Book CCXVI, OPMN Book 30
Sarajevo 2024, **p. 33–37** DOI: 10.5644/PI2024.216.05

CONNECTIVITY ESTIMATES IN THE HOMOLOGICAL TAYLOR TOWER FOR THE SPACE OF REDUCED EMBEDDINGS IN $\mathbb{R}^n$

FRANJO ŠARČEVIĆ

ABSTRACT. Define $\overline{\mathrm{Emb}}(M,\mathbb{R}^n)$, the space of reduced embeddings of a smooth manifold M in $\mathbb{R}^n$, to be the homotopy fiber of the inclusion map $\mathrm{Emb}(M,\mathbb{R}^n) \to \mathrm{Imm}(M,\mathbb{R}^n)$, where $\mathrm{Imm}(M,\mathbb{R}^n)$ is the space of immersions of M in $\mathbb{R}^n$, and denote by $\underline{H}\mathbb{Z}$ the Eilenberg-MacLane spectrum. The Taylor tower for the space $\underline{H}\mathbb{Z} \wedge \overline{\mathrm{Emb}}(M,\mathbb{R}^n)$, which is the homological version of the tower for the space $\mathrm{Emb}(M,\mathbb{R}^n)$, is known to converge under certain dimensional assumptions, meaning that the connectivity of the map from $\underline{H}\mathbb{Z} \wedge \overline{\mathrm{Emb}}(M,\mathbb{R}^n)$ to its k^{th} polynomial approximation $T_k \underline{H}\mathbb{Z} \wedge \overline{\mathrm{Emb}}(M,\mathbb{R}^n)$ approaches ∞ as k approaches ∞. Here we give a brief exposition of the known results and derive a slightly better connectivity estimate using a recent result obtained for the space of r-immersions.

1. INTRODUCTION

Manifold calculus of functors, or Goodwillie calculus, studies *good* (meaning *finitary* and *isotopy*) contravariant functors $F\colon \mathcal{O}(M) \to \mathcal{C}$, where $\mathcal{O}(M)$ is the category of open subsets of a smooth manifold M with inclusions as morphisms, and $\mathcal{C}$ is a suitable category (usually Top or Spectra).

The central question in the theory is that of the convergence of the *Taylor tower*

$$F(-) \to (T_\infty F(-) \to \cdots \to T_k F(-) \to \cdots \to T_0 F(-))$$

associated to the functor. Here $T_k F(-)$, k-th stage of the tower, is a k-th polynomial approximation of the functor, and $T_\infty F(-)$ is the inverse limit of the tower. There are two convergence questions: *intrinsic* convergence of the tower, which means that the connectivity of the map between two succesive stages $T_{k+1}F(-) \to T_k F(-)$ approaches ∞ as k approaches ∞, and the convergence *to* the tower, which means that there exists a weak equivalence between $F(-)$ and $T_\infty F(-)$.

Define $\mathrm{Emb}(M,\mathbb{R}^n)$ to be the space of embedding of M in $\mathbb{R}^n$. The central result of the Goodwillie calculus is the Goodwillie-Klein-Weiss theorem which, in a special case, says that the map

$$T_{k+1}\,\mathrm{Emb}(M,\mathbb{R}^n) \to T_k\,\mathrm{Emb}(M,\mathbb{R}^n)$$

2020 *Mathematics Subject Classification.* Primary: 18F50; Secondary: 57R40.
Key words and phrases. embeddings, r-immersions, functor calculus.

is $(k(n-m-2)-m+1)$-connected and that the map

$$\mathrm{Emb}(M,\mathbb{R}^n) \to T_k\,\mathrm{Emb}(M,\mathbb{R}^n)$$

is $(k(n-m-2)-m+1)$-connected. Therefore, as long as $n > m+2$, the Taylor tower

$$\mathrm{Emb}(M,\mathbb{R}^n) \to (T_\infty\,\mathrm{Emb}(M,\mathbb{R}^n) \to \cdots \to T_k\,\mathrm{Emb}(M,\mathbb{R}^n) \to \cdots \to T_0\,\mathrm{Emb}(M,\mathbb{R}^n))$$

converges intrinsically and to the tower. The details can be found in [2, 3, 5, 6].

This convergence is actually homotopical convergence, because the connectivity in question here is the homotopical one. We can also consider the homological version of the Taylor tower for $\mathrm{Emb}(M,\mathbb{R}^n)$. Taking the smash product $\wedge$ of the Eilenberg-MacLane spectrum $\underline{H}\mathbb{Z}$ with a based space X produces the spectrum $\underline{H}\mathbb{Z}\wedge X$ whose homotopy is equivalent to the reduced homology of the space X; more precisely, there exists an isomorphism $\pi_i(\underline{H}\mathbb{Z}\wedge X) \cong \widetilde{\mathrm{H}}_i(X;\mathbb{Z})$.

Thus, we consider the Taylor tower for the space

$$\underline{H}\mathbb{Z}\wedge\overline{\mathrm{Emb}}(M,\mathbb{R}^n)$$

to be the *homological Taylor tower* for $\mathrm{Emb}(M,\mathbb{R}^n)$, where we have replaced the space of embeddings with embeddings modulo immersions defined by

$$\overline{\mathrm{Emb}}(M,\mathbb{R}^n) = \mathrm{hofiber}\,(\mathrm{Emb}(M,\mathbb{R}^n) \to \mathrm{Imm}(M,\mathbb{R}^n))\,,$$

which is convenient (to cancel the tangential data of the immersion).

As shown in [8], the connectivity of the map

$$\underline{H}\mathbb{Z}\wedge\overline{\mathrm{Emb}}(M,\mathbb{R}^n) \to T_k\underline{H}\mathbb{Z}\wedge\overline{\mathrm{Emb}}(M,\mathbb{R}^n) \tag{1.1}$$

is

$$(k+1)\left(\frac{n}{2}-m-\frac{1}{2}\right) \tag{1.2}$$

and the tower converges for $n > 2m+1$.

Actually, when we write $\underline{H}\mathbb{Z}\wedge\overline{\mathrm{Emb}}(M,\mathbb{R}^n)$ we really mean the *taming* of the functor $\underline{H}\mathbb{Z}\wedge\overline{\mathrm{Emb}}(-,\mathbb{R}^n)$, evaluated on a *tame* manifold M, which is the interior of a compact manifold with boundary. Namely, even if a cofunctor $F(-)\colon \mathcal{O}(M) \to \mathrm{Top}$ is good, the cofunctor $\underline{J}\wedge F(-)\colon \mathcal{O}(M) \to \mathrm{Spectra}$ for a fixed spectrum $\underline{J}$ is not good [4, 8], but the taming of this functor *is* good. Therefore, when evaluated on a *tame subset* of M – an element of $\mathcal{O}(M)$ which is the interior of a compact smooth codimension zero submanifold od M – the taming of a functor is equivalent to the functor. So, when evaluated on tame manifolds, there is no difference between $\underline{J}\wedge F(M)$ and the taming of it.

Here we will provide a stronger connectivity estimate for the map (1.1) (Proposition 2.1). It is

$$(k+1)\left(\frac{n}{2}-m-\frac{1}{2}\right)+(n-1)\left(\frac{1}{2}+\frac{k \bmod 2}{2}\right).$$

Prior to that, let us present two of the three results on which this story is based. The notion of analyticity is explained in the cited literature.

Theorem 1.1 ([5]). *Let F be a ρ-analytic good cofunctor with excess c and U the interior of a smooth compact codimension 0 submanifold of M of handle index $q < \rho$. Then $F(U) \to T_k F(U)$ is $(c + (k+1)(\rho - q))$-connected.*

Weiss provided the following result.

Theorem 1.2 ([8]). *If a good cofunctor $F\colon O(M) \to$ Top is ρ-analytic with excess $c < 0$, where $\rho + \frac{c}{l} > m$, such that $T_{l-1}F(-)$ vanishes for some $l > 0$, and $\underline{J}$ is a (-1)-connected CW-spectrum, then the taming of the functor $\underline{J} \wedge F(-)$ is $\left(\rho + \frac{c}{l}\right)$-analytic with excess 0.*

2. Estimates

It is known that the functor $F(-) = \overline{\mathrm{Emb}}(-, \mathbb{R}^n)$ is $(n-2)$-analytic with excess $3-n$ [2, 3, 5].

The spectrum $\underline{H}\mathbb{Z}$ is a (-1)-connected CW-spectrum, because it does not have nontrivial homotopy groups in negative dimensions (actually, it is nontrivial $\mathbb{Z}$ only in the 0-th dimension).

Also, $T_1 F(-)$ vanishes because $T_1 \mathrm{Emb}(-, \mathbb{R}^n) \simeq \mathrm{Imm}(-, \mathbb{R}^n)$ and $T_1 \mathrm{Imm}(-, \mathbb{R}^n) \simeq \mathrm{Imm}(-, \mathbb{R}^n)$ [7], so $l = 2$ in terms of Theorem 1.2.

It follows from Theorem 1.2 that the functor $\underline{H}\mathbb{Z} \wedge \overline{\mathrm{Emb}}(-, \mathbb{R}^n)$ is $\left(\frac{n}{2} - \frac{1}{2}\right)$-analytic with excess 0.

Now from Theorem 1.1 and the remarks on tameness it follows that the map

$$\underline{H}\mathbb{Z} \wedge \overline{\mathrm{Emb}}(M, \mathbb{R}^n) \to T_k \underline{H}\mathbb{Z} \wedge \overline{\mathrm{Emb}}(M, \mathbb{R}^n)$$

is

$$(k+1)\left(\frac{n}{2} - m - \frac{1}{2}\right)\text{-connected.}$$

In [1] the authors study the space of r-immersions, which are the immersions without r-fold self intersections. That is, the space $\mathrm{rImm}(M, \mathbb{R}^n)$ of r-immersions of M in $\mathbb{R}^n$ is the space of immersions of M in $\mathbb{R}^n$ with the property that at most $r-1$ points of M are mapped to the same point in $\mathbb{R}^n$.

The part of the central result is the following:

Theorem 2.1 ([1]). *When $r \leq n+1$, the map*

$$T_k \underline{H}\mathbb{Z} \wedge \overline{\mathrm{rImm}}(M, \mathbb{R}^n) \to T_{k-1} \underline{H}\mathbb{Z} \wedge \overline{\mathrm{rImm}}(M, \mathbb{R}^n)$$

is

$$k\left(n\frac{r-1}{r} - m - \frac{1}{r}\right) - \frac{k \bmod r}{r}(r - n - 1)\text{-connected.}$$

The tower converges intrinsically if

$$n > \frac{rm+1}{r-1}.$$

As is clear from the definition, injective immersions are just 2-immersions. If M is compact, then injective immersions are the same thing as embeddings. That is, for M compact,

$$\mathrm{Emb}(M, \mathbb{R}^n) = 2\,\mathrm{Imm}(M, \mathbb{R}^n).$$

The same is true in a more general case relevant to us: when M is tame, the space $2\mathrm{Imm}(M,\mathbb{R}^n)$ is equivalent to the space $\mathrm{Emb}(M,\mathbb{R}^n)$.

So, in our consideration, letting $r = 2$ in Theorem 2.1 we get a result for the space of reduced embeddings.

Corollary 2.1. *The connectivity of the map*

$$T_{k+1}\underline{H}\mathbb{Z}\wedge\overline{\mathrm{Emb}}(M,\mathbb{R}^n)\to T_k\underline{H}\mathbb{Z}\wedge\overline{\mathrm{Emb}}(M,\mathbb{R}^n)$$

is

$$(k+1)\left(\frac{n}{2}-m-\frac{1}{2}\right)+(n-1)\left(\frac{1}{2}+\frac{k \bmod 2}{2}\right).$$

The tower converges intrinsically if

$$n > 2m+1.$$

Proposition 2.1. *Let M be a tame manifold. The connectivity of the map*

$$\underline{H}\mathbb{Z}\wedge\overline{\mathrm{Emb}}(M,\mathbb{R}^n)\to T_k\underline{H}\mathbb{Z}\wedge\overline{\mathrm{Emb}}(M,\mathbb{R}^n)$$

is

$$(k+1)\left(\frac{n}{2}-m-\frac{1}{2}\right)+(n-1)\left(\frac{1}{2}+\frac{k \bmod 2}{2}\right).$$

Proof. It is known and easy to prove that, if $f\colon X\to Y$ and $g\colon Y\to Z$ are k-connected maps, then $g\circ f\colon X\to Z$ is also a k-connected map. If a map is k-connected, then it is j-connected for all $j\le k$. Now, if f is ∞-connected (i.e. a weak equivalence), then f is also k-connected for all k, so g and $g\circ f$ have the same connectivity.

Using the fact that the map

$$\underline{H}\mathbb{Z}\wedge\overline{\mathrm{Emb}}(M,\mathbb{R}^n)\to T_\infty\underline{H}\mathbb{Z}\wedge\overline{\mathrm{Emb}}(M,\mathbb{R}^n)$$

is a weak equivalence, this means that the connectivities c_1 and c_2 in the diagram

$$\begin{array}{ccc}\underline{H}\mathbb{Z}\wedge\overline{\mathrm{Emb}}(M,\mathbb{R}^n) & \xrightarrow{\sim} & T_\infty\underline{H}\mathbb{Z}\wedge\overline{\mathrm{Emb}}(M,\mathbb{R}^n)\\ & \searrow^{c_2} & \downarrow{\scriptstyle c_1}\\ & & T_k\underline{H}\mathbb{Z}\wedge\overline{\mathrm{Emb}}(M,\mathbb{R}^n)\end{array}$$

are the same.

Also, if the map

$$T_{k+1}\underline{H}\mathbb{Z}\wedge\overline{\mathrm{Emb}}(M,\mathbb{R}^n)\to T_k\underline{H}\mathbb{Z}\wedge\overline{\mathrm{Emb}}(M,\mathbb{R}^n)$$

is c-connected, then the connectivity of the map

$$T_\infty\underline{H}\mathbb{Z}\wedge\overline{\mathrm{Emb}}(M,\mathbb{R}^n)\to T_k\underline{H}\mathbb{Z}\wedge\overline{\mathrm{Emb}}(M,\mathbb{R}^n)$$

is at least c.

This, together with Corollary 2.1 implies that the map

$$T_\infty\underline{H}\mathbb{Z}\wedge\overline{\mathrm{Emb}}(M,\mathbb{R}^n)\to T_k\underline{H}\mathbb{Z}\wedge\overline{\mathrm{Emb}}(M,\mathbb{R}^n)$$

is also

$$(k+1)\left(\frac{n}{2}-m-\frac{1}{2}\right)+(n-1)\left(\frac{1}{2}+\frac{k \bmod 2}{2}\right)\text{-connected.}$$

That finally implies that the map

$$\underline{H}\mathbb{Z}\wedge\overline{\mathrm{Emb}}(M,\mathbb{R}^n)\to T_k\underline{H}\mathbb{Z}\wedge\overline{\mathrm{Emb}}(M,\mathbb{R}^n)$$

is

$$(k+1)\left(\frac{n}{2}-m-\frac{1}{2}\right)+(n-1)\left(\frac{1}{2}+\frac{k \bmod 2}{2}\right)\text{-connected.}$$

□

The connectivity estimate (1.2) is improved by the number

$$(n-1)\left(\frac{1}{2}+\frac{k \bmod 2}{2}\right).$$

If k is odd, this number is $n-1$; if k is even, this number is $\frac{1}{2}(n-1)$.

REFERENCES

[1] G. Arone and F. Šarčević, *Intrinsic convergence of the homological Taylor tower for r-immersions in* $\mathbb{R}^n$, Homol. Homotopy Appl., 26(2), 163–192, 2024.
[2] T. G. Goodwillie and J. R. Klein. *Multiple disjunction for spaces of Poincaré embeddings*, J. Topol., 1(4):761–803, 2008.
[3] T. G. Goodwillie and J. R. Klein, *Multiple disjunction for spaces of smooth embeddings*, J. Topol., 8(3):651–674, 2015.
[4] T. G. Goodwillie, J. R. Klein, and M. S. Weiss, *Spaces of smooth embeddings, disjunction and surgery*, In Surveys on surgery theory, Vol. 2, volume 149 of Ann. of Math. Stud., pages 221–284, Princeton Univ. Press, Princeton, NJ, 2001.
[5] T. G. Goodwillie and M. Weiss, *Embeddings from the point of view of immersion theory II*, Geom. Topol., 3:103–118 (electronic), 1999.
[6] F. Šarčević and I. Volić, *A streamlined proof of the convergence of the Taylor tower for embeddings in* $\mathbb{R}^n$, Colloq. Math., 156(1), 91–122, 2019.
[7] M. Weiss, *Embeddings from the point of view of immersion theory I*, Geom. Topol., 3:67–101 (electronic), 1999.
[8] M. S. Weiss, *Homology of spaces of smooth embeddings*, Q. J. Math. 55(4), 499–504, 2004.

(Received: April 24, 2024)
(Revised: September 09, 2024)

Franjo Šarčević
University of Sarajevo
Department of Mathematics and Computer Science
Zmaja od Bosne 33-35, 71000 Sarajevo
e-mail: *franjo.sarcevic@pmf.unsa.ba*
URL: *pmf.unsa.ba/franjos*

PROCEEDINGS OF THE CONFERENCE ON MARCH 14 – INTERNATIONAL DAY OF MATHEMATICS
Special Editions ANUBiH, Book CCXVI, OPMN Book 30
Sarajevo 2024, **p. 38–47** DOI: 10.5644/PI2024.216.06

ASYMPTOTIC BEHAVIOR OF NON-AUTONOMOUS COMPETITIVE SYSTEMS OF DIFFERENCE EQUATIONS

MEHMED NURKANOVIĆ

Dedicated to the 75th birthday of our dear Professor Mirjana Vuković

ABSTRACT. The problem of the behavior (convergence and stability) of the solutions of non-autonomous systems of difference equations with asymptotically constant coefficients is still open. In previous research, the main interest was related to the results on global attractiveness for some classes of non-autonomous competitive and cooperative systems. In this paper, using those results and the same methods in the partial ordering of the space $\mathbb{R}_+^2$, we prove the general theorem for any non-autonomous competitive system with asymptotically constant coefficients. The obtained results are also illustrated with concrete examples.

1. INTRODUCTION

An autonomous system of difference equations has constant coefficients, while a non-autonomous system has variable coefficients (sequences). In this paper, we will consider the behavior of non-autonomous competitive systems of difference equations whose coefficients are asymptotically constant. This problem is still open and has yet to have a general result covering all cases. Nevertheless, we will give such a result here for the case of competitive systems. It relies on previously obtained results for some general classes of these systems.

In [9] the following non-autonomous competitive systems whose coefficients are asymptotically constant:

$$X_{n+1} = \begin{bmatrix} x_{n+1} \\ y_{n+1} \end{bmatrix} = \begin{bmatrix} a_n f(x_n, y_n) \\ b_n g(x_n, y_n) \end{bmatrix}, \quad n = 0, 1, \ldots,$$

$$X_{n+1} = \begin{bmatrix} x_{n+1} \\ y_{n+1} \end{bmatrix} = \begin{bmatrix} \dfrac{x_n}{a_n + y_n} \\ \dfrac{y_n}{b_n + x_n} \end{bmatrix}, \quad n = 0, 1, \ldots,$$

2020 *Mathematics Subject Classification.* 39A22, 39A30.

Key words and phrases. non-autonomous systems, discrete dynamical systems, difference equations, stability.

$$X_{n+1} = \begin{bmatrix} x_{n+1} \\ y_{n+1} \end{bmatrix} = \begin{bmatrix} \dfrac{\alpha_n x_n}{a_n + y_n} \\ \dfrac{\beta_n y_n}{b_n + x_n} \end{bmatrix}, \quad n = 0, 1, ...,$$

and the following non-autonomous *Leslie-Gower model*

$$X_{n+1} = \begin{bmatrix} \dfrac{a_n x_n}{1 + c_n^{(11)} x_n + c^{(12)} y_n} \\ \dfrac{b_n y_n}{1 + c^{(21)} x_n + c_n^{(22)} y_n} \end{bmatrix}, \quad n = 0, 1, 2, ... ,$$

were considered, and a theorem of the form Theorem 2.1 was proved in that case. After that, the corresponding Leslie-Gower evolutionary model with two Fisher's equations was considered separately.

Also, in [10] the following non-autonomous cooperative systems:

$$x_{n+1}^{(i)} = A_n^{(i)} x_n^{(i)} \frac{\prod\limits_{i \neq j=1}^{k} x_n^{(j)}}{1 + \prod\limits_{i \neq j=1}^{k} x_n^{(j)}}, \quad n = 0, 1, ...; i = 1, 2, ..., k,$$

and

$$\left.\begin{aligned} x_{n+1} &= \frac{a_n x_n}{\delta_1 + x_n} + \frac{b_n y_n}{\delta_2 + y_n}, \\ y_{n+1} &= \frac{c_n x_n}{\delta_2 + x_n} + \frac{d_n y_n}{\delta_1 + y_n}, \end{aligned}\right\} n = 0, 1, \ldots$$

were considered.

All obtained results are based on the behavior of the corresponding autonomous competitive and cooperative systems. Regarding autonomous competitive and cooperative systems, see [2–5, 7, 8, 11–15].

In this paper, we use the method of difference inequalities to prove global attractivity results for two-dimensional competitive systems in [9]. The map $F : \mathbb{R}_+^2 \to \mathbb{R}_+^2$, $F = (f, g)$ is called a competitive map if f is non-decreasing in the first variable and non-increasing in the second variable, and g is non-increasing in the first variable and non-decreasing in the second variable. However, the results in [9] are two-dimensional, and it is not clear how to extend them to the k-dimensional case for $k > 2$.

Also, we will use the so-called "north-east" partial ordering of the space $\mathbb{R}_+^2$, defined it in the following way:

$$X_n = \begin{bmatrix} x_n^{(1)} \\ x_n^{(2)} \end{bmatrix} \preccurlyeq_{ne} Y_n = \begin{bmatrix} y_n^{(1)} \\ y_n^{(2)} \end{bmatrix} \Longleftrightarrow \left(x_n^{(1)} \leq y_n^{(1)} \text{ and } x_n^{(2)} \leq y_n^{(2)} \right),$$

and the so-called "south-east" partial ordering of the space $\mathbb{R}_+^2$ defined by

$$X_n = \begin{bmatrix} x_n^{(1)} \\ x_n^{(2)} \end{bmatrix} \preccurlyeq_{se} Y_n = \begin{bmatrix} y_n^{(1)} \\ y_n^{(2)} \end{bmatrix} \Longleftrightarrow \left(x_n^{(1)} \leq y_n^{(1)} \text{ and } x_n^{(2)} \geq y_n^{(2)} \right).$$

If we replace "$\leq$" and "$\geq$" with "$<$" and "$>$" in the above relations, then "$\preceq$" also changes to "$\prec$".

2. Main results

In [9], the following lemma is proved.

Lemma 2.1. *([9], Lemma 1) Assume that*

a) $F:\mathbb{R}_+^2\to\mathbb{R}_+^2$, $F=\begin{bmatrix} f \\ g \end{bmatrix}$ *is a competitive map.*

b) $\{X_n\}$, $\{Y_n\}$, $\{Z_n\}$ *are sequences of the real components in* $\mathbb{R}_+^2$ *such that*

$$X_0 \preccurlyeq_{se} Y_0 \preccurlyeq_{se} Z_0$$

and

$$\left.\begin{array}{c} X_{n+1}\preccurlyeq_{se} F(X_n) \\ Y_{n+1}=F(Y_n) \\ Z_{n+1}\succcurlyeq_{se} F(Z_n) \end{array}\right\},\quad n=0,1,2,\ldots.$$

Then,

$$X_n \preccurlyeq_{se} Y_n \preccurlyeq_{se} Z_n,\quad n=0,1,2,\ldots. \tag{2.1}$$

Lemma 2.1 is necessary to obtain the following general result on the behavior of non-autonomous competitive systems of difference equations whose coefficients are asymptotically constant.

Theorem 2.1. *Consider the following non-autonomous system of difference equations*

$$X_{n+1}=\begin{bmatrix} f(a_1(n),\ldots,a_k(n);x_n,y_n) \\ g(b_1(n),\ldots,b_l(n);x_n,y_n) \end{bmatrix},\quad n=0,1,2,\ldots, \tag{2.2}$$

where $A_n=[a_1(n),\ldots,a_k(n),b_1(n),\ldots,b_l(n)]^T$, *$k$ and l are positive integers, and* $F=\begin{bmatrix} f \\ g \end{bmatrix}:\mathbb{R}_+^2\to\mathbb{R}_+^2$ *is a competitive map. Assume that*

$$\lim_{n\to\infty} A_n=\lim_{n\to\infty}[a_1(n),\ldots,a_k(n),b_1(n),\ldots,b_l(n)]^T=[a_1,\ldots,a_k,b_1,\ldots,b_l]^T=A. \tag{2.3}$$

Also, assume that there exists $\varepsilon_0=\left[\varepsilon_0^{(1)},\ldots,\varepsilon_0^{(k)},\varepsilon_0^{(k+1)},\ldots,\varepsilon_0^{(k+l)}\right]^T \succ_{ne} \left[\underbrace{0,\ldots,0}_{k+l}\right]^T$ *such that for every* $A_\varepsilon=[\alpha_1,\ldots,\alpha_k,\beta_1,\ldots,\beta_l]^T$, *with*

$$\alpha_i\in\left(a_i-\varepsilon_0^{(i)},a_i+\varepsilon_0^{(i)}\right),\ \beta_j\in\left(b_j-\varepsilon_0^{(j)},b_j+\varepsilon_0^{(j)}\right),\ i=1,\ldots,k;j=1,\ldots,l,$$

all the solutions of the system

$$Y_{n+1}=\begin{bmatrix} f(\alpha_1,\ldots,\alpha_k;u_n,v_n) \\ g(\beta_1,\ldots,\beta_l;u_n,v_n) \end{bmatrix},\quad n=0,1,2,\ldots,;k,l\in\mathbb{Z}^+ \tag{2.4}$$

converge to a constant solution $\overline{Y}_{A_\varepsilon}=\begin{bmatrix} \overline{x}_{A_\varepsilon} \\ \overline{y}_{A_\varepsilon} \end{bmatrix}$.

Additionally, suppose that $\lim\limits_{A_\varepsilon\to A}\overline{Y}_{A_\varepsilon}=\overline{Y}_A$.

Then, every solution of the system (2.2) converges to $\overline{Y}_A$.

Proof. According to (2.3), for any

$$\varepsilon_1 = \left[\varepsilon_{1,1},...,\varepsilon_{1,k}\right]^T \succ_{ne} \left[\underbrace{0,...,0}_{k}\right]^T \text{ and } \varepsilon_2 = \left[\varepsilon_{2,1},...,\varepsilon_{2,l}\right]^T \succ_{ne} \left[\underbrace{0,...,0}_{l}\right]^T,$$

there exists $N = N(\varepsilon_1, \varepsilon_2)$ such that for $n \geq N$ the following holds:

$$\begin{aligned} a_i - \varepsilon_{1,i} &< a_i(n) < a_i + \varepsilon_{1,i}, \quad i = 1,2,...,k, \\ b_j - \varepsilon_{2,j} &< b_j(n) < b_j + \varepsilon_{2,j}, \quad j = 1,2,...l. \end{aligned}$$

Thus, for $n \geq N$, we get

$$\begin{aligned} \begin{bmatrix} f\left(a_{L,1},...,a_{L,k};x_n,y_n\right) \\ g\left(b_{L,1},...,b_{L,l};x_n,y_n\right) \end{bmatrix} \preccurlyeq_{se} X_{n+1} = \begin{bmatrix} f\left(a_1(n),...,a_k(n);x_n,y_n\right) \\ g\left(b_1(n),...,b_l(n);x_n,y_n\right) \end{bmatrix} \\ \preccurlyeq_{se} \begin{bmatrix} f\left(a_{D,1},...,a_{D,k};x_n,y_n\right) \\ g\left(b_{D,1},...,b_{D,l};x_n,y_n\right) \end{bmatrix}, \end{aligned} \tag{2.5}$$

for $a_{L,i} = a_i - \varepsilon_{1,i}$, or $a_{L,i} = a_i + \varepsilon_{1,i}$ $(i = 1,...,k)$ and $b_{L,j} = b_j - \varepsilon_{2,j}$ or $b_{L,j} = b_j + \varepsilon_{2,j}$ $(j = 1,...,l)$, and

$$a_{D,i} = \begin{cases} a_i - \varepsilon_{1,i} & \text{if } a_{L,i} = a_i + \varepsilon_{1,i} \\ a_i + \varepsilon_{1,i} & \text{if } a_{L,i} = a_i - \varepsilon_{1,i} \end{cases} \quad (i = 1,...,k),$$

and

$$b_{D,j} = \begin{cases} b_j - \varepsilon_{2,j} & \text{if } b_{L,j} = b_j + \varepsilon_{2,j} \\ b_j + \varepsilon_{2,j} & \text{if } b_{L,j} = b_j - \varepsilon_{2,j} \end{cases} \quad (j = 1,...,l).$$

Since $F = \begin{bmatrix} f\left(a_1(n),...,a_k(n);x_n,y_n\right) \\ g\left(b_1(n),...,b_l(n);x_n,y_n\right) \end{bmatrix}$ is a competitive map, Lemma 2.1 implies

$$L_n \preccurlyeq_{se} X_n \preccurlyeq_{se} U_n, \quad n \geq N(\varepsilon), \tag{2.6}$$

where $\{L_n\} = \left\{\begin{bmatrix} l_n^{(1)} \\ l_n^{(2)} \end{bmatrix}\right\}$ satisfies

$$L_{n+1} = \begin{bmatrix} f\left(a_{L,1},...,a_{L,k};l_n^{(1)},l_n^{(2)}\right) \\ g\left(b_{L,1},...,b_{L,l};l_n^{(1)},l_n^{(2)}\right) \end{bmatrix},$$

and $\{U_n\} = \left\{\begin{bmatrix} u_n^{(1)} \\ u_n^{(2)} \end{bmatrix}\right\}$ satisfies

$$U_{n+1} = \begin{bmatrix} f\left(a_{D,1},...,a_{D,k};u_n^{(1)},u_n^{(2)}\right) \\ g\left(b_{D,1},...,b_{D,l}u_n^{(1)},u_n^{(2)}\right) \end{bmatrix}.$$

By using (2.6), we obtain

$$\lim_{n\to\infty} L_n \preccurlyeq_{se} \liminf_{n\to\infty} X_n \preccurlyeq_{se} \limsup_{n\to\infty} X_n \preccurlyeq_{se} \lim_{n\to\infty} U_n,$$

i.e.,

$$\overline{Y}_{\alpha_{\varepsilon_1,\varepsilon_2}} \preccurlyeq_{se} \liminf_{n\to\infty} X_n \preccurlyeq_{se} \limsup_{n\to\infty} X_n \preccurlyeq_{se} \overline{Y}_{\beta\varepsilon_1,\varepsilon_2}, \tag{2.7}$$

where $\alpha_{\varepsilon_1,\varepsilon_2} = \begin{bmatrix} \mathbf{a}_L \\ \mathbf{b}_L \end{bmatrix}$, $\beta_{\varepsilon_1,\varepsilon_2} = \begin{bmatrix} \mathbf{a}_D \\ \mathbf{b}_D \end{bmatrix}$, and

$$\mathbf{a}_L = \begin{bmatrix} a_{L,1} \\ \vdots \\ a_{L,k} \end{bmatrix}, \mathbf{b}_L = \begin{bmatrix} b_{L,1} \\ \vdots \\ b_{L,l} \end{bmatrix}, \mathbf{a}_D = \begin{bmatrix} a_{D,1} \\ \vdots \\ a_{D,k} \end{bmatrix}, \mathbf{b}_D = \begin{bmatrix} b_{D,1} \\ \vdots \\ b_{D,l} \end{bmatrix}.$$

Since $\lim\limits_{\substack{\varepsilon_1\to\mathbf{0}\\ \varepsilon_2\to\mathbf{0}}} \overline{Y}_{\alpha_{\varepsilon_1,\varepsilon_2}} = \lim\limits_{\substack{\varepsilon_1\to\mathbf{0}\\ \varepsilon_2\to\mathbf{0}}} \overline{Y}_{\beta_{\varepsilon_1,\varepsilon_2}} = \overline{Y}_A$, where $\mathbf{0} = \left[\underbrace{0,\ldots,0}_{k+l}\right]^T$, (2.7) implies that the sequence $\{X_n\}$ is convergent and that

$$\lim_{n\to\infty} X_n = \overline{Y}_A. \qquad \square$$

Remark 2.1. The condition on the system (2.4) means that the map associated with the system (2.2) is structurally stable.

Now, we will state a more general non-autonomous Leslie-Gower model and demonstrate the individual steps of the proof of Theorem 2.1.

Consider the following general non-autonomous Leslie-Gower model (see [6], [9], [16])

$$X_{n+1} = \begin{bmatrix} \dfrac{a_n x_n}{1 + c_n^{(11)} x_n + c_n^{(12)} y_n} \\ \dfrac{b_n y_n}{1 + c_n^{(21)} x_n + c_n^{(22)} y_n} \end{bmatrix}, \quad n = 0, 1, 2, \ldots, \tag{2.8}$$

and assume that

$$\lim_{n\to\infty} A_n = \lim_{n\to\infty} \left[a_n, c_n^{(11)}, c_n^{(12)}, b_n, c_n^{(21)}, c_n^{(22)}\right]^T = \left[a, c^{(11)}, c^{(12)}, b, c^{(21)}, c^{(22)}\right]^T = A.$$

Note that the condition (2.5) for the system (2.8) has the form:

$$L_n \preccurlyeq_{se} X_{n+1} = \begin{bmatrix} \dfrac{a_n x_n}{1 + c_n^{(11)} x_n + c^{(12)} y_n} \\ \dfrac{b_n y_n}{1 + c^{(21)} x_n + c_n^{(22)} y_n} \end{bmatrix} \preccurlyeq_{se} U_n,$$

where

$$L_n = \begin{bmatrix} \dfrac{\left(a - \varepsilon_{1,1}\right) x_n}{1 + \left(c^{(11)} + \varepsilon_{1,2}\right) x_n + \left(c^{(12)} + \varepsilon_{1,3}\right) y_n} \\ \dfrac{\left(b + \varepsilon_{2,1}\right) y_n}{1 + \left(c^{(21)} - \varepsilon_{2,2}\right) x_n + \left(c^{(22)} - \varepsilon_{2,3}\right) y_n} \end{bmatrix},$$

$$U_n = \begin{bmatrix} \dfrac{\left(a + \varepsilon_{1,1}\right) x_n}{1 + \left(c^{(11)} - \varepsilon_{1,2}\right) x_n + \left(c^{(12)} - \varepsilon_{1,3}\right) y_n} \\ \dfrac{\left(b - \varepsilon_{2,1}\right) y_n}{1 + \left(c^{(21)} + \varepsilon_{2,2}\right) x_n + \left(c^{(22)} + \varepsilon_{2,3}\right) y_n} \end{bmatrix},$$

and

$$\begin{aligned} a-\varepsilon_{1,1} < a_n < a+\varepsilon_{1,1},\\ c^{(11)}-\varepsilon_{1,2} < c_n^{(11)} < c^{(11)}+\varepsilon_{1,2},\\ c^{(12)}-\varepsilon_{1,3} < c_n^{(12)} < c^{(12)}+\varepsilon_{1,3},\\ b-\varepsilon_{2,1} < b_n < b+\varepsilon_{2,1},\\ c^{(21)}-\varepsilon_{2,2} < c_n^{(21)} < c^{(21)}+\varepsilon_{2,2},\\ c^{(22)}-\varepsilon_{2,3} < c_n^{(22)} < c^{(22)}+\varepsilon_{2,3}, \end{aligned}$$

for $n \geq N$.

By Theorem 2.1 the non-autonomous system (2.8) is asymptotic to the limiting system

$$\begin{aligned} x_{n+1} &= \frac{ax_n}{1+c^{(11)}x_n+c^{(12)}y_n},\\ y_{n+1} &= \frac{by_n}{1+c^{(21)}x_n+c^{(22)}y_n}, \end{aligned} \qquad (n=0,1,2,...). \tag{2.9}$$

Note that the system (2.9) has four equilibrium points (see [6], [16]):

$$E_0=(0,0)\,,\ E_x=\left(\frac{a-1}{c^{(11)}},0\right),\ E_y=\left(0,\frac{b-1}{c^{(22)}}\right),$$

and

$$E_+=\left(\frac{(a-1)\,c^{(22)}-(b-1)\,c^{(12)}}{c^{(11)}c^{(22)}-c^{(21)}c^{(12)}},\frac{(b-1)\,c^{(11)}-(a-1)\,c^{(21)}}{c^{(11)}c^{(22)}-c^{(21)}c^{(12)}}\right).$$

Based on the results in [16], Theorem 4.4 in [9], Remark 2, and using Theorem 2.1, we obtain the following result on the stability of the model (2.8).

Corollary 2.1. *For the non-autonomous Lesli-Gower model (2.8) the following statements are true:*

(i) *If $0<a<1$ and $0<b<1$, then all solutions of the system (2.8) converge to E_0, for all points (x_0,y_0) in the interior of $\mathbb{R}^2_+$; more precisely, E_0 is globally asymptotically stable in $\mathbb{R}^2_+$.*

(ii) *If $c^{(12)}-c^{(22)}>0$ and $c^{(21)}-c^{(11)}>0$, then all solutions of the system (2.8) converge to E_x, for all points (x_0,y_0) in the interior of $\mathbb{R}^2_+$.*

(iii) *If $c^{(12)}-c^{(22)}<0$ and $c^{(21)}-c^{(11)}<0$, then all solutions of the system (2.8) converge to E_+, for all points (x_0,y_0) in the interior of $\mathbb{R}^2_+$.*

(iv) *If $c^{(12)}-c^{(22)}>0$ and $c^{(21)}-c^{(11)}<0$, then all solutions of the system (2.8) converge to E_y, for all points (x_0,y_0) in the interior of $\mathbb{R}^2_+$.*

In [1], the following competitive system of difference equations

$$x_{n+1}=\frac{x_n}{a+y_n^2},\ \ y_{n+1}=\frac{y_n}{b+x_n^2},\ \ n=0,1,..., \tag{2.10}$$

was considered, where the parameters a and b are positive numbers, and initial conditions x_0 and y_0 are arbitrary non-negative numbers. Using linearized theory and sequence theory, it was proved that the zero equilibrium $E_0=(0,0)$ is globally asymptotically stable. Here, we will prove it using the method of Lyapunov functions, taking that

$V : \mathbb{R}^2_+ \to \mathbb{R}$ of the form $V\left(\left[\begin{array}{c} x \\ y \end{array}\right]\right) = x^2 + y^2$ of the map F associated with the system (2.10). Namely, if $x \geq 0$, $y \geq 0$, $(x, y) \neq (0, 0)$, $0 < a < 1$, and $0 < b < 1$, we have that

$$\begin{aligned} \Delta V &= V\left(F\left(\left[\begin{array}{c} x \\ y \end{array}\right]\right)\right) - V\left(\left[\begin{array}{c} x \\ y \end{array}\right]\right) = \left(x\frac{1}{a+y^2}\right)^2 + \left(y\frac{1}{b+x^2}\right)^2 - x^2 - y^2 \\ &= x^2\left(\left(\frac{1}{a+y^2}\right)^2 - 1\right) + y^2\left(\left(\frac{1}{b+x^2}\right)^2 - 1\right) \\ &\leq x^2\left(\frac{1}{a^2} - 1\right) + y^2\left(\frac{1}{b^2} - 1\right) < 0. \end{aligned}$$

Since $V\left(\left[\begin{array}{c} x \\ y \end{array}\right]\right) = x^2 + y^2 \to \infty$, as $\left\|\left[\begin{array}{c} x \\ y \end{array}\right]\right\| \to \infty$ the equilibrium point $E_0 = (0, 0)$ is globally asymptotically stable when $0 < a < 1$ and $0 < b < 1$.

If we consider the following non-autonomous system

$$x_{n+1} = \frac{x_n}{a_n + y_n^2}, \quad y_{n+1} = \frac{y_n}{b_n + x_n^2}, \quad n = 0, 1, \ldots, \tag{2.11}$$

where $\lim_{n\to\infty} a_n = a$ and $\lim_{n\to\infty} b_n = b$, then, by using Theorem 2.1, for which the system (2.10) is a limiting system, we obtain the following result.

Corollary 2.2. *All solutions of the system (2.10) globally asymptotically converge to* $E_0 = (0, 0)$ *for* $0 < a < 1$ *and* $0 < b < 1$*, and for all* $x_0 \geq 0$ *and* $y_0 \geq 0$*.*

Now, consider the following autonomous competitive system of difference equations:

$$\begin{aligned} x_{n+1} &= a x_n e^{-\alpha y_n}, \\ y_{n+1} &= b y_n e^{-\beta x_n}, \end{aligned} \quad (n = 0, 1, 2, \ldots). \tag{2.12}$$

The equilibrium points $(\overline{x}, \overline{y})$ of the system (2.12) satisfy the following system of algebraic equations:

$$\begin{aligned} \overline{x} &= a\overline{x}e^{-\alpha\overline{y}}, \\ \overline{y} &= a\overline{y}e^{-\beta\overline{x}}. \end{aligned}$$

It is easy to see that the system (2.12) has the equilibrium $E_0 = (0, 0)$ for all values of the parameters. This equilibrium point is unique if $0 < a < 1$ and $0 < b < 1$. For $a > 1$, $b > 1$, $\alpha > 0$ and $\beta > 0$ the system (2.12) has a positive equilibrium $E_+ = \left(\frac{\ln b}{\beta}, \frac{\ln a}{\alpha}\right)$. If $a = 1$, then there exist infinitely many equilibrium points $E_{\overline{x}} = (\overline{x}, 0)$, $\overline{x} \geq 0$, but if $b = 1$, then there exist infinitely many equilibrium points $E_{\overline{y}} = (\overline{y}, 0)$, $\overline{y} \geq 0$.

The map associated with the system (2.12) has the following form:

$$T\left(\begin{array}{c} x \\ y \end{array}\right) = \left(\begin{array}{c} axe^{-\alpha y} \\ bye^{-\beta x} \end{array}\right). \tag{2.13}$$

Based on the Jacobian matrix associated with the map (2.13),

$$\begin{pmatrix} ae^{-\alpha y} & -a\alpha x e^{-\alpha y} \\ -b\beta y e^{-\beta x} & be^{-\beta x} \end{pmatrix},$$

we obtained the following result about the local stability of the equilibrium point E_0.

Lemma 2.2. *The following statements hold for the equilibrium point E_0:*

(a) If $0 < a < 1$ and $0 < b < 1$, then E_0 is globally asymptotically stable.
(b) If $a = 1$ or $b = 1$, then E_0 is a non-hyperbolic.
(c) If $a > 1$ or $b > 1$, then E_0 is unstable (a saddle point or a repeller).

Proof. The Jacobian of the map T at the equilibrium $E_0 = (0,0)$ is of the following form

$$J_T(0,0) = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}.$$

The eigenvalues of the Jacobian at the equilibrium $E_0 = (0,0)$ are $\lambda_1 = a$ and $\lambda_2 = b$, which implies that $E_0 = (0,0)$ is locally asymptotically stable for $0 < a < 1$ and $0 < b < 1$, but is unstable (a saddle point or a repeller) if $a > 1$ or $b > 1$ and a non-hyperbolic point for $a = 1$ or $b = 1$.

If $0 < a < 1$ and $0 < b < 1$, then the first equation of the system (2.13) implies that $x_{n+1} < ax_n < a^{n+1}x_0$, which means that $x_n \to 0$ as $n \to \infty$ (since $x_n \geq 0$ for all $n = 0, 1, \ldots$) . From the second equation of the system (2.13), we have that $y_{n+1} < bx_n$, which implies that $y_n \to 0$ as $n \to +\infty$ (since $y_n \geq 0$ for all $n = 0, 1, \ldots$), that is, $E_0 = (0,0)$ is a global attractor. Since $E_0 = (0,0)$ is locally asymptotically stable, we conclude it is globally asymptotically stable. □

Remark 2.2. By using the Jacobian matrix associated with the map (2.13), we have that the following statements are true:

1. If $a = 1$, then every equilibrium point $E_{\overline{x}}$, $\overline{x} \geq 0$ is non-hyperbolic.
2. If $b = 1$, then every equilibrium point $E_{\overline{y}}$, $\overline{y} \geq 0$ is non-hyperbolic.
3. If $a > 1$ and $b > 1$, then E_+ is unstable (a saddle point or a repeller).

Note that the system (2.13) is a limiting system of the following non-autonomous competitive system:

$$\begin{aligned} x_{n+1} &= a_n x_n e^{-\alpha_n y_n}, \\ y_{n+1} &= b_n y_n e^{-\beta_n x_n}, \end{aligned} \quad (n = 0, 1, 2, \ldots), \tag{2.14}$$

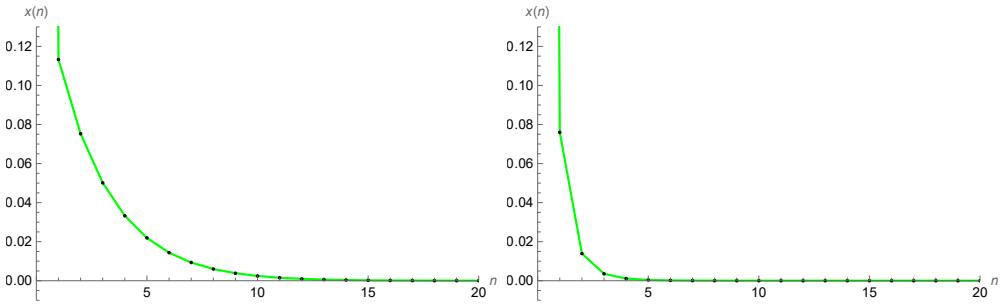
where $\lim_{n\to\infty} a_n = a$, $\lim_{n\to\infty} b_n = b$, $\lim_{n\to\infty} \alpha_n = \alpha$ and $\lim_{n\to\infty} \beta_n = \beta$.

We obtain the following result using Lemma 2.2 and Theorem 2.1.

Corollary 2.3. *If $0 < a < 1$, $0 < b < 1$, $\alpha > 0$ and $\beta > 0$, then all solutions of the system (2.14) globally asymptotically converge to $E_0 = (0,0)$ for all $x_0 \geq 0$ and $y_0 \geq 0$.*

Example 2.1. *It seems interesting to compare the behavior of the autonomous system solutions (2.10) for $a = 0.9$ and $b = 0.99$:*

$$x_{n+1} = \frac{x_n}{0.9 + y_n^2}, \quad y_{n+1} = \frac{y_n}{0.99 + x_n^2}, \quad n = 0, 1, \ldots, \tag{2.15}$$



(A) (B)

FIGURE 1. Time series of the components x_n of the systems in Example 2.1: (A) autonomous case; (B) non-autonomous case (with initial values $x_0 = 2.1, y_0 = 4.2$).

with the solutions of the corresponding non-autonomous system (2.11) with the coefficients $a_n = \frac{0.9n+10}{n+1}$ *and* $b_n = 0.99 + \frac{1}{n}$, *that is:*

$$x_{n+1} = \frac{x_n}{\frac{0.9n+10}{n+1} + y_n^2}, \quad y_{n+1} = \frac{y_n}{0.99 + \frac{1}{n} + x_n^2}, \quad n = 0, 1, \ldots . \tag{2.16}$$

What is unexpected in this case is the faster convergence of the solution of the non-autonomous system compared to the autonomous system, especially of the components x_n. *In both cases, the components* y_n *converge to* 0 *quickly (Figure 1).*

3. CONCLUSION

Relying on previous research, where theorems of global attractiveness of some classes of non-autonomous competitive systems of difference equations with asymptotically constant coefficients were proved, this paper presents a general theorem for an arbitrary non-autonomous competitive system. The obtained results were applied to three typical cases. In the end, the rate of convergence of the solution of a non-autonomous competitive system of difference equations was compared with the convergence of the solution of the corresponding limiting autonomous system. In doing so, the unexpected conclusion was reached that the solutions of a non-autonomous competitive system can converge to the equilibrium point even faster than the solutions of its limiting autonomous system.

REFERENCES

[1] Dž. Burgić, M.R.S. Kulenović and M. Nurkanović, Global Dynamics of a Rational System of Difference Equations in the Plane, *Communications on Applied Nonlinear Analysis*, **15** (1) (2008), 71-84.

[2] M. Garić-Demirović, M.R.S. Kulenović and M.Nurkanović, Global Behavior of Four Competitive Rational Systems of Difference Equations in the Plane, *Discrete Dynamics in Nature and Society*, Volume 2009, Article ID 153058, 34 pages.

[3] M. Garić-Demirović, M.R.S. Kulenović and M. Nurkanović, Global Behavior of Two Competitive Rational Systems of Difference Equations in the Plane, *Communications on Applied Nonlinear Analysis*, 16 (2009), No. 3, 1-18.

[4] M.R.S. Kulenović and E. Bertrand, Global Dynamic Scenarios for Competitive Maps in the Plane, *Advances in Difference Equations*, (2018), 2018: 28p.

[5] M.R.S. Kulenović, J. Marcotte, and O. Merino, Properties of Basins of Attraction for Planar Discrete Cooperative Maps, *Discrete Contin. Dyn. Syst. B*, 26 (2021), 2721–2737.
[6] M.R.S. Kulenović and D.T. McArdle, Global Dynamics of Leslie-Gower Competitive Systems in the Plane, *Mathematics,* **7** (1), 76 (2019). https://doi.org/10.3390/math7010076
[7] M.R.S. Kulenović and O. Merino, Invariant Curves for Planar Competitive and Cooperative Maps, *Journal of Difference Equations and Applications*, 24 (2018) 898-915.
[8] M.R.S. Kulenović, O. Merino, and M. Nurkanović, Global Dynamics of Certain Competitive Systems in the Plane, *Journal of Difference Equations and Applications*, Vol. 18, No. 12 (2012), 1951-1966.
[9] M.R.S. Kulenović, M. Nurkanović, Z. Nurkanović, and S. Trolle, Asymptotic Behavior of Certain Non-autonomous Planar Competitive Systems of Difference Equations, *Mathematics* (MDPI), 11, 3909 (2023), 22 p. https://doi.org/10.3390/math11183909
[10] M.R.S. Kulenović, M. Nurkanović, Z. Nurkanović, and S. Trolle, Stability of certain non-autonomous cooperative systems of difference equations with the application to evolutionary dynamics, (2024) (will appear).
[11] M.R.S. Kulenović and M. Nurkanović, Global asymptotical behavior of a two dimensional system of difference equations modeling cooperation, *Journal of Difference Equations and Applications*, 9 (2003), 149-159.
[12] M.R.S. Kulenović and M. Nurkanović, Asymptotical behavior of a system of linear fractional difference equations, *Journal of Inequalities and Applications*, Vol. 2005, No.2 (2005), 127-143.
[13] M.R.S. Kulenović and M. Nurkanović, Asymptotic Behavior of a Competitive System of Linear Fractional Difference Equations, *Advances in Difference Equations*, Volume 2006, Article ID 19756, Pages 1-13.
[14] M.R.S. Kulenović and M. Nurkanović, Global Behavior of a Two-dimensional Competitive System of Difference Equations with Stocking, *Mathematical and Computer Modelling,* 55 (2012), 1998-2011.
[15] M.R.S. Kulenović and S. Van Beaver, Global Dynamics of a Cooperative System with Ceiling Density Dependence, *International J. Difference Equations*, 14 (2019), 59-74.
[16] K. Mokni, S. Elaydi, M. CH-Chaui, and A. Eladdadi, Discrete evolutionary population models: a new approach, *Journal of Biological Dynamics*, **14** (1), (2020), 454-478. https://doi.org/10.1080/17513758.2020.1772997

(Received: May 14, 2024)
(Revised: August 20, 2024)

Mehmed Nurkanović
University of Tuzla
Department of Mathematics
U. Vejzagića 4, 75000 Tuzla
Bosnia and Herzegovina
email: *mehmed.nurkanovic@untz.ba*

PROCEEDINGS OF THE CONFERENCE ON MARCH 14 – INTERNATIONAL DAY OF MATHEMATICS
Special Editions ANUBiH, Book CCXVI, OPMN Book 30
Sarajevo 2024, **p. 48–62** DOI: 10.5644/PI2024.216.07

SOLVING FIRST-ORDER AND SECOND-ORDER DIFFERENCE EQUATIONS USING LIE SYMMETRIES

MEHMED NURKANOVIĆ AND MIRSAD TRUMIĆ

Dedicated to the 75th birthday of our dear Professor Mirjana Vuković

ABSTRACT. There are well-developed algorithms for solving certain nonlinear difference equations with constant coefficients. On the other hand, nonlinear difference equations, especially with variable coefficients, are very complex. Namely, there are no universal methods of solving them. Nevertheless, difference equation methods, especially Lie symmetry groups, have been successfully used for certain classes of these equations. Using Lie symmetries, it is possible to construct the characteristics of a given equation. Then, with the help of canonical coordinates, it is possible to successfully solve some linear and non-linear difference equations of the first and second order with variable coefficients. The method of reducing the order of nonlinear difference equations can also, with certain specificities, be used successfully when solving difference equations.

The mentioned methods are illustrated in several characteristic examples.

1. INTRODUCTION AND PRELIMINARIES

It is well known that Lie symmetries can sometimes be successfully used to solve differential equations. In this paper, we will show that this is also possible in the case of difference equations by analogy with differential equations. For this reason, we will first familiarize ourselves with the essential characteristics of Lie symmetries, [1, 2].

The symmetry of a geometric object is an invertible transformation that maps the object into itself. The set of all symmetries $\mathcal{G}$ of a geometric object is a group. Symmetries $\Gamma_1, ..., \Gamma_k$ are generators of the group $\mathcal{G}$ if each symmetry can be written as a product of some symmetries Γ_i and their inversions. A differential or difference equation transformation is a symmetry if every solution of the transformed equation is also a solution of the initial equation and vice versa.

We will only consider translations, reflections, and rotations (each of which is rigid). Let us look at scaling transformations

$$\Gamma_\varepsilon : u_n \longmapsto \widehat{u_n} = e^{\varepsilon} u_n \tag{1.1}$$

2020 *Mathematics Subject Classification.* 39A05, 39A06.

Key words and phrases. difference equation, characteristic, compatible canonical coordinate, Lie point symmetries.

to a scalar linear homogeneous difference equation of order p. If $U_1(n), \ldots, U_p(n)$ are linearly independent solutions, then the general solution is given by

$$u_n = \sum_{i=1}^{p} c_i U_i(n).$$

Scaling (1) maps this solution into

$$\widehat{u_n} = \sum_{i=1}^{p} \widehat{c_i} U_i(n) \quad (\widehat{c_i} = e^{\varepsilon} c_i),$$

so that the set of all solutions is mapped (invertible) into itself; thus, Γ_ε is a symmetry of a difference equation for each $\varepsilon \in \mathbb{R}$.

Here, $\widehat{u_n}$ is a smooth function of u_n. Really, $\Gamma : \mathbb{R} \to \mathbb{R}$ is a diffeomorphism, a smooth invertible map whose inverse is also smooth. The set of transformations $G = \{\Gamma_\varepsilon : \varepsilon \in \mathbb{R}\}$ is a group with a composition $\Gamma_\delta \Gamma_\varepsilon = \Gamma_{\delta+\varepsilon}$, for all $\delta, \varepsilon \in \mathbb{R}$. Here, Γ_0 is the identical map, and $\Gamma_\varepsilon^{-1} = \Gamma_{-\varepsilon}$ holds. In addition, $\widehat{u_n}$ is an analytic function of the parameter ε. An important feature of this group is: Γ_ε is close identity transformation for every small enough ε. Suppose these close identities of the symmetry transformation are given by different equations of the pth order. In that case, the individual solution will be mapped into a one-parameter family of close solutions whose arbitrary constants depend analytically on ε. This property can be used to solve various first-order equations, which need not be linear, as will be demonstrated later.

