[И Ш Н И К
НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“

Факултет по математика и информатика

ANNUAL
OF SOFIA UNIVERSITY "ST. KLIMENT OHRIDSKI"

Faculty of Mathematics and Informatics

СОФИЯ 2017 SOFIA 2017

ТОМ/VOLUME 104

УНИВЕРСИТЕТСКО ИЗДАТЕЛСТВО „СВ. КЛИМЕНТ ОХРИДСКИ“
ST. KLIMENT OHRIDSKI UNIVERSITY PRESS

ГОДИШНИК

НА

СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА

Том 104
2017

ANNUAL

OF

SOFIA UNIVERSITY
"ST. KLIMENT OHRIDSKI"

FACULTY OF MATHEMATICS AND INFORMATICS

Volume 104
2017

СОФИЯ • 2017 • SOFIA
УНИВЕРСИТЕТСКО ИЗДАТЕЛСТВО „СВ. КЛИМЕНТ ОХРИДСКИ“
"ST. KLIMENT OHRIDSKI" UNIVERSITY PRESS

Annual of Sofia University "St. Kliment Ohridski"
Faculty of Mathematics and Informatics

Годишник на Софийския университет „Св. Климент Охридски"

Факултет по математика и информатика

Address for correspondence:

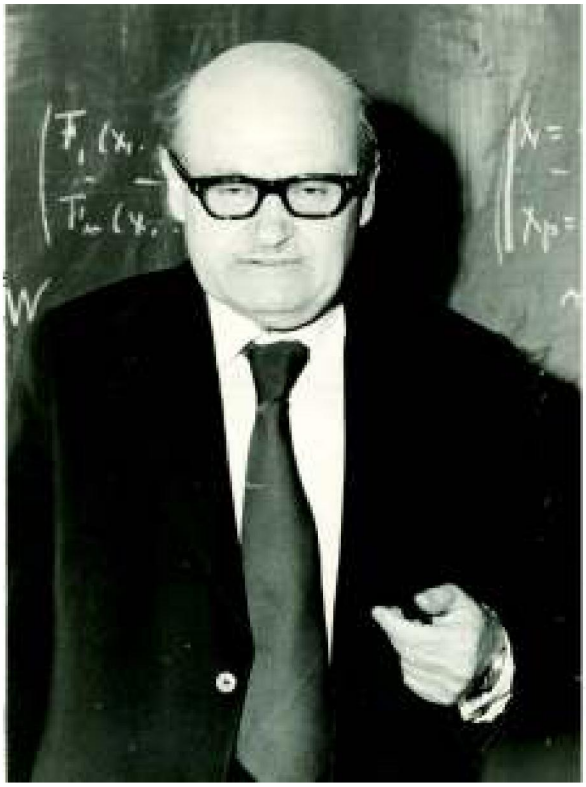
Faculty of Mathematics and Informatics
"St. Kliment Ohridski" University of Sofia
5, J. Bourchier Blvd., P.O. Box 48
BG-1164 Sofia, Bulgaria

Fax xx(359 2) 8687 180
Electronic mail:
annuaire@fmi.uni-sofia.bg

Aims and Scope. The *Annual* is the oldest Bulgarian journal, founded in 1904, devoted to pure and applied mathematics, mechanics and computer science. It is reviewed by *Zentralblatt für Mathematik*, *Mathematical Reviews* and the Russian *Referativnii Jurnal*. The *Annual* publishes significant and original research papers of authors both from Bulgaria and abroad in some selected areas that comply with the traditional scientific interests of the Faculty of Mathematics and Informatics at the "St. Kliment Ohridski" University of Sofia, i.e., algebra, geometry and topology, analysis, mathematical logic, theory of approximations, numerical methods, computer science, classical, fluid and solid mechanics, and their fundamental applications.

ISSN 0205-0808

CONTENTS



Yaroslav Tagamlitzki (1917–1983)

Editorial Note

In 2017 the mathematical community in Bulgaria celebrated one century from the birth of Professor Yaroslav Tagamlitzki. On this occasion, Faculty of Mathematics and Informatics of Sofia University "St. Kliment Ohridski", in cooperation with Institute of Mathematics and Informatics of Bulgarian Academy of Sciences organized a Jubilee Conference "100 years from the birth of Professor Jaroslav Tagamlitzki". The conference, which took place in the Department of Mathematics and Informatics of Sofia University during the period of 15–17 September, 2017, provided a nice opportunity for colleagues, friends and students to commemorate the distinguished Bulgarian mathematician, dedicated teacher and remarkable person Yaroslav Tagamlitzki. Most of the papers in this volume of the Annual are based on talks, given by participants in the conference.

Geno Nikolov

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 104

ANNUAL OF SOFIA UNIVERSITY „ST. KLIMENT OHRIDSKI“
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 104

100 YEARS FROM THE BIRTH OF YAROSLAV TAGAMLITZKI *

DIMITER SKORDEV

The life and the deed of the eminent Bulgarian mathematician Yaroslav Tagamlitzki (1917–1983) are considered.

Keywords: analysis, biography, convex, deed, distinguished, education, eminent, generalization, life, mathematician, pedagogical, publications, research, scientific, topological.

2000 Math. Subject Classification: 01A60, 01A70.

About a century ago, less than two months before the October Revolution, a Russian family acquired a son whom fate has ordered to become after several decades one of the distinguished Bulgarian mathematicians (when mentioning fate, the consequences of the revolution are also taken into account). The child was born in the southern Russian city of Armavir on September 11, 1917.[1] The father Eng. Alexander Mihailovich Tagamlitzki (1881–1945) and the mother Vera Leonidovna (1894–1976) chose the name Yaroslav for their son. He was, however, named Yaroslav-Roman, because the priest who performed the baptism said the name Yaroslav is not a Christian one (this has not prevented later actually using only the first part of the official double name).

There was another child in the family - the daughter Galina, born in 1916,

*This is a revised and extended version of the talk [19].

[1]Perhaps it is difficult to determine whether this date is Old or New Style. Such a problem does not arise for people born in 1917 in Bulgaria, where the transition from Old to New Style happened in 1916. In Russia, however, it was in 1918.

a future Bulgarian specialist in Russian philology[2]. After a few hard years that followed the revolution, the Tagamlitzki family moved to Bulgaria in 1921 and settled in Sofia. Here, however, as seen by Galina's recollections [21], their lives were not at all easy. A few years after the immigration, the father became seriously ill and, in addition, the company where he found a job went bankrupt. In the sequel, the father's health continued to deteriorate, and finding a permanent job became impossible. The family care services lay mainly on the shoulders of the mother, who started work as a seamstress and ironer, and later as an embroidress. The penury became an everyday occurrence in the life of the four for decades. Many changes of abode occurred aiming at reducing the burden of the rent expenses.[3]

Absorbed in their troubles, the parents missed sending children in time to school. Fortunately, the American elementary school in Sofia which was opened shortly before allowed them to go straight into classes corresponding to their ages – Yaroslav into the second grade, and Galina into the third one (thanks to the fact that the two children already had become literate and had sufficiently educated themselves, no essential difficulties arose for them at school classes).

Concerning the school years of Yaroslav Tagamlitzki, let us quote [4] (with a footnote added here): "During the education of the young Tagamlitzki in the primary school no indication about the great talents hidden in him could be observed, but when he entered the famous Second Boys' High School in Sofia the things changed radically. Obviously an exceptionally favorable combination has arisen of, on the one hand, the great innate ability of the already mature pupil and his irresistible pursuit of science and, on the other hand, the high professional level of teachers and their genuine love towards their profession and care for the trainees. Already in this period the mathematical interests of Tagamlitzki far exceeded the matter studied in the secondary school, and he had also serious manifestations of his own scientific work, although for his disappointment the results turned out to be already known. Again in this period Tagamlitzki was a regular listener of the guest lectures in 1935 in Sofia of the prominent German mathematician Otto Blumenthal.[4] Not only, however, in mathematics and not only in science has shown his abilities the gifted and studious young man. To that time, for example, goes back his great attraction to music, the interest in which from aesthetic and scientific point of view does not leave him for the rest of his life."

In 1936, Yaroslav Tagamlitzki graduated from secondary education and became a mathematics student at the then Faculty of Physics and Mathematics of the Sofia University. After his second year of study there, his name appeared in

[2]Brief information about her can be found for instance in the calendar accessible from the web page [53] (the information is in the July 2016 section of the calendar; the link to that section is currently on the second page of the calendar).

[3]The dwelling places mentioned in [53] are indicated there by the corresponding street names. More complete information about the location of the fourth of them is present in the return address 46 Milin Kamak Str. of a letter sent in 1939 by the student Yaroslav Tagamlitzki to Professor Lyubomir Chakalov.

[4]Ludwig Otto Blumenthal (1876–1944). Some information about his lectures in Sofia is given on page 10 of the comprehensive biographical paper [2].

the papers [11, 12], where some results obtained by the student Tagamlitzki were included by his mathematical analysis professor, the prominent mathematician Professor Kiril Popov. A little later, still as a student, Tagamlitzki himself wrote three papers which appeared in editions of the Bulgarian Association for Physics and Mathematics, namely the articles [22, 23, 24]. The third of them, quite different in spirit from the other Bulgarian publications of that time, shows a profound knowledge of Lebesgue's integration theory.[5] It is worth noting that Tagamlitzki's letter to Chakalov mentioned in a preceding footnote concerns the subject matter of [24].

Tagamlitzki graduated from the University in 1940, and was seconded by the Ministry of Education to the then existing Mathematical Institute of the Sofia University. In 1942 and 1943 he had an academic specialization at Leipzig University and completed it with the defense of a doctoral dissertation in the field of complex analysis, namely [25].[6]

Here are the autobiographical data given by Yaroslav Tagamlitzki at the end of his dissertation:[7]

"I, Yaroslav Alexandrov Tagamlitzki, am a Bulgarian citizen, born on 11 September 1917 in Armavir, and I am the son of the engineer Alexander Tagamlitzki and his wife Vera. In Sofia I attended primary school and the semi-classical department of Second Boys' High School, which I graduated in 1936. Eight semesters (academic years from 1936/37 to 1939/40) I studied mathematics at Sofia University (Bulgaria). I attended there the lectures on mathematics of Messrs. Popov, Chakalov, Obreshkov, Tabakov, Tsenov and Stoyanov, the lectures on physics of Messrs. Nadzhakov, Penchev, Manev and Raynov and those of astronomy of Mr. Bonev.[8] During the academic year 1940/41 I was assisting at the Sofia Univer-

[5]However, a statement in [19] connected with this turns out to be not correct. It is claimed there that lectures on the theory in question started to be read in Sofia University much later. Actually such ones were read by Professor Kiril Popov already in the academic year 1939/1940, and a corresponding book by him appeared in 1941.

[6]Information about the fact of the defense and some other data can be found for instance by means of a search in the website [54] by the word *Tagamlitski* (this is the used there transcription of his family name). A scanned copy of the dissertation itself (as well as of almost all other works of him) is accessible from the page "A bibliography of Yaroslav Tagamlitzkis works" of the website [52] (in the dissertation, its author's name is transcribed as *Jaroslaw Tagamlizki*, and several other transcriptions of that name are indicated on the page "Web resources about Tagamlitzki" of [52]).

[7]The translation from German and the footnotes are mine, the Bulgarian names in the fourth sentence and in the footnote to it being transliterated according to the present-day official Bulgarian transliteration system (most of these names had other transcriptions in publications of the time – for instance the transcriptions Kyrille Popoff and Ljubomir Tschakaloff of the names Kiril Popov and Lyubomir Chakalov, as well as the transcription Obrechkoff of the family name Obreshkov). Let us note that we depart in this paper from the above-mentioned transliteration system in the case of Tagamlitzki's name by adhering to its transcription which is most frequently used in his Western language publications.

[8]These are Kiril Popov (1880–1966), Lyubomir Chakalov (1886–1963), Nikola Obreshkov (1896-1963), Dimitar Tabakov (1879–1973), Ivan Tsenov (1883–1967), Arkadi Stoyanov (1896–1963), Georgi Nadzhakov (1896–1981), Petar Penchev (1873–1956), Georgi Manev (1884–1965), Ruscho Raynov (1886–1965) and Nikola Bonev (1898–1979). A lot of information about the first six of them can be found in [5].

sity. During three semesters in the academic years 1941/42 and 1942/43 I attended as a regular student in Leipzig the lectures by Messrs. Koebe, van der Waerden, Schnee, Hopf, Heisenberg, Hund and Hopmann.[9] Then I started developing this treatise."

The thesis defended by Tagamlitzki in Leipzig is on a topic suggested to him by Paul Koebe, one of the world's leading experts in complex analysis at that time. The results presented in the dissertation generalize some important Koebe's results in the conformal mapping theory.

We will continue with a brief listing of further facts from the biography of Yaroslav Tagamlitzki, following closely [20, p. 231–232].

After his military service in the turbulent 1943 and 1944, Tagamlitzki was appointed in 1945 as Assistant Professor in the Faculty of Physics and Mathematics of the Sofia University – at the Department of Differential and Integral Calculus which was then led by Professor Kiril Popov (academician from 1947 on). In 1947 and 1949, Tagamlitzki was consecutively elected as a private and a regular associate professor at the same department. Since 1954 he has been Professor, Head of the Department of Differential and Integral Calculus. In 1958, Tagamlitzki was awarded a second doctor grade – this time according to the new rules for the scientific degrees in Bulgaria.[10] In 1961 he was elected as a corresponding member of the Bulgarian Academy of Sciences. Besides the Department on Differential and Integral Calculus, he also leads the Section of Functional Analysis at the Mathematical Institute of the Bulgarian Academy of Sciences, and after the unification of the two units in the early 1970s – the resulting sector of Real and Functional Analysis in the then established United Center of Mathematics and Mechanics at the Sofia University and the Bulgarian Academy of Sciences.

For Tagamlitzki's research on Dirichlet series and on the Laplace integral equation, he received in 1947 the Award for Science from the Committee for Science, Art and Culture[11]. In 1952 he was awarded the Dimitrov Prize[12] for the work [32]. For his scientific and teaching activities he was awarded the first grade "Cyrille and Methodius" order in 1953 and 1967, as well as a jubilee medal in 1969. In 1982 he was awarded the title Merited Scientist.

The active and versatile activity of Professor Yaroslav Tagamlitzki was cut short by his sudden death in Sofia on 28 November 1983.

Without any attempt at completeness, we will further consider the major scientific achievements of Tagamlitzki. But before going to them we will say, essentially

[9] Envisaged are Paul Koebe (1882-1945), Bartel Leendert van der Waerden (1903-1996), Walter Schnee (1885–1958), Eberhard Friedrich Ferdinand Hopf (1902–1983), Werner Heisenberg (1901–1976), Friedrich Hund (1896–1997) and Josef Hopmann (1890–1975). Biographical information about each of them can be found on the website [55].

[10] The degree was awarded for the work [32]; the reviewers were the academicians Lyubomir Chakalov, Nikola Obreshkov and Kiril Popov.

[11] This was a state institution which existed in the period from 1947 to 1954 and had the rank of a ministry.

[12] A high Bulgarian state award at that time.

following [20, p. 246], a few words about his teaching activity and his care for improvement of mathematical education. Their place in his life was very essential. From the beginning of his academic career to his last days Tagamlitzki was engaged most actively in the process of students education. Only a few years after this beginning, he developed the first modern calculus course in Bulgaria, and he taught it for the rest of his life, steadily bringing improvements and refinements. Typical of this course is the skillful combining of logical rigor with accessibility. The corresponding textbook which appeared first in a cyclostyle edition and then in six regular print editions stands out with its high qualities. Tagamlitzki delivered his lectures with remarkable pedagogical ability and took great care to ensure a thorough mastering of the taught material. Throughout several decades, he also read advanced lecture courses in various other fields of mathematics as integral equations, combinatorial topology, Fourier series, interpolation series, real functions theory, generalized functions. Tagamlitzki's functional analysis lectures which were read for more than a quarter of a century had a special place in his lecturing. They were quite different from the traditional functional analysis courses because aimed at displaying the results of the research of the lecturer himself.

In the talk [43] delivered in 1978 at a conference of the Union of Mathematicians in Bulgaria, Tagamlitzki presents his views on the teaching of mathematics at the university, supporting them by instructive examples.

Throughout the decades of his teaching at the Sofia University Yaroslav Tagamlitzki maintained close connections with secondary school, reading repeatedly there lectures on appropriate mathematical questions.[13] In the years 1963–1965 he conducted systematic work with bright secondary school students. The main subjects were the method of mathematical induction and a new method developed by him for building a certain essential part of calculus without using limit transition. Later he examined the applicability of the latter method for teaching the basics of calculus in secondary school, and during the school year 1973/1974 personally participated in the experimental application of this method in a school in Sofia. The method is described briefly in [39, 40], and a detailed description of it is given in [44] (the article [40] is a paper presented at the 1974 Spring Conference of the Bulgarian Mathematical Society[14]; besides the method in question also other important issues are considered in this paper).

The book [47] gives a fairly complete picture of the principles of Tagamlitzki's pedagogical activity in the mass education in mathematics, the depth of its implementation and his innovative approach to it. However, the devotional work of Tagamlitzki with capable mathematics students was especially fruitful. He did it with an indisputable talent. By attracting such students to research, he forwarded the growth as highly qualified mathematicians of many people from the next generations. The road to this for most of them went through Tagamlitzki's seminar.

[13] In the early fifties, being a secondary school student, the present author had the chance to be among the listeners of one of these lectures. Its title was "Solvable and unsolvable problems in mathematics".

[14] So then it was called the Union of Mathematicians in Bulgaria.

Despite of being modestly named 'Calculus Study Group', this seminar operated at a very high scientific level. In the paper [3], after a description of the scarcity of opportunities for contacts of the Bulgarian students with scientific life during the first post-war decade, the following is written about the seminar: "... we can imagine what an impression the work did in a study group where the participants were proposed to be full and equal associates of their enthusiastic teacher, and how easily that enthusiasm could find its way to their hearts. The unique atmosphere created determined the further path of many of the participants in the group". Eight PhD theses and many master theses created by Tagamlitzki's students contain essential scientific contributions on a variety of subjects in the stream of his investigations.

An exciting picture of the most essential of the above things and of Tagamlitzki's personality can be taken from the recollections [51, 14, 6, 7, 15] (all in the first chapter of the book [10]).

As has already been said, three publications of Tagamlitzki are from the time when he was a student. The next one is his PhD dissertation. Several years later, during the period from 1946 to 1951, a series of works of Tagamlitzki appeared, which make him the pioneer of functional analysis in Bulgaria. These works are internally united by the idea of indecomposability – in part of them it is present implicitly, and in others it occurs in an explicit form by the notion of prime vector. We will briefly mention some of the first ones and some of the others.

The above-mentioned series of works begins with the papers [26, 27, 28] inspired by certain investigations of N. Obreshkov. In the first of these papers, the following result is established: *if the function $f(x)$ is defined and has derivatives of any order for every x to the left of a fixed point a of the real axis, and A is a constant such that $|f^{(k)}(x)| \leq Ae^x$ for all $x < a$ and any nonnegative integer k, then the function $f(x)$ has the form Be^x, where B is a constant satisfying the inequality $|B| \leq A$* (the opposite is, of course, trivial). Of course, the result can be reformulated for the case of functions defined for $x > a$ – with e^{-x} instead of e^x; the considered property of the function e^{-x} is indicated and other proofs of it are given in [27, 30]. An analogous result, but for infinite sequences of numbers is established in [27, 28], wherein finite differences and infinite geometric progressions with terms between 0 and 1 are considered instead of derivatives and the exponential function.

In the paper [29], which is Tagamlitzki's habilitation work for the position of a regular associate professor[15], Tagamlitzki considers for each real number a a linear space $K(a)$ of functions defined in the interval $(a, +\infty)$,[16] and proves that among them are the functions $f(x)$, defined for $x > a$, tending to 0 as $x \to +\infty$, having derivatives of any order and satisfying the inequalities

$$(-1)^k f^{(k)}(x) \geq 0, \quad k = 0, 1, 2, \ldots,$$

for each $x > a$ (obviously the functions $e^{-\lambda x}$, where $\lambda > 0$, have the above properties; other functions with these properties are, for example, the functions of the

[15] For the position of private associate professor, his habilitation work is [28].

[16] The specific definition of this space will be not important here.

form $(x-b)^n$, where $b \leq a$ and $n < 0$). The functions with the listed properties are called *positively definite*. A partial ordering is defined in the space $K(a)$ by the convention that a function f majorizes a function g if and only if the difference $f - g$ is positively definite. The functions of the form $e^{-\lambda x}$, where $\lambda > 0$, turn out to be simple vectors of the so obtained partially ordered linear space in the sense that these positively definite functions are different from the zero element of the space, but any positively definite function majorized by such a function is a product of it with a constant. This fact is used to prove a theorem which, for an arbitrary sequence $\{\lambda_n\}$ of distinct positive real numbers gives a necessary and sufficient condition for the expandability in a series of the form $\sum A_n e^{-\lambda_n x}$, absolutely convergent in the interval $(a, +\infty)$ (the expandability in such a series could be intuitively regarded as a generalized representability as linear combination of prime vectors). This and other results in [29] are generalized in the article [31] for the case of arbitrary partially ordered linear spaces satisfying certain natural requirements (as is close to mind, the definition of a prime vector in this case is the following: *an element p of the space is called its prime vector if p majorizes the zero element, p is different from it, and p is collinear with each of its minorants which majorize the zero element*).

The above-mentioned intuitive idea plays an important heuristic role in obtaining the main results in the extremely deep and contentful work [32]. Let the infinite sequence of real numbers $x_0, x_1, x_2, \ldots$ be an arithmetic progression with positive difference τ. Interesting is the question about expandability of functions in series of the form $\sum_{n=0}^{\infty} a_n P_n(x)$, where the coefficients a_n do not depend on x, and $P_0(x), P_1(x), P_2(x), \ldots$ are the Abel interpolation polynomials defined as follows: $P_n(x) = (x_0 - x)(x_n - x)^{n-1}/n!, \quad n = 0, 1, 2, \ldots$ The answer to this question is obviously positive for a function $f(x)$ which is a polynomial (in this case we have $a_n = (-1)^n f^{(n)}(x_n)$ for $n = 0, 1, 2, \ldots$). Generally, however, the situation is much more complicated. The results proven in [32] reveal a substantial reason for this. Let a be a number less than x_0, and A be the linear space whose elements are the functions defined and infinitely many times differentiable in the interval $(a, +\infty)$. A function $f(x)$ from A is called *positively definite*, if, for each nonnegative integer k, we have $(-1)^k f^{(k)}(x) \geq 0$ when $a < x \leq x_k$ (the positively definite functions of the space $K(a)$ from the paper [29] obviously satisfy this condition, but some other functions in A also meet it – it can be shown, for example, that it is satisfied also by the Abel interpolation polynomials). Let a partial order in the space A be defined by the convention that a function f majorizes a function g when the difference $f - g$ is positively definite. It turns out that the Abel interpolation polynomials are prime vectors of the so obtained partially ordered linear space, but besides them and their products with positive constants there are many other prime vectors – it is proven in the paper that the set of prime vectors of that space consists of the products with positive constants of the Abel interpolation polynomials and of the

functions of x of the form $R(x,t)$, $0<t\leq 1$, wherein

$$R(x,t)=\begin{cases}\dfrac{e^{\lambda(x_0-x)}-e^{\mu(x_0-x)}}{\lambda-\mu} & \text{for } 0<t<1,\ \tau\lambda e^{1-\tau\lambda}=\tau\mu e^{1-\tau\mu}=t,\ \tau\mu<1<\tau\lambda,\\ (x_0-x)e^{(x_0-x)/\tau} & \text{for } t=1.\end{cases}$$

Due to this, the realization of the intuitive idea about generalized representability as a linear combination of prime vectors becomes more complicated – it naturally uses not only series but also integrals. One of the main results in [32] is exactly in this spirit. It concerns the representability of functions $f(x)$ in the form

$$f(x)=\sum_{n=0}^{\infty}a_nP_n(x)+\int_0^1 R(x,t)d\theta(t), \tag{1}$$

where the coefficients a_n are nonnegative numbers not depending on x, and $\theta(t)$ is a monotonically increasing function which also does not depend on x. The result is that a function $f(x)$ of the space A is representable in this form if and only if $f(x)$ is positively definite. In that case a uniqueness theorem holds according to which the coefficients a_n are uniquely determined, and the function $\theta(t)$ is substantially uniquely determined. These results can be interpreted as an indication that the expandability in a series on the Abel polynomials is a rare and unusual case, and the natural task is that of representation in the form (1) (possibly without requirements about nonnegativity of the coefficients a_n and monotony of the function $\theta(t)$).

The representability in the form (1) with nonnegative a_k and monotone increasing $\theta(t)$ is obvious for the Abel interpolation polynomials and the functions of x of the form $R(x,t)$, where $0<t\leq 1$, so it is present for each prime vector of the space A. If we denote by K the set of the elements of a partially ordered linear space which majorize the zero element of the space, then the prime vectors of this space are actually those nonzero elements of K which cannot be represented as the sum of two noncollinear elements of K. The set K is a cone, i.e. belonging to K is preserved under multiplying with nonnegative numbers. In the work [33], which appeared a few years later, Tagamlitzki calls *indecomposable* elements of a cone those of its nonzero elements which cannot be represented as a sum of two noncollinear its elements, and proves under certain assumptions that *if all indecomposable elements of a cone belong to a convex cone*[17] *then all elements of the first cone belong to the second one.* The most restrictive of the assumptions is maybe that the considered cone is located in a linear space with scalar product and there exists a fixed nonzero element of the space forming acute angles with all nonzero elements of the cone (such a cone cannot for example have nonzero mutually opposite elements). Nevertheless, the said result provides a general method for proving a number of statements that are in the spirit of the statement about the representability of the positively definite elements of the space A in the form (1). Several examples are

[17] A cone is called convex if belonging to it is preserved under addition (it is easy to show that this is equivalent to the requirement the cone in question to be a convex set).

given in the paper which illustrate this method, namely proofs of the positive case of the Hausdorff moment theorem, of Bernstein's theorem about the integral representation of completely monotone functions in an infinite interval, of a statement equivalent to the theorem of Bernstein about the analyticity of functions completely monotone in a finite interval, and of a theorem about representability by integral, similar to the one in the representation (1). As indicated in [9], another application made then by Tagamlitzki remained unpublished, namely an elegant proof of the classical Bochner's theorem on the positive definite functions (one of the first applications of Tagamlitzki's theorem that also remained unpublished is its application to the positive case of F. Riesz's representation theorem for linear functionals on $C[a,b]$).

The above-mentioned restrictive assumption is removed in the published in [34] two years later another version of the above theorem. Let K be a convex cone, and P be a nonnegative real-valued function such that $P(\lambda a) = \lambda P(a)$ for each nonnegative number λ and each $a \in K$, $P(a+b) \leq P(a) + P(b)$ for any $a, b \in K$, and the value of P can be 0 only for the zero element of the cone. Deviating slightly from the terminology of [34], we will call such a function a *norm of* K (the definition of norm in [34] imposes also a certain semicontinuity requirement). By definition, an *indecomposable element of* K *with respect to* P is such a nonzero element a of K that no noncollinear elements b and c of K exist satisfying the equations $a = b + c$ and $P(a) = P(b) + P(c)$. In the new version of the theorem, it is assumed to be given a convex cone K with a norm P and a contained therein convex cone L with a norm Q, and it is proved under some additional assumptions that *if the conditions* $x \in L$ *and* $P(x) \geq Q(x)$ *are satisfied whenever* x *is an indecomposable element of* K *with respect to* P, *then these conditions are met for each* x *of* K. It is shown in the article that the method derived from this theorem (now known as *Cones Theorem*), allows proving with its help the Hausdorff moment theorem (first in the general case and then in the positive one), as well as Widder's theorem about the integral representation of the functions $f(x)$, defined and infinitely many times differentiable for $x > 0$, which satisfy the condition $\frac{1}{n!}\int_0^\infty x^n \left| f^{(n+1)}(x) \right| dx \leq A$, $n = 0, 1, 2, \ldots$, with constant A, independent of n (it is indicated how then one can get also Bernstein's theorem about integral representation of the absolutely monotone functions). Numerous other applications of the method are listed in the summary [35] of a talk presented by Tagamlitzki in 1956 at an international mathematical conference in Sofia (in the following years appeared also many other applications). The application of the theorem to the general case of Riesz's theorem on the representation of the bounded linear functionals in $C[a,b]$ was the first among the applications listed in the summary. However, a detailed presentation of this application was published only thirty years later in [8], and such presentations of some other ones remained unpublished at all, although all these results were considered in detail at Tagamlitzki's seminar or on his functional analysis lectures.

In 1956 and 1957, the three parts of the work [36] appeared. They present a way suggested by Tagamlitzki for building the theory of generalized functions.

Here is the abstract of the third part (with references made independent of its context):[18]

"The present work is the last part of a research, whose first two parts were published under the same title and contain the general principles of completion of cones, the definition of the space S_n^σ of pseudofunctions, and their basic properties.

Pseudofunctions, which include all summable functions and the Dirac functions on a given interval, can be differentiated arbitrary many times; moreover, the derivative of an element of S_n^σ is in $S_{n+1}^{\sigma+1}$. For $\sigma > n$ we have $S_n^\sigma \subset S_{n+1}^{\sigma+1}$. The space S_n^σ possesses a countable coordinate system. Unlike the space of the Schwartz distributions, it possesses a semicontinuous norm and it is compact with respect to it.

It follows from our earlier work *On a generalization of the notion of indecomposability* that the space S_n^σ contains indecomposable elements. The present third part is dedicated primarily to the indecomposable elements of these spaces. We establish that the indecomposable elements of S_n^σ with respect to the corresponding norm are the n-th derivatives of the Dirac functions."

The paper [50] which appeared a little later is also in such a spirit. As indicated in a footnote on its first page, it reproduces, with relevant generalizations and additions made by the second author, the main points of the investigations which Tagamlitzki expounded in his functional analysis lectures in the academic year 1955/56.

Thanks to the participation of foreign mathematicians in the 1956 conference more people abroad became aware of the method developed by Tagamlitzki. The French mathematician Choquet[19] offered Tagamlitzki to publish in France a more systematic presentation of the results obtained. A publishing house in East Germany started negotiations with Tagamlitzki to print a monograph on these results. He accepted these proposals in principle but did not hurry to implement them – both because of the continued intensive emergence of new results obtained by him and his disciples, and because of another important reason to which we will turn now.

At some point after 1956 Tagamlitzki realized that his Cones Theorem can be obtained as a consequence of the Krein-Milman Theorem. The delay in recognition of this fact is explained by the limited possibility for contacts of the Bulgarian mathematicians with the international mathematical community during the Second World War and the first postwar decade. Despite the fact in question, however, many of the applications of the Cones Theorem are scientific contributions – all its applications can be considered as applications of the Krein-Milman Theorem, and many of them turn out to be new. However, Tagamlitzki was still considering desirable a further reflection and an expansion of the achieved before proceeding to its monographic exposure. A few years later, in 1962, he was again invited to write

[18] The terminology of the earlier work mentioned in the abstract (i.e., of the work [34]) is used in it.

[19] Gustave Choquet (1915–2006).

a monograph – this time from the American publishing house Van Nostrand (the invitation was inspired by a recommendation of the famous mathematician Marshall Stone[20]). Tagamlitzki tended to accept the invitation and was considering a plan for the monograph in question, but unfortunately it remained unwritten. This is probably due to the fact that at that time he worked on a far-reaching generalization of the Krein-Milman Theorem. This generalization was reported in [37] and improved further – subsequent versions of this generalization are set forth in detail in [49] and, eight years after Tagamlitzki's death, in [46] (three years after his death, also a specific other version of the generalization was published in [13]). In the final form of the generalization in question, an arbitrary compact topological space with a collection of open sets satisfying certain conditions is considered instead of a compact subset of a linear space with a locally convex topology (the Krein-Milman Theorem is obtained by applying the general result to this subset and the collection of all its intersections with convex open sets). An application of this generalization produces a more general form of Bauer's maximum principle (cf. [49, Application 2] and [41]).

In the applications of the Cones Theorem, the search for the indecomposable elements is often done by means of so-called decomposing operators, and they are usually linear ones. In the paper [42], an analog of the linear decomposing operators is considered which could be useful for the search of extreme points in the applications of Krein-Milman's theorem. A characterization of the topological simplexes is given by means of the introduced notion.

Besides the results concerning indecomposability, another contribution of Tagamlitzki in functional analysis is the published in [38] generalization of certain theorems on separability of convex sets. Instead of sets in a linear space he considers such ones in mathematical structures with an appropriate notion, generalizing the notion of segment. Tagamlitzki proved two theorems of this kind, the second one concerning the case when one of the considered convex sets is open in a topology co-ordinated with the above-mentioned notion. It became known later that the first of these theorems follows from a theorem published by Ellis in 1952, but the second one could not be derived in such a way. In addition, as shown in [17], without essentially changing the proof of the first theorem, its assumptions can be weakened in a way hindering its derivation from Ellis's one. Tagamlitzki's investigations on separability of convex sets were essentially extended further in the works of Ivan Prodanov.

As already mentioned, some of Tagamlitzki's results in functional analysis remained unpublished. The results obtained in the last two decades of his life had such a fate especially often. However, some of them are mentioned by other authors who anyhow became aware of these results. For instance, the following is written in [1, p. 255]:

„Positive definite functions on semigroups have been studied by Tagamlitzki and his pupils in Bulgaria. In lectures at the University of Sofia during the years

[20]Marshal Harvey Stone (1903–1989).

1965–1969, Tagamlitzki introduced the operators $W_{\xi,a}$ for functions f on an abelian group G by

$$W_{\xi,a}f(s) = 2f(s) + \xi f(s+a) + \overline{\xi} f(s-a),$$

where $a, s \in G$ and $\xi \in \mathbf{C}$.

He used these operators for proving Bochner's theorem for discrete groups as a direct consequence of Krein-Milman's theorem."

The quoted author indicated further that a similar approach is published by Choquet in a paper from 1969, and it is mentioned that the method in question is extended to semigroups with involution by T. Tonev in his Master Thesis, written under the supervision of Tagamlitzki and defended in 1969 (Tonev's publication on this subject from 1979 in Semigroup Forum is also referred to). Some information on a paper of D. Shopova from 1970 is also given.

Among the results of Tagamlitzki in abstract areas of mathematics, other than functional analysis, we can point to a broad generalization of Tychonoff's theorem on compactness of topological products. Unfortunately the only materials in paper-like form on this subject left from Tagamlitzki are three texts in the annual scientific reports of the Sector of Real and Functional Analysis – for the years 1977, 1980 and 1982. The paper [45] was written later on the base of these texts.

Tagamlitzki developed also some new approaches to the theory of manifolds and to some mathematical questions of theoretical physics. The corresponding materials in paper-like form left by him are again texts in the annual scientific reports of the Sector of Real and Functional Analysis – for the years 1973, 1974, 1976, 1978–1981.

Besides in mathematics, Tagamlitzki carried out systematic research also on questions of archaeology, linguistics and medicine. He offered new ideas in the doctrine of tonality in music.

A lot of information waiting for its investigation can be found in Tagamlitzki's archive in the Bulgarian Academy of Sciences. An overview of the items of mathematical character in this archive is given in [16].

The person and the deed of Yaroslav Tagamlitzki left a deep and bright trace in the history of our science, in our education and in the souls of many people of several generations. The centenary of his birth is an appropriate occasion to express our admiration to the memory of this remarkable man.

REFERENCES

[1] Berg, C.: Positive definite and related functions on semigroups. In: *The Analytical and Topological Theory of Semigroups* (K. H. Hofmann, J. D. Lawson and J. S. Pym, eds.), Walter de Gruyter, 1990, pp. 253–278.

[2] Butzer, P., Volkmann, L.: Otto Blumenthal (1876–1944) in retrospect. *J. Approx. Theory*, **138**, 2006, 1–36.

[3] Chakalov, V., Skordev, D.: The scientific and pedagogical deed of Yaroslav Tagamlitzki. In: [10, pp. 89–102] (in Bulgarian).

[4] Chakalov, V., Skordev, D.: The life and deed of Yaroslav Tagamlitzki. *Ann. Univ. Sofia, Fac. Math. Inf.*, **91**, 1999, 13–19 (in Bulgarian).

[5] Chobanov, I., Rusev, P. (eds.): *Bulgarian Mathematicians*. Sofia, Narodna Prosveta, 1987 (in Bulgarian).

[6] Doitchinov, D.: Prof. Tagamlitzki in my recollections. In: [10, pp. 58–64] (in Bulgarian).

[7] Genchev, T.: Yaroslav Tagamlitzki – a teacher and an educator of scientists. In: [10, pp. 65–76] (in Bulgarian).

[8] Genchev, T.: An unpublished proof of Y. Tagamlitzki (the Riesz Theorem on the linear functionals in $C[a,b]$ as a corollary of the Cones Theorem). In: [10, pp. 134–140] (in Bulgarian).

[9] Genchev, T.: A solution of the trigonometric moment problem via Tagamlitzki's „Theorem of the Cones". *Pliska Studia Mathematica Bulgarica*, **11**, 1991, 35–39.

[10] Genchev, T., Prodanov, I., Skordev, D., Todorov, I., Chakalov, V. (eds.): *Yaroslav Tagamlitzki – Scholar and Teacher*. Sofia, Nauka i Izkustvo, 1986 (in Bulgarian).
Remark. The web page at
`https://www.fmi.uni-sofia.bg/fmi/logic/skordev/errata_Tagamlitzki.htm`
yields an errata list (in Bulgarian) for the book.

[11] Popov, K. (Popoff, K.): Sur une extension de la notion de dérivée. *C. R. Acad. Sci. Paris*, **207**, 1938, 110–112.

[12] Popov, K.: On a generalization of the notion of derivative. *Ann. Univ. Sofia, Fac. Phys.-Math.*, **35**, livre 1, 1939, 225–245 (in Bulgarian)

[13] Prodanov, I.: A note on the unpublished investigations of Y. Tagamlitzki on the extreme elements method. In: [10, pp. 141–149] (in Bulgarian).
Remark. See [18] in connection with an editor's oversight affecting the applicability of the theorem on p. 147.

[14] Sendov, Bl.: Student's recollections on Prof. Y. Tagamlitzki. In: [10, pp. 54–57] (in Bulgarian).

[15] Skordev, D.: My earliest and my latest recollections on Y. Tagamlitzki. In: [10, pp. 77–82] (in Bulgarian).

[16] Skordev, D.: Brief information about certain scientific materials in Professor Yaroslav Tagamlitzki's archive. In: [10, pp. 262–266] (in Bulgarian).

[17] Skordev, D.: A separation theorem of Y. Tagamlitzki in its natural generality. *Ann. Univ. Sofia, Fac. Math. Inf.*, **91**, 1999, 73–78.

[18] Skordev, D.: An oversight of mine and a way for its correction.
`https://www.fmi.uni-sofia.bg/fmi/logic/skordev/correction.htm`
(in Bulgarian)

[19] Skordev, D.: 100 years from the birth of Professor Yaroslav Tagamlitzki. In: *Mathematics and Mathematical Education, Proc. of the 46th Spring Conference of UMB, Borovets, 9–13.4.2017*, Bulg. Acad. Sci., 2017, 7–16 (in Bulgarian).

[20] Skordev, D., Genchev, T., Prodanov, I., Chakalov, V.: Yaroslav Tagamlitzki. In: [5, pp. 231–259] (in Bulgarian).

[21] Tagamlitzka, G.: My brother Yaroslav Tagamlitzki. In: [10, pp. 11–38] (in Bulgarian).

[22] Tagamlitzki, Y.: On the mean value theorem. *Journal of the Association for Physics and Mathematics*, **24**, no. 3–4, 1938, 95–98 (in Bulgarian).

[23] Tagamlitzki, Y.: A generalization of the mean value theorem. *Journal of the Association for Physics and Mathematics*, **24**, no. 5–6, 1939, 173–175 (in Bulgarian).

[24] Tagamlitzki, Y.: A property of the summable functions in Lebesgues sense. In: *Anniversary Collection of the Association for Physics and Mathematics*, Part 2, Sofia, 1939, 73–74 (in Bulgarian).

[25] Tagamlitzki, Y. (Tagamlizki, J.): Zum allgemeinen Kreisnormierungsprinzip der konformen Abbildung. *Sitzungsberichte der Sächsischen Akademie der Wissenschaften, Mathematisch-physische Klasse*, **95**, 1943, 111–132. Reproduced in: [48, pp. 9–30].

[26] Tagamlitzki, Y.: Functions satisfying certain inequalities on the real axis. *Ann. Univ. Sofia, Fac. Phys.-Math.*, **42**, livre 1, 1946, 239–256 (in Bulgarian). English translation in: [48, pp. 31–48].

[27] Tagamlitzki, Y.: Sur les suites vérifiant certaines inégalités. *C. R. Acad. Sci. Paris*, **223**, 1946, 940–942. Reproduced in: [48, pp. 49–51].

[28] Tagamlitzki, Y.: On sequences satisfying certain inequalities. *Ann. Univ. Sofia, Fac. Phys.-Math.*, **43**, livre 1, 1947, 193–237 (in Bulgarian). English translation in: [48, pp. 52–95].

[29] Tagamlitzki, Y.: Investigation of a class of functions. *Ann. Univ. Sofia, Fac. des Sciences*, **44**, livre 1, 1948, 317–356 (in Bulgarian). English translation in: [48, pp. 117–158].

[30] Tagamlitzki, Y.: Sur une propriété de la fonction exponentielle. *C. R. Acad. Bulg. Sci.*, **1**, no. 1, 1948, 33–34. Reproduced in: [48, pp. 159–161]. Bulgarian translation in: [10, pp. 103–105].

[31] Tagamlitzki, Y.: On certain applications of the general theory of partially ordered linear spaces. *Ann. Univ. Sofia, Fac. des Sciences*, **45**, livre 1, 1949, 263–286 (in Bulgarian). English translation in: [48, pp. 162–187].

[32] Tagamlitzki, Y.: A research on the Abel interpolation series. *Ann. Univ. Sofia, Fac. des Sciences*, **46**, livre 1, 1950, 385–443 (in Bulgarian). English translation in: [48, pp. 193–249].

[33] Tagamlitzki, Y.: On the geometry of the cones in Hilbert spaces. *Ann. Univ. Sofia, Fac. des Sciences Phys. Math.*, **47**, livre 1, partie 2, 1952, 85–107 (in Bulgarian). English translation in: [48, pp. 256–281].

[34] Tagamlitzki, Y.: On a generalization of the notion of indecomposability. *Ann. Univ. Sofia, Fac. des Sciences Phys. Math.*, **48**, livre 1, partie 1, 1954, 69–85 (in Bulgarian). English translation in: [48, pp. 300–317].

[35] Tagamlitzki, Y.: The indecomposable elements and their applications. In [10, pp. 106–108] (in Bulgarian; translated from the book of abstracts of 1956, where it is in Russian and German – cf. [52, Works] for links to scanned copies of all versions).

[36] Tagamlitzki, Y.: Cone completion and its application to the function notion generalization problem (I, II, III). *Ann. Univ. Sofia, Fac. des Sciences Phys. Math.*, **49**, livre 1, partie 1, 1956, 23–48, **49**, livre 1, partie 2, 1956, 41–54, **50**, livre 1, partie 1,

1957, 135–163 (in Bulgarian). English translation in: [48, pp. 318–345, pp. 346–359, pp. 360–389].

[37] Tagamlitzki, Y.: On the extreme points method. *Studia Mathematica*, Serija Specjalna, Zeszyt 1, 1963, 129–130 (in Russian). English translation in: [48, pp. 498–499].

[38] Tagamlitzki, Y.: On the separation principle in Abelian associative spaces. *Bulletin de l'Institut de mathématiques, Académie des sciences de Bulgarie*, **7**, 1963, 169–183 (in Bulgarian). English translation in: [48, pp. 500–513].

[39] Tagamlitzki, Y.: On the modernization of the mathematics curricula in the secondary schools by introducing ephodics. *Mathematics and Physics*, 1973, no. 6, 6–8 (in Bulgarian).

[40] Tagamlitzki, Y.: Some issues in teaching mathematics in the secondary schools. In: *Mathematics and Mathematical Education, Proc. of the 3rd Spring Conference of UMB, Burgas, 2–4.4.1974*, Bulg. Acad. Sci., 1976, 87–93 (in Bulgarian). Reproduced in: [10, pp. 159–167] and [47, pp. 9–17])

[41] Tagamlitzki, Y.: Sur le principe du maximum. In: *Mathematical Structures – Computational Mathematics – Mathematical Modelling, Papers dedicated to Professor L. Iliev's 60th anniversary*, Bulg. Acad. Sci., 1975, 471–477. Reproduced in: [48, pp. 520–527].

[42] Tagamlitzki, Y.: A boundary value problem in linear spaces. *C. R. Acad. Bulg. Sci.*, **29**, no. 3, 1976, 307–309. Reproduced in: [48, pp. 528–531].

[43] Tagamlitzki, Y.: Lecture and creativity. In: *Mathematics and Mathematical Education, Proc. of the 7th Spring Conference of UMB, Sunny Beach, 5–8.4.1978*, Bulg. Acad. Sci., 1978, 114–124 (in Bulgarian). Reproduced in: [10, pp. 222–232] and [47, pp. 18–27]).

[44] Tagamlitzki, Y.: A method for introducing some elements of calculus without limit transition (prepared by V. Chakalov). In: [10, pp. 170–221] (in Bulgarian). Reproduced in [47, pp. 28–78].

[45] Tagamlitzki, Y.: A diagonal principle for generalized sequences and some of its applications (prepared by D. Skordev). *Serdica*, **16**, 1990, 201-209. Reproduced in: [48, pp. 532–543].

[46] Tagamlitzki, Y.: The principle of topological induction (prepared by O. Kounchev). *Ann. Univ. Sofia, Fac. Math. Méc.*, **80**, livre 1, 1991, 53–58. Reproduced in: [48, pp. 544–550].

[47] Tagamlitzki, Y.: *On the Teaching of Mathematics (from the pedagogical legacy of Professor Tagamlitzki)*. "St. Kliment Ohridski" University Press, 2016 (in Bulgarian).

[48] Tagamlitzki, Y.: *Selected Papers*. "St. Kliment Ohridski" University Press, 2017.
Remark. The web page `https://www.fmi.uni-sofia.bg/fmi/logic/skordev/errata_selected.htm` contains an errata list for the book.

[49] Tagamlitzki, Y. (Tagamlicki, Ja. A.), Dehen, M.: L'induction topologique. *Séminaire Choquet. Initiation à l'analyse*, **10**, no. 1, 1970–1971, 1–7. Reproduced in: [48, pp. 514–519].

[50] Tagamlitzki, Y., Doitchinov, D.: Investigation of a class of generalized functions. *Ann. Univ. Sofia, Fac. des Sciences Phys. Math.*, **52**, livre 1, 1959, 23–95 (in Bulgarian). English translation in: [48, pp. 421–497].

[51] Todorov, I.: My teacher Professor Yaroslav Tagamlitzki. In: [10, pp. 42–53] (in Bulgarian).

[52] *In memory of Prof. Yaroslav Tagamlitzki.* http://www.tagamlitzki.com/

[53] *Language and Linguistics Heritage.* http://www.e-nasledstvo.com/ (in Bulgarian)

[54] *Mathematics Genealogy Project.* http://www.genealogy.ams.org/

[55] *Professorenkatalog der Universität Leipzig — catalogus professorum lipsiensium.* http://www.uni-leipzig.de/unigeschichte/professorenkatalog/

Received on October 13, 2017

Dimiter Skordev

Faculty of Mathematics and Informatics
"St. Kl. Ohridski" University of Sofia
5, J. Bourchier blvd., BG-1164 Sofia
BULGARIA

E-mail: skordev@fmi.uni-sofia.bg

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 104

ANNUAL OF SOFIA UNIVERSITY „ST. KLIMENT OHRIDSKI“
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 104

AUTOMORPHISMS OF ALGEBRAS AND BOCHNER'S PROPERTY FOR DISCRETE VECTOR ORTHOGONAL POLYNOMIALS

EMIL HOROZOV

We construct new families of discrete vector orthogonal polynomials that have the property to be eigenfunctions of some difference operator. They are extensions of Charlier, Meixner and Kravchuk polynomial systems. The ideas behind our approach lie in the studies of bispectral operators. We exploit automorphisms of associative algebras which transform elementary (vector) orthogonal polynomial systems which are eigenfunctions of a difference operator into other systems of this type. While the extension of Charlier polynomilas is well known it is obtained by different methods. The extension of Meixner polynomial system is new.

Keywords: vector orthogonal polynomials, finite recurrence relations, bispectral problem.

2000 Math. Subject Classification: 34L20 (Primary); 30C15, 33E05 (Secondary).

1. INTRODUCTION

The present paper is a continuation of [16] but could be read independently. Both papers are devoted to vector orthogonal polynomials with Bochner's property.

S. Bochner [7] has classified all systems of orthogonal polynomials $P_n(x)$, $n = 0, \ldots,$ that are also eigenfunctions of a second order differential operator

$$L(x, \partial_x) = A(x)\partial_x^2 + B(x)\partial_x + C(x) \tag{1.1}$$

with eigenvalues λ_n. Here the coefficients A, B, C of the differential equation do not depend on the index n.

A similar problem was solved by O. Lancaster [19] and P. Lesky [21], although earlier E. Hildebrandt [15] has found all needed components of the proof. For more information see the excellent review article by W. Al-Salam [1].

The statement of Lancaster's theorem is that all polynomial systems with such properties are the discrete orthogonal polynomials of Hahn, Meixner, Charlier and Kravchuk.[1]

In recent times there is much activity in generalizations and versions of the classical result of Bochner as well as their discrete counterparts. The first one was the generalization by H. L. Krall [18]. He classified all order 4 differential operators which have a family of orthogonal polynomials as eigenfunctions. Later many authors found new families of orthogonal polynomials that are eigenfunctions of a differential operator (see, e.g. [12, 13]).

The classical discrete orthogonal polynomials are also a source from which new orthogonal polynomials have been obtained. In particular A. Durán and M. de la Iglesia [9] have obtained extensions of the classical polynomial systems of Hahn, Meixner and Charlier.

An important role in some of these generalizations plays the ideology of the bispectral problem which was initiated in [8]. Translating the Bochner and Krall results (cf. [11]) into this language already gives a good basis to continue investigations. We formulate it for the case of discrete orthogonal polynomials. Let us introduce the function $\psi(x, n) = P_n(x)$. Denote by D the shift operator acting on functions of x as $Df(x) = f(x + 1)$. Also let T be the shift operator in n, i.e. $Th(n) = h(n + 1)$. Recall that the orthogonality condition, due to a classical theorem by Favard-Shohat is equivalent to the well known 3-terms recurrence relation

$$xP_n = P_{n+1} + \gamma_0(n)P_n + \gamma_1(n)P_{n-1}, \tag{1.2}$$

where $\gamma_j(n)$ are constants, depending on n. Here we use the polynomials normalized by the condition that their highest order coefficient is 1.

If we write the right-hand side of the 3-term recurrence relation as a difference operator $\Lambda(n)$ acting on the variable n then the 3-term recurrence relation can be written as

$$\Lambda(n)\psi(x, n) = x\psi(x, n).$$

On the other hand we want $\psi(x, n)$ to be an eigenfunction of a difference operator L in x:

$$L\psi(x, n) = \lambda(n)\psi(x, n).$$

[1]Below when speaking about orthogonality we mean orthogonality with respect to a nondegenerate functional, which does not need to be positive definite.

Hence we can formulate the discrete version of Bochner-Krall problem as follows.

Find all systems of orthogonal polynomials $P_n(x)$ (with respect to some functional u) which are eigenvalues of a difference operator.

We also use some ideas relevant to the studies of bispectral operators. Before explaining them and the main results of the present paper let us introduce one more concept which is central for us. This is the notion of vector orthogonal polynomials (VOP), introduced by J. van Iseghem [24]. Let $\{P_n(x)\}$ be a family of monic polynomials such that $\deg P_n = n$. A theorem of P. Maroni [22] $\{P_n(x)\}$ gives an equivalent condition, which we use as definition.

Definition 1.1. We will say that the set of polynomials (P_n) are Vector Orthogonal polynomials (VOP) iff they satisfy a $d+2$-term recurrence relation, $d \geq 1$, of the form

$$xP_n(x) = P_{n+1} + \sum_{j=0}^{d} \gamma_j(n) P_{n-j}(x) \tag{1.3}$$

with constants (independent of x) $\gamma_j(n)$, $\gamma_d(n) \neq 0$.

In the last 20-30 years there is much activity in the study of vector orthogonal polynomials and the broader class of multiple orthogonal polynomials.

Applications of the VOP include the simultaneous Padé approximation problem [2] and random matrix theory [2, 6]. The VOP can be obtained from general multiple orthogonal polynomials under some restrictions upon their parameters.

One problem that deserves attention is to find vector orthogonal analogs of the classical orthogonal polynomials. Several authors [3, 14] have found multiple orthogonal polynomials, that share a number of properties with the classical orthogonal polynomials - they have a raising operators, Rodrigues type formulas, Pearson equations for the weights, etc. However one of the features of the classical orthogonal polynomials - a differential or difference operator for which the polynomials are eigenfunctions is missing. Sometimes this property is relaxed to the property that the polynomials satisfy linear differential/difference equation, whose coefficients may depend on the index of the polynomial, see [20].

In the present paper we are looking for polynomials $P_n(x)$, $n = 0, 1, \ldots$ that are eigenfunctions of a difference operator $L(x, D)$ with eigenvalues depending on the variable n (the index) and which at the same time are eigenfunctions of a difference operator in n, i.e. finite-term recurrence relation with an eigenvalues, depending only on the variable x. Hence we find families $\{P_n(x)\}, n = 0, 1, \ldots$ of discrete VOP that possess Bochner's property - they are simultaneously eigenfunctions of two discrete operators:

$$L(x, D)P_n(x) = \lambda(n)P_n(x), \quad \Lambda(n, T)P_n(x) = xP_n(x).$$

Our main results include an extension of Meixner polynomials. We construct systems of vector orthogonal polynomials $\{P_n(x)\}$ which are eigenfunctions of a difference operator $L(x, D)$. It is different from the family found in [10] and [5] except for the first member. Our approach uses ideas of the bispectral theory from [4] but does not use Darboux transformations, which is usually the case, see e.g. [12, 13, 9]. We use methods introduced in [4]. Also a well known extension of Charlier polynomials (see [5, 25]) is presented. The reason to repeat it is that our construction is a new one in comparison to the techniques of [5, 25]. However, there are some similarities with [25]. The authors also use automorphisms of algebras and make a beautiful connection with representation theory. Our construction is simpler and quite straightforward. The same method was recently applied to extensions of Hermite and Laguerre polynomials as well as to a family that has no classical analog [16].

The methods from the present paper and [16] can be applied to various versions of vector orthogonal polynomials as well as to matrix, multivariate, etc. This will be done elsewhere.

Acknowledgements. The author is deeply grateful to Boris Shapiro for showing and discussing some examples of systems of VOP and in particular the examples from [23]. They helped me to guess that the methods from [4] can be useful for the study of vector orthogonal polynomials. The author is grateful to the Mathematics Department of Stockholm University for the hospitality in April 2015. The research is partially supported by Grant DN 02-/05 of the Bulgarian Fund "Scientific research".

Last but not least I am extremely grateful to Prof. T. Tanev, Prof. K. Kostadinov, and Mrs. Z. Karova from the Bulgarian Ministry of Education and Science and Prof. P. Dolashka, BAS who helped me in the difficult situation when I was sacked by Sofia university in violations of the Bulgarian laws. This was a retaliation for my attempt to reveal a large -scale corruption, that involves highest university and science officials in Bulgaria.[2]

2. ELEMENTS OF BISPECTRAL THEORY

The following introductory material is mainly borrowed from [4]. Below we present the difference-difference version of the general bispectral problem which is suitable in the set-up of discrete orthogonal polynomial sequences.

For $i = 1, 2$, let Ω_i be two subsets of $\mathbb{C}$ such that Ω_1 is invariant under the translation operator

$$D\colon x \mapsto x + 1,\ x \in \Omega_1$$

[2]See, e.g. EMS NEWSLETTER, http://www.ems-ph.org/journals/newsletter/pdf/2015-12-98.pdf

and its inverse D^{-1}, while Ω_2 is invariant under the translation operator

$$T\colon n \mapsto n+1$$

and its inverse T^{-1}.

A difference operator on Ω_1 is a finite sum of the form

$$\sum_{k\in\mathbb{Z}} c_k(x)D^k,$$

where $c_k\colon \Omega_1 \to \mathbb{C}$ are some functions in x. In the same way we define difference operators on Ω_2 to be finite sums of the form

$$\sum_{k\in\mathbb{Z}} s_k(n)T^k,$$

where $s_k\colon \Omega_1 \to \mathbb{C}$ are functions in n.

By $\mathcal{B}_1$ we denote an algebra with unit, consisting of difference operators $L(x,D)$ in the variable x. By $\mathcal{B}_2$ we denote an algebra of difference operators $\Lambda(n,T)$. Denote by $\mathcal{M}$ the space of functions on $\Omega_1\times\Omega_2$. The space $\mathcal{M}$ is naturally equipped with the structure of bimodule over the algebra of difference operators $L(x,D)$ on Ω_1 and the difference operators $\Lambda(n,T)$ on Ω_2.

Assume that there exists an algebra map $b\colon \mathcal{B}_1 \to \mathcal{B}_2$ and an element $\psi \in \mathcal{M}$ such that

$$P\psi = b(P)\psi, \quad \forall P \in \mathcal{B}_1.$$

We call $\psi \in \mathcal{M}$ *a discrete-discrete bispectral function*, i.e., if there exist difference operators $L(x,D)$ and $\Lambda(n,T)$ on Ω_1 and Ω_2, and functions

$$\theta(x) \quad \text{and} \quad \lambda(n),$$

such that

$$\begin{aligned} L(x,D)\psi(x,n) &= \lambda(n)\psi(x,n),\\ \Lambda(n,T)\psi(x,n) &= \theta(x)\psi(x,n), \end{aligned} \tag{2.1}$$

on $\Omega_1\times\Omega_2$. In fact, as we would be interested in VOP, we will consider only the case when $\theta(x)\equiv x$. We will assume that $\psi(x,n)$ is a nonsplit function of x and n in the sense that it satisfies the condition

$(**)$ there are no nonzero difference operators $L(x,\partial_x)$ and $\Lambda(n,T)$ that satisfy one of the above conditions with $f(n)\equiv 0$ or $\theta(x)\equiv 0$.

The assumption $(**)$ implies that the map $b\colon \mathcal{B}_1 \to \mathcal{B}_2$, given by $b(P(x,\partial_x)) := S(n,T)$ is a well defined algebra anti-isomorphism. Let us introduce the subalgebras K_i $i=1,2$ of $\mathcal{B}_i$ to be the algebras of functions in x (respectively in n). The algebra

$$A_1 := b^{-1}(K_2)$$

consists of the bispectral operators corresponding to $\psi(x,z)$ (i.e., difference operators in x having the properties (2.1)) and the algebra

$$A_2 := b(K_1)$$

consists of the bispectral operators corresponding to $\psi(x,n)$, i.e. difference operators in n having the properties (2.1)).

For the goals of VOP we are interested in the case when, for any fixed n, the function $\psi(x,n)$ defining the map b is a polynomial in x.

Let $\mathcal{R}_1$ be the algebra spanned over $\mathbb{C}$ by the operator $\hat{x}$ (multiplication by x), D and D^{-1}. Needless to say, the commutation relations in $\mathcal{R}_1$

$$[D,x] = D, \;\; [D^{-1},x] = -D^{-1}, \;\; [D,D^{-1}] = 0$$

play a crucial role.

In the same way we define another algebra $\mathcal{R}_2$, using the operators T, its inverse T^{-1} and the operator n of multiplication by the variable n. Finally the module $\mathcal{M}$ is a linear space of bivariate functions $f(x,n)$, where x and n are discrete variables. Next we define a subalgebra $\mathcal{B}_1 \subset \mathcal{R}_1$ as follows. Introduce the operators $\Delta = D-1$ and $\nabla = D^{-1}-1$. $\mathcal{B}_1$ will be spanned by the generators $\Delta, L = -x\nabla, \hat{x}$. It would also be convenient to introduce the element $f = \hat{x} - L$.

$$[\Delta, L] = \Delta, \;\; [L,f] = f, \;\; [\Delta,f] = 1. \tag{2.2}$$

□

In what follows we use the notation of the falling factorial:

$$(x)_k = x(x-1)\dots(x-k+1) \;\text{ for }\; k \in \mathbb{N}, \;\text{ and }\; (x)_0 = 1.$$

We notice that the notation $(x)_k$ is quite often used with a different meaning but here we will use it only in the above sense. Let $\psi(x,n) := S_n(x) := (x)_n$. Obviously

$$LS(x,n) = nS(x,n), \;\; (T+n)S(x,n) = xS(x,n).$$

In this way we can define the anti-automorphism b by

$$\begin{cases} b(f) = T \\ b(L) = n \\ b(\Delta) = nT^{-1}. \end{cases} \tag{2.3}$$

The image of $\mathcal{B}_1$ under the map b will be the algebra $\mathcal{B}_2$.

Next, following [4], we recall how to construct new bispectral operators from already known ones. The method is quite general and does not depend on the specific form of the operators. First, we remind the reader that for an operator

$L \in \mathcal{B}$ it is said that ad_L acts locally nilpotently on $\mathcal{B}$ when for any element $a \in \mathcal{B}$ there exists $k \in \mathbb{N}$, such that

$$\mathrm{ad}_L^k(a) = 0.$$

We formulate the simple observation from [4], needed in the present paper, in a form suitable for the discrete VOP.

Proposition 2.1. *Let $\mathcal{B}_1, \mathcal{B}_2$ be unital algebras with the properties described above. Let $L \in \mathcal{B}_1$ such that $\mathrm{ad}_L : \mathcal{B}_1 \to \mathcal{B}_1$ is a locally nilpotent operator and let $b\colon \mathcal{B}_1 \to \mathcal{B}_2$ be a bispectral involution. Suppose that, for any fixed n, $e^L\psi(x,n)$ is a polynomial in x of degree n. Define a new map $b^{'} : \mathcal{B}_1 \to \mathcal{B}_2$ via the new polynomial function $\psi'(x,n) := e^{\mathrm{ad}L}\psi(x,n)$. Then $b^{'} : \mathcal{B}_1 \to \mathcal{B}_2$ is a bispectral anti-involution.*

3. CHARLIER TYPE VECTOR ORTHOGONAL POLYNOMIALS

Here the algebras $\mathcal{B}_i$ are the ones defined in the previous section. Let $P(X)$ be a polynomial of degree $d \geq 1$ without a free term. We define the automorphism $\sigma : \mathcal{B}_1 \to \mathcal{B}_1$ by

$$\sigma = e^{\mathrm{ad}_{P(\Delta)}}.$$

Let us compute explicitly its action on the generators.

Lemma 3.1. *The automorphism σ acts on the generators as*

$$\begin{cases} \sigma(f) = f + P'(\Delta) \\ \sigma(L) = L + P'(\Delta)\Delta \\ \sigma(\Delta) = \Delta. \end{cases}$$

Proof. Starting with the relation $[\Delta, f] = 1$ we prove by induction that for each m

$$[\Delta^m, f] = m\Delta^{m-1}. \tag{3.1}$$

Really for $m = 1$ it is obvious. Assuming (3.1) is verified for $m = j - 1$, we have for $j = m$

$$\begin{aligned} [\Delta^m, f] &= \Delta^m f - f\Delta^m = \Delta\Delta^{m-1}f - f\Delta^m \\ &= \Delta(f\Delta^{m-1} + (m-1)\Delta^{m-1}) - f\Delta^m \\ &= [\Delta, f]\Delta^{m-1} + (m-1)\Delta^{m-1}) = m\Delta^{m-1}. \end{aligned}$$

Hence

$$e^{\mathrm{ad}_{P(\Delta)}}(f) = f + P'(\Delta),$$

as the rest of the terms vanish.

To prove the second formula we start with the identity $[\Delta, L] = \Delta$. By induction we see that

$$[\Delta^m, L] = m\Delta^m.$$

This proves the second formula. The last formula is obvious. □

A direct consequences of the lemma is

Corollary 3.1. *The image of x under the automorphism σ is:*

$$\sigma(x) = x + P'(\Delta)(1 + \Delta). \tag{3.2}$$

Proof. We use $x = f + L$. Hence

$$\sigma(f + L) = f + P'(\Delta) + L + P'(\Delta)\Delta = x + P'(\Delta)(1 + \Delta).$$

□

Let us define the anti-involution $b_1 = b(\sigma^{-1})$. Below we use that

$$\sigma^{-1} = \sum_{j=0}^{\infty} \frac{(-\mathrm{ad}_{P(\Delta)})^j}{j!}.$$

Also we define the difference operator

$$L_1 = \sigma(L) = -x\nabla + P'(\Delta)\Delta. \tag{3.3}$$

From Lemma 3.1 and Corollary 3.1 it follows almost immediately that

Lemma 3.2. *The anti-involution b_1 acts as*

$$\begin{cases} b_1(x) = T + n + P'(nT^{-1})(1 + nT^{-1}) \\ b_1(L_1) = n \\ b_1(\Delta) = nT^{-1}. \end{cases}$$

Proof. We have

$$b_1(x) = b(\sigma^{-1}(x)) = b(x + P'(\Delta)(1 + \Delta) = T + n + P'(nT^{-1})(1 + nT^{-1}).$$

Next,

$$b_1(L_1) = b(\sigma^{-1} \circ \sigma(L)) = b(L) = n$$

Finally,

$$b_1(\Delta) = b(\Delta) = nT^{-1}.$$

□

Let us define the "wave function"

$$C_n^P(x) = e^{P(\Delta))}\psi(x,n) = \sum_{j=0}^{\infty} \frac{P(\Delta)^j(x)_n}{j!}. \tag{3.4}$$

Notice that the operator Δ reduces the degrees of the polynomials. The same is true for $P(\Delta)$ (we recall that $P(X)$ has no free term). This shows that the sum (3.4) is finite and for this reason $C_n^P(x)$ is a polynomial.

Let us write explicitly $P(\Delta)$ as

$$P(\Delta) = \sum_{j=1}^{d} \beta_j \Delta^j.$$

We will list the basic properties of the polynomials $C_n^P(x)$ in terms of the polynomial $P(\Delta)$ in the next theorem.

Theorem 3.1. *The polynomials $C_n^P(x)$ have the following properties:*

(i) They satisfy $d+2$-term recurrence relation

$$xC_n^P(x) = C_{n+1}^P(x) + n(1+\beta_1)C_n^P(x) + \sum_{j=1}^{d}[(j+1)\beta_{j+1} + j\beta_j](n)_j C_{n-j}^P,$$

where $\beta_{d+1} = 0$.

(ii) They are eigenfunctions of the difference operator L_1 (3.3)

$$L_1 C_n^P(x) = nC_n^P(x).$$

(iii) They have a lowering operator

$$\Delta C_n^P(x) = nC_{n-1}^P(x)$$

Proof. (i) From Lemma 3.1 we have that

$$xC_n^P(x) = \left\{T + n + P'(nT^{-1})(1+nT^{-1})\right\}C_n^P.$$

Let us work out the expression $E = P'(nT^{-1})(1-nT^{-1})C_n^P$. We have

$$\begin{aligned} E &= \sum_{j=1}^{d} j\beta_j (nT^{-1})^{j-1} C_n^P + \sum_{j=1}^{d} j\beta_j (nT^{-1})^j C_n^P \\ &= \sum_{j=0}^{d}[(j+1)\beta_{j+1} + j\beta_j](nT^{-1})^j C_n^P. \end{aligned}$$

Using that $(nT^{-1})^j = (n)_j T^{-j}$ we obtain

$$E = \sum_{j=1}^{d} [(j+1)\beta_{j+1} + j\beta_j](n)_j C_{n-j}^P + \beta_1 C_n^P.$$

(ii) From the definitions of L_1 and $C_n^P(x)$ we obtain

$$L_1 C_n^P(x) = e^P L e^{-P} e^P \psi(x,n) = e^P n\psi(x,n) = nC_n^P(x).$$

(iii) follows directly from Lemma 3.2. □

4. MEIXNER TYPE VECTOR ORTHOGONAL POLYNOMIALS

We use the notation of the previous section $\hat{x}$, Δ, $L = -x\nabla$ to define an algebra $\mathcal{B}_1$ of discrete operators. It will be spanned by the operators $\hat{x}$, L, $G = (L+\beta)\Delta$, where β is a constant. Again it would be convenient to work with the element $f = \hat{x} - L$. They satisfy the following commutation relations

$$\begin{cases} [L, f] = f \\ [G, f] = 2L + \beta \\ [G, L] = G. \end{cases} \tag{4.1}$$

Also the wave function would be the same as in the previous section, namely $\psi(x,n) = (x)_n$. It has the properties:

$$\begin{cases} L\psi(x,n) = n\psi(x,n) \\ x\psi(x,n) = \psi(x,n+1) + n\psi(x,n) \\ G\psi(x,n) = (n-1+\beta)\psi(x,n-1). \end{cases}$$

We sum up these properties in terms of the following anti-involution b.

$$\begin{cases} b(L) = n \\ b(f) = T \\ b(G) = n(n-1+\beta)T^{n-1}. \end{cases}$$

The algebra $\mathcal{B}_2$ will be the image $b(\mathcal{B}_1)$. This gives our initial bispectral problem.

Let $P(X)$ be a polynomial of degree $d \geq 1$ without a free term. We define the automorphism $\sigma : \mathcal{B}_1 \to \mathcal{B}_1$ by

$$\sigma = e^{\mathrm{ad}_{P(G)}}.$$

In the next lemma we compute it on the generators.

Lemma 4.1. *The automorphism σ acts on the generators as*

$$\begin{cases} \sigma(f) = f + (2L + \beta)P'(G) + P''(G)G + P'^2(G)G \\ \sigma(L) = L + P'(G)G \\ \sigma(G) = G. \end{cases}$$

Proof. Let us start with the second formula. We have $[G, L] = G$. Then by induction we find

$$[G^m, L] = mG^m \tag{4.2}$$

Hence

$$\mathrm{ad}_{P(G)}L = P'(G)G.$$

which proves the second formula. Next we prove by induction that for each m

$$[G^m, f] = 2\sum_{j=0}^{m-1} G^j L G^{m-1-j} + m\beta G^{m-1}.$$

We use the above formula (4.2) in the form $G^j L = LG^j + jG^j$ to transform the first sum into

$$\sum_{j=0}^{m-1} G^j L G^{m-1-j} = 2\sum_{j=0}^{m-1}(L + j)G^{m-1}$$

This shows that

$$[G^m, f] = m(2L + \beta)G^{m-1} + m(m-1)G^{m-1},$$

which yields

$$\mathrm{ad}_{P(G)}(f) = (2L + \beta)P'(G) + P''(G)G.$$

Now we easily compute $\mathrm{ad}_P^2(f)$:

$$\mathrm{ad}_P^2(f) = [P(G), 2LP'(G)] = 2P'^2(G)G.$$

From the expressions for $\mathrm{ad}_P^j, j = 0, 1, 2$ we obtain the first identity.

The last identity is obvious. □

We define the anti-involution $b_1 = b(\sigma^{-1})$. Our bispectral operator L_1 will be given by

$$L_1 = \sigma(L) = -x\nabla + P'(G)G. \tag{4.3}$$

The next lemma computes the action of b_1 on the needed elements.

Lemma 4.2. *The anti-involution b_1 acts as*

$$\begin{cases} b_1(x) = T + n - [P'(u)(2n+\beta) + P''(u)u + P'(u)u - P'^2(u)u]_{|u=n(n-1+\beta)T^{-1}} \\ b_1(L_1) = n \\ b_1(G) = n(n-1+\beta)T^{-1}. \end{cases}$$

Proof. The last two identities are direct consequences of the definitions of L_1 and G together with the formulas for b. The more involved first identity follow from the last two and Lemma 4.1. Really, we have

$$\begin{aligned} b_1(x) = b_1(L) + b_1(f) &= b(L - P'(G)G) + \\ &+ b(f - [(2L+\beta)P'(G) + P''(G)G] + P'^2(G)G) \\ &= b(f+L) - b([(2L+\beta+G)P'(G) + P''(G)G] - P'^2(G)G)). \end{aligned}$$

after which we put the expressions for $b(G), b(f)$ and $b(L)$. □

We come to the definition of the VOP, i.e. the "wave function"

$$M_n^P(x) = e^{P(G))}\psi(x,n) = \sum_{j=0}^{\infty} \frac{P(G)^j(x)_n}{j!}. \tag{4.4}$$

We assume that $P(G) = \alpha G^m + \ldots$ with $\alpha \neq 0$.

Notice that the operator G reduces the degrees of the polynomials by one unit. This shows that the sum (4.4) is finite (we recall that $P(X)$ has no free term) and for this reason $M_n^P(x)$ is a polynomial.

The basic properties of the polynomials $M_n^P(x)$ are listed in the following theorem

Theorem 4.1. *Let $\beta \notin \mathbb{N}$. Then the polynomials $M_n^P(x)$ have the following properties:*

(i) They satisfy the recurrence relation

$$\begin{aligned} xM_n^P(x) &= M_{n+1}^P(x) + nM_n^P(x) \\ &- [P'(u)(2n+\beta+u) + P''(u)u - P'^2(u)u]_{u=n(n-1+\beta)T^{-1}} M_n^P. \end{aligned}$$

(ii) They are eigenfunctions of the difference operator L_1 (4.3)

$$L_1 M_n^P(x) = nM_n^P(x).$$

(iii) The operator G acts on them as lowering operator

$$GM_n^P(x) = n(n-1+\beta)M_{n-1}^P(x).$$

Proof. The proof is similar to the proof Theorem 3.1 and follows easily from Lemma 4.2 . Therefore it is omitted. □

Remark 4.1. Notice that the above polynomial system is well defined for all values of β but it is not always VOP. For example, when $\beta = -N,\ N \in \mathbb{N}$ and $d = 1$ we come to Kravchuk system of orthogonal polynomials, which contains only finite number of members. Similar situation occurs when $d > 1$. This is discussed in the next section.

5. EXAMPLES

Example 5.1. For the definitions and properties of orthogonal polynomials we follow mainly [17].

Let $P(\Delta) = -a\Delta$. We find that

$$L_1 = -a\Delta - x\nabla.$$

This is the difference operator which has Charlier polynomials as eigenfunctions. The latter can be defined according to our scheme by

$$C_n^P = e^{-a\Delta}(x)_n = \sum_{j=0}^{n} \frac{(-a)^j \Delta^j (x)_n}{j!} = \sum \frac{(-a)^j (-n)_j (x)_{n-j}}{j!}.$$

We see that these are the normed Charlier polynomials denoted in [17] by $p_n(x)$. This example, together with the construction of Appell polynomials in [16], motivates the name "Charlier-Appell polynomials" for general P.

Example 5.2. In the second example we take the algebra from section 4. Let us take $P(G) = \alpha G$, G $= (-x\nabla + \beta)\Delta$. Then

$$L_1 = \alpha(-x\nabla + \beta)\Delta - x\nabla.$$

Let c be a constant, $c \neq 0, 1$. Put $\hat{L} = (c-1)L_1$. Take the constant α to be

$$\alpha = \frac{c}{c-1}$$

and β to be different from a negative integer. We obtain exactly the Meixner operator

$$\hat{L} = c(x+\beta)\Delta + x\nabla.$$

as given in [17]. It has eigenvalues $(c-1)n$.

In case $\beta = -N,\ N \in \mathbb{N}$ we obtain Kravchuk polynomials $K_0, \ldots, K_N$, which form a finite set of orthogonal polynomials.

Example 5.3. Let us again use the settings from section 4. We present here the simplest new example. Let us take $P(G) = G^2/2$. Then the new polynomials

$$M_n^P(x) = \sum_{j=0}^{\infty} \frac{(G^2)^j (x)_n}{j!}$$

are eigenfunctions of the operator L_1

$$L_1 M_n^P(x) = (-x\nabla + (x\nabla\Delta + \beta\Delta)^2) M_n^P(x) = -n M_n^P(x)$$

The recurrence relation reads

$$\begin{aligned} xM_n^P &= M_{n+1}^P + nM_n^P - n(n+\beta-1)(2n-1+\beta)M_{n-1}^P \\ &- (n)_2(n+\beta)_2 M_{n-2}^P + (n)_3(n+\beta)_3 M_{n-3}^P. \end{aligned}$$

Example 5.4. Kravchuk-like polynomials. In this example we investigate the case when $\beta = -N$, $N \in \mathbb{N}$. We take $P(G) = G^2/2$. The recurrence relation is as above.

We see that the polynomials satisfy 5-term recurrence relation. However the coefficient at M_{n-3} is zero for $n = N$, thus violating the condition of P. Maroni's theorem [22]. This shows that the vector orthogonality is valid only for the polynomials $M_n, n = 0, \ldots, N$. The situation with the general polynomial $P(G)$ is similar.

6. REFERENCES

[1] AI-Salam, W. A.: Characterization theorems for orthogonal polynomials. In: *Orthogonal Polynomials: Theory and Practice*, edited by P. Nevai, with the assistance of M. E. H. Ismail, Proceedings of the NATO Advanced Study Institute on Orthogonal Polynomials and Their Applications Colombus, Ohio, U.S.A. May 22 - June 3, 1989. Kluwer Academic publishers, Dordrecht / Boston / London.

[2] Aptekarev, A. I., Kuijlaars, A.: Hermite-Padé approximations and multiple orthogonal polynomial ensembles. *Uspekhi Mat. Nauk*, **66**, no. 6(402), 2011, 123–190.

[3] Arvesú, J., Coussement, J., Van Assche, W.: Some discrete multiple orthogonal polynomials. *J. Comp. Appl. Math.*, **153**, 2003, 19–45.

[4] Bakalov, B., Horozov, E., Yakimov, M.: General methods for constructing bispectral operators. *Phys. Lett. A*, **222**, no. 1-2, 1996, 59–66.

[5] Ben Cheikh, Y., Zaghouani, A.: Some discrete d-orthogonal polynomial sets. *J. Comput. Appl. Math.*, **156**, 2003, 253–263.

[6] Bleher, P., Delvaux, S., Kuijlaars, A. B. J.: Random matrix model with external source and a constrained vector equilibrium problem. *Comm. Pure Appl. Math.*, **64**, 2011, 116–160.

[7] Bochner, S.: Über Sturm-Liouvillesche Polynomsysteme. *Math. Z.*, **29**, 1929, 730–736.

[8] Duistermaat, J. J., Grünbaum, F. A.: Differential equation in the spectral parameter. *Commun. Math. Phys.*, **103**, 1998, 177–240.

[9] Durán, A. J., de la Iglesia, M. D.: Constructing bispectral orthogonal polynomials from the classical discrete families of Charlier, Meixner and Kravchuk. *Constr. Approx.*, **42**, 2015, 49–91.

[10] Genest, V. X., Vinet, L., Zhedanov, A.: d-Orthogonal polynomials and su(2). *J. Math. Anal. Appl.*, **390**, no. 2, 2012, 472–487.

[11] Grünbaum, F. A., Haine, L.: A theorem of Bochner, revisited. In: *Algebraic Aspects of Integrable Systems* (I. M. Gel'fand and T. Fokas, eds.), Progress in Nonlinear Differential Equations and Their Applications, **26**, Volume in Honor of I. Dorfman, Birkhäuser Boston - Basel - Berlin, 1997, pp. 143–172.

[12] Grünbaum, F. A., Haine, L., Horozov, E.: Some functions that generalize the Krall-Laguerre polynomials. *J. Comp. Appl. Math.*, **106**, no. 2, 1999, 271–297.

[13] Grünbaum, F. A., Yakimov, M.: Discrete bispectral Darboux transformations from Jacobi operators. *Pacific J. Math.*, **204**, no. 2, 2002, 395–431.

[14] Van Assche, W.: Difference Equations for Multiple Charlier and Meixner Polynomials. In: *Proceedings of the Sixth International Conference on Difference Equations Augsburg, Germany 2001* (S. Elaydi , B. Aulbach , and G. Ladas, eds.), New Progress in Difference Equations, CRC Press 2004, pp. 549–557.

[15] Hildebrandt, E. H.: Systems of polynomials connected with the Charlier expansion and the Pearson differential equation. *Ann. Math. Stat.*, **2**, 1931, 379–439.

[16] Horozov, E.: Automorphisms of algebras and Bochner's property for vector orthogonal polynomials. *SIGMA*, **12**, 2016, 050, 14 pages; arXiv:1512.03898 http://dx.doi.org/10.3842/SIGMA.2016.050, Special Issue on Orthogonal Polynomials, Special Functions and Applications.

[17] Koekoek, R., Lesky, P. A., Swarttouw, R. F.: *Hypergeometric Orthogonal Polynomials and Their q-Analogues*, Springer Heidelberg Dordrecht London New York, 2010.

[18] Krall, H. L.: *On orthogonal polynomials satisfying a certain fourth order differential equation.* The Pennsylvania State College Studies, no.6, The Pennsylvania State College, State College, PA, 1940.

[19] Lancaster, O. E.: Orthogonal polynomials defined by difference equations. *American J. Math.*, **63**, 1941, 185–207.

[20] Lee, D. W.: Difference equations for discrete classical multiple orthogonal polynomials. *J. Approx. Theory*, **150**, no. 2, 2008, 132–152.

[21] Lesky, P.: Über Polynomsysteme, die Sturm-Liouvilleschen Differenzengleichungen genügen. *Math. Z.*, **78**, 1962, 439–445.

[22] Maroni, P.: L'orthogonalité et les récurrences de polynômes d'ordre supérieur à deux. *Ann. Fac. Sci. Toulouse*, **10**, no. 1, 1989, 105–139.

[23] Tater, M. (joint with B. Shapiro): Asymptotic zero distribution of polynomial solutions to degenerate exactly-solvable equations. Talk at CRM-ICMAT Workshop on Exceptional Orthogonal polynomials and exact solutions in Mathematical Physics, Segovia, 2014 (private communication).

[24] Van Iseghem, J.: Vector orthogonal relations. Vector QD-algorithm. *J. Comp. Appl. Math.*, **19**, no. 1, suppl. 1, 1987, 141–150.

[25] Vinet, L., Zhedanov, A.: Automorphisms of the Heisenberg-Weyl algebra and d-orthogonal polynomials. *J. Math. Phys.*, **50**, 2009, 033511–033511-19.

Received on December 23, 2016

Emil Horozov

Department of Mathematics and Informatics
University of Sofia
5 James Bourchier Blvd.
1164 Sofia
BULGARIA

E-mail: horozov@fmi.uni-sofia.bg

and

Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Acad. G. Bonchev Str., Block 8
1113 Sofia, BULGARIA

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 104

ANNUAL OF SOFIA UNIVERSITY „ST. KLIMENT OHRIDSKI“
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 104

LOWER BOUNDING THE FOLKMAN NUMBERS $F_v(a_1, \ldots, a_s; m-1)$

ALEKSANDAR BIKOV, NEDYALKO NENOV

For a graph G the expression $G \overset{v}{\to} (a_1, \ldots, a_s)$ means that for every s-coloring of the vertices of G there exists $i \in \{1, \ldots, s\}$ such that there is a monochromatic a_i-clique of color i. The vertex Folkman numbers

$$F_v(a_1, \ldots, a_s; m-1) = \min\{|V(G)| : G \overset{v}{\to} (a_1, \ldots, a_s) \text{ and } K_{m-1} \not\subseteq G\}.$$

are considered, where $m = \sum_{i=1}^{s}(a_i - 1) + 1$. We know the exact values of all the numbers $F_v(a_1, \ldots, a_s; m-1)$ when $\max\{a_1, \ldots, a_s\} \leq 6$ and also the number $F_v(2, 2, 7; 8) = 20$. In [1] we present a method for obtaining lower bounds on these numbers. With the help of this method and a new improved algorithm, in the special case when $\max\{a_1, \ldots, a_s\} = 7$ we prove that $F_v(a_1, \ldots, a_s; m-1) \geq m + 11$ and this bound is exact for all m. The known upper bound for these numbers is $m + 12$. At the end of the paper we also prove the lower bounds $19 \leq F_v(2, 2, 2, 4; 5)$ and $29 \leq F_v(7, 7; 8)$.

Keywords: Folkman number, clique number, independence number, chromatic number.

2000 Math. Subject Classification: 05C35.

1. INTRODUCTION

Only finite, non-oriented graphs without loops and multiple edges are considered in this paper. $G_1 + G_2$ denotes the graph G for which $\mathrm{V}(G) = \mathrm{V}(G_1) \cup \mathrm{V}(G_2)$ and $\mathrm{E}(G) = \mathrm{E}(G_1) \cup \mathrm{E}(G_2) \cup E'$, where $E' = \{[x, y] : x \in \mathrm{V}(G_1), y \in \mathrm{V}(G_2)\}$, i.e.

G is obtained by connecting with an edge every vertex of G_1 to every vertex of G_2. All undefined terms can be found in [19].

Let $a_1, \ldots, a_s$ be positive integers. The expression $G \overset{v}{\to} (a_1, \ldots, a_s)$ means that for every coloring of $\mathrm{V}(G)$ in s colors (s-coloring) there exists $i \in \{1, \ldots, s\}$ such that there is a monochromatic a_i-clique of color i. In particular, $G \overset{v}{\to} (a_1)$ means that $\omega(G) \geq a_1$. Further, for convenience, instead of $G \overset{v}{\to} (\underbrace{2, \ldots, 2}_{r})$ we write $G \overset{v}{\to} (2_r)$ and instead of $G \overset{v}{\to} (\underbrace{2, \ldots, 2}_{r}, a_1, \ldots, a_s)$ we write $G \overset{v}{\to} (2_r, a_1, \ldots, a_s)$.

Set

$$\mathcal{H}(a_1, \ldots, a_s; q) := \left\{ G : G \overset{v}{\to} (a_1, \ldots, a_s) \text{ and } \omega(G) < q \right\};$$
$$\mathcal{H}(a_1, \ldots, a_s; q; n) := \{ G : G \in \mathcal{H}(a_1, \ldots, a_s; q) \text{ and } |\mathrm{V}(G)| = n \}.$$

The vertex Folkman number $F_v(a_1, \ldots, a_s; q)$ is defined by the equality:

$$F_v(a_1, \ldots, a_s; q) = \min \{ |\mathrm{V}(G)| : G \in \mathcal{H}(a_1, \ldots, a_s; q) \}.$$

The graph G is called an extremal graph in $\mathcal{H}(a_1, \ldots, a_s; q)$ if $G \in \mathcal{H}(a_1, \ldots, a_s; q)$ and $|\mathrm{V}(G)| = F_v(a_1, \ldots, a_s; q)$. We denote by $\mathcal{H}_{extr}(a_1, \ldots, a_s; q)$ the set of all extremal graphs in $\mathcal{H}(a_1, \ldots, a_s; q)$.

Folkman proved in [6] that:

$$F_v(a_1, \ldots, a_s; q) \text{ exists } \Leftrightarrow q > \max \{a_1, \ldots, a_s\}. \tag{1.1}$$

Other proofs of (1.1) are given in [5] and [8]. In the special case $s = 2$, a very simple proof of this result is given in [12] with the help of corona product of graphs.

Obviously, $F_v(a_1, \ldots, a_s; q)$ is a symmetric function of $a_1, \ldots, a_s$, and if $a_i = 1$, then

$$F_v(a_1, \ldots, a_s; q) = F_v(a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_s; q).$$

Therefore, it suffices to consider only such Folkman numbers $F_v(a_1, \ldots, a_s; q)$ for which

$$2 \leq a_1 \leq \cdots \leq a_s. \tag{1.2}$$

We call the numbers $F_v(a_1, \ldots, a_s; q)$ for which inequalities (1.2) hold canonical vertex Folkman numbers.