Definition 1.1. *A parameterized set of transformations by points*

$$\Gamma_\varepsilon : X \longmapsto \widehat{X}(X;\varepsilon), \quad \varepsilon \in (\varepsilon_0, \varepsilon_1), \quad \varepsilon_0 < 0, \ \varepsilon_1 > 0,$$

is called a one-parameter local Lie group if the following conditions apply:

1. Γ_0 *is an identical map, so* $\widehat{X} = X$*, for* $\varepsilon = 0$,

2. $\Gamma_\delta \Gamma_\varepsilon = \Gamma_{\delta+\varepsilon}$ *for every* δ, ε *close enough to zero,*

3. *Each* $\widehat{x}^\alpha$ *can be represented as a Taylor series in* ε *(about* $\varepsilon = 0$ *which is determined by* X*)*

$$\widehat{x}^\alpha(X;\varepsilon) = x^\alpha + \varepsilon \xi^\alpha(X) + O\left(\varepsilon^2\right), \quad \alpha = 1, \ldots, N.$$

It follows from conditions **1.** and **2.** that $\Gamma_\varepsilon^{-1} = \Gamma_{-\varepsilon}$ when $|\varepsilon|$ is small enough. Despite its name, a local Lie group does not need to be a group; it is only necessary to satisfy the axioms of the group for small enough parameter values.
In general, the local one-parameter Lie group of symmetries of a given scalar difference equation will depend on both n and the variable u_n [1].

Example 1.1. *The general solution of the difference equation*

$$u_{n+1} = \frac{n+1}{n} u_n, \quad n \geq 1, \tag{1.2}$$

is $u_n = c_1 n$*. Any transformation of the form*

$$(\widehat{n}, \widehat{u_n}) = (n, u_n + \varepsilon n) \tag{1.3}$$

is the symmetry that maps $u_n = c_1 n$ *into* $\widehat{u_n} = (c_1 + \varepsilon) n$*. For each* $n \geq 1$*, (1.3) defines a one-parameter local Lie group of translations.*

2. Characteristics and canonical coordinates

The following considerations will be limited to Lie symmetries for which $\widehat{u_n}$ depends only on n and u_n. These are the so-called Lie point symmetries that are of the form

$$\widehat{n} = n, \quad \widehat{u_n} = u_n + \varepsilon \mathcal{K}(n, u_n) + O\left(\varepsilon^2\right). \tag{2.1}$$

To see how such symmetries transform the shifted variable u_{n+k}, we simply replace n with $n+k$ in (2.2):

$$\widehat{u_{n+k}} = u_{n+k} + \varepsilon \mathcal{K}(n+k, u_{n+k}) + O\left(\varepsilon^2\right). \tag{2.2}$$

The formula (2.2) represents the *formula prolongations* for pointwise Lie symmetries.

The function $\mathcal{K}(n, u_n)$ is the *characteristic* of the local Lie group with respect to the coordinates (n, u_n). For example, the characteristic of vertical translation $(\widehat{n}, \widehat{u_n}) = (n, u_n + \varepsilon)$ is of the form

$$\mathcal{K}(n, u_n) = 1. \tag{2.3}$$

Let us observe the effect of changing coordinates from (n, u_n) in (n, v_n), where $v'(n, u_n) \neq 0$ $\left(v' = \dfrac{\partial v}{\partial u_n}\right)$. When (2.2) is a symmetry for every ε close enough to zero, we can apply Taylor's theorem to get

$$\begin{aligned} \widehat{v_n} &= v\left(n, \widehat{u_n}\right) = v\left(n, u_n + \varepsilon \mathcal{K}(n, u_n) + O\left(\varepsilon^2\right)\right) \\ &= v_n + \varepsilon v'(n, u_n) \mathcal{K}(n, u_n) + O\left(\varepsilon^2\right) \end{aligned} \tag{2.4}$$

Therefore, the characteristic with respect to (n, v_n) is equal to $\widetilde{\mathcal{K}}(n, v_n)$, where

$$\widetilde{\mathcal{K}}(n, v(n, u_n)) = v'(n, u_n) \mathcal{K}(n, u_n). \tag{2.5}$$

The values of $\widetilde{\mathcal{K}}$ and $\mathcal{K}$ will differ at most points (n, u_n), where $v'(n, u_n) \neq 1$, including only points where $\mathcal{K}(n, u_n) = 0$. When $\mathcal{K}(n, u_n) \neq 0$, then it is especially useful to introduce the *canonical coordinate* s_n, so that the translation symmetries of s_n are:

$$(\widehat{n}, \widehat{s}) = (n, s_n + \varepsilon), \quad \varepsilon \in \mathbb{R}. \tag{2.6}$$

The characteristic in relation to (n, s_n) is $\widetilde{\mathcal{K}}(n, s_n) = 1$, so due to (2.5)

$$s(n, u_n) = \int \frac{du_n}{\mathcal{K}(n, u_n)}. \tag{2.7}$$

For each n, the possible values of u_n lie on the real line, which is (typically) divided into intervals where we have omitted each value u_n for which $\mathcal{K}(n, u_n) = 0$. The equality (2.7) defines the canonical coordinate s (locally) on each interval, but it is to be expected that different coordinates correspond to different intervals. For example, if $\mathcal{K}(n, u_n) = u_n^2 - 1$, the appropriate real-valued canonical coordinate depends on u_n, as follows:

$$s(n, u_n) = \int \frac{du_n}{u_n^2 - 1} = \begin{cases} \frac{1}{2} \ln \frac{u_n - 1}{u_n + 1}, & |u_n| > 1 \\ \frac{1}{2} \ln \frac{1 - u_n}{1 + u_n}, & |u_n| < 1. \end{cases} \tag{2.8}$$

In this case $s\left(n, \frac{1}{u_n}\right) = s(n, u_n)$ for every non-zero u_n so that the map from u_n in s is not injective, which cannot be invertible unless it is predetermined that $|u_n| \geq 1$. The most significant benefit of canonical coordinates is that they simplify or even solve the given difference equation. The idea is to write the difference equation in a simpler form for s; if a more straightforward difference equation can be solved, all that remains is to write the solution over the original variables. To use this approach, one must be able to invert the map from u_n to s (at least for all points (n, u_n) that occur in any solution of the original differential equation and satisfy $\mathcal{K}(n, u_n) \neq 0$). Any coordinate s that meets this requirement will be called *compatible* with the given difference equation [1, 2, 7].

For any compatible canonical coordinate we can replace n with $n + k$, to obtain

$$s_{n+k} = s(n + k, u_{n+k}) = E^k s, \quad k \in \mathbb{Z}. \tag{2.9}$$

According to the prolongation formula, Lie symmetry $\widehat{s} = s + \varepsilon$ prolongs into

$$\widehat{s_{n+k}} = s_{n+k} + \varepsilon. \tag{2.10}$$

3. SOLVING FIRST-ORDER DIFFERENCE EQUATIONS USING LIE SYMMETRIES

Consider the following first-order difference equation

$$u_{n+1} = w(n, u_n). \tag{3.1}$$

To map the set of solutions of (3.1) into itself, the following symmetry condition must be satisfied

$$\widehat{u_{n+1}} = w(\widehat{n}, \widehat{u_n}) \quad \text{when} \quad u_{n+1} = w(n, u_n). \tag{3.2}$$

Example 3.1. *([1], Problem 2.2) Find the characteristic $\mathcal{K}(n, u_n)$ for the difference equation*

$$u_{n+1} = \frac{nu_n - 1}{u_n + n}, \tag{3.3}$$

and then solve this equation.

Solution. Equation (3.1) is the well-known Riccati equation [4–6]. Starting from the formula

$$\mathcal{K}(n + 1, u_{n+1}) = w'(n, u_n)\mathcal{K}(n, u_n)$$

where

$$w'(n, u_n) = \frac{d}{du_n}\left(\frac{nu_n - 1}{u_n + n}\right) = \frac{nu_n + n^2 - nu_n + 1}{(u_n + n)^2} = \frac{n^2 + 1}{(u_n + n)^2},$$

then we have

$$\mathcal{K}(n + 1, u_{n+1}) = \frac{n^2 + 1}{(u_n + n)^2}\mathcal{K}(n, u_n). \tag{3.4}$$

We look for the characteristic $\mathcal{K}(n, u_n)$ in the form

$$\mathcal{K}(n, u_n) = \alpha_n u_n^2 + \beta_n u_n + \gamma_n,$$

where $\alpha_n, \beta_n, \gamma_n$ are the coefficients which should be determined from (3.4). Thus, we have

$$\alpha_{n+1}u_{n+1}^2+\beta_{n+1}u_{n+1}+\gamma_{n+1}=\frac{n^2+1}{(u_n+n)^2}\left(\alpha_n u_n^2+\beta_n u_n+\gamma_n\right),$$

that is

$$\alpha_{n+1}\left(\frac{nu_n-1}{u_n+n}\right)^2+\beta_{n+1}\frac{nu_n-1}{u_n+n}+\gamma_{n+1}=\frac{n^2+1}{(u_n+n)^2}\left(\alpha_n u_n^2+\beta_n u_n+\gamma_n\right),$$

or

$$\alpha_{n+1}\left(nu_n-1\right)^2+\beta_{n+1}\left(nu_n-1\right)\left(u_n+n\right)+\gamma_{n+1}\left(u_n+n\right)^2=\left(n^2+1\right)\left(\alpha_n u_n^2+\beta_n u_n+\gamma_n\right).$$

By equalizing the coefficients that are found with u_n^2, u_n, and free members, we get a system

$$\begin{aligned}
n^2\alpha_{n+1}+n\beta_{n+1}+\gamma_{n+1}&=\left(n^2+1\right)\alpha_n\\
-2n\alpha_{n+1}+\left(n^2-1\right)\beta_{n+1}+2n\gamma_{n+1}&=\left(n^2+1\right)\beta_n\\
\alpha_{n+1}-n\beta_{n+1}+n^2\gamma_{n+1}&=\left(n^2+1\right)\gamma_n.
\end{aligned}$$

By adding the first and third equations of the last system, we have

$$\alpha_{n+1}+\gamma_{n+1}=\alpha_n+\gamma_n=c_1. \tag{3.5}$$

Subtracting from the first equation of the system the second equation multiplied by n, and then subtracting the third, also multiplied by n, we obtain

$$n\left(n^2+1\right)\alpha_{n+1}+\left(n^2+1\right)\beta_{n+1}-n\left(n^2+1\right)\gamma_{n+1}=\left(n^2+1\right)\left(n\alpha_n-\beta_n-n\gamma_n\right),$$

i.e.,

$$n\left(\alpha_{n+1}-\gamma_{n+1}\right)-n\left(\alpha_n-\gamma_n\right)=-\left(\beta_{n+1}+\beta_n\right). \tag{3.6}$$

By replacing (3.5) in (3.6), we get

$$2n\left(\alpha_{n+1}-\alpha_n\right)=-\left(\beta_{n+1}+\beta_n\right). \tag{3.7}$$

If we include (3.5) in the first equation of the system, we have

$$n^2\left(\alpha_{n+1}-\alpha_n\right)+n\beta_{n+1}+c_1-\alpha_{n+1}=\alpha_n,$$

and now, considering (3.7),

$$-n^2\frac{1}{2n}\left(\beta_{n+1}+\beta_n\right)+n\beta_{n+1}+c_1-\alpha_{n+1}=\alpha_n,$$

from which

$$\beta_{n+1}-\beta_n=\frac{2}{n}\left(\alpha_{n+1}-\alpha_n-c_1\right). \tag{3.8}$$

Adding (3.7) and (3.8) gives

$$-\frac{n^2+1}{n}\alpha_{n+1}+\frac{n^2-1}{n}\alpha_n+\frac{1}{n}c_1=\beta_n. \tag{3.9}$$

Finally, by substituting (3.9) into the first equation of the system, we obtain

$$n^2\alpha_{n+1}+n\left(-\frac{(n+1)^2+1}{n+1}\alpha_{n+2}+\frac{(n+1)^2-1}{n+1}\alpha_{n+1}+\frac{1}{n}c_1\right)+$$
$$+c_1-\alpha_{n+1}=\left(n^2+1\right)\alpha_n.$$

After arranging the last expression, we get the following inhomogeneous linear difference equation

$$n\left(n^2+2n+2\right)\alpha_{n+2}-\left(2n^3+3n^2-n-1\right)\alpha_{n+1}+(n+1)\left(n^2+1\right)\alpha_n=(2n+1)\,c_1.$$

One particular solution to this equation is $\alpha_n=\frac{c_1}{2}$. Namely, if we assume that $\alpha_n=(An+B)\,c_1$, the particular solution will be obtained. It implies that $\gamma_n=\frac{c_1}{2}$, and that $\beta_n=0$ must hold (considering all three equations of the given system). Therefore, the characteristic is of the form

$$\mathcal{K}(n,u_n)=\frac{c_1}{2}\left(u_n^2+1\right).$$

If $c_1=2$, then we have

$$s_n=\int\frac{1}{\mathcal{K}(n,u_n)}du_n=\int\frac{1}{u_n^2+1}du_n=\arctan u_n. \tag{3.10}$$

Thus,

$$s_{n+1}-s_n=\arctan u_{n+1}-\arctan u_n=\arctan\frac{u_{n+1}-u_n}{1+u_{n+1}u_n}$$
$$=\arctan\frac{\frac{nu_n-1}{u_n+n}-u_n}{1+\frac{nu_n^2-u_n}{u_n+n}}=\arctan\left(-\frac{1}{n}\right)=-\arctan\frac{1}{n}.$$

from which

$$s_n=s_1-\sum_{i=1}^{n-1}\arctan\frac{1}{i}. \tag{3.11}$$

Using the property for the sum of the function arctan, we have

$$\sum_{i=1}^{n-1}\arctan\frac{1}{i}=\arctan A\,(n)\,.$$

From (3.11), due to (3.10), it follows that

$$\arctan u_n=\arctan u_1-\arctan A\,(n)=\arctan\frac{u_1-A\,(n)}{1+u_1A\,(n)},$$

that is

$$u_n=\frac{u_1-A\,(n)}{1+u_1A\,(n)}\quad(n=2,3,\ldots)\,. \tag{3.12}$$

Let us check for the first few members using the iteration procedure and then the formula (3.12).

$$\arctan\frac{1}{1}=\arctan A(1)\Longrightarrow A(1)=1$$

$$\arctan\frac{1}{1}+\arctan\frac{1}{2}=\arctan A(2)\Longrightarrow\arctan A(2)=\arctan\frac{1+\frac{1}{2}}{1-\frac{1}{2}}=\arctan 3$$

$$\Longrightarrow A(2)=3$$

$$\arctan\frac{1}{1}+\arctan\frac{1}{2}+\arctan\frac{1}{3}=\arctan A(3)\Longrightarrow A(3)=\frac{3+\frac{1}{3}}{1-3\cdot\frac{1}{3}}=\infty.$$

a) Iterative

$$n=1\Longrightarrow u_2=\frac{u_1-1}{u_1+1}$$

$$n=2\Longrightarrow u_3=\frac{2u_2-1}{u_2+2}\,\frac{\frac{2u_1-2}{u_1+1}-1}{\frac{u_1-1}{u_1+1}+2}=\frac{u_1-3}{3u_1+1}$$

$$n=3\Longrightarrow u_4=\frac{3u_3-1}{u_3+3}\,\frac{\frac{3u_1-9}{3u_1+1}-1}{\frac{u_1-3}{3u_1+1}+3}=-\frac{1}{u_1}.$$

b) According to the formula (3.12), it follows that

$$n=1\Longrightarrow u_2=\frac{u_1-1}{u_1+1}$$

$$n=2\Longrightarrow u_3=\frac{u_1-A(2)}{1+u_1A(2)}=\frac{u_1-3}{1+3u_1}$$

$$n=3\Longrightarrow u_4=\frac{u_1-A(3)}{1+u_1A(3)}=\frac{\frac{u_1}{A(3)}-1}{\frac{1}{A(3)}+u_1}=-\frac{1}{u_1}.$$

So, the formula (3.12) really gives the solution of the considered equation.

4. Solving second-order difference equations using Lie symmetries

Consider the following second-order difference equation

$$u_{n+2}=w(n,u_{n+1},u_n). \tag{4.1}$$

The so-called LSC condition for the difference equation (4.1) is of the form

$$\mathcal{K}(n+2,w)-D_2w\mathcal{K}(n+1,u_{n+1})-D_1w\mathcal{K}(n,u_n)=0. \tag{4.2}$$

The first term in LSC (4.2), with assumptions $D_1w\neq 0$, $D_2w\neq 0$, is eliminated by applying the following differential operator

$$\left(\frac{1}{D_1w}\right)\frac{\partial}{\partial u_n}-\left(\frac{1}{D_2w}\right)\frac{\partial}{\partial u_{n+1}}, \tag{4.3}$$

which gives

$$\mathcal{K}'(n+1,u_{n+1})+D_2\eta\mathcal{K}(n+1,u_{n+1})-\mathcal{K}'(n,u_n)+D_1\eta\mathcal{K}(n,u_n)=0, \tag{4.4}$$

where $\eta(n,u_n,u_{n+1}) = \ln\left|\dfrac{D_2 w}{D_1 w}\right|$. Then $\mathcal{K}'(n+1,u_{n+1})$ can be eliminated by differentiating (4.4) with respect to u_n, which gives

$$D_{12}\eta\mathcal{K}(n+1,u_{n+1}) - \mathcal{K}''(n,u_n) + D_1\eta\mathcal{K}'(n,u_n) + D_{11}\eta\mathcal{K}(n,u_n) = 0. \tag{4.5}$$

At this stage, we have two possibilities.

First, if $D_{12}\eta = 0$, then (4.5) can be integrated to obtain

$$\mathcal{K}'(n,u_n) - D_1\eta\mathcal{K}(n,u_n) = \alpha(n). \tag{4.6}$$

Now, replacing (4.6) with (4.4) (to eliminate $\mathcal{K}'(n,u_n)$ and $\mathcal{K}'(n+1,u_{n+1})$) we get

$$(ED_1\eta + D_2\eta)\mathcal{K}(n+1,u_{n+1}) = \alpha(n) - \alpha(n+1). \tag{4.7}$$

If $D_2\eta = -ED_1\eta$, then from (4.7) it follows that $\alpha(n) = c_1$ and

$$\mathcal{K}(n,u_n) = \frac{\alpha(n-1) - \alpha(n)}{D_1\eta + E^{-1}D_2\eta}, \tag{4.8}$$

which can be replaced in (4.6), thus obtaining (at most) a difference equation of the first order in $\alpha(n)$. In the second case, the function $\mathcal{K}(n,u_n)$ which results from (4.4) must be replaced by (4.2), and any additional constraints this creates must be resolved.

Another possibility is to be $D_{12}\eta \neq 0$, when the equation (4.5) needs to be divided by $D_{12}\eta$ and then differentiated once more by u_n. The coefficients of the resulting difference equation may depend on u_{n+1}. If this happens, then the equation should be separated into a system of equations whose coefficients depend only on n and u_n. Then, continue the solving process as before.

Example 4.1. *([1], Ex. 2.17) Determine all characteristics of Lie point symmetries for the difference equation*

$$u_{n+2} = \frac{1}{u_{n+1}+u_n} - u_{n+1} - 2(-1)^n \quad (n \geq 0). \tag{4.9}$$

Solution. Here, we give a much more detailed solution than in [1].
The LSC for a given difference equation is of the form

$$\mathcal{K}\left(n+2, \frac{1}{u_{n+1}+u_n} - u_{n+1} - 2(-1)^n\right) + \left(1 + \frac{1}{(u_{n+1}+u_n)^2}\right)\mathcal{K}(n+1,u_{n+1}) + \\ + \frac{1}{(u_{n+1}+u_n)^2}\mathcal{K}(n,u_n) = 0, \tag{4.10}$$

because $D_1 w = -\dfrac{1}{(u_{n+1}+u_n)^2}$, $D_2 w = -\dfrac{1}{(u_{n+1}+u_n)^2} - 1$, so

$$\eta = \ln\left|\frac{D_2 w}{D_1 w}\right| = \ln\left((u_{n+1}+u_n)^2\right) + 1.$$

Since $D_{12}\eta \neq 0$, then the equation

$$D_{12}\eta\mathcal{K}(n+1,u_{n+1}) - \mathcal{K}''(n,u_n) + D_1\eta\mathcal{K}'(n,u_n) + D_{11}\eta\mathcal{K}(n,u_n) = 0, \tag{4.11}$$

should be divided by $D_{12}\eta = \dfrac{2\left(1-(u_{n+1}+u_n)^2\right)}{\left((u_{n+1}+u_n)^2+1\right)^2}$, and we get (due to $D_{11}\eta = D_{12}\eta$)

$$\mathcal{K}(n+1,u_{n+1}) + \frac{\left((u_{n+1}+u_n)^2+1\right)^2}{2\left((u_{n+1}+u_n)^2-1\right)}\mathcal{K}''(n,u_n) -$$

$$-\frac{(u_{n+1}+u_n)\left((u_{n+1}+u_n)^2+1\right)}{(u_{n+1}+u_n)-1}\mathcal{K}'(n,u_n) + \mathcal{K}(n,u_n) = 0. \tag{4.12}$$

After differentiation by u_n, $\mathcal{K}(n+1,u_{n+1})$ is lost, and we get

$$\left((u_{n+1}+u_n)^4-1\right)\mathcal{K}'''(n,u_n) - 4(u_{n+1}+u_n)\mathcal{K}''(n,u_n) + 4\mathcal{K}'(n,u_n). \tag{4.13}$$

All coefficients in the equation (4.13) depend on u_{n+1} so the equation can be decomposed into a system of differential equations, each of which can be multiplied by a special power of $(u_{n+1}+u_n)$

$$\mathcal{K}'''(n,u_n) = 0,\ \ \mathcal{K}''(n,u_n) = 0,\ \ \mathcal{K}'(n,u_n) = 0.$$

From here, $\mathcal{K}(n,u_n) = \alpha(n)$, so by replacing it in (4.12), we obtain

$$\alpha(n+1) + \alpha(n) = 0 \Longrightarrow \alpha(n) = c_1(-1)^n.$$

Now, we can express the characteristic $\mathcal{K}(n,u_n) = \alpha(n) = c_1(-1)^n$, and substituting into (4.10) we have

$$c_1(-1)^{n+2} + \left(1+\frac{1}{(u_{n+1}+u_n)^2}\right)c_1(-1)^{n+1} + \frac{1}{(u_{n+1}+u_n)^2}c_1(-1)^n = 0,$$

$$c_1(-1)^n\left(1-1-\frac{1}{(u_{n+1}+u_n)^2}+\frac{1}{(u_{n+1}+u_n)^2}\right) = 0,$$

which is true for every $c_1 \in \mathbb{R}$, so $\mathcal{K}(n,u_n) = c_1(-1)^n$ for each $c_1 \in \mathbb{R}$, and that is the general solution for LSC of the given difference equation.

Remark 4.1. ([1], Note on page 58) In each of the previous examples, we have eliminated w, then u_{n+1}, to leave a difference equation for $\mathcal{K}(n,u_n)$. In general, this is a good approach, but it is not always the simplest way to derive a difference equation for the characteristic. For some difference equations, the calculations are simpler if one eliminates $\mathcal{K}(n,u_n)$ in order to find a difference equation by $\mathcal{K}(n+i,u_{n+i})$, for special $i \geq 1$.

5. Reducing the order of nonlinear difference equations

It is well known that if we want to reduce the order of a linear differential equation or difference equation of the k^{th} order, it is necessary to know a non-trivial solution of the corresponding homogeneous equation. Sophus Lie extended this method to the case of nonlinear differential equations by exploiting one-parameter Lie groups of point

symmetries. It turns out that this method can be successfully applied in the case of nonlinear difference equations by determining the compatible canonical coordinate.

First, let us apply the corresponding Lie method to the second-order difference equation

$$u_{n+2} = w\left(n, u_n, u_{n+1}\right). \tag{5.1}$$

Let us assume that we managed to determine the characteristic $K(n;u_n)$ for the given difference equation (5.1) and that s_n is a compatible canonical coordinate. Let

$$r_n = s_{n+1} - s_n = \int \frac{du_{n+1}}{\mathcal{K}(n+1,u_{n+1})} - \int \frac{du_n}{\mathcal{K}(n,u_n)}. \tag{5.2}$$

By using the shift operator on (5.2), we get

$$r_{n+1} = \int \frac{du_{n+2}}{\mathcal{K}(n+2,u_{n+2})} - \int \frac{du_{n+1}}{\mathcal{K}(n+1,u_{n+1})}. \tag{5.3}$$

On solutions of the difference equation (5.1), we can replace u_{n+2} in (5.3) by w and treat r_{n+1} as a function of n, u_n i u_{n+1}. Then, we obtain

$$\begin{aligned}\frac{\partial r_{n+1}}{\partial u_{n+1}} &= \frac{D_2 w}{\mathcal{K}(n+2,u_{n+2})} - \frac{1}{\mathcal{K}(n+1,u_{n+1})} = -\frac{D_1 w \mathcal{K}(n,u_n)}{\mathcal{K}(n+1,u_{n+1})\mathcal{K}(n+2,w)} \\ &= -\frac{\mathcal{K}(n,u_n)}{\mathcal{K}(n+1,u_{n+1})}\frac{\partial r_{n+1}}{\partial u_n}.\end{aligned} \tag{5.4}$$

If we now consider r_{n+1} as a function of n, s_n i s_{n+1}, then

$$\frac{\partial r_{n+1}}{\partial s_{n+1}} = \frac{\partial r_{n+1}}{\partial u_{n+1}}\frac{\partial u_{n+1}}{\partial s_{n+1}} = -\frac{\mathcal{K}(n,u_n)}{\mathcal{K}(n+1,u_{n+1})}\frac{\partial r_{n+1}}{\partial u_n}\frac{\partial u_{n+1}}{\partial s_{n+1}}.$$

From the fact that $s_{n+1} = \int \frac{du_{n+1}}{\mathcal{K}(n+1,u_{n+1})}$, it follows $\frac{\partial u_{n+1}}{\partial s_{n+1}} = \mathcal{K}(n+1,u_{n+1})$, so we have that

$$\frac{\partial r_{n+1}}{\partial s_{n+1}} = -\mathcal{K}(n,u_n)\frac{\partial r_{n+1}}{\partial u_n}. \tag{5.5}$$

On the other hand, using that $\frac{\partial u_n}{\partial s_n} = \mathcal{K}(n,u_n)$, we obtain

$$\frac{\partial r_{n+1}}{\partial s_n} = \frac{\partial r_{n+1}}{\partial u_n}\frac{\partial u_n}{\partial s_n} = \mathcal{K}(n,u_n)\frac{\partial r_{n+1}}{\partial u_n}. \tag{5.6}$$

Adding (5.5) and (5.6), gives

$$\frac{\partial r_{n+1}}{\partial s_{n+1}} + \frac{\partial r_{n+1}}{\partial s_n} = 0. \tag{5.7}$$

Equation (5.7) implies that r_{n+1} depends only on n, s_n and s_{n+1}. Thus, we have reduced the equation (5.1) to a first-order difference equation of the following form:

$$r_{n+1} = F\left(n, r_n\right). \tag{5.8}$$

We can continue the process if the equation (5.1) can be solved. Let its general solution be of the form

$$r_n = f\left(n; c_1\right).$$

Then, the general solution of the equation (5.1) is given by

$$\int \frac{du_n}{\mathcal{K}(n,u_n)} = s_n = \sum_{k=n_0}^{n-1} f(k;c_1) + c_2. \tag{5.9}$$

As s_n is a canonical compatible coordinate, this solution can be inverted (in principle) to yield u_n.

Example 5.1. *([1], Ex. 2.19) The Lie point symmetries of the following difference equation*

$$u_{n+2} = \frac{2u_n u_{n+1}}{u_{n+1} + u_n} \tag{5.10}$$

include symmetries whose characteristic is $\mathcal{K}(n,u_n) = u_n^2$. Use this result to reduce the equation (5.10) to a first-order difference equation, and hence find the general solution of the equation (5.10).

Solution. Equation (5.10) also appears in [3] as a special case and is reduced to linear equation by substituting $z_n = 1/u_n$. However, using Lie's method, we will demonstrate a slightly more detailed solution than in [1]. From (5.2) let

$$r_{n+1} = \int \frac{du_{n+1}}{u_{n+1}^2} - \int \frac{du_{n+1}}{u_n^2} = \frac{1}{u_n} - \frac{1}{u_{n+1}},$$

and it is obvious that $s_n = -\dfrac{1}{u_n}$ is compatible. Then, based on the solutions of (5.10),

$$r_{n+1} = \frac{1}{u_{n+1}} - \frac{1}{u_{n+2}} = \frac{1}{u_{n+1}} - \frac{u_{n+1}+u_n}{2u_n u_{n+1}} = \frac{1}{2u_{n+1}} - \frac{1}{2u_n} = -\frac{r_n}{2}, \tag{5.11}$$

from which $r_n = c_1\left(-\dfrac{1}{2}\right)^n$. Since $r_n = s_{n+1} - s_n = \Delta s_n$, then

$$s_n = \sum_{k=0}^{n-1} r_k + c_2 = c_1 \sum_{k=0}^{n-1}\left(-\frac{1}{2}\right)^k + c_2 = \frac{1-\left(-\frac{1}{2}\right)^n}{1-\left(-\frac{1}{2}\right)} + c_2 = \frac{2c_1}{3}\left(1-\left(-\frac{1}{2}\right)^n\right) + c_2.$$

From $s_n = -\dfrac{1}{u_n}$ we obtain the general solution of (5.10):

$$u_n = \frac{1}{C_1\left(-\frac{1}{2}\right)^n + C_2}, \quad (C_1, C_2 - \text{new constants}).$$

Remark 5.1. The above reduction process is a generalization of order reduction methods for linear difference equations.

To see the correctness of the previous remark, consider the following second-order linear difference equation:

$$u_{n+2} + a_1(n)\,u_{n+1} + a_0(n)\,u_n = b_n. \tag{5.12}$$

The following is a detailed explanation of all the steps in contrast to the summary overview in [1]. Let us determine the characteristic $\mathcal{K}(n,u_n)$ symmetry of the Lie point for the equation (5.12) using the already described algorithm for the case of a second-

order difference equation. Here, since $w(n,u_n,u_{n+1}) = b_n - a_1(n)u_{n+1} - a_0(n)u_n)$, the LSC form is

$$\mathcal{K}(n+2,w) + a_1(n)\mathcal{K}(n+1,u_{n+1}) + a_0(n)\mathcal{K}(n,u_n) = 0, \tag{5.13}$$

because $D_1 w = -a_0(n)$ i $D_2 w = -a_1(n)$. Then, due to the elimination of the first term in (5.13)), we have

$$\left(\frac{1}{D_1 w}\frac{\partial}{\partial u_n} - \frac{1}{D_2 w}\frac{\partial}{\partial u_{n+1}}\right)\mathcal{K}(n+2, b_n - a_1(n)u_{n+1} - a_0(n)u_n) = 0,$$

that is,

$$\left(-\frac{1}{a_0(n)}\frac{\partial}{\partial u_n} - \frac{1}{a_1(n)}\frac{\partial}{\partial u_{n+1}}\right)\left(-a_1(n)\mathcal{K}(n+1,u_{n+1}) - a_0(n)\mathcal{K}(n,u_n)\right) = 0,$$

thus,

$$-\frac{1}{a_0(n)}(-a_0(n))\mathcal{K}'(n,u_n) + \frac{1}{a_1(n)}(-a_1(n))\mathcal{K}'(n+1,u_{n+1}) = 0.$$

Applying the operator $\dfrac{d}{du_n}$ to the last equation, we get $\mathcal{K}''(n,u_n) = 0$, which implies that $\mathcal{K}(n,u_n) = c_1 n + c_2$. Taking $c_1 = 1,\ c_2 = 0$ and $u_n = f(n)$, where $f(n)$ is a non-zero solution of the associated homogeneous equation (5.12), we get $\mathcal{K}(n,u_n) = f(n)$. Note that:

$$f(n+2) + a_1(n)f(n+1)\,a_0(n)f(n) = 0. \tag{5.14}$$

Further, we have

$$\begin{aligned} r_n = s_{n+1} - s_n &= \int \frac{du_{n+1}}{\mathcal{K}(n+1,u_{n+1})} - \int \frac{du_n}{\mathcal{K}(n,u_n)} = \int \frac{du_{n+1}}{f(n+1)} - \int \frac{du_n}{f(n)} \\ &= \frac{1}{f(n+1)}\int du_{n+1} - \frac{1}{f(n)}\int du_n = \frac{u_{n+1}}{f(n+1)} - \frac{u_n}{f(n)}, \end{aligned}$$

so that,

$$\begin{aligned} r_{n+1} &= \frac{u_{n+2}}{f(n+2)} - \frac{u_{n+1}}{f(n+1)} = \frac{b(n) - a_1(n)u_{n+1} - a_0(n)u_n}{f(n+2)} - \frac{u_{n+1}}{f(n+1)} \\ &= \frac{b(n)}{f(n+2)} - \left(a_1(n)\frac{u_{n+1}}{f(n+2)} + a_0(n)\frac{u_n}{f(n+2)} + \frac{u_{n+1}}{f(n+1)}\right) \\ &= \frac{b(n)}{f(n+2)} - \left(a_1(n)\frac{u_{n+1}}{f(n+1)}\frac{f(n+1)}{f(n+2)} + a_0(n)\frac{u_n}{f(n)}\frac{f(n)}{f(n+2)} + \frac{u_{n+1}}{f(n+1)}\right) \\ &= \frac{b(n)}{f(n+2)} - \left[\frac{u_{n+1}}{f(n+1)}\left(a_1(n)\frac{f(n+1)}{f(n+2)} + 1\right) + a_0(n)\frac{u_n}{f(n)}\frac{f(n)}{f(n+2)}\right] \\ &= \frac{b(n)}{f(n+2)} - \left[\frac{u_{n+1}}{f(n+1)}\left(\frac{-f(n+2) - a_0(n)f(n)}{f(n+2)} + 1\right) + a_0(n)\frac{u_n}{f(n)}\frac{f(n)}{f(n+2)}\right] \\ &= \frac{b(n)}{f(n+2)} + a_0(n)\frac{u_{n+1}}{f(n+1)}\frac{f(n)}{f(n+2)} - a_0(n)\frac{u_n}{f(n)}\frac{f(n)}{f(n+2)}\Bigg] \\ &= \frac{b(n)}{f(n+2)} + a_0(n)\frac{f(n)}{f(n+2)}\left(\frac{u_{n+1}}{f(n+1)} - \frac{u_n}{f(n)}\right). \end{aligned}$$

Thus,

$$r_{n+1} = \frac{b(n)}{f(n+2)} + \frac{a_0(n) f(n)}{f(n+2)} r_n,$$

which is a first-order linear difference equation whose general solution is given by

$$r_n = \left(\prod_{i=0}^{n-1} \frac{a_0(i) f(i)}{f(i+2)} \right) r_0 + \sum_{k=0}^{n-1} \left(\prod_{i=k+1}^{n-1} \frac{a_0(i) f(i)}{f(i+2)} \right) \frac{b(k)}{f(k+2)}.$$

Example 5.2. *([1], Example 2.21 - modified here into a more complex problem) Determine the characteristic of Lie point symmetries for the difference equation*

$$u_{n+2} = \frac{2u_{n+1} - u_n + u_n u_{n+1}^2}{1 - u_{n+1}^2 + 2u_n u_{n+1}}. \tag{5.15}$$

Use this characteristic for the reduction of order and solve the difference equation (5.15).

Solution. For the difference equation (5.15), we have

$$D_1 w = \frac{\partial w}{\partial u_n} = \frac{\left(-1 + u_{n+1}^2\right)\left(1 - u_{n+1}^2 + 2u_n u_{n+1}\right) - 2u_{n+1}\left(2u_{n+1} - u_n + u_n u_{n+1}^2\right)}{\left(1 - u_{n+1}^2 + 2u_n u_{n+1}\right)^2}$$

$$= -\frac{(1 + u_{n+1})^2}{\left(1 - u_{n+1}^2 + 2u_n u_{n+1}\right)^2},$$

$$D_2 w = \frac{\partial w}{\partial u_{n+1}} = \frac{\left(2 + 2u_n u_{n+1}\right)\left(1 - u_{n+1}^2 + 2u_n u_{n+1}\right) - \left(-2u_{n+1} + 2u_n\right)\left(2u_{n+1} - u_n + u_n u_{n+1}^2\right)}{\left(1 - u_{n+1}^2 + 2u_n u_{n+1}\right)^2}$$

$$= \frac{2\left(1 + u_n^2\right)\left(1 + u_{n+1}^2\right)}{\left(1 - u_{n+1}^2 + 2u_n u_{n+1}\right)^2}.$$

From here, we get $\eta = \ln\left|\frac{D_2 w}{D_1 w}\right| = \ln \frac{2\left(1 + u_n^2\right)}{1 + u_{n+1}^2}$, so

$$D_1 \eta = \frac{1 + u_{n+1}^2}{2\left(1 + u_n^2\right)} \frac{4u_n}{1 + u_{n+1}^2} = \frac{2u_n}{1 + u_n^2} \Longrightarrow D_{12}\eta = 0,$$

$$D_2 \eta = \frac{1 + u_{n+1}^2}{2\left(1 + u_n^2\right)} \frac{-2u_{n+1} 2\left(1 + u_n^2\right)}{\left(1 + u_{n+1}^2\right)^2} = \frac{-2u_{n+1}}{1 + u_{n+1}^2}.$$

Since here $D_{12}\eta = 0$ and $D_2\eta = -E D_1 \eta$, we can apply the following formula

$$\mathcal{K}'(n, u_n) - D_1 \eta \mathcal{K}(n, u_n) = \alpha(n),$$

where $\alpha(n) = c_1$. If we take $c_1 = 0$, we get

$$\mathcal{K}'(n,u_n) - \frac{2u_n}{1+u_n^2}\mathcal{K}(n,u_n) = 0,$$

that is,

$$\frac{d\mathcal{K}(n,u_n)}{\mathcal{K}(n,u_n)} = \frac{2u_n}{1+u_n^2}du_n,$$

from which

$$\mathcal{K}(n,u_n) = 1+u_n^2.$$

By using (5.2), we obtain

$$r_n = s_{n+1} - s_n = \int \frac{du_{n+1}}{\mathcal{K}(n+1,u_{n+1})} - \int \frac{du_n}{\mathcal{K}(n,u_n)} = \int \frac{du_{n+1}}{1+u_{n+1}^2} - \int \frac{du_n}{1+u_n^2}$$
$$= \arctan u_{n+1} - \arctan u_n = \arctan \frac{u_{n+1}-u_n}{1+u_nu_{n+1}}.$$

Therefore,

$$r_{n+1} = \arctan u_{n+2} - \arctan u_{n+1} = \arctan \frac{u_{n+2}-u_{n+1}}{1+u_{n+1}u_{n+2}} = \arctan \frac{\frac{2u_{n+1}-u_n+u_nu_{n+1}^2}{1-u_{n+1}^2+2u_nu_{n+1}} - u_{n+1}}{1+u_{n+1}\frac{2u_{n+1}-u_n+u_nu_{n+1}^2}{1-u_{n+1}^2+2u_nu_{n+1}}}$$

$$= \frac{u_{n+1}-u_n-u_nu_{n+1}^2+u_{n+1}^3}{1+u_{n+1}^2+u_nu_{n+1}+u_nu_{n+1}^3} = \arctan \frac{(u_{n+1}-u_n)\left(1+u_{n+1}^2\right)}{\left(1+u_{n+1}^2\right)(1+u_nu_{n+1})},$$

that is,

$$r_{n+1} = \arctan \frac{u_{n+1}-u_n}{1+u_nu_{n+1}} = r_n \qquad (n=0,1,2,\ldots),$$

so $r_n = c_1$ (c_1 is a constant). Since $r_n = s_{n+1} - s_n = \Delta s_n$, we obtain $s_n = \Delta^{-1}c_1 = c_1n + c_2$ (c_2 is a constant). It implies that

$$s_n = \arctan u_n \Longrightarrow \arctan u_n = c_1n + c_2,$$

that is, $u_n = \tan(c_1n + c_2)$, taking care that the domain of the function arctan is the interval $(-\pi/2, \pi/2)$.

Remark 5.2. The consideration carried out for the equation (5.12) shows that if a particular characteristic is common to a class of difference equations, every equation in the class can be reduced by the same r_n (subject to the canonical coordinate s being compatible with the difference equation). This is also true for ordinary difference equations; most methods for solving a particular class of ordinary differential equations exploit a Lie symmetry group that is shared by all differential equations in that class.

Reduction of order is not restricted to second-order difference equations. By the same process a difference equation of the p^{th}-order with a known non-zero characteristic $\mathcal{K}(n,u_n)$ reduces to a $(p-1)^{th}$-order difference equation, for $r_n = s_{n+1} - s_n$, where s is any compatible canonical coordinate. It is easy to check that LSC yields

$$\frac{\partial r_{n+p-1}}{\partial s_{n+p-1}} + \ldots + \frac{\partial r_{n+p-1}}{\partial s_{n+1}} + \frac{\partial r_{n+p-1}}{\partial s_n} = 0,$$

and consequently there exists a function F such that

$$r_{n+p-1} = F\left(n, s_{n+1} - s_n, \ldots, s_{n+p-1} - s_{n+p-2}\right) = F\left(n, r_n, \ldots, r_{n+p-2}\right). \tag{5.16}$$

If the general solution of the equation (5.16) is of the form

$$r_n = f\left(n; c_1, c_2, \ldots, c_{p-1}\right),$$

the general solution of the original p^{th}-order difference equation is

$$\int \frac{du_n}{\mathcal{K}(n, u_n)} = s_n = \sum_{k=n_0}^{n-1} f\left(k; c_1, c_2, \ldots, c_{p-1}\right) + c_p.$$

Since s is compatible, this solution defines u_n uniquely (for given c_1).

References

[1] P.E. Hydon, *Difference Equations by Differential Equation Methods*, Cambridge University Press, Cambridge, 2014.

[2] P.E. Hydon, Symmetries and first integrals of ordinary difference equations, *Proc. R. Soc. Lond. Ser. A Mah. Phys. g. Sci.,* 456 (2000), 2835-2855.

[3] S. Jašarević and M.R.S. Kulenović, Basins of attraction of equilibrium and boundary points of second-order difference equations, *Journal of Difference Equations and Applications*, Volume 20, Issue 5-6 (2014), 1-13. http://www.tandfonline.com/loi/gdea20

[4] M.R.S. Kulenović and O. Merino, *Discrete dynamical systems and difference equations with Mathematica*, Chapman&Hall/CRC, Boca Raton, USA, 2002.

[5] M. Nurkanović, *Diferentne jednadžbe - Teorija i primjene,* Denfas, Tuzla, 2008.

[6] M. Nurkanović, Z. Nurkanović, *Linearne diferentne jednadžbe - Teorija i zadaci sa primjenama,* PrintCom, Tuzla, 2016.

[7] M. Trumić, *Primjena diferentnih jednadžbi u nastavi matematike,* Doktorska disertacija, PMF Sarajevo, 2023.

(Received: May 15, 2024)
(Revised: May 30, 2024)

Mehmed Nurkanović
University of Tuzla
Department of Mathematics
U. Vejzagića 4
75000 Tuzla
Bosnia and Herzegovina
email: *mehmed.nurkanovic@untz.ba*
and
Mirsad Trumić
High school: Agricultural and Medical school
Brčko Disdtrict
Bosnia and Herzegovina
email: *trumicmirsad@yahoo.com*

PROCEEDINGS OF THE CONFERENCE ON MARCH 14 – INTERNATIONAL DAY OF MATHEMATICS
Special Editions ANUBiH, Book CCXVI, OPMN Book 30
Sarajevo 2024, **p. 63–67** DOI: 10.5644/PI2024.216.08

PROPERTIES OF SOME SPECTRA OF SUPERPOSITION OPERATORS

SANELA HALILOVIĆ

ABSTRACT. We consider the nonlinear superposition operator F in Banach spaces of sequences l_p, generated by the function $f(s,u)$. We analyze the Rhodius spectra $\sigma_R(F)$ and the Neuberger spectra $\sigma_N(F)$ of these operators F generated by the function

$$f(s,u) = a(s) + \phi(u),$$

where $(a(s))_{s\in\mathbb{N}}$ is a sequence from l_p $(1 \le p \le \infty)$, and $\phi(u)$ is a continuous function in $\mathbb{R}$. Some connections between the property of the function $\phi(u)$ and the corresponding spectra $\sigma_R(F)$ and $\sigma_N(F)$ are given in this paper. There are also a few examples that verify proposed theorems.

1. INTRODUCTION

In this paper we consider the Rhodius and Neuberger spectra of superposition operators in Banach spaces l_p $(1 \le p \le \infty)$. The superposition operators have an important place in many mathematical problems and also there are various applications in mathematical physics, mathematical economics, mathematical biology and so on. Let $f(s,u)$ be a function defined on $\mathbb{N}\times\mathbb{R}$ with values in $\mathbb{R}$. Given a function $x = x(s)$, by applying f, we get the function $f(s,x(s))$ and this function generates an operator F

$$F(x(s)) = f(s,x(s)). \tag{1.1}$$

This operator (1.1) is called the superposition operator, composition operator or Nemytskii operator. We take $x = x(s)$ a sequence from the l_p spaces of sequences $(1 \le p \le \infty)$, which are the Banach spaces equipped with the standard norm. It is known that the spectrum of a linear operator has many useful properties. For nonlinear operators F, the notion spectrum of F is a wider concept, based on the property of $\lambda I - F$ being a regular map.

For the class of all continuous operators F on a Banach space X, denoted by $\mathfrak{C}(X)$, the following definition of Rhodius spectrum has been introduced.

Definition 1.1. *For the continuous operator $F : X \to X$ the set*

$$\rho_R(F) = \{\lambda \in \mathbb{R} : \lambda I - F \text{ is bijective and } (\lambda I - F)^{-1} \in \mathfrak{C}(X)\}$$

2020 *Mathematics Subject Classification.* 47J10, 47H30.

Key words and phrases. superposition operator, spectra, nonlinear operator, Rhodius spectrum, Neuberger spectrum.

is called the Rhodius resolvent set and

$$\sigma_R(F) = \mathbb{R} \setminus \rho_R(F)$$

is the Rhodius spectrum.

Thus, a point $\lambda \in \mathbb{R}$ belongs to $\rho_R(F)$ if and only if $\lambda I - F$ is a homeomorphism on X. The Neuberger spectrum of nonlinear operators was proposed for $\mathfrak{C}^1(X)$, the class of continuously Fréchet differentiable operators on Banach space X.

Definition 1.2. *For an operator $F : X \to X$, which is continuously Fréchet differentiable, the Neuberger resolvent set is defined by*

$$\rho_N(F) = \left\{\lambda \in \mathbb{R} : \lambda I - F \text{ is bijective and } (\lambda I - F)^{-1} \in \mathfrak{C}^1(X)\right\}$$

and the set $\sigma_N(F) = \mathbb{R} \setminus \rho_N(F)$ is called the Neuberger spectrum of F.

A point $\lambda \in \mathbb{R}$ belongs to $\rho_N(F)$ if and only if $\lambda I - F$ is a diffeomorphism on X.
Some useful properties of these spectra are:

- If F is a linear operator $\Rightarrow \sigma_R(F) = \sigma_N(F) = \sigma(F)$,
- If $F0 = 0 \Rightarrow \sigma_R(F) \subseteq \sigma_N(F)$,
- If the underlying space is complex, then $\sigma_N(F)$ is nonempty.

More information about various nonlinear spectral theories can be found in [1]. The conditions of acting, continuity and differentiability of the superposition operator in Banach spaces l_p, are given in the following three theorems from [2].

Theorem 1.1. *Let $1 \le p, q < \infty$. Then the following properties are equivalent:*

- *the operator F acts from l_p to l_q;*
- *there are functions $a(s) \in l_q$ and constants $\delta > 0, n \in \mathbb{N}, b \ge 0$, for which*

$$|f(s,u)| \le a(s) + b|u|^{\frac{p}{q}} \ (s \ge n, |u| < \delta);$$

- *for any $\epsilon > 0$ there exists a function $a_\epsilon \in l_q$ and constants $\delta_\epsilon, > 0, n_\epsilon \in \mathbb{N}$, $b_\epsilon \ge 0$, for which $\|a_\epsilon(s)\|_q < \epsilon$ and*

$$|f(s,u)| \le a_\epsilon(s) + b_\epsilon|u|^{\frac{p}{q}} \ (s \ge n_\epsilon, |u| < \delta_\epsilon).$$

Theorem 1.2. *Let $1 \le p, q < \infty$ and let the superposition operator* (1.1)*, generated by the function $f(s,u)$, acting from l_p to l_q. Then this operator is continuous if and only if each of the functions is continuous for every $s \in \mathbb{N}$.*

Theorem 1.3. *Let $1 \le p, q < \infty$ and the operator F generated by the function $f(s,u)$ acting from l_p into l_q. The operator F is differentiable at $x_0 \in l_p$ if and only if $f'_u(s,\cdot)$ is continuous at x_0 for almost all $s \in \mathbb{N}$.*

2. Main Results

The Rhodius and Neuberger spectra are defined for a continuous operator F, that is, $F \in \mathfrak{C}(l_p)$, so the function $\phi(u)$ has to be continuous function on $\mathbb{R}$.

Theorem 2.1. *Let the superposition operator $F : l_p \to l_p$ be generated by the function $f(s,u) = a(s) + \phi(u)$, where $(a(s))_s$ is a sequence from the space l_p $(1 \le p \le \infty)$. If $\phi(u)$ is a bijective function, then $0 \in \rho_R(F)$ and $0 \notin \sigma_R(F)$*

Proof. For $\lambda = 0$ the operator $\lambda I - F$ becomes $-F$. If $\phi(u)$ is a bijection then the functions $f(s,u) = a(s) + \phi(u)$ and $-f(s,u) = -a(s) - \phi(u)$ are bijective for every $s \in \mathbb{N}$. It follows that the operator F, generated by f and the operator $-F$, generated by $-f$ are bijective. Since $-f$ is a bijective function, it follows that its inverse $(-f)^{-1}$ exists and it is also a bijective function. The function $-f(s,u)$ is a bijective and continuous for every s and hence $(-f)^{-1}(s,u)$ is a bijective and continuous function for every s. From the Theorem 1.2 we conclude the operator $(-F)^{-1}$ generated by $(-f)^{-1}$ is a continuous operator. So, we show that operator $-F$ is bijective and $(-F)^{-1} \in \mathfrak{C}(l_p)$. It means that $0 \in \rho_R(F)$ and $0 \notin \sigma_R(F)$. □

The Neuberger spectrum is defined for $F \in \mathfrak{C}^1(l_p)$, so the function $\phi(u)$ has to be continuously differentiable on $\mathbb{R}$. Then the derivative $f'_u(s,u) = \phi'(u)$ is a continuous function for all $s \in \mathbb{N}$ and from Theorem 1.3 it follows the operator F is continuously Fréchet differentiable.

Theorem 2.2. *Let the superposition operator $F : l_p \to l_p$ be generated by the function $f(s,u) = a(s) + \phi(u)$, where $(a(s))_s$ is a sequence from the space l_p $(1 \le p \le \infty)$. Let $\phi(u)$ be a bijective and continuously differentiable function.*
a) If $\phi'(u) \neq 0, \forall u$, then $0 \in \rho_N(F)$ and $0 \notin \sigma_N(F)$.
b) If there exists u_0 such that $\phi'(u_0) = 0$, then $0 \notin \rho_N(F)$ and $0 \in \sigma_N(F)$.

Proof. a) The function $\phi(u)$ is a continuously differentiable function, so $\phi'(u)$ exists for every $u \in \mathbb{R}$ and it is a continuous function. Since the function ϕ is bijective, its inverse ϕ^{-1} exists and it is also bijective. If $\phi'(u) \neq 0, \forall u$, then in virtue of the inverse function theorem (see [6]), the ϕ^{-1} is a differentiable function and

$$(\phi^{-1})'(y) = \frac{1}{\phi'(u)} \tag{2.1}$$

holds, where $\phi(u) = y$ and $\phi^{-1}(y) = u$.
We have $f'_u(s,u) = (a(s) + \phi(u))'_u = \phi'(u) \neq 0$ and

$$(f^{-1})'_u = \frac{1}{f'_u(s,u)} = \frac{1}{\phi'_u(u)}. \tag{2.2}$$

This derivative $(f^{-1})'_u$ is continuous in u because $\phi'(u)$ is a continuous function and $\phi'(u) \neq 0$ for every u. That is why, according to the Theorem 1.3, $(f^{-1})'_u$ generates a continuous operator $(F^{-1})'$, i.e. $(F^{-1}) \in \mathfrak{C}^1(l_p)$. Then clearly, $((-f)^{-1})'_u$ is also a continuous function and $(-F)^{-1} \in \mathfrak{C}^1(l_p)$. In the proof of Theorem 2.1 we have already shown that $-F$ is a bijective operator and now we see that $(-F)^{-1} \in \mathfrak{C}^1(l_p)$, so we have proved that $0 \in \rho_N(F)$ and $0 \notin \sigma_N(F)$.
b) If there exists u_0 such that $\phi'(u_0) = 0$, then the function ϕ^{-1} is not differentiable at $y_0 = \phi(u_0)$ and consequently, the partial derivatives $(f^{-1})'_u$ and $(-f^{-1})'_u$ are not continuously differentiable functions. Hence, $(-F)^{-1} \notin \mathfrak{C}^1(l_p)$ and this means that $0 \notin \rho_N(F)$ and $0 \in \sigma_N(F)$. □

2.1. **Examples**

Here we give two examples of nonlinear superposition operators and their Rhodius and Neuberger spectra, which illustrate and verify these theorems (Theorem 2.1 and Theorem 2.2).

Example 2.1. *Let the operator $F : l_p \to l_p$ be generated by the function*

$$f(s,u) = a(s) + u^n, \tag{2.3}$$

where $n \geq 3$ and n is an odd number. This function $\phi(u) = u^n$ is bijective. In [4] it was found that $\sigma_R(F) = (0,\infty)$, so $0 \notin \sigma_R(F)$. So, in this example we see that the function $\phi(u)$ is bijective and $0 \notin \sigma_R(F)$ and this confirms Theorem 2.1.
The Fréchet derivative of the operator F generated by (2.3)*, at $x_0 = (x_1, x_2, \ldots)$ along $h = (h_1, h_2, \ldots)$ is:*

$$F'(x_0)h = (nx_1^{n-1}h_1, nx_1^{n-1}h_2, \ldots).$$

In [3] we found $\sigma_N(F) = [0,\infty)$, so $0 \in \sigma_N(F)$. Here, the function $\phi(u) = u^n$ is continuously differentiable ($\phi'(u) = n \cdot u^{n-1}$), but its inverse $\phi^{-1}(u) = \sqrt[n]{u}$ is not differentiable at $u = 0$ $\left((\phi^{-1})'(u) = \frac{1}{n} \cdot \frac{1}{\sqrt[n]{u^{n-1}}}\right)$. In this case the function $\phi(u) = u^n$ is bijective and $\phi'(u) = n \cdot u^{n-1}$. Therefore, there exists $u_0 = 0$, such that $\phi'(u_0) = \phi'(0) = 0$. In virtue of the Theorem 2.2 it follows that $0 \in \sigma_N(F)$. Hence, in this example Theorem 2.2 is verified.

Example 2.2. *Let the operator $F : l_p \to l_p$ be generated by the function*

$$f(s,u) = a(s) + \sqrt[n]{u}, \tag{2.4}$$

where $n \geq 3$ and n is an odd number. This function $\phi(u) = \sqrt[n]{u}$ is bijective and the function $f(s,u)$ is bijective for every $s \in \mathbb{N}$, so the operator F is bijective. In [4] we found that $\sigma_R(F) = (0,\infty)$, so $0 \notin \sigma_R(F)$. Hence,in this example the function $\phi(u) = \sqrt[n]{u}$, (n-odd number) is bijective and $0 \notin \sigma_R(F)$ and it agrees with Theorem 2.1.
The Fréchet derivative of the operator F generated by (2.4)*, at $x_0 = (x_1, x_2, \ldots)$ along $h = (h_1, h_2, \ldots)$ is:*

$$F'(x_0)h = \left(\frac{1}{n\sqrt[n]{x_1^{n-1}}}h_1, \frac{1}{n\sqrt[n]{x_2^{n-1}}}h_2, \ldots\right).$$

The function $-f(s,u)$ is bijective and the operator $-F$ is bijective. We have

$$-f(s,u) = -a(s) - \sqrt[n]{u}$$

and

$$(-f)^{-1}(s,u) = (-a(s) - u)^n.$$

Let us find the partial derivative of $(-f)^{-1}$ with respect to the variable u :

$$((-f)^{-1})'_u = n \cdot (-u - a(s))^{n-1} \cdot (-u - a(s))'_u = -n \cdot (-u - a(s))^{n-1}. \tag{2.5}$$

The function (2.5) *is continuous for all $s \in \mathbb{N}$ and from Theorem 1.3 it follows that the operator $(-F)^{-1}$ is continuously differentiable. Here we see that $-F$ is a bijective operator and $(-F)^{-1} \in \mathfrak{C}(l_p)$. Now from Definition 1.2 it follows $0 \in \rho_N(F)$ and $0 \notin$*

$\sigma_N(F)$. In this case we have the function $\phi(u) = \sqrt[n]{u}$ which is bijective and

$$\phi'(u) = (\sqrt[n]{u})'_u = \frac{1}{n} \cdot \frac{1}{\sqrt[n]{u^{n-1}}} \neq 0, \forall u \in \mathbb{R}. \tag{2.6}$$

Therefore, from Theorem 2.2 it follows that $0 \in \rho_N(F)$ and $0 \notin \sigma_N(F)$. So, in this example we also get the validation of Theorem 2.2.