In [9] for arbitrary positive integers $a_1, \ldots, a_s$ the following terms are defined

$$m(a_1, \ldots, a_s) = m = \sum_{i=1}^{s} (a_i - 1) + 1 \quad \text{and} \quad p = \max \{a_1, \ldots, a_s\}. \tag{1.3}$$

It is easy to see that $K_m \overset{v}{\to} (a_1, \ldots, a_s)$ and $K_{m-1} \overset{v}{\not\to} (a_1, \ldots, a_s)$. Therefore

$$F_v(a_1, \ldots, a_s; q) = m, \quad q \geq m + 1.$$

The following theorem for the numbers $F_v(a_1, \ldots, a_s; m)$ is true:

Theorem 1.1. *Let $a_1, \ldots, a_s$ be positive integers and let m and p be defined by the equalities* (1.3). *If $m \geq p+1$, then:*

(a) $F_v(a_1, \ldots, a_s; m) = m + p$, ([9, 8]);

(b) $K_{m+p} - C_{2p+1} = K_{m-p-1} + \overline{C}_{2p+1}$ *is the only extremal graph in* $\mathcal{H}(a_1, \ldots, a_s; m)$, ([8]).

The condition $m \geq p+1$ is necessary according to (1.1). Other proofs of Theorem 1.1 are given in [13] and [14].

Very little is known about the numbers $F_v(a_1, \ldots, a_s; m-1)$. According to (1.1) we have

$$F_v(a_1, \ldots, a_s; m-1) \text{ exists } \Leftrightarrow m \geq p+2. \tag{1.4}$$

The following general bounds are known:

$$m + p + 2 \leq F_v(a_1, \ldots, a_s; m-1) \leq m + 3p, \tag{1.5}$$

where the lower bound is true if $p \geq 2$ and the upper bound is true if $p \geq 3$. The lower bound is obtained in [13] and the upper bound is obtained in [7]. In the border case $m = p+2$ the upper bounds in (1.5) are significantly improved in [18].

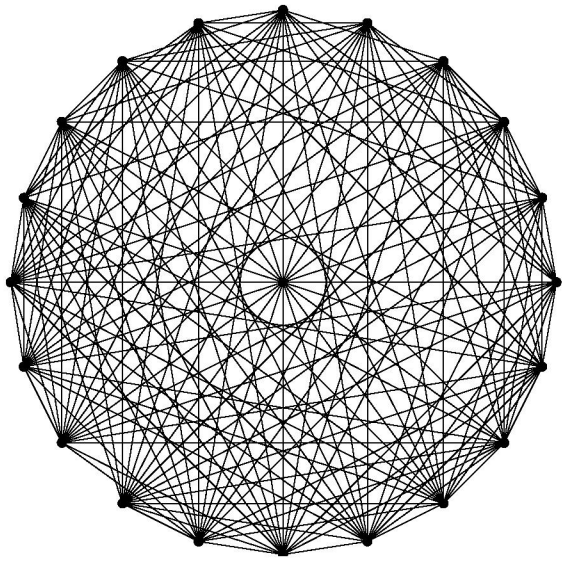


Figure 1: 20-vertex graph in $\mathcal{H}(2, 2, 7; 8)$

We know all the numbers $F_v(a_1, \ldots, a_s; m-1)$ when $\max\{a_1, \ldots, a_s\} \leq 6$, see [4] for details. When $\max\{a_1, \ldots, a_s\} = 7$ it is known that $F_v(2, 2, 7; 8) = 20$ and

$$m + 10 \leq F_v(a_1, \ldots, a_s; m-1) \leq m + 12.$$

The lower bound $F_v(2,2,7;8) \geq 20$ is obtained with the help of Algorithm 3.5, and the upper bound is obtained by constructing 20-vertex graphs in $\mathcal{H}(2,2,7;8)$. An example for such a graph is given in Figure 1.

In this paper we present an algorithm (Algorithm 3.9), with the help of which we can obtain lower bounds on the numbers $F_v(a_1,\ldots,a_s;m-1)$. Using Algorithm 3.9 and $F_v(2,2,7;8) = 20$, we improve the lower bound on the numbers $F_v(a_1,\ldots,a_s;m-1)$ when $\max\{a_1,\ldots,a_s\} = 7$ by proving the following:

Main Theorem. *Assume that $a_1,\ldots,a_s$ are positive integers such that $\max\{a_1,\ldots,a_s\} = 7$ and $m = \sum\limits_{i=1}^{s}(a_i-1)+1 \geq 9$. Then*

$$F_v(a_1,\ldots,a_s;m-1) \geq m+11.$$

Remark 1.2. As is seen from (1.4), the condition $m \geq 9$ in the Main Theorem is necessary.

2. BOUNDS ON THE NUMBERS $F_v(a_1,\ldots,a_s;q)$

Let m and p be positive integers. Denote by $\mathcal{S}(m,p)$ the set of all $(b_1,\ldots,b_r)$ (r is not fixed), where b_i are positive integers such that $\max\{b_1,\ldots,b_r\} = p$ and $\sum_{i=1}^{r}(b_i-1)+1 = m$. Let $(a_1,\ldots,a_s) \in \mathcal{S}(m,p)$. Then obviously

$$\min_{(b_1,\ldots,b_r)\in\mathcal{S}(m,p)} F_v(b_1,\ldots,b_r;q) \leq F_v(a_1,\ldots,a_s;q) \leq \max_{(b_1,\ldots,b_r)\in\mathcal{S}(m,p)} F_v(b_1,\ldots,b_r;q).$$

Note that $(2_{m-p},p) \in \mathcal{S}(m,p)$, $p \geq 2$ and it is easy to prove that

$$\min_{(b_1,\ldots,b_r)\in\mathcal{S}(m,p)} F_v(b_1,\ldots,b_r;q) = F_v(2_{m-p},p;q) \quad \text{(see [1])}.$$

We see that the lower bounding of the vertex Folkman numbers can be achieved by computing or lower bounding the numbers $F_v(2_{m-p},p;q)$. In general, this is a hard problem. However, in the case $q = m-1$, in [1] we presented a method for the computation of these numbers, which is based on the following:

Theorem 2.1 ([1]). *Let $r_0 = r_0(p)$ be the smallest positive integer for which*

$$\min_{r\geq 2}\{F_v(2_r,p;r+p-1)-r\} = F_v(2_{r_0},p;r_0+p-1)-r_0.$$

Then:

(a) $F_v(2_r,p;r+p-1) = F(2_{r_0},p;r_0+p-1)+r-r_0,\ r \geq r_0$.

(b) If $r_0 = 2$, *then* $F_v(2_r,p;r+p-1) = F_v(2,2,p;p+1)+r-2,\ r \geq 2$.

(c) If $r_0 > 2$ and G is an extremal graph in $\mathcal{H}(2_{r_0}, p; r_0 + p - 1)$, then $G \overset{v}{\to} (2, r_0 + p - 2)$.

(d) $r_0 < F_v(2, 2, p; p + 1) - 2p$.

From this theorem it becomes clear that, for a fixed p, the computation of the members of the infinite sequence $F_v(2_{m-p}, p; m - 1)$, $m \geq p + 2$, is reduced to the computation of its first r_0 members, where $r_0 < F_v(2, 2, p; p + 1) - 2p$. We conjecture that it is enough to know only its first member $F_v(2, 2, p; p + 1)$.

Conjecture 2.2 ([1]). *If $p \geq 4$, then*

$$\min_{r \geq 2} \{F_v(2_r, p; r + p - 1) - r\} = F_v(2, 2, p; p + 1) - 2,$$

i.e., $r_0(p) = 2$, and therefore

$$F_v(2_r, p; r + p - 1) = F_v(2, 2, p; p + 1) + r - 2, \quad r \geq 2.$$

This conjecture is proved for $p = 4, 5$ and 6 in [16], [1] and [4], respectively. In [4] it is also proved that the conjecture is true when $F_v(2, 2, p; p + 1) \leq 2p + 5$. In this paper we will prove that Conjecture 2.2 is also true when $p = 7$:

Theorem 2.3. $F_v(2_{m-7}, 7; m - 1) = m + 11$.

The Main Theorem follows easily from Theorem 2.3.

Remark 2.4. This method is not suitable for obtaining upper bounds for the vertex Folkman numbers, as it is not clear how $\max_{(b_1,\ldots,b_r) \in \mathcal{S}(m,p)} F_v(b_1, \ldots, b_r; q)$ can be computed or bounded. In [2] we present another method for upper bounding of the vertex Folkman numbers (see also [1] and [4]).

3. ALGORITHMS

Finding all graphs in $\mathcal{H}(a_1, \ldots, a_s; q; n)$ using a brute force approach is practically impossible for $n > 13$. In this section we present algorithms for obtaining these graphs.

We say that G is a maximal graph in $\mathcal{H}(a_1, \ldots, a_s; q)$ if $G \in \mathcal{H}(a_1, \ldots, a_s; q)$ but $G + e \notin \mathcal{H}(a_1, \ldots, a_s; q), \forall e \in \mathrm{E}(\overline{G})$, i.e. $\omega(G + e) = q, \forall e \in \mathrm{E}(\overline{G})$. The graphs in $\mathcal{H}(a_1, \ldots, a_s; q)$ can be obtained by removing edges from the maximal graphs in this set.

For convenience, we also define the following term:

Definition 3.1. The graph G is called a $(+K_t)$-graph if $G + e$ contains a new t-clique for all $e \in \mathrm{E}(\overline{G})$.

Obviously, $G \in \mathcal{H}(a_1, \ldots, a_s; q)$ is a maximal graph in $\mathcal{H}(a_1, \ldots, a_s; q)$ if and only if G is a $(+K_q)$-graph. We shall denote by $\mathcal{H}_{+K_t}(a_1, \ldots, a_s; q)$ the set of all $(+K_t)$-graphs in $\mathcal{H}(a_1, \ldots, a_s; q)$, and by $\mathcal{H}_{max}(a_1, \ldots, a_s; q)$ all maximal K_q-free graphs in this set. The sets $\mathcal{H}_{max}(a_1, \ldots, a_s; q; n)$ and $\mathcal{H}_{+K_t}(a_1, \ldots, a_s; q; n)$ are defined in the same way as $\mathcal{H}(a_1, \ldots, a_s; q; n)$.

We shall denote by $\mathcal{H}^t_{max}(a_1, \ldots, a_s; q; n)$ and $\mathcal{H}^t_{+K_t}(a_1, \ldots, a_s; q; n)$ the subsets of all graphs with independence number not greater than t in the sets $\mathcal{H}_{max}(a_1, \ldots, a_s; q; n)$ and $\mathcal{H}_{+K_t}(a_1, \ldots, a_s; q; n)$, respectively.

Remark 3.2. In the special case $s = 1$ we have

$\mathcal{H}(a_1; q; n) = \{G : a_1 \leq \omega(G) < q \text{ and } |\,\mathrm{V}(G)| = n\}$.

Obviously, if $a_1 \leq n \leq q - 1$, then $\mathcal{H}_{max}(a_1; q; n) = \{K_n\}$.

If $a_1 \leq q - 1 \leq n$, then $\mathcal{H}_{max}(a_1; q; n) = \mathcal{H}_{max}(q - 1; q; n)$.

Further, we shall use the following propositions, which are easy to prove:

Proposition 3.3 ([4]). *Let G be a graph, $G \xrightarrow{v} (a_1, ..., a_s)$ and $a_i \geq 2$. Then for every independent set A in G*

$$G - A \xrightarrow{v} (a_1, \ldots, a_{i-1}, a_i - 1, a_{i+1}, \ldots, a_s).$$

Proposition 3.4 ([4]). *Let $G \in \mathcal{H}_{max}(a_1, \ldots, a_s; q; n)$ and A be an independent set of vertices of G. Then $G - A \in \mathcal{H}_{+K_{q-1}}(a_1 - 1, \ldots, a_s; q; n - |A|)$.*

The following algorithm for finding all graphs $G \in \mathcal{H}_{max}(a_1, \ldots, a_s; q; n)$ with $r \leq \alpha(G) \leq t$ is given in [4]:

Algorithm 3.5 ([4]). The set $\mathcal{A} = \mathcal{H}^t_{max}(a_1 - 1, \ldots, a_s; q; n - r)$ is the input of the algorithm. The output of the algorithm is the set $\mathcal{B}$ of all graphs $G \in \mathcal{H}^t_{max}(a_1, \ldots, a_s; q; n)$ with $\alpha(G) \geq r$.

1. By removing edges from the graphs in $\mathcal{A}$ obtain the set

$\mathcal{A}' = \mathcal{H}^t_{+K_{q-1}}(a_1 - 1, \ldots, a_s; q; n - r)$.

2. For each graph $H \in \mathcal{A}'$:

2.1. Find the family $\mathcal{M}(H) = \{M_1, \ldots, M_l\}$ of all maximal K_{q-1}-free subsets of $\mathrm{V}(H)$.

2.2. Find all r-element multisets $N = \{M_{i_1}, M_{i_2}, \ldots, M_{i_r}\}$ of elements of $\mathcal{M}(H)$, which fulfill the conditions:

(a) $K_{q-2} \subseteq M_{i_j} \cap M_{i_k}$ for every $M_{i_j}, M_{i_k} \in N$.

(b) $\alpha(H - \bigcup_{M_{i_j} \in N'} M_{i_j}) \leq t - |N'|$ for every subset N' of N.

2.3. For each r-element multiset $N = \{M_{i_1}, M_{i_2}, \ldots, M_{i_r}\}$ of elements of $\mathcal{M}(H)$ found in step 2.2 construct the graph $G = G(N)$ by adding new independent vertices $v_1, v_2, \ldots, v_r$ to $\mathrm{V}(H)$ such that $N_G(v_j) = M_{i_j}, j = 1, \ldots, r$. If $\omega(G + e) = q, \forall e \in \mathrm{E}(\overline{G})$, then add G to $\mathcal{B}$.

3. Remove the isomorphic copies of graphs from $\mathcal{B}$.

4. Remove from the obtained in step 3 set $\mathcal{B}$ all graphs G for which $G \overset{v}{\not\to} (a_1, \ldots, a_s)$.

Theorem 3.6 ([4]). *After the execution of Algorithm 3.5, the obtained set $\mathcal{B}$ coincides with the set of all graphs $G \in \mathcal{H}^t_{max}(a_1, \ldots, a_s; q; n)$ with $\alpha(G) \geq r$.*

Algorithm 3.5 is based on a very similar algorithm that we used in [3] to prove the lower bound $F_e(3, 3; 4) > 19$. It is possible to prove the Main Theorem using Algorithm 3.5, but it would take months of computational time. For this reason, we will present an algorithm which is a modification of Algorithm 3.5 and helped us prove the Main Theorem in less than a week work of a personal computer.

Further we shall use the following term:

Definition 3.7. We say that v is a cone vertex in the graph G if v is adjacent to all other vertices in G.

Suppose that $G \in \mathcal{H}_{max}(a_1, \ldots, a_s; q; n)$ and G has a cone vertex, i.e. $G = K_1 + H$. According to Proposition 3.3, $H \in \mathcal{H}_{max}(a_1 - 1, \ldots, a_s; q - 1; n - 1)$. Therefore, if we know all the graphs in $\mathcal{H}_{max}(a_1 - 1, \ldots, a_s; q - 1; n - 1)$, we can easily obtain the graphs in $\mathcal{H}_{max}(a_1, \ldots, a_s; q; n)$, which have a cone vertex. We will use this fact to modify Algorithm 3.5 and make it faster in the case where all graphs in $\mathcal{H}_{max}(a_1 - 1, \ldots, a_s; q - 1; n - 1)$ are already known. The new modified algorithm is based on the following:

Proposition 3.8. *Let $G \in \mathcal{H}_{max}(a_1, \ldots, a_s; q; n)$ be a graph without cone vertices and A be an independent set in G such that $G - A$ has a cone vertex, i.e. $G - A = K_1 + H$. Then $G = \overline{K}_{r+1} + H$, where $r = |A|$, H has no cone vertices and $K_1 + H \in \mathcal{H}_{max}(a_1, \ldots, a_s; q; n - r)$.*

Proof. Let $A = \{v_1, \ldots, v_r\}$ be an independent set in G and $G - A = K_1 + H = \{u\} + H$. Since G has no cone vertices, there exist $v_i \in A$ such that v_i is not adjacent to u. Then $N_G(v_i) \subseteq N_G(u)$ and since G is a maximal K_q-free graph, we obtain $N_G(v_i) = N_G(u) = \mathrm{V}(H)$. Hence, u is not adjacent to any of the vertices in A, and therefore $N_G(v_j) = N_G(u) = \mathrm{V}(H), \forall v_j \in A$. We derived $G = \overline{K}_{r+1} + H$. The graph H has no cone vertices, since any cone vertex in H would be a cone vertex in G. It is easy to see that if $\overline{K}_{r+1} + H \overset{v}{\to} (a_1, \ldots, a_s)$, then $K_1 + H \overset{v}{\to} (a_1, \ldots, a_s)$. Therefore $K_1 + H \in \mathcal{H}_{max}(a_1, \ldots, a_s; q; n - r)$. □

Now we present the main algorithm used in this paper, which is a modification of Algorithm 3.5.

Algorithm 3.9. The input of the algorithm are the set $\mathcal{A}_1 = \mathcal{H}^t_{max}(a_1 - 1, \ldots, a_s; q; n - r)$ and the set $\mathcal{A}_2 = \mathcal{H}^t_{max}(a_1 - 1, \ldots, a_s; q - 1; n - 1)$. The output of the algorithm is the set $\mathcal{B}$ of all graphs $G \in \mathcal{H}^t_{max}(a_1, \ldots, a_s; q; n)$ with $\alpha(G) \geq r$.

1. By removing edges from the graphs in $\mathcal{A}_1$ obtain the set

$$\mathcal{A}'_1 = \left\{ H \in \mathcal{H}^t_{+K_{q-1}}(a_1 - 1, \ldots, a_s; q; n - r) : \ H \text{ has no cone vertices} \right\}.$$

2. Repeat step 2 of Algorithm 3.5.

3. Repeat step 3 of Algorithm 3.5.

4. Repeat step 4 of Algorithm 3.5.

5. If $t > r$, find the subset $\mathcal{A}_1''$ of $\mathcal{A}_1$ containing all graphs with exactly one cone vertex. For each graph $H \in \mathcal{A}_1''$, if $K_1 + H \xrightarrow{v} (a_1, ..., a_s)$, then add $\overline{K}_{r+1} + H$ to $\mathcal{B}$.

6. For each graph H in $\mathcal{A}_2$ such that $\alpha(H) \geq r$, if $K_1 + H \xrightarrow{v} (a_1, \ldots, a_s)$, then add $K_1 + H$ to $\mathcal{B}$.

Theorem 3.10. *After the execution of Algorithm 3.9, the obtained set $\mathcal{B}$ coincides with the set of all graphs $G \in \mathcal{H}^t_{max}(a_1, \ldots, a_s; q; n)$ with $\alpha(G) \geq r$.*

Proof. Suppose that after the execution of Algorithm 3.9, $G \in \mathcal{B}$. If after step 4 $G \in \mathcal{B}$, then according to Theorem 3.6, $G \in \mathcal{H}^t_{max}(a_1, \ldots, a_s; q; n)$ and $\alpha(G) \geq r$. If G is added to $\mathcal{B}$ in step 5 or step 6, then clearly $G \in \mathcal{H}^t_{max}(a_1, \ldots, a_s; q; n)$ and $\alpha(G) \geq r$.

Now let $G \in \mathcal{H}^t_{max}(a_1, \ldots, a_s; q; n)$ and $\alpha(G) \geq r$. If $G = K_1 + H$ for some graph H, then, according to Proposition 3.3, $H \in \mathcal{A}_2$ and in step 6 G is added to $\mathcal{B}$. Suppose that G has no cone vertices and G has an independent set A such that $|A| = r$ and $G - A$ has a cone vertex, i.e. $G - A = K_1 + H$. Then, according to Proposition 3.8, $G = \overline{K}_{r+1} + H$, $K_1 + H$ has exactly one cone vertex and $K_1 + H \xrightarrow{v} (a_1, \ldots, a_s)$. It is clear that $t > r$ and hence in step 5 G is added to $\mathcal{B}$. Finally, if $G - A$ has no cone vertices, then according to Proposition 3.4, $G - A \in \mathcal{A}_1'$ and it follows from Theorem 3.6 that after the execution of step 4, $G \in \mathcal{B}$. □

Remark 3.11. Note that if $n \geq q$ and $r = 2$, then Algorithms 3.5 and 3.9 obtain all graphs in $G \in \mathcal{H}^t_{max}(a_1, ..., a_s; q; n)$.

The *nauty* programs [10] have an important role in this paper. We use them for fast generation of non-isomorphic graphs and isomorphic rejection.

4. PROOF OF THE MAIN THEOREM AND THEOREM 2.3

We will first prove Theorem 2.3 by proving Conjecture 2.2 in the case $p = 7$. Since $F_v(2, 2, 7; 8) = 20$ [4], in view of Theorem 2.1(d), to prove the conjecture in this case we need to prove the inequalities $F_v(2, 2, 2, 7; 9) > 20$, $F_v(2, 2, 2, 2, 7; 10) > 21$ and $F_v(2, 2, 2, 2, 2, 7; 11) > 22$. It is easy to see that it is enough to prove only the last of the three inequalities (see [4] for details). Using Algorithm 3.5 it can be proved that $F_v(2, 2, 2, 2, 2, 7; 11) > 22$, but it would require a lot of computational time. Instead, we will prove the three inequalities successively using Algorithm 3.9. Only the proof of the first inequality is presented in details, since the proofs of the others are very similar. We will show that $\mathcal{H}(2, 2, 7; 8; 19) = \emptyset$. The proof uses the graphs $\mathcal{H}^3_{max}(4; 8; 8)$, $\mathcal{H}^3_{max}(5; 8; 10)$, $\mathcal{H}^3_{max}(6; 8; 12)$, $\mathcal{H}^3_{max}(7; 8; 14)$,

$\mathcal{H}^3_{max}(2,7;8;16)$, $\mathcal{H}^3_{max}(2,2,7;8;19)$, $\mathcal{H}^2_{max}(4;8;9)$, $\mathcal{H}^2_{max}(5;8;11)$, $\mathcal{H}^2_{max}(6;8;13)$, $\mathcal{H}^2_{max}(7;8;15)$, $\mathcal{H}^2_{max}(2,7;8;17)$, $\mathcal{H}^2_{max}(2,2,7;8;19)$ obtained in [4] in the proof of the lower bound $F_v(2,2,7;8) \geq 20$ (see Table 1).

For positive integers $a_1,\ldots,a_s$ and m and p defined by (1.3), Nenov proved in [15] that if $G \in \mathcal{H}(a_1,\ldots,a_s;m-1;n)$ and $n < m+3p$, then $\alpha(G) < n-m-p+1$. Suppose that $G \in \mathcal{H}(2,2,2,7;9;20)$. It follows that $\alpha(G) \leq 3$ and it is clear that $\alpha(G) \geq 2$. Therefore, it is enough to prove that there are no graphs with independence number 2 or 3 in $\mathcal{H}_{max}(2,2,2,7;9;20)$.

First we prove that there are no graphs in $\mathcal{H}_{max}(2,2,2,7;9;20)$ with independence number 3. It is clear that K_7 is the only graph in $\mathcal{H}_{max}(4;9;7)$. By applying Algorithm 3.9($r=2;t=3$) with $\mathcal{A}_1 = \mathcal{H}^3_{max}(4;9;7) = \{K_7\}$ and $\mathcal{A}_2 = \mathcal{H}^3_{max}(4;8;8)$ were obtained all graphs in $\mathcal{H}^3_{max}(5;9;9)$ (see Remark 3.11). In the same way, we successively obtained all graphs in $\mathcal{H}^3_{max}(6;9;11)$, $\mathcal{H}^3_{max}(7;9;13)$, $\mathcal{H}^3_{max}(2,7;9;15)$ and $\mathcal{H}^3_{max}(2,2,7;9;17)$ (see Remark 3.11). In the end, by applying Algorithm 3.9($r=3;t=3$) with $\mathcal{A}_1 = \mathcal{H}^3_{max}(2,2,7;9;17)$ and $\mathcal{A}_2 = \mathcal{H}^3_{max}(2,2,7;8;19) = \emptyset$, no graphs with independence number 3 in $\mathcal{H}_{max}(2,2,2,7;9;20)$ were obtained.

Next we prove that there are no graphs in $\mathcal{H}_{max}(2,2,2,7;9;20)$ with independence number 2. Clearly, K_8 is the only graph in $\mathcal{H}_{max}(4;9;8)$. By applying Algorithm 3.9($r=2;t=2$) with $\mathcal{A}_1 = \mathcal{H}^2_{max}(4;9;8) = \{K_8\}$ and $\mathcal{A}_2 = \mathcal{H}^2_{max}(4;8;9)$ were obtained all graphs in $\mathcal{H}^2_{max}(5;9;10)$ (see Remark 3.11). In the same way, we successively obtained all graphs in $\mathcal{H}^2_{max}(6;9;12)$, $\mathcal{H}^2_{max}(7;9;14)$, $\mathcal{H}^2_{max}(2,7;9;16)$ and $\mathcal{H}^2_{max}(2,2,7;9;18)$ (see Remark 3.11). In the end, by applying Algorithm 3.9($r=2;t=2$) with $\mathcal{A}_1 = \mathcal{H}^2_{max}(2,2,7;9;18)$ and $\mathcal{A}_2 = \mathcal{H}^2_{max}(2,2,7;8;19) = \emptyset$, no graphs with independence number 2 in $\mathcal{H}_{max}(2,2,2,7;9;20)$ were obtained.

We proved that $\mathcal{H}_{max}(2,2,2,7;9;20) = \emptyset$ and $F_v(2,2,2,7;9) > 20$.

Similarly, the graphs obtained in the proof of the inequality $F_v(2,2,2,7;9) > 20$ are used to prove $F_v(2,2,2,2,7;10) > 21$ and the graphs obtained in the proof of the inequality $F_v(2,2,2,2,7;10) > 21$ are used to prove $F_v(2,2,2,2,2,7;11) > 22$. The number of graphs obtained in each step of the proofs is shown in Table 2, Table 3 and Table 4. Notice that the number of graphs without cone vertices is relatively small, which reduces the computation time significantly.

Thus, $r_0(7) = 2$ and

$$F_v(2_{m-7};7;m-1) = F_v(2,2,7;8) + m - 9 = m + 11,$$

which completes the proof of Theorem 2.3. The Main Theorem now follows easily. Indeed, let $a_1,\ldots,a_s$ be positive integers such that $\max\{a_1,\ldots,a_s\} = 7$ and $m = \sum_{i=1}^{s}(a_i-1)+1$. Then

$$F_v(a_1,\ldots,a_s;m-1) \geq F_v(2_{m-7};7;m-1) = m+11. \qquad \square$$

5. CONCLUDING REMARKS

The proposed method for obtaining of lower bounds for $F_v(a_1,\ldots,a_s;q)$ produces good and accurate results when $q=m-1$. However, when $q<m-1$, the bounds are not exact. We will consider the most interesting case $q=p+1$, where $p=\max\{a_1,\ldots,a_s\}$. In [1] we prove the inequality

$$F_v(a_1,\ldots,a_s;p+1)\geq F_v(2,2,p;p+1)+\sum_{i=3}^{m-p}\alpha(i,p), \tag{5.1}$$

where $\alpha(i,p)=\max\{\alpha(G):G\in\mathcal{H}_{extr}(2_i,p;p+1)\}$. Since $\alpha(i,p)\geq 2$, from (5.1) it follows that

$$F_v(a_1,...,a_s;p+1)\geq F_v(2,2,p;p+1)+2(m-p-2).$$

In the special case $p=7$, since $F_v(2,2,7;8)=20$, we obtain

$$F_v(a_1,\ldots,a_s;8)\geq 2m+2. \tag{5.2}$$

In particular, when $m=13$ we have $F_v(a_1,\ldots,a_s;8)\geq 28$. Since the Ramsey number $R(3,8)=28$, it follows that $\alpha(i,7)\geq 3$, when $i\geq 6$. Now from (5.1) we obtain easily the following result:

Theorem 5.1. *If $m\geq 13$, and $\max\{a_1,...,a_s\}=7$, then*

$$F_v(a_1,\ldots,a_s;8)\geq 3m-10.$$

It is clear that when $3m-10\geq R(4,8)$, these bounds for $F_v(a_1,\ldots,a_s;8)$ can be improved significantly.

In [21] is proved the inequality $F_v(p,p;p+1)\geq 4p-1$. From this result it follows that $F_v(7,7;8)\geq 27$. From (5.2) we deduce that $F_v(7,7;8)\geq 28$, and from Theorem 5.1 we obtain $F_v(7,7;8)\geq 29$.

The numbers $F_v(p,p;p+1)$ are of significant interest, but so far we know very little about them. Only two of these numbers are known, $F_v(2,2;3)=5$ (obvious), and $F_v(3,3;4)=14$ ([11, 17]). It is also known that $17\leq F_v(4,4;5)\leq 23$, [20], $F_v(5,5;6)\geq 23$, [1], $28\leq F_v(6,6;7)\leq 70$, [4], and $F_v(7,7;8)\geq 29$ from this paper. Using Algorithm 3.5, we managed to improve the known lower bound $F_v(2,2,2,4;5)\geq 17$ and thus improved the lower bound on $F_v(4,4;5)$ as well:

Theorem 5.2. $F_v(4,4;5)\geq F_v(2,3,4;5)\geq F_v(2,2,2,4;5)\geq 19$.

Proof. The inequalities $F_v(4,4;5)\geq F_v(2,3,4;5)\geq F_v(2,2,2,4;5)$ are easy to prove (see eq. (4.1) in [1]). It remains to prove that $F_v(2,2,2,4;5)\geq 19$. Suppose that $\mathcal{H}_{max}(2,2,2,4;5;18)\neq\emptyset$ and let $G\in\mathcal{H}_{max}(2,2,2,4;5;18)$. Since the Ramsey number $R(3,5)=14$, $\alpha(G)\geq 3$. In [20] it is proved that $F_v(2,2,4;5)=13$ and

$\mathcal{H}(2,2,4;5;13) = \{Q\}$, where Q is the unique 13-vertex K_5-free graph with independence number 2. From Proposition 3.3 and the equality $F_v(2,2,4;5) = 13$ it follows that $\alpha(G) \leq 5$. By applying Algorithm 3.5 to the graph Q it follows that there are no graphs in $\mathcal{H}_{max}(2,2,2,4;5;18)$ with independence number 5. It remains to prove that there are no graphs in $\mathcal{H}_{max}(2,2,2,4;5;18)$ with independence number 3 or 4. Using *nauty* it is easy to obtain the sets $\mathcal{H}^4_{max}(3;5;8)$ and $\mathcal{H}^3_{max}(3;5;9)$. By applying Algorithm 3.5 ($r = 2, t = 4$) starting from the set $\mathcal{H}^4_{max}(3;5;8)$ were successively obtained all graphs in the sets $\mathcal{H}^4_{max}(4;5;10)$, $\mathcal{H}^4_{max}(2,4;5;12)$, $\mathcal{H}^4_{max}(2,2,4;5;14)$ (see Remark 3.11), and by applying Algorithm 3.5 ($r = 4, t = 4$) were found no graphs in $\mathcal{H}_{max}(2,2,2,4;5;18)$ with independence number 4. Next, we applied Algorithm 3.5 ($r = 2, t = 3$) starting from the set $\mathcal{H}^3_{max}(3;5;9)$ to successively obtain all graphs in the sets $\mathcal{H}^3_{max}(4;5;11)$, $\mathcal{H}^3_{max}(2,4;5;13)$, $\mathcal{H}^3_{max}(2,2,4;5;15)$ (see Remark 3.11), and by applying Algorithm 3.5 ($r = 3, t = 3$) were found no graphs in $\mathcal{H}_{max}(2,2,2,4;5;18)$ with independence number 3. The number of graphs obtained in each of the steps is shown in Table 5. We obtained $\mathcal{H}_{max}(2,2,2,4;5;18) = \emptyset$ and therefore $F_v(2,2,2,4;5) \geq 19$. □

The upper bound $F_v(4,4;5) \leq 23$ is proved in [20] with the help of a 23-vertex transitive graph. We were not able to obtain any other graphs in $\mathcal{H}(4,4;5;23)$, which leads us to believe that this bound may be exact. We did find a large number of 23-vertex graphs in $\mathcal{H}(2,2,2,4;5)$, but so far we have not obtained smaller graphs in this set.

Concluding this section, let us pose the following question:

Question 5.1. Is it true that the sequence $F_v(p,p;p+1), p \geq 2$, is increasing?

ACKNOWLEDGEMENT. The authors were partially supported by the Sofia University Research Fund through Contract 80-10-74/20.04.2017.

A. RESULTS OF COMPUTATIONS

set	ind. number	maximal graphs	$(+K_7)$-graphs
$\mathcal{H}(2,7;8;15)$	≤ 4	1	1
$\mathcal{H}(2,2,7;8;19)$	$=4$	0	
$\mathcal{H}(3;8;6)$	≤ 3	1	1
$\mathcal{H}(4;8;8)$	≤ 3	1	4
$\mathcal{H}(5;8;10)$	≤ 3	3	45
$\mathcal{H}(6;8;12)$	≤ 3	12	3 104
$\mathcal{H}(7;8;14)$	≤ 3	169	4 776 518
$\mathcal{H}(2,7;8;16)$	≤ 3	34	22 896
$\mathcal{H}(2,2,7;8;19)$	$=3$	0	
$\mathcal{H}(3;8;7)$	≤ 2	1	1
$\mathcal{H}(4;8;9)$	≤ 2	1	8
$\mathcal{H}(5;8;11)$	≤ 2	3	84
$\mathcal{H}(6;8;13)$	≤ 2	10	5 394
$\mathcal{H}(7;8;15)$	≤ 2	102	4 984 994
$\mathcal{H}(2,7;8;17)$	≤ 2	2760	380 361 736
$\mathcal{H}(2,2,7;8;19)$	$=2$	0	
$\mathcal{H}(2,2,7;8;19)$		0	

Table 1: Steps in finding all maximal graphs in $\mathcal{H}(2,2,7;8;19)$

set	ind. number	max. graphs	max. graphs no cone v.	$(+K_8)$-graphs	$(+K_8)$-graphs no cone v.
$\mathcal{H}(2,2,7;9;16)$	≤ 4	1	0	1	0
$\mathcal{H}(2,2,2,7;9;20)$	$=4$	0	0		
$\mathcal{H}(4;9;7)$	≤ 3	1	0	1	0
$\mathcal{H}(5;9;9)$	≤ 3	1	0	4	0
$\mathcal{H}(6;9;11)$	≤ 3	3	0	45	0
$\mathcal{H}(7;9;13)$	≤ 3	12	0	3 113	9
$\mathcal{H}(2,7;9;15)$	≤ 3	169	0	4 783 615	7 097
$\mathcal{H}(2,2,7;9;17)$	≤ 3	36	2	22 918	22
$\mathcal{H}(2,2,2,7;9;20)$	$=3$	0	0		
$\mathcal{H}(4;9;8)$	≤ 2	1	0	1	0
$\mathcal{H}(5;9;10)$	≤ 2	1	0	8	0
$\mathcal{H}(6;9;12)$	≤ 2	3	0	85	1
$\mathcal{H}(7;9;14)$	≤ 2	10	0	5 474	80
$\mathcal{H}(2,7;9;16)$	≤ 2	103	1	5 346 982	361 988
$\mathcal{H}(2,2,7;9;18)$	≤ 2	2845	85	387 948 338	7 586 602
$\mathcal{H}(2,2,2,7;9;20)$	$=2$	0	0		
$\mathcal{H}(2,2,2,7;9;20)$		0	0		

Table 2: Steps in finding all maximal graphs in $\mathcal{H}(2,2,2,7;9;20)$

set	ind. number	max. graphs	max. graphs no cone v.	$(+K_9)$-graphs	$(+K_9)$-graphs no cone v.
$\mathcal{H}(2,2,2,7;10;17)$	≤ 4	1	0	1	0
$\mathcal{H}(2,2,2,2,7;10;21)$	$= 4$	0	0		
$\mathcal{H}(5;10;8)$	≤ 3	1	0	1	0
$\mathcal{H}(6;10;10)$	≤ 3	1	0	4	0
$\mathcal{H}(7;10;12)$	≤ 3	3	0	45	0
$\mathcal{H}(2,7;10;14)$	≤ 3	12	0	3 115	2
$\mathcal{H}(2,2,7;10;16)$	≤ 3	169	0	4 784 483	868
$\mathcal{H}(2,2,2,7;10;18)$	≤ 3	36	0	22 919	1
$\mathcal{H}(2,2,2,2,7;10;21)$	$= 3$	0	0		
$\mathcal{H}(5;10;9)$	≤ 2	1	0	1	0
$\mathcal{H}(6;10;11)$	≤ 2	1	0	8	0
$\mathcal{H}(7;10;13)$	≤ 2	3	0	85	0
$\mathcal{H}(2,7;10;15)$	≤ 2	10	0	5 495	21
$\mathcal{H}(2,2,7;10;17)$	≤ 2	103	0	5 371 651	24 669
$\mathcal{H}(2,2,2,7;10;19)$	≤ 2	2848	3	387 968 658	20 320
$\mathcal{H}(2,2,2,2,7;10;21)$	$= 2$	0	0		
$\mathcal{H}(2,2,2,2,7;10;21)$		0	0		

Table 3: Steps in finding all maximal graphs in $\mathcal{H}(2,2,2,2,7;10;21)$

set	ind. number	max. graphs	max. graphs no cone v.	$(+K_{10})$-graphs	$(+K_{10})$-graphs no cone v.
$\mathcal{H}(2,2,2,2,7;11;18)$	≤ 4	1	0	1	0
$\mathcal{H}(2,2,2,2,2,7;11;22)$	$= 4$	0	0		
$\mathcal{H}(6;11;9)$	≤ 3	1	0	1	0
$\mathcal{H}(7;11;11)$	≤ 3	1	0	4	0
$\mathcal{H}(2,7;11;13)$	≤ 3	3	0	45	0
$\mathcal{H}(2,2,7;11;15)$	≤ 3	12	0	3 116	1
$\mathcal{H}(2,2,2,7;11;17)$	≤ 3	169	0	4 784 638	155
$\mathcal{H}(2,2,2,2,7;11;19)$	≤ 3	36	0	22 919	0
$\mathcal{H}(2,2,2,2,2,7;11;22)$	$= 3$	0	0		
$\mathcal{H}(6;11;10)$	≤ 2	1	0	1	0
$\mathcal{H}(7;11;12)$	≤ 2	1	0	8	0
$\mathcal{H}(2,7;11;14)$	≤ 2	3	0	85	0
$\mathcal{H}(2,2,7;11;16)$	≤ 2	10	0	5 502	7
$\mathcal{H}(2,2,2,7;11;18)$	≤ 2	103	0	5 374 143	2 492
$\mathcal{H}(2,2,2,2,7;11;20)$	≤ 2	2848	0	387 968 676	18
$\mathcal{H}(2,2,2,2,2,7;11;22)$	$= 2$	0	0		
$\mathcal{H}(2,2,2,2,2,7;11;22)$		0	0		

Table 4: Steps in finding all maximal graphs in $\mathcal{H}(2,2,2,2,2,7;11;22)$

set	ind. number	maximal graphs	$(+K_4)$-graphs
$\mathcal{H}(2,2,4;5;13)$	≤ 5	1	1
$\mathcal{H}(2,2,2,4;5;18)$	$=5$	0	
$\mathcal{H}(3;5;8)$	≤ 4	7	274
$\mathcal{H}(4;5;10)$	≤ 4	44	65 422
$\mathcal{H}(2,4;5;12)$	≤ 4	1 059	18 143 174
$\mathcal{H}(2,2,4;5;14)$	≤ 4	13	71
$\mathcal{H}(2,2,2,4;5;18)$	$=4$	0	
$\mathcal{H}(3;5;9)$	≤ 3	11	2 252
$\mathcal{H}(4;5;11)$	≤ 3	135	1 678 802
$\mathcal{H}(2,4;5;13)$	≤ 3	11 439	2 672 047 607
$\mathcal{H}(2,2,4;5;15)$	≤ 3	1 103	78 117
$\mathcal{H}(2,2,2,4;5;18)$	$=3$	0	
$\mathcal{H}(2,2,2,4;5;18)$		0	

Table 5: Steps in finding all maximal graphs in $\mathcal{H}(2,2,2,4;5;18)$

6. REFERENCES

[1] Bikov, A., Nenov, N.: The vertex Folkman numbers $F_v(a_1,...,a_s;m-1)=m+9$, if $\max\{a_1,...,a_s\}=5$. To appear in the *Journal of Combinatorial Mathematics and Combinatorial Computing*, preprint: arxiv:1503.08444, August 2015.

[2] Bikov, A., Nenov, N.: Modified vertex Folkman numbers. *Mathematics and Education. Proceedings of the 45th Spring Conference of the Union of Bulgarian Mathematicians*, **45**, 2016, 113–123. preprint: arxiv:1511.02125, November 2015.

[3] Bikov, A., Nenov, N.: The edge Folkman number $F_e(3,3;4)$ is greater than 19. *GEOMBINATORICS*, **27**, no. 1, 2017, 5–14. preprint: arxiv:1609.03468, September 2016.

[4] Bikov, A., Nenov, N.: On the vertex Folkman numbers $F_v(a_1,...,a_s;m-1)$ when $\max\{a_1,...,a_s\}=6$ or 7. To appear in the *Journal of Combinatorial Mathematics and Combinatorial Computing*, preprint: arxiv:1512.02051, April 2017.

[5] Dudek, A., Rödl, V.: New upper bound on vertex Folkman numbers. *Lecture Notes in Computer Science*, **4557**, 2008, 473–478.

[6] Folkman, J.: Graphs with monochromatic complete subgraphs in every edge coloring. *SIAM J. Appl. Math*, **18**, 1970, 19–24.

[7] Kolev, N., Nenov. N.: New upper bound for a class of vertex Folkman numbers. *The Electron. J. Comb.*, **13**, 2006.

[8] Luczak, T., Ruciński, A., Urbański, S.: On minimal vertex Folkman graphs. *Discrete Math.*, **236**, 2001, 245–262.

[9] Luczak, T., Urbański, S.: A note on restricted vertex Ramsey numbers. *Periodica Math. Hungarica*, **33**, 1996, 101–103.

[10] McKay, B. D., Piperino, A.: Practical graph isomorphism, II. *J. Symb. Comp.*, **60**, 2013, 94–112. Preprint available at arxiv.org.

[11] Nenov, N.: An example of a 15-vertex (3, 3)-Ramsey graph with clique number 4. *C. R. Acad. Bulg. Sci.*, **34**(11), 1981, 1487–1489 (in Russian).

[12] Nenov, N.: Application of the corona-product of two graphs in Ramsey theory. *Ann. Univ. Sofia Fac. Math. Inform.*, **79**, 1985, 349–355 (in Russian).

[13] Nenov, N.: On a class of vertex Folkman graphs. *Ann. Univ. Sofia Fac. Math. Inform.*, **94**, 2000, 15–25.

[14] Nenov, N.: A generalization of a result of Dirac. *Ann. Univ. Sofia Fac. Math. Inform.*, **95**, 2001, 59–69.

[15] Nenov, N.: Lower bound for a number of vertices of some vertex Folkman graphs. *C. R. Acad. Bulg. Sci.*, **55**, n0. 4, 2002, 33–36.

[16] Nenov, N.: On a class of vertex Folkman numbers. *Serdica Math. J.*, **28**, 2002, 219–232.

[17] Piwakowski, K., Radziszowski, S., Urbanski, S.: Computation of the Folkman number $F_e(3,3;5)$. *J. Graph Theory*, **32**, 1999, 41–49.

[18] Shao, Z., Xu, X., Pan, L.: New upper bounds for vertex Folkman numbers $F_v(3,k;k+1)$. *Utilitas Math.*, **80**, 2009, 91–96.

[19] West, D.: *Introduction to Graph Theory.* Prentice Hall, Inc., Upper Saddle River, Second edition, 2001.

[20] Xu, X., Luo, H., Shao, Z.: Upper and lower bounds for $F_v(4,4;5)$. *Electron J. Combinatorics*, **17**, 2010.

[21] Xu, X., Shao, Z.: On the lower bound for $F_v(k,k;k+1)$ and $F_e(3,4;5)$. *Utilitas Math.*, **81**, 2010, 187–192.

Received on October 13, 2017

Aleksandar Bikov, Nedyalko Nenov

Faculty of Mathematics and Informatics
"St. Kl. Ohridski" University of Sofia
5, J. Bourchier blvd., BG-1164 Sofia
BULGARIA

E-mails: asbikov@fmi.uni-sofia.bg
nenov@fmi.uni-sofia.bg

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 104

ANNUAL OF SOFIA UNIVERSITY „ST. KLIMENT OHRIDSKI“
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 104

ESTIMATES FOR THE BEST CONSTANT IN A MARKOV L_2–INEQUALITY WITH THE ASSISTANCE OF COMPUTER ALGEBRA

GENO NIKOLOV, RUMEN ULUCHEV

We prove two-sided estimates for the best (i.e., the smallest possible) constant $c_n(\alpha)$ in the Markov inequality

$$\|p_n'\|_{w_\alpha} \le c_n(\alpha)\|p_n\|_{w_\alpha}\,, \qquad p_n \in \mathcal{P}_n\,.$$

Here, $\mathcal{P}_n$ stands for the set of algebraic polynomials of degree $\le n$, $w_\alpha(x) := x^\alpha\, e^{-x}$, $\alpha > -1$, is the Laguerre weight function, and $\|\cdot\|_{w_\alpha}$ is the associated L_2-norm,

$$\|f\|_{w_\alpha} = \left(\int_0^\infty |f(x)|^2 w_\alpha(x)\,dx\right)^{1/2}.$$

Our approach is based on the fact that $c_n^{-2}(\alpha)$ equals the smallest zero of a polynomial Q_n, orthogonal with respect to a measure supported on the positive axis and defined by an explicit three-term recurrence relation. We employ computer algebra to evaluate the seven lowest degree coefficients of Q_n and to obtain thereby bounds for $c_n(\alpha)$. This work is a continuation of a recent paper [5], where estimates for $c_n(\alpha)$ were proven on the basis of the four lowest degree coefficients of Q_n.

Keywords: Markov type inequalities, Laguerre polynomials, three-term recurrence relation, Newton identities, computer algebra.

2000 Math. Subject Classification: 41A17.

1. INTRODUCTION AND STATEMENT OF THE RESULTS

Throughout this paper $\mathcal{P}_n$ will stand for the set of algebraic polynomials of degree at most n, assumed, without loss of generality, with real coefficients. Let

$w_\alpha(x) := x^\alpha\, e^{-x}$, where $\alpha > -1$, be the Laguerre weight function, and $\|\cdot\|_{w_\alpha}$ be the associated L_2-norm,

$$\|f\|_{w_\alpha} = \left(\int_0^\infty |f(x)|^2 w_\alpha(x)\, dx\right)^{1/2}.$$

We study the best constant $c_n(\alpha)$ in the Markov inequality in this norm

$$\|p_n'\|_{w_\alpha} \le c_n(\alpha)\|p_n\|_{w_\alpha}\,, \qquad p_n \in \mathcal{P}_n\,, \tag{1.1}$$

namely the constant

$$c_n(\alpha) := \sup_{p_n\in\mathcal{P}_n} \frac{\|p_n'\|_{w_\alpha}}{\|p_n\|_{w_\alpha}}\,.$$

Before formulating our results, let us give a brief account on the results known so far.

It is only the case $\alpha = 0$ where the best Markov constant is known, namely, Turán [9] proved that

$$c_n(0) = \left(2\sin\frac{\pi}{4n+2}\right)^{-1}.$$

Dörfler [2] showed that $c_n(\alpha) = \mathcal{O}(n)$ for every fixed $\alpha > -1$ by proving the estimates

$$c_n^2(\alpha) \ge \frac{n^2}{(\alpha+1)(\alpha+3)} + \frac{(2\alpha^2+5\alpha+6)\,n}{3(\alpha+1)(\alpha+2)(\alpha+3)} + \frac{\alpha+6}{3(\alpha+2)(\alpha+3)}\,, \tag{1.2}$$

$$c_n^2(\alpha) \le \frac{n(n+1)}{2(\alpha+1)}\,, \tag{1.3}$$

see [3] for a more accessible source. In the same paper, [3], Dörfler proved for the asymptotic constant

$$c(\alpha) := \lim_{n\to\infty} \frac{c_n(\alpha)}{n}\,, \tag{1.4}$$

that

$$c(\alpha) = \frac{1}{j_{(\alpha-1)/2,1}}\,, \tag{1.5}$$

where $j_{\nu,1}$ is the first positive zero of the Bessel function $J_\nu(z)$.

Nikolov and Shadrin obtained in [5] the following result:

Theorem A ([5, Theorem 1]). *For all $\alpha > -1$ and $n \in \mathbb{N}$, $n \ge 3$, the best constant $c_n(\alpha)$ in the Markov inequality (1.1) admits the estimates*

$$\frac{2\big(n+\frac{2\alpha}{3}\big)\big(n-\frac{\alpha+1}{6}\big)}{(\alpha+1)(\alpha+5)} < c_n^2(\alpha) < \frac{\big(n+1\big)\big(n+\frac{2(\alpha+1)}{5}\big)}{(\alpha+1)\big[(\alpha+3)(\alpha+5)\big]^{1/3}}\,, \tag{1.6}$$

where for the left-hand inequality it is additionally assumed that $n > (\alpha+1)/6$.

Theorem A implies some inequalities for the asymptotic Markov constant $c(\alpha)$ and, through (1.5), inequalities for $j_{\nu,1}$, the first positive zero of the Bessel function J_ν (see [5, Corollaries 1, 3]). It was also shown in [5, Theorem 2] that $c(\alpha) = \mathcal{O}(\alpha^{-1})$, which indicates that the upper estimate for $c_n(\alpha)$ in Theorem A, though rather good for moderate α, is not optimal.

In a recent paper [7] Nikolov and Shadrin proved an upper bound for $c_n(\alpha)$ which is of the correct order with respect to both n and α as they tend to infinity.

Theorem B ([7, Theorem 1.1]). *For all $n \in \mathbb{N}$, $n \geq 3$, the best constant $c_n(\alpha)$ in the Markov inequality* (1.1) *satisfies the inequality*

$$c_n^2(\alpha) \leq \frac{4n\big(n+2+\frac{3(\alpha+1)}{4}\big)}{\alpha^2+10\alpha+8}, \qquad \alpha \geq 2. \tag{1.7}$$

As a consequence of Theorem B and Dörfler's lower bound (1.2) for $c_n(\alpha)$ Nikolov and Shadrin showed that

$$c_n^2(\alpha) \asymp \frac{n(n+\alpha+3)}{(\alpha+1)(\alpha+8)}, \qquad n \geq 3,\ \alpha \geq 2.$$

Corollary C ([7, Corollary 1.1]). *For all $\alpha \geq 2$ and $n \geq 3$ the best constant $c_n(\alpha)$ in the Markov inequality* (1.1) *satisfies*

$$\frac{2n(n+\alpha+3)}{3(\alpha+1)(\alpha+8)} \leq c_n^2(\alpha) \leq \frac{4n(n+\alpha+3)}{(\alpha+1)(\alpha+8)}. \tag{1.8}$$

In addition, Nikolov and Shadrin found the limit value of $(\alpha+1)c_n^2(\alpha)$ as $\alpha \to -1$, and proved asymptotic inequalities for $\alpha\, c_n^2(\alpha)$ as $\alpha \to \infty$.

Corollary D ([7, Corollary 1.2]). *The best constant $c_n(\alpha)$ in the Markov inequality* (1.1) *satisfies:*

(i) $\displaystyle\lim_{\alpha\to-1}(\alpha+1)c_n^2(\alpha) = \frac{n(n+1)}{2}$;

(ii) $\displaystyle\frac{2n}{3} \leq \lim_{\alpha\to\infty} \alpha\, c_n^2(\alpha) \leq 3n$.

A combination of Theorems A and B implies bounds for $c(\alpha)$ defined in (1.4):

Corollary E ([7, Corollary 1.3]). *The asymptotic Markov constant $c(\alpha)$ satisfies*

$$\frac{2}{(\alpha+1)(\alpha+5)} < c^2(\alpha) < \begin{cases} \dfrac{1}{(\alpha+1)\sqrt[3]{(\alpha+3)(\alpha+5)}}, & -1 < \alpha \leq \alpha^*, \\ \dfrac{4}{\alpha^2+10\alpha+8}, & \alpha > \alpha^*, \end{cases}$$

where $\alpha^* \approx 43.4$.

The ratio of the upper and the lower bound for $c(\alpha)$ in Corollary E is less than $\sqrt{2}$ for all $\alpha > -1$.

In this paper we investigate the best Markov constant $c_n(\alpha)$ following the approach from [5]. It is known (see Proposition 1 below) that $c_n^{-2}(\alpha)$ is equal to the smallest zero of a polynomial Q_n, which is orthogonal with respect to a measure supported on $\mathbb{R}_+$. Since $\{Q_n\}_{n\in\mathbb{N}}$ are defined by an explicit three-term recurrence relation, one can evaluate (at least theoretically) as many coefficients of Q_n as necessary. With the assistance of Wolfram's *Mathematica* we find the seven lowest degree coefficients of the polynomial Q_n, and thereby the six highest degree coefficients of R_n, the monic polynomial reciprocal to Q_n. Then we apply a simple technique for estimating the largest zero x_n of R_n on the basis of its k highest degree coefficients, $3 \le k \le 6$, thus obtaining lower and upper bounds for $c_n^2(\alpha)$. Our main result in this paper is:

Theorem 1. *For* $3 \le k \le 6$ *and for all* $n \ge k$, *the best constant* $c_n(\alpha)$ *in the Markov inequality* (1.1) *admits the estimates*

$$\underline{c}_{n,k}(\alpha) \le c_n(\alpha) \le \overline{c}_{n,k}(\alpha)\,, \qquad \alpha > -1\,, \tag{1.9}$$

where

$$\underline{c}^{\,2}_{n,3}(\alpha) = \frac{2\,n\big(n+\frac{3(\alpha+1)}{8}\big)}{(\alpha+1)(\alpha+5)}\,, \tag{1.10}$$

$$\overline{c}^{\,2}_{n,3}(\alpha) = \frac{(n+1)\big(n+\frac{2(\alpha+1)}{5}\big)}{(\alpha+1)\big[(\alpha+3)(\alpha+5)\big]^{1/3}}\,, \tag{1.11}$$

$$\underline{c}^{\,2}_{n,4}(\alpha) = \frac{(5\alpha+17)\,n\big(n+\frac{8(\alpha+1)}{25}\big)}{2(\alpha+1)(\alpha+3)(\alpha+7)}\,, \tag{1.12}$$

$$\overline{c}^{\,2}_{n,4}(\alpha) = \frac{(5\alpha+17)^{1/4}(n+1)\big(n+\frac{3(\alpha+1)}{7}\big)}{(\alpha+1)(\alpha+3)^{1/2}\big[2(\alpha+5)(\alpha+7)\big]^{1/4}}\,, \tag{1.13}$$

$$\underline{c}^{\,2}_{n,5}(\alpha) = \frac{2(7\alpha+31)n\big(n+\frac{25(\alpha+1)}{84}\big)}{(\alpha+1)(\alpha+9)(5\alpha+17)}\,, \tag{1.14}$$

$$\overline{c}^{\,2}_{n,5}(\alpha) = \frac{(7\alpha+31)^{1/5}(n+1)\big(n+\frac{4(\alpha+1)}{9}\big)}{(\alpha+1)(\alpha+3)^{2/5}\big[(\alpha+5)(\alpha+7)(\alpha+9)\big]^{1/5}}\,, \tag{1.15}$$

$$\underline{c}^{\,2}_{n,6}(\alpha) = \frac{\big(21\alpha^3+299\alpha^2+1391\alpha+2073\big)n\big(n+\frac{2(\alpha+1)}{7}\big)}{(\alpha+1)(\alpha+3)(\alpha+5)(\alpha+11)(7\alpha+31)}\,, \tag{1.16}$$

$$\overline{c}^{\,2}_{n,6}(\alpha) = \frac{\big(21\alpha^3+299\alpha^2+1391\alpha+2073\big)^{1/6}(n+1)\big(n+\frac{5(\alpha+1)}{11}\big)}{(\alpha+1)(\alpha+3)^{1/2}(\alpha+5)^{1/3}\big[(\alpha+7)(\alpha+9)(\alpha+11)\big]^{1/6}}\,. \tag{1.17}$$

Remark 1. For $3 \le k \le 6$, the pair $\big(\underline{c}_{n,k}(\alpha), \overline{c}_{n,k}(\alpha)\big)$ of bounds for $c_n(\alpha)$ is deduced with the use of the k highest degree coefficients of the polynomial R_n (and (1.11) is also proved in [5]). Generally, the bounds for $c_n(\alpha)$ obtained with larger k are better, though some exceptions are observed for small n and α.

Clearly, inequalities (1.9) imply bounds for the asymptotic Markov constant $c(\alpha)$. Here, it is not difficult to prove that the larger k, the better the implied lower and upper bounds for $c(\alpha)$, hence the best bounds for $c(\alpha)$ are obtained from (1.9) with $k = 6$.

Thus, Theorem 1 yields an improvement of the estimates for the asymptotic Markov constant $c(\alpha)$ in Corollary E.

Corollary 1. *The asymptotic Markov constant* $c(\alpha) = \lim\limits_{n\to\infty} n^{-1}c_n(\alpha)$ *satisfies the inequalities*

$$\underline{c}(\alpha) < c(\alpha) < \overline{c}(\alpha)\,,$$

where

$$\underline{c}^{\,2}(\alpha) := \frac{21\alpha^3 + 299\alpha^2 + 1391\alpha + 2073}{(\alpha+1)(\alpha+3)(\alpha+5)(\alpha+11)(7\alpha+31)}$$

and

$$\overline{c}^{\,2}(\alpha) := \begin{cases} \dfrac{\big(21\alpha^3 + 299\alpha^2 + 1391\alpha + 2073\big)^{1/6}}{(\alpha+1)(\alpha+3)^{1/2}(\alpha+5)^{1/3}\big[(\alpha+7)(\alpha+9)(\alpha+11)\big]^{1/6}}\,, & -1<\alpha\le\alpha^\star\,, \\ \dfrac{4}{\alpha^2+10\alpha+8}\,, & \alpha>\alpha^\star, \end{cases}$$

with $\alpha^\star \approx 172$.

It is worth noticing that the ratio of the upper and the lower bound for $c(\alpha)$ in Corollary 1 does no exceed $\frac{2\sqrt{3}}{3} \approx 1.1547$ for all $\alpha > -1$.

Theorem 1, in particular inequality (1.16), implies an improvement of the lower bound in Corollary D(ii).

Corollary 2. *The best constant* $c_n(\alpha)$ *in the Markov inequality* (1.1) *satisfies:*

$$\frac{6n}{7} \le \lim_{\alpha\to\infty} \alpha\, c_n^2(\alpha) \le 3n\,.$$

The rest of the paper is organized as follows. Section 2 contains some preliminaries. In Section 2.1 we characterize the squared best Markov constant as the largest zero of an n-th degree monic polynomial R_n with positive roots, and propose a recursive procedure for the evaluation of its coefficients (Proposition 2). Two-sided estimates for the largest zero of polynomials with only positive roots in terms of few of their coefficients are proposed in Sect. 2.2 (Proposition 2.3). The assisted by Wolfram's *Mathematica* proof of our results is given in Section 3.

In Section 4 we give some final remarks and conclusions, and formulate two conjectures concerning the asymptotic behavior of the best Markov constant and the coefficients of the characteristic polynomial R_n.

2. PRELIMINARIES

2.1. AN ORTHOGONAL POLYNOMIAL RELATED TO $c_n(\alpha)$

It is well-known that the squared best constant in a Markov-type inequality in L_2-norm is equal to the largest eigenvalue of a related positive definite $n \times n$ matrix $\mathbf{A}_n$, thus the problem of finding the best Markov constant is equivalent to evaluating the largest eigenvalue of $\mathbf{A}_n$. Perhaps, a less known fact is that for a wide class of L_2-norms, the inverse matrix $\mathbf{A}_n^{-1}$ is tri-diagonal, see [1, Sect. 2]. In the particular case of the L_2-norm induced by the Laguerre weight function w_α this connection is given by the following proposition:

Proposition 1 ([**3, p. 85**]). *The quantity $c_n^{-2}(\alpha)$ is equal to the smallest zero of the polynomial $Q_n(x) = Q_n(x, \alpha)$, which is defined recursively by*

$$
\begin{aligned}
&Q_{n+1}(x) = (x - d_n)Q_n(x) - \lambda_n^2 Q_{n-1}(x), \quad n \geq 0\,;\\
&Q_{-1}(x) := 0, \quad Q_0(x) := 1\,;\\
&d_0 := 1 + \alpha, \quad d_n := 2 + \frac{\alpha}{n+1}\,, \quad n \geq 1\,;\\
&\lambda_0 > 0 \ \text{ arbitrary}, \ \lambda_n^2 := 1 + \frac{\alpha}{n}\,, \quad n \geq 1\,.
\end{aligned}
$$

By Favard's theorem, for any $\alpha > -1$, $\{Q_n(x, \alpha)\}_{n=0}^{\infty}$ form a system of monic orthogonal polynomials. Since Q_n is the characteristic polynomial of the inverse of a positive definite matrix (which is also positive definite), it follows that all the zeros of Q_n are positive (and distinct). Consequently, $\{Q_n\}_{n=0}^{\infty}$ are orthogonal with respect to a measure supported on $\mathbb{R}_+$.

By Proposition 1, we have

$$
Q_{n+1}(x) = \Big(x - 2 - \frac{\alpha}{n+1}\Big)Q_n(x) - \Big(1 + \frac{\alpha}{n}\Big)Q_{n-1}(x)\,, \quad n \geq 1\,, \tag{2.1}
$$

$$
Q_0(x) = 1\,, \quad Q_1(x) = x - \alpha - 1\,. \tag{2.2}
$$

If we write Q_n in the form

$$
Q_n(x) = x^n - a_{n-1,n}\, x^{n-1} + a_{n-2,n}\, x^{n-2} - \cdots + (-1)^n\, a_{0,n}\,,
$$

then

$$
a_{0,n} = \binom{n+\alpha}{n}\,, \qquad n \in \mathbb{N}_0\,, \tag{2.3}
$$

with the convention that the right-hand side is equal to 1 for $n = 0$. The proof is by induction with respect to n. For $n = 0,\, 1$, (2.3) follows from (2.2). Assuming (2.3) is true for all $m \le n$, we verify it for $m = n+1$ by putting $x = 0$ in (2.1) and using the induction hypothesis:

$$\begin{aligned}(-1)^{n+1} a_{0,n+1} &= \Big(2 + \frac{\alpha}{n+1}\Big)(-1)^{n+1}\binom{n+\alpha}{n} + \Big(1 + \frac{\alpha}{n}\Big)(-1)^n \binom{n-1+\alpha}{n-1} \\ &= (-1)^{n+1}\binom{n+1+\alpha}{n}.\end{aligned}$$

Now, instead of $\{Q_n\}_{n=0}^{\infty}$, we consider the sequence of orthogonal polynomials $\{\widetilde{Q}_n\}_{n=0}^{\infty}$ normalized so that $\widetilde{Q}_n(0) = 1$, $n \in \mathbb{N}_0$, i.e.,

$$Q_n(x) = (-1)^n \binom{n+\alpha}{n} \widetilde{Q}_n(x), \qquad n \in \mathbb{N}_0.$$

It follows from (2.1) and (2.2) that $\{\widetilde{Q}_n\}_{n\in\mathbb{N}_0}$ are determined by

$$\Big(1 + \frac{\alpha}{n+1}\Big)\widetilde{Q}_{n+1}(x) = \Big(2 + \frac{\alpha}{n+1} - x\Big)\widetilde{Q}_n(x) - \widetilde{Q}_{n-1}(x), \quad n \ge 1, \tag{2.4}$$

$$\widetilde{Q}_0(x) = 1, \quad \widetilde{Q}_1(x) = 1 - \frac{x}{\alpha+1}. \tag{2.5}$$

Writing $\widetilde{Q}_n$ in the form

$$\widetilde{Q}_n(x) = 1 - A_{1,n}\, x + A_{2,n}\, x^2 - \cdots + (-1)^n A_{n,n}\, x^n$$

and rewriting (2.4) as

$$\widetilde{Q}_{n+1}(x) - \widetilde{Q}_n(x) = \frac{n+1}{n+\alpha+1}\big(\widetilde{Q}_n(x) - \widetilde{Q}_{n-1}(x)\big) + \frac{n+1}{n+\alpha+1}\, x\, \widetilde{Q}_n(x), \quad n \in \mathbb{N},$$

we deduce the following recurrence relation for the evaluation of the coefficients $\{A_{i,m}\}$:

$$\begin{aligned}&A_{i,n+1} - A_{i,n} = \frac{n+1}{n+\alpha+1}\big(A_{i,n} - A_{i,n-1}\big) + \frac{n+1}{n+\alpha+1}\, A_{i-1,n}, \quad n \ge k \ge 1, \\ &\text{with } A_{0,n} = 1 \text{ and } A_{1,1} = \frac{1}{\alpha+1}.\end{aligned} \tag{2.6}$$

Since, by Proposition 1, $c_n^{-2}(\alpha)$ is equal to the smallest zero of $\widetilde{Q}_n$, it follows that $c_n^2(\alpha)$ equals the largest zero of the reciprocal polynomial of $\widetilde{Q}_n$,

$$R_n(x) = x^n\, \widetilde{Q}_n(1/x). \tag{2.7}$$

The above observations allow us to reformulate Proposition 1 in the following equivalent form:

Proposition 2. *The squared best Markov constant* $c_n^2(\alpha)$ *is equal to the largest zero of the polynomial*

$$R_n(x) = x^n - A_{1,n}\, x^{n-1} + A_{2,n}\, x^{n-2} - \cdots + (-1)^n A_{n,n}\,. \tag{2.8}$$

The coefficients of R_n *are evaluated recursively by the following procedure:*

- $A_{1,1} = \frac{1}{\alpha+1}$ *;*
- *Set* $A_{0,m} = 1\,,\ m = 0, \ldots, n$ *;*
- *For* $i = 1$ *to* n*:*
  1. *Find the sequence* $\{D_{i,m}\}_{m=i-1}^{n}$ *as solution of the recurrence equation*

$$D_{i,m+1} = \frac{m+1}{m+\alpha+1}\, D_{i,m} + \frac{m+1}{m+\alpha+1}\, A_{i-1,m} \tag{2.9}$$

  with the initial condition $D_{i,i-1} = 0$ *;*

  2. *Evaluate*

$$A_{i,n} = \sum_{m=i}^{n} D_{i,m}\,. \tag{2.10}$$

2.2. POLYNOMIALS WITH POSITIVE ROOTS: BOUNDS FOR THE LARGEST ZERO

Let P be a monic polynomial of degree n with zeros $\{x_i\}_{i=1}^{n}$,

$$P(x) = \prod_{i=1}^{n}(x - x_i) = x^n - b_1\, x^{n-1} + b_2\, x^{n-2} - \cdots + (-1)^n b_n\,.$$

The coefficients $b_r = b_r(P)\,,\ r = 1, \ldots, n$, are given by the elementary symmetric functions of $\{x_i\}_{i=1}^{n}$,

$$b_r = s_r = s_r(P) = \sum_{1 \le i_1 < i_2 < \cdots < i_r \le n} x_{i_1} x_{i_2} \cdots x_{i_r}\,, \qquad r = 1, \ldots, n\,.$$

It is well known that the elementary symmetric functions $\{s_r\}$ and the Newton functions (sums of powers of x_i)

$$p_r = p_r(P) = \sum_{i=1}^{n} x_i^r\,, \qquad r = 1, 2, 3, \ldots\,,$$

are connected by the Newton identities:

$$p_r + \sum_{i=1}^{r-1} (-1)^i p_{r-i}\, s_i + (-1)^r r\, s_r = 0\,, \qquad \text{if } 1 \le r \le n\,, \tag{2.11}$$

$$p_r + \sum_{i=1}^{n} (-1)^i p_{r-i}\, s_i = 0\,, \qquad \text{if } r > n\,. \tag{2.12}$$

For a proof, see e.g. [10] or [4].

Our interest in the Newton functions is motivated by the fact that they provide tight bounds for the largest zero of a polynomial whose roots are all positive. For any such polynomial P, we set

$$\ell_k(P) := \frac{p_k(P)}{p_{k-1}(P)}\,, \qquad u_k(P) := \big[p_k(P)\big]^{1/k}\,, \qquad k \in \mathbb{N}\,,$$

with the convention that $p_0(P) := \deg(P)$.

Proposition 3. *Let* $P(x) = x^n - b_1\,x^{n-1} + b_2\,x^{n-2} - \cdots + (-1)^{n-1}b_{n-1}\,x + (-1)^n b_n$ *be a polynomial with positive zeros* $x_1 \le x_2 \le \cdots \le x_n$.

Then the largest zero x_n *of* P *satisfies the inequalities*

$$\ell_k(P) \le x_n < u_k(P)\,, \qquad k \in \mathbb{N}\,. \tag{2.13}$$

Moreover, the sequence $\{\ell_k(P)\}_{k=1}^{\infty}$ *is monotonically increasing while the sequence* $\{u_k(P)\}_{k=1}^{\infty}$ *is monotonically decreasing, and*

$$\lim_{k\to\infty} \ell_k(P) = \lim_{k\to\infty} u_k(P) = x_n\,. \tag{2.14}$$

Proof. For $i = 1, \ldots, n-1$, we set $a_i := \frac{x_i}{x_n}$, then $0 < a_i \le 1$. Now both inequalities (2.13) and the limit relations (2.14) readily follow from the representations

$$\ell_k(P) = \frac{a_1^k + \cdots + a_{n-1}^k + 1}{a_1^{k-1} + \cdots + a_{n-1}^{k-1} + 1}\,x_n\,, \qquad u_k(P) = \big(a_1^k + \cdots + a_{n-1}^k + 1\big)^{1/k} x_n\,.$$

The monotonicity of the sequence $\{\ell_k(P)\}_{k=1}^{\infty}$ follows easily from Cauchy-Bouniakowsky's inequality. Indeed, we have

$$\Big(\sum_{i=1}^{n} x_i^k\Big)^2 = \Big(\sum_{i=1}^{n} x_i^{\frac{k-1}{2}}\, x_i^{\frac{k+1}{2}}\Big)^2 \le \Big(\sum_{i=1}^{n} x_i^{k-1}\Big)\Big(\sum_{i=1}^{n} x_i^{k+1}\Big)\,,$$

whence $p_k^2(P) \le p_{k-1}(P)\,p_{k+1}(P)$, and consequently

$$\ell_k(P) = \frac{p_k(P)}{p_{k-1}(P)} \le \frac{p_{k+1}(P)}{p_k(P)} = \ell_{k+1}(P)\,.$$

To prove monotonicity of the sequence $\{u_k(P)\}_{k=1}^{\infty}$, we recall that $0 < a_i \le 1$ and therefore $a_i^{k+1} \le a_i^k$. We have

$$\big(a_1^{k+1} + \cdots + a_{n-1}^{k+1} + 1\big)^{1/(k+1)} < \big(a_1^{k+1} + \cdots + a_{n-1}^{k+1} + 1\big)^{1/k} \le \big(a_1^k + \cdots + a_{n-1}^k + 1\big)^{1/k},$$

which yields

$$u_{k+1}(P) < u_k(P)\,.$$

□

3. COMPUTER ALGEBRA ASSISTED PROOF OF THE RESULTS

Here we give the algorithms, the source code and the results of the computer algebra assisted proof of estimates (1.10)-(1.17) in Theorem 1. While the case $k = 3$ and to a certain extent $k = 4$ could be studied by hand, it seems impossible to provide similar calculations for larger k. We implement the idea from [5] for estimating $c_n(\alpha)$ using $k = 3$ highest degree coefficients of the polynomial $R_n(x)$ and with the assistance of Wolfram's *Mathematica* v. 10 software we investigate the cases $k = 4, 5, 6$, as well. Software based on the algorithms described below failed with calculations for $k > 6$.

Henceforth, we write the polynomial R_n from (2.7) and (2.8) in the form

$$R_n(x) = x^n - b_1 x^{n-1} + b_2 x^{n-2} + \cdots + (-1)^n b_n .$$

3.1. LOWER BOUNDS FOR $c_n(\alpha)$

We apply Proposition 3 to estimate the largest zero $x_n = c_n^2(\alpha)$ of the polynomial $R_n(x)$ from below,

$$x_n \geq \ell_k(R_n) = \frac{p_k(R_n)}{p_{k-1}(R_n)}, \qquad k = 3,\, 4,\, 5,\, 6,$$

and then with the help of computer algebra obtain a further estimation of the form

$$\ell_k(R_n) \geq c\, n(n + \sigma(\alpha + 1)),$$

with the optimal (i.e., the largest possible) constants $c = c(k)$ and $\sigma = \sigma(k)$.

Algorithm 1 Estimating $c_n(\alpha)$ from below

Input:	$k \in \{3, 4, 5, 6\}$ – the number of the highest degree coefficients of $R_n(x)$
Step 1.	Express the power sums $p_{k-1}(R_n)$ and $p_k(R_n)$ in terms of $\{b_i\}_{i=1}^k$
Step 2.	Find coefficients $\{b_i\}_{i=1}^k$ in terms of n and α using Proposition 2
Step 3.	Find a proper value σ for parameter s in $p_k - c\, n(n + s(\alpha + 1))p_{k-1}$, where c is the coefficient of n^2 in the quotient p_k/p_{k-1}
Step 4.	Represent the numerator of $f = p_k - c\, n(n + \sigma(\alpha + 1))p_{k-1}$ in powers of n and $(\alpha + 1)$
Step 5.	Estimate from below the expression f to prove that $f \geq 0$

Step 1: Let $\{x_i\}_{i=1}^n$ be all the zeros of the polynomial $R_n(x)$ from (2.7). In order to express a power sum $p_r = \sum_{i=1}^r x_i^r$, $1 \leq r \leq n$, by $\{b_i\}_{i=1}^r$, we apply the direct formula

$$p_r = \begin{vmatrix} b_1 & 1 & 0 & \cdots & 0 \\ 2b_2 & b_1 & 1 & \cdots & 0 \\ 3b_3 & b_2 & b_1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ rb_r & b_{r-1} & b_{r-2} & \cdots & b_1 \end{vmatrix}$$

which easily follows from the Newton identities (2.11).

Below is the code of the programme and the results for $k = 1, \dots, 6$:

```
k = 6;
Do[p_κ = Det[Table[Which[j == 1, i b_i, 1 < j ≤ i, b_{i+1-j}, j == i+1, 1, j > i+1, 0], {i, κ}, {j, κ}]];
   Print[Subscript["p", κ], "=", TraditionalForm[p_κ]], {κ, k}]
```

$$p_1=b_1$$
$$p_2=b_1^2-2\,b_2$$
$$p_3=b_1^3-3\,b_2\,b_1+3\,b_3$$
$$p_4=b_1^4-4\,b_2\,b_1^2+4\,b_3\,b_1+2\,b_2^2-4\,b_4$$
$$p_5=b_1^5-5\,b_2\,b_1^3+5\,b_3\,b_1^2+5\,b_2^2\,b_1-5\,b_4\,b_1-5\,b_2\,b_3+5\,b_5$$
$$p_6=b_1^6-6\,b_2\,b_1^4+6\,b_3\,b_1^3+9\,b_2^2\,b_1^2-6\,b_4\,b_1^2-12\,b_2\,b_3\,b_1+6\,b_5\,b_1-2\,b_2^3+3\,b_3^2+6\,b_2\,b_4-6\,b_6$$

Step 2: We find coefficients $\{b_i\}_{i=1}^k$ of the polynomial $R_n(x)$ using Proposition 2. The source and the results for $k = 1, \dots, 6$ follow below:

```
k = 6;
fb[κ_, n_] :=
  If[κ == 1, Sum[FullSimplify[RSolveValue[{ru[q+1] == (ru[q]+1) (q+1)/(q+1+α), ru[1] == 1/(α+1)}, ru[q], q]], {q, 1, n}],
    Sum[Simplify[RSolveValue[{rv[q+1] == (rv[q]+fb[κ-1, q]) (q+1)/(q+1+α), rv[1] == 0}, rv[q], q]], {q, 1, n}]]
Do[If[κ == 1, b_κ = fb[κ, n],
    b_κ = Factor[Part[FactorTermsList[Numerator[fb[κ, n]], α], 2]] *
        Collect[Part[FactorTermsList[Numerator[fb[κ, n]], α], 3], n, FullSimplify] / Denominator[fb[κ, n]]];
  Print[Subscript["b", κ], "=", TraditionalForm[b_κ]], {κ, 1, k}]
```

$$b_1=\frac{n\,(n+1)}{2\,(\alpha+1)}$$
$$b_2=\frac{(n-1)\,n\,(n+1)\,(3\,n\,(\alpha+2)+2\,(\alpha+6))}{24\,(\alpha+1)\,(\alpha+2)\,(\alpha+3)}$$
$$b_3=\frac{(n-2)\,(n-1)\,n\,(n+1)\left(5\,(\alpha+2)\,(\alpha+4)\,n^2+(\alpha\,(5\,\alpha+86)+200)\,n+12\,(\alpha+20)\right)}{240\,(\alpha+1)\,(\alpha+2)\,(\alpha+3)\,(\alpha+4)\,(\alpha+5)}$$
$$b_4=\Big((n-3)\,(n-2)\,(n-1)\,n\,(n+1)\left(105\,(\alpha+2)\,(\alpha+4)\,(\alpha+6)\,n^3+3\,(\alpha\,(7\,\alpha\,(5\,\alpha+204)+9316)+15\,120)\,n^2 + (131\,040-2\,\alpha\,(7\,\alpha\,(5\,\alpha+44)-17\,244))\,n-8\,(\alpha\,(7\,\alpha\,(\alpha+28)+2244)-15\,120)\right)\Big)\Big/ (40\,320\,(\alpha+1)\,(\alpha+2)\,(\alpha+3)\,(\alpha+4)\,(\alpha+5)\,(\alpha+6)\,(\alpha+7))$$
$$b_5=\Big((n-4)\,(n-3)\,(n-2)\,(n-1)\,n\,(n+1)\left(21\,(\alpha+2)\,(\alpha+4)\,(\alpha+6)\,(\alpha+8)\,n^4 + 2\,(\alpha\,(\alpha\,(7\,\alpha\,(\alpha+108)+9956)+42\,928)+56\,448)\,n^3+(\alpha\,(\alpha\,(17\,988-7\,\alpha\,(7\,\alpha+212))+248\,496)+572\,544)\,n^2 + (1\,241\,856-2\,\alpha\,(\alpha\,(\alpha\,(21\,\alpha+1096)+26\,468)-34\,832))\,n-240\,(\alpha\,(\alpha\,(\alpha+38)+1528)-4032)\right)\Big)\Big/ (80\,640\,(\alpha+1)\,(\alpha+2)\,(\alpha+3)\,(\alpha+4)\,(\alpha+5)\,(\alpha+6)\,(\alpha+7)\,(\alpha+8)\,(\alpha+9))$$
$$b_6=\Big((n-5)\,(n-4)\,(n-3)\,(n-2)\,(n-1)\,n\,(n+1) \left(3465\,(\alpha+2)\,(\alpha+4)\,(\alpha+6)\,(\alpha+8)\,(\alpha+10)\,n^5+360\,(13\,\alpha\,(11\,\alpha\,(\alpha\,(7\,\alpha+164)+1348)+49\,936)+739\,200)\,n^4 + 9\,(\alpha\,(131\,884\,640-11\,\alpha\,(\alpha\,(35\,\alpha\,(5\,\alpha+278)-10\,644)-1\,805\,704))+229\,152\,000)\,n^3 - 8\,(\alpha\,(11\,\alpha\,(\alpha\,(5\,\alpha\,(28\,\alpha+2685)+620\,812)+2\,759\,292)-220\,067\,280)-964\,656\,000)\,n^2 + 44\,(\alpha\,(\alpha\,(\alpha\,(5\,\alpha\,(35\,\alpha+1014)-37\,756)-20\,283\,336)-53\,575\,200)+315\,705\,600)\,n + 96\,(\alpha\,(11\,\alpha\,(\alpha\,(5\,\alpha\,(\alpha+66)+7714)+237\,564)-62\,191\,440)+99\,792\,000)\right)\Big)\Big/ (159\,667\,200\,(\alpha+1)\,(\alpha+2)\,(\alpha+3)\,(\alpha+4)\,(\alpha+5)\,(\alpha+6)\,(\alpha+7)\,(\alpha+8)\,(\alpha+9)\,(\alpha+10)\,(\alpha+11))$$

Step 3: The quotient p_k/p_{k-1} is a quadratic polynomial in n, and we denote by c its leading coefficient.

The goal of this step is to find a proper value (say σ) for parameter s in the expression

$$f_s = p_k - c\,n(n + s(\alpha+1))p_{k-1}\,,$$

such that $f_\sigma \geq 0$ for all admissible α and n. For a fixed k quantity f_s depends on α, n and s. It is a polynomial of degree $2k-1$ in n and a rational function in α. Let us write the numerator of f_s in the form

$$\sum_{i=1}^{2k-1}\sum_{j=0}^{d}\mu_{i,j}(s)(\alpha+1)^{d-j}n^{2k-i}\,.$$

The highest order coefficients in $\sum_j \mu_{i,j}(s)(\alpha+1)^{d-j}$ are linear functions in s of the form $A_i - B_i s$, with $A_i > 0$ and $B_i > 0$. We denote their zeros by s_i for each i and set $\sigma = \min_i s_i$. Since we seek estimates valid for all $\alpha > -1$, our choice of σ guarantee that for α sufficiently large the inequality $\sum_j \mu_{i,j}(s)(\alpha+1)^{d-j} > 0$ holds true.

The code is as follows:

```
r = PolynomialQuotient[pk, pk−1, n];
c = Factor[Coefficient[r, n, 2]];
fs = pk − c n (n + s (α + 1)) pk−1;
numfs = Numerator[Together[Apart[fs, α]]]
Do[gs = Factor[Coefficient[numfs, n, i]];
    num = Normal[Series[ gs , {α, −1, Exponent[gs, α]}]];
    sols = Solve[Coefficient[num, α, Exponent[gs, α]] == 0, s, Reals];
    ss[i] = s /. Flatten[sols], {i, 2 k − 1, 1, −1}]
σ = Min[Table[ss[i], {i, 2, 2 k − 1}]];
```

Table 1 gives results for the optimal values of c and σ for $k = 3, 4, 5, 6$.

Table 1: The optimal values of c and σ in the lower bounds for $c_n^2(\alpha)$.

k	c	σ
3	$\dfrac{2}{(\alpha+1)(\alpha+5)}$	$\dfrac{3}{8}$
4	$\dfrac{5\alpha+17}{2(\alpha+1)(\alpha+3)(\alpha+7)}$	$\dfrac{8}{25}$
5	$\dfrac{2(7\alpha+31)}{(\alpha+1)(\alpha+9)(5\alpha+17)}$	$\dfrac{25}{84}$
6	$\dfrac{21\alpha^3+299\alpha^2+1391\alpha+2073}{(\alpha+1)(\alpha+3)(\alpha+5)(\alpha+11)(7\alpha+31)}$	$\dfrac{2}{7}$

Step 4: We set

$$f = p_k - c\,n(n+\sigma(\alpha+1))p_{k-1} =: \frac{\varphi(n,\alpha)}{\psi(\alpha)}$$

with c and σ determined in Step 3. Here, $\varphi(n,\alpha)$ is a bivariate polynomial in n and α, and $\psi(\alpha)$ is a polynomial in α. More precisely, $\varphi(n,\alpha)$ has degree $2k-1$ in n, and degree d in α which our programme calculates for each fixed k.

Note that $\psi(\alpha) > 0$ for $\alpha > -1$ since it is a product of powers of $\alpha + j$, $j \geq 1$ and multipliers $A\alpha + B$, $0 < A < B$. Therefore, $\operatorname{sign} f = \operatorname{sign} \varphi$.

We expand $\varphi(n,\alpha)$ in the form

$$\varphi(n,\alpha) = \sum_{i=1}^{2k-1} \sum_{j=0}^{d} \mu_{i,j} (\alpha+1)^{d-j} n^{2k-i} = \begin{pmatrix} n^{2k-1} \\ n^{2k-2} \\ \vdots \\ n \end{pmatrix}^{\top} \mathbf{M} \begin{pmatrix} (\alpha+1)^{d} \\ (\alpha+1)^{d-1} \\ \vdots \\ 1 \end{pmatrix},$$

where $\mathbf{M} = \left(\mu_{i,j}\right)_{i=1,j=0}^{2k-1,d}$ and all entries $\mu_{i,j}$ are integer numbers.

The source for computation of the matrix $\mathbf{M}$ is listed below.

```
f = pk − c n (n + σ (α + 1)) pk−1;
φ = Numerator[Together[Apart[f, α]]];
ψ = Denominator[Together[Apart[f, α]]];
Do[g = Factor[Coefficient[φ, n, i]]; dag[i] = Exponent[g, α] , {i, 2 k − 1, 1, −1}]
d = Max[Table[dag[i], {i, 1, 2 k − 1}]] + 1;
μ = ConstantArray[0, {2 k − 1, d}];
Do[g = Factor[Coefficient[φ, n, i]];
    Table[μ[[2 k − i, d − j]] = SeriesCoefficient[Series[g, {α, −1, dag[i]}], j], {j, 0, dag[i]}],
    {i, 2 k − 1, 1, −1}];
```

If $\mu_{i,j} \geq 0$ for all i,j, then $\varphi(n,\alpha) \geq 0$ and $f \geq 0$ for all $\alpha > -1$ and $n \geq k$. In a case some of coefficients $\mu_{i,j} < 0$ we apply the next step of the algorithm.

The results for $k = 3,4,5,6$ are given together with the estimates from Step 5.

Step 5: If there are coefficients $\mu_{i,j} < 0$ we need additional arguments to verify that $f \geq 0$ for all $\alpha > -1$ and $n \geq k$. We bring into use a new $(2k-1) \times (d+1)$ matrix $\mathbf{\Lambda}$ which elements we put initially $\lambda_{i,j} := \mu_{i,j}$, for $i = 1, \ldots, 2k-1$ and $j = 0, \ldots, d$.

The procedure described below checks recursively all coefficients $\lambda_{i,j}$ and makes the corresponding estimations. We need not introduce a new matrix after each iteration, but only replace a pair of elements in a column of $\mathbf{\Lambda}$ with new entries in such a manner that the value of the function

$$\Phi(\mathbf{\Lambda}) = \sum_{i=1}^{2k-1} \sum_{j=0}^{d} \lambda_{i,j} (\alpha+1)^{d-j} n^{2k-i} = \begin{pmatrix} n^{2k-1} \\ n^{2k-2} \\ \vdots \\ n \end{pmatrix}^{\top} \mathbf{\Lambda} \begin{pmatrix} (\alpha+1)^{d} \\ (\alpha+1)^{d-1} \\ \vdots \\ 1 \end{pmatrix}$$

decreases. At the end of the procedure we get a matrix $\mathbf{\Lambda}$ satisfying $\mathbf{0} \leq \mathbf{\Lambda} \leq \mathbf{M}$ (in the sense that $0 \leq \lambda_{i,j} \leq \mu_{i,j}$ for all i,j) and therefore

$$0 \leq \Phi(\mathbf{\Lambda}) \leq \Phi(\mathbf{M}) = \varphi(n,\alpha).$$

Suppose that $\lambda_{i,j} < 0$ for some pair of indices i, j. Then we set

$$h := \min\{i - \eta : \lambda_{\eta,j} > 0,\ 1 \le \eta \le i-1\} \quad \text{and} \quad \delta := \frac{\lambda_{i,j}}{k^{i-h}} \quad (\delta < 0)\,.$$

If $\lambda_{h,j} + \delta \ge 0$, for $n \ge k$ we have

$$\begin{aligned}(\lambda_{h,j} + \delta)n^{2k-h} + 0\, n^{2k-i} &= \Big(\lambda_{h,j} + \frac{\lambda_{i,j}}{k^{i-h}}\Big)n^{2k-h} = \lambda_{h,j}n^{2k-h} + \lambda_{i,j}\frac{n^{2k-h}}{k^{i-h}} \\ &\le \lambda_{h,j}n^{2k-h} + \lambda_{i,j}\frac{n^{2k-h}}{n^{i-h}} = \lambda_{h,j}n^{2k-h} + \lambda_{i,j}n^{2k-i}\,.\end{aligned}$$

Otherwise, if $\lambda_{h,j} + \delta < 0$, for $n \ge k$ we have

$$\begin{aligned}0\, n^{2k-h} + \big(\lambda_{h,j}k^{i-h} + \lambda_{i,j}\big)n^{2k-i} &= \lambda_{h,j}n^{2k-i}k^{i-h} + \lambda_{i,j}n^{2k-i} \\ &\le \lambda_{h,j}n^{2k-i}n^{i-h} + \lambda_{i,j}n^{2k-i} \\ &\le \lambda_{h,j}n^{2k-h} + \lambda_{i,j}n^{2k-i}\,.\end{aligned}$$

So, replacing only two elements in $\mathbf{\Lambda}$,

$$\begin{cases}\lambda_{h,j} := \lambda_{h,j} + \lfloor\delta\rfloor & \text{and} \quad \lambda_{i,j} := 0\,, \quad \text{if } \lambda_{h,j} + \delta \ge 0, \\ \lambda_{i,j} := \lambda_{h,j}\, k^{i-h} + \lambda_{i,j} & \text{and} \quad \lambda_{h,j} := 0\,, \quad \text{otherwise}\,,\end{cases}$$

we obtain that

$$\lambda_{h,j}(\alpha+1)^{d+1-j}n^{2k-h} + \lambda_{i,j}(\alpha+1)^{d+1-j}n^{2k-i}$$

decreases for the new values of $\lambda_{h,j}$ and $\lambda_{i,j}$, and hence $\Phi(\mathbf{\Lambda})$ also decreases.

Applying recursively the above iteration process for $i = 2k-1, 2k-2, \dots, 1$ and $j = 0, 1, \dots, d$ we finally obtain a matrix $\mathbf{\Lambda}$ satisfying $\mathbf{0} \le \mathbf{\Lambda} \le \mathbf{M}$. Then $\varphi(n, \alpha) \ge 0$, $f \ge 0$ and therefore

$$c_n^2(\alpha) \ge \frac{p_k}{p_{k-1}} \ge c\, n(n + \sigma(\alpha+1))$$

for the optimal c and σ evaluated in Step 3. For $k = 3, 4, 5, 6$ we obtain estimates (1.10), (1.12), (1.14), and (1.16), respectively.

The following source implements the procedure described in Step 5.

```
λ = μ;
For[i = 2 k − 1, i > 1, i−−,
    For[j = 1, j ≤ d, j++, If[λ[[i, j]] ≥ 0, Continue[]];
        h = i − First[FirstPosition[Positive[λ[[i − 1 ;; 1 ;; −1, j]]], True]];
        δ = λ[[i, j]] / (k^(i − h));
        If[λ[[h, j]] + δ ≥ 0, λ[[h, j]] = λ[[h, j]] + Floor[δ]; λ[[i, j]] = 0,
                λ[[i, j]] = λ[[h, j]] * k^(i − h) + λ[[i, j]]; λ[[h, j]] = 0; i = i + 1]]]
Print["Λ = ", MatrixForm[λ]]
Print["M = ", MatrixForm[μ]]
```

Next, we give matrices $\mathbf{M}$ from Step 4 and $\mathbf{\Lambda}$ from Step 5 obtained with *Mathematica*.

Case $k = 3$:

This partial case needs a special attention as we have to assume strict inequality $n > k$, i.e., $n \geq 4$, to obtain estimate (1.10). This causes a minor modification in Step 5 of Algorithm 1, namely, replacement of k^{i-h} with $(k+1)^{i-h}$. Namely, we determine $\delta := \lambda_{i,j}/(k+1)^{i-h}$ and set

$$\begin{cases} \lambda_{h,j} := \lambda_{h,j} + \lfloor \delta \rfloor & \text{and } \lambda_{i,j} := 0, \quad \text{if } \lambda_{h,j} + \delta \geq 0, \\ \lambda_{i,j} := \lambda_{h,j}\,(k+1)^{i-h} + \lambda_{i,j} & \text{and } \lambda_{h,j} := 0, \quad \text{otherwise}. \end{cases}$$

Matrices $\mathbf{M}$ and $\mathbf{\Lambda}$ in this case are

$$\mathbf{\Lambda} = \begin{pmatrix} 0 & 4 & -4 & 225 & 360 \\ 0 & 0 & 390 & 510 & 720 \\ 15 & 155 & 205 & 1185 & 360 \\ 15 & 270 & 495 & 900 & 0 \\ 0 & 36 & 684 & 0 & 0 \end{pmatrix} \qquad \mathrm{M} = \begin{pmatrix} 0 & 19 & -4 & 225 & 360 \\ 0 & -60 & 390 & 510 & 720 \\ 15 & 155 & 205 & 1185 & 360 \\ 15 & 270 & 495 & 900 & 0 \\ 0 & 36 & 684 & 0 & 0 \end{pmatrix}.$$

Although there is a negative element of $\mathbf{\Lambda}$, from $4(\alpha+1)^2 - 4(\alpha+1) + 225 \geq 0$ for all $\alpha > -1$ we conclude that $4(\alpha+1)^3 - 4(\alpha+1)^2 + 225(\alpha+1) + 360 > 0$ and consequently $\Phi(\mathbf{\Lambda}) \geq 0$ for $n \geq 4$.

By a direct verification one can see that inequality (1.10) holds also in the case $n = k = 3$.

Case $k = 4$:

$$\mathbf{\Lambda} = \begin{pmatrix} 0 & 0 & 10200 & 72480 & 323700 & 1413060 & 3602340 & 4340700 & 1890000 \\ 0 & 4882 & 30891 & 359695 & 2625259 & 7966210 & 13275570 & 12707100 & 5670000 \\ 0 & 0 & 229110 & 1642830 & 6282570 & 16699200 & 24837120 & 18692100 & 5670000 \\ 2100 & 46515 & 120645 & 2404465 & 10159765 & 20026720 & 25810890 & 16625700 & 1890000 \\ 2756 & 106120 & 876330 & 2582090 & 7616630 & 17567550 & 18060000 & 6300000 & 0 \\ 0 & 11060 & 662604 & 2653840 & 6215776 & 11121880 & 7413000 & 0 & 0 \\ 0 & 0 & 0 & 1120600 & 4777900 & 3435000 & 0 & 0 & 0 \end{pmatrix}$$

$$\mathrm{M} = \begin{pmatrix} 0 & 0 & 10200 & 72480 & 323700 & 1413060 & 3602340 & 4340700 & 1890000 \\ 0 & 8715 & 30891 & 359695 & 2625259 & 7966210 & 13275570 & 12707100 & 5670000 \\ 0 & -15330 & 229110 & 1642830 & 6282570 & 16699200 & 24837120 & 18692100 & 5670000 \\ 2100 & 46515 & 120645 & 2404465 & 10159765 & 20026720 & 25810890 & 16625700 & 1890000 \\ 2800 & 106120 & 876330 & 2582090 & 7616630 & 17567550 & 18060000 & 6300000 & 0 \\ 0 & 15960 & 722904 & 2653840 & 6215776 & 11121880 & 7413000 & 0 & 0 \\ -700 & -19600 & -241200 & 1120600 & 4777900 & 3435000 & 0 & 0 & 0 \end{pmatrix}$$

Case $k = 5$:

$$\mathbf{\Lambda} = \begin{pmatrix} 0 & 0 & 0 & 64925 & 1064665 & 8138830 & 43256150 & 172898565 & 474925185 & 805850640 & 734423760 & 266716800 \\ 0 & 0 & 91665 & 1204470 & 9699090 & 71280390 & 373661895 & 1241223900 & 2610599670 & 3473555400 & 2804336640 & 1066867200 \\ 0 & 19824 & 130578 & 3408188 & 48487642 & 313463920 & 1271550350 & 3522779568 & 6544523790 & 7686433440 & 5117787360 & 1600300800 \\ 0 & 0 & 1451982 & 16288020 & 114900450 & 672910770 & 2546690160 & 6152610870 & 9859721760 & 10218685680 & 5871579840 & 1066867200 \\ 3675 & 128835 & 0 & 24490445 & 226233910 & 991504675 & 3153540110 & 7169071245 & 10438959825 & 9013742640 & 3935025360 & 266716800 \\ 6027 & 381850 & 6416795 & 22404550 & 169885205 & 1005110890 & 2985302145 & 5744010510 & 7716554370 & 5584488840 & 1111320000 & 0 \\ 0 & 52297 & 5062484 & 58263912 & 213196158 & 589342950 & 1804792500 & 3787471002 & 4038237000 & 1770703200 & 0 & 0 \\ 0 & 0 & 0 & 15084950 & 144208510 & 409403975 & 1057769610 & 1931913900 & 1309770000 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 256255650 & 690284700 & 417538800 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$\mathrm{M} = \begin{pmatrix} 0 & 0 & 0 & 64925 & 1064665 & 8138830 & 43256150 & 172898565 & 474925185 & 805850640 & 734423760 & 266716800 \\ 0 & 0 & 91665 & 1204470 & 9699090 & 71280390 & 373661895 & 1241223900 & 2610599670 & 3473555400 & 2804336640 & 1066867200 \\ 0 & 27804 & 130578 & 3408188 & 48487642 & 313463920 & 1271550350 & 3522779568 & 6544523790 & 7686433440 & 5117787360 & 1600300800 \\ 0 & -39900 & 1500030 & 16288020 & 114900450 & 672910770 & 2546690160 & 6152610870 & 9859721760 & 10218685680 & 5871579840 & 1066867200 \\ 3675 & 128835 & -240240 & 24490445 & 226233910 & 991504675 & 3153540110 & 7169071245 & 10438959825 & 9013742640 & 3935025360 & 266716800 \\ 6125 & 381850 & 6416795 & 22404550 & 169885205 & 1005110890 & 2985302145 & 5744010510 & 7716554370 & 5584488840 & 1111320000 & 0 \\ 0 & 77616 & 5699022 & 58263912 & 213196158 & 589342950 & 1804792500 & 3787471002 & 4038237000 & 1770703200 & 0 & 0 \\ -2450 & -123445 & -3055430 & 20292530 & 152590030 & 409403975 & 1057769610 & 1931913900 & 1309770000 & 0 & 0 & 0 \\ 0 & -15750 & -636300 & -26037900 & -41907600 & 256255650 & 690284700 & 417538800 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Case $k = 6$:

$$\Lambda^{\mathrm{T}} = \begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 48510 & 95223 & 0 & 0 & 0 & 16170 \\
0 & 0 & 0 & 425810 & 0 & 2817045 & 9741270 & 1348462 & 0 & 0 & 1252020 \\
0 & 0 & 3476550 & 6110115 & 48434732 & 0 & 336258384 & 215861747 & 0 & 0 & 38848656 \\
0 & 6055665 & 95465370 & 190273710 & 1221447150 & 1171139970 & 2726237052 & 7293843195 & 0 & 0 & 1158647028 \\
3128160 & 2045531 95 & 1490047030 & 5336244870 & 15771654360 & 32618391960 & 21628131756 & 73442566505 & 29020437724 & 0 & 0 \\
116263280 & 3318028175 & 18873326010 & 77596724865 & 174458095350 & 356484794820 & 298526144072 & 992659895320 & 419332019003 & 0 & 0 \\
1988081620 & 35746404925 & 197029544250 & 726747795015 & 1551387171180 & 2562636437130 & 2754345379016 & 1907270574440 & 2403530867450 & 384166681454 & 0 \\
21102099620 & 2903069613 29 & 15582069402 90 & 4960832042100 & 10157416978170 & 14397653320512 & 15830016304564 & 10335346465675 & 8454107203080 & 3721378395370 & 0 \\
157521933940 & 1842856573327 & 9151953918030 & 25828662738780 & 49078270584420 & 64463871381756 & 64790433176084 & 46215397662665 & 25413887653770 & 13866272170542 & 3624993260826 \\
879576036500 & 9097329993521 & 40326294432270 & 103410904320900 & 179681190528840 & 223720508502183 & 206003553429058 & 148200002432020 & 74515561079190 & 35329760467746 & 21352330210512 \\
3768921407020 & 34425402760287 & 134937782918400 & 317406347163180 & 506548245985320 & 592781349468231 & 516092578758680 & 349390783269990 & 182484813042840 & 86130722275092 & 36302262824520 \\
12408373123020 & 9858556853145 0 & 34452362690130 0 & 74272033283380 & 109886468687920 & 119510121225033 0 & 98829965916158 4 & 62262637618104 0 & 31519582447160 & 12781702216800 0 & 27594339093216 \\
30888195414300 & 21122313449001 86 & 66716115336486 0 & 131409020225480 & 180585150908500 & 181525834506277 8 & 139899372069394 0 & 80848519546410 0 & 35077036074696 0 & 10351097620617 6 & 79599114201 60 \\
56418683248620 & 33395606966106 0 & 96143321993796 0 & 173065873703184 0 & 218454701519974 0 & 201147075904698 0 & 137161413358200 0 & 69102103177880 & 22702113846720 0 & 33927611477760 & 0 \\
72303376012560 & 38059749615888 0 & 99694890095328 0 & 164078227738656 0 & 186198023315304 0 & 152094450250512 0 & 87654535646832 0 & 33019934580864 0 & 64746646272000 & 0 & 0 \\
60744201708960 & 29766966158160 0 & 70488555207840 0 & 104554813698720 0 & 103655345991120 0 & 69581039075856 0 & 30053374686720 0 & 6368877123840 0 & 0 & 0 & 0 \\
29689237670400 & 14316519571200 0 & 30745436194800 0 & 39195024422400 0 & 31743795283200 0 & 15115206839040 0 & 31685955840000 & 0 & 0 & 0 & 0 \\
6337191168000 & 31685955840000 & 63371911680000 & 63371911680000 & 31685955840000 & 6337191168000 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}$$

$$\mathrm{M}^{\mathrm{T}} = \begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 48510 & 97020 & 0 & −64680 & 0 & 16170 \\
0 & 0 & 0 & 544005 & −709170 & 2817045 & 9741270 & 2279970 & −5453910 & −810810 & 1252020 \\
0 & 0 & 3476550 & 6110115 & 51415980 & −17887485 & 336258384 & 259149660 & −23328459 0 & −50657310 & 38848656 \\
0 & 6055665 & 95465370 & 190273710 & 1221447150 & 1171139970 & 2726237052 & 7523825695 & −427583080 & −8310051520 & 1158647028 \\
3128160 & 2045531 95 & 1490047030 & 5336244870 & 15771654360 & 32618391960 & 21628131756 & 73442566505 & 36182631870 & −42053593230 & −5517429876 \\
116263280 & 3318028175 & 18873326010 & 77596724865 & 174458095350 & 356484794820 & 298526144072 & 992659895320 & 439184120760 & −87653879280 & −188752387572 \\
1988081620 & 35746404925 & 197029544250 & 726747795015 & 1551387171180 & 2562636437130 & 2754345379016 & 1907270574440 & 2403530867430 & 568697131530 & −1107182700456 \\
21102099620 & 2903069613 29 & 15582069402 90 & 4960832042100 & 10157416978170 & 14397653320512 & 15830016304564 & 10335346465675 & 8454107203080 & 4052140160904 & −1984570575204 \\
157521933940 & 1842856573327 & 9151953918030 & 25828662738780 & 49078270584420 & 64463871381756 & 64790433176084 & 46215397662665 & 25413887653770 & 13866272170542 & 3624993260826 \\
879576036500 & 9097329993521 & 40326294432270 & 103410904320900 & 179681190528840 & 223720508502183 & 206003553429058 & 148200002432020 & 74515561079190 & 35329760467746 & 21352330210512 \\
3768921407020 & 34425402760287 & 134937782918400 & 317406347163180 & 506548245985320 & 592781349468231 & 516092578758680 & 349390783269990 & 182484813042840 & 86130722275092 & 36302262824520 \\
12408373123020 & 9858556853145 0 & 34452362690130 0 & 74272033283380 & 109886468687920 & 119510121225033 0 & 98829965916158 4 & 62262637618104 0 & 31519582447160 & 12781702216800 0 & 27594339093216 \\
30888195414300 & 21122313449001 86 & 66716115336486 0 & 131409020225480 & 180585150908500 & 181525834506277 8 & 139899372069394 0 & 80848519546410 0 & 35077036074696 0 & 10351097620617 6 & 79599114201 60 \\
56418683248620 & 33395606966106 0 & 96143321993796 0 & 173065873703184 0 & 218454701519974 0 & 201147075904698 0 & 137161413358200 0 & 69102103177880 & 22702113846720 0 & 33927611477760 & 0 \\
72303376012560 & 38059749615888 0 & 99694890095328 0 & 164078227738656 0 & 186198023315304 0 & 152094450250512 0 & 87654535646832 0 & 33019934580864 0 & 64746646272000 & 0 & 0 \\
60744201708960 & 29766966158160 0 & 70488555207840 0 & 104554813698720 0 & 103655345991120 0 & 69581039075856 0 & 30053374686720 0 & 6368877123840 0 & 0 & 0 & 0 \\
29689237670400 & 14316519571200 0 & 30745436194800 0 & 39195024422400 0 & 31743795283200 0 & 15115206839040 0 & 31685955840000 & 0 & 0 & 0 & 0 \\
6337191168000 & 31685955840000 & 63371911680000 & 63371911680000 & 31685955840000 & 6337191168000 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}$$

3.2. UPPER BOUNDS FOR $c_n(\alpha)$

We apply Proposition 3 to estimate the largest zero $x_n = c_n^2(\alpha)$ of the polynomial $R_n(x)$ from above,

$$x_n \le u_k(R_n) = p_k(R_n)^{1/k}\,, \qquad k = 3,\ 4,\ 5,\ 6\,.$$

Then with the assistance of computer algebra we obtain a further estimation of the form

$$u_k(R_n) \le c^{1/k}\,(n+1)(n+\sigma(\alpha+1)),$$

with the optimal (i.e., the smallest possible) constants $c = c(k)$ and $\sigma = \sigma(k)$.

The algorithm is analogous to Algorithm 1, and the code has only a few differences which are specified later.

Algorithm 2 Estimating $c_n(\alpha)$ from above

Input:	$k \in \{3, 4, 5, 6\}$ – the number of the highest degree coefficients of $R_n(x)$
Step 1.	Express the power sum $p_k(R_n)$ in terms of $\{b_i\}_{i=1}^k$
Step 2.	Find $\{b_i\}_{i=1}^k$ in terms of n and α using Proposition 2
Step 3.	Find a proper value σ for parameter s in the expression $c\,(n+1)^k(n+s(\alpha+1))^k - p_k\,,$ where c is the coefficient of n^{2k} in p_k
Step 4.	Represent the numerator of $f = c\,(n+1)^k(n+\sigma(\alpha+1))^k - p_k$ in powers of n and $(\alpha+1)$
Step 5.	Estimate from below the expression f to prove that $f \ge 0$

Step 1: The same as in Algorithm 1.

Step 2: Identical to that in Algorithm 1.

Step 3: The only differences with Algorithm 1 are that we set c to be the coefficient of n^{2k} in p_k and

$$f_s = c\,(n+1)^k(n+s(\alpha+1))^k - p_k\,.$$

The highest order coefficients in $\sum_j \mu_{i,j}(s)(\alpha+1)^{d-j}$ are functions in s of the form $A_i s^\nu - B_i$, with $A_i > 0$ and $B_i \geq 0$. We denote their non-negative zeros by s_i for each i and choose $\sigma = \max_i s_i$.

The results for $k = 3, 4, 5, 6$ obtained by symbolic computations are given in Table 2.

Table 2: The optimal values of c and σ in the upper bounds for $c_n^2(\alpha)$.

k	c	σ
3	$\dfrac{1}{(\alpha+1)^3(\alpha+3)(\alpha+5)}$	$\dfrac{2}{5}$
4	$\dfrac{5\alpha+17}{2(\alpha+1)^4(\alpha+3)^2(\alpha+5)(\alpha+7)}$	$\dfrac{3}{7}$
5	$\dfrac{(7\alpha+31)}{(\alpha+1)^5(\alpha+3)^2(\alpha+5)(\alpha+7)(\alpha+9)}$	$\dfrac{4}{9}$
6	$\dfrac{21\alpha^3+299\alpha^2+1391\alpha+2073}{(\alpha+1)^6(\alpha+3)^3(\alpha+5)^2(\alpha+7)(\alpha+9)(\alpha+11)}$	$\dfrac{5}{11}$

Step 4: With c and σ determined in the previous Step 3 we set

$$f = c\,(n+1)^k(n+\sigma(\alpha+1))^k - p_k =: \frac{\varphi(n,\alpha)}{\psi(\alpha)}\,.$$

The rest of the source has no difference with Step 4 of Algorithm 1.

Step 5: The same as in Algorithm 1. Using the same recursive procedure we find a matrix $\mathbf{\Lambda}$ satisfying $\mathbf{0} \leq \mathbf{\Lambda} \leq \mathbf{M}$. Then $\varphi(n,\alpha) \geq 0$, $f \geq 0$ and therefore

$$c_n^{2k}(\alpha) \leq p_k \leq c\,(n+1)^k(n+\sigma(\alpha+1))^k$$

for the corresponding c and σ evaluated in Step 3. For $k = 3, 4, 5, 6$ we obtain estimations (1.11), (1.13), (1.15), and (1.17), respectively.

The matrices $\mathbf{M}$ from Step 4 and $\mathbf{\Lambda}$ from Step 5 obtained with *Mathematica* are given below.

Case $k = 3$:

$$\mathbf{\Lambda} = \begin{pmatrix} 0 & 0 & 0 & 1500 & 3300 \\ 0 & 115 & 1885 & 4170 & 4233 \\ 32 & 598 & 3026 & 6360 & 0 \\ 96 & 979 & 2143 & 850 & 0 \\ 96 & 624 & 1098 & 0 & 0 \end{pmatrix} \qquad \mathrm{M} = \begin{pmatrix} 0 & 0 & 0 & 1500 & 3300 \\ 0 & 115 & 1885 & 4170 & 4650 \\ 32 & 598 & 3026 & 6360 & -600 \\ 96 & 979 & 2143 & 1560 & -1950 \\ 96 & 624 & 1098 & -2130 & 0 \end{pmatrix}$$

Case $k = 4:$

$$\Lambda = \begin{pmatrix} 0 & 0 & 0 & 0 & 905520 & 8808240 & 29717520 & 41571600 & 19756800 \\ 0 & 0 & 54390 & 2038890 & 16676660 & 60285680 & 115770830 & 117031110 & 48774600 \\ 0 & 42294 & 1237572 & 10966494 & 52723608 & 141477042 & 198565500 & 127823850 & 24194362 \\ 6075 & 266115 & 3694950 & 25364010 & 85166735 & 157047575 & 154257320 & 46893642 & 0 \\ 24300 & 617510 & 5700800 & 26734470 & 72437020 & 97039330 & 34815501 & 0 & 0 \\ 36450 & 678780 & 4979940 & 16392810 & 28823750 & 17907835 & 0 & 0 & 0 \\ 24300 & 360421 & 2131108 & 6792156 & 5246162 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$\mathrm{M} = \begin{pmatrix} 0 & 0 & 0 & 0 & 905520 & 8808240 & 29717520 & 41571600 & 19756800 \\ 0 & 0 & 54390 & 2038890 & 16676660 & 60285680 & 115770830 & 117031110 & 48774600 \\ 0 & 42294 & 1237572 & 10966494 & 52723608 & 141477042 & 198565500 & 127823850 & 27783000 \\ 6075 & 266115 & 3694950 & 25364010 & 85166735 & 157047575 & 154257320 & 52558380 & -11730600 \\ 24300 & 617510 & 5700800 & 26734470 & 72437020 & 97039330 & 38636640 & -18088350 & -10495800 \\ 36450 & 678780 & 4979940 & 16392810 & 28823750 & 20280800 & -12849340 & -18282390 & 0 \\ 24300 & 360421 & 2131108 & 6792156 & 5246162 & -9491857 & -9740850 & 0 & 0 \end{pmatrix}$$

Case $k = 5:$

$$\Lambda = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 85424220 & 1436596560 & 8988832440 & 26097558480 & 34662943980 & 16203045600 \\ 0 & 0 & 0 & 4261005 & 260814330 & 3617057430 & 22250151630 & 73071107235 & 134891273160 & 134642808090 & 56710659600 \\ 0 & 0 & 5436720 & 241567920 & 3235204800 & 22246774740 & 91003127400 & 223063050420 & 312360753600 & 222393230640 & 64812182400 \\ 0 & 1982358 & 88937982 & 1392482448 & 12340605438 & 63755213760 & 194677526736 & 357163148790 & 375802372260 & 186521488020 & 12638375568 \\ 200704 & 14563010 & 340432890 & 4020858058 & 25446365294 & 99455228208 & 241336266948 & 338611016520 & 235926284580 & 44541786567 & 0 \\ 1003520 & 42390775 & 693405300 & 6004806185 & 31876009900 & 96870254355 & 175080003840 & 176585507595 & 54286938720 & 0 & 0 \\ 2007040 & 63580160 & 829630410 & 5638883530 & 22495811450 & 57112266330 & 77686343280 & 30853075478 & 0 & 0 & 0 \\ 2007040 & 52428341 & 568553244 & 3375204826 & 9950248616 & 17535199185 & 13032227178 & 0 & 0 & 0 & 0 \\ 1003520 & 22758400 & 207566490 & 998218460 & 3486984100 & 3092469120 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$\mathrm{M} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 85424220 & 1436596560 & 8988832440 & 26097558480 & 34662943980 & 16203045600 \\ 0 & 0 & 0 & 4261005 & 260814330 & 3617057430 & 22250151630 & 73071107235 & 134891273160 & 134642808090 & 56710659600 \\ 0 & 0 & 5436720 & 241567920 & 3235204800 & 22246774740 & 91003127400 & 223063050420 & 312360753600 & 222393230640 & 64812182400 \\ 0 & 1982358 & 88937982 & 1392482448 & 12340605438 & 63755213760 & 194677526736 & 357163148790 & 375802372260 & 186521488020 & 16203045600 \\ 200704 & 14563010 & 340432890 & 4020858058 & 25446365294 & 99455228208 & 241336266948 & 338611016520 & 235926284580 & 51689001420 & -16203045600 \\ 1003520 & 42390775 & 693405300 & 6004806185 & 31876009900 & 96870254355 & 175080003840 & 176585507595 & 59214803760 & -31849915230 & -8101522800 \\ 2007040 & 63580160 & 829630410 & 5638883530 & 22495811450 & 57112266330 & 77686343280 & 32878980540 & -21278795580 & -19430795160 & 0 \\ 2007040 & 52428341 & 568553244 & 3375204826 & 9950248616 & 17535199185 & 14090589072 & -8987585040 & -16802648100 & 0 & 0 \\ 1003520 & 22758400 & 207566490 & 998218460 & 3486984100 & 3092469120 & -5291809470 & -5709701340 & 0 & 0 & 0 \end{pmatrix}$$

Case $k = 6:$

$$\Lambda^{\mathrm{T}} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 137812500 & 826875000 & 2067187500 & 2756250000 & 2067187500 & 826875000 \\ 0 & 0 & 0 & 0 & 1712831340 & 16225847190 & 61013597190 & 121629375000 & 141073384270 & 96161625000 & 35900159390 \\ 0 & 0 & 0 & 7214978925 & 129414811020 & 682431963570 & 1909641192060 & 3176038395495 & 3233829060980 & 1990746645930 & 685515044860 \\ 0 & 0 & 11987252200 & 496609292025 & 3975634217280 & 15545520413640 & 34393961720958 & 48625624140345 & 44026671755710 & 24295459922370 & 7604368197232 \\ 0 & 6757992780 & 889711262440 & 13291579355670 & 71880656791680 & 218617776740580 & 407202449242428 & 488239798537050 & 391807489052700 & 198688598846220 & 53934381373052 \\ 0 & 678222070680 & 23583750724380 & 204179315141340 & 848078318563740 & 2083353346583670 & 3322865459747106 & 3448886831378940 & 2400975627767010 & 1108865878357650 & 280425349855044 \\ 1566575287720 & 19589397975840 & 339201489248160 & 2064129233162535 & 6887056367079900 & 14142830057025810 & 19301411805553332 & 17606432763060855 & 10537484120843400 & 4243406630751540 & 1099649486098068 \\ 5737886375760 & 285634042298220 & 3098176511479280 & 14556440238205635 & 39911599747186200 & 69897865311080520 & 81762211401327198 & 65213552886377595 & 33832857309375390 & 11331273200216130 & 2850997375559272 \\ 90962828787600 & 2522414547688728 & 19367097472935520 & 73609280525046420 & 168732748048142520 & 253532902195144824 & 255297208204811052 & 174944618786085540 & 78706656458079300 & 21128915605090608 & 3776309107237628 \\ 821967442647120 & 14663875740671388 & 85966291726204120 & 270402692712289830 & 525000679384868280 & 675267348901207524 & 584508472246139364 & 337073370501014070 & 126378098307624240 & 25176310282435158 & 100310771384846 \\ 4683403445822640 & 58501440227708016 & 275002627721514000 & 724008553578199890 & 1198318687198670400 & 1310711831601226488 & 960008914211603088 & 452338940602722600 & 125459625652811660 & 10624247696162964 & 0 \\ 17541804701337840 & 163077338127380292 & 633513827175290600 & 1404602858405626110 & 1976258187012861840 & 1815905828323340796 & 1083067213944785244 & 386368902093765990 & 50117292233704841 & 0 & 0 \\ 43597315400007600 & 317712831386568156 & 1035759237211329400 & 1939380788446101000 & 2286621788179676640 & 1716811248944619828 & 768783125513892720 & 147265658296879789 & 0 & 0 & 0 \\ 70706025735594480 & 424476237961133820 & 1166286653916203820 & 1837004374966081860 & 1763385304898752920 & 1006399215292377060 & 256729802815313776 & 0 & 0 & 0 & 0 \\ 71282695085965440 & 370662526010533680 & 857715825988763280 & 1112754260403129960 & 820925242403999040 & 277604251541734810 & 0 & 0 & 0 & 0 & 0 \\ 40151827863100800 & 190176873072832800 & 371893024262944800 & 374764348674777600 & 176396296906922400 & 0 & 0 & 0 & 0 & 0 & 0 \\ 9572830210944000 & 43261828837920000 & 72716691025440000 & 49617294906564000 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$\mathrm{M}^{\mathrm{T}} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 137812500 & 826875000 & 2067187500 & 2756250000 & 2067187500 & 826875000 \\ 0 & 0 & 0 & 0 & 1712831340 & 16225847190 & 61013597190 & 121629375000 & 141073384270 & 96161625000 & 35900159390 \\ 0 & 0 & 0 & 7214978925 & 129414811020 & 682431963570 & 1909641192060 & 3176038395495 & 3233829060980 & 1990746645930 & 685515044860 \\ 0 & 0 & 11987252200 & 496609292025 & 3975634217280 & 15545520413640 & 34393961720958 & 48625624140345 & 44026671755710 & 24295459922370 & 7604368197232 \\ 0 & 6757992780 & 889711262440 & 13291579355670 & 71880656791680 & 218617776740580 & 407202449242428 & 488239798537050 & 391807489052700 & 198688598846220 & 53934381373052 \\ 0 & 678222070680 & 23583750724380 & 204179315141340 & 848078318563740 & 2083353346583670 & 3322865459747106 & 3448886831378940 & 2400975627767010 & 1108865878357650 & 280425349855044 \\ 1566575287720 & 19589397975840 & 339201489248160 & 2064129233162535 & 6887056367079900 & 14142830057025810 & 19301411805553332 & 17606432763060855 & 10537484120843400 & 4243406630751540 & 1099649486098068 \\ 5737886375760 & 285634042298220 & 3098176511479280 & 14556440238205635 & 39911599747186200 & 69897865311080520 & 81762211401327198 & 65213552886377595 & 33832857309375390 & 11331273200216130 & 2850997375559272 \\ 90962828787600 & 2522414547688728 & 19367097472935520 & 73609280525046420 & 168732748048142520 & 253532902195144824 & 255297208204811052 & 174944618786085540 & 78706656458079300 & 21128915605090608 & 3776309107237628 \\ 821967442647120 & 14663875740671388 & 85966291726204120 & 270402692712289830 & 525000679384868280 & 675267348901207524 & 584508472246139364 & 337073370501014070 & 126378098307624240 & 25176310282435158 & 100310771384846 \\ 4683403445822640 & 58501440227708016 & 275002627721514000 & 724008553578199890 & 1198318687198670400 & 1310711831601226488 & 960008914211603088 & 452338940602722600 & 125459625652811660 & 11652636597863226 & -6170333410201568 \\ 17541804701337840 & 163077338127380292 & 633513827175290600 & 1404602858405626110 & 1976258187012861840 & 1815905828323340796 & 1083067213944785244 & 386368902093765990 & 52601940894289780 & -13776737928943668 & -6786924207395784 \\ 43597315400007600 & 317712831386568156 & 1035759237211329400 & 1939380788446101000 & 2286621788179676640 & 1716811248944619828 & 768783125513892720 & 152704407947442120 & -28750678275282040 & -22906285043773824 & -2307796348667040 \\ 70706025735594480 & 424476237961133820 & 1166286653916203820 & 1837004374966081860 & 1763385304898752920 & 1006399215292377060 & 264555646816493340 & -39077291676501540 & -45675842674954800 & -9544747851001440 & 0 \\ 71282695085965440 & 370662526010533680 & 857715825988763280 & 1112754260403129960 & 820925242403999040 & 283521195216151200 & -25185070761515280 & -59085225024570360 & -16885936111968000 & 0 & 0 \\ 40151827863100800 & 190176873072832800 & 371893024262944800 & 374764348674777600 & 178350638317790400 & -40686454333221600 & -43598489040136800 & -14075574910689600 & 0 & 0 & 0 \\ 9572830210944000 & 43261828837920000 & 72716691025440000 & 49705079941440000 & 1840928886720000 & -13438780873056000 & -4602322216800000 & 0 & 0 & 0 & 0 \end{pmatrix}$$

4. CONCLUDING REMARKS

1. In our computer algebra approach for derivation of bounds for the best Markov constant $c_n(\alpha)$ we perform some optimization with respect to parameter s.

Our motivation for searching lower bounds for $c_n^2(\alpha)$ with a factor depending on n of the special form $n\big(n+\sigma(\alpha+1)\big)$ is Corollary D(ii).

An interesting observation about the lower bounds $\underline{c}_{n,k}(\alpha)$ in Theorem 1 is that they imply

$$\frac{k\,n}{k+1} = \lim_{\alpha\to\infty} \alpha\,\underline{c}_{n,k}^2(\alpha) \le \lim_{\alpha\to\infty} \alpha\,c_n^2(\alpha)\,, \qquad 3\le k\le 6$$

(the lower bound in Corollary 2 follows from the case $k=6$). This observation and Proposition 3 give rise for the following

Conjecture 1. The best Markov constant $c_n(\alpha)$ satisfies:

$$\lim_{\alpha\to\infty} \alpha\,c_n^2(\alpha) = n\,.$$

We also performed a search for lower bounds for $c_n^2(\alpha)$ with a factor depending on n of the form $(n+1)\big(n+\sigma(\alpha+1)\big)$. Such a choice is reasonable, as the resulting lower bounds preserve the limit relation in Corollary D (i). The optimal value then is $\sigma=-1/3$ (the same for all k, $3\le k\le 6$), and we obtain lower bounds as in Theorem 1 with $n\big(n+\sigma(\alpha+1)\big)$ replaced by $(n+1)\big(n-(\alpha+1)/3\big)$. These lower bounds make sense only for $n>(\alpha+1)/3$, and are better than those in Theorem 1 only for α close to -1.

2. The bounds $\big(\underline{c}_{n,k}(\alpha),\overline{c}_{n,k}(\alpha)\big)$ $(3\le k\le 6)$ in Theorem 1 imply bounds $(\ell_k(\alpha),u_k(\alpha))$ (occurring in the middle columns of Tables 1 and 2) for the asymptotic Markov constant $c(\alpha)$, and the bounds deduced with a larger k are superior. While the lower bounds $\ell_k(\alpha)$ are of the correct order $\mathcal{O}(\alpha^{-1})$ as $\alpha\to\infty$, for the upper bound $u_k(\alpha)$ we have $u_k(\alpha)=\mathcal{O}(\alpha^{-1+\frac{1}{2k}})$ as $\alpha\to\infty$, $(3\le k\le 6)$. The ratio

$$\rho_k(\alpha) := \frac{u_k(\alpha)}{\ell_k(\alpha)}\,, \qquad 3\le k\le 6,$$

tends to 1 as $\alpha\to -1$, which indicates that for moderate α the bounds $\ell_k(\alpha)$ and $u_k(\alpha)$ are rather tight. This observation is clearly seen in the particular case $\alpha=0$, where, according to Turán's result, we have $c(0)=\frac{2}{\pi}$. We give the lower and the upper bounds for $c(0)$ and the overestimation factors in Table 3.

3. Another interesting observation, concerning the coefficients of R_n inspires the following

Conjecture 2. For every fixed $k\in\mathbb{N}$, the coefficient $b_{k,n}$, $n>k$, of the polynomial $R_n(x)=x^n-b_{1,n}\,x^{n-1}+b_{2,n}\,x^{n-2}-\cdots+(-1)^n\,b_{n,n}$, satisfies

$$b_{k,n} = \frac{n^{2k}}{2^k\,k!(\alpha+1)\cdots(\alpha+2k-1)} + \mathcal{O}(n^{2k-1})\,. \tag{4.1}$$

Conjecture 2 is verified with our computer algebra approach for $1\le k\le 6$, but so far we do not have a proof for the general case. Having (4.1) proved,

we could try to find the explicit form of d_k, the coefficient of n^{2k} in Newton's function $p_k(R_n)$, and consequently to obtain two sequences $\{\ell_k\}$ and $\{u_k\}$ defined by $\ell_k = \sqrt{d_k/d_{k-1}}$ and $u_k = \sqrt[2k]{d_k}$ which converge monotonically from below and from above, respectively, to $c(\alpha)$, the sharp asymptotic Markov constant.

Table 3: The lower and the upper bounds for the asymptotic Markov constant $c(0)$ and the overestimation factors.

k	$\ell_k(0)$	$u_k(0)$	$\frac{c(0)}{\ell_k(0)}$	$\frac{u_k(0)}{c(0)}$
3	$\sqrt{\frac{2}{5}} \approx 0.63245553$	$\sqrt[6]{\frac{1}{15}} \approx 0.63677321$	1.006584242	1.00024103
4	$\sqrt{\frac{17}{42}} \approx 0.63620901$	$\sqrt[8]{\frac{17}{630}} \approx 0.63663212$	1.00064564	1.00001939
5	$\sqrt{\frac{62}{153}} \approx 0.63657580$	$\sqrt[10]{\frac{31}{2835}} \approx 0.63662085$	1.00006906	1.00000170
6	$\sqrt{\frac{2073}{5115}} \approx 0.63661494$	$\sqrt[12]{\frac{2073}{467775}} \approx 0.63661987$	1.00000757	1.00000015

Although the ratios ρ_k, $3 \le k \le 6$, satisfy $\rho_k(\alpha) \to \infty$ as $\alpha \to \infty$, they grow rather slowly. For instance, $\rho_6(\alpha) < 2$ for $\alpha < 140000$, see Figure 1.

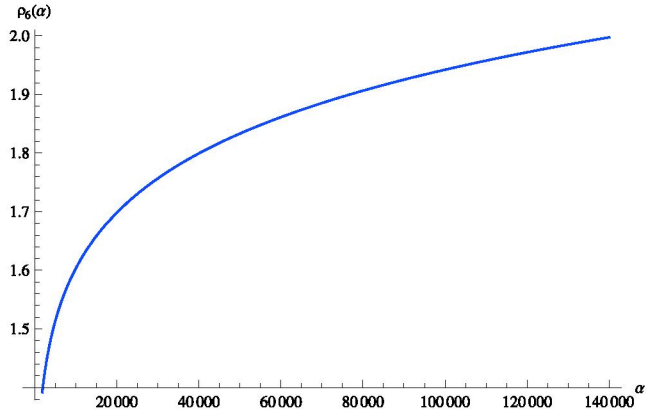


Figure 1: The graph of $\rho_6(\alpha) < 2$.

ACKNOWLEDGEMENT. The authors are partially supported by the Bulgarian National Research Fund under Contract DN 02/14 and by the Sofia University Research Fund under Contract 80.10-11/2017.

5. REFERENCES

[1] Aleksov, D., Nikolov, G.: Markov L_2 inequality with the Gegenbauer weight. *J. Approx. Theory*, **225**, 2018, 224–241, https://doi.org//10.1016/j.jat.2017.10.008.

[2] Dörfler, P.: Über die bestmögliche Konstante in Markov-Ungleichungen mit Laguerre Gewicht. *Österreich. Akad. Wiss. Math.-Natur. Kl. Sitzungsber. II*, **200**, 1991, 13–20.

[3] Dörfler, P.: Asymptotics of the best constant in a certain Markov-type inequality. *J. Approx. Theory*, **114**, 2002, 84–97.

[4] Mead, D. G.: Newton Identities. *Amer. Math. Monthly* **99**, 1992, 749–751.

[5] Nikolov, G., Shadrin, A.: On the L_2 Markov inequality with Laguerre weight. In: *Progress in Approximation Theory and Applicable Complex Analysis*, (N. K. Govil et al., eds.), Springer Optimization and Its Applications **117**, 2017, pp. 1–17. DOI: 10.1007/978-3-319-49242-1_1.

[6] Nikolov, G., Shadrin, A.: On the Markov inequality in the L_2 norm with the Gegenbauer weight. *Constr. Approx.*, 2018, https://doi.org//10.1007/s00365-017-9406-2.

[7] Nikolov, G., Shadrin, A.: Markov L_2–inequality with the Laguerre weight. In: *Constructive Theory of Functions, Sozopol 2016*, (K. Ivanov et al., eds.), Prof. Marin Drinov Publishing House, Sofia, 2017, pp. 197–211. Also availabe as: arXiv:1705.03824v1 [math.CA]

[8] Szegő, G.: *Orthogonal Polynomials*, AMS Colloq. Publ. **23**, AMS, Providence, RI, 1975.

[9] Turán, P.: Remark on a theorem of Ehrhard Schmidt. *Mathematica (Cluj)*, **2**, 1960, 373–378.

[10] Van der Waerden, B. L.: *Modern Algebra*, Vol. 1, New York, Frederick Ungar Publishing Co., 1949.

Received on May 17, 2017

Geno Nikolov, Rumen Uluchev

Department of Mathematics and Informatics
University of Sofia
5 James Bourchier Blvd.
1164 Sofia
BULGARIA

E-mails: geno@fmi.uni-sofia.bg
rumenu@fmi.uni-sofia.bg

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 104

ANNUAL OF SOFIA UNIVERSITY „ST. KLIMENT OHRIDSKI“
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 104

WEIGHTED APPROXIMATION IN UNIFORM NORM BY MEYER-KÖNIG AND ZELLER OPERATORS

IVAN GADJEV, PARVAN E. PARVANOV

The weighted approximation errors of Meyer-König and Zeller operator is characterized for weights of the form $w(x) = x^{\gamma_0}(1-x)^{\gamma_1}$, where $\gamma_0 \in [-1, 0], \gamma_1 \in \mathbb{R}$. Direct inequalities and strong converse inequalities of type A are proved in terms of the weighted K-functional.

Keywords: Meyer-König and Zeller operator, K-functional, direct theorem, strong converse theorem, weighted approximation.

2000 Math. Subject Classification: 41A36, 41A25, 41A27, 41A17.

1. INTRODUCTION AND STATEMENT OF THE RESULTS

The classical Meyer-Konig and Zeller (MKZ) operator is defined for functions $f \in C[0, 1)$ by the formula

$$M_n(f, x) = \sum_{k=0}^{\infty} f\left(\frac{k}{n+k}\right) m_{n,k}(x), \tag{1.1}$$

where

$$m_{n,k}(x) = \binom{n+k}{k} x^k (1-x)^{n+1}.$$

Right after their appearance, the MKZ operators became a subject of serious investigations. A reason for this is that they allow approximation of functions

unbounded at the point 1 (which is not the case with Bernstein polynomials). However, the fact that the function values are taken at the points $\frac{k}{n+k}$ creates some additional difficulties when working with these operators.

In this paper we investigate the weighted approximation of functions by the classical variant of MKZ operator in uniform norm $\|.\|_{[0,1)}$, i.e. we want to characterize the weighted error of approximation $\sup_{x\in[0,1)} |w(x)f(x)|$, where

$$w(x) = x^{\gamma_0}(1-x)^{\gamma_1} \tag{1.2}$$

are the Jacobi weights.

In the unweighted case $w(x) = 1$ a direct theorem was proved in [4], and a strong converse inequality of type A (in the terminology of [3]) was proved in [5]. Regarding the weighted case, the first results were obtained by Becker and Nessel in [2], where they proved direct theorems for some symmetrical weights $w(x) = \varphi^{\alpha}(x)$. Here, $\varphi(x) = x(1-x)^2$ is the weight function naturally connected with the second derivative of MKZ operators.

In [10] Totik established that for $0 < \alpha \leq 1$ and $\varphi(x) = x(1-x)^2$ the condition

$$\varphi^{\alpha}|\Delta_h^2(f,x)| \leq Kh^{2\alpha}$$

is equivalent to

$$M_n f - f = \mathcal{O}\left(n^{-\alpha}\right).$$

In [9] the authors proved that for $0 \leq \lambda \leq 1$ and $0 < \alpha < 2$ the condition

$$|M_n f(x) - f(x)| = \mathcal{O}\left(\left(\frac{\varphi^{(1-\lambda)/2}(x)}{\sqrt{n}}\right)^{\alpha}\right)$$

is equivalent to

$$\omega^2_{\varphi^{\lambda/2}}(f,t) = \mathcal{O}(t^{\alpha}).$$

Here $\omega^2_{\varphi^{\lambda/2}}(f,t)$ are the modulus of Ditzian-Totik of second order

$$\omega^2_{\varphi^{\lambda/2}}(f,t) = \sup_{0<h\leq t} \sup_{x\pm h\varphi^{\lambda/2}(x)\in[0,1)} |\Delta^2_{h\varphi^{\lambda/2}(x)} f(x)|.$$

In [7] Holhoş proved the next direct inequality for weights $\gamma_0 = 0, \gamma_1 > 0$:

$$\|w(M_n f - f)\|_{[0,1)} \leq 2\omega\left(f(1-e^{-t})e^{-\gamma_1 t}, \frac{1}{\sqrt{n}}\right) + \frac{\gamma_1 C(\gamma_1)}{\sqrt{n}}\|wf\|_{[0,1)}.$$

In this paper we prove better results than the results mentioned above. But before stating our main result, let us introduce some notation and definitions. The first derivative operator is denoted by $D = \frac{d}{dx}$. Thus, $Dg(x) = g'(x)$ and $D^2 g(x) = g''(x)$. By $C[0,1)$ we denote the space of functions continuous on $[0,1)$. The functions from $C[0,1)$ are not expected to be continuous or bounded at 1. By

$L_\infty[0,1)$ we denote the space of Lebesgue measurable and essentially bounded in $[0,1)$ functions equipped with the uniform norm $\|\cdot\|_{[0,1)}$. For a weight function w we set

$$C(w)[0,1) = \{g \in C[0,1); \quad wg \in L_\infty[0,1)\},$$
$$W^2(w\varphi)[0,1) = \left\{g, Dg \in AC_{loc}(0,1) \ \& \ w\varphi D^2 g \in L_\infty[0,1)\right\},$$
$$W^3(w\varphi^{3/2})[0,1) = \left\{g, Dg, D^2 g \in AC_{loc}(0,1) \ \& \ w\varphi^{3/2} D^3 g \in L_\infty[0,1)\right\},$$

where $AC_{loc}(0,1)$ is the set of functions which are absolutely continuous in $[a,b]$ for every $[a,b] \subset (0,1)$.

The weighted approximation error $\|w(f - M_n f)\|_{[0,1)}$ will be compared with the K-functional between the weighted spaces $C(w)[0,1)$ and $W^2(w\varphi)[0,1)$, which for every

$$f \in C(w)[0,1) + W^2(w\varphi)[0,1) := \{f_1 + f_2 : f_1 \in C(w)[0,1), f_2 \in W^2(w\varphi)[0,1)\}$$

and $t > 0$ is defined by

$$K_w(f,t)_{[0,1)} = \inf_{g \in W^2(w\varphi),\ f-g \in C(w)} \left\{\|w(f-g)\|_{[0,1)} + t\|w\varphi D^2 g\|_{[0,1)}\right\}. \tag{1.3}$$

Our main result is the following theorem, which establishes a full equivalence between the K-functional $K_w\left(f, \frac{1}{n}\right)_{[0,1)}$ and the weighted error $\|w(M_n f - f)\|_{[0,1)}$.

Theorem 1. *For w defined by (1.2), where $\gamma_0 \in [-1,0], \gamma_1 \in \mathbb{R}$, there exist positive constants C_1, C_2 and L such that for every natural $n \geq L$ and for all*

$$f \in C(w)[0,1) + W^2(w\varphi)[0,1)$$

there holds

$$C_1\|w(M_n f - f)\|_{[0,1)} \leq K_w\left(f, \frac{1}{n}\right)_{[0,1)} \leq C_2\|w(M_n f - f)\|_{[0,1)}. \tag{1.4}$$

The proof is based on a method, used for the first time in [8]. In short, its idea is the following: by making an appropriate transformation, we move to Baskakov operators, for which we have the needed estimations, and then go back by the inverse transformation.

2. A CONNECTION BETWEEN BASKAKOV AND MKZ OPERATORS

Following [8], we introduce a transformation T mapping functions defined on $[0,\infty)$ into functions defined on $[0,1)$. We make the agreement that, from now on, we shall denote variables, functions and operators, defined in $[0,1)$ the usual way, and their analogs, defined in $[0,\infty)$, with tilde.

Now we give some notation and definitions. The uniform norm on the interval $[0,\infty)$ is denoted by $\|\cdot\|_{[0,\infty)}$, and we define the following function spaces:

$$C(\tilde{w})[0,\infty) = \{\tilde{g} \in C[0,\infty); \quad \tilde{w}\tilde{g} \in L_\infty[0,\infty)\},$$
$$W^2(\tilde{w}\tilde{\varphi})[0,\infty) = \left\{\tilde{g}, \tilde{D}\tilde{g} \in AC_{loc}(0,\infty) \ \& \ \tilde{w}\tilde{\varphi}\tilde{D}^2\tilde{g} \in L_\infty[0,\infty)\right\},$$
$$W^3(\tilde{w}\tilde{\varphi}^{3/2})[0,\infty) = \left\{\tilde{g}, \tilde{D}\tilde{g}, \tilde{D}^2\tilde{g} \in AC_{loc}(0,\infty) \ \& \ \tilde{w}\tilde{\varphi}^{3/2}\tilde{D}^3\tilde{g} \in L_\infty[0,\infty)\right\}.$$

The weighted error by Baskakov operators will be characterized by the next K-functional, defined for every function $\tilde{f} \in C(\tilde{w})[0,\infty) + W^2(\tilde{w}\tilde{\varphi})[0,\infty)$ and for every $t > 0$ by the formula

$$K_{\tilde{w}}(\tilde{f}, t)_{[0,\infty)} = \inf\left\{\|\tilde{w}(\tilde{f}-\tilde{g})\|_{[0,\infty)} + t\left\|\tilde{w}\tilde{\varphi}\tilde{D}^2\tilde{g}\right\|_{[0,\infty)}\right\}, \tag{2.1}$$

where the infimum is taken over functions $\tilde{g} \in W^2(\tilde{w}\tilde{\varphi})[0,\infty)$ such that $\tilde{f}-\tilde{g} \in C(\tilde{w})[0,\infty)$.

We start with the change of variable $\sigma : [0,1) \to [0,\infty)$ (used for the first time by V.Totik in [10]) given by

$$\tilde{x} = \sigma(x) = \frac{x}{1-x}. \tag{2.2}$$

Then the inverse change of variable $\sigma^{-1} : [0,\infty) \to [0,1)$ is

$$x = \sigma^{-1}(\tilde{x}) = \frac{\tilde{x}}{1+\tilde{x}}.$$

The transformation operator T, transforming a function $\tilde{f}$ defined on $[0,\infty)$ to a function f defined on $[0,1)$ is defined by

$$f(x) = T(\tilde{f})(x) = \lambda(x)(\tilde{f}\circ\sigma)(x), \quad \lambda(x) = 1-x. \tag{2.3}$$

Then the inverse operator T^{-1}, transforming a function f defined on $[0,1)$ to a function $\tilde{f}$ defined on $[0,\infty)$ is

$$\tilde{f}(\tilde{x}) = T^{-1}(f)(\tilde{x}) = \frac{1}{(\lambda\circ\sigma^{-1})(\tilde{x})}(f\circ\sigma^{-1})(\tilde{x}).$$

We want to estimate the weighted error by MKZ, so we define a new transformation operator S by

$$w(x) = S(\tilde{w})(x) = \frac{1}{\lambda(x)}(\tilde{w}\circ\sigma)(x), \tag{2.4}$$

and its inverse S^{-1} is

$$\tilde{w}(\tilde{x}) = S^{-1}(w)(\tilde{x}) = (\lambda\circ\sigma^{-1})(\tilde{x})(w\circ\sigma^{-1})(\tilde{x}). \tag{2.5}$$

Obviously we have:

$$\begin{aligned} wf &= S(\tilde{w})T(\tilde{f}) = (\tilde{w}\circ\sigma)(\tilde{f}\circ\sigma), \\ \tilde{w}\tilde{f} &= S^{-1}(w)T^{-1}(f) = (w\circ\sigma^{-1})(f\circ\sigma^{-1}). \end{aligned} \tag{2.6}$$

In the next lemmas, w is a weight in $[0,1)$ and $\tilde{w} = S^{-1}(w)$ is the corresponding weight in $[0,\infty)$.

Lemma 1. *The operators T and its inverse T^{-1} are linear positive operators and the next equalities are true:*

$$\begin{aligned} T(\tilde{\varphi}\tilde{D}^2\tilde{f}) &= \varphi D^2(T\tilde{f}), \\ T^{-1}(\varphi D^2 f) &= \tilde{\varphi}\tilde{D}^2(T^{-1}f). \end{aligned} \tag{2.7}$$

Proof. We prove only the first equality, as the proof of the second one is similar. For the right-hand side of the first equality we have

$$\begin{aligned} D(T\tilde{f}) &= D\left(\lambda(\tilde{f}\circ\sigma)\right) = -\tilde{f}\circ\sigma + \lambda D\tilde{f}\circ\sigma \\ &= -\tilde{f}\circ\sigma + \lambda\tilde{D}\tilde{f}\circ\sigma.\lambda^{-2} = -\tilde{f}\circ\sigma + \lambda^{-1}\tilde{D}\tilde{f}\circ\sigma \end{aligned}$$

and

$$\begin{aligned} D^2(T\tilde{f}) &= D\left(-\tilde{f}\circ\sigma + \lambda^{-1}\tilde{D}\tilde{f}\circ\sigma\right) \\ &= -\tilde{D}\tilde{f}\circ\sigma.\lambda^{-2} + D(\lambda^{-1})\tilde{D}\tilde{f}\circ\sigma + \lambda^{-1}D\left(\tilde{D}\tilde{f}\circ\sigma\right) \\ &= -\lambda^{-2}\tilde{D}\tilde{f}\circ\sigma + \lambda^{-2}\tilde{D}\tilde{f}\circ\sigma + \lambda^{-1}\tilde{D}^2\tilde{f}\circ\sigma.\lambda^{-2} = \lambda^{-3}\tilde{D}^2\tilde{f}\circ\sigma. \end{aligned}$$

Consequently,

$$\varphi D^2(T\tilde{f}) = \lambda\frac{\varphi}{\lambda^4}\tilde{D}^2\tilde{f}\circ\sigma = \lambda\tilde{\varphi}\tilde{D}^2\tilde{f}\circ\sigma = T(\tilde{\varphi}\tilde{D}^2\tilde{f}). \qquad \square$$

Lemma 2. *The operator $T : C(\tilde{w})[0,\infty) \to C(w)[0,1)$ is an one-to-one correspondence with*

$$\|wT(\tilde{f})\|_{[0,1)} = \|\tilde{w}\tilde{f}\|_{[0,\infty)}, \quad \|\tilde{w}T^{-1}(f)\|_{[0,\infty)} = \|wf\|_{[0,1)}.$$

Proof. The above equalities are easily obtainable from the definition (2.3) of the operator T and from the equalities (2.6). $\square$

Lemma 3. *The operator $T : W^2(\tilde{w}\tilde{\varphi})[0,\infty) \to W^2(w\varphi)[0,1)$ is an one-to-one correspondence with*

$$\|w\varphi D^2(T(\tilde{f}))\|_{[0,1)} = \|\tilde{w}\tilde{\varphi}\tilde{D}^2\tilde{f}\|_{[0,\infty)}, \quad \|\tilde{w}\tilde{\varphi}\tilde{D}^2(T^{-1}(f))\|_{[0,\infty)} = \|w\varphi D^2 f\|_{[0,1)}.$$

Proof. From the definition (2.3) of tne operator T and from the equalities (2.6) and (2.7) we have

$$\begin{aligned}\tilde{w}\tilde{\varphi}\tilde{D}^2\tilde{f} &= \tilde{w}T^{-1}\left(\varphi D^2(T\tilde{f})\right) = \tilde{w}\frac{1}{\lambda\circ\sigma^{-1}}\left(\varphi D^2(T\tilde{f})\right)\circ\sigma^{-1}\\ &= \left(\lambda\circ\sigma^{-1}\right)\left(w\circ\sigma^{-1}\right)\frac{1}{\lambda\circ\sigma^{-1}}\left(\varphi D^2(T\tilde{f})\right)\circ\sigma^{-1}\\ &= \left(w\circ\sigma^{-1}\right)\left(\varphi D^2(T\tilde{f})\right)\circ\sigma^{-1} = \left(w\varphi D^2(T(\tilde{f}))\right)\circ\sigma^{-1}.\end{aligned}$$

Consequently

$$\tilde{w}\tilde{\varphi}\tilde{D}^2\tilde{f}(\tilde{x}) = \left(w\varphi D^2(T(\tilde{f}))\right)\circ\sigma^{-1}(\tilde{x}) = w\varphi D^2(T(\tilde{f}))(x)$$

or

$$\|w\varphi D^2(T(\tilde{f}))\|_{[0,1)} = \|\tilde{w}\tilde{\varphi}\tilde{D}^2\tilde{f}\|_{[0,\infty)}.$$

The proof of the second equality is similar, and therefore is omitted. □

Lemma 4. *For every $f \in C(w)[0,1) + W^2(w\varphi)[0,1)$, $\tilde{f} = T^{-1}f$ and $t > 0$ we have*

$$K_w(f,t)_{[0,1)} = K_{\tilde{w}}(\tilde{f},t)_{[0,\infty)}.$$

Proof. From the definition of the K-functional (2.1) we have

$$K_{\tilde{w}}(\tilde{f},t)_{[0,\infty)} = \inf_{\tilde{g}\in W^2(\tilde{w}\tilde{\varphi}),\ \tilde{f}-\tilde{g}\in C(\tilde{w})}\left\{\|\tilde{w}(\tilde{f}-\tilde{g})\|_{[0,\infty)} + t\|\tilde{w}\tilde{\varphi}\tilde{D}^2\tilde{g}\|_{[0,\infty)}\right\}.$$

Now, from (2.6)

$$\tilde{w}(\tilde{f}-\tilde{g}) = \left(w\circ\sigma^{-1}\right)\left((f-g)\circ\sigma^{-1}\right)$$

and consequently

$$\|\tilde{w}(\tilde{f}-\tilde{g})\|_{[0,\infty)} = \|w(f-g)\|_{[0,1)}.$$

From Lemma 4 we have

$$\|\tilde{w}\tilde{\varphi}\tilde{D}^2\tilde{g}\|_{[0,\infty)} = \|w\varphi D^2(T(\tilde{g}))\|_{[0,1)} = \|w\varphi D^2 g\|_{[0,1)}.$$ □

The classical Baskakov operator $V_nf(x)$ (see [1]) is defined for bounded functions $f(x)$ in $[0,\infty)$ by the formula

$$V_nf(x) = (V_nf,x) = V_n(f,x) = \sum_{k=0}^{\infty} f\left(\frac{k}{n}\right)v_{n,k}(x), \tag{2.8}$$

where

$$v_{n,k}(x) = \binom{n+k-1}{k}x^k(1+x)^{-n-k}.$$

The next two lemmas give the connection between the MKZ operators M_n and the Baskakov operators V_n.

Lemma 5. *For every f such that one of the series below is convergent and for every $n \in \mathbb{N}$ we have*

$$M_n(f)(x) = T(V_n(T^{-1}(f)))(x), \quad x \in [0,1). \tag{2.9}$$

Proof. From the definition of T we get

$$\begin{aligned} T\left(V_n\left(T^{-1}(f)\right)\right)(x) &= \lambda(x)(V_n(T^{-1}(f)) \circ \sigma^{-1})(x) \\ &= \frac{1}{1+\tilde{x}}(V_n(T^{-1}(f)(\tilde{x}) = \frac{1}{1+\tilde{x}} V_n(\tilde{f}, \tilde{x}) \\ &= \frac{1}{1+\tilde{x}} \sum_{k=0}^{\infty} \binom{n+k-1}{k} \frac{\tilde{x}^k}{(1+\tilde{x})^{n+k}} \tilde{f}\left(\frac{k}{n}\right) \\ &= \sum_{k=0}^{\infty} \binom{n+k-1}{k} \frac{\tilde{x}^k}{(1+\tilde{x})^{n+k+1}} \frac{1}{(\lambda \circ \sigma^{-1})\left(\frac{k}{n}\right)} (f \circ \sigma^{-1})\left(\frac{k}{n}\right). \end{aligned}$$

Since

$$\sigma^{-1}\left(\frac{k}{n}\right) = \frac{k/n}{1+k/n} = \frac{k}{n+k},$$

we have

$$(\lambda \circ \sigma^{-1})\left(\frac{k}{n}\right) = \lambda\left(\frac{k}{n+k}\right) = \frac{n}{n+k}$$

and

$$(f \circ \sigma^{-1})\left(\frac{k}{n}\right) = f\left(\frac{k}{n+k}\right).$$

Also,

$$\frac{\tilde{x}^k}{(1+\tilde{x})^{n+k+1}} = \left(\frac{\tilde{x}}{1+\tilde{x}}\right)^k \frac{1}{(1+\tilde{x})^{n+1}} = x^k(1-x)^{n+1}.$$

Consequently,

$$\begin{aligned} T(V_n(T^{-1}(f)))(x) &= \sum_{k=0}^{\infty} \binom{n+k-1}{k} \frac{n+k}{k} x^k(1-x)^{n+1} f\left(\frac{k}{n+k}\right) \\ &= \sum_{k=0}^{\infty} \binom{n+k}{k} x^k(1-x)^{n+1} f\left(\frac{k}{n+k}\right) = M_n(f,x). \end{aligned}$$

□

Lemma 6. *For every $f \in C(w)[0,1)$ and for every $n \in \mathbb{N}$ we have*

$$\left\| w(M_n f - f) \right\|_{[0,1)} = \left\| \tilde{w}(V_n \tilde{f} - \tilde{f}) \right\|_{[0,\infty)}.$$

Proof. From Lemma 5 we have

$$M_n(f)(x) = T(V_n(T^{-1}(f)))(x) = \lambda(x)(V_n(T^{-1}(f)) \circ \sigma^{-1})(x)$$
$$= \frac{1}{1+\tilde{x}}(V_n(T^{-1}(f)(\tilde{x}) = \frac{1}{1+\tilde{x}} V_n(\tilde{f}, \tilde{x}).$$

Since

$$f(x) = T(\tilde{f})(x) = \lambda(x)(\tilde{f} \circ \sigma)(x) = \frac{1}{1+\tilde{x}}\tilde{f}(\tilde{x}),$$

it follows that

$$M_n(f)(x) - f(x) = \frac{1}{1+\tilde{x}}\left(V_n(\tilde{f}, \tilde{x}) - \tilde{f}(\tilde{x})\right).$$

Also, from (2.4) we have

$$w(x) = S(\tilde{w})(x) = \frac{1}{\lambda(x)}(\tilde{w} \circ \sigma)(x) = (1+\tilde{x})\tilde{w}(\tilde{x}).$$

Consequently

$$w(x)\big(M_n f - f\big)(x) = (1+\tilde{x})\tilde{w}(\tilde{x})\frac{1}{1+\tilde{x}}\left(V_n(\tilde{f}, \tilde{x}) - \tilde{f}(\tilde{x})\right)$$
$$= \tilde{w}(\tilde{x})\left(V_n\tilde{f} - \tilde{f}\right)(\tilde{x})$$

i.e.

$$\left\|w(M_n f - f)\right\|_{[0,1)} = \left\|\tilde{w}(V_n\tilde{f} - \tilde{f})\right\|_{[0,\infty)}. \qquad \square$$

3. PROOF OF THEOREM 1 AND SOME OTHER RESULTS FOR MKZ

From Lemma 5 we have

$$M_n(f)(x) = T(V_n(T^{-1}(f)))(x) = \lambda(x)(V_n(T^{-1}(f)) \circ \sigma^{-1})(x)$$
$$= \frac{1}{1+\tilde{x}}(V_n(T^{-1}(f)(\tilde{x}) = \frac{1}{1+\tilde{x}} V_n(\tilde{f}, \tilde{x}).$$

Since

$$f(x) = T(\tilde{f})(x) = \lambda(x)(\tilde{f} \circ \sigma)(x) = \frac{1}{1+\tilde{x}}\tilde{f}(\tilde{x}),$$

it follows that

$$M_n(f)(x) - f(x) = \frac{1}{1+\tilde{x}}\left(V_n(\tilde{f}, \tilde{x}) - \tilde{f}(\tilde{x})\right).$$

Also, from (2.4) we have

$$w(x) = S(\tilde{w})(x) = \frac{1}{\lambda(x)}(\tilde{w} \circ \sigma)(x) = (1+\tilde{x})\tilde{w}(\tilde{x}).$$

Consequently,

$$\begin{aligned} w(x)\big(M_n f - f\big)(x) &= (1+\tilde{x})\tilde{w}(\tilde{x})\frac{1}{1+\tilde{x}}\left(V_n(\tilde{f},\tilde{x}) - \tilde{f}(\tilde{x})\right) \\ &= \tilde{w}(\tilde{x})\left(V_n\tilde{f} - \tilde{f}\right)(\tilde{x}) \end{aligned}$$

i.e.,

$$\left\|w(M_n f - f)\right\|_{[0,1)} = \left\|\tilde{w}(V_n\tilde{f} - \tilde{f})\right\|_{[0,\infty)}.$$

From [6, Theorem 1] we have that for weights $\tilde{w}(\tilde{x}) = \tilde{x}^{\beta_0}(1+\tilde{x})^{\beta_\infty}$, where $\beta_0 \in [-1,0]$, $\beta_\infty \in \mathbb{R}$, the next equivalency is true:

There exists an absolute constant L such that, for every natural number $n > L$,

$$\left\|\tilde{w}(V_n\tilde{f} - \tilde{f})\right\|_{[0,\infty)} \sim K_{\tilde{w}}\left(\tilde{f}, \frac{1}{n}\right)_{[0,\infty)}.$$

From Lemma 4 we have

$$K_w(f,t)_{[0,1)} = K_{\tilde{w}}(\tilde{f},t)_{[0,\infty)},$$

and consequently

$$\left\|w(M_n f - f)\right\|_{[0,1)} \sim K_w\left(f, \frac{1}{n}\right)_{[0,1)}.$$

For the weights $\tilde{w}(\tilde{x}) = \tilde{x}^{\beta_0}(1+\tilde{x})^{\beta_\infty}$ we have

$$\begin{aligned} w(x) &= \frac{1}{\lambda(x)}(\tilde{w} \circ \sigma)(x) = (1+\tilde{x})\tilde{w}(\tilde{x}) = \tilde{x}^{\gamma_0}(1+\tilde{x})^{\gamma_\infty+1} \\ &= x^{\gamma_0}(1-x)^{-(\gamma_\infty+\gamma_0+1)} = x^{\gamma_0}(1-x)^{\gamma_1}. \end{aligned}$$

Since $\beta_0 \in [-1,0]$, $\beta_\infty \in \mathbb{R}$, we have $\gamma_0 \in [-1,0], \gamma_1 \in \mathbb{R}$.

The proof of Theorem 1 is complete. □

From Lemma 6, Lemma 3 and Lemma 5 in [6] we obtain the following Jackson-type inequality.

Theorem 2. *For w, defined by (1.2) there exists a constant C such that for every natural $n \geq |1+\gamma_0+\gamma_1|$ we have*

$$\left\|w(M_n f - f)\right\|_{[0,1)} \leq \frac{C}{n}\left\|w\varphi D^2 f\right\|_{[0,1)}$$

for every function $f \in W^2(w\varphi)[0,1)$.

From the definition of T, Lemma 3, Lemma 5 and Lemma 7 in [6] we obtain the following Bernstein-type inequality.

Theorem 3. *For w, defined by* (1.2) *there exists a constant C such that for every natural $n \geq |1 + \gamma_0 + \gamma_1|$ we have*

$$\left\|w\varphi D^2 M_n f\right\|_{[0,1)} \leq Cn\|wf\|_{[0,1)}$$

for every function $f \in C(w)[0,1)$.

ACKNOWLEDGEMENT. The first–named author is partially supported by the Bulgarian National Research Fund under Contract DN 02/14.

4. REFERENCES

[1] Baskakov, V. A.: An instance of a sequence of the linear positive operators in the space of continuous functions. *Dokl. Akad. Nauk SSSR*, **113**, 1957, 249–251.

[2] Becker, M., Nessel, R. J.: A global approximation theorem for Meyer-König and Zeller operators. *Math. Z.*, **160**, 1978, 195–206.

[3] Ditzian, Z., Ivanov, K. G.: Strong converse inequalities. *J. Anal. Math.*, **61**, 1993, 61–111.

[4] Ditzian, Z., Totik, V.: *Moduli of Smoothness.* Springer, Berlin, New York, 1987.

[5] Gadjev, I.: Strong converse result for uniform approximation by Meyer-König and Zeller operator. *J. Math. Anal. Appl.*, **428**, 2015, 32–42.

[6] Gadjev, I.: Weighted approximation by Baskakov operators. *J. Math. Inequal. Appl.*, **18**, no. 4, 2015, 1443–1461.

[7] Holhoş, A.: Uniform approximation of functions by Meyer-König and Zeller operators. *J. Math. Anal. Appl.*, **393**, 2012, 33–37.

[8] Ivanov, K. G., Parvanov, P. E.: Weighted approximation by Meyer-König and Zeller-Type operators. In: *Constructive Theory of Functions, Sozopol 2010, in memory of Borislav Bojanov* (G. Nikolov and R. Uluchev, eds.), Prof. Marin Drinov Academic Publishing House, Sofia, 2012, pp. 150–160.

[9] Lixia Liu Shunsheng Guo, Zhiming Wang: Pointwise approximation by Meyer-König and Zeller operators. *Numerical Functional Analysis and Optimization*, **29**, no. 7-8, 2008, 770–778.

[10] Totik, V.: Uniform aproximation by Baskakov and Meyer-König and Zeller-type operators. *Period. Math. Hungar*, **14**, no. 3-4, 1983, 209–228.

Received on October 30, 2017

Ivan Gadjev, Parvan E. Parvanov

Faculty of Mathematics and Informatics
"St. Kl. Ohridski" University of Sofia
5, J. Bourchier blvd., BG-1164 Sofia
BULGARIA

E-mails: gadjevivan@hotmail.com
pparvan@fmi.uni-sofia.bg

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 104

ANNUAL OF SOFIA UNIVERSITY „ST. KLIMENT OHRIDSKI“
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 104

ON THE REPRESENTATION OF MODULES OVER FINITE CHAIN RINGS

NEVYANA GEORGIEVA, IVAN LANDJEV

We define a standard form for matrices over finite chain rings and describe some basic operations on modules over such rings. These results are used as a tool for the investigation of network codes over finite chain rings and spreads in projective Hjelmslev geometries.

Keywords: finite chain rings, modules over finite chain rings, standard form of a matrix over a finite chain ring.

2000 Math. Subject Classification: 16D10,16P10 (Primary); 51C05,94B05 (Secondary).

1. INTRODUCTION

The aim of this paper is to define a standard representation for modules over finite chain rings and demonstrate its application to certain basic operations over modules. The problem of defining a standard representation arises in connection with various problems. These include the problem of the construction of spreads of projective Hjelmslev spaces by subspaces of various shapes, the problem of the construction of R-analogues of designs, as well as the construction of network codes over finite chain rings.

The paper is structured as follows. In section 2 we summarize some basic facts about the structure of finite chain rings. We introduce a linear order on the ring elements which is used in the definition of the standard form. In section 3 we

present structure results about modules over finite chain rings as well as a counting formula for the number of submodules of given shape contained in a fixed module. In section 4 we introduce the standard form of a matrix over a finite chain ring. We prove one of our central results that for every module ${}_RM \leq {}_RR^n$ there exists a unique matrix in standard form whose rows generate ${}_RM$. In section 5 we obtain the standard form of the matrix whose rows generate the right orthogonal $M_R^{\perp}$ of a given left module ${}_RM$. In section 6 we discuss how to generate all submodules of a given module spanned by the rows of a matrix in standard form.

2. FINITE CHAIN RINGS

In this section, we give some facts about finite chain rings. An associative ring with identity is called a left (right) chain ring if the lattice of its left (right) ideals is a chain. The general structure of finite chain rings is given in the following theorem.

Theorem 1. *For a finite chain ring R the following conditions are equivalent*

(i) R is a left chain ring;

(ii) the principal left ideals of R form a chain;

(iii) R is a local ring and $\operatorname{Rad} R = R\theta$ *for any* $\theta \in \operatorname{Rad} R/(\operatorname{Rad} R)^2$*;*

(iv) R is a right chain ring.

If R satisfies the above conditions then every proper left(right) ideal of R has the form $(\operatorname{Rad} R)^i = R\theta^i = \theta^i R$ *for some positive integer i.*

It is well-known that the factor-ring $R/\operatorname{Rad} R$ is a field. We denote its cardinality by $q = p^r$. The smallest positive integer m for which $(\operatorname{Rad} R)^m = (0)$ is called the length of R. Furthermore, for each $i = 0, \ldots, m-1$, $(\operatorname{Rad} R)^i/(\operatorname{Rad} R)^{i+1}$ is a vector space of dimension 1 over $R/\operatorname{Rad} R$, and we have $|(\operatorname{Rad} R)^i/(\operatorname{Rad} R)^{i+1}| = q$. This implies that $|R| = q^m$. The characteristic of R is $\operatorname{char} R = p^s$ for some positive integer s.

Let $\Gamma = \{\gamma_0 = 0, \gamma_1 = 1, \gamma_2, \ldots, \gamma_{q-1}\}$ be a set of elements of R with $\gamma_i \not\equiv \gamma_j \pmod{\operatorname{Rad} R}$ for all i, j with $0 \leq i < j \leq q-1$. Let us fix a generator θ of R. Every element a from R can be written in a unique way as

$$a = a_0 + a_1\theta + \cdots + a_{m-1}\theta^{m-1},$$

for some $a_i \in \Gamma$. We fix the following linear order on Γ:

$$\gamma_0 \prec \gamma_1 \prec \cdots \prec \gamma_{m-1}.$$

This order is extended to the elements of R as follows. For the elements $a = a_0 + a_1\theta + \cdots + a_{m-1}\theta^{m-1}$ and $b = b_0 + b_1\theta + \cdots + b_{m-1}\theta^{m-1}$, $a_i, b_i \in \Gamma$, we write $a \prec b$ if and only if

$$a_{m-1} = b_{m-1}, \ldots, a_{j+1} = b_{j+1}, a_j \prec b_j,$$

for some $0 \leq j \leq m-1$. We can define a bijection $\varphi : R \to \{0, 1, \ldots, q^m - 1\}$ which is consistent with the linear order of the elements of R given above. Set $\varphi(\gamma_i) = i$. Further for $a = a_0 + a_1\theta + \cdots + a_{m-1}\theta^{m-1}$, $a_i \in \Gamma$, we let $\varphi(a) = \sum_{i=0}^{m-1} \varphi(a_i)q^i$. The following lemma contains some straightforward properties of φ.

Lemma 1. *(1) $a \in \Gamma$ if and only if $\varphi(a) < q$; more generally, the elements a with $\varphi(a) < q^i$ form a system of distinct representatives modulo $(\operatorname{Rad} R)^i$;*

(2) for each $i \in \mathbb{N}$, $a \in (\operatorname{Rad} R)^i$, i.e. $a = b\theta^i$, $b \in R^*$, if and only if $\varphi(a)$ divides q^i;

(3) if q^i divides $\varphi(b)$ then $b = a\theta^i$ with $a = \varphi^{-1}\left(\frac{\varphi(b)}{q^i}\right)$.

Throughout the paper, the letters $\theta, \Gamma, p, q, m, r, s$ will have the meaning fixed above. For a more detailed study of finite chain rings we refer to [2,3,4,5,7,8].

3. MODULES OVER FINITE CHAIN RINGS

Let ${}_RM$ be a finitely generated left R-module. We say that the element $x \in {}_RM$ has period θ^i if $i \geq 0$ is the smallest integer with $\theta^i x = 0$. The element $x \in {}_RM$ is said to have height j if j is the largest integer with $x = \theta^j y$ for some $y \in M$. We set

$$M^* = \{x \in M \mid x \text{ has period } \theta^m\}.$$

An integer partition of the positive integer N is a sequence $\lambda = (\lambda_1, \lambda_2, \ldots)$ with $\lambda_i \in \mathbb{Z}$, $\lambda_1 \geq \lambda_2 \geq \cdots$, $\lambda_i = 0$ for all but finitely many i, and $N = \lambda_1 + \lambda_2 + \cdots$. We write this as $\lambda \vdash N$. Sometimes it is convenient to suppress the trailing zeros in the sequence λ. Partitions can be written multiplicatively as $\lambda = 1^{s_1}2^{s_2}3^{s_3}\ldots$ where s_i is the number of λ_j's equal to i. Denote by λ'_j the number of parts greater or equal to j. Then $\lambda' = (\lambda'_1, \lambda'_2, \ldots)$ is again a partition of N and it is called the conjugate partition of λ.

The following theorem describes the structure of finite R-modules.

Theorem 2. *Let R be a finite chain ring. For every finite module ${}_RM$ there exists a uniquely determined partition $\lambda = (\lambda_1, \ldots, \lambda_k) \vdash \log_R |M|$ into parts $1 \leq \lambda_i \leq m$ such that*

$${}_RM \cong R/(\operatorname{Rad} R)^{\lambda_1} \oplus \cdots \oplus R/(\operatorname{Rad} R)^{\lambda_k}.$$

The parts of the conjugate partition $\lambda' = (\lambda'_1, \lambda'_2, \ldots) \vdash \log_q |M|$ are the Ulm-Kaplansky invariants $\lambda'_i = \dim_{R/\operatorname{Rad} R}(M[\theta] \cap \theta^{i-1}M)$.

The partitions λ and λ' are called the shape, resp. conjugate shape, of ${}_RM$. The integer $k = \lambda'_1 = \dim_{R/\operatorname{Rad} R} M[\theta]$ is called the rank of ${}_RM$ and the integer λ'_m is called the free rank of ${}_RM$.

Denote by $\boldsymbol{M}_{m,n}(R)$ the set of all m-by-n matrices over the chain ring R.

Theorem 3 ([3]). *Let $A \in \boldsymbol{M}_{m,n}(R)$ be a matrix over R. Then the left module ${}_RL \leq {}_RR^n$ generated by the rows of A and the right module $M_R \leq {}_RR^n$ generated by the columns of A have the same shape.*

It is known that an n-dimensional vector space over the finite field $\mathbb{F}_q$ has exactly

$$\begin{bmatrix} n \\ k \end{bmatrix}_q = \frac{(q^n-1)(q^{n-1}-1)\dots(q^{n-k+1}-1)}{(q^k-1)(q^{k-1}-1)\dots(q-1)}$$

k-dimensional vector subspaces.

A similar counting formula holds true for modules over finite chain rings. Let ${}_RM$ be a module of shape λ and $U \leq {}_RM$ be a submodule of shape μ. The conjugate partitions λ', μ' are related by $\mu' \leq \lambda'$ which is equivalent to $\mu \leq \lambda$. The next theorem is our main counting tool. For the special case of $R = \mathbb{Z}_m$ it is known from [1]. For the case of general R we refer to [7].

Theorem 4. *Let ${}_RM$ be a module of shape λ. For every partition μ satisfying $\mu \leq \lambda$ the module ${}_RM$ has exactly*

$$\begin{bmatrix} \lambda \\ \mu \end{bmatrix}_q := \prod_{i=1}^{\infty} q^{\mu'_{i+1}(\lambda'_i - \mu'_i)} \cdot \begin{bmatrix} \lambda'_i - \mu'_{i+1} \\ \mu'_i - \mu'_{i+1} \end{bmatrix}_q \tag{3.1}$$

submodules of shape μ. In particular, the number of free rank s submodules of ${}_RM$ equals

$$q^{s(\lambda'_1 - s) + \dots + s(\lambda'_{m-1} - s)} \cdot \begin{bmatrix} \lambda'_m \\ s \end{bmatrix}_q.$$

Corollary 1. *Let $\boldsymbol{m} = (\underbrace{m, \dots, m}_{n})$ and let $\mu = (\mu_1, \dots, \mu_n)$, where $m \geq \mu_1 \geq \dots \geq \mu_n \geq 0$. Then*

$$\begin{bmatrix} \boldsymbol{m} \\ \mu \end{bmatrix}_q = \begin{bmatrix} \boldsymbol{m} \\ \overline{\mu} \end{bmatrix}_q,$$

where $\overline{\mu} = (m - \mu_n, \dots, m - \mu_1)$.

Remark 1. The formula in Corollary 1 can be viewed as analogue of the usual binomial identity $\binom{n}{k} = \binom{n}{n-k}$.

4. THE STANDARD FORM OF A MATRIX OVER A FINITE CHAIN RING

Let R be a finite chain ring of cardinality q^m and with residue field $R/\operatorname{Rad} R \cong \mathbb{F}_q$, $q = p^r$, where p is a prime. Given a finite set of generators of a submodule M of ${}_RR^n$, we consider the problem of finding a standard generating set for M, which can be easily operated on, i.e., from the standard form we expect to be able to find easily the dual module, the span of two modules, as well as their intersection.

We denote by $\boldsymbol{M}_{k,n}$ the set of all k-by-n matrices over R.

Definition 1. We say that the matrix $A = (a_{ij}) \in \boldsymbol{M}_{k,n}$ is in standard form if

(1) $a_{ij_i} = \theta^{m-t_i}$ for some $t_i \in \{0, \ldots, m\}$;

(2) $a_{is} = \theta^{m-t_i+1}\beta$, $\beta \in R$, for all $s < j_i$;

(3) $a_{is} = \theta^{m-t_i}\beta$, $\beta \in R$, for all $s > j_i$;

(4) $a_{sj_i} \prec a_{ij_i}$ for all $s \neq i$ (here $\prec$ is the lexicographic order defined in section);

(5) $i_1 < i_2 < i_3 < \cdots$.

The integer t_i is called the type of row i, $i = 1, \ldots, k$. Let $\boldsymbol{a} = (a_1, \ldots, a_n) \in {}_RR^n$. The smallest $i \in \{0, \ldots, m\}$ such that $\theta^i a = 0$ is called the type of $\boldsymbol{a}$. The leftmost component a_j with $a_j \in (\operatorname{Rad} R)^{m-i} \setminus (\operatorname{Rad} R)^{m-i+1}$ is called the leader of $\boldsymbol{a}$. For a matrix $A \in \boldsymbol{M}_{k,n}(R)$ in standard form we denote the set of coordinate positions of the row-leaders of A by $J(A) = \{j_1, j_2, \ldots, j_k\}$.

Lemma 2. *Let ${}_RM \leq {}_RR^n$ be a module and let A be a matrix in standard form whose rows generate ${}_RM$. For an arbitrary element $\boldsymbol{v} \in {}_RM$ denote by s the position of its leader. Then $s \in J(A)$.*

Proof. Denote the rows of A by $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$. Let further $J(A) = \{j_1, \ldots, j_k\}$, and let the respective leaders be $\theta^{m-t_1}, \ldots, \theta^{m-t_k}$. Without loss of generality we can assume that $t_1 \geq t_2 \geq \cdots \geq t_k \geq 1$. Then all elements in column j_s, $s = 1, \ldots, k$, that are under the leader of row s, are zeros.

Set $\boldsymbol{v} = \lambda_1\boldsymbol{v}_1 + \ldots + \lambda_k\boldsymbol{v}_k$. Let the leader of $\boldsymbol{v}$ be in position l. We consider three cases.

(1) Let $l < j_1$. Assume that $s \in \{1, \ldots, k\}$ is such an index that the type of $\lambda_s\boldsymbol{v}_s$ is the largest among the types of the vectors $\lambda_i\boldsymbol{v}_i$. If $\lambda_s \in (\operatorname{Rad} R)^{\tau_s}$ then the type of $\boldsymbol{v}$ is at most $t_s - 1 - \tau_s$. On the other hand, the element in the j_s-th coordinate of $\boldsymbol{v}$ is λ_s+(terms which are a linear combination of $1, \theta, \ldots, \theta^{m-t_s-1}$). Therefore the type of $\boldsymbol{v}$ is at least $t_s - \tau_s$, a contradiction.

(2) Let $j_{i-1} < l < j_i$. Assume $\lambda_s \neq 0$ for some $s \leq i-1$ and $\lambda - s\boldsymbol{v}_s$ is the largest type of a vector from $\{\lambda_1\boldsymbol{v}_1, \ldots, \lambda_i\boldsymbol{v}_i\}$. If $\lambda_s \in (\operatorname{Rad} R)^{\tau_s}$, this largest type is at most $t_s - \tau_s$. On the other hand, $\boldsymbol{v}$ has in position j_s the element

$\lambda_s\theta^{m-t_s}$+(terms which are a linear combination of $1, \theta, \ldots, \theta^{m-t_s-1}$). The first term is from $R\theta^{m-t_s+\tau_s}$, but is to the left of the leader, a contradiction. We have proved so far that $\lambda_j = 0$ for all $j \leq i-1$. Now we can use the argument from (1) to complete this case.

(3) Now let $l > j_k$. Now we can use the argument from the first part of (2).

By (1–3) l should be a coordinate position which is from $J(A)$. □

Theorem 5. *For every module $M \leq {}_RR^n$ there exists a unique matrix B in standard form such that M is spanned by the rows of B.*

Proof. 1) *Existence.* We prove the existence by induction on $k = \operatorname{rk} M$. There is nothing to prove for $k = 1$. One has only to note that by a suitable multiplication one can make the leader have the form θ^{m-t} for some t.

Let $\boldsymbol{v}'_1 \in M$ be an element of the maximal possible type in M, $m - t_1$ say. Without loss of generality we may assume that the leader is in position j_1 and is the leftmost among all leaders of elements of M. By a suitable multiplication, we can make this leader equal to θ^{m-t_1}. Now ${}_RM = {}_R\boldsymbol{v}'_1 \oplus {}_RM'$, where ${}_RM'$ has rank $k-1$ and is the submodule of ${}_RM$ containing all vectors having 0 in position j_1. This follows by the fact that for every vector $\boldsymbol{v} \in {}_RM$ one can find a $\lambda \in R$ such that $\boldsymbol{v} - \lambda_1\boldsymbol{v}'_1$ has zero in position j_1.