3. CONCLUSION

In this paper we observe the class of superposition operators $F : l_p \to l_p$, generated by the function $f(s,u) = a(s) + \phi(u)$. We find out how the fact that the function $\phi(u)$ is bijective affects the Rhodius and Neuberger spectra of the operator F, generated by the function $f(s,u)$. We conclude that it affects the corresponding spectra in regards to whether $\sigma_R(F)$ and $\sigma_N(F)$ contain 0. In [5] we found that if the function $\phi(u)$ is not a bijection, then $0 \in \sigma_R(F)$ and $0 \in \sigma_N(F)$. Our further goal is to investigate how some other properties of these spectra of nonlinear superposition operators, such as closedness, boundedness etc, depend on the properties of their generating function $f(s,u)$.

REFERENCES

[1] J. Appell, E. De Pascale and A. Vignoli, *Nonlinear Spectral Theory*, Walter de Gruyter, 2004.
[2] F. Dedagić and P. P. Zabreiko, *On the superposition operator in l_p spaces*, Sibir. Mat. Zhurn., vol.28, No.1, (1987), pp. 86-98.
[3] S. Halilović and R. Vugdalić, *The Neuberger Spectra of Nonlinear Superposition Operators in the Spaces of Sequences*, Journal of the International Mathematical Virtual Institute, vol.4, (2014), pp. 97-119.
[4] S. Halilović and R. Vugdalić, *The Rhodius Spectra of Some Nonlinear Superposition Operators in the Spaces of Sequences*, Adv. Math., Sci.J., vol.3, no. 2, (2014), pp. 83-96.
[5] S. Halilović, *Some Spectra of Superposition Operators Generated by an Exponential Function*, Communications in Mathematics and Applications, vol.12, no. 1 (2021), pp. 221-229.
[6] W. Rudin, *Principles of Matemathematical Analysis*, Mc Graw-Hill, 1976.

(Received: May 18, 2024)
(Revised: July 27, 2024)

Sanela Halilović
University of Tuzla
Faculty of Natural Sciences and Mathematics
Urfeta Vejzagića 4
75000 Tuzla
Bosnia and Herzegovina
e-mail: *sanela.halilovic@untz.ba*

PROCEEDINGS OF THE CONFERENCE ON MARCH 14 – INTERNATIONAL DAY OF MATHEMATICS
Special Editions ANUBiH, Book CCXVI, OPMN Book 30
Sarajevo 2024, **p. 68–80** DOI: 10.5644/PI2024.216.09

ERGODICITY OF UNIFORMLY DIFFERENTIABLE FUNCTIONS MODULO p ON $\mathbb{Z}_p$ AND SOME CLASSES OF 1-LIPSCHITZ MEASURE PRESERVING FUNCTIONS ON $\mathbb{Z}_p$

JASMINA MUMINOVIĆ HUREMOVIĆ

ABSTRACT. Various applications in physics, cognitive science and cryptography often lead to the study of the behavior of dynamical systems on $\mathbb{Z}_p$. For example, in the theory of pseudorandom number generation, it is useful to have a mapping, defined on the set of integers, that gives large cycles modulo n for a given integer n . Given that minimal mappings have only one cycle of maximal length modulo p^n, for each n, good candidates are precisely minimal mappings in the set of p-adic integers (maps whose orbits are all dense).
In this paper we give theoretical research in the field of p-adic analysis, p-adic ergodic functions and dynamical systems defined on the set of p-adic integers $\mathbb{Z}_p$. In [5] it was shown that for 1-Lipschitz functions, which are measure preserving, the notions of minimality and ergodicity are equivalent. Guided by the results from the mentioned paper, here we give some results about the necessary and sufficient conditions for the ergodicity of uniformly differentiable functions modulo p on p-adic integers $\mathbb{Z}_p$. The class of uniformly differentiable functions modulo p includes the space of rational functions, so we also give the application of the obtained results to rational functions in $\mathbb{Z}_3$ and $\mathbb{Z}_5$. Finally, some classes of 1-Lipschitz functions that preserve measure on the group of p-adic integers $\mathbb{Z}_p$ are considered, and necessary and sufficient conditions for their ergodicity in terms of their Van der Put coefficients are established.

1. INTRODUCTION

We recall some facts about the ring of p-adic integers $\mathbb{Z}_p$. Let p be a fixed prime number. The p-adic ordinal or valuation of $0 \neq x \in \mathbb{Z}$ we define as

$$ord_p x = max\{r : p^r | x\} \geq 0.$$

If $a/b \in \mathbb{Q}$, then we define p-adic valuation as

$$ord_p \frac{a}{b} = ord_p a - ord_p b.$$

2020 *Mathematics Subject Classification.* 11S82,37A05.
Key words and phrases. p-adic dynamical system, ergodic function, uniformly differentiable functions modulo p, Van der Put basis, 1-Lipschitz function.

Definition 1.1. *Let $x \in \mathbb{Q}$. p-adic absolute value of x is given by*

$$|x|_p = \begin{cases} p^{-ord_p x}, & x \neq 0, \\ 0, & x = 0. \end{cases}$$

The p-adic absolute value is non-Archimedean and it induces a metric

$$\rho(x, y) = |x - y|_p.$$

Definition 1.2. *The completion of $\mathbb{Q}$, with respect to the p-adic norm is the field of p-adic numbers, $\mathbb{Q}_p$.*

Definition 1.3. *The unit disk about $0 \in \mathbb{Q}_p$ is called the set of p-adic integers and it is denoted by $\mathbb{Z}_p$, i.e.*

$$\mathbb{Z}_p = \{\alpha \in \mathbb{Q}_p : |\alpha|_p \leq 1\}.$$

Theorem 1.1. *Every p-adic number $\alpha \in \mathbb{Q}_p$ has unique p-adic representation*

$$\alpha = \alpha_{-r} p^{-r} + \alpha_{1-r} p^{1-r} + \alpha_{2-r} p^{2-r} + \ldots + \alpha_{-1} p^{-1} + \alpha_0 + \alpha_1 p^1 + \alpha_2 p^2 + \ldots$$

with $\alpha_n \in \mathbb{Z}$ i $0 \leq \alpha_n \leq (p-1)$. Moreover, $\alpha \in \mathbb{Z}_p$ if and only if $\alpha_{-r} = 0$, for all $r > 0$.

Definition 1.4. *A function $f : \mathbb{Z}_p \to \mathbb{Z}_p$ is said to be 1-Lipschitz if for all $x, y \in \mathbb{Z}_p$ we have*

$$|f(x) - f(y)|_p \leq |x - y|_p.$$

Definition 1.5. *Let $\mathbb{Z}_p$ be the ring of p-adic integers endowed with its ultra-metric norm $|.|$ and natural probability measure μ.*

(1) *A bijective function $f : \mathbb{Z}_p \to \mathbb{Z}_p$ is said to be measure preserving if and only if $\mu(f^{-1}(S)) = \mu(S)$ for every measurable subset S of $\mathbb{Z}_p$.*
(2) *A measure preserving function is said to be ergodic if it has no proper invariant subset, i.e. $\mu(S) = 1$ or $\mu(S) = 0$ for every measurable subset $S \subset \mathbb{Z}_p$ such that $f^{-1}(S) = S$.*

Definition 1.6. *The dynamical system $(\mathbb{Z}_p, \mu, f)$ is called minimal if $S = \emptyset$ or $S = \mathbb{Z}_p$ whenever is S a closed invariant set, or equivalently, every orbit $Orb_f(x) = \{f^n(x) | n \in \mathbb{Z}\}$ is dense in $\mathbb{Z}_p$.*

Definition 1.7. *A function $f : \mathbb{Z}_p \to \mathbb{Z}_p$ is said to be bijective modulo p^n, where n is a positive integer if for arbitrary $x \in \mathbb{Z}_p$ the elements $x, f(x), \ldots, f^{p^n-1}(x)$ are representatives of distinct classes of $\mathbb{Z}_p / p^n \mathbb{Z}_p$.*

Definition 1.8. *A function $f : \mathbb{Z}_p \to \mathbb{Z}_p$ is said to be transitive modulo p^n if it is bijective modulo p^n and the set $x, f(x), \ldots, f^{p^n-1}(x)$ is composed of only one cycle. In other words, $f^{p^n}(x) = x \pmod{p^n}$, but $f^r(x) \neq x \pmod{p^n}$, for all $r < p^n$.*

We recall that in [2, Theorem 1.1.] and [3, Proposition 4.35.] it is proved that a 1-Lipschitz measure preserving function is ergodic if and only if it is transitive modulo p^n for every positive integer n. Some equivalent definitions of 1-Lipschitz measure preserving and ergodic functions are presented in [2], [1], [3], and [5].

Proposition 1.1. Let $f:\mathbb{Z}_p\to\mathbb{Z}_p$ be a polynomial. Then $(\mathbb{Z}_p,f)$ is minimal if and only if $(\mathbb{Z}/p^{\delta}\mathbb{Z},f_{/\delta})$ is minimal, where $\delta=2$, if $p>3$ and $\delta=3$, if $p\in\{2,3\}$.

Proof. See [5]. □

In [5] it is proved that the 1-Lipschitz function $f:\mathbb{Z}_p\to\mathbb{Z}_p$ is minimal if and only if it is ergodic for Haar measure.

We recall the Van der Put representation for functions on $\mathbb{Z}_p$ (see [7]). If the p-adic expansion of the positive integer k is given by

$$k=\sum_{i=0}^{s}k_ip^i,\ \ 0\le k_i<p,\ \ k_s\neq 0,$$

then we define $q(k)=k_sp^s$.

For every function $f:\mathbb{Z}_p\to\mathbb{Z}_p$ we define the coefficients

$$B_k=\begin{cases} f(k), & k\in\{0,\dots,p-1\};\\ f(k)-f(k-q(k)), & k\ge p.\end{cases}$$

In this way the function f can be represented in the so called Van der Put basis as follows

$$f(x)=\sum_{k=0}^{\infty}B_k\chi(k,x),$$

where if $k>0$,

$$\chi(k,x)=\begin{cases}1, & |x-k|\le p^{-\lfloor\log_p k\rfloor-1};\\ 0, & \text{otherwise.}\end{cases}$$

For $k=0$ we have

$$\chi(0,x)=\begin{cases}1, & |x|\le p^{-1};\\ 0, & \text{otherwise.}\end{cases}$$

A function $f:\mathbb{Z}_p\to\mathbb{Z}_p$ is said to be uniformly differentiable modulo p^k if there exists a positive integer N and a function $\partial_k f:\mathbb{Z}_p\to\mathbb{Q}_p$ such that for all $r\ge N$ and $h\in\mathbb{Z}_p$, we have

$$f(u+p^rh)=f(u)+p^rh\partial_kf(u)\ \ (mod\ p^{k+r}),\forall u\in\mathbb{Z}_p.$$

The smallest integer N satisfying this property is denoted by $N_k(f)$. In [3, Proposition 3.41.] it was proved that if f is 1-Lipschitz , then $\partial_k f$ takes its values in $\mathbb{Z}_p$.

2. Ergodic uniformly differentiable functions modulo p on $\mathbb{Z}_p$

Theorem 2.1. *Let f be an isometric and uniformly differentiable function modulo p on $\mathbb{Z}_p$, where $N_1(f)=1$. Then, f is ergodic on $\mathbb{Z}_p$ if and only if the following conditions are satisfied:*

(1) *f is transitive modulo p.*

(2) *For every positive integer k, $f^{p^k}(0)\neq 0\ \ (mod\ p^{k+1})$.*

(3) *For every positive integer k,*

$$\frac{\prod_{j=0}^{p^k-1} B_{j+p^k}}{(p^k)^{p^k}} = 1 \pmod p.$$

Proof. Conditions (1) and (2) are obviously necessary. According to [3, Proposition 4.35.] it suffices to prove that for every fixed positive integer k, if f is transitive modulo p^k, then it is transitive modulo p^{k+1} if and only if

$$\frac{\prod_{j=0}^{p^k-1} B_{j+p^k}}{(p^k)^{p^k}} = 1 \pmod p. \tag{2.1}$$

Namely, it suffices to prove that if $f^{p^k}(0) \neq 0 \pmod{p^{k+1}}$, then (2.1) holds if and only if for every $l \in \{2,\dots,p-1\}$, $f^{lp^k}(0) \neq 0 \pmod{p^{k+1}}$.

Assume that f is transitive modulo p^k for some arbitrary and fixed $k \geq 1$. Let $\{t_0,\dots,t_{p^k-1}\}$ be representatives of $p^k\mathbb{Z}_p$-cosets such that $f(t_i) = t_{i+1} \pmod{p^k}$, for $0 \leq i \leq p^k-2$ and $f(t_{p^k-1}) = t_0 \pmod{p^k}$. We may choose

$$\{t_0,\dots,t_{p^k-1}\} = \{0,\dots,p^k-1\}. \tag{2.2}$$

Our first task is to prove that for all $l \in \{1,\dots,p-1\}$ and $s \in \{0,\dots,p^k-1\}$

$$\begin{aligned} f^{lp^k}(0) \;=\; & f(t_{p^k-1}) + \sum_{i=1}^{s} \frac{f(t_{p^k-i-1}) - t_{p^k-i}}{p^{ki}} \prod_{j=1}^{i} B_{t_{p^k-j}+p^k} \\ & + \frac{f^{lp^k-s-1}(0) - t_{p^k-s-1}}{p^{k(s+1)}} \prod_{j=1}^{s+1} B_{t_{p^k-j}+p^k} \pmod{p^{k+1}}. \end{aligned} \tag{2.3}$$

We know from [6, Formula (4.3)] that for all $j < p^k$, $r \in \{1,\dots,p-1\}$, $B_{j+rp^k} = rB_{j+p^k} \pmod{p^{k+1}}$. It follows that

$$\begin{aligned} f^{lp^k}(0) \;&=\; f(f^{lp^k-1}(0)) = f(t_{p^k-1} + p^k \frac{f^{lp^k-1}(0) - t_{p^k-1}}{p^k}) \\ &=\; f(t_{p^k-1}) + \frac{f^{lp^k-1}(0) - t_{p^k-1}}{p^k} B_{t_{p^k-1}+p^k} \pmod{p^{k+1}}. \end{aligned}$$

Hence, (2.3) holds for $s = 0$. Assume it is true for some $s \in \{0,\dots,p^k-2\}$. Applying [6, Formula (4.3)] we get

$$\begin{aligned} f^{lp^k-s-1}(0) \;&=\; f(f^{lp^k-s-2}(0)) = f(t_{p^k-s-2} + p^k \frac{f^{lp^k-s-2}(0) - t_{p^k-s-2}}{p^k}) \\ &=\; f(t_{p^k-s-2}) + \frac{f^{lp^k-s-2}(0) - t_{p^k-s-2}}{p^k} B_{t_{p^k-s-2}+p^k} \pmod{p^{k+1}}. \end{aligned}$$

Hence, (2.3) becomes

$$f^{lp^k}(0) = f(t_{p^k-1}) + \sum_{i=1}^{s+1} \frac{f(t_{p^k-i-1}) - t_{p^k-i}}{p^{ki}} \prod_{j=1}^{i} B_{t_{p^k-j}+p^k} + \frac{f^{lp^k-s-2}(0) - t_{p^k-s-2}}{p^{k(s+2)}} \prod_{j=1}^{s+2} B_{t_{p^k-j}+p^k} \pmod{p^{k+1}}.$$

Then, (2.3) holds for every $s \in \{0, \dots, p^k - 1\}$.

Now, for $l = 1$ and $s = p^k - 2$, (2.3) takes the form

$$\begin{aligned} f^{p^k}(0) &= f(t_{p^k-1}) + \sum_{i=1}^{p^k-2} \frac{f(t_{p^k-i-1}) - t_{p^k-i}}{p^{ki}} \prod_{j=1}^{i} B_{t_{p^k-j}+p^k} \\ &\quad + \frac{f(0) - t_1}{p^{k(p^k-1)}} \prod_{j=1}^{p^k-1} B_{t_{p^k-j}+p^k} \pmod{p^{k+1}} \\ &= f(t_{p^k-1}) + \sum_{i=1}^{p^k-1} \frac{f(t_{p^k-i-1}) - t_{p^k-i}}{p^{ki}} \prod_{j=1}^{i} B_{t_{p^k-j}+p^k}. \end{aligned} \tag{2.4}$$

On the other hand for $l \geq 2$ and $s = p^k - 2$, (2.3) takes the form

$$f^{lp^k}(0) = f(t_{p^k-1}) + \sum_{i=1}^{p^k-2} \frac{f(t_{p^k-i-1}) - t_{p^k-i}}{p^{ki}} \prod_{j=1}^{i} B_{t_{p^k-j}+p^k} + \frac{f^{(l-1)p^k+1}(0) - t_1}{p^{k(p^k-1)}} \prod_{j=1}^{p^k-1} B_{t_{p^k-j}+p^k} \pmod{p^{k+1}}.$$

Applying the same techniques as above we get

$$\frac{f^{(l-1)p^k+1}(0) - t_1}{p^{k(p^k-1)}} = \frac{f(0 + p^k \frac{f^{(l-1)p^k}(0)}{p^k}) - t_1}{p^{k(p^k-1)}} = \frac{f(0) - t_1 + \frac{f^{(l-1)p^k}(0)}{p^k} B_{p^k}}{p^{k(p^k-1)}} \pmod{p^{k+1}}.$$

Therefore, applying (2.2), (2.3) becomes

$$f^{lp^k}(0) = f(t_{p^k-1}) + \sum_{i=1}^{p^k-1} \frac{f(t_{p^k-i-1}) - t_{p^k-i}}{p^{ki}} \prod_{j=1}^{i} B_{t_{p^k-j}+p^k} + \frac{f^{(l-1)p^k}(0)}{p^{kp^k}} \prod_{j=0}^{p^k-1} B_{j+p^k} \pmod{p^{k+1}}.$$

Then, (2.4) yields

$$f^{lp^k}(0) = f^{p^k}(0) + \frac{f^{(l-1)p^k}(0)}{p^{kp^k}} \prod_{j=0}^{p^k-1} B_{j+p^k} \pmod{p^{k+1}}.$$

Replacing l by $l-1$, then $l-1$ by $l-2$,...gives

$$f^{lp^k}(0) = f^{p^k}(0)\left(1 + \frac{\prod_{j=0}^{p^k-1} B_{j+p^k}}{p^{kp^k}} + \ldots + \left(\frac{\prod_{j=0}^{p^k-1} B_{j+p^k}}{p^{kp^k}}\right)^{l-1}\right).$$

We conclude that if $f^{p^k}(0) \neq 0 \pmod{p^{k+1}}$, then (2.1) holds if and only if for every $l \in \{2,\ldots,p-1\}$, $f^{lp^k}(0) \neq 0 \pmod{p^{k+1}}$. □

Remark 2.1. Notice that for any analytic function f we have $\frac{B_{i+p^k}}{p^k} = f'(i) \pmod{p^k}$, for every positive integer k and every $i \in \{0,\ldots,p^k-1\}$, because $f(i+p^k) = f(i) + p^k f'(i) \pmod{p^{2k}}$.

2.1. **Ergodic rational functions on $\mathbb{Z}_3$**

In the following corollaries we study ergodic rational functions $R = \frac{P}{Q}$ on $\mathbb{Z}_3$ where the numerator P is not an ergodic polynomial. We study cases where the denominator is always a unit. Without loss of generality we may assume that $P(0) = Q(0) = 1$.

Corollary 2.1. *Let P be an isometric polynomial on $\mathbb{Z}_3$. Assume that P is transitive modulo 3, $P(0) = 1$,*

$$P^3(0) = 0 \pmod{9}$$

and

$$P'(0)P'(1)P'(2) = 1 \pmod{3}.$$

Then, $R = \dfrac{P}{Q}$ is ergodic if the following conditions are satisfied

(1) $Q(\mathbb{Z}_3) \subseteq 1 + 3\mathbb{Z}_3$,
(2) $Q'(x) = 0 \pmod{3}$, *for every* $x \in \mathbb{Z}_3$,
(3) $Q(1) \neq 1 \pmod{9}$,
(4) $P^3(0) + P^3(3) + P^3(6) + 2P'(2)\left(\frac{1}{Q(1)} + \frac{1}{Q(4)} + \frac{1}{Q(7)} - 3\right)$
$+P'(1)P'(2)\left(\frac{1}{Q(0)} + \frac{1}{Q(3)} + \frac{1}{Q(6)} - 3\right) \neq 0 \pmod{3^3}$.

Proof. See [9]. □

Example 2.1. *Let $P(x) = 3x^2 + x + 1$. This is an isometric polynomial, transitive modulo 3, and it satisfies the conditions*

(i) $P(0) = 1$,
(ii) $P^3(0) = 0 \pmod{9}$,
(iii) $P'(0)P'(1)P'(2) = 1 \pmod{3}$.

According to Corollary 2.1, the function $\dfrac{3x^2+x+1}{Q(x)}$ would be an ergodic function if the polynomial $Q(x)$ satisfies the following conditions:

(1) $Q(\mathbb{Z}_3) \subseteq 1+3\mathbb{Z}_3$,
(2) $Q'(x) = 0 \pmod 3$ *for* $\forall x \in \mathbb{Z}_3$,
(3) $Q(1) \neq 1 \pmod 9$, *and*
(4) $9 - \left(\frac{1}{Q(1)} + \frac{1}{Q(4)} + \frac{1}{Q(7)}\right) + 10\left(\frac{1}{Q(0)} + \frac{1}{Q(3)} + \frac{1}{Q(6)}\right) \neq 0 \pmod{27}$.

One of the polynomials that satisfy these conditions is $Q(x) = 3x^3 + 1$. *So, the function* $R(x) = \dfrac{3x^2+x+1}{3x^3+1}$ *is an ergodic function.*

Corollary 2.2. *Let* P *be an isometric polynomial on* $\mathbb{Z}_3$. *Assume that* $P(0) = 1$, P *is transitive modulo 9, but not modulo* 3^3.

Then, $R = \dfrac{P}{Q}$ *is ergodic if the following conditions are satisfied*

(1) $Q(\mathbb{Z}_3) \subseteq 1+3\mathbb{Z}_3$,
(2) $Q'(x) = 0 \pmod 3$, *for every* $x \in \mathbb{Z}_3$,
(3) $Q(1) = 1 + P(2)P'(2) + P(1) - 2 \pmod 9$,
(4) $2P'(2)\left(\frac{1}{Q(1)} + \frac{1}{Q(4)} + \frac{1}{Q(7)} - 3\right) + P'(1)P'(2)\left(\frac{1}{Q(0)} + \frac{1}{Q(3)} + \frac{1}{Q(6)} - 3\right) \neq 0 \pmod{3^3}$.

Proof. See [9]. □

2.2. **Ergodic rational functions on** $\mathbb{Z}_5$

Corollary 2.3. *Let* $P(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + 1$ *be an isometric polynomial on* $\mathbb{Z}_5$. *Let* t_i *be representatives of* $5\mathbb{Z}_5$*-cosets such that* $P(t_i) = t_{i+1} \pmod 5$ *and* $P(t_4) = t_0 = 0 \pmod 5$. *Assume that*

$$P^5(t_0) = t_0 \pmod{25}$$

and

$$P'(t_0)P'(t_1)P'(t_2)P'(t_3)P'(t_4) = 1 \pmod 5.$$

Then $R = \dfrac{P}{Q}$ *is ergodic if the polynomial* $Q(x)$ *satisfies the following conditions*

(1) $Q(\mathbb{Z}_5) \subseteq 1+5\mathbb{Z}_5$,
(2) $Q'(x) = 0 \pmod 5$, *for all* $x \in \mathbb{Z}_5$,
(3) $t_4\left(1 - \frac{1}{Q(t_3)}\right) + t_3P'(t_3)\left(1 - \frac{1}{Q(t_2)}\right) + t_2P'(t_3)P'(t_2)\left(1 - \frac{1}{Q(t_1)}\right) \neq 0 \pmod{25}$.

Proof. See [11]. □

Remark 2.2. Let $P(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + 1$. If we introduce the notation

$$\sum_{i \in 1+4\mathbb{N}} a_i = A_1, \quad \sum_{i \in 2+4\mathbb{N}} a_i = A_2, \quad \sum_{i \in 3+4\mathbb{N}} a_i = A_3, \quad \sum_{i \in 4\mathbb{N}} a_i = A_4,$$

then, according [4, Proposition 4.2], we have six classes of transitive polynomials modulo 5. Hence, the third condition from the previous Corollary for all six classes , can be written in the following way:

$$\begin{cases} 4\Big(1-\frac{1}{Q(3)}\Big)+3P'(3)\Big(1-\frac{1}{Q(2)}\Big)+2P'(2)P'(3)\Big(1-\frac{1}{Q(1)}\Big)\neq 0 \pmod{25}, \\ \text{if} \quad A_1\equiv 1, \quad A_2\equiv 0, \quad A_3\equiv 0 \quad \text{i} \quad A_4\equiv 0 \pmod 5; \\ 3\Big(1-\frac{1}{Q(4)}\Big)+4P'(4)\Big(1-\frac{1}{Q(2)}\Big)+2P'(2)P'(4)\Big(1-\frac{1}{Q(1)}\Big)\neq 0 \pmod{25}, \\ \text{if} \quad A_1\equiv 4, \quad A_2\equiv 4, \quad A_3\equiv 3 \quad \text{i} \quad A_4\equiv 0 \pmod 5, \\ 4\Big(1-\frac{1}{Q(2)}\Big)+2P'(2)\Big(1-\frac{1}{Q(3)}\Big)+3P'(2)P'(3)\Big(1-\frac{1}{Q(1)}\Big)\neq 0 \pmod{25}, \\ \text{if} \quad A_1\equiv 1, \quad A_2\equiv 3, \quad A_3\equiv 3 \quad \text{i} \quad A_4\equiv 0 \pmod 5, \\ 2\Big(1-\frac{1}{Q(4)}\Big)+4P'(4)\Big(1-\frac{1}{Q(3)}\Big)+3P'(3)P'(4)\Big(1-\frac{1}{Q(1)}\Big)\neq 0 \pmod{25}, \\ \text{if} \quad A_1\equiv 1, \quad A_2\equiv 4, \quad A_3\equiv 2 \quad \text{i} \quad A_4\equiv 0 \pmod 5, \\ 3\Big(1-\frac{1}{Q(2)}\Big)+2P'(2)\Big(1-\frac{1}{Q(4)}\Big)+4P'(2)P'(4)\Big(1-\frac{1}{Q(1)}\Big)\neq 0 \pmod{25}, \\ \text{if} \quad A_1\equiv 4, \quad A_2\equiv 2, \quad A_3\equiv 2 \quad \text{i} \quad A_4\equiv 0 \pmod 5, \\ 2\Big(1-\frac{1}{Q(3)}\Big)+3P'(2)\Big(1-\frac{1}{Q(4)}\Big)+4P'(2)P'(3)\Big(1-\frac{1}{Q(1)}\Big)\neq 0 \pmod{25}, \\ \text{if} \quad A_1\equiv 0, \quad A_2\equiv 0, \quad A_3\equiv 3 \quad \text{i} \quad A_4\equiv 0 \pmod 5. \end{cases}$$

Example 2.2. *Let $P(x)=2x^7+3x^6+5x^5+5x^4+3x^3+2x^2+x+1$. This is an isometric polynomial, transitive modulo 5 and it satisfies conditions*

(i) $P(i)=i+1$ *for all* $i\in\{0,1,2,3,4\}$,
(ii) $P^5(0)=0 \pmod{25}$,
(iii) $P'(0)P'(1)P'(2)P'(3)P'(4)=1 \pmod 5$.

The function $R=\dfrac{P}{Q}$ would be ergodic if the polynomial $Q(x)$ satisfies the conditions of Corollary 2.3. One of the polynomials which satisfies these conditions is $Q(x)=10x^4+5x^2+1$. Hence, such a function $R(x)$ is ergodic.

The next result is in the case when the numerator is not transitive modulo 5.

Corollary 2.4. *Let P be an isometric polynomial on $\mathbb{Z}_5$. Assume that P is not transitive modulo 5 and let $2\leq i\leq 4$ be a fixed number such that*

(i) $P(k)=(k+1)i \pmod 5$, $0\leq k\leq 4$ *and*
(ii) $P'(0)P'(1)P'(2)P'(3)P'(4)=i \pmod 5$.

Then $R=\dfrac{P}{Q}$ is ergodic if the polynomial $Q(x)$ satisfies conditions

(1) $Q(\mathbb{Z}_5)\subseteq i+5\mathbb{Z}_5$,
(2) $Q'(x)=0 \pmod 5$, *for all* $x\in\mathbb{Z}_5$,
(3) $1+\sum_{s=0}^{4}\frac{l_{4-s}}{i^s}\prod_{j=1}^{s}P'(t_{5-j})\neq 0 \pmod 5$, *where* $l_k\in\{0,\ldots,4\}$ *satisfy*
$P(k)=(k+1+5l_k)Q(k) \pmod{25}$.

Proof. See [11]. □

3. On some classes of 1-Lipschitz measure preserving ergodic functions on $\mathbb{Z}_p$

Lemma 3.1. *Let f be a 1-Lipschitz measure preserving function on $\mathbb{Z}_p$. Assume that f is transitive modulo p and satisfies*

$$B_{i+lp^k} = lB_{i_0+p^k} \pmod{p^{k+1}},\ \forall k \geq 1,\ \forall l \in \{1,\dots,p-1\},\ \forall i < p^k, \tag{3.1}$$

where $i_0 \in \{0,\dots,p-1\}$ is a unique integer depending on i and satisfying $i = i_0 \pmod p$. Then,

(1) *for every $x \in \mathbb{Z}_p$, $l \in \{1,\dots,p-1\}$ and $k \geq 1$,*

$$f(x+lp^k) = f(x) + lB_{x_0+p^k} \pmod{p^{k+1}}, \tag{3.2}$$

where $x_0 \in \{0,\dots,p-1\}$ is a unique integer depending on x and satisfying $x = x_0 \pmod p$,

(2) *for every $x \in \mathbb{Z}_p$, $l \in \{1,\dots,p-1\}$ and $n,k \geq 1$,*

$$f^n(x+lp^k) = f^n(x) + l\left(\prod_{i=0}^{p-1}\frac{B_{i+p^k}}{p^k}\right)^{n-m}\left(\prod_{i=s}^{m+s-1}\frac{B_{y_i+p^k}}{p^k}\right)p^k \pmod{p^{k+1}}, \tag{3.3}$$

where the sequence $(y_i)_i$ is such that $\{y_i, i \geq 0\} = \{0,\dots,p-1\}$, $y_0 = 0$ and $y_{i+1} = f(i) \pmod p$, for every nonnegative integer i. The numbers $s,m \in \{0,\dots,p-1\}$ are such that $m = n \pmod p$ and $x = y_s \pmod p$. The second product is taken to be equal to 1 if $m = 0$.

Proof. See [10]. □

Theorem 3.1. *Let f be a function satisfying the conditions of Lemma 3.1. Then, under the notation of Lemma 3.1, f is ergodic if and only if the following conditions are satisfied*

(1)
$$\prod_{i=0}^{p-1}\frac{B_{i+p^k}}{p^k} = 1 \pmod p, \forall k \geq 1,$$

(2)
$$\sum_{s=0}^{p-1}\prod_{t=s+1}^{p-1}\frac{B_{y_t+p}}{p}B_{y_s} \neq \sum_{s=0}^{p-1}\prod_{t=s}^{p-1}\frac{B_{y_t+p}}{p}y_s \pmod{p^2},$$

and

$$\sum_{s=0}^{p-1}\prod_{t=s+1}^{p-1}\frac{B_{y_t+p^k}}{p^k}\left(p^{k-1}B_{y_s} + \sum_{l=1}^{k-1}p^{k-l-1}\sum_{\substack{m\in\{p^l,\dots,p^{l+1}-1\}\\ m=y_s \pmod p}}B_m\right)$$
$$\neq \sum_{s=0}^{p-1}\prod_{t=s}^{p-1}\frac{B_{y_t+p^k}}{p^k}\left(\frac{p}{2}(p-1)+y_s\right)p^{k-1} \pmod{p^{k+1}},\ \forall k \geq 2.$$

Proof. Since f is transitive modulo p, it can be easily seen ([9]) that f is ergodic if and only if

$$f^{lp^k}(0) \neq 0 \pmod{p^{k+1}},\ \forall k \geq 1,\ \forall l \in \{1,\dots,p-1\}.$$

Following the steps made in the proof of [9, Theorem 2.1], we can see that for every $l \in \{1,\dots,p-1\}$ and $k \geq 1$,

$$f^{lp^k}(0) = f^{p^k}(0)\left(1 + \prod_{j=0}^{p^k-1} \frac{B_{j+p^k}}{p^k} + \dots + \left(\prod_{j=0}^{p^k-1} \frac{B_{j+p^k}}{p^k}\right)^{l-1}\right).$$

Hence, f is ergodic if and only if for every $l \in \{1,\dots,p-1\}$ and $k \geq 1$

$$f^{p^k}(0) \neq 0 \pmod{p^{k+1}}, \tag{3.4}$$

and

$$\prod_{j=0}^{p^k-1} \frac{B_{j+p^k}}{p^k} = 1 \pmod{p}. \tag{3.5}$$

We first prove that (3.5) is equivalent to condition (1). According to (3.1), identity (3.5) is equivalent to

$$\left(\prod_{j=0}^{p-1} \frac{B_{j+p^k}}{p^k}\right)^{p^{k-1}} = 1 \pmod{p}.$$

Since

$$\left(\prod_{j=0}^{p-1} \frac{B_{j+p^k}}{p^k}\right)^{p-1} = 1 \pmod{p},$$

then, (3.5) is equivalent to

$$\left(\prod_{j=0}^{p-1} \frac{B_{j+p^k}}{p^k}\right)^{(p-1)(1+\dots+p^{k-2})+1} = \prod_{j=0}^{p-1} \frac{B_{j+p^k}}{p^k} \pmod{p} = 1 \pmod{p},\ \forall k \geq 1.$$

It remains to verify that (3.4) is equivalent to condition (2).

By induction on $k \geq 1$ it can be seen that

$$\sum_{\substack{m\in\{0,\dots,p^k-1\}\\ m=y_s \pmod{p}}} f(m) = p^{k-1}B_{y_s} + \sum_{l=1}^{k-1} p^{k-l-1} \sum_{\substack{m\in\{p^l,\dots,p^{l+1}-1\}\\ m=y_s \pmod{p}}} B_m. \tag{3.6}$$

On the other hand it can be easily seen that

$$\sum_{\substack{m\in\{0,\dots,p^k-1\}\\ m=y_s \pmod{p}}} m = \left(\frac{p}{2}(p-1)+y_s\right)p^{k-1} \pmod{p^{k+1}}. \tag{3.7}$$

If we prove that for every $k \geq 1$,

$$f^{p^k}(0)=\sum_{s=0}^{p-1}\prod_{t=s+1}^{p-1}\frac{B_{y_t+p^k}}{p^k}\sum_{\substack{m\in\{0,\ldots,p^k-1\}\\ m=y_s \pmod p}} f(m)-\sum_{s=0}^{p-1}\prod_{t=s}^{p-1}\frac{B_{y_t+p^k}}{p^k}\sum_{\substack{m\in\{0,\ldots,p^k-1\}\\ m=y_s \pmod p}} m \pmod{p^{k+1}}, \tag{3.8}$$

then condition (2) can be obtained by a combination of (3.4), (3.8), (3.6) and (3.7).

We first consider the case when $k=1$. Formula (3.3) yields

$$\begin{aligned} f^p(0) &= f^{p-1}(f(0)) = f^{p-1}(y_1+f(0)-y_1)\\ &= f^{p-1}(y_1)+(f(0)-y_1)\prod_{t=1}^{p-1}\frac{B_{y_t+p}}{p} \pmod{p^2}. \end{aligned} \tag{3.9}$$

In a similar way, for every $r\in\{1,\ldots,p-2\}$,

$$f^{p-r}(y_r)=f^{p-r-1}(y_{r+1})+(f(y_r)-y_{r+1})\prod_{t=r+1}^{p-1}\frac{B_{y_t+p}}{p} \pmod{p^2}. \tag{3.10}$$

Combining (3.9) and (3.10) gives

$$\begin{aligned} f^p(0) &= f(y_{p-1})+\sum_{r=0}^{p-2}(f(y_r)-y_{r+1})\prod_{t=r+1}^{p-1}\frac{B_{y_t+p}}{p} \pmod{p^2}\\ &= \sum_{r=0}^{p-1}\prod_{t=r+1}^{p-1}\frac{B_{y_t+p}}{p}f(y_r)-\sum_{r=1}^{p-1}\prod_{t=r}^{p-1}\frac{B_{y_t+p}}{p}y_r \pmod{p^2}. \end{aligned}$$

Let $k\geq 2$ be such that f is transitive modulo p^k. Proceeding as in the proof of [8, Theorem 2.2], we put $i_j^s=f^s(j\cdot p^k) \pmod{p^{k+1}}$, for $j\in\{0,\ldots,p-1\}$ and $s\in\{0,\ldots,p^{k-1}-1\}$, where $\{i_j^s, j\in\{0,\ldots,p-1\}, s\in\{0,\ldots,p^{k-1}-1\}\}=\{0,\ldots,p^k-1\}$. Let $\{s_1,\ldots,s_{p-1}\}=\{1,\ldots,p-1\}$ be such that

$$f^{p^{k-1}}(0)=s_1p^{k-1} \pmod{p^k}, \tag{3.11}$$

and for $i\in\{1,\ldots,p-2\}$,

$$f^{p^{k-1}}(s_ip^{k-1})=s_{i+1}p^{k-1} \pmod{p^k}. \tag{3.12}$$

It is clear that

$$f^{p^{k-1}}(s_{p-1}p^{k-1})=0 \pmod{p^k}. \tag{3.13}$$

Combining (3.3), (3.11) and condition (1) gives

$$\begin{aligned} f^{p^k}(0) &= f^{p^{k-1}(p-1)}(f^{p^{k-1}}(0)) = f^{p^{k-1}(p-1)}(s_1p^{k-1}+f^{p^{k-1}}(0)-s_1p^{k-1})\\ &= f^{p^{k-1}(p-1)}(s_1p^{k-1})+f^{p^{k-1}}(0)-s_1p^{k-1} \pmod{p^{k+1}}. \end{aligned}$$

Similarly, combining (3.12) in a recursive way with (3.3) and condition (1), we obtain

$$\begin{aligned} f^{p^k}(0) &= f^{p^{k-1}}(s_{p-1}p^{k-1})+f^{p^{k-1}}(s_{p-2}p^{k-1})\\ &\quad -s_{p-1}p^{k-1}+\ldots+f^{p^{k-1}}(0)-s_1p^{k-1} \pmod{p^{k+1}}\\ &= \sum_{j=0}^{p-1}f^{p^{k-1}}(i_j^0)-\sum_{j=0}^{p-1}i_j^0 \pmod{p^{k+1}}. \end{aligned} \tag{3.14}$$

For every $j \in \{0,\dots,p-1\}$, since $i_j^1 = 1 \pmod p$, an application of (3.3) and condition (1) gives

$$\begin{aligned} f^{p^{k-1}}(i_j^0) &= f^{p^{k-1}-1}(i_j^1 + f(i_j^0) - i_j^1) \\ &= f^{p^{k-1}-1}(i_j^1) + (f(i_j^0) - i_j^1) \prod_{t=1}^{p^{k-1}-1} \frac{B_{y_t+p^k}}{p^k} \pmod{p^{k+1}}. \end{aligned} \tag{3.15}$$

In a similar way, for all $r \in \{1,\dots,p^{k-1}-2\}$,

$$\begin{aligned} f^{p^{k-1}-r}(i_j^r) &= f^{p^{k-1}-r-1}(i_j^{r+1} + f(i_j^r) - i_j^{r+1}) \\ &= f^{p^{k-1}-r-1}(i_j^{r+1}) + (f(i_j^r) - i_j^{r+1}) \prod_{t=r+1}^{p^{k-1}-1} \frac{B_{y_t+p^k}}{p^k} \pmod{p^{k+1}}. \end{aligned} \tag{3.16}$$

Combining (3.15) and (3.16) we obtain

$$f^{p^{k-1}}(i_j^0) = f(i_j^{p^{k-1}-1}) + \sum_{r=0}^{p^{k-1}-2} (f(i_j^r) - i_j^{r+1}) \prod_{t=r+1}^{p^{k-1}-1} \frac{B_{y_t+p^k}}{p^k} \pmod{p^{k+1}}.$$

Therefore,

$$\begin{aligned} \sum_{j=0}^{p-1} f^{p^{k-1}}(i_j^0) &= \sum_{j=0}^{p-1} f(i_j^{p^{k-1}-1}) + \sum_{r=0}^{p^{k-1}-2} \prod_{t=r+1}^{p^{k-1}-1} \frac{B_{y_t+p^k}}{p^k} \sum_{j=0}^{p-1} f(i_j^r) \\ &\quad - \sum_{r=0}^{p^{k-1}-2} \prod_{t=r+1}^{p^{k-1}-1} \frac{B_{y_t+p^k}}{p^k} \sum_{j=0}^{p-1} i_j^{r+1} \pmod{p^{k+1}}. \end{aligned}$$

Since for all $r \in \{1,\dots,p^{k-1}-2\}$, $j \in \{0,\dots,p-1\}$, $i_j^r = y_r \pmod p$, we get

$$\begin{aligned} \sum_{j=0}^{p-1} f^{p^{k-1}}(i_j^0) &= \sum_{r=0}^{p^{k-1}-1} \prod_{t=r+1}^{p^{k-1}-1} \frac{B_{y_t+p^k}}{p^k} \sum_{j=0}^{p-1} f(i_j^r) \\ &\quad - \sum_{r=1}^{p^{k-1}-1} \prod_{t=r}^{p^{k-1}-1} \frac{B_{y_t+p^k}}{p^k} \sum_{j=0}^{p-1} i_j^r \pmod{p^{k+1}} \\ &= \sum_{s=0}^{p-1} \prod_{t=s+1}^{p-1} \frac{B_{y_t+p^k}}{p^k} \sum_{\substack{m\in\{0,\dots,p^k-1\} \\ m=y_s \pmod p}} f(m) \\ &\quad - \sum_{s=0}^{p-1} \prod_{t=s}^{p-1} \frac{B_{y_t+p^k}}{p^k} \sum_{\substack{m\in\{0,\dots,p^k-1\} \\ m=y_s \pmod p \\ m\neq 0 \pmod{p^k}}} m \pmod{p^{k+1}}, \end{aligned}$$

where the latter equality follows from the properties of the sequence $(y_i)_i$ and condition (1).

Hence, (3.14) yields (3.8). □

Corollary 3.1. *Let f be a* 1*-Lipschitz measure preserving function on $\mathbb{Z}_p$. Assume that f is transitive modulo p and satisfies*

$$B_{i+lp^k} = lp^k \pmod{p^{k+1}},\ \forall k \geq 1,\ \forall l \in \{1,\dots,p-1\},\ \forall i < p^k. \tag{3.17}$$

Then, f is ergodic if and only if

$$\sum_{m=0}^{p-1} p^{k-1}B_m + \sum_{l=1}^{k-1} p^{k-l-1} \sum_{m=p^l}^{p^{l+1}-1} B_m \neq \frac{p^k}{2}(p-1) \pmod{p^{k+1}},\ \forall k \geq 1. \tag{3.18}$$

Proof. See [10]. □

References

[1] V. Anashin, *Ergodic transformations in the space of p-adic integers*, p-Adic Mathematical Physics, AIP Conf. Proc. **826**, 3–24, 2006.

[2] V. S. Anashin, *Uniformly distributed sequences of p-adic integers*, Math. Notes, **55**, No. 1-2, 109-133, 1994.

[3] V. Anashin, A. Khrennikov, *Applied Algebraic Dynamics*, de Gruyter Expositions in Mathematics 49. Berlin, 2009.

[4] K. Donggyun, K. Youngwoo and S. Kyunghwan, *Minimality of 5-adic polynomial dynamics, Dyn. Syst.* **35**, No. 4, 584-596, 2020.

[5] F. Durand and F. Paccaut, *Minimal polynomial dynamics on the set of 3-adic integers*, Bull. Lond. Math. Soc., **41**, No. 2, 302-314, 2009.

[6] S. Jeong, *Measure-preservation and the existence of a root of p-adic* 1*-Lipschitz functions in Mahler's expansion*, p-Adic Numbers Ultrametric Anal. Appl. **10**, No. 3, 192–208, 2018.

[7] K. Mahler, *p-Adic Numbers and Their Functions*, Cambridge Univ. Press, Cambridge, 1981.

[8] N. Memić, *On some compatible functions on the set of* 3*-Adic integers*, Colloq. Math. **155**, No. 2, 197-214, 2019.

[9] N. Memić, J. Muminović Huremović, *Ergodic uniformly differentiable functions modulo p on $\mathbb{Z}_p$*, p-Adic Numbers Ultrametric Anal. Appl. **12**, No. 1, 49-59, 2020.

[10] N. Memić, J. Muminović Huremović, *On some classes of* 1*-Lipschitz measure-preserving ergodic functions on $\mathbb{Z}_p$*, Asian-European Journal of Mathematics, **16**, No. 9, 2250167, 2022.

[11] J. Muminović Huremović, *On some ergodic rational functions on $\mathbb{Z}_5$*, Advances in Mathematics: Scientific Journal, **12**, No. 10, 2023.

(Received: 17 May, 2024)
(Revised: 16 September, 2024)

Jasmina Muminović Huremović
University of Tuzla
Department of Mathematics
Urfeta Vejzagića 4
75000 Tuzla
Bosnia and Herzegovina
e-mail: *jasmina.muminovic@untz.ba*

PROCEEDINGS OF THE CONFERENCE ON MARCH 14 – INTERNATIONAL DAY OF MATHEMATICS
Special Editions ANUBiH, Book CCXVI, OPMN Book 30
Sarajevo 2024, **p. 81–89** DOI: 10.5644/PI2024.216.10

FINDING A MINIMAL DOMINATING SET OF A GRAPH COMBINING VARIOUS HEURISTIC APPROACHES BASED ON VARIABLE NEIGHBORHOOD SEARCH

ANTON VRDOLJAK

Dedicated to the 75th birthday of our dear Professor Mirjana Vuković

ABSTRACT. Many complex combinatorial optimization problems cannot be solved by traditional optimization techniques. Therefore, these types of problems often require the application and sometimes a combination of other scientific approaches to obtain, at the very least, adequate solutions. As demonstrated in numerous papers, heuristic approaches are usually more efficient and effective compared to classical methods, especially in managing neighboring uncertainty. Therefore, this paper focuses on combining various heuristic approaches based on the basic variable neighborhood search (BVNS) metaheuristic. This paper aims to find a minimal dominating set of a graph, which is a well-known NP-complete problem.

1. INTRODUCTION

Combinatorial optimization problems are often modeled using graphs because many such problems can be represented as searching for an optimal structure within a graph. Next, their invariance to permutation is crucial for such quests.

Definition 1.1. *A graph is an ordered triple $G = (V, E, \varphi)$, where V is a non-empty set of vertices, E is a set of edges that is disjoint from V, and φ is a function that assigns to each edge from E two, not necessarily different, vertices from V. A graph is usually represented by the ordered pair $G = (V, E)$ or simply G.*

Definition 1.2. *Let $G = (V, E)$ be any graph. A set $D \subseteq V$ is a dominating set of a graph G if for every vertex $v \in V \setminus D$ there is at least one $u \in D$ that is adjacent to v, i.e. for which $uv \in E$ holds.*

Definition 1.3. *A dominating set D of a graph G is called a minimal dominating set, if there is no proper subset $D' \subset D$ that is a dominating set of the graph G. A dominating set of the smallest cardinality (size) is called a minimum dominating set, and its cardinal number determines the (lower) domination number $\gamma(G)$ for the given graph G.*

Hence, a dominating set of a graph $G = (V, E)$ is a subset D of V, such that every vertex from the graph G is either in a dominating set or its neighbor is in a dominating

2020 *Mathematics Subject Classification.* 05C69.
Key words and phrases. Dominating set, greedy algorithm, variable neighborhood search.

set [7]. A minimum dominating set is always a minimal dominating set, but the converse does not necessarily hold [15] (see Figure 1 for a counterexample). Furthermore, every graph has at least one dominating set: if $D = V$ = the set of all vertices of the graph G then the set D is by definition a dominating set of a graph G, since $V \setminus D$ is necessarily the empty set. Clearly, a graph without edges has the whole vertex set V as its unique dominating set. For graphs without isolated vertices (an isolated vertex is a vertex with degree zero), there are always two disjoint dominating sets: if D_m = a minimal dominating set of a graph G, then $V \setminus D_m$ is a dominating set.

Unfortunately, finding a dominating set of size k represents an NP-complete decision problem in computational complexity theory, and for now, there is no efficient algorithm that finds a minimal dominating set for a selected graph [8]. Accordingly, checking whether the domination number

$$\gamma(G) = \min\{|D| : D \subseteq V \text{ and } D \text{ is a dominating set of a graph } G\}$$

of an arbitrary graph G is less than a given integer is also an NP-complete problem. If the graph G does not contain isolated vertices, then surely $\gamma(G) \leq \frac{n}{2}$ holds, where $n = |V|$. The domination number $\gamma(G)$ for the graph G from Figure 1, i.e. for the Petersen graph, is 3.

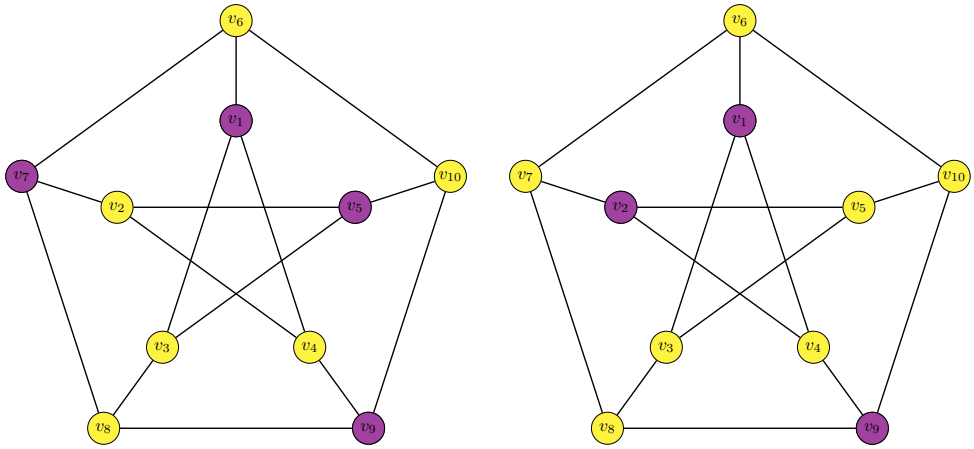


FIGURE 1. A minimal (on the left) and a minimum dominating sets of the Petersen graph.

The primary research problem of this paper is finding a minimal dominating set of a simple undirected graph, i.e. our primary goal is to devise an efficient algorithm for solving such an NP-complete problem. We propose a method that links several heuristic approaches based on variable neighborhood search metaheuristic.

2. ALGORITHM

Graphs are an unparalleled mathematical way to represent a network of interconnected objects that model real-life problems. One such problem was addressed as the

second research problem in the dissertation submitted to the Faculty of Science (Department of Mathematics and Computer Sciences) at the University of Sarajevo [14]. The main results, as well as the algorithm, regarding this problem, i.e. the construction (finding) of the minimal dominating set of amino-acid scales, are submitted to Mathematical Medicine and Biology: A Journal of the Institute of Mathematics & its Applications [13]. To summarize the proposed approaches, we will provide a brief review.

Clearly, to achieve the primary goal, it is necessary to create some initial dominating set. This task (optimization problem) is correctly solved using the greedy algorithm (first approach) and cluster analysis (second approach).

A greedy algorithm is any algorithm that uses a metaheuristics to solve a problem, in a way that it solves the problem by choosing a locally optimal solution at each step (hoping to reach the global optimum that way) [3]. In most combinatorial optimization problems it proceeds, starting from the partial solution $X = \emptyset$, by repeatedly adding to X an element x from the ground set E [1]. Although such an approach can be disastrous for some computational tasks, it is still optimal for many others. Several advantages of greedy algorithms are of particular importance, for example, their design and implementation are usually simple, execution is fast, and we practically never revisit previous choices.

Cluster analysis is a set of methods (algorithms) that allows us to classify (divide) observed data (objects) into groups (classes or clusters) that are meaningful, useful or both. In other words, it allows us to conceptualize similarities and differences between observed data. Cluster analysis is closely related to different fields of research and has been present in the literature for several decades, so it undoubtedly possesses one of the desirable features that any clustering algorithm should have: the ability to work with different types of data. It is also insensitive to the order of steps in the algorithm, it enables the detection of clusters of arbitrary sizes and shapes, as well as the identification of high-quality clusters in the presence of noise.

Greedy approach. In the first approach, we begin with an initial graph $G_0 = G$ of size n and a partial solution $D = \emptyset$. In each step i, we select a vertex v_i with the maximal degree in a graph G_{i-1} and then construct the graph

$$G_i = G_{i-1} - v_i - N_{G_{i-1}}(v_i),$$

where $N_{G_{i-1}}(v_i)$ denotes the set of neighbors of v_i in a graph G_{i-1}. If there are multiple vertices with an equal largest degree, we randomly select one of them. Hence, in each step, we need to determine a maximal degree which takes $o(v(G_i)) \leq o(n)$ operations and eliminate all its neighbors which takes $o(d_{G_i}(v_i)) \leq o(n)$ operations. Hence, this algorithm can be executed in less than or equal to $o(n^2)$ operations. A briefly introduced custom-tailored greedy algorithm (greedy approach) is formally described by Algorithm 1.

The algorithm's correct operation is guaranteed by Proposition 2.1 and its execution in polynomial time is guaranteed by Proposition 2.2, which is stated without proof (see the notes in the paragraph above).

Algorithm 1 Pseudocode for a custom-tailored greedy algorithm [13]

Require: Graph $G = (V(G), E(G))$
Ensure: D = Graph-Dominating Set
1: $G_0 \leftarrow G$
2: $i \leftarrow 0$
3: $D \leftarrow \emptyset$
4: **repeat**
5: $\quad i \leftarrow i + 1$
6: $\quad$ Find a vertex v_i with the maximal degree in a graph G_{i-1}
7: $\quad D \leftarrow v_i$
8: $\quad G_i \leftarrow G_{i-1} - v_i -$ **AllNeighbors**(v_i)
9: **until** $(G_i = \emptyset)$
10: **return** (D)

Proposition 2.1. The set D is a subset of $V(G)$ and a dominating set of a graph $G = (V(G), E(G))$ at the end of the Algorithm 1 (custom-tailored greedy algorithm).

Proof. According to the initial declarations and repeat loop of Algorithm 1 at Step 7, the only elements of the set D are vertices v_i taken from $V(G)$, hence D is necessarily a subset of $V(G)$. Next, assume that D is not a dominating set in G at the end of Algorithm 1. Let $H = V \setminus D$, then there is at least one vertex $u \in H \subseteq V$ whose neighbors are not in the D. According to the stop condition for repeat loop, H is an empty set, but H is not. Therefore, according to the performance of Algorithm 1, the algorithm should not be ended, which is a contradiction. □

Proposition 2.2. The time complexity of Algorithm 1 is at most $o(n^2)$.