By the induction hypothesis, there exists a matrix $\boldsymbol{B}$ in standard form whose rows generate ${}_RM'$. Denote these rows by $\boldsymbol{v}_2, \ldots, \boldsymbol{v}_k$. Set $J(B) = \{j_2, \ldots, j_k\}$. Further, let the j_s-th component of $\boldsymbol{v}'_1$ be $\alpha_s + \beta_s\theta^{m-t_s}$, $s = 2, \ldots, k$, where $\alpha_s = x_0 + x_1\theta + x_{m-t_s-1}$, $\beta_s = y_0 + \cdots + y_{t_s-1}$, $x_i, y_i \in \Gamma$.

The element $\boldsymbol{v}_1 = \boldsymbol{v}'_1 - \beta_2\boldsymbol{v}_2 - \cdots - \beta_k\boldsymbol{v}_k$ has the property that the element in position j_s is smaller (with respect to $\prec$) than θ^{m-t_s} for all $s = 2, \ldots, k$. It is also clear that the components of $\boldsymbol{v}_1$ to the left of the leader belong to $\operatorname{Rad} R^{m-t+1}$ (since the i-th, $i < j_1$, component in each one of $\boldsymbol{v}'_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k$ is in $\operatorname{Rad} R^{m-t+1}$). Hence the matrix A having as rows the vectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k$ is the desired matrix.

2) *Uniqueness.* Assume $A' = (\boldsymbol{v}'_1, \ldots, \boldsymbol{v}'_k)^T$ and $A'' = (\boldsymbol{v}''_1, \ldots, \boldsymbol{v}''_k)^T$ are two matrices in standard form whose rows generate the same module ${}_RM$. By Lemma 2 $J(A') = J(A'') = \{j_1, \ldots, j_k\}$. Let the leaders of $\boldsymbol{v}'_i$ (resp. $\boldsymbol{v}''_i$) be $\theta^{m-t'_i}$ (resp. $\theta^{m-t''_i}$). With no loss of generality $t'_1 \geq t'_2 \geq \cdots \geq t'_k$. In particular, this means that all elements in A' in the columns $j_1, j_2, \ldots, j_k$ below the leader of the corresponding row are zeros, i.e. we have.

$$
\begin{array}{lccccccccc}
\boldsymbol{v}'_1 = & (\ldots & \theta^{m-t'_1} & v'_{1,j_2} & \ldots & v'_{1,j_i} & \ldots & v'_{1j_k} & \ldots) \\
\boldsymbol{v}'_2 = & (\ldots & 0 & \theta^{m-t'_2} & \ldots & v'_{2,j_i} & \ldots & v'_{2j_k} & \ldots) \\
\vdots & & \vdots & \vdots & & \vdots & & \vdots & \\
\boldsymbol{v}'_i = & (\ldots & 0 & 0 & \ldots & \theta^{m-t'_i} & \ldots & v'_{ij_k} & \ldots) \\
\vdots & & \vdots & \vdots & & \vdots & & \vdots & \\
\boldsymbol{v}'_k = & (\ldots & 0 & 0 & \ldots & 0 & \ldots & \theta^{m-t'_k} & \ldots)
\end{array}
$$

Now we can express $\boldsymbol{v}_i''$ as $\boldsymbol{v}_1'' = \lambda_1 \boldsymbol{v}_1' + \cdots + \lambda_k \boldsymbol{v}_k'$. Since the leader of $\boldsymbol{v}_i''$ is in position j_i, we get that

$$\theta^{m-t_i''} = \lambda_i \theta^{m-t_i'} + \sum_{s=1}^{i-1} \lambda_s v'_{sj_s}.$$

Let us note that $\lambda_s = 0$ for all $s < i$; otherwise the leader of $\boldsymbol{v}_i''$ is in a position with a smaller number than j_i. Thus the above equality simplifies to $\theta^{m-t_i''} = \lambda_i \theta^{m-t_i'}$, which implies that $m - t_i' \leq m - t_i''$, i.e., $t_i' \geq t_i''$ for all $i = 1, \ldots, k$. Since the sets $\{t_i'\}$ and $\{t_i''\}$ (taken in non-increasing order) give the shape of ${}_R M$ we have $t_i' = t_i''$ for all i.

Now we can conclude that $\boldsymbol{v}_k$ have zeros in positions $j_1, \ldots, j_{k-1}$ and $\theta^{m-t_k'} = \theta^{m-t_k''}$ in position j_k. Then $\boldsymbol{v}_k' - \boldsymbol{v}_k''$ has zeros in positions $j_1, \ldots, j_k$. Let $\boldsymbol{v}_k' - \boldsymbol{v}_k'' \neq 0$. Then its leader is in position different from $j_1, \ldots, j_k$, a contradiction to Lemma 2. Hence $\boldsymbol{v}_k' = \boldsymbol{v}_k''$ and the proof is completed by induction on the rank of ${}_R M$. □

Corollary 2. *Let A be a $(k \times n)$-matrix in standard form over the chain ring R. There exist permutation matrices T_1 of size $(k \times k)$ and T_2 of size $(n \times n)$ such that*

$$T_1 A T_2 = \begin{pmatrix} I_{k_0} & A_{01} & A_{02} & \ldots & A_{0,m-1} & A_{0,m} \\ 0 & \theta I_{k_1} & \theta A_{12} & \ldots & \theta A_{1,m-1} & \theta A_{1,m} \\ 0 & 0 & \theta^2 I_{k_2} & \ldots & \theta^2 A_{2,m-1} & \theta^2 A_{2,m} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & \theta^{m-1} I_{k_{m-1}} & \theta^{m-1} A_{m-1,m} \end{pmatrix}, \tag{4.1}$$

where the entries in the matrices A_{ij} are from Γ.

5. THE ORTHOGONAL MODULE

Let R be a finite chain ring and consider a left module ${}_R M \leq {}_R R^n$. For two vectors $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$ we define their inner product by

$$\boldsymbol{x}\boldsymbol{y} = x_1 y_1 + \cdots + x_n y_n.$$

The right orthogonal to ${}_R M$ is defined by

$$M_R^{\perp} = \{\boldsymbol{y} \in R^n \mid \boldsymbol{x}\boldsymbol{y} = 0 \text{ for all } \boldsymbol{x} \in M\}.$$

Analogously, we define the left orthogonal to right module $M_R \leq R_R^n$. The following theorem summarizes some basic properties of orthogonal modules [4, 5].

Theorem 6. *Let R be a chain ring with $|R| = q^m$, $R/\operatorname{Rad} R \cong \mathbb{F}_q$, and let ${}_RM \leq {}_RR^N$ be a left submodule of shape $\lambda = (\lambda_1, \ldots, \lambda_n)$.*

(1) The right module $M_R^\perp$ has shape $\overline{\lambda} = (m - \lambda_n, \ldots, m - \lambda_1)$. In particular $|M||M^\perp| = |R^n|$.

(2) ${}^\perp(M^\perp) = M$.

(3) $M \to M^\perp$ defines an antiisomorphism between the lattices of left (resp. right) submodules of R^n and hence

$$(M_1 \cap M_2)^\perp = M_1^\perp + M_2^\perp, (M_1 + M_2)^\perp = M_1^\perp \cap M_2^\perp,$$

for $M_1, M_2 \leq R^n$.

Assume A is a matrix over the chain ring R in (upper) standard form. Let ${}_RM$ be the left module generated by the rows of A. We are going to describe a method of finding a matrix B in lower standard form, whose rows generate the right orthogonal module $M_R^\perp$.

Theorem 7. *Let ${}_RM$ be a submodule of ${}_RR^n$ generated by the rows of the matrix A of the form (4.1). Then $M_R^\perp$ is generated by the matrix*

$$B = \begin{pmatrix} B_{01}\theta^{m-1} & I_{k_1}\theta^{m-1} & 0 & 0 & \ldots & 0 \\ B_{02}\theta^{m-2} & B_{12}\theta^{m-2} & I_{k_2}\theta^{m-2} & 0 & \ldots & 0 \\ B_{03}\theta^{m-3} & B_{13}\theta^{m-3} & B_{23}\theta^{m-3} & I_{k_2}\theta^{m-3} & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ B_{0,m} & B_{1,m-1} & B_{2,m-1} & B_{3,m-1} & \ldots & I_{k_{m-1}} \end{pmatrix}, \tag{5.1}$$

where

$$B_{ij} = -\big(A_{ij} - \sum_{1<k<j+1} A_{ik}A_{k,j+1} + \sum_{i<k<l<j+1} A_{ik}A_{kl}A_{l,j+1} - \cdots + (-1)^{j-i+1}A_{i,i+1}A_{i+1,i+2}\ldots A_{j,j+1}\big)^T.$$

Proof. We have to show that the dot product of any row of A with any row of B is zero. □

Corollary 3. *Let $A \in \boldsymbol{M}_{k,n}$ be a matrix over a chain ring R whose rows generate the module ${}_RM$. Let $A' = T_1AT_2$, where T_1 and T_2 are permutation matrices of orders k and n, respectively, be of the form (4.1). The module $M_R^\perp$ is generated by the rows of*

$$B = T_1^T B' T_2^T,$$

where B' is the matrix given by (5.1).

6. GENERATION OF ALL SUBMODULES OF ${}_RM$ OF FIXED SHAPE

Let ${}_RM$ be a module of shape

$$\lambda = (\underbrace{m,\dots,m}_{k_0},\underbrace{m-1,\dots,m-1}_{k_1},\dots,\underbrace{1,\dots,1}_{k_{m-1}}) = m^{k_0}(m-1)^{k_1}\dots 1^{k_{m-1}}$$

and let ${}_RN$ be a submodule of M of shape $\mu \le \lambda$. Assume M is generated by the rows of a matrix A in standard form. With no loss of generality, A has the form (4.1). Let further N be generated by the rows of another matrix B that is also in standard form. Under the above assumptions B can be represented as

$$B = CA,$$

where C is a matrix in standard form with the following properties:

(1) if the leader in row i of B is in position j_i then the leader of row i in C is in position l_i with $l_i \ge j_i$;

(2) the components of C contained in the j-th column where

$$k_0 + k_1 + \dots + k_{s-1} + 1 \le j \le k_0 + k_1 + \dots + k_s, \quad k_{-1} = 0,$$

are from $\Gamma + \theta\Gamma + \dots + \theta^{m-s-1}\Gamma$.

The proof of this observation is straightforward. It allows us to generate all submodules with a fixed shape of a module generated by the rows of some matrix A. We demonstrate this by the following example.

Example 1. Let $R = \mathbb{Z}_4$ and let A be the matrix

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

The module M generated by the rows (but also by the columns) of A is of shape $\lambda = (2,2,1,1)$. By Theorem 4 the number of all submodule $N \le M$ of shape $\mu = (2,1)$ is

$$\begin{bmatrix} \lambda \\ \mu \end{bmatrix}_{2^2} = 2^{1\cdot(4-2)} \begin{bmatrix} 4-1 \\ 2-1 \end{bmatrix}_2 2^{0(2-1)} \begin{bmatrix} 2-0 \\ 1-0 \end{bmatrix}_2 = 84.$$

We are going to construct the possible matrices C satisfying the conditions described above. Note that the last two columns can contain only entries from $\Gamma = \{0,1\}$.Thus we have the following possibilities for C:

$$\begin{pmatrix} 1 & \Gamma & \Gamma & \Gamma \\ 0 & 2 & 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & \Gamma & \Gamma \end{pmatrix}, \quad \begin{pmatrix} 1 & R & 0 & \Gamma \\ 0 & \operatorname{Rad} R & 1 & \Gamma \end{pmatrix}$$

$$\begin{pmatrix} 1 & R & \Gamma & 0 \\ 0 & \operatorname{Rad} R & 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} \operatorname{Rad} R & 1 & 0 & \Gamma \\ \operatorname{Rad} R & 0 & 1 & \Gamma \end{pmatrix}, \quad \begin{pmatrix} \operatorname{Rad} R & 1 & \Gamma & 0 \\ \operatorname{Rad} R & 0 & 0 & 1 \end{pmatrix}$$

Here $R = \{0, 1, 2, 3\}$, $\Gamma = \{0, 1\}$, and $\operatorname{Rad} R = \{0, 2\}$. Thus the number of matrices C of the first type is 8, of the second type – 4, of the third type – 32 etc, giving a total of

$$8 + 4 + 32 + 16 + 16 + 8 = 84,$$

as given by Theorem 4.

ACKNOWLEDGEMENTS. This research has been supported by the Science Research Fund of Sofia University under Contract No. 29/11.04.2016.

7. REFERENCES

[1] Birkhof, G.: Subgroups of abelian groups. *Proc. London Math. Society*, **38**, no. 2, 1934/35, 385–401.

[2] Clark, W. E., Drake, D. A.: Finite chain rings. *Abh. Math. Sem. der Univ. Hamburg*, **39**, 1974, 147–153.

[3] Honold, Th., Landjev, I.: Linearly representable codes over chain rings. *Abh. Math. Sem. der Univ. Hamburg*, **69**, 1999, 187–203.

[4] Honold, Th., Landjev, I.: Linear codes over finite chain rings. *Electron. J. Combinatorics*, **7**, 2000, #11.

[5] Honold, Th., Landjev, I.: Linear codes over finite chain rings and projective Hjelmslev geometries. In: *Codes over Rings* (P. Solé, ed.), World Scientific, 2009, pp. 60–123.

[6] van Lint, J. H., Wilson, R. M.: *A Course in Combinatorics*. Cambridge University Press, 1992.

[7] MacDonald, I. G.: *Symmetric Functions and Hall Polynomilas*. Oxford University Press, Second edition, 1995.

[8] McDonald, B. R.: *Finite Rings with Identity*. Marcel Dekker, New York, 1974.

[9] Nechaev, A. A.: *Finite Principal Ideal Rings*, Russian Acad. Sciences, Sbornik Mathematics 2091973, 364–382.

Received on December 19, 2016

Nevyana Georgieva
Department of Informatics
New Bulgarian University
21 Montevideo str., 1618 Sofia
BULGARIA
E-mail: nevyanag@fmi.uni-sofia.bg

Ivan Landjev
Department of Informatics
New Bulgarian University
21 Montevideo str., 1618 Sofia
BULGARIA
E-mail: i.landjev@nbu.bg

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 104

ANNUAL OF SOFIA UNIVERSITY „ST. KLIMENT OHRIDSKI“
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 104

RIEMANN HYPOTHESIS ANALOGUE FOR LOCALLY FINITE MODULES OVER THE ABSOLUTE GALOIS GROUP OF A FINITE FIELD

AZNIV KASPARIAN, IVAN MARINOV

The article provides a sufficient condition for a locally finite module M over the absolute Galois group $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$ of a finite field $\mathbb{F}_q$ to satisfy the Riemann Hypothesis Analogue with respect to the projective line $\mathbb{P}^1(\overline{\mathbb{F}_q})$. The condition holds for all smooth irreducible projective curves of positive genus, defined over $\mathbb{F}_q$. We give an explicit example of a locally finite module, subject to the assumptions of our main theorem and, therefore, satisfying the Riemann Hypothesis Analogue with respect to $\mathbb{P}^1(\overline{\mathbb{F}_q})$, which is not isomorphic to a smooth irreducible projective curve, defined over $\mathbb{F}_q$.

Keywords: ζ-function of a locally finite $\mathfrak{G}$-module; Riemann Hypothesis Analogue with respect to the projective line; finite unramified coverings of locally finite $\mathfrak{G}$-modules with Galois closure.

2000 Math. Subject Classification: 14G15, 94B27, 11M38.

1. INTRODUCTION

A set M with an action of a group G will be called a G-module. Most of the time we consider modules over the absolute Galois group $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$ of a finite field $\mathbb{F}_q$.

Definition 1. A $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$-module M is locally finite if all $\mathfrak{G}$-orbits on M are finite and for any $n \in \mathbb{N}$ there are at most finitely many $\mathfrak{G}$-orbits on M of cardinality n.

The cardinality of a $\mathfrak{G}$-orbit $\mathrm{Orb}_{\mathfrak{G}}(x)$, $x \in M$ is referred to as its degree and denoted by $\deg \mathrm{Orb}_{\mathfrak{G}}(x)$.

The smooth irreducible projective curves $X/\mathbb{F}_q \subseteq \mathbb{P}^n(\overline{\mathbb{F}_q})$, defined over a $\mathbb{F}_q$ are examples of locally finite $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$-modules.

Definition 2. If M is a locally finite $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$-module then the formal power series

$$\zeta_M(t) := \prod_{\nu \in \mathrm{Orb}_{\mathfrak{G}}(M)} \left(\frac{1}{1 - t^{\deg \nu}} \right) \in \mathbb{C}[[t]]$$

is called the ζ-function of M.

By its very definition, $\zeta_M(0) = 1$. In the case of a smooth irreducible curve $X/\mathbb{F}_q \subseteq \mathbb{P}^n(\overline{\mathbb{F}_q})$, the ζ-function $\zeta_X(t)$ of X as a locally finite $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$-module coincides with the local Weil ζ-function of X. We fix the projective line $\mathbb{P}^1(\overline{\mathbb{F}_q})$ as a basic model, to which we compare the locally finite $\mathfrak{G}$-modules M under consideration and recall its ζ-function

$$\zeta_{\mathbb{P}^1(\overline{\mathbb{F}_q})}(t) = \frac{1}{(1-t)(1-qt)}.$$

Definition 3. If M is a locally finite $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$-module then the ratio

$$P_M(t) := \frac{\zeta_M(t)}{\zeta_{\mathbb{P}^1(\overline{\mathbb{F}_q})}(t)}$$

of the ζ-function of M by the ζ-function of $\mathbb{P}^1(\overline{\mathbb{F}_q})$ is called briefly the ζ-quotient of M. We say that M has a polynomial ζ-quotient if $P_M(t) \in \mathbb{Z}[t]$ is a polynomial with integral coefficients.

A locally finite $\mathfrak{G}$-module M satisfies the Riemann Hypothesis Analogue with respect to the projective line $\mathbb{P}^1(\overline{\mathbb{F}_q})$ if M has a polynomial ζ-quotient

$$P_M(t) = \sum_{i=0}^{d} a_i t^i = \prod_{i=1}^{d} (1 - \omega_i t) \in \mathbb{C}[t]$$

with $|\omega_i| = \sqrt[d]{|\omega_1| \dots |\omega_d|} = \sqrt[d]{|a_d|}$, $\forall 1 \leq i \leq d$.

In order to explain the etymology of the notion, let us plug in q^{-s}, $s \in \mathbb{C}$ in the ζ-function $\zeta_M(t) = \zeta_{\mathbb{P}^1(\overline{\mathbb{F}_q})}(t) \prod_{i=1}^{d} (1 - \omega_i t)$ of M and view

$$\zeta_M\left(q^{-s}\right) = \frac{\prod_{i=1}^{d} (q^s - \omega_i)}{q^{sd-2s+1}(1-q^s)(1-q^{s-1})}$$

as a meromorphic function of $s \in \mathbb{C}$ with poles $2\pi i\mathbb{Z} \cup (1 + 2\pi i\mathbb{Z})$. If $\lambda := \log_q \sqrt[d]{|a_d|} \in \mathbb{R}^{\geq 0}$ then M satisfies the Riemann Hypothesis Analogue with respect to $\mathbb{P}^1(\overline{\mathbb{F}_q})$ exactly when the complex zeros $s_o \in \mathbb{C}$ of $\zeta_M(q^{-s})$ have $\mathrm{Re}(s_o) = \lambda$. All smooth irreducible curves $X/\mathbb{F}_q \subset \mathbb{P}^n(\overline{\mathbb{F}_q})$ of genus $g \geq 1$ satisfy the Riemann Hypothesis Analogue with respect to $\mathbb{P}^1(\overline{\mathbb{F}_q})$ by the Hasse - Weil Theorem (cf. [1] or [2]). Namely, $P_X(t) = \frac{\zeta_X(t)}{\zeta_{\mathbb{P}^1(\overline{\mathbb{F}_q})}(t)} = \prod_{i=1}^{2g}(1-\omega_i t)$ with $|\omega_i| = q^{\frac{1}{2}}$, $\forall 1 \leq i \leq 2g$, which is equivalent to $\mathrm{Re}(s_o) = \frac{1}{2}$ for all the complex zeros $s_o \in \mathbb{C}$ of $\zeta_X(q^{-s})$. That resembles the original Riemann Hypothesis $\mathrm{Re}(z_o) = \frac{1}{2}$ for the non-trivial zeros $z_o \in \mathbb{C} \setminus (-2\mathbb{N})$ of Riemann's ζ-function $\zeta(z) := \sum_{n=1}^{\infty} \frac{1}{n^z}$, $z \in \mathbb{C}$.

The present article translates Bombieri's proof of the Hasse - Weil Theorem from [1] in terms of the locally finite $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$-action on $X/\mathbb{F}_q \subset \mathbb{P}^n(\overline{\mathbb{F}_q})$ and provides a sufficient condition for an abstract locally finite $\mathfrak{G}$-module M to satisfy the Riemann Hypothesis Analogue with respect to $\mathbb{P}^1(\overline{\mathbb{F}_q})$. Grothendieck has classified the finite etale coverings of a connected scheme by the continuous action of a profinite group on their generic fibre (see [3]). In analogy with his treatment, we introduce the notion of a finite unramified covering of locally finite $\mathfrak{G}$-modules and study the deck transformation group of such a covering. One can look for an arithmetic objects A, whose reductions modulo prime integers p are locally finite $\mathrm{Gal}(\overline{\mathbb{F}_p}/\mathbb{F}_p)$-modules and study the global ζ-functions of A. Another topic of interest is the Grothendieck ring of a locally finite $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$-module and the construction of a motivic ζ-function. Our study of the Riemann Hypothesis Analogue for a locally finite $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$-module is motivated also by Duursma's notion of a ζ-function $\zeta_C(t)$ of a linear code $C \subset \mathbb{F}_q^n$ and the Riemann Hypothesis Analogue for $\zeta_C(t)$, discussed in [4]. Recently, ζ-functions have been used for description of the subgroup growth or the representations of a group, as well as of some properties of finite graphs.

The main result of the article is Theorem 29, which provides a criterion for a locally finite $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$-module M to satisfy the Riemann Hypothesis Analogue with respect to $\mathbb{P}^1(\overline{\mathbb{F}_q})$. The criterion is based on three assumptions, which are shown to be satisfied by the smooth irreducible projective curves $X/\mathbb{F}_q \subset \mathbb{P}^N(\overline{\mathbb{F}_q})$ of genus $g \geq 1$. The first assumption is the presence of a polynomial ζ-quotient $P_M(t) = \frac{\zeta_M(t)}{\zeta_{\mathbb{P}^1(\overline{\mathbb{F}_q})}(t)} = \sum_{i=0}^{d} a_i t^i \in \mathbb{Z}[t]$. The second one is the existence of locally finite $\mathfrak{G}_m = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_{q^m})$-submodules $M_o \subseteq M$, $L_o \subseteq \mathbb{P}^1(\overline{\mathbb{F}_q})$ for some $m \in \mathbb{N}$ with at most finite complements $M \setminus M_o$, $\mathbb{P}^1(\overline{\mathbb{F}_q}) \setminus L_o$, which are related by a finite unramified covering $\xi : M_o \to L_o$ of $\mathfrak{G}_m$-modules with a Galois closure (N, H, H_1), defined over $\mathbb{F}_{q^m}$. This means that N is a locally finite $\mathfrak{G}_m$-module, H is a finite fixed-point free subgroup of the automorphism group $\mathrm{Aut}_{\mathfrak{G}_m}(N)$ of N and H_1 is a subgroup of H, such that there are isomorphisms of $\mathfrak{G}_m$-modules $L_o \simeq \mathrm{Orb}_H(N) = N/H$, $M_o \simeq \mathrm{Orb}_{H_1}(N) = N/H_1$ and the finite unramified H-Galois covering $\xi_H : N \to N/H$, $\xi_H(x) = \mathrm{Orb}_H(x)$, $\forall x \in N$ has factorization

$\xi_H = \xi\xi_{H_1}$ through ξ and the finite H_1-Galois covering $\xi_{H_1} : N \to N/H_1$, $\xi_{H_1}(x) = \mathrm{Orb}_{H_1}(x)$. Finally, we assume that $\lambda := \log_q \sqrt[d]{|a_d|} \in \mathbb{R}^{\geq 0}$ is an upper bound of the Hasse - Weil order $\mathrm{ord}_{\mathfrak{G}}(M/\mathbb{P}^1(\overline{\mathbb{F}_q}))$ of M with respect to $\mathbb{P}^1(\overline{\mathbb{F}_q})$ and the Hasse - Weil H-order $\mathrm{ord}^H_{\mathfrak{G}_m}(N/\mathbb{P}^1(\overline{\mathbb{F}_q}))$ of N with respect to $\mathbb{P}^1(\overline{\mathbb{F}_q})$. We observe that the Riemann Hypothesis Analogue for M with respect to $\mathbb{P}^1(\overline{\mathbb{F}_q})$ implies a specific functional equation for the ζ-polynomial $P_M(t)$. An explicit example, constructed in Proposition 30 illustrates the existence of locally finite $\mathfrak{G}$-modules M, which are not isomorphic as $\mathfrak{G}$-modules to a smooth irreducible curve $X/\mathbb{F}_q \subset \mathbb{P}^n(\overline{\mathbb{F}_q})$ of genus $g \geq 1$ and satisfy the assumptions of our criterion for the Riemann Hypothesis Analogue with respect to $\mathbb{P}^1(\overline{\mathbb{F}_q})$.

Here is a brief synopsis of the paper. The next section 2 collects some trivial immediate properties of the locally finite $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$-modules M and their morphisms. Section 3 supplies several expressions of the ζ-function $\zeta_M(t)$ of M and shows that $\zeta_M(t)$ determines uniquely the structure of M as a $\mathfrak{G}$-module. It studies the ζ-quotient $P_M(t) = \frac{\zeta_M(t)}{\zeta_{\mathbb{P}^1(\overline{\mathbb{F}_q})}(t)} \in \mathbb{Z}[[t]]$ of M and provides two necessary and sufficient conditions for $P_M(t) \in \mathbb{Z}[t]$ to be a polynomial. An arbitrary smooth irreducible curve $X/\mathbb{F}_q \subset \mathbb{P}^n(\overline{\mathbb{F}_q})$ of genus $g \geq 1$ is shown to contain a $\mathfrak{G}_m = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_{q^m})$-submodule $X_o \subseteq X$ with $|X \setminus X_o| < \infty$, which admits a finite unramified covering $f : X_o \to L_o$ of $\mathfrak{G}_m$-modules and quasi-affine varieties onto a $\mathfrak{G}_m$-submodule $L_o \subseteq \mathbb{P}^1(\overline{\mathbb{F}_q})$ with $\left|\mathbb{P}^1(\overline{\mathbb{F}_q}) \setminus L_o\right| < \infty$. The fixed-point free automorphisms $h : M \to M$ of $\mathfrak{G}$-modules, preserving the fibres of a finite unramified covering $\xi : M \to L$ are called deck transformations of ξ. If a deck transformation group $H < \mathrm{Aut}_{\mathfrak{G}}(M)$ of ξ acts transitively on one and, therefore, on any fibre of ξ, then ξ is said to be an H-Galois covering. In order to explain the etymology of this notion, we show that if the finite separable extension $\overline{\mathbb{F}_q}(X) = \overline{\mathbb{F}_q}(X_o) \supset \overline{\mathbb{F}_q}(L_o) = \overline{\mathbb{F}_q}(\mathbb{P}^1(\overline{\mathbb{F}_q}))$ of function fields, induced from $f : X_o \to L_o$ is Galois then f is an unramified $\mathrm{Gal}(\overline{\mathbb{F}_q}(X)/\overline{\mathbb{F}_q}(\mathbb{P}^1(\overline{\mathbb{F}_q})))$-Galois covering of locally finite $\mathfrak{G}_m$-modules. For an arbitrary locally finite $\mathfrak{G}$-module M and an arbitrary finite fixed-point free subgroup $H < \mathrm{Aut}_{\mathfrak{G}}(M)$ we establish that the correspondence $\xi_H : M \to \mathrm{Orb}_H(M) = M/H$, associating to a point $x \in M$ its H-orbit $\mathrm{Orb}_H(x)$ is an H-Galois covering of locally finite $\mathfrak{G}$-modules. Moreover, $\xi_H : M \to \mathrm{Orb}_H(M)$ turns to be equivariant with respect to the pro-finite completion $\widehat{\langle\varphi\rangle}$ of the infinite cyclic subgroup of $\mathrm{Aut}_{\mathfrak{G}}(M)$, generated by $\varphi := h\Phi_q^r$ for any $h \in H$, any $r \in \mathbb{N}$ and the Frobenius automorphism Φ_q, which is a topological generator of $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q) = \widehat{\langle\Phi_q\rangle}$. Our notion of a Galois closure (N, H, H_1) of a finite unramified covering $\xi : M \to L$ of locally finite $\mathfrak{G}$-modules arises from the fact that if the function field $\overline{\mathbb{F}_q}(Z)$ of an irreducible quasi-projective curve $Z \subset \mathbb{P}^r(\overline{\mathbb{F}_q})$ is the Galois closure of the finite separable extension $\overline{\mathbb{F}_q}(X_o) \supset \overline{\mathbb{F}_q}(L_o)$, induced from $f : X_o \to L_o$ then $(Z, \mathrm{Gal}(\overline{\mathbb{F}_q}(Z)/\overline{\mathbb{F}_q}(L_o)), \mathrm{Gal}(\overline{\mathbb{F}_q}(Z)/\overline{\mathbb{F}_q}(X_o)))$ is a Galois closure of the restriction $f : X' \to L'$ of f to some locally finite $\mathfrak{G}_s$-submodules $X' \subseteq X_o$, $L' \subseteq L_o$ with $|X_o \setminus X'| < \infty$, $|L' \setminus L_o| < \infty$. The final, fifth section is devoted to the main result of the article. After reducing the Riemann Hypothesis Analogue with respect to $\mathbb{P}^1(\overline{\mathbb{F}_q})$ for a locally finite $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$-module M to lower and upper

bounds on the number of rational points of M, we introduce the notion of a Hasse - Weil order $\mathrm{ord}_{\mathfrak{G}}(M/L)$ of a locally finite $\mathfrak{G}$-module M with respect to a locally finite $\mathfrak{G}$-module L, as well as the notion of a Hasse - Weil H-order $\mathrm{ord}^{H}_{\mathfrak{G}}(N/L)$ of a locally finite $\mathfrak{G}$-module N with a finite fixed-point free subgroup $H < \mathrm{Aut}_{\mathfrak{G}}(N)$ with respect to a locally finite $\mathfrak{G}$-module L. These definitions are motivated by the celebrated Hasse - Weil bound on the number of rational points of a smooth irreducible curve $X/\mathbb{F}_q \subset \mathbb{P}^n(\overline{\mathbb{F}_q})$, which can be stated as an upper bound $\frac{1}{2}$ on the Hasse - Weil order of X with respect to the projective line $\mathbb{P}^1(\overline{\mathbb{F}_q})$. For an arbitrary finite fixed-point free subgroup $H < \mathrm{Aut}_{\mathfrak{G}}(X)$ we establish that the Hasse - Weil H-order $\mathrm{ord}^{H}_{\mathfrak{G}}(X/\mathbb{P}^1(\overline{\mathbb{F}_q})) \leq \frac{1}{2}$. The Hasse - Weil order and the Hasse - Weil H-order are shown to be preserved when passing to submodules with finite complements. The existence of a finite unramified covering $\xi : M \to L$ of locally finite $\mathfrak{G}$-modules guarantees $\mathrm{ord}_{\mathfrak{G}}(M/L) \leq 1$, while the presence of an H-Galois covering $\xi : N \to L$ suffices for $\mathrm{ord}^{H}_{\mathfrak{G}}(N/L) \leq 1$. Our main Theorem 29 provides a sufficient condition for a locally finite $\mathfrak{G}$-module M to satisfy the Riemann Hypothesis Analogue with respect to $\mathbb{P}^1(\overline{\mathbb{F}_q})$. By a specific example we establish that the assumptions of Theorem 29 hold for a class of locally finite $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$-modules, which contains strictly the smooth irreducible curves $X/\mathbb{F}_q \subset \mathbb{P}^n(\overline{\mathbb{F}_q})$ of genus $g \geq 1$. We observe also that the Riemann Hypothesis Analogue for M with respect to $\mathbb{P}^1(\overline{\mathbb{F}_q})$ implies a functional equation for the ζ-polynomial $P_M(t) := \frac{\zeta_M(t)}{\zeta_{\mathbb{P}^1(\overline{\mathbb{F}_q})}(t)} \in \mathbb{Z}[t]$ of M.

2. PRELIMINARIES ON LOCALLY FINITE GAL($\overline{\mathbb{F}_Q}/\mathbb{F}_Q$)-MODULES AND THEIR MORPHISMS

The algebraic and the separable closure of a finite field $\mathbb{F}_q$ is $\overline{\mathbb{F}_q} = \cup_{m=1}^{\infty}\mathbb{F}_{q^m}$. The absolute Galois group $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q) = \varprojlim \mathrm{Gal}(\mathbb{F}_{q^m}/\mathbb{F}_q)$ is the projective limit of the finite Galois groups $\mathrm{Gal}(\mathbb{F}_{q^m}/\mathbb{F}_q) = \langle\Phi_q\rangle = \{\Phi_q^i \,|\, 0 \leq i \leq m-1\}$, generated by the Frobenius automorphism $\Phi_q : \overline{\mathbb{F}_q} \to \overline{\mathbb{F}_q}$, $\Phi_q(a) = a^q$, $\forall a \in \overline{\mathbb{F}_q}$. Namely,

$$\mathfrak{G} = \left\{ \left(\Phi_q^{l_m (\mathrm{mod}\, m)}\right)_{m\in\mathbb{N}} \in \prod_{m=1}^{\infty} (\mathbb{Z}_m, +) \,\middle|\, l_n \equiv l_m (\mathrm{mod}\, m) \ \text{ for } \ m/n \right\}$$

is the pro-finite completion $\mathfrak{G} = \widehat{\langle\Phi_q\rangle} \simeq (\widehat{\mathbb{Z}}, +)$ of the infinite cyclic group $\langle\Phi_q\rangle \simeq (\mathbb{Z}, +)$. For an arbitrary $n \in \mathbb{N}$, note that

$$\mathfrak{G} \times \mathbb{P}^n(\overline{\mathbb{F}_q}) \longrightarrow \mathbb{P}^n(\overline{\mathbb{F}_q}),$$

$$(\Phi_q^{l_s (\mathrm{mod}\, s)})_{s\in\mathbb{N}}[a_0 : \ldots : a_i : \ldots a_n] = [a_0^{q^{l_s}} : \ldots : a_n^{q^{l_s}}] \quad \text{if} \quad a_0, \ldots, a_n \in \mathbb{F}_{q^s}$$

is a correctly defined action with finite orbits by Remark 2.1.10 (i) and Lemma 2.1.9 from [5]. By Lemma 2.1.11 from [5], the degree of $\mathrm{Orb}_{\mathfrak{G}}(a) = \mathrm{Orb}_{\langle\Phi_q\rangle}(a)$, $a \in$

$\mathbb{P}^n(\overline{\mathbb{F}_q})$ is the minimal $m \in \mathbb{N}$ with $\left[a_0^{q^m} : \ldots : a_n^{q^m}\right] = \Phi_q^m(a) = a = [a_0 : \ldots : a_n]$. If $a_i \neq 0$ then $\Phi_q^m(a) = a$ amounts to $\left(\frac{a_j}{a_i}\right)^{q^m} = \frac{a_j}{a_i}$, $\forall 0 \leq j \leq n$ and holds exactly when $\frac{a_j}{a_i} \in \mathbb{F}_{q^m}$, $\forall 0 \leq j \leq n$. Thus, $\forall m \in \mathbb{N}$ there are finitely many $\mathrm{Orb}_{\mathfrak{G}}(a) \subset \mathbb{P}^n(\overline{\mathbb{F}_q})$ of $\deg \mathrm{Orb}_{\mathfrak{G}}(a) = m$ and $\mathbb{P}^n(\overline{\mathbb{F}_q})$ is a locally finite $\mathfrak{G}$-module.

If $X = V(f_1, \ldots, f_l) \subset \mathbb{P}^n(\overline{\mathbb{F}_q})$ is a smooth irreducible curve, cut by homogeneous polynomials $f_1, \ldots, f_l \in \mathbb{F}_q[x_0, \ldots, x_n]$ with coefficients from $\mathbb{F}_q$, X is said to be defined over $\mathbb{F}_q$ and denoted by $X/\mathbb{F}_q \subset \mathbb{P}^n(\overline{\mathbb{F}_q})$. The $\mathfrak{G}$-action on $\mathbb{P}^n(\overline{\mathbb{F}_q})$ restricts to a locally finite $\mathfrak{G}$-action on X, due to the $\mathfrak{G}$-invariance of $f_1, \ldots, f_l$.

Here are some trivial properties of the locally finite $\widehat{\mathbb{Z}}$-actions.

Lemma 4. *Let $\mathfrak{G} = \widehat{\langle\varphi\rangle}$ be the profinite completion of an infinite cyclic group $\langle\varphi\rangle \simeq (\mathbb{Z}, +)$, M be a locally finite $\mathfrak{G}$-module with closed stabilizers, $\mathrm{Orb}_{\mathfrak{G}}(x) \subseteq M$ be a $\mathfrak{G}$-orbit on M of degree $m = \deg \mathrm{Orb}_{\mathfrak{G}}(x)$ and $\mathfrak{G}_m = \widehat{\langle\varphi^m\rangle}$ be the profinite completion of $\langle\varphi^m\rangle \simeq (\mathbb{Z}, +)$. Then:*

(i) any $y \in \mathrm{Orb}_{\mathfrak{G}}(x)$ has stabilizer $\mathrm{Stab}_{\mathfrak{G}}(y) = \mathrm{Stab}_{\mathfrak{G}}(x) = \mathfrak{G}_m$;

(ii) the orbits $\mathrm{Orb}_{\mathfrak{G}}(x) = \mathrm{Orb}_{\langle\varphi\rangle}(x) = \{x, \varphi(x), \ldots, \varphi^{m-1}(x)\}$ coincide;

(iii) $\forall r \in \mathbb{N}$ with greatest common divisor $\mathrm{GCD}(r, m) = d \in \mathbb{N}$, the $\mathfrak{G}$-orbit

$$\mathrm{Orb}_{\mathfrak{G}}(x) = \coprod_{j=1}^{d} \mathrm{Orb}_{\mathfrak{G}_r}(\varphi^{i_j}(x))$$

of x decomposes into a disjoint union of d orbits of degree $m_1 = \frac{m}{d}$ with respect to the action of $\mathfrak{G}_r = \widehat{\langle\varphi^r\rangle}$.

Proof. If $\mathfrak{G}' := \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q) = \widehat{\langle\Phi_q\rangle}$ is the absolute Galois group of the finite field $\mathbb{F}_q$, then the group isomorphism $f : \langle\varphi\rangle \longrightarrow \langle\Phi_q\rangle$, $f(\varphi^s) = \Phi_q^s$, $\forall s \in \mathbb{N}$ extends uniquely to a group isomorphism

$$f : \mathfrak{G} = \widehat{\langle\varphi\rangle} \longrightarrow \widehat{\langle\Phi_q\rangle} = \mathfrak{G}', \quad f(\varphi^{l_s(\mathrm{mod}\, s)})_{s\in\mathbb{N}} = (\Phi_q^{l_s(\mathrm{mod}\, s)})_{s\in\mathbb{N}} \in \prod_{s\in\mathbb{N}} (\langle\Phi_q\rangle/\langle\Phi_q^s\rangle)$$

of the corresponding pro-finite completions. That is why it suffices to prove the lemma for $\mathfrak{G}' = \widehat{\langle\Phi_q\rangle}$.

(i) By assumption, $\mathrm{Stab}_{\mathfrak{G}}(x)$ is a closed subgroup of $\mathfrak{G}$ of index

$$[\mathfrak{G} : \mathrm{Stab}_{\mathfrak{G}}(x)] = \deg \mathrm{Orb}_{\mathfrak{G}}(x) = m.$$

According to $\mathrm{Gal}(\mathbb{F}_{q^m}/\mathbb{F}_q) = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)/\mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_{q^m}) = \mathfrak{G}'/\mathfrak{G}'_m$ for $\mathfrak{G}'_m = \widehat{\langle\Phi_q^m\rangle}$, the closed subgroup $\mathfrak{G}'_m$ of $\mathfrak{G}'$ is of index m and the closed subgroup $\mathfrak{G}_m$ of $\mathfrak{G}$ is of index $[\mathfrak{G} : \mathfrak{G}_m] = m$. If $\mathcal{H}$ is a closed subgroup of $\mathfrak{G}$ of $[\mathfrak{G} : \mathcal{H}] = m$ then $\mathfrak{G}/\mathcal{H}$ is an abelian group of order m and $\varphi^m \in \mathcal{H}$, $\forall \varphi \in \mathfrak{G}$. Therefore the closure $\mathfrak{G}_m = \widehat{\langle\varphi^m\rangle}$ of $\langle\varphi^m\rangle$ in $\mathfrak{G}$ is contained in $\mathcal{H}$ and $[\mathcal{H} : \mathfrak{G}_m] = \frac{[\mathfrak{G}:\mathfrak{G}_m]}{[\mathfrak{G}:\mathcal{H}]} = 1$. Thus, $\mathcal{H} = \mathfrak{G}_m$ is the

only closed subgroup of $\mathfrak{G}$ of index m and $\mathrm{Stab}_{\mathfrak{G}}(x) = \mathfrak{G}_m$. Since $\mathfrak{G}$ is an abelian group, any $y \in \mathrm{Orb}_{\mathfrak{G}}(x)$ has the same stabilizer $\mathrm{Stab}_{\mathfrak{G}}(y) = \mathrm{Stab}_{\mathfrak{G}}(x) = \mathfrak{G}_m$ as x.

(ii) The inclusion $\langle\varphi\rangle \subset \widehat{\langle\varphi\rangle} = \mathfrak{G}$ of groups implies the inclusion $\mathrm{Orb}_{\langle\varphi\rangle}(x) \subseteq \mathrm{Orb}_{\mathfrak{G}}(x)$ of the corresponding orbits. It suffices to show that $x, \varphi(x), \ldots, \varphi^{m-1}(x)$ are pairwise different, in order to conclude that $\deg \mathrm{Orb}_{\langle\varphi\rangle}(x) \geq m = \deg \mathrm{Orb}_{\mathfrak{G}}(x)$, whereas $\mathrm{Orb}_{\langle\varphi\rangle}(x) = \mathrm{Orb}_{\mathfrak{G}}(x)$. Indeed, if $\varphi^i(x) = \varphi^j(x)$ for some $0 \leq i < j \leq m-1$ then $x = \varphi^{j-i}(x)$ implies $\varphi^{j-i} \in \mathrm{Stab}_{\mathfrak{G}}(x) \cap \langle\varphi\rangle = \widehat{\langle\varphi^m\rangle} \cap \langle\varphi\rangle = \langle\varphi^m\rangle$ and m divides $0 < j - i \leq m - 1$. This is an absurd, justifying $\mathrm{Orb}_{\langle\varphi\rangle}(x) = \mathrm{Orb}_{\mathfrak{G}}(x)$.

(iii) It suffices to check that $\forall y \in \mathrm{Orb}_{\mathfrak{G}}(x)$ has stabilizer $\mathrm{Stab}_{\mathfrak{G}_r}(y) = \mathfrak{G}_{rm_1}$, in order to apply (i) and to conclude that $\deg \mathrm{Orb}_{\mathfrak{G}_r}(y) = m_1$. Bearing in mind that $\mathrm{Stab}_{\mathfrak{G}_r}(y) = \mathrm{Stab}_{\mathfrak{G}}(y) \cap \mathfrak{G}_r = \mathfrak{G}_m \cap \mathfrak{G}_r$ and the least common multiple of m and r is $\mathrm{LCM}(m, r) = rm_1 = mr_1 \in \mathbb{N}$ for $r_1 = \frac{r}{d}$, we reduce the statement to $\mathfrak{G}_m \cap \mathfrak{G}_r = \mathfrak{G}_{\mathrm{LCM}(m,r)}$. According to

$$\mathfrak{G}_r/(\mathfrak{G}_m \cap \mathfrak{G}_r) \simeq \mathfrak{G}_r\mathfrak{G}_m/\mathfrak{G}_m < \mathfrak{G}/\mathfrak{G}_m,$$

the index $s := [\mathfrak{G} : \mathfrak{G}_m \cap \mathfrak{G}_r] = [\mathfrak{G} : \mathfrak{G}_r][\mathfrak{G}_r : (\mathfrak{G}_m \cap \mathfrak{G}_r)] \leq rm$ is finite and $\mathfrak{G}_m \cap \mathfrak{G}_r = \mathfrak{G}_s$. By $\mathfrak{G}_s < \mathfrak{G}_m < \mathfrak{G}$ and $\mathfrak{G}_s < \mathfrak{G}_r < \mathfrak{G}$ the integer $s \in \mathbb{N}$ is a common multiple of m, r, so that $\mathrm{LCM}(m, r) \in \mathbb{N}$ divides s. Since $\mathfrak{G}_{\mathrm{LCM}(m,r)} = \mathfrak{G}_{rm_1} = \mathfrak{G}_{r_1 m}$ is contained in $\mathfrak{G}_m$ and $\mathfrak{G}_r$, there follows $\mathfrak{G}_{\mathrm{LCM}(m,r)} \leq \mathfrak{G}_m \cap \mathfrak{G}_r = \mathfrak{G}_s$, so that s divides $\mathrm{LCM}(m, r)$ and $s = \mathrm{LCM}(m, r)$. □

If M and L are modules over a group G then the G-equivariant maps

$$\xi : M \longrightarrow L, \quad g\xi(x) = \xi(gx) \quad \forall g \in G, \quad \forall x \in M$$

are called morphisms of G-modules. Let $\xi : M \to L$ be a morphism of locally finite $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$-modules. The next proposition provides a numerical description of the restriction of ξ on a preimage of a $\mathfrak{G}$-orbit, by the means of the inertia indices of ξ. Note that the image $\xi(M)$ is $\mathfrak{G}$-invariant and for any complete set $\Sigma_{\mathfrak{G}}(\xi(M)) \subseteq \xi(M)$ of $\mathfrak{G}$-orbit representatives on $\xi(M)$, the $\mathfrak{G}$-orbit decomposition $\xi(M) = \coprod_{x \in \Sigma_{\mathfrak{G}}(\xi(M))} \mathrm{Orb}_{\mathfrak{G}}(x)$ pulls back to a disjoint $\mathfrak{G}$-module decomposition

$$M = \coprod_{x \in \Sigma_{\mathfrak{G}}(\xi(M))} \xi^{-1}\mathrm{Orb}_{\mathfrak{G}}(x). \tag{2.1}$$

Thus, the morphism $\xi : M \to L$ of $\mathfrak{G}$-modules is completely determined by the surjective morphisms $\xi : \xi^{-1}\mathrm{Orb}_{\mathfrak{G}}(x) \longrightarrow \mathrm{Orb}_{\mathfrak{G}}(x)$ of $\mathfrak{G}$-modules $\forall x \in \Sigma_{\mathfrak{G}}(\xi(M))$.

Proposition 5. *Let $\xi : M \to L$ be a morphism of locally finite modules with closed stabilizers over the pro-finite completion $\mathfrak{G} = \widehat{\langle\varphi\rangle}$ of an infinite cyclic group $\langle\varphi\rangle \simeq (\mathbb{Z}, +)$,*

$$\delta = \deg \mathrm{Orb}_{\mathfrak{G}} : L \longrightarrow \mathbb{N}, \quad \delta(x) = \deg \mathrm{Orb}_{\mathfrak{G}}(x) \quad \textit{for} \quad \forall x \in L \quad \textit{and}$$

$$e_\xi : M \longrightarrow \mathbb{Q}^{>0}, \quad e_\xi(y) = \frac{\deg \mathrm{Orb}_{\mathfrak{G}}(y)}{\deg \mathrm{Orb}_{\mathfrak{G}}(\xi(y))} \quad \forall y \in M.$$

Then:

(i) $\mathrm{Stab}_{\mathfrak{G}}(y)$ *is a subgroup of* $\mathrm{Stab}_{\mathfrak{G}}(\xi(y))$ *for all the points* $y \in M$*, so that* $e_\xi(y) = [\mathrm{Stab}_{\mathfrak{G}}(\xi(y)) : \mathrm{Stab}_{\mathfrak{G}}(y)] \in \mathbb{N}$ *takes natural values;*

(ii) for any $x \in \xi(M)$ *there is a subset* $S_x \subseteq \xi^{-1}(x)$*, such that*

$$\xi^{-1}\mathrm{Orb}_{\mathfrak{G}}(x) = \coprod_{y \in S_x} \mathrm{Orb}_{\mathfrak{G}}(y) \quad \textit{with} \quad \deg \mathrm{Orb}_{\mathfrak{G}}(y) = \delta(x) e_\xi(y); \tag{2.2}$$

(iii) $\forall x \in \xi(M)$ *the fibre* $\xi^{-1}(x)$ *is a* $\mathfrak{G}_{\delta(x)}$*-module with orbit decomposition*

$$\xi^{-1}(x) = \coprod_{y \in S_x} \mathrm{Orb}_{\mathfrak{G}_{\delta(x)}}(y) \quad \textit{of} \quad \deg \mathrm{Orb}_{\mathfrak{G}_{\delta(x)}}(y) = e_\xi(y). \tag{2.3}$$

The correspondence $e_\xi : M \to \mathbb{N}$ *is called the inertia map of* $\xi : M \to L$*. The values* $e_\xi(y)$*,* $y \in M$ *of* e_ξ *are called inertia indices of* ξ*.*

Proof. (i) The $\mathfrak{G}$-equivariance of ξ implies that $\mathrm{Stab}_{\mathfrak{G}}(y) \leq \mathrm{Stab}_{\mathfrak{G}}(\xi(y)) \leq \mathfrak{G}$. Combining with Lemma 4 (i), one expresses

$$e_\xi(y) = \frac{[\mathfrak{G} : \mathrm{Stab}_{\mathfrak{G}}(y)]}{[\mathfrak{G} : \mathrm{Stab}_{\mathfrak{G}}(\xi(y))]} = [\mathrm{Stab}_{\mathfrak{G}}(\xi(y)) : \mathrm{Stab}_{\mathfrak{G}}(y)] \in \mathbb{N}.$$

(ii) We claim that $\forall x \in \xi(M)$ all $\mathfrak{G}$-orbits on $\xi^{-1}\mathrm{Orb}_{\mathfrak{G}}(x)$ intersect the fibre $\xi^{-1}(x)$. Indeed, assuming $\xi(z) = \varphi^s(x)$ for some $z \in M$ and $0 \leq s \leq \delta(x) - 1$, one observes that $\xi(\varphi^{\delta(x)-s} z) = \varphi^{\delta(x)-s}\xi(z) = x$, whereas $y := \varphi^{\delta(x)-s}(z) \in \xi^{-1}(x)$ with $\mathrm{Orb}_{\mathfrak{G}}(z) = \mathrm{Orb}_{\mathfrak{G}}(y)$. That allows to choose a complete set $S_x \subseteq \xi^{-1}(x)$ of $\mathfrak{G}$-orbit representatives on $\xi^{-1}\mathrm{Orb}_{\mathfrak{G}}(x)$ and to obtain (2.2) by the very definition of $e_\xi(y)$ with $y \in S_x \subseteq \xi^{-1}(x)$.

(iii) If $x \in \xi(M)$, $y \in \xi^{-1}(x)$ then $\xi(\varphi^{\delta(x)} y) = \varphi^{\delta(x)}\xi(y) = \varphi^{\delta(x)}(x) = x$ implies $\varphi^{\delta(x)}(y) \in \xi^{-1}(x)$, so that $\xi^{-1}(x)$ is acted by $\mathfrak{G}_{\delta(x)} = \widehat{\langle \varphi^{\delta(x)} \rangle}$. That justifies the inclusion $\cup_{y \in S_x} \mathrm{Orb}_{\mathfrak{G}_{\delta(x)}}(y) \subseteq \xi^{-1}(x)$. For any $y, y' \in S_x$ the assumption $y' \in \mathrm{Orb}_{\mathfrak{G}_{\delta(x)}}(y) \subseteq \mathrm{Orb}_{\mathfrak{G}}(y)$ implies that $y' = y$, so that the union $\coprod\limits_{y \in S_x} \mathrm{Orb}_{\mathfrak{G}_{\delta(x)}}(y)$ is disjoint. By the very definition of S_x, any

$$z \in \xi^{-1}(x) \subset \xi^{-1}\mathrm{Orb}_{\mathfrak{G}}(x) = \coprod_{y \in S_x} \mathrm{Orb}_{\mathfrak{G}}(y)$$

is of the form $z = \varphi^s(y)$ for some $y \in S_x$ and $0 \leq s < \delta(x) e_\xi(y) - 1$. Due to $x = \xi(z) = \xi(\varphi^s(y)) = \varphi^s\xi(y) = \varphi^s(x)$, there follows $\varphi^s \in \mathrm{Stab}_{\mathfrak{G}}(x) \cap \langle \varphi \rangle = \widehat{\langle \varphi^{\delta(x)} \rangle} \cap \langle \varphi \rangle = \langle \varphi^{\delta(x)} \rangle$, whereas $s = \delta(x) r$ for some $r \in \mathbb{Z}^{\geq 0}$. Thus, $z = \varphi^{\delta(x) r}(y) \in \mathrm{Orb}_{\mathfrak{G}_{\delta(x)}}(y)$ and $\xi^{-1}(x) \subseteq \coprod\limits_{y \in S_x} \mathrm{Orb}_{\mathfrak{G}_{\delta(x)}}(y)$. That justifies the $\mathfrak{G}_{\delta(x)}$-orbit decomposition (2.3). By (ii) and the proof of Lemma 4 (iii), one has $\mathrm{Stab}_{\mathfrak{G}_{\delta(x)}}(y) =$

$\mathrm{Stab}_{\mathfrak{G}}(y)\cap\mathfrak{G}_{\delta(x)}=\mathfrak{G}_{\delta(x)e_\xi(y)}\cap\mathfrak{G}_{\delta(x)}=\mathfrak{G}_{\delta(x)e_\xi(y)}$, as far as $\mathrm{LCM}(\delta(x)e_\xi(y),\delta(x))=\delta(x)e_\xi(y)$. Now, Lemma 4(i) applies to provide $\deg \mathrm{Orb}_{\mathfrak{G}_{\delta(x)}}(y)=e_\xi(y)$. □

3. LOCALLY FINITE MODULES WITH A POLYNOMIAL ζ-QUOTIENT

In order to provide two more expressions for the ζ-function of a locally finite module M over $\mathfrak{G}=\mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$, let us recall that on an arbitrary smooth irreducible curve $X/\mathbb{F}_q\subseteq\mathbb{P}^n(\overline{\mathbb{F}_q})$, defined over $\mathbb{F}_q$, the fixed points

$$X^{\Phi_q^r}:=\{x\in X\,|\,\Phi_q^r(x)=x\}=X(\mathbb{F}_{q^r})$$

of an arbitrary power Φ_q^r, $r\in\mathbb{N}$ of the Frobenius automorphism Φ_q coincide with the $\mathbb{F}_{q^r}$-rational ones. That is why, for an arbitrary locally finite module M over the pro-finite completion $\mathfrak{G}=\widehat{\langle\varphi\rangle}$ of an infinite cyclic group $\langle\varphi\rangle\simeq(\mathbb{Z},+)$, the fixed points

$$M^{\varphi^r}:=\{x\in M\,|\,\varphi^r(x)=x\}$$

of φ^r with $r\in\mathbb{N}$ are called φ^r-rational. Note that if $\deg\mathrm{Orb}_{\mathfrak{G}}(x)=m$ then $x\in M^{\varphi^r}$ if and only if $\varphi^r\in\mathrm{Stab}_{\mathfrak{G}}(x)=\mathfrak{G}_m=\widehat{\langle\varphi^m\rangle}$ and this holds exactly when m divides r. Since any fixed $r\in\mathbb{N}$ has finitely many natural divisors m and for any $m\in\mathbb{N}$ there are at most finitely many $\mathfrak{G}$-orbits on M of degree m, the sets M^{φ^r} are finite.

Let us consider the free abelian group $(\mathrm{Div}(M),+)$, generated by the $\mathfrak{G}$-orbits $\nu\in\mathrm{Orb}_{\mathfrak{G}}(M)$. Its elements are called divisors on M and are of the form $D=a_1\nu_1+\ldots+a_s\nu_s$ for some $\nu_j\in\mathrm{Orb}_{\mathfrak{G}}(M)$, $a_j\in\mathbb{Z}$. The terminology arises from the case of a smooth irreducible curve $X/\mathbb{F}_q\subseteq\mathbb{P}^n(\overline{\mathbb{F}_q})$, in which the $\mathfrak{G}=\mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$-orbits ν are in a bijective correspondence with the places $\widetilde{\nu}$ of the function field $\mathbb{F}_q(X)$ of X over $\mathbb{F}_q$. If $R_{\widetilde{\nu}}$ is the discrete valuation ring, associated with the place $\widetilde{\nu}$ then the residue field $R_{\widetilde{\nu}}/\mathfrak{M}_{\widetilde{\nu}}$ of $R_{\widetilde{\nu}}$ is of degree $[R_{\widetilde{\nu}}/\mathfrak{M}_{\widetilde{\nu}}:\mathbb{F}_q]=\deg\nu$.

Note that the degree of a $\mathfrak{G}$-orbit extends to a group homomorphism

$$\deg:(\mathrm{Div}(M),+)\longrightarrow(\mathbb{Z},+),\quad \deg\left(\sum_{\nu\in\mathrm{Orb}_{\mathfrak{G}}(M)}a_\nu\nu\right)=\sum_{\nu\in\mathrm{Orb}_{\mathfrak{G}}(M)}a_\nu\deg\nu.$$

A divisor $D=a_1\nu_1+\ldots+a_s\nu_s\geq 0$ is effective if all of its non-zero coefficients are positive. Let $\mathrm{Div}_{\geq 0}(M)$ be the set of the effective divisors on M. Note that the effective divisors $D=a_1\nu_1+\ldots+a_s\nu_s\geq 0$ on M of fixed degree $\deg D=a_1\deg\nu_1+\ldots+a_s\deg\nu_s=m\in\mathbb{Z}^{\geq 0}$ have bounded coefficients $1\leq a_j\leq m$ and bounded degrees $\deg\nu_j\leq m$ of the $\mathfrak{G}$-orbits from the support of D. Bearing in mind that M has at most finitely many $\mathfrak{G}$-orbits ν_j of degree $\deg\nu_j\leq m$, one concludes that there are at most finitely many effective divisors on M of degree $m\in\mathbb{Z}^{\geq 0}$ and denotes their number by $\mathcal{A}_m(M)$.

The following statement generalizes two of the well known expressions of the local Weil ζ-function $\zeta_X(t)$ of a smooth irreducible curve $X/\mathbb{F}_q \subset \mathbb{P}^n(\overline{\mathbb{F}_q})$ to the ζ-function of any locally finite $\mathfrak{G} = \widehat{\langle\varphi\rangle}$-module M. The proofs are similar to the ones for $X/\mathbb{F}_q \subset \mathbb{P}^n(\overline{\mathbb{F}_q})$, given in [5] or in [2].

Proposition 6. *Let $\mathfrak{G} = \widehat{\langle\varphi\rangle}$ be the pro-finite completion of an infinite cyclic group $\langle\varphi\rangle$ and M be a locally finite $\mathfrak{G}$-module. Then the ζ-function of M equals*

$$\zeta_M(t) = \exp\left(\sum_{r=1}^{\infty} \left|M^{\varphi^r}\right| \frac{t^r}{r}\right) = \sum_{m=0}^{\infty} \mathcal{A}_m(M) t^m,$$

where $\left|M^{\varphi^r}\right|$ is the number of φ^r-rational points on M and $\mathcal{A}_m(M)$ is the number of the effective divisors on M of degree $m \in \mathbb{Z}^{\geq 0}$.

Proof. If $B_k(M)$ is the number of $\mathfrak{G}$-orbits on M of degree k then

$$\zeta_M(t) := \prod_{\nu \in \mathrm{Orb}_{\mathfrak{G}}(M)} \left(\frac{1}{1-t^{\deg \nu}}\right) = \prod_{k=1}^{\infty} \left(\frac{1}{1-t^k}\right)^{B_k(M)}.$$

Therefore

$$\begin{aligned}\log \zeta_M(t) &= -\sum_{k=1}^{\infty} B_k(M) \log(1-t^k) = \sum_{k=1}^{\infty} B_k(M) \left(\sum_{n=1}^{\infty} \frac{t^{kn}}{n}\right) \\ &= \sum_{r=1}^{\infty} \left(\sum_{k/r} k B_k(M)\right) \frac{t^r}{r},\end{aligned}$$

according to the equality of formal power series

$$\log(1-z) = -\sum_{r=1}^{\infty} \frac{z^r}{r} \in \mathbb{Q}[[z]]. \tag{3.1}$$

If $M^{\varphi^r} = \coprod_{\deg \mathrm{Orb}_{\mathfrak{G}}(x)/r} \mathrm{Orb}_{\mathfrak{G}}(x)$ is the decomposition of M^{φ^r} into a disjoint union of $\mathfrak{G}$-orbits then the number of the φ^r-rational points on M is

$$\left|M^{\varphi^r}\right| = \sum_{k/r} k B_k(M), \tag{3.2}$$

whereas $\log \zeta_M(t) = \sum_{r=1}^{\infty} \left|M^{\varphi^r}\right| \frac{t^r}{r}$.

On the other hand, there is an equality of formal power series

$$\zeta_M(t) = \prod_{\nu \in \mathrm{Orb}_{\mathfrak{G}}(M)} \left(\sum_{n=0}^{\infty} t^{\deg(n\nu)}\right) = \sum_{D \in \mathrm{Div}_{\geq 0}(M)} t^{\deg D} = \sum_{m=0}^{\infty} \mathcal{A}_m(M) t^m. \qquad \square$$

For an arbitrary group G, the bijective morphisms $\xi : M \to L$ of G-modules are called isomorphisms of G-modules.

Corollary 7. *Locally finite $\mathfrak{G} = \widehat{\langle\varphi\rangle}$-modules M, L admit an isomorphism of $\mathfrak{G}$-modules $\xi : M \to L$ if and only if their ζ-functions $\zeta_M(t) = \zeta_L(t)$ coincide.*

Proof. Let $\xi : M \to L$ be an isomorphism of $\mathfrak{G}$-modules and $x \in L$ be a point with $\deg \mathrm{Orb}_{\mathfrak{G}}(x) = \delta(x)$. Then (2.3) from Proposition-Definition 5 (iii) provides a decomposition $\xi^{-1}(x) = \coprod\limits_{y \in S_x} \mathrm{Orb}_{\mathfrak{G}_{\delta(x)}}(y)$ of the fibre $\xi^{-1}(x)$ in a disjoint union of $\mathfrak{G}_{\delta(x)}$-orbits of $\deg \mathrm{Orb}_{\mathfrak{G}_{\delta(x)}}(y) = e_\xi(y)$. Therefore $|S_x| = 1$, $\forall x \in L$, $e_\xi(y) = 1$, $\forall y \in M$ and $\xi^{-1}\mathrm{Orb}_{\mathfrak{G}}(x) = \mathrm{Orb}_{\mathfrak{G}}\xi^{-1}(x)$ is of degree $\delta(x)$ by (2.2) from Proposition-Definition 5 (ii). As a result, (2.1) takes the form $M = \coprod\limits_{x \in \Sigma_{\mathfrak{G}}(L)} \mathrm{Orb}_{\mathfrak{G}}\xi^{-1}(x)$ for any complete set $\Sigma_{\mathfrak{G}}(L)$ of $\mathfrak{G}$-orbit representatives on L and $\zeta_M(t) = \prod\limits_{x \in \Sigma_{\mathfrak{G}}(L)} \left(\frac{1}{1-t^{\delta(x)}}\right) = \zeta_L(t)$.

Conversely, assume that the locally finite $\mathfrak{G}$-modules M and L have one and a same ζ-function $\zeta_M(t) = \zeta_L(t)$. Then by Proposition 6, there follows the equality

$$\sum_{r=1}^{\infty} \left|M^{\varphi^r}\right| \frac{t^r}{r} = \log \zeta_M(t) = \log \zeta_L(t) = \sum_{r=1}^{\infty} \left|L^{\varphi^r}\right| \frac{t^r}{r} \in \mathbb{Q}[[t]]$$

of formal power series of t, whereas the equalities

$$\sum_{d/r} dB_d(M) = \left|M^{\varphi^r}\right| = \left|L^{\varphi^r}\right| = \sum_{d/r} dB_d(L)$$

of their coefficients $\forall r \in \mathbb{N}$. By an induction on r, one derives that $B_d(M) = B_d(L)$, $\forall d \in \mathbb{N}$. For any $k \in \mathbb{N}$ note that $M^{(\leq k)} := \{x \in M \mid \deg \mathrm{Orb}_{\mathfrak{G}}(x) \leq k\}$ is a finite $\mathfrak{G}$-submodule of M and the locally finite $\mathfrak{G}$-module $M = \cup_{k=1}^{k} M^{(\leq k)}$ is exhausted by $M^{(\leq k)}$. If $L^{(\leq k)} := \{y \in L \mid \deg \mathrm{Orb}_{\mathfrak{G}}(y) \leq k\}$ then by an induction on $k \in \mathbb{N}$ one constructs isomorphisms $\xi : M^{(\leq k)} \to L^{(\leq k)}$ of $\mathfrak{G}$-modules and obtains an isomorphism of $\mathfrak{G}$-modules $\xi : M = \cup_{k=1}^{\infty} M^{(\leq k)} \to \cup_{k=1}^{\infty} L^{(\leq k)} = L$. □

Lemma 8. *If M is a locally finite $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$-module with ζ-function $\zeta_M(t) \in \mathbb{Z}[[t]]$ then the quotient*

$$P_M(t) = \frac{\zeta_M(t)}{\zeta_{\mathbb{P}^1(\overline{\mathbb{F}_q})}(t)} = \sum_{i=0}^{\infty} a_i t^i \in \mathbb{Z}[[t]]^*$$

is a formal power series with integral coefficients $a_m \in \mathbb{Z}$, which is invertible in $\mathbb{Z}[[t]]$. Its coefficients $a_m \in \mathbb{Z}$ satisfy the equality

$$\mathcal{A}_m(M) = \sum_{i=0}^{m} a_i \left|\mathbb{P}^{m-i}(\mathbb{F}_q)\right|$$

and can be interpreted as "multiplicities" of the projective spaces $\mathbb{P}^{m-i}(\mathbb{F}_q)$, "exhausting" the effective divisors on M of degree m.

Proof. If $P_M(t) = \sum\limits_{m=0}^{\infty} a_m t^m \in \mathbb{C}[[t]]$ is a formal power series with complex coefficients $a_m \in \mathbb{C}$ then the comparison of the coefficients of

$$\sum_{m=0}^{\infty} a_m t^m = P_M(t) = \zeta_M(t)(1-t)(1-qt) = \left(\sum_{m=0}^{\infty} \mathcal{A}_m(M) t^m\right)[1-(q+1)t+qt^2]$$

yields

$$a_m = \mathcal{A}_m(M) - (q+1)\mathcal{A}_{m-1}(M) + q\mathcal{A}_{m-2}(M) \in \mathbb{Z} \ \ \forall m \in \mathbb{Z}^{\geq 0}, \tag{3.3}$$

as far as $\mathcal{A}_m(M) \in \mathbb{Z}^{\geq 0}$, $\forall m \in \mathbb{Z}^{\geq 0}$ and $\mathcal{A}_{-1}(M) = \mathcal{A}_{-2}(M) = 0$. In particular, $a_0 = \mathcal{A}_0(M) = \zeta_M(0) = 1$ and $P_M(t) = 1 + \sum\limits_{i=1}^{\infty} a_i t^i \in \mathbb{Z}[[t]]^*$ is invertible by a formal power series $P_M^{-1}(t) = 1 + \sum\limits_{m=1}^{\infty} b_m t^m \in \mathbb{Z}[[t]]$ with integral coefficients. (The existence of $b_m \in \mathbb{Z}$ with $[1 + \sum\limits_{m=1}^{\infty} a_m t^m][1 + \sum\limits_{m=1}^{\infty} b_m t^m] = 1$ follows from $b_m + \sum\limits_{i=1}^{m-1} b_i a_{m-i} + a_m = 0$ by an induction on $m \in \mathbb{N}$.)

The comparison of the coefficients of

$$\sum_{m=0}^{\infty} \mathcal{A}_m(M) t^m = \zeta_M(t) = P_M(t)\zeta_{\mathbb{P}^1(\overline{\mathbb{F}_q})}(t) = \left(\sum_{m=0}^{\infty} a_m t^m\right)\left(\sum_{s=0}^{\infty} t^s\right)\left(\sum_{r=0}^{\infty} q^r t^r\right)$$

provides

$$\mathcal{A}_m(M) = \sum_{i=0}^{m} a_i \left(\sum_{j=0}^{m-i} q^j\right) = \sum_{i=0}^{m} a_i \left(\frac{q^{m-i+1}-1}{q-1}\right) = \sum_{i=0}^{m} a_i \left|\mathbb{P}^{m-i}(\mathbb{F}_q)\right|. \tag{3.4}$$

□

According to the Riemann-Roch Theorem for a divisor D of degree $\deg D = n \geq 2g-1$ on a smooth irreducible curve $X/\mathbb{F}_q \subseteq \mathbb{P}^n(\overline{\mathbb{F}_q})$ of genus $g \geq 0$, the linear equivalence class of D is isomorphic to $\mathbb{P}^{n-g}(\mathbb{F}_q)$. For any $n \in \mathbb{Z}^{\geq 0}$ there exist one and a same number h of linear equivalence classes of divisors on X of degree n. The natural number $h = P_X(1)$ equals the value of the ζ-polynomial $P_X(t) = \frac{\zeta_X(t)}{\zeta_{\mathbb{P}^1(\overline{\mathbb{F}_q})}(t)} = \sum\limits_{j=0}^{2g} a_j t^j \in \mathbb{Z}[t]$ of X at 1 and is called the class number of X. Thus, for any natural number $n \geq 2g-1$ there are

$$\mathcal{A}_n(X) = P_X(1)\left|\mathbb{P}^{n-g}(\mathbb{F}_q)\right| = P_X(1)\left(\frac{q^{n-g+1}-1}{q-1}\right)$$

effective divisors of X of degree n. Note that the ζ-function $\zeta_X(t) = \frac{P_X(t)}{(1-t)(1-qt)}$ has residua $\mathrm{Res}_{\frac{1}{q}}(\zeta_X(t)) = \frac{P_X(\frac{1}{q})}{1-q}$, $\mathrm{Res}_1(\zeta_X(t)) = \frac{P_X(1)}{q-1}$ at its simple poles $\frac{1}{q}$,

respectively, 1. The ζ-polynomial $P_X(t)$ of X satisfies the functional equation $P_X(t) = P_X\left(\frac{1}{qt}\right) q^g t^{2g}$, according to Theorem 4.1.13 from [5] or to Theorem V.1.15 (b) from [2]. In particular, $P_X\left(\frac{1}{q}\right) = q^{-g} P_X(1)$ and

$$\mathcal{A}_n(X) = -q^{n+1}\mathrm{Res}_{\frac{1}{q}}(\zeta_X(t)) - \mathrm{Res}_1(\zeta_X(t)) \quad \forall n \geq 2g-1.$$

Definition 9. A locally finite module M over $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$ satisfies the Generic Riemann-Roch Conditions if M has

$$\mathcal{A}_n(M) = -q^{n+1}\mathrm{Res}_{\frac{1}{q}}(\zeta_M(t)) - \mathrm{Res}_1(\zeta_M(t))$$

effective divisors of degree n for sufficiently large natural numbers $n \geq n_o$.

One can compare the Generic Riemann-Roch Conditions with the Polarized Riemann-Roch Conditions from [6], which are shown to be equivalent to Mac Williams identities for linear codes over finite fields. A generalized version of [6], concerning additive codes will appear elsewhere.

Here is a characterization of the locally finite $\mathfrak{G}$-modules M with a polynomial ζ-quotient $P_M(t) = \frac{\zeta_M(t)}{\zeta_{\mathbb{P}^1(\overline{\mathbb{F}_q})}(t)} \in \mathbb{Z}[t]$.

Proposition 10. *The following conditions are equivalent for the ζ-function $\zeta_M(t)$ of a locally finite module M over $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$:*

(i) $P_M(t) := \frac{\zeta_M(t)}{\zeta_{\mathbb{P}^1(\overline{\mathbb{F}_q})}(t)} \in \mathbb{Z}[t]$ is a polynomial of $\deg P_M(t) = d \leq \delta \in \mathbb{N}$;

(ii) M satisfies the Generic Riemann-Roch Conditions

$$\mathcal{A}_n(M) = -q^{n+1}\mathrm{Res}_{\frac{1}{q}}(\zeta_M(t)) - \mathrm{Res}_1(\zeta_M(t)) = \frac{q^{n+1}P_M\left(\frac{1}{q}\right) - P_M(1)}{q-1} \tag{3.5}$$

for all $n \geq \delta - 1$;

$$\text{(iii)} \quad \left|\mathbb{P}^1(\overline{\mathbb{F}_q})^{\Phi_q^r}\right| - \left|M^{\Phi_q^r}\right| = \sum_{j=1}^{d} \omega_j^r \quad \text{for} \quad \forall r \in \mathbb{N} \tag{3.6}$$

and some $\omega_j \in \mathbb{C}^*$, which turn out to satisfy $P_M(t) = \prod_{j=1}^{d}(1-\omega_j t)$.

Proof. *(i) $\Rightarrow$ (ii)* If $P_M(t) = \frac{\zeta_M(t)}{\zeta_{\mathbb{P}^1(\overline{\mathbb{F}_q})}(t)} = \sum_{j=0}^{d} a_j t^j \in \mathbb{Z}[t]$ is a polynomial of $\deg P_M(t) = d \leq \delta \in \mathbb{N}$ then (3.4) reduces to

$$\mathcal{A}_m(M) = \sum_{i=0}^{d} a_i \left(\frac{q^{m-i+1}-1}{q-1}\right) = \frac{q^{m+1}P_M\left(\frac{1}{q}\right) - P_M(1)}{q-1} \quad \forall m \geq \delta.$$

Moreover, (3.4) implies that

$$\mathcal{A}_{\delta-1}(M)=\frac{q^\delta\left[P_M\left(\frac{1}{q}\right)-\frac{a_\delta}{q^\delta}\right]-[P_M(1)-a_\delta]}{q-1}=\frac{q^\delta P_M\left(\frac{1}{q}\right)-P_M(1)}{q-1}.$$

Now (3.5) follows from the fact that the residua of $\zeta_M(t)=\frac{P_M(t)}{(1-t)(1-qt)}$ at its simple poles are $\mathrm{Res}_{\frac{1}{q}}\left(\zeta_M(T)\right)=\frac{P_M\left(\frac{1}{q}\right)}{1-q}$, respectively, $\mathrm{Res}_1\left(\zeta_M(t)\right)=\frac{P_M(1)}{q-1}$.

$(ii)\Rightarrow(i)$ Plugging (3.5) in (3.3), one obtains $a_m(M)=0$, $\forall m\geq\delta+1$. Therefore $P_M(t)=\sum\limits_{i=0}^{\delta}a_i(M)t^i\in\mathbb{Z}[t]$ is a polynomial of degree $\deg P_M(t)\leq\delta$.

$(i)\Rightarrow(iii)$ If $P_M(t)=\frac{\zeta_M(t)}{\zeta_{\mathbb{P}^1(\overline{\mathbb{F}_q})}(t)}\in\mathbb{Z}[t]$ is a polynomial of degree $\deg P_M(t)=d\leq\delta$, then $P_M(0)=\frac{\zeta_M(0)}{\zeta_{\mathbb{P}^1(\overline{\mathbb{F}_q})}(0)}=1$ allows to express $P_M(t)=\prod\limits_{j=1}^{d}(1-\omega_jt)$ by some complex numbers $\omega_j\in\mathbb{C}^*$. According to Proposition 6,

$$\zeta_M(t)=\exp\left(\sum_{r=1}^{\infty}\left|M^{\Phi_q^r}\right|\frac{t^r}{r}\right)\quad\text{and}\quad\zeta_{\mathbb{P}^1(\overline{\mathbb{F}_q})}(t)=\exp\left(\sum_{r=1}^{\infty}\left|\mathbb{P}^1(\overline{\mathbb{F}_q})^{\Phi_q^r}\right|\frac{t^r}{r}\right),\ (3.7)$$

whereas

$$\sum_{j=1}^{d}\log(1-\omega_jt)=\log P_M(t)=\log\zeta_M(t)-\log\zeta_{\mathbb{P}^1(\overline{\mathbb{F}_q})}(t)=\sum_{r=1}^{\infty}\left(\left|M^{\Phi_q^r}\right|-\left|\mathbb{P}^1(\overline{\mathbb{F}_q})^{\Phi_q^r}\right|\right)\frac{t^r}{r}.$$

Making use of (3.1), one obtains $-\sum\limits_{r=1}^{\infty}\left(\sum\limits_{j=1}^{d}\omega_j^r\right)\frac{t^r}{r}=\sum\limits_{r=1}^{\infty}\left(\left|M^{\Phi_q^r}\right|-\left|\mathbb{P}^1(\overline{\mathbb{F}_q})^{\Phi_q^r}\right|\right)\frac{t^r}{r}$. The comparison of the coefficients of $\frac{t^r}{r}$, $\forall r\in\mathbb{N}$ provides (3.6).

$(iii)\Rightarrow(i)$ Multiplying (3.6) by $\frac{t^r}{r}$, summing $\forall r\in\mathbb{N}$ and making use of (3.1), one obtains $\log\zeta_{\mathbb{P}^1(\overline{\mathbb{F}_q})}(t)-\log\zeta_M(t)=-\sum\limits_{j=1}^{d}\log(1-\omega_jt)$. The change of the sign and an exponentiation provides $P_M(t)=\frac{\zeta_M(t)}{\zeta_{\mathbb{P}^1(\overline{\mathbb{F}_q})}(t)}=\prod\limits_{j=1}^{d}(1-\omega_jt)\in\mathbb{Z}[t]$. □

Corollary 11. *Let M and L be locally finite $\mathfrak{G}=\mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$-modules with polynomial ζ-quotients $P_M(t)=\frac{\zeta_M(t)}{\zeta_{\mathbb{P}^1(\overline{\mathbb{F}_q})}(t)}$, $P_L(t)=\frac{\zeta_L(t)}{\zeta_{\mathbb{P}^1(\overline{\mathbb{F}_q})}(t)}\in\mathbb{Z}[t]$ of degree $\deg P_M(t)\leq\delta$, $\deg P_L(t)\leq\delta$. Then M and L are isomorphic (as $\mathfrak{G}$-modules) if and only if they have one and the same number $B_k(M)=B_k(L)$ of $\mathfrak{G}$-orbits of degree k for all $1\leq k\leq\delta$.*

Proof. According to Corollary 7, it suffices to prove that $B_k(M)=B_k(L)$ for all $1\leq k\leq\delta$ is equivalent to the coincidence $\zeta_M(t)=\zeta_L(t)$ of the corresponding

ζ-functions. The infinite product expressions

$$\zeta_M(t) = \prod_{k=1}^{\infty} \left(\frac{1}{1-t^k}\right)^{B_K(M)}, \quad \zeta_L(t) = \prod_{k=1}^{\infty} \left(\frac{1}{1-t^k}\right)^{B_K(L)}$$

reveals that $\zeta_M(t) = \zeta_L(t)$ if and only if $B_k(M) = B_k(L)$, $\forall k \in \mathbb{N}$. There remains to be shown that if $\deg P_M(t) \leq \delta$ then $B_k(M)$ with $1 \leq k \leq \delta$ determine uniquely $B_k(M)$ for $\forall k \in \mathbb{N}$. Let $P_M(t) = \prod_{j=1}^{d}(1-\omega_j t)$ for some $d \leq \delta$, $\omega_j \in \mathbb{C}^*$ and denote $S_r := \sum_{j=1}^{d} \omega_j^r$, $\forall r \in \mathbb{N}$. By (3.6) from Proposition 10 and (3.2) from the proof of Proposition 6 one has

$$S_r = (q^r+1) - \left|M^{\Phi_q^r}\right| = (q^r+1) - \sum_{k/r} kB_k(M) \quad for \quad \forall r \in \mathbb{N}. \tag{3.8}$$

Thus $B_k(M)$ with $1 \leq k \leq \delta$ determine uniquely S_r, $\forall 1 \leq r \leq \delta$. Since $P_M(t)$ is of $\deg P_M(t) = d \leq \delta$, S_r with $1 \leq r \leq \delta$ determine uniquely S_r, $\forall r \in \mathbb{N}$ by Newton formulae. By an induction on $r \in \mathbb{N}$ and making use of (3.8), S_r with $r \in \mathbb{N}$ determine uniquely $B_r(M)$, $\forall r \in \mathbb{N}$. □

Proposition 12. *Let M be a locally finite module over the pro-finite completion $\mathfrak{G} = \widehat{\langle\varphi\rangle}$ of $\langle\varphi\rangle \simeq (\mathbb{Z},+)$ and M_r be the locally finite $\mathfrak{G}_r = \widehat{\langle\varphi^r\rangle}$-module, supported by M for some $r \in \mathbb{N}$. Then the ζ-functions of M and M_r are related by the equality*

$$\zeta_{M_r}(t^r) = \prod_{k=0}^{r-1} \zeta_M\left(e^{\frac{2\pi i k}{r}} t\right). \tag{3.9}$$

In particular, if M has polynomial ζ-quotient $P_M(t) = \frac{\zeta_M(t)}{\zeta_{\mathbb{P}^1(\overline{\mathbb{F}_q})}(t)} = \prod_{j=1}^{d}(1-\omega_j t)$ of $\deg P_M(t) = d$ then M_r has $P_{M_r}(t) := \frac{\zeta_{M_r}(t)}{\zeta_{\mathbb{P}^1(\overline{\mathbb{F}_q})_r}(t)} = \prod_{j=1}^{d}(1-\omega_j^r t)$ of $\deg P_{M_r}(t) = d$ and M satisfies the Riemann Hypothesis Analogue with respect to $\mathbb{P}^1(\overline{\mathbb{F}_q})$ as a $\mathfrak{G}$-module if and only if M_r satisfies the Riemann Hypothesis Analogue with respect to $\mathbb{P}^1(\overline{\mathbb{F}_q})_r$ as a $\mathfrak{G}_r$-module.

Proof. According to (1.7) from subsection V.1 of [2], for any $m, r \in \mathbb{N}$ with greatest common divisor $\mathrm{GCD}(m,r) = d \in \mathbb{N}$ there holds the equality of polynomials

$$\left(1-t^{r\frac{m}{d}}\right)^d = \prod_{k=0}^{r-1}\left[1-\left(e^{\frac{2\pi i k}{r}} t\right)^m\right].$$

By Lemma 4 (iii), any $\mathfrak{G}$-orbit ν of $\deg\nu = m$ splits in d orbits $\nu = \nu_1 \coprod \ldots \coprod \nu_d$ over $\mathfrak{G}_r$ of $\deg\nu_j = \frac{m}{d}$, $\forall 1 \leq j \leq d$. The contribution of ν to $\left[\prod_{k=0}^{r-1} \zeta_M\left(e^{\frac{2\pi i k}{r}} t\right)\right]^{-1}$ is

$\prod_{k=0}^{r-1}\left[1-\left(e^{\frac{2\pi ik}{r}}t\right)^m\right]=\left(1-t^r\frac{m}{d}\right)^d=\prod_{j=1}^{d}\left(1-t^{r\deg\nu_j}\right)$ and equals the contribution of $\nu_1\coprod\ldots\coprod\nu_d$ to $\zeta_{M_r}(t^r)^{-1}$. That justifies the equality of power series (3.9).

For any $\omega\in\mathbb{C}^*$ note that

$$\prod_{k=0}^{r-1}\left(1-e^{\frac{2\pi ik}{r}}\omega t\right)=(\omega t)^r\prod_{k=0}^{r-1}\left(\frac{1}{\omega t}-e^{\frac{2\pi ik}{r}}\right)=(\omega t)^r\left[\frac{1}{(\omega t)^r}-1\right]=1-\omega^r t^r. \quad (3.10)$$

If $P_M(t):=\frac{\zeta_M(t)}{\zeta_{\mathbb{P}^1(\overline{\mathbb{F}_q})}(t)}=\prod_{j=1}^{d}(1-\omega_j t)\in\mathbb{Z}[t]$ with $a_d:=\mathrm{LC}(P_M(t))=(-1)^d\omega_1\ldots\omega_d$ for some $\omega_j\in\mathbb{C}^*$ and $\mathbb{P}^1(\overline{\mathbb{F}_q})_r$ is the $\mathfrak{G}_r$-module, supported by $\mathbb{P}^1(\overline{\mathbb{F}_q})=\mathbb{P}^1(\overline{\mathbb{F}_{q^r}})$ then (3.9) and (3.10) yield

$$\begin{aligned}P_{M_r}(t^r)&=\frac{\zeta_{M_r}(t^r)}{\zeta_{\mathbb{P}^1(\overline{\mathbb{F}_q})_r}(t^r)}=\prod_{k=0}^{r-1}\frac{\zeta_M\left(e^{\frac{2\pi ik}{r}}t\right)}{\zeta_{\mathbb{P}^1(\overline{\mathbb{F}_q})}\left(e^{\frac{2\pi ik}{r}}t\right)}=\prod_{k=0}^{r-1}P_M\left(e^{\frac{2\pi ik}{r}}t\right)\\&=\prod_{k=0}^{r-1}\prod_{j=1}^{d}\left(1-\omega_j e^{\frac{2\pi ik}{r}}t\right)=\prod_{j=1}^{d}\prod_{k=0}^{r-1}\left(1-\omega_j e^{\frac{2\pi ik}{r}}t\right)=\prod_{j=1}^{d}\left(1-\omega_j^r t^r\right).\end{aligned}$$

Thus, $P_{M_r}(t)=\prod_{j=1}^{d}(1-\omega_j^r t)$ is a polynomial of $\deg P_{M_r}(t)=d\in\mathbb{N}$ with $|\mathrm{LC}(P_{M_r}(t))|$ $=|\omega_1\ldots\omega_d|^r=|a_d|^r$ and $|\omega_j|=\sqrt[d]{|a_d|}$ if and only if $\left|\omega_j^r\right|=\sqrt[d]{|\mathrm{LC}(P_{M_r}(t))|}$. That justifies the equivalence of the Riemann Hypothesis Analogue for M and M_r with respect to the projective line, whenever M has a polynomial ζ-quotient $P_M(t)$. □

4. FINITE UNRAMIFIED COVERING OF LOCALLY FINITE MODULES

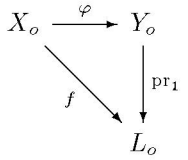
Extracting some properties of the finite unramified coverings $f:X\to Y$ of quasi-projective curves X,Y or topological spaces X,Y, we introduce the notion of a finite unramified covering of locally finite $\mathfrak{G}=\mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$-modules.

Definition 13. A surjective morphism $\xi:M\to L$ of $\mathfrak{G}=\mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$-modules is an unramified covering of degree $\deg\xi=k$ if all the fibres $\xi^{-1}(x)$, $x\in L$ of ξ are of one and a same cardinality $\left|\xi^{-1}(x)\right|=k$.

The inertia map $e_\xi:M\to\mathbb{N}$ of an unramified covering $\xi:M\to L$ of $\deg\xi=k$ takes values in $\{1,\ldots,k\}$. This follows from Proposition-Definition 5 (iii), according to which $\xi^{-1}(x)=\coprod_{y\in S_x}\mathrm{Orb}_{\mathfrak{G}_{\delta(x)}}(y)$, $\forall x\in M$, $\delta(x)=\deg\mathrm{Orb}_{\mathfrak{G}}(x)$, $\deg\mathrm{Orb}_{\mathfrak{G}_{\delta(x)}}(y)=e_\xi(y)$, whereas $k=\left|\xi^{-1}(x)\right|=\sum_{y\in S_x}e_\xi(y)$ with $e_\xi(y)\in\mathbb{N}$.

The next proposition establishes that an arbitrary irreducible quasi-projective curve $X \subset \mathbb{P}^n(\overline{\mathbb{F}_q})$ of genus $g \geq 1$ contains a locally finite $\mathfrak{G}_m = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_{q^m})$-submodule X_o with at most finite complement $X \setminus X_o$, which admits a finite unramified covering $f : X_o \to f(X_o)$ onto a $\mathfrak{G}_m$-submodule $f(X_o) \subseteq \mathbb{P}^1(\overline{\mathbb{F}_q})$ with $\left|\mathbb{P}^1(\overline{\mathbb{F}_q}) \setminus f(X_o)\right| < \infty$ for some $m \in \mathbb{N}$.

Proposition 14. *For any irreducible quasi-projective curve $X \subset \mathbb{P}^n(\overline{\mathbb{F}_q})$ of positive genus there exist $m \in \mathbb{N}$ and locally finite $\mathfrak{G}_m = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_{q^m})$-submodules $X_o \subseteq X \cap \overline{\mathbb{F}_q}^n \subset \mathbb{P}^n(\overline{\mathbb{F}_q})$, $L_o \subseteq \overline{\mathbb{F}_q} \subset \mathbb{P}^1(\overline{\mathbb{F}_q})$ with at most finite complements $X \setminus X_o$, $\mathbb{P}^1(\overline{\mathbb{F}_q}) \setminus L_o$, related by a finite unramified covering $f : X_o \to L_o$ of $\mathfrak{G}_m$-modules and quasi-affine curves, which induces the identical inclusion $f^* = \mathrm{Id} : \overline{\mathbb{F}_q}(L_o) = \overline{\mathbb{F}_q}(\mathbb{P}^1(\overline{\mathbb{F}_q})) \hookrightarrow \overline{\mathbb{F}_q}(X) = \overline{\mathbb{F}_q}(X_o)$ of the corresponding function fields. Moreover, there exist a plane quasi-affine curve $Y_o \subset \overline{\mathbb{F}_q}^2$, which is a locally finite $\mathfrak{G}_m$-module, as well as an isomorphism $\varphi : X_o \to Y_o$ of quasi-affine curves and $\mathfrak{G}_m$-modules, such that f factors through φ and the first canonical projection $\mathrm{pr}_1 : Y_o \to L_o$, $\mathrm{pr}_1(u_o, v_o) = u_o$, $\forall (u_o, v_o) \in Y_o$ along the commutative diagram*



Proof. According to Proposition 1 from 4 of Algebraic Preliminaries of [7], there exist such generators u, v of the function field $\overline{\mathbb{F}_q}(X) = \overline{\mathbb{F}_q}(u, v)$ of X over $\overline{\mathbb{F}_q}$ that u is transcendental over $\overline{\mathbb{F}_q}$ and v is separable over $\overline{\mathbb{F}_q}(u)$. If $\widetilde{g}(x) = \sum\limits_{i=0}^{k} \frac{\alpha_i(u)}{\beta_i(u)} x^i \in \overline{\mathbb{F}_q}(u)[x]$ with $\alpha_i(u), \beta_i(u) \in \overline{\mathbb{F}_q}[u]$, $\alpha_k(u) = \beta_k(u) \equiv 1$ is the minimal polynomial of v over $\overline{\mathbb{F}_q}(u)$ and $q(u) \in \overline{\mathbb{F}_q}[u]$ is a least common multiple of the denominators $\beta_i(u)$ of the coefficients of $\widetilde{g}(x)$ then

$$q(u)\widetilde{g}(x) = \sum_{i=0}^{k} \frac{q(u)\alpha_i(u)}{\beta_i(u)} x^i \in \overline{\mathbb{F}_q}[u, x]$$

is a polynomial in two variables u, x of positive degree $k := \deg_x(q(u)\widetilde{g}(x)) \in \mathbb{N}$ with respect to x. Dividing by the greatest common divisor of the coefficients $\frac{q(u)\alpha_i(u)}{\beta_i(u)} \in \overline{\mathbb{F}_q}[u]$, $0 \leq i \leq k$ of $q(u)\widetilde{g}(x)$, one obtains a primitive and therefore irreducible polynomial $g(u, x) \in \overline{\mathbb{F}_q}[u, x]$. The affine curve

$$Y := V(g(u, x)) = \{(u_o, v_o) \in \overline{\mathbb{F}_q}^2 \mid g(u_o, v_o) = 0\}$$

has function field $\overline{\mathbb{F}_q}(Y) = \overline{\mathbb{F}_q}(u, v) = \overline{\mathbb{F}_q}(X)$. That suffices for the existence of a birational map $\varphi : X \dashrightarrow Y$, inducing the identity $\varphi^* = \mathrm{Id} : \overline{Y} = \overline{\mathbb{F}_q}(u, v) \to \overline{\mathbb{F}_q}(u, v) = \overline{\mathbb{F}_q}(X)$ of $\overline{\mathbb{F}_q}$-algebras. In other words, there are quasi-affine curves

$X_1 \subseteq X$, $X_1 \subseteq \overline{\mathbb{F}_q}^n$, respectively, $Y_1 \subseteq Y \subset \overline{\mathbb{F}_q}^2$ with an isomorphism $\varphi : X_1 \to Y_1$ of quasi-affine varieties. For any $1 \le j \le 2$ let $\mathrm{pr}_j : \overline{\mathbb{F}_q}^2 \to \overline{\mathbb{F}_q}$, $\mathrm{pr}_j(x_1, x_2) = x_j$ be the canonical projection on the j-th component. Then $\varphi_j := \mathrm{pr}_j\varphi : X_1 \longrightarrow \overline{\mathbb{F}_q}$, $1 \le j \le 2$ are regular functions on X_1 and there are such polynomials $g_j(x_1, \ldots, x_n), h_j(x_1, \ldots, x_n) \in \overline{\mathbb{F}_q}[x_1, \ldots, x_n]$ that $\varphi_j\Big|_{X_1} = \frac{g_j(x_1,\ldots,x_n)}{h_j(x_1,\ldots,x_n)}\Big|_{X_1}$, after replacing X_1 by its sufficiently small Zariski open subset. The proper Zariski closed subvarieties of curves are finite sets of points, so that $|X \setminus X_1| < \infty$, $|Y \setminus Y_1| < \infty$. If $Y \setminus Y_1 = \{y_1, \ldots, y_s\}$ then $Y_2 := Y \setminus \mathrm{pr}_1^{-1}\{\mathrm{pr}_1(y_1), \ldots, \mathrm{pr}_1(y_s)\} \subseteq Y_1$ is a quasi-affine curve, on which the fibres $\mathrm{pr}_1^{-1}(u_o) = \{(u_o, v_o) \in \overline{\mathbb{F}_q}^2 \,|\, g(u_o, v_o) = 0\} \simeq \{v_o \in \overline{\mathbb{F}_q} \,|\, g(u_o, v_o) = 0\}$ of $\mathrm{pr}_1 : Y_2 \to \mathrm{pr}_1(Y_2)$ coincide with the corresponding fibres of $\mathrm{pr}_1 : Y \to \overline{\mathbb{F}_q}$ and are of cardinality $\left|\mathrm{pr}_1^{-1}(u_o)\right| \le k$. Note that $X_2 := \varphi^{-1}(Y_2)$ is a quasi-affine curve, $|X_1 \setminus X_2| < \infty$, $|Y_1 \setminus Y_2| < \infty$ and $\varphi : X_2 \to Y_2$ is an isomorphism of quasi-affine curves. The discriminant $D_x(g) \in \overline{\mathbb{F}_q}[u]$ of $g(u, x)$ with respect to x is a polynomial of u and has a finite set of zeroes $V(D_x(g)) \subset \mathrm{pr}_1(Y_2)$. All the fibres of

$$\mathrm{pr}_1 : Y_o = Y_2 \setminus \mathrm{pr}_1^{-1}(V(D_x(g))) \longrightarrow \overline{\mathbb{F}_q}$$

are of cardinality k and $\varphi : X_o = \varphi^{-1}(Y_o) \to Y_o$ is an isomorphism of quasi-affine varieties with $|X_1 \setminus X_o| < \infty$, $|Y_1 \setminus Y_o| < \infty$. If $X_o = V(g'_1, \ldots, g'_s) \setminus V(h'_1, \ldots, h'_r)$ consists of the common zeroes of the polynomials $g'_i(x_1, \ldots, x_n) \in \overline{\mathbb{F}_q}[x_1, \ldots, x_n]$, which are not a common zero of $h'_1(x_1, \ldots, x_n), \ldots, h'_r(x_1, \ldots, x_n) \in \overline{\mathbb{F}_q}[x_1, \ldots, x_n]$, then the minimal finite extension $\mathbb{F}_{q^\mu} \supseteq \mathbb{F}_q$, which contains the coefficients of all $g'_i(x_1, \ldots, x_n), h'_j(x_1, \ldots, x_n)$ is called the definition field of X_o. One sees immediately that for any $\mathbb{F}_{q^s} \supseteq \mathbb{F}_{q^\mu}$ the quasi-affine curve X_o is a locally finite $\mathfrak{G}_s = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_{q^s})$-module. The minimal finite extension $\mathbb{F}_{q^\nu} \supseteq \mathbb{F}_q$, containing the coefficients of the numerators $g_j(x_1, \ldots, x_n) \in \overline{\mathbb{F}_q}[x_1, \ldots, x_n]$ and the denominators $h_j(x_1, \ldots, x_n) \in \overline{\mathbb{F}_q}[x_1, \ldots, x_n]$ of the components φ_j of $\varphi = (\varphi_1, \varphi_2) : X_o \to Y_o \subset \overline{\mathbb{F}_q}^2$ is said to be the definition field of φ. We choose such $m \in \mathbb{N}$ that $\mathbb{F}_{q^m}$ contains the definition fields of X_0, Y_0, φ and observe that $\varphi : X_0 \to Y_0$ is an isomorphism of locally finite $\mathfrak{G}_m = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_{q^m})$-modules.

Moreover, $L_o := \mathrm{pr}_1(Y_o) \subseteq \overline{\mathbb{F}_q} \subset \mathbb{P}^1(\overline{\mathbb{F}_q})$ is a quasi-affine curve since $\left|\overline{\mathbb{F}_q} \setminus L_o\right| < \infty$ and $\mathrm{pr}_1 : Y_o \to L_o$ is an unramified covering of quasi-affine varieties. If $\mathbb{F}_{q^m}$ contains the definition field of L_o then $\mathrm{pr}_1 : Y_o \to L_o$ is a finite unramified covering of locally finite $\mathfrak{G}_m$-modules of degree k. We put $f := \mathrm{pr}_1\varphi : X_o \to L_o$ and note that under the aforementioned choices $f : X_o \to L_o$ is a finite unramified covering of locally finite $\mathfrak{G}_m$-modules and quasi-affine varieties, inducing the identical inclusion $f^* = \varphi^*\mathrm{pr}_1^* = \mathrm{pr}_1^* : \overline{\mathbb{F}_q}(L_o) = \overline{\mathbb{F}_q}(u) \hookrightarrow \overline{\mathbb{F}_q}(u, v) = \overline{\mathbb{F}_q}(X_o)$. □

An automorphism α of a $\mathfrak{G}$-module M is a self-isomorphism $\alpha : M \to M$ of $\mathfrak{G}$-modules. We denote by $\mathrm{Aut}_{\mathfrak{G}}(M)$ the automorphism group of M. Since $\mathfrak{G}$ is an abelian group, any $\varphi \in \mathfrak{G}$ induces an automorphism $\varphi : M \to M$. In such a way there arises a group homomorphism $\Psi : \mathfrak{G} \to \mathrm{Aut}_{\mathfrak{G}}(M)$. If Ψ is injective, the $\mathfrak{G}$-module M is said to be faithful and $\mathfrak{G}$ is identified with $\Psi(\mathfrak{G}) \le \mathrm{Aut}_{\mathfrak{G}}(M)$.

Lemma 15. *A locally finite module M over $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$ with closed stabilizers is faithful if and only if M is an infinite set.*

Proof. By the very definition of the homomorphism $\Psi : \mathfrak{G} \to \mathrm{Aut}_{\mathfrak{G}}(M)$, its kernel

$$\ker \Psi = \cap_{x \in M} \mathrm{Stab}_{\mathfrak{G}}(x)$$

is the intersection of the stabilizers of all the points of M. In the proof of Lemma 4 (iii) we have established that $\mathfrak{G}_m \cap \mathfrak{G}_n = \mathfrak{G}_{\mathrm{LCM}(m,n)}$. If $M = \{x_1, \ldots, x_r\}$ is a finite set then the map $\deg \mathrm{Orb}_{\mathfrak{G}} : M \to \mathbb{N}$ has finitely many values $m_1, \ldots, m_\nu$, $\nu \leq r$. As a result, $\ker \Psi = \cap_{j=1}^{\nu} \mathfrak{G}_{m_j} = \mathfrak{G}_{\mathrm{LCM}(m_j \,|\, 1 \leq j \leq \nu)} \neq \{0\}$ and M is not a faithful $\mathfrak{G}$-module.

Suppose that M is an infinite locally finite $\mathfrak{G}$-module and

$$\begin{aligned} \alpha = (\Phi_q^{l_s \pmod s})_{s \in \mathbb{N}} &\in \ker \Psi = \cap_{x \in M} \mathrm{Stab}_{\mathfrak{G}}(x) \\ &= \cap_{x \in M} \mathfrak{G}_{\deg \mathrm{Orb}_{\mathfrak{G}}(x)} = \cap_{x \in M} \left\{ \Phi_q^{\deg \mathrm{Orb}_{\mathfrak{G}}(x) m_s \pmod s} \right\}_{s \in \mathbb{N}}. \end{aligned}$$

Then for any point $x \in M$ and any $s \in \mathbb{N}$ the degree $\deg \mathrm{Orb}_{\mathfrak{G}}(x)$ of the $\mathfrak{G}$-orbit of x divides l_s. For an infinite locally finite $\mathfrak{G}$-module M the map $\deg \mathrm{Orb}_{\mathfrak{G}} : M \to \mathbb{N}$ has an infinite image, so that any l_s is divisible by infinitely many different natural numbers $\deg \mathrm{Orb}_{\mathfrak{G}}(x)$, $x \in M$. That implies $l_s = 0$, $\forall s \in \mathbb{N}$, whereas $\ker \Psi = \{0\}$. Thus, any infinite locally finite $\mathfrak{G}$-module M is faithful. □

Definition 16. If $\xi : M \to L$ is a finite unramified covering of locally finite $\mathfrak{G}$-modules then the fixed-point free automorphisms of $\mathfrak{G}$-modules $\alpha : M \to M$ with $\xi\alpha = \xi$ are called deck transformations of ξ.

Any subgroup H of $\mathrm{Aut}_{\mathfrak{G}}(M)$, which consists of deck transformations of $\xi : M \to L$ is called a deck transformation group of ξ.

Note that an automorphism $\alpha : M \to M$ of a locally finite $\mathfrak{G}$-module M and a finite unramified covering $\xi : M \to L$ of $\mathfrak{G}$-modules are subject to the equality $\xi\alpha = \xi$ if and only if α restricts to a bijection $\alpha : \xi^{-1}(x) \to \xi^{-1}(x)$ on any fibre $\xi^{-1}(x)$, $x \in L$ of ξ. Namely, $y \in \xi^{-1}(x)$ maps to $\alpha(y) \in \xi^{-1}(x)$ exactly when $\xi\alpha(y) = x = \xi(y)$. Thus, for any deck transformation group H of $\xi : M \to L$ and any point $x \in L$ there arises a group homomorphism

$$\Psi_x : H \to \mathrm{Sym}(\xi^{-1}(x)) = \mathrm{Sym}(k),$$

where $k = \deg(\xi)$. Due to the lack of fixed points of H, Ψ_x are injective and H is a finite group, whose orbits on $\xi^{-1}(x)$ are of one and a same cardinality $|H| \leq k!$. In particular, H acts transitively on some fibre $\xi^{-1}(x_o)$, $x_o \in L$ of a finite unramified covering $\xi : M \to L$ exactly when $|H| = k = \deg(\xi)$. If so, then H acts transitively on all the fibres $\xi^{-1}(x)$, $x \in L$ of ξ.

Definition 17. A finite unramified covering $\xi : M \to L$ of locally finite $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$-modules is H-Galois if there is a deck transformation group $H < \mathrm{Aut}_{\mathfrak{G}}(M)$, acting transitively on one and, therefore, on any fibre $\xi^{-1}(x)$, $x \in L$ of ξ.

Proposition 18. *In the notations from Proposition 14, the Galois group*

$$H = \mathrm{Gal}(\overline{\mathbb{F}_q}(X)/\overline{\mathbb{F}_q}(\mathbb{P}^1(\overline{\mathbb{F}_q})))$$

of the finite separable function fields extension $\overline{\mathbb{F}_q}(\mathbb{P}^1(\overline{\mathbb{F}_q})) \subset \overline{\mathbb{F}_q}(X)$ *is a deck transformation group of the finite unramified covering* $f = \mathrm{pr}_1\varphi : X_o \to L_o$ *of locally finite* $\mathfrak{G}_m = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_{q^m})$*-modules. If* $\overline{\mathbb{F}_q}(\mathbb{P}^1(\overline{\mathbb{F}_q})) \subset \overline{\mathbb{F}_q}(X)$ *is a Galois extension then* $f = \mathrm{pr}_1\varphi : X_o \to L_o$ *is an* H*-Galois covering. If* $f = \mathrm{pr}_1\varphi : X_o \to L_o$ *has a deck transformation group* H*, which consists of birational maps* $h : X_o \dashrightarrow X_o$ *and acts transitively on the fibres of* $f : X_o \to L_o$ *then the finite separable extension of function fields* $\overline{\mathbb{F}_q}(\mathbb{P}^1(\overline{\mathbb{F}_q})) \subset \overline{\mathbb{F}_q}(X)$ *is Galois and* $H \simeq \mathrm{Gal}(\overline{\mathbb{F}_q}(X)/\overline{\mathbb{F}_q}(\mathbb{P}^1(\overline{\mathbb{F}_q})))$.

Proof. As far as $\varphi : X_o \to Y_o$ is an isomorphism of locally finite $\mathfrak{G}_m$-modules, inducing the identity $\varphi^* = \mathrm{Id} : \overline{\mathbb{F}_q}(Y_o) = \overline{\mathbb{F}_q}(u, v) \to \overline{\mathbb{F}_q}(X_o) = \overline{\mathbb{F}_q}(X)$ of the corresponding function fields, it suffices to prove the corresponding statements for $\mathrm{pr}_1 : Y_o \to L_o$. More precisely, we claim that $H = \mathrm{Gal}(\overline{\mathbb{F}_q}(Y_o)/\overline{\mathbb{F}_q}(L_o))$ with $\overline{\mathbb{F}_q}(\mathbb{P}^1(\overline{\mathbb{F}_q})) = \overline{\mathbb{F}_q}(L_o) = \overline{\mathbb{F}_q}(u)$ is a deck transformation group of the finite unramified covering $\mathrm{pr}_1 : Y_o \to L_o$ of locally finite $\mathfrak{G}_m$-modules. If $\overline{\mathbb{F}_q}(u) \subset \overline{\mathbb{F}_q}(u, v)$ is a Galois extension then $\mathrm{pr}_1 : Y_o \to L_o$ is a Galois covering. If $\mathrm{pr}_1 : Y_o \to L_o$ has a deck transformation group H, which consists of birational maps $h : Y_o \dashrightarrow Y_o$ and acts transitively on the fibres of $\mathrm{pr}_1 : Y_o \to L_o$ then the finite separable extension $\overline{\mathbb{F}_q}(u) \subset \overline{\mathbb{F}_q}(u, v)$ of function fields is Galois.

Note that for any fixed $u_o \in L_o$ the Galois group $H = \mathrm{Gal}(\overline{\mathbb{F}_q}(u, v)/\overline{\mathbb{F}_q}(u))$ acts without fixed points on the fibre $\mathrm{pr}_1^{-1}(u_o) = \{(u_o, v_o) \in \overline{\mathbb{F}_q}^2 \mid g(u_o, v_o) = 0\}$ of the projection $\mathrm{pr}_1 : Y_o \to L_o$. That allows to view H as a fixed-point free subgroup of the symmetric group $\mathrm{Sym}(Y_o)$ of Y_o. If $\deg_x g(u, x) = k$ then $\overline{\mathbb{F}_q}(u, v)$ is a k-dimensional vector space over $\overline{\mathbb{F}_q}(u)$ with basis $1, v, \ldots, v^{k-1}$. The Frobenius automorphism $\Phi_{q^m} : \overline{\mathbb{F}_q}(u, v) \to \overline{\mathbb{F}_q}(u, v)$ acts on the coefficients of the rational functions $\frac{g_1(u)}{g_2(u)} \in \overline{\mathbb{F}_q}(u)$ with $g_1(u), g_2(u) \in \overline{\mathbb{F}_q}[u]$, $g_2(u) \not\equiv 0$ and fixes v^i for $\forall 0 \leq i \leq k-1$. By their very definition, all $h \in H = \mathrm{Gal}(\overline{\mathbb{F}_q}(u, v)/\overline{\mathbb{F}_q}(u))$ act identically on $\overline{\mathbb{F}_q}(u)$ and permute the roots $x_i \in \overline{\mathbb{F}_q}$ of $g(u, x) = 0$. That is why $h\Phi_{q^m} = \Phi_{q^m}h$ as an automorphism of the function field $\overline{\mathbb{F}_q}(u, v) = \overline{\mathbb{F}_q}(Y_o)$ and of the affine coordinate ring $\overline{\mathbb{F}_q}[Y_o] = \overline{\mathbb{F}_q}[u, x]/\langle g(u, x)\rangle = \overline{\mathbb{F}_q}[u, v] = \overline{\mathbb{F}_q}[u] + \overline{\mathbb{F}_q}[u]v + \ldots + \overline{\mathbb{F}_q}[u]v^{k-1}$ of Y_o. The affine closure $Y = V(g(u, x)) \subset \overline{\mathbb{F}_q}^2$ of Y_o in $\overline{\mathbb{F}_q}^2$ has the same affine coordinate ring $\overline{\mathbb{F}_q}[Y] = \overline{\mathbb{F}_q}[Y_o]$ as Y_o. The $\overline{\mathbb{F}_q}$-algebra automorphisms of $\overline{\mathbb{F}_q}[Y]$ are in a bijective correspondence with the automorphisms $Y \to Y$ of the affine curve Y, so that $h\Phi_{q^m} = \Phi_{q^m}h$ coincide as automorphisms of Y. By the very choice of $m \in \mathbb{N}$, the quasi-affine curve Y_o is Φ_{q^m}-invariant. According to $Y_o = Y \setminus \mathrm{pr}_1^{-1}\{u_1, \ldots, u_r\}$ for some $u_1, \ldots, u_r \in \overline{\mathbb{F}_q}$, the fibres of $\mathrm{pr}_1 : Y_o \to \mathrm{pr}_1(Y_o)$ coincide with the fibres of

$\mathrm{pr}_1 : Y \to \overline{\mathbb{F}_q}$ over $\mathrm{pr}_1(Y_o)$. Since h acts on the fibres of $\mathrm{pr}_1 : Y \to \overline{\mathbb{F}_q}$ without fixed points, the curve Y_o is preserved by h and $h\Phi_{q^m} = \Phi_{q^m} h$ coincide as automorphisms of Y_o. In such a way we have justified that H is a deck transformation group of the unramified covering $\mathrm{pr}_1 : Y_o \to L_o$ of $\mathfrak{G}_m$-modules.

If the finite separable extension $\overline{\mathbb{F}_q}(u) \subset \overline{\mathbb{F}_q}(u,v)$ is normal, i.e., Galois, then its Galois group $H = \mathrm{Gal}(\overline{\mathbb{F}_q}(u,v)/\overline{\mathbb{F}_q}(u))$ is of order $|H| = [\overline{\mathbb{F}_q}(u,v) : \overline{\mathbb{F}_q}(u)] = \deg_x g(u,x) = k = \deg(\mathrm{pr}_1)$. Therefore H acts transitively on the fibres of $\mathrm{pr}_1 : Y_o \to L_o$ and $\mathrm{pr}_1 : Y_o \to L_o$ is an H-Galois covering of locally finite $\mathfrak{G}_m$-modules.

Let H be a deck transformation group of $\mathrm{pr}_1 : Y_o \to L_o$, which consists of birational maps $h : Y_o \dashrightarrow Y_o$ and acts transitively on the fibres of pr_1. After replacing Y_o by a non-empty Zariski open subset $Y_1 \subseteq Y_o$, one can assume that all $h \in H$ are injective morphisms $h : Y_1 \to Y_o$. Any such $h = (h_1, h_2)$ is a pair of regular functions $h_i : Y_1 \to \overline{\mathbb{F}_q}$, $1 \le i \le 2$. The equality $\mathrm{pr}_1 h = \mathrm{pr}_1$, $\forall h = (h_1, h_2)$ is equivalent to $h_1(u,v) = u$, so that $h_1 = \mathrm{pr}_1$. Any birational map $h : Y_o \to Y_o$ induces an isomorphism $h^* : \overline{\mathbb{F}_q}(Y_o) = \overline{\mathbb{F}_q}(u,v) \to \overline{\mathbb{F}_q}(u,v) = \overline{\mathbb{F}_q}(Y_o)$ of $\overline{\mathbb{F}_q}$-algebras. According to $u = \mathrm{pr}_1(u,v)$ one has $h^*(u) = h^*(\mathrm{pr}_1)(u,v) = \mathrm{pr}_1 h(u,v) = h_1(u,v) = u$, $\forall h \in H$. Moreover, h^* acts identically on the constant field $\overline{\mathbb{F}_q}$ and, therefore, fixes any element of $\overline{\mathbb{F}_q}(u)$. That allows to view $h^* \in \mathrm{Gal}(\overline{\mathbb{F}_q}(u,v)/\overline{\mathbb{F}_q}(u))$ as an element of the Galois group of the finite separable extension $\overline{\mathbb{F}_q}(u) \subset \overline{\mathbb{F}_q}(u,v)$. The group H, acting transitively on the fibres of $\mathrm{pr}_1 : Y_o \to L_o$ is of order $|H| = \deg(\mathrm{pr}_1) = k = \deg_x g(u,x) = [\overline{\mathbb{F}_q}(u,v) : \overline{\mathbb{F}_q}(u)]$ and the extension $\overline{\mathbb{F}_q}(u) \subset \overline{\mathbb{F}_q}(u,v)$ is Galois. □

Note that, in general, if the finite coverings $\mathrm{pr}_1 : Y_o \to L_o$, $f = \mathrm{pr}_1 \varphi : X_o \to L_o$ of locally finite $\mathfrak{G}_m$-modules are H-Galois for some deck transformation group H of pr_1 and f then the finite separable extension $\overline{\mathbb{F}_q}(L_o) = \overline{\mathbb{F}_q}(u) \subset \overline{\mathbb{F}_q}(u,v) = \overline{\mathbb{F}_q}(Y_o) = \overline{\mathbb{F}_q}(X_o)$ is not supposed to be Galois. The reason is that the automorphisms $h \in H$ of the $\mathfrak{G}_m$-modules Y_o, X_o are not necessarily birational maps of Y_o, X_o.

Let $\xi : M \to L$ be a finite unramified covering of locally finite $\mathfrak{G}$-modules. Then any deck transformation group H of ξ is a finite fixed-point free subgroup of the automorphism group $\mathrm{Aut}_{\mathfrak{G}}(M)$ of M. The next lemma establishes that the orbit space $\mathrm{Orb}_H(M)$ of an arbitrary finite fixed-point free subgroup $H < \mathrm{Aut}_{\mathfrak{G}}(M)$ has natural structure of a locally finite $\mathfrak{G}$-module, with respect to which the map $\xi_H : M \to \mathrm{Orb}_H(M)$, $\xi_H(x) = \mathrm{Orb}_H(x)$, associating to a point $x \in M$ its H-orbit $\mathrm{Orb}_H(x)$ is an H-Galois covering.

Lemma 19. *Let M be an infinite locally finite $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$-module and H be a finite fixed-point free subgroup of $\mathrm{Aut}_{\mathfrak{G}}(M)$. Then:*

(i) the product $H\mathfrak{G} \simeq H \times \mathfrak{G}$ of the subgroups H and $\mathfrak{G}$ of $\mathrm{Aut}_{\mathfrak{G}}(M)$ is direct;

(ii) the set $\mathrm{Orb}_H(M) = \{\mathrm{Orb}_H(x) \,|\, x \in M\}$ of the H-orbits on M is a locally finite $\mathfrak{G}$-module with respect to the action

$$\begin{aligned} &\mathfrak{G} \times \mathrm{Orb}_H(M) \longrightarrow \mathrm{Orb}_H(M), \\ &(\varphi, \mathrm{Orb}_H(x)) \mapsto \varphi \mathrm{Orb}_H(x) = \mathrm{Orb}_H \varphi(x) \quad \forall \varphi \in \mathfrak{G}, \quad \forall x \in M; \end{aligned} \tag{4.1}$$

(iii) the correspondence

$$\xi_H : M \to \mathrm{Orb}_H(M), \quad \xi_H(x) = \mathrm{Orb}_H(x) \quad \forall x \in M$$

is a finite unramified H-Galois covering of degree $\deg \xi_H = |H|$.

Proof. (i) According to Lemma 15, the infinite locally finite $\mathfrak{G}$-module M is faithful and one can view $\mathfrak{G}$ as a subgroup of $\mathrm{Aut}_{\mathfrak{G}}(M)$. By its very definition, $\mathrm{Aut}_{\mathfrak{G}}(M)$ centralizes $\mathfrak{G}$. In particular, $h\varphi = \varphi h$, $\forall h \in H$ and $\forall \varphi \in \mathfrak{G}$. The isomorphism $\mathfrak{G} \simeq (\widehat{\mathbb{Z}}, +) \simeq \prod\limits_{\text{prime } p} (\widehat{\mathbb{Z}}_p, +)$ with the direct product of the additive groups $(\widehat{\mathbb{Z}}_p, +)$ of the p-adic integers reveals that any $\varphi \in \widehat{\mathbb{Z}} \setminus \{0\}$ is of infinite order. As far as any entry h of the finite group H is of finite order in $\mathrm{Aut}_{\mathfrak{G}}(M)$, there follows $H \cap \mathfrak{G} = \{\mathrm{Id}_M\}$ and the product $H\mathfrak{G} \simeq H \times \mathfrak{G}$ of subgroups of $\mathrm{Aut}_{\mathfrak{G}}(M)$ is direct.

(ii) Note that the map (4.1) is correctly defined, as far as $\forall x \in M$, $\forall \varphi \in \mathfrak{G}$, $\forall h \in H$ one has $\varphi \mathrm{Orb}_H(hx) = \mathrm{Orb}_H(\varphi h(x)) = \mathrm{Orb}_H(h\varphi(x)) = \mathrm{Orb}_H(\varphi(x)) = \varphi \mathrm{Orb}_H(x)$. The axioms for a $\mathfrak{G}$-action on $\mathrm{Orb}_H(M)$ follow from the ones for the $\mathfrak{G}$-action on M. Since H centralizes $\mathfrak{G}$ the $\mathfrak{G}$-orbits $\mathrm{Orb}_{\mathfrak{G}}\xi_H(x) = \mathrm{Orb}_{\mathfrak{G}}\mathrm{Orb}_H(x) = \mathrm{Orb}_H\mathrm{Orb}_{\mathfrak{G}}(x) = \xi_H\mathrm{Orb}_{\mathfrak{G}}(x)$ on $\mathrm{Orb}_H(M)$ are the images of the $\mathfrak{G}$-orbits on M under ξ_H, so that $\deg \mathrm{Orb}_{\mathfrak{G}}\xi_H(x) < \infty$, $\forall x \in M$. If $\deg \mathrm{Orb}_{\mathfrak{G}}\xi_H(x) = |\xi_H\mathrm{Orb}_{\mathfrak{G}}(x)| = m$ then the restriction $\xi_H|_{\mathrm{Orb}_{\mathfrak{G}}(x)} : \mathrm{Orb}_{\mathfrak{G}}(x) \to \mathrm{Orb}_{\mathfrak{G}}\xi_H(x)$ of $\xi_H : M \to \mathrm{Orb}_H(M)$ is of degree $\deg(\xi_H|_{\mathrm{Orb}_{\mathfrak{G}}(x)}) \leq \deg(\xi_H) = |H|$, so that

$$\deg \mathrm{Orb}_{\mathfrak{G}}(x) = \deg(\xi_H|_{\mathrm{Orb}_{\mathfrak{G}}(x)}) \deg \mathrm{Orb}_{\mathfrak{G}}\xi_H(x) \leq m|H|.$$

By assumption, the $\mathfrak{G}$-action on M is locally finite and there are finitely many $\mathfrak{G}$-orbits $\mathrm{Orb}_{\mathfrak{G}}(x)$ on M of degree $\leq m|H|$. Therefore, there are finitely many $\mathfrak{G}$-orbits $\mathrm{Orb}_{\mathfrak{G}}\xi_H(x)$ on $\mathrm{Orb}_H(M)$ of degree m and $\mathrm{Orb}_H(M)$ is a locally finite $\mathfrak{G}$-module.

(iii) The $\mathfrak{G}$-equivariance of ξ_H is an immediate consequence of the definition of the $\mathfrak{G}$-action on $\mathrm{Orb}_H(M)$ □

Let M be an infinite locally finite $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$-module. The next proposition describes the "twist" of the $\mathfrak{G}$-action on M by a fixed-point free automorphism $h \in \mathrm{Aut}_{\mathfrak{G}}(M)$ of finite order.

Proposition 20. *Let M be an infinite locally finite $\mathfrak{G} = \mathfrak{G}(\Phi_q) = \widehat{\langle \Phi_q \rangle}$-module with closed stabilizers, H be a finite fixed-point free subgroup of* $\mathrm{Aut}_{\mathfrak{G}}(M)$ *and $\varphi = h\Phi_q^r$ for some $h \in H$ and some natural number $r \in \mathbb{N}$. Then:*

(i) the pro-finite completion $\mathfrak{G}(\varphi) = \widehat{\langle \varphi \rangle}$ of the infinite cyclic group $\langle \varphi \rangle \simeq (\mathbb{Z}, +)$ is a subgroup of $H\mathfrak{G} \simeq H \times \mathfrak{G}$;

(ii) M is a locally finite $\mathfrak{G}(\varphi) = \widehat{\langle \varphi \rangle}$-module;

(iii) the second canonical projection $\mathrm{pr}_2 : H \times \mathfrak{G} \to \mathfrak{G}$, $\mathrm{pr}_2(h', \gamma) = \gamma$, $\forall h' \in H$, $\forall \gamma \in \mathfrak{G}$ *provides a locally finite* $\mathfrak{G}(\varphi)$*-action*

$$\mathfrak{G}(\varphi) \times \mathrm{Orb}_H(M) \longrightarrow \mathrm{Orb}_H(M),$$
$$(\gamma, \mathrm{Orb}_H(x)) \mapsto \mathrm{pr}_2(\gamma)\mathrm{Orb}_H(x) = \mathrm{Orb}_H(\mathrm{pr}_2(\gamma)x);$$

(iv) the map

$$\xi_H : M \longrightarrow \mathrm{Orb}_H(M), \quad \xi_H(x) = \mathrm{Orb}_H(x) \quad \forall x \in M$$

is an H*-Galois covering of locally finite* $\mathfrak{G}(\varphi)$*-modules.*

Proof. (i) First of all, $\varphi = h\Phi_q^r$ is of infinite order. Otherwise, for h of order m and φ of order l, one has $\mathrm{Id}_N = \varphi^{ml} = h^m\Phi_q^{rml} = \Phi_q^{rml}$ and the Frobenius automorphism $\Phi_q : M \to M$ turns to be of finite order. This is an absurd, justifying $\langle\varphi\rangle \simeq (\mathbb{Z}, +)$. Note that $\varphi = h\Phi_q^r \in H\mathfrak{G}$ suffices for $\langle\varphi\rangle$ to be a subgroup of the compact group $H\mathfrak{G}$. The pro-finite completion $\mathfrak{G}(\varphi) = \widehat{\langle\varphi\rangle}$ is the closure of $\langle\varphi\rangle$ with respect to the discrete topology, so that $\mathfrak{G}(\varphi) = \widehat{\langle\varphi\rangle} \leq H\mathfrak{G}$ since $H\mathfrak{G}$ is closed with respect to the discrete topology.

(ii) In order to show that all the $\mathfrak{G}(\varphi)$-orbits on M are of finite degree, let us consider a point $x \in M$ with $\deg \mathrm{Orb}_{\mathfrak{G}}(x) = \delta$. If $h \in H < \mathrm{Aut}_{\mathfrak{G}}(M)$ is of order m then

$$\mathfrak{G}(\varphi^{m\delta}) := \widehat{\langle\varphi^{m\delta}\rangle} = \widehat{\langle\Phi_q^{mr\delta}\rangle} = \mathfrak{G}(\Phi_q^{mr\delta}) \leq \mathfrak{G}(\Phi_q^{\delta}) = \mathrm{Stab}_{\mathfrak{G}}(x) \leq \mathrm{Stab}_{H\times\mathfrak{G}}(x),$$

whereas $\mathfrak{G}(\varphi^{m\delta}) \leq \mathfrak{G}(\varphi) \cap \mathrm{Stab}_{H\times\mathfrak{G}}(x) = \mathrm{Stab}_{\mathfrak{G}(\varphi)}(x) \leq \mathfrak{G}(\varphi)$. Therefore

$$\deg \mathrm{Orb}_{\mathfrak{G}(\varphi)}(x) = [\mathfrak{G}(\varphi) : \mathrm{Stab}_{\mathfrak{G}(\varphi)}(x)] = \frac{[\mathfrak{G}(\varphi) : \mathfrak{G}(\varphi^{m\delta})]}{[\mathrm{Stab}_{\mathfrak{G}(\varphi)}(x) : \mathfrak{G}(\varphi^{m\delta})]}$$
$$= \frac{m\delta}{[\mathrm{Stab}_{\mathfrak{G}(\varphi)}(x) : \mathfrak{G}(\varphi^{m\delta})]} \in \mathbb{N}$$

and all the $\mathfrak{G}(\varphi)$-orbits on M are finite. Let $n \in \mathbb{N}$ and $y \in M$ be a point with $\deg \mathrm{Orb}_{\mathfrak{G}(\varphi)}(y) = n$ or, equivalently, with $\mathrm{Stab}_{\mathfrak{G}(\varphi)}(y) = \mathfrak{G}(\varphi^n)$. If $\delta := \deg \mathrm{Orb}_{\mathfrak{G}}(y)$ and $h \in H < \mathrm{Aut}_{\mathfrak{G}}(M)$ is of order m then

$$\mathfrak{G}(\varphi^{nm}) = \mathfrak{G}(\Phi_q^{nmr}) < \mathfrak{G} \cap \mathrm{Stab}_{H\times\mathfrak{G}}(y) = \mathrm{Stab}_{\mathfrak{G}}(x) = \mathfrak{G}(\Phi_q^{\delta}).$$

Therefore δ is a natural divisor of nmr. By assumption, M contains finitely many $\mathfrak{G}$-orbits of degree δ. For any fixed $n \in \mathbb{N}$ there are finitely many natural divisors δ of nmr and, therefore, finitely many $\mathfrak{G}(\varphi)$-orbits on M of degree n. In such a way we have checked that the $\mathfrak{G}(\varphi)$-action on M is locally finite.

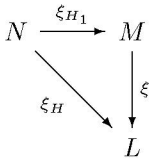
(iii) is an immediate consequence of Lemma 19 (ii).

(iv) Towards the $\mathfrak{G}(\varphi)$-equivariance of $\xi_H : M \to \mathrm{Orb}_H(M)$, $\xi_H(x) = \mathrm{Orb}_H(x)$, $\forall x \in M$, let us consider the first canonical projection $\mathrm{pr}_1 : H\times\mathfrak{G} \to \mathfrak{G}$, $\mathrm{pr}_1(h', \gamma) = h'$,

$\forall h' \in H$, $\forall \gamma \in \mathfrak{G}$. An arbitrary $\rho \in \mathfrak{G}(\varphi) < H\mathfrak{G} \simeq H \times \mathfrak{G}$ has an unique factorization $\rho = \mathrm{pr}_1(\rho)\mathrm{pr}_2(\rho)$ into a product of $\mathrm{pr}_1(\rho) \in H$ and $\mathrm{pr}_2(\rho) \in \mathfrak{G}$. Then $\xi_H(\rho x) = \xi_H(\mathrm{pr}_1(\rho)\mathrm{pr}_2(\rho)x) = \xi_H(\mathrm{pr}_2(\rho)x) = \mathrm{pr}_2(\rho)\xi_H(x)$, $\forall x \in M$ verifies that ξ_H is an H-Galois covering of locally finite $\mathfrak{G}(\varphi)$-modules. $\square$

From now on, we identify the isomorphic locally finite $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$-modules, in order to avoid cumbersome notations.

Definition 21. A Galois closure of a finite unramified covering $\xi : M \to L$ of locally finite $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$-modules is a triple (N, H, H_1), which consists of a locally finite $\mathfrak{G}_m = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_{q^m})$-module N for some $m \in \mathbb{N}$, a finite fixed-point free subgroup H of $\mathrm{Aut}_{\mathfrak{G}_m}(N)$ and a subgroup H_1 of H, such that $\mathrm{Orb}_{H_1}(N)$ is isomorphic to M as a $\mathfrak{G}_m$-module, $\mathrm{Orb}_H(N)$ is isomorphic to L as a $\mathfrak{G}_m$-module and the H-Galois covering $\xi_H : N \to L$, $\xi_H(x) = \mathrm{Orb}_H(x)$, $\forall x \in N$ factors through the H_1-Galois covering $\xi_{H_1} : N \to M$, $\xi_{H_1}(x) = \mathrm{Orb}_{H_1}(x)$, $\forall x \in N$ and ξ along a commutative diagram



of finite unramified coverings of $\mathfrak{G}_m$-modules.

We say that (N, H, H_1) is defined over $\mathbb{F}_{q^m}$.

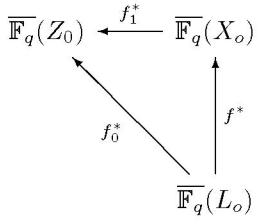
Proposition 22. *For any irreducible quasi-projective curve X of positive genus over $\overline{\mathbb{F}_q}$ there exist $s \in \mathbb{N}$, locally finite $\mathfrak{G}_s = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_{q^s})$-submodules $X' \subseteq X$, $L \subseteq \mathbb{P}^1(\overline{\mathbb{F}_q})$ with at most finite complements $X \setminus X'$, $\mathbb{P}^1(\overline{\mathbb{F}_q}) \setminus L$, a finite unramified covering $f : X' \to L$ of $\mathfrak{G}_s$-modules and a Galois closure (Z, H, H_1) of f, such that Z is an irreducible quasi-projective curve $Z \subseteq \mathbb{P}^r(\overline{\mathbb{F}_q})$, $H = \mathrm{Gal}(\overline{\mathbb{F}_q}(Z)/\overline{\mathbb{F}_q}(\mathbb{P}^1(\overline{\mathbb{F}_q})))$, $H_1 = \mathrm{Gal}(\overline{\mathbb{F}_q}(Z)/\overline{\mathbb{F}_q}(X))$.*

Proof. Let $f : X_o \to L_o$ be the finite unramified covering of locally finite $\mathfrak{G}_m$-modules from Proposition 14. The finite separable extension

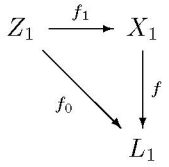
$$\overline{\mathbb{F}_q}(X) = \overline{\mathbb{F}_q}(X_o) = \overline{\mathbb{F}_q}(u, v) \supset \overline{\mathbb{F}_q}(u) = \overline{\mathbb{F}_q}(L_o) = \overline{\mathbb{F}_q}(\mathbb{P}^1(\overline{\mathbb{F}_q}))$$

of the corresponding function fields admits a Galois closure $K \supseteq \overline{\mathbb{F}_q}(u, v) \supseteq \overline{\mathbb{F}_q}(u)$ of finite degree $[K : \overline{\mathbb{F}_q}(u)] < \infty$, i.e., K is normal over $\overline{\mathbb{F}_q}(u)$ and $\overline{\mathbb{F}_q}(u, v)$. Then there is an irreducible quasi-projective curve $Z_0 \subset \mathbb{P}^r(\overline{\mathbb{F}_q})$ with function field $\overline{\mathbb{F}_q}(Z_0) = K$ and dominant rational maps $f_0 : Z_0 \dashrightarrow L_0$, $f_1 : Z_0 \dashrightarrow X_0$, inducing the identical inclusions $f_0^* = \mathrm{Id} : \overline{\mathbb{F}_q}(L_o) = \overline{\mathbb{F}_q}(u) \hookrightarrow \overline{\mathbb{F}_q}(Z_0)$, respectively, $f_1^* = \mathrm{Id} : \overline{\mathbb{F}_q}(X_o) = \overline{\mathbb{F}_q}(u, v) \hookrightarrow \overline{\mathbb{F}_q}(Z_0)$ of the associated function fields. Bearing in mind that the finite covering $f : X_o \to L_o$ induces the identity $f^* = \mathrm{Id} : \overline{\mathbb{F}_q}(L_o) = \overline{\mathbb{F}_q}(u) \hookrightarrow \overline{\mathbb{F}_q}(u, v) =$

$\overline{\mathbb{F}_q}(X_o)$, one obtains a commutative diagram



of identical inclusions of function fields over $\overline{\mathbb{F}_q}$. Therefore, the composition ff_1 coincides with the dominant rational map f_0. Let $Z_1' \subset \overline{\mathbb{F}_q}^r$ be a quasi-affine curve, contained in the regularity domains of f_0 and f_1. Then $f_0 : Z_1' \to f_0(Z_1')$ is a finite covering of affine curves. Removing from Z_1' the branch locus of $f_0|_{Z_1'}$, one obtains a quasi-affine curve $Z_1'' \subseteq Z_1' \subseteq Z_0$. The finite set $Z_0 \setminus Z_1''$ has finite image $f(Z_0 \setminus Z_1'')$, so that $L_1 := L_o \setminus f_0(Z_0 \setminus Z_1'')$, $X_1 := f^{-1}(L_1)$, $Z_1 := f_0^{-1}(L_1) = f_1^{-1}(X_1) \subseteq Z_1''$ are quasi-affine curves, subject to a commutative diagram



of finite unramified coverings of quasi-affine curves. In particular, $Z_0 \setminus Z_1$, $X_o \setminus X_1$, $L_o \setminus L_1$ are finite sets.

The normal separable extension $\overline{\mathbb{F}_q}(L_o) \subset \overline{\mathbb{F}_q}(Z_0)$ is finite, so that its Galois group $H := \mathrm{Gal}(\overline{\mathbb{F}_q}(Z_0)/\overline{\mathbb{F}_q}(L_o)) = \mathrm{Gal}(\overline{\mathbb{F}_q}(Z_1)/\overline{\mathbb{F}_q}(L_1))$ is finite. Any $h \in H$ transforms the affine coordinates z_j, $1 \leq j \leq r$ on $Z_1 \subset \overline{\mathbb{F}_q}^r$ to rational functions $h(z_j) \in \overline{\mathbb{F}_q}(Z_1)$. Let Z_2' be the intersection of the regularity domains of $h(z_j) : Z_1 \dashrightarrow \overline{\mathbb{F}_q}$, $\forall h \in H$ and $\forall 1 \leq j \leq r$. Then for any $h \in H$ the map

$$\widetilde{h} : Z_2' \longrightarrow \widetilde{h}(Z_2') \subseteq Z_1 \subset \overline{\mathbb{F}_q}^r,$$

$$\widetilde{h}(u_1, \ldots, u_r) := (h(z_1)(u_1), \ldots, h(z_r)(u_r)) \quad \forall u = (u_1, \ldots, u_r) \in Z_2'$$

is a morphism of quasi-affine varieties. Since H is a finite group, $Z_2'' := \cap_{h \in H} \widetilde{h}(Z_2')$ is a quasi-affine curve, so that $|Z_2' \setminus Z_2''| < \infty$. Moreover, $\forall u \in Z_2''$ and $\forall h_o, h \in H$ one has $u \in \widetilde{h_o}^{-1}\widetilde{h}(Z_2')$, whereas $\widetilde{h_o}(u) \in \widetilde{h}(Z_2')$. Thus, $\widetilde{h_o}(u) \in \cap_{h \in H} \widetilde{h}(Z_2') = Z_2''$, $\forall u \in Z_2''$, $\forall h_o \in H$ and Z_2'' is H-invariant. Note that for any $h \in H$ the equation $\widetilde{h}(u) = u$ has at most finitely many solutions on Z_2''. Therefore H has at most finitely many fixed points on Z_2''. After removing the H-orbits of the H-fixed points on Z_2'', one obtains a quasi-affine curve $Z_2 \subseteq Z_2''$, acted by H without fixed points.

By the very construction of Z_0, the function fields extensions

$$\overline{\mathbb{F}_q}(Z_0) = \overline{\mathbb{F}_q}(Z_1) = \overline{\mathbb{F}_q}(Z_2) \supseteq \overline{\mathbb{F}_q}(X_1) = \overline{\mathbb{F}_q}(X_o) \quad \text{and}$$

$$\overline{\mathbb{F}_q}(Z_0) = \overline{\mathbb{F}_q}(Z_1) = \overline{\mathbb{F}_q}(Z_2) \supset \overline{\mathbb{F}_q}(L_1) = \overline{\mathbb{F}_q}(L_o)$$

are Galois. Therefore the Galois groups

$$H = \mathrm{Gal}(\overline{\mathbb{F}_q}(Z_2)/\overline{\mathbb{F}_q}(L_1)) \quad \text{and} \quad H_1 := \mathrm{Gal}(\overline{\mathbb{F}_q}(Z_0)/\overline{\mathbb{F}_q}(X_o)) = \mathrm{Gal}(\overline{\mathbb{F}_q}(Z_2)/\overline{\mathbb{F}_q}(X_1))$$

have invariant fields $\overline{\mathbb{F}_q}(Z_2)^H = \overline{\mathbb{F}_q}(L_1)$, respectively, $\overline{\mathbb{F}_q}(Z_2)^{H_1} = \overline{\mathbb{F}_q}(X_1)$. The correspondence

$$f_H : Z_2 \longrightarrow \mathrm{Orb}_H(Z_2) = Z_2/H, \quad f_H(z) = \mathrm{Orb}_H(z) \quad \forall z \in Z_2,$$

associating to $z \in Z_2$ its H-orbit is a surjective morphism of algebraic curves, which induces an isomorphism $f_H^* : \overline{\mathbb{F}_q}(Z_2/H) \to \overline{\mathbb{F}_q}(Z_2)^H = \overline{\mathbb{F}_q}(L_1)$ of $\overline{\mathbb{F}_q}$-algebras. Therefore there is a birational map $\varphi_0 : L_1 \dashrightarrow Z_2/H$ with $\varphi_0^* = f_H^*$. Similarly,

$$f_{H_1} : Z_2 \longrightarrow \mathrm{Orb}_{H_1}(Z_2) = Z_2/H_1, \quad f_{H_1}(z) = \mathrm{Orb}_{H_1}(z) \quad \forall z \in Z_2$$

is a surjective morphism of algebraic curves, inducing an isomorphism of $\overline{\mathbb{F}_q}$-algebras $f_{H_1}^* : \overline{\mathbb{F}_q}(Z_2/H_1) \to \overline{\mathbb{F}_q}(Z_2)^{H_1} = \overline{\mathbb{F}_q}(X_1)$. Let $\varphi_1 : X_1 \dashrightarrow Z_2/H_1$ be the birational map with $\varphi_1^* = f_{H_1}^*$. The commutative diagrams

$$\begin{array}{ccc} \overline{\mathbb{F}_q}(Z_2) & & \\ {\scriptstyle \mathrm{Id}}\uparrow & \nwarrow^{f_H^*} & \\ \overline{\mathbb{F}_q}(L_1) & \xleftarrow{\varphi_0^*} & \overline{\mathbb{F}_q}(Z_2/H) \end{array} \quad , \quad \begin{array}{ccc} \overline{\mathbb{F}_q}(Z_2) & & \\ {\scriptstyle \mathrm{Id}}\uparrow & \nwarrow^{f_{H_1}^*} & \\ \overline{\mathbb{F}_q}(X_1) & \xleftarrow{\varphi_1^*} & \overline{\mathbb{F}_q}(Z_2/H_1) \end{array}$$

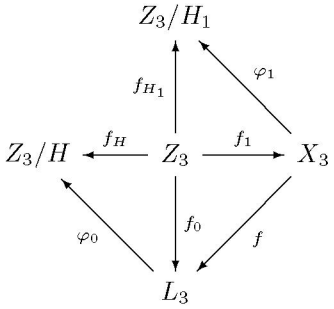
of embeddings Id, f_H^*, $f_{H_1}^*$ of $\overline{\mathbb{F}_q}$-algebras and isomorphisms φ_0^*, φ_1^* of $\overline{\mathbb{F}_q}$-algebras induce commutative diagrams

$$\begin{array}{ccc} Z_2 & & \\ {\scriptstyle f_0}\downarrow & \searrow^{f_H} & \\ L_1 & \xrightarrow{\varphi_0} & Z_2/H \end{array} \quad , \quad \begin{array}{ccc} Z_2 & & \\ {\scriptstyle f_1}\downarrow & \searrow^{f_{H_1}} & \\ X_1 & \xrightarrow{\varphi_1} & Z_2/H_1 \end{array}$$

of morphisms f_0, f_H, f_1, f_{H_1} and birational maps φ_0, φ_1.

There is a quasi-affine curve $L_2' \subseteq L_1$, such that $\varphi_0 : L_1 \dashrightarrow Z_2/H$ restricts to an isomorphism $\varphi_0 : L_2' \to \varphi_0(L_2') \subseteq Z_2/H$ of algebraic varieties. Similarly, one can choose a quasi-affine curve $X_2' \subseteq X_1$, such that $\varphi_1 : X_2' \to \varphi_1(X_2') \subseteq Z_2/H_1$ is an isomorphism of algebraic curves. Since $L_1 \setminus L_2'$ and $X_1 \setminus X_2'$ are finite sets and

$f_0 : Z_1 \to L_1$, $f_1 : Z_1 \to X_1$ are finite coverings, $S := f_0^{-1}(L_1 \setminus L_2') \cup f_1^{-1}(X_1 \setminus X_2')$ is a finite subset of Z_2. Removing from Z_2 the H-orbit of S, one obtains a quasi-affine curve $Z_3 \subseteq Z_2$, acted by H without fixed points. The factorization $f_H|_{Z_3} = \varphi_0 f_0|_{Z_3}$ with a biregular $\varphi_0 : f_0(Z_3) \to f_H(Z_3)$ implies the coincidence of the fibres of f_H and f_0. Therefore, $f_H : Z_3 \to f_H(Z_3)$ and $f_0 : Z_3 \to L_3 := f_0(Z_3)$ are finite unramified coverings of algebraic curves of degree $|H|$. Similarly, $f_{H_1}|_{Z_3} = \varphi_1 f_1|_{Z_3}$ with biregular $\varphi_1 : f_1(Z_3) \to f_{H_1}(Z_3)$ reveals that $f_{H_1} : Z_3 \to f_{H_1}(Z_3)$ and $f_1 : Z_3 \to X_3 := f_1(Z_3)$ are finite unramified coverings of algebraic curves of degree $|H_1|$. There exists a sufficiently large $s \in \mathbb{N}$, such that $\mathbb{F}_{q^s}$ contains the definition fields of the curves Z_3, X_3, L_3, Z_3/H, Z_3/H_1, as well as the coefficients of the components of the regular maps f, f_0, f_1, f_H, f_{H_1}. Then



turns out to be a commutative diagram of finite unramified coverings of locally finite $\mathfrak{G}_s$-modules with bijective φ_0, φ_1, H-Galois covering f_H, H_1-Galois covering f_{H_1}. Introducing $Z := Z_3$, $X' := X_3$, $L := L_3$, one concludes that (Z, H, H_1) is a Galois closure of the finite unramified covering $f : X' \to L$. □

5. RIEMANN HYPOTHESIS ANALOGUE FOR LOCALLY FINITE MODULES

The next proposition provides a numerical necessary and sufficient condition for a locally finite $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$-module with a polynomial ζ-quotient to satisfy the Riemann Hypothesis Analogue with respect to the projective line $\mathbb{P}^1(\overline{\mathbb{F}_q})$.

Proposition 23. *The following conditions are equivalent for a locally finite module M over $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$ with a polynomial ζ-quotient $P_M(t) = \frac{\zeta_M(t)}{\zeta_{\mathbb{P}^1(\overline{\mathbb{F}_q})}(t)} \in \mathbb{Z}[t]$ of $deg P_M(t) = d \in \mathbb{N}$ with leading coefficient $\mathrm{LC}(P_M(t)) = a_d \in \mathbb{Z} \setminus \{0\}$ and for $\lambda := \log_q \sqrt[d]{|a_d|} \in \mathbb{R}^{\geq 0}$:*

(i) M satisfies the Riemann Hypothesis Analogue with respect to $\mathbb{P}^1(\overline{\mathbb{F}_q})$ as a $\mathfrak{G}$-module;

(ii) $q^r + 1 - dq^{\lambda r} \leq |M^{\Phi_q^r}| \leq q^r + 1 + dq^{\lambda r}$, $\forall r \in \mathbb{N}$;

(iii) there exist constants $C_1, C_2 \in \mathbb{R}^{>0}$, $\nu, r_1, r_2 \in \mathbb{N}$, *such that*

$$|M^{\Phi_q^{\nu r}}| \leq q^{\nu r} + 1 + C_1 q^{\lambda \nu r} \quad \forall r \in \mathbb{N}, \quad r \geq r_1 \quad \textit{and}$$

$$|M^{\Phi_q^{\nu r}}| \geq q^{\nu r} + 1 - C_2 q^{\lambda \nu r} \quad \forall r \in \mathbb{N}, \quad r \geq r_2.$$

Proof. $(i) \Rightarrow (ii)$ If $P_M(t) = \prod_{j=1}^{d}(1 - q^{\lambda} e^{i\varphi_j} t)$ for some $\varphi_j \in [0, 2\pi)$ then

$$\left|\mathbb{P}^1(\overline{\mathbb{F}_q})^{\Phi_q^r}\right| - \left|M^{\Phi_q^r}\right| = \sum_{j=1}^{d} q^{\lambda r} e^{ir\varphi_j} \quad for \quad \forall r \in \mathbb{N}$$

by (3.6) from Proposition 10. Therefore,

$$\left|\left|M^{\Phi_q^r}\right| - (q^r + 1)\right| = \left|\sum_{j=1}^{d} q^{\lambda r} e^{ir\varphi_j}\right| \leq \sum_{j=1}^{d} \left|q^{\lambda r} e^{ir\varphi_j}\right| = \sum_{j=1}^{d} q^{\lambda r} = dq^{\lambda r},$$

hence (ii) holds.

$(ii) \Rightarrow (iii)$ is trivial

$(iii) \Rightarrow (i)$ Let $P_M(t) = \prod_{j=1}^{d}(1 - \omega_j t) \in \mathbb{Z}[t]$. The formal power series

$$H(t) := \sum_{j=1}^{d} \frac{\omega_j^{\nu} t}{1 - \omega_j^{\nu} t}$$

has radius of convergence $\rho = \min\left(\frac{1}{|\omega_1|^{\nu}}, \ldots, \frac{1}{|\omega_d|^{\nu}}\right)$, i.e., $H(t) < \infty$ converges $\forall t \in \mathbb{C}$ with $|t| < \rho$ and $H(t) = \infty$ diverges $\forall t \in \mathbb{C}$ with $|t| > \rho$. Making use of the formal series expansion $\frac{1}{1-\omega_j^{\nu} t} = \sum_{i=0}^{\infty} \omega_j^{\nu i} t^i$ and exchanging the summation order, one represents

$$H(t) = \sum_{i=0}^{\infty} \Big(\sum_{j=1}^{d} \omega_j^{\nu(i+1)} \Big) t^{i+1}.$$

Let $C := \max(C_1, C_2)$, $r_0 := \max(r_1, r_2)$ and note that assumption (iii) implies that

$$\left|\sum_{j=1}^{d} \omega_j^{\nu r}\right| = \left|\left|M^{\Phi_q^{\nu r}}\right| - (q^{\nu r} + 1)\right| \leq C q^{\lambda \nu r} \quad \forall r \in \mathbb{N}, \quad r \geq r_0,$$

according to (3.6) from Proposition 10. Thus, $\left|\sum_{j=1}^{d} \omega_j^{\nu(i+1)}\right| \leq C q^{\lambda \nu (i+1)}$, $\forall i \in \mathbb{Z}$, $i \geq r_0 - 1$ and

$$|H(t)| \leq \sum_{i=0}^{\infty} \left|\sum_{j=1}^{d} \omega_j^{\nu(i+1)}\right| t^{i+1} \leq C \sum_{i=0}^{\infty} q^{\lambda \nu (i+1)} t^{i+1} = C \sum_{i=0}^{\infty} (q^{\lambda \nu} t)^{i+1}.$$

As a result, $H(t) < \infty$, $\forall t \in \mathbb{C}$ with $|t| < \frac{1}{q^{\lambda\nu}}$, whereas $\frac{1}{q^{\lambda\nu}} \leq \rho \leq \frac{1}{|\omega_j|^{\nu}}$, $\forall 1 \leq j \leq d$. Bearing in mind that for any fixed $\nu \in \mathbb{N}$ the function $f(x) = x^{\nu}$ is non-decreasing on $x \in [0, \infty) \subset \mathbb{R}$, one concludes that $q^{\lambda} \geq |\omega_j|$. Therefore, the leading coefficient $a_d := \mathrm{LC}(P_M(t)) = \prod\limits_{j=1}^{d}(-\omega_j) \in \mathbb{Z} \setminus \{0\}$ has modulus $|a_d| = \prod\limits_{j=1}^{d}|\omega_j| \leq q^{\lambda d} = |a_d|$, whereas $|\omega_j| = q^{\lambda}$, $\forall 1 \leq j \leq d$ and M satisfies the Riemann Hypothesis Analogue with respect to $\mathbb{P}^1(\overline{\mathbb{F}_q})$ as a module over $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$. $\square$

In the case of a smooth irreducible projective curve $X/\mathbb{F}_q \subset \mathbb{P}^n(\overline{\mathbb{F}_q})$ of genus g, defined over $\mathbb{F}_q$, condition (ii) from Proposition 23 reduces to the celebrated Hasse - Weil bound

$$\left|\left|X^{\Phi_q^r}\right| - (q^r + 1)\right| \leq 2g\sqrt{q^r} \quad \forall r \in \mathbb{N} \tag{5.1}$$

on the number $\left|X^{\Phi_q^r}\right| = |X(\mathbb{F}_{q^r})| = |X \cap \mathbb{P}^n(\mathbb{F}_{q^r})|$ of the $\mathbb{F}_{q^r}$-rational points of X. The equivalence of the conditions (i) and (iii) from Proposition (23) is well known and shown by Theorem V.2.3 and Lemma V.2.5 from Stichtenoth's monograph [2]. The proof of the Riemann Hypothesis Analogue for X with respect to $\mathbb{P}^1(\overline{\mathbb{F}_q})$ from [2] makes use of the bound

$$\left|X^{\Phi_q^{2r}}\right| < q^{2r} + 1 + (2g+1)q^r \quad \forall r \in \mathbb{N}, \tag{5.2}$$

which is established in [2, Proposition V.2.6]. Bearing in mind that $\left|\mathbb{P}^1(\overline{\mathbb{F}_q})^{\Phi_q^{2r}}\right| = q^{2r} + 1 > q^{2r}$, we note that (5.2) implies

$$\left|X^{\Phi_q^{2r}}\right| - \left|\mathbb{P}^1(\overline{\mathbb{F}_q})^{\Phi_q^{2r}}\right| < (2g+1)\left|\mathbb{P}^1(\overline{\mathbb{F}_q})^{\Phi_q^{2r}}\right|^{\frac{1}{2}} \quad \forall r \in \mathbb{N}$$

and think of $\lambda := \log_q \sqrt[2g]{\mathrm{LC}(P_X(t))} = \log_q \sqrt[2g]{q^g} = \frac{1}{2}$ as of the Hasse - Weil order of X with respect to $\mathbb{P}^1(\overline{\mathbb{F}_q})$. That motivates the following

Definition 24. Let M and L be locally finite $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$-modules. If there exist constants $\rho \in \mathbb{R}^{\geq 0}$, $C \in \mathbb{R}^{>0}$, $\nu, r_o \in \mathbb{N}$, such that

$$\left|M^{\Phi_q^{\nu r}}\right| - \left|L^{\Phi_q^{\nu r}}\right| \leq C\left|L^{\Phi_q^{\nu r}}\right|^{\rho} \quad \forall r \in \mathbb{N}, \quad r \geq r_o, \tag{5.3}$$

M is said to be of finite Hasse - Weil order with respect to L.

The minimal $\rho \in \mathbb{R}^{\geq 0}$, subject to (5.3) for some $C \in \mathbb{R}^{>0}$, $\nu, r_o \in \mathbb{N}$ is called the Hasse - Weil order of M with respect to L and denoted by $\mathrm{ord}_{\mathfrak{G}}(M/L)$.

The following simple lemma collects some properties of the Hasse - Weil order of locally finite $\mathfrak{G}$-modules.

Lemma 25. *(i) If M, L are infinite locally finite $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$-modules and $M_o \subseteq M$, $L_o \subseteq L$ are $\mathfrak{G}$-submodules with at most finite complements $M \setminus M_o$, $L \setminus L_o$, then*

$$\mathrm{ord}_{\mathfrak{G}}(M/L) = \mathrm{ord}_{\mathfrak{G}}(M_o/L_o).$$

(ii) If $\xi : M \to L$ is a finite unramified covering of locally finite $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$-modules, then $\mathrm{ord}_{\mathfrak{G}}(M/L) \leq 1$.

(iii) Let M be a locally finite $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$-module such that $\zeta_M(t) = P_M(t)\zeta_{\mathbb{P}^1(\overline{\mathbb{F}_q})}(t)$ for a polynomial $P_M(t) \in \mathbb{Z}[t]$ of $\deg P_M(t) = d \in \mathbb{N}$ with $\mathrm{LC}(P_M(t)) = a_d$ and $\lambda := \log_q \sqrt[d]{|a_d|}$. If M satisfies the Riemann Hypothesis Analogue with respect to $\mathbb{P}^1(\overline{\mathbb{F}_q})$, then $\mathrm{ord}_{\mathfrak{G}}(M/\mathbb{P}^1(\overline{\mathbb{F}_q})) \leq \lambda$.

Proof. (i) It suffices to show that if there exist $C \in \mathbb{R}^{>0}$, $\nu, r' \in \mathbb{N}$ with

$$\left|M^{\Phi_q^{\nu r}}\right| \leq \left|L^{\Phi_q^{\nu r}}\right| + C\left|L^{\Phi_q^{\nu r}}\right|^{\rho} \quad \forall r \in \mathbb{N}, \quad r \geq r', \tag{5.4}$$

then there exist $C_o \in \mathbb{R}^{>0}$, $\nu_o, r'_o \in \mathbb{N}$ with

$$\left|M_o^{\Phi_q^{\nu_o r}}\right| \leq \left|L_o^{\Phi_q^{\nu_o r}}\right| + C_o\left|L_o^{\Phi_q^{\nu_o r}}\right|^{\rho} \quad \forall r \in \mathbb{N}, r \geq r'_o \tag{5.5}$$

and if there are $\widetilde{C_o} \in \mathbb{R}^{>0}$, $\widetilde{\nu_o}, \widetilde{r'_o} \in \mathbb{N}$ with

$$\left|M_o^{\Phi_q^{\widetilde{\nu_o} r}}\right| \leq \left|L_o^{\Phi_q^{\widetilde{\nu_o} r}}\right| + \widetilde{C_o}\left|L_o^{\Phi_q^{\widetilde{\nu_o} r}}\right|^{\rho} \quad \forall r \in \mathbb{N}, \quad r \geq \widetilde{r'_o}, \tag{5.6}$$

then there are $\widetilde{C} \in \mathbb{R}^{>0}$, $\widetilde{\nu}, \widetilde{r'} \in \mathbb{N}$ with

$$\left|M^{\Phi_q^{\widetilde{\nu} r}}\right| \leq \left|L^{\Phi_q^{\widetilde{\nu} r}}\right| + \widetilde{C}\left|L^{\Phi_q^{\widetilde{\nu} r}}\right|^{\rho} \quad \forall r \in \mathbb{N}, \quad r \geq \widetilde{r'}. \tag{5.7}$$

To this end, let us denote $m := |M \setminus M_o|$, $s := |L \setminus L_o| \in \mathbb{Z}^{\geq 0}$ and observe that

$$\left|L^{\Phi_q^{\nu_o r}}\right| = \left|L_o^{\Phi_q^{\nu_o r}}\right| + \left|L^{\Phi_q^{\nu_o r}} \setminus L_o\right| \leq \left|L_o^{\Phi_q^{\nu_o r}}\right| + s,$$

$$\left|M^{\Phi_q^{\widetilde{\nu} r}}\right| = \left|M_o^{\Phi_q^{\widetilde{\nu} r}}\right| + \left|M^{\Phi_q^{\widetilde{\nu} r}} \setminus M_o\right| \leq \left|M_o^{\Phi_q^{\widetilde{\nu} r}}\right| + m, \quad \forall r \in \mathbb{N}.$$

Since L_o is an infinite locally finite $\mathfrak{G}$-module, the map

$$\deg \mathrm{Orb}_{\mathfrak{G}} : L_o \to \mathbb{N}, \quad x \mapsto \deg \mathrm{Orb}_{\mathfrak{G}}(x)$$

takes infinitely many values and there exists $\sigma_o \in \mathbb{N}$ with $\sigma_o \geq \max\left(s, \sqrt[\rho]{s}\right)$ from the image of $\deg \mathrm{Orb}_{\mathfrak{G}} : L_o \to \mathbb{N}$. In other words, the number $B_{\sigma_o}(L_o) \geq 1$ of the $\mathfrak{G}$-orbits on L_o of degree σ_o is positive. If $\nu_o := \nu\sigma_o \in \mathbb{N}$, then by (3.2) one has

$$\left|L_o^{\Phi_q^{\nu_o r}}\right| = \sum_{k/\nu_o r} kB_k(L_o) \geq \sigma_o B_{\sigma_o}(L_o) \geq \sigma_o \geq \max\left(s, \sqrt[\rho]{s}\right) \quad \forall r \in \mathbb{N}.$$

Similarly, there exists $\sigma \in \mathbb{N}$ with $\sigma > \sqrt[\rho]{m}$ and $B_\sigma(L_o) \geq 1$. Thus, for $\widetilde{\nu} := \widetilde{\nu_o}\sigma \in \mathbb{N}$ there holds

$$\left|L_o^{\Phi_q^{\widetilde{\nu} r}}\right| = \sum_{k/\widetilde{\nu} r} kB_k(L_o) \geq \sigma B_\sigma(L_o) \geq \sigma \geq \sqrt[\rho]{m} \quad \forall r \in \mathbb{N}.$$

Now (5.4) implies

$$\left|M_o^{\Phi_q^{\nu_o r}}\right| \le \left|M^{\Phi_q^{\nu_o r}}\right| \le \left|L^{\Phi_q^{\nu_o r}}\right| + C\left|L^{\Phi_q^{\nu_o r}}\right|^\rho \le \left|L_o^{\Phi_q^{\nu_o r}}\right| + s + C\left(\left|L_o^{\Phi_q^{\nu_o r}}\right| + s\right)^\rho$$
$$\le \left|L_o^{\Phi_q^{\nu_o r}}\right| + \left|L_o^{\Phi_q^{\nu_o r}}\right|^\rho + C\left(2\left|L_o^{\Phi_q^{\nu_o r}}\right|\right)^\rho = \left|L_o^{\Phi_q^{\nu_o r}}\right| + (2^\rho C + 1)\left|L_o^{\Phi_q^{\nu_o r}}\right|^\rho$$

$\forall r \in \mathbb{N}$, $r \ge \frac{r'}{\sigma_o}$, which is equivalent to (5.5) with $C_o = 2^\rho C + 1$ and some $r'_o \in \mathbb{N}$, $r'_o \ge \frac{r'}{\sigma_o}$. Similarly, (5.6) yields

$$\left|M^{\Phi_q^{\tilde{\nu} r}}\right| \le \left|M_o^{\Phi_q^{\tilde{\nu} r}}\right| + m \le \left|L_o^{\Phi_q^{\tilde{\nu} r}}\right| + \widetilde{C_o}\left|L_o^{\Phi_q^{\tilde{\nu} r}}\right|^\rho + \left|L_o^{\Phi_q^{\tilde{\nu} r}}\right|^\rho$$
$$= \left|L_o^{\Phi_q^{\tilde{\nu} r}}\right| + (\widetilde{C_o} + 1)\left|L_o^{\Phi_q^{\tilde{\nu} r}}\right|^\rho \le \left|L^{\Phi_q^{\tilde{\nu} r}}\right| + (\widetilde{C_o} + 1)\left|L^{\Phi_q^{\tilde{\nu} r}}\right|^\rho$$

$\forall r \in \mathbb{N}$, $r \ge \frac{\widetilde{r'_o}}{\sigma}$, and hence (5.7) holds with $\widetilde{C} := \widetilde{C_o} + 1$ and some $\widetilde{r'} \in \mathbb{N}$, $\widetilde{r'} \ge \frac{\widetilde{r'_o}}{\sigma}$.

(ii) The $\mathfrak{G}$-equivariance of ξ implies that $\xi(M^{\Phi_q^r}) \subseteq L^{\Phi_q^r}$, $\forall r \in \mathbb{N}$. The cardinalities of the fibres of $\xi|_{M^{\Phi_q^r}}$ do not exceed $k := \deg \xi$, so that

$$\left|L^{\Phi_q^r}\right| \ge \left|\xi(M^{\Phi_q^r})\right| \ge \frac{\left|M^{\Phi_q^r}\right|}{k}$$

and

$$\left|M^{\Phi_q^r}\right| - \left|L^{\Phi_q^r}\right| \le (k-1)\left|L^{\Phi_q^r}\right|.$$

That suffices for $\mathrm{ord}_{\mathfrak{G}}(M/L) \le 1$.

(iii) By Proposition 23, if M satisfies the Riemann Hypothesis Analogue with respect to $\mathbb{P}^1(\overline{\mathbb{F}_q})$ as a $\mathfrak{G}$-module, then

$$\left|M^{\Phi_q^r}\right| \le q^r + 1 + dq^{\lambda r} = \left|\mathbb{P}^1(\overline{\mathbb{F}_q})^{\Phi_q^r}\right| + d\left(\left|\mathbb{P}^1(\overline{\mathbb{F}_q})^{\Phi_q^r}\right| - 1\right)^\lambda$$
$$< \left|\mathbb{P}^1(\overline{\mathbb{F}_q})^{\Phi_q^r}\right| + d\left|\mathbb{P}^1(\overline{\mathbb{F}_q})^{\Phi_q^r}\right|^\lambda \quad \forall r \in \mathbb{N},$$

so that $\mathrm{ord}_{\mathfrak{G}}(M/\mathbb{P}^1(\overline{\mathbb{F}_q})) \le \lambda$. □

Definition 26. Let M and L be locally finite $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$-modules and H be a finite fixed-point free subgroup of $\mathrm{Aut}_{\mathfrak{G}}(M)$. If there exist constants $\rho \in \mathbb{R}^{\ge 0}$, $C \in \mathbb{R}^{>0}$, $\nu, r_o \in \mathbb{N}$, such that

$$\left|M^{h\Phi_q^{\nu r}}\right| - \left|L^{\Phi_q^{\nu r}}\right| \le C\left|L^{\Phi_q^{\nu r}}\right|^\rho \quad \text{for} \quad \forall r \in \mathbb{N}, \quad r \ge r_o \quad \text{and} \quad \forall h \in H, \qquad (5.8)$$

then M is said to be of finite Hasse - Weil H-order with respect to L.

The minimal $\rho \in \mathbb{R}^{\ge 0}$, subject to (5.8) for some $C \in \mathbb{R}^{>0}$, $\nu, r_o \in \mathbb{N}$ is called the Hasse - Weil H-order of M with respect to L and denoted by $\mathrm{ord}^H_{\mathfrak{G}}(M/L)$.

Proposition 27. *(i) If M is an infinite locally finite $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$-module, $H < \mathrm{Aut}_{\mathfrak{G}}(M)$ is a finite fixed-point free subgroup and $M_o \subset M$ is an $H \times \mathfrak{G}$-submodule of M with $|M \setminus M_o| < \infty$, then*

$$\mathrm{ord}^H_{\mathfrak{G}}(M/\mathrm{Orb}_H(M)) = \mathrm{ord}^H_{\mathfrak{G}}(M_o/\mathrm{Orb}_H(M_o)).$$

(ii) If M is a locally finite $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$-module and $H < \mathrm{Aut}_{\mathfrak{G}}(M)$ is a finite fixed-point free subgroup, then $\mathrm{ord}^H_{\mathfrak{G}}(M/\mathrm{Orb}_H(M)) \leq 1$.

(iii) Let $X/\mathbb{F}_q \subset \mathbb{P}^n(\overline{\mathbb{F}_q})$ be a smooth irreducible curve of genus $g \geq 1$ and H be a finite fixed-point free subgroup of $\mathrm{Aut}_{\mathfrak{G}}(X)$. *Then* $\mathrm{ord}^H_{\mathfrak{G}}(X/\mathbb{P}^1(\overline{\mathbb{F}_q})) \leq \frac{1}{2}$.

Proof. (i) As in the proof of Lemma 25 (i), one has to check that if there exist $\rho \in \mathbb{R}^{\geq 0}$, $C \in \mathbb{R}^{>0}$, $\nu, r' \in \mathbb{N}$ with

$$\left|M^{h\Phi_q^{\nu r}}\right| \leq \left|\mathrm{Orb}_H(M)^{\Phi_q^{\nu r}}\right| + C\left|\mathrm{Orb}_H(M)^{\Phi_q^{\nu r}}\right|^{\rho} \ \forall h \in H, \forall r \in \mathbb{N}, r \geq r', \quad (5.9)$$

then there exist $C_o \in \mathbb{R}^{>0}$, $\nu_o, r'_o \in \mathbb{N}$ with

$$\left|M_o^{h\Phi_q^{\nu_o r}}\right| \leq \left|\mathrm{Orb}_H(M_o)^{\Phi_q^{\nu_o r}}\right| + C_o\left|\mathrm{Orb}_H(M_o)^{\Phi_q^{\nu_o r}}\right|^{\rho} \ \forall h \in H, \forall r \in \mathbb{N}, r \geq r'_o \quad (5.10)$$

and if there are $\widetilde{C}_o \in \mathbb{R}^{>0}$, $\widetilde{\nu}_o, \widetilde{r}_o \in \mathbb{N}$ with

$$\left|M_o^{h\Phi_q^{\widetilde{\nu}_o r}}\right| \leq \left|\mathrm{Orb}_H(M_o)^{\Phi_q^{\widetilde{\nu}_o r}}\right| + \widetilde{C}_o\left|\mathrm{Orb}_H(M_o)^{\Phi_q^{\widetilde{\nu}_o r}}\right|^{\rho} \ \forall h \in H, \forall r \in \mathbb{N}, r \geq \widetilde{r}_o \quad (5.11)$$

then there are $\widetilde{C} \in \mathbb{R}^{>0}$, $\widetilde{\nu}, \widetilde{r} \in \mathbb{N}$ with

$$\left|M^{h\Phi_q^{\widetilde{\nu} r}}\right| \leq \left|\mathrm{Orb}_H(M)^{\Phi_q^{\widetilde{\nu} r}}\right| + \widetilde{C}\left|\mathrm{Orb}_H(M)^{\Phi_q^{\widetilde{\nu} r}}\right|^{\rho} \ \forall h \in H, \forall r \in \mathbb{N}, r \geq \widetilde{r}. \quad (5.12)$$

Note that if $|M \setminus M_o| = m$, then $\mathrm{Orb}_H(M) \setminus \mathrm{Orb}_H(M_o) = \mathrm{Orb}_H(M \setminus M_o)$ is of cardinality $|\mathrm{Orb}_H(M \setminus M_o)| = \frac{m}{|H|}$ and $\mathrm{Orb}_H(M_o)$ is an infinite locally finite $\mathfrak{G}$-module. As in the proof of Lemma 25 (i), one has

$$\left|\mathrm{Orb}_H(M)^{\Phi_q^{\nu_o r}}\right| \leq \left|\mathrm{Orb}_H(M_o)^{\Phi_q^{\nu_o r}}\right| + \frac{m}{|H|} \quad \text{and} \quad \left|M^{h\Phi_q^{\widetilde{\nu} r}}\right| \leq \left|M_o^{h\Phi_q^{\widetilde{\nu} r}}\right| + m \ \ \forall r \in \mathbb{N}.$$

Further, there exist $\nu_o := \nu\sigma_o$ and $\widetilde{\nu} := \widetilde{\nu}_o\sigma$ with $\sigma_o, \sigma \in \mathbb{N}$, such that

$$\left|\mathrm{Orb}_H(M_o)^{\Phi_q^{\nu_o r}}\right| \geq \sigma_o \geq \max\left(\frac{m}{|H|}, \sqrt[\rho]{\frac{m}{|H|}}\right),$$

respectively,

$$\left|\mathrm{Orb}_H(M_o)^{\Phi_q^{\widetilde{\nu} r}}\right| \geq \sigma \geq \sqrt[\rho]{m} \ \ \forall r \in \mathbb{N}.$$

Then from

$$\begin{aligned}\left|M_o^{h\Phi_q^{\nu_o r}}\right| &\leq \left|M^{h\Phi_q^{\nu_o r}}\right| \leq \left|\mathrm{Orb}_H(M)^{\Phi_q^{\nu_o r}}\right| + C\left|\mathrm{Orb}_H(M)^{\Phi_q^{\nu_o r}}\right|^{\rho}\\ &\leq \left|\mathrm{Orb}_H(M_o)^{\Phi_q^{\nu_o r}}\right| + \frac{m}{|H|} + C\left(\left|\mathrm{Orb}_H(M_o)^{\Phi_q^{\nu_o r}}\right| + \frac{m}{|H|}\right)^{\rho}\\ &\leq \left|\mathrm{Orb}_H(M_o)^{\Phi_q^{\nu_o r}}\right| + \left|\mathrm{Orb}_H(M_o)^{\Phi_q^{\nu_o r}}\right|^{\rho} + C\left(2\left|\mathrm{Orb}_H(M_o)^{\Phi_q^{\nu_o r}}\right|\right)^{\rho},\end{aligned}$$

$\forall r \in \mathbb{N}$, $r \geq \frac{r'}{\sigma_o}$, we deduce (5.10) with $C_o := 2^{\rho}C + 1$, and from

$$\begin{aligned}\left|M^{h\Phi_q^{\tilde{\nu}r}}\right| &\leq \left|M_o^{h\Phi_q^{\tilde{\nu}r}}\right| + m\\ &\leq \left|\mathrm{Orb}_H(M_o)^{\Phi_q^{\tilde{\nu}r}}\right| + \widetilde{C}_o\left|\mathrm{Orb}_H(M_o)^{\Phi_q^{\tilde{\nu}r}}\right|^{\rho} + \left|\mathrm{Orb}_H(M_o)^{\Phi_q^{\tilde{\nu}r}}\right|^{\rho}\\ &\leq \left|\mathrm{Orb}_H(M)^{\Phi_q^{\tilde{\nu}r}}\right| + (\widetilde{C}_o + 1)\left|\mathrm{Orb}_H(M)^{\Phi_q^{\tilde{\nu}r}}\right|^{\rho},\end{aligned}$$

$\forall r \in \mathbb{N}$, $r \geq \frac{\widetilde{r}_o}{\sigma}$, we obtain (5.12) with $\widetilde{C} := \widetilde{C}_o + 1$.

(ii) For any $h \in H$ and $r \in \mathbb{N}$ the map $\xi_H : M \to \mathrm{Orb}_H(M)$ is an H-Galois covering of locally finite modules over $\mathfrak{G}(h\Phi_q^r) = \widehat{\langle h\Phi_q^r \rangle}$ by Proposition 20. If $y \in M^{h\Phi_q^r}$, then the $\mathfrak{G}(h\Phi_q^r)$-equivariance of ξ_H implies $\Phi_q^r\xi_H(y) = \xi_H(\Phi_q^r y) = \xi_H(h\Phi_q^r y) = \xi_H(y)$, so that $\xi_H(y) \in \mathrm{Orb}_H(M)^{\Phi_q^r}$ and $\xi_H(M^{h\Phi_q^r}) \subseteq \mathrm{Orb}_H(M)^{\Phi_q^r}$. Bearing in mind that the restriction $\xi_H : M^{h\Phi_q^r} \to \mathrm{Orb}_H(M)^{\Phi_q^r}$ has fibres of cardinality $\leq |H|$, one concludes that $\left|\mathrm{Orb}_H(M)^{\Phi_q^r}\right| \geq \left|\xi_H(M^{h\Phi_q^r})\right| \geq \frac{\left|M^{h\Phi_q^r}\right|}{|H|}$. Therefore

$$\left|M^{h\Phi_q^r}\right| - \left|\mathrm{Orb}_H(M)^{\Phi_q^r}\right| \leq (|H| - 1)\left|\mathrm{Orb}_H(M)^{\Phi_q^r}\right|,$$

$\forall h \in H$, $\forall r \in \mathbb{N}$ and $\mathrm{ord}_{\mathfrak{G}}^H(M/\mathrm{Orb}_H(M)) \leq 1$.