To illustrate the functioning of Algorithm 1, we will refer to the graph presented in Figure 1. In a given example with the Petersen graph, all vertices have equal vertex degrees, i.e. $d_G(v_i) = 3$, $\forall i \in \{1, 2, \ldots, 10\}$. Hence, the algorithm starts with an initial graph $G_0 = G$, a dominating set $D = \emptyset$ and in step 1 it will choose one vertex randomly among all vertices. For instance, let vertex v_2 be the chosen one. Therefore, at the end of step 1, we have $G_1 = G_0 \setminus \{v_2, v_4, v_5, v_7\} \neq \emptyset$ and $D = \{v_2\}$. Since G_1 is not an empty set, the algorithm will continue and in step 2 it will search for a vertex with maximal degree in a "surviving" graph G_1. In this particular graph, all vertices also have equal vertex degree, but this time $d_{G_1}(v_j) = 2$, $j \in \{1, 3, 6, 8, 9, 10\}$. So, the algorithm has to choose again randomly one of them. For instance, let vertex v_1 be chosen in step 2. Consequently, at the end of step 2, we will have $G_2 = G_1 \setminus \{v_1, v_3, v_6\} \neq \emptyset$ and $D = \{v_1, v_2\}$. The algorithm will still operate and in step 3 search for a vertex with a maximal degree in the new surviving graph G_2. In this particular case, only one vertex has a maximal degree. It is vertex v_9, so by adding it to the set D, and then eliminating it and all its neighbours from graph G_2, a new surviving graph G_3 will necessarily be an empty set. Therefore, according to the performance of Algorithm 1, the procedure should be finished, and as an output we will have $D = \{v_1, v_2, v_9\}$.

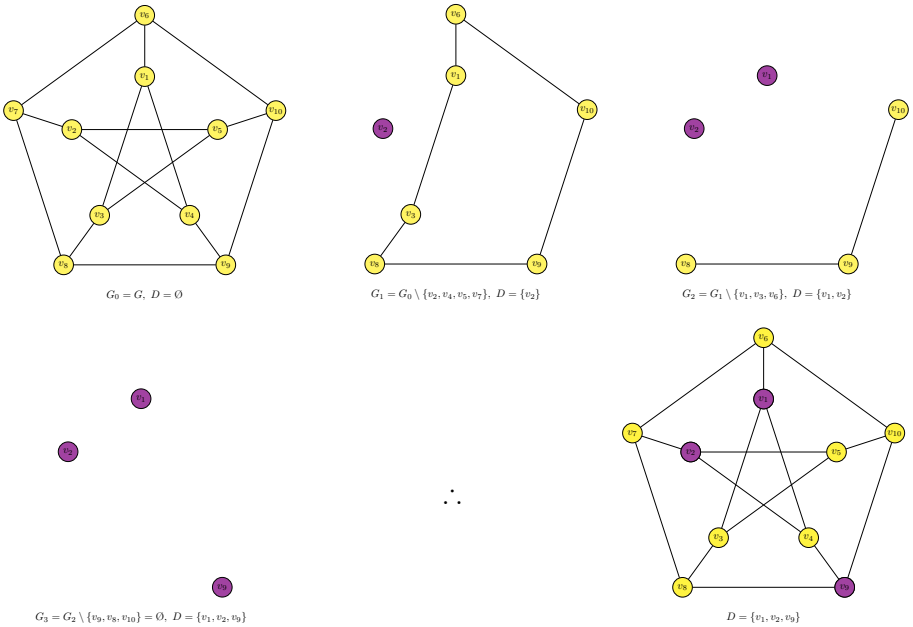


FIGURE 2. A demonstration of Algorithm 1 in action.

The Figure 2 showcases a demonstration of Algorithm 1 in action. We shall notice that in this simple example the set D is the most optimal solution candidate, i.e. it is the minimum dominating set for a given graph. Otherwise, the greedy algorithm does not have to find a solution to the problem. Even if it finds one, that solution may not be optimal, because the greedy algorithm does not use the objective function anywhere, but only the choice function.

Upon creating the initial dominating set, we apply another heuristic method – VNS or variable neighborhood search [11], [6] in order to reduce the size of that primal solution found by a greedy algorithm. Variable neighborhood search is a relatively recently developed metaheuristic. It was introduced in the nineties of the twentieth century, after which it underwent many applications and thus expansion. In this paper, only the basic variable neighborhood search (BVNS) was used, and in the following text, it will mainly be called variable neighborhood search, or VNS metaheuristic for the sake of simplicity. VNS is formally described by Algorithm 2.

The basic idea of the method is very simple (see Figure 3): a systematic change of neighborhoods within a local search (LS). The VNS metaheuristic is based on the following three key principles [5]:

[1]: A local minimum with respect to one neighborhood structure is not necessary so with another;
[2]: A global minimum is a local minimum with respect to all possible neighborhood structures;
[3]: For many problems, local minima with respect to one or several neighborhoods are relatively close to each other.

Algorithm 2 Pseudocode for VNS [2], [13]

Require: Neighborhoods
Ensure: D_{best}
$D_{\text{best}} \leftarrow$ RandomInitialSolution()
while (¬StopCondition()) **do**
  foreach (Neighborhood$_i \in$ Neighborhoods) **do**
    Neighborhood$_{\text{current}} \leftarrow$ CalculateNeighborhood(D_{best}, Neighborhood$_i$)
    $D_{\text{perturbare}} \leftarrow$ RandomPerturbareInNeighborhood(Neighborhood$_{\text{current}}$)
    $D_{\text{perturbare}} \leftarrow$ LocalSearch($D_{\text{perturbare}}$)
    if ($D_{\text{perturbare}}$ **is better than** D_{best}) **then**
      $D_{\text{best}} \leftarrow D_{\text{perturbare}}$
      break
    end if
  end foreach
end while
return (D_{best})

A part of the algorithm used to deal with the determination of solution quality tightly relies on the VNS metaheuristic. We propose two criteria to determine the quality of the solution obtained by a greedy algorithm:

[1]: One of the two solutions (dominating sets) is considered better than the other if it has a smaller number of vertices;
[2]: If two solutions have the same number of vertices, the one whose vertices have a larger sum of squared degrees in the original set is better.

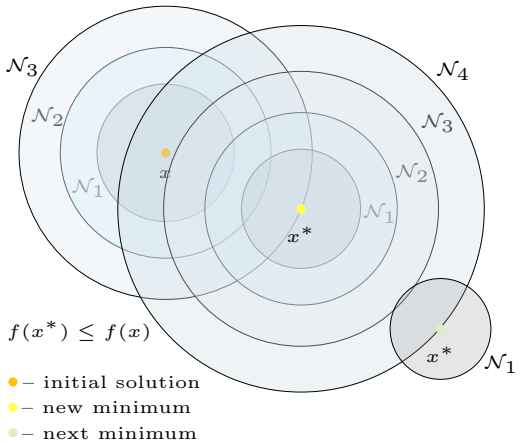


FIGURE 3. VNS introduces multiple neighborhoods for local search.

The argument for both conditions (criteria) is apparent since we are searching for the smallest dominating set. Next, considering that VNS more exhaustively analyzes

a set of greater similarities, this pushes searching for the optimal solution in the right direction.

The distance between sets of vertices is the cardinality of their symmetric difference. Since we can binary code sets of vertices (with 0-1 characters), Hamming distance is chosen as the metric. If some vertex set of size n has x ones and $n-x$ zeros, then there are $\binom{n}{d}$ vertex sets with distance d from that set. Clearly, in the case that we are dealing with large sets of vertices, the number $\binom{n}{d}$ will grow rapidly with increasing the distance d, i.e. it tends to grow in the order of n^d. Hence, this rationale leads us to exhaustively search neighborhoods at distances up to d_{min} and probabilistically neighborhoods at distances up to d_{max}. Determining if a selected set is a dominating set of vertices, is very often an action within the VNS improvement algorithm. The standard algorithm used here has a complexity of $o\left(n+\sum_{v\in D} d_G(v)\right)$ which is less than or equal to $o(n+m)$ where m is the number of edges in the observed graph. Hence, the VNS improvement algorithm has complexity $o(n^d \cdot (n+m)) = o(n^{d+1}+n^d \cdot m)$.

Proposition 2.3. The time complexity of Algorithm 2 is at most $o(n^d \cdot (n+m))$.

Clustering approach. In the second approach, we created the initial dominating set using a $k-$means iterative clustering algorithm with the Euclidean metric. Usually, we can fix the number of iterations required for convergence (to get stable centroid values). In a data-set that does have a clustering structure it is often small, as most changes typically occur in the first few iterations. For a fixed number of iterations i, the overall complexity of this algorithm is $o(k \cdot n \cdot m \cdot i)$ for a data-set with n objects ($m-$dimensional vectors), each with m attributes [10]. Thus, in practice, $k-$means is linear in all relevant factors, although it is in the worst case superpolynomial when performed until convergence. The usage of $k-$means is formally described by Algorithm 3.

Algorithm 3 Pseudocode for basic $k-$means algorithm [12]

Require: Data-set with n objects (each with m features), fixed number k
Ensure: k distinct non-overlapping clusters
1: Select k objects as initial centroids
2: **repeat**
3: Form k clusters by assigning each object to its closest centroid
4: Re-compute the centroid of each cluster
5: **until** Centroids do not change

However, this algorithm has several significant drawbacks [4]. One is manifested in the fact that the number of clusters k (in the vast majority of problems) is fixed and must be chosen by the user (inappropriate k may yield misleading results). Furthermore, this algorithm often converges to a local optimum. Also, the running time of the algorithm is unbounded (unlimited), although it usually works well on real-life problems (real

networks) [12]. Sometimes we only have a set of data, but we don't know how many different groups to expect in that data. In general, there is no method for determining the exact value of k, and it is often an *ad hoc* decision based on prior knowledge, assumptions, and practical experience. Nevertheless, if we want to determine the number of clusters for some data-set, that is, if we want to achieve a balance between the accuracy of joining objects to a cluster and the minimization of the objective function as a function of k, an adequate estimate can be obtained using the *average silhouette* method [9], *gap statistic* method, as well as the so-called *elbow* method [4].

After determining the desired number of clusters, we apply again a greedy algorithm from the first approach, but this time to each cluster. If there were singleton clusters (clusters with only one object), the method was slightly modified. The algorithm will merge such clusters with another cluster to obtain more meaningful initial sets. This is achieved through a single function that can be executed in $o(k \cdot n)$ operations. Ending up with a singleton cluster is frequent if a data-set contains many *outliers*, objects that are far from any other objects and therefore do not fit well into any cluster. Often, if an outlier is chosen as an initial seed, then no other vector is assigned to it during subsequent iterations [10]. For our second approach, we combine all initial sets of vertices (a union) obtained from each cluster by a greedy algorithm. The initial set, i.e. dominating set, found by a clustering algorithm is then fixed using the VNS strategy too.

3. Conclusions

In this paper, we have concisely reviewed the summary of the findings of a minimal graph-dominating set combining various heuristic approaches based on variable neighborhood search. With the paper's algorithms, a minimal dominating set of a simple undirected graph can be efficiently constructed by applying the VNS strategy on an initial dominating set obtained either by a greedy approach or by combining clustering and greedy approaches. Both implemented algorithms for constructing the initial set of vertices are relatively fast, i.e. require a very small amount of time, while the algorithm with regard to the VNS strategy operates at a quite leisurely pace. This is reasonable since NP-complete does not necessarily imply unsolvable. It just means any proposed solution will be slow.

Moreover, some results and the efficiency of the given approaches when applied on the particular real-life problem have been reported in an article [13] and a dissertation [14]. There a simple mathematical framework regarding modeling of a reduction to a dominating set of a graph with the set of 507 amino-acid scales (indices), constructed by Kawashima[1] and collaborators, is presented . Concerning guidelines for future research, given in the dissertation [14], it might be interesting to see whether similar generalized reduction in other contexts, especially in those major real-world instances (networks) where computationally intensive modeling is a dominant issue, will benefit from techniques used in these papers. Furthermore, it is also interesting to study is it possible to achieve even more significant improvements through the use of different settings, mainly regarding the desired number of clusters k and wanted numbers for dis-

[1] Amino acid index database: `https://www.genome.jp/aaindex/`

tances $d_{\min}$, $d_{\max}$. Finally, one can consider changing a default metric within the VNS improvement algorithm.

ACKNOWLEDGMENTS

The author is thankful to academician Mirjana Vuković and professor Amela Muratović-Ribić, supervisor for the 3rd cycle of study "Mathematical Sciences in Southeast Europe" at the Faculty of Science University of Sarajevo, for the invitation to participate in this special Conference on the occasion of March 14 – International Day of Mathematics.

REFERENCES

[1] G. Bendall and F. Margot, *Greedy-type resistance of combinatorial problems*, Discrete Optim., **3** (2006), 288–298, https://doi.org/10.1016/j.disopt.2006.03.001.

[2] J. Brownlee, *Clever Algorithms: Nature-Inspired Programming Recipes*, Open source book, 2012., http://www.cleveralgorithms.com/.

[3] T. H. Cormen, C. E. Leiserson, R. L. Rivest and C. Stein, *Introduction to Algorithms*, MIT Press, Massachusetts, 2001.

[4] B. S. Everitt, S. Landau, M. Leese and D. Stahl, *Cluster Analysis*, 5th edn. Wiley, New York, 2011, https://doi.org/10.1002/9780470977811.

[5] P. Hansen and N. Mladenović, *Variable neighborhood search*. In: F. Glover and G. A. Kochenberger (eds) Handbook of Metaheuristics. Springer, New York, 2003, pp. 145–184, https://doi.org/10.1007/b101874.

[6] P. Hansen, N. Mladenović and J. A. Moreno Pérez, *Variable neighbourhood search: methods and applications*, Ann. Oper. Res., **175** (2010), 367–407, https://doi.org/10.1007/s10479-009-0657-6.

[7] T. W. Haynes, S. T. Hedetniemi and P. J. Slater, *Fundamentals of Domination in Graphs*, Marcel Dekker, New York, 1998.

[8] R. M. Karp, *Reducibility among Combinatorial Problems*. In: Miller, R. E., Thatcher, J. W., Bohlinger, J. D. (eds) Complexity of Computer Computations. The IBM Research Symposia Series. Springer, Boston, Massachusetts, 1972, pp. 85–103, https://doi.org/10.1007/978-1-4684-2001-2_9.

[9] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, 1st edn. Wiley, New York, 1990, https://doi.org/10.1002/9780470316801.

[10] C. D. Manning, P. Raghavan and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, Cambridge, 2009, http://www.informationretrieval.org/

[11] N. Mladenović and P. Hansen, *Variable neighborhood search*, Comput. Oper. Res., **24** (1997), 1097–1100, https://doi.org/10.1016/S0305-0548(97)00031-2.

[12] P-N. Tan, M. Steinbach and V. Kumar, *Introduction to Data Mining*, 1st edn. Pearson Addison-Wesley, Boston, 2005.

[13] A. Vrdoljak and D. Vukičević, *Selector of Amino-Acid Scales Set*, Math. Med. Biol., **41** (2024), 157–168, https://doi.org/10.1093/imammb/dqae007.

[14] A. Vrdoljak, *Automati grafova*, disertacija, Univerzitet u Sarajevu, Prirodno–matematički fakultet, Odsjek za matematičke i kompjuterske nauke, 2024.

[15] E. W. Weisstein, "Minimum Dominating Set". From MathWorld–A Wolfram Web Resource. https://mathworld.wolfram.com/MinimumDominatingSet.html

(Received: May 06, 2024)
(Revised: August 08, 2024)

Anton Vrdoljak
University of Mostar
Faculty of Civil Engineering, Architecture and Geodesy
Matice hrvatske b.b., 88000 Mostar, BiH
e-mail: *anton.vrdoljak@fgag.sum.ba*

PROCEEDINGS OF THE CONFERENCE ON MARCH 14 – INTERNATIONAL DAY OF MATHEMATICS
Special Editions ANUBiH, Book CCXVI, OPMN Book 30
Sarajevo 2024, **p. 90–99** DOI: 10.5644/PI2024.216.11

USING DIRECTED GRAPHS TO DESCRIBE AUTOMATION PROCESSES AND ANALYZE CONTROL SYSTEMS

IVANA ZUBAC, SNJEŽANA REZIĆ, AND JADRANKO BATISTA

Dedicated to the 75th birthday of our dear Professor Mirjana Vuković

ABSTRACT. In many ways, engineering has always been dependent on mathematical background. This was usually mathematical analysis. However, many problems in natural, technical and social science can be successfully formulated in terms of graph theory. Today graph theory is well developed, strongly stimulated by technical and chemical applications. Graphs have been increasingly used in many fields, such as automation, programming or algorithms. In the automation process directed graphs are used for description and analysis of the control system, where nodes represent states and arrows direct edges or transitions between states automation that manages processes. In this paper we show how to use graphs to simplify the process of automation.

1. INTRODUCTION

Many problems in the natural, technical and social sciences can be successfully formulated in terms of graph theory. Today, graph theory is well developed, strongly motivated by technical and chemical applications, and it has established itself as an important mathematical tool in a wide variety of subjects, from operational research to genetics and linguistics, and from electrical and mechanical engineering and geography to sociology and architecture. For general background on graph theory and terminology, we refer the reader to the classic book by West D.B. [7], [4]. For the theory of directed graphs or digraphs, which is not defined here, we also recommend [1], [3]. Graph theory studies the ways in which sets of points, called vertices, can be connected by lines or arcs, called edges. The term graph in this context differs from graphs that show mathematical relationships and functions in coordinate systems. A directed graph has oriented edges, which we show with arrows. In practice, there is a great need to display various systems using such graphs (e.g. traffic regulation in a city, liquid flow in a system, transport of goods, process automation, etc.)

1.1. Basic definition of graphs and digraphs

A graph G consists of a finite nonempty set V of p vertices together with a defined set X of q unordered pairs of distinct vertices of V. Each pair $x = \{u, v\}$ of points in X is an edge of G, and we say that x joins u and v. A graph with p vertices and q edges is called a (p, q) graph.

2020 *Mathematics Subject Classification.* 05C20, 05C90, 94C15.
Key words and phrases. graph theory; digraph; automation; matrix.

A directed graph, or **digraph** D consists of a non-empty finite set $V(D)$ of elements called vertices, and a finite family $A(D)$ of ordered pairs of elements of $V(D)$ called arcs arc or directed edge. We call $V(D)$ the vertex set and $A(D)$ the arc family of D. An arc (v, w) is usually abbreviated to vw, the ordering of the vertices in an arc being indicated by an arrow. If D is a digraph, the graph obtained from D by 'removing the arrows' (that is, by replacing each arc of the form vw by a corresponding edge vw) is the underlying graph of D.

Two or more arcs with the same beginning and end are called multiple arcs. So, D is the orientation of G and we write $D = \overrightarrow{G}$. A directed graph consists of a finite set of V vertices (vertices) and a collection of ordered pairs. A strict graph is a directed graph whose associated graph is simple and complete. The associated graph $D(G)$ of a graph G is a directed graph obtained from G by replacing each edge with two oppositely oriented arcs with the same endpoints. The associated graph $G(D)$ of a directed graph D is the graph obtained from D by deleting all arrows. A tournament is a directed graph whose graph is simple and complete.

Many terms used in graphs are also used in directed graphs. For example directed cycle or dicycle, directed path or dipath. A directed graph is acyclic if it does not contain a dicycle.

On the other hand, oriented edges between two vertices are allowed in simple directed graphs, and such a pair of edges is called a digon. We say that a digraph D is weakly connected (or connected for short) if the associated graph is connected, and strongly connected if for every two vertices $u, v \in V(D)$ there is a (u,v)-diput.

A component of a digraph is a connectivity component of an associated graph.

Unlike graphs, digraphs have two types of vertex degrees. The *in-degree* of a vertex in a digraph is the number of arrows coming into it, and similarly its *out-degree* is the number of arrows out of it. More precisely, for $v \in V(D)$ we define:

- **indeg(v),** $d_D^-(v)$ and
- **outdeg(v)** $d_D^+(v)$.

The following proposition is analogous to the proposition about the degree of an undirected graph:

Proposition 1.1. *Let D be a directed graph with a set of vertices $V(D)$ and a set of arcs $A(D)$. Then*

$$\sum_{v \in V(D)} d^-(v) = |A(D)| = \sum_{v \in V(D)} d^+(v).$$

In this paper, all directed graphs are finite and can have loops and multiple edges (edges with the same starting and ending vertices). A directed graph D is simple if D has no loops and there is at most one edge from v_i to v_j for any $v_i, v_j \in V(D)$.

1.2. **Matrix representation of a graph**

It is common to represent a graph using a graphical and matrix display. The graph is completely determined by adjacency or incidence matrices. The adjacency matrix $A = \left[a_{ij}\right]$ of a labeled graph G with p points is a $p \times p$ matrix in which $a_{ij} = 1$ if v_i is

adjacent to v_j and $a_{ij} = 0$ otherwise. The second matrix, associated with the graph G in which vertices and edges are marked, is the incidence matrix $B = \left[b_{ij}\right]$. This $p \times q$ matrix has $b_{ij} = 1$ if v_i and x_j are incident and $b_{ij} = 0$ otherwise.

The adjacency matrix of the directed graph D on the set of vertices $V(D) = \{v_1, v_2, \ldots, v_n\}$ is the square matrix $A(D) = \left[a_{ij}\right]$ of order n, where a_{ij} is the number of arcs in D starting at v_i and ending at v_j .

If there are no multiple arcs, then the elements of the adjacency matrix are only zeros and ones, otherwise they are non-negative integers. Each adjacency matrix uniquely defines a directed graph D.

It is often possible to use these matrices to identify certain properties of a graph. For a digraph, the degree of a vertex is the sum of the column and row entries corresponding to its adjacency matrix. Row values corresponding to the vertex v represent edges with v as the starting vertex, and column values corresponding to v represent edges with v as the final vertex. The sum of all entries in the adjacency matrix is, of course, the total number of edges in the digraph.

A binary relation on a set can be represented by a digraph. Let R be a binary relation on the set A, that is, R is a subset of $A \times A$. Then the digraph representing R can be constructed as follows:

Let the elements of A be vertices of a digraph G, and let $\langle x, y\rangle$ be an arc of G from vertex x to vertex y if and only if $\langle x, y\rangle$ is in R.

2. Automation process

Automation of the process is replacing human labor with machines, not only in terms of strength, but also intellectual work. Technically, the automatic machine (technical system) consists of three groups of elements:

- senses (sensors, cameras, microphones, etc.)
- controllers (processors who process information)
- executive elements.

For theoretical background on of automation of the process not defined here we also recommend [2], [5], [6]. *The management process* is a system, in which one or more input variables, over the legality which characterizes this system, affect other variables as output values. *The process* is quantitative and/or qualitative change dependent on the weather, and it takes place in nature, society and technology.

Automating processes involve integrating disparate systems for seamless data flow across an organization. Additionally, it's a critical part of continuous optimization, and core to many organization's digital transformation initiatives.

Information is data about a particular phenomenon, concept or event. The holder of information is the signal. The signal is a changeable and measurable variable at the entry or exit of the system, and can take a variety of physical forms. Classification of signals depends primarily on the respective parameters of amplitude and time. We distinguish: continuous, discrete and binary signals. Digital automat is a universal sequential circuit, whose behaviour depends only on the current and previous input data-events. Working machines can be explained by the theory of systems management.

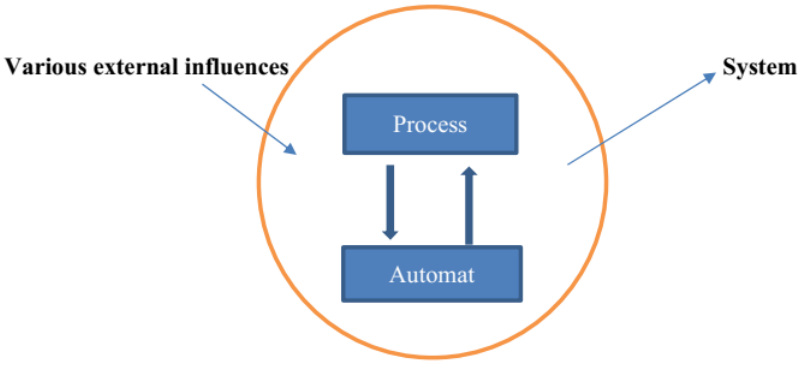


Figure 1. Automation process

An automat measures the state of the process and resolves all its essential conditions. On the basis of the current and previous events in the process we can determine the optimal action or series of actions. These actions should result in an automat process in the optimal mode. To make this possible, the automat must have a sufficient set of actions, in order to compensate for any predictable impact of the environment. We say that the process must be controllable, so that the management could work. If the value of the output variable depends not only on the current values of the input variable, but also on the past values (total input-output) we says that the digital system is a sequential system or is automatic.

The properties of automata:

- *Define finality (there are a finite number of states, the final memory);*
- *Define discretion (working in discrete time);*
- *Define the digital mode (available digital inputs and outputs);*
- *Define determination (unambiguously perform its function);*
- *Define the specificity (completely - all possible sequences of input events - expects an arbitrary set of inputs; incomplete - possible are only a series of input events);*
- *Define synchronicity (discrete time defined by phased signal).*

3. AUTOMATION OF THE SELF-SERVICE DEVICE FOR BEVERAGES

Before starting the design of the automation process, it is important to determine the inputs and outputs of the automaton and state automata.

We will assume that the self-service machine dispenses four drinks, which means that we need to have four inputs (select1, select2, select3, select4) for the drinks. We also assume that the machine only accepts 0.50 KM coins and one convertible mark, which we mark as km_50 and km_100. We must also implement an input for canceling the order if the customer wants to return the money. Return, drink and change will represent our exits. A refund returns the money back to the person, a drink out throws out the selected drink, while a change returns the excess money.

In the juice vending machine, one bottle costs 1.50 KM. The operation of the machine is controlled by a sequential circuit with two inputs X_1 and X_2.

A logic unit appears at input X_1 (for the duration of one clock pulse) when a 1 KM coin is inserted, and a logic unit appears at input X_2 (for the same duration) when a 2 KM coin is inserted. At all other times, inputs X_1 and X_2 are at logic zero. The sequential circuit has two outputs Y_1 and Y_2. The logic unit at output Y_1 starts the motor that ejects the bottle of juice, while the logic unit at output Y_2 orders the device to eject change of 0.50 KM.

It's apparently a sequential circuit, since the device has to keep track of how much money was previously inserted in order to know how to react to the insertion of the next coin. It is important to note that it is not necessary to keep records of the total amount of money in the device, but only of the amount of money that was inserted after the last ejected bottle of juice. Namely, after each ejected bottle, the device, from the user's point of view, behaves as if it had just started working. This fact is of crucial importance, because based on it we can conclude that, from the aspect of behavior, the assembly can only be in one of three different states:

S_0 - No coins were inserted after the last bottle was thrown

S_1 - After throwing out the last bottle, a total of 0.50 KM was put

S_2 - After throwing out the last bottle, a total of 1 KM was put

The bottle is ejected if a 1 KM coin is inserted in state S_1, or if either a 0.50 KM coin or a 1 KM coin is inserted in state S_2 (in the latter case, change is also thrown out). In doing so, the device returns to the initial state S_0 .

Based on previous thinking and analysis, it is quite easy to draw a graph that describes the circuit's operation.

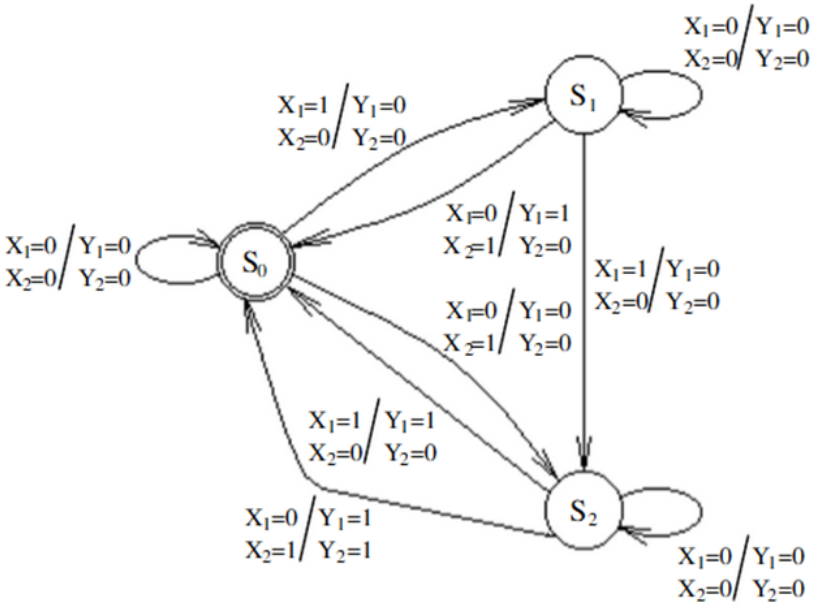


Figure 2. Automaton graph

Alternatively, instead of a graph, the operation of the circuit can be described with the following transition table, which is easy to compile by directly reading the graph.

TABLE 1. Transition table

X_1	X_2	The old state	The new state	Y_1	Y_2
0	0	S_0	S_0	0	0
0	0	S_1	S_1	0	0
0	0	S_2	S_2	0	0
0	1	S_0	S_2	0	0
0	1	S_1	S_0	1	0
0	1	S_2	S_0	1	1
1	0	S_0	S_1	0	0
1	0	S_1	S_2	0	0
1	0	S_2	S_0	1	0

The displayed states must be coded with binary numbers. Let's adopt the following encoding:

TABLE 2. Table of code

X_1	X_2	$Q_1(n)$	$Q_2(n)$	$Q_1(n+1)$	$Q_2(n+1)$	Y_1	Y_2
0	0	0	0	0	0	0	0
0	0	0	1	0	1	0	0
0	0	1	0	x	x	x	x
0	0	1	1	1	1	0	0
0	1	0	0	1	1	0	0
0	1	0	1	0	0	1	0
0	1	1	0	x	x	x	x
0	1	1	1	0	0	1	1
1	0	0	0	0	1	0	0
1	0	0	1	1	1	0	0
1	0	1	0	x	x	x	x
1	0	1	1	0	0	1	0
1	1	0	0	x	x	x	x
1	1	0	1	x	x	x	x
1	1	1	0	x	x	x	x
1	1	1	1	x	x	x	x

To implement the transition/output table, JK and D bistable will be used for memory and state change.

We can see the implementation of the automaton using combination circuits in Figure 3.

3.1. **Logical scheme**

In order to simplify the entire cycle of buying a beverage, let's look at the flow diagram of our device in Figure 4.

The state diagram consists of four states (User's selection, Waiting for the money to be entered, ejecting the product and servicing if the selection is not available). Primarily,

TABLE 3. Table of transition/exits

X_1	X_2	$Q_1(n)$	$Q_2(n)$	$Q_1(n+1)$	$Q_2(n+1)$	Y_1	Y_2	J_1	K_1	D_2
0	0	0	0	0	0	0	0	0	x	0
0	0	0	1	0	1	0	0	0	x	1
0	0	1	0	x	x	x	x	x	x	x
0	0	1	1	1	1	0	0	x	0	1
0	1	0	0	1	1	0	0	1	x	1
0	1	0	1	0	0	1	0	0	x	0
0	1	1	0	x	x	x	x	x	x	x
0	1	1	1	0	0	1	1	x	1	0
1	0	0	0	0	1	0	0	0	x	1
1	0	0	1	1	1	0	0	1	x	1
1	0	1	0	x	x	x	x	x	x	x
1	0	1	1	0	0	1	0	x	1	0
1	1	0	0	x	x	x	x	x	x	x
1	1	0	1	x	x	x	x	x	x	x
1	1	1	0	x	x	x	x	x	x	x
1	1	1	1	x	x	x	x	x	x	x

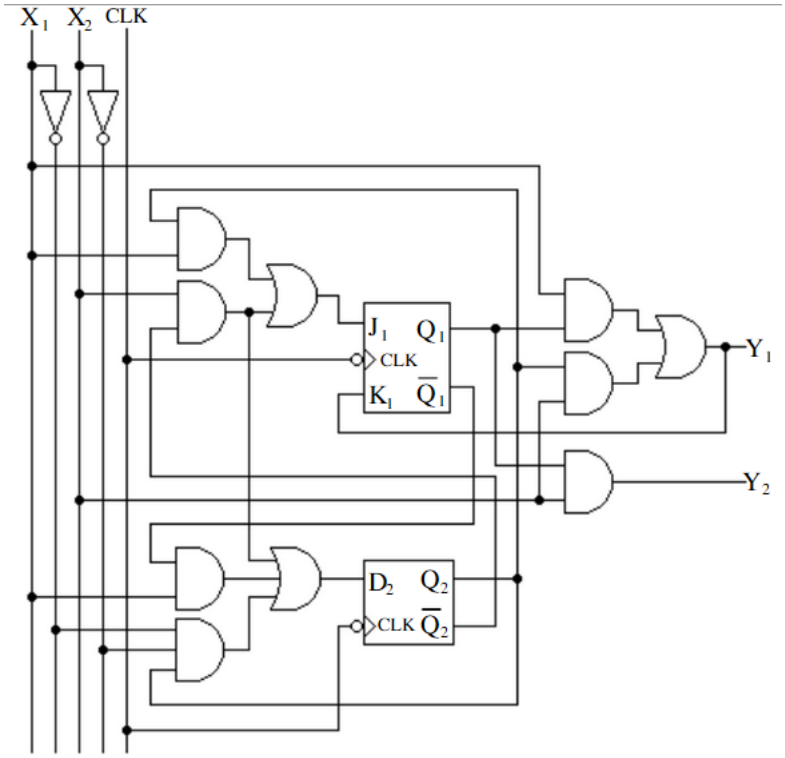


Figure 3. Logical scheme

when the reset button is pressed, the device will be ready to select a beverage. After this state the user chooses a drink, this state can be one of the four listed selections. The device accepts two types of coins: 0.50 KM and 1 KM. Let's assume that the useful one

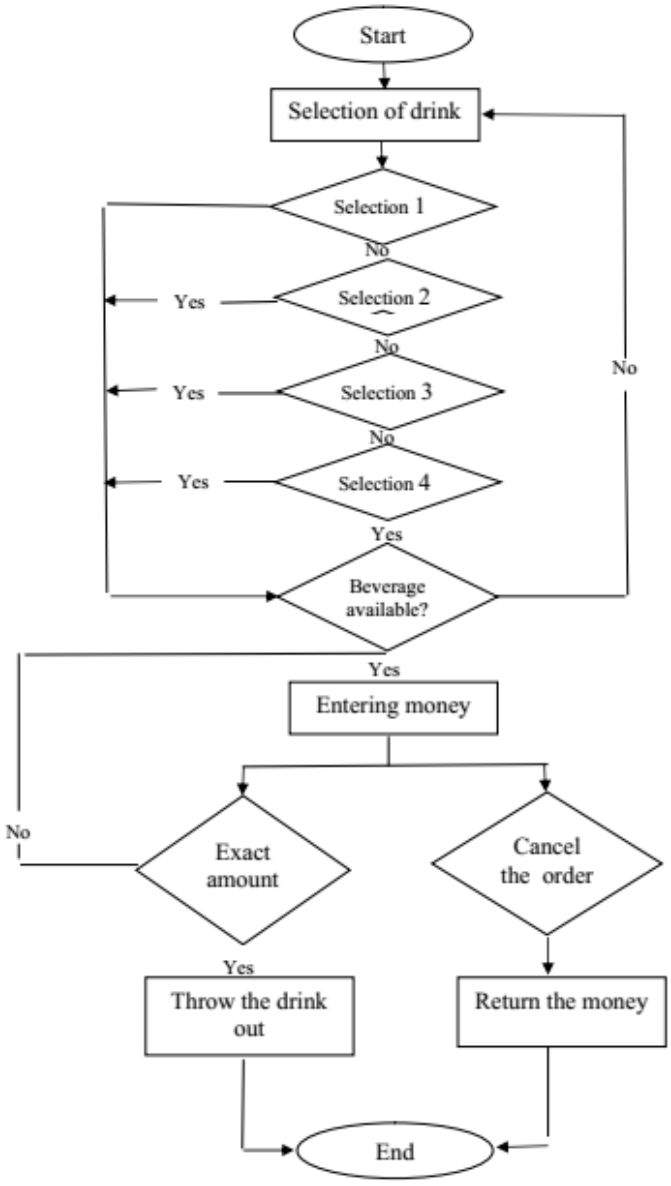


Figure 4. Flow diagram

chooses the drink selection1. The device primarily checks whether the drink is available or not. After that, the control unit switches to another state, where it waits for the money to be entered. If a 0.50 KM coin is entered, it goes to state_1 where it waits until the desired amount is entered. If one convertible mark is entered, it switches to state_2 and waits until one and a half convertible marks are entered. When the desired amount of money is filled, the device ejects the drink.

The obtained graph of the operation of the automaton clearly shows the connection between all its internal states and transitions, i.e. the flow of information and signals. All this enables us to describe and analyze the entire management system, but also to facilitate the implementation and simulation of theoretical and real systems, or their synthesis.

Among the classical methods of mathematical description of the control system and graphoanalytical methods, there is a correlation of both models, and the developed graphoanalytical methods easily lead to a complex mathematical model, which significantly shortens the time of setting up the model, synthesizing the control and developing the control algorithm. Also, we usually write graphs in the computer memory via the adjacency matrix, which enables easier management and monitoring of the automated process.

4. Conclusion

Unlike conventional methods of mathematical modeling and mathematical description of automatic control systems, graphoanalytical methods have taken precedence not only in practice but also in research, because they lead to a simpler and more comprehensible description of the system, and thus to its analysis and ultimately to the synthesis of optimal solutions.

The fact that an edge connects node A to node B without requiring feedback is the reason for using a directed graph in the development of automation for control processes, because feedback is not always needed in the description, analysis and synthesis of automation for process control.

References

[1] J. Bang-Jensen, G. Gutin, *Digraphs Theory, Algorithms and Applications*, Springer-Verlag, London, 2007.
[2] M. Balach, *Complete Digital Design*, McGraw-Hill, New York, 2003.
[3] R. Garnier, J. Taylor, *Discrete Mathematics for New Technology*, Institute of Physics Publishing, Bristol and Philadelphia, 2002.
[4] F. Harary, *Graph theory*, Adison Wesley Longman Publishing Co., Massachusetts, 1972.
[5] J. Hopcroft, J. Ullman, *Introduction to Automata Theory, Languages, and Computation*, Addison-Wesley, Boston, 1979.
[6] Edited by A. D. Rodic, *Automation Control – Theory and Practice*, InTech, 2009.
[7] D. B. West, *Introduction to Graph Theory*, Prentice Hall Inc., Upper Saddle River, NJ, 2007.

(Received: May 17, 2024)
(Revised: September 01, 2024)

Ivana Zubac
University of Mostar
Faculty of Mechanical Engineering, Computing and Electrical Engineering,
Matice hrvatske b.b.
88000 Mostar, BiH
e-mail: *ivana.zubac@fsre.sum.ba*
and
Snježana Rezić
University of Mostar
Faculty of Mechanical Engineering, Computing and Electrical Engineering,
Matice hrvatske b.b.
88000 Mostar, BiH
e-mail: *snjezana.rezic@fsre.sum.ba*
and
Jadranko Batista
University of Mostar
Faculty of Science and Education
Matice hrvatske b.b.
88000 Mostar, BiH
e-mail: *jadranko.batista@fpmoz.sum.ba*

PROCEEDINGS OF THE CONFERENCE ON MARCH 14 – INTERNATIONAL DAY OF MATHEMATICS
Special Editions ANUBiH, Book CCXVI, OPMN Book 30
Sarajevo 2024, **p. 100–113** DOI: 10.5644/PI2024.216.12

THE MATHEMATICS OF ARTIFICIAL INTELLIGENCE

ERVIN MACIĆ, TARIK HUBANA AND MIGDAT HODŽIĆ

Dedicated to the 75th birthday of our dear Professor Mirjana Vuković

ABSTRACT. Although artificial intelligence (AI) is often perceived as the field of computer science that experienced the greatest development, without the wide scope of mathematical fields and methods, AI would not be possible. Mathematical areas in AI include standard methods of analysis (continuous or discrete), set theory and various logic's (propositional, first-order logic, ineffable logic), statistics, probability theory and random processes, vector theory and linear algebra, matrix, optimization, estimation theory and filtering, harmonic analysis, information theory, entropy analysis, graph theory, search methods and other related fields. All the mentioned areas and methods are intertwined in many ways, and specific applications and research determine which specific methods and areas are used. In this sense, this paper provides an overview of a wide range of mathematical fields, methods and concrete applications in a comprehensive manner. Therefore, this paper contributes to the existing knowledge base by summarizing the main mathematical areas, methods and applications in AI, providing mathematicians and artificial intelligence engineers with a basis for further research.

1. INTRODUCTION

Artificial intelligence (AI) is a branch of computing that deals with the development of computer systems capable of performing tasks and solving problems that typically require human intelligence. These include abilities such as learning, reasoning, pattern recognition, decision making and communicating with people. Mathematics forms an integral foundation of AI by providing the key tools and concepts needed to understand, model and solve complex problems. Mathematical methods play a key role in the development of algorithms and models that enable computers to learn, reason, make decisions and work with data. In addition, statistical methods in mathematics are essential for analyzing data, studying patterns and drawing conclusions based on them. Logical formalization enables precise representation of knowledge, drawing conclusions and making decisions in AI systems. Geometry and topology find applications in areas such as computer vision, where they are used to analyze shapes and spatial relationships in images and 3D models. Information theory, another branch of mathematics, plays a key role in signal processing, data compression, and the analysis of complex systems.

2020 *Mathematics Subject Classification.* 68T01.

Key words and phrases. Artificial intelligence, mathematical methods, machine learning, deep learning, large language models.

Because of this key role of mathematics in the large scope of the field of artificial intelligence, the field of research is very active. The greatest achievements and improvements in the field of artificial intelligence were actually inspired by research works. For example, the current interest and advancement of large language models (the model behind the popular application ChatGPT [1]) is a direct consequence of the popular research paper "Attention is all that is needed" [2] where a new encoder-decoder configuration is proposed. A similar effect was achieved by the research that proposed ADAM, a gradient optimization method for stochastic objective functions based on adaptive estimates of lower order moments [3]. The introduction of the perceptron model, which represents a simple mathematical model of a neural network [4] and the backpropagation algorithm [5] provided the basis for the further development of neural network models leading to more complex mathematical models and network structures such as recurrent neural networks that can analyze the flow of information between layers as a two-way neural network [6], deep learning networks [7], Long Short-Term Memory (LSTM) networks [8] and models such as Support vector machine (SVM) models [9] and Random forest (RF) models [10]. Therefore, it is clear that mathematics is essential to the understanding, development and application of algorithms and models in various areas of artificial intelligence, providing the foundation for progress in this rapidly growing field.

2. INTELLIGENT AGENTS

2.1. **What is an intelligent agent?**

An agent can be considered anything that can perceive its environment through sensors and act in that environment. On the other hand, intelligent agents usually represent software that has the ability to perform a task flexibly, independently and without user intervention, and informs the end user about the completion of the task or the very occurrence of the expected event. The agent itself interacts with the environment in order to perform the set task as precisely as possible. Intelligence in this context can be seen as the agent's ability to accept given goals, and the way in which they execute them. Intelligence reflects the level of quality of thinking and learned behavior [11]. A simple agent can be mathematically defined with an agent function that maps each possible sequence of perception to a possible action that the agent can perform or to a coefficient, feedback, function or constant that affects the eventual action [11]:

$$f : P^* \to A. \tag{2.1}$$

The function of an agent is an abstract concept that includes various decision-making principles such as the calculation of individual possibilities, deductions beyond logical rules, and the like, so there is a whole range of diverse agents. However, what they all have in common is that they can improve their performance through learning.

2.2. **Problem solving by searching**

When agents are faced with uncertainty regarding a certain action, it is necessary to carry out planning in advance, and determine the sequence of actions that form a road map to the desired state. This group of agents are called problem-solving agents, and the

problem-solving process is called searching. Currently, there are a number of search algorithms, such as best-first search, breadth-first search, uniform-cost search, depth-first search, depth-first iterative search, and bidirectional search, all of which attempt to find solutions in environments that are episodic, contain one agent, observable, deterministic, static, discrete and fully known. These algorithms make a trade-off between search time, memory usage and solution quality. The search is preceded by a well-defined problem formulation, consisting of an initial state, a set of actions, a transition model, target states, and an action cost function. Search algorithms traverse state space graphs, treating states and actions atomically. Evaluation criteria for search algorithms include completeness, cost optimality, time and space complexity. Uninformed methods, such as breadth-first search and depth-first search, operate solely on problem definitions. Informed methods, such as greedy best-first search, A* search, two-way A* search, IDA*, RBFS (recursive first-best search), SMA* (simplified memory bound A* search), and weighted A* search use heuristics functions for estimating solution costs. The quality of a heuristic search depends on the accuracy of the heuristic, often improved by problem relaxation, pre-calculated solution costs, landmark identification, or learning from problem experience. One of the most frequently used algorithms in the field of artificial intelligence is the A* search, which mathematically determines its main loop in each iteration where paths of the weighted graph extend starting from the initial node. This is done on the basis of the journey cost and the assessment of the costs required to extend the journey to the destination. Mathematically, A* chooses the path that minimizes the following expression [11]:

$$f(n) = g(n) + h(n) \tag{2.2}$$

where n is the next node on the path, $g(n)$ is the cost of the path from the starting node to n, and $h(n)$ is a heuristic function that estimates the cost of the cheapest path from n to the goal.

2.3. **Searching in complex spaces**

In contrast to the search algorithms presented in the previous chapter, search in complex situations explores the relaxation of constraints in search problems, extending beyond fully observable, deterministic, and static environments. In this case, it is necessary to deal with the task to find optimal states without considering the path to them, covering discrete and continuous states. Local search methods such as the peak-climbing algorithm, simulated annealing, and evolutionary algorithms are aimed at direct state optimization, which is useful for both discrete and continuous problems, and offer efficient solutions to optimization tasks with appropriate parameterization. Linear programming and convex optimization excel in certain forms of state space, while evolutionary algorithms maintain a population of states, using mutation and crossover to generate states. In nondeterministic environments, the AND–OR search facilitates contingency planning. When the environment is partially observable, the belief state method represents the set of possible states the agent could be in, where standard search algorithms can be used for sensorless problems, while the belief state AND-OR search can generally solve partially observable problems [11]. Often the objective functions are expressed in mathematical form in such a way that it is possible to use calculus to

solve the problem analytically rather than empirically. There are a number of methods that use gradient environments to find the maximum. However, for many problems, the most efficient algorithm is the Newton-Raphson method, which represents a general technique for finding the roots of a function, i.e. solving an expression in the form $g(x) = 0$. The method works by computing a new estimate for the root using Newton's formula [11]:

$$x \leftarrow x - \frac{g(x)}{g'(x)} \tag{2.3}$$

In order to find the maximum, it is necessary to find such x for which the gradient is a zero vector (ie $\nabla \mathrm{f}(x) = 0$). Therefore, $g(x)$ in the previous formula becomes $\nabla \mathrm{f}(x)$, and the updated equation can be written in matrix form as:

$$x \leftarrow x - H_f^{-1}(x)\nabla \mathrm{f}(x) \tag{2.4}$$

where $H_f(x)$ is the Hessian matrix of second derivatives, whose elements H_{i_j} are given as $\partial^2 f / \partial x_i \partial x_j$.

2.4. **Adversial search and games**

When the agent environment becomes competitive, adversarial search problems arise where multiple agents pursue conflicting goals. Therefore, as a solution for an optimal result, games that represent practical strategies and behavior of agents in adversarial environments are imposed. Key concepts include defining a game in terms of initial state, legal actions, outcomes of actions, terminal conditions, and utility function. Well-known methods include the minimax algorithm, the Alpha-beta algorithm and the Monte Carlo algorithm. Also, many programs use pre-calculated tables of moves that help decisions in certain games. For games based on chance, the Expectiminimax algorithm is used. Using these methods, artificial intelligence has triumphed in games like chess and GO, while humans still maintain superiority in some imperfect information games. In video games, AI competes effectively, using rapid decision-making capabilities [11]. In the context of artificial intelligence, a commonly used method is Monte Carlo Tree Search (MCTS) which is an algorithm that searches the state space and makes statistical evidence of the decisions available in the corresponding states. Since it can be modeled as a Markov decision process, the process is modeled as an ordered sequence (S, A_s, P_a, R_a), where [12]:

S– set of states that are possible in the environment (state space).
A_s -set of actions available in state s.
$P_a(s, s\prime)$ – transition function modeled as the probability that action a taken in state s will lead to state s'.
$P_a(s, s\prime)$ – transition function modeled as the probability that action a taken in state s will lead to state s'.
$R_a(s)$– current reward for reaching state s through action a.

MCTS represents an algorithm that is time-limited, so it can be stopped at any time while finding the best action (decision) at that moment using the following formula [12]:

$$a^* = argmax_{a \in A(s)} Q(s, a) \tag{2.5}$$

where A(s) is the set of actions available in state s in which a decision needs to be made, and Q(s,a) is the empirical average result of executing action a in state s.

2.5. **Constraint Satisfaction Problems**

Constraint satisfaction problems (CSP) represent a state by variable/value pairs and represent the conditions for a solution by setting constraints on the variables. The backward search algorithm, often used for CSPs, uses minimum residual, degree, and least limiting value heuristics to guide variable selection and value assignment. A conflict-oriented backtracking algorithm goes back to the source of the conflict, while a constraint learning algorithm records the conflicts it encounters to prevent them from recurring. Local search, especially using the minimum conflict heuristic, has been shown to be successful for CSPs. The complexity of solving CSPs is strongly related to the structure of the constraint graph, so reduction techniques such as cut conditioning and tree decomposition are often applied, which enable a better ratio of memory consumption and algorithm execution time.

3. KNOWLEDGE AND REASONING

3.1. **Logical agents**

Understanding logical agents is essential for a deeper understanding in the field of artificial intelligence. Logical agents are computer programs or systems that use logic to make decisions in their environment. They use formal logical languages and reasoning rules to process information and draw conclusions. After that, based on these conclusions, they make certain actions or decisions. Logical agents can be applied in various areas such as recommendation systems, process management, diagnostics, planning and many other areas where it is necessary to make decisions based on available information. One of the main advantages of logical agents is their ability to make clear and transparent reasoning in contrast to non-transparent neural networks. These agents use formal logical rules, which allow humans to understand and verify the way the agent makes decisions. Despite their advantages, logical agents face challenges in situations where the environment is dynamic or when it is necessary to handle large amounts of vague or imprecise information. Also, the complexity of the problem can limit the application of pure logical rules. Further development of logical agents includes research into advanced inference techniques, such as probabilistic logical reasoning or combining logic with machine learning techniques to improve their adaptability and robustness in different environments.

3.2. **First-order logic**

First-order logic, also known as predicate logic, is a formal system for representing and reasoning about the relationships and properties of objects in the world. This type of logic makes it possible to precisely express statements about individual elements, their relationships and quantification. The basic elements of first-order logic include objects, relations (or predicates) that describe relationships between objects, functions that map objects to other objects, quantifiers that denote general statements about sets of objects,

and logical conjunctions that connect statements. Expressions in first-order logic consist of terms that represent individual objects or functions of objects, atoms that represent basic propositions about relationships between objects, and complex expressions that consist of atoms, logical conjunctions, and quantifiers. An example of the use of quantifiers in the logic of the first order "All kings are persons": $\forall x\ King(x) \Rightarrow Person(x)$.

3.3. **Inference in first-order logic**

First-order logic uses rules of inference, such as Modus Ponens or rules of inference by generalization, to derive new conclusions from existing propositions. First-order logic is widely used in artificial intelligence for formally expressing knowledge about a problem domain, making inferences, and solving problems such as diagnostics, planning, and reasoning about agent behavior. Although first-order logic is a powerful tool, it faces challenges in efficiently handling large amounts of information and complex problem domains. Also, it is sometimes necessary to extend first-order logic to deal with uncertainty or temporal aspects in problems.

3.4. **Knowledge representation**

Knowledge representation plays a key role in artificial intelligence because it allows computers to represent, manipulate, and understand the information needed to make decisions and solve problems. Knowledge representation is the process of converting information from the real world into a form that computer systems can understand and process. This includes the identification of essential entities, relationships and properties in the problem domain and their formal expression. There are different approaches to representing knowledge, including declarative (through statements and facts), procedural (through algorithms and procedures), and example-based (through sets of examples or patterns). In AI, formal languages such as first-order logic, ontology's, knowledge graphs, or rule-based languages are often used to precisely describe the problem domain and knowledge about it. Ontology's are formal models that describe concepts in a certain area, their properties and relationships between them. They enable the precise definition of conceptual structures and common understanding of information within a domain. As an example, knowledge graphs represent knowledge in the form of a graph where objects or subjects are represented by nodes, and the relationships between them are shown by branches. This structure enables an intuitive understanding of the relationships and connections between different objects.

4. UNCERTAIN KNOWLEDGE AND REASONING

4.1. **Quantitfying uncertainty**

Uncertainty measurement refers to the process of assessing and managing the degree of uncertainty or imprecision in the information we possess. It is used to quantify and understand the degree of reliability or unreliability in data or claims, which is essential for making informed decisions. Uncertainty measurement methods include different approaches, such as fuzzy logic, possibility theory or Dempster-Shafer theory, which allow modeling and treating different degrees of vagueness or indeterminacy in data.

4.2. **Probabilistic reasoning**

Probabilistic reasoning is the process of making inferences or predictions based on the likelihood of an event or outcome. This method takes into account not only the available information, but also its probability, which allows computers to make better decisions even in situations where there is uncertainty.

4.3. **Probabilistic reasoning over Time**

Reasoning based on probability and time extends the concept of reasoning based on probability to dynamic situations where events unfold over time. This method makes it possible to model and predict the probability of an event or outcome in the future taking into account both the current state and temporal features and patterns. Applications include time series forecasting, dynamic planning and resource management in time-sensitive environments. Weighing uncertainty and probabilistic reasoning, including the aspect of time, enable computers to make informed decisions even in situations where there is uncertainty or changing conditions, which is key to creating robust and adaptive artificial intelligence systems.

$$f(x) = a_0 + \sum_{n=1}^{\infty} \left(a_n \cos \frac{n\pi x}{L} + b_n \sin \frac{n\pi x}{L} \right) \tag{4.1}$$

$$F(k) = \int f(x)\, e^{-2\pi i k x} dx. \tag{4.2}$$

Using Fourier transforms in time series forecasting allows analysts to better understand periodic patterns and seasonal components in data and to develop efficient models for predicting future time series values. The ways in which the Fourier transform is used are: identification of periodic patterns, noise filtering, time series decomposition, spectral analysis and prediction based on frequency components.

4.4. **Probabilistic based programming**

Probability-based programming (PBP) is a technique used to solve decision-making problems where there are uncertainties and the probabilities of different outcomes are known or can be estimated. Graphical models are models that represent probabilistic dependencies between different variables using graphical structures such as Bayesian Networks or Markov Random Fields. Bayesian networks are based on conditional probabilities. As an example, let's look at the case of lung cancer diagnosis. Based on the established conditional probabilities in relation to the parameters of whether the patient is a smoker, whether he lives in a polluted environment, lung imaging results and symptoms of shortness of breath, the model can determine the probability that the patient really suffers from cancer. Various inference algorithms are used to draw conclusions based on models and data, such as exact inference, Monte Carlo methods, or variational methods.

4.5. **Making simple decisions**

This chapter explores decision-making strategies in situations where there is a clear problem structure and little uncertainty. Utility or value is a measure of utility that quan-

tifies agents' preferences regarding different outcomes. Utility can be defined mathematically as a function that maps different outcomes (that is, states) to real numbers, with higher values reflecting greater utility or satisfaction. Formally: $U(x)$ where x represents the outcome or state. Utility can be a function of a single outcome or an aggregated function that takes into account multiple outcomes. The basic principle of decision-making based on the selection of actions that maximize expected utility. Analogous to expected values, expected utility can be written as:

$$EU(a) = \sum_i P(s_i) \cdot U(s_i) \tag{4.3}$$

where s_i represents possible outcomes (states), and $P(s_i)$ is the probability of a particular outcome, and $U(s_i)$ is the utility or value of that outcome.