(iii) The argument is a slight modification of Grothedieck's proof of the Hasse - Weil Theorem (see Theorem 3.6 from Mustaţă's book [8]). Namely, let $S := X \times X$ be the Cartesian square of X, $\Delta := \{(x, x) \in S \,|\, x \in X\}$ be the diagonal of S, $L_1 := X \times \{x_2\}$ be a generic fibre of the second canonical projection $\mathrm{pr}_2 : S \to X$, $\mathrm{pr}_2(x_1, x_2) = x_2$ and $L_2 := \{x_1\} \times X$ be a generic fibre of the first canonical projection $\mathrm{pr}_1 : S \to X$, $\mathrm{pr}_1(x_1, x_2) = x_1$. For arbitrary $h \in H$ and $r \in \mathbb{N}$ put $\varphi := h\Phi_q^r$ and denote by $\Gamma(\varphi) := \{(x, \varphi(x)) \,|\, x \in X\}$ the graph of $\varphi : X \to X$. Then the intersection number $\Gamma(\varphi).\Delta = |X^{\varphi}|$ equals the number of the φ-rational points of X. One checks immediately that $L_1^2 = L_2^2 = 0$, $L_1.L_2 = 1$, $\Delta.L_1 = \Delta.L_2 = 1$, $\Gamma(\varphi).L_2 = 1$ and $\Gamma(\varphi).L_1 = \Gamma(\Phi_q^r).L_1 = q^r$, as far as the equation $h\Phi_q^r(x) = x_2$ is equivalent to $\Phi_q^r(x) = h^{-1}(x_2)$ and has q^r solutions on a smooth irreducible projective curve X, defined over $\mathbb{F}_q$. The canonical class K_S of S is numerically equivalent to $(2g-2)(L_1 + L_2)$ and the application of the Adjunction Formula to Δ and $\Gamma(\varphi)$ provides

$$2g - 2 = \Delta.(\Delta + K_S) = \Delta^2 + 2(2g - 2),$$

$$2g-2=\Gamma(\varphi).(\Gamma(\varphi)+K_S)=\Gamma(\varphi)^2+(q^r+1)(2g-2),$$

whereas $\Delta^2=-(2g-2)$, $\Gamma(\varphi)^2=-q^r(2g-2)$. The Hodge Index Theorem on $S=X\times X$ asserts that if a divisor $E\subset S$ has vanishing intersection number $E.H=0$ with an ample divisor $H\subset S$ then E has non-positive self-intersection $E^2\leq 0$. For an arbitrary divisor $D\subset S$ let us put $E:=D-(D.L_1)L_2-(D.L_2)L_1$, $H:=L_1+L_2$ and note that H is an ample divisor on S with $E.H=0$. Therefore

$$0\geq E^2=D^2-2(D.L_1)(D.L_2). \tag{5.13}$$

If $D:=a\Delta+b\Gamma(\varphi)$ for some $a,b\in\mathbb{Z}$, $b\neq 0$ and $f(z):=gz^2+(q^r+1-|X^\varphi|)z+gq^r\in\mathbb{Z}[z]$, then (5.13) is equivalent to $f\left(\frac{a}{b}\right)\geq 0$, $\forall\frac{a}{b}\in\mathbb{Q}$ and holds exactly when the discriminant $D(f)=(q^r+1-|X^\varphi|)^2-4q^rg^2\leq 0$. Thus,

$$-2gq^{\frac{r}{2}}\leq|X^\varphi|-(q^r+1)\leq 2gq^{\frac{r}{2}}\quad\forall r\in\mathbb{N}$$

and, in particular,

$$\begin{aligned}\left|X^{h\Phi_q^{2r}}\right|&\leq(q^{2r}+1)+2gq^r=\left|\mathbb{P}^1(\overline{\mathbb{F}_q})^{\Phi_q^{2r}}\right|+2g\left(\left|\mathbb{P}^1(\overline{\mathbb{F}_q})^{\Phi_q^{2r}}\right|-1\right)^{\frac{1}{2}}\\&\leq\left|\mathbb{P}^1(\overline{\mathbb{F}_q})^{\Phi_q^{2r}}\right|+2g\left|\mathbb{P}^1(\overline{\mathbb{F}_q})^{\Phi_q^{2r}}\right|^{\frac{1}{2}}\quad\forall r\in\mathbb{N}.\end{aligned}$$

That establishes the inequality $\mathrm{ord}_{\mathfrak{G}}^H(X/\mathbb{P}^1(\overline{\mathbb{F}_q}))\leq\frac{1}{2}$. □

The following simple lemma is crucial for the proof of the main Theorem 29.

Lemma 28. *Let $\xi_H:N\to L$ be an H-Galois covering of infinite locally finite modules over $\mathfrak{G}=\mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$ for some finite fixed-point free subgroup $H<\mathrm{Aut}_{\mathfrak{G}}(N)$. Then*

$$\sum_{h\in H}\left|N^{h\Phi_q}\right|=|H|\left|L^{\Phi_q}\right|.$$

Proof. The lack of fixed points of H implies that $N^{h_1\Phi_q}\cap N^{h_2\Phi_q}=\emptyset$ for all $h_1,h_2\in H$, $h_1\neq h_2$. It suffices to check that $\xi_H^{-1}(L^{\Phi_q})=\coprod_{h\in H}N^{h\Phi_q}$, in order to conclude that

$$|H||L^{\Phi_q}|=\left|\xi_H^{-1}(L^{\Phi_q})\right|=\sum_{h\in H}\left|N^{h\Phi_q}\right|.$$

If $y\in\xi_H^{-1}(L^{\Phi_q})$, then $\xi_H(y)=\Phi_q\xi_H(y)=\xi_H(\Phi_q(y))$ implies the existence of $h\in H$ with $h(y)=\Phi_q(y)$. Therefore $y\in N^{h^{-1}\Phi_q}$ and $\xi_H^{-1}(L^{\Phi_q})\subseteq\coprod_{h\in H}N^{h\Phi_q}$.

Conversely, for any $y\in N^{h\Phi_q}$ one has $h^{-1}(y)=\Phi_q(y)$, whereas

$$\xi_H(y)=\xi_H(h^{-1}(y))=\xi_H(\Phi_q(y))=\Phi_q\xi_H(y).$$

That justifies $N^{h\Phi_q}\subseteq\xi_H^{-1}(L^{\Phi_q})$ and $\xi_H^{-1}(L^{\Phi_q})=\coprod_{h\in H}N^{h\Phi_q}$. □

Here is the main result of the article.

Theorem 29. *Let M be an infinite locally finite module over $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$ with closed stabilizers and a polynomial ζ-quotient $P_M(t) = \frac{\zeta_M(t)}{\zeta_{\mathbb{P}^1(\overline{\mathbb{F}_q})}(t)} = \sum\limits_{j=0}^{d} a_j t^j \in \mathbb{Z}[t]$ of $\deg P_M(t) = d \in \mathbb{N}$ with leading coefficient $\mathrm{LC}(P_M(t)) = a_d \in \mathbb{Z} \setminus \{0\}$ and $\lambda := \log_q \sqrt[d]{|a_d|} \in \mathbb{R}^{>0}$. Suppose that there exist $m \in \mathbb{N}$ and $\mathfrak{G}_m = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_{q^m})$-submodules $M_o \subseteq M$, $L_o \subseteq \mathbb{P}^1(\overline{\mathbb{F}_q})$ with $|M \setminus M_o| < \infty$, $\left|\mathbb{P}^1(\overline{\mathbb{F}_q}) \setminus L_o\right| < \infty$, which are related by a finite unramified covering $\xi : M_o \to L_o$ of $\mathfrak{G}_m$-modules with a Galois closure (N, H, H_1), defined over $\mathbb{F}_{q^m}$.*

(i) If $\lambda \geq 1$, then M satisfies the Riemann Hypothesis Analogue with respect to the projective line $\mathbb{P}^1(\overline{\mathbb{F}_q})$ as a $\mathfrak{G}$-module.

(ii) If

$$\max\left(\mathrm{ord}_{\mathfrak{G}}(M/\mathbb{P}^1(\overline{\mathbb{F}_q})), \mathrm{ord}^H_{\mathfrak{G}_m}(N/\mathbb{P}^1(\overline{\mathbb{F}_q})\right) \leq \lambda < 1,$$

then M satisfies the Riemann Hypothesis Analogue with respect to $\mathbb{P}^1(\overline{\mathbb{F}_q})$ as a $\mathfrak{G}$-module.

Proof. It suffices to prove that if

$$\max(\mathrm{ord}_{\mathfrak{G}}(M/\mathbb{P}^1(\overline{\mathbb{F}_q})), \mathrm{ord}^H_{\mathfrak{G}_m}(N/\mathbb{P}^1(\overline{\mathbb{F}_q})) \leq \lambda, \tag{5.14}$$

then M satisfies the Riemann Hypothesis Analogue with respect to $\mathbb{P}^1(\overline{\mathbb{F}_q})$ as a $\mathfrak{G}$-module. Namely, if $\lambda \geq 1$, then by Lemma 25 (i), (ii) one has

$$\mathrm{ord}_{\mathfrak{G}}(M/\mathbb{P}^1(\overline{\mathbb{F}_q})) = \mathrm{ord}_{\mathfrak{G}}(M_o/L_o) \leq 1 \leq \lambda,$$

while Proposition 27 (i), (ii) guarantee that

$$\mathrm{ord}^H_{\mathfrak{G}_m}(N/\mathbb{P}^1(\overline{\mathbb{F}_q})) = \mathrm{ord}^H_{\mathfrak{G}_m}(N/L_o) \leq 1 \leq \lambda,$$

whence (5.14) holds.

Since $f(x) = a^x$ is an increasing function on $x \in \mathbb{R}$ for $a \in \mathbb{N}$, $a \geq 2$, the assumption $\mathrm{ord}_{\mathfrak{G}}(M/\mathbb{P}^1(\overline{\mathbb{F}_q})) \leq \lambda$ implies the existence of constants $C_1 \in \mathbb{R}^{>0}$, $\nu_1, r_1 \in \mathbb{N}$, such that

$$\left|M^{\Phi_q^{\nu_1 r}}\right| \leq (q^{\nu_1 r}+1)+C_1(q^{\nu_1 r}+1)^{\lambda} < (q^{\nu_1 r}+1)+C_1(2q^{\nu_1 r})^{\lambda} = (q^{\nu_1 r}+1)+(2^{\lambda}C_1)q^{\lambda\nu_1 r},$$

$\forall r \in \mathbb{N}$, $r \geq r_1$. Similarly, $\mathrm{ord}^H_{\mathfrak{G}_m}(N/\mathbb{P}^1(\overline{\mathbb{F}_q})) \leq \lambda$ provides the presence of constants $C_2 \in \mathbb{R}^{>0}$, $\nu_2, r_2 \in \mathbb{N}$ with

$$\left|N^{h\Phi_q^{\nu_2 r}}\right| \leq (q^{\nu_2 r}+1) + C_2(q^{\nu_2 r}+1)^{\lambda} < (q^{\nu_2 r}+1) + (2^{\lambda}C_2)q^{\lambda\nu_2 r},$$

$\forall r \in \mathbb{N}$, $r \geq r_2$. For an arbitrary common multiple $\nu \in \mathbb{N}$ of ν_1 and ν_2, one has

$$\left|M^{\Phi_q^{\nu r}}\right| < (q^{\nu r}+1) + (2^{\lambda}C_1)q^{\lambda\nu r} \quad \forall r \in \mathbb{N}, \quad r \geq \frac{r_1\nu_1}{\nu} \tag{5.15}$$

and

$$\left|N^{h\Phi_q^{\nu r}}\right| < (q^{\nu r}+1) + (2^\lambda C_2) q^{\lambda\nu r} \quad \forall r \in \mathbb{N}, \quad r \geq \frac{r_2\nu_2}{\nu}.$$

If $\left|\mathbb{P}^1(\overline{\mathbb{F}_q}) \setminus L_o\right| = s$, then the decomposition $\mathbb{P}^1(\overline{\mathbb{F}_q})^{\Phi_q^{\nu r}} = L_o^{\Phi_q^{\nu r}} \coprod (\mathbb{P}^1(\overline{\mathbb{F}_q})^{\Phi_q^{\nu r}} \setminus L_o)$ into a disjoint union provides the inequality $q^{\nu r}+1 \leq \left|L_o^{\Phi_q^{\nu r}}\right| + s$, whereas

$$\left|N^{h\Phi_q^{\nu r}}\right| < \left|L_o^{\Phi_q^{\nu r}}\right| + s + (2^\lambda C_2) q^{\lambda\nu r} \leq \left|L_o^{\Phi_q^{\nu r}}\right| + (2^\lambda C_2 + 1) q^{\lambda\nu r}, \tag{5.16}$$

$\forall r \in \mathbb{N}$, $r \geq r_o$ and a fixed natural number $r_o \geq \max\left(\frac{r_2\nu_2}{\nu}, \frac{\log_q(s)}{\lambda\nu}\right)$. By Proposition 23, it suffices to show the existence of constants $C \in \mathbb{R}^{>0}$, $r_o \in \mathbb{N}$ with

$$\left|M^{\Phi_q^{\nu r}}\right| \geq (q^{\nu r}+1) - C q^{\lambda\nu r} \quad \forall r \in \mathbb{N}, \quad r \geq r_o \tag{5.17}$$

and to combine with (5.15), in order to conclude that M satisfies the Riemann Hypothesis Analogue with respect to $\mathbb{P}^1(\overline{\mathbb{F}_q})$ as a module over $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$.

To this end, note that Lemma 28 implies

$$\sum_{h\in H} \left|N^{h\Phi_q^{\nu r}}\right| = |H| \left|L_o^{\Phi_q^{\nu r}}\right| \quad \text{and} \quad \sum_{h\in H_1} \left|N^{h\Phi_q^{\nu r}}\right| = |H_1| \left|M_o^{\Phi_q^{\nu r}}\right| \quad \forall r \in \mathbb{N}.$$

Putting together with (5.16), one obtains that

$$\begin{aligned}
|H_1| \left|M_o^{\Phi_q^{\nu r}}\right| &= \sum_{h\in H_1} \left|N^{h\Phi_q^{\nu r}}\right| + |H| \left|L_o^{\Phi_q^{\nu r}}\right| - \sum_{h\in H} \left|N^{h\Phi_q^{\nu r}}\right| \\
&= |H| \left|L_o^{\Phi_q^{\nu r}}\right| - \sum_{h\in H\setminus H_1} \left|N^{h\Phi_q^{\nu r}}\right| \\
&\geq |H| \left|L_o^{\Phi_q^{\nu r}}\right| - (|H| - |H_1|) \left|L_o^{\Phi_q^{\nu r}}\right| - (|H| - |H_1|)(2^\lambda C_2 + 1) q^{\lambda\nu r} \\
&= |H_1| \left|L_o^{\Phi_q^{\nu r}}\right| - (|H| - |H_1|)(2^\lambda C_2 + 1) q^{\lambda\nu r} \quad \forall r \in \mathbb{N},\ r \geq r_o.
\end{aligned}$$

Denoting $C_3 := \left(\frac{|H|-|H_1|}{|H_1|}\right)(2^\lambda C_2 + 1) \in \mathbb{R}^{\geq 0}$ and dividing by $|H_1|$, one obtains

$$\left|M_o^{\Phi_q^{\nu r}}\right| \geq \left|L_o^{\Phi_q^{\nu r}}\right| - C_3 q^{\lambda\nu r} \quad \forall r \in \mathbb{N}, \quad r \geq r_o.$$

Bearing in mind $\left|L_o^{\Phi_q^{\nu r}}\right| \geq (q^{\nu r}+1) - s \geq (q^{\nu r}+1) - q^{\lambda\nu r}$ for $r \geq \frac{\log_q(s)}{\lambda\nu}$, one concludes that

$$\left|M_o^{\Phi_q^{\nu r}}\right| \geq (q^{\nu r}+1) - (C_3+1) q^{\lambda\nu r} \quad \forall r \in \mathbb{N}, \quad r \geq r_o.$$

Combining with $\left|M^{\Phi_q^{\nu r}}\right| \geq \left|M_o^{\Phi_q^{\nu r}}\right|$, one verifies (5.17) with $C := C_3 + 1$ and concludes the proof of the theorem. □

According to Proposition 22, Lemma 25 (iv) and Proposition 27 (iii), any smooth irreducible curve $X/\mathbb{F}_q \subset \mathbb{P}^n(\overline{\mathbb{F}_q})$ of genus $g \geq 1$ satisfies the assumptions of Theorem 29 with $\lambda = \frac{1}{2}$ as a locally finite $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$-module. Here is an example of a locally finite $\mathfrak{G}$-module M, which is subject to the assumptions of Theorem 29 with $\lambda = 0$. Therefore M satisfies the Riemann Hypothesis Analogue with respect to $\mathbb{P}^1(\overline{\mathbb{F}_q})$ as a $\mathfrak{G}$-module and is not isomorphic (as a $\mathfrak{G}$-module) to a smooth irreducible projective curve, defined over $\mathbb{F}_q$.

Proposition 30. *For any finite field $\mathbb{F}_q$ and $\forall x_1 \in \mathbb{F}_{q^2} \setminus \mathbb{F}_q$ the quasi-affine curve $M := \overline{\mathbb{F}_q} \setminus \{x_1, x_1^q\}$, defined over $\mathbb{F}_{q^2}$ is a locally finite $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$-module with*

$$\zeta_M(t) = \frac{(1-t)(1+t)}{1-qt}, \tag{5.18}$$

which satisfies the assumptions of Theorem 29. Thus, M is subject to the Riemann Hypothesis Analogue with respect to $\mathbb{P}^1(\overline{\mathbb{F}_q})$ as a module over $\mathfrak{G}$ and M is not isomorphic (as a $\mathfrak{G}$-module) to a smooth irreducible projective curve $X/\mathbb{F}_q \subset \mathbb{P}^n(\overline{\mathbb{F}_q})$ of genus $g \geq 1$, defined over $\mathbb{F}_q$.

Proof. The identical inclusion $\mathrm{Id} : M \hookrightarrow \mathbb{P}^1(\overline{\mathbb{F}_q}) = \overline{\mathbb{F}_q} \cup \{\infty\}$ is a finite unramified covering of $\mathfrak{G}$-modules of degree 1 over its image. It has a Galois closure $(M, \{\mathrm{Id}_M\}, \{\mathrm{Id}_M\})$. If $\zeta_M(t)$ is given by (5.18) then

$$P_M(t) := \frac{\zeta_M(t)}{\zeta_{\mathbb{P}^1(\overline{\mathbb{F}_q})}(t)} = (1-t)^2(1+t) \in \mathbb{Z}[t]$$

is a polynomial of $\deg P_M(t) = 3$ with $a_3 = \mathrm{LC}(P_M(t)) = 1$, so that $\lambda := \log_q \sqrt[3]{|a_3|} = 0$. Since M is a $\mathfrak{G}$-submodule of $\mathbb{P}^1(\overline{\mathbb{F}_q})$ with $\left|\mathbb{P}^1(\overline{\mathbb{F}_q}) \setminus M\right| = 3 < \infty$, the relative order $\mathrm{ord}_{\mathfrak{G}}(M/\mathbb{P}^1(\overline{\mathbb{F}_q})) = \mathrm{ord}_{\mathfrak{G}}(M/M) = 0 = \lambda$ by Lemma 25 (i) and M is subject to the assumptions of Theorem 29. If M were isomorphic to a smooth irreducible curve $X/\mathbb{F}_q \subset \mathbb{P}^n(\overline{\mathbb{F}_q})$ as a module over $\mathfrak{G}$ then $P_M(t) = P_X(t) \in \mathbb{Z}[t]$ would have an even degree $\deg P_M(t) = 2g \in \mathbb{N}$ and $\lambda := \log_q \sqrt[2g]{|\mathrm{LC}(P_M(t))|} = \frac{1}{2}$, which contradicts (5.18).

Towards the calculation of $\zeta_M(t)$, let us note that $\overline{\mathbb{F}_q}$ is a locally finite $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$-module and $\mathrm{Orb}_{\mathfrak{G}}(x_1) = \{x_1, x_1^q\}$, in order to conclude that M is a locally finite $\mathfrak{G}$-module. Moreover, $x_1, x_1^q \in \overline{\mathbb{F}_q}^{\Phi_q^{2r}} = \mathbb{F}_{q^{2r}}$ and $x_1, x_1^q \notin \overline{\mathbb{F}_q}^{\Phi_q^{2r+1}} = \mathbb{F}_{q^{2r+1}}$ for $\forall r \in \mathbb{Z}^{\geq 0}$. Therefore $\left|M^{\Phi_q^{2r}}\right| = \left|\overline{\mathbb{F}_q}^{\Phi_q^{2r}}\right| - 2 = q^{2r} - 2$, $\forall r \in \mathbb{N}$, $\left|M^{\Phi_q^{2r+1}}\right| = \left|\overline{\mathbb{F}_q}^{\Phi_q^{2r+1}}\right| = q^{2r+1}$, $\forall r \in \mathbb{Z}^{\geq 0}$, whereas

$$\begin{aligned}
\log \zeta_M(t) &= \sum_{r=1}^{\infty} \left|M^{\Phi_q^r}\right| \frac{t^r}{r} = \sum_{r=1}^{\infty} (q^{2r} - 2)\frac{t^{2r}}{2r} + \sum_{r=0}^{\infty} q^{2r+1} \frac{t^{2r+1}}{2r+1} \\
&= \sum_{r=1}^{\infty} q^r \frac{t^r}{r} - \sum_{r=1}^{\infty} \frac{t^{2r}}{r} = \log\left(\frac{1}{1-qt}\right) - \log\left(\frac{1}{1-t^2}\right) = \log\left(\frac{1-t^2}{1-qt}\right),
\end{aligned}$$

by (3.1). That suffices for(5.18). □

The next corollary establishes that the Riemann Hypothesis Analogue with respect to the projective line $\mathbb{P}^1(\overline{\mathbb{F}_q})$ for a locally finite $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$-module M implies a functional equation for the polynomial ζ-quotient $P_M(t) = \frac{\zeta_M(t)}{\zeta_{\mathbb{P}^1(\overline{\mathbb{F}_q})}(t)} \in \mathbb{Z}[t]$.

Corollary 31. *Let M be an infinite locally finite module over $\mathfrak{G} = \mathrm{Gal}(\overline{\mathbb{F}_q}/\mathbb{F}_q)$, which satisfies the Riemann Hypothesis Analogue with respect to $\mathbb{P}^1(\overline{\mathbb{F}_q})$. Then the polynomial ζ-quotient $P_M(t) = \frac{\zeta_M(t)}{\zeta_{\mathbb{P}^1(\overline{\mathbb{F}_q})}(t)} = \sum\limits_{j=0}^{d} a_j t^j \in \mathbb{Z}[t]$ of M satisfies the functional equation*

$$P_M(t) = \mathrm{sign}(a_d) P_M\left(\frac{1}{q^{2\lambda}t}\right) q^{\lambda d} t^d \quad \textit{for} \quad \lambda := \log_q \sqrt[d]{|a_d|}.$$

Proof. If $P_M(t) = \prod\limits_{j=1}^{d}(1 - q^{\lambda} e^{i\varphi_j} t)$ for some $\varphi_j \in [0, 2\pi)$ then the leading coefficient $\mathrm{LC}(P_M(t)) = a_d = (-1)^d q^{\lambda d} e^{i\left(\sum\limits_{j=1}^{d} \varphi_j\right)}$, whereas

$$P_M\left(\frac{1}{q^{2\lambda}t}\right) = \frac{a_d}{q^{2\lambda d} t^d} \prod_{j=1}^{d}(1 - q^{\lambda} e^{-i\varphi_j} t).$$

The polynomial $P_M(t) \in \mathbb{Z}[t]$ has real coefficients and is invariant under the complex conjugation. Thus, the sets $\left\{e^{i\varphi_j} \mid 1 \le j \le d\right\} = \left\{e^{-i\varphi_j} \mid 1 \le j \le d\right\}$ coincide when counted with multiplicities and $P_M(t) = \prod\limits_{j=1}^{d}(1 - q^{\lambda} e^{-i\varphi_j} t)$. That allows to express

$$P_M\left(\frac{1}{q^{2\lambda}t}\right) = \frac{a_d}{q^{2\lambda d}} P_M(t) t^{-d}.$$

Making use of $|a_d| = q^{\lambda d}$ and $a_d = \mathrm{sign}(a_d)\,|a_d|$, one concludes that

$$P_M\left(\frac{1}{q^{2\lambda}t}\right) = \frac{\mathrm{sign}(a_d)}{q^{\lambda d}} P_M(t) t^{-d}.$$ □

ACKNOWLEDGEMENT. This work was partially supported by the Sofia University Research Fund under Contract 144/2015, Contract 57/12.04.2016 and Contract 80-10-74/20.04.2017.

6. REFERENCES

[1] Bombieri, E.: Counting Points on Curves over Finite Fields. *Sém. Bourbaki* **430**, 1972/73.

[2] Stichtenoth, H.: *Algebraic Function Fields and Codes.* Springer-Verlag, Berlin, Heidelberg, 1993.

[3] Grothendieck, A.: *Séminaire de géométrie algébrique, 1: Revêtements étales et groupe fondamental, 1960-1961.* Lecture Notes in Mathematics **224**, Springer-Verlag, Berlin, 1971.

[4] Duursma, I.: From weight enumerators to zeta functions. *Discrete Appl. Math.*, **111**, 2001, 55-73.

[5] Niederreiter, H., Xing, Ch.: *Algebraic geometry in Coding Theory and Cryptography.* Princeton University Press, Princeton and Oxford, 2009).

[6] Kasparian, A., Marinov, I.: Mac Williams identities for linear codes as Riemann-Roch Conditions. *Electronic Notes in Discrete Mathematics*, **57**, 2017, 121–126.

[7] Shafarevich, I. R.: *Basics of Algebraic Geometry.* Science, Moscow, 1988.

[8] Mustaţă, M.: *Zeta Functions in Algebraic Geometry.* Lecture Notes of Mihnea Popa.

Received on July 17, 2017

Azniv Kasparyan, Ivan Marinov

Faculty of Mathematics and Informatics
"St. Kl. Ohridski" University of Sofia
5, J. Bourchier blvd., BG-1164 Sofia
BULGARIA

E-mails: kasparia@fmi.uni-sofia.bg
ivanm@fmi.uni-sofia.bg

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 104

ANNUAL OF SOFIA UNIVERSITY „ST. KLIMENT OHRIDSKI“
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 104

ON A DIFFERENTIAL INEQUALITY

ROSSEN NIKOLOV

We show that the existing methods for estimating the distance of a two-dimensional normed space with modulus of smoothness of power type 2 to the Euclidean space generated by the John sphere of the former, are not exact. To this end, we construct a class of simple counterexamples in the plane.

Keywords: Banach spaces geometry, moduli of convexity and smoothness, Banach–Mazur distance.

2000 Math. Subject Classification: 46B03, 46B20.

1. INTRODUCTION

The moduli of convexity and smoothness of a Banach space X:

$$\delta_X(\varepsilon) := \inf\left\{1 - \left\|\frac{x+y}{2}\right\| : \ \|x\| = \|y\| = 1, \ \|x-y\| = \varepsilon\right\}, \quad 0 \le \varepsilon \le 2,$$

and

$$\rho_X(\tau) := \sup\left\{\frac{\|x+\tau y\| + \|x-\tau y\| - 2}{2} : \ \|x\| = \|y\| = 1\right\}, \quad \tau \ge 0,$$

respectively, are fundamental concepts in Banach space theory. The duality between them is given by Lindenstrauss formula, see e.g. [6, p. 61]

$$\rho_{X^*}(\tau) = \sup\left\{\frac{\tau\varepsilon}{2} - \delta_X(\varepsilon) : \ 0 \le \varepsilon \le 2\right\}.$$

According to the Nordlander Theorem [7], a Hilbert space H is in a sense the most convex and the most smooth among Banach spaces, that is, for any Banach space X

$$\delta_X(\varepsilon) \leq \delta_H(\varepsilon) = 1 - \sqrt{1 - \varepsilon^2/4} = \varepsilon^2/8 + \mathcal{O}(\varepsilon^4)$$

and

$$\rho_X(\tau) \geq \rho_H(\tau) = \sqrt{1 + \tau^2} - 1 = \tau^2/2 + \mathcal{O}(\tau^4).$$

The Taylor expansion is written down not only for sake of greater clarity, but also because the asymptotic behaviors at 0 play an important role.

For technical reasons, further we concentrate on the asymptotic behavior at 0 of the modulus of smoothness. The results concerning the modulus of convexity can be derived through the Lindenstrauss formula.

Let $a \geq 0$ and let $\mathcal{X}_a$ be the class of all Banach spaces X such that

$$\rho_X(\tau) = \frac{1+a}{2}\tau^2 + o(\tau^2).$$

Several authors independently showed that $\mathcal{X}_0$ contains only Hilbert spaces [5, 9, 8].

So, it stands to reason that for small $a > 0$ the spaces in $\mathcal{X}_a$ might be close to a Hilbert space in some sense. This is indeed so and the sense is made precise below.

Recall that the Banach-Mazur distance between two isomorphic Banach spaces Y and Z is given by

$$d(Y, Z) = \inf\{\|T\|.\|T^{-1}\| : \ T : Y \to Z \text{ arbitrary isomorphism}\}.$$

Now, for a Banach space X one defines

$$d_2(X) := \inf\{d(Y, l_2^{(2)}) : \ Y \subset X, \ \dim Y = 2\},$$

where $l_2^{(2)}$ denotes the two-dimensional Hilbert space, or, in other words, the Euclidean plane.

Obviously, $d_2(X)$ measures how far from an ellipse the two-dimensional sections of the sphere of X are. The famous Jordan-von Neumann Theorem [3] reads $d_2(X) = 1$ if and only if X is a Hilbert space. The measure $d_2(X)$ is very useful for estimating the type and cotype of X.

The standard way of estimating $d_2(X)$ is through the use of the John Sphere, [4]. In the pioneering work [4] John showed that $d_2(X) \leq \sqrt{2}$ for any X.

Elaborating on this idea, let Y be a two-dimensional space. It is clear that there is an ellipse , say $\mathcal{E}$, of maximal volume contained in the unit ball of Y. Define

$$j(Y) := \max_{x \in \mathcal{E}} \|x\|^{-1}. \tag{1}$$

Then, of course,

$$d_2(X) \le \sup\{j(Y) : \ Y \subset X, \ \dim Y = 2\}.$$

Let for $a \ge 0$

$$g(a) := \sup\{j(Y) : \ Y \in \mathcal{X}_a, \ \dim Y = 2\}.$$

Then

$$d_2(X) \le g(a), \quad \forall a \ge 0, \ \forall X \in \mathcal{X}_a.$$

From the above considerations we know that $g(0) = 1$. Rakov [8] estimated $j(Y)$ for $Y \in \mathcal{X}_a$, his estimate as $a \to 0$ reads

$$g(a) \le 1 + k\sqrt{a}$$

with some constant k which is not important here. An asymptotically sharp estimate was given in [1, 2]:

$$g(a) \le 1 + \frac{a}{\sqrt{2} + (1+\sqrt{2})a}. \tag{2}$$

From the method of [1, 2] it is not clear if (2) is exact. In this work we show that it is not.

We will explain briefly the method of [1, 2].

Let Y be a two-dimensional space in $\mathcal{X}_a$ for some $a > 0$. We may assume that Y is realized in such a way in $\mathbb{R}^2$ that the unit circle $x_1^2 + x_2^2 = 1$ of $\mathbb{R}^2$ is the John sphere of X. Denote the standard basis of $\mathbb{R}^2$ by $e_1 = (1, 0)$ and $e_2 = (0, 1)$. Define

$$r(\sigma) := \|e_1 \cos\sigma + e_2 \sin\sigma\|.$$

Then in this polar annotation the unit sphere S_Y of Y is

$$S_Y = \left\{ \frac{1}{r(\sigma)}(\cos\sigma, \sin\sigma) : \ \sigma \in [-\pi, \pi] \right\}.$$

If $x, y \in S_Y$, $x = r^{-1}(\theta)(e_1 \cos\theta + e_2 \sin\theta)$, $y = r^{-1}(\varphi)(e_1 \cos\varphi + e_2 \sin\varphi)$, then Lemma 3.2 from [1] states that if $r''(\theta)$ exists, then

$$\lim_{\tau \to 0} \frac{\|x + \tau y\| + \|x - \tau y\| - 2}{\tau^2} = \frac{\sin^2(\theta - \varphi)}{r^2(\varphi)} r(\theta)(r(\theta) + r''(\theta)).$$

Since $Y \in \mathcal{X}_a$ we have that for almost all $\theta \in [0, 2\pi]$

$$\sup_{\varphi \in [0, 2\pi]} \frac{\sin^2(\theta - \varphi)}{r^2(\varphi)} r(\theta)(r(\theta) + r''(\theta)) \le 1 + a. \tag{3}$$

From John Theorem it follows that on each arc of the unit circle of length $\pi/2$ there is a contact point, that is such that $r = 1$. Therefore, without loss of generality, there is $\alpha \in (0, \pi/2]$ such that

$$r(0) = r(\alpha) = 1, \tag{4}$$

and

$$j(Y) = \max_{\theta\in[0,\alpha]} \frac{1}{r(\theta)}. \tag{5}$$

So, we may consider the system (3), (4) and try to estimate $j(Y)$.

However, with φ in (3) the problem is non-local and probably very difficult. In order to handle it [1, 2] substitute $\varphi = \theta + \pi/2$ and use $r(\varphi) \leq 1$ to derive from (3)

$$r(\theta)(r(\theta) + r''(\theta)) \leq 1 + a \quad \text{for almost all } \theta \in [0, 2\pi]. \tag{6}$$

Then the system (6), (4) is used to estimate $j(Y)$ through (5).

We will demonstrate that for a close to zero the outlined approach can never produce the exact value of $\sup j(Y)$ for $Y \in \mathcal{X}_a$, denoted above as $g(a)$.

For sake of clarity, for $a \geq 0$ denote by G_a the class of all π-periodic functions $r = r(\theta)$, $0 \leq \theta \leq \pi$, such that

(i) $0 < r(\theta) \leq 1, \quad r(0) = r(\pi/2) = 1;$

(ii) $r'(\theta)$ is absolutely continuous and $0 \leq r(\theta)(r(\theta) + r''(\theta)) \leq 1 + a$ almost everywhere;

(iii) the region B_r inside the curve

$$S_r = \left\{ \frac{1}{r(\theta)}(\cos\theta, \sin\theta) : \ \sigma \in [-\pi, \pi] \right\}$$

is convex.

(It is easy to check that (iii) follows from $r + r'' \geq 0$, which is contained in (ii), but we do not need this fact.)

Finally we introduce the class F_a of all π-periodic functions, which satisfy (i), (ii), (iii) and additionally

(iv)

$$\sup_{\varphi,\theta} \frac{\sin^2(\varphi - \theta)}{r^2(\varphi)} r(\theta)\big(r(\theta) + r''(\theta)\big) = 1 + a.$$

It is clear that $F_a \subset G_a$.

Theorem 1.1. *There exist an interval I and a class $X_a\left(\mathbb{R}^2, \|.\|_a\right) \in \mathcal{X}_a, \ a \in I$ of Banach spaces, such that for all $a \in I$ there are $b = b(a) > a$ which satisfy:*

(i) $r_b \in G_a, \ r_b \in F_b, \ r_b \notin F_a$;

(ii)

$$\max_{\sigma} \frac{1}{r_a(\sigma)} = d_2\left(X_a\right) < \max_{\sigma} \frac{1}{r_b(\sigma)} = d_2\left(X_b\right),$$

where $r_a(\sigma) = \|\cos\sigma e_1 + \sin\sigma e_2\|_a$, $r_b(\sigma) = \|\cos\sigma e_1 + \sin\sigma e_2\|_b$.

2. CONSTRUCTION OF A CLASS OF TWO-DIMENSIONAL SPACES

Pick $\lambda \in [0,1]$ and set $\mu = 1-\lambda$, $\nu = 2\mu^2 - \lambda^2 = \lambda^2 - 4\lambda + 2$.
For $\theta_1, \theta_2 = \pm 1$ we denote with D_{θ_1,θ_2} the Euclidean disk of radius λ centered at $O_{\theta_1,\theta_2} = (\theta_1\mu, \theta_2\mu)$, i.e.

$$D_{\theta_1,\theta_2} = \left\{x = (x_1, x_2) \in \mathbb{R}^2 : \ (x_1 - \theta_1\mu)^2 + (x_2 - \theta_2\mu)^2 \le \lambda^2\right\}.$$

Also let

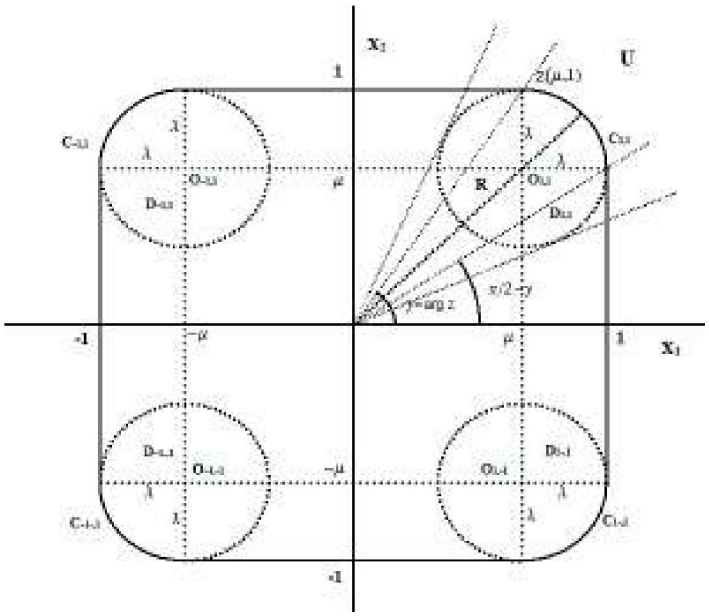
$$C_{\theta_1,\theta_2} = \left\{x = (x_1, x_2) \in \mathbb{R}^2 : \ (x_1 - \theta_1\mu)^2 + (x_2 - \theta_2\mu)^2 = \lambda^2\right\}.$$

and

$$D = D_{1,1} \cup D_{1,-1} \cup D_{-1,1} \cup D_{-1,-1}, \quad B_\lambda = convD$$

We can say that B_λ is a rotund square (see Figure 1). Clearly B_1 is the unit (Euclidean) disk.

Figure 1: A rotund square



It is easy to see that $B_0 = Q$ where Q is the unit square, i.e.

$$Q = \left\{x = (x_1, x_2) \in \mathbb{R}^2 : \ |x_1| \leq 1, \ |x_2| \leq 1\right\}.$$

We set $Y_\lambda = \left(\mathbb{R}^2, \|.\|_\lambda\right)$, where $\|.\|_\lambda$ is the Minkowski functional of B_λ, i.e.

$$\|x\|_\lambda = \inf\left\{t > 0 : \ \frac{x}{t} \in B_\lambda\right\}.$$

Since B_λ is symmetric with respect to the coordinate system, the line $x_1 = x_2$ and the origin $O(0,0)$, we have that (x_1, x_2), $(\theta_1 x_1, \theta_2 x_2) \in B_\lambda$, $\theta_1, \theta_2 = \pm 1$ provided $(x_2, x_1) \in B_\lambda$. So

$$\|x_2 e_1 + x_1 e_2\|_\lambda = \|\theta_1 x_1 e_1 + \theta_2 x_2 e_2\|_\lambda,$$

where e_1, e_2 is the unit vector basis in $\mathbb{R}^2$, $e_1 = (1,0)$, $e_2 = (0,1)$.

This implies monotonicity of the basis, i.e. $\|x_1 e_1 + x_2 e_2\|_\lambda \leq \|y_1 e_1 + y_2 e_2\|_\lambda$ whenever $|x_1| \leq |y_1|$, $|x_2| \leq |y_2|$.

Let us mention that $d\left(Y_\lambda, l_2^{(2)}\right)$ is the radius of the circumcircle of B_λ centered at the origin. Thus

$$d_2\left(Y_\lambda\right) = d\left(Y_\lambda, l_2^{(2)}\right) = R = \sqrt{2}\mu + \lambda = \sqrt{2} + \left(1 - \sqrt{2}\right)\lambda.$$

In order to find the asymptotic behavior of $\rho_{Y_\lambda}(\tau)$ at O, we need an explicit formula for the norm of Y_λ. Fix $\lambda \in (0,1]$. Further we omit the index λ, i.e. we write Y instead of Y_λ, $\|.\|$ instead of $\|.\|_\lambda$, B and S stand for the unit ball and the unit sphere of Y_λ. Let $x = (\rho\cos\sigma, \rho\sin\sigma)$ be the representation of $x \in \mathbb{R}^2$ in polar coordinates. We set $\sigma = \arg x$. Having in mind the symmetry of B, it is enough to find an explicit formula for $\|x\|, x = (x_1, x_2)$, when $0 \leq x_1 \leq x_2$, i.e. $\frac{\pi}{4} \leq \arg x \leq \frac{\pi}{2}$. Denote by $z(\mu, 1)$ the unique common point of the circle

$$C_{1,1} = \left\{(x_1, x_2) \in \mathbb{R}^2 : \ (x_1 - \mu)^2 + (x_2 - \mu)^2 = \lambda^2\right\}$$

and the straight line $l = \left\{(x_1, x_2) \in \mathbb{R}^2 : \ x_2 = 1\right\}$.

Set $\gamma = \arg z$. Obviously $\frac{\pi}{4} < \gamma \leq \frac{\pi}{2}$ and $\|x\| = x_2$ for all points x with $\arg x \in [\gamma, \frac{\pi}{2}]$. If $x \in [\frac{\pi}{2} - \gamma, \gamma]$, then the vector $x/\|x\|$ belongs to the circle $C_{1,1}$. Setting $f(x_1, x_2) = \|x\| = \|(x_1, x_2)\|$, we obtain

$$\left(\frac{x_1}{f} - \mu\right)^2 + \left(\frac{x_2}{f} - \mu\right)^2 = \lambda^2.$$

Calculating the roots of the above equation we get :

$$f(x_1, x_2) = \begin{cases} \frac{1}{\nu}\left(\mu(x_1 + x_2) - \sqrt{\lambda^2(x_1^2 + x_2^2) - \mu^2(x_1 - x_2)^2}\right), & \lambda \neq 2 - \sqrt{2} \\ \frac{1}{2\mu}\left(x_1 + x_2 - \frac{2x_1 x_2}{x_1 + x_2}\right), & \lambda = 2 - \sqrt{2}. \end{cases}$$

So :

$$\|x\| = \begin{cases} x_1 & if \ \ 0 \le \arg x \le \frac{\pi}{2} - \gamma \\ f(x_1, x_2) & if \ \ \frac{\pi}{2} - \gamma \le \arg x \le \gamma \\ x_2 & if \ \ \gamma \le \arg x \le \frac{\pi}{2}. \end{cases}$$

We mention here that the function f is defined not only on the sector $\left\{x \in \mathbb{R}^2 : \frac{\pi}{2} - \gamma \le \arg x \le \gamma\right\}$. Actually f is defined on the set

$$E = \left\{x \in \mathbb{R}^2 : \ \lambda^2(x_1^2 + x_2^2) \ge \mu^2(x_1 - x_2)^2\right\} \supset \left\{x \in \mathbb{R}^2 : \ \frac{\pi}{2} - \gamma \le \arg x \le \gamma\right\}.$$

It is easy to see that $E \supset \left\{x \in \mathbb{R}^2 : \ 0 \le \arg x \le \frac{\pi}{2}\right\}$ for $\lambda \ge 1/2$, while

$$E = \left\{x \in \mathbb{R}^2 : \ k_1 \le \frac{x_2}{x_1} \le k_2\right\},$$

where $k_1 < k_2$ are the roots of

$$\left(\mu^2 - \lambda^2\right) k^2 - 2\mu^2 k + \left(\mu^2 - \lambda^2\right) = 0 \quad \text{for} \quad 0 < \lambda < 1/2.$$

Fact 2.1. *For all $x \in E$ we have:*

(i) $f(x) \ge \|x\|$ *for* $x \in E \cap \left\{x : \ 0 \le \arg x \le \frac{\pi}{2}\right\}$

(ii) If $x(x_1, x_2) \in S \cap \left\{x : \ \frac{\pi}{2} - \gamma \le \arg x \le \gamma\right\}$, *then*

$$f''_{11}(x_1, x_2) = \frac{\lambda^2 x_2^2}{g(x_1, x_2)}; \quad f''_{12}(x_1, x_2) = -\frac{\lambda^2 x_1 x_2}{g(x_1, x_2)};$$
$$f''_{22}(x_1, x_2) = \frac{\lambda^2 x_1^2}{g(x_1, x_2)},$$

where $g(x_1, x_2) = (\mu(x_1 + x_2) - \nu)^3$.

Proof. We prove only (i). Since $f(x_1, x_2)$ is homogeneous, i.e. $f(kx_1, kx_2) = kf(x_1, x_2)$ for all $k > 0$, it suffices to check (i) only for $x \in S$.

If $x \in \left\{u \in \mathbb{R}^2 : \ \frac{\pi}{2} - \gamma \le \arg u \le \gamma\right\}$, then $f(x) = 1 = \|x\|$.

If $x \in E \cap \left\{u \in \mathbb{R}^2 : \ \gamma \le \arg u \le \frac{\pi}{2}\right\}$, then $x_2 = 1$. From $\left(\frac{x_1}{f}, \frac{x_2}{f}\right) \in C_{1,1}$ it follows $\frac{1}{f} < 1$, i.e. $f(x) = f(x_1, x_2) > 1 = \|x\|$. □

Set

$$\Delta_2(x, y, \tau) = \frac{1}{2}\left(\|x + \tau y\| + \|x - \tau y\| - 2\|x\|\right).$$

Evidently,

$$\rho_Y(\tau) = \sup\left\{\Delta_2(x, y, \tau) : \ x, y \in S\right\}.$$

Due to the symmetry of S:

$$\rho_Y(\tau) = \sup\left\{\Delta_2(x, y, \tau) : \ x, y \in S, \ \arg x \in [\frac{\pi}{4}, \frac{\pi}{2}]\right\}.$$

Fact 2.2. *Let $x, y \in S$, $\arg x \in [\gamma, \frac{\pi}{2}]$, $|\tau| \le \frac{\mu}{2}$. Then*

$$\Delta_2(x, y, \tau) \le \Delta_2(z, y, \tau).$$

Proof. Since $\|y\| = 1$, we get $|y_1| \le 1$. So $|\tau y_1| \le \frac{\mu}{2}$. Since $\arg x \in [\gamma, \frac{\pi}{2}]$, we have $0 \le x_1 \le \mu$ and $x_2 = 1$, $|x_1 \pm \tau y_1| \le \mu \pm \tau y_1$. The monotonicity of the basis implies:

$$\begin{aligned}\|x \pm \tau y\| &= \|(x_1 \pm \tau y_1)e_1 + (1 \pm \tau y_2)e_2\| \\ &\le \|(\mu \pm \tau y_1)e_1 + (1 \pm \tau y_2)e_2\| = \|z \pm \tau y\|.\end{aligned}$$

~ □

Corollary 2.3. *If $|\tau| \le \frac{\mu}{2}$, then*

$$\rho_Y(\tau) = \sup\{\Delta_2(x, y, \tau) : \ x \in A, \ y \in S\},$$

where A is the arc$\{x \in S, \ \frac{\pi}{4} \le \arg x \le \gamma\}$.

Proposition 2.4. *The following estimate holds:*

$$\overline{\lim_{\tau \to 0}} \frac{\rho_Y(\tau)}{\tau^2} \le \frac{\lambda^2}{2} \sup\left\{ \left| \begin{matrix} x_1 & x_2 \\ y_1 & y_2 \end{matrix} \right|^2 \frac{1}{g(x_1, x_2)} : \ x \in A, \ y \in S \right\}.$$

Proof. Pick a convex compact set $F \subset E$ such that its interior contains the arc A. Choose $\tau_F \in \left(0, \frac{\mu}{2}\right)$ in such a way that $x \pm \tau y \in F$ whenever $x \in A$, $y \in S$, $|\tau| \le \tau_F$.

Set

$$\Delta_2 f(x, y, \tau) = \frac{1}{2}\big(f(x + \tau y) + f(x - \tau y) - 2f(x)\big).$$

From Fact 2.1(i) we have

$$f(x \pm \tau y) \ge \|x \pm \tau y\|, \quad x \in A, \quad y \in S, \quad |\tau| \le \tau_F.$$

Since $f(x) = \|x\|$ for $x \in A$ we get

$$\Delta_2(x, y, \tau) \le \Delta_2 f(x, y, \tau)$$

whenever $x \in A$, $y \in S$, $|\tau| \le \tau_F$. Using that, we get for $\tau \in (0, \tau_F]$

$$\rho_Y(\tau) \le \sup\{\Delta_2 f(x, y, \tau) : \ x \in A, \ y \in S\}.$$

Take $x \in A, \ y \in S$. Applying Taylor's formula to $\varphi(\tau) = f(x + \tau y) - f(x)$ and $\psi(\tau) = f(x - \tau y) - f(x)$, we can find $\theta_1 = \theta_1(x, y, \tau)$, $\theta_2 = \theta_2(x, y, \tau) \in (0, 1)$ in such a way that

$$\begin{aligned}\frac{\Delta_2 f(x, y, \tau)}{\tau^2} =& \frac{1}{4}\left\{\left(f''_{11}(x + \theta_1 \tau y)y_1^2 + 2f''_{12}(x + \theta_1 \tau y)y_1 y_2 + f''_{22}(x + \theta_1 \tau y)y_2^2\right)\right\} \\ &+ \frac{1}{4}\left\{\left(f''_{11}(x + \theta_2 \tau y)y_1^2 + 2f''_{12}(x + \theta_2 \tau y)y_1 y_2 + f''_{22}(x + \theta_2 \tau y)y_2^2\right)\right\}.\end{aligned}$$

Having in mind that the second derivatives are uniformly continuous on F, we get :

$$\overline{\lim_{\tau\to 0}}\,\frac{\rho_Y(\tau)}{\tau^2} \le \frac{1}{2}\left\{\left(f''_{11}(x)y_1^2 + 2f''_{12}(x)y_1y_2 + f''_{22}(x)y_2^2 :\ x\in A,\ y\in S\right)\right\}$$

To finish the proof, it is enough to use Fact 2.1 (ii). □

Lemma 2.5. *Let $x = (x_1, x_2) \in A$. Then*

$$\sup_{(y_1,y_2)\in S}\begin{vmatrix} x_1 & x_2 \\ y_1 & y_2 \end{vmatrix}^2 = \left(\mu(x_1+x_2) + \lambda\sqrt{x_1^2+x_2^2}\right)^2.$$

Proof. The determinant represents the oriented area of the parallelogram, defined by the vectors x and y. Therefore, for a fixed x, the left-hand side achieves its greatest value when the distance from the point (y_1, y_2) to the support of the vector x is maximal. This is satisfied for some $\bar{y} \in C_{1,-1} \cap S$ or $\bar{y} \in C_{-1,1} \cap S$. Without loss of generality we assume that $\bar{y} \in C_{1,-1} : (s-\mu)^2 + (t+\mu)^2 = \lambda^2$. The tangent to $C_{1,-1}$ at the point $(\bar{y}_1, \bar{y}_2)$ is parallel to the support of x, i.e. the normal to $C_{1,-1}$ is orthogonal to x. Therefore the scalar product $< v, x >= 0$, where $v = (\bar{y}_1 - \mu, \bar{y}_2 + \mu)$. We get for $\bar{y}_1, \bar{y}_2$ the system:

$$\begin{cases} x_1(\bar{y}_1 - \mu) + x_2(\bar{y}_2 + \mu) = 0 \\ (\bar{y}_1 - \mu)^2 + (\bar{y}_2 + \mu)^2 = \lambda^2 \end{cases}$$

with solution:

$$\begin{cases} \bar{y}_1 = \mu + \frac{\lambda x_2}{\sqrt{x_1^2+x_2^2}} \\ \bar{y}_2 = -\mu - \frac{\lambda x_1}{\sqrt{x_1^2+x_2^2}}. \end{cases}$$

Hence,

$$\begin{vmatrix} x_1 & x_2 \\ \bar{y}_1 & \bar{y}_2 \end{vmatrix} = x_1\bar{y}_2 - x_2\bar{y}_1 = -\mu(x_1+x_2) - \lambda\sqrt{x_1^2+x_2^2}.$$

□

Proposition 2.6. *Let $0 < \lambda < 2 - \sqrt{2}$. Then*

$$\overline{\lim_{\tau\to 0}}\,\frac{\rho_Y(\tau)}{\tau^2} \le \frac{1}{2}h(\lambda),$$

where

$$h(\lambda) = \frac{1}{\lambda}\left(\lambda^2 - 3\lambda + 2 + \lambda\sqrt{\lambda^2 - 2\lambda + 2}\right)^2.$$

Proof. From Proposition 2.4 and Lemma 2.5 it follows

$$\overline{\lim_{\tau\to 0}}\,\frac{\rho_Y(\tau)}{\tau^2} \le \frac{\lambda^2}{2}\sup\left\{\frac{\left(\mu(x_1+x_2) + \lambda\sqrt{x_1^2+x_2^2}\right)^2}{g(x_1,x_2)} : (x_1,x_2)\in A\right\},$$

where $g(x_1, x_2) = (\mu(x_1 + x_2) - \nu)^3$ is defined in Fact 2.1. For brevity, we denote

$$M(x_1, x_2) = \frac{\left(\mu(x_1 + x_2) + \lambda\sqrt{x_1^2 + x_2^2}\right)^2}{g(x_1, x_2)}.$$

On the arc A we have :

$$x_1^2 + x_2^2 - 2\mu(x_1 + x_2) + 2\mu^2 = \lambda^2,$$

i.e.

$$x_1^2 + x_2^2 = 2\mu(x_1 + x_2) - \left(2\mu^2 - \lambda^2\right) = 2\mu(x_1 + x_2) - \nu.$$

Set $\sqrt{x_1^2 + x_2^2} = t$, then $\mu(x_1 + x_2) = \frac{t^2+\nu}{2}$, and after substituting we get

$$M(x_1, x_2) = \frac{\left(\frac{t^2+\nu}{2} + \lambda t\right)^2}{\left(\frac{t^2+\nu}{2} - \nu\right)^3} = \frac{2\left(t^2 + 2\lambda t + \nu\right)^2}{\left(t^2 - \nu\right)^3}.$$

We need to examine the function

$$m(t) = \frac{\left(t^2 + 2\lambda t + \nu\right)^2}{\left(t^2 - \nu\right)^3}.$$

By the cosine formula we get

$$\sqrt{1 + \mu^2} \leq t \leq \sqrt{2}\mu + \lambda = \sqrt{2} + (1 - \sqrt{2})\lambda.$$

The left-hand side expression represents the distance from the origin $O(0, 0)$ to $z(\mu, 1)$, while the right-hand side expression is the distance to the middle of arc A. From

$$t^2 \geq 1 + \mu^2 = \lambda^2 - 2\lambda + 2 > \lambda^2 - 4\lambda + 2 = \nu$$

it is clear that $m(t)$ is defined in this interval. Calculating m' and simplifying, we obtain:

$$m'(t) = -\frac{2\left(t^2 + 2\lambda t + \nu\right)\left(t^3 + 4\lambda t^2 + 5\nu t + 2\lambda\nu\right)}{\left(t^2 - \nu\right)^4}.$$

We now show that $m' < 0$ for $0 < \lambda < 2 - \sqrt{2}$, i.e. m is decreasing in the interval $I = \left[\sqrt{1 + \mu^2}, \sqrt{2} + (1 - \sqrt{2})\lambda\right]$. Obviously $I \subset \left[1, \sqrt{2}\right]$. The quadratic polynomial $u(t) = t^2 + 2\lambda t + \nu$ is increasing in $[-\lambda, \infty]$, whence

$$u(t) > u(1) = 1 + 2\lambda + \lambda^2 - 4\lambda + 2 = \lambda^2 - 2\lambda + 3 \geq 2 > 0.$$

Obviously $\nu = \lambda^2 - 4\lambda + 2 > 0$. It follows that the coefficients of $v(t) = t^3 + 4\lambda t^2 + 5\nu t + 2\lambda\nu$ are positive, which implies $v(t) > 0$ when $t \in I$. Finally, in order to find the greatest value of M we use :

$$1 + \mu^2 + \nu = 1 + (1 - \lambda)^2 + 2(1 - \lambda)^2 - \lambda^2 = 2\lambda^2 - 6\lambda + 4,$$
$$1 + \mu^2 - \nu = 1 + (1 - \lambda)^2 - 2(1 - \lambda)^2 + \lambda^2 = 2\lambda,$$

i.e.

$$M(\mu,1)=2m\left(\sqrt{1+\mu^2}\right)=\frac{2\left(1+\mu^2+2\lambda\sqrt{1+\mu^2}+\nu\right)^2}{\left(1+\mu^2-\nu\right)^3}$$
$$=\frac{1}{\lambda^3}\left(\lambda^2-3\lambda+2+\lambda\sqrt{\lambda^2-2\lambda+2}\right)^2.$$

The above and the remark at the beginning complete the proof. □

Theorem 2.7. *For $0<\lambda<2-\sqrt{2}$ we have*

$$\lim_{\tau\to 0}\frac{\rho_Y(\tau)}{\tau^2}=\frac{1}{2}h(\lambda).$$

Proof. According to Proposition 2.6, the function m is continuous and decreasing in the interval $\left[\sqrt{1+\mu^2},\sqrt{2}+\left(1-\sqrt{2}\right)\lambda\right]$. Let

$$\epsilon\in\left(0,m\left(\sqrt{1+\mu^2}\right)-m\left(\sqrt{2}+\left(1-\sqrt{2}\right)\lambda\right)\right).$$

There exists

$$z_\epsilon(x_1,x_2)\in A,\quad (x_1=x_1(\epsilon),x_2=x_2(\epsilon)),$$

such that

$$m\left(\sqrt{x_1^2+x_2^2}\right)=m\left(\sqrt{1+\mu^2}\right)-\epsilon.$$

Whence

$$M(z_\epsilon)=M(x_1,x_2)=2m\left(\sqrt{1+\mu^2}\right)-2\epsilon.$$

Choose $\tau_\epsilon>0$, such that

$$\{(p,q):\max(|p-x_1|,|q-x_2|)\le\tau_\epsilon\}\subset\left\{u\in\mathbb{R}^2:\frac{\pi}{2}-\gamma\le\arg u\le\gamma\right\}.$$

If $|\tau|<\tau_\epsilon$, then $\Delta_2(z_\epsilon,y,\tau)=\Delta_2 f(z_\epsilon,y,\tau)$ for all $y\in S$. By Lemma 2.5, similarly as in Proposition 2.4 we get :

$$\underline{\lim}_{\tau\to 0}\frac{\rho_Y(\tau)}{\tau^2}\ge\lim_{\tau\to 0}\sup_{y\in S}\frac{\Delta_2(z_\epsilon,y,\tau)}{\tau^2}=\lim_{\tau\to 0}\sup_{y\in S}\frac{\Delta_2 f(z_\epsilon,y,\tau)}{\tau^2}$$
$$=\frac{\lambda^2}{2}M(x_1,x_2)=\frac{\lambda^2}{2}\left(2m\left(\sqrt{1+\mu^2}\right)-2\epsilon\right)=\frac{1}{2}h(\lambda)-\lambda^2\epsilon,$$

which combined with Proposition 2.6 concludes the proof. □

Remark 2.8. Let us point out that for arbitrary small τ,

$$\rho_Y(\tau)=\sup\left\{\frac{\|x+\tau y\|+\|x-\tau y\|-2}{2},\quad x\in A,\ y\in S\right\}$$

is not attained at the point $z(\mu, 1)$. Indeed, for $\tau \in (0, \tau_\epsilon)$ there holds either

$$\frac{\pi}{2} - \gamma \leq \arg(z + \tau y) \leq \gamma, \quad \gamma \leq \arg(z - \tau y) \leq \frac{\pi}{2}$$

or

$$\frac{\pi}{2} - \gamma \leq \arg(z - \tau y) \leq \gamma, \quad \gamma \leq \arg(z + \tau y) \leq \frac{\pi}{2}.$$

Similarly as in Proposition 2.4, we have

$$\Delta_2(z, y, \tau) = \frac{\tau^2}{4} \left[\left(f''_{11}(z + \theta\tau y) y_1^2 + 2 f''_{12}(z + \theta\tau y) y_1 y_2 + f''_{22}(z + \theta\tau y) y_2^2 \right) \right],$$

where $\theta = \theta(y, \tau) \in (0, 1))$. Thus

$$\limsup_{\tau \to 0} \left\{ \frac{\|z + \tau y\| + \|z - \tau y\| - 2}{2\tau^2}, \ \ y \in S \right\} = \frac{1}{4} h(\lambda).$$

This is because $r''(\sigma)$ does not exists at $\sigma = \gamma$ ($r(\sigma)$ is defined in the Introduction).

3. PROOF OF THE MAIN THEOREM

We start by establishing

Fact 3.1. *The function*

$$h(\lambda) = \frac{1}{\lambda} \left(\lambda^2 - 3\lambda + 2 + \lambda \sqrt{\lambda^2 - 2\lambda + 2} \right)^2$$

is decreasing in $(0, 1]$.

Proof. It is sufficient to check that

$$\tilde{h}(\lambda) = \lambda^2 - 3\lambda + 2 + \lambda \sqrt{\lambda^2 - 2\lambda + 2}$$

is decreasing. The derivative

$$\tilde{h}'(\lambda) = \frac{(2\lambda - 3)\sqrt{\lambda^2 - 2\lambda + 2} + 2\lambda^2 - 3\lambda + 2}{\sqrt{\lambda^2 - 2\lambda + 2}}$$

is negative if

$$(2\lambda - 3)\sqrt{\lambda^2 - 2\lambda + 2} + 2\lambda^2 - 3\lambda + 2 < 0, \quad \lambda \in (0, 1).$$

The latter is equivalent to the inequality

$$\sqrt{\lambda^2 - 2\lambda + 2} \left[\sqrt{\lambda^2 - 2\lambda + 2} + 2\lambda - 3 \right] + \lambda(\lambda - 1) < 0, \qquad \lambda \in (0, 1),$$

which is true because both summands are negative for $\lambda \in (0, 1)$. □

Lemma 3.2. *Let $0 < \lambda < 1$, $\|.\|_\lambda$ correspond to the space Y_λ and*

$$r_\lambda(\theta) = \|\cos\theta e_1 + \sin\theta e_2\|_\lambda$$

be the function which describes the sphere of "rotund square". Then

$$s(\lambda) = \sup_\theta r_\lambda(\theta)\left(r_\lambda(\theta) + r''_\lambda(\theta)\right) \leq \frac{1}{\lambda^2 - 2\lambda + 2}\left(1 + a\left(Y_\lambda\right)\right) < 1 + a\left(Y_\lambda\right),$$

where we have set

$$1 + a\left(Y_\lambda\right) = 2\lim_{\rho\to 0}\frac{\rho_{Y_\lambda}(\tau)}{\tau^2} = h(\lambda).$$

Above, we assumed that $\theta \neq \gamma = \arg z$. Also, θ does not correspond to any other common point of the circle and the straight line, because for such points $r''(\theta)$ does not exist.

Proof. If $x = \frac{1}{r_\lambda(\theta)}(\cos\theta, \sin\theta)$ belongs to a segment of S_λ, then $r_\lambda(\theta) + r''_\lambda(\theta) = 0$. Let $x \in A$ (see Corollary 2.3) and $x \neq z(\mu, 1)$. From

$$\sup_{\theta,\varphi}\frac{\sin^2(\theta-\varphi)}{r^2_\lambda(\varphi)} r_\lambda(\theta)\left(r_\lambda(\theta) + r''_\lambda(\theta)\right) = 1 + a\left(Y_\lambda\right),$$

by substituting $\varphi = \theta - \frac{\pi}{2}$ we get

$$\frac{1}{r^2_\lambda(\varphi)} r_\lambda(\theta)\left(r_\lambda(\theta) + r''_\lambda(\theta)\right) \leq 1 + a\left(Y_\lambda\right).$$

Hence,

$$r_\lambda(\theta)\left(r_\lambda(\theta) + r''_\lambda(\theta)\right) \leq r^2_\lambda(\varphi)\left(1 + a\left(Y_\lambda\right)\right) < \frac{1}{\lambda^2 - 2\lambda + 2}\left(1 + a\left(Y_\lambda\right)\right).$$

Above we have used the inequality

$$r_\lambda(\varphi) < \frac{1}{\|z\|_2}, \quad \text{where} \quad \|z\|_2 = \sqrt{1+\mu^2} = \sqrt{\lambda^2 - 2\lambda + 2}.$$

□

Proof of Theorem 1.1

At the beginning we note that $d_2\left(Y_\lambda\right) = \sqrt{2} + \left(1 - \sqrt{2}\right)\lambda$ is a decreasing function of λ. From Lemma 3.2,

$$s(\lambda) = \sup_\theta r_\lambda(\theta)\left(r_\lambda(\theta) + r''_\lambda(\theta)\right) \leq \frac{1}{\lambda^2 - 2\lambda + 2} h(\lambda) = \frac{1}{\lambda(\lambda^2 - 2\lambda + 2)}\tilde{h}^2(\lambda).$$

We denote the right-hand side with $k(\lambda)$. As $\frac{1}{\lambda(\lambda^2-2\lambda+2)}$ is decreasing in $(0,1)$, $k(\lambda)$ decreases in this interval too, due to Fact 3.1.

Thus for all $\lambda \in (0, 2-\sqrt{2})$ we have

$$s(\lambda) \leq k(\lambda) < h(\lambda) \tag{7}$$

and

$$\lim_{\lambda \to 0^+} k(\lambda) = \lim_{\lambda \to 0^+} h(\lambda) = \infty.$$

Let $a \in I = \left(h(2-\sqrt{2}) - 1, \infty\right)$. There exists a unique $\lambda = \lambda(a) < 2-\sqrt{2}$, such that $a = h(\lambda) - 1 = h(\lambda(a)) - 1$, i.e. $\lambda(a)$ is the inverse function of $a = h(\lambda) - 1$, considered in the interval $(0, 2-\sqrt{2})$. We define $X_a = Y_{\lambda(a)}$, which means $X_a = \left(\mathbb{R}^2, |||.|||_a\right)$, where $|||.|||_a = ||.||_{\lambda(a)}$. Respectively let

$$\tilde{r}_a(\sigma) = |||\cos\sigma e_1 + \sin\sigma e_2|||_a = ||\cos\sigma e_1 + \sin\sigma e_2||_{\lambda(a)} = r_\lambda(\sigma).$$

By definition it is clear that $X_a \in \mathcal{X}_a$ for all $a \in I$. Let $a \in I$ is fixed and $\lambda = \lambda(a)$ is as above. From (7) it follows that there exists a unique $\lambda_1 : 0 < \lambda_1 < \lambda$, for which $a = h(\lambda) - 1 = k(\lambda_1) - 1$. Let $b = h(\lambda_1) - 1$, i.e. $\lambda_1 = \lambda(b)$. Obviously $b > a$. For $r_{\lambda_1}(\sigma)$ we have:

$$r_{\lambda_1}(\sigma)\left(r_{\lambda_1}(\sigma) + r''_{\lambda_1}(\sigma)\right) \leq s(\lambda_1) \leq k(\lambda_1) \leq h(\lambda) = 1 + a = 1 + a\left(Y_\lambda\right).$$

But this is equivalent to $\tilde{r}_b(\sigma) = r_{\lambda_1}(\sigma) \in G_a$. Also it is clear that $\tilde{r}_b(\sigma) \in F_b$ whence $\tilde{r}_b(\sigma) \notin F_a$. From the note in the beginning

$$\max_\sigma \frac{1}{\tilde{r}_a(\sigma)} = d_2\left(X_a\right) < d_2\left(X_b\right) = \max_\sigma \frac{1}{\tilde{r}_b(\sigma)}.$$

In the wording of the theorem we write r_a and $\|.\|_a$, instead of $\tilde{r}_a$ and $|||.|||_a$. □

4. REFERENCES

[1] Ivanov, M., Troyanski, S.: Uniformly smooth renorming of Banach spaces with modulus of convexity of power type 2. *J. Funct. Anal.*, **237**, 2006, 373–390.

[2] Ivanov, M., Parales, A. J., Troyanski, S.: On the geometry of Banach spaces with modulus of convexity of power type 2. *Studia Mathematica*, **197**, no. 1, 2010, 81–91.

[3] Jordan, P., von Neumann, J.: On inner products in linear, metric spaces. *Ann. Math.*, **36**, 1935, 719–723.

[4] John, F.: Extremum problems with inequalities as subsidary conditions. In: *Studies and Essays Presented to R. Courant*, Interscience, 1948, 187–204.

[5] Kirchev, K., Troyanski, S.: On some characterisations of spaces with scalar product. *C. R. Acad. Bulgare Sci.*, **28**, 1975, 445–447.

[6] Lindenstrauss, J., Tzafriri, L.: *Classical Banach Spaces. II: Function Spaces*, Springer, 1979.

[7] Nordlander, G.: The modulus of convexity in normed linear spaces. *Ark. Mat.*, **4**, 1960, 15–17.

[8] Rakov, S.: Uniformly smooth renormings of uniformly convex Banach spaces. *J. Soviet Math.*, **31**, 1985, 2713–2721.

[9] Senechalle, D.: Euclidean and non-Euclidean norms in a plane. *Illinois J. Math.*, **15**, 1971, 281–289.

Received on March 20, 2017

Rosen Nikolov

Faculty of Mathematics and Informatics
"St. Kl. Ohridski" University of Sofia
5, J. Bourchier blvd., BG-1164 Sofia
BULGARIA

E-mail: rumpo1959@abv.bg

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 104

ANNUAL OF SOFIA UNIVERSITY „ST. KLIMENT OHRIDSKI“
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 104

DEFINITE QUADRATURE FORMULAE OF ORDER THREE WITH EQUIDISTANT NODES

ANA AVDZHIEVA, VESSELIN GUSHEV, GENO NIKOLOV

A sequence of definite quadrature formulae of order three with equidistant nodes is constructed. The error constants of these quadratures are evaluated and simple a posteriori error estimates derived under the assumption that the integrand's third derivative does not change its sign in the integration interval.

Keywords: Definite quadrature formulae, Peano kernels, Euler-MacLaurin summation formulae, a posteriori error estimates.
2000 Math. Subject Classification: 41A55, 65D30, 65D32.

1. INTRODUCTION AND STATEMENT OF THE RESULTS

For the purposes of numerical integration, the definite integral

$$I[f] := \int_0^1 f(x)\,dx \tag{1.1}$$

is approximated by a quadrature formula, i.e., a linear functional of the form

$$Q[f] = \sum_{i=0}^{n} a_i\, f(x_i), \qquad 0 \le x_0 < x_1 < \cdots < x_n \le 1\,. \tag{1.2}$$

Quadrature formula (1.2) is said to have algebraic degree of precision m (in short, $ADP(Q) = m$), if its remainder

$$R[Q; f] := I[f] - Q[f]$$

vanishes whenever $f \in \pi_m$, and $R[Q; f] \neq 0$ when f is a polynomial of degree $m+1$. Here and henceforth, π_k stands for the set of algebraic polynomials of degree not exceeding k.

We are interested in definite quadrature formulae.

Definition 1. Quadrature formula (1.2) is said to be *definite of order* r, $r \in \mathbb{N}$, if there exists a real non-zero constant $c_r(Q)$ such that its remainder functional admits the representation

$$R[Q; f] = I[f] - Q[f] = c_r(Q)\, f^{(r)}(\xi)$$

for every real-valued function $f \in C^r[0,1]$, with some $\xi \in [0,1]$ depending on f.

Furthermore, Q is called positive definite (resp., negative definite) of order r, if $c_r(Q) > 0$ ($c_r(Q) < 0$).

The importance of the definite quadrature formulae of order r stems in the fact that they provide one-sided approximation to $I[f]$ whenever $f^{(r)}$ has a permanent sign in the integration interval. For brevity sake, we adopt the following

Definition 2. A real-valued function $f \in C^r[0,1]$ is called $r-$positive (resp., $r-$negative) if $f^{(r)}(x) \geq 0$ (resp. $f^{(r)}(x) \leq 0$) for every $x \in [0,1]$.

If $\{Q^+,\ Q^-\}$ is a pair of a positive and a negative definite quadrature formula of order r and f is an r-positive function, then for the true value of $I[f]$ we have the inclusion $Q^+[f] \leq I[f] \leq Q^-[f]$. This simple observation serves as a base for derivation of a posteriori error estimates and rules for termination of calculations (stopping rules) in the algorithms for automatic numerical integration (see [3] for a survey). Most of quadratures used in practice (e.g., quadrature formulae of Gauss, Radau, Lobatto, Newton-Cotes) are definite of certain order.

Perhaps, the best known definite quadrature formulae are the midpoint and the trapezium rules,

$$Q_n^{Mi}[f] = \frac{1}{n}\sum_{k=1}^{n} f\Big(\frac{2k-1}{2n}\Big), \qquad Q_{n+1}^{Tr}[f] = \frac{1}{2n}\big(f(0)+f(1)\big) + \frac{1}{n}\sum_{k=1}^{n-1} f\Big(\frac{k}{n}\Big),$$

they are respectively positive and negative definite of order two. Moreover, Q_n^{Mi} and Q_{n+1}^{Tr} are the optimal definite quadrature formulae of order two. The latter means that $c_2(Q_n^{Mi}) = \frac{1}{24n^2}$ is the smallest possible error constant of a n-point positive definite quadrature formula of order two, and $c_2(Q_{n+1}^{Tr}) = -\frac{1}{12n^2}$ is the

largest possible error constant of a $(n+1)$-point negative definite quadrature formulae of order two. Additional advantages of Q_n^{Mi} and Q_{n+1}^{Tr} are that they use equispaced nodes and equal weights.

The optimal definite quadrature formulae of higher order are not known explicitly, although their existence and uniqueness is known, see [10, 4, 6, 7]. In [10] Schmeisser [10] constructed optimal definite quadrature formulae of even order *with equidistant nodes*. Köhler and Nikolov [5] showed that certain Gauss-type quadratures for spaces of polynomial splines with double equidistant knots are asymptotically optimal definite quadrature formulae, and based on this result, Nikolov [8] proposed an algorithm for the construction of asymptotically optimal definite quadrature formulae of order four. In a recent paper [1] two of the authors constructed sequences of asymptotically optimal definite quadrature formulae of order four with all but few boundary nodes being equidistant; moreover, for suitable pairs of such definite quadrature formulae they derived a posteriori error estimates.

The simplest example of a pair of definite quadrature formulae of odd order is the left- and the right- rectangle rules,

$$Q_+[f] = \frac{1}{n}\sum_{k=0}^{n-1} f\Big(\frac{k}{n}\Big)\,, \qquad Q_-[f] = \frac{1}{n}\sum_{k=1}^{n} f\Big(\frac{k}{n}\Big)\,,$$

which are a positive and a negative definite quadrature formula, respectively, of order one. Indeed, if f is an $1-$positive (or simply nondecreasing) function, then $R[Q_+; f] \geq 0$, $R[Q_-; f] \leq 0$, and

$$|R[Q_\pm[f]| \leq Q_-[f] - Q_+[f] = \frac{1}{n}\big(f(1) - f(0)\big)\,. \tag{1.3}$$

We observe some differences with the definite quadrature formulae of even order: while, most often, definite quadrature formulae of even order are symmetrical, the left and the right rectangles formulae are non-symmetrical. Furthermore, each of them is obtained from the other one by a *reflection*.

Definition 3. Quadrature formula (1.2) is called:

- *symmetrical*, if

$$a_k = a_{n-k}\,, \qquad k = 0, \ldots, n\,; \tag{1.4}$$
$$x_k = 1 - x_{n-k}\,, \qquad k = 0, \ldots, n\,; \tag{1.5}$$

- *nodes-symmetrical*, if only condition (1.5) is satisfied;

- Quadrature formula

$$\widetilde{Q}[f] = \widetilde{Q}[Q; f] := \sum_{k=0}^{n} a_k\, f(x_{n-k}) \tag{1.6}$$

is called the *reflected quadrature formula* to (1.2).

Thus, a quadrature formula Q is symmetrical if and only if it coincides with its reflected, $\widetilde{Q}$. By adding (if necessary) nodes with weights equal to zero, each quadrature formula may be considered as nodes-symmetrical.

The following simple statement shows that our observations about the left- and the right- rectangle rules apply to a more general situation.

Proposition 1. *(i) If Q is a positive definite quadrature formula of order r, r - odd, then its reflected quadrature formula $\widetilde{Q}$ is negative definite of order r and vice versa. Moreover,*

$$c_r(\widetilde{Q}) = -c_r(Q)\,. \tag{1.7}$$

(ii) If Q is a nodes-symmetrical definite quadrature formula of order r, r - odd, and f is an $r-$ positive or $r-$negative function, then, with Q^* standing for either Q or $\widetilde{Q}$ we have

$$\big|R[Q^*;f]\big| \le B[Q;f] := \left|\sum_{k=0}^{\lfloor\frac{n}{2}\rfloor}\big(a_k - a_{n-k}\big)\big(f(x_{n-k}) - f(x_k)\big)\right|. \tag{1.8}$$

(iii) Under the same assumptions for Q and f as in (ii), for the remainder of quadrature formula $\hat{Q} = \frac{1}{2}\,(Q + \widetilde{Q})$ we have

$$\big|R[\hat{Q};f]\big| \le \frac{1}{2}\,B[Q;f]\,. \tag{1.9}$$

Proof. (i) Let $\widetilde{f}(x) = f(1-x)$, then $I[f] = I[\widetilde{f}]$. If Q is a definite quadrature formula of order r, r - odd, and f is an $r-$positive or $r-$negative function, then $\widetilde{Q}[f] = Q[\widetilde{f}]$ and

$$R[\widetilde{Q};f] = I[f] - \widetilde{Q}[f] = I[\widetilde{f}] - Q[\widetilde{f}] = c_r(Q)\widetilde{f}^{(r)}(\xi) = -c_r(Q)f^{(r)}(1-\xi)\,,$$

which shows that $\widetilde{Q}$ is also definite of order r and $c_r(\widetilde{Q}) = -c_r(Q)$.

Now we prove (ii) and (iii). If, e.g., Q is a nodes-symmetrical positive definite quadrature formulae of order r and f is an $r-$positive function, then

$$Q[f] \le I[f] \le \widetilde{Q}[f]\,, \tag{1.10}$$

and consequently

$$\begin{aligned} 0 \le R[Q;f] \le \widetilde{Q}[f] - Q[f] &= \sum_{k=0}^{n} a_k\big(f(x_{n-k}) - f(x_k)\big) = \sum_{k=0}^{n}\big(a_{n-k} - a_k\big)f(x_k) \\ &= \sum_{k=0}^{\lfloor\frac{n}{2}\rfloor}\big(a_k - a_{n-k}\big)\big(f(x_{n-k}) - f(x_k)\big) = B[Q;f]\,. \end{aligned}$$

Inequality (1.9) is an obvious consequence of (1.10). The proof of the other cases is completely analogous, and therefore is omitted. □

Proposition 1 implies, in particular, that definite quadrature formulae of odd order are never symmetrical. The error estimate (1.8) is especially simple when Q is of almost Chebyshev type, i.e. almost all weights of Q are equal to each other. The definite quadrature formulae constructed in this paper enjoy this property.

Before formulating our main result, let us introduce some notation.

For $n \in \mathbb{N}$ and a function f defined on the interval $[0,1]$, we set

$$x_{i,n} = \frac{i}{n}\,, \quad f_i = f(x_{i,n})\,, \qquad i = 0\,\ldots, n\,.$$

Recall that the finite differences $\Delta^k f_i$ are defined recursively by

$$\Delta^1 f_i = \Delta f_i := f_{i+1} - f_i \quad \text{and} \quad \Delta^{k+1} f_i = \Delta\big(\Delta^k f_i\big)\,, \quad k \geq 1\,.$$

Theorem 1. *For every $n \geq 8$, quadrature formula*

$$Q_n[f] = \sum_{k=0}^{n-1} A_{k,n}\, f(x_{k,n})\,, \qquad x_{k,n} = \frac{k}{n}\,,$$

with coefficients $A_{k,n} = \dfrac{1}{n}$, $3 \leq k \leq n-4$, and

$$A_{0,n} = \frac{81+\sqrt{3}}{216n}, \qquad A_{1,n} = \frac{126-\sqrt{3}}{108n}, \qquad A_{2,n} = \frac{207+\sqrt{3}}{216n},$$
$$A_{n-3,n} = \frac{297-\sqrt{3}}{216n}, \quad A_{n-2,n} = \frac{\sqrt{3}-18}{108n}, \quad A_{n-1,n} = \frac{495-\sqrt{3}}{216n},$$

is positive definite of order 3 with the error constant

$$c_3(Q_n) = \frac{\sqrt{3}}{216n^3} + \frac{27-\sqrt{3}}{72n^4}\,. \tag{1.11}$$

If f is a 3-positive or 3-negative function, then

$$\big|R[Q_n; f]\big| \leq \frac{1}{216n}\,\Big|81\big(\Delta^3 f_{n-3} - \Delta^3 f_0\big) + \sqrt{3}\big(\Delta^2 f_{n-2} + \Delta^2 f_{n-3} - \Delta^2 f_0 - \Delta^2 f_1\big)\Big|\,.$$

As an immediate consequence of Theorem 1 and Proposition 1 we have:

Corollary 1. *The reflected to Q_n quadrature formula $\widetilde{Q}_n$ is negative definite of order 3 with the error constant $c_3(\widetilde{Q}_n) = -c_3(Q_n)$.*

If f is a 3-positive or 3-negative function and $\hat{Q}_n = \dfrac{1}{2}\,\big(Q_n + \widetilde{Q}_n\big)$, then

$$\big|R[\widetilde{Q}_n; f]\big| \leq \frac{1}{216n}\,\Big|81\big(\Delta^3 f_{n-3} - \Delta^3 f_0\big) + \sqrt{3}\big(\Delta^2 f_{n-2} + \Delta^2 f_{n-3} - \Delta^2 f_0 - \Delta^2 f_1\big)\Big|\,,$$

$$\big|R[\hat{Q}_n; f]\big| \leq \frac{1}{432n}\,\Big|81\big(\Delta^3 f_{n-3} - \Delta^3 f_0\big) + \sqrt{3}\big(\Delta^2 f_{n-2} + \Delta^2 f_{n-3} - \Delta^2 f_0 - \Delta^2 f_1\big)\Big|\,.$$

Let us point out that, while error estimates of the form

$$|R[Q_n; f]| \le c_3(Q_n)\|f'''\|_{C[0,1]}$$

and alike require knowledge about the magnitude of integrand's derivative, the bounds in Theorem 1 and Corollary 1 in terms of finite differences involve only eight values of the integrand and may serve as a simple criteria for the number of nodes n needed to guarantee the evaluation of $I[f]$ with a prescribed tolerance. In this respect, it is preferable to use quadrature formula $\hat{Q}_n$ rather than the definite quadrature formulae Q_n and $\widetilde{Q}_n$.

The rest of the paper is organised as follows. Section 2 provides known facts about the Peano kernel representation of linear functionals and the Euler-MacLaurin expansion formula for the remainder of the trapezium quadrature formula. In Sections 3 we present the proof of Theorem 1. In our construction of quadrature formula (1.10) we perform some optimization, minimizing its error constant and at the same time trying to preserve its almost Chebyshev structure.

2. PRELIMINARIES

By $W_1^r = W_1^r[0,1]$, $r \in \mathbb{N}$, we denote the Sobolev class of functions

$$W_1^r[0,1] := \{f \in C^{r-1}[0,1] \,:\, f^{(r-1)} \text{ abs. continuous}, \int_0^1 |f^{(r)}(t)|\,dt < \infty\}\,.$$

In particular, $W_1^r[0,1]$ contains the class $C^r[0,1]$.

If $\mathcal{L}$ is a linear functional defined in $W_1^r[0,1]$ which vanishes on π_{r-1}, then, by a classical result of Peano [9], $\mathcal{L}$ admits the integral representation

$$\mathcal{L}[f] = \int_0^1 K_r(t) f^{(r)}(t)\,dt, \qquad K_r(t) = \mathcal{L}\Big[\frac{(\cdot - t)_+^{r-1}}{(r-1)!}\Big], \quad t \in [0,1]\,,$$

where

$$u_+(t) = \max\{t, 0\}\,, \quad t \in \mathbb{R}\,.$$

In the case when $\mathcal{L}$ is the remainder $R[Q;\cdot]$ of a quadrature formula Q with $ADP(Q) \ge r-1$, the function $K_r(t) = K_r(Q;t)$ is referred to as the r-th Peano kernel of Q. For Q as in (2.1), explicit representations for $K_r(Q;t)$, $t \in [0,1]$, are

$$K_r(Q;t) = \frac{(1-t)^r}{r!} - \frac{1}{(r-1)!}\sum_{i=0}^{n} a_i (x_i - t)_+^{r-1}, \tag{2.1}$$

$$K_r(Q;t) = (-1)^r\Big[\frac{t^r}{r!} - \frac{1}{(r-1)!}\sum_{i=0}^{n} a_i (t - x_i)_+^{r-1}\Big]. \tag{2.2}$$

Since for $f \in C^r[0,1]$ we have

$$R[Q;f] = \int_0^1 K_r(Q;t)\, f^{(r)}(t)\, dt\,,$$

it is clear that Q is a positive (negative) definite quadrature formula of order r if and only if $ADP(Q) = r-1$ and $K_r(Q;t) \geq 0$ (resp. $K_r(Q;t) \leq 0$) for all $t \in [0,1]$.

Throughout this paper, $\{x_{k,n}\}_{k=0}^n$ will stand for the nodes of the n-th compound trapezium formula Q_{n+1}^{Tr},

$$x_{k,n} = \frac{k}{n}, \quad k = 0, \ldots, n\,,$$

so that

$$Q_{n+1}^{Tr}[f] = \frac{1}{2n}\big(f(x_{0,n}) + f(x_{n,n})\big) + \frac{1}{n}\sum_{k=1}^{n-1} f(x_{k,n})\,. \tag{2.3}$$

Our definite quadrature formulae are obtained by an appropriate modification of Q_{n+1}^{Tr}. The following lemma gives a particular case of the Euler-Maclaurin formula, see, e.g., [2, Satz 98]:

Lemma 1. *Assume that $f \in W_1^3$. Then*

$$I[f] = Q_{n+1}^{Tr}[f] - \frac{1}{12n^2}\big[f'(1) - f'(0)\big] - \frac{1}{n^3}\int_0^1 \widetilde{B}_3(nx) f'''(x)\, dx\,, \tag{2.4}$$

where $\widetilde{B}_3$ is the 1-periodic extension of the third Bernoulli polynomial

$$B_3(x) = \frac{x^3}{6} - \frac{x^2}{4} + \frac{x}{12}\,.$$

Note that $\widetilde{B}_3(x) = B_3(\{x\})$, $x \in \mathbb{R}$, where $\{x\}$ stands for the fractional part of x. In the sequel, we shall use the fact that

$$-\frac{\sqrt{3}}{216} \leq \widetilde{B}_3(x) \leq \frac{\sqrt{3}}{216}\,, \qquad x \in \mathbb{R}\,. \tag{2.5}$$

3. PROOF OF THE RESULTS.

We rewrite formula (2.4) in Lemma 1 in the following form:

$$\begin{aligned} I[f] =& Q_{n+1}^{Tr}[f] - \frac{1}{12\,n^2}\big[f'(1) - f'(0)\big] - \frac{\sqrt{3}}{216\,n^3}\big[f''(1) - f''(0)\big] \\ &+ \frac{1}{n^3}\int_0^1 \Big(\frac{\sqrt{3}}{216} - \widetilde{B}_3(nx)\Big) f^{(3)}(x)\, dx \\ =:& \widetilde{Q}[f] + R[\widetilde{Q};f]\,, \end{aligned} \tag{3.1}$$

where

$$\widetilde{Q}[f] = Q_{n+1}^{Tr}[f] + \frac{1}{12n^2} f'(0) + \frac{\sqrt{3}}{216n^3} f''(0) - \frac{1}{12n^2} f'(1) - \frac{\sqrt{3}}{216n^3} f''(1)\,. \quad (3.2)$$

By (3.1) and (2.5) it follows that $\widetilde{Q}$ is a positive definite quadrature formula, however, it is not of the desired form as it involves values of integrand's derivatives. That is why we approximate the derivatives values at the end-points appearing in $\widetilde{Q}$ by pairs of formulae for numerical differentiation involving values at the closest nodes. The reason for not using single formulae for numerical differentiation is that it is not a priory clear whether they will result in a positive definite quadrature formula, so we need some flexibility to achieve definiteness.

Thus, $f'(0)$ is approximated as follows:

$$f'(0) \approx \frac{n}{2}\big[-3f(x_{0,n}) + 4f(x_{1,n}) + f(x_{2,n}) =: D_{1,1}[f]\,,$$

$$f'(0) \approx \frac{n}{2}\big[-5f(x_{1,n}) + 8f(x_{2,n}) - 3f(x_{3,n}) =: D_{1,2}[f]\,,$$

and for any $\alpha \in \mathbb{R}$ we have

$$\begin{aligned} &f'(0) \approx \alpha D_{1,1}[f] + (1-\alpha) D_{1,2}[f] =: D_1^{\alpha}[f]\,,\\ &L_1[f] := f'(0) - D_1^{\alpha}[f] \ \text{ vanishes on } \ \pi_2\,. \end{aligned} \quad (3.3)$$

Likewise, $f''(0)$ is approximated by

$$f''(0) \approx n^2\big[f(x_{0,n}) - 2f(x_{1,n}) + f(x_{2,n}) =: D_{2,1}[f]\,,$$

$$f''(0) \approx n^2\big[f(x_{1,n}) - 2f(x_{2,n}) + f(x_{3,n}) =: D_{2,2}[f]\,,$$

and for any $\beta \in \mathbb{R}$

$$\begin{aligned} &f''(0) \approx \beta D_{2,1}[f] + (1-\beta) D_{2,2}[f] =: D_2^{\beta}[f]\,,\\ &L_2[f] := f''(0) - D_2^{\beta}[f] \ \text{ vanishes on } \ \pi_2\,. \end{aligned} \quad (3.4)$$

For the approximation of $f'(1)$ and $f''(1)$ we use the above formulae for numerical differentiation, applied to $-f(1-x)$ and $f(1-x)$, respectively. For the first derivative this yields

$$f'(1) \approx \frac{n}{2}\big[3f(x_{n,n}) - 4f(x_{n-1,n}) - f(x_{n-2,n}) =: \widetilde{D}_{1,1}[f]\,,$$

$$f'(1) \approx \frac{n}{2}\big[5f(x_{n-1,n}) - 8f(x_{n-2,n}) + 3f(x_{n-3,n}) =: \widetilde{D}_{1,2}[f]\,,$$

and for any $\gamma \in \mathbb{R}$ we have

$$\begin{aligned} &f'(1) \approx \gamma \widetilde{D}_{1,1}[f] + (1-\gamma) D_{1,2}[f] =: \widetilde{D}_1^{\gamma}[f]\,,\\ &\widetilde{L}_1[f] := f'(1) - \widetilde{D}_1^{\gamma}[f] \ \text{ vanishes on } \ \pi_2\,. \end{aligned} \quad (3.5)$$

Similarly,

$$f''(1) \approx n^2 \left[f(x_{n,n}) - 2f(x_{n-1,n}) + f(x_{n-2,n}) =: \widetilde{D}_{2,1}[f] \right],$$

$$f''(1) \approx n^2 \left[f(x_{n-1,n}) - 2f(x_{n-2,n}) + f(x_{n3,n}) =: \widetilde{D}_{2,2}[f] \right],$$

and for any $\delta \in \mathbb{R}$

$$\begin{aligned} &f''(1) \approx \delta \widetilde{D}_{2,1}[f] + (1-\delta)\widetilde{D}_{2,2}[f] =: \widetilde{D}_2^{\delta}[f], \\ &\widetilde{L}_2[f] := f''(1) - \widetilde{D}_2^{\delta}[f] \ \text{ vanishes on } \ \pi_2 . \end{aligned} \tag{3.6}$$

The replacement of $f'(0)$, $f''(0)$, $f'(1)$ and $f''(1)$ in (3.2) by $D_1^{\alpha}[f]$, $D_2^{\beta}[f]$, $\widetilde{D}_1^{\gamma}[f]$ and $\widetilde{D}_2^{\delta}[f]$, respectively, yields a quadrature formula

$$Q[f] = \sum_{k=0}^{n} A_{k,n} f\Big(\frac{k}{n}\Big), \tag{3.7}$$

which, by construction, evaluates $I[f]$ to the exact value whenever $f \in \pi_2$.

Assuming that $n \geq 8$, we have $A_{k,n} = 1/n$ for $4 \leq k \leq n-4$. Formally, coefficients $A_{k,n}$, $0 \leq k \leq 3$, depend on parameters α and β, while coefficients $A_{k,n}$, $n-3 \leq k \leq n$ depend on parameters γ and δ. In fact, it is not difficult to see that $\{A_{k,n}\}_0^3$ depend on a single parameter, say θ, while $\{A_{k,n}\}_{n-3}^n$ depend on another single parameter, say ϱ, where

$$\theta := 27\alpha - \sqrt{3}\beta, \qquad \varrho := 27\gamma + \sqrt{3}\delta .$$

Specifically, we have

$$\begin{aligned} A_{0,n} &= \frac{108-\theta}{216n}, & A_{1,n} &= \frac{171+\sqrt{3}+3\theta}{216n}, \\ A_{2,n} &= \frac{288-2\sqrt{3}-3\theta}{216n}, & A_{3,n} &= \frac{189+\sqrt{3}+\theta}{216n}, \\ A_{n-3,n} &= \frac{189-\sqrt{3}+\varrho}{216n}, & A_{n-2,n} &= \frac{288+2\sqrt{3}-3\varrho}{216n}, \\ A_{n-1,n} &= \frac{171-\sqrt{3}+3\varrho}{216n}, & A_{n,n} &= \frac{108-\varrho}{216n}, \\ A_{k,n} &= \frac{1}{n}, \quad 4 \leq k \leq n-4. \end{aligned}$$

Our next goal is to determine the values of parameters θ and ϱ which ensure that quadrature formula (3.7) is positive definite of order 3. Not only want we (3.7) to be positive definite, but also require θ and ϱ to be chosen in such a way that its error constant, $c_3(Q)$, is as small as possible. To this end, let us look closer at the third Peano kernel of Q.

From (3.2)–(3.6) we observe that

$$R[Q;f] = R[\widetilde{Q};f] + \frac{1}{12n^2}\,L_1[f] + \frac{\sqrt{3}}{216n^3}\,L_2[f] + \frac{1}{12n^2}\,\widetilde{L}_1[f] + \frac{\sqrt{3}}{216n^3}\,\widetilde{L}_2[f]\,,$$

therefore

$$\begin{aligned} K_3(Q;t) = K_3(\widetilde{Q};t) &+ \frac{1}{12n^2}\,K_3(L_1;t) + \frac{\sqrt{3}}{216n^3}\,K_3(L_2;t) \\ &+ \frac{1}{12n^2}\,K_3(\widetilde{L}_1;t) + \frac{\sqrt{3}}{216n^3}\,K_3(\widetilde{L}_2;t)\,. \end{aligned} \tag{3.8}$$

Based on the definition of Peano kernels, it is not difficult to see that $K_3(L_1;\cdot)$ and $K_3(L_2;\cdot)$ vanish identically on the interval $[x_{3,n},1]$ whereas $K_3(\widetilde{L}_1;\cdot)$ and $K_3(\widetilde{L}_2;\cdot)$ vanish identically on the interval $[0,x_{n-3,n}]$. Hence, in view of (3.1),

$$K_3(Q;t) = K_3(\widetilde{Q};t) = n^{-3}\left[\frac{\sqrt{3}}{216} - \widetilde{B}_3(n\,t)\right] \geq 0\,, \qquad t\in[x_{3,n},x_{n-3,n}]\,, \tag{3.9}$$

therefore we have to verify condition $K_3(Q;t)\geq 0$ only on the intervals $[0,x_{3,n}]$ and $[x_{n-3,n},1]$. Assuming that this condition is satisfied, for the error constant of Q we have

$$c_3(Q) = \int_0^{x_{3,n}} K_3(Q;t)\,dt + \int_{x_{n-3,n}}^{1} K_3(Q;t)\,dt + \frac{\sqrt{3}(n-6)}{216n^4}\,, \tag{3.10}$$

where the last summand comes from the integral of $K_3(Q;\cdot)$ over the interval $[x_{3,n},x_{n-3,n}]$, and we have used that $\widetilde{B}_3$ has mean value zero on the period.

We aim to minimize $c_3(Q)$, i.e., to minimize the integrals in (3.10) with respect to parameters θ and ϱ, respectively, subject to the requirement $K_3(Q;t)\geq 0$ on the intervals $[0,x_{3,n}]$ and $[x_{n-3,n},1]$.

3.1. POSITIVITY OF $K_3(Q;t)$ ON $[0,x_{3,n}]$

We make use of formula (2.2) for Peano kernels, with $r=3$, and after the change of variable $t=\frac{u}{n}$ arrive at the following representation of $K_3(Q;t)$ for $t\in[0,x_{3,n}]$:

$$\begin{aligned} K_3(Q;t) &= -\frac{1}{6n^3}\Big[u^3 - \frac{108-\theta}{72}\,u^2 - \frac{171+\sqrt{3}+3\theta}{72}(u-1)_+^2 - \frac{288-2\sqrt{3}-3\theta}{72}\,(u-2)_+^2\Big] \\ &=: -\frac{1}{6n^3}\,\varphi(\theta;u) = -\frac{1}{6n^3}\,\varphi(u)\,, \qquad u\in[0,3]\,. \end{aligned}$$

Thus, we have the equivalence

$$K_3(Q;t)\geq 0, \quad t\in[0,x_{3,n}] \quad\Leftrightarrow\quad \varphi(u)\leq 0, \quad u\in[0,3]. \tag{3.11}$$

Before verifying what values of θ ensure condition $\varphi(u) \le 0, u \in [0,3]$, we evaluate the first integral in (3.10):

$$\int_0^{x_{3,n}} K_3(Q;t)\,dt = -\frac{1}{6n^4}\int_0^3 \varphi(u)\,du = \frac{33+\sqrt{3}-\theta}{216n^4}\,. \tag{3.12}$$

To minimize the latter integral and thereby, in view of (3.10), $c_3(Q)$, we have to find the largest value of θ ensuring that $\varphi(u) \le 0,\ u \in [0,3]$.

Case 1: $u \in [0,1]$. In this case

$$\varphi(u) = \frac{u^2}{72}(72u - 108 + \theta),$$

and condition $\varphi(u) \le 0,\ u \in [0,1]$ is equivalent to $\theta \le 36$.

Case 2: $u \in [1,2]$. We set $v = u-1,\ v \in [0,1]$, and using Wolfram's *Mathematica*, find

$$\begin{aligned}\varphi(u) &= u^3 - \frac{108-\theta}{72}\,u^2 - \frac{171+\sqrt{3}+3\theta}{72}\,(u-1)^2\\ &= v^3 - \frac{63+\sqrt{3}}{72}\,v^2 - \frac{1}{2} + \frac{\theta}{36}\Big(-v^2+v-\frac{1}{2}\Big) =: \varphi_1(v)\,.\end{aligned}$$

Since $-v^2+v-1/2 < 0$ for all $v \in [0,1]$, it follows that $\varphi(u) < 0$ for every $u \in [1,2]$ provided θ is big enough; in addition, if the latter condition holds for some θ_0, it will hold also for all $\theta > \theta_0$. The largest value of θ such that $\varphi_1(v) \le 0$ for all $v \in [0,1]$ should be such that φ_1 has a double zero in $(0,1)$, i.e. θ is a zero of $D(\varphi_1)$, the discriminant of φ_1. Using Wolfram's *Mathematica*, we find $D(\varphi_1)$, which is a quintic polynomial of θ with four distinct real zeros: $\theta_1 = -57.5774$, $\theta_2 = 28.0556$, $\theta_3 = 30.7503$ and $\theta_4 = 92.4621$. Only for $\theta = \theta_2$ the polynomial φ_1 has a double zero in $(0,1)$. Therefore, in this case we have the restriction $\theta \le \theta_2 = 28.0556$.

Case 3: $u \in [2,3]$. We set $v = u-2,\ v \in [0,1]$, and find with *Mathematica*

$$\begin{aligned}\varphi(u) &= u^3 - \frac{108-\theta}{72}\,u^2 - \frac{171+\sqrt{3}+3\theta}{72}\,(u-1)^2 - \frac{288-2\sqrt{3}-3\theta}{72}\,(u-2)^2\\ &= \frac{1}{72}\big[72v^3 - (135-\sqrt{3})\,v^2 + (90-2\sqrt{3})\,v - 27 - \sqrt{3} + \theta(v-1)^2\big]\\ &=: \varphi_2(v) = \varphi_2(\theta;v)\,.\end{aligned}$$

Since $\varphi_2(1) = -\sqrt{3}/36 < 0$, it is clear that $\varphi_2(v) < 0$ for all $v \in [0,1]$ provided θ is small enough; moreover, if the above condition on φ_2 is satisfied for some θ_0, then it is satisfied for all $\theta < \theta_0$. The critical value θ^* should be such

that φ_2 has a double zero in $(0,1)$, i.e., θ^* is a zero of $D(\varphi_2)$, the discriminant of φ_2. With the help of *Mathematica*, we find

$$\frac{D(\varphi_2)}{8} = \sqrt{3}\,\theta^3 + (171 + 243\sqrt{3})\theta^2 + (27702 + 8352\sqrt{3})\theta - 802134 - 386370\sqrt{3}\,.$$

By Descartes' rule of signs, the latter polynomial has a unique positive root θ^*. In fact, using again *Mathematica*, we find that $\theta^* = 27 - \sqrt{3} = 25.2679\ldots$ and

$$D(\varphi_2) = 8(\theta - \theta^*)\big[\sqrt{3}\theta^2 + 6(28 + 45\sqrt{3})\theta + 6(5238 + 2579\sqrt{3})\big]\,.$$

Thus, the optimal value of θ in *Case 3* is $\theta = \theta^* = 27 - \sqrt{3}$. Just for one more check, we verify that

$$\varphi_2(\theta^*; v) = v^3 - \frac{3}{2}\,v^2 + \frac{1}{2}\,v - \frac{\sqrt{3}}{36} = \Big(v - \frac{3 + 2\sqrt{3}}{6}\Big)\Big(v - \frac{3 - \sqrt{3}}{6}\Big)^2 \le 0, \quad v \in [0,1].$$

Summarizing the three cases considered above, we see that the optimal value of θ ensuring that $\varphi(\theta; u) \le 0$ for all $u \in [0,3]$ is $\theta = \theta^* = 27 - \sqrt{3}$. We have

$$\varphi(\theta^*; u) = u^3 - \frac{81 + \sqrt{3}}{72}\,u^2 - \frac{126 - \sqrt{3}}{36}\,(u-1)_+^2 - \frac{207 + \sqrt{3}}{72}\,(u-2)_+^2\,.$$

The graph of $-\varphi(\theta^*; u)$ is depicted in Figure 1.

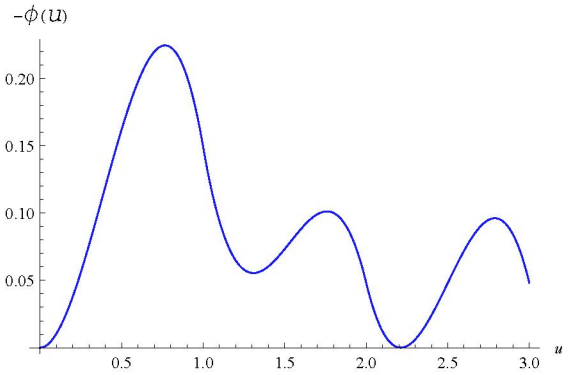


Figure. 1. The graph of $-\varphi(\theta^*; u)$, $0 \le u \le 3$.

In view of (3.11), $\theta = \theta^*$ ensures that $K_3(Q; t) \ge 0$ for all $t \in [0, x_{3,n}]$.

With the optimal value $\theta = \theta^*$, the coefficients $\{A_{k,n}\}_0^3$ of quadrature formula (3.7) are given by

$$A_{0,n} = \frac{81 + \sqrt{3}}{216n}, \quad A_{1,n} = \frac{126 - \sqrt{3}}{108n}, \quad A_{2,n} = \frac{207 + \sqrt{3}}{216n}, \quad A_{3,n} = \frac{1}{n}\,,$$

i.e., $\{A_{k,n}\}_0^3$ coincide with the coefficients of quadrature formula Q_n in Theorem 1. Moreover, (3.12) with $\theta = \theta^*$ yields

$$\int_0^{x_{3,n}} K_3(Q;t)\,dt = \frac{3+\sqrt{3}}{108n^4}\,. \tag{3.13}$$

3.2. POSITIVITY OF $K_3(Q;t)$ ON $[x_{n-3,n}, 1]$

We apply (2.1) with $r = 3$ and Q being quadrature formula (3.7) to obtain:

$$\begin{aligned} K_3(Q;t) &= \frac{(1-t)^3}{6} - \frac{1}{2}\sum_{k=0}^{n} A_{k,n}(x_{k,n}-t)_+^2 \\ &= \frac{(1-t)^3}{6} - \frac{1}{2}\sum_{k=0}^{n} A_{k,n}(1-t-x_{n-k,n})_+^2 \\ &\overset{x=1-t}{=} \frac{x^3}{6} - \frac{1}{2}\sum_{k=0}^{n} A_{n-k,n}(x-x_{k,n})_+^2 := \widetilde{K}_3(Q;x)\,. \end{aligned}$$

Hence,

$$\int_{x_{n-3,n}}^{1} K_3(Q;t)\,dt = \int_0^{x_{3,n}} \widetilde{K}_3(Q;x)\,dx \overset{x=u/n}{=} \frac{1}{n^4}\int_0^3 \psi(\varrho;u)\,du\,,$$

with $\psi(u) = \psi(\varrho;u)$ given by

$$\psi(u) = u^3 - \frac{108-\varrho}{72}\,u^2 - \frac{171-\sqrt{3}+3\varrho}{72}\,(u-1)_+^2 - \frac{288+2\sqrt{3}-3\varrho}{72}\,(u-2)_+^2\,.$$

Now we have the equivalence

$$K_3(Q;t) \geq 0, \quad t \in [x_{n-3,n}, 1] \quad \Leftrightarrow \quad \psi(\varrho;u) \geq 0, \quad u \in [0,3]\,.$$

By a straightforward calculation we obtain

$$\int_{x_{n-3,n}}^{1} K_3(Q;t)\,dt = \frac{\varrho+\sqrt{3}-33}{216n^4}\,, \tag{3.14}$$

therefore, to minimize the error constant $c_3(Q)$, we need to find the smallest ϱ such that $\psi(\varrho;u) \geq 0$ for all $u \in [0,3]$. A necessary condition for the latter requirement to hold is $\varrho \geq 108$, since

$$\psi(\varrho;u) = u^2\Big(u + \frac{\varrho-108}{72}\Big)\,, \qquad u \in [0,1]\,.$$

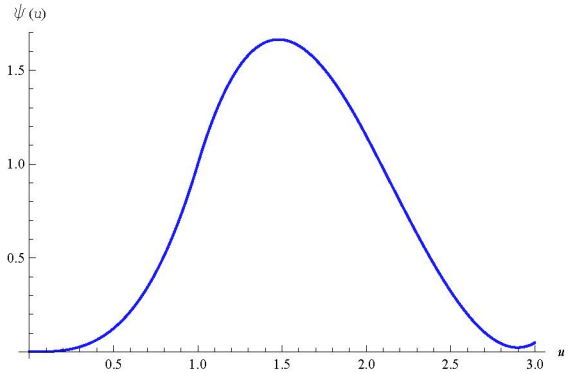


Figure. 2. The graph of $\psi(\varrho^*;u)$, $0 \le u \le 3$.

It turns out that the choice $\varrho = \varrho^* = 108$ is optimal, as it guarantees the non-negativity of ψ on the interval $[0,3]$, see the graph of $\psi(\varrho^*;\cdot)$ in Figure 2.

With $\varrho = \varrho^*$, coefficients $\{A_{k,n}\}_{k=n-3}^{n}$ of quadrature formula (3.7) are given by

$$A_{n-3,n} = \frac{297-\sqrt{3}}{216n}, \quad A_{n-2,n} = \frac{\sqrt{3}-18}{108n}, \quad A_{n-1,n} = \frac{495-\sqrt{3}}{216n}, \quad A_{n,n} = 0\,.$$

Let us summarize: with the optimal values $(\theta\,,\varrho) = (\theta^*\,,\varrho^*) = (27-\sqrt{3},108)$ the nodes of quadrature formula (3.7) are given in Table 1:

Table 1. The coefficients of quadrature formula (3.7).

$A_{0,n}$	$A_{1,n}$	$A_{2,n}$	$A_{k,n},\ \ 3 \le k \le n-1$
$\frac{81+\sqrt{3}}{216n}$	$\frac{126-\sqrt{3}}{108n}$	$\frac{207+\sqrt{3}}{216n}$	$\frac{1}{n}$
$A_{n-3,n}$	$A_{n-2,n}$	$A_{n-1,n}$	$A_{n,n}$
$\frac{297-\sqrt{3}}{216n}$	$\frac{\sqrt{3}-18}{108n}$	$\frac{495-\sqrt{3}}{216n}$	0

It is clear now that (3.7) is the positive definite quadrature formula Q_n in Theorem 1.

From (3.14) with $\varrho = \varrho^*$ we find

$$\int_{x_{n-3,n}}^{1} K_3(Q;t)\,dt = \frac{75+\sqrt{3}}{216n^4}\,. \tag{3.15}$$

Now (3.10), (3.13) and (3.15) yields

$$c_3(Q_n) = \frac{3+\sqrt{3}}{108n^4} + \frac{75+\sqrt{3}}{216n^4} + \frac{\sqrt{3}(n-6)}{216n^4} = \frac{\sqrt{3}}{216n^3} + \frac{27-\sqrt{3}}{72n^4}\,,$$

which proves (1.10).

To prove the last claim of Theorem 1, we apply Proposition 1(ii) with $Q = Q_n$. Since $A_{k,n} = 1/n$ for $3 \le k \le n-4$, we have, with $f_i = f(x_{i,n})$,

$$\big|R[Q_n; f]\big| \le B[Q_n; f] = \Big|\sum_{k=0}^{3} \big(A_{k,n} - A_{n-k,n}\big)\big(f_{n-k} - f_k\big)\Big|\,. \tag{3.16}$$

On using

$$\begin{aligned} A_{0,n} - A_{n,n} &= \tfrac{81+\sqrt{3}}{216n}\,, & A_{1,n} - A_{n-1,n} &= -\tfrac{243+\sqrt{3}}{216n}\,,\\ A_{2,n} - A_{n-2,n} &= \tfrac{243-\sqrt{3}}{216n}\,, & A_{3,n} - A_{n-3,n} &= -\tfrac{81-\sqrt{3}}{216n} \end{aligned}$$

and the explicit form of finite differences,

$$\Delta^m f_i = \sum_{k=0}^{m} (-1)^{m-k} \binom{m}{k} f_{i+k}\,,$$

we obtain from (3.16), after rearrangement,

$$B[Q_n; f] = \frac{1}{216n}\,\big|81\big(\Delta^3 f_{n-3} - \Delta^3 f_0\big) + \sqrt{3}\big(\Delta^2 f_{n-2} + \Delta^2 f_{n-3} - \Delta^2 f_0 - \Delta^2 f_1\big)\big|\,.$$

The proof of Theorem 1 is complete. □

Proof of Corollary 1. Since $\widetilde{Q}_n[f] = Q_n[\widetilde{f}]$ and $\widetilde{f}_i = f_{n-i}$, $0 \le i \le n$, we deduce from (3.16) that $B[\widetilde{Q}_n; f] = B[Q_n; f]$, which proves the first claim of Corollary 1. The second claim follows from Proposition 1(iii). □

ACKNOWLEDGEMENT. The authors were supported by the Sofia University Research Fund under Contract 80-10-11/2017. The third author acknowledges the support of the Bulgarian National Research Fund through Contract DN 02/14.

4. REFERENCES

[1] Avdzhieva, A., Nikolov, G.: Asymptotically optimal definite quadrature formulae of 4th order. *J. Comput. Appl. Math.*, **311**, 2017, 565–582.

[2] Braß, H.: *Quadraturverfahren.* Vandenhoeck & Ruprecht, Göttingen, 1977.

[3] Förster, K.-J.: Survey on stopping rules in quadrature based on Peano kernel methods. *Suppl. Rend. Circ. Math. Palermo*, Ser. II **33**, 1993, 311–330.

[4] Jetter, K.: Optimale Quadraturformeln mit semidefiniten Peano-Kernen. *Numer. Math.* **25**, 1976, 239–249.

[5] Köhler, P., Nikolov, G.: Error bounds for optimal definite quadrature formulae. *J. Approx. Theory*, **81**, 1995, 397–405.

[6] Lange, G.: *Beste und optimale definite Quadraturformel.* Ph.D. Thesis, Technical University Clausthal, Germany, 1977.

[7] Lange, G.: Optimale definite Quadraturformel. In: *Numerische Integration* (G. Hämmerlin, ed.), ISNM vol. 45, Birkhäuser, Basel, Boston, Stuttgart, 1979, pp. 187–197.

[8] Nikolov, G.: On certain definite quadrature formulae. *J. Comput. Appl. Math.*, **75**, 1996, 329–343.

[9] Peano, G.: Resto nelle formule di quadratura espresso con un integrale definito. *Atti della Reale Accademia dei Lincei: Rendiconti* (Ser. 5), **22**, 1913, 562–569.

[10] Schmeisser, G.: Optimale Quadraturformeln mit semidefiniten Kernen. *Numer. Math.* **20**, 1972, 32–53.

Received on September 8, 2017

Ana Avdzhieva, Vesselin Gushev, Geno Nikolov

Faculty of Mathematics and Informatics
"St. Kl. Ohridski" University of Sofia
5, J. Bourchier blvd., BG-1164 Sofia
BULGARIA

e-mails: aavdzhieva@fmi.uni-sofia.bg
v_gushev@fmi.uni-sofia.bg
geno@fmi.uni-sofia.bg

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 104

ANNUAL OF SOFIA UNIVERSITY „ST. KLIMENT OHRIDSKI“
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 104

ON AN EQUATION INVOLVING FRACTIONAL POWERS WITH PRIME NUMBERS OF A SPECIAL TYPE

ZHIVKO PETROV

We consider the equation $[p_1^c] + [p_2^c] + [p_3^c] = N$, where N is a sufficiently large integer, and $[t]$ denotes the integer part of t. We prove that if $1 < c < \frac{17}{16}$, then it has a solution in prime numbers p_1, p_2, p_3 such that each of the numbers $p_1 + 2$, $p_2 + 2$, $p_3 + 2$ has at most $\left[\frac{95}{17-16c}\right]$ prime factors, counted with their multiplicities.

Keywords: Waring's problem, sieve methods.
2000 Math. Subject Classification: 11P05 (Primary); 11N36 (Secondary).

1. INTRODUCTION AND STATEMENT OF THE RESULT

In 1937 I. M. Vinogradov [16] proved that for every sufficiently large odd integer N the equation

$$p_1 + p_2 + p_3 = N \tag{1.1}$$

has a solution in prime numbers p_1, p_2, p_3.

Analogous problem was considered in 1952 by Piatetski-Shapiro [9]. If $H(c)$ denotes the least integer s such that the diophantine inequality

$$|p_1^c + \cdots + p_s^c - N| < \varepsilon,$$

has a solution in primes $p_1, \ldots, p_s$, where $c > 1$ is not an integer, $\varepsilon > 0$ is small, and N is large real number, then Piatetski-Shapiro proved that

$$\limsup_{c \to \infty} \frac{H(c)}{c \log c} \leq 4.$$

He also proved that if $1 < c < 3/2$, then $H(c) \leq 5$. In 1992, Tolev [14] established that if $1 < c < \frac{15}{14}$, then the diophantine inequality

$$|p_1^c + p_2^c + p_3^c - N| < N^{-\kappa}$$

has a solution in prime numbers p_1, p_2, p_3 for certain $\kappa = \kappa(c) > 0$. Several improvements were made and the strongest of them is due to Baker and Weingartner [1], who improved Tolev's result with $1 < c < \frac{10}{9}$.

In 1995, M. B. Laporta and D. I. Tolev [7] considered the equation

$$[p_1^c] + [p_2^c] + [p_3^c] = N, \tag{1.2}$$

where $c \in \mathbb{R}$, $c > 1$, $N \in \mathbb{N}$ and $[t]$ denotes the integer part of t. They showed that if $1 < c < \frac{17}{16}$ and N is a sufficiently large integer, then the equation (1.2) has a solution in prime numbers p_1, p_2, p_3.

For any natural number r, let $\mathcal{P}_r$ denote the set of r-almost primes, i.e. the set of natural numbers having at most r prime factors counted with multiplicities. There are many papers devoted to the study of problems involving primes and almost primes. For example, in 1973 J. R. Chen [4] established that there exist infinitely many primes p such that $p + 2 \in \mathcal{P}_2$. In 2000 Tolev [12] proved that for every sufficiently large integer $N \equiv 3 \pmod 6$ the equation (1.1) has a solution in prime numbers p_1, p_2, p_3 such that $p_1 + 2 \in P_2$, $p_2 + 2 \in P_5$, $p_3 + 2 \in P_7$. Thereafter this result was improved by Matomäki and Shao [8], who showed that for every sufficiently large integer $N \equiv 3 \pmod 6$ the equation (1.1) has a solution in prime numbers p_1, p_2, p_3 such that $p_1 + 2, p_2 + 2, p_3 + 2 \in \mathcal{P}_2$.

Recently Tolev [15] established that if N is sufficiently large, $E > 0$ is an arbitrarily large constant and $1 < c < \frac{15}{14}$, then the inequality

$$|p_1^c + p_2^c + p_3^c - N| < (\log N)^{-E}$$

has a solution in prime numbers p_1, p_2, p_3, such that each of the numbers $p_1 + 2$, $p_2 + 2$, $p_3 + 2$ has at most $\left[\frac{369}{180-168c}\right]$ prime factors, counted with their multiplicities.

In this paper, we prove the following

Theorem 1.1. *Suppose that $1 < c < \frac{17}{16}$. Then for every sufficiently large N the equation (1.2) has a solution in prime numbers p_1, p_2, p_3, such that each of the numbers $p_1 + 2$, $p_2 + 2$, $p_3 + 2$ has at most $\left[\frac{95}{17-16c}\right]$ prime factors, counted with their multiplicities.*

We note that the integer $\left[\frac{95}{17-16c}\right]$ is equal to 95 if c is close to 1 and it is large if c is close to $\frac{17}{16}$.

To prove Theorem 1.1 we combine ideas developed by Laporta and Tolev [7] and Tolev [15]. First we apply a version of the vector sieve and then the circle method. In section 4 we find an asymptotic formula for the integrals Γ_1' and Γ_4' (defined by

(3.11) and (3.14) respectively). In section 5 we estimate Γ_1'' and Γ_4'' (defined by (3.12) and (3.15) respectively) and we then complete the proof of Theorem 1.1.

2. NOTATION AND SOME LEMMAS

We use the following notations: with $\{t\} = t - [t]$ we denote the fractional part of t. With $||t||$ we denote the distance from t to the nearest integer. As usual with $\mu(n)$, $\varphi(n)$ and $\Lambda(n)$ we denote respectively, Möbius' function, Euler's function and von Mangoldt's function. Also $e(t) = e^{2\pi i t}$.

We use Vinogradov's notation $A \ll B$, which is equivalent to $A = O(B)$. If we have simultaneously $A \ll B$ and $B \ll A$, then we shall write $A \asymp B$.

We reserve p, p_1, p_2, p_3 for prime numbers. By ϵ we denote an arbitrarily small positive number, which is not necessarily the same in the different formulae.

With $\mathbb{N}$, $\mathbb{Z}$ and $\mathbb{R}$ we will denote respectively the set of natural numbers, the set of integer numbers and the set of real numbers.

Now we quote some lemmas, which shall be used later.

Lemma 2.1. *Suppose that $D \in \mathbb{R}, D > 4$. There exist arithmetical functions $\lambda^{\pm}(d)$ (Rosser's functions of level D) with the following properties:*

1. *For any positive integer d we have*

$$|\lambda^{\pm}(d)| \leq 1, \qquad \lambda^{\pm}(d) = 0 \quad \text{if} \quad d > D \quad \text{or} \quad \mu(d) = 0.$$

2. *If $n \in \mathbb{N}$, then*

$$\sum_{d|n} \lambda^{-}(d) \leq \sum_{d|n} \mu(d) \leq \sum_{d|n} \lambda^{+}(d).$$

3. *If $z \in \mathbb{R}$ is such that $z^2 \leq D \leq z^3$ and if*

$$\begin{aligned} P(z) &= \prod_{2<p<z} p, \qquad \mathcal{B} = \prod_{2<p<z} \left(1 - \frac{1}{p-1}\right), \\ \mathcal{N}^{\pm} &= \sum_{d|P(z)} \frac{\lambda^{\pm}(d)}{\varphi(d)}, \quad s_0 = \frac{\log D}{\log z}, \end{aligned} \tag{2.1}$$

then we have

$$\mathcal{B} \leq \mathcal{N}^{+} \leq \mathcal{B}\left(F(s_0) + O\left((\log D)^{-\frac{1}{3}}\right)\right), \tag{2.2}$$

$$\mathcal{B} \geq \mathcal{N}^{-} \geq \mathcal{B}\left(f(s_0) + O\left((\log D)^{-\frac{1}{3}}\right)\right), \tag{2.3}$$

where $F(s)$ and $f(s)$ satisfy

$$f(s) = 2e^{\gamma} s^{-1} \log(s-1), \quad F(s) = 2e^{\gamma} s^{-1} \quad \text{for} \quad 2 \leq s \leq 3. \tag{2.4}$$

Here γ is Euler's constant.

Proof. See Greaves [5, Chapter 4, Theorem 3]. □

Lemma 2.2. *Suppose that* $\Lambda_i, \Lambda_i^{\pm}$ *are real numbers satisfying* $\Lambda_i = 0$ *or* 1, $\Lambda_i^- \leq \Lambda_i \leq \Lambda_i^+$, $i = 1, 2, 3$. *Then*

$$\Lambda_1\Lambda_2\Lambda_3 \geq \Lambda_1^-\Lambda_2^+\Lambda_3^+ + \Lambda_1^+\Lambda_2^-\Lambda_3^+ + \Lambda_1^+\Lambda_2^+\Lambda_3^- - 2\Lambda_1^+\Lambda_2^+\Lambda_3^+. \tag{2.5}$$

Proof. The proof is similar to the proof of Lemma 13 in [2]. □

Lemma 2.3. *Suppose that* x, $y \in \mathbb{R}$ *and* $M \in \mathbb{N}$, $M \geq 3$. *Then*

$$e(-x\{y\}) = \sum_{|m|\leq M} c_m e(my) + O\left(\min\left(1, \frac{1}{M||y||}\right)\right),$$

where

$$c_m = \frac{1 - e(-x)}{2\pi i(x+m)}. \tag{2.6}$$

Proof. Proof can be find in Buriev [3, Lemma 12]. □

Lemma 2.4. *Consider the integral*

$$I = \int_a^b e(f(x))dx,$$

where $f(x)$ *is real function with continuous second derivative and monotonous first derivative. If* $|f'(x)| \geq h > 0$ *for all* $x \in [a, b]$, *then* $I \ll h^{-1}$.

Proof. See [10, Lemma 4.3]. □

3. BEGINNING OF THE PROOF

Let η, δ, ξ and μ be positive real numbers depending on c. We shall specify them later. Now we only assume that they satisfy the conditions

$$\xi + 3\delta < \frac{12}{25}, \qquad 2 < \frac{\delta}{\eta} < 3, \qquad \mu < 1. \tag{3.1}$$

We denote

$$X = N^{\frac{1}{c}}, \qquad z = X^{\eta}, \qquad D = X^{\delta}, \qquad \Delta = X^{\xi - c} \tag{3.2}$$

and

$$P(z) = \prod_{2<p<z} p. \tag{3.3}$$

Consider the sum

$$\Gamma = \sum_{\substack{\mu X < p_1,p_2,p_3 \le X \\ [p_1^c]+[p_2^c]+[p_3^c]=N \\ (p_i+2,P(z))=1,\ i=1,2,3}} (\log p_1)(\log p_2)(\log p_3). \tag{3.4}$$

If we prove the inequality

$$\Gamma > 0, \tag{3.5}$$

then equation (1.2) would have a solution in primes p_1, p_2, p_3 satisfying conditions in the sum Γ. Suppose that $p_i + 2$ has l prime factors, counted with multiplicities. From (3.2), (3.3) and $(p_i + 2, P(z)) = 1$ we have

$$X + 2 \ge p_i + 2 \ge z^l = X^{\eta l}$$

and then $l \le \frac{1}{\eta}$. This means that $p_i + 2$ has at most $[\eta^{-1}]$ prime factors counted with multiplicities. Therefore, to prove Theorem 1.1 we have to establish (3.5) for an appropriate choice of η.

For $i = 1, 2, 3$ we define

$$\Lambda_i = \sum_{d|(p_i+2,P(z))} \mu(d) = \begin{cases} 1 & \text{if } (p_i + 2, P(z)) = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{3.6}$$

Then we find that

$$\Gamma = \sum_{\substack{\mu X < p_1,p_2,p_3 \le X \\ [p_1^c]+[p_2^c]+[p_3^c]=N}} \Lambda_1\Lambda_2\Lambda_3(\log p_1)(\log p_2)(\log p_3).$$

We can write Γ as

$$\Gamma = \sum_{\mu X < p_1,p_2,p_3 \le X} \Lambda_1\Lambda_2\Lambda_3(\log p_1)(\log p_2)(\log p_3) \int_{-\frac{1}{2}}^{\frac{1}{2}} e(\alpha([p_1^c] + [p_2^c] + [p_3^c] - N))d\alpha.$$

Suppose that $\lambda^{\pm}(d)$ are the Rosser functions of level D. Let also denote

$$\Lambda_i^{\pm} = \sum_{d|(p_i+2,P(z))} \lambda^{\pm}(d), \qquad i = 1, 2, 3. \tag{3.7}$$

Then from Lemma 2.1, (3.6) and (3.7) we find that

$$\Lambda_i^- \le \Lambda_i \le \Lambda_i^+.$$

We use Lemma 2.2 and find that

$$\Gamma \ge \Gamma_1 + \Gamma_2 + \Gamma_3 - 2\Gamma_4,$$

where $\Gamma_1, \ldots, \Gamma_4$ are the contributions coming from the consecutive terms of the right-hand side of (2.5). We have $\Gamma_1 = \Gamma_2 = \Gamma_3$ and

$$\Gamma_1 = \sum_{\mu X < p_1, p_2, p_3 \leq X} \Lambda_1^- \Lambda_2^+ \Lambda_3^+ (\log p_1)(\log p_2)(\log p_3) \int_{-\frac{1}{2}}^{\frac{1}{2}} e(\alpha([p_1^c]+[p_2^c]+[p_3^c]-N))d\alpha,$$

$$\Gamma_4 = \sum_{\mu X < p_1, p_2, p_3 \leq X} \Lambda_1^+ \Lambda_2^+ \Lambda_3^+ (\log p_1)(\log p_2)(\log p_3) \int_{-\frac{1}{2}}^{\frac{1}{2}} e(\alpha([p_1^c]+[p_2^c]+[p_3^c]-N))d\alpha.$$

Hence, we get

$$\Gamma \geq 3\Gamma_1 - 2\Gamma_4. \tag{3.8}$$

Let us first consider Γ_1. We have

$$\Gamma_1 = \int_{-\frac{1}{2}}^{\frac{1}{2}} e(-N\alpha) L^-(\alpha) L^+(\alpha)^2 d\alpha, \tag{3.9}$$

where

$$L^{\pm}(\alpha) = \sum_{\mu X < p \leq X} (\log p) e(\alpha[p^c]) \sum_{d | (p+2, P(z))} \lambda^{\pm}(d).$$

Changing the order of summation we get

$$L^{\pm}(\alpha) = \sum_{d | P(z)} \lambda^{\pm}(d) \sum_{\substack{\mu X < p \leq X \\ p+2 \equiv 0 (\bmod d)}} (\log p) e(\alpha[p^c]).$$

We divide the integral from (3.9) into two parts:

$$\Gamma_1 = \Gamma_1' + \Gamma_1'', \tag{3.10}$$

where

$$\Gamma_1' = \int_{|\alpha| < \Delta} e(-N\alpha) L^-(\alpha) L^+(\alpha)^2 d\alpha, \tag{3.11}$$

$$\Gamma_1'' = \int_{\Delta < |\alpha| < \frac{1}{2}} e(-N\alpha) L^-(\alpha) L^+(\alpha)^2 d\alpha, \tag{3.12}$$

with Δ defined by (3.2).

Similarly, for Γ_4 we have

$$\Gamma_4 = \Gamma_4' + \Gamma_4'', \tag{3.13}$$

where

$$\Gamma_4' = \int\limits_{|\alpha|<\Delta} e(-N\alpha)L^+(\alpha)^3 d\alpha, \tag{3.14}$$

$$\Gamma_4'' = \int\limits_{\Delta<|\alpha|<\frac{1}{2}} e(-N\alpha)L^+(\alpha)^3 d\alpha, \tag{3.15}$$

and Δ is defined by (3.2).

4. THE INTEGRALS Γ_1' AND Γ_4'

We shall find an asymptotic formula for the integrals Γ_1' and Γ_4' defined by (3.11) and (3.14), respectively. The arithmetic structure of the Rosser weights $\lambda^{\pm}(d)$ is not important here, so we consider a sum of the form

$$L(\alpha) = \sum_{d\le D} \lambda(d) \sum_{\substack{\mu X<p\le X \\ p+2\equiv 0(\bmod d)}} (\log p) e(\alpha[p^c]), \tag{4.1}$$

where $\lambda(d)$ are real numbers satisfying

$$|\lambda(d)| \le 1, \qquad \lambda(d) = 0 \quad \text{if} \quad 2|d \quad \text{or} \quad \mu(d) = 0. \tag{4.2}$$

It is easy to see that

$$\begin{aligned} L(\alpha) &= \sum_{d\le D} \lambda(d) \sum_{\substack{\mu X<p\le X \\ p+2\equiv 0(\bmod d)}} (\log p) e(\alpha p^c + O(|\alpha|)) \\ &= \sum_{d\le D} \lambda(d) \sum_{\substack{\mu X<p\le X \\ p+2\equiv 0(\bmod d)}} (\log p) e(\alpha p^c)(1 + O(|\alpha|)) \\ &= \overline{L}(\alpha) + O(\Delta X(\log X)), \end{aligned} \tag{4.3}$$

where

$$\overline{L}(\alpha) = \sum_{d\le D} \lambda(d) \sum_{\substack{\mu X<p\le X \\ p+2\equiv 0(\bmod d)}} (\log p) e(\alpha p^c).$$

For $\overline{L}(\alpha)$ we use the asymptotic formula from Lemma 10 in [15]. From (3.1) and (3.2) we see that, when $|\alpha| < \Delta$, then for every constant $A > 0$, we have

$$\overline{L}(\alpha) = \sum_{d\le D} \frac{\lambda(d)}{\varphi(d)} I(\alpha) + O(X(\log X)^{-A}), \tag{4.4}$$

where

$$I(\alpha) = \int\limits_{\mu X}^{X} e(\alpha t^c)dt. \tag{4.5}$$

Hence from (3.2), (4.3) and (4.4) we see that

$$L(\alpha) = \sum_{d\le D} \frac{\lambda(d)}{\varphi(d)} I(\alpha) + O(X(\log X)^{-A}). \tag{4.6}$$

From (2.1) and (4.6) we find

$$L^{\pm}(\alpha) = \mathcal{N}^{\pm} I(\alpha) + O(X(\log X)^{-A}), \qquad \text{for } |\alpha| < \Delta. \tag{4.7}$$

Let

$$\mathcal{M}^{\pm} = \mathcal{N}^{\pm} I(\alpha). \tag{4.8}$$

It is easy to see that

$$\mathcal{N}^{\pm} \ll \log X. \tag{4.9}$$

We use (4.7), (4.8) and the identity

$$L^-(L^+)^2 = (L^- - \mathcal{M}^-)(L^+)^2 + (L^+ - \mathcal{M}^+)\mathcal{M}^- L^+ + (L^+ - \mathcal{M}^+)\mathcal{M}^+\mathcal{M}^- + \mathcal{M}^-(\mathcal{M}^+)^2$$

to find that

$$|L^-(L^+)^2 - \mathcal{M}^-(\mathcal{M}^+)^2| \ll X(\log X)^{-A}\left(|L^+|^2 + |\mathcal{M}^-|^2 + |\mathcal{M}^+|^2\right). \tag{4.10}$$

Let

$$B = \int\limits_{|\alpha|<\Delta} e(-N\alpha)\mathcal{M}^-(\alpha)(\mathcal{M}^+(\alpha))^2 d\alpha. \tag{4.11}$$

From (3.11), (4.9) – (4.11) we have

$$\Gamma_1' - B \ll X(\log X)^{2-A}\left(\int\limits_{|\alpha|<\Delta} |L^+(\alpha)|^2 d\alpha + \int\limits_{|\alpha|<\Delta} |I(\alpha)|^2 d\alpha\right).$$

We need the next lemma, which is an analog of Lemma 11 in [15].

Lemma 4.5. *If $\Delta \le X^{1-c}$, then for the sum $L(\alpha)$ defined by* (4.1) *and for the integral $I(\alpha)$ defined by* (4.5) *we have*

$$\int\limits_{|\alpha|<\Delta} |L(\alpha)|^2 d\alpha \ll X^{2-c}(\log X)^6,$$

$$\int\limits_{|\alpha|<\Delta} |I(\alpha)|^2 d\alpha \ll X^{2-c}(\log X)^6,$$

$$\int\limits_{|\alpha|<1} |L(\alpha)|^2 d\alpha \ll X(\log X)^5.$$

Proof. The proof is similar to the proof of Lemma 11 in [15]. □

Hence

$$\Gamma_1' - B \ll X^{3-c}(\log X)^{8-A}. \tag{4.12}$$

Consider now the integral

$$B_1 = \int\limits_{-\infty}^{\infty} e(-N\alpha)I(\alpha)^3 d\alpha. \tag{4.13}$$

Using the method in Lemma 5.6.1 in [11] we find

$$B_1 \gg X^{3-c}. \tag{4.14}$$

For $I(\alpha)$ we apply Lemma 2.4 and see that $I(\alpha) \ll |\alpha|^{-1}X^{1-c}$. Then from (3.2), (4.8), (4.11) and (4.13) we find

$$|\mathcal{N}^-(\mathcal{N}^+)^2 B_1 - B| \ll (\log X)^3 \int\limits_{|\alpha|>\Delta} |I(\alpha)|^3 d\alpha \ll (\log x)^3 X^{3-c-2\xi}. \tag{4.15}$$

If $A = 12$, then using (4.12) and (4.15) we find

$$\Gamma_1' = \mathcal{N}^-(\mathcal{N}^+)^2 B_1 + O(X^{3-c}(\log X)^{-4}). \tag{4.16}$$

We proceed with Γ_4' in the same way and prove that

$$\Gamma_4' = (\mathcal{N}^+)^3 B_1 + O(X^{3-c}(\log X)^{-4}). \tag{4.17}$$

5. ESTIMATION OF INTEGRALS Γ_1'' AND Γ_4'' AND COMPLETION OF THE PROOF

In this section we consider the integrals Γ_1'' and Γ_4'' defined by (3.12) and (3.15) respectively. We shall show that Γ_1'' and Γ_4'' are small enough. Now we assume that

$$\xi = \frac{16c-5}{32}, \qquad \delta = \frac{17-16c}{32}. \tag{5.1}$$

It is obvious that for Γ_1'' defined by (3.12) we have

$$\Gamma_1'' \ll \max_{\Delta \le |\alpha| \le \frac{1}{2}} |L^-(\alpha)| \int\limits_0^1 |L^+(\alpha)|^2 d\alpha.$$

We use Lemma 4.5 and find that

$$\Gamma_1'' \ll X(\log X)^5 \max_{\Delta \le |\alpha| \le \frac{1}{2}} |L^-(\alpha)|. \tag{5.2}$$

From (4.1) we see that

$$L(\alpha) = L_1(\alpha) + O\left(X^{\frac{1}{2}+\varepsilon}\right), \tag{5.3}$$

where

$$L_1(\alpha) = \sum_{d\le D}\lambda(d) \sum_{\substack{\mu X<n\le X\\ n+2\equiv 0(\bmod d)}} \Lambda(n)e(\alpha[n^c]).$$

Let $M = X^\kappa$ for some κ, which will be specified later. Now for $L_1(\alpha)$ we apply Lemma 2.3 with parameters $x = \alpha$, $y = n^c$ and M (note that $[t] = t - \{t\}$). We obtain

$$\begin{aligned} L_1(\alpha) = &\sum_{|m|\le M} c_m \sum_{d\le D}\lambda(d) \sum_{\substack{\mu X<n\le X\\ n+2\equiv 0(\bmod d)}} \Lambda(n)e((\alpha+m)n^c) \\ &+ O\left(X^\varepsilon \sum_{\mu X<n\le X} \min\left(1, \frac{1}{M||n^c||}\right)\right). \end{aligned} \tag{5.4}$$

We need the following

Lemma 5.6. *Suppose that D, Δ are defined by (3.2) and ξ, δ are specified by (5.1). Suppose also that $\lambda(d)$ satisfy (4.2) and c_m are defined by (2.6). Then*

$$\begin{aligned} &\max_{\Delta\le\alpha\le M+1}\left|\sum_{|m|\le M} c_m \sum_{d\le D}\lambda(d) \sum_{\substack{\mu X<n\le X\\ n+2\equiv 0(\bmod d)}} \Lambda(n)e(\alpha n^c)\right| \\ &\ll x^\varepsilon\left(X^{\frac{1}{3}+\frac{c}{2}}DM^{\frac{1}{2}} + X^{1-\frac{c}{2}}\Delta^{-\frac{1}{2}} + X^{\frac{3}{4}+\frac{c}{6}}D^{\frac{2}{3}}M^{\frac{1}{6}} + X^{\frac{5}{6}} + X^{1-\frac{c}{6}}D^{\frac{1}{3}}\Delta^{-\frac{1}{6}} + X^{1-\frac{c}{4}}\Delta^{-\frac{1}{4}}\right). \end{aligned}$$

Proof. See Lemma 15 in [15]. □

We also need the following result.

Lemma 5.7. *One has*

$$\sum_{\mu X<n\le X} \min\left(1, \frac{1}{M||n^c||}\right) \ll X^\varepsilon\left(XM^{-1} + M^{\frac{1}{2}}X^{\frac{c}{2}}\right). \tag{5.5}$$

Proof. From [13, Lemma 5.2.3] we know that the Fourier series

$$\min\left(1, \frac{1}{M||n^c||}\right) = \sum_{k\in\mathbb{N}} b_M(k)e(kn^c), \tag{5.6}$$

has Fourier coefficients satisfying

$$|b_M(k)| \le \begin{cases} \frac{4\log M}{M} & \text{if } k\in\mathbb{Z}, \\ \frac{M}{k^2} & \text{if } k\in\mathbb{Z}, k\ne 0. \end{cases} \tag{5.7}$$

From (5.6) we get

$$\sum_{\mu X<n\leq X} \min\left(1, \frac{1}{M||n^c||}\right) = \sum_{\mu X<n\leq X} \sum_{k\in\mathbb{N}} b_M(k)e(kn^c). \tag{5.8}$$

Changing the order of summation in last formula we obtain

$$\sum_{\mu X<n\leq X} \min\left(1, \frac{1}{M||n^c||}\right) = \sum_{k\in\mathbb{N}} b_M(k)H(k),$$

where

$$H(k) = \sum_{\mu X<n\leq X} e(kn^c).$$

Now using (5.7) and (5.8) and the identity $|H(k)| = |H(-k)|$ we find

$$\sum_{\mu X<n\leq X} \min\left(1, \frac{1}{M||n^c||}\right) \ll \frac{X\log M}{M} + \frac{\log M}{M} \sum_{1\leq k\leq M} |H(k)| + M \sum_{k>M} \frac{|H(k)|}{k^2}. \tag{5.9}$$

If $\theta(x) = kx^c$, then $\theta''(x) = c(c-1)kx^{c-2} \asymp kX^{c-2}$ uniformly for $x \in [\mu X, X]$. Hence, we can apply Van der Corput's theorem (see [6, Chapt. 1, Theorem 5] to obtain

$$H(k) \ll k^{\frac{1}{2}} X^{\frac{c}{2}} + k^{-\frac{1}{2}} X^{1-\frac{c}{2}}. \tag{5.10}$$

Hence from (5.9) and (5.10) we prove (5.5). □

When combining Lemma 5.6, Lemma 5.7 and (5.3) – (5.4) we find that

$$\max_{\Delta\leq\alpha\leq M+1} |L(\alpha)| \ll x^{\varepsilon}\Big(X^{\frac{1}{3}+\frac{c}{2}} D M^{\frac{1}{2}} + X^{1-\frac{c}{2}}\Delta^{-\frac{1}{2}} + X^{\frac{3}{4}+\frac{c}{6}} D^{\frac{2}{3}} M^{\frac{1}{6}} + \\ + X^{\frac{5}{6}} + X^{1-\frac{c}{6}} D^{\frac{1}{3}}\Delta^{-\frac{1}{6}} + X^{1-\frac{c}{4}}\Delta^{-\frac{1}{4}} + XM^{-1}\Big).$$

Then from last formula, (3.2) and (5.2) we find

$$\Gamma_1'' \ll x^{\varepsilon}\left(X^{\frac{4}{3}+\frac{c}{2}+\delta+\frac{\kappa}{2}} + X^{\frac{7}{4}+\frac{c}{6}+\frac{2\delta}{3}+\frac{\kappa}{6}} + X^{\frac{11}{6}} + X^{2+\frac{\delta}{3}-\frac{\xi}{8}} + X^{2-\kappa}\right). \tag{5.11}$$

If we choose $\kappa = \frac{8c-5}{56}$, then from (5.1) and (5.11) we conclude that if $1 < c < \frac{17}{16}$ then

$$\Gamma_1'' \ll X^{3-c-\varepsilon}.$$

From (3.8), (3.10), (3.13) and (4.14) – (4.17) we conclude that

$$\Gamma \geq |3\mathcal{N}^- - 2\mathcal{N}^+|(N^+)^3 B_1 + O(X^{3-c}(\log x)^{-4}). \tag{5.12}$$

Now we shall find a lower bound for the difference $3\mathcal{N}^- - 2\mathcal{N}^+$. It is easy to see that

$$\mathcal{B} \asymp (\log X)^{-1}. \tag{5.13}$$

From (2.2) and (2.3) we see that

$$3\mathcal{N}^- - 2\mathcal{N}^+ \geq \mathcal{B}(3f(s_0) - F(s_0)) + O\left((\log X)^{-\frac{4}{3}}\right),$$

where s_0 is defined by (2.1) and $F(s)$ and $f(s)$ are defined by (2.4). If we choose $s_0 = 2.95$, then from (2.1), (3.2) and (5.1) we find

$$\eta = \frac{\delta}{2.95} = \frac{17 - 16c}{94.4}$$

and also from (2.4) we find $3f(s_0) - F(s_0) > 0$.

Now from (2.2), (4.14), (5.12) and (5.13) we obtain

$$\Gamma \gg X^{3-c}(\log X)^{-3}.$$

Therefore $\Gamma > 0$ and this proves Theorem 1.1. □

ACKNOWLEDGEMENTS. The author wishes to express his thanks to Professor D. Tolev for suggesting the problem and for the helpful conversations. The research is partially supported by the Sofia University Research Fund through Grant 80-10-215/2017.

6. REFERENCES

[1] Baker, R., Weingartner, A.: A ternary diophantine inequality over primes. *Acta Arith.*, **162**, 2014, 159-196.

[2] Brüdern, J., Fouvry, E.: Lagrange's four squares theorem with almost prime variables. *J. Reine Angew. Math.*, **454**, 1994, 59-96.

[3] Buriev, K.: *Additive Problems with Prime Numbers.* Ph.D. Thesis, Moscow State University, 1989 (in Russian).

[4] Chen, J. R.: On the representation of a large even integer as the sum of a prime and the product of at most two primes. *Sci. Sinica*, **16**, 1973, 157-167.

[5] Greaves, G.: *Sieves in Number Theory*, Springer, 2001.

[6] Karatsuba, A. A.: *Basic Analytic Number Theory*, Springer, 1993.

[7] Laporta, M. B., Tolev, D. I.: On an equation with prime numbers. *Mat. Zametki*, **57**, 1995 (in Russian).

[8] Matomäki, K., Shao, X.: Vinogradov's three prime theorem with almost twin primes. arXiv:1512.03213, 2017.

[9] Piatetski-Shapiro, I. I.: On a variant of Waring-Goldbach's problem. *Mat. Sb.*, **30** (72), no. 1, 1952, 105–120 (in Russian).

[10] Titchmarsh, E. G.: *The Theory of the Riemann Zeta-function* (revised by D. R. Heath-Brown), Clarendon Press, Oxford, 1986.

[11] Todorova, T.: *Three Problems of Analytic Number Theory*, Ph.D. Thesis, Sofia University, Sofia, 2015 (in Bulgarian).

[12] Tolev, D. I.: Additive problems with prime numbers of special type. *Acta Arith.*, **96**, no. 11, 2000, 53–88.

[13] Tolev, D. I.: *Lecture on Elementary and Analytic Number Theory*, Sofia University Press, 2016 (in Bulgarian).

[14] Tolev, D. I.: On a diophantine inequality involving prime numbers. *Acta Arith.*, **61**, no. 3, 1992, 289–306.

[15] Tolev, D. I.: On a diophantine inequality with prime numbers of a special type. arXiv:1701.07652, 2017.

[16] Vinogradov, I. M.: Representation of an odd number as a sum of three primes. *Dokl. Akad. Nauk SSSR*, **15**, 1937, 169–172 (in Russian).

Received on July 13, 2017

Zhivko Petrov

Department of Mathematics and Informatics
University of Sofia
5 James Bourchier Blvd.
1164 Sofia
BULGARIA

E-mail: zhpetrov@fmi.uni-sofia.bg

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 104

ANNUAL OF SOFIA UNIVERSITY „ST. KLIMENT OHRIDSKI“
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 104

FAST CONVERGING SEQUENCE TO EULER-MASCHERONI CONSTANT

IVAN GEORGIEV

The aim of the paper is to apply an exponential trapezoidal quadrature rule to an integral representation of the Euler-Mascheroni constant. The resulting sequence has subexponential convergence rate and is particularly useful for estimating the subrecursive complexity of the constant.

Keywords: computable real number, subrecursive complexity, Euler-Mascheroni constant, exponential trapezoidal rule.
2000 Math. Subject Classification: 03F60, 03D15.

1. INTRODUCTION

The Euler-Mascheroni constant is usually denoted by γ and is defined by the equality

$$\gamma = \lim_{n\to\infty}\left(1+\frac{1}{2}+\cdots+\frac{1}{n}-\ln n\right).$$

Since this sequence converges to γ very slowly, it is not suitable for the effective computation of γ. This is why other much faster methods are invented, like the method of Karatsuba in [2], which is suitable to prove polynomial-time computability of γ.

Our aim is to study the complexity of γ in another context, namely the subrecursive class $\mathcal{M}^2$, contained in the third level $\mathcal{E}^2$ of Grzegorczyk's hierarchy. It

appears that the known methods for effective computation of γ are not suitable in this context.

The author has proven in [1] that γ is $\mathcal{M}^2$-computable, as a consequence of some results on $\mathcal{M}^2$-computability of integration. The aim of the present paper is to extract an actual sequence from this proof, which converges to γ with subexponential convergence rate. This will be done by a careful estimation of the error of approximation.

The starting point is the following well-known representation

$$-\gamma = \int_0^\infty e^{-x} \ln x \, dx.$$

Let us define

$$I_1 = \int_0^1 e^{-x} \ln x \; dx = -\int_1^\infty t^{-2} \, e^{-\frac{1}{t}} \ln t \, dt,$$

$$I_2 = \int_1^\infty e^{-x} \ln x \, dx,$$

so that $\gamma = -I_1 - I_2$.

2. EXPONENTIAL TRAPEZOIDAL RULE

The trapezoidal quadrature rule is a famous method for numerical integration. It approximates the definite integral of a real function over an interval with the area of a trapezoid. In practice, the initial interval is split into sufficiently many subintervals of equal length, the rule is applied to each one of them and the obtained results are summed. It is known that in certain situations the result of this method approaches the exact answer very quickly when increasing the number of subintervals. This phenomenon is described in great detail in the paper [3]. The results from Section 1.5 in [3] can be used to estimate the subrecursive complexity of the integration operation. More concretely, the author has proved in [1] that the definite integral of an analytic real function, belonging to a certain low subrecursive class of computable real functions, is itself computable in this class. The steps in this proof proceed as follows:

1. We start with a function θ, which is analytic on an open set of the complex plane containing the interval $[\alpha, \beta]$ and we wish to approximate $I = \int_\alpha^\beta \theta(x) \; dx$.
2. By a linear change of variables we may assume $[\alpha, \beta] = [-1, 1]$.
3. We apply the transformation $x = \tanh(t)$ and thus we obtain the integral

$$I = \int_{-\infty}^{+\infty} \theta(\tanh(t)) \frac{1}{\cosh^2(t)} \, dt.$$

(This is the so-called *tanh-rule.*)

4. We discretise the integral by the trapezoidal quadrature rule with step h and then truncate the obtained infinite series to its (two-way) n-th partial sum. Finally, we put $h = \dfrac{1}{\sqrt{n}}$ and the result is

$$I_h^{[n]} = h \sum_{k=-n}^{n} \theta(\tanh(kh)) \frac{1}{\cosh^2(kh)} = \frac{1}{\sqrt{n}} \sum_{k=-n}^{n} \theta\left(\tanh(\frac{k}{\sqrt{n}})\right) \frac{1}{\cosh^2(\frac{k}{\sqrt{n}})}$$

for every positive integer number n.

5. The error of the approximation is

$$|I - I_h^{[n]}| \leq \frac{M(2\pi + 4)}{e^{\frac{1}{C}\sqrt{n}} - 1},$$

where C, M depend on θ, α, β only. (See the proof of Theorem 5.3 in [1].)

3. APPROXIMATION OF I_1

Let $\phi : [1, \infty) \to \mathbb{R}$ be defined by

$$\phi(t) = t^{-2} e^{-\frac{1}{t}} \ln t.$$

Lemma 1. *For any fixed $\xi > 0$, the function ϕ has an analytic continuation to an open set of the complex plane $\mathbb{C}$ containing the set*

$$D_\xi = \left\{ z \in \mathbb{C} \,\middle|\, \operatorname{Re} z \in [1, \xi + 1],\ \operatorname{Im} z \in \left[-\frac{\xi}{2}, \frac{\xi}{2}\right] \right\}. \tag{1}$$

Moreover, $|\phi(z)| \leq \ln(\xi + 1) + \frac{7}{2}$ for $z \in D_\xi$.

Proof. The first claim is obviously true, assuming the principal value of the logarithmic function with branch cut the non-positive real numbers. In fact, ϕ has an analytic continuation defined on the whole halfplane $\operatorname{Re} z > 0$.

Now we estimate $|\phi(z)|$, $z \in D_\xi$. Let $\xi > 0$ and $z \in D_\xi$, where $z = x + Bi$ with $1 \leq x \leq \xi + 1$ and $|B| \leq \frac{\xi}{2}$. We have

$$|\phi(z)| = \frac{1}{|z|^2} |e^{-\frac{1}{z}}| . |\ln z|.$$

For the first two factors we have

$$\frac{1}{|z|^2} = \frac{1}{x^2 + B^2} \leq \frac{1}{1 + B^2} \leq 1,$$

$$|e^{-\frac{1}{z}}| = e^{\operatorname{Re}(-\frac{1}{z})} = e^{-\frac{x}{x^2 + B^2}} < 1.$$

The third factor is estimated as follows:

$$\begin{aligned}|\ln z| &= \sqrt{\ln^2|z| + \operatorname{Arg}^2 z} \le |\ln|z|| + |\operatorname{Arg} z| \\ &\le \frac{1}{2}\ln(x^2+B^2) + \pi \le \frac{1}{2}\ln((\xi+1)^2 + \frac{\xi^2}{4}) + \pi \\ &< \frac{1}{2}\ln(\frac{5}{4}(\xi+1)^2) + \pi < \ln(\xi+1) + \frac{7}{2}.\end{aligned}$$

The result follows trivially. □

We replace the integral $-I_1 = \int_1^\infty \phi(t)\,dt$ by

$$J_\xi^1 = \int_1^{\xi+1} \phi(t)\,dt,$$

where $\xi > 0$ will be specified later. For the truncation error $e_1(\xi)$ we have (using that $e^{-x} \le 1$ for any $x \ge 0$)

$$\begin{aligned}e_1(\xi) &= \int_{\xi+1}^\infty t^{-2}\,e^{-\frac{1}{t}}\ln t\,dt = \int_0^{\frac{1}{\xi+1}} (-\ln x)e^{-x}\,dx \\ &\le \int_0^{\frac{1}{\xi+1}} (-\ln x)\,dx = \frac{\ln(\xi+1)+1}{\xi+1}.\end{aligned}$$

Following the steps from the previous section we have

$$J_\xi^1 = \frac{\xi}{2}\int_{-1}^1 \phi_1(u,\xi)\,du,$$

where

$$\phi_1(u,\xi) = \phi\left(\frac{\xi}{2}.u + \frac{\xi+2}{2}\right).$$

Then we approximate J_ξ^1 by

$$J_{\xi,n}^1 = \frac{\xi}{2}\frac{1}{\sqrt{n}}\sum_{k=-n}^{n} \phi_1\left(\tanh(\frac{k}{\sqrt{n}}),\xi\right)\frac{1}{\cosh^2(\frac{k}{\sqrt{n}})}.$$

In the proof of Theorem 5.3 in [1] we can arrange $A' = 1, a = \frac{\pi}{4}, C = 1$ and by Lemma 1 we obtain

$$\left|J_\xi^1 - J_{\xi,n}^1\right| \le \frac{\xi}{2}\frac{(\ln(\xi+1)+\frac{7}{2})(2\pi+4)}{e^{\sqrt{n}}-1}$$

for any $\xi > 0$ and positive integer number n.

4. APPROXIMATION OF I_2

Let $\psi : [1, \infty) \to \mathbb{R}$ be defined by

$$\psi(x) = e^{-x} \ln x.$$

Lemma 2. *For any fixed $\xi > 0$, the function ψ has an analytic continuation to an open set in $\mathbb{C}$ containing the set D_ξ defined in (1).*

Moreover, $|\psi(z)| \leq \ln(\xi + 1) + \frac{7}{2}$ for $z \in D_\xi$.

Proof. Analogous to the proof of Lemma 1, this time using that

$$|e^{-z}| = e^{\operatorname{Re}(-z)} = e^{-\operatorname{Re}(z)} < 1,$$

for any complex number z with $\operatorname{Re} z > 0$. □

We replace I_2 by

$$J_\xi^2 = \int_1^{\xi+1} \psi(x)\, dx,$$

where $\xi > 0$. For the error $e_2(\xi)$ after this replacement we have (using $\ln x \leq x$ for any real number $x \geq 1$)

$$e_2(\xi) = \int_{\xi+1}^{\infty} e^{-x} \ln x\, dt \leq \int_{\xi+1}^{\infty} e^{-x} x\, dx = \frac{\xi + 2}{e^{\xi+1}}.$$

Following the steps from Section 2 we have

$$J_\xi^2 = \frac{\xi}{2} \int_{-1}^{1} \psi_1(u, \xi)\, du,$$

where

$$\psi_1(u, \xi) = \psi\left(\frac{\xi}{2}.u + \frac{\xi + 2}{2}\right).$$

Then we approximate J_ξ^2 by

$$J_{\xi,n}^2 = \frac{\xi}{2} \frac{1}{\sqrt{n}} \sum_{k=-n}^{n} \psi_1\left(\tanh(\frac{k}{\sqrt{n}}), \xi\right) \frac{1}{\cosh^2(\frac{k}{\sqrt{n}})}.$$

Again we can arrange $C = 1$ in the proof of Theorem 5.3 in [1] and similarly obtain

$$\left|J_\xi^2 - J_{\xi,n}^2\right| \leq \frac{\xi}{2} \frac{(\ln(\xi + 1) + \frac{7}{2})(2\pi + 4)}{e^{\sqrt{n}} - 1}$$

for any $\xi > 0$ and positive integer number n.

5. MAIN RESULT

After approximating $-I_1$ by $J^1_{\xi,n}$ and I_2 by $J^2_{\xi,n}$, we are ready to approximate $\gamma = -I_1 - I_2$ and estimate the error of the approximation by choosing a suitable ξ depending on n.

Let $p(x) = \dfrac{\ln x + 1}{x}$. It is easy to see that p is decreasing in the interval $[1, +\infty)$. Therefore,

$$e_2(\xi) = p(e^{\xi+1}) \leq p(\xi + 1) = e_1(\xi),$$

since $e^{\xi+1} > \xi + 1$ for $\xi > 0$.

Now the approximation of γ by $J^1_{\xi,n} - J^2_{\xi,n}$ leads to an error, which is bounded above by

$$e_1(\xi) + e_2(\xi) + 2\frac{\xi}{2}\frac{(\ln(\xi+1) + \frac{7}{2})(2\pi + 4)}{e^{\sqrt{n}} - 1} \leq 2e_1(\xi) + \frac{\xi(\ln(\xi+1) + \frac{7}{2})(2\pi + 4)}{e^{\sqrt{n}} - 1}$$

for any $\xi > 0$ and any positive integer number n. To produce the desired sequence A, we choose $\xi = \sqrt{e}^{\sqrt{n}} - 1$ to obtain

$$A(n) = J^1_{\xi,n} - J^2_{\xi,n} = \frac{\sqrt{e}^{\sqrt{n}} - 1}{2\sqrt{n}} \sum_{k=-n}^{n} \theta\left(\tanh(\frac{k}{\sqrt{n}}), \sqrt{e}^{\sqrt{n}} - 1\right) \frac{1}{\cosh^2(\frac{k}{\sqrt{n}})},$$

where $\theta(u, \xi) = \phi_1(u, \xi) - \psi_1(u, \xi)$.

For any positive integer n, the error of approximation of γ by $A(n)$ satisfies

$$\begin{aligned} |A(n) - \gamma| &\leq 2e_1(\sqrt{e}^{\sqrt{n}} - 1) + \frac{(\sqrt{e}^{\sqrt{n}} - 1)(\frac{1}{2}\sqrt{n} + \frac{7}{2})(2\pi + 4)}{e^{\sqrt{n}} - 1} \\ &= \frac{2(\frac{1}{2}\sqrt{n} + 1)}{\sqrt{e}^{\sqrt{n}}} + \frac{(\sqrt{n} + 7)(\pi + 2)}{\sqrt{e}^{\sqrt{n}} + 1} \leq \frac{(\pi + 3)\sqrt{n} + 7\pi + 16}{\sqrt{e}^{\sqrt{n}}}. \end{aligned}$$

6. CONCLUSION

The sequence A is suitable for proving $\mathcal{M}^2$-computability of γ. It turns out that the sequence B, defined by

$$B(m) = A(\lfloor \log_2(m+1) \rfloor^2),$$

is $\mathcal{M}^2$-computable and has polynomial convergence rate.

Unfortunately, the expression for the general term of A is too complex to be used in practice for computation of many decimal digits of γ. Simple numerical

experiments with Simulink/Matlab using high precision calculations give 18 correct decimal digits for $n = 10000$ and 30 correct decimal digits for $n = 25000$.

ACKNOWLEDGEMENTS. This paper is supported by the Bulgarian National Science Fund through the project "Models of computability", DN-02-16/19.12.2016.

7. REFERENCES

[1] Georgiev, I.: On subrecursive complexity of integration (submitted for publication in *Ann. of Pure and Appl. Log.*).

[2] Karatsuba, E.: On the computation of the Euler constant γ. *Numer. Alg.*, **24**, 2000, 83–97.

[3] Trefethen, L., Weideman, J.: The exponentially convergent trapezoidal rule. *SIAM Review*, **56**, no. 3, 2014, 385–458.

Received on October 15, 2017

Ivan Georgiev

Department of Mathematics and Physics
Faculty of Natural Sciences
"Prof. d-r Asen Zlatarov" University
1 Prof. Yakimov Str., 8010, Burgas
BULGARIA

E-mail: ivandg@yahoo.com

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 104

ANNUAL OF SOFIA UNIVERSITY „ST. KLIMENT OHRIDSKI“
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 104

MULTITYPE BRANCHING PROCESSES IN CONTINUOUS TIME AS MODELS OF CANCER

KALOYAN VITANOV, MAROUSSIA SLAVTCHOVA-BOJKOVA

A generalization of the results for two-type decomposable branching processes model in continuous time derived in [9] is obtained. More precisely, the system of integral equations for probability generating functions (p.g.f.) of the $(n+1)-$type processes is obtained. Another important result is the recursive equations satisfied by the p.g.f. of the number of mutations. The results obtained are the base for further research of probabilities of extinction and estimation of the risk of cancer recurrence.

Keywords: decomposable multi–type branching processes, continuous time, mutations.

2000 Math. Subject Classification: 60J85, 60J80.

1. INTRODUCTION

The aim of this paper stems from the attempts to model mathematically the behavior of cancer cell populations subjected to some treatment, i.e. chemotherapy, radiotherapy or another type of medical treatment. As a result of the treatment, the reproduction rate of the cancer cells decreases. In terms of the branching process theory, the reproduction of these cells acquires subcritical characteristics meaning that the mean of the offspring per progenitor is less than 1. It is well-known (see e.g. [1]) that a subcritical population goes extinct with probability 1 (or almost surely) given a sufficiently long period of time. The empirical experience, however, shows that during the division of the subcritical cells it is possible for mutations to

occur. These mutant cells may have supercritical characteristics, i.e. the mean of cells per progenitor is greater than 1, meaning that after a sufficiently long period of time either the population completely disappears with probability $0 < q < 1$, or it increases (theoretically) indefinitely with probability $1 - q$. Models in discrete time with more than one subcritical cell type and one supercritical cell type are considered in [7], [8] and others. Similar models, but in a continuous time are studied in [2], [9] and [10].

The framework of classical branching theory, on which the results in the current paper are based, is developed in the classical books [3, 1, 4, 6]. Further development on the application of branching processes in biology can be found in [5] and [2].

This work aims to expand the results in [9] and [10], in the case of more than one subcritical cell types. Let us note that considering a branching process with more than one type of subcritical cancer cells is actually of interest from a practical point of view. Cancer is a multi-stage disease and frequently metastases are observed after local eradication of the tumor and subsequent adequate treatment meaning that the cancer spreads to different parts of the body, depending on the particular type of cancer. The differences in the environment may encourage the variation in the characteristics of the cancer cells, leading to the differentiation of different types. This motivates studying a more complicated model allowing more than one type of subcritical cells.

In what follows, we will discuss the general case in which the offspring of a subcritical cell of an arbitrary type may be of any other type, i.e. from one of the subcritical types or from the supercritical one. On the other hand, we will limit ourselves to the case of one supercritical type, where the generation of supercritical type cells can only be of the supercritical type. This means we will explore a decomposable branching process (for this type of branching processes, the theoretical results are far less abundant).

The paper is organized as follows: Section 2 introduces the branching process model with $n + 1$ types of cells in continuous time. Section 3 contains the main results and proofs. In Theorem 1 we prove the basic functional equations for p.g.f. of the process itself. In Theorem 2 we obtain the p.g.f. of both the number of mutations occurred up to time t and the number of mutations to the escape type cells in the whole process.

2. FORMULATION OF THE MODEL

Before proceeding with the main model in Definition 2 let us recall:

Definition 1. An age–dependent branching process $\{Z(t), t \geq 0\}$ with one type of cells, being the number of cells alive at time t, starting at time 0 with a single progenitor of age 0, i.e. $Z(0)$, is called a Bellman–Harris branching process (BHBP), if:

(i) The life–length τ of the progenitor has distribution $G(t) = P(\tau \leq t)$, $G(0^+) = 0$;

(ii) Each cell produces k, $k \geq 0$, similar cells of age 0 at the end of its life with probability $p_k, 0 \leq p_k \leq 1$, which have the same life–length distribution $G(t)$ and reproduce independently according to $\{p_k\}$, $\sum_{k=0}^{\infty} p_k = 1$.

The single–type Bellman–Harris branching process together with proper biological applications is studied by Jagers [4] and more theoretically by Athreya and Ney [1].

Now we will state the constructive definition of the main model, as a model of a multi-stage cancer process, where type 0 is reserved for the malignant-type (supercritical) cells, characterized by high capacity of division, and the types i, $i = 1, \ldots, n$, correspond to the successfully treated (subcritical) types of cancer cells.

Definition 2. Define the multi-type branching model of mutations with $n+1$ types of cells as follows:

(i) There are $n+1$, $n > 1$ different types of cells;

(ii) Each cell type has the properties stated in the definition of the single–type BHBP (although it is not necessary for the offspring to be of the same type as the mother cell, see (iii)). Each type has a (possibly) distinctive (continuous) distribution $G_i(t) = P(\tau_i \leq t)$, $G_i(0^+) = 0$, of the life–length τ_i, and (possibly) distinctive (discrete) distribution $\{p_{ik}\}$, $\sum_{k=0}^{\infty} p_{ik} = 1$ of the number of cells in the offspring ν_i, $i = 0, \ldots, n$;

(iii) Each of the descendants of a subcritical type i, $i = 1, \ldots, n$, can mutate at birth, independently of other cells, to any other type, with probabilities u_{ik}, $0 \leq u_{ik} \leq 1$, $k = 0, \ldots, n$, $\sum_{k=0}^{n} u_{ik} = 1$. Descendants of the supercritical type 0 can not mutate to another type, i. e. $u_{00} = 1$, meaning that there is no backward mutation. Because of the mutation scheme of type 0 the branching process is decomposable.

(iv) Cells of type i, $i = 1, \ldots, n$, have subcritical reproduction, i.e. for the offspring mean m_i, we have $0 < m_i < 1$. Cells of type 0 have supercritical reproduction, i.e. have reproduction mean m_0, with $1 < m_0 < \infty$;

(v) Formally, we denote $\{\mathbf{Z}(t) = (Z^0(t), Z^1(t), \ldots, Z^n(t)), t \geq 0\}$, where $\{Z^i(t), t \geq 0\}$ stands for the number of cells of type i, $i = 0, \ldots, n$, at time t, respectively.

From now on, unless stated otherwise, we assume that the process starts with just one cell of type i, $i = 1, \ldots, n$, i.e. $Z^0(0) = 0$, $Z^i(0) = 1$ and $Z^j(0) = 0$, $j \neq i$.

The p.g.f. of the offspring ν_i of type i cells will be denoted by $f_i(s)$, $i = 0, \ldots, n$, and

$$f_i(s) = E(s^{\nu_i}) = \sum_{k=0}^{\infty} p_{ik} s^k, \quad |s| \leq 1.$$

In addition, we introduce the following notation for the p.g.f. of the process:

1. For each type $i = 0, \ldots, n$, we denote

$$F_i(t; s_0, \ldots, s_n) = \mathbb{E}(s_0^{Z^0(t)} \ldots s_n^{Z^n(t)} | Z^i(0) = 1, Z^j(0) = 0, j \neq i);$$

2. The p.g.f. of the whole process is

$$\mathbf{F}(t; \mathbf{s}) = (F_0(t; \mathbf{s}), \ldots, F_n(t; \mathbf{s})), \mathbf{s} = (s_0, \ldots, s_n).$$

3. MAIN RESULTS

3.1. BASIC FUNCTIONAL EQUATIONS

In the following theorem we will obtain the basic non-linear integral equations for the p.g.f. of the age–dependent branching process defined in Section 2.

Theorem 1. *The probability generating function* $\mathbf{F}(t; s_0, \ldots, s_n)$ *satisfies the following non-linear integral equations*

1. *For type* 0*:*

$$\begin{aligned} F_0(t; s_0, s_1, \ldots, s_n) &= F_0(t; s_0) \\ &= s_0(1 - G_0(t)) + \int_0^t f_0(F_0(t-y; s_0)) dG_0(y). \end{aligned} \tag{3.1}$$

2. *For type* i, $1 \leq i \leq n$*:*

$$\begin{aligned} F_i(t; s_0, s_1, \ldots, s_n) &= s_i(1 - G_i(t)) \\ &+ \int_0^t f_i\Big(u_{i0} F_0(t-y; s_0) + \sum_{k=1}^{n} u_{ik} F_k(t-y; s_0, s_1, \ldots, s_n)\Big) dG_i(y). \end{aligned} \tag{3.2}$$

Proof. 1). Let us consider the case when the process starts with one cell of type 0. The independence assumption of the cells' evolution allows us to consider our process as consisting of k separate processes, after first splitting of the initial cell, which gives us the following relation:

$$F_0(t;s_0,s_1,\dots,s_n) = E(E(s_0^{Z^0(t)}s_1^{Z^1(t)}\dots s_n^{Z^n(t)}|Z^0(0)=1,Z^j(0)=0,j\neq 0,(\tau_0,\nu_0)))$$

$$= s_0(1-G_0(t)) + \int_0^t \sum_{k=0}^{\infty} p_{0k} E(s_0^{Z^0(t-y)}s_1^{Z^1(t-y)}\dots s_n^{Z^n(t-y)}|Z^0(0)=k,Z^j(0)=0,j\neq 0)dG_0(y)$$

$$= s_0(1-G_0(t)) + \int_0^t \sum_{k=0}^{\infty} p_{0k}(E(s_0^{Z^0(t-y)}|Z^0(0)=1,Z^j(0)=0,j\neq 0))^k dG_0(y)$$

$$= s_0(1-G_0(t)) + \int_0^t \sum_{k=0}^{\infty} p_{0k} F_0(t-y;s_0)^k dG_0(y)$$

$$= s_0(1-G_0(t)) + \int_0^t f_0(F_0(t-y;s_0))dG_0(y).$$

Notice that this equation is the integral equation obtained for the classical BHBP.

2). Consider the case where the process starts with one cell of type i, $1\leq i\leq n$. Again, using the independence assumption, a decomposition of the sample space Ω in accordance with the life–length τ_i and number ν_i of offspring of the initial cell of type i and multinomial distribution yields the relation:

$$F_i(t;s_0,s_1,\dots,s_n) = E(E(s_0^{Z^0(t)}s_1^{Z^1(t)}\dots s_n^{Z^n(t)}|Z^i(0)=1,Z^j(0)=0,j\neq i,(\tau_i,\nu_i)))$$
$$= s_i(1-G_i(t))$$
$$+ \int_0^t \sum_{k=0}^{\infty} p_{ik} \sum_{\sum_0^n k_\ell = k} \frac{u_{i0}^{k_0}u_{i1}^{k_1}\dots u_{in}^{k_n}}{k_1!k_2!\dots k_n!} k! E(s_0^{Z^0(t-y)}s_1^{Z^1(t-y)}\dots s_n^{Z^n(t-y)}|Z^j(0)=k_j,\forall j)dG_i(y)$$

$$= s_i(1-G_i(t)) + \int_0^t \sum_{k=0}^{\infty} p_{ik} \sum_{\sum_0^n k_\ell = k} \Bigg[\frac{u_{i0}^{k_0}u_{i1}^{k_1}\dots u_{in}^{k_n}}{k_1!k_2!\dots k_n!} k! E(s_0^{Z^0(t-y)}|Z^0(0)=1,Z^j(0)=0,j\neq 0)^{k_0}$$
$$\times \prod_{m=1}^{n} E(s_0^{Z^0(t-y)}s_1^{Z^1(t-y)}\dots s_n^{Z^n(t-y)}|Z^m(0)=1,Z^j(0)=0,j\neq m)^{k_m}\Bigg] dG_i(y)$$

$$= s_i(1-G_i(t)) + \int_0^t \sum_{k=0}^{\infty} p_{ik} \sum_{\sum_0^n k_\ell = k} \Bigg[\binom{k}{k_0,k_1,\dots,k_n} u_{i0}F_0(t-y;s_0)]^{k_0}$$
$$\times \prod_{m=1}^{n} [u_{im}F_m(t-y;s_0,s_1,\dots,s_n)]^{k_m}\Bigg] dG_i(y)$$

$$= s_i(1-G_i(t)) + \int_0^t \sum_{k=0}^{\infty} p_{ik}\Big[u_{i0}F_0(t-y;s_0) + \sum_{\nu=1}^{n} u_{i\nu}F_\nu(t-y;s_0,s_1,\dots,s_n)\Big]^k dG_i(y)$$

$$= s_i(1-G_i(t)) + \int_0^t f_i\Big(u_{i0}F_0(t-y;s_0) + \sum_{\nu=1}^{n} u_{i\nu}F_\nu(t-y;s_0,s_1,\ldots,s_n)\Big)dG_i(y).$$

□

3.2. NUMBER OF MUTANTS

Definition 3. In the context of the model under discussion a "mutant" cell is each cell of type 0, whose mother cell is of type i, $1{\leq}i{\leq}n$.

It is worth noticing that, at any moment of time, the random variable (r.v.) "number of cells of type 0" is rather different from the r.v. "number of mutants".

Let us denote by $I_i(t)$, $1{\leq}i{\leq}n$, the r.v. "number of mutants that occurred in the process until time t, for a process starting with one cell of type i". We denote the p.g.f. of $I_i(t), 1{\leq}i{\leq}n$ as:

$$h_{I_i(t)}(s) = E(s^{I_i(t)}), \quad |s|{\leq}1. \tag{3.3}$$

Let $I_i, 1{\leq}i{\leq}n$ be the r.v. "total number of mutant cells, that occurred in a process with one initial cell of type i, for the duration of the whole process".

The p.g.f. of $I_i, 1{\leq}i{\leq}n$ is denoted by:

$$h_{I_i}(s) = E(s^{I_i}), |s|{\leq}1. \tag{3.4}$$

Theorem 2. *The probability generating functions $h_{I_i(t)}(s)$ of $I_i(t)$ and $h_{I_i}(s)$ of I_i satisfy the integral equations:*

$$h_{I_i(t)}(s) = 1{-}G_i(t){+}\int_0^t f_i(u_{i0}s{+}u_{i1}h_{I_1(t-y)}(s){+}\cdots{+}u_{in}h_{I_n(t-y)}(s))dG_i(y), \tag{3.5}$$

$$h_{I_i}(s) = f_i(u_{i0}s + u_{i1}h_{I_1}(s) + \cdots + u_{in}h_{I_n}(s)). \tag{3.6}$$

Proof. 1). Let us consider $h_{I_i(t)}(s)$. We have

$$h_{I_i(t)}(s) = E(s^{I_i(t)}) = E(E(s^{I_i(t)}|(\tau_i,\nu_i)))$$

$$= 1 - G_i(t) + \int_0^t \sum_{k=0}^{\infty} p_{ik} \sum_{\sum_0^n k_\ell = k} \Bigg[\binom{k}{k_0,k_1,\ldots,k_n} u_{i0}^{k_0}u_{i1}^{k_1}\ldots u_{in}^{k_n}\, s^{k_0} \times E(s^{I_i(t-y)}|Z^j(0) = k_j, j{\neq}0)\Bigg] dG_i(y)$$

$$= 1 - G_i(t) + \int_0^t \sum_{k=0}^{\infty} p_{ik} \sum_{\sum_0^n k_\ell = k} \Bigg[\binom{k}{k_0,k_1,\ldots,k_n} u_{i0}^{k_0}u_{i1}^{k_1}\ldots u_{in}^{k_n}\, s^{k_0} \prod_{m=1}^{n} E(s^{I_m(t-y)})^{k_m}\Bigg] dG_i(y)$$

$$= 1 - G_i(t)$$
$$+ \int_0^t \sum_{k=0}^{\infty} p_{ik} \sum_{\sum_0^n k_\ell = k} \left[\binom{k}{k_0, k_1, \ldots, k_n} u_{i0}^{k_0} u_{i1}^{k_1} \ldots u_{in}^{k_n} \, s^{k_0} \prod_{m=1}^{n} (h_{I_m(t-y)}(s))^{k_m} \right] dG_i(y)$$
$$= 1 - G_i(t) + \int_0^t \sum_{k=0}^{\infty} p_{ik} (u_{i0}s + u_{i1} h_{I_1(t-y)}(s) + \cdots + u_{in} h_{I_n(t-y)}(s))^k dG_i(y)$$
$$= 1 - G_i(t) + \int_0^t f_i(u_{i0}s + u_{i1} h_{I_1(t-y)}(s) + \cdots + u_{in} h_{I_n(t-y)}(s)) dG_i(y).$$

2). In a similar manner, for $h_{I_i}(s)$ we obtain:

$$h_{I_i}(s) = E(s^{I_i}) = E(E(s^{I_i} | (\tau_i, \nu_i)))$$
$$= \sum_{k=0}^{\infty} p_{ik} \sum_{\sum_0^n k_\ell = k} \left[\binom{k}{k_0, k_1, \ldots, k_n} u_{i0}^{k_0} u_{i1}^{k_1} \ldots u_{in}^{k_n} \, s^{k_0} E(s^{I_i} | Z^j(0) = k_j, j \neq 0) \right]$$
$$= \sum_{k=0}^{\infty} p_{ik} \sum_{\sum_0^n k_\ell = k} \left[\binom{k}{k_0, k_1, \ldots, k_n} u_{i0}^{k_0} u_{i1}^{k_1} \ldots u_{in}^{k_n} \, s^{k_0} \prod_{m=1}^{n} E(s^{I_m})^{k_m} \right]$$
$$= \sum_{k=0}^{\infty} p_{ik} \sum_{\sum_0^n k_\ell = k} \left[\binom{k}{k_0, k_1, \ldots, k_n} u_{i0}^{k_0} u_{i1}^{k_1} \ldots u_{in}^{k_n} \, s^{k_0} \prod_{m=1}^{n} h_{I_m}(s))^{k_m} \right]$$
$$= \sum_{k=0}^{\infty} p_{ik} (u_{i0}s + u_{i1} h_{I_1}(s) + \cdots + u_{in} h_{I_n}(s))^k$$
$$= f_i(u_{i0}s + u_{i1} h_{I_1}(s) + \cdots + u_{in} h_{I_n}(s)).$$

□

ACKNOWLEDGEMENT. This research is partially supported by the National Fund for Scientific Research at the Ministry of Education and Science of Bulgaria, grant no. DFNI-I02/17, and by the Research Fund of the Sofia University (grant 80-10-146/21.04.2017).