Bayesian theory is a theory that integrates Bayes' theorem with the concept of utility to make optimal decisions under uncertainty. Bayes' theorem allows updating the probability of hypotheses based on new evidence or information. If hypothesis H and evidence D are given, the probability of hypothesis H after taking into account evidence D (the so-called Posterior) can be calculated using the Bayesian formula:

$$P(H|D) = \frac{P(D|H) \cdot P(H)}{P(D)} \tag{4.4}$$

where: $P(H|D)$ posterior probability of hypothesis H after seeing evidence D, $P(D|H)$ probability of evidence D assuming that is the hypothesis H true (the so-called Probability of the evidence), $P(H)$ the prior value of the hypothesis H before we took into account the evidence D and $P(D)$ the probability of the evidence D regardless of the hypothesis.

4.6. **Making complex decisions**

Rational decision-making is a decision-making process that takes into account complex aspects of a problem, including multiple interests, multiple variables, and various aspects of uncertainty. There are many approaches to making complex decisions in AI, and some of the most well-known are decision schemes, Markov processes, probabilistic-graphical models, and Bayesian networks. Decision Trees are graphical representations that model the sequence of decisions and their consequences, helping to analyze and make decisions in complex situations. Markov Decision Processes are formal models that describe a sequence of decisions in a dynamic environment with uncertainty, and enable the optimization of long-term decisions.

4.7. **Making complex decisions with multiple agents**

There is a group of problems where different decision-making strategies need to be explored in situations where multiple agents interact with each other and have different goals. Nash equilibrium is a concept from game theory that describes a state in which no agent can increase his utility by changing his strategy, assuming that the other agents keep their strategies. The mathematical significance of the Nash equilibrium lies in its stability and relevance in the study of interactions between different agents or players. Cooperative vs. non-cooperative decision making are different approaches to

decision making in situations where agents may cooperate or compete with each other. Cooperative decision-making typically involves negotiation and resource sharing, while non-cooperative decision-making typically involves strategies to maximize benefits independent of the behavior of other agents.

5. Machine learning

When the agent is a computer, its learning process is called machine learning where the computer receives relevant data, builds a model based on the data, and uses this model both as a hypothesis about the environment and as software that can solve problems.

5.1. **Learning from Examples**

Machine learning can be supervised, where feedback provides accurate answers, or unsupervised, where patterns are inferred from data. Supervised learning includes regression for continuous values and classification for categorical outcomes. On the other hand, decision trees, informed by information gain, effectively represent Boolean functions. Linear regression and logistic regression are widely used models for linear and probabilistic classification. Non parametric models use the data to make each estimate instead of generalizing the data with parameters. Examples of these methods are nearest neighbor, support vector, and kernel methods. Ensemble methods such as bagging and boosting improve model performance by combining weak classifiers with the goal of obtaining a stronger classifier [11]. A very often used method in the field of artificial intelligence is linear regression with several variables. If the constant representing the intersection of the axis in the equation is multiplied by the "dummy" variable $x_{j,0}$ which we define as always equal to the number 1, the expression can be generalized as follows:

$$h_w\left(x_j\right) = \sum_i w_i x_{j,i} \tag{5.1}$$

where x are inputs, y are outputs and w are weight factors that need to be learned in the machine learning process. The best vector of weight factors, w^*, minimizes the squared loss error:

$$w^* = \underset{w}{argmin} \sum_j L_2(y_j, w\, x_j) \tag{5.2}$$

Using gradient descent, the minimum of the loss function will be achieved, and the equation of the updated coefficients will be:

$$w_i \leftarrow w_i + \alpha \sum_j \left(y_j - h_w\left(x_j\right)\right) \times x_{j,i}. \tag{5.3}$$

5.2. **Learning based on probabilistic models**

Statistical learning methods have a wide scope, from simple average calculations to complex models such as Bayesian networks. Bayesian methods use probabilistic inference to update hypotheses, while maximum posterior (MAP) learning selects the most likely hypotheses. The maximum likelihood learning method selects a hypothesis that maximizes the likelihood of the data, while naive Bayesian learning is effectively and

widely used. The expectation maximization (EM) algorithm facilitates learning using hidden variables. Statistical learning is a very lively research field that is advancing both theoretically and practically, with the goal of getting to the point where it is possible to learn any model for which inference is feasible [11]. An effective and widely used probabilistic model in the field of artificial intelligence is the Naive Bayesian model. The model is described by the following equation:

$$P(Cause, Consequence_1, \ldots, Consequence_n) = P(Cause) \prod_i P(Cause_i|Consequence). \tag{5.4}$$

This type of distribution is called naive, because it is usually used in cases where the *Consequence* variables are not strictly independent in relation to the causal variable.

5.3. **Deep learning**

Deep learning is a broad family of machine learning techniques in which hypotheses take the form of a complex algebraic circuit with adjustable connection strengths. The word "deep" refers to the fact that cars are usually organized in multiple layers. The basic model used for deep learning is artificial neural network, while for the minimization of the error function algorithms are used that propagate the error backward using gradient descent in the parameter space. Deep learning is proven to work well for object and speech recognition, and as reinforcement learning in complex environments. Convolutional networks stand out especially for image recognition, while recurrent networks are very efficient for processed strings of values/sequences [11]. The mathematical model of neural networks lies in the background of all described methods, and the basic unit within the network (neuron) can be described by the following equation:

$$a_j = g_j(\sum_i w_{i,j} a_i) \equiv g_j\left(in_j\right) \tag{5.5}$$

Where a_j is the output of unit $j, w_{i,j}$ is the weighting factor of the connection between units i and j, g_j is the nonlinear activation function associated with unit j, and in_j is the weighted sum of the inputs to unit j. The choice of activation functions is also diverse, but the most commonly used are logistic or sigmoid, ReLU (corrective activation function), softplus functions and the hyperbolic tangent [11] function.

5.4. **Reinforcment learning**

In supervised learning, the agent learns passively by observing examples of input and output pairs available to it. For real problems, there is often not enough data to learn from which the agent will be ready for new problems. An alternative to this approach is incentive learning, where an agent interacting with the environment periodically receives a reward/incentive that indicates the agent is on the right track. Agent design dictates the type of information that must be learned, so we distinguish between model-based agents where agents possess or acquire a model for the environment and a global goal, and model-free agents that learn a global goal or policy. Global objectives can be learned using multiple approaches, such as direct estimation, adaptive dynamic programming, temporal difference, Q-learning, deep support learning, reward shaping, and

policy search [11]. The simplest case of supported learning is a fully observable environment with a small number of actions and states, in which the agent already has a fixed policy $\pi(s)$ that determines the actions. The agent in this case tries to learn the global objective function $U^{\pi}(s)$ which represents the expected total reward if the policy π is executed starting in state s. This type of agent is called a passive learning agent. The global objective function can be defined as follows [11]:

$$U^{\pi}(s) = E\left[\sum_{t=0}^{\infty} \gamma^{r} R(S_t, \pi(S_t), S_{t+1})\right] \tag{5.6}$$

where $R(S_t, \pi(S_t), S_{t+1})$ is the reward/incentive received when the action $\pi(S_t)$ is taken in state S_t and reaches state S_{t+1}. Here S_t is a random variable that indicates the state reached at the moment t when policy π is executed, starting from the state $S_0 = s$. Given the nature of supported learning agents, and how environments become more and more complex, the advantages of this approach are likely to become more pronounced [11].

5.5. **Communicating, perceiving and acting**

To make actions in environments, computer systems communicate with users, perceive the environment, and make decisions based on this information. Key concepts in this area include: Understanding users' natural language and being able to respond to queries or commands in a way that is understandable and useful to users. The ability of computer systems to perceive and interpret their environment through sensors, cameras, microphones, and other input devices The process of making decisions based on collected information about the environment and system goals.

5.6. **Natural language processing**

Natural language processing is an area of artificial intelligence that deals with the understanding, generation and interpretation of human language. Text analysis involves understanding the structure and meaning of a text through techniques such as tokenization, lemmatization, entity extraction, and syntactic analysis. Understanding language involves understanding the semantic connections and contexts between words, sentences and text segments. Language generation means creating text outputs that are grammatically correct and meaningful based on entered queries or conditions.

5.7. **Natural Language Processing with deep learning**

When deep learning with neural networks with hundreds or thousands of layers is applied to the field of natural language processing, this technique enables computers to better understand and generate human language. The representation of words that a computer can understand is achieved as a vector representation that enables computers to efficiently process and understand the semantic connections between words. For Large Language Models (LLMs), vectors with multiple components are most often used, that is, members of an n-dimensional continuous vector space. By training these representation vectors over the corpus of words we use, we get that words that appear in

similar contexts within a language tend to be closer to each other. The process of learning vector representation is most often performed using unsupervised learning. Within the space of representations, there is a trained algebraic relationship between words, so that we can add and subtract words like for example

$$king - man + woman = queen. \tag{5.7}$$

Deep neural networks include various architectures such as: recurrent neural networks, transformers, deep convolutional neural networks, autoencoders, gated recurrent units (GRU) and generative adversarial network (GAN) models.

5.8. **Computer Vision**

Computer Vision is a popular subfield of AI or machine learning that deals with the ability of computer systems to analyze, interpret and understand visual information, such as images and videos. The goal is to enable computers to use visual perception in order to solve tasks or make decisions. This area encompasses a variety of tasks including object recognition, face detection, image classification, motion tracking, object segmentation, and more. At the beginning of the 21st century, with the advent of modern and capable graphics cards, deep learning strategies have shown record results in the fields of object detection and image classification. Among the most commonly used strategies in building models capable of real-time object detection from camera videos are deep neural networks. In the field of computer vision, convolutional neural networks are the most represented technique in the development of these models, next to newly created transformers. Convolutional neural networks consist of convolutional layers and pooling layers and can be represented by equations:

$$z_{ij} = (w * x)_{ij} + b \tag{5.8}$$

$$a_{ij} = f\left(z_{ij}\right) \tag{5.9}$$

$$a_{ij} = pooling\left(x_{ij}\right) \tag{5.10}$$

where x,w,b,f,z,a respectively are input data (eg image), convolution filter weights, bias, activation function, convoluted output, activation layer output. The inspiration for using convolution is partly the human brain and the ways in which we recognize the basic shapes of an image (edges, corners, colors), but also the mathematical operation of convolution. It is based on the concept of filtering and processing input data using filters. Let us denote the input data as X where X can be a $2D$ image or a multichannel image tensor ($X \in R^{C\times H\times W}$) where C is the number of channels (eg RGB). A filter is a smaller field or tensor applied to the input data to perform a convolution operation. Let K be the filter (kernel), where K is usually smaller than the input data and has dimension $C \times F \times F$ where F is the size of the filter. The convolution operation in the discrete case can be written as:

$$Y(i,j) = \sum_{m=0}^{F-1}\sum_{m=0}^{F-1} X(i+m, j+n)\cdot K(m,n) \tag{5.11}$$

where Y(i,j) is the map output value at position (i,j).

This form is a special case of the continuous, general form of the convolution of two functions in t:

$$(f * g)(t) = \int_{-\infty}^{\infty} f(\tau)\, g(t-\tau)\, d\tau. \tag{5.12}$$

6. Philosophy, ethics and safety of AI

The rapid development of powerful AI models brings with it risks and ethical questions that have yet to be resolved. Questions like how to ensure an equal set of data that adequately represents the population that will be modeled on, and how to act in situations when the AI model is not safe or the user himself doubts its output. These and many other issues of security and ethics are dealt with by the fields of AI security and AI ethics. The field of AI security is also concerned with discerning why the model produced the appropriate output for a given input. Various methods of interpretation of the structure of the interior of the neural network and approximation of individual parts of the network are used.

7. Future of AI and conclusions

The fusion of mathematics and artificial intelligence (AI) has the power and potential to transform the world and advance many industries. Artificial intelligence is deeply dependent on mathematics and logic, with the ability to process large amounts of data and make complex decisions. The mathematical framework of AI contains the results of linear algebra, statistics, probability theory, Bayes theorem, random processes, optimization, game theory, fractal mathematics, chaos theory, logic, vector and matrix theory, various methods of discrete and continuous mathematics. Together, mathematics and AI are transforming society, industries, education, productivity, and innovation. Mathematics becomes even more important and represents a constant "fuel" for the continuous development of AI. By providing an overview in this paper as a kind of catalog of leading mathematical fields, methods and applications in the field of AI, this paper provides mathematicians and artificial intelligence engineers with a basis for further research.

References

[1] OpenAI, *OpenAI*, OpenAI, [Online]. Available: https://openai.com/. [Last access 15 05 2024].
[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, *Attention is All you Need*, at 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 2017.
[3] D. P. Kingma and J. Ba, *Adam: A Method for Stochastic Optimization*, at 3rd International Conference for Learning Representations, San Diego, 2015.
[4] F. Rosenblatt, *The perceptron: A probabilistic model for information storage and organization in the brain*, Psychological Review, vol 65, no. 6, pp. 386-408, 1958.
[5] D. E. Rumelhart, G. E. Hinton and R. J. Williams, *Learning representations by back-propagating errors*, Nature, tom 323, pp. 533-536, 1986.
[6] M. I. Jordan, *Serial order: a parallel distributed processing approach*, Technical report, 1986.
[7] Y. LeCun, Y. Bengio and G. Hinton, *Deep Learning*, Nature, vol 521, pp. 436-444, 2015
[8] S. Hochreiter and J. Schmidhuber, *Long Short-term Memory*, Neural Computation, vol 9, 1997.

[9] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt and B. Scholkopf, *Support vector machines*, IEEE Intelligent Systems and their Applications, vol 13, no. 4, pp. 12-28, 1998.
[10] L. Breiman, *Random Forests*, Machine Learning, vol 45, pp. 5-32, 2001.
[11] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, London, Great Britain, Pearson, 2021.
[12] M. Swiechowski, K. Godlewski, B. Sawicki and J. Mańdziuk, *Monte Carlo Tree Search: a review of recent modifications and applications*, Artificial Intelligence Review, vol 56, no. 3, p. 2497–2562, 2023.

(Received: May 17, 2024)
(Revised: September 13, 2024)

Ervin Macić
ARTI Analytics Inc.
European Operations
Marsala Tita 6
71000 Sarajevo
Bosnia and Herzegovina
e-mail: *ervin.macic@artianalytics.com*
and
Tarik Hubana
ARTI Analytics Inc.
European Operations
Marsala Tita 6
71000 Sarajevo
Bosnia and Herzegovina
e-mail: *tarik.hubana@artianalytics.com*
and
Migdat Hodžić
ARTI Analytics Inc.
European Operations
Marsala Tita 6
71000 Sarajevo
Bosnia and Herzegovina
e-mail: *migdat.hodzic@artianalytics.com*

PROCEEDINGS OF THE CONFERENCE ON MARCH 14 – INTERNATIONAL DAY OF MATHEMATICS
Special Editions ANUBiH, Book CCXVI, OPMN Book 30
Sarajevo 2024, **p. 114–136** DOI: 10.5644/PI2024.216.13

MATHEMATICAL MODEL OF THE LORENZ CURVE: ON BALANCING THE WEALTH OF COMMUNITIES

MIGDAT HODŽIĆ

Dedicated to the 75th birthday of our dear Professor Mirjana Vuković

ABSTRACT. This paper presents a new (i) mathematical approach to modeling the wealth of a community and (ii) a method of balancing this wealth, with the idea of fair distribution among community members. The paper evaluates how far the wealth of a community (country, the world) is from the ideal, measured by the so-called Lorenz curve, known from the domain of economics. The work can also be interpreted as a detailed mathematical model of the Lorenz curve with a recursive algorithm for "correcting" that curve, which is ideally a line in a two-dimensional space (number of community members, their wealth). In economic literature, it is easy to find descriptions of Lorenz curves for various countries, in the form of a set of straight lines (which approximate an otherwise non-linear function) for various groups of wealth. Another method based on the Lorenz curve is the so-called The Gini index, which also measures differences in wealth. Unlike such standard models, our work presents a detailed mathematical model of Lorenc curve as well as wealth balancing. Mathematically, the algorithm describes in detail a recursive method that "corrects" the central part of the Lorenz curve, until it becomes a linear function, thereby illustrating the "balancing" of wealth in the community. The central part of the curve represents the middle class in a community. The described model also "mathematizes" various wealth groups and precisely defines them.

1. INTRODUCTION

In studies of the inequality of the financial distribution of world wealth, the basic tools used are the Lorenz curve and the Gini index. The Lorenz curve was first introduced back in 1905 when the American researcher Max O. Lorenz devised a curve that represents a measure of the financial inequality of the distribution of wealth in a given community. The curve represents the functional relationship between the number of community members and their wealth. The Lorenz curve becomes a line (completely "straightens out") when wealth is balanced (eg 50% of people own 50% of the wealth). A few years later, in 1912, the Italian statistician and sociologist Corrado Gini defined an index that numerically, with a single number, indicates the level of inequality [1], [2], [3]. The Gini index is usually defined in terms of the Lorenz curve, but it can be defined in other ways. A community in which only one person owns all the wealth would correspond to a Gini index of 1, and a just society would correspond to a Gini index of 0. At the world level, the Gini index is currently around 0.6, depending on

2020 *Mathematics Subject Classification.* 39A06.

Key words and phrases. Community wealth, Lorenz curve, wealth distribution model, wealth balancing algorithm.

the data source [4]. The simplicity of a single number is the main value of the Gini index. On the other hand, the same index can be produced by different distributions, which can be considered a disadvantage. The Lorenz curve and the Gini index can be applied to many areas, financial and non-financial (ecology, health), where there is a lack of balance in some parameters. Indices can be applied to a specific industry, group of people, or even species in an ecology. References [5] and [6] describe financial applications. Other applications, [7], [8] and [9], describe education and healthcare models. There are other measures of inequality. such as Robin Hood index (Figure 2.2, also known as Hoover or Schutz index), Atkinson and Suits indices [10]. Our interest in this work is the development of (i) a mathematical model of the Lorenz curve and (ii) a recursive model of balancing wealth with social giving. The results indicate (iii) how far the world is from financial balance (156 years), as well as (iv) the importance of the middle class in balancing. The balancing algorithm shows how the middle class "pulls" wealth towards the balance. There is a huge imbalance between the rich and the poor in the world, in most human societies and countries, whether small or large. Furthermore, unfortunately, the recent events in world finance (the last crisis of 2008-2009 as well as the previous similar credit crises of the 1980s) testify to the growing rift between the richest and the poorest, and there is no visible action plan or even any desire of the strongest and richest to change this trend. In the annual world financial reports, Credit Suisse (and many other similar organizations, including the UN), [11]-[15], published numerous reports indicating a trend of fewer people owning more and more wealth. Reports show that less than 1% of the richest own more than 50% of the world's wealth. In this paper, we consider a mathematical model for community wealth, based on which wealth is balanced by social giving. This is a naive impractical assumption, but it is the first step in understanding how far a community is from balance in wealth. In our other works, we continue this model with the addition of investing. The ultimate goal of this research is to show that wealth is greater if investing and social giving are combined, compared to investing alone. Our interest is not only mathematical but to some extent also ideological, because we believe in a just world. In such a world, the Lorenz curve is a straight line (even distribution of wealth) and the Gini index is 0.

The paper is organized as follows. Section 2 provides a brief summary of the basics of the Lorenz curve and the Gini index, illustrated using data for several countries. Other indices are also mentioned, e.g. Robin Hood Index. Section 3 describes a new wealth distribution model, the "mean halved" (MH) wealth distribution model, which is based on a set of linear approximations from the richest to the poorest. The model is general and illustrated with examples. Section 4 deals with the specifics of the new model of wealth distribution, where it is assumed that all but the poorest group give a fixed percentage of their wealth to the community. As the wealth balancing algorithm progresses (per a given period of giving, for example annually), the initial unfavorable curve changes step by step, eventually leading to a perfect straight line. Various boundary conditions are satisfied as the algorithm progresses for each new dispensing cycle. At the end of Section 4, examples of the "balancing table" as well as the current wealth curve in the world, using the MH model presented in this paper, are given. The conclusion is in Section 5. At the end of the paper is a list of references.

2. LORENZ CURVE AND INEQUALITY INDICES

The first part of the paper presents the Lorenz curve and gives several examples of countries and their curves. From a mathematical point of view, all curves are non-linear functions (wealth on the ordinate, number of community members on the abscissa, or vice versa). The "greater" the nonlinearity, the greater the difference in wealth between the poorest and the richest.

2.1. **Lorenz curve, Gini index**

Examples of Lorenz curves taken from different sources [12] are shown in Figure 2.1. All curves are only rough approximations. Our mathematical model is much more accurate. Although not visually equivalent to the standard Lorenz curve. the working model has the same information with the same line of balanced wealth. Figure 2.1 shows the Lorenz curves for Bangladesh, the UK, Brazil and the world. Interestingly, Bangladesh is ahead of Great Britain and Brazil, well ahead of the world average, but with less overall wealth. Figure 2.2 shows the general shape of the Lorenz curve and defines the Gini and Robin Hood indices. In our work, the shape of the Lorenz curve is adapted for the needs of precise mathematical analysis.

2.2. **Index of social giving**

Definition 2.1. *The social giving index "D" is the portion of the total wealth of the community (W_T) that can be given by all working community members for the common good of the poorer in the community:*

$$D = W_T / Z \tag{2.1}$$

Examples: Z=20, 40, 100, i.e. 5%, 2.5%, and 1% allowance. Index D can be constant or variable. Index D can be used to balance the wealth of the community, and a method to improve the economic situation. Our work presents the problem of inequality and balancing in a precise mathematical way, provides a method for explaining and assessing how far each community is from the ideal Lorenz line, and lays out the foundation for economic policies that reduce damaging economic booms and busts.

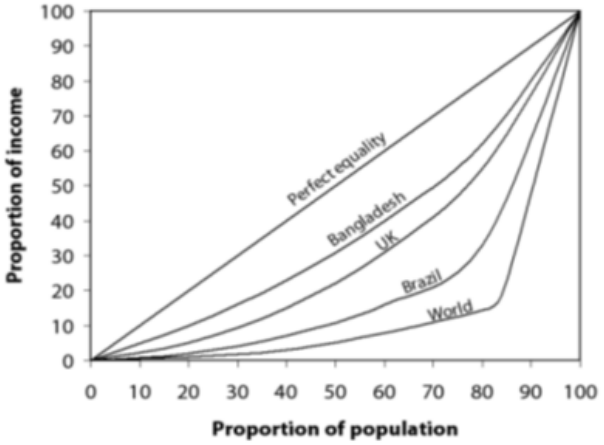


FIGURE 2.1. *Lorenz Curve Examples*

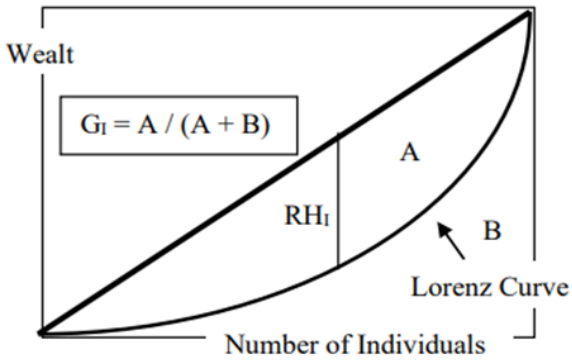


FIGURE 2.2. *Gini and Robin Hood indexes*

2.3. **Wealth Investment Index**

Our research is focused on how to combine social giving and wealth investing to achieve individual financial goals and social equity. Similar to the giving index D, we have:

Definition 2.2. *The wealth investment index "U" is the part of wealth W_T that can be invested:*

$$U = W_T/V \tag{2.2}$$

where V is the investment ratio. E.g. V=20, 40, 100 corresponding to 5%, 2.5%, 1% investment. In future work, we show that the combination of giving and investing can lead to very positive economic strategies, i.e. (i) increasing the total wealth of W_T and (ii) reducing the total risk of R_T:

$$W_T(investing + giving) > W_T(investing) \tag{2.3}$$

$$R_T(investing + giving) < R_T(investing) \tag{2.4}$$

3. WEALTH DISTRIBUTION MODEL

In this paper, we present a new wealth model that can be used for a variety of applications, from wealth balancing, to modeling constant (normalized) and variable community wealth, including analysis of the "distance" of wealth from the ideal Lorenz line.

3.1. **Basic assumptions of the model**

We will now describe a simple wealth distribution model in which we assume a q-quantile distribution of wealth with q = 2, and apply it repeatedly to the desired level of wealth granularity. Simply put, we divide the total wealth W_T of the community into two halves, i.e. we find the middle point, as the initial distribution of wealth. We then divide one of the halves representing the poorer end of the wealth spectrum into two halves (two quarters of the original half) and so on, cutting the poorer end of the wealth in half. We assume that the distribution of wealth within each group is uniform. It can be generalized.

Definition 3.1. *The distribution model of the "mean halved" (MH) total wealth of W_T is:*

$$\begin{aligned} W_T &= W_T/2 + W_T/2 \\ &= W_T/2 + W_T/4 + W_T/4 \\ &= W_T/2 + W_T/4 + W_T/8 + W_T/8 \\ &= W_T(1/2 + 1/4 + 1/8 + 1/16 + \ldots + 1/(2L) + 1/(2L)) \end{aligned} \tag{3.1}$$

where L+1 is the total number of groups in the given community. This can also be written as:

$$W_T = W1 + W2 + \ldots + W_L + W_{L+1} = \sum W_m, m = 1, 2, ..., L+1 \tag{3.2}$$

with the i-th group:

$$W_i = W_T/2^i = \sum W_n, i = 1, 2, ..., L; n = i+1, ..., L+1 \tag{3.3}$$

that is, each previous wealth is the sum of the remaining wealth, with the boundary condition $W_{L+1} = W_L$. In addition, the total number of community members can be divided into the corresponding sum:

$$N_T = N_1 + N_2 + \ldots + N_{L+1} = \sum N_m, m = 1, 2, ..., L+1 \tag{3.4}$$

The last group W_{L+1} can be further divided into two smaller groups as desired. Finally, we come to the point in equation 3.3 when we decide that W_{L+1} is the poorest group, for practical and mathematical reasons. The last N_{L+1} corresponds to the last W_{L+1}. Additionally, $N_L \neq N_{L+1}$, with $N_{L+1} > N_L$ for most practical cases. For the whole world, for the poorest individual, L is 31, for about 3.4 billion working adults in 2023 [16], which is between 231 and 232. For a large city of about 1.05 million working people, L = 20. Practically we are not going to the level of an individual, but to the level of a large group of individuals. Through several examples we came to L = 6 as a practical number, so W_7 is the poorest group. Our MH model is very flexible and can be adjusted to any desired level of precision. Our main goal here is to use the wealth model for:

(1) Defining recursion for social giving and determining when group wealth balance is achieved. In further works we will describe:
(2) Defining a giving index, constant, variable, or sectoral, as a measure of the group's contribution
(3) A variable group wealth model that includes social giving as well as wealth investment to understand their relationship and changes in wealth between two or more periods of giving and investment
(4) Analysis of sensitivity, robustness and risk of the wealth model
(5) Demonstrate the beneficial effect of combined social giving and investment.

This is the key result of our research project, the first step of which is described in this paper.

Figure 3.1 shows the (non-linear) distribution of wealth approximated with linear segments. with the property of "halved middle" based on equations 3.1 - 3.4. In our model, we plot individual wealth along the vertical axis, not total wealth as in the stan-

dard Lorenz curve. Total wealth in our model is the area under the line segments in Figure 3.1. This is done to clearly define the balancing algorithm in Section 4. The horizontal axes show the number of community members (poorest on the right, richest on the left) divided into groups, so that each area corresponds to the total wealth in that group, according to equation 3.3. The size of the groups is not in the precise relationship in Figure 3.1.

The shape of the distribution function is general and can be applied to any practical situation. For example, if Figure 3 represents the world wealth, then w0 = 212 billion dollars is the current wealth of Frenchman Bernard Arnault (not Elon Musk anymore!) [16], and N1 < 1% is the number of the richest people in the world. The value of w1 can be calculated using a simple geometric calculation from Figure 3, given w0 and W1. The variable w1 represents the smallest amount of wealth in the W1 group. If the 1-line model is not valid, then the 2- or 3-line model is modified (Section 3.4). Values along the axes can be normalized to 1 or 100 to work with percentages. Normalized values to 100 are defined as:

$$\begin{aligned} &\text{Individual wealth} : w_n or = 100(w/w0), \\ &\text{Number of community members} : n_n or = 100(n/nL+1). \end{aligned} \tag{3.5}$$

3.2. **Description of the basic model**

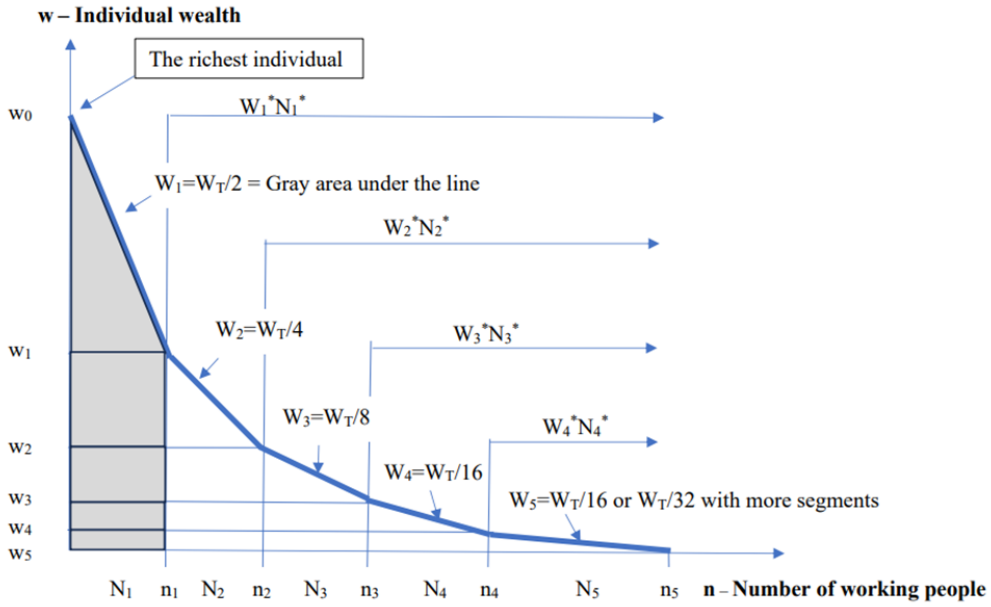


FIGURE 3.1. *MH Wealth Distribution Model*

The total wealth in Figure 3.1 is the area under the line segments. We start with group W_1. The wealth of that group is equal to the area under the first line

$$W_1 = W_T/2 = n1w1 + (n1/2)(w0 - w1) = N1(w1 + w0)/2 \tag{3.6}$$

where $N1 = n1 - n0$, $n0 = 0$. Also, W_1 is the sum of all individual wealth in the group:

$$W_1 = W_T/2 = \sum w1^m \tag{3.7}$$

where m = 1,2,...,N1 , $w1^1$ = w0 , w1 = $w1^{N1}$. This represents the sum of all the individual wealths ($1xw1^m$) Figure 3.1. The next wealth W_2 is calculated as:

$$W_2 = W_T/4 = (n2 - n1)w2 + (n2 - n1)(w1 - w2)/2 = N2(w2 + w1)/2 \tag{3.8}$$

where $N_2 = n_2 - n_1$. As in 3.8 we have:

$$W_2 = W_T/4 = \sum w2^m \tag{3.9}$$

and m = 1,2,...,N_2 i $w_2^1 = w_1$, $w_2 = w_1^{N2}$. Using induction we have a general result for W_i:

Result 3.1. Wealth W_i is the vertical area under the corresponding linear segment in Figure 3.1:

$$W_i = W_T/2^i = (n_i - n_{i-1})(w_i + w_{i-1})/2 = N_i(w_i + w_{i-1})/2 \tag{3.10}$$

$$N_i = n_i - n_{i-1}, n_0 = 0 \tag{3.11}$$

where N_i represents the number of members of the group W_i. Each W_i, $i = 1, 2, 3, \ldots, L + 1$ also represents the corresponding sum of individual wealths:

$$W_i = W_T/2^i = \sum wi^m \tag{3.12}$$

$$wi^1 = w_i, w_i = wi^{Ni}, m = 1, 2, \ldots, N_i \tag{3.13}$$

From 3.13 we obtain:

$$W_i = W_{i-1}/2 \tag{3.14}$$

and combined with 3.12 we get:

$$Wi_1 = W_T/2^{i-1} = N_i(wi + w_{i-1}) \tag{3.15}$$

The above equations are used when normalizing recursive formulas. Plus, W_T can also be expressed as integrals:

$$W_T = \int W(n)dn = \int N(w)dw \tag{3.16}$$

$$= \sum W_i = \sum \sum wi^m, i = 1, 2, \ldots, L + 1, m = 1, 2, \ldots, N_i. \tag{3.17}$$

The limits of the first integral are w_{L+1} (practically very close to 0) to w0, and of the second are from 0 do n_{L+1}. The two functions W(n) and N(w) are inverses of each other, and the two integrals can be calculated by some numerical methods starting from $\sum W_i$ u 3.17. From Figure 3.1 we see that the function W(n) consists of individual lines:

$$W_i(n) = K_i(n + ni*) \tag{3.18}$$

where K_i, is line W_i(n) slope:

$$K_i = (w_i - w_{i-1})/(n_i - n_{i-1}) = w_i^*/(n_i - n_{i-1}) \tag{3.19}$$

$$w_i^* = (w_i n_i - 1 - w_{i-1} w_i)/(w_i - w_{i-1}) \tag{3.20}$$

and w_i^* is the point of intersection of the vertical axes, given W_T, w_0 and N_i while w_i is

calculated recursively from:

$$w_i = (2W_i)/N_i - w_{i-1} = (W_T/2^{i-1})/N_i - w_{i-1}. \tag{3.21}$$

The corresponding inverse functions $N_i(w)$ can be found by solving 3.18 for w as a function of n. The above calculations are used in future works where we calculate the number of community members, i.e. balancing the horizontal axis in Figure 3.1, after wealth has been balanced (Chapter 4). These two balancing operations form the basis for the wealth normalization process. In the normalized case, the ideal Lorenz line corresponds to the function W(n):

$$W(n) = -Kn + w^* = -n + 100 \tag{3.22}$$

$$W_i(n) = K_i n + N_i^* = -n + N_i^* \tag{3.23}$$

where K_i = - 1, i = 1,2,..., L+1 represents the slope of the local Lorenz line, which allows us to accurately categorize different groups into terms of rich, middle class and poor (or even more precisely).

3.3. **Definition of wealth classes**

From Definition 2.2 and the equations above, we can obtain a precise mathematical definition of wealth class. We formally define different classes as one or more groups with a certain relationship to the ideal Lorenz line. Using 3.22 i 3.23 above, we have:

Definition 3.2. *The classes of wealth in the MH model of wealth distribution are:*

$$\begin{aligned} &\textit{Rich Classes: } |K_i| = |K_R| >> 1 \\ &\textit{Middle Classes: } |K_i| = |K_M|, |K_P| < |K_M| < |K_R| \\ &\textit{Poor Classes: } |K_i| = |K_P| << 1 \end{aligned} \tag{3.24}$$

where "i" denotes some range of values, in each of the above 3 general wealth groups. Furthermore, we can divide the middle class into three separate groups:

$$\begin{aligned} &\textit{Upper Middle Class: } 1 < |K_M| < |K_R| \\ &\textit{"Middle" Middle Class: } |K_M| \approx 1 \ \ (\textit{closest to the ideal value of 1}) \\ &\textit{Lower Middle Class: } |K_P| < |_K M| < 1 \end{aligned} \tag{3.25}$$

In this paper, after testing several examples, we selected 7 *groups Wi with* $i = 1, 2, \ldots, L + 1$, $L = 6$.

The importance of the middle class lies in the fact that it is the closest to the ideal Lorenz line. It is known from economic theory that the more numerous the middle class is, the better it is for the community. We also divided the rich and poor classes into two groups. With L = 6, the MH wealth distribution model can be precisely defined as:

Definition 3.3. *Wealth groups W_i, i=1,2,...,7, i=1,2,...,7, in the MH model are 100% in total, i.e.:*

$W1$ – *Super rich group* (50% *of wealth*)
$W2$ – *Very rich group (25% of wealth)*
$W3$ – *Upper middle class group (12,5% of wealth)*
$W4$ – *Middle middle class group (6,25% of wealth)*

$W5$ – *Lower middle class group (3,125% of wealth)*
$W6$ – *Poor group (1,5625% of wealth)*
$W7$ – *VVery poor group (1,5625% of wealth)* (3.26)

with the corresponding number of group members N_i *and line slopes* K_i *and* $i = 1, 2, \ldots, 6, 7$.

A few comments are in order:

(1) The super-rich group can be further divided into $W_1 = W_1^1 + W_1^2$, where W_1^1 is a few thousand extremely rich individuals, who own about 16.7% of the world's total wealth, and W_1^2 is a very rich group with 1% of the world's population who owns 1/ 3 or 33.3%.
(2) Rich W_2 is a "transient" group, give as much as you get (see Section 4)
(3) When all the W_3, W_4 and W_5 middle class percentages are added together, we get 12.5% + 6.25%+ 3.125% = 21.875% which agrees very well with the UN estimate of middle class wealth at around 22%. This may be a coincidence, or an indication of the general validity of our wealth model.

By combining Definitions 3.2 i 3.3, with Figure 5.1 (which will be further explained below), we get:

Result 3.2. Coefficients of Lorenz lines and W groups (as of 2016, updated for 2024):

$W1$ – "Super Rich" group: $|K_1| = 50$
$W2$ – "Rich" group: $|K_2| = 10$
$W3$ – "Upper Middle Class" group: $|K_3| = 3$
$W4$ – Middle "Middle Class" group: $|K_4| = 1,08$
$W5$ – "Lower Middle Class" group: $|K_5| = 0,45$
$W6$ – "Poor group": $|K_6| = 0,043$
$W7$ – "Very Poor" group: $|K_7| = 0,028$ (3.27)

with $|K4|$ = 1,08 for W4 as the closest to the ideal Lorenz line $|K| = 1$, per Definition 3.1.

3.4. **Modified linear model**

If the calculation in 3.21 does not give a reasonable value for wi, it implies that the 1-line Wi model in Figure 3.1 should be corrected. If we look at group W1, less than 1% of individuals own more than 50% of the wealth. When we made the first drafts of our model in 2016, the wealth situation was a little "better", namely about 1% owned about 50%. If we calculate w1 from 3.21, we get a negative value. Therefore, the W_1 group should be broken down into several lines (the same area under W_1 is kept) to calculate w1. Figure 4.1 shows any W_i modeled with 2 - line segments instead of one.

3.4.1. ***Linearity test.*** To check the validity of the 1-line model we need a linearity test. First, w_i is calculated from equation 3.21. If the value is positive (and economically "reasonable"), the linearized model is valid. Otherwise w_i is broken into multiple linear segments as shown in Figure 4.2, with two linear segments. If necessary, the process is repeated. With or without the linearity test, the balancing algorithm of Section 4

along the "w" axis holds either way because W_i, $W_T/2^i$, with one or more lines.For balancing along the "n" axis, we must satisfy the boundary conditions between the groups $(n_1, N_1, w_1; n_2, N_2, w_2; ...,)$. Details are given in our future works.

4. RECURSIVE WEALTH BALANCING MODEL WITH SOCIAL GIVING

In this paper, we present several reasonable and practical assumptions regarding social giving and wealth balancing. Due to limitations in the size of the text, this paper deals only with the MH model and the mathematics of social giving. Investment, as another key part of the model, is described in other texts.

4.1. **Basic assumptions**

(1) During the cycle of giving and investing (one year) the total wealth of W_T is either constant (normalized to 100) or variable, which can also be normalized in each cycle.
(2) The poorest members of the community do not contribute to social giving. This is the W_i group that "does not give". This assumption is logical and fair in practice, and is often practiced or at least recommended in various world religions. On the other side of the model, the super rich group only gives to the lower classes. These assumptions are reflected in our mathematical model.
(3) Social giving is evenly distributed in the community and reaches all corners of the community, with contributions from the rich to the poor, along defined wealth groups. The richest group only gives, the other groups give and receive.
(4) Before each new cycle, we assume that total wealth is either (i) normalized or (ii) not normalized. Normalization can be done in several ways, as discussed below. This can include new wealth generated between two cycles (say through investing)
(5) We assume that each member of the community (except the poorest group) gives an equal share (percentage) of his wealth, Z, which is distributed for the common good, equally throughout the community. Our model allows variable giving Z as well as "sectoral" giving say for agriculture or other areas of interest.
(6) Individual groups can invest an arbitrary part of their wealth according to their needs and wishes.

4.2. **Normalized wealth notation**

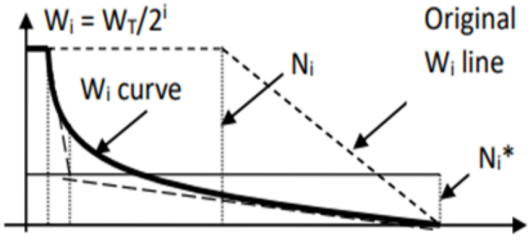


FIGURE 4.1. *W_i with 2-line approximation*

With the assumptions from Section 3, we continue with the description of the recursive wealth balancing algorithm. We introduce the normalized dynamical notation for

the group W_i from Figure 3.1: for the kth cycle, and 1/2 part of the wealth in equation (6). Figure 4.2 showing the normalized W_T, divided into mean halved groups. The notation on the right-hand side of (32) is a bit complicated, but serves the purpose of explaining the algorithm at this point. We will return to a simpler notation shortly.

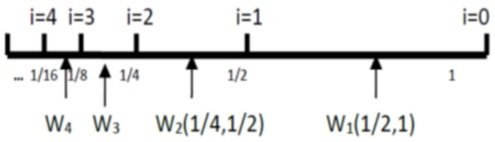


FIGURE 4.2. *Wealth normalization*

The initial condition before any social giving corresponds to k = 1, and for the richest half of the wealth, i.e. i = 1, we have W1 in Figure 3 which corresponds to W1(k)=Wk(1/2,1) in (32) notation. This corresponds to the richest half of wealth from 1/2 to 1 in normalized terms in Figure 4.2 when total normalized wealth is:

$$W_i = W_i(k) = W_k(1/2^i, 1/2^{i-1}) \tag{4.1}$$

for the kth cycle, and 1/2 part of the wealth in equation 3.2. Figure 4.2 showing the normalized W_T, divided into mean halved groups. The notation on the right-hand side of 4.1 s a bit complicated, but serves the purpose of explaining the algorithm at this point. We will return to a simpler notation shortly. The initial condition before any social giving corresponds to k = 1, and for the richest half of the wealth, i.e. i = 1, we have W_1 in Figure 3 which corresponds to W_1(k)=W_k(1/2,1) in 4.1 notation. This corresponds to the richest half of wealth from 1/2 to 1 in normalized terms in Figure 4.2 when total normalized wealth is:

$$W_T/W_T = 1 = \sum W_i/W_T. \tag{4.2}$$

For i = 2, we have W_2(k)=W_k(1/4,1/2) which corresponds to the first quarter of the wealth remaining from W_k(1/2,1). When i = 3 we have W_k(1/8,1/4) which corresponds to the first 8^{th} of the remaining wealth of W_k(1/4,1/2), etc. We also note that after each normalization step , for any k:

$$\begin{aligned} W_k(1/2,1) &= W_1(k) = 1/2 \\ W_k(0,1/2) &= W_1^*(k) = 1\text{–}W_1(k) = 1/2 \end{aligned} \tag{4.3}$$

where W_1(k) represents 50% of the total wealth of the community owned by the richer, and the complementary W_1^*(k) is 50% owned by the other less rich and poorer (see Figure 3.1 for W_1^*(k)). We then divide W_1^*(k) into two equal quarters representing the two quarters of the poorest 50%:

$$\begin{aligned} W_k(1/4,1/2) &= W_2(k) = 1/4 \\ W_k(0,1/4) &= W_2^*(k) = W_1^*(k) - W_2(k) = 1/4 \end{aligned} \tag{4.4}$$

followed by dividing W_2^*(k) into equal 8s representing the two 8ths of the poorest 25%:

$$\begin{aligned} W_k(1/8,1/4) &= W_3(k) = 1/8 \\ W_k(0,1/8) &= W_3^*(k) = W_2^*(k) - W_3(k) = 1/8 \end{aligned} \tag{4.5}$$

etc., through the general i-th group W_i:

$$W_k(1/2^i, 1/2^{i-1}) = W_i(k) = 1/2^i$$
$$W_k(0, 1/2^i) = W_i^*(k) = W_{i-1}^*(k) - W_i(k) = 1/2^i \tag{4.6}$$

The final step is to divide the last group into two equal parts:

$$W_k(1/2^{L+1}, 1/2^L) = W_{L+1}(k) = 1/2^{L+1}$$
$$W_k(0, 1/2^L) = W_L^*(k) = W_{L-1}^*(k) - W_L(k) = 1/2^L \tag{4.7}$$

where $W_L^*(k)=W_{L+1}(k)$ is the last group that does not give, its members only receive from other groups. Each group in 4.1-4.7, i.e. $W_1, W_2, \ldots, W_L$ (richer halves) gives a fixed proportion to their complementary pairs $W_1^*, W_2^*, \ldots, W_L^*$ (poorer halves), while the first group W1 only gives and receives nothing. As for the W_i^*s, those "complementary" groups also give and receive, except for the last $W_L^* = W_{L+1}$ which only receives. This holds under a fixed total wealth and fixed giving assumption. New wealth produced (or lost) between giving periods can easily be incorporated into our model, normalized or not. Our future work elaborates a variable giving index. We also note that in the normalized case $W_i(k) = W_i^*(k)$, i=1,2,...,L, the right and left sides of the center 2^i in Figure 4.2, plus $W_L^*(k) = W_{L+1}(k)$. If normalization is not performed, this is no longer the case as the wealth values of the groups change. In essence, total wealth is normalized if the assumptions in Definition 1 hold. More on that below.

4.3. **Algorithm for balancing the MH wealth model**

The recursive social giving algorithm starts with $W_k(1/2,1)$, the richest 50% of the total wealth of W_T. Giving starts here. We assume in the kth period that a fixed percentage Z of $W_k(1/2,1)$ wealth is given uniformly across the poorer group $W_k(0,1/2)$. In practical terms this can be given to different social needs across $W_k(0,1/2)$. We have two situations: (i) k-iteration without wealth normalization, i.e. we do not fit $W_i(k)$ to Definition 1 for each k. (ii) With normalization, all k (Figure 3.1) the groups $W_i(k)$ and Ni are recalculated per Definition 1. In both cases we use the same simplified notation W_i(k).With normalization, we get very interesting (perhaps fundamental) relationships between wealth groups.

4.3.1. ***Model of social giving without normalization.*** This corresponds to n_i (N_i) and w_i fixed in Figure 3.1. The following equations describe the give/receive in the (k+1)-th period as a function of the previous kth cycle. Here we switch to simplified notation:

$$W_1(k+1) = W_1(k) - W_1(k)/Z = (Z-1)W_1(k)/Z$$
$$W_1^*(k+1) = W_1^*(k) + W_1(k)/Z. \tag{4.8}$$

Then the poor portion of 50% of W_1(k) is shifted and divided into two quarters W_2(k) and W_2^*(k), where a quarter of W_2(k) gives and a quarter of $W_2^*(k)$ receives from both W_1(k) and W_2(k). The contribution of W_1(k) is equally divided between two quarters

of W_2(k) and $W_2^*(k)$, so $W_2^*(k)$ receives only half of W_1(k)/Z and full W_2(k)/Z:

$$W_2(k+1) = W_2(k) + W_1(k)/(2Z) – W_2(k)/Z \tag{4.9}$$

$$W_2^*(k+1) = W_2^*(k) + W_1(k)/(2Z) + W_2(k)/Z. \tag{4.10}$$

Here W_1(k) decreases, W_1^*(k),W_2(k),W_2^*(k) increase. We continue in the same way and using induction we get:

Result 4.1. The recursive wealth of the i-th group W_i(k) in the non-normalized balancing model of the "mean halved" MH wealth model with the assumption of a constant endowment index is:

$$W_i(k+1) = (Z-1)W_i(k)/Z + \sum W_{i-n}(k)/(2^n Z) \tag{4.11}$$

$$W_i^*(k+1) = W_i^*(k) + \sum W_{i-m}(k)/(2^m Z) \tag{4.12}$$

$$W_L^*(k) = W_{L+1}(k+1) \tag{4.13}$$

$$W_T = \sum W_p(k) + W_i^*(k) \tag{4.14}$$

with n = 1,2,...,i-1, and m = 0,1,...,i-1 i p = 1,2,...,i. The algorithm goes to i = L, when we have a final giving from W_L(k) and receiving at W_L^*(k) = W_{L+1}(k).

The equations in Result 4.1 can be simplified with the notation:

$$W_i(k+1) = (1-A_Z)W_i(k) + A_Z \sum B_{in} W_n(k) \tag{4.15}$$

$$W_i^*(k+1) = W_i^*(k) + A_Z \sum B_{im} W_m(k) \tag{4.16}$$

where $A_Z = 1/Z, B_{in} = 1/2^{i-n}$, with $n = 1,2,\ldots,i-1$, $m = 0,1,\ldots,i-1$, $i = 1,2,\ldots,L$. Then, a new cycle of giving begins and we go to $k+2$, etc., until the Lorentz curve is an ideal line when a uniform distribution of wealth is achieved. This happens in the ideal case when K_i u 3.22 has the same value for all i.

Tests for achieving balance are in Section 4.6.1. The balancing algorithm stops at this point. Before renormalization, equations 4.2 -4.6 applied to the iterated wealth values do not hold. Not yet, not until the new renormalization is complete, as described in the next section. The expression $\sum B_{in} W_n(k)$, $n = 1,2,\ldots,i-1$, in Result 4.1 in the normalized case for $W_i(k) = W_T/2_i$ reduces to:

$$\begin{aligned}\sum B_{in} W_n(k) &= W_1(k)/2^{i-1} + W_2(k)/2^{i-2} + \ldots W_i - 1(k)/2^i \\ &= W_T/2^i + W_T/2^i + \ldots + W_T/2^i = (i-1)W_i(k).\end{aligned} \tag{4.17}$$

The mentioned equation is very important, because it connects the general and normalized case.

4.3.2. ***Model of social giving with normalization.*** We proceed with further simplifications of the above equations, when normalization is performed between each k and k+1 administration cycle. This case corresponds to the variables n_i (N_i) and w_i. As in the non-normalized case 4.3.1, we start from the same equations:

$$W_1(k+1) = W_1(k) - W_1(k)/Z = (Z-1)W_1(k)/Z \tag{4.18}$$

$$W_1^*(k+1) = W_1^*(k) + W_1(k)/Z = (Z+1)W_1(k)/Z \tag{4.19}$$

where we used the fact that the two terms $W_1(k)$ and $W_1^*(k)$ are equal when normalization is performed. Then we use Definition 3.1 to calculate:

$$\begin{aligned} W_2(k+1) &= W_2(k) + W_1(k)/(2Z) - W_2(k)/Z \\ &= W_2(k) + W_2(k)/Z - W_2(k)/Z = W_2(k) \end{aligned} \tag{4.20}$$

$$\begin{aligned} W_2*(k+1) &= W_2^*(k) + W_1(k)/(2Z) + W_2(k)/Z \\ &= W_2*(k) + W_2(k)/Z + W_2(k)/Z \\ &= W_2^*(k) + W_2^*(k)/Z + W_2^*(k)/Z \\ &= (Z+2)W_2^*(k)/Z. \end{aligned} \tag{4.21}$$

We see that $W_2(k+1) = W_2(k)$, and that the W_2 group gives the same as it receives ("passive group"). For W_3 we have: We see that $W_2(k+1) = W_2(k)$, and that the W_2 group gives the same as it receives ("passive group"). For W_3 we have::

$$\begin{aligned} W_3(k+1) &= W_3(k) - W_3(k)/Z + W_1(k)/(4Z) + W_2(k)/(2Z) \\ &= W_3(k) - W_3(k)/Z + W_3(k)/Z + W_3(k)/Z = W_3(k) + W_3(k)/Z \end{aligned} \tag{4.22}$$

Complementary wealth W_3^* is given below:

$$\begin{aligned} W_3^*(k+1) &= W_3^*(k) + W_1(k)/(4Z) + W_2(k)/(2Z) + W_3(k)/Z \\ &= W_3^*(k) + W_3(k)/Z + W_3(k)/Z + W_3(k)/Z \\ &= W_3^*(k) + W_3^*(k)/Z + W_3^*(k)/Z + W_3^*(k)/Z = (Z+3)W_3^*(k)/Z. \end{aligned} \tag{4.23}$$

We continue in the same way and using induction, with $A_Z = 1/Z$, we conclude the following general result for the i-th Wi group with normalization between k and k+1 cycles:

Result 4.2. The recursive wealth of the i-th group in the normalized wealth balancing model with a "mean halved" (MH) is:

$$W_i(k+1) = [1 + (i-2)A_Z]W_i(k) \tag{4.24}$$

$$W_i^*(k+1) = (1 + iA_Z)W_i^*(k) \tag{4.25}$$

(a special case of Result 4.1), plus:

$$W_L^*(k) = W_{L+1}(k+1), i = 1, 2, \ldots, L. \tag{4.26}$$

The algorithm terminates at i = L, with the final step giving $W_L(k)$ do $W_L^*(k) = W_L(k+1)$. At this point the algorithm stops. Before continuing for k+1, wealth normalization is done using the corrected values at k. Equations 4.2 - 4.6 hold. Figure 4.3 summarizes the process. Thick arrows indicate giving and block arrows indicate equal halving. The upper shaded block (richest half) only gives and the lower shaded block (poorest group) only receives. Between iteration steps, normalization (remodeling) is performed so that the next iteration of balancing starts with normalized Figure 3, with new iterated values of $w_i, n_i and W_i$. Normalization can be done analytically or numerically for the integration of N(w) or W(n), and by dividing the area under any function in

groups, with $W_T = W_T(1/2 + 1/4 + 1/ 8 + \ldots + 1/(2L) + 1/(2L))$. Finally, normalization of W_T i N_T to 1 or 100 follows.

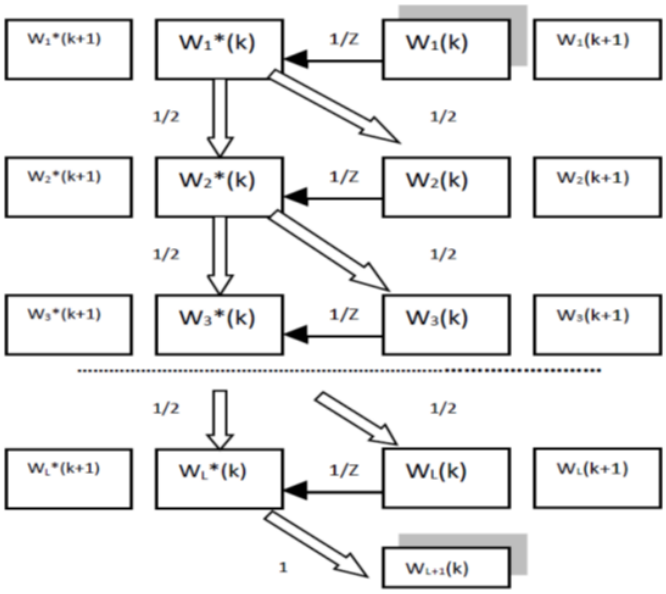


FIGURE 4.3. *MH wealth model balancing process*

4.3.3. ***Verification of total wealth.*** It should be checked whether the total wealth of W_T is preserved after the balancing algorithm (algorithm integrity check). The sum of all corrected (i =1,2,...,L+1) wealth values should be equal to the initial W_T so that the balancing algorithm is consistent. For the normalized case, the check confirms the consistency:

$$
\begin{aligned}
W_T &= \sum_i W_i(k+1) + W_{L+1}(k+1) \\
&= \sum_i \{[Z+(i-2)]/Z\} W_i(k) + W_L^*(k) \\
&= (1/Z)[(Z-1)W_1(k) + (Z)W_2(k) + (Z+1)W_3(k) + \ldots \\
&\quad + (Z+L-2)W_L(k) + (Z+L)W_L(k) \\
&= (1/Z)[(Z-1)W_1(k) + (Z/2)W_1(k) + (Z+1)W_1(k)/4 + \ldots \\
&\quad + (Z+L-2)W_1(k)/2L - 1 + (Z+L)W_1(k)/2^{L-1}] \\
&= (1/Z)W_1(k)[(Z-1) + (Z/2) + (Z+1)/4 + \ldots \\
&\quad + (Z+L-2)/2^{L-1} + (Z+L)/2^{L-1}] \\
&= W_1(k)[(Z/Z)(1 + 1/2 + 1/4 + \ldots + 1/2L + 1/2L) \\
&\quad + (1/Z)(-1 + 1/4 + 2/8 + 3/16 + 4/32 + \ldots + (L-2)/2^{L-1} + L/2^{L-1})] \\
&= 2W_1(k) + (1/Z)(0) = W_T. \qquad (4.27)
\end{aligned}
$$

4.4. **Average wealth received**

As the Giving Index causes wealth to be balanced in different groups, these groups give and receive additional wealth according to Outcomes 4.1 and 4.2. We need to calculate what the averages are for each individual in each group so that we can recalculate and renormalize group richness as well as the number of group members. From the previous results we get the following: **Result 4.3** The total received net wealth (received minus given) for the group $W_i(k)$, with Ni members is:

(1) General non-normalized case:

$$\Delta W_i^T(k) = A_Z[\sum B_{im}W_m(k) - W_i(k)] \tag{4.28}$$

(2) Normalized case:

$$\begin{aligned}\Delta W_i^T(k) &= A_Z[\sum_m B_{im}W_m(k) - W_i(k)] \\ &= A_Z[(i\text{–}1)W_i(k) - W_i(k)] = A_Z(i\text{–}2)W_i(k)\end{aligned} \tag{4.29}$$

(3) The average net wealth per wealth group member and the total average wealth per N_i members are:

$$\Delta W_i^A(k) = \Delta W_i^T(k)/N_i, W_i^A(k) = W_i(k+1)/N_i,\ i = 1,2,\ldots,L \ \text{ i } \ m = 1,2,\ldots,i-1 \tag{4.30}$$

For the last poorest group $W_L + 1$ we have a general and normalized case for average wealth:

$$\begin{aligned}\Delta W_{L+1}^T(k) &= A_Z[\sum B_{(L+1)m}W_m(k)] \\ \Delta W_{L+1}^T(k) &= A_Z[(L+1\text{–}1)W_{L+1}(k) = LA_ZW_L + 1(k)\end{aligned} \tag{4.31}$$

4.5. **Normalization**

To understand the normalization process, we assume that in some period k all groups W_i(k) are normalized and a new (k+1) giving cycle begins. All newly calculated W_i(k+1) are denormalized as a result of their giving and receiving.

4.5.1. ***Denormalization of wealth.*** Then, to calculate some group's wealth values for the (k+1)th period, we recall that W_1(k) (normalized over period k) gives W_1(k)/Z over N_1^*,which means giving all groups starting from W_2 to W_{L+1} uniformly. Due to this assignment, W_1 is denormalized and is equal to $W_1(k+1)$. Similarly, $W_2(k)$ receives its part W_1(k)/(2Z) = W_T/(4Z) from the normalized W_1(k) and gives its part W_2(k)/Z = W_T/(4Z) to W_3 through W_{L+1}. So the two amounts are equal, and W_2(k+1) remains normalized. The other half of W_1(k)/(2Z) of the total W_1 given W_1(k)/Z distributes through W_3 through W_{L+1}. Other groups for i > 2 receive more than they give and are thus denormalized. Group W_2 is a special group, because it gives the same as it receives, and on one side is W1, which reduces its wealth due to giving, and groups W_3 - W_{L+1} increase its wealth due to overall giving and receiving. Once the process is complete for all i = 1,2,..., L+1, we need to go back to W1(k+1) and see how to normalize it again. This is important to understand because of the elegant and simple properties of

normalized groups. The term W_1(k)/Z (not just W_1(k)/(2Z)) must be borrowed from the neighboring group W_2(k+1) to compensate for giving W_1(k). This compensation comes from the members in group W2 that need to be moved in the renormalization from the still normalized W_2(k+1) to the denormalized W_1(k+1). But, due to member wealth borrowing from W_2 to re-normalize W_1, W_2 is also denormalized. This process moves through the group hierarchy.

4.5.2. ***Re-normalization of wealth.*** We now analyze the renormalization of each W_i group. The other side of renormalization, that of the number of members of each group, is dealt with in another paper. From the above we can write for W_1(k+1) the renormalizing term:

$$\Delta W_1^C(k) = W_1(k)/Z = A_Z W_1(k) = W_T/(2Z). \tag{4.32}$$

This amount is borrowed from $W_2(k+1)$ and moved to W1(k+1) to normalize it. This obviously denormalizes $W_2(k+1)$, so the same amount must be borrowed from $W_3(k+1)$ to renormalize $W_2(k+1)$. So we have to consider what is net received in W_3. From Section 4.4 and Result 2.1, for W_3 we can write:

$$\begin{aligned}\Delta W_3^T(k) &= A_Z[\sum_m B_{3m}W_m(k)–W_3(k)] \\ &= A_Z[(3-1)W_3(k) - W_3(k)] = A_Z(3–2)W_3(k).\end{aligned} \tag{4.33}$$

From the above amount one needs to subtract $\Delta W_1^C(k)$ to give to $W_2(k+1)$, i.e:

$$\begin{aligned}\Delta W_3^T(k) - \Delta W_1^C(k) &= A_Z(3–2)W_3(k) - A_Z W_1(k) \\ &= A_Z[W_3(k)–4W_3(k)] = -3A_Z W_3(k)\end{aligned} \tag{4.34}$$

where we used $W_1(k) = 2W_2(k) = 4W_3(k)$ for normalized wealth groups. Hence, we can further write modified $W_3^M(k+1)$ as in:

$$W_3^M(k+1) = W_3(k) + \Delta W_3^T(k) - \Delta W_1^C(k) = W_3(k) - 3A_Z W_3(k) \tag{4.35}$$

where $W_3(k)$ is normalized. The next step is to normalize $W_3^M(k+1)$ taking $3A_Z W_3(k)$ from $W_4(t+1)$. The above amount needs to reduced by $3A_Z W_3(k)$ to be given to $W_3^M(k+1)$, i.e.:

$$\begin{aligned}\Delta W_4^T(k) &= A_Z[\sum_m B_{4m}W_m(k)–W_4(k)] \\ &= A_Z[(4–1)W_4(k)–W_4(k)] = A_Z(4–2)W_4(k)\end{aligned} \tag{4.36}$$

where we used $W_3(k) = 2W_4(k)$ for normalized groups. Further we write modified $W_4^M(k+1)$ as:

$$\begin{aligned}\Delta W_4^T(k) - 3A_Z W_3(k) &= A_Z(4–2)W_4(k) - 3A_Z W_3(k) \\ &= A_Z[2W_4(k)–6W_4(k)] = -4A_Z W_3(k).\end{aligned} \tag{4.37}$$

Hence, we can further write $W_4^M(k+1)$ as:

$$W_4^M(k+1) = W_4(k) + \Delta W_4^T(k) - 3A_Z W_4(k) = W_4(k) - 4A_Z W_4(k) \tag{4.38}$$

where $W_4(k)$ is normalized. The next step is the normalization of $W_4^M(k+1)$ borrowing $4A_Z W_4(k)$ from $W_5(t+1)$, etc. General results then are as follows:

Result 4.4. Renormalization of $W_i(k+1)$ requires adding to it the amount:

$$iA_Z W_i(k) \tag{4.39}$$

from $Wi+1(k+1)$, ending with i = L+1, until $W_{L+1}(k+1)$ which produces $LA_Z W_L(k)$ for $W_L(k+1)$ to be normalized again. This can be easily verified from Result 4.2 of Section 4.3.2. After re-normalization, all $W_i(k+1)$ are normalized again, and the same procedure is repeated for (k+2), (k+3), etc. Due to limited space, we do not deal with the renormalization of the number of group members. This is given in our other works.

4.6. **Wealth balancing tables**

Using the Results of this section we can generate balancing tables, each for a specific index of giving. This is illustrated by an example.

4.6.1. ***When is balance achieved?*** As the balancing algorithm progresses, k = 1, 2, 3, ..., there is a point where the rich will become poor and the poor will become rich, and this is not the intention of the giving process (despite the fact that the poor might like it). Instead, the idea is to have a balanced distribution across all groups. Therefore, we must determine when to stop the recursion, which is when the ideal Lorenz line is reached, and the giving process should be stopped. In theory this is achieved when all segments in Figure 3.1 are at the same angle and K_i in 3.19 has the same value for all W_i. This corresponds to the average total W_{TA} wealth among N_T members as well as the average of each W_i:

$$W_{TA} = W_T/N_T = W_i(k+1)/N_i \tag{4.40}$$

Given the average wealth of W_{TA} we can calculate the giving period to stop the algorithm for each group W_i as:

$$W_i(k_B) = (W_T/N_T)N_i = W_{TA}N_i \tag{4.41}$$

where k_B is the wealth balancing period. This assumes that all N_i are known. If not, what are known as stopping criteria can be used. See examples in Section. 4.7. Another method to stop is:

$$|W_i(k+1) - W_i(k)|/N_i = |\Delta W_i(k+1)|/N_i = |W_i^A(k+1)| < \epsilon_W \tag{4.42}$$

for a small ϵ_W . Another option is to test the sums of the wealth against some given thresholds:

$$\sum W_i(k+1) < \delta W. \tag{4.43}$$

The range of values ϵ_W and δ_W is related to average values of wealth and can be determined by some numerical and statistical analysis. One of the potential problems with the above methods is that we may not achieve the true average wealth. Another simple test can be used to avoid this, as we illustrate in the examples in this section:

$$|W_i^A(k_{min})| > W_A > |W_i^A(k_{min}+1)|; k_{min} \le k_B \le k_{min}+1. \tag{4.44}$$

The algorithm can be stopped as soon as one or more of the above conditions are met. In the case of 4.44, we perform an additional calculation step to make sure that the algorithm has reached only one step above the threshold. In Section 4.7 we illustrate the stopping method 4.44 using several concrete examples.

4.6.2. ***Interpretation of the balance table.*** The numerical examples in this section demonstrate a number of important features of our wealth balancing algorithm. Tables carry universality as a theoretical and numerical tool. Here are the important points:

(1) All calculations are normalized to $W_T = 100$ and $N_T = 100$. These tables are universal normalized wealth balancing tables and apply to ANY N_i, given $\sum N_i = N_T$ and ANY W_i given $\sum W_i = W_T$, both normalized to 100, hence, the results can be interpreted as percentages.
(2) Each table for a given giving index Z starts with the same set of numbers (ideally 50%, 25%, 12.5%, 6.25%, etc., but the algorithm is very robust and it works for other numbers as well). These numbers indicate the percentages of groups from which balancing starts. At this time we do not state anything about how many people own 50% or 25% etc.
(3) The examples illustrate L = 6. Recall that the (L+1) group is the group that does not give (the poorest), and the first group is the group that only gives (the richest).
(4) When balancing is initiated, the period k begins to increase and each new row in the balancing table represents a new and changed value of group wealth, due to giving and receiving, for the same number of people who owned the initial percentages.
(5) Balancing should stop when average wealth is reached. For example, if we use an example "1% ($N_1 = 1$) people own 50% of the wealth", when do we end the balancing? From (35) we have $W_1(k_B) = (W_T/N_T)N_1 = (100/100)1 = 1$ hence we look for $W_1(k_B)$ which is either 1, for $k = k_B$ (very unlikely) or two consecutive entries, one of which is slightly greater than 1, for kmin, and the other slightly smaller than 1 for $k = k_min + 1$, per (63). Plus $W1 = 1$ indicates the percentage of the total wealth after the balancing, ie. "1% people now own 1% of the total wealth". The period kB indicates the number of giving cycles (years) to reach the wealth balance. If $N1 = 5$, the algorithm stops at $W_1(k_B) = 5$, and similarly for other $N1 = 5$.
(6) We can even start from some middle point in the balancing table and check how many cycles it would take to reach a certain average wealth. In other words, when specifying an individual N_i, we can interpret ANY entry in the table as $W_i(k_B) = W_A N_i$=certain percentage of $W_T - a$. Therefore, normalized wealth tables are universal and can be applied to any community wealth normalized to 100, with a set of L groups, with arbitrary N_i.
(7) Ideally, we would know all N_i. If not, we can use the ones we know. Most likely the value for N1 will be accessible because it indicates some key information, such as "1% owns 50%". Given a specific community, these numbers can be determined by some statistical methods.

4.7. **Numerical example**

We now illustrate the "mean halved" MH methodology with an example for two indices, Z = 20 and 40, i.e. members of the community give one 20th or one 40th of their wealth to the community. We assume $W_T = 100$, $N_T = 100$, L=5 (6 groups), with $N_1 = 1$ (1% owns 50%) and that is the current situation of the world's wealth. Tables 1 and 2 show the normalized results. In both cases, the average wealth is $W_T/N_T =$

100/100 = 1 per community member, and that should hold for Ni in balance, when all Ni are defined. Balanced group W1 requires an average $W_A = W_1(k_B)/N_1 = 1$, i.e. $W_1(k_B) = W_A N_1 = 1$. Grey areas in the tables indicate k_{min} and $k_{min}+1$ per (63) when the algorithm stops. Table 1 shows that it is required between $k_{min} = 77$ and $k_{min}+1 = 78$ years. Table 2 for Z = 40 shows that it takes about two times longer for the balance compared to Z = 20, i.e. $k_{min}+1 = 155$ and $k_{min}+1 = 156$ years. Continuing with the example, we assign the remaining values N2 through N6 using world inequality data, per references at the back of he paper. The details follow:

$$N_1 = 1, N_2 = 2.5, N_3 = 4.5, N_4 = 6.5, N_5 = 6.5, N_6 = 79 \tag{4.45}$$

i	1	2	3	4	5	6	
k							W_T
1	50	25	12.5	6.25	3.125	3.125	100
2	47.5	25	13.125	6.875	3.59375	3.90625	100
3	45.125	24.9375	13.6875	7.46875	4.0546875	4.7265625	100
4	42.86875	24.81875	14.190625	8.03125	4.506640625	5.583984375	100
5	40.7253125	24.64953125	14.63742188	8.562617188	4.948554688	6.4765625	100
6	38.68904688	24.4351875	15.03085547	9.063074219	5.379486328	7.402349609	100
.....	.....	.....	.....	.....	.....	.....	.....
75	1.123354413	2.749262116	4.569656386	6.26005174	7.589933109	77.70774224	100
76	1.067186692	2.63988287	4.42394705	6.102677305	7.444751823	78.32155426	100
77	1.013827358	2.534568394	4.282086603	5.947810568	7.300214729	78.92149235	100
78	0.96313599	2.433185658	4.144019325	5.795490731	7.156434602	79.50773369	100
79	0.91497919	2.335604775	4.0096872	5.645751098	7.013517592	80.08046014	100
80	0.869230231	2.241699016	3.879030199	5.498619403	6.87156342	80.63985773	100

TABLE 1. $W(k+1)$ *za* $W_T = 100, N_T = 100, N_1 = 1, Z = 20, L = 5$

i	1	2	3	4	5	6	
k							W_T
1	50	25	12.5	6.25	3.125	3.125	100
2	48.75	25	12.8125	6.5625	3.359375	3.515625	100
3	47.53125	24.984375	13.109375	6.8671875	3.591796875	3.916015625	100
4	46.34296875	24.95390625	13.39101563	7.1640625	3.822119141	4.325927734	100
5	45.18439453	24.9093457	13.65780762	7.453132324	4.050202637	4.745117188	100
6	44.05478467	24.85141699	13.91013171	7.734411255	4.275915344	5.173340027	100
.....	.....	.....	.....	.....	.....	.....	.....
153	1.065789876	2.609818799	4.353727118	5.995139921	7.313418592	78.66210569	100
154	1.039145129	2.557895702	4.284167862	5.919324972	7.242554151	78.95691218	100
155	1.013166501	2.506937624	4.215532019	5.844128123	7.171874997	79.24836074	100
156	0.987837338	2.456928764	4.147812729	5.769553576	7.101394052	79.53647354	100
157	0.963141405	2.407853512	4.081003004	5.695605192	7.031123848	79.82127304	100
158	0.93906287	2.359696442	4.015095731	5.622286501	6.961076536	80.10278192	100

TABLE 2. $W(k+1)$ *za* $W_T = 100, N_T = 100, N_1 = 1, Z = 40, L = 5$

Balance Table 3 was created based on these data.

i	1	2	3	4	5	6	
N_i	N_1	N_2	N_3	N_4	N_5	N_6	N_T
	1	2.5	4.5	6.5	6.5	79	100
k							W_T
155	1.013166501	2.506937624	4.215532019	5.844128123	7.171874997	79.24836074	100
156	0.987837338	2.456928764	4.147812729	5.769553576	7.101394052	79.53647354	100

TABLE 3. $W(k+1)$ *za* $W_T = 100, N_T = 100, Z = 40, L = 5$

As an illustration of the flexibility of the MH model, in the following example we choose L = 6 (7 groups) and disassemble W_6 from the above example into two groups,

”poor” and ”very poor”. The corresponding N numbers are:

$$N_1 = 1, N_2 = 2.5, N_3 = 4.5, N_4 = 6.5, N_5 = 6.5, N_6 = 71, N_7 = 18. \tag{4.46}$$

Figure 5.1 (Lorenz curve) shows all the details for the 7 groups according to the model of this paper. Comments follow.

(1) The MH balancing algorithm transforms the original Lorenz curve in Figure 5.1 into an ideal Lorenz line. Mathematically, the MH balancing algorithm changes a non-linear curve to a linear one using recursion in a specified number of steps. Depending on the value of Z, the balancing lasts from 6-7 years (Z=2, 50% giving), 155-156 years (Z=40, 2.5% giving) to 390-391 years (Z=100, 1% benefits).

(2) Figure 5.1 clearly identifies the corresponding middle class W_3, W_4 and W_5, between very rich (W_1), rich (W_2), and poor (W_6), and very poor (W_7), (¡ \$1 per day [16]), no giving group. W_4 can be considered as a middle of the middle class. As the number of wealth groups is increased (larger L) the results are more reliable. In this paper we show results for L =6 and L= 7. Figure 5.1 gives a precise world Lorenz curve and it also shows our MH model with W_i groups clearly indicated.

(3) Figure 5.1 shows middle class in the middle of the curve. Around 17,5% uf the population is in the middle class and they own 22% of W_4. “Middle“ middle class W_4 is very close to the ideal line w = pn + 100. The slope p for W_4 can be calculated from Figure 5.1 as p = (12,5-6,25)/(7,6-13,4) = - 1,0776 which is very close to the ideal p = -1 Lorenz line. On the other hand the whole middle class (W_3, W_4 and W_5) has p = (25-3.125)/(3.5-20.4)=-1.29, relatively close to p=-1. The importance of middle class is well known in economics. Between 1971 and 2021 the USA middle class was reduced from 61% to 50%. World-wide at this moment there is around 17.1% middle class of the total population [16]. We conclude that our MH wealth model is a reliable mathematical model of the current Lorenz curve.

5. CONCLUSION

In this paper, we present a new simple mathematical model for the Lorenz curve, which represents the distribution of the wealth of a given community. Our “mean halved” (MH) robust model of wealth distribution is easily adapted to a variety of practical situations both in the field of wealth and in some other applications, such as education, health, ecology and climate modeling, wherever there is an uneven distribution of some resource. The model is defined with the idea of defining a wealth redistribution algorithm. The algorithm actually corrects the Lorenz curve to an ideal line (and improves the Gini index). The investment index is also mentioned in the paper, and the results are in a future paper. The idea of the whole project is to show that the combination of giving and investing is better than just investing. The paper also shows the analysis of the convergence of the allocation algorithms. The algorithm assumes a fixed total amount of wealth between giving periods, as well as a fixed giving percentage. The first condition can be removed by the wealth normalization process between giving cycles, and the second giving condition can also be adjusted on the variable.

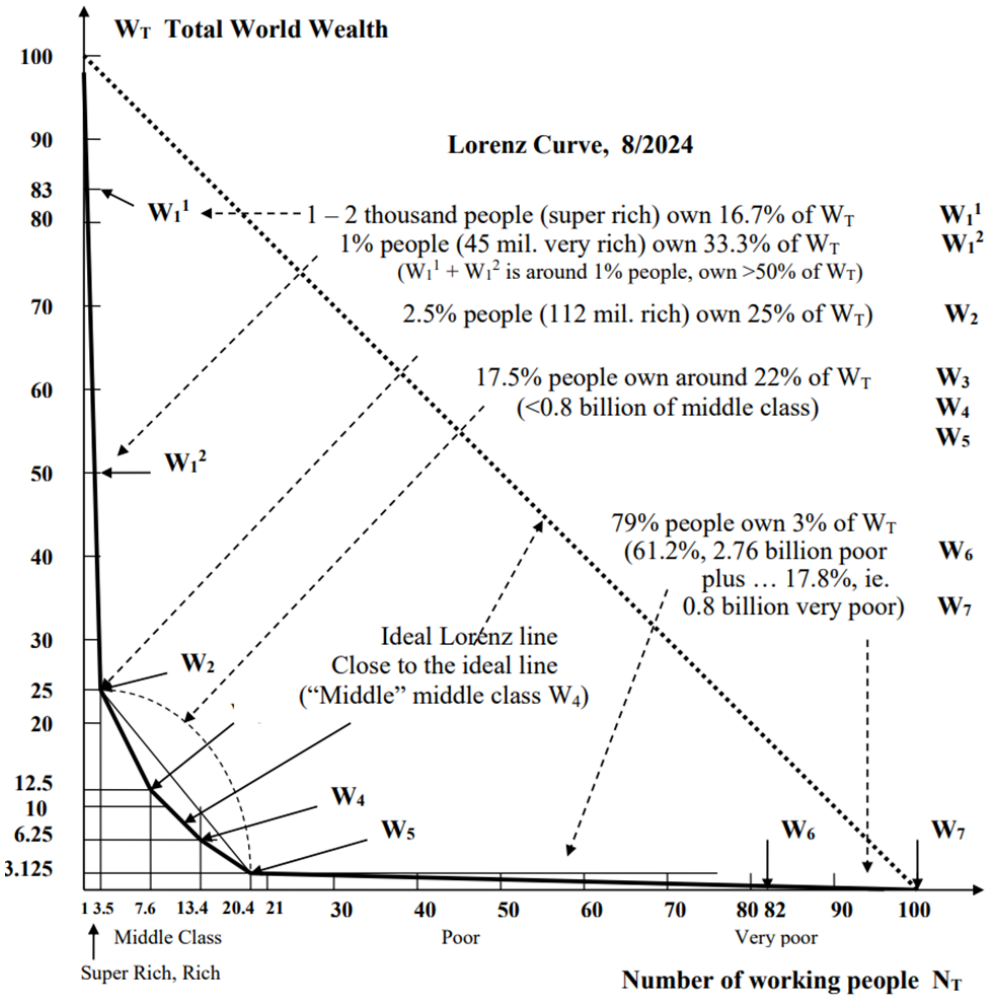


FIGURE 5.1. *World Lorenz curve per our MH model*

REFERENCES

[1] Mauguen, Audrey, Begg Colin B., Epidemiology, *Using the Lorenz Curve to Characterize Risk Predictiveness and Etiologic Heterogeneity*, July 2016, Volume 27, Issue 4, pp. 531–537

[2] Gastwirth, Joseph and Modarres, Reza and Bura, Efstathia, *The use of the Lorenz curve, Gini index and related measures of relative inequality and uniformity in securities law*, 2005, METRON, International Journal of Statistics, No. 3, pp. 451-469

[3] Chotikapanich, D., *Modeling Income Distributions and Lorenz Curves*, 2016, Springer.

[4] World bank 2024, *Gini Indeks po zemljama*, https://data.worldbank.org/indicator/SI.POV.GINI

[5] CIA, *Lorenz Curve, Gini Coefficient*, https://www.cia.gov/library/publications/the-world-factbook/

[6] Rolf Aaberge, *Characterization of Lorenz Curves and income distributions*,Springer Verlag, Social Choice Welfare, Vol. 17, pp. 639-653, 2000.

[7] Michael T. Catalano, Tanya L. Leise, Thomas J. Pfaf, Numeracy, *Measuring Resource Inequality: The Gini Coefficient*, Advancing Education in Quantitative Literacy, Volume 2, Issue 2, June 2009

[8] Joseph Gastwirth, Reza Modarres, Efstathia Bura, *The use of the Lorenz curve, Gini index and related measures of relative inequality and uniformity in securities law*, METRON, International Journal of Statistics, 2005, Vol. LXIII, No. 3, pp. 451-469
[9] Uwe E. Reinhardt, *The construct of Lorenz Curves and of the Gini-coeficient to depict degrees in inequality in health care.*
[10] A. B. Atkinson, *On the Measurement of Inequality*, Journal of Economic Theory 2, pp. 244-263, 1970
[11] Arjun Kharpal, http://cnbc.com/2013/10/11/global-wealth-hith-241-trillion-but-distribution-skewed.html, Thomson Reuters, CNBC, 10/13/2016.
[12] Wealth distribution by Country, 2024. https://en.wikipedia.org/wiki/Wealth_distribution_by_country
[13] W. Lu, Bloomberg, *What If the Richest Person in Every Country Gave All Their Money to the Poor?*, June 10, 2016.
[14] Annual Report 2022 https://www.credit- suisse.com/media/ib/docs/ investment-banking/financial-regulatory/international/csi-annual-report-2022.pdf, Credit Suisse Group AG, 2022.
[15] Annual Report 2023, https://www.credit-suisse.com/media/assets/corporate/docs/-/about-us/investor-relations/financial-disclosures/results/csg-financialreport-1q23.pdf, 1st Quarter, 2023.
[16] visual capitalist, https://www.visualcapitalist.com/the-richest-people-in-the-world-in-2024

(Received: May 10, 2024)
(Revised: November 11, 2024)

Migdat Hodžić
Univerzitet Dzemal Bijedić
Analytics and FIT
88000 Mostar
Bosnia and Herzegovina
e-mail: *migdat.hodzic@artianalytics.com*

PROCEEDINGS OF THE CONFERENCE ON MARCH 14 – INTERNATIONAL DAY OF MATHEMATICS
Special Editions ANUBiH, Book CCXVI, OPMN Book 30
Sarajevo 2024, **p. 137–147** DOI: 10.5644/PI2024.216.14

STOCHASTIC CALCULATION OF NET LIFE INSURANCE PREMIUMS

SUADA ALIĆ – MEŠANOVIĆ

Dedicated to the 75th birthday of our dear Professor Mirjana Vuković

ABSTRACT. The features of life insurance are: death risk coverage (the risk of death is covered if it occurs during the contracted insurance term), long-term (life insurance contracts are concluded for several years), fixed premium (the amount of the premium is the same for the entire insurance period), savings (in many forms of this insurance, savings are also included). Life insurance contracts, among other things, differ according to the method of premium payment, namely [9], [10]:

1) insurance with premium payment at once
2) insurance with premium payment in installments - monthly, quarterly, semi-annually, annually.

Using mathematical methods based on probability and statistics, financial mathematics, stochastic models, risk theory and credibility theory, actuarial mathematics determines insurance prices, required reserves, self-retention amounts and other elements of business policy [1], [4], [5].

Therefore, regardless of the life insurance model, the principle of equivalence must be realized throughout the obligation period [7].

1. INTRODUCTION

Before moving on to stochastic approaches to calculating net premiums in life insurance, let's recall certain definitions from probability theory that we will use or rely on in our paper [8].

1.1. **Random variable and discrete random variable**

A random variable X is a function that assigns real numbers to the outcomes of an experiment (elements of the set Ω). The set of all values (X) that the random variable X can take is called the image of the random variable (elements are usually denoted by x_i). We are often interested in the probability p_i that the random variable X is realized by values from some set $A \subseteq \mathbb{R}$.

A discrete random variable X can have a finite $\{x_1, x_2, \ldots, x_n\}$ or a countable $\{x_1, x_2, \ldots, x_n, \ldots\}$ set of values (X). If the probabilities $P(X = x_i)$ are known for all possible values $x_i \in \mathcal{R}(X)$, we say that the distribution of the discrete random variable X is known.

2020 *Mathematics Subject Classification.* 91G05, 91G30, 62P05.

Key words and phrases. life insurance, premium, random variable, expectation, interest rate, stochastic model, actuarial mathematics, financial mathematics.

Discrete random variables are fully determined:

- by their own picture
- and by the probability function (ie the probability $p_i = P(X = x_i), x_i \in \mathcal{R}(X)$)

1.2. Continuous random variable and distribution function

A continuous random variable is characterized by the image (X) which is not a discrete set. (a subset S of the topological space X in which every point $x \in S$ has a neighborhood in X to which no other point from S belongs is a discrete set). For example, (X) can be some interval or even the whole set of real numbers.

We define **the distribution of the continuous random variable** X by a non-negative function f, **the density function**, for which the area between the graph of this function and the x axis is equal to 1. The probability that the continuous random variable X is realized by values from some set $A \subseteq \mathbb{R}$ is equal to the area under the graph of the density function f over the set A, as shown in the following Figure 1.:

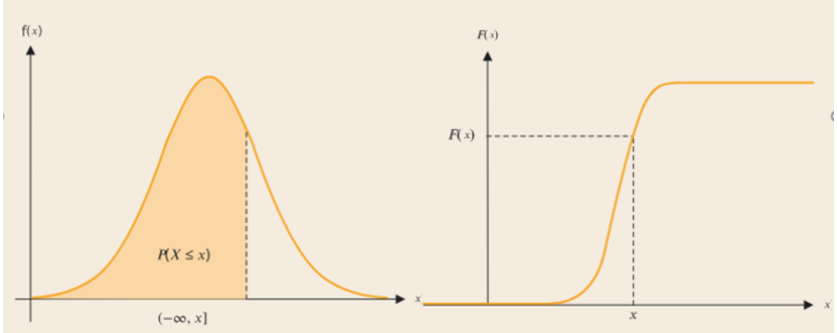


Figure 1.

Continuous random variables, which have as their image an uncountable set in the set of real numbers $\mathbb{R}$, are called **continuous random variables**.

A random variable$X : \Omega \to \mathbb{R}$, is continuous if there is a (measurable) function $f : \mathbb{R} \to \mathbb{R}$ for which:

(i) $f(x) \geq 0, x \in \mathbb{R}$,
(ii) $\int_{-\infty}^{\infty} f(x)dx = 1$,
(iii) $P(X \leq a) = \int_{-\infty}^{a} f(x)dx,\ a \in \mathbb{R}$.

The function $f(x)$ is called zovemo **the density function** of X.

It follows from (iii) that for all $a,\ b \in \mathbb{R},\ a \leq b,\ P(a < X \leq b) = \int_a^b f(x)dx$.

The same applies to $P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b)$.

The cumulative distribution function of X is the function $F : \mathbb{R} \to \mathbb{R}$, given by $F(x) = P(X \leq x)$.

The following applies:

(i) $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$,
(ii) $F(x)$ is nondecreasing
(iii) $F(x) = \int_{-\infty}^{x} f(t)dt,\ x \in \mathbb{R}$
(iv) $P(a < X \leq b) = F(b) - F(a)$,

(v) if $f(x)$ is piecewise continuous, then $F'(x) = f(x)$ except perhaps at points of discontinuity of $f(x)$.

1.3. **Deterministic characteristics of continuous random variables**

Let introduce the deterministic characteristics of continuous random variables, namely:

1. expectation
2. variance
3. standard deviation

For a continuous random variable X and its density function $f(x)$, we define **the expectation** of X (if the lower integral exists) with

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx$$

The following applies:

(i) $\mathbb{E}(\lambda X) = \lambda \mathbb{E}(X),\ \lambda \in \mathbb{R}$,
(ii) $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$.

The variance of X (if the lower integral exists) is defined by

$$Var(X) = \mathbb{E}\left[(X - \mathbb{E}(X))^2\right] = \int_{-\infty}^{\infty} (X - \mathbb{E}(X))^2 f(x)dx.$$

We can easily derive:

$$Var(X) = \mathbb{E}\left(X^2\right) - \mathbb{E}(X)^2 = \int_{-\infty}^{\infty} x^2 f(x)dx - \mathbb{E}\left(X^2\right).$$

The following applies:

(i) $Var(\lambda X) = \lambda^2 Var(X),\ \lambda \in \mathbb{R}$
(ii) $Var(X + \lambda) = Var(X)$.

The standard deviation of X (if X has variance) is defined by

$$\sigma(X) = \sqrt{Var(X)}.$$

Let's note the following:

If X is a continuous random variable with image (X) and density function $f(x)$ and $g\colon \mathbb{R} \to \mathbb{R}$, some function, then $g(x)$ is a random variable defined on the same probability space as X, has image $g(\mathcal{R}(X))$ and holds (if the lower integrals exist)

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx,$$

$$Var\,(g(X)) = \int_{-\infty}^{\infty} (g(x) - \mathbb{E}\,(g(X)))^2 f(x)dx = \int_{-\infty}^{\infty} g^2(x)f(x)dx - \mathbb{E}\,(g(X))^2.$$

1.4. **Life insurance**

Life insurance is a long-term business and carries with it long-term risks, but much of modern actuarial risk management is focused on short-term modeling approaches [2]. A life insurance model in which the insurance premium is paid once is called single

premium insurance, and a model in which the sum insured is paid multiple times at equal time intervals and in the same amount is called multiple premium insurance premium payments. Premiums can be paid multiple times in equal time intervals with different amounts, so it is insurance with multiple variable premium payments. Variability of premiums should be based on arithmetic or geometric progression. According to the duration of premium payments in relation to the duration of life, the premium is divided into lifetime and temporary. The premium is lifetime if the insured person pays it for the rest of their life, while the insured person pays the temporary premium only for a period specified in the contract.

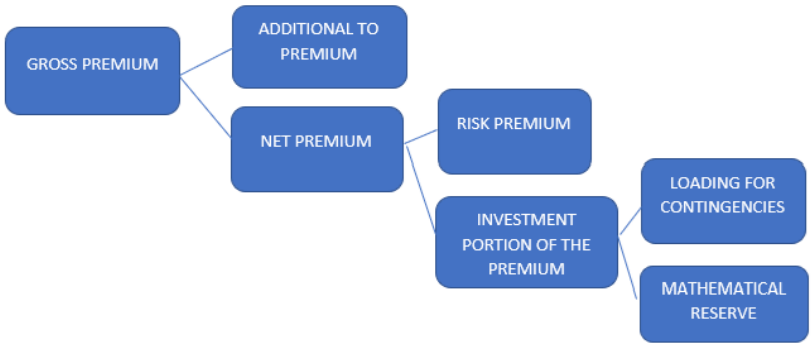


Figure 2 [11]

According to the number of payments of the insured sum, insurance is divided into capital insurance and annuity insurance. If the insured sum is paid to the insured or the beneficiary once, it is capital insurance, and if the payment occurs in several amounts and at equal time intervals, it is annuity insurance. In accordance with the differences in individual insurance models, different combinations of insurance payments and payments are created.

The success and safety of the life insurer's business depends primarily on the calculation or mathematical basis, namely the mortality tables and the interest rate. They are used to determine the net premiums, from which the funds sufficient to cover the obligations to the insured (that is, the beneficiaries) are formed.

2. STOCHASTIC APPROACH TO THE CALCULATION OF NET PREMIUM IN LIFE INSURANCE

2.1. **Preliminaries**

Mathematical laws of life insurance are based on the law of large numbers, calculus of probability and statistics, stochastic processes and credibility theory and hedging strategy. The part of mathematics used to solve and explain insurance calculation problems is actuarial mathematics. Actuarial mathematics is based on the principle, with respect for the age of the persons who enter the life insurance portfolio, the legality of stochastic processes, with the application of the time value of money.

Actuarial mathematics solves the problems of expectation of realization of the in-

sured event using the strong law of large numbers. This law is made special by the large number of observed cases, and the greater the number of observations, the more accurate the conclusion - data, and the smaller the deviations. If an event is observed individually, it is a case, and in a large number of observations, it is a law. There are several theorems about this law, but each asserts that empirical average values converge to the expected value. These theorems are often called laws of averages.

Insurance companies are paying more attention to new phenomena that are happening and are also reflected in insurance (increased mortality of (old) persons). Variability in mortality rates within different demographic groups and/or populations within plans results in the need to review assumptions and better adapt them to specific groups. This increases the interest in new theories such as, for example, the classical theory of credibility, but also newer theories of c-credibility, such as, for example, means for adapting standard mortality tables to plans, or the portfolio included in those plans.

In addition to forecasting the occurrence of an insured event, it is also important to know the probability of the occurrence of certain insured events, so in this segment, actuarial mathematics relies on probability calculations, which are used to create mortality tables and commutative numbers (which is not the subject of this paper). Determining the probability of an adverse event in life insurance is the basis for determining the insurance premium, which follows below. Insurers with a high claims ratio usually charge high premiums. Other "competitors" set a competitive premium or accept a fixed premium to stay "in the game", otherwise they will operate below the "optimal point". Such dynamic systems, which develop in time, can also be described by non-linear Lotka-Volterra differential equations, given that it is a matter of the interaction of two types. [12] can also be applied to insurance. Regarding the type of data that affects the price of the service (insurance), the COVID-19 pandemic had an important impact on the change in the financial performance of insurance companies and on the global results in correlations between the period before and after the pandemic. [13] and [14] can also be applied to insurance.

2.2. **Mathematical model**

For life cycle modeling it is essential to know how long an individual will live. For this reason, insurers use life expectancy models to be able to calculate the probability of an individual's death at a certain age.

We start life from birth. The length of life is marked with X, x is the age of the individual. The variable X is a continuous random variable. Let $F(x)$ be the distribution function of the length of life, i.e. the distribution function of X, holds

$$F(x) = P(X \leq x),\ x \geq 0.$$

Let's define the inverse function of the distribution function $F(x)$, $s(x)$ – the distribution survival function,

$$s(x) = 1 - F(x) = 1 - P(X \leq x) = P(X \succ x),\ x \geq 0.$$

The random variable X (expected life expectancy) is completely determined by the life span distribution function $F(x)$ or the distribution survival function $s(x)$.

Thus, $F(x)$ represents the probability that the newborn will live less than x years, i.e. the probability that he will not live to age x, and $s(x)$ is the probability that the newborn will live more than x years, .e. the probability that he will live to x years. The survival function is the basis in actuarial science, and in statistics it plays the role of the mortality distribution function. The probability that a newborn dies between the years x and z, $(x < z)$ is

$$P(x \le X \le z) = F(z) - F(x) = s(x) - s(z),$$

of course, on the condition that the newborn lived to be x years old, that is

$$P(x < X < z)/X \succ x) = \frac{F(z) - F(x)}{1 - F(x)} = \frac{s(x) - s(z)}{s(x)} \tag{1}$$

The label $T(x)$ is introduced for the random variable remaining life expectancy of a person aged x in the manner

$$T(x) = X - x$$

and let ${}_tp_x$ survival probability $x+t$ for a person aged x and let ${}_tq_x = P\left(T(x) \le t\right),\ t \ge 0$ the probability that a person aged x will die during the next t years, i.e. the distribution of the function $T(x)$ is

$${}_tp_x = 1 - {}_tq_x = P\left(T(x) \succ t\right),\ t \ge 0$$

the probability that a person aged x will live to the age of $x+t$, i.e. the survival function for a person aged x.

We also introduce simpler labels

q_x - probability of death of a person aged x during the next year,

p_x - the probability that a person aged x will live to be $x+1$ years old,

${}_{t/u}q_x$ - the probability that a person aged x will live for the next t years and die in the next u, u, i.e. the probability of death occurring in the time interval $(x+t, x+t+u)$, i.e. it is valid,

$${}_{t/u}q_x = P(t < T(x) \le t = {}_{t+u}\,q_x - {}_tq_x = {}_tp_x - {}_{t+u}\,p_x$$

and according (1) is

$${}_tp_x = \frac{{}_{x+t}p_0}{{}_xp_0} = \frac{s(x+t)}{s(x)}$$

and

$${}_tq_x = 1 - \frac{s(x+t)}{s(x)}.$$

Let's also find the connection between conditional and unconditional probability

$${}_{t/u}q_x = \frac{s(x+t) - s(x+t+u)}{s(x)} = \frac{s(x+t)}{s(x)} \cdot \frac{s(x+t) - s(x+t+u)}{s(x+t)} = {}_tp_x \cdot {}_uq_{x+t}.$$

Also

$$P\left(T(x) = k\right) = P\left(T(x) = k+1\right) = 0, \quad \text{for } k = 0, 1, 2, 3, \ldots$$

because $T(x)$ is a continuous random variable.

Formula (1) is an equation for the conditional probability of the death of a newborn between the years x and z, with the condition that he lives to the age of x. When $x \succ z$

the probability in equation (1) still retains the property of continuity, so we can observe it as a function of x. It then describes the distribution the probability of mortality in the near future for a person who lives to age x (between time 0 and z). Analogously, the function for immediate death is obtained using the probability frequency of mortality for the case of living for the year x. Applying equation (1) fo $z = x + \triangle x$ we get

$$P\left(x < X \leq x + \triangle \frac{x}{X} \succ x\right) = \frac{F(x+\triangle x) - F(x)}{1 - F(x)} = \frac{\frac{F(x+\triangle x)-F(x)}{\triangle x} \cdot \triangle x}{1 - F(x)}$$
$$\cong \frac{F^{'}(x) \cdot \triangle x}{1 - F(x)} = \frac{f(x) \cdot \triangle x}{1 - F(x)}$$

where $f(x) = F'(x)$ is the distribution density of a continuous random variable, and the function $\frac{f(x)}{1-F(x)}$ represents the conditional probability density. For each year x it gives the value of the conditional density of the distribution of the random variable X in case of survival the same year. The function $f(x)$ is called the mortality intensity and represents the mortality rate. If we introduce the notation μ_x we have

$$\mu_x = \frac{f(x)}{1 - F(x)} = -\frac{s^{'}(x)}{s(x)} \geq 0.$$

Next, we have (with replacement)

$$\mu_y = -\frac{s^{'}(y)}{s(y)} \Longrightarrow -\mu_y dy = d(\ln s(y)) \Longrightarrow -\int_x^{x+t} \mu_y dy = \ln\left(\frac{s(x+t)}{s(x)}\right) = \ln {}_t p_x \Longrightarrow$$
$${}_t p_x = \exp\left(-\int_x^{x+t} \mu_y dy\right).$$

If we assume that $y = x + s$ we get:

$${}_t p_x = \exp\left(-\int_0^t \mu_{x+s} ds\right) \text{ i.e. } {}_t p_x = e^{-\int_0^t \mu_{x+s} ds}.$$

If the surviving years of life are compared with the value zero and the survival time with x we obtain:

$${}_n p_x = s(x) = \exp\left(-\int_0^n \mu_s ds\right),$$
$$F(x) = 1 - s(x) = 1 - \exp\left(-\int_0^x \mu_s ds\right)$$
$$F^{'(x)} = f(x) = \exp\left(-\int_0^x \mu_s ds\right) \cdot \mu_x = {}_x p_0 \cdot \mu_x$$

The following marks are introduced:

$\Phi(t)$ - distribution function of a continuous random variable remaining life time of a person x years of age,

$T(x) =$ x - X, respecting $\Phi(t) = {}_t q_x$ and

$\varphi(t)$ - density of distribution of continuous random variable $T(x)$.

$$\varphi(x) = \frac{d}{dt}{}_tq_x = \frac{d}{dt}\left(1 - \frac{s(x+t)}{s(x)}\right) = \frac{s(x+t)}{s(x)} \cdot \left(-\frac{s^{'}(x+t)}{s(x)}\right) = {}_tp_x \cdot \mu_{x+t}, \text{ for } t \geq 0$$

or

$$\varphi(x) = \frac{d}{dt}\left(1 - {}_tp_x\right) = \frac{d}{dt}{}_tp_x = {}_tp_x \cdot \mu_{x+t} \text{ otherwise.}$$

The product ${}_tp_x \cdot \mu_{x+t}$ represents the probability of mortality between the years x and $x+t$, for a person aged x years, i.e.

$$\int_0^\infty {}_tp_x \cdot \mu_{x+t}dt = 1,\ t \geq 0 \text{ holds.}$$

For a continuous random variable, the expected value is equal to a definite integral [3]:

$$\mathbb{E}\left[f(t)\right] = \int_0^\infty f(t)g(t)dt.$$

In the stochastic model, it is assumed that the interest rate is constant (the interest rate is a relative measure that describes interest, that is, the difference between the final sum of money at the end of the compounding period and the nominal value of the principal). Enter the indicator b_t in the following way:

$b_t = 1$ – if the insured risk occurs while the contract is in force

$b_t = 0$ – if the insured risk does not occur while the contract is in force

Let ν_t be the discounted sum insured (the present value of the amount at which the insurance contract was concluded, i.e. the present value of the amount that will be paid to the insurance beneficiary when the insured event occurs), for the discount factor ν o which the time t (from the beginning of the insurance) is related , until the liability of the insurer). Additional clarification: determining the present value with a known final (future) value is often called discounting.

It is valid $\nu_t = \nu^t$. The variables b_t and ν_t are dependent on time and directly determine the random variable remaining life time of a person x years - $T(x)$. Let the nominal value of the sum insured be the random variable Z, $Z = z(t) = b_t \cdot \nu_t$. The expected value of the discount value of the sum insured is $E(Z)$ one-time premium in life insurance.

2.3. **Present value of one-time net premium payments**

If it is a temporary capital insurance in the event of death, the insurer's obligation is to pay the insured sum to the insured in the event of the death of the insured within the term defined in the contract. That is, if death occurs before the expiration of the term, the insurer has the obligation to pay the insured amount, otherwise it does not. Let n be the symbol for the duration of the insurance, so the nominal value of the insured sum is equal to:

$$Z = \begin{cases} \nu_t,\ T \leq n \\ 0, T \succ n \end{cases} \quad \text{when} \quad b_t = \begin{cases} 1, t \leq n \\ 0, t \succ n \end{cases} \qquad \text{and} \qquad \nu_t = \nu^t.$$

The notation $\ddot{A}_{x\overline{n}|}$ [6] is introduced for the one-time net premium of n annual capital insurance for the death of a person aged x It is equal to the expected nominal value of the sum insured.

The function $Z = z(t)$ is the density function of the random variable $T(x)$ so:

$$\ddot{A}_{x\overline{n}|} = \mathbb{E}(Z) = \mathbb{E}(z_t) = \int_0^{\infty} z_t \cdot g(t)dt = \int_0^{n} v^t {}_t p_x \mu_{x+1} dt.$$

The distribution of the random variable Z at j moment can be determined from:

$$\mathbb{E}\left(Z^j\right) = \mathbb{E}(z_T) = \int_0^{n} (v_t)^j {}_t p_x \mu_{x+1} dt = \int_0^{n} e^{-(\delta \cdot j)t} {}_t p_x \mu_{x+1} dt.$$

It follows from this equation that the j moment of the distribution of Z is equal to the one-time premium of n annual insurance in the event of death for an interest rate that is j times higher than δ, that is, for the decursive factor in the continuous increase $e^{-\delta \bullet j}$. he statement is also valid for interest at the effective interest rate.

The variance, as a measure of the dispersion of the expected value, is:

$$Var(Z) = \mathbb{E}\left(Z^2\right) - (\mathbb{E}(Z))^2 = {}^2\ddot{A}_{x\overline{n}|} - \left(\ddot{A}_{x\overline{n}|}\right)^2$$

where ${}^2\ddot{A}_{x\overline{n}|}$ is a one-time net premium for an n – year period with an interest rate of 2δ.

In the case of lifetime capital insurance, the insurer must pay the sum insured to the beneficiaries upon the occurrence of the insured event. The assumptions of this stochastic model are:

$b_t = 1$ for $t \geq 0$
$v_t = v^t$ for $t \geq 0$
$Z = v^t$ for $T \geq 0$.

The symbol for the one-time net premium for life insurance $\ddot{A}_x$ is introduced and is determined:

$$\ddot{A}_x = \mathbb{E}(Z) = \mathbb{E}(z_t) = \int_0^{\infty} z_t \cdot g(t)dt = \int_0^{n} v^t {}_t p_x \mu_{x+1} dt$$

Life insurance lasts until the end of the insured person's life, so the number of years of insurance is considered infinitely large, i.e. $n \to \infty$. We know from experience that there are few people who live more than 100 years, but in general the marginal value of life expectancy, and thus of insurance, is an infinite value. This assumption does not significantly change the value of the one-time premium because:

$$\int_{100}^{\infty} v^t {}_t p_x \mu_{x+1} dt \to 0$$

If we observe the intensity of mortality μ_x as a constant μ with a certain constant interest rate $\delta = \frac{p}{100}$, the one-time lifetime insurance premium can be expressed using the following equation:

$$\ddot{A}_x = \mathbb{E}(Z) = \mathbb{E}(z_t) = \int_0^{\infty} e^{-\delta t} e^{-\mu t} \mu dt = \frac{\mu}{\mu + \delta}.$$

In the case of life insurance, the sum insured is paid to the beneficiary if the insured lives to the age for which the contract was concluded. The following applies:

$$Z = \begin{cases} 0,\ T \le n \\ v^n, T \succ n \end{cases} \qquad \text{when} \quad b_t = \begin{cases} 1, t \le n \\ 0, t \succ n \end{cases} \quad \text{and} \quad v_t = v^t,\ t \ge 0$$

The one-time net premium for the case of survival is denoted by $A_{x\overline{n}|}$ and is equal to:

$$A_{x\overline{n}|} = \mathbb{E}(Z) = v^n {}_np_x$$

with variance

$$Var(Z) =^2 A_{x\overline{n}|} - \left(A_{x\overline{n}|}\right)^2 = v^{2n} \cdot {}_np_x \cdot {}_nq_x.$$

2.4. **Present value of multiple payments**

If the premiums are paid continuously until the occurrence of the insured event, and if the symbol a_x is introduced for the expected present value of all payments, we have

$$a_x = \int_0^\infty e^{-\delta t} dt.$$

Let's assume for simplicity that the payments are unitary and that they are paid over n years, we have

$$a_x = \int_0^n e^{-\delta t} dt.$$

If we denote by A_x the expectation of the stochastic discounted present value of the sum insured (in the amount of one monetary unit), we have

$$A_x = \int_0^\infty e^{-\delta t} f(t) dt,$$

where $f(t)$ is the probability density function for the remaining lifetime of the random variable T_x.

From the last two equalities we have

$$A_x = \int_0^\infty e^{-\delta t} f(t) dt = 1 - \delta a_x.$$

When the insurance is paid in multiple equal payments, the net periodic premium is calculated by the ratio,

$$NPP = \frac{A_x}{a_x} = \frac{1}{a_x} - \delta.$$

With this approach, the discount values of payments (payments) that are a function of time are equated with expected values that depend on time and the interest rate.

3. CONCLUSION

The stochastic model allows the determination of variance for selected functions related to mortality, and variance is by definition a deviation from the expected value and certainly one of the measures of risk. This means the possibility of determining the error for the calculated single premium. A positive characteristic of the stochastic model is certainly risk minimization, but it is not acceptable for practical application.

REFERENCES

[1] Faculty and Institute of Actuaries, Schohastic Modeling - Core Reading for subject 103.
[2] Curry B., *Long-term stochastic risk models*, Institute and Faculty of Actuaries, 2021.
[3] Slud F. V., *Actuarial Mathematics and Life-Table Statistics*, Mathematics Department University of Maryland, College Park.
[4] Faculty & Institute of Actuaries, *Core Reading for Subject 302.*
[5] Institute and Faculty of Actuaries, *Actuarial Mathematics for Modelling (CM1) Core Principles.*
[6] Šain, Ž., *Aktuarski modeli životnih osiguranja, I.dio, Osnove aktuarske matematike*, Ekonomski fakultet u Sarajevu, Sarajevo, 2009.
[7] Šain, Ž., *Aktuarski modeli životnih osiguranja II. Dio Primjena aktuarske matematike*, Ekonomski fakultet u Sarajevu, Sarajevo, 2009.
[8] Sarapa N., *Teorija vjerojatnosti*, Školska knjiga – Zagreb 1987.
[9] Andrijašević, S., Petranović, V., *Ekonomika osiguranja*, Alfa d.d., Zagreb, 1999.
[10] Kočović J., Šulejić P. Rakonjac-Antić T, O*siguranje*, Ekonomski fakultet Univerziteta u Beogradu.
[11] Kočović J., *Aktuarske osnove formiranja tarifa u osiguranju lica*, Ekonomski fakultet Univerziteta u Beogradu.
[12] Hodžić M., *Some Extensions to Classic Lotka-Volterra Modeling For Predator Prey Applications*, Southeast Europe Journal of Soft Computing, 2014.
[13] Brkić S., Hodžić M., Džanić E., *Soft-hard data fusion using uncertainty balance principle -corporate credit risk in commercial banking*, Periodicals of Engineering and Natural Sciences, 2019.
[14] Hodžić M., Saračević N., *Credit Risk Assessment for an Islamic Bank in Bosnia and Herzegovina*, Islamic Finance Practices Experiences from South Eastern Europe, Palgrave Macmillan, 2019.

(Received: May 16, 2024)
(Revised: September 13, 2024)

Suada Alić – Mešanović
Appointed actuary
Asa Central osiguranje
Trg međunarodnog prijateljstva 25
71000 Sarajevo
Bosnia and Herzegovina
e-mail: *suadaalic@yahoo.com*

PROCEEDINGS OF THE CONFERENCE ON MARCH 14 – INTERNATIONAL DAY OF MATHEMATICS
Special Editions ANUBiH, Book CCXVI, OPMN Book 30
Sarajevo 2024, **p. 148–160** DOI: 10.5644/PI2024.216.15

USING MATHEMATICAL SOFTWARE FOR HYPERBOLIC PARABOLOIDS IN BUILDING DESIGN

IRMA IBRIŠIMOVIĆ, SELMA PLAVŠIĆ AND AJŠA HRUSTIĆ

Dedicated to the 75th birthday of our dear Professor Mirjana Vuković

ABSTRACT. In the modern world it is impossible to imagine an environment without concrete, buildings, asphalt tracks, banks, and the like. Throughout history, the environment has changed and developed due to human desire for intellectual and material progress. From flat to minimal surfaces, knowledge of surfaces is closely related to construction. Builders and planners used new forms in construction, and the ball, circle, sphere, and their parts were used as a basis. As forms were developed, so were new materials, but stone remained a building material for thousands of years. The emergence of new building materials is often combined with new forms. Until then, the usual forms were spheres, rollers, planes, and surfaces. The 20th century was characterized by both theoretical study and the use of plate surfaces in construction. The special feature of these surfaces is that they support themselves. One such surface is a hyperbolic paraboloid, and the reason for this is simple. The hyperbolic paraboloid is a surface of great application, great possibilities, and enviable aesthetic value. Its application is wide; however, it is mostly used as a roof surface. The application of mathematics as a science in various fields with the help of programming mathematical tools and the like is becoming increasingly common. This paper was created as a result of multiple applications of mathematics as a science, where two software packages, MATLAB and Wolfram Mathematica, were used. A hyperbolic paraboloid in space, mutual combinations of two or more equal paraboloids with different dimensions, and their application in construction are presented. In this work, a hyperbolic paraboloid is presented in a new form, a form that carries a new age of construction. The work is divided into three parts: the mathematical, the software, and the structural part.

1. INTRODUCTION

In our rapidly evolving world, using mathematics as a fundamental science is increasingly pervasive across diverse fields. Accompanied by the aid of mathematical programming tools and similar technologies its application extends into realms such as engineering, architecture, and construction. Central to our discourse is exploring plate surfaces and their geometric constructs within the context of construction endeavors.

Among the repertoire of plate surfaces, the hyperbolic paraboloid emerges as a prominent representative of second-order surfaces. Its inherent properties and geometric characteristics render it a compelling subject for further investigation. In this exploration,

2020 *Mathematics Subject Classification.* 16W20.

Key words and phrases. Mathematics, Informatics, architecture hyperbolic paraboloid, Wolfram Mathematica, MATLAB.

we delve into the conceptualization and application of the hyperbolic paraboloid in spatial contexts, examining both singular surfaces and their amalgamations in construction practice.

This inquiry unveils the simplicity of constructing hyperbolic paraboloids juxtaposed against their broad utility and aesthetic appeal. As we showcase these surfaces in novel configurations, we illuminate their role as harbingers of a new era in architectural and construction paradigms. Through meticulous examination and creative application, we embark on a journey to elucidate the transformative potential of the hyperbolic paraboloid within the fabric of contemporary building construction.