4. REFERENCES

[1] Athreya, K. B., Ney, P.: *Branching Processes*, Springer, Berlin, 1972.

[2] Durrett, R.: *Branching Process Models of Cancer*, Springer, 2015.

[3] Harris, T.: *The Theory of Branching Processes*, Springer, 1963.

[4] Jagers, P.: *Branching Processes with Biological Applications*, First edition, John Wiley & Sons Ltd, 1975.

[5] Kimmel, M., Axelrod, D.: *Branching Processes in Biology*, Springer, 2002.

[6] Mode, C.: *Stochastic Processes In Demography and Their Computer Implementation*, Springer, 1985.

[7] Serra, M.: On waiting time to escape. *J. Appl. Prob.*, **43**, 2006, 296–302.

[8] Serra, M., Haccou, P.: Dynamics of escape mutants. *Theor. Popul. Biol.*, **72**, 2007, 167–178.

[9] Slavtchova-Bojkova, M.: (2016). On two-type decomposable branching processes in continuous time and time to escape extinction. In: *Branching Processes and their Applications* (del Puerto, I.M., et al, eds.), Lecture Notes in Statistics-Proceedings **219**, pp. 319–329.

[10] Slavtchova-Bojkova, M., Trayanov, P., Dimitrov, S.: Branching processes in continuous time as models of mutations: Computational approaches and algorithms. *Comp. Stat. Data Anal.*, 2017, `http://dx.doi.org/10.1016/j.csda.2016.12.013l`

Received on October 15, 2017

Kaloyan Vitanov, Maroussia Slavtchova-Bojkova

Faculty of Mathematics and Informatics
"St. Kl. Ohridski" University of Sofia
5, J. Bourchier blvd., BG-1164 Sofia
BULGARIA

E-mails: kalovitanov@gmail.com
bojkova@fmi.uni-sofia.bg

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 104

ANNUAL OF SOFIA UNIVERSITY „ST. KLIMENT OHRIDSKI“
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 104

SPECTRAL CLUSTERING OF MULTIDIMENSIONAL GENETIC DATA

TSVETELIN ZAEVSKI, OGNYAN KOUNCHEV, DEAN PALEJEV, EUGENIA STOIMENOVA

The main purpose of the present paper is to initiate the application of methods from Spectral graph theory to the analysis of multidimensional genetic data, and in particular to the problem of detecting differential expression based on RNA-Seq data. Here we introduce a new algorithm, that is based on the method of Spectral Clustering and integrates an additional information about a priori given relations among the genes.

Keywords: spectral clustering, multidimensional data analysis, RNA-Seq data analysis.

2010 Math. Subject Classification: 62-07, 62H30, 91C20, 92D20.

1. INTRODUCTION

In recent years large amounts of DNA-Seq and RNA-Seq data were produced as a consequence of the advancements of the high-throughput sequencing technologies. One of the most interesting questions that can be answered by analyzing RNA-Seq data is finding differentially expressed genes or transcripts, for which the overall levels in one group of subjects (e.g. patients with a particular disease) is significantly different than the overall levels in another group (e.g. healthy controls). Ever since the first RNA-Seq datasets became available, researchers started developing different methods for analyzing it in order to find differentially expressed

genes and nowadays there are dozens such methods. Further in this article we will show that even for the same dataset, some of these methods produce very different results than the others.

Such differences of the results naturally raise the questions whether some methods are better than others, and more generally how to compare and evaluate such methods.

Comparing the results of differentially expressed genes is difficult because typically researchers are only able to biologically validate some portion of the genes being determined as differentially expressed by the method. In addition, very few or even none of the ones not being determined as differentially expressed are validated as such. There are issues even in cases in which generated or in-silico data is used and therefore we know the true differentially expressed genes, e.g. Soneson and Delorenzi (2013). In these cases the data generation assumes certain distributions, e.g. NB or Poisson, or includes artificially added outliers, which gives and advantage to differential expression methods that assume the respective distributions.

Here we discuss some general methods, in particular Spectral Clustering, for calibrating binary classifications that can also be used to compare and evaluate such classifications. Starting with an initial guess for the clusters (a split of number of points into two groups, i.e. initial classification) and an a priori information about the correlations, the method "moves" some of them between the clusters in order to improve the classification, in a sense that the resulting (calibrated) classification is closer to the "true" classification.

The research is structured in the following way: in section 2 we give a brief review of the spectral clustering algorithm, in section 3 we explain the used methodology and the corresponding results, and finally in section 4 in a short Appendix, we provide a curious relation between the method of Spectral Clustering and kernel PCA.

More details of the present research and the experimental results will be provided in subsequent publication.

2. INTRODUCTION TO THE SPECTRAL CLUSTERING

In the present introduction we provide a short description of the method of Spectral Clustering (SC) and provide some useful references related to recent developments and applications of the method.

The search for clusters is called traditionally clustering, but more recently synonyms were introduced as community detection or modularity maximization, cf. Clauset et al. (2004), Newman (2006), Newman (2008), and Fortunato and Barthelemy (2007). It is one of the main problems in Data Analysis, when studying data which are identified as points not only in a Euclidean space but also in an abstract graph where the weights of the edges may be used to generate a similarity

matrix. The role of the similarity matrix is to reflect the neighborhood relations between data points.

Unlike the usual methods for data clustering and graph partitions, as e.g. $k-$means, the method of Spectral Clustering is based on a completely different view on partition of graphs. In principle, SC may be applied to graphs where one has a naturally defined similarity matrix; in particular, if the data may be embedded into an Euclidean space, then we may use various approaches to defining a similarity matrix. Hence, we may apply the SC to very abstract situations.

Whereas the standard approach to clustering, as the method of $k-$means emphasizes upon the "compactness" of the data points, the SC makes the point on the "connectivity" or the "modularity" of the data points. The method of SC may be considered as a method for partitioning of graphs. Assume that the vertices of an undirected graph are enumerated as $x_j \in V$ (the set of vertices) and the *similarity* between them is defined by a weight matrix

$$W = (w_{ij})_{i,j}$$

with coefficients

$$w_{ij} := \omega(x_i, x_j) \geq 0$$

where the function ω regulates the size of the neigbourhoods. Then the set of edges is defined as those couples $E_{ij} := (x_i, x_j) \in E$ for which $w_{ij} > 0$. The main idea of the graph partitioning is to subdivide it into groups of vertices, so that edges E_{ij} for which x_i and x_j belong to the same group have large weights w_{ij}, while edges E_{ij} with x_i and x_j in different groups have small weight w_{ij}.

The simplest example would be if we consider a graph consisting of points $x_j \in \mathbb{R}^n$. One may take a weight function of the form

$$w_{ij} := g(x_i - x_j)$$

in particular, the Gaussian one

$$w_{ij} := \exp\left(\frac{\|x_i - x_j\|^2}{2\sigma^2}\right).$$

A standard method for clustering is the Min-Cut: For every two sets A and B we define the "strength of interaction" as

$$\operatorname{cut}(A, B) := \sum_{i \in A, j \in B} w_{ij}$$

Now the intuitive idea of the method of Min-Cut is based on minimizing the weight of edges connecting vertices in A to vertices in B. This very intuitive algorithm takes $O(|V||E|)$ time for the calculations, where $|V|$ denotes the set of elements in

the set V. However, it is not a very successful algorithm as it often isolates vertices. It is essentially improved by the method of Normalized-Cut defined as

$$\mathrm{Ncut}\,(A,B) := \mathrm{cut}\,(A,B)\left(\frac{1}{\|A\|}+\frac{1}{\|B\|}\right)$$

where for every subset A in the graph we have put

$$\|A\| := \sum_{i\in A} d_i$$

and d_i is the degree of the vertex i. The method of Normalized Cut is based on the minimization of $\mathrm{Ncut}\,(A,B)$, i.e. the weights of edges connecting vertices in A to vertices in B, while keeping the sizes of A and B very similar. However it is NP-hard to solve.

An interesting approach to understanding the idea of the SC method is by first introducing the Normalized Cut. A main observation is that if we are given two sets A and B and define now the vector $f = (f_j)_j$ by putting

$$f_j := \begin{cases} \frac{1}{\|A\|} & \text{for } j \in A \\ \frac{-1}{\|B\|} & \text{for } j \in B \end{cases}$$

then we have

$$f^T L f = \sum_{i,j} w_{ij}\,(f_i - f_j)^2 = \sum_{i,j} w_{ij}\left(\frac{1}{\|A\|}+\frac{1}{\|B\|}\right)^2$$

and

$$f^T D f = \sum_{i,j} d_i f_i^2 = \frac{1}{\|A\|}+\frac{1}{\|B\|}$$

Here we see that the important notions appear in a natural way: the diagonal matrix D has its diagonal given by the vector $(d_j)_j$ and L is the *unnormalized Laplacian* matrix defined by

$$L := D - W.$$

We see easily that

$$\mathrm{Ncut}\,(A,B) = \frac{f^T L f}{f^T D f}$$

hence

$$\min_{A,B} \mathrm{Ncut}\,(A,B) = \min_{A,B} \frac{f^T L f}{f^T D f}$$

where the minimum is taken over the sets A, B. Obviously, we may apply a relaxation by considering only those f for which $f^T D 1 = 0$. Hence, we obtain the solutions to the above problems as a solution to the generalized eigenvalue problem

$$Lf = \lambda D f.$$

For details, we refer to Chung (1997) and von Luxburg (2007).

The spectral properties of the Laplacian are closely related to the topological properties of the graph, as the following classical result shows.

Proposition 1. *The matrix L is symmetric and positive semi-definite; as such it has n non-negative, real eigenvalues, and the smallest one satisfies $\lambda_1 = 0$; the corresponding eigenvector has all elements equal to 1. If G is an undirected graph with nonnegative weights $w_{ij} \geq 0$, then the multiplicity k of the eigenvalue λ_1 is equal to the number of connected components of G.*

We see that the eigenvalue λ_1 gives the basic information about the clustering of the graph into disconnected components, hence, it is very natural to ask for a deeper knowledge of the cluster structure by inspecting the next eigenvalues. Thus, these thought follow naturally the historical steps undertaken in 1973 in the paper of Donath and Hoffman (1973) and the paper of Fiedler (1973), who considered the second eigenvalue.

2.1. SPECTRAL CLUSTERING ALGORITHM

The general scheme of the SC algorithm is given by the following steps (see e.g. von Luxburg (2007)):

1. Let $W \in R^{n\times n}$ be the similarity matrix with elements w_{ij}. Let also put $d_i = \sum_{j=1,...,n} w_{ij}$ and define D as the diagonal matrix having diagonal elements d_i. Let us assume that the number of clusters is k.

   Compute the Laplacian matrix by putting $L = D - W$.

2. Compute the first k eigenvectors $u_1, ..., u_k$ of L.

   Let $U \in R^{n\times k}$ be the matrix constructed from the vectors $u_1, ..., u_k$ as columns.

3. Let $y_i \in R^k$, for $i = 1, ..., n$, be the corresponding i-th row of U.

   Cluster the vectors $y_i, i = 1, ..., n$, in k clusters, $C_1, C_2, ..., C_k$, using the k-means algorithm.

4. The clusters, $A_1, A_2, ..., A_k$, of the initial data are recomputed by $A_i = \{j, y_j \in C_i\}$.

The success of the SC method is usually illustrated by a relatively simple toy example, with data points located on the real axis, cf. von Luxburg (2007), p. 399. This toy data set consists of a random sample of 200 points $x_1, ..., x_{200} \in \mathbb{R}$ drawn according to a mixture of four Gaussians. Since the main applications which we intend are in the area of genetic analysis using gene expressions which are at

least two-dimensional, we will be more interested in demonstrating the power of the SC method for simulated two-dimensional data. In Figure 1 below we have examples of two-dimensional graphs which are generated in a way very analogous to the one-dimensional. In Figure 1a we have generated a random sample of 20000 points in the plane drawn according to a mixture of four two-dimensional Gaussians which are located on four ellipses (5000 points on each one). The semiaxes of the ellipses are $(1,1)$, $(2,3)$, $(5,4)$, and $(30,5)$, respectively. In Figure 1b these are seven mixtures with total of 35000 points located in seven ellipses (again 5000 points on each one). In addition to the ellipses in Figure 1a we generate three other with semiaxes $(2,3)$, $(7,2)$, and $(6,1)$, respectively. After the deterministic generation of each point, we move it on random distance in every axes (normally distributed with parameters $(0, 0.1)$).

In fact, in Figure 1 one sees the result of the application of SC - it provides a perfect clustering.

3. ENHANCEMENT OF INITIAL CLUSTERING BY INCORPORATING A PRIORI INFORMATION

The purpose of this section is to introduce our methodology for enhancing an already available clustering, which is based on the appropriate usage of additional information in the form of a priori given correlations between the elements. We present the performance of our method on simulated data.

3.1. ALGORITHM

Suppose we are given data which is already clustered using some method. For simplicity we shall use two clusters, the sets I and NI with significantly different sizes – the smaller one I will be considered to be containing the significant elements (important genes), and for this reason will be called "important set", while NI will be bigger and will contain the not important elements. If we have in addition some information for the relations between the elements of the graph, especially in the form of correlations, we will use it to improve the initial clustering. Our aim is (1) to incorporate in a proper way the a priori correlation information by means of defining an appropriate similarity matrix (2) to keep as many elements as possible in the set I and (3) to move to the set I those elements in NI which have a high value of similarity w.r.t. any element in I.

In Statistical data analysis, the correlation matrix is an important statistical technique which measures the relation between two variables. For our methodology we develop a model for which we need to know (1) which elements are important, i.e. the set I, and (2) a "good enough" correlation matrix, which will be used for creating a similarity matrix. The proposed SC algorithm will enhance every initially given clustering defined by initial sets I and NI in two respects: First, it

will add some new elements to the set I, which have a high correlation with the elements in I. Second, it will remove from the set I some elements which were thought initially to be important, because of their large correlation with the set NI.

For simplicity sake we will explain our methodology on an example which appears in the analysis of expression levels of genes for RNA-Seq data. The most important problem when analyzing RNA-Seq data, is to find differentially expressed genes or transcripts, for which the overall levels in one group of subjects (e.g. patients with a particular disease) is significantly different than the overall levels in another group (e.g. healthy controls). On the other hand, and important ingredient of this difficult problem is a matrix with historically available correlations between the genes which is however not positive-definite. It is important to find an appropriate way to incorporate this a priori information in the algorithm. In the present example, we assume that the number of expressed genes is 8824.[1]

Let the number of all subjects studied be n and n_1 be the disease patients, while $n_2 = n - n_1$ be the number of the healthy controls. Thus the graph G we have to study is the subset of all 8824 points in the euclidean space $\mathbb{R}^n$. To simplify this setting, we calculate the average of the expression levels for each of the 8824 genes for the n_1 subjects, and on the other hand, calculate the average of the expression levels for each of the 8824 genes for the n_2 subjects. Thus we obtain 8824 points in the real plane $\mathbb{R}^2$ which reduces the problem to a clustering problem in the plane. Although this situation seems to be too simplified, it remains very non-trivial. It still makes deep sense to identify which are the important genes since the intuitive expectation is that the averaged gene expression levels for the disease patients would be in principle different from the averaged gene expression levels for the healthy controls. Such identification of the important genes in the plane would be very helpful to solve the genuine clustering problem in $\mathbb{R}^n$.

Our algorithm runs as follows:

1. *Generation of simulated clustering*

   First, we generate a simulated clustering given by a partition of the graph G given by $G = I \cup NI$. We generate the set I_{500} by selecting randomly 500 (respectively, the set I_{1000} with 1000) points in the plane $\mathbb{R}^2$ – normally distributed with expectation one and standard deviation 0.1. This will be defined as the important set I, and it is visualized in the top right corner in Figure 2. We generate in a similar way the set of not important elements NI (NI_{500} with 8324 points, and respectively NI_{1000} with 7824 points) – the center of their normal distribution is -1. This set is placed in the bottom left corner in Figure 2.

[1]Here the number 8824 is not accidentally chosen, but is the number of genes with average expressions at least 8 in the widely-used dataset by Bottomly et al. (2011).

2. *Correlation matrix*

   We generate a correlation matrix which would mimic the historically available correlations between the genes. We generate a correlation matrix C by using an algorithm described in Numpacharoen and Atsawarungruangkit (2012), modified by an introduction of a beta distribution. We provide two experimental settings by generating two correlation matrices, C_1 and C_2:

   (a) The matrix C_1 is generated by using a beta distribution $Beta(2, 5)$ with parameters 2 and 5

   (b) The matrix C_2 is generated in a similar way by the beta distribution $Beta(2, 2)$.

   The main difference between them is that C_1 has a relatively low large values.

3. *Similarity matrix*

   (a) Let us note that there are some elements in the cluster NI, which have a very low correlations with the others (less than 0.03). The spectral clustering algorithm can not decide correctly if such element is important or not. For this reason, we state that such elements are closer to the elements in NI, by assigning higher correlation levels.

   (b) We will modify the correlation matrix C by introducing the so-called level of significance l. The meaning of this parameter is to increase the role of the correlations which are higher than l. Here we use a power function of the form

$$f(x) := \begin{cases} x^p & \text{for } x < l \\ x^{1/p} & \text{for } x \geq l \end{cases}$$

   for an appropriate integer number p. One may use also different functions f which have similar "amplification behavior". We replace the matrix C with elements $c_{i,j}$ by the matrix C' with elements $c'_{i,j} = f(c_{i,j})$. We carry out experiments with different significance levels l. Since the maximal correlation of the first correlation matrix C_1 is 0.8807, it makes sense to make experiments with three different values $l = \{0.6, 0.7, 0.8\}$. For the same reason, for the second matrix C_2 we make experiments with five values $l = \{0.6, 0.7, 0.8, 0.9, 0.98\}$.

   (c) The core of our algorithm is the definition of a proper similarity matrix which takes into account the correlation matrix C. We define the similarity matrix W by putting:

      i.

$$w_{i,j} := \exp\left[-\frac{d(x_i, x_j)^2}{2\sigma^2}\right]$$

for the elements x_i, x_j in I; this is the Gaussian similarity coefficient which preserves the geometrical closeness of the elements, as here $d(x_i, x_j)$ denotes the euclidean distance.

ii. for taking into account the a priori given correlations C' we put

$$w_{i,j} := c'_{i,j}$$

for the rest of the pairs (x_i, x_j).

3.2. RESULTS

The results for the model with clustering sets I_{500} and NI_{500} are presented in Figure 2 and Tables 1, 2. The Figure representing the model with clustering sets I_{1000} and NI_{1000} looks similar, and we do not provide it here. The set of important elements after clustering are colored in red. The new important elements are the red points in the bottom left corner. As we can expect, their number varies for different levels of significance l – these elements are more for smaller values. This can be easily viewed in Figure 2. The initially accepted for important elements in I, which after clustering are changed to not important, can not be seen clearly in Figure 2 since they are only few, however one can observe their number in the fourth column of Tables 1 and 2. These tables contain the following values:

1. The first column contains the values of the parameters – in the brackets are the parameters of the beta distribution used for the generation of the correlation matrix; the other parameter is the level of significance l.

2. The second column contains the number of the expected important elements before the clustering – respectively 500 and 1000.

3. The third column contains the number of those expected important elements which are important again after the clustering.

4. The fourth column contains the number of those expected important elements which are NOT important after the clustering (column 2 - column 3).

5. The fifth column contains the number of the expected not important elements before the clustering –respectively 8324 and 7824.

6. The sixth column contains the number of those expected not important elements which have moved to the important set after the clustering.

7. The last column contains the number of those expected not important elements which are again not important after the clustering. (column 5 - column 6).

Also, it is reasonable to expect that the total number of the important elements after clustering varies for different levels of significance – they are more for lower values of l. Table 1 shows that for beta distribution $Beta(2,2)$ they vary, respectively in the following ranges:

1. between 499 and 3434, for the model with clustering given by the sets I_{500} and NI_{500} ,
2. between 1000 and 3753, for the model with clustering given by the sets I_{1000} and NI_{1000}.

The same observation is true when the correlation matrix C is generated using a beta distribution $Beta(2,5)$ – we can see in the Table 2 that the number of important elements varies in the following ranges:

1. between 515 and 844, for the model with clustering given by the sets I_{500} and NI_{500} ,
2. between 1010 and 1322, for the model with clustering given by the sets I_{1000} and NI_{1000}.

We can see immediately that the number of the important elements when we use beta distribution $Beta(2,5)$ are significantly less than the corresponding number in the model with beta distribution $Beta(2,2)$. This is true because the high levels in the $(2,2)$-correlation matrix are significantly more than those in the $(2,5)$-matrix.

We will only briefly explain the idea of our algorithm. Let us assume that we have after clustering a set of important elements I. On the other hand, let $I_1 \subset I$ be the set of those important elements, which before we perform clustering are not expected to be important. And finally, let NI be the set of not important elements after clustering. Now, the logic of our algorithm becomes clear from the following inequalities:

$$\min_i \left\{ \max_j \{|C(m_i, m_{1,j})|\} \right\} > l, \qquad m_i \in I, m_{1,j} \in I_1 \tag{1}$$

$$\max_i \left\{ \max_j \{|C(m_i, n_j)|\} \right\} < l, \qquad m_i \in I, n_j \in NI \tag{2}$$

where l is the level of significance and C is the corresponding correlation matrix introduced above. This means, that

1. For every important element, for which we initially thought that it is not important, there exists at least one important element such that the correlation between them is larger than the level of significance l.
2. For every not important element, there is no important one such that the correlation between them is larger than the level of significance l.

4. APPENDIX ON SPECTRAL CLUSTERING AND KERNEL PCA

For a reader more used to the traditional methods for dimensionality reduction in data analysis, we provide below a short comment about the relation between the method of Spectral Clustering and the so-called kernel Principal Component Analysis (PCA). This has been observed apparently for the first time by Bengio et al. (2003), where the authors show how both methods are special cases of a more general learning problem, that of learning the principal eigenfunctions of a kernel. An essential role is played by the fact that the smallest eigenvectors of graph Laplacians can also be interpreted as the largest eigenvectors of kernel matrices.

Before defining *kernel PCA*, let us remind that PCA is a basis transformation to diagonalize an estimate of the covariance matrix of the data. Given N points in d dimensions PCA essentially projects the data points onto p, directions ($p < d$) which capture the maximum variance of the data. These directions correspond to the eigenvectors of the covariance matrix of the training data points. Intuitively PCA fits an ellipsoid in d dimensions and uses the projections of the data points on the first p major axes of the ellipsoid. The "classic" PCA approach is a linear projection technique that works well if the data is linearly separable. However, in the case of linearly inseparable data, a nonlinear technique is required if the task is to reduce the dimensionality of a dataset. An here we come to the *Kernel PCA*. It is another unsupervised learning method that was proposed earlier and that is based on the simple idea of performing PCA in the feature space of a kernel by Schoelkopf, Smola and Muller in 1998. Schölkopf et al. (1997) propose the use of integral operator kernel functions, for computing principal components in high dimensional feature spaces, related to input space by some nonlinear map.

The basic idea of kernel PCA to deal with linearly inseparable data is to project it onto a (much) higher dimensional space where it becomes linearly separable. Let ϕ be this nonlinear mapping function so that a sample x can be mapped as $x \to \phi(x)$. The term "kernel" represents a function that calculates the dot product of the images of the samples x under ϕ, namely,

$$\kappa(x_i, x_j) = \phi(x_i)\phi(x_j)^T.$$

In other words, the function ϕ maps the original d-dimensional features into a larger, k-dimensional feature space by creating nonlinear combinations of the original features. Often, the mathematical definition of the Gaussian basis kernel function is written and implemented as

$$\kappa(x_i, x_j) = \exp(-\gamma||xi - xj||^2)$$

where $\gamma = 1/2\sigma^2$ is a free parameter that is to be optimized.

5. REFERENCES

Yoshua Bengio, Pascal Vincent, Jean-Franois Paiement, Olivier Delalleau, Marie Ouimet, and Nicolas Le Roux. Spectral clustering and kernel pca are learning eigenfunctions. Technical report, CIRANO, 2003.

Daniel Bottomly, Nicole A. Walter, Jessica Ezzell E. Hunter, Priscila Darakjian, Sunita Kawane, Kari J. Buck, Robert P. Searles, Michael Mooney, Shannon K. McWeeney, and Robert Hitzemann. Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays. *PLOS One*, 6(3):e17820+, March 2011. ISSN 1932-6203.

F.R.K. Chung. *Spectral Graph Theory*. Number 92 in CBMS Regional Conference Series. Conference Board of the Mathematical Sciences, 1997. ISBN 9780821889367. URL https://books.google.bg/books?id=YUc38_MCuhAC.

Aaron Clauset, M. E. J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Phys. Rev. E*, 70:066111, Dec 2004. doi: 10.1103/PhysRevE.70.066111. URL https://link.aps.org/doi/10.1103/PhysRevE.70.066111.

W. E. Donath and A. J. Hoffman. Lower bounds for the partitioning of graphs. *IBM J. Res. Dev.*, 17(5):420–425, September 1973. ISSN 0018-8646. doi: 10.1147/rd.175.0420. URL http://dx.doi.org/10.1147/rd.175.0420.

Miroslav Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23(2):298–305, 1973. URL http://eudml.org/doc/12723.

Santo Fortunato and Marc Barthelemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41, 2007.

M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006. doi: 10.1073/pnas.0601602103. URL http://www.pnas.org/content/103/23/8577.abstract.

M.E.J. Newman. Mathematics of networks. In Steven N. Durlauf and Lawrence E. Blume, editors, *The New Palgrave Dictionary of Economics*. Palgrave Macmillan, Basingstoke, 2008.

K. Numpacharoen and A. Atsawarungruangkit. Generating correlation matrices based on the boundaries of their coefficients. *PLOS ONE*, 7(11):1–7, 11 2012. doi: 10.1371/journal.pone.0048902. URL https://doi.org/10.1371/journal.pone.0048902.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. *Kernel principal component analysis*, pages 583–588. Springer Berlin Heidelberg, Berlin, Heidelberg, 1997. URL https://doi.org/10.1007/BFb0020217.

Charlotte Soneson and Mauro Delorenzi. A comparison of methods for differential expression analysis of rna-seq data. *BMC Bioinformatics*, 14(1):91, Mar 2013. ISSN 1471-2105. doi: 10.1186/1471-2105-14-91. URL https://doi.org/10.1186/1471-2105-14-91.

Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4): 395–416, Dec 2007. doi: 10.1007/s11222-007-9033-z. URL https://doi.org/10.1007/s11222-007-9033-z.

ACKNOWLEDGEMENT. The second and third of the authors were supported by Project I02/19, while the first- and the fourth-named authors were supported by project DH02-13 with Bulgarian NSF.

A. TABLES AND FIGURES

Table 1: Clustering results for data with 500 initially important elements

parameters	Expected important			Expected not important		
	Total	Imp.	Not imp.	Total	Imp.	Not imp.
(2,2) 0.6	500	500	0	8324	2934	5390
(2,2) 0.7	500	490	10	8324	1778	6546
(2,2) 0.8	500	499	1	8324	850	7474
(2,2) 0.9	500	487	13	8324	239	8085
(2,2) 0.98	500	489	11	8324	10	8314
(2,5) 0.6	500	500	0	8324	344	7980
(2,5) 0.7	500	499	1	8324	80	8244
(2,5) 0.8	500	500	0	8324	15	8309

Table 2: Clustering results for data with 1000 initially important elements

parameters	Expected important			Expected not important		
	Total	Imp.	Not imp.	Total	Imp.	Not imp.
(2,2) 0.6	1000	999	1	7824	2754	5070
(2,2) 0.7	1000	996	4	7824	1669	6155
(2,2) 0.8	1000	999	1	7824	801	7023
(2,2) 0.9	1000	990	10	7824	255	7599
(2,2) 0.98	1000	991	9	7824	9	7815
(2,5) 0.6	1000	1000	0	7824	322	7502
(2,5) 0.7	1000	996	4	7824	74	7750
(2,5) 0.8	1000	997	3	7824	13	7811

Figure 1: SC succeeds to separate all ellipses in the Gaussian mix example

(a) (b)

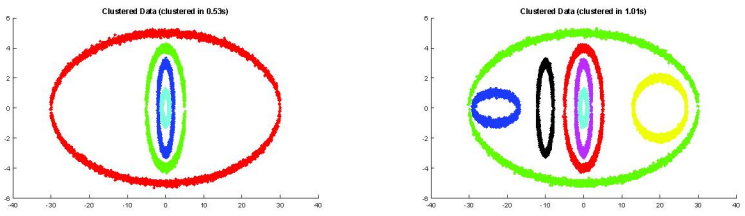
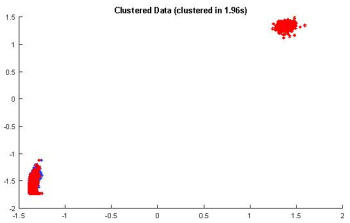
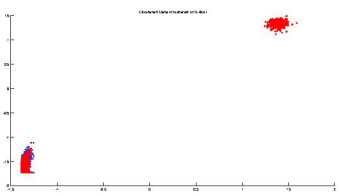


Figure 2: Clustering with 500 initial important elements

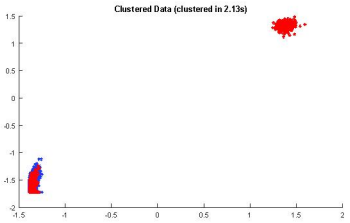
(a) Parameters: *Beta(2,2), l = 0.6*



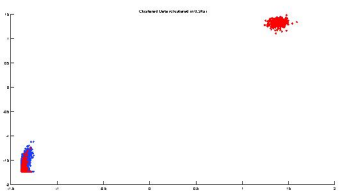
(b) Parameters: *Beta(2,2), l = 0.7*



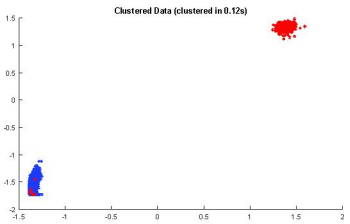
(c) Parameters: *Beta(2,2), l = 0.8*



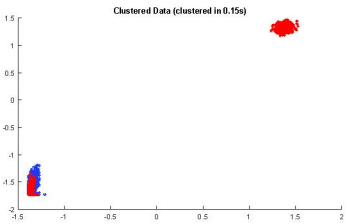
(d) Parameters: *Beta(2,2), l = 0.9*



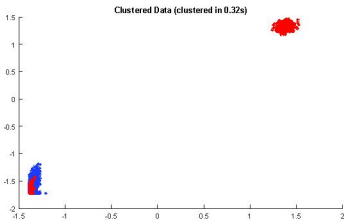
(e) Parameters: *Beta(2,2), l = 0.98*



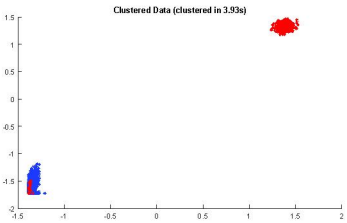
(f) Parameters: *Beta(2,5), l = 0.6*



(g) Parameters: *Beta(2,5), l = 0.7*



(h) Parameters: *Beta(2,5), l = 0.8*



Received on November 2, 2017

Tsvetelin Zaevski, Ognyan Kounchev, Dean Palejev, Evgenia Stoimenova

Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Acad. G. Bonchev st., bl. 8, BG-1113 Sofia
BULGARIA

E-mails: t_s_zaevski@math.bas.bg
okounchev@gmail.com
palejev@math.bas.bg
jenistoimenova@gmail.com

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 104

ANNUAL OF SOFIA UNIVERSITY „ST. KLIMENT OHRIDSKI“
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 104

EM ALGORITHM FOR MAXIMUM LIKELIHOOD ESTIMATION OF CORRELATED PROBIT MODEL FOR TWO LONGITUDINAL ORDINAL OUTCOMES

DENITSA GRIGOROVA

Correlated probit models (CPMs) are widely used for modeling of ordinal data or joint analyses of ordinal and continuous data which are common outcomes in medical studies. When we have clustered or longitudinal data CPMs with random effects are used to take into account the dependence between clustered measurements. When the dimension of the random effects is large, finding of the maximum likelihood estimates (MLEs) of the model parameters via standard numerical approximations is computationally cumbersome or in some cases impossible. EM algorithms for CPM for one ordinal longitudinal variable [13] and a joint CPM for one ordinal and one continuous longitudinal variable [14] are recently developed. ECM algorithm for ML estimation of the parameters of a joint CPM for two longitudinal ordinal variables will be presented. The algorithm is applied to estimation of CPM for the longitudinal ordinal outcomes self-rated health and categorized body mass index from the Health and Retirement Study (http://hrsonline.isr.umich.edu/, HRS). Results from fitting the model to the data and also results from some simulation studies will be reported.

Keywords: correlated probit model, EM algorithm, ordinal data, random effects.
2000 Math. Subject Classification: 62H12, 62J10, 62P10.

1. INTRODUCTION

Bliss [3, 4] and Gaduum [11] were the first to introduce the probit models for binary data. These models are also suitable for ordinal data as Aitchison and Silvey [1] proposed. The main characteristic of the probit models is the assumption

of a latent normally distributed variable behind the observed ordinal outcome. The density of the latent variable is divided into as many pieces as the number of the levels of the ordinal response through thresholds. The area of each density piece represents the probability of observing the corresponding level of the ordinal outcome. The usefulness of the model is not affected when the existence of the latent variable does not seem natural.

Ashford and Sowden [2] introduced a multivariate extension of the probit model based on an underlying multivariate normal distribution. Ochi and Prentice [24] first introduced a correlated probit model but only for exchangeable binary data. Extensions of this model were proposed by Hedeker and Gibbons [12], Catalano [6], Grilli and Rampichini [15], Gueorguieva and Sanacora [17] among others. Gueorguieva [16] has a detailed overview on correlated probit models. Correlated probit models are widely used for modelling of multiple categorical variables or clustered/longitudinal ordinal outcomes for these models have two main advantages. They are easy for interpretation and they allow rich correlation structure of the latent variables via random effects and/or correlated errors. That allows to take into account the natural dependence of the measurements on the same subject or within cluster.

The correlated probit model does not have closed form expression for the likelihood function. Approximations need to be used in order to obtain estimates of the unknown parameters. There are several methods of statistical inference based on numerical, stochastic or analytical approximations. Most popular appear to be extensions of numerical approximations such as Gauss-Hermite quadrature ([10] pp. 306-307) or adaptive Gaussian Quadrature [19]. Another approach is based on analytical approximations (Breslow and Clayton [5], Wolfinger and O'Connell [29]) but it has been shown to produce bias in the parameter estimates especially for binary data or ordinal data with few categories. A third approach is the Expectation-Maximization (EM) algorithm [8]. An extension of the EM algorithm is the Expectation/Conditional Maximization (ECM) algorithm [23] which is used in cases of complicated M-step.

Ruud [26] is the first to apply the EM algorithm for the estimation of the parameters of probit models. Kawakatsu and Largey [18] extend Ruud's work to a joint model of a single ordinal and multivariate normal outcomes. Chan and Kuk [7] consider a correlated model for a clustered binary variable and propose an ECM algorithm for parameter estimation.

Our algorithm is a modification of the algorithm of Chan and Kuk [7] and Grigorova and Gueorguieva [13] to estimation of a joint model for two longitudinal ordinal outcomes by using the parameter transformation proposed by Ruud [26] for estimation of the threshold parameters.

We apply the model to data on 12543 individuals from the Health and Retirement Study (http://hrsonline.isr.umich.edu/, HRS). HRS is a longitudinal survey among American citizens born between 1931 and 1941 and their spouses that assesses changes in labor force participation and health status over the transition

period from working to retirement and the years after. The launch of the study was in 1992 and data were collected at intervals of two years. The study provides a wealth of information to address important questions about aging. In our work the goal was to assess gender-related differences and the effects of smoking on measures of physical health in this representative sample of individuals over 50 years of age. We considered two repeatedly measured dependent variables: categorized body mass index (CBMI) and self-rated health (SRH). CBMI was selected because values different from normal weight might be predictive of a variety of health problems. CBMI is also easy to measure and is objective. CBMI has four levels: underweight (BMI<18.5, coded as 1), normal (18.5<BMI<25, coded as 2), overweight (25<BMI<30, coded as 3), obese (BMI>30, coded as 4). SRH is an ordinal measure that takes the following possible values: excellent (coded as 1), very good (2), good (3), fair (4) and poor (5). This is a more direct measure of health but is based on self-report and is more subjective. The two measures are expected to be positively correlated and joint modeling would allow to estimate this correlation cross-sectionally and over time and to test for overall effects of smoking and gender on these measures over time.

The paper is organized as follows. Section 2 defines the correlated probit model and outlines the estimation of the parameters and of their standard errors. Section 3 describes the simulation studies that were performed in order to examine the performance of the proposed algorithm. An application of the model to the data from the first seven waves of the HRS is included in Section 4. Section 5 contains concluding remarks and discussion about possible extensions of the algorithm.

2. MODEL

From now on bold typeset is used for vectors and matrices.

Let y^*_{1ij} is the measurement of the first ordinal variable with m_1 levels on the ith subject at time j and y^*_{2ij} is the observation on second ordinal outcome with m_2 levels on the same subject at the same time, $j = 1, \ldots, n_i, i = 1, \ldots, n$. We assume that there are two latent normal variables y_{1ij} and y_{2ij} that generated the observed ordinal variables. We consider the following random effects model:

$$\begin{aligned} y_{1ij} &= \boldsymbol{x'_{1ij}}\boldsymbol{\beta_1} + \boldsymbol{z'_{1ij}}\boldsymbol{b_{1i}} + \epsilon_{1ij}, \\ y_{2ij} &= \boldsymbol{x'_{2ij}}\boldsymbol{\beta_2} + \boldsymbol{z'_{2ij}}\boldsymbol{b_{2i}} + \epsilon_{2ij}. \end{aligned} \tag{2.1}$$

The relation between the observed ordinal variable and the latent normal variable is the following:

$$y^*_{kij} = \begin{cases} 1, \ y_{kij} \leq \alpha_{k,1}; \\ l, \ \alpha_{k,l-1} < y_{kij} \leq \alpha_{k,l}, \quad l = 2, \ldots, m_k - 1; \\ m_k, \ y_{kij} > \alpha_{k,m_k-1}; \end{cases} \tag{2.2}$$

for some unknown thresholds $\alpha_{k,1}, \ldots, \alpha_{k,m_k-1}, \quad k = 1, 2$.

We assume a normal distribution of the q-dimensional vector of the random effects $\boldsymbol{b_i} = (\boldsymbol{b'_{1i}}, \boldsymbol{b'_{2i}})' \sim N(\mathbf{0_q}, \boldsymbol{\Sigma})$. The covariance matrix $\boldsymbol{\Sigma}$ is a quadratic $q \times q$ positive semi-definite matrix:

$$\boldsymbol{\Sigma} = Var \begin{pmatrix} \boldsymbol{b_{1i}} \\ \boldsymbol{b_{2i}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma_{11}} & \boldsymbol{\Sigma_{12}} \\ \boldsymbol{\Sigma_{21}} & \boldsymbol{\Sigma_{22}} \end{pmatrix}.$$

The error terms on the same subject at the same time are not necessarily assumed independent $(\epsilon_{1ij}, \epsilon_{2ij})' \sim N(\mathbf{0_2}, \boldsymbol{\Sigma_\epsilon})$, where

$$\boldsymbol{\Sigma_\epsilon} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}.$$

The error terms among individuals and different time points are assumed to be independent. We also assume that the random effects and the error terms are independent of each other.

The regression parameters for the fixed effects in model (2.1) are denoted by the vectors $\boldsymbol{\beta_k}, k = 1, 2$. The vectors of predictors for the fixed effects are $\boldsymbol{x_{kij}}, k = 1, 2$ and the predictors for the random effects are $\boldsymbol{z_{kij}}, k = 1, 2$.

From the observed data it is not possible to uniquely estimate all of the unknown parameters, so we pose the following restrictions: the first thresholds $\alpha_{k,1}, k = 1, 2$ are set to zero, the variance of the first normal error term σ_{11} is set to 1 and the variance of the second error term given the first error term is also 1, i.e. $\sigma_{2|1} = \sigma_{22} - \sigma_{12}^2/\sigma_{11} = 1$. Some other restrictions and reparameterisations are possible.

2.1. EM ALGORITHM FOR MLE

We propose an EM algorithm [8] for estimation of the unknown model parameters in (2.1) and thresholds in (2.2).

The EM algorithm is an iterative procedure for obtaining maximum likelihood estimates for models that depend on unobserved data. In our model the unobserved data are the latent variables and the random effects. Each iteration of the EM algorithm consists of two steps: E-step (Expectation step) and M-step (Maximisation step). Let us denote with $\mathbf{X}$ the observed data, with $\mathbf{Z}$ the unobserved data and with $\boldsymbol{\Gamma}$ the vector of unknown parameters of the model. The two steps at the $(k+1)$-st iteration of the algorithm are:

- E-step: $Q(\boldsymbol{\Gamma}|\boldsymbol{\Gamma^{(k)}}) = \mathrm{E}_{\mathbf{Z}|\mathbf{X},\boldsymbol{\Gamma^{(k)}}} [\ln L(\boldsymbol{\Gamma}; \mathbf{X}, \mathbf{Z})]$, where the 'complete data' likelihood function is $L(\boldsymbol{\Gamma}; \mathbf{X}, \mathbf{Z}) = p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\Gamma})$, where $p(.)$ is a density function,
- M-step: $\boldsymbol{\Gamma^{(k+1)}} = \arg\max_{\boldsymbol{\Gamma}} Q(\boldsymbol{\Gamma}|\boldsymbol{\Gamma^{(k)}})$.

The algorithm starts with initial values for the unknown parameters $\boldsymbol{\Gamma^{(0)}}$, iterates between the E-step and the M-step and stops when a converging criterion is met. Our choice for converging criterion is when $|\boldsymbol{\Gamma^{(k+1)}} - \boldsymbol{\Gamma^{(k)}}| < \epsilon$ for each element of the vector, where ϵ is a preselected small number.

The first difficulty in applying the EM algorithm to our model is the introduction of the thresholds in the complete data likelihood. We adopt the approach by Kawakatsu and Largey [18] who extend Ruud's work [26]. According to their method, we define the differences between consecutive thresholds with $\delta_{k,i} = \alpha_{k,i} - \alpha_{k,i-1},\ i = 2, \ldots, m_k - 1, k = 1, 2$ (we define additionally $\delta_{k,1} = \delta_{k,m_k} = 1$). It follows the connection $\alpha_{k,i} = \sum_{j=2}^{i} \delta_{k,j},\ k = 1, 2,\ i = 2, \ldots, m_k - 1$. Then we consider new variables, which are a linear transformation of the latent variables. The new variables are denoted by $y_{kij_{new}} = (y_{kij} - \alpha_{k,y^*_{kij}-1})/\delta_{k,y^*_{kij}}, k = 1, 2$, where $\alpha_{k,0} = 0, k = 1, 2$ and $\boldsymbol{y_{ki_{new}}} = (y_{ki1_{new}}, y_{ki2_{new}}, \ldots, y_{kin_{i_{new}}})'$.

Since the new variables are a linear transformation of the latent variables, they are also normally distributed. But given the observed ordinal variables, the transformed variables have truncated multivariate normal distribution with boundaries of truncation independent of the unknown threshold parameters. For example, if we observe the first level of y^*_{1ij}, the new variable $y_{1ij_{new}}$ is truncated at $(-\infty, 0]$, if y^*_{1ij} is between the first and the last level, the new variable is truncated at $(0, 1]$, and if we observe the last level of y^*_{1ij}, the new variable is truncated at $(0, \infty)$.

We use the approach by Chan and Kuk [7] in order to find closed form expressions for the unknown parameters $\boldsymbol{\Gamma} = (\boldsymbol{\beta_1'}, \boldsymbol{\beta_2'}, \boldsymbol{\delta_1'}, \boldsymbol{\delta_2'}, \boldsymbol{vect(\Sigma)'}, \lambda)'$, where $\boldsymbol{\delta_k} = (\delta_{k,2}, \ldots, \delta_{k,m_k-1}),\ k = 1, 2$, $\boldsymbol{vect(\Sigma)}$ is the vector of unique elements in the covariance matrix $\boldsymbol{\Sigma}$ and $\lambda = \sigma_{12}$.

2.1.1. COMPLETE DATA LOG-LIKELIHOOD

Complete data log-likelihood is the joint density of the transformed latent variables and the random effects. It has the following form:

$$\begin{aligned} \ln L &= \ln f(\boldsymbol{b}, \boldsymbol{y_{1_{new}}}, \boldsymbol{y_{2_{new}}}) = \sum_{i=1}^{n} \ln f(\boldsymbol{b_i}) f(\boldsymbol{y_{1i_{new}}} | \boldsymbol{b_i}) f(\boldsymbol{y_{2i_{new}}} | \boldsymbol{b_i}, \boldsymbol{y_{1i_{new}}}) \\ &= \sum_{i=1}^{n} \ln[f(\boldsymbol{b_i}) \prod_{j=1}^{n_i} f(y_{1ij_{new}} | \boldsymbol{b_i}) f(y_{2ij_{new}} | \boldsymbol{b_i}, y_{1ij_{new}})], \end{aligned}$$

where $f(.)$ denotes a normal density function.

From the model definition and the assumption for the distribution of the ran-

dom effects it follows that apart from the constants the log-likelihood is:

$$\begin{aligned}\ln L = &-0.5\sum_{i=1}^{n}\ln|\boldsymbol{\Sigma}| - 0.5\sum_{i=1}^{n}\boldsymbol{b_i'}\boldsymbol{\Sigma^{-1}}\boldsymbol{b_i} - 0.5\sum_{i=1}^{n}\sum_{j=1}^{n_i}\ln\sigma_{11} + \sum_{i=1}^{n}\sum_{j=1}^{n_i}\ln\delta_{1,y^*_{1ij}}\\ &-\frac{1}{2\sigma_{11}}\sum_{i=1}^{n}\sum_{j=1}^{n_i}(\delta_{1,y^*_{1ij}}y_{1ij_{new}}-\mu_{1ij_{new}})^2 - 0.5\sum_{i=1}^{n}\sum_{j=1}^{n_i}\ln\sigma_{2|1} + \sum_{i=1}^{n}\sum_{j=1}^{n_i}\ln\delta_{2,y^*_{2ij}}\\ &-\frac{1}{2\sigma_{2|1}}\sum_{i=1}^{n}\sum_{j=1}^{n_i}[\delta_{2,y^*_{2ij}}y_{2ij_{new}} - \mu_{2ij_{new}} - \lambda(\delta_{1,y^*_{1ij}}y_{1ij_{new}} - \mu_{1ij_{new}})]^2,\end{aligned}$$

where

$$\begin{aligned}\mu_{1ij_{new}} &= \boldsymbol{x'_{1ij}}\boldsymbol{\beta_1} + \boldsymbol{z'_{1ij}}\boldsymbol{b_{1i}} - \alpha_{1,y^*_{1ij}-1},\\ \mu_{2ij_{new}} &= \boldsymbol{x'_{2ij}}\boldsymbol{\beta_2} + \boldsymbol{z'_{2ij}}\boldsymbol{b_{2i}} - \alpha_{2,y^*_{2ij}-1}.\end{aligned}$$

2.1.2. CLOSED FORM EXPRESSIONS FOR THE ESTIMATORS

We obtain closed form expressions for the estimators of the unknown parameters by setting the first derivatives of the complete data log-likelihood to zero.

The estimator for the covariance matrix $\boldsymbol{\Sigma}$ of the random effects is:

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=0}^{n}\boldsymbol{b_i}\boldsymbol{b_i'}.$$

The regression parameters for the fixed effects for the first variable $\boldsymbol{\beta_1}$ satisfy the following system of equations:

$$\begin{aligned}(1+\lambda^2)\sum_{i=1}^{n}\sum_{j=1}^{n_i}\boldsymbol{x_{1ij}}\boldsymbol{x'_{1ij}}\boldsymbol{\beta_1} =&(1+\lambda^2)\sum_{i=1}^{n}\sum_{j=1}^{n_i}(\delta_{1,y^*_{1ij}}y_{1ij_{new}} - \boldsymbol{z'_{1ij}}\boldsymbol{b_{1i}} + \alpha_{1,y^*_{1ij}-1})\boldsymbol{x_{1ij}}\\ &-\lambda\sum_{i=1}^{n}\sum_{j=1}^{n_i}[\delta_{2,y^*_{2ij}}y_{2ij_{new}} - (\boldsymbol{x'_{2ij}}\boldsymbol{\beta_2} + \boldsymbol{z'_{2ij}}\boldsymbol{b_{2i}} - \alpha_{2,y^*_{2ij}-1})]\boldsymbol{x_{1ij}}.\end{aligned}$$

The regression parameters for the fixed effects for the second variable $\boldsymbol{\beta_2}$ satisfy the following system of equations:

$$\sum_{i=1}^{n}\sum_{j=1}^{n_i}\boldsymbol{x_{2ij}}\boldsymbol{x'_{2ij}}\boldsymbol{\beta_2} = \sum_{i=1}^{n}\sum_{j=1}^{n_i}[\delta_{2,y^*_{2ij}}y_{2ij_{new}} - \boldsymbol{z'_{2ij}}\boldsymbol{b_{2i}} + \alpha_{2,y^*_{2ij}-1} - \lambda(\delta_{1,y^*_{1ij}}y_{1ij_{new}} - \mu_{1ij_{new}})]\boldsymbol{x_{2ij}}.$$

It follows that the regression parameters $\boldsymbol{\beta_2}$ are a least square solution of the regression of $\tilde{y}_{2ij}$ on $\boldsymbol{x_{2ij}}$, where

$$\tilde{y}_{2ij} = \delta_{2,y^*_{2ij}}y_{2ij_{new}} - \boldsymbol{z'_{2ij}}\boldsymbol{b_{2i}} + \alpha_{2,y^*_{2ij}-1} - \lambda(\delta_{1,y^*_{1ij}}y_{1ij_{new}} - \mu_{1ij_{new}}).$$

The equation for λ is:

$$\lambda \sum_{i=1}^{n} \sum_{j=1}^{n_i} (\delta_{1,y^*_{1ij}} y_{1ij_{new}} - \mu_{1ij_{new}})^2 = \sum_{i=1}^{n} \sum_{j=1}^{n_i} (\delta_{2,y^*_{2ij}} y_{2ij_{new}} - \mu_{2ij_{new}})(\delta_{1,y^*_{1ij}} y_{1ij_{new}} - \mu_{1ij_{new}}).$$

The equations for $\delta_{1,l},\ l = 2, \ldots, m_1 - 1$, are quadratic equations of the form: $a_1 \delta_{1,l}^2 + b_1 \delta_{1,l} + c_1 = 0,$ which always have real roots and the bigger root is always positive. The constants a_1, b_1, c_1 are given as follows:

$$\begin{aligned}
a_1 =& (1+\lambda^2) \sum_{i,j} \sum_{y^*_{1ij}=l} (y^2_{1ij_{new}}) + (1+\lambda^2)(n_{1,l+1} + n_{1,l+2} + \ldots + n_{1,m_1}),\\
b_1 =& - \sum_{i,j} \sum_{y^*_{1ij}=l} y_{1ij_{new}} [\mu_{1ij_{new}} + \lambda(\delta_{2,y^*_{2ij}} y_{2ij_{new}} - \mu_{2ij_{new}} + \lambda \mu_{1ij_{new}})]\\
&+ \sum_{i,j} \sum_{y^*_{1ij}>l} [\delta_{1,y^*_{1ij}} y_{1ij_{new}} - (\boldsymbol{x'_{1ij}\beta_1} + \boldsymbol{z'_{1ij}b_{1i}} - \alpha_{1,y^*_{1ij}-1,-l})]\\
&- \sum_{i,j} \sum_{y^*_{1ij}>l} \{\lambda[\delta_{2,y^*_{2ij}} y_{2ij_{new}} - \mu_{2ij_{new}} - \lambda(\boldsymbol{x'_{1ij}\beta_1} + \boldsymbol{z'_{1ij}b_{1i}} - \alpha_{1,y^*_{1ij}-1,-l})]\},\\
c_1 =& - n_{1,l},
\end{aligned}$$

where $n_{1,l}$ is the number of the observations of the categorical variable y_1^* at l-th level and $\alpha_{1,y^*_{1ij}-1,-l} = \delta_{1,1} + \cdots + \delta_{1,l-1} + \delta_{1,l+1} + \cdots + \delta_{1,y^*_{1ij}-1}$.

Analogously, the equations for $\delta_{2,l},\ l = 2, \ldots, m_2 - 1$, are quadratic equations of the form: $a_2 \delta_{2,l}^2 + b_2 \delta_{2,l} + c_2 = 0,$ which always have real roots and the bigger root is always positive. The constants a_2, b_2, c_2 are given as follows:

$$\begin{aligned}
a_2 =& \sum_{i,j} \sum_{y^*_{2ij}=l} (y^2_{2ij_{new}}) + n_{2,l+1} + n_{2,l+2} + \ldots + n_{2,m_2},\\
b_2 =& - \sum_{i,j} \sum_{y^*_{2ij}=l} y_{2ij_{new}} [\mu_{2ij_{new}} + \lambda(\delta_{1,y^*_{1ij}} y_{1ij_{new}} - \mu_{1ij_{new}})]\\
&+ \sum_{i,j} \sum_{y^*_{2ij}>l} [\delta_{2,y^*_{2ij}} y_{2ij_{new}} - (\boldsymbol{x'_{2ij}\beta_2} + \boldsymbol{z'_{2ij}b_{2i}} - \alpha_{2,y^*_{2ij}-1,-l})\\
&\qquad - \lambda(\delta_{1,y^*_{1ij}} y_{1ij_{new}} - \mu_{1ij_{new}})],\\
c_2 =& - n_{2,l},
\end{aligned}$$

where $n_{2,l}$ is the number of the observations of the categorical variable y_2^* at l-th level and $\alpha_{2,y^*_{2ij}-1,-l} = \delta_{2,1} + \cdots + \delta_{2,l-1} + \delta_{2,l+1} + \cdots + \delta_{2,y^*_{2ij}-1}$.

In order to update the new estimates of the parameters, we need to express the conditional expectations in the closed form expressions for the estimators. We will

show that all of the conditional expectations depend only on the first two moments of truncated multivariate normal distribution.

Let us introduce the following notation:

$$\boldsymbol{X_{ki}} = \begin{pmatrix} \boldsymbol{x'_{ki1}} \\ \boldsymbol{x'_{ki2}} \\ \vdots \\ \boldsymbol{x'_{kin_i}} \end{pmatrix}, \quad \boldsymbol{Z_{ki}} = \begin{pmatrix} \boldsymbol{z'_{ki1}} \\ \boldsymbol{z'_{ki2}} \\ \vdots \\ \boldsymbol{z'_{kin_i}} \end{pmatrix}, \quad \boldsymbol{\beta_k} = \begin{pmatrix} \beta_{k1} \\ \beta_{k2} \\ \vdots \\ \beta_{kp_k} \end{pmatrix}, \qquad k = 1, 2,$$

$$\boldsymbol{\alpha_{k,i}} = \begin{pmatrix} \alpha_{k,y^*_{ki1}-1} \\ \alpha_{k,y^*_{ki2}-1} \\ \vdots \\ \alpha_{k,y^*_{kin_i}-1} \end{pmatrix}, \qquad \boldsymbol{\delta^{-1}_{k,i}} = \begin{pmatrix} 1/\delta_{k,y^*_{ki1}} \\ 1/\delta_{k,y^*_{ki2}} \\ \vdots \\ 1/\delta_{k,y^*_{kin_i}} \end{pmatrix}, \qquad k = 1, 2,$$

$$\boldsymbol{\delta_i^{-1}} = \begin{pmatrix} \boldsymbol{\delta^{-1}_{1,i}} \\ \boldsymbol{\delta^{-1}_{2,i}} \end{pmatrix}, \quad \boldsymbol{\alpha_i} = \begin{pmatrix} \boldsymbol{\alpha_{1,i}} \\ \boldsymbol{\alpha_{2,i}} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \end{pmatrix},$$

$$\boldsymbol{y_{i_{new}}} = \begin{pmatrix} \boldsymbol{y_{1i_{new}}} \\ \boldsymbol{y_{2i_{new}}} \end{pmatrix}, \quad \boldsymbol{Z_i} = \begin{pmatrix} \boldsymbol{Z_{1i}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Z_{2i}} \end{pmatrix}, \quad \boldsymbol{X_i} = \begin{pmatrix} \boldsymbol{X_{1i}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X_{2i}} \end{pmatrix}.$$

Then the joint distribution of $\boldsymbol{y_{1i_{new}}}$, $\boldsymbol{y_{2i_{new}}}$ and $\boldsymbol{b_i}$ is multivariate normal:

$$\begin{pmatrix} \boldsymbol{y_{1i_{new}}} \\ \boldsymbol{y_{2i_{new}}} \\ \boldsymbol{b_i} \end{pmatrix} \sim N \left[\begin{pmatrix} (\boldsymbol{X_{1i}\beta_1} - \boldsymbol{\alpha_{1,i}}) \circ \boldsymbol{\delta^{-1}_{1,i}} \\ (\boldsymbol{X_{2i}\beta_2} - \boldsymbol{\alpha_{2,i}}) \circ \boldsymbol{\delta^{-1}_{2,i}} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} & \boldsymbol{C} \\ \boldsymbol{B'} & \boldsymbol{D} & \boldsymbol{E} \\ \boldsymbol{C'} & \boldsymbol{E'} & \boldsymbol{\Sigma} \end{pmatrix} \right],$$

where $\circ$ is the Hadamard (element-wise) product, the elements of the covariance matrix are:

$$\begin{aligned} \boldsymbol{A} &= (\boldsymbol{Z_{1i}\Sigma_{11}Z'_{1i}} + \sigma_{11}\boldsymbol{I}_{n_i}) \circ \boldsymbol{\delta^{-1}_{1,i}\delta^{-1'}_{1,i}}, \\ \boldsymbol{B} &= (\boldsymbol{Z_{1i}\Sigma_{12}Z'_{2i}} + \sigma_{12}\boldsymbol{I}_{n_i}) \circ \boldsymbol{\delta^{-1}_{1,i}\delta^{-1'}_{2,i}}, \\ \boldsymbol{C} &= \boldsymbol{Z_{1i}(\Sigma_{11}\Sigma_{12})} \circ \boldsymbol{J_{n_i\times q}\delta^{-1}_{1,i}}, \\ \boldsymbol{D} &= (\boldsymbol{Z_{2i}\Sigma_{22}Z'_{2i}} + \sigma_{22}\boldsymbol{I}_{n_i}) \circ \boldsymbol{\delta^{-1}_{2,i}\delta^{-1'}_{2,i}}, \\ \boldsymbol{E} &= \boldsymbol{Z_{2i}(\Sigma_{21}\Sigma_{22})} \circ \boldsymbol{J_{n_i\times q}\delta^{-1}_{2,i}}, \end{aligned}$$

$\boldsymbol{J_{n_i\times q}\delta^{-1}_{k,i}}$ is $n_i \times q$ matrix with columns $\boldsymbol{\delta^{-1}_{k,i}}$, $k = 1, 2$, and $\boldsymbol{I_{n_i}}$ is the $n_i \times n_i$ identity matrix.

Let us denote

$$\boldsymbol{M_i} = \begin{bmatrix} \boldsymbol{y_{1i_{new}}} - (\boldsymbol{X_{1i}\beta_1} - \boldsymbol{\alpha_{1,i}}) \circ \boldsymbol{\delta^{-1}_{1,i}} \\ \boldsymbol{y_{2i_{new}}} - (\boldsymbol{X_{2i}\beta_2} - \boldsymbol{\alpha_{2,i}}) \circ \boldsymbol{\delta^{-1}_{2,i}} \end{bmatrix} = \boldsymbol{y_{i_{new}}} - (\boldsymbol{X_i\beta} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta_i^{-1}}$$

and

$$\boldsymbol{\Sigma}_{B_i} = \begin{pmatrix} \boldsymbol{C}' & \boldsymbol{E}' \end{pmatrix} \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}' & \boldsymbol{D} \end{pmatrix}^{-1}.$$

Then the conditional distribution of $\boldsymbol{b_i}$ given $\boldsymbol{y_{1i_{new}}}$ and $\boldsymbol{y_{2i_{new}}}$ is again normal:

$$\boldsymbol{b_i}|\boldsymbol{y_{1i_{new}}}, \boldsymbol{y_{2i_{new}}} \sim N[\boldsymbol{\Sigma}_{B_i}\boldsymbol{M_i}, \boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{B_i}\begin{pmatrix} \boldsymbol{C} \\ \boldsymbol{E} \end{pmatrix}].$$

In the expressions for the estimators we have to calculate the following conditional expectations: $E(\boldsymbol{b_i}|\boldsymbol{y_i^*}), E(\boldsymbol{b_i b_i'}|\boldsymbol{y_i^*}), E(\boldsymbol{b_i y'_{i_{new}}}|\boldsymbol{y_i^*})$. We will show that they depend only on the first two moments of the distribution of the transformed latent variables given the observed variables, i.e. they depend on the first two moments of $\boldsymbol{y_{i_{new}}}|\boldsymbol{y_i^*}$, which distribution is truncated multivariate normal.

The expectation of the random effects given the observed variables is:

$$\begin{aligned} E(\boldsymbol{b_i}|\boldsymbol{y_i^*}) &= E[E(\boldsymbol{b_i}|\boldsymbol{y_{i_{new}}})|\boldsymbol{y_i^*}] \\ &= E[\boldsymbol{\Sigma}_{B_i}(\boldsymbol{y_{i_{new}}} - (\boldsymbol{X_i\beta} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta_i^{-1}})|\boldsymbol{y_i^*}] \\ &= \boldsymbol{\Sigma}_{B_i}[E(\boldsymbol{y_{i_{new}}}|\boldsymbol{y_i^*}) - (\boldsymbol{X_i\beta} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta_i^{-1}}]. \end{aligned}$$

The expectation of the second moment of the random effects given the observed variables is:

$$\begin{aligned} E(\boldsymbol{b_i b_i'}|\boldsymbol{y_i^*}) &= E[E(\boldsymbol{b_i b_i'}|\boldsymbol{y_{i_{new}}})|\boldsymbol{y_i^*}] \\ &= E[Var(\boldsymbol{b_i}|\boldsymbol{y_{i_{new}}}) + E(\boldsymbol{b_i}|\boldsymbol{y_{i_{new}}})E(\boldsymbol{b_i'}|\boldsymbol{y_{i_{new}}})|\boldsymbol{y_i^*}]. \end{aligned}$$

The last expectation that we need is:

$$\begin{aligned} E(\boldsymbol{b_i y'_{i_{new}}}|\boldsymbol{y_i^*}) &= E[E(\boldsymbol{b_i y'_{i_{new}}}|\boldsymbol{y_{i_{new}}})|\boldsymbol{y_i^*}] \\ &= E\{\boldsymbol{\Sigma}_{B_i}[\boldsymbol{y_{i_{new}}} - (\boldsymbol{X_i\beta} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta_i^{-1}}]\boldsymbol{y'_{i_{new}}}|\boldsymbol{y_i^*}\} \\ &= \boldsymbol{\Sigma}_{B_i}[E(\boldsymbol{y_{i_{new}} y'_{i_{new}}}|\boldsymbol{y_i^*}) - (\boldsymbol{X_i\beta} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta_i^{-1}} E(\boldsymbol{y'_{i_{new}}}|\boldsymbol{y_i^*})]. \end{aligned}$$

2.1.3. $(K+1)$-ST ITERATION OF THE EM ALGORITHM

We use an extension of the EM algorithm called Expectation/Conditional Maximization (ECM) algorithm [23]. The E-step at the $(k+1)$-st iteration of the proposed algorithm consists of finding of the following expectations: $E(\boldsymbol{b_i}|\boldsymbol{y_i^*}; \boldsymbol{\Gamma^k})$, $E(\boldsymbol{b_i b_i'}|\boldsymbol{y_i^*}; \boldsymbol{\Gamma^k})$, $E(\boldsymbol{b_i y'_{i_{new}}}|\boldsymbol{y_i^*}; \boldsymbol{\Gamma^k})$, where $\boldsymbol{\Gamma^k}$ are the k-th estimates of the unknown parameters $\boldsymbol{\Gamma}$. The M-step consists of several computationally simpler CM-steps. In each CM-step we maximise the expectation of the complete data log-likelihood function in respect to some parameters while the other parameters are held fixed. The algorithm starts with initial values for the uknown parameters $\boldsymbol{\Gamma^0}$, iterates between the E-step and M-step and stops when $|\boldsymbol{\Gamma^{k+1}} - \boldsymbol{\Gamma^k}| < \epsilon$ for each element of the vector, where ϵ is a preselected small number (for example $\epsilon = 0.0001$).

We use the bootstrap method for standard errors approximation described in [22, pp. 130–131]. The steps are as follows:

1. We fit model (2.1) to the observed data set consisting of n individuals using the proposed ECM algorithm and obtain the estimates of the unknown parameters $\hat{\mathbf{\Gamma}} = (\hat{\boldsymbol{\beta}}_{\mathbf{1}}', \hat{\boldsymbol{\beta}}_{\mathbf{2}}', \hat{\boldsymbol{\delta}}_{\mathbf{1}}', \hat{\boldsymbol{\delta}}_{\mathbf{2}}', \boldsymbol{vect}(\hat{\boldsymbol{\Sigma}})', \hat{\lambda})'$. To generate a bootstrap sample first we generate n random effects $\boldsymbol{b}_{\boldsymbol{i}}^{\boldsymbol{b}}$ from $N(\mathbf{0}, \hat{\boldsymbol{\Sigma}})$, $i = 1, \ldots, n$. Next we simulate normal values $\boldsymbol{y}_{\mathbf{1}\boldsymbol{i}}^{\boldsymbol{b}}$ and $\boldsymbol{y}_{\mathbf{2}\boldsymbol{i}}^{\boldsymbol{b}}$ of dimension n_i according to the fitted model for every random effect $\boldsymbol{b}_{\boldsymbol{i}}^{\boldsymbol{b}}$. We use the estimated via $\hat{\boldsymbol{\delta}}_{\mathbf{1}}$ and $\hat{\boldsymbol{\delta}}_{\mathbf{2}}$ thresholds to determine in which interval the normal data $\boldsymbol{y}_{\mathbf{1}\boldsymbol{i}}^{\boldsymbol{b}}$ and $\boldsymbol{y}_{\mathbf{2}\boldsymbol{i}}^{\boldsymbol{b}}$ fall and determine the levels of the bootstrap categorical variables $\boldsymbol{y}_{\mathbf{1}\boldsymbol{i}}^{\boldsymbol{b}*}$ and $\boldsymbol{y}_{\mathbf{2}\boldsymbol{i}}^{\boldsymbol{b}*}$. The bootstrap sample consists of the categorical variables $\boldsymbol{y}_{\mathbf{1}\boldsymbol{i}}^{\boldsymbol{b}*}$ and $\boldsymbol{y}_{\mathbf{2}\boldsymbol{i}}^{\boldsymbol{b}*}$, $i = 1, \ldots, n$.

2. We apply the ECM algorithm to the bootstrap data $\boldsymbol{y}_{\mathbf{1}\boldsymbol{i}}^{\boldsymbol{b}*}$ and $\boldsymbol{y}_{\mathbf{2}\boldsymbol{i}}^{\boldsymbol{b}*}$, $i = 1, \ldots, n$ to obtain estimates for the generated bootstrap data set $\mathbf{\Gamma}^{\boldsymbol{b}}$.

3. We use Monte Carlo method to approximate the bootstrap covariance matrix. That means that we repeat step 1 and step 2 B times and calculate the covariance matrix of the B estimated parameters $\mathbf{\Gamma}^{\boldsymbol{b}}$, $b = 1, \ldots, B$:

$$Cov(\hat{\mathbf{\Gamma}}) \approx \sum_{b=1}^{B} \frac{(\mathbf{\Gamma}^{\boldsymbol{b}} - \bar{\mathbf{\Gamma}})(\mathbf{\Gamma}^{\boldsymbol{b}} - \bar{\mathbf{\Gamma}})'}{B-1},$$

where $\bar{\mathbf{\Gamma}} = \sum_{b=1}^{B} \mathbf{\Gamma}^{\boldsymbol{b}}/B$.

3. SIMULATIONS

We simulated values from the following random intercept model:

$$\begin{aligned} y_{1ij} &= \beta_{10} + \beta_{11} t_{ij} + b_{1i} + \epsilon_{1ij}, \ j = 1, \ldots, 6, \\ y_{2ij} &= \beta_{20} + \beta_{21} t_{ij} + b_{2i} + \epsilon_{2ij}, \ j = 1, \ldots, 6, \end{aligned} \tag{3.1}$$

where $\beta_{10} = -0.5$, $\beta_{11} = 1$, $\beta_{20} = 1$, $\beta_{21} = -0.5$, $\alpha_{1,1} = 0$, $\alpha_{1,2} = 1.2$, $\alpha_{1,3} = 3$, $\alpha_{2,1} = 0$, $\alpha_{2,2} = 2$, $\lambda = 0.8$. The covariance matrices of errors $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ and of the random effects $\boldsymbol{\Sigma}$ are:

$$\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = \begin{pmatrix} \boldsymbol{\sigma_{11}} & \boldsymbol{\sigma_{12}} \\ \boldsymbol{\sigma_{12}} & \boldsymbol{\sigma_{22}} \end{pmatrix} = \begin{pmatrix} \mathbf{1} & \mathbf{0.8} \\ \mathbf{0.8} & \mathbf{1.64} \end{pmatrix},$$

$$\boldsymbol{\Sigma} = \boldsymbol{Var} \begin{pmatrix} \boldsymbol{b_{1i}} \\ \boldsymbol{b_{2i}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\sigma}_{\mathbf{11}}^{\boldsymbol{b}} & \boldsymbol{\sigma}_{\mathbf{12}}^{\boldsymbol{b}} \\ \boldsymbol{\sigma}_{\mathbf{21}}^{\boldsymbol{b}} & \boldsymbol{\sigma}_{\mathbf{22}}^{\boldsymbol{b}} \end{pmatrix} = \begin{pmatrix} 1 & -0.8 \\ -0.8 & 1 \end{pmatrix}.$$

Table 1: Estimates and standard errors in the simulation model 3.1

parameter	β_{10}	β_{11}	β_{20}	β_{21}	$\delta_{1,2}$	$\delta_{1,3}$	$\delta_{2,2}$	λ	σ^b_{11}	σ^b_{12}	σ^b_{22}
	Sample size **1500**										
true value	-0.5	1	1	-0.5	1.2	1.8	2	0.8	1	-0.8	1
mean est.	-0.51	1	0.99	-0.50	1.20	1.80	1.99	0.79	1.00	-0.79	0.99
stand.dev. of estim.	0.041	0.010	0.047	0.012	0.018	0.024	0.039	0.026	0.047	0.039	0.055
mean of boot.st.er.	0.038	0.010	0.046	0.012	0.020	0.023	0.040	0.026	0.049	0.042	0.058
	Sample size **3000**										
mean est.	-0.51	1.00	1.01	-0.50	1.20	1.80	2.00	0.79	1.00	-0.79	0.98
stand.dev. of estim.	0.025	0.006	0.036	0.009	0.011	0.019	0.029	0.023	0.033	0.027	0.041
mean of boot.st.er.	0.027	0.007	0.032	0.008	0.013	0.016	0.028	0.019	0.034	0.029	0.040

We simulated 100 samples with two different sample sizes ($\boldsymbol{n = 1500}$ and $\boldsymbol{n = 3000}$). For each approximation of the standard errors we used 50 bootstrap samples which is within the recommended range of 50 to 100 bootstrap replications (Efron and Tibshirani [9]). The results are presented in Table 1.

Note that due to the re-parametrization we estimate the differences in the thresholds rather than the thresholds themselves, but they coincide in the case of only three levels of the categorical variables. In both simulation studies most of the averages of the estimated parameters are equal to the parameter values from which the samples were generated and where they differ the difference is smaller than **0.02**.

As expected the standard errors get smaller when we increase the sample size. All of the estimates are statistically significantly different from zero.

The approximate equality of the standard deviations of the estimates and the means of the bootstrap standard errors confirms that the algorithm is converging as expected. However, larger simulation study that varies the parameter settings is necessary to confirm the above observations.

3.1. IMPLEMENTATION OF THE ALGORITHM

For the implementation of the algorithm we used the free software environment for statistical computing and graphics R [25]. The R code for fitting the presented models is available from the author.

We want to point out several things regarding the implementation of the proposed ECM algorithm. In the package **tmvtnorm** [28] there are functions for analytical finding of the first two moments of multivariate truncated normal distribution based on the work by Manjunath and Wilhelm [21]. There are also functions for generating random numbers using Gibbs sampling [27] which allows stochastic approximation of the first two moments of the truncated normal distribution. But for these models we recommend stochastic approximation because the analytical calculation could be very slow when we have many observations per subject.

A good choice for starting points for the regression parameters in model (2.1) and thresholds in (2.2) for the proposed ECM algorithm are estimates from model without random effects. Selecting large values as starting points for the variances of the random effects should be avoided. Problems with performance of the algorithm may occur with starting points corresponding to a multivariate truncated normal distribution for which the truncation area is close to 0. In such cases finding analytical solutions for the moments of the truncated normal distribution may fail. Generating random numbers via Gibbs sampling may also fail.

4. APPLICATION OF THE MODEL

We analyzed the first seven waves of HRS data with 12,543 individuals. We fitted the following correlated probit model to the data:

$$\begin{aligned}
y_{1ij} =& \beta_{10} + \beta_{11} t_{ij} + \beta_{12} I(smoker) + \beta_{13} I(female) \\
&+ \beta_{14} t_{ij} I(smoker) + \beta_{15} t_{ij} I(female) \\
&+ \beta_{16} I(smoker) I(female) + \beta_{17} t_{ij} I(smoker) I(female) \\
&+ b_{1i1} + b_{1i2} t_{ij} + \epsilon_{1ij}, \\
y_{2ij} =& \beta_{20} + \beta_{21} t_{ij} + \beta_{22} I(smoker) + \beta_{23} I(female) \\
&+ \beta_{24} t_{ij} I(smoker) + \beta_{25} t_{ij} I(female) \\
&+ \beta_{26} I(smoker) I(female) + \beta_{27} t_{ij} I(smoker) I(female) \\
&+ b_{2i1} + b_{2i2} t_{ij} + \epsilon_{2ij}, \\
y^*_{kij} =& \begin{cases} 1, & y_{kij} \leq \alpha_{k,1} = 0, \\ l, & \alpha_{k,l-1} < y_{kij} \leq \alpha_{k,l},\ l = 2, \ldots, m_k - 1 \\ m_k, & y_{kij} > \alpha_{k,m_k-1}, \end{cases}
\end{aligned} \tag{4.1}$$

where $k = 1, 2, m_1 = 5, m_2 = 4$ and where the covariance matrix of the errors is: $\Sigma_\epsilon = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$ and the covariance matrix of the random effects is:

$$\Sigma = Var \begin{pmatrix} b_{1i1} \\ b_{1i2} \\ b_{2i1} \\ b_{2i2} \end{pmatrix} = \begin{pmatrix} \sigma^b_{11} & \sigma^b_{12} & \sigma^b_{13} & \sigma^b_{14} \\ \sigma^b_{21} & \sigma^b_{22} & \sigma^b_{23} & \sigma^b_{24} \\ \sigma^b_{31} & \sigma^b_{32} & \sigma^b_{33} & \sigma^b_{34} \\ \sigma^b_{41} & \sigma^b_{42} & \sigma^b_{43} & \sigma^b_{44} \end{pmatrix}.$$

The estimates of the parameters, their standard errors and z-scores are presented in Table 2 and Table 3. Z-scores are computed before rounding off the estimates and their standard errors, and then rounded to the second decimal point in Table 2 and third decimal point in Table 3.

The results show that all of the parameters in the model are statistically significantly different from zero, except the regression parameters for the three-way interactions between time, smoking and gender in both sub-models (β_{17} and β_{27}),

Table 2: Table of estimates, standard errors and z-scores of the regression parameters and threshold differences in model 4.1 fitted to the first seven waves of HRS data

Regression parameters for latent self-rated health								
parameter	β_{10}	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{16}	β_{17}
estimate	1.37	0.13	0.64	0.06	0.03	-0.03	-0.26	0.01
stand. error	0.019	0.004	0.039	0.025	0.009	0.006	0.053	0.012
z-score	71.11	30.61	16.46	2.44	2.81	-4.73	-4.91	1.09
Regression parameters for latent categorized body mass index								
parameter	β_{20}	β_{21}	β_{22}	β_{23}	β_{24}	β_{25}	β_{26}	β_{27}
estimate	6.00	0.05	-0.88	-0.28	-0.01	0.05	-0.09	0.01
stand. error	0.053	0.006	0.038	0.030	0.010	0.007	0.061	0.015
z-score	113.4	8.94	-23.17	-9.36	-0.77	6.89	-1.53	0.53
Threshold parameters for both latent variables								
parameter	$\delta_{1,2}$	$\delta_{1,3}$	$\delta_{1,4}$		$\delta_{2,2}$	$\delta_{2,3}$		
estimate	1.64	1.56	1.45		4.97	3.31		
stand. error	0.010	0.010	0.012		0.145	0.074		
z-score	172.18	164.15	120.11		34.28	44.91		

Table 3: Table of estimates, standard errors and z-scores of the covariance parameters in model 4.1 fitted to the first seven waves of HRS data

parameter	σ^b_{11}	σ^b_{22}	σ^b_{33}	σ^b_{44}	σ^b_{12}	σ^b_{13}	σ^b_{14}
estimate	3.541	0.038	8.185	0.082	-0.194	1.308	-0.074
stand. error	0.065	0.001	0.311	0.005	0.006	0.071	0.007
z-score	54.872	47.948	26.294	17.507	-30.277	18.345	-10.151
parameter	σ^b_{23}	σ^b_{24}	σ^b_{34}		λ		
estimate	-0.020	0.002	-0.160		-0.002		
stand. error	0.008	0.001	0.020		0.007		
z-score	-2.619	2.638	-8.122		-0.267		

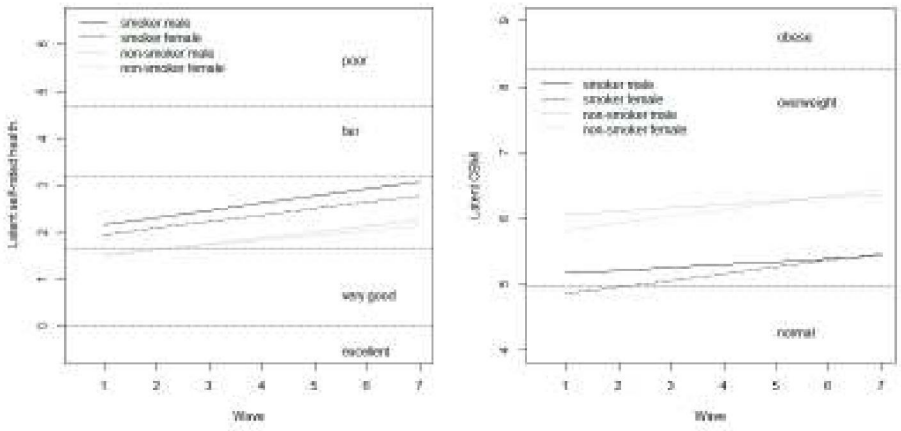
the regression parameter $\boldsymbol{\beta_{24}}$ for the two-way interaction between time and smoking and the regression parameter $\boldsymbol{\beta_{26}}$ for the two-way interaction between smoking and gender in the sub-model for the latent CBMI.

The correlation between the random intercept and random slope for the latent self-rated health is estimated to $\mathbf{-0.53}$ and for the BMI is estimated to $\mathbf{-0.20}$. The estimate for the correlation between the random intercepts is 0.24.

The trajectories of the latent variables for SRH and CBMI over time for four individuals with zero random effects are presented in Figure 1. They reflect the progress of the variables on average over time for each group of people: smoker male, smoker female, non-smoker male and non-smoker female. As expected, for the four groups of people on average the self-evaluation of health is worsening over time and people are gaining weight with time. The group with the most gentle slope for the self-rated health is the group of non-smoker female and with the steepest

slope - smoker male, which means that, according to their own opinion on their own health, smoker males are worsening most quickly, while non-smoker females are worsening most slowly on average. For the BMI, smoker males are gaining weight most slowly and non-smoker females are gaining weight most quickly on average over time.

Figure 1: Latent SRH and latent CBMI over time for four individuals with zero random effects



5. DISCUSSION

In this paper we considered a correlated probit model for a joint analysis of two longitudinal ordinal outcomes. We proposed an extension of the EM algorithm of Chan and Kuk [7] and the ECM algorithm of Grigorova and Gueorguieva [13] for obtaining maximum likelihood estimates. The algorithm is implemented in the free software environment for statistical computing and graphics R [25]. We studied its performance via simulations. We illustrated the approach on the data from the Health and Retirement Study. Our approach has advantages over alternative estimation methods in that it can handle a large dimension of the multivariate outcome, it can be easily extended to any combination of binary, ordinal and continuous outcomes and it provides asymptotically unbiased estimates. It is also easily implemented in the free open-source software environment R.

We used bootstrap method for standard error estimation which is computationally very intensive. While the bootstrap algorithm can always be applied, it is not efficient. Other approaches may be possible. For example, one might consider the Louis's approximation method [20].

Further research is needed to extend the algorithm to the estimation of a joint model for time to drop-out and combinations of ordinal and continuous longitudinal outcomes. Model selection and model diagnostics are also open areas of research.

ACKNOWLEDGEMENT. This work is partially supported by the financial funds allocated to the Sofia University "St. Kl. Ohridski", Grants No. 013/2016 and No. 80-10-146/21.04.2017.

6. REFERENCES

[1] Aitchison, J., Silvey, S. D.: The generalization of probit analysis to the case of multiple responses. *Biometrika*, **44**, no. 1-2, 1957, 131–140.

[2] Ashford, J. R., Sowden, R. R.: Multi-variate probit analysis. *Biometrics*, **26**, no. 3, 1970, 535–546.

[3] Bliss, C. I.: The method of probits. *Science*, **79**, 2037, 1934, 38–39.

[4] Bliss, C. I.: The method of probits - a correction. *Science*, **79**, 2053, (1934), 409–410.

[5] Breslow, N. E., Clayton, D. G.: Approximate inference in generalized linear mixed models. *J. Amer. Stat. Assoc.*, **88**, 1993, 9–25.

[6] Catalano, P. J.: Bivariate modelling of clustered continuous and ordered categorical outcomes. *Statistics in Medicine*, **16**, no. 8, 1997, 883–900.

[7] Chan, J. S. K., Kuk, A. Y. C.: Maximum likelihood estimation for probit-linear mixed models with correlated random effects. *Biometrics*, **53**, 1997, 86–97.

[8] Dempster, A. P., Laird, N. M., Rubin, D. B.: Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Stat. Soc.*, Series B (Methodological), **39**, no. 2, 1977, 1–22.

[9] Efron, B., Tibshirani, R. J.: *An Introduction to the Bootstrap*, First edition, Monographs on Statistics & Applied Probability, **57**, Chapman & Hall, New York, 1994.

[10] Fahrmeir, L., Tutz, G.: *Multivariate Statistical Modelling Based on Generalized Linear Models*, Second edition, Springer-Verlag, New York, 2001.

[11] Gaddum, J. H.: Methods of biological assay depending on a quantal response. *Reports on biological standards*, III, 1933.

[12] Gibbons, R. D., Hedeker, D.: Application of random-effects probit regression models. *J. Consulting and Clinical Psychology*, **62**, no. 2, 1994, 285–296.

[13] Grigorova, D., Gueorguieva, R.: Implementation of the EM algorithm for maximum likelihood estimation of a random effects model for one longitudinal ordinal outcome. *Pliska Stud. Math. Bulgar.*, **22**, 2013, 41–56.

[14] Grigorova, D., Gueorguieva, R.: Correlated probit analysis of repeatedly measured ordinal and continuous outcomes with application to the Health and Retirement Study. *Statistics in Medicine*, **35**, no. 23, 2016, 4202–4225, sim. 6982.

[15] Grilli, L., Rampichini, C.: Alternative specifications of multivariate multilevel probit ordinal response models. *J. Educational and Behavioral Statistics*, **28**, 2003, 31–44.

[16] Gueorguieva, R. V.: Correlated probit model. In: *Encyclopedia of Biopharmaceutical Statistics*, 2006, Chapt. 59, pp. 355–362.

[17] Gueorguieva, R. V., Sanacora, G.: Joint analysis of repeatedly observed continuous and ordinal measures of disease severity. *Statistics in Medicine*, **25**, 2006, 1307–1322.

[18] Kawakatsu, H., Largey, A. G.: EM algorithms for ordered probit models with endogenous regressors. *Econometrics Journal*, **12**, 2009, 164–186.

[19] Liu, Q., Pierce, D. A.: A note on Gauss-Hermite quadrature. *Biometrika*, **81**, no. 3, 1994, 624–629.

[20] Louis, T. A.: Finding the observed information matrix when using the EM algorithm. *J. Royal Stat. Soc.*, Series B, **44**, no. 2, 1982, 226–233.

[21] Manjunath, B. G.,Wilhelm, S.: Moments calculation for the double truncated multivariate normal density. `http://ssrn.com/abstract=1472153`, September 11, 2009.

[22] McLachlan, G. J., Krishnan, T.: *The EM Algorithm and Extensions*, Wiley Series in Probability and Statistics, Second edition, Wiley-Interscience, 2008.

[23] Meng, X.-L., Rubin, D. B.: Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, **80**, no. 2, 1993, 267–278.

[24] Ochi, Y., Prentice, R. L.: Likelihood inference in a correlated probit regression model. *Biometrika*, **71**, no. 3, 1984, 531–543.

[25] R Core Team: R: *A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013, ISBN 3-900051-07-0.

[26] Ruud, P. A.: Extensions of estimation methods using the EM algorithm. *Journal of Econometrics*, **49**, no. 3, 305–341.

[27] Wilhelm, S.: Gibbs sampler for the truncated multivariate normal distribution. `http://cran.r-project.org/web/packages/tmvtnorm/vignettes/GibbsSampler.pdf`, April 6, 2012.

[28] Wilhelm, S., Manjunath, B. G.: *tmvtnorm: Truncated Multivariate Normal and Student t Distribution*, 2012, R package version 1.4-7.

[29] Wolfinger, R., O'Connell, M.: Generalized linear mixed models: A pseudo-likelihood approach. *J. Stat. Comp. Sim.*, *48*, 1993, 233–243.

Received on October 17, 2017

Denitsa Grigorova

Faculty of Mathematics and Informatics
"St. Kl. Ohridski" University of Sofia
Department of Probability, Operations Research and Statistics
5, J. Bourchier blvd., BG-1164 Sofia
BULGARIA

E-mail: dgrigorova@fmi.uni-sofia.bg