2. THE HYPERBOLIC PARABOLOID AND ITS REAL-WORLD APPLICATION

Hyperbolic paraboloids are a canonical example of a surface with a "saddle point", i.e., a stationary point that is neither a maximum nor a minimum. At such points on the surfaces the Gaussian curvature is negative. The name "hyperboloid" derives from the fact that the vertical sections of this surface are parabolas, while the horizontal cross sections are hyperbolas. However, even the vertical sections are more complex than the elliptical paraboloid. The general form of the hyperboloid equation is given by:

$$z = \frac{x^2}{a^2} - \frac{y^2}{b^2}. \tag{2.1}$$

Now let's explore a couple of real-world applications of the hyperbolic paraboloid. In construction, hyperbolic buildings use less material compared to other conical shapes. They offer greater stability against external forces than flat buildings. Despite their decorative effect, hyperbolic structures often exhibit low space efficiency [1].

For instance, hyperbolic structures find widespread use in the cooling towers of power plants and industrial facilities. Their shape facilitates efficient air circulation and heat dissipation. The upward draft created by the hyperboloid's conical shape enables effective cooling of water or gases, making it an indispensable component in thermal power plants and industrial processes. Additionally, the hyperbolic shape enhances airflow through the cooling tower. It contributes to the strength and stability of tall structures, as cooling towers must release steam into the atmosphere from a significant height. The Kobe Port Tower boasts an hourglass shape featuring two hyperbolas. This symmetry ensures that views from one side mirror those from the opposite side (Figure 1).

FIGURE 1. The Kobe Port Tower

In antenna systems, the hyperboloid shape offers advantages for telecommunications and radar applications (Figure 2).

FIGURE 2. The antenna system

It provides a wide radiation pattern, improving signal coverage. Hyperboloid reflectors and arrays are utilized in radio astronomy, satellite communications, and wireless networks for efficient signal transmission and reception over long distances. Regarding lamp design, bed lights typically have a cylindrical shape. However, when illuminated, they cast a unique, often hyperbolic shade on the wall behind them. This effect occurs because these lights usually open at both the top and bottom, resulting in circular light scattering intersected by an ordinary wall, creating a hyperbolic shade. Such forms are frequently employed for wall decoration (Figure 3).

FIGURE 3. Lamp

3. APPLICATION OF THE HYPERBOLIC PARABOLOID WITH THE HELP OF MATHEMATICAL SOFTWARE

This paper was created as a result of various applications of mathematics as a science. In this paper we used two software packages, MATLAB and Wolfram Mathematica, which is why we will talk about them briefly.

3.1. **Wolfram Mathematica**

The hyperbolic paraboloid is a surface characterized by its saddle shape, where the curvature is negative along one axis and positive along the other. In this significance test we utilize Wolfram Mathematica to analyze critical points on the surface and determine their nature-whether they represent local minima, local maxima, or saddle points. First we define the equation of the hyperbolic paraboloid and visualize it using Mathematica's plotting capabilities. Then we compute the partial derivatives of the hyperbolic paraboloid concerning its variables, x and y. Next we identify critical points by solving for the points where both partial derivatives are equal to zero. These critical points represent potential extrema or saddle points on the surface. To determine the nature of each critical point we compute the Hessian matrix-a square matrix of second-order partial derivatives-at each critical point. The eigenvalues of the Hessian matrix provide valuable information about the curvature of the surface at each critical point. If both eigenvalues are positive, the critical point represents a local minimum. Conversely, if both eigenvalues are negative, the critical point is a local maximum. If the eigenvalues have opposite signs, the critical point is a saddle point. By performing this significance test using Wolfram Mathematica we gain valuable insights into the geometric properties of the hyperbolic paraboloid and its critical points, facilitating further analysis and understanding of this important mathematical surface. One example of a hyperbolic paraboloid in the Wolfram Mathematica software, where the function is given in the form $z = x^2 - y^2$, ranging from -1 to 0.1, is provided in parametric form (Figure 4). The details can be found in [3], [4] and [5].

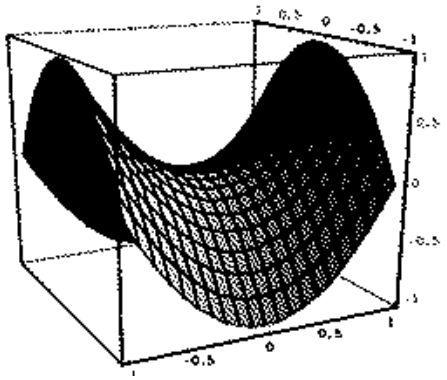


FIGURE 4. Hyperbolic paraboloid

3.2. **MATLAB**

As a computational tool widely used in engineering and scientific research, MATLAB plays a significant role in our analysis and manipulation of hyperbolic paraboloids. MATLAB provides us with a comprehensive environment for numerical analysis, allowing us to perform calculations and simulations related to hyperbolic paraboloids with precision and efficiency. Its extensive library of built-in functions and toolboxes enables us to solve complex mathematical problems associated with hyperbolic paraboloids. In addition to numerical analysis, MATLAB offers us robust visualization capabilities, making it easier to visualize and interpret the geometric properties of hyperbolic paraboloids. We can generate 3D plots, contour plots, and surface plots to gain insights into the shape, curvature and behavior of hyperbolic paraboloids (see Figure 5, [5]).

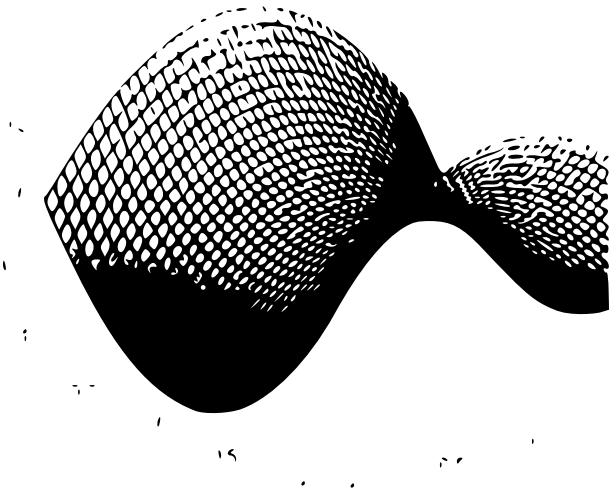


FIGURE 5. Hyperbolic paraboloid

MATLAB allows us to perform parametric modeling of hyperbolic paraboloids, enabling us to define and manipulate the parameters that govern the shape and characteristics of the surface. This flexibility facilitates our exploration of various configurations and designs of hyperbolic paraboloids for different applications. Furthermore, MATLAB includes optimization algorithms that we can apply to optimize parameters such as dimensions, curvature and orientation of hyperbolic paraboloids for specific objectives or constraints. This capability is particularly valuable in engineering and design tasks where maximizing performance or minimizing costs is essential. MATLAB seamlessly integrates with other software tools and programming languages, allowing us to combine the capabilities of MATLAB with those of other software packages for comprehensive analysis and design of hyperbolic paraboloids. This interoperability enhances our productivity and facilitates interdisciplinary collaboration.

4. APPLICATION OF THE HYPERBOLIC PARABOLOID TO THE CONSTRUCTION OF THE ROOF SYSTEM OF RESIDENTIAL BUILDINGS

To have a clear process of how to apply a hyperbolic paraboloid to the roof of an object it is necessary to first draw a 3D situation and a project of the roof (Figure 6).

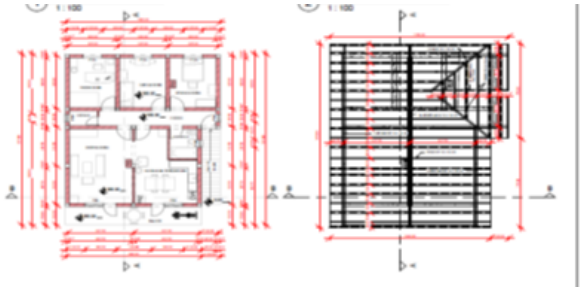


FIGURE 6. 3D situation of the roof of the buildings [6]

In the context of a roof structure, the equation can be slightly modified to suit the orientation and dimensions of the roof. Let's say the roof is aligned with the x and y axes, and the highest point (top) of the roof is at the start line (0,0,0). Then the equation of the hyperbolic paraboloid becomes

$$z = \frac{x^2}{a^2} - \frac{y^2}{b^2} + h, \tag{4.1}$$

h is the height of the peak above the $x - y$ plane (see Figure 7, [2]).

FIGURE 7. Hyperbolic paraboloid of the roof structure of a residential building [6]

This equation describes the shape of the roof surface. To create the physical structure, we should define the dimensions of the roof (length, width, and height) and then use this equation to generate the appropriate curvature for the roof surface. In an architectural and engineering context, hyperbolic paraboloid roofs are often constructed using flat beams or cables arranged in a transverse pattern to form a hyperbolic paraboloid shape. The mathematical equations that determine the geometry of these beams or cables would be more complex and would involve concepts from structural engineering and statics. The core is precisely reflected in the mathematical application of geometry to the object. By observing the project of a building it was noticed that the roof construction is very unstable and over time the roof would have to be changed in the next 3-5 years due to various everyday influences. For example, if you hold a sheet of paper in your hand, it bends and cannot support its weight. That same sheet of paper,

if squeezed or slightly curved upwards, becomes able to support its weight. The upward curvature increases the stiffness and load-bearing capacity, thereby moving part of the material away from the neutral axis. The same can be applied to the roof of the building. The load-bearing capacity of the roof structure depends on the curvature and specific curvature. When a hyperbolic paraboloid is tilted in the direction of its generating lines, it essentially means that it is rotated or tilted along one or both of its axes. Let's consider the case when it is tilted along one axis, let's say the x-axis. We can achieve this slope by introducing a rotation matrix into the equation, using the general equation of the hyperbolic paraboloid (Figure 8).

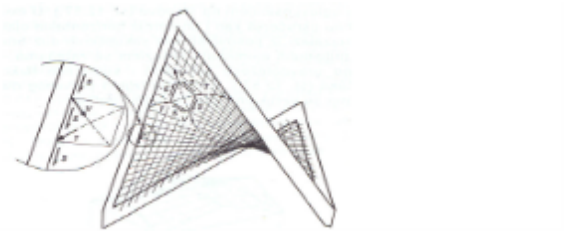


FIGURE 8. A hyperbolic paraboloid leaning in the direction of the directions that generate it [6]

A roof composed of hyperbolic paraboloids (hypers) usually includes multiple hyperunits arranged together to form an overall structure. Each hyperunit itself can be described mathematically using the equation of a hyperbolic paraboloid. The parameters would have to be determined based on the specific design requirements and constraints of the roof structure. In order to create a continuous roof surface additional considerations such as how the hyperunits are connected and joined together would also have to be solved mathematically, often through computational techniques, geometry and architectural design (Figure 9).

FIGURE 9. A view of the roof composed of hypers [6]

A roof composed of conoidal hyperbolic paraboloids refers to a structure in which multiple hyperbolic paraboloid units are arranged to form a conical shape. Each unit of the hyperbolic paraboloid can be described mathematically by the equation of the hyperbolic paraboloid. However, to represent the conoid shape we need to introduce additional parameters to control the position, orientation, and scale of each hyperbolic paraboloid unit (Figure 10).

FIGURE 10. Colloidal hyperbolic paraboloid roof [6]

5. IMPLEMENTATION OF HYPERBOLIC PARABOLOID ON ROOF SYSTEMS USING MATHEMATICAL SOFTWARE

Now based on the above consideration, in our paper we can create animations that will solve the observed problems. For the sake of simplicity, animations coded in mathematical software are shown in the paper in the form of figures.
An animation made in MATLAB showing the construction of a hyperbar (hyperbolic paraboloid) in three directions proceeds as follows. The animation begins with the initialization of the plot window in MATLAB. All the necessary parameters and variables are defined, including the dimensions of the hyper, the number of steps for each direction and all other parameters relevant to the construction. This grid consists of points in 3D space that will define the shape of the hyper. These points are calculated based on the equations that describe the hyper in three directions. The animation then proceeds to gradually construct the hyper surface. This can be done by iteratively adjusting the parameters of the hyper equations to create a smooth transition from the flat mesh to the final hyper shape. Each iteration updates the position of the grid points according to the evolving hyper equations. As the hyper is constructed, it is visualized in the MATLAB plot window. The graph is updated at each iteration to show the current state of the hyper surface. The visualization could include wireframe rendering or surface rendering to show the shape of the hyper more clearly. When the hyper construction is complete the animation ends showing the fully constructed surface of the hyper. At this stage any finishing touches or visualization adjustments can be made to ensure the clarity and accuracy of the final result (Figure 11).

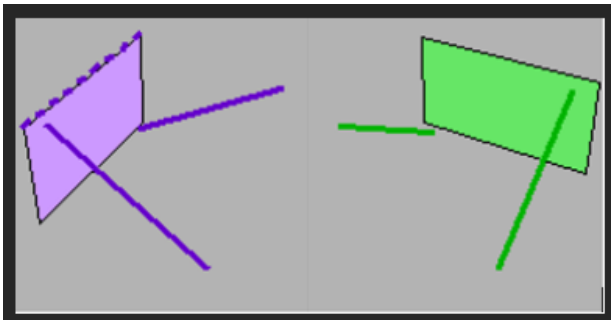


FIGURE 11. Hyper construction using three directions

To begin constructing the hyperbolic paraboloid (hyper) using MATLAB, we would initialize the MATLAB environment and set up the graphics window. This entails defining parameters such as hyperdimensions, the number of steps for each plane and any other relevant variables. Next we would generate three planes in 3D space to intersect at right angles, forming a framework for constructing the hyper. Each plane would be represented by a set of points or vertices. The points of intersection of these three planes would then be calculated, as they lie on the surface of the hyper and are crucial for its construction. Subsequently, a grid or matrix of points on the surface of the hyper would be created based on the intersection points of the three planes, effectively defining the

shape of the hyper. Through iterative processes, each point on the grid would be processed to calculate its position using the equations of the hyperbolic paraboloid. This calculation determines the height (z-coordinate) of each point based on its x and y coordinates. Throughout this process, the MATLAB graph would be continuously updated to visualize the evolving shape of the hypersurface. Various rendering techniques such as wireframe or surface rendering can be employed to enhance clarity. By leveraging MATLAB's animation functions the plot would be updated at each iteration to depict the current state of the hyper-construction. This would create a smooth transition between frames, effectively illustrating how the hyper gradually takes shape over time (Figure 12).

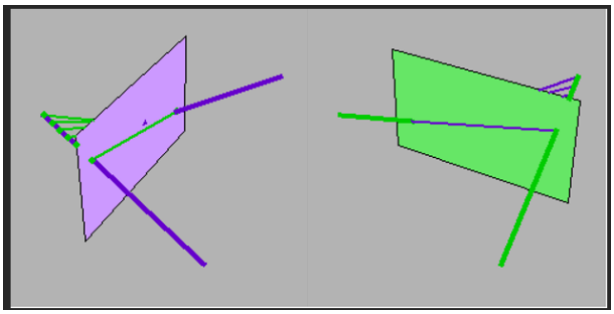


FIGURE 12. Construction using three planes

MATLAB Code for Animating the Hyperbolic Paraboloid:

Define the range and parameters for the plot

$$x = linspace(-5,5,100);$$
$$y = linspace(-5,5,100);$$
$$[X,Y] = meshgrid(x,y);$$

Hyperbolic paraboloid equation

$$Z = X.^2 - Y.^2;$$

Create a figure for the animation

$$figure;$$
$$axis([-55-55-2525]);$$
$$hold\ \ on;$$

Plot the hyperbolic paraboloid surface

$$hSurface = surf(X,Y,Z);$$
$$shading\ \ interp;$$
$$colormap(jet);$$

Define animation parameters

$$numFrames = 100;$$
$$angleStep = 360/numFrames;$$

Example construction of rulings (generatrices) for two systems

System 1 rulings

$$for\ \ t = -5 : 0.5 : 5$$
$$lineX = linspace(-5, 5, 100);$$
$$lineY = t * ones(1, 100);$$
$$lineZ = lineX.^2 - lineY.^2;$$
$$plot3(lineX, lineY, lineZ, 'k', 'LineWidth', 1);$$
$$end$$

System 2 rulings

$$t = -5 : 0.5 : 5$$
$$lineX = t * ones(1, 100);$$
$$lineY = linspace(-5, 5, 100);$$
$$lineZ = lineX.^2 - lineY.^2;$$
$$plot3(lineX, lineY, lineZ, 'r', 'LineWidth', 1);$$
$$end$$

Animation loop

$$for\ \ k = 1 : numFrames$$

Rotate the view

$$view(angleStep * k, 30);$$

Update the plot

$$drawnow;$$

Pause for a short duration to control the speed of the animation

$$pause(0.05);$$
$$end$$

Optionally, save the animation as a video

$$v = VideoWriter('hyperbolic_paraboloid.avi');$$
$$open(v);$$
$$for\ \ k = 1 : numFrames$$
$$view(angleStep * k, 30);$$
$$frame = getframe(gcf);$$
$$writeVideo(v, frame);$$
$$drawnow;$$
$$end$$
$$close(v);$$

In the process of animating a hyperbolic paraboloid (hyper) construct using winged four-pointers in MATLAB, our team begins by setting up the environment, initializing parameters and defining variables relevant to the animation (Figure 13). These parameters include the dimensions of the hyper, the number of steps for the animation stages

and other important variables. We then proceed to define the geometry of the four-pointer wing structure, typically consisting of four intersecting planes arranged in a specific configuration to create a hyper shape. Each plane is represented by its equation or set of points in 3D space. Subsequently, we calculate the intersection points of the four planes. These points serve as the vertices of the hyper, determining its overall shape and structure. Using these intersection points, we generate a grid or matrix of points representing the hyper surface. These points are crucial for defining the surface of the hyper and form the basis for its construction. Throughout the animation process, we progress through each stage of the construction, gradually adjusting the parameters or positions of the intersection points to ensure a smooth transition from the initial state to the final hyper shape. At each iteration, the MATLAB plot is updated to visualize the ongoing construction of the hyper. This visualization may employ techniques such as wireframe or surface rendering to effectively display the evolving hyperstructure. To achieve fluid animation, the plot is updated at regular intervals, illustrating the progressive development of the hyper construction over time. MATLAB's animation functions provide control over frame timing and appearance, ensuring coherent and visually appealing animation. Once the construction of the hyper is complete, we finalize the animation, showcasing the fully constructed surface of the hyper.

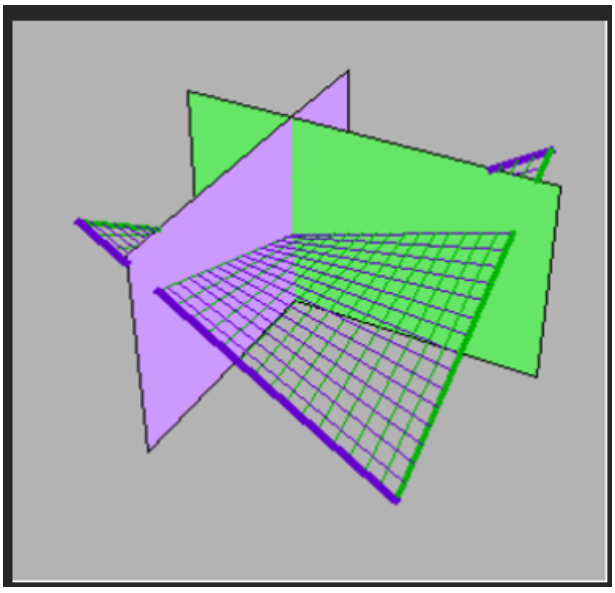


FIGURE 13. Construction using four tops

In the context of MATLAB animation, the construction of a hyperbolic paraboloid (hyper) using translation surfaces involves several key steps (Figure 14):

1. It starts by setting up the MATLAB environment and initializing parameters such as hyper dimensions, number of animation steps and any other relevant variables.
2. Create translation surfaces that serve as a framework for constructing the hyper. These surfaces can be generated by translating a curve or profile along two

orthogonal directions in 3D space. These translation surfaces will intersect to form a hypershape.
3. Calculate the intersection points of the translational surfaces. These points will define the peaks of the hipper and determine its overall shape.
4. Generate a grid or grid of points on the hypersurface. These points will be used to construct the hyper and provide the basis for its visualization.
5. Iterate through each step of the construction process, adjusting the parameters or positions of the intersection points to gradually build the hyper shape. This involves translating and transforming translation surfaces to create a hyperbolic paraboloid.
6. Update the MATLAB plot at each iteration to visualize the ongoing hyper-construction. Animation functions in MATLAB are used to create smooth transitions between frames, illustrating the progressive development of the hyper structure.
7. When the hyper structure is complete, integrate it into the roof structure by adding support elements such as beams or columns.

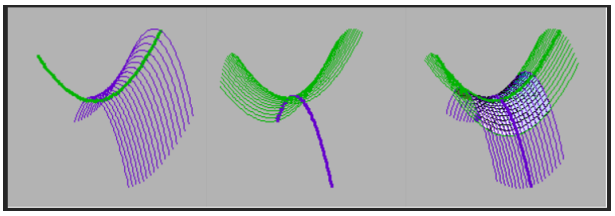


FIGURE 14. Construction using translation surfaces

All animations featured in this paper can be accessed via the following https://drive.google.com/drive/folders/1qhKCIr_Nz7qPNUbSuXrMMwGbV8n-57M3?usp=drive_link

6. CONCLUSION

Hyperbolic paraboloids offer a versatile and aesthetically pleasing solution for roof construction capable of covering large areas with minimal support structures due to their inherent structural stability. Their unique geometric shape not only adds architectural interest but also enhances the visual appeal of buildings, contributing to sophisticated design in both modern and historic architecture. Hyperbolic paraboloids efficiently distribute the load and enable economical construction while maintaining structural integrity, making them particularly suitable for large-span roof structures in residential and commercial buildings. Additionally, their geometric shape facilitates the integration of skylights and terrace windows, allowing sufficient natural light and ventilation, which can improve people's comfort and energy efficiency. Hyper roofs can be designed to accommodate green roof systems, solar panels, and rainwater harvesting systems, promoting sustainability and environmental responsibility in building design. Looking ahead, further research could explore advanced computational methods to optimize hyperdesign, innovative construction materials, and integration of hyper with new technologies such as parametric design and digital manufacturing. Studies on

the long-term performance and maintenance of hyper roofs could provide valuable insight for future architectural practice and urban planning. By embracing the potential of hyperbolic paraboloids, we can continue to push the boundaries of architectural innovation and create more sustainable and aesthetically pleasing built environments for generations to come.

REFERENCES

[1] R. Vugdalić, *Mathematics III*, Tuzla, 2014.
[2] M. Prvanović, *Projective geometry*, University of Belgrade, Scientific book, 1986.
[3] Z. Lučić, *Euclidean and hyperbolic geometry*, Second Edition, Total Design and Math Faculty, 1997.
[4] T. Došlić, N. Sandrić, *Mathematics 1*, Faculty of Civil Engineering Zagreb, 2008.
[5] Predrag S. Stanimirović, Gradimir V. Milovanović, *Mathematica software package and applications*, Niš, 2002.
[6] M. Davidović, I. Buhin *Covering buildings with hyperbolic paraboloids*, University of Zagreb,2006.

(Received: May 14, 2024)
(Revised: September 11, 2024)

Irma Ibrišimović
University of Tuzla
Department of Mathematics
Faculty of Science and Mathematics
Urfeta Vejzagića 4
75000 Tuzla
Bosnia and Herzegovina
e-mail: *irma.ibrisimovic@untz.ba*
and
Selma Plavšić
University of Tuzla
Department of Mathematics
Faculty of Science and Mathematics
Urfeta Vejzagića 4
75000 Tuzla
Bosnia and Herzegovina
e-mail: *selma.plavsic@untz.ba*
and
Ajša Hrustić
University of Tuzla
Department of Mathematics
Faculty of Science and Mathematics
Urfeta Vejzagića 4
75000 Tuzla
Bosnia and Herzegovina
e-mail: *ajsa.hrustic@untz.ba*

PROCEEDINGS OF THE CONFERENCE ON MARCH 14 – INTERNATIONAL DAY OF MATHEMATICS
Special Editions ANUBiH, Book CCXVI, OPMN Book 30
Sarajevo 2024, **p. 161–171** DOI: 10.5644/PI2024.216.16

ASSESSMENT OF MATHEMATICS STUDENTS' KNOWLEDGE AND SKILLS

KARMELITA PJANIĆ AND SANELA NESIMOVIĆ

This article is dedicated to dear Professor Mirjana Vuković on the occasion of her 75th birthday

ABSTRACT. Evaluating and assessing university students' knowledge and skills is a complex process and for many professors it is the most challenging aspect of their job. Mathematics curricula highlight the learning outcomes and competencies that students will acquire at the end of an educational cycle, which should guide professors in designing assessments of students' knowledge and skills. When designing assessments, consideration should be given to the objectives of the assessment - what is to be assessed or measured. In addition, questions and tasks should be relevant, varied in form, varied in difficulty, clear and understandable, without double meaning or confusion, with clear and precise instructions. Any test should be reliable and valid. The scoring of the results and their interpretation should be clear. Combining the above will ensure that the test is of high quality and measures what needs to be assessed. Teachers should ask themselves what exactly the tasks and questions they use in exams are measuring. Are they using tests and are they made up of questions and tasks that meet all the criteria for a test? Do these tests, questionnaires and sets of objective tasks provide answers about the results obtained? The aim of this paper is to provide an overview of recent research on the assessment of mathematics students' knowledge and skills with a particular focus on the assessment of student performance in proving mathematical statements, and e-assessment in mathematics at university level.

1. INTRODUCTION

The landscape of assessment in university level mathematics education is undergoing significant change, driven by advances in educational theory, technology and changing educational goals. Traditional assessment paradigms are being re-evaluated in the light of student-centered approaches to learning, and online learning environments have introduced new challenges and opportunities for assessing mathematical understanding. This paper explores different assessment paradigms, methods and the current state of research on assessment in university-level mathematics. In particular, we focus on online environments and the assessment of mathematical proofs.

2020 *Mathematics Subject Classification.* 97D60.
Key words and phrases. assessment, university-level mathematics, e-assessment, proof.

2. Assessment paradigms in university-level mathematics

Assessment paradigms in education serve as basic frameworks that guide the evaluation of student learning and the effectiveness of teaching. In the context of university mathematics, these paradigms are central to aligning assessment practices with educational goals and learning outcomes. Traditionally, the assessment of learning has been the norm, predominantly through summative means such as standardized tests and final examinations, emphasizing the measurement of learning outcomes against pre-defined standards [6]. However, such traditional methods, including written and multiple-choice tests, have been criticized for their limited ability to inform teaching and promote deeper understanding [36], [35]. They often focus on procedural knowledge rather than conceptual understanding and may not fully capture students' reasoning or creativity [34]. Over the past three decades, a paradigm shift has been advocated, emphasizing assessment for learning rather than assessment of learning [5], [8].

Assessment for learning, characterized by formative approaches that focus on continuous feedback to students and teachers, has been shown to have a significant positive impact on student achievement [6]. It encourages self-assessment, reflection and autonomy in the learning process [13]. In mathematics education, this shift could include iterative problem-solving sessions with feedback not only on the correctness of solutions but also on the underlying reasoning [7]. In addition, alternative assessment methods such as portfolios, projects and collaborative assessments have been proposed to promote student autonomy and responsibility [55], [57], [1]. These learner-centered approaches promote deeper engagement, collaboration and increased interaction between students and teachers [64]. They are consistent with the aims of initiatives such as the Bologna Process, which emphasize student autonomy and responsibility in learning [44]. Research suggests that learner-centered methods, such as the use of portfolios, lead to deeper learning outcomes compared to traditional assessments [57], [47].

Furthermore, learner-centered methods aim to develop students' autonomy and sense of responsibility, in line with the aims of initiatives such as the Bologna Process [44]. They promote autonomous learning and enable students to take responsibility for their learning process [51]. Research suggests that the use of portfolios for student assessment, as opposed to methods such as multiple-choice tests, leads to deeper learning outcomes [57], [47]. Overall, these findings highlight the potential benefits of moving towards learner-centered assessment methods to improve student engagement, learning outcomes and self-regulation in higher education.

3. Research of assessment in mathematics at university level

In general, research on assessment in mathematics at university level has taken several directions, such as research on tasks/problems given in written examinations, case studies of single classroom assessment methods, assessment practices, impact of assessment on learning, e-assessment, assessment of evidence. Much of the research has focused on a single classroom case study exploring a less traditional form of assessment (e.g. [16], [33], [54].

Another subset of the literature on assessment in mathematics education focuses specifically on the items included in traditional written examinations. These studies typically take a sample of written exams and analyze the items within each assessment. These studies consistently show a tendency for such exams to prioritize procedural knowledge over conceptual understanding. Notable findings include those of [4], [22], [56], [65] and [42]. Tallman et al. [56] found that the majority of Calculus I final exams sampled across the US required minimal cognitive demands, focusing primarily on recall and application of procedures. Furthermore, they showed that there has been little change in exams that require students to use higher-order thinking (e.g., applying understanding) over the course of approximately 25 years. Mac an Bhaird et al.'s [22] analysis of calculus exams in Irish universities echoed this trend, indicating a lack of emphasis on conceptual understanding. Despite calls for alternative assessment methods, such as those advocated by [21], there has been little evolution in exam design over time. Reed et al. [42] proposed criteria for high quality exam items, emphasizing the importance of assessing conceptual understanding and higher order thinking skills, a recommendation supported by findings from Mac an Bhaird et al. [22] and Tallman et al. [56].

Only a few studies address the observation of assessment practices. Iannone and Simpson's [19] study offered a comprehensive examination of assessment in undergraduate mathematics courses in England and Wales, providing valuable insights into the assessment methods used in these settings. Through careful data collection from 43 representative courses and in-depth interviews with 27 senior members of mathematics departments, Iannone and Simpson [19] uncovered notable trends in assessment practices. The study revealed a striking prevalence of closed-book examinations, with over a quarter of modules assessed entirely in this format.

Furthermore, almost 70% of modules allocated a significant proportion of the final mark to closed-book examinations, indicating the widespread use of this assessment method across different mathematics courses. The research also elucidated the perspectives of heads of mathematics departments on assessment practices. These senior members expressed support for closed-book exams, but also raised concerns about alternative assessment methods. Key concerns included issues of fairness, plagiarism, collusion, satisfaction with existing assessment patterns, institutional pressures, employability and promoting student learning. A decade later, Iannone and Simpson [21] revisited their seminal research to assess whether significant changes had occurred in assessment practices across the UK higher education landscape. Despite the intervening years being marked by changes at a policy level and an increased emphasis on training and supervision in higher education teaching, the findings revealed a surprising continuity in assessment practices. Closed-book examinations continued to be the predominant method of assessment in universities, suggesting a remarkable persistence of traditional assessment methods despite changing educational contexts.

Moreover, the prevalence of traditional exams in mathematics education reflects broader trends observed in various disciplines worldwide. Studies by researchers such as Meyer et al. [27], Postareff et al. [38] and Rawlusyk [39] have documented a similar dominance of traditional assessment methods in higher education institutions world-

wide. This comparative perspective underscores the pervasiveness of traditional assessment practices, and highlights the need for further research and potential reform of assessment methods to better align with contemporary educational goals and student needs.

In their study, Zheng et al. [70] conducted a comprehensive review of syllabi from courses offered during the spring 2021 semester at a large university in the southwestern United States. Their findings provide valuable insights into the prevailing assessment practices at this educational institution, building on previous research [19], [21] conducted by Iannone and Simpson in 2012 and 2022. The study shows that written tests were the most commonly used form of assessment, used in approximately 91.9% of the sampled courses. This was closely followed by traditional homework, which was used in 93% of courses. Quizzes were the third most common form of assessment, although significantly less common than written tests and homework, used in 64% of courses. In addition, participation in the final assessment was included in approximately 53.5% of course sections. Further analysis of the data revealed variations in assessment practices across subjects, years and class sizes, with patterns generally in line with practical expectations. For example, applied mathematics courses tended to include projects more frequently than pure mathematics courses. In addition, lower-division undergraduate courses had the greatest variety of assessment methods, with an average of 3.8 methods per course. In contrast, upper-division undergraduate and postgraduate courses tended to use fewer assessment methods, with an average of 2.7 and 3.0 methods respectively. The study also highlighted the relatively low use of learner-centered approaches to assessment. For example, portfolio assessment, oral exams, case studies, peer assessment, self-assessment and reading notes/questions were rarely used in the sampled courses, indicating a potential gap between current assessment practices and the principles of learner-centered education. Overall, the study provides a nuanced understanding of assessment practices within a large university setting, highlighting the prevalence of traditional assessment methods such as written tests and homework, as well as providing insights into variations by subject area, year level and class size. It also highlights the limited adoption of learner-centered approaches to assessment and suggests potential areas for further research and development of assessment practices in higher education.

4. Research on impact of assessment on learning

Early studies such as [29] ”found unexpectedly that what influenced students most was not the teaching but the assessment” [12]. Snyder [53] suggested that assessment dominates how students budget time and effort to focus on the more point-bearing learning activities. He brought the term ”hidden curriculum” to the attention of the higher education community. Rowntree [43] described, ”if we wish to discover the truth about an educational system, we must first look to its assessment procedures.”

Research by [23] and [9] had firmly established assessment as a central element in higher education, with a significant impact on students' perceptions of learning and their study priorities. Assessments play a crucial role in directing students' focus, potentially leading them to superficial learning strategies aimed at passing exams rather than fos-

tering deep understanding and knowledge retention. Brown and Knight [10] highlight the dual nature of this influence, emphasizing that assessment can either help or hinder student learning, depending on the design and implementation of assessment methods.

Building on this foundational research, more recent studies have provided empirical evidence of the profound impact of assessment on students' learning strategies. Researchers such as [57], [46], [60] and [15] have contributed to our understanding of the complex relationship between assessment practices and student learning outcomes. The collective literature underlines the urgent need for educators and curriculum designers to reassess their assessment strategies. On the other hand, in contrast to the message of the general literature, Iannone and Simpson's study [20] showed that mathematics students perceive traditional assessment as the best discriminator of ability. Meyers and Nulty [28] argue that students' perceptions of assessment shape their understanding of the curriculum, highlighting the importance of designing assessments that are aligned with educational goals. Furthermore, trust in assessors, as emphasized by [67], highlights the importance of transparency and expertise in the assessment process.

Another important aspect of assessment is feedback. Beaumont, O'Doherty and Shannon [2] suggest the need to improve the quality of feedback to students in higher education, with implications for curriculum redesign. Scholars argue for a 'new culture at university' that includes faculty competencies that encompass methodological, evaluative and supportive dimensions, as proposed by [68]. The study [14] sheds light on undergraduate students' perceptions of assessment methods and feedback, providing insights for improving assessment practices in higher education. In Education, learner-centered methods are more prevalent than in other disciplines, which often rely on traditional assessment methods. In addition, the study finds that engineering students have distinct preferences for assessment methods, especially those related to team projects. Participants who frequently use learner-centered assessment methods perceive assessment as fairer and more effective than those who prefer traditional methods. However, there were no statistically significant differences between the two groups in the importance attached to feedback or the reliability of its sources.

In summary, studies of the impact of assessment on learning highlight the critical role of assessment in shaping student learning experiences and outcomes in higher education. It calls for a re-evaluation of assessment practices to ensure that they effectively support deeper learning and are aligned with educational goals, emphasizing transparency and expertise in the assessment process to foster trust and meaningful learning experiences for students.

5. E-ASSESSMENT IN MATHEMATICS AT UNIVERSITY LEVEL

The advent of online education has necessitated a re-evaluation of assessment strategies in mathematics at university level. Online assessment offers unique opportunities for scalability, flexibility and innovative assessment methods. However, it also poses significant challenges, including issues of academic integrity, accessibility, and the adequacy of the technology to fully capture students' mathematical understanding. Online assessment has evolved from simple quizzes and automated grading systems to sophisti-

cated platforms that incorporate adaptive learning technologies, real-time feedback and collaborative problem-solving tasks. Studies [25] and [3] have highlighted the growth of online assessment tools that can accommodate a wide range of mathematical tasks, from basic arithmetic to complex calculus problems.

Methods used in online assessment include computer-assisted assessment (CAA), which often includes automated feedback mechanisms, and dynamic assessment tools, which adjust the level of difficulty based on the student's performance. A large body of research emphasize the importance of these tools in providing immediate feedback and personalized learning experiences, which are crucial for mastering mathematical concepts (e.g. [50], [11], [59]).

E-assessments offer several advantages, especially when teaching large cohorts: they offer the possibility of automatic marking and feedback. However, most e-learning systems are poorly adapted for use in mathematics, a language in its own right [17]. According to [52], current e-learning systems do not adequately support the necessary notations and diagrams, "the very building blocks of mathematical communication". The effectiveness of online assessment is measured by its ability to accurately assess student understanding, promote engagement and support learning outcomes. A meta-analysis [25] suggests that online and blended learning environments, when properly designed, can be as effective as traditional classrooms in terms of student achievement in mathematics. However, the transition to online assessment requires careful consideration of assessment design, technological infrastructure, and pedagogical strategies to ensure that assessments are fair, reliable, and aligned with learning objectives. Research [41] suggests that well-designed online assessments can enhance the teaching and learning of complex mathematical skills by incorporating interactive simulations, visualizations and problem-solving tasks.

Despite its potential, online assessment faces several challenges. Academic integrity is a major concern, with research [61] exploring the prevalence of cheating in online environments and strategies to mitigate this problem. In addition, technical difficulties, accessibility issues and the digital divide can hinder the effectiveness of online assessment, potentially exacerbating inequalities between students [49]. The assessment of higher order thinking skills, such as mathematical reasoning and proof construction, remains a complex area in online environments. Traditional assessment methods may not translate well to digital formats, and innovative approaches are needed to accurately assess these skills [63].

6. Assessment of mathematical proof

The foundation of pure mathematics lies in mathematical proofs. At university level, the assessment of a mathematics student's performance revolves primarily around evaluating their ability to construct proofs [62]. However, assessing proofs is challenging, as mathematicians need to assess not only the quality of the written proof, but also the depth of understanding behind it [30]. Furthermore, there is a noticeable lack of assessment tools specifically designed to assess the understanding of proofs [26].

Several studies have been conducted to measure proof comprehension skills. In the study [66], authors refer to it as "reading proof comprehension" because understanding proof is more dominant in the reading aspect. Yang and Lin [66] attempted to compile instruments to measure reading proof comprehension skills for secondary school students. They conceptualized proof comprehension on the basis of previous studies [18] and [48]. Yang and Lin [66] formulated five facets in reading proof comprehension: basic knowledge, logical status, integration or summation, generality, and application or extension.

Nevertheless, the question of how to effectively assess a student's understanding of proofs remains open [26]. Feedback is emerging as an important tool for supporting students' learning of proof construction [31], [32]. However, students often fail to understand feedback from lecturers and rarely receive further feedback on their revisions [37]. Furthermore, proof validation, which involves critically evaluating proofs to determine their correctness [48], has been shown to have a positive impact on students' ability to construct their own proofs [40]. This process requires students to engage deeply with their learning material, such as asking and answering questions, constructing subproof, and interpreting definitions and theorems [48].

Nevertheless, the question of how to effectively assess a student's understanding of proofs remains open [26]. Feedback is emerging as an important tool for supporting students' learning of proof construction [31], [32]. However, students often fail to understand feedback from lecturers and rarely receive further feedback on their revisions [37]. Furthermore, proof validation, which involves critically evaluating proofs to determine their correctness [48], has been shown to have a positive impact on students' ability to construct their own proofs [40]. This process requires students to engage deeply with their learning material, such as asking and answering questions, constructing subproofs, and interpreting definitions and theorems [48].

7. CONCLUSION

The landscape of assessment in mathematics at university level is complex and diverse, encompassing a range of paradigms and methods that reflect the diverse goals of mathematics education. As the field continues to evolve, particularly with the increasing prevalence of online learning environments, educators and researchers must continue to explore innovative assessment strategies that not only measure mathematical knowledge and skills, but also foster deeper engagement with the material. Central to these efforts will be the development of assessment methods that are equitable, inclusive and capable of capturing the nuanced and sophisticated thinking that characterizes advanced mathematics. Ultimately, the aim of assessment in mathematics at university level should be not only to evaluate learning outcomes, but also to enhance the educational experience by promoting a deeper understanding and appreciation of mathematics.

REFERENCES

[1] R. J. Almond, Group assessment: comparing group and individual undergraduate module marks. *Assessment & Evaluation in Higher Education*, 34, no. 2, 2009. pp. 141–48.

[2] C. Beaumont, M. O. O'Doherty, L. Shannon, Reconceptualising assessment feedback: a key to improving student learning? *Studies in Higher Education*, 36, no. 6, 2011. pp. 671–87.

[3] R. E. Bennett, H. Persky, A. Weiss, F. Jenkins, Measuring Problem Solving with Technology: A Demonstration Study for NAEP. *Journal of Technology, Learning, and Assessment*, 8(8), 2010. Retrieved 8.3.2024. from `http://www.jtla.org`.

[4] E. Bergqvist, Types of reasoning required in university exams in mathematics. *The Journal of Mathematical Behavior*, 26(4), 2007. pp. 348–370.

[5] M. Birenbaum, New insights into learning and teaching and their implications for assessment. In M. Segers, F. Dochy, and E. Cascallar (Eds.), *Optimising new modes of assessment: In search of qualities and standards*, Dordrecht: Kluwer, 2003.

[6] P. Black, D. Wiliam, Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 1998. pp. 7–74.

[7] P. Black, C. Harrison, C. Lee, B. Marshall, D. Wiliam, *Assessment for learning: Putting it into practice*, Buckingham: Open University Press, 2003.

[8] P. Black, Assessment for learning: where is it now? Where is it going? In C. Rust (Ed.) *Improving student learning through the curriculum*, Oxford: Oxford Centre for Staff and Learning Development, 2006. pp. 9–20.

[9] D. Boud, R. Cohen, J. Sampson, Peer learning and assessment. *Assessment & Evaluation in Higher Education* , 24, no. 4, 1999. pp. 413–26.

[10] S. Brown, P. Knight, *Assessing learners in higher education.* , London: KoganPage, 1994.

[11] G. Conole, B. Warburton, A review of computer-assisted assessment, *Research in Learning Technology*, Vol. 13, No. 1, 2005. pp. 17–31.

[12] F. Dochy, M. Segers, D. Gijbels, Assessment engineering: Breaking down barriers between teaching and learning, and assessment. In D. Boud and N. Falchikov (Eds.), *Assessment & Evaluation in Higher Education*, New York, NY: Routledge, 2007. pp. 97–110.

[13] L. M. Earl, *Assessment as Learning: Using Classroom Assessment to Maximize Student Learning* , Corwin Press, 2003.

[14] M. A. Flores, A. M. Veiga Simão, A. Barros, D. Pereira, Perceptions of effectiveness, fairness and feedback of assessment methods: a study in higher education. *Studies in Higher Education*, 40(9), 2014. pp. 1–12.

[15] S. Fernandes, M. A. Flores, R. M. Lima, Students' views of assessment in project-led engineering education: findings from a case study in Portugal. *Assessment & Evaluation in Higher Education*, 37(2), 2012. 163–178.

[16] B. Gold, S. Z. Keith, W. A. Marion (Eds.), *Assessment Practices in Undergraduate Mathmatics*, The Mathematical Association of America, 1999.

[17] S. Gruttmann, D. Böhm, H. Kuchen, E-assessment of mathematical proofs: chances and challenges for students and tutors, *2008 International Conference on Computer Science and Software Engineering*, Vol. 5, IEEE, 2008. pp. 612–615.

[18] L. Healy, C. Hoyles, A Study of Proof Conceptions in Algebra. *Journal for Research in Mathematics Education*, Vol. 31, No. 4, 2000. pp. 396–428.

[19] P. Iannone, A. Simpson (Eds.), *Mapping university mathematics assessment practices*, Norwich: University of East Anglia, 2012.

[20] P. Iannone, A. Simpson, Students' perceptions of assessment in undergraduate mathematics. *Research in Mathematics Education*, 15(1), 2013. pp. 17–33.

[21] P. Iannone, A. Simpson, How we assess mathematics degrees: the summative assessment diet a decade on. *Teaching Mathematics and its Applications: an International Journal of the IMA*, 41(1), 2022. pp. 22–31.

[22] C. Mac an Bhaird, B. C. Nolan, A. O'Shea, K. Pfeiffer, A study of creative reasoning opportunities in assessments in undergraduate calculus courses. *Research in Mathematics Education*, 19(2), 2017. pp. 147–162.

[23] F. Marton, R. Saljo, Approaches to learning. In F. Marton, D. Hounsell, and N. Entwistle (Eds.) *AThe experience of learning. Implications for teaching and studying in higher education*, Edinburgh: Scottish Academic Press, 1997. pp. 39–58.

[24] G. Mayrhofer, S. Saminger, W. Windsteiger, CreaComp: Computer-Supported Experiments and Automated Proving in Learning and Teaching Mathematics, *Proceedings of ICTMT 8*, Corwin Press, 2007.

[25] B. Means, Y. Toyama, R. Murphy, M. Bakia, K. Jones, *Evaluation of Evidence-Based Practices in Online Learning: A Meta-analysis and Review of Online Learning Studies*, US Department of Education. https://www2.ed.gov/rschstat/eval/tech/evidence-based-practices/finalreport.pdf, 2010.

[26] J. P. Mejia-Ramos, E. Fuller, K. Weber, K. Rhoads, A. Samkof, An assessment model for proof comprehension in undergraduate mathematics. *Educational Studies in Mathematics*, 79(1), 2012. pp. 3–18.

[27] L. Meyer, S. Davidson, L. Mckenzie, M. Rees, H. Anderson, R. Fletcher, P. Johnston, An investigation of tertiary assessment policy and practice: Alignment and contradictions. *Higher Education Quarterly*, 64(3), 2010. pp. 331–350.

[28] N. M. Meyers, D.D. Nulty, How to use (five) curriculum design principles to align authentic learning environments, assessment, students' approaches to thinking and learning outcomes. *Assessment & Evaluation in Higher Education*, 34, no 5, 2008. pp: 565–577.

[29] C. M. I. Miller, M. Parlett, *Up to the mark: A study of the examination game*, Guildford: Society for Research into Higher Education, 1974.

[30] D. Miller, N. Infante, K. Weber, How mathematicians assign points to student proofs. *The Journal of Mathematical Behavior*, 49, 2018. pp: 24–34.

[31] R. C. Moore, Mathematics professors' evaluations of students' proofs: A complex teaching process. *International Journal of Research in Undergraduate Mathematics Education*, 2(2), 2016. pp: 246–278.

[32] R. C. Moore, M. Byrne, S. Hanusch, T. Fukawa-Connelly, When we grade students' proofs, do they understand our feedback? *Faculty Publications*, 422, 2016. Retrieved 8.3.2024. from https://digitalcommons.andrews.edu/pubs/422.

[33] M. Munakata, C. Monahan, E. Krupa, A. Vaidya, Non-traditional assessments to match creative instruction in undergraduate mathematics courses. *International Journal of Mathematical Education in Science and Technology*, 54(7), 2023. pp: 1272–1287.

[34] J. Pellegrino, N. Chudowsky, R. Glaser, *Knowing What Students Know: The Science and Design of Educational Assessment*, National Academy Press, 2001.

[35] D. Pereira, M. Flores, L. Niklasson, Assessment revisited: A review of research in assessment and evaluation in higher education. *Assessment & Evaluation in Higher Education*, 41(7), 2016. pp: 1008–1032.

[36] P. Perrenoud, *Avaliação: da excelência à regulação das aprendizagens: entreduaslógicas*, Porto Alegre: Artmed., 1999.

[37] A. Pinto, J. Cooper, Formative assessment of proof comprehension in undergraduate mathematics: Affordances of iterative lecturer feedback. *Eleventh Congress of the European Society for Research in Mathematics Education*, Utrecht University, Utrecht, 2019.

[38] L. Postareff, V. Virtanen, N. Katajavuori, S. Lindblom-Ylänne, Academics' conceptions of assessment and their assessment practices. *AStudies in Educational Evaluation*, 38(3-4), 2012. pp: 84–92.

[39] E. P. Rawlusyk, Assessment in higher education and student learning. *Journal of Instructional Pedagogies*, 21, 2018. pp: 1–34.

[40] R. Powers, C. Craviotto, R. Grassl, Impact of proof validation on proof writing in abstract algebra. *International Journal of Mathematical Education In Science and Technology*, 41(4), 2010. pp: 501–514.
[41] E. S. Quellmalz, J. W. Pellegrino, Technology and testing. *Science*, 323(5910), 2009. pp: 75–79.
[42] Z. Reed, M. A. Tallman, M. Oehrtman, M. P. Carlson, Characteristics of Conceptual Assessment Items in Calculus. *PRIMUS*, 32(8), 2022. pp: 881–901.
[43] D. Rowntree, *Assessing Students: How Shall We Know Them?*, London: Kogan Page, 1987.
[44] K. Sambell, L. McDowell, The values of self and peer assessment to the developing lifelong learner. In C. Rust (ed.) *Improving student learning – Improving students as learners*, Oxford, UK: Oxford Center for Staff and Learning Development, 1998. pp: 56–66.
[45] C. Sangwin, *Computer aided assessment of mathematics*, OUP Oxford, 2013.
[46] K. Scouller, The influence of assessment method on students' learning approaches: multiple choice question examination versus assignment essay, *Higher Education*, 35, 1998. pp: 453–472.
[47] M. Segers, D. Gijbels, M. Thurlings, The relationship between students' perceptions of portfolio assessment practice and their approaches to learning. *Educational Studies*, 34 no.1, 2008. pp: 35–44.
[48] A. Selden, J. Selden, Validations of Proofs Considered as Texts: Can Undergraduates Tell Whether an Argument Proves a Theorem? *Journal for Research in Mathematics Education* , 34(1), 2003. pp: 4–36.
[49] N. Selwyn, Digital downsides: exploring university students' negative engagements with digital technology. *Teaching in Higher Education*, 21:8, 2016. pp: 1006–1021.
[50] G. Sim, P Holifield, M. Brown, Implementation of computer assisted assessment: lessons from the literature, *Research in Learning Technology*, Vol. 12, No. 3, 2004. pp: 215–229.
[51] D. Sluijsmans, F. Dochy, G. Moerkerke, Creating a learning environment by using self-, peer- and co-assessment. *Learning Environment Research*, 1, 1999. pp: 293–319.
[52] G. G. Smith, D. Ferguson, Student attrition in mathematics e-learning*Australasian Journal of Educational Technology*, 21(3), 2005. pp: 323–334.
[53] B. R. Snyder, *The Hidden Curriculum*, Cambridge, MA: MIT Press, 1971.
[54] L. A. Steen, B. Gold, L. Hopkins, D. Jardine, W. A. Marion (Eds.)., *Supporting assessment in undergraduate mathematics*, The Mathematical Association of America., 2006.
[55] K. Struyven, F. Dochy, S. Janssens,, Students' perceptions about evaluation and assessment in higher education: A review. *Assessment & Evaluation in Higher Education* , 30, no.4, 2005. pp: 331–347.
[56] M. A. Tallman, M. P. Carlson, D. M. Bressoud, M. Pearson, A characterization of calculus I final exams in US colleges and universities. *International Journal of Research in Undergraduate Mathematics Education*, 2, 2016. pp: 105–133.
[57] C. Tang, Effects of modes of assessment on students' preparation strategies, In G. Gibbs (ed.) *Improving Student Learning: Theory and Practice*, Oxford: Oxford Centre for Staff Development, 1994. pp: 151–170.
[58] C. Tang, P. Lai., D. Arthur, S. F. Leung, How do students prepare for traditional and portfolio assessment in a problem-based learning curriculum? In J. Conway and A. Williams (Eds.), *Themes and Variations in PBL: Refereed proceedings of the 1999 Biennial PBL Conference*, Vol. 1, Australia: Australia Problem-Based Learning Network, 1999. pp: 206–217.
[59] A. E. Tshibalo, The potential impact of computer-aided assessment technology in higher education, *SAJHE*, 22(6), 2007. pp: 684–693.
[60] G. van der Watering, D. Gijbels, F. Dochy, L. van der Rijt, Students' assessment preferences, perceptions of assessment and their relationships to study results. *Higher Education*, 56(6), 2008. pp: 645–658.
[61] G. Watson, J. Sottile, Cheating in the Digital Age: Do Students Cheat More in Online Courses? , *Online Journal of Distance Learning Administration* , 13(1), 2010.
[62] K. Weber, Student difficulty in constructing proofs: The need for strategic knowledge. *Educational Studies in Mathematics* , 48 , 2001. pp: 101–119.

[63] K. Weber, Problem-solving, proving, and learning: The relationship between problem-solving processes and learning opportunities in the activity of proof construction. *The Journal of Mathematical Behavior*, 24, 2003. pp: 351–360.

[64] K. L. Webber, The use of learner-centered assessment in US colleges and universities. *Research in Higher Education*, 53, 2012. pp: 201–228 .

[65] N. White, V. Mesa, Describing cognitive orientation of Calculus I tasks across different types of coursework. *ZDM*, 46(4), 2014. pp: 675–690.

[66] K-L. Yang, F-L. Lin, A model of reading comprehension of geometry proof. *Educational Studies in Mathematics*, 67, 2008. pp: 59–76.

[67] M. Yorke, Summative assessment: dealing with the 'measurement fallacy'. *Studies in Higher Education*, 36 no.3, 2011. pp: 251–273.

[68] M. Zabalza, *Competencias docentes del professorado universitario. Calidad y desarrolo profesional*, Madrid: Narcea., 2007.

[69] S. Zegowitz, Evaluating the use of e-assessment in a first-year pure mathematics module. 2022. Retrieved on 8.3.2024. from https://arxiv.org/pdf/1908.01028.pdf.

[70] Y. Zheng, F. Van Vliet, J. I. Jin, Case Study of the Use of Learner-Centered Assessment in the Math School of a Large University in the United States, 2023. Retrieved on 8.3.2024. from https://doi.org/10.35542/osf.io/7gzrk

(Received: April 15, 2024)
(Revised: July 24, 2024)

Karmelita Pjanić
Pedagoški fakultet Univerziteta u Bihaću
Luke Marjanovića b.b.
Bihać
Bosnia and Herzegovina
e-mail: *kpjanic@gmail.com*
karmelita.pjanic@unbi.ba
and
Sanela Nesimović
Univerzitet u Sarajevu
Pedagoški fakultet
Skenderija 72
71000 Sarajevo
Bosnia and Herzegovina
e-mail: *snesimovic@pf.unsa.ba*

PROCEEDINGS OF THE CONFERENCE ON MARCH 14 – INTERNATIONAL DAY OF MATHEMATICS
Special Editions ANUBiH, Book CCXVI, OPMN Book 30
Sarajevo 2024, **p. 172–186** DOI: 10.5644/PI2024.216.17

MULTIMEDIA LEARNING THROUGH ONLINE MATHEMATICS EDUCATION IN ELEMENTARY AND SECONDARY SCHOOLS TO REDUCE COGNITIVE LOAD

AZRA HADŽIOMEROVIĆ

ABSTRACT. Due to the COVID-19 pandemic, the transition to online teaching has posed numerous challenges, particularly in mathematics education, which requires active interaction between teachers and students, as well as task demonstrations and feedback. Mathematics teachers and professors have attempted to adapt their methods to improve understanding of the material, reduce the burden of less critical content, and enhance concentration and retention of new information. This paper explores how cognitive learning and cognitive load theory can help improve online mathematics instruction through the use of well-designed multimedia educational content. According to cognitive load theory, learning is limited by the capacity of working memory, which must have enough space to process information and build long-term knowledge structures. Additionally, this paper examines ways to reduce working memory overload and emphasizes the importance of well-designed multimedia content for effective mathematics learning. Connecting mathematics to everyday life and using multimedia techniques can increase student interest and engagement, while excessive use of certain methods may lead to monotony and boredom.

1. INTRODUCTION

Due to the Coronavirus pandemic, we were forced to transition to online teaching. Students with online mathematics classes had the most difficulties. Since mathematics lessons require dialogue between teachers and students, demonstration of task-solving techniques, and feedback on comprehension, students of all ages encountered difficulties, particularly those in elementary schools. Students have struggled with understanding mathematical concepts since early grades, and these challenges became more pronounced during online classes. Teachers implemented various methods to conduct online mathematics classes, aiming to improve understanding of the material, reduce cognitive load from less essential content, and enhance concentration and retention of new material, often integrating it with previously learned concepts.

In this paper, we will explain what cognitive learning is, the principles of cognitive load theory, and how online mathematics education can be enhanced using multimedia educational materials.

2020 *Mathematics Subject Classification.* 97U50.
Key words and phrases. mathematics, multimedia teaching, STEM, online teaching, cognitive load.

It is crucial for teachers to carefully plan their presentations so that students can easily grasp the educational content. Understanding the material involves not only comprehending key concepts but also establishing connections between these concepts and previously acquired knowledge, and applying them in tasks. According to cognitive load theory, learning is limited by the capacity of working memory, where acquired information must be stored while leaving enough space for processing in order to form personal knowledge structures within specific areas in long-term memory. Overloading working memory, either due to the large amount of information or through ineffective teaching and testing methods, complicates the retention of information. We will discuss methods for reducing working memory overload.

Good understanding and memorizing of mathematics teaching materials can be achieved through well-designed multimedia educational content. We will outline the fundamental principles of designing multimedia educational e-content. Effective implementation requires a thorough understanding of mathematics teaching by the instructor to create high-quality multimedia content for mathematics and present it in adequate way, aimed at enhancing understanding and memory retention of mathematical concepts through online teaching. Creating educational multimedia content for mathematics instruction necessitates utilizing appropriate principles, techniques, and tools.

2. COGNITIVE LEARNING THEORIES

Cognitive psychology considers the learning process internal, suggesting that the quantity of learned material depends on cognitive processing abilities, effort invested in learning, depth of processing, and prior knowledge. Focusing on the study of learning processes, cognitive psychology includes the examination of internal processes such as memory, motivation, thinking, and reasoning.

Cognitive theories, unlike behavioral ones, are grounded in mental processes (e.g., perception, recognition, understanding, memory, problem-solving, etc.). The origins of cognitivism can be traced to Tolman, who, through experiments with various animals, found that a hungry animal seeking food in a maze gradually makes fewer errors, forming a cognitive map of where the food is located. Through these experiments, Edward Tolman (1932) demonstrated that research on internal mental processes, alongside the behavioral approach, must be included in studies for more effective learning (Širanović, 2012).

Cognitive learning theories offer a productive mechanism for knowledge acquisition, requiring a certain level of intelligence for long-term application in various situations. Through introspection, or observing one's own mental processes, new knowledge is acquired. Thus, we have insight learning and hidden learning.

In insight learning, we have a connection between the situation we find ourselves in, the means, and the goal. When faced with a particular situation, reasoning leads to a solution. This method is more pronounced in intelligent individuals who can gain insight and find unique solutions. However, insight can be trained by observing various situations, asking new questions, and encouraging critical thinking.

In implicit learning, knowledge is used only when needed, forming a cognitive map in the mind. An example of this is the hungry animal in the maze. A similar example

is when arriving in a new city, men typically look at a city map, while women notice observable details like shops and markets.

Connecting these learning approaches with mathematics instruction, there are problems that we sketch, and by observing the sketches, we determine how to reach the solution. Some students use different approaches when solving problems, demonstrating insight learning. By working on multiple problems, students can recognize similar problems and know the path to the solution, thereby training this type of learning. Mathematics problems can also require implicit learning, where students observe given data, create a cognitive map, and identify necessary formulas to find the solution.

There are two approaches to designing multimedia educational e-content: technology-focused and learner-focused (Mateljan et al., 2009).

The technology-focused approach emphasizes technological functionality for successful multimedia delivery. This approach focuses on achieving effective transmission of multimedia content, aiming to use technology efficiently. The learner-focused approach starts with the recipient's cognitive abilities, aiming to aid understanding and retention. This approach designs and tailors multimedia content to facilitate better and faster retention (Mayer, 2001).

Designing instructional content in online mathematics education is crucial for acquiring mathematical concepts. Capturing and maintaining the attention of both elementary and secondary school students is a significant challenge, even in traditional classroom settings. In addition to video discussions between the teacher and students, it is necessary to provide materials for learning and reviewing new lessons. To capture students' attention, materials should include multimedia content such as video clips, animations, and graphics. This helps retain lessons in memory longer, as multimedia elements remind students of the topics covered, allowing them to connect and apply memorized content.

2.1. **Cognitive Load Theory**

Cognitive Load Theory (CLT) is a cognitive learning theory introduced by John Sweller, an Australian educational psychologist, in the mid-1980s. The key premise of this theory is the focus on human cognitive architecture: the characteristics and interactions between long-term memory and working memory, and how cognitive load affects learning. Working memory is a critical component of this system as it allows new information to be integrated into long-term memory.

John Sweller's Cognitive Load Theory addresses techniques for reducing the load on working memory to facilitate changes in long-term memory related to schema acquisition (Schwartz et al., 2013).

One important aspect of John Sweller's Cognitive Load Theory is that heavy cognitive load can have negative effects on task completion. Experience with cognitive load shows varying effects. For example, older adults, students, and children experience different and often higher levels of cognitive load (“John Sweller's Cognitive Load Theory,” 2018).

Cognitive Load Theory is based on the relationship between our working memory and long-term memory and explains how the learning process changes depending on

cognitive load. Sweller argued that instructional design can be used to reduce students' cognitive load. Much later, other researchers developed methods for measuring perceived mental effort, which indicates cognitive load.

2.1.1. *Working memory*

In the past two decades, extensive research has been conducted on how people actually learn with respect to cognitive capacity, i.e., specificially through Cognitive Load Theory, which posits that information processing and knowledge construction are limited by the capacity of working memory. According to this theory, only a portion of the information will be processed and retained in working memory, while the rest will lead to overload, impeding information retention. Working memory is the part of the brain where information is temporarily held, worked on, and organized to achieve understanding (Sweller & Chandler, 1991).

We use working memory when performing new tasks, drawing on numerous pieces of information to successfully complete it. As previously emphasized, the amount of information that can be stored and duration it can be held in working memory are both limited. Working memory is crucial for integrating new information into long-term memory. The goal is to transfer information from working memory to long-term memory as quickly as possible and to free up space for the acquisition new information in working memory.

In order for information to be stored in long-term memory, it is organized in a specific way, allowing for vast amounts of information to be stored effectively. Once we have adopted these cognitive schemas, we can more easily access them in working memory. Therefore, our ability to manipulate vast amounts of information, which is necessary for task execution, depends on our familiarity with the task to be performed.

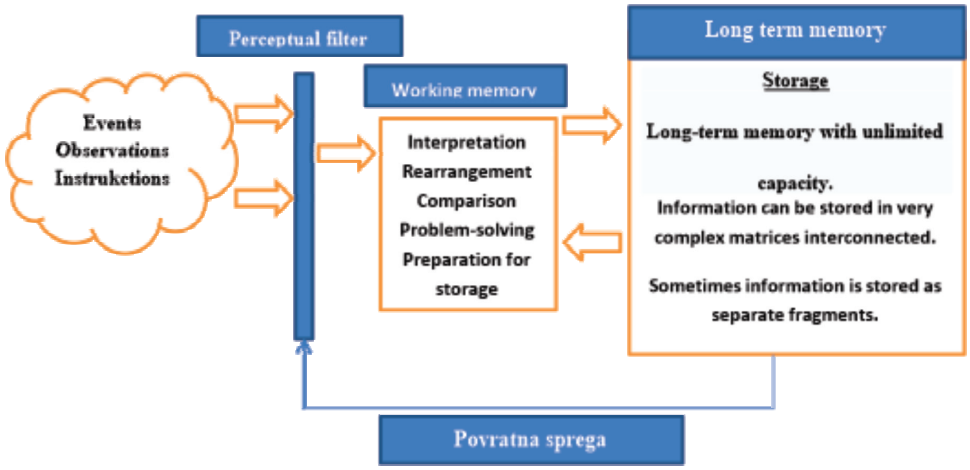


Figure 1: *Information Processing Model*

The Information Processing Model in Figure 1 shows that new information must first be perceived, then processed in working memory, and assimilated into long-term memory. If the schema is not fully adopted, it is necessary to keep all key information for task completion in working memory, which leads to greater conscious effort in per-

forming the task. For the learning process to be effective, it is important that the amount of cognitive load does not exceed the capacity of working memory (Đuras, 2020).

We know that not all students have the same prior knowledge in a given area, so it is crucial to present new material in a meaningful way that builds upon previously learned content. However, preparing students in advance for the new lesson is also important. There are several ways to facilitate learning.

Since every lesson should include an introductory part aimed at activating students' prior knowledge of the chosen topic, this can be achieved by asking questions related to the new lesson's content, with the possibility of presenting a brainstorming session (one of the teaching methods), or solving a specific example on the board where previously learned material will be applied. If the lesson involves a new area, this part should define the basic concepts to be used and provide explanations before they are employed in the instructional unit, so that students are not confused when these terms are mentioned during the lesson. Using these methods helps students feel more confident with new material, participate more actively in the lesson, and engage more comfortably in providing answers and completing tasks.

Regardless of the amount of instructional content planned for the lesson, this method of preparation for the main part of the lesson provides a foundation for interactive learning, trains the students' minds, benefiting both those with weaker prior knowledge and those with stronger knowledge in mathematics. It shifts their attention to the new instructional unit while connecting it with previous material.

In mathematics, an important part of the lesson also includes the concluding section, in which the instructor highlights the most important concepts. These main points from the lesson need to be processed by students in working memory after being learned, so they can be stored in long-term memory. This creates a prerequisite for solving tasks from that area and for completing homework assignments.

Therefore, it is crucial to invest effort in every part of the lesson to ensure it is productive and successful, enabling instructors to meet the lesson's objectives and students to achieve the outcomes set for that instructional unit.

2.2. **Cognitive Theory of Multimedia Learning**

The Cognitive Theory of Multimedia Learning is one of the cognitive learning theories introduced by American psychology professor Richard Mayer in the 1990s. This theory is based on John Sweller's Cognitive Load Theory and is specifically applied to multimedia learning, thus sharing many similarities with it. Mayer's theory posits that human working memory has two subsystems (visual and verbal/auditory) that operate in parallel, and that learning can be more effective if both data processing channels are used simultaneously.

Mayer's theory is based on three assumptions suggested by cognitive research:
The assumption of dual channels - Verbal and visual channels (similar to what Baddeley referred to as the phonological loop and the visuospatial sketchpad) in our working memory are separate and can be used to process information simultaneously, thereby enhancing the learning process. The working memory model with multiple subcomponents was first introduced by Alan Baddeley and Graham Hitch in 1974 and revised

by Baddeley in 1992. These findings were later incorporated into Allan Paivio's Dual Coding Theory and subsequently into the work of Mayer and his colleagues.

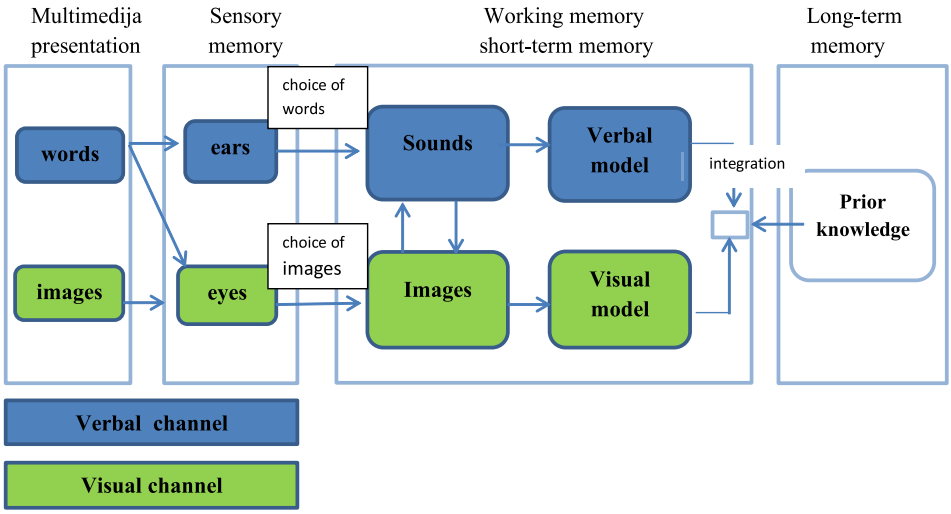


Figure 2: *Cognitive theory of multimedia learning*

1. **The assumption of limited capacity** - As Miller's information processing theory has shown, these channels have limited capacity and duration for holding information. Therefore, too much information can lead to cognitive overload.
2. **The assumption of active processing** - Learning is an active process of collecting, organizing, and integrating new information. This definition shows a similarity with constructivist learning (constructivist learning). (Mayer, 2001).

Online mathematics instruction provides us with a great opportunity to use multimedia content when presenting lessons. The platform should be designed to include not only mandatory video calls but also presentations with texts, images, animations, and video clips. Previous research indicates that combining multiple multimedia elements in order to transfer material from short-term memory to long-term memory.

For example, a trigonometry lesson can be presented more effectively through multimedia instruction compared to traditional classroom teaching. Along with the derived formulas, incorporating an image or animation from GeoGebra to show how the formula is obtained can capture students' attention. Drawing graphs of trigonometric functions will be more engaging through multimedia content. Additionally, demonstrating the practical applications of the subject area helps students understand why they are learning a particular lesson, how broadly it is applied in other fields, and in everyday life. When students see photographs and video clips of trigonometry's applications in architecture, astronomy, meteorology, economics, electronics, music, medicine, and other fields, they will get answers to the frequently asked question, “Why do we need sin(x), cos(x), tan(x), and cot(x)?”

The curriculum in our schools, including textbooks, is based on formulating lessons and creating tasks that do not integrate mathematical content with other areas, preventing students from understanding the purpose of learning the lesson. Online instruction

offers greater access to resources, making it easier to present multimedia content and applications of the material during lessons. The reason for this is that students can independently access certain resources during class, whereas in traditional classroom teaching, they only have access to what the instructor has prepared in advance.

The PISA testing conducted in our country has shown very poor results (Džumhur, 2020) because our students learn certain topics but are unable to apply them, causing the instructional content to be erased from their memory. Therefore, we emphasize the importance of students learning the practical application of each subject area, and that lessons should be presented using multimedia content that will remain in their long-term memory. This way, the application will help them recall the material they have covered. By encountering these applications daily, students will associate them with the content they have learned. Online instruction offers a greater possibility for this compared to classroom teaching, as everything is readily accessible with a single click. It is evident that the most crucial factor is the design and creation of the lesson by the instructor.

The application of modern information and communication technologies (ICT) has become inevitable in all areas of life, including the learning process. In 2006, the European Parliament issued a Recommendation on Key Competences for Lifelong Learning, which includes eight key competencies, including digital competence. The National Curriculum Framework for Preschool Education and General Compulsory and Secondary Education of the Republic of Croatia has provided for the systematic treatment of the cross-curricular theme of the Use of Information and Communication Technology through the content of all subjects. As the most advanced available teaching tool and aid, ICT contributes to the development of students' abilities for independent learning and collaboration with others, as well as their communication skills, the development of a positive attitude towards learning, improvement in how students present their work, and enhancement of their approaches to problem-solving and exploration.

The use of ICT to enhance the quality of learning is commonly referred to as e-learning by most authors (Ćukušić and Jadrić, 2012).

Given that online teaching is a new development in Bosnia and Herzegovina with the arrival of Covid-19, both for teachers and students, all participants view it as a burden, not recognizing its endless positive aspects. Looking at it from a different perspective, alongside the improvement of digital literacy and learning various programs, online mathematics instruction can be more beneficial and effective in delivering content. Teachers may require more time to plan and execute lessons, but multimedia content can simplify everything, from drawing graphs to explaining lessons and performing formulas. Many online training programs on the digitalization of teaching are available, where teachers can enhance their skills. Successfully created content can be utilized in subsequent years. We will outline principles and effects that connect cognitive theories and facilitate the creation of effective online mathematics lessons.

3. PRINCIPLES AND EFFECTS OF CONNECTING COGNITIVE LOAD THEORY AND COGNITIVE THEORY OF MULTIMEDIA LEARNING THROUGH ONLINE INSTRUCTION

By integrating cognitive load theory and cognitive theory of multimedia learning through online teaching, and utilizing the principles and effects outlined below, we develop a more functional approach to presenting mathematics instruction via multimedia content. The goal is to reduce cognitive load and enhance the processing and retention of newly acquired information. **In addition to outlining the principles and effects, we provide explanations and applications within the context of mathematics instruction**.

3.1. **Description of the principles**

These principles facilitate the creation of higher-quality multimedia content in mathematics instruction. Below, we list and present them within the framework of mathematics.

- Modality Principle: Learning will be enhanced if textual information is presented in an auditory format, rather than solely in a visual format, when accompanied by other visual information such as charts, diagrams, or animations.
  - If students only watch a presentation, it can lead to monotony, causing them to lose interest in acquiring the information. The importance of classroom teaching lies in the interaction between the instructor and the students. This can also be achieved in online teaching by using a variety of multimedia content in the presentation along with video conversations between students and the instructor.
- The Modality Principle suggests that multimedia messages are more effective when students encounter spoken words and graphics. When instructors include text on the screen, they risk occupying the students' visual channel with both images and words, where students might inadvertently direct their cognitive processes to comparing the spoken and written text.
  - For example, defining trigonometric functions on the trigonometric circle and the signs of trigonometric functions can be more effectively presented through animation along with additional explanations from the professor that accompany the animation. This way, the student learns the material more quickly than by just reading text alongside charts, animations, etc. The focus remains on the animation, allowing the student to avoid spending time reading text next to it.
- Redundancy Principle**:** The capacity of both human information channels can be overloaded by redundant information, which negatively affects the learning process.
  - This principle applies when the on-screen text and the audio narration of graphics are the same. Adding text on the screen to a narrated image can lead to cognitive overload for students, as they are processing more information simultaneously. The material presented in online mathematics instruction should be concrete and avoid lengthy explanations of concepts, as this negatively affects student interest and can make the lesson monotonous and unengaging, which is not the goal. Therefore, it is

crucial to carefully design the lesson and find a presentation method that helps students remember the instructional content more effectively. The focus is on applying theory to practical problem-solving. Mathematics instruction requires student engagement and active participation, so it is very important not to overload students with unnecessary information. When students experience working memory overload, they may have difficulty learning and understanding mathematical content.

- Split-Attention Principle: When each source of information is crucial for understanding the presented topic, learning is enhanced when multiple sources of information are presented both spatially and temporally integrated, rather than in separate formats.
  - This principle indicates that in designing a lesson, including multimedia instruction, it is important to avoid materials that require students to divide their attention between multiple sources of information. Materials should be designed so that different sources of information are physically and temporally integrated, thereby eliminating the need for students to engage in mental integration. Removing the need for mental integration of multiple sources of information reduces cognitive load. When discussing a specific part of a mathematics lesson, it is preferable to simultaneously demonstrate solving a problem with narration, present a formula, or explain an animation with the instructor's voice, so that the lesson is structured to complete the instructional unit within the planned class time.
- Spatial Contiguity Principle: Processing information is easier when two related visual sources of information are closer to each other, for example, text placed near the relevant part of a diagram results in more effective learning than if the text is placed below the diagram.
  - This principle suggests that instructors should place text (such as labels or captions) close to the graphics they describe. By doing so, they minimize the cognitive effort students must expend to align the meaning of the text with the graphics themselves. Therefore, instead of wasting time scanning the screen to find connections between the presented content, students can devote all their cognitive effort to integration and building connections. Additionally, it is beneficial for students to read the text before the animated graphics so they know where to direct their attention. It is also useful to have the animated graphics displayed in parallel with the instructor's audio explanation.
- Temporal Contiguity Principle: Simultaneous presentation of information should align with the way the human mind functions, connecting what is logically related during the presentation. This yields good experimental results, as does presenting related multimedia information with very short time intervals between them.
  - Students learn better when relevant words and images are presented simultaneously rather than sequentially. This principle dictates that narration and animation should be delivered at the same time. A good example of this is when the instructor solves a problem and simultaneously explains the steps through speech. Similarly, when presenting a specific formula, it would be more useful to immediately show the graph from which the final formula is derived. If we want to highlight the application of certain material, it is desirable to include a video or images that demonstrate the words along with the narration.

- Coherence Principle (also known as the Seductive Details Principle): States that extraneous material, although it may be interesting, is irrelevant and consumes learning resources.
  - The material we wish to convey to students through online mathematics instruction should not contain irrelevant information that will distract students' attention unnecessarily. Before presentation, the content should be carefully aligned. Since learning is an active process, all details can interfere with students' construction of mental models to represent the material. To adequately address this principle, it is necessary to use graphics, brief text (such as formulas), and the instructor's speech that support learning objectives (avoiding extraneous text, decorative images, and background music).
- Personalization Principle suggests that students learn better from multimedia presentations when the words are presented in a conversational style rather than a formal one.
  - According to the Personalization Principle, using a relaxed tone in online classes can positively impact students. Instructors should avoid rigid, academic language and instead use more approachable words to explain mathematical concepts. Informal language has the effect of activating a social response in students, such as engagement in trying to understand what the instructor has said. As a result, students will pay closer attention to the lesson and problem-solving process, which is essential for them to solve problems independently. During instruction, it is advisable to use first and second person pronouns (”I”, ”we”, ”you”, ”our”, etc.), as well as polite language (”please”, ”could you”, ”let's”, etc.).
- Voice Principle: Indicates that students learn better when narration is spoken by a human voice rather than a machine-generated one. (Mayer, 2001)
  - The importance of this principle has increased with advancements in technology. It suggests that it is better for the instructor to use spoken narration rather than recorded sound, as this helps maintain students' attention more effectively. The lesson presentation can be interrupted with a question from the instructor to check students' engagement. In online mathematics instruction, this principle is crucial for understanding instructional units, particularly during task demonstrations and when assisting students with independent problem-solving. Additionally, when displaying graphics, animations, and background sound, it is advisable to omit the use of a video of the instructor to avoid distracting from the content. Multimedia content can enhance learning outcomes when used effectively.

The presentation framework is the area within which educational e-content is displayed. This framework can encompass the entire screen or the window of an application where multimedia elements and objects such as text, graphics, images, animations, etc., are arranged. It can be argued that the multimedia principle is the starting point for all other principles, as students are more engaged when exposed to both words and images, rather than just words. Effective use of images and words simultaneously encourages generative processing. Including images should complement the explanation of mathematical content. Images, graphs, and animations enhance the meaning of spoken words, i.e., provide additional clarification of concepts. In mathematics, animations are preferred over static images.

The problem of misunderstanding mathematics through online instruction can be reduced by applying the mentioned principles. Instead of presenting students with just "dry text and numbers" on the platform, materials can be enriched with various colored charts and accompanying explanations which will arouse their curiosity. Certain topics can be demonstrated using animations, such as showing volumes of solids by pouring a liquid from one bowl to another to illustrate the relationship between the volumes of two bodies. The surfaces of figures, presented alongside formulas and standard problems, can be shown practically, for example, how a specific area is tiled, including its dimensions, etc. In this way, students learn the practical application of mathematical content, allowing the material to be stored in long-term memory by connecting it to its application.

3.2. **Effects**

Effects indicate how certain ways of creating multimedia content in mathematics instruction can influence student attention. They serve as guidelines for active student participation in online mathematics classes. Below, we provide effects and examples of their application in mathematics.

- Signaling Effect: Refers to the enhancement of learning outcomes by directing attention to relevant information. Signals are based on natural attention grabbers such as motion. In multimedia, this can be achieved through underlining, arrows, or color coding.
  - During the presentation of new mathematical content, students are faced with a plethora of information that can sometimes be difficult to connect. What will help them focus on the most important parts is the use of signaling. Signaling in mathematics can be achieved by underlining key concepts, framing formulas, connecting graphs and task steps with arrows, highlighting each part of a task with explanations, etc. This way, students will know what to pay attention to and how to integrate information to build their own mental models. However, moderation is key. Too much signaling can be confusing for students.
- Segmentation Effect: Learning could be more effective if continuous animation or narration is broken down into smaller segments.
  - Students learn better when multimedia messages are presented in segments or "microparts" that are tailored to the specific lesson and student age group, rather than as a continuous whole. Once they grasp a smaller part, they can more easily focus on a new segment by connecting it to what they have already learned. Additionally, it is easier for instructors to review a specific part due to misunderstandings rather than the entire content. There are two ways to implement the segmentation effect. The first is to insert pauses between segments, which gives students time to perform necessary cognitive processes. The second way is to divide animations into meaningful segments, which provides students with learning support and makes it easier for them to master the material by displaying animations in partitions with specific time frames. This approach helps students understand the procedures shown in the animation. The segmentation process has a positive effect on memory and application of the material.

- Effect of Worked Examples: Reduces cognitive load by providing detailed demonstrations of how to solve a task or problem.
  - It is well known that demonstration is a crucial method in mathematics when solving tasks. The instructor's presentation of examples shows how to use formulas. Sometimes students find theory or formula derivation unclear because the material in mathematics is interconnected. Therefore, if students have not previously mastered or do not remember some part of the material used in a new mathematical unit, they can more easily recall it through examples. The demonstration method is the best for mastering lessons in mathematics, both in classroom instruction and in online teaching. The instructor can use a smart board or tablets, which make it much easier to write, erase, add signaling marks, and change colors compared to a traditional teaching. Additionally, if a student does not understand a part of the lesson and asks at the end of the class, the instructor can go back to the previous screen where the task was done, whereas in a classroom setting, the example would need to be redone.
- Reverse Effect: Teaching techniques that are highly effective with weaker students can have the opposite effect, meaning negative consequences, when used with stronger students. For these students, explaining only the steps or providing hints during task completion has a more positive impact on conceptual understanding.
  - This effect refers to the reversal of the effectiveness of teaching techniques for students with different levels of prior knowledge. The primary recommendation arising from the reverse effect is that instructional design methods must be adapted to the students' knowledge acquisition in a given area. The goal of learning is to construct integrated mental representations of relevant information, which requires significant working memory resources. To solve a task without overloading working memory, some form of guidance or direction is necessary. Teaching techniques that help students create schemas in long-term memory are more useful for beginners or students with low levels of prior knowledge. Conversely, for students with higher levels of knowledge, or those with more prior knowledge of the material, the opposite is true. Thus, the same instructions for solving a task may not have the same effect on all students. One reason is that they may already have their own schemas for solving tasks, so instructions from the instructor might confuse them. It would be beneficial to use a variety of methods and approaches based on students' prior knowledge. Students with low levels of knowledge lack schema-based understanding in the targeted domain and therefore need instructions to support them in reducing cognitive load when tackling new tasks. Otherwise, their working memory will be overloaded. In contrast, students with higher levels of knowledge enter task-solving with pre-existing schemas. Providing them with additional instructions may result in processing redundant information, thereby increasing cognitive load. They need to integrate components from long-term memory with the additional instructions. Such integration processes burden working memory. In this case, giving additional instructions becomes unnecessary.
- Effect of Collective Working Memory: When the complexity of the material being learned is low, individual learning is more effective than collaborative learning. For

complex materials, collaborative learning is more successful as it allows for the distribution of working memory load among participants (Schneider et al., 2018).

- Group work in mathematics has proven to be effective for solving tasks. Participants can divide the tasks among themselves and each work on solving their own example, then present their task to the group. All group members explain the process they used. This way, working memory facilitates more productive processing and easier understanding of the tasks. This principle is useful for more complex materials and tasks. However, for simpler tasks, individual learning is preferred, as it is essential for everyone to master the basics of a particular area. This has been experimentally confirmed as well as observed in practice. In online mathematics instruction, dividing students into separate virtual "rooms" and then bringing them back to a common video chat to demonstrate their solutions could be implemented. The groups should consist of students with varying levels of mathematical knowledge.

Regarding the effects and their impact through online instruction, we can conclude that it is important to present the material to students in an engaging manner. For example, definitions and theorems should be written in red letters, important terms should be underlined, and so on. Utilizing animations, charts, and incorporating all elements that can capture students' attention and help retain the material in long-term memory is essential. What is presented through animations and videos should be shown in segments with specific pauses between each segment to facilitate easier processing of the acquired information.

In mathematics, the application of the demonstration method is crucial. This involves solving tasks with step-by-step instructions and detailed explanations to help students better grasp a particular lesson. In subsequent examples, it is beneficial to provide only verbal instructions to encourage students to engage in the work. The best way to assess the outcomes of a lesson is through students' independent problem-solving. During this process, they utilize everything retained in their memory from the instructor's presentation. Different methods should be used for students with varying levels of prior mathematical knowledge. When introducing new material and tasks from a new area, group work also enhances learning. Students should be divided into groups, each receiving specific tasks, with each group member responsible for a particular part of the assignment. This approach ensures that all students actively participate in mastering the material, assisting each other in reaching solutions. It encourages them to explore on their own and simplify the material, which they then present to the other students. This method yields good results both in traditional classrooms and in online instruction.

4. Conclusion

Creating high-quality educational multimedia content for mathematics instruction requires disciplined, educated teachers, instructors, and professors who invest time in the creative preparation of lessons. It is essential for the teacher to be familiar with the principles and effects of multimedia content design, grounded in cognitive psychology, to ensure the content is both high-quality and effective. The goal of well-structured multimedia mathematical content is to help recipients understand, comprehend, and re-

member a particular area as effectively as possible. This can be achieved by applying principles that guide us in creating as good as possible materials for mathematics instruction in primary and secondary schools. Therefore, the use of these principles and effects greatly facilitates the creation of online mathematics content.

A significant impact is achieved by connecting mathematics instruction with other areas and everyday life, which helps students process the lesson more easily in working memory and transfer it to long-term memory. Memorization causes cognitive overload for students, making it difficult for them to meet the demands placed on them. The main task of teachers and instructors is to simplify the material for online instruction, break it into micro-parts for easier student assimilation, and meaningfully present the content with the help of the mentioned effects and examples within the lesson, to facilitate the construction of new knowledge in mathematics. This approach will increase students' interest and make them more engaged participants in the learning process.

We have observed that working memory is quite overloaded, which slows down the acquisition, processing, and retention of material. Therefore, it is essential to enhance its function by incorporating lectures in the introductory part of the lesson, where the teacher introduces students to the topic to be covered, boosts their confidence, and improves their concentration. Whether through quizzes where students compete with each other in answering questions, providing examples, or defining new concepts, the introductory part of the lesson plays a key role in motivating students. The purpose of this section is to increase individual student participation in the class, strengthen confidence, facilitate interactive learning, and improve success. This has a very positive effect as it enhances understanding of a specific mathematics lesson, which motivates students to explore further.

A multimedia introductory part of the lesson improves conceptual understanding of fundamental concepts, especially the more abstract ones that require visualization. By using high-quality online materials, students approach learning more seriously with a higher level of interest, as they feel more prepared for the task, which helps them master the material in a shorter period of time. Information must be presented clearly with all necessary explanations so that students can see the purpose of learning the specific lesson.

Applying the principles and effects of creating multimedia content in mathematics lessons, we found that students were more engaged, actively participated in the class, and responded positively to the assigned tasks. All students completed their homework, presented it in the next class, and collaboratively created posters where they visually demonstrated the application of a specific instructional unit. When creating multimedia mathematical content, it is important to apply each principle judiciously. Excessive reliance on any single principle can lead to student disengagement.

REFERENCES

[1] Sweller, J., and Chandler, P. (1991). Evidence for cognitive load theory. *Cognition and Instruction*, *8*(4), 351-362.

[2] Džumhur, Ž. (2020). *PISA 2018.* Report for Bosnia and Herzegovina. Agency for Preschool, Primary, and Secondary Education, Tuzla.

[3] Dindia, L. (2013). Pre-lecture activities in undergraduate science courses. *Teaching Innovation Projects, 3*(1), 1-8.
[4] Ćukušić, M., and Jadrić, M. (2012). *E-Learning: Concept and Application*. Zagreb: School's book.
[5] Mateljan, V., Širanović, Ž., and Šimović, V. (2009). Proposal for a Model for Designing Multimedia Web-Based Educational Content According to Pedagogical Practice in Croatia. *Informatologia, 42*(1), 38- 44.
[6] Mayer, R.E.. (2001). *Multimedia Learning (2nd ed.)*. Cambridge University Press, Cambridge.
[7] Mishra, S. and Sharma, R. C. (2005). *Interactive multimedia in education and training*. Idea Group Publishing.
[8] Schwartz, R.N., Milne, C., Homer, B.D., and Plass, J.L. (2013). Designing and implementing effective animations and simulations for chemistry learning. u J.P. Suits and M.J. Sanger (Eds.) *Pedagogic Roles of Animations and Simulations in Chemistry Courses*, pp. 43-76. Washington, DC: American Chemical Society.
[9] Schneider, S., Beege, M., Nebel, S., and Rey, G.D. (2018). A meta-analysis of how signaling affects learning with media. *Educational Research Review, 23*, 1-24.
[10] Širanović, Ž. (2012). *Model for designing multimedia educational web content*. Doctoral Dissertation, University in Zagreb.
[11] Cognitive Load Theory. (2020) Downloaded 24.12.2020 from `https://www.learning-theories.org/doku.php?id=hr:learning_theories:cognitive_load_theory`
[12] John Sweller's Cognitive Load Theory. (2018.). Downloaded 25.12.2020. from `https://hr.sainte-anastasie.org/articles/psicologia/la-teora-de-la-carga-cognitiva-de-john-sweller.html`

(Received: June 30, 2024)
(Revised: August 20, 2024)

Azra Hadžiomerović
Mostar Gymnasium
Bulevar Meše Selimovića 59
71000 Sarajevo
Bosnia and Herzegovina
e-mail: *azrahagi@gmail.com*

PROCEEDINGS OF THE CONFERENCE ON MARCH 14 – INTERNATIONAL DAY OF MATHEMATICS
Special Editions ANUBiH, Book CCXVI, OPMN Book 30
Sarajevo 2024, **p. 187–196** DOI: 10.5644/PI2024.216.18

CONSTRUCTIVIST APPROACH TO EDUCATION WITH REFERENCE TO CONSTRUCTIVISM IN THE TEACHING OF MATHEMATICS

AMRA ALIKADIĆ FAZLIĆ

Dedicated to 75th years of life and 50 years of scientific work of Academician Mirjana Vuković

ABSTRACT. Constructivism as a philosophy of education can be said to be based on the work of Giambattista Vico , the philosopher of the 18th century who believed that people only understand what they themselves make. Many philosophers and educators have worked on this idea, but the ideas behind constructivism were developed by Jean Piaget and John Dewey. Today, when rapid changes are taking place in all countries with the goal of improving education, teachers show great interest in education with a constructivist approach, because constructivism is an approach by which we learn new things we encounter by connecting them and putting them in certain relationships with already acquired knowledge. The Ministry of Education of our country should start with radical changes and harmonize the changes with the most common approach in new programs in the world, which is precisely the constructivist one. The impact of this fundamental change in the understanding of education seems inevitable for the education and upbringing of both students and teachers in our country.

1. THEORETICAL FRAMEWORK

Constructivism has its roots in philosophy, and was applied in sociology and anthropology, as well as in cognitive psychology and education. In literature, constructivism is treated as a scientific theory and a theory of cognition, but also as a theoretical paradigm (in sociology, cognitive science and psychology), that is, as an image of man and as a didactic position or learning strategy. Constructivism, from the point of view of learning theory, is an approach that tries to explain how people learn, and from a philosophical point of view, it is a term that is related to epistemology . A lot has been said about the constructivist approach in education in recent years, and it is quite old. Giambattista Vico (1668 - 1744), Jean Jacques Rousseau (1712 - 1778) and Immanuel Kant (1724 - 1804) are considered representatives of constructivism in the past centuries.

Giambattista Vico, a thinker of the 18th century, gave constructivism the modern "look" it has today. With the work, "De antiquissima Italorum sapientia ", published in 1710, he opened a new perspective to epistemology, defended the point of view that people understand only what they themselves have built and said that "a person understands something only when he can explain it". [8]

2020 *Mathematics Subject Classification.* 16W20.
Key words and phrases. constructivism, education, mathematics, student, teacher.

Immanuel Kant advocated similar views and emphasized that people are not passive recipients of knowledge. According to him, they receive it actively, connect with previous knowledge, and via adding their own thoughts, adopt it. [15] According to Rousseau, the most suitable program for children is one that opens the door to the child's natural curiosity and desire for learning and knowledge. The curriculum should not be imposed by adults, but should have the student at the center and reflect the child's interests and occupations. [16]

Constructivism can be linked to the clarification of the very nature of knowledge. It is believed that the learner should build knowledge for himself, because each person individually and socially acquires his knowledge. We can take two things from this:

1. Teachers should not concentrate on the topic or the lesson, but on the individual thinking about his task (learning);

2. There is no knowledge independent of the knowledge created from the experience of the learner. [9]

According to Von Glasersfeld [17], knowledge, no matter how we define it, is in a person's head and is shaped based on and related to his personal experiences. The constructivist approach implies that language learners should form the meanings of phrases, sentences and texts individually. [14] [7]

We can say the following about the basic principles of this approach: Students add meaning to the new concepts they encounter only within their existing understandings. Therefore, this is an active process in which students connect their existing knowledge with new ideas and create new knowledge [10]. According to the more education-oriented definition of constructivism, the association of knowledge is closely related to experience. Students come to class with their own experience and thought structures formed by that experience. These previously created constructs can be good, bad or incomplete. The student reorganizes his thought structures creating a connection between the new knowledge and experience and the old one. In order for new knowledge to be useful and complete part of the student's memory, conclusions, details and relationships between older understandings and new ideas must be drawn and created by the student himself. Otherwise, new, memorized information, unrelated to previous experiences, will be forgotten very quickly. In short, in order for learning to be with understanding, the student must add new knowledge to existing material in an active way.

Even though many 20th century scientists worked on the application of constructivism in the classroom and in the development of children, the following stand out: Piaget, Dewey, Bruner, and Vygotsky.

Constructivism of Piaget relies on cognitivism. In one of his works, he requires teachers to take into account the stages in the development of a child's mind. He believes that the basis of understanding is discovery. "To understand is to discover or rediscover in order to build again. In creating creative future members of society, building knowledge plays an important role." Children, in their free time, in classes that provide them with opportunities for activities that interest them, should discover connections and ideas. Understanding is built, step by step, through active participation. In this way, Piaget says that knowledge is not a "passive copy of reality", but that it is a construction that an individual manages to build over time. [11] [12]

For Dewey, education is about activity. Knowledge and ideas are born from situations that are significant for the learner and from which he can draw significant experiences. These situations occur in settings such as the classroom where students who master the materials form a community that learns together by building their knowledge. [3]

Another important name for constructivism is Lev Vygotsky. While some critics argue that he is not a constructivist because he insists too much on the importance of the social environment in learning, others believe that he actually insists that children as builders should create their own views. He believes that children learn scientific concepts as a result of comparing their own views with those of adults. A child can only memorize a concept that has just been brought to him from the world of adults. In order to be able to turn it into his product, the child must use the connection between the concept and the idea presented to him.

Bruner also has views that shed light on the constructivist approach. For him, learning is a social process in which students can apply new concepts to existing knowledge. The student, with the aim of combining the new experience with the existing mental constructions, chooses information/knowledge, creates a hypothesis and makes a decision. A sense of independence, which is developed by encouraging students to learn new elements on their own, is the core of effective learning. In addition to this, educational programs should be designed in the form of a spiral structure that allows students to build on newly acquired knowledge.

The principles that briefly characterize Bruner's theory are:

- *Education should support experiences that will bring students to a state of interest and openness to work.*
- *Education should be built in a way that students can easily understand (spiral constructivism).*
- *Education should be created in such a way that it facilitates the use of acquired knowledge in different situations.*

In addition to the ones mentioned above, modern representatives of constructivism include: Ernst von Glasersfeld (1917), Heinz von Foerster (1911-2002), Paul Watzlawick (1921), Francisco J. Varela (1946-2001) and Humberto R. Maturana (1928).

Only a few works on constructivism were published in the territory of the former Yugoslavia until 2020: Miljak, 1998, Mušanović, 2000, Palekčić, 1999, 2001, 2002, Babić/Irović, 2001, Krstović J. 2001, Gojkov G. in 2002, and in the field of psychology, Dušan Stojnov's book "From psychology of personality to psychology of persons" appeared in 2005.

In recent years, more works have appeared on this topic.

There are different types of constructivism: radical constructivism, moderate constructivism, operational constructivism, methodical constructivism, new constructivism (in pedagogical psychology) and others. The reception of constructivism in pedagogy, i.e. didactics, is more recent, especially in Bosnia and Herzegovina.

Constructivism is one of the theories of learning and teaching that has had the greatest impact on education in practice. One of the most important reasons for this is the search for a solution to serious qualitative problems in education. Research shows that

students from developed countries such as the USA and Germany, especially when it comes to reading comprehension , and success in mathematics and physics, give poorer results than children from developing countries (Pisa-Schock, 2002). Again, research shows that even the most successful children on standardized tests cannot show the same success when their knowledge needs to be compared, integrated or used in everyday practical life [18]. Today, when all countries are busy looking for new changes that will bring education out of this situation and improve the quality of education provided in schools, teachers, primarily those from developed countries, show a great interest in the understanding of education that relies on a constructivist approach [13]. The reason for this is that education based on constructivism has the educational goal for the student to learn. In short, constructivism implies receiving new knowledge by putting it in relation with previous knowledge and thus creating new knowledge related to the already existing one.

Students should be able to apply what they have learned in school in different and unexpected situations in their lives. Classical education, where the teacher is the one who transmits knowledge, and the student is tied to the book, has proven to be extremely unsuccessful in raising students who think, criticize, comment, and interpret. If so, the focus of the classroom should be shifted from the dominance of the teacher and bring the student to the center with a constructivist approach [8].

Although constructivism was talked about throughout the 20th century, it only became relevant at the end of that century. One of the reasons is brain research, the number of which increased sharply in the 1990s. The results obtained from research in the field of neurophysiology have interested experts who deal with education, and they have tried to use this knowledge in organizing the learning and teaching process, i.e. of the education process. Constructivism is one of the concepts that stood out the most during these attempts. Although constructivism was among the topics explored much earlier by philosophy and psychology, constructivism in mathematics and science programs and education has been attracting attention since the 1990s.

Açıkgöz [1] believes that the term constructivism began to be used simultaneously with the term that has been used very often in recent years, ”active learning”. The theoretical foundations of active learning are based on constructivism and cognitivism.

2. Problem and subject of research

If we talk about schooling processes, two terms are necessarily imposed on us: ”traditional” and ”modern school”. The traditional school, which was connected to the industrial society until the end of the Second World War, is shown in a simplified form: school year - subject - class - lesson. The foundation, the key of a traditional school is the ”amount of material adopted”, which is measured by the degree of state usefulness. After the Second World War, there were intense scientific and technological changes, including a change in the school process. Learning becomes a communication process, and in the new ”modern school” the important terms ”quality school” and ”quality of knowledge” are mentioned. With the collapse of communism and as a result of changes in the civil society itself, a postmodern school or an innovative modern school appears

on the scene, in which, in addition to the technological moment (specific to the modern school), there is also a moral - civility.

School is not what it used to be. It does not lose its importance, but simply changes and becomes different. The school is becoming more and more a center of education, and less an institution of knowledge transmission or a communication link between student and source of knowledge. At school age, students not only acquire a general education in literacy and traditional subjects covered by knowledge, but also have the opportunity to learn using mass media.

If we observe the changes taking place in education, we notice changes at the global and lower school level. The attention of modern society is concentrated on the lower level, where children learn (schools, family, mass media, institutions and programs of free time, etc.).

These changes are followed by regional, national and international research . PISA (Programme for International Student Assessment) is an international test that tries to answer the question of how effectively schools prepare students for the challenges of the future. This program assesses the extent to which 15-year-old students have thoroughly mastered the knowledge and skills essential for participating in social life. OECD member countries (Organization for Economic Cooperation and Development) participate in the research and other countries can also access.

PISA measures knowledge and competences that are important in the workplace and in the private life of individuals, and which are also important for society at the same time. The purpose of the research is to assess the readiness of 15-year-old students for the challenges of life in today's society. Data collection in the PISA test is done in cycles of three years. Data were collected for the first time in 2000, then in 2003. The PISA survey in 2006 included students from 57 countries. The cyclical collection of data in PISA enables effective research into trends in student achievement and the development of educational system reforms. The Agency for Standards of Bosnia and Herzegovina reported on the conducted PISA survey in 2019, but abandoned the same in 2022, when 85 countries participated. Many countries in Europe, especially Germany, initially improved their PISA results, but in recent years the results have deteriorated. Those countries have a large amount of research and are trying to use the results of the research to improve education.

Testing is a function of the efforts of all nations to develop their human capital, which the OECD defines as: "knowledge, skills, abilities and other qualities embodied in individuals that are important for personal, social and economic well-being". This research examines achievements in three areas: natural sciences, mathematics, and comprehension and interpretation of texts.

The school in our country is traditional and we have a long way to go towards establishing a modern school. That path is even more difficult if we remember that our country is underdeveloped, with low productivity, with a low technological base, a large "brain drain" and that it does not provide the conditions for the rapid development of a modern school. Furthermore, traditional schooling cannot satisfactorily address individual differences among students and their developmental needs. [5]

In a traditional school, teaching is directed towards the average student and its scheme is lecture and memorization of teaching material. The modern school is primarily different in terms of learning environment. In it, the ideas of freedom come to the fore, manifested through the autonomy and individuality of students, plurality and tolerance. The teacher changes roles and becomes a student leader.

Modernization of the teaching process by introducing a constructivist approach, in which the teaching process is interactive and the role of the student is emphasized, is a necessary consequence of new knowledge about the nature of learning during the last decades.

For the success of teaching, it is important to realize the interest and activity of students in the teaching process. Constructivist learning offers a bold departure from the traditional, objectivist classroom. The goal for the learner is to play an active role in assimilating knowledge into their existing mindset. The ability of students to apply their knowledge, learned in school, in the real world is valued more than memorization bits and pieces of knowledge that may seem unrelated. Constructivist teaching requires the teacher to relinquish his role as information dispenser and instead continuously analyze his curriculum and instructional methodology. It is probably best for the constructivist teacher to have ""instantaneous and intuitive vision of the pupil's mind as it gropes and fumbles to grasp a new idea" [2]. Of course, the constructivist view opens up new approaches to learning, as well as challenges for teachers trying to design it.

Some ideas for teachers to implement constructivism are suggested by Yager [18]:

- *seek and use students' questions and ideas to guide the lesson and the entire unit;*
- *accept and encourage students to present new ideas;*
- *promote student leaders, collaboration, locating information and performing actions as a result of the learning process;*
- *use the student's thinking, experience and interests during the presentation of the lesson;*
- *encourage the use of alternative sources of information;*
- *ask students to present their ideas before presenting the teacher's ideas or ideas from books, etc.;*
- *encourage students to challenge the conceptualizations and ideas of others;*
- *use and respect all ideas presented by students;*
- *encourage self-analysis , evidence that supports ideas and reformulation of ideas in the light of new knowledge;*
- *use student problem identifications;*
- *use local resources as original sources of information that can be used in solving problems,*
- *engage students in searching for information that can be used to solve real-life problems;*
- *extend learning beyond class, class and school time;*
- *focus on the impact of science on each individual student;*
- *refrain from seeing scientific facts as something only students need to master in order to pass tests.*

How students learn is more important than how the teacher teaches is the constructivist didactic credo.

CT Fosnot [6], putting the student in focus, defines constructivist learning as an active process that:

- *implies student independence;*
- *the focus is more on learning than teaching ;*
- *the student has the will to learn;*
- *the student intends to learn;*
- *learning significantly depends on prior knowledge;*
- *the student owns his beliefs, attitudes and knowledge, and new ideas are developed in the process of adaptation and change of old ideas;*
- *learning is a process of creating new ideas, not a mechanical accumulation of data;*
- *motivation is the key to quality learning and a student motivated to learn is ready to explore, possess curiosity and initiative.*

It can be concluded that the constructivist model of learning includes the research activity of students and the development of specific socio-educational communication.

As a researcher, I focused on individualized teaching of mathematics, because it provides breadth in the methodological sense, and at the same time respects the individual capabilities of students. It seems that the constructivist approach to learning is mostly achieved by students in high classes.

Considering greater international success of some of these mathematics major high school students, it is worth investigating the way they learn. That research, in its content and results, would be significant for improving the quality of educational work with children of high school age, and at the same time a valuable contribution to the methodology of teaching mathematics.

There is almost no such research conducted in Bosnia and Herzegovina. Considering the development of modern technologies, and the possibility to monitor research in Bosnia and Herzegovina and surrounding countries and in the world, it is necessary to investigate every segment that is important for the development of divergent thinking, and which is in direct relation to the creativity of students. In connection with what was said about constructivist learning, there is a need to also in our environment explore this learning model.

Here I have in mind the thought of the management guru Peter Drucker, who speaking about the ”age of knowledge” says that *the creativity and talent of employees is the basic economic resource that replaces the former capital.* [4]

If we look back at the goals of mathematics teaching, it is suggested that the way to their realization leads through a constructivist approach to learning and teaching .

The goal of teaching mathematics in elementary school is to ensure that all students acquire basic linguistic and mathematical literacy and progress towards the realization of appropriate standards of educational achievement, as well as to:

- *Enable students to solve problems and tasks in new, unfamiliar situations;*
- *Enable students to express and justify their opinion and discuss with others;*
- *Develop creative and critical thinking;*

- *They develop motivation for learning and interest in subject contents;*
- *It ensures that students acquire elementary mathematical knowledge that is necessary for understanding phenomena and laws in nature and society;*
- *Train students to apply acquired mathematical knowledge in solving various tasks in real life situations;*
- *It represents the basis for successful continuation of mathematical education and for self-education;*
- *It contributes to the development of mental abilities, the formation of a scientific view of the world and the all-round development of the student's personality.*

3. SIGNIFICANCE OF RESEARCH

Research results should have both social, pedagogical and methodological significance, which would contribute to pedagogical practice and theory.

The social importance of research stems from the very importance of mathematical education for the overall development of an individual. When we talk about the upbringing and education of high school students, there is no doubt that the teaching of mathematics as a subject in which upbringing and education are realized with mathematical content occupies an important place. Through the teaching of mathematics, students acquire the knowledge needed for everyday life, as well as the knowledge needed for professional education and performing many activities. Bosnia and Herzegovina has a long tradition of institutional high school education, but there is very little scientific research in this area.

Pedagogical significance of research is reflected in the search for an answer to the question, what are the effects of the constructivist approach to learning mathematical content, among high school students. The effects of the constructivist approach to the learning of mathematical content should be reflected in the shift in the field of mathematics teaching methodology and definitely help the student in our schools to find himself in focus and become a more active participant in the teaching process than is the case today. The results of the research should contribute to the improvement of educational work with students of high school age and be at least a small step in the transition from a "traditional" to a "modern" school.

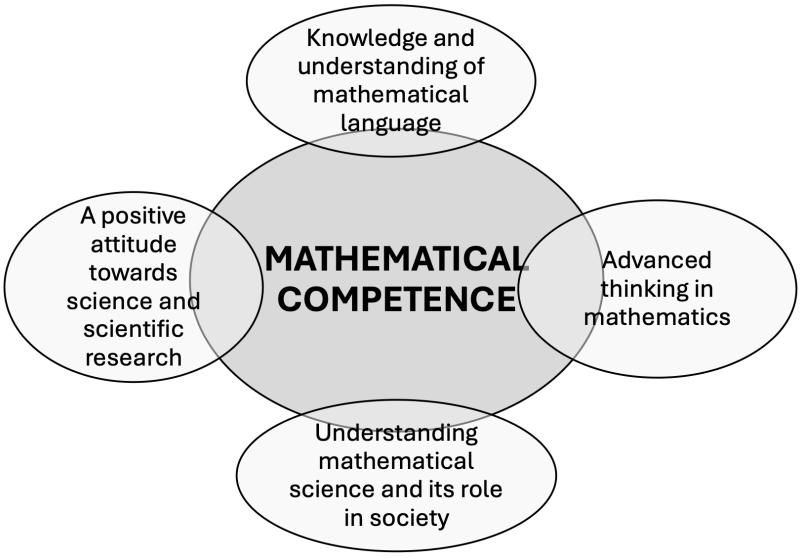
4. THE AIM OF THE RESEARCH

As one of the basic tasks of teaching and learning at school is the acquisition and assimilation of knowledge, a number of requirements are set for the methodology of all subjects, and especially for the methodology of teaching mathematics. They are aimed at finding interesting ways of presenting the material, adequate teaching aids, learning models, and all with the aim of active participation of students in the teaching process, and the acquisition of quality, permanent knowledge that can be used in everyday life.

The main goal of the research would be to examine, analyze and determine the effects of the constructivist approach to learning on the mastery of teaching content in the field of mathematics.

We propose the following schematic representation of how to achieve mathematical competence with a constructivist learning model:

5. FIGURES AND TABLES



6. CONCLUSION

= The fact is that the high school curriculum in Bosnia and Herzegovina is based in a way that requires the teacher to convey as much information as possible to the students. The question arises as to how far teachers are able to go a step further and separate the important from the less important through autonomous reasoning, help children to search for the essence and enable children to develop into people who are eager for knowledge equipped with the tools to obtain knowledge. And a good way of teaching mathematics should contribute to that. Many teachers have not yet dared to do so, but there are quite a few who have bravely stepped onto the path of constructivist learning. The traditional way of teaching will still be present, but if we include elements of constructivism in it, it will not continue to be so rigid and focused only on the content that needs to be processed.

The effects of the constructivist approach to the learning of mathematical content should be reflected in the shift in the field of mathematics teaching methodology and definitely help the student in our schools to find himself in focus and become a more active participant in the teaching process than is the case today.

REFERENCES

[1] K. Açıkgöz, *Effective learning and teaching*, Education World Publications, Izmir, 2003.
[2] J. Brooks, M. Brooks, *The cases for the constructivist classrooms*, Alexandria,Va., ASCD,1993.
[3] J. Dewey, *Democracy und Erziehung (Eine Einleitung die philosophische Padagogik)* , Beltz Tascehnbuch , Weinheim , 2004.

[4] P. Drucker, *The New Realities*, Transaction Publishers , New Jersey (reprint), 2003.
[5] A. Fazlić-Alikadić, *Educational values of individualized basic mathematics in the lower grades of elementary school*, Naša škola no. 25/03
[6] C.T. Fosnot, *Constructivism: Theory, Perspectives and Practice* Teachers College Press, New York, 1996.
[7] Foerster, Glasersfeld, Hejl, Schmidt, Watzlawick, *Einfuhrung in den Konsrtruktivismus*, Munich, 1998.
[8] S. Hanley, *On constructivism*, Maryland Collaborative for Teacher Preparation, The University of Maryland at College Park, 2005.
[9] G.E. Hein, *Constructivist Learning Theory*, CECA (International Committee of Museum Educators) Conference, Jerusalem Israel, 1991.
[10] S. Naylor, B. Keogh *Constructivism in the Classroom: Theory into Practice*, Journal of Science Teacher Education, 1999.
[11] J. Piaget, *Part I: Cognitive development in children: Piaget development and learning*, Journal Research in Science Teaching, 2(3), 176–186, 1964.
[12] J. Piaget, *The theory of stages in cognitive development* In D. R. Green, M. P. Ford, & G. B. Flamer, Measurement and Piaget, McGraw-Hill, 1971.
[13] A. G. Powell, E. Farrar, David K. Cohen, *The Shopping Mall High School: Winners and Losers in the Educational Marketplace*, National Association of Secondary School Principals (U.S.), National Association of Independent Schools. Commission on Educational Issues Houghton Mifflin, 1985.
[14] W. A. Suchting, *Constructivism Deconstructed*, Book chapter in Constructivism in Science Education, Dordrecht , The Netherlands, 1998.
[15] F. Thilly, *Felsefenin Öyküsü*, Çağdaş Felsefe , İzdüşüm Yayınları, Istanbul, 2002.
[16] J.A. Vadeboncouer, *Child Development and the Purpose of Education : A Historical Context for Constructivism in Teacher Education* , 1997.
[17] E. Von Glasersfeld, *Radical Constructivism : A Way of Knowing and Learning*, Falmer Press, London, 1995.
[18] R. Yager, *The Constructivist Learning Model: Towards Real Reform in Science Education*, The Science Teacher 58, No: 6, 1991.

(Received: 29 March, 2024)
(Revised: 15 September, 2024)

Amra Alikadić Fazlić
e-mail: *amraaf@gmail.com